CONSERVED VIRUS PROTEIN FAMILIES IN BACTERIOPHAGE GENOMES AND IN METAGENOMES OF HUMANS

By

Xixu Cai

Submitted to the graduate degree program in Microbiology, Molecular Genetics and Immunology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

_____

Chairperson, Arcady Mushegian, Ph.D.

_____

Joe Lutkenhaus, Ph.D.

_____

Philip Hardwidge, Ph.D.

Date Defended: July 1st, 2011

The Thesis Committee for Xixu Cai

certifies that this is the approved version of the following thesis:

CONSERVED VIRUS PROTEIN FAMILIES IN BACTERIOPHAGE GENOMES AND IN METAGENOMES OF HUMANS

_____

Chairperson, Arcady Mushegian, Ph.D.

Date approved: July 22, 2011

Abstract

Viruses are likely to be the most abundant genomes in the biosphere, displaying remarkable molecular diversity. Their fast-evolving genomes and lack of universal marker genes make phylogenetic and taxonomic studies more difficult than with other organisms.

A detailed determination of gene conservation between virus genomes should facilitate the study of virus evolution and function. Here we used sequence similarity methods to build a phage orthologous groups (POGs) resource. The number of POGs has grown significantly in the past decade, while the percentage of genes in phage genomes that have orthologs in other phages has also been increasing, and the percentage of unknown "ORFans" - phage genes that are not in POGs - is decreasing. Other properties of phage genomes remain stable, in particular the high fraction of genes that are never or only rarely observed in their cellular hosts. This suggests that despite the role of phages in transferring cellular genes, a large fraction of the genes in phage genomes maintain an evolutionary trajectory that is distinct from that of host genes.

Next generation sequencing technologies provide new opportunities to study viruses, their diversity and evolution, directly from environmental samples. The standards of sensitivity and specificity appropriate for analysis of these relatively short shotgun sequence reads are still evolving. In another part of our work, we used sensitive sequence similarity methods to identify more than 400 virus-related genes in 3,280 libraries derived from patients and environmental samples after low-complexity reads were removed. These identifications serve as a starting point to isolate viruses potentially associated with disease and outbreaks of unknown etiology.

Acknowledgement

I would like to thank all people that have helped me during my graduate study at Stowers Institute for Medical Research and Kansas University Medical Center.

I am especially thankful to my advisor, Dr. Arcady Mushegian for his guidance, support, advice and encouragement through my whole graduate study. I learned a lot from Dr. Mushegian which will help my future career and life. I would like to thank all past and current members in the Mushegian's lab: Dr. David Kristensen, Samuel Chapman, Dr. Hua Li, Dr. Lavanya Kannan, Olga Tsoy and Andrei Kucharavy. Particularly: Dr. David Kristensen gave me a jump start in comparative genomics and programming; Samuel Chapman helped me in many different things. I also thank Dr. Hua Li, Dr. Lavanya Kannan, Olga Tsoy and Andrei Kucharavy for extensive discussions on my projects.

I am also grateful to my committee members: Drs. Joe Lutkenhaus and Philip Hardwidge for their advice and suggestions on every committee meeting. I also thank them for providing opportunities to conduct my rotations in their labs and encouraging me during the difficult times.

I would like to thank all members in the Bioinformatics group at the Stowers Institute: Malcolm Cook, Madelaine Gogol, Ariel Paulson and Amy Ubben. Malcolm, Madelaine and Ariel helped me a lot on programing and gave me very good advice on my research. Amy helped me to get used to different systems in the Institute and also trips to conferences.

Finally, I would like to thank my parents and my whole family for their support and love throughout my life.

Table of Contents

List of Figures

Page

List of Appendices

Introduction

Analysis of similarities between biological sequences was one of the earliest applications of computational molecular biology and still remains a lively area of research. In this work, I applied some of the modern, sensitive methods of protein sequence analysis to two problems in virus genomics. I begin this thesis with a short review of the sequence analysis techniques that we used in our two projects. In Part I, I discuss the first project --- Phage Orthologous Groups (POGs), in which we identified conserved orthologs in completely sequenced genomes of double-strand DNA phages and arranged them into evolutionary families. In Part II, I discuss the second project of discovering new virus-related sequences in metagenomic libraries obtained by deep sequencing of various biological samples.

Protein sequence conservation in genomes and in databases, and use of conserved families for functional and phylogenetic inference

Sequence similarity searches aim at identifying the homologs of the query sequence among all the sequences in the databases. The importance of studying similarities between biological sequences was emphasized by Pauling and Zuckerkandl in a series of influential papers in early 1960s [1, 2].

In the 1960s and 1970s, the main focus of the sequence similarity analysis was to understand the trends of evolution of protein sequences whose functions were already known. Indeed, the first protein families, collected in the pioneering series of works by M.O. Dayhoff

[3], consisted of many closely related proteins with the same functions but from different biological species, such as globins or cytochromes C, and often sequenced directly, at the amino acid sequence level. There was no concern that we would not know what the function of these proteins was. However, the development of DNA sequencing and the ability to construct genomic and cDNA libraries, together with the establishment of DNA protein sequence databases, led to a new application of sequence analysis, namely comparing a novel sequence to entries in a sequence database, in a hope that it will turn out to be similar to something already known, and that this similarity will provide a clue to the origin and function of the new sequence. Thus, one of the most important questions for bioinformatics in the last decades has been to develop ways of making inferences about the function of uncharacterized proteins on the basis of their similarity with better-studied proteins.

Definition of homologous sequences within one organism and between different organisms is also the cornerstone of molecular-evolutionary studies. As emphasized by Pauling and Zuckerkandl, molecular genic characters, e.g., nucleotides, amino acids, genes, operons, etc., provide ultimate clues about the ancestral genes (and, as we know now, ancestral genomes), as well as about the tempo and mode of the evolutionary processes that led from these ancestral molecules to their currently existing offspring.

The advent of the high-throughput genome sequencing technology, resulting in the 1990s in deciphering complete sequences of many prokaryote genomes, and then dozens of diverse eukaryotic genomes, takes all this to a new level. Many organisms, their genes and proteins have been studied in considerable detail, and this information is available in online sequence databases. However, our ability to obtain genome sequences of many other, diverse organisms is far ahead of our capacity to study gene functions in all these species. Therefore,

extracting information about studied organisms and their genes and computationally relating these genes to sequences in an unstudied organism is the main practical approach to understand more about the newly sequenced species, and, eventually, about the function, evolution, and ecological relationships of all species on Earth.

Amino acid vs. nucleotide sequence analysis

In sequence analysis studies discussed in this work, we usually translate DNA into amino acid sequences. Some of the advantages of protein sequence analysis over DNA sequence analysis for our purpose are, first, that the sensitivity, resolution and statistical significance are better for a twenty amino acid alphabet versus only a four nucleotide alphabet. Second, because of the redundancy of codons, a large fraction of the bases in protein-coding regions are under less evolutionary pressure (especially the third base), so that nucleotide sequences tend to diverge faster than the protein sequences they encode. Third, the nucleic acid databases are much larger than protein databases, as the nucleic acid databases contain both coding and non-coding DNA sequences, and the entries in the nucleotide database are longer. Analyzing this dataset is computationally much more expensive than scanning the databases of putative proteins.

The widely used amino acid substitution matrices, such as PAM (Point Accepted Mutations) series constructed by Dayhoff and Eck in 1968 ([3]) and the BLOSUM series developed by Henikoff and Henikoff [4, 5] capture, in somewhat different ways, the probabilities of every possible type of amino acid replacement in the course of protein sequence evolution. BLOSUM62 has been the default substitution matrix in most sequence comparison search tools. Based on the amino acid substitution matrix and the Karlin-Altschul

statistics of database searches [6] embodied in the BLAST family of algorithms [7, 8], we can estimate the probability that a query sequence is a homolog of the database sequences.

Gene Identification

When we obtain genomic sequences by either Sanger sequencing or one of the Next Generation Sequencing (NGS) technologies, one of the first steps in sequence analysis (immediately following the contig assembly and quality control) is to identify genes. The most direct way to do this in prokaryotes and viruses, whose genes have no spliceosomal introns, is to look for a start codon (usually ATG, but in some cases, GTG, TTG, or CTG) and an in-frame stop codon (TAA, TGA, or TAG; some bacteria, like mycoplasmas, have only two stop codons as UGA codes for Trp). The sequence between the start codon and the stop codon will be a potential open reading frame (ORF). Usually, there are no restrictions on the gene length. However, in practice, prokaryotic mRNAs with ORFs that are less than 30 codons tend to be poorly translated [9], and, with not much other guidance available, many authors suggest 30 amino acids as a practical cutoff that does not lead to excessive under-prediction of coding regions. In the phage project, however, we chose a lower minimal ORF length, as phage proteins are often shorter than the cellular ones, and also because we could validate many of these short ORFs by sequence similarity analysis. In the metagenomics project, on the other hand, we chose 45 amino acids as a length limit to reduce the number of ORFs to a manageable amount, even as we realized that we were losing some good virus-related matches.

In practice, we translate the DNA sequences in six frames and then look for the longest amino acid sequence beginning with a start codon and ending with a stop codon. However, DNA sequencing errors and the special cases such as nested, overlapping or

recoded ORFs, require additional evidence that an ORF actually encodes a protein. The database search for homologs of the query ORF is used to look for the additional evidence that the ORF is not spurious. Other methods examine the GC content, codon frequency or oligonucleotide composition (no bias), and also the precedence of a ribosome-binding sequence or promoter in the given species would give extra support for a protein coding sequence [9]. These approaches were not used in our study. In the metagenomic project, we did not always require a stop codon and a start codon for our ORFs, because of the shorter average read length in these libraries.

Identification of Homologous Genes on the basis of Sequence Similarity

Homologs are inheritable characters descending from the same ancestral character. In the context of sequence analysis, we usually deal with homologous genes and their products, homologous proteins. The homologs diverge as they evolve, and, as a rule, their sequences continue to be more similar than two random sequences of the same length and amino acid composition, which allows us to identify homologs by their statistically significant sequence similarity. If two or more proteins have, for example, eighty percent sequence similarity, then there is a very small chance that they have not emerged from the common ancestor. Sequence similarity statistics, however, allow us to infer common ancestry even for pairs of sequences that are much less similar than that, with sometimes as low as 10-15 % identity. There are also cases when two homologous proteins share no sequence similarity that can be detected by the search engine, or are related at such a distant level that their similarity is recovered by the search program without statistical significance. Such homologs that have low, “gray-zone” sequence similarity are particularly common among fast-evolving genes

and genomes. Phages and viruses, the subject of this study, are an example of such fast-evolving genomes. Finally, there are pairs of homologs that have evolved beyond all recognition, and this would be missed by the sequence analysis approaches.

In this study, we examined sequence homology that is easier to establish by unequivocal similarity statistics, and also such that is reported by the sensitive search programs with borderline significance and requires additional analysis to support the homology inference. We undertook this analysis, paying particular attention to three factors: (1) the aligned proteins have approximately the same length, and the alignment extends to most of the length of both proteins, or, if one protein is shorter than the other, along the length of the shorter protein (this criterion could only be applied in the first project, where full-length phage proteins were predicted); (2) if there are several proteins related with insignificant sequence similarity, but they all share conserved sequence motifs; (3) the conserved motif has known functional importance, i.e. DNA-binding motif, metal-binding motif, etc.

Orthology and paralogy: two classes of homology

To predict the biochemical activities and biological functions of unknown proteins and also to reconstruct protein evolutionary history, it is critical to distinguish two principal types of homologous relationships. The two categories of homologs are orthologs and paralogs, which differ in their evolution history and functional implications. The distinctions between two types of homologs was first introduced by W.M.Fitch [10] . Orthologs are evolutionary counterparts in two lineages, one in each of the lineages, which arose by speciation at the most recent point of origin of these lineages. In this thesis, we focus on

orthologous gene products, although orthologs apply to other characters like chromosomal segments [11]. Essentially, orthologs are "the same genes" in different species. In contrast, paralogs are homologous genes that are evolved through duplication within the same lineage. In a word, speciation leads to the divergence of orthologs while duplications give rise to paralogs. For example, human myoglobins are orthologous to chimpanzee myoglobins, but paralogous to both human and chimp hemoglobins. More generally, as shown in Figure 1, gene 1 in species C and gene 1 in species A and B are orthologs. Gene 1' in species C and gene 1 in species A and B are also orthologous. Gene 1 and gene 1' in species C are paralogous because they are related by duplication, not speciation, at their most recent point of origin.

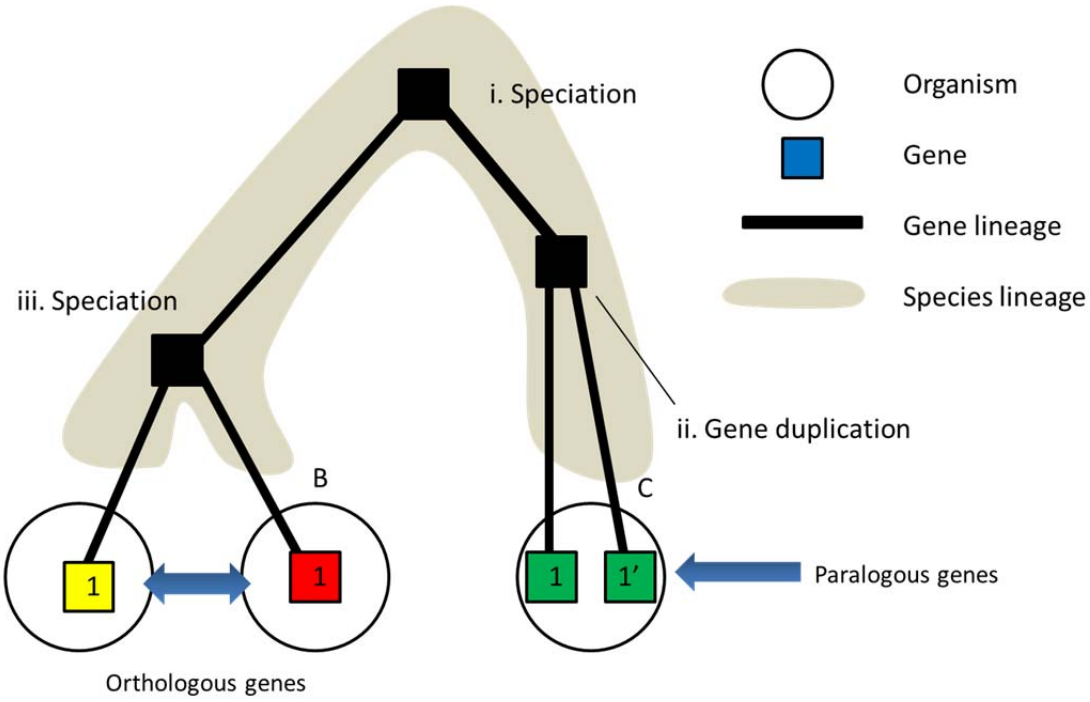


Figure 1. Two kinds of homology: orthologous and paralogous genes. The descendants of an ancestral gene are present in both lineages after speciation (i) occurs. Another speciation event occurs at point (iii), again with one descendant present in both species A and B. These descendants, i.e., gene 1 in organism A and gene 1 in organism B, are orthologs of each other by definition. In event (ii), the gene has duplicated within one lineage, producing two gene copies in the absence of speciation. In this case, the two genes 1 and 1' are called paralogs. Because 1 and 1' are in the same lineage, they are called in-paralogs and are considered co-orthologous to gene 1 in organisms A and B.

There are many practical challenges in identifying genome-wide sets of orthologous and paralogous genes, especially for distantly related organisms. This is because many events such as horizontal gene transfer, lineage-specific gene loss, gene fusion and fission, etc. complicate the routes of gene evolution and our ability to reconstruct these past events from the present-day data. A practical need to define orthologs and paralogs comes from the observation that orthologs tend to preserve the same molecular and biological function as their immediate ancestral gene. By contrast, paralogs are more likely to evolve different, though maybe related functions over time via subfunctionalization or neofunctionalization routes [12, 13], as the pressure of purifying selection acting upon at least some of the copies decreases. Thus, scientists may use the functional information of well-studied genes for annotation of their uncharacterized orthologs [14-16], but borrowing functional annotation from a paralog may be more problematic.

In an era when more and more complete genomes are available, different methods for orthology detection are available, and more of them are being developed [17]. Typically, orthologs are identified by a phylogenetic tree-based approach, or by a heuristic similarity-based approach, or a combination of the two, with synteny information for more evidence.

Methods for Orthology Identification

Phylogenetic tree-based approaches

The tree-based methods use the definition of orthologs directly and rely on a model of the evolutionary history of the given set of genes, in the form of a gene family tree. Most approaches compare this gene family tree to the evolutionary tree of the species the genes

belong to. Methods to construct the gene family tree can be based on computing distances between sequences, defined in some way, or on comparing sequence characters directly. The distance-based methods are much faster [12]. Character-based approaches are more accurate when they rely on more sophisticated evolutionary models, but usually lack in efficiency, though recent fast approximations of the Maximum Likelihood are practical enough [18-21]. The major advantage of the tree-based approach is that it models the evolution of all genes under consideration by using a multiple sequence alignment. These methods can therefore distinguish orthologous gene duplicates arising from speciation and paralogous genes arising from within-species gene duplication. Tree-based approaches are also better than pairwise-match approaches when trying to account for gene losses in multiple lineages, as long as some of the examined species still have that gene [22, 23].

Although phylogenetic analysis is the most accurate method to identify orthologs and paralogs, it has several practical disadvantages. Despite recent performance improvements, trees are still computationally intensive to construct when the number of organisms and genes is large [24]. The phylogenetic inference is sensitive to noise and bias in the data, such as unequal evolutionary rates at different sites and along different branches of the tree, saturation effects at large evolutionary distances, etc. [25, 26]. Trees are also sensitive to the accuracy of multiple sequence alignment which is error-prone when dealing with fast-evolving proteins, multi-domain proteins, etc. [27, 28].Finally, it may not be possible to use a bifurcating tree to accurately construct the evolution of prokaryotes and viruses, due to horizontal gene transfer [29-32]. These events need to be modeled in graphs consisting of both vertical and horizontal branches; but so far, no algorithms exist that can do this [33-39].

Heuristic best match methods

The heuristic similarity-based methods rely on the assumption that the sequences of two or more orthologous genes, one per organism, are more similar to each other than they are to other genes from the compared organisms. In practice, the most common method to infer probable orthologous genes is the use of symmetric best match relationships [40], sometimes also called Bidirectional Best Hits (BBHs)[41]. The BBHs are usually determined by taking the top-ranking matches from a pairwise sequence similarity search like BLAST, or by other sequence similarity measures, such as the similarity scores derived from Smith-Waterman alignments[42] or Maximum Likelihood estimates from significant scoring pairwise alignments (Reciprocal Smallest Distance, RSD). Many best match algorithms group together pairs of BBHs orthologous genes from multiple genomes. Each group of genes represents all the descendants from a common ancestral gene within the studied organisms [43]. Grouping BBHs from multiple genomes verifies their homologous relationship, because there is a very small chance that a gene/protein would randomly form a BBH with each of the two proteins in two other genomes, while these other two proteins are also each other BBHs.

The major advantage of heuristic best match methods is that they are much faster and easier to automate than the tree-based methods. Efficient implementation has been developed for assembling consistent pairs of BBHs in very large numbers of genomes. As these methods do not rely on multiple sequence alignments nor on phylogenetic trees, they avoid many multiple alignment pitfalls and artifacts associated with constructing phylogenies [27, 44].

The typical disadvantage of heuristic approaches is that they fail to detect differential gene loss [45, 46]. By contrast, the phylogenetic methods can pinpoint the lineage-specific gene loss by noting that genomes in some lineages contain both paralogs even if most lineages contain just one of them. Additionally, orthologous groups from BBHs tend to be overly inclusive and may create mixed groups in large complicated families which do not accurately represent the evolutionary history of the questioned genes, or combine two or more evolutionary families together when dealing with multi-domain proteins.

Synteny

The comparison of the general structures of many different species' genomes found that the closely related organisms show similar blocks of genes in the same relative positions in the genome. This situation is called synteny [47-49]. Homologs surrounded by the sets of orthologous genes in recently diverged species are likely to be orthologous themselves [50]. However, at least in animals, the rate of loss of syntenic neighborhoods is roughly proportional to the rate of amino acid sequence divergence in orthologs, and synteny becomes undetectable when the average protein identity is lower than fifty percent [51, 52]. Prokaryotes show a higher rate of synteny loss, which may occur even at more than ninety percent identity [48, 49]. Thus, it is not a powerful approach by itself to identifying orthologs, but it can help distinguish true orthologs, e.g., by breaking the ties between best and second-best matches with very close degree of similarity to the query.

Since the phylogenetic and heuristic approaches have their own advantages and disadvantages, in practice, the three methods can be combined to predict orthologs more accurately.

In the work here we use the heuristic similarity-based approach – the use of symmetric best matches - to build orthologous gene groups for phages. We use several additional steps to reduce some of the artifacts associated with the heuristic approach as described in Project I section.

Protein sequence complexity

Studies have shown that the distribution of amino acids in sequences of proteins with known three-dimensional structure in the Brookhaven Protein Data Bank is close to random [53, 54]. However, many proteins in all divisions of Life contain low-complexity regions, which have a biased sequence composition. That is to say, the distribution of amino acids in such regions is non-random. For example, some of them may be rich in glycine or proline, or in acidic, or basic amino acids. Other low-complexity sequences have a certain amino acid periodicity. These low-complexity sequences usually correspond to non-globular regions and are nearly always devoid of enzymatic functions, but some of them have important other biological functions. Most often they are involved in protein-protein interactions or cellular adhesion to various surfaces and to each other.

Inconveniently, the low complexity regions violate the statistical hypothesis that is the basis of sequence comparison algorithms. In general, the sequence comparison programs will report high similarity scores to non-homologous regions which are by chance enriched in one and the same amino acid or otherwise share similar composition biases, and this may lead to gross errors in sequence analysis. The SEG algorithm and the corresponding program partition the low-complexity and high (normal) complexity regions automatically [54]. In practice, the SEG program can run with a set of parameters to predict globular and non-globular regions with

considerable accuracy. SEG filtering is used as the default to mask low complexity regions in the query sequence in BLAST. The example of a virus sequence that is predicted to contain both high-complexity and low-complexity regions, is shown in Figure 2. Studies have shown that the measles nucleoprotein is divided into two regions: a structured N-terminal moiety, NCORE (aa 1–400), which contains all the regions necessary for self-assembly and RNA binding, and a C-terminal domain, NTAIL (aa 401–525). NTAIL lacks any stable secondary and tertiary structure [55]. As shown in Figure 2, the SEG program (window length 45, trigger complexity 3.3, extension complexity 3.8) partitioned this protein into low complexity (left, lower case) and high complexity (right, upper case) regions, which correspond quite well with the biochemical data.

```
>gi|13397897|emb|CAC34604.1| nucleoprotein [Measles virus]

                                  1-400  MATLLRSLALFKRNKDKPPITSGSGGAIRG
                                         IKHIIIVPIPGDSSITTRSRLLDRLVRLIG
                                         NPDVSGPKLTGALIGILSLFVESPGQLIQR
                                         ITDDPDVSIRLLEVVQSDQSQSGLTFASRG
                                         TNMEDEADQYFSHDDPISGDQSRSGWFENK
                                         EISDIEVQDPEGFNMILGTILAQIWVLLAK
                                         AVTAPDTAADSELRRWIKYTQQRRVVGEFR
                                         LERKWLDVVRNRIAEDLSLRRFMVALILDI
                                         KRTPGNKPRIAEMICDIDTYIVEAGLASFI
                                         LTIKFGIETMYPALGLHEFAGELSTLESLM
                                         NLYQQMGETAPYMVILENSIQNKFSAGSYP
                                         LLWSYAMGVGVELENSMGGLNFGRSYFDPA
                                         YFRLGQEMVRRSAGKVSSTLASELGITAED
                                         ARLVSEIAMH
ttedrisravgprqaqvsfihgdqgenelp  401-516
rlggkedrrvkqsrgearesyretgssras
daraahlpistpldidtasesgqdpqdsrr
    sadallrlqamagileeqgsdtdisr
                                  517-525  VYNDKDLLD
```

Figure 2. SEG analysis of measles nucleoprotein. E and S residues are most overrepresented in the low-complexity region .

In this work, we used the SEG program in a different way, namely to remove DNA sequences whose low complexity is manifest only at the protein level from the original read

library, in order to reduce the size of this library and remove simply, possibly non-unique regions that interfere with successful assembly.

PSI-BLAST and HHsearch

BLAST (Basic Local Alignment Search Tool) is the most popular method for sequence similarity search, which also combines the high-scoring subsequences into longer gapped alignments. BLAST first searches for a short peptide sequences ("words") with length W (usually 3 amino acids) and a score equal or greater than T , then extending them in both directions to achieve a score greater than S, where both T and S are given by adding the score for aligning each pair of residues in two sequences using a modification of the Smith-Waterman algorithm and a particular substitution matrix [8]. The choice of the program in the BLAST family depends on the nature of the query, the purpose of the search, and the database intended as the target. For example, in this work we used BLASTX in order to search for proteins similar to translated nucleotide sequence query in a peptide database. We also used BLASTP (the standard protein BLAST) to identify the query protein sequence or find protein sequences similar to the query in a peptide database. Since virus genomes are highly diverse, in this study, we also made use of PSI-BLAST, the powerful, sensitive enhancement of BLASTP. It is useful for finding very distantly related proteins or new members of a protein family. PSI-BLAST can not only find distant homologous relationship but also improve the confidence of each prediction [56], by iteratively producing probabilistic models of sequence family members found at an earlier step of database search. The first run of the program is a standard protein-protein BLAST search. The program then uses all the high-scoring matches above a certain s score or E-value cut-off to generate a

multiple alignment and builds a position-specific scoring matrix (PSSM or profile) representing this alignment. The PSSM is used to produce the alignments in the next iteration. In each iteration, any new database matches below the inclusion threshold are included in the construction of the new PSSM. When no more matches to new database sequences are found in subsequent iterations, a PSI-BLAST search halts.

HHsearch algorithm is even more sensitive than PSI-BLAST for detecting remote homologs of proteins [57, 58]. By matching sequence profiles, rather than single sequences, to the sequence targets in the dastabase, PSI-BLAST has improved homology detection sensitivity a lot over BLAST which uses sequence-sequence matching. HHsearch takes this further by using Hidden Markov Models (HMMs) of multiple sequence alignments instead of PSSM. HMMs model the amino acid frequencies at each position and also consider the possibility of deletion and insertion in each position. In addition, HHsearch recruits profile-profile searches, which improves sensitivity significantly, as sequence profiles contain much more information about the protein family than the single sequence itself. HHsearch also predicts protein secondary structures and includes the structure information into HMMs. [59].

Motifs and domains

Protein sequence motifs and protein domains are widely used terms but are not easy to define rigorously. In our projects, we define motif as a short stretch (usually spanning 10 to 30 amino acids) of conserved residues with a particular biological function. These motifs provide clues for uncharacterized proteins' function. For example, Walker A motif is present in nucleotide-binding proteins, which interacts with the gamma-phosphate group of the

nucleoside triphosphate. The motifs has the pattern (A/G)XXXXGK(T/S), which has an approximate probability of chance occurrence: (1/10)(1/20)(1/20) (1/10)=2.5 × 10^5. In a database with 3.2 × 10^8 residues, the expected matches are over 8,000. In this case, we need more information to distinguish true homologs from false homologs. Examination of the crystal structures in the Protein Data Bank showed that about half of the sequences with this minimal version of Walker A motif do not bind or use nucleotides [60]. We can, however, provide other information such as adjacent residues that tend to have conserved properties to increase the specificity (e.g., Walker A motif is always located in a loop between a predicted beta-strand and a predicted alpha-helix). Figure 3 shows a small part (Walker A motif part) of the HHsearch result for POG125, a particular phage family within the P loop NTPase superfamily. This POG consists of 20 proteins/domains. From the alignment, we can see the conserved Walker A motif GXXXXGKT is in a loop between a predicted strand and a helix.

```
Q ss_pred       CEEEEECCCCCCHHHHHHHHHHHHHhcCC-ceEEEECCCcHHHHHHHHh
Q POG125        RHITIRGPAGTGKTTLTKFLLDRLFQTGQ-QGIVLAAPTHQAKKELSKH
Q Consensus     ~~~~i~G~~GtGKT~l~~~l~~~~~~~~~~-~~v~~~a~t~~a~~~l~~~
                 +.+|+|+|||||||+++.++..+...+. .+|.++||||+||.+|++.
T Consensus     ~~~iltGGpGTGKTT~i~~il~~~~~~~~~~i~L~APTGrAAkRl~e~
T TIGR01448     KVVILTGGPGTGKTTITRAIIELAEELGGLLPVGLAAPTGRAAKRLGEV
T ss_pred       CCeEEecCCCCCchhHHHHHHHHHHHHcCCCceEEEeCchHHHHHHHHh
Confidence      899999999999999999999987653 3899999999999999
```

Figure 3. A portion of the HHsearch result for POG125, which is predicted to be an ATP-binding protein on the basis of the conserved Walker A and Walker B (not shown) sequence motifs. Q is short for Query and T is short for Target. The ss_pred value denotes the secondary structure prediction, and a confidence value is included. "H" stands for helix; "E" stands for elongated structure (i.e., a beta-strand); "C" stands for coiled region. In the consensus sequences, upper and lower case amino acids indicate high (≥ 60%) and moderate (≥ 40%) conservation, respectively. Symbols indicating the quality of the column-column match: '|' very good, '+' good, '.' Neutral and '-' bad.

In structural biology, a protein domain is described as a structurally compact, independently folding unit that forms a stable three-dimensional structure and shows a certain level of evolutionary conservation. In comparative genomics and sequence similarity

searches, the domain is a part of the protein that has the same evolutionary trajectory. Domains usually comprise 100 to 300 amino acids with two or more motifs. Two or more domains with distinct evolutionary trajectories can fuse into one protein which will cause the false joining of two or more orthologous groups. To deal with the multi-domain problem, in the POGs project, we used a heuristic approach (see page 25-28) to split proteins into domains and constructed the orthologous groups for domains instead of proteins.

Bioinformatics of Virus Proteins

Viruses are intracellular parasitic genomes, whose genetic material (DNA, RNA, double or single stranded) is surrounded by a protein coat. Viruses exist wherever life is found [44]. Viruses are the most abundant life form in the biosphere [61]. It has been estimated that there are 10^4-10^7 prokaryotes per milliliter of water depending on the location and counting methods [62], and that phages, the viruses of prokaryotes, outnumber their host at least by one order of magnitude [63]. The observations by transmission electron microscopy found up to 2.5×10^8 virus particles per milliliter in natural waters [64], and, even though the abundance decreases with depth and distance from the shore [65], given the huge volume of the ocean, the overall number of particles works out to be 10^{31} . The first direct counts of soil bacteriophage showed that virus numbers in soil averaged 10^7 - 10^8 g^{-1}. Given the huge volume of soil, this further proved that viruses are the most abundant entities on Earth [61, 66].

Viruses infect almost every clade of cellular organisms that are known. Notwithstanding the existence of "virus hallmark genes", i.e., families of distantly related proteins found in diverse groups of viruses, but not commonly found in cellular organisms [67], the degree of sequence diversity and the variation of gene repertoire beyond this conserved core is astounding.

This variability is one of the reasons that it is not easy to find a good treatment of many human diseases caused by viruses.

Far beyond their medical, veterinary and agricultural importance, viruses, especially those infecting bacteria, influence many biogeochemical and ecological processes such as organic nutrient cycle and algal blooms, etc. [63]. Viruses have long been suspected to play a major role in horizontal gene transfer in the biosphere. Recent studies of general transmission agents (GTAs), virus-like particles that appear to encapsidate fragments of host DNA biased towards less-conserved genes [68, 69] gives credence to this hypothesis.

It is not easy to study the biology of diverse viruses revealed from large-scale sequencing efforts, because laboratory or natural hosts are not known for most of these viruses [70]. On the other hand, progress in sequencing technology, currently extending to single-molecule sequencing [71-73] allows us to obtain complete genomes of uncultivated viruses and to use comparative computational genomics to decipher many aspects of virus biology.

In the case of cellular organisms, a universal gene, such as ribosomal DNA (rDNA), can be used to derive phylogenetic and taxonomic relationships. However, this technique cannot be applied to viruses because of the absence of a universal marker gene, as even one of the virus hallmark genes is present in all virus genomes. *In silico* studies have proved that the similarity based search such as BLAST can be used for taxonomic classification of viral metagenomic sequences [61]. Researchers tend to use stringent e-values for BLAST searches in order to minimize the false positive similarity identifications. However, the downside of such conservative approach is that many of the more divergent, but homologous, sequences are not recognized.

Project I: Phage Orthologous Groups (POGs)

Background

Bacteriophages (phages for short), discovered in 1910s [74, 75], are viruses of bacteria, or more generally, viruses of prokaryotes (sometimes viruses of archaea are called archaeophages). At their extracellular stage, phage particles always have two basic components: the genetic material inside and a protein coat outside. Some phages have a more elaborate particle structure, which may include several protein and lipid layers. These particles, however, are not reproducing by themselves: phages are obligatory intracellular parasites, which need to penetrate a living cell and exploit its biosynthetic machinery to produce copies of its genome and capsid.

Most phages detected in the environment are head-tail phages. Early studies focused on organismal, population, and community aspects, for example, virion survival and physic-chemical properties, plaque growth, impact on bacteria, etc. Phage genetics and biochemistry studies exploded in 1950-1960s, defining the era of molecular biology [76] . More recently, important roles of phages in ecosystems, in bacterial pathogenesis, and in host genome dynamics have renewed interest in phage biology, and novel research techniques allowed us to characterize many phage genomes and to start better understanding the mechanisms by which they interact with their hosts.

Because of the short length of phage genomes, they were easier to sequence than complete cellular genomes. Phage φX174, the first DNA genome to be sequenced, was a landmark achievement in molecular biology [77]. As shown in Figure 4, the number of complete dsDNA phage genomes stored in NCBI has been increasing, reaching a particularly fast pace

after 2001. By June 2011, there were almost 600 fully sequenced genomes of phages in NCBI genome database. Phage genetic material may be represented by DNA or RNA, in either case in double or single stranded form, ranging from ~3.5 kb (*Enterobacteria phage HK022*) to ~300 kb (*Pseudomonas phage 201phi2-1*) with a median of 60 kb. The available complete phage genomes encode various numbers of proteins, from 4 to 461 with mean 76 (s.d. 63) (Figure 5). In this project, we focused on the phages that have a double-stranded DNA (dsDNA) genome; this group comprises eighty-five percent of the fully sequenced phages.

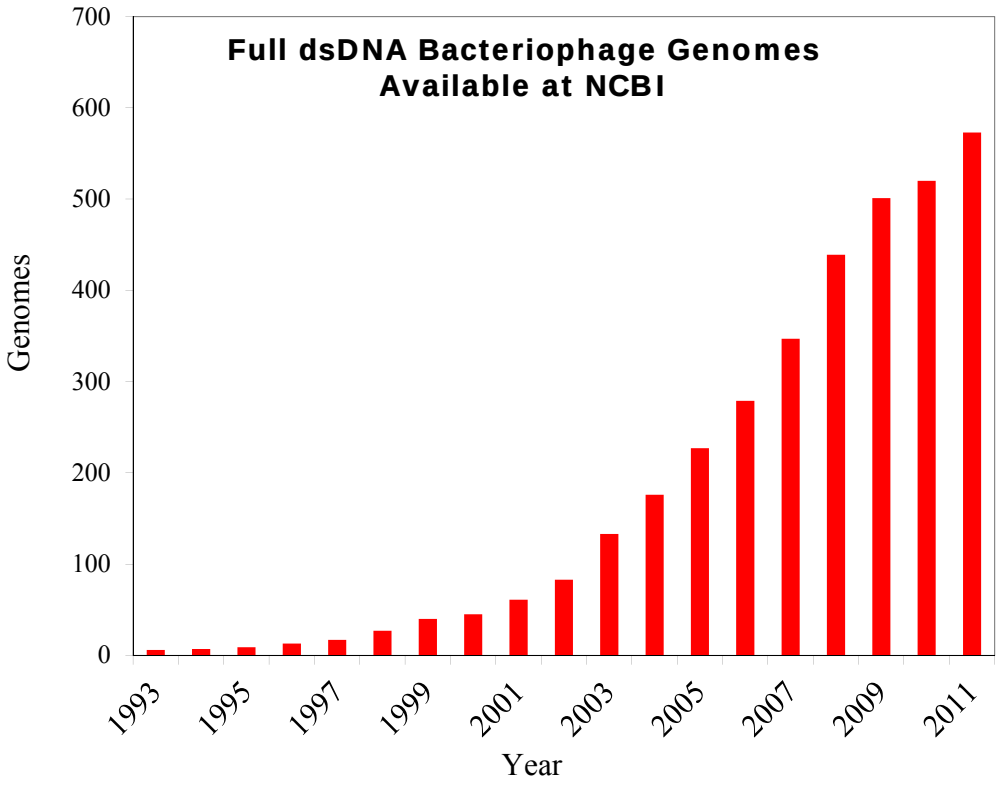


Figure 4. The growth of the number of complete dsDNA phage genomes in the NCBI Genome Database.

The high diversity and fast evolution complicates our reconstruction of the evolutionary history of phages. Because not a single gene is shared by all phage genomes, no marker is available to construct the universal phylogenetic tree, in a way that is possible for cellular

genomes, all of which share ribosomal RNA genes and genes encoding ribosomal proteins, RNA polymerase and several other universal genes. One way to overcome this limitation is to analyze the extent of shared orthologous genes in different phages. The phage orthologous groups (POGs), a natural system of phage protein families, was introduced in 2006 [78] (we call it as POGs-2004 because it included all available dsDNA phage genomes by the end of 2004). Figure 4 shows that the number of phage genomes continues to rise rapidly, which requires a strategy to update the POG resource.

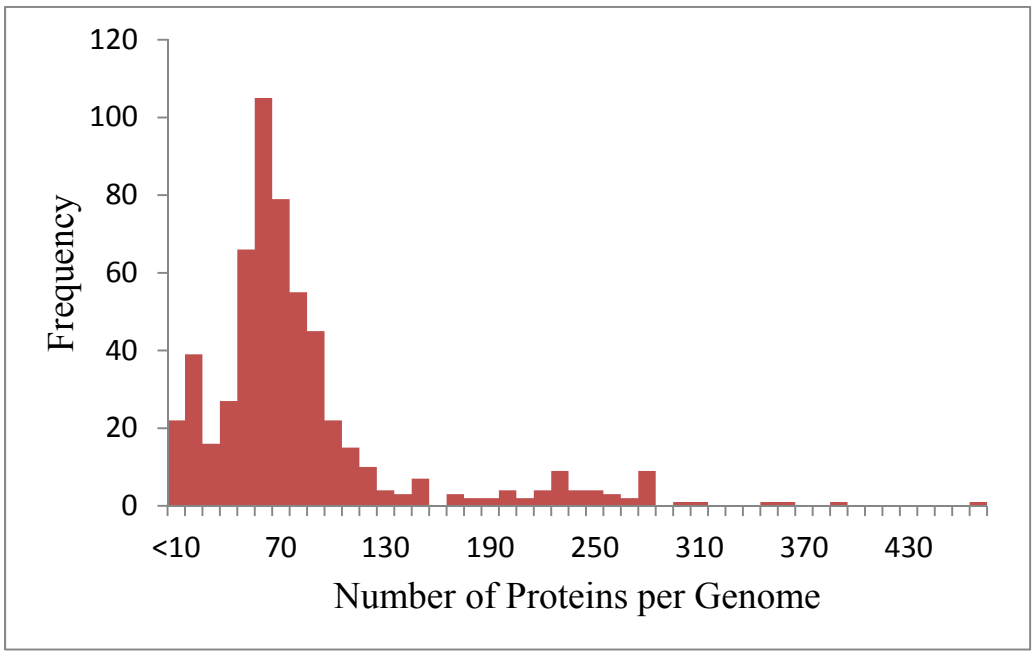


Figure 5. The coding capacity of phage genomes varies by two orders of magnitude.

Dataset: genomes and proteins

The POGs-2004 consisted of 6,378 genes from 164 completely sequenced dsDNA bacteriophage genomes that were available by 2004. In this project, we expanded the POGs system to include 323 genomes, i.e., all available dsDNA phage genomes by 2007, which is

twice the number of genomes available for the original POGs, followed by a fully automated update to account for even more recent additions to the phage genome databases.

The POGs-2007 dataset, which is the main subject of the first part of my thesis, includes 169 *Siphoviridae*, 67 *Myoviridae*, 57 *Podoviridae* (and 8 more *Caudovirales* not assigned to a family), 8 *Tectiviridaviridae*, 1 *Corticoviridae*, and 1 *Plasmaviridae*.

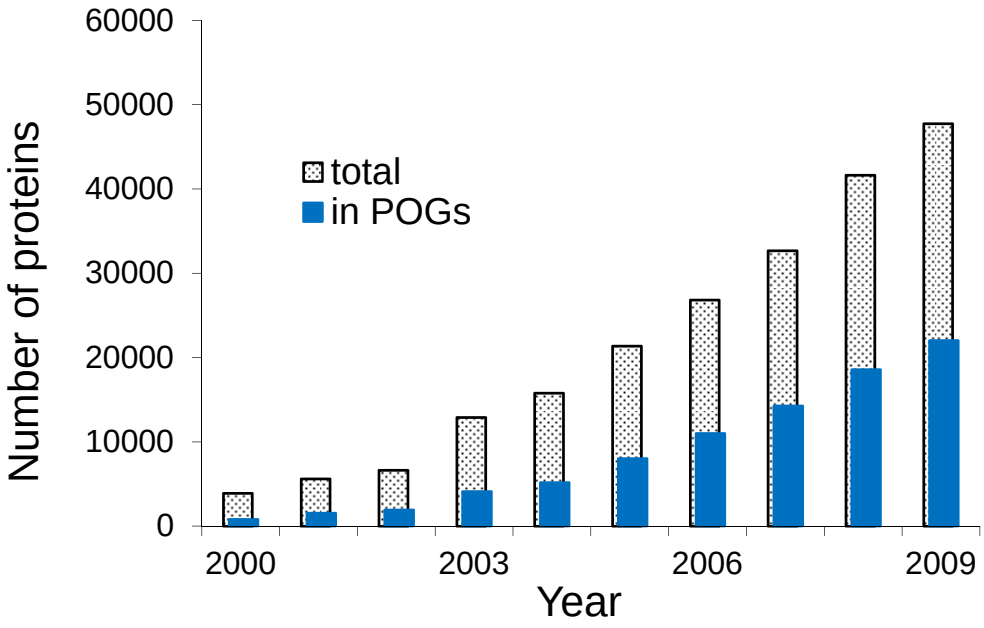
The hosts of these phages represent a broad range of prokaryotes. Among them, *Staphylococcus aureus, Escherichia coli, Lactococcus lactis*, and members of the genus *Mycobacterium* are hosts for more than 20 phages each. About 100 other listed hosts belong to a variety of phyla, such as *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Cyanobacteria*, the *Bacteroidetes/Chlorobi* group, *Tenericutes*, the *Deinococcus-Thermus* group, *Crenarchaeota*, and *Euryarchaeota*. Most of these are only hosts for one or two phages.

The 323 complete genomes encoded 27,254 putative proteins. Among them, 25,675 (ninety-four percent) were recorded at NCBI, and 1,579 (~5 ORFs per genome on average) were predicted by us using the GeneMarkS program [79], which started from the codon frequency model and then updated on subsequent iterations with the model derived from already-found homologous sequences in the public databases. The proteomes of the well-studied T7, T4, and lambda phages, as well as 25 finished, expert-annotated phage genomes were used as gold standard to test the sensitivity and specificity of the program. On average, GeneMarkS has recall of ninety-three percent and precision of ninety-five percent among the 28 phages (recall is the number of true positives divided by the number of actual positives, and precision is the number true positives divided by the number of elements labeled as positive).

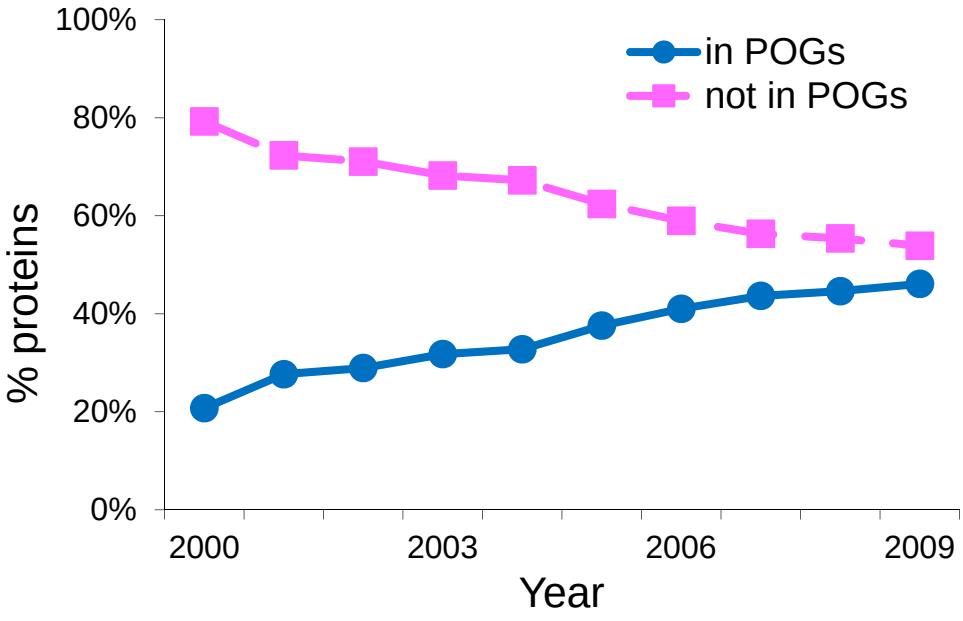
Construction of Orthologous Groups

POGs were constructed from three or more proteins that are each other's best-scoring matches (BBHs). A draft set of POGs was constructed using the original POGs-2004 procedure for the 27,254 proteins in 323 dsDNA phages. We first applied the COGtrinagles program which was used to construct COGs in cellular organisms at NCBI [80]. Later in the process of this work, our group developed and applied the efficient EdgeSearch algorithm [81]. The new algorithm scales much better with a large number of genomes and reduces the run-time about an order of magnitude. After this process, 13,470 proteins formed 2,015 candidate POGs. About 49 percent of all proteins in completely sequenced dsDNA phage genomes are conserved in 3 or more phages each. Of the 1,579 proteins that were identified by GeneMarkS, 357 were included in POGs, and six POGs contain proteins entirely from this source. Another 1,017 of these proteins had at least one database match. However, these matches are not necessarily in a completed phage genome. This indicates that many of the newly predicted ORFs are conserved gene products that will probably form new POGs in the future when more phage genomes become available.

There were a total of 2,015 POG candidates, which spanned all 323 genomes. The first main result of our study is that the number of POGs continues to grow sharply. On the other hand, the proportion of genes that are conserved between several phage genomes (definition of a POG, with some qualifications below) continues to rise, too. Figure 6 (a) and (b) show that the number and percentage of proteins conserved in POGs increase each year, while the percentage of proteins not in POGs decreases. This is because new proteins are added to the existing POGs and new POGs are created from singletons or pairs of proteins that had no mates to form the initial triangle before.



(a)



(b)

Figure 6. Properties of the conserved phage proteins over the past decade. (a) Total number of proteins in genomes and number that are in POGs; (b) percentage of phage proteins in and outside of POGs.

The number of candidate POG proteins in a genome varies. Some organisms, such as *Mycoplasma phage P1*, had just one, whereas others, such as *Enterobacteria phage T4*, had

more than 200. The average number of POG proteins per phage is 42, (s.d. 35) and the average number of proteins in a phage genome that are in POGs is fifty four percent . POGs contained an average of seven proteins from seven phages, but some had as few as 3 proteins/3 phages and as many as 141 proteins/136 phages. Also, POG proteins tended to be about fifteen percent longer than other proteins.

Most POGs do not contain paralogs. In contrast to COGs encoded by cellular genomes, in which sixty-four percent contain at least one paralog and thirty-seven percent have more, only seven percent POGs contain at least one paralogs and zero point two percent have more than one. The maximum number of paralogs in a phage genome is four, while the maximum is 122 in cellular genomes. Proteins associated with phage gene regulation, such as Cro/cI repressors and Roi/ANT-type phage antirepressors, more often have multiple paralogs. Other examples are "selfish" proteins like DNA methylases and HNH homing endonucleases.

Domains

The evolutionary histories of each domain within a protein may be different. Thus two separate orthologous families might be mistakenly merged together because some domains might fuse together during evolution. An extreme example is the modular signaling domains in eukaryotes. Although phage proteins are relatively short, still some of them are multi-domain proteins. Figure 7 shows that one POG, which contained 19 proteins under the old method, was split into three POGs--POG125, POG126 and POG 1629--when studying domains instead of whole protein sequences. POG125, consisting of 20 proteins/domains, is a DNA helicase, P loop NTPase superfamily. POG1629, consisting of three domains, is the

C-terminal of exonuclease/ RecD family, and POG126, consisting of 11 domains, is a Ti-type conjugative transfer relaxase (TraA) family.

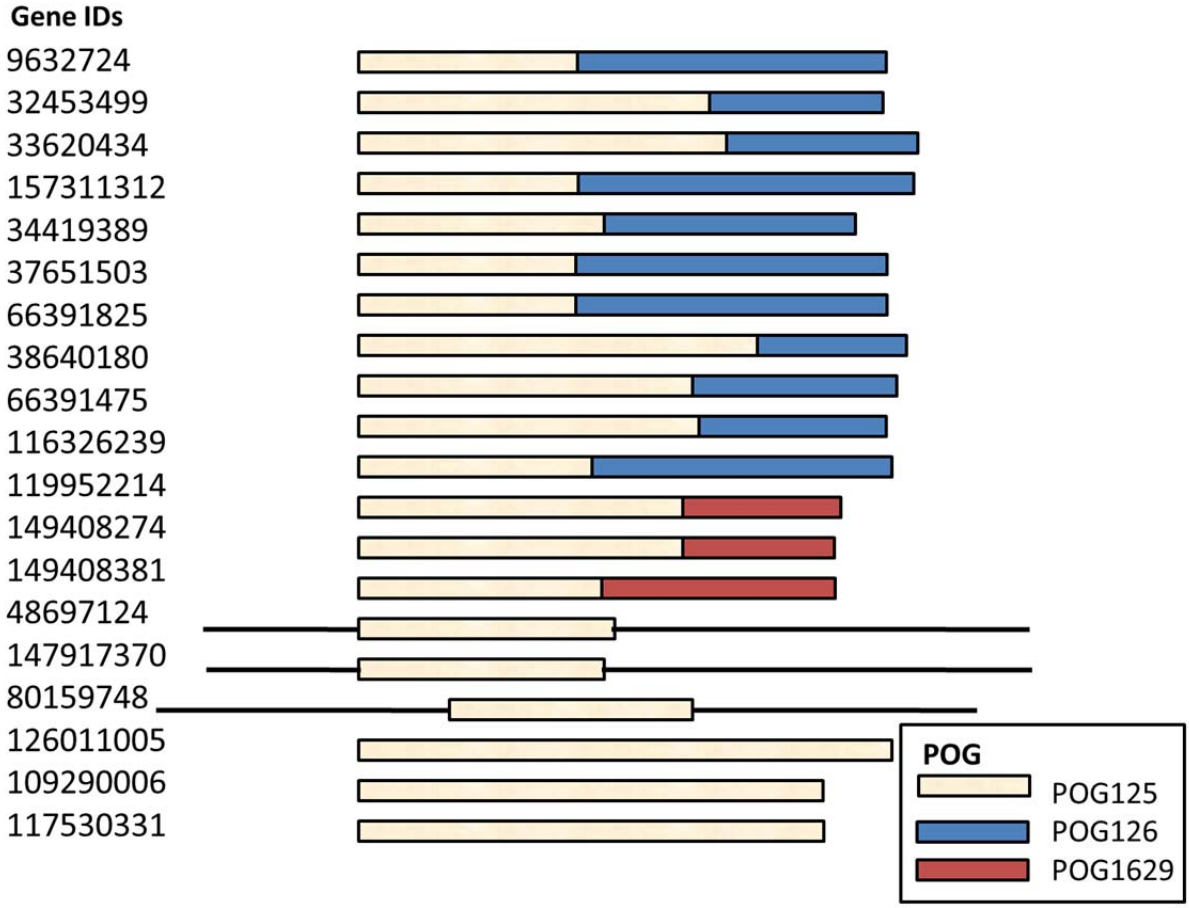


Figure 7. The relationship among POG125, POG126 and POG1629. These three POGS are created when considering domains, but would exist as a single POG when considering proteins whole proteins.

To avoid mistakenly joining evolutionarily separate multi-domain proteins together, we split proteins into domains and examine the evolutionary histories of domains instead of proteins. We used a heuristic approach based on the hidden Markov model (HHM) matching method HHsearch to identify domains. We used the SMART, PFAM, LOAD and CD subsets of the NCBI's Conserved Domain Database [82] as the search space. Various extent of curation has been applied to these databases to reduce the number of multi-domain protein families in them. When we detected matches to multiple distinct domains in one of the phage proteins, the unmatched regions of the protein were split evenly between the closest domains

if the unmatched region is shorter than 100 amino acids otherwise the unmatched region was treated as another domain. Applying this approach to the 2,199 phage proteins, we obtained 5,128 candidate domains. On average, there are seven multi-domain proteins per genome, that is, eight percent of the proteins in each phage genome. This ratio is much lower than the estimated sixty percent of multi-domain proteins in unicellular organisms from the three kingdoms of life and more than eighty percent of multi-domain proteins in metazoans [83].

Using domains instead of proteins to construct POGs, the total number of POGs increases by eleven percent to 2,227. As a result, the overall number of phage proteins/domains that are in POGs increases by fourteen percent, but as with the previous dataset, this is still about half (fifty-one percent). Eighty-six percent of the original 2,015 POGs remain unchanged and the other fourteen percent are either adding or removing members. The average number of POG proteins per phage is 48, (s.d. 38), which is a little higher than 42 (s.d. 35) observed prior. On average every POG still has seven proteins/domains from seven phages.

To understand the effect of gene fusion, we looked for examples in the new POGs built with proteins/domains where a POG built with full-length proteins was split into multiple POGs. By doing this, one can approximate the number of chimeric POGs that were not joined properly using the original full-length protein method. Only about one percent of proteins in the dataset appear to be in this category, and many of those have domains that tend to co-occur, such as variable N- and C-terminal domains surrounding a central catalytic domain. Examples of such domains include Ig-like domains found in structural proteins; some anti-repressors that contain different combinations of ANT, Bro-N, HTH_XRE, KilA, Phage_pRha, and others; some domains in several cell wall-associated lytic enzymes, such as the peptidase_M23 superfamily,

bacterial SH3, amidase, and also peptidoglycan-binding domains; and the set of cI repressors artificially joined to other DNA-binding proteins on account of a common XRE-type Helix-Turn-Helix domain. Therefore, while it is advantageous to use the automated domain dissection method to annotate COGs in organisms comprised of cells, especially metazoans, there are few instances of this method affecting POGs in dsDNA phages.

Phage isolates

The POG-2004 resource treated each phage genome as a separate evolutionary lineage. In POG-2007, we decided to join closely related phage genomes, in order to avoid inflation of POG counts (e.g., three closely related *Bordetella* phage isolates BPP-1, BMP-1 and BIP-1 share more than ninety-nine percent identity at the nucleotide level and therefore each of their 50 shared genes gives rise to a POGs, even though many of these genes are not seen in any other phages). To remove this bias, we grouped phage isolates by joining all genomes that share ≥ ninety percent of their genes. As a result, strains of essentially the same phage with minor genetic rearrangements become a single entity. The use of lineages based on groups of isolates instead of individual phage genomes for POGs construction is designed not to assemble phages into high level hierarchies but only to alleviate the redundancy by grouping together closely related phages. Proteins must be conserved outside this group (in at least two other lineages) to be considered a POG. After this step, the 323 individual phages yielded a total of 280 lineages. About eighty-nine percent lineages consist only a single phage and the rest contain multiple phages isolates (from two to six).

Using isolates instead of the original lineages reduced the number of POGs by twenty-four percent to 1,689. In the new set, POGs covered about forty-three percent of

phage genes. Each lineage has at least one conserved protein or domain, and on average each phage genome has about 47 (s.d.40) such proteins/domains. Also, each POG has from 3-219 lineages, with an average of 7.9 (s.d.12.8) and median 4. No POG includes proteins/domains from more than half of lineages. However, the percentage of phage proteins that are in POGs is still more than forty percent.

Functional annotation

To better study phage conserved proteins/domains and also provide a resource for experimental analysis of viral protein function, we annotated the function of phage families on the basis of sequence similarities by PSI-BLAST and HHsearch. About half (905) of the POGs can be functionally annotated to some extent. One-third of the 905 POGs are related to virion structure and assembly, such as head, tail, scaffold, packaging, portal, etc. Another one-third of POGs function in genome maintenance and expression, such as replication, repair, recombination, transcription, regulation and so on. The remaining POGs are involved in other activities, i.e., host lysis, metabolism, host ribosomal control, etc. The annotations that I produced are included into this thesis as Appendix A.

Paralogy

Different from genomes of cellular organisms, phage genomes have few paralogs. Only three point five percent POGs have even one paralog, and only zero point three percent POGs have more than one paralogs. The HNH endonuclease family has five paralogs, which has the maximum number of paralogs among all POGs.

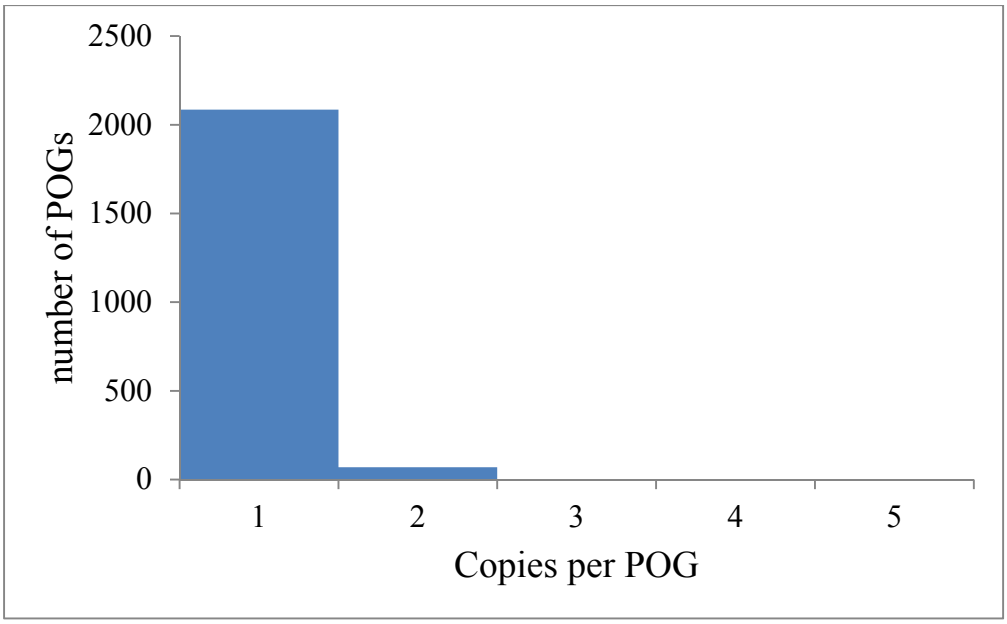


Figure 8. The number of POGs containing different copies of paralogs.

Phageness Quotient

Phages can transduce genes between their hosts in the lab experiments, and there is no reason to think that transduction is limited to man-made situations. The virus-like Gene Transfer Agents (GTAs) may play an important role in Horizontal Gene Transfer (HGT) between hosts[69]. Phages, especially the temperate ones, may contribute to gene flux between different species and lineages of prokaryotes. The extent of this transfer is not fully known, but it is thought to be an important factor of horizontal gene transfer in populations of bacteria and archaea [84, 85].

A related but different question is what impact host gene transduction is having on the phage genomes: are phage and host genes completely jumbled in evolution, or can we find phage genes that have little involvement in exchanges with the hosts? We have defined a Phageness Quotient (PQ) of each conserved family to measure the outcome of this exchange.

At first, we defined PQ as the log ratio of two hypotheses: the null hypothesis H_0 (a member of this family comes from a phage genome) versus the alternative hypothesis H_1 (a member of this protein family comes from a cellular genome. In other words, PQ equals log of the frequency of occurrence of family i in phages (f_v) over the frequency of occurrence of family i in prokaryotes (*fc*). The frequency in each organism (phages or prokaryotes) is calculated as the ratio of the number of organisms which has at least one match to the total number of organisms. PSI-BLAST matches to the 323 ds DNA phage genomes which were used to build POGs were counted. Suppose this protein matches to x dsDNA phage genomes. Then the numerator *fv*=*x*/323. PSI-BLAST matches to the non-prophage regions of 727 completely sequence microbial genomes (by August 2007) were summed in the denominator. Suppose the result is *y*. Then the denominator equals *fc*=*y*/727. So this original PQ= *fv*/*fc*=(*x*/323)/(*y*/727). In this formulation, PQ goes from negative infinity to positive infinity.

Another way to defined the Phageness Quotient (PQ) is PQ=f_v/(f_v+f_c). In this new definition, PQ goes from 0 to 1. *x* and *y* have the same meaning as above, *m* is the number of complete phage genomes, and *n* is the number of complete microbe genomes. The new PQ can be calculated as: $PQ = \frac{f_v}{f_v+f_c} = \frac{\frac{x}{m}}{\frac{x}{m}+\frac{y}{n}} = \frac{1}{1+\frac{f_c}{f_v}}$

As phage proteins are widely and sparsely distributed in phage genomes, the number of matches of POG proteins to phage genomes is relatively low. Only one POG (a TMP repeat-containing tail protein) matched more than half of the genomes. An average POG matched eleven phage genomes. There are not many very large POGs (>50 proteins/domains per POG), and many of them tend to have PQs near 0.5. These POGs have ubiquitous and promiscuous functions, such as integrase, which are shared by both phages and cellular mobile elements. Noticeably, 851 POGs (fifty percent), consisting of 3 to 32 proteins/domains (5.5 on average),

have a PQ of 1. These POGs are not observed in cellular genomes outside of the recently integrated prophages. This fact suggests that even though dsDNA phages have the ability to share and transduce the host genes, they at the same time also maintain a large fraction of unique, phage-specific genes that may be regarded as "dsDNA phage hallmark genes". Another 402 POGs (twenty-four percent) have more matches in phages than cellular hosts.

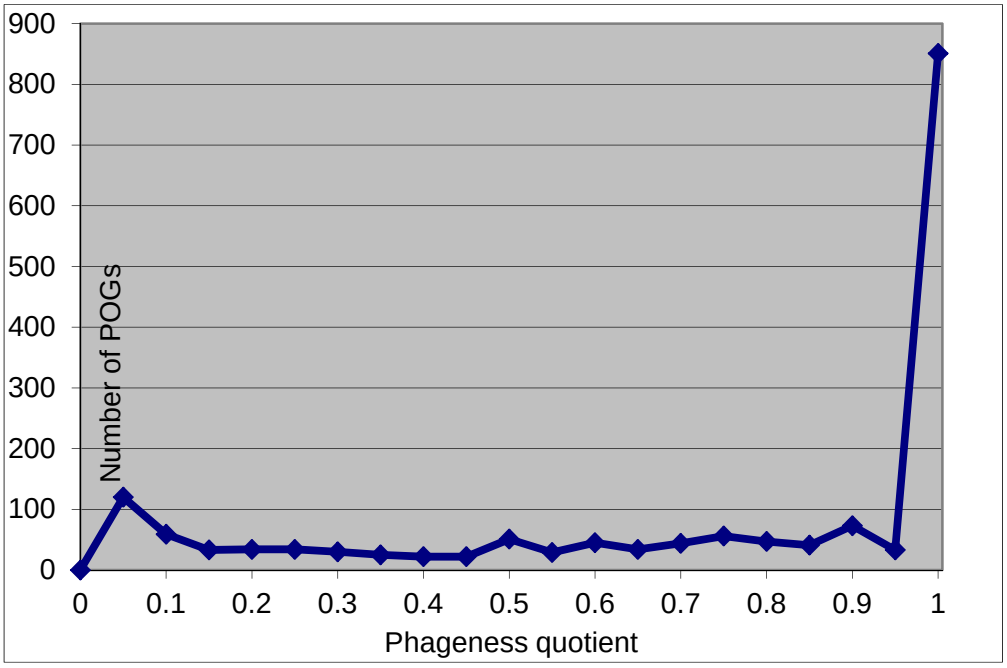


Figure 9. Distribution of the new phageness quotient values by the number of POGs.

Most of the POGs with a PQ of one have phage essential functions, such as virion structure (tail, head, portal, components, and assembly factors), transcriptional control (transcriptional activator rinB and late promoter transcription accessory protein), translational control (sigma factor for T4 late transcription and RegA translational repressor), and virulence (cell wall hydrolase). Many other high PQ POGs have unknown functions. As expected, the enzymes of cellular metabolism tend to belong to POGs with lower PQs. However, there are

exceptions. For example, S-adenosyl-L methionine hydrolase has a PQ of one. This protein is presumably used by the phage in restriction-modification warfare with the host to make sure that its own DNA is unmodified. This is useful when the phage encodes its own methylation-dependent restriction enzymes [86].

We also did a reverse search of cellular COGs (publicly available NCBI COG resource ftp://ftp.ncbi.nih.gov/pub/COG/COG/) to phage proteins to estimate the degree to which the cases of gene sharing between phages and their cellular hosts. In the cellular COGs, only eleven percent match dsDNA phage proteins with E values less than 0.01, and four percent match a conserved protein in a POG. In other words, even if the number of shared phage/host genes is large, it is not high when expressed in terms of a percentage of phage genes, especially when considering orthologs that are conserved across a broad range of phage lineages.

Project II: Metagenomics

Background

The next-generation sequencing technology has broken the technical and economic bottlenecks inherent in Sanger sequencing, enabling an in-depth comprehensive view into the diversity of DNA in complex samples. Metagenomics, the study of the gene content in a complex sample, provide a more unbiased view of the phylogenetic composition and functional diversity within a community and discovers new genes from uncultured species [87]. For diseases with no known etiology, sequencing and analyzing DNA content in samples from patients may help identify the responsible pathogens. However, the development of technology also challenges bioinformaticians to analyze the massive amount of data in an efficient way.

In the second part of my work, I studied the large and heterogeneous collection of metagenomic data provided to us by Herbert W. Virgin and David Wang of the Washington University at St. Louis (WashU). They are interested in describing the known and especially novel viruses in various samples of medical and public-health concern. For example, for patients with diarrhea, WashU colleagues obtained fecal samples; for patients with fever, our colleagues collected serum or a nasopharyngeal (NP) swab. By collecting different samples for different diseases, they were able to better characterize the pathogens. Environmental samples were also sequenced, such as sewage samples from different locations.

The WashU colleagues followed the standard methods to extract whole DNA from these samples [88]. In order to distinguish different samples, the colleagues amplified each sample using the PCR primers barcoded by 6 nt unique sequences. They pooled samples together and

sent them to the Washington University Genome Center for sequencing using the Roche 454 platform and Illumina GS FLX Titanium platforms. Within two years, our colleagues generated a total of 75,145,532 reads.

To identify putative virus-related sequences in the data, our colleagues have built a standard pipeline to process the dataset, centered around RepeatMasker and BLAST. RepeatMasker is a program that screens DNA sequences for interspersed repeats and low-complexity DNA sequences [89], and the output of this program can be given to the BLAST programs as well as the programs that perform sequence assembly, gene finding, etc. The pipeline includes three BLAST steps: BLASTN against human genome (e-value ≤ e-10), BLASTN against nt (e-value ≤ e-10) and BLASTX against nr (e-value ≤ e-5). Any sequences that did not have any significant matches at any the steps are called "unassigned" and pooled together, repeating the analysis several times over the years.

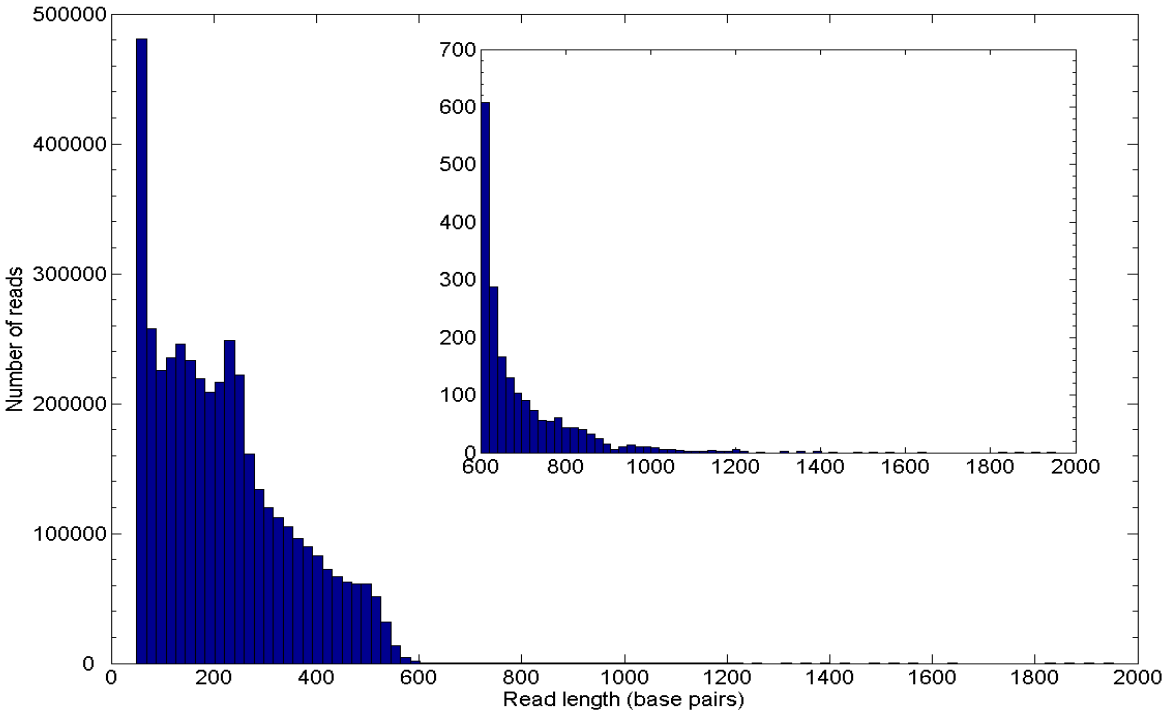
The unassigned reads from different samples were finally sent to us. As this dataset contains sequences from about 3,280 samples from a wide range of habitats, we consider it a metagenomic dataset and we call it Dataset0.

Since metagenomic data analysis is still in its infancy, there is no standard protocol to deal with such data. Our colleagues have noticed only a low proportion of their assigned reads had significant similarity to virus DNA or proteins (though with very large number of reads, this low proportion actually worked out to be a high absolute number). Moreover, many of the matches were only in a short region, because reads generated from the next generation sequencing (NGS) technology were themselves short (see Figure 10(a) for the length distribution). Therefore, we decided to assemble the unassigned data in an attempt to produce

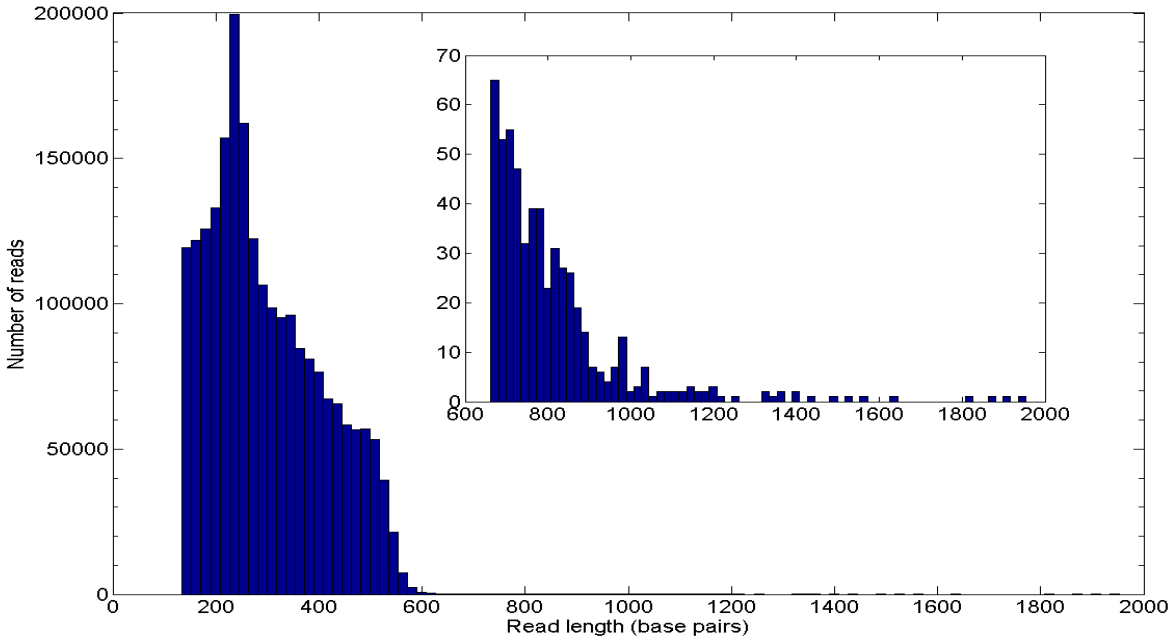
longer putative protein products which, hopefully, would match more virus sequences in the database.

The properties of the unassembled dataset and a pilot study of sequence similarity

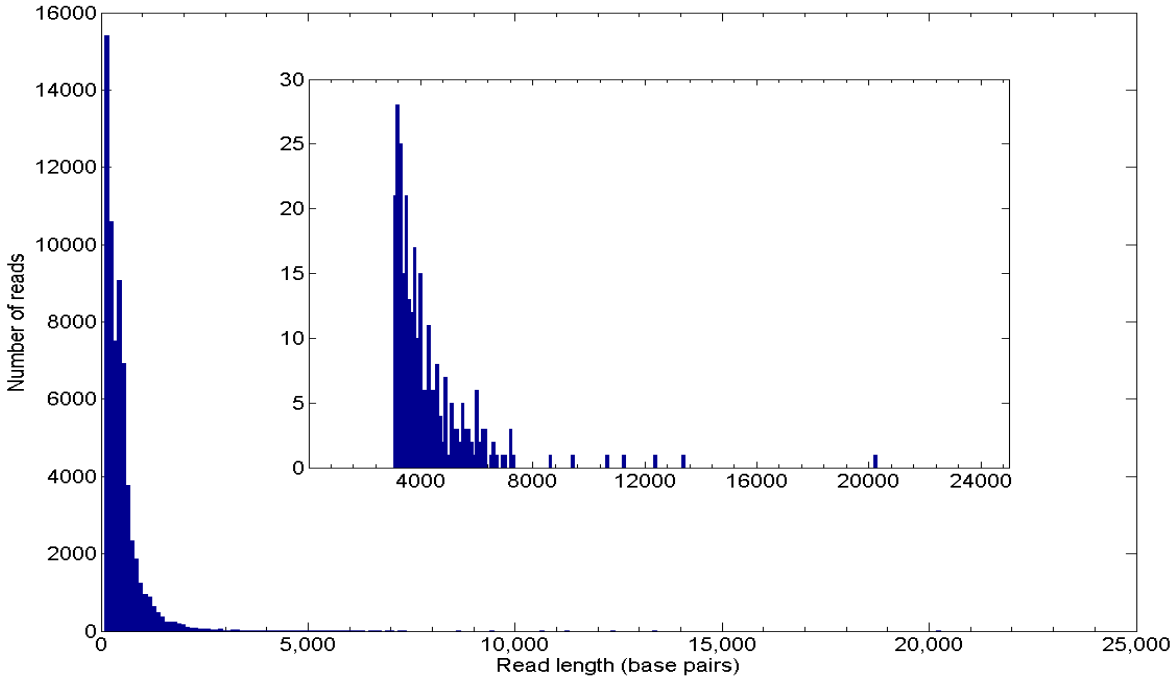
Dataset 0 consisted of 4,123,105 Roche 454 sequence reads. The length of the reads was from 50 bp to 1,953 bp, with an average of 219 bp (Figure 10(a)) and more than 75% of the reads shorter than 298 bp (the very thin tail of the long reads is not visible at all at the resolution of Figure 10(a) and is shown on the inset).



(a) Dataset 0



(b) Dataset 1



(c) Dataset 2

Figure 10. Length distribution of Dataset 0 (a), Dataset 1 (b) and Dataset 2 (c).

Removal of low-complexity sequences at the amino acid level

After inspecting the reads in Dataset0, we found many low-complexity sequences at the amino acid level. Because of the degeneracy of the genetic code, low-complexity and quasi-repetitive amino acid sequence regions cannot be detected and removed easily by masking simple or interspersed repeats at the nucleotide level. Therefore, we translated all four million reads in six frames, collected Open Reading Frames (ORFs) longer than 45 amino acids and applied the SEG filter to detect the potential non-globular regions [53, 54]. Any reads that did not encode a globular peptide longer than 45 residues (in total, 1, 213,122 reads) were removed. The length of these removed reads was from 50 bp to 1,383 bp with average of 126 bp. On the other hand, 2,209,983 reads (Dataset1, Figure 10(b)) passed the SEG filter. They were longer than the removed low-complexity reads, with length ranging from 135 bp to 1,953 with an average of 300 bp.

We realize that this filtering must have removed some of the biologically interesting sequences, e.g., those encoded by bacterial or viral pathogens, but we traded that for the ability to focus on globular proteins from these sources, which were more informative and more suitable for phylogenetics analysis.

Assembly

ROCHE 454 recommends their proprietary Newbler program (also called gsAssembler) for read assembly, but this engine has mostly been tested on and optimized for the assembly of shotgun genome sequences, not for the more complex metagenomic libraries. The running time of the program is proportional to the length of the dataset and highly sensitive to the extent of repeats in the data. This might be the reason that our attempts to assemble Dataset0 using various Newbler settings

failed. After we removed the reads encoding the low-complexity peptides, we applied the Newbler program to assemble Dataset1. We used 'CPU = 0' option that employed 24 CPU cores of our powerful Linux server. The total computation time was roughly two hours. The detailed parameters are given in Appendix B.

The 2,209,983 reads of Dataset1 were assembled into 63,967 contigs (Dataset 2, Figure 10(c)). The length of the resulting mix of contigs and singletons was from 100 to 20,279 bp with mean 484 bp. Assembly greatly increased the length of the sequences. Figure 10 (c) shows the length distribution of the contigs (note that the scale on the abscissa is different from Figures 10(a) and 10(b)).

Translation

All the 63,967 contigs were translated in six frames and produced 295,806 putative protein-coding ORFs longer than 45 amino acids (the range is from 45 amino acids to 2,016 amino acids with an average of 79 amino acids). On these 22,894 contigs, we detected 57,564 non-overlapping putative protein-coding regions, with one contig encoding as many as 27 ORFs.

Virus Genes

We ran PSI-BLAST and HHsearch with the ORFs that have more than 100 amino acids and examined the results. The matches were verified by curation of conserved sequence motifs manually. There are 55,347 ORFs in total, with the shortest having 100 amino acids and the longest 1,301 amino acids. The analysis of the taxonomic origin of the nearest database matches shows clear prevalence of proteins from prokaryotes (almost exclusively bacteria, with occasional archaea), followed by bacteriophages and then viruses of eukaryotes, and a nearly-complete absence of eukaryotic genes. As discussed before, we have focused on matches to the eukaryotic viruses,

especially animal viruses. Appendix C includes matches to all the putative virus proteins that we were able to identify with some confidence.

The information in Columns 3 and 4 of the Table is for the best database match of the query sequence, not for the query itself (e.g., Cell 3:2 and Cell 4:2 have "GKS/GKT ATPase / replication protein E1" "Human papillomavirus type 55", which is the proper annotation of gi 1020269 in Cell 2:2, not of the query in Cell 1:2). This has to be remembered because some of these best database matches may be close or identical to the new sequence that we were analyzing, whereas other best matches may be their distant homologs.

In total, we were confident about 411 matches. Among them, 74 ORFs were homologous to proteins of 11 dsDNA viruses; 235 ORFs to 21 ssDNA virus proteins; 83 ORFs are homologous to 31 ssRNA viruses; 7 ORFs to 4 dsRNA viruses; 8 to 2 unclassified viruses (in both cases, these viruses are only partially sequenced).

Some ORFs matched to the non-overlapping regions from the same virus, but we had no way to distinguish computationally whether these ORFs are from the same virus species or from several closely related virus species. For example, we identified three ORFs (contig31601_5_1, contig32622_2_1 and contig41498_2_1) that matched to the same protein --- replicase polyprotein [*Drosophila C virus*] (gi|9629651). Figure 11 (a) shows the sequence alignment for these three ORFs. However, we cannot say whether the three ORFs were from the same virus or from more than one virus. Another example is shown in Figure 11 (b). In the same vein, we identified eight ORFs matched to different portions of the *Gryllus bimaculatus nudivirus* genome. Figure 11 (b) shows a part of the *Gryllus bimaculatus nudivirus* genome with two ORFs matched to it. Again we could not use bioinformatics to distinguish whether these two ORFs were from the same virus species or two different species. As shown in Figure 11 (c), when two or more ORFs match to the same virus

genome, as the red ORF and the blue ORF both do, full-length genome sequencing and closure is required to determine the number of distinct genomes in the sample.

```
Query= contig32622_2_1         (396 letters)
>gi|9629651|ref|NP_044945.1| replicase polyprotein [Drosophila C virus]
 gi|2388673|gb|AAC58807.1| replicase polyprotein [Drosophila C virus]
          Length = 1759

 Score = 68.9 bits (167), Expect = 1e-09,   Method: Composition-based stats.
 Identities = 84/321 (26%), Positives = 147/321 (45%), Gaps = 36/321 (11%)

Query: 72   QGCVD-ANQYQMIEMVARQQYQILGAYPSGVK--PMGNGIVVKGSIFLTCQHFIYVM--N 126
            QGC D A    MI++  +  Y++  +Y  G K   +GN   V+G  F+   HF+  +
Sbjct: 945  QGCSDDAAHNLMIDVFQKNTYRM--SYFRGDKRYQLGNCTFVRGWSFIMPYHFVQAVFAR 1002

Query: 127  RNPPTHVIFRNSYLPNGIV---------ATYDDFK------RKIVEVENKQDLVLVDM-T 170
            R PP  +I  +  +   ++          A  D+F       R   +  + +D V+V++ +
Sbjct: 1003 RLPPNTIISLSQQMSEDLMQIPLSHFFSAGVDNFYLTDNCVRLPFKNGDFRDCVMVNLHS 1062

Query: 171  RCLPPGKDITKHFISRDSLPRLT-SFRAVLSGHRHVKESLFLSASSGSAEVSTEIRYTLA 229
            R   P +D+ +HFI      +L  SF   ++         L+   +  +A    + +  +
Sbjct: 1063 RMCTPHRDLVRHFILTSDQGKLKGSFSGAMATFHVNNMGLYRVYNWLNAVRPCDKKIEIF 1122

Query: 230  H-DDGTK-QQVNAIKTICYDID--TKDGDCGMILTVSDYQLPAKIVGLHVAGTDKIAFYN 285
            H +DG +  + + I+  CY+ +  T+ GDCG I+ + +  L  KI+G+H+AG D    +
Sbjct: 1123 HPEDGFEYPEESYIQRDCYEYNAPTRTGDCGSIIGLYNKYLERKIIGMHIAGNDA-EEHG 1181

Query: 286  FASLTTAETLRETVSKHFPLEAQVGQS---FEVFNSVVDNAPLSEDLPFKGEFLPIGK-V 341
            +A   T E L    S       +   S   +E+ N V    PL +    +G+F  +GK
Sbjct: 1182 YACPLTQECLETAFSALVNKNKKNISSQFYYEIPNMV---DPLGDSSVPEGKFYALGKSS 1238

Query: 342  FTSGEALKTRIKPSIIWGALS 362
               G+A+ + I PS I+G LS
Sbjct: 1239 IRVGQAVNSSIIPSRIYGKLS 1259
```

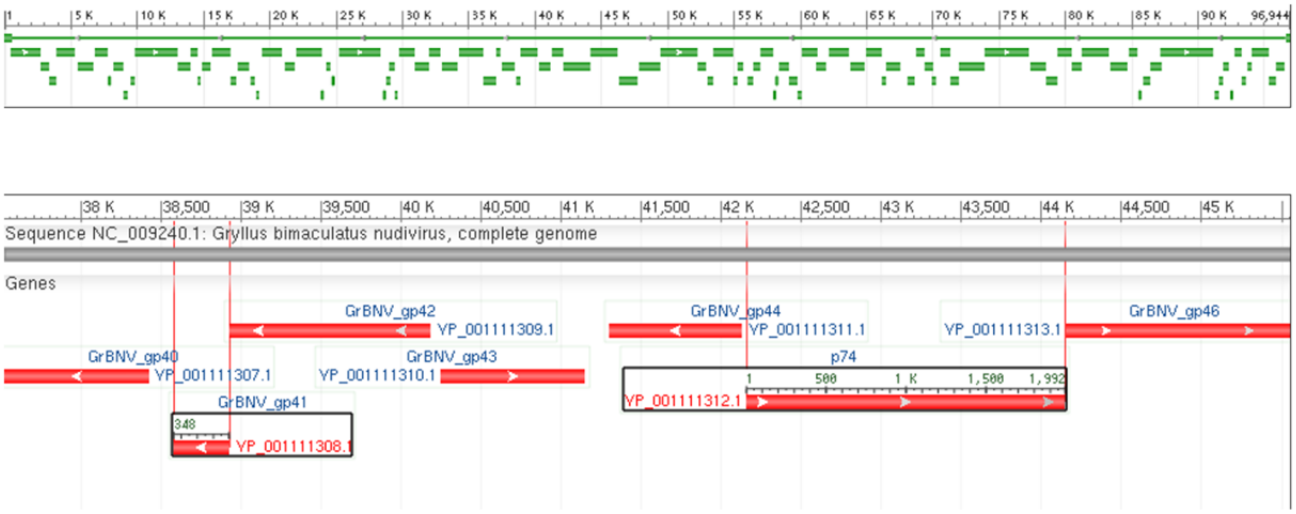
```
Query= contig31601_5_1         (154 letters)
>gi|9629651|ref|NP_044945.1| replicase polyprotein [Drosophila C virus]
 gi|2388673|gb|AAC58807.1| replicase polyprotein [Drosophila C virus]
          Length = 1759

 Score = 42.4 bits (98), Expect = 0.018,   Method: Composition-based stats.
 Identities = 33/113 (29%), Positives = 57/113 (50%), Gaps = 13/113 (11%)

Query: 32  LYSRIQELAT-------LEQQLKLRLDPALISEIDKLFITSLNLSTGLRPSEARDHNNY- 83
           LY  IQ  A        L+Q+ K+ LD    + +++L+I  L   +   P  +++ +
Sbjct: 357 LYGEIQAWAQEVRQYLELDQRNKIDLDTETANRVEQLWIKGLKFKS--EPLLSKEMSALV 414

Query: 84  HATL---TRLHRACITSPARGAKMRVLPVQVQLFGDAGVGKTKLVSLLALTPL 133
           H TL    +L+     SP +G   R+ P+ + L G++GVGKT++V  L +  L
Sbjct: 415 HTTLLPAKQLYEYVSCSPVKGGGPRMRPICLWLVGESGVGKTEMVYPLCIDVL 467
```

```
Query= contig41498_2_1         (574 letters)

>gi|9629651|ref|NP_044945.1| replicase polyprotein [Drosophila C virus]
 gi|2388673|gb|AAC58807.1| replicase polyprotein [Drosophila C virus]
          Length = 1759

 Score = 40.4 bits (93), Expect = 0.69,   Method: Composition-based stats.
 Identities = 38/135 (28%), Positives = 58/135 (42%), Gaps = 26/135 (19%)

Query: 142 NNKIFLED----LNKFSTPCERQLHNVFVIAETPSEYPSWSTVEEQGRVIRFKNGCVIKH 197
           N  +F+ED    L   +   ++++   F   E+P E P +          RF    ++
Sbjct: 16  NKMMFVEDKISTLKMVADYYQKEVKYDFDAVESPREAPVFRCT------CRFLGYTIMTQ 69

Query: 198 FAAKNKIIAEIDTSFLLLLSGDVETNPGP----------HSKYCNEDKDREQR-RKMSL 245
              K     E     LLLLSGDVETNPGP             +Y   +K  E+R  K+
Sbjct: 70  GIGKKNPKQEAARQMLLLLSGDVETNPGPVQSRPVYYRYNDPRYTRLEKAIERRDDKIKT 129

Query: 246 MLKEISKLKREQMKH 260
           ++KE+    R Q+K+
Sbjct: 130 LIKEL----RRQIKN 140
```

(a)

Gryllus bimaculatus nudivirus, complete genome

gi|134303398|ref|NC_009240.1|



(b)

(c)

Figure 11. Complications in virus identification. (a) Alignments of three ORFs to different parts of the same protein. (b) Two ORFs match to different parts of the *Gryllus bimaculatus nudivirus* genome. (c) Full-length genome sequencing and closure should help to distinguish sequences from closely related virus species.

Conclusion

In this thesis, I describe my work on two different projects, which dealt with different classes of viruses – bacteriophages in one case, and animal viruses in the other. The unifying theme in both of these cases was the attention to remote sequence similarities, which allowed us to expand the range of detection of novel virus proteins in sequence databases and in freshly sequenced metagenomic libraries.

In the first project, we used whole genome information to classify phage genes into orthologous families, which we hope will become a useful resource for phage phylogeny studies and phage molecular biology. For the highly diverse virus genomes which do not share any universal genes, the traditional marker gene methods for cellular organisms cannot be applied. The Phage Orthologous Groups can be used to assert the relationship between phage genomes and to annotate phage genomes. POGs will also help to identify unknown phage genes from metagenomics data.

In our studies, we found some differences between double-stranded DNA phage genomes, the kind we studied, and genomes of cellular organisms. Unlike eukaryotic genomes which have many multi-domain proteins, only a very small portion of our studied phage genes encode multi-domain proteins. As a result, examining phage proteins in terms of domains affects POGs to a lesser extent than in the same case involving cellular organisms. In addition, POGs contain few paralogs. The relatively simple structure of the POGs shows that intra-genomic duplication is not a frequent event in phage evolutionary history.

Phages can transfer genes from one cellular host to another. However, our results show that about half of the known conserved protein families encoded by phage genomes rarely show up in the genomes of cellular hosts. These phage-only POGs contain three to thirty-two proteins

each, most of which have functions associated with virion structure and assembly, though some are enzymes involved in phage genome maintenance.

Over the years, as additional complete phage genomes have become available, new POGs have been created from the proteins discovered in these genomes. Also, the discovery of these proteins has allowed some known proteins to be grouped into new or existing POGs since their orthologous relationship to other proteins is now known. In this case, as more complete genomes become available in the future, a larger and larger proportion of their proteins should be matched to their orthologs to form new POGs. Our results (Figure 6b) indicate a trend: over time, the proportion of phage genes that have been assigned to POGs has grown to about half. We would like to know that as enough complete genomes become available, will it be possible that one day most phage proteins will find their orthologous partners and group into a POG, or if there are some phage proteins that could never find their own orthologs to form a POG.

The second project was the discovery of virus-related sequences in metagenomic samples from human patients and samples from diverse environmental sources. Although we were given the sequence libraries after the "low-hanging fruit", i.e., obviously high matches to the sequence databases, were recognized and removed, we still reported 411 proteins that have homologs to eukaryotic virus proteins in databases, and found many other ORFs that aligned to bacterial and phage proteins (not discussed here).

In this project, we applied the SEG algorithm, which is typically used with BLAST before using Newbler to assemble the reads. By removing low complexity sequences at the amino acid level, we were able to assemble the short reads into longer contigs, which could not be done previously.

In the sequence similarity searches, we used sensitive methods to look for remote homologs, eschewing an artificial e-value cutoff. We also manually checked every possible result to look for additional evidence, such as short conserved motifs of known function. Our results suggest that an across-the-board e-value cutoff cannot be recommended for viral proteins which are subjected to a greater mutation rate during evolutionary history.

The methods we used to identify new virus proteins from the metagenomic samples may be used on other metagenomic studies. In addition, the POGs can be used as a resource to identify new phage proteins. With more phages being identified and sequenced in the future, more POGs will be formed, telling us even more about phage genomes.

References

1. Zuckerkandl, E. and L. Pauling, *Molecules as documents of evolutionary history.* J Theor Biol, 1965. **8**(2): p. 357-66.
2. Zuckerkandl, E., R.T. Jones, and L. Pauling, *A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis.* Proc Natl Acad Sci U S A, 1960. **46**(10): p. 1349-60.
3. Dayhoff MO., E.R., *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD, 1968. **3**.
4. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
5. Henikoff, S. and J.G. Henikoff, *Performance evaluation of amino acid substitution matrices.* Proteins, 1993. **17**(1): p. 49-61.
6. Karlin, S. and S.F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.* Proc Natl Acad Sci U S A, 1990. **87**(6): p. 2264-8.
7. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
8. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
9. Koonin, E.V. and M.Y. Galperin, in *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*2003: Boston.
10. Fitch, W.M., *Distinguishing homologous from analogous proteins.* Syst Zool, 1970. **19**(2): p. 99-113.
11. Hachiya, T., et al., *Accurate identification of orthologous segments among multiple genomes.* Bioinformatics, 2009. **25**(7): p. 853-60.
12. Lynch, M. and V. Katju, *The altered evolutionary trajectories of gene duplicates.* Trends Genet, 2004. **20**(11): p. 544-9.
13. Ohno, S., *Evolution by Gene Duplication*1970, New York: Springer-Verlag.
14. Park, D., et al., *IsoBase: a database of functionally related proteins across PPI networks.* Nucleic Acids Res, 2011. **39**(Database issue): p. D295-300.
15. Hulsen, T., et al., *Benchmarking ortholog identification methods using functional genomics data.* Genome Biol, 2006. **7**(4): p. R31.
16. Eisen, J.A., *Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.* Genome Res, 1998. **8**(3): p. 163-7.
17. Gabaldon, T., et al., *Joining forces in the quest for orthologs.* Genome Biol, 2009. **10**(9): p. 403.
18. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
19. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.* Mol Biol Evol, 2009. **26**(7): p. 1641-50.
20. Liu, K., et al., *Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees.* Science, 2009. **324**(5934): p. 1561-4.
21. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol, 2003. **52**(5): p. 696-704.

22. Hughes, A.L. and R. Friedman, *Differential loss of ancestral gene families as a source of genomic divergence in animals.* Proc Biol Sci, 2004. **271 Suppl 3**: p. S107-9.
23. Gout, J.F., L. Duret, and D. Kahn, *Differential retention of metabolic genes following whole-genome duplication.* Mol Biol Evol, 2009. **26**(5): p. 1067-72.
24. Vilella, A.J., et al., *EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.* Genome Res, 2009. **19**(2): p. 327-35.
25. Felsenstein, J., *Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.* Methods Enzymol, 1996. **266**: p. 418-27.
26. Hahn, M.W., *Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution.* Genome Biol, 2007. **8**(7): p. R141.
27. Liu, K., C.R. Linder, and T. Warnow, *Multiple sequence alignment: a major challenge to large-scale phylogenetics.* PLoS Curr, 2010. **2**: p. RRN1198.
28. Thorne, J.L. and H. Kishino, *Freeing phylogenies from artifacts of alignment.* Mol Biol Evol, 1992. **9**(6): p. 1148-62.
29. Schliep, K., et al., *Harvesting evolutionary signals in a forest of prokaryotic gene trees.* Mol Biol Evol, 2011. **28**(4): p. 1393-405.
30. Olendzenski, L. and J.P. Gogarten, *Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer.* Ann N Y Acad Sci, 2009. **1178**: p. 137-45.
31. Koonin, E.V. and Y.I. Wolf, *The fundamental units, processes and patterns of evolution, and the tree of life conundrum.* Biol Direct, 2009. **4**: p. 33.
32. Bapteste, E., et al., *Prokaryotic evolution and the tree of life are two different things.* Biol Direct, 2009. **4**: p. 34.
33. Treangen, T.J. and E.P. Rocha, *Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes.* PLoS Genet, 2011. **7**(1): p. e1001284.
34. Puigbo, P., Y.I. Wolf, and E.V. Koonin, *The tree and net components of prokaryote evolution.* Genome Biol Evol, 2010. **2**: p. 745-56.
35. Dagan, T., Y. Artzy-Randrup, and W. Martin, *Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution.* Proc Natl Acad Sci U S A, 2008. **105**(29): p. 10039-44.
36. Gogarten, J.P. and J.P. Townsend, *Horizontal gene transfer, genome innovation and evolution.* Nat Rev Microbiol, 2005. **3**(9): p. 679-87.
37. Boucher, Y., et al., *Lateral gene transfer and the origins of prokaryotic groups.* Annu Rev Genet, 2003. **37**: p. 283-328.
38. Gogarten, J.P., W.F. Doolittle, and J.G. Lawrence, *Prokaryotic evolution in light of gene transfer.* Mol Biol Evol, 2002. **19**(12): p. 2226-38.
39. Koonin, E.V., K.S. Makarova, and L. Aravind, *Horizontal gene transfer in prokaryotes: quantification and classification.* Annu Rev Microbiol, 2001. **55**: p. 709-42.
40. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.
41. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.
42. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
43. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.
44. Suttle, C.A., *Viruses in the sea.* Nature, 2005. **437**(7057): p. 356-61.

45. Wolf, Y.I., et al., *The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages.* Proc Natl Acad Sci U S A, 2009. **106**(18): p. 7273-80.
46. Koonin, E.V. and Y.I. Wolf, *Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.* Nucleic Acids Res, 2008. **36**(21): p. 6688-719.
47. *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.
48. Koonin, E.V., *Evolution of genome architecture.* Int J Biochem Cell Biol, 2009. **41**(2): p. 298-306.
49. Koonin, E.V. and Y.I. Wolf, *Constraints and plasticity in genome and molecular-phenome evolution.* Nat Rev Genet, 2010. **11**(7): p. 487-98.
50. Jun, J., Mandoiu, II, and C.E. Nelson, *Identification of mammalian orthologs using local synteny.* BMC Genomics, 2009. **10**: p. 630.
51. Zdobnov, E.M. and P. Bork, *Quantification of insect genome divergence.* Trends Genet, 2007. **23**(1): p. 16-20.
52. Zdobnov, E.M., et al., *Consistency of genome-based methods in measuring Metazoan evolution.* FEBS Lett, 2005. **579**(15): p. 3355-61.
53. Wootton, J.C. and S. Federhen, *Analysis of compositionally biased regions in sequence databases.* Methods Enzymol, 1996. **266**: p. 554-71.
54. Wootton, J.C., *Non-globular domains in protein sequences: automated segmentation using complexity measures.* Comput Chem, 1994. **18**(3): p. 269-85.
55. Bourhis, J.M., et al., *The intrinsically disordered C-terminal domain of the measles virus nucleoprotein interacts with the C-terminal domain of the phosphoprotein via two distinct sites and remains predominantly unfolded.* Protein Sci, 2005. **14**(8): p. 1975-92.
56. Jones, D.T. and M.B. Swindells, *Getting the most from PSI-BLAST.* Trends Biochem Sci, 2002. **27**(3): p. 161-4.
57. Soding, J., M. Remmert, and A. Biegert, *HHrep: de novo protein repeat detection and the origin of TIM barrels.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W137-42.
58. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
59. Park, J., et al., *Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.* J Mol Biol, 1998. **284**(4): p. 1201-10.
60. Ramakrishnan, C., V.S. Dani, and T. Ramasarma, *A conformational analysis of Walker motif A [GXXXXGKT (S)] in nucleotide-binding and other proteins.* Protein Eng, 2002. **15**(10): p. 783-98.
61. Edwards, R.A. and F. Rohwer, *Viral metagenomics.* Nat Rev Microbiol, 2005. **3**(6): p. 504-10.
62. Whitman, W.B., D.C. Coleman, and W.J. Wiebe, *Prokaryotes: the unseen majority.* Proc Natl Acad Sci U S A, 1998. **95**(12): p. 6578-83.
63. Fuhrman, J.A., *Marine viruses and their biogeochemical and ecological effects.* Nature, 1999. **399**(6736): p. 541-8.
64. Bergh, O., et al., *High abundance of viruses found in aquatic environments.* Nature, 1989. **340**(6233): p. 467-8.
65. Paul, J.H., et al., *Distribution of viral abundance in the reef environment of Key Largo, Florida.* Appl Environ Microbiol, 1993. **59**(3): p. 718-24.

66. Hendrix, R.W., *Bacteriophages: evolution of the majority.* Theor Popul Biol, 2002. **61**(4): p. 471-80.
67. Koonin, E.V., T.G. Senkevich, and V.V. Dolja, *The ancient Virus World and evolution of cells.* Biol Direct, 2006. **1**: p. 29.
68. McDaniel, L.D., et al., *High frequency of horizontal gene transfer in the oceans.* Science, 2010. **330**(6000): p. 50.
69. Kristensen, D.M., et al., *New dimensions of the virus world discovered through metagenomics.* Trends Microbiol, 2010. **18**(1): p. 11-9.
70. Ackermann, H.W., Dubow, M. S., *Viruses of Prokarytoes Vol. 1, General Properties of Bacteriophages.* CRC, Boca Raton, 1987.
71. Harris, T.D., et al., *Single-molecule DNA sequencing of a viral genome.* Science, 2008. **320**(5872): p. 106-9.
72. Braslavsky, I., et al., *Sequence information can be obtained from single DNA molecules.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3960-4.
73. Pushkarev, D., N.F. Neff, and S.R. Quake, *Single-molecule sequencing of an individual human genome.* Nat Biotechnol, 2009. **27**(9): p. 847-50.
74. d'Hérelle, F., *Sur un microbe invisible antagoniste des bacilles dysentériques.* C. R. Acad. Sci. Ser. D, 1917. **164**: p. 373-375.
75. Twort, F.W., *An investigation on the nature of the ultra-microscopic viruses.* Lancet, 1915. **11**: p. 1241-1243.
76. John Cairns, G.S.S., James D. Watson, ed. *Phage and the Origins of Molecular Biology, The Centennial Edition*. 2007, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
77. Sanger, F., et al., *Nucleotide sequence of bacteriophage phi X174 DNA.* Nature, 1977. **265**(5596): p. 687-95.
78. Liu, J., G. Glazko, and A. Mushegian, *Protein repertoire of double-stranded DNA bacteriophages.* Virus Res, 2006. **117**(1): p. 68-80.
79. Besemer, J., A. Lomsadze, and M. Borodovsky, *GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.* Nucleic Acids Res, 2001. **29**(12): p. 2607-18.
80. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.
81. Kristensen, D.M., et al., *A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches.* Bioinformatics, 2010. **26**(12): p. 1481-7.
82. Marchler-Bauer, A., et al., *CDD: specific functional annotation with the Conserved Domain Database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D205-10.
83. Apic, G., J. Gough, and S.A. Teichmann, *Domain combinations in archaeal, eubacterial and eukaryotic proteomes.* J Mol Biol, 2001. **310**(2): p. 311-25.
84. Davison, J., *Genetic exchange between bacteria in the environment.* Plasmid, 1999. **42**(2): p. 73-91.
85. Sinkovics, J.G., *Horizontal gene transfers and cell fusions in microbiology, immunology and oncology (Review).* Int J Oncol, 2009. **35**(3): p. 441-65.
86. Bist, P., et al., *S-adenosyl-L-methionine is required for DNA cleavage by type III restriction enzymes.* J Mol Biol, 2001. **310**(1): p. 93-109.
87. Roche, *Metagenomics and Microbial Diversity Flier.*

88. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins.* Nature, 2009. **457**(7228): p. 480-4.
89. Smit, A., Hubley, R & Green, P. , *RepeatMasker Open-3.0.* http://www.repeatmasker.org, 1996-2010.

Appendix A: the annotation of POGs-2007

name	function	comments	
POG1	gp28 baseplate hub distal subunit	679, 678, 819	POG 678 doesn't have a good probability (88) but nealy full length, POG 819 with extremely low score (29.5) but nearly full length
POG2	lipoprotein	680	
POG3	Arn.4 conserved hypothetical protein	960	
POG4	dNMP kinase	681	
POG5	gp10 baseplate wedge subunit and tail pin	911, 805, 912	805 matches about 1/3 of the length
POG6	baseplate wedge subunit and tail pin	806	
POG7	gp12 short tail fibers	912	
POG8	gp13 neck protein	736	
POG9	gp14 neck protein	808	
POG10	gp15 tail sheath stabilizer and completion protein	809	
POG11	gp16 terminase DNA packaging enzyme small subunit	810	
POG12	terminase large subunit	901, 5, 47, 240, 66, 646	all old pogs connect to terminase large subunit
POG13	tail sheath protein	733/317	Matches to tail lysozyme, but not good alignment and short length.
POG14	gp19 tail tube protein	732, 796, 844	maybe a homology for Hemolysin-coregulated protein
POG15	gp20 portal vertex protein of head	734	This group of peptidases belongs to MEROPS peptidase family U9 (phage prohead processing peptidase family, clan U-), which play a role in the head assembly of bacteriophage T4.
POG16	gp21 prohead core scaffold protein and protease	677	
POG17	gp22 prohead core scaffold protein	812	
POG18	gp23 major head protein	738, 739, 898	
POG19	gp24 head vertex protein	738/739	
POG20	RnlB RNA ligase 2	813	
POG21	gp24.2 conserved hypothetical protein	957	
POG22	gp24.3 conserved hypothetical protein	814 /// they all matches 'DNA primase helicase subunit' but short length	
POG23	GPW_gp25 baseplate	313/631	connect to POG 14.

	wedge subunit, tail lysozyme, putative		
POG24	gp26 baseplate hub subunit	678/819	
POG25	gp27 baseplate hub subunit	820	triple-stranded beta-helix, OB fold, pseudohexamer, T4 tail lysozyme
POG26	gp29 baseplate hub subunit tail length determinator	821	
POG27	gp2 DNA end protector protein	797	
POG28	gp3 tail completion and sheath stabilizer protein	796/732/844	
POG29	DNA ligase N-terminal domain	342	link to POG30, 1417
POG30	ATP dependent DNA ligase domain, catalytic domain	342	link to POG29, 1417
POG31	gp30.1 conserved hypothetical protein		
POG32	5' nucleotidase	916	psiblast finds only 'gp30.2 conserved hypothetical protein'; but HHsearch finds 5' nucleotidase, also matches hydrolase, phosphotase (not as good as nucleotidase matches)
POG33	gp30.3 conserved hypothetical protein	825	
POG34	gp30.4 conserved hypothetical protein		
POG35	gp30.6 conserved hypothetical protein	826	
POG36	gp30.7 conserved hypothetical protein		
POG37	gp30.8 conserved hypothetical protein		
POG38	gp30.9 conserved hypothetical protein	958	
POG39	gp31 head assembly cochaperone with GroEL	828	
POG40	gp31.1 conserved hypothetical protein	829	
POG41	gp32 single-stranded DNA binding protein	833	
POG42	gp33 late promoter transcription accessory protein	835	not finish yet. antitoxin
POG43	tail fiber protein	913/254/914	
POG44	gp35 hinge connector of long tail fiber, may contain C-terminal carbohydrate binding domain	837	
POG45	hinge connector of long tail	987/289/254	

	fiber distal connector; it may contains inner repeats.		
POG46	tail fiber assembly protein	425, 316, 398	
POG47	topoisomerase II large subunit	766	large subunit, ATPase
POG48	gp39.1 hypothetical protein		
POG49	gp39.2 conserved hypothetical protein	767	Connection to FmdB transcriptional regulator family
POG50	gp40 head vertex assembly chaperone	772	
POG51	gp41 DNA primase-helicase subunit, N terminal	650	
POG52	DNA primase/helicase, ATPase domain	650	? One protein matches thymidylate synthase, others matches hydroxymethylase
POG53	gp42 dCMP hydroxymethylase / thymidylate synthase	674	
POG54	DNA_polB, catalitic domain	406	
POG55	gp44 clamp loader subunit, DNA polymerase accessory protein	762	
POG56	gp45 sliding clamp DNA polymerase accessory protein, N terminal	775	different domains
POG57	gp45 sliding clamp DNA polymerase accessory protein, C terminal, adopt a ATPase clamp flold	775	different domains
POG58	gp45.2 conserved hypothetical protein	952	
POG59	gp46 recombination endonuclease subunit, P-loop ATPase	686	link to POG59. some are exonuclease? Or ABC type transport system
POG60	gp46.1 hypothetical protein		
POG61	gp46.2 hypothetical protein		
POG62	gp47 recombination endonuclease subunit, N terminal	759	needs to manually split domains. Link to POG63 and 1529; some are exonuclease
POG63	gp47 recombination endonuclease subunit, C terminal	759	needs to manually join with POG1529 and some domains from POG62. Link to POG62 and 1529;
POG64	gp48 baseplate tail tube cap	822	
POG65	Nuclease	98, one relates to 207	Most are endonuclease, several are (putative) exonuclease
POG66	gp49.1 conserved protein of unknown function		
POG67	gp4 head completion protein	798	homologous with TnsA endonuclease, N-terminal domain which is catalytic
POG68	lysozyme	65/933/934/588/311	

POG69	baseplate hub subunit and tail lysozyme	933,15	
POG70	gp5.1 conserved hypothetical protein	800	
POG71	gp5.4 conserved hypothetical protein	801	PAAR motif
POG72	gp51 baseplate hub assembly catalyst	819/678	
POG73	gp52 topo II medium subunit	841 DNA gyrase/topoisomerase small subunit	
POG74	gp53 baseplate wedge subunit	799	
POG75	gp54 baseplate tail tube initiator	844, 732, 796	
POG76	gp55 Sigma factor for T4 late transcription	778	
POG77	gp55.1 hypothetical protein	178	
POG78	gp55.2 conserved hypothetical protein	779	Related to antirepressor and Middle operon regulator, Mor, may have DNA binding activity
POG79	gp55.3 conserved hypothetical protein	780	
POG80	gp55.4 conserved hypothetical protein	781	
POG81	gp55.5 conserved protein of unknown function		
POG82	gp55.6 conserved hypothetical protein		
POG83	gp55.8 conserved hypothetical predicted membrane protein		
POG84	gp56 dCTPase	770	
POG85	gp57B conserved hypothetical protein	795	
POG86	gp61 DNA primase subunit, zinc finger	196864	HHsearch: pfam01807:zf-CHC2 CHC2 zinc fi; smart00400:ZnF_CHCC zinc finge
POG87	gp61 DNA primase subunit, catalytic core, contains a Toprim-N domain	196	
POG88	gp61 DNA primase subunit, TOPRIM_DnaG_primases	196	
POG89	gp59 loader of gene 41 DNA helicase	834	
POG90	gp6 baseplate wedge subunit,	802/382/314	strange POGs with 91
POG91	gp6 baseplate wedge subunit	802	
POG92	gp61.1 hypothetical	738/898	connected to gp23 major head protein in

	protein; Major capsid protein		T4-like phages
POG93	gp61.2 hypothetical protein		hhsearch doesn't work well in this example because only find 3 sequences in psiblast and only 5 in hhsensor.
POG94	sp spackle periplasmic protein		
POG95	gp61.4 hypothetical protein		
POG96	Dmd discriminator of mRNA degradation	771	
POG97	clamp loader subunit DNA polymerase accessory protein	774	
POG98	RNA ligase, catalytic domain, contains motif KXDGSL	737	Check POG98 and 99's domains, need to separate some domains in 98 manuallly.
POG99	RnlA RNA ligase 1 and tail fiber attachment catalyst	737	Can I call it C-terminal domain?
POG100	gp67 prohead core protein		
POG101	gp68 prohead core protein, involved in head size determination	811	
POG102	gp7 baseplate wedge initiator	803	
POG103	gp8 base plate wedge component	804	
POG104	gp9 baseplate wedge tail fiber connector	805/911	
POG105	UvsY.-1 conserved hypothetical protein		
POG106	UvsY.-2 conserved hypothetical protein	817	
POG107	gp52.1 conserved hypothetical predicted membrane protein		
POG108	a-gt.3 hypothetical protein		
POG109	a-gt.5 conserved hypothetical protein		
POG110	Alc inhibitor of host transcription	831	Aligns to Sp38, Zona-pellucida-binding protein. Weird.
POG111	Alt RNA polymerase ADP-ribosylase	823	unknown domain
POG112	Alt RNA polymerase ADP-ribosylase	823	matches to VIP2 (vegetative insectcidal protein
POG113	Alt.-3 conserved hypothetical protein		
POG114	Alt.1 conserved hypothetical protein		
POG115	AsiA anti-sigma 70 protein	839	
POG116	Arn.1 conserved hypothetical protein		
POG117	Arn.2 conserved		

	hypothetical protein		
POG118	Arn.3 conserved hypothetical protein	675	
POG119	deoxycytidinylate deaminase; Cd dCMP deaminase	684	
POG120	Cd.2 conserved hypothetical protein	959	
POG121	Cd.3 conserved hypothetical protein		aligns to PF06287 DUF1039, which encoded in pathogenicity islands for bacterail type III secretion systems in various strains, but has not yet to have been characterised as part of the apparatus or as an effector protein.
POG122	goF mRNA metabolism modulator	949	
POG123	DNA adenine methyltransferase	337	
POG124	exonuclease; RNaseH	351	
POG125	DNA helicase, P loop NTPase superfamily	294	check relationship between 125, 126, and 1629
POG126	Ti-type conjugative transfer relaxase TraA	294	
POG127	Dda.1 hypothetical protein		
POG128	putative Srd anti-sigma factor	768	
POG129	DenA endonuclease II	832	GIY-YIG GIY-YIG catalytic domain
POG130	DenB DNA endonuclease IV		
POG131	DenV endonuclease V, N-glycosylase UV repair enzyme	792	belongs to Pyrimidine dimer DNA glycosylase (CDD)
POG132	exonuclease A	593/ 665	
POG133	dsDNA binding protein	836	
POG134	nudix hydrolase	608	
POG135	e.3 conserved hypothetical		predicted membrane protein
POG136	e.4 conserved hypothetical		
POG137	e.5 conserved hypothetical protein		
POG138	e.6 conserved hypothetical protein	794	Hhsearch found PspA_IM30 family.
POG139	e.8 conserved hypothetical protein		
POG140	Frd dihydrofolate reductase	672	
POG141	Frd.1 conserved hypothetical protein		
POG142	Frd.2 conserved hypothetical protein		Bacteriophoge FRD2 protein
POG143	Frd.3 hypothetical protein		Bacteriophoge FRD3 protein
POG144	Hoc head outer capsid	905	

	protein		
POG145	Imm immunity to superinfection membrane protein		
POG146	Imm.1 hypothetical predicted membrane protein		
POG147	Inh inhibitor of prohead protease gp21	815/905	
POG148	cef modifier of suppressor tRNAs	950	
POG149	HNH nuclease	25/729	The c terminal may align to group I intron endonuclease.
POG150	MobD.1 conserved hypothetical protein		
POG151	MobD.2 conserved hypothetical protein		
POG152	hypothetical protein	845	Bacterioferritin
POG153	rl.1 conserved hypothetical protein	954	Ribosomal protein L29P
POG154	ADP-ribosylase	769	
POG155	ModA.2 hypothetical protein		
POG156	MotA activator of middle period transcription	840	
POG157	MotB modifier of transcription		
POG158	MotB.2 hypothetical protein		
POG159	Mrh.2 hypothetical protein	951	Raf-like Ras-binding domain, or RNA_pol_Rbp7_N RNA polymerase Rpb7, N-terminal domain
POG160	Ndd nucleoid disruption protein	961	
POG161	Ndd.1 conserved hypothetical protein		
POG162	aerobic ribonucleoside diphosphate reductase, diphosphate or triphosphate; ATP cone domain		
POG163	ribonucleotide reductase, RNR_PFL super-family	408	
POG164	NrdA.1 conserved hypothetical protein	456	
POG165	NrdA.2 conserved hypothetical protein		
POG166	NrdB aerobic NDP reductase	676	**Ferritin_like Super-family**
POG167	NrdC thioredoxin	785	contain a redox active CXXC motif in a TRX fold
POG168	NrdC.10 conserved	787	AAA domain of cell division protein FtsH

	hypothetical protein		
POG169	NrdC.11 conserved hypothetical protein	728	
POG170	NrdC.2 conserved hypothetical protein		might be DNA binding protein
POG171	NrdC.3 conserved hypothetical protein		
POG172	NrdD anaerobic NTP reductase large subunit	784	
POG173	NrdG anaerobic NTP reductase small subunit	783	
POG174	NrdH glutaredoxin	782	
POG175	NrdC.7 conserved hypothetical		
POG176	NrdC.8 conserved hypothetical protein		
POG177	NrdC.9 conserved hypothetical protein		
POG178	Pin protease inhibitor		
POG179	PseT polynucleotide 5'-kinase and 3'-phosphatase	707	P-loop NTPase superfamily
POG180	PseT polynucleotide 5'-kinase and 3'-phosphatase	707	acid phosphotase
POG181	PseT.1 conserved hypothetical protein		
POG182	rI lysis inhibition regulator membrane protein	953	
POG183	rIIA protector from prophage-induced early lysis, N terminal, APTase	765	Histidine kinase like ATPase
POG184	rIIA protector from prophage-induced early lysis, C terminal	765	
POG185	rIIB protector from prophage-induced early lysis, N terminal	842	one protein 116326223 has three domains associating with 183 and 184 instead of 186. pretty interesting.
POG186	rIIB protector from prophage-induced early lysis, C terminal	842	
POG187	rIII lysis inhibition accessory protein	827	
POG188	RegA translational repressor protein	773	
POG189	RegB site-specific RNA endonuclease	758	
POG190	RpbA RNA polymerase binding protein	776	
POG191	SegC homing endonuclease	832	GIY-YIG endonuclease [range 1 .. 109] HHsearch against CDD matches: pfam01541:GIY-YIG GIY-YIG cata; smart00465:GIYc GIY-YIG type n
POG192	Soc small outer capsid		

	protein		
POG193	Stp activator of host PrrC lysyl-tRNA endonuclease	too short	
POG194	t holin lysis mediator	838	
POG195	dTMP (thymidylate) synthase	674	
POG196	thymidine kinase	735	
POG197	Tk.2 conserved hypothetical protein		GvpK, Gas Vesicle protein K?
POG198	Macro_H2A_like Marcro domain, a high-affinity ADP-ribose binding module	788	
POG199	UvsW RNA-DNA and DNA-DNA helicase		
POG200	helicase UvsW	295	P-loop NTPase superfamily
POG201	helicase UvsW	295	P-loop NTPase superfamily
POG202	UvsW.1 conserved hypothetical protein	816	9632837-4 matches helicase in psiblast while matches helicase and UvsW.1 in Hhsearch.
POG203	Vs valyl-tRNA synthetase modifier	955	
POG204	Vs.1 conserved hypothetical protein	789	most hhsearch results have hit ' transglycosylase SLT domain'
POG205	Vs.4 conserved hypothetical protein		might relate to Poxvirus All protein because of the presence of DLEXXXXEL/ID motif. However, Pox_all protein has a consevered whole length domain
POG206	Vs.6 hypothetical protein, formate acetyltransferase 2	790	Pyruvate-formate lyase; actually only aligns to C terminal, but contains active sites SGY
POG207	Vs.7 conserved hypothetical protein		might relate to RCC_reductase because of EFI motif
POG208	Vs.8 conserved hypothetical protein	791	Terminase, DNA packaging protein GP17, nucleotide-biding fold, but two proteins 33620488 and 157311402 don't align to terminase, instead, they align to Gnd 6-phosphogluconate dehydrogenase
POG209	Wac fibritin neck whiskers	807	
POG210	Trna.2 conserved hypothetical protein		
POG211	Trna.3 conserved hypothetical protein		
POG212	Trna.4 conserved hypothetical predicted membrane protein		
POG213	RecA-like recombination protein / recombinase A	709	
POG214	UvsY recombination repair and ssDNA binding protein	818	
POG215	RB69ORF029w hypothetical protein		
POG216	hypothetical protein		
POG217	Replication Protein GPA	373	

POG218	putative structural protein	755	Amidase_4 Mannosyl-glycoprotein endo-beta-N-acetylglucosamidase
POG219	Hypothetical protein	436	
POG220	Hypothetical protein	437/658	
POG221	Hypothetic protein / Apaf-1 related killer DARK	441	
POG222	Endonuclease	25	Hhsearch found 'endonuclease'; while psiblase found mostly hypothetic
POG223	putative terminase small subunit	134	
POG224	putative terminase large subunit	5/901	
POG225	putative head-tail joining protein	180	
POG226	(putative) portal protein	6	
POG227	major head protein, N terminal	135	
POG228	major head protein	7/9/135/335	
POG229	putative DNA packaging protein	181	phage capsid family
POG230	putative head-tail joining protein	182	
POG231	(putative) tail component protein	183/570	
POG232	major tail protein	184/140	
POG233	putative tail component	184	
POG234	(putative) phage related minor tail protein	903	
POG235	tail length tape measure protein / transglycosylase SLT domain	910	NlpC/P60 family
POG236	tail length tape measure protein / transglycosylase SLT domain	910	lysozyme_like domain
POG237	putative tail protein	186	
POG238	structure protein, tail-host specificity protein; PblB Super-family, phage related protein	900/50/922/141	
POG239	tail-host specificity protein, Collagen triple helix repeat	50/919	
POG240	putative minor structural protein	270	
POG241	Hypothetical protein	291	
POG242	cl01989: phage_holin_4 super-family	566/62	
POG243	cl0344: Phage_holin_1 super-family	216	
POG244	lysin, cl11438: NLPC_P60 super-family	174/63	function unknown, but found in several lipoproteins
POG245	lysin, C-terminal	174	
POG246	Hypothetical protein	449	Phage_Gp111, Streptococcus

			thermophilus bacteriophage Gp111 protein
POG247	Hypothetical protein	895	
POG248	repressor	937/112/109/894	DNA binding domain, possible cro, CI-like, etc
POG249	Sipho_Gp157	255	Sipho_Gp157 It is thought that bacteria possessing the gene coding for this protein have an increased resistance to the bacteriophage
POG250	NTP-binding motif protein	256/326	
POG251	putative helicase		helicase UvsW, C-terminal
POG252	hypothetical protein, possible single strand DNA binding protein	931	Hhsearch supports SSB
POG253	putative replication protein	195	
POG254	(putative) primase	196/195	
POG255	Hypothetical protein	589	
POG256	hypothetical protein	28/661	VRR_NUC superfamily
POG257	Hypothetical protein	438	transcriptional repressor protein
POG258	Hypothetical protein	162/167	HTH?
POG259	Hypothetical protein	442	
POG260	Hypothetical protein	165	
POG261	DNA binding protein	112/109/894	huge pog
POG262	DNA binding domain	276/894/275	different function
POG263	hypothetical protein	439	
POG264	hypothetical protein	440	
POG265	(putative) ATP-dependent ClpP protein	26	
POG266	putative tail component protein	27	
POG267	unknown		short sequences
POG268	Hypothetical protein	194	
POG269	integrase	299	
POG270	hypothetical protein, similar to cI-like repressor	151/111	
POG271	anti repressor	113	
POG272	(putative) anti repressor	113	
POG273	putative DNA methyltransferase	945	small pog, short sequences
POG274	transcription factors	198	Because of short sequence, it may be DNA binding domain
POG275	hypothetical protein	874	
POG276	hypothetical protein	896/125	
POG277	hypothetical protein, putative primase	196/195	
POG278	hypothetical protein	261	
POG279	hypothetic protein	284/338	maybe ArpU family transcriptional regulator
POG280	hypothetical protein	18	
POG281	hypothetical protein	171/186/214/33/670	Hhsearch found so many old POGs?
POG282	hypothetical protein	334	short sequences

POG283	hypothetical protein	875	
POG284	lysin	65	
POG285	amidase	63	probably hydrolysis--lysin, amidase, peptidoglygan hydrolase
POG286	hypothetic protein, probable zinc finger	31	Hhsearhc found DnaK supressor but short length
POG287	putative excisionase		short length
POG288	hypothetical protein	522	
POG289	hypothetical protein	219	
POG290	hypothetical protein	447	
POG291	hypothetical protein	525	
POG292	hypothetical protein	526	
POG293	hypothetical protein	528	
POG294	hypothetical protein	529	
POG295	hypothetical protein	99	short length with 99
POG296	hypothetical protein	532	
POG297	hypothetical protein	533	
POG298	hypothetical protein	535	
POG299	lambada exonuclease	536	
POG300	Bet, RecT homolog	539	
POG301	host-nuclease inhibitor protein Gam	543	
POG302	Kil protein	222	
POG303	regulatory protein CIII	223	
POG304	ea10, putative ssDNA binding protein	544	
POG305	antitermination protein N	224	
POG306	cI-like repressor	112	
POG307	regulatory protein CII	226	
POG308	replication protein O	547	
POG309	DNA replication helicase, gpP, DnaB	650	
POG310	hypothetical protein	764	
POG311	hypothetical protein	36	
POG312	hypothetical protein	551	
POG313	Protein Nin B	228	
POG314	DNA adenine-methylase, dam	396/945	
POG315	NinE protein	229	short length
POG316	DNA binding protein Roi	113	
POG317	NinG	553	
POG318	NinH	34	
POG319	antitermination protein Q	409	
POG320	shiga-like toxin type II A subunit	410	
POG321	shiga-like toxin type II B subunit	411	
POG322	hypothetical protein	434	
POG323	hypothetical protein	434	
POG324	hypothetical protein	434	

POG325	hypothetical protein	558	
POG326	lysis protein	559	
POG327	anti-repressor protein Ant	236	
POG328	endopeptidase Rz	606	
POG329	hypothetic protein	563	
POG330	putative terminase small subunit	564	
POG331	putative phage portal protein	401/241	
POG332	putative phage portal protein	401/241	
POG333	hypothetical protein	460	
POG334	putative virion structural protein	403	
POG335	hypothetical protein	464	
POG336	hypothetical protein	466	
POG337	hypothetical protein	467	
POG338	hypothetical protein	404	
POG339	putative tail fiber protein	917/918	
POG340	hypothetical protein	22	
POG341	hypothetical protein	477	
POG342	hypothetical protein	402	
POG343	putative tail tip fiber protein	8	
POG344	hypothetical protein	481	
POG345	hypothetical protein	444/445	
POG346	outer membrane protein, Lom	21	
POG347	hypothetical protein	484	
POG348	hypothetical protein	486	
POG349	hypothetical protein	489	
POG350	hypothetical protein	492	
POG351	hypothetical protein	494/399	
POG352	hypothetical protein	512	
POG353	hypothetical protein	513	
POG354	hypothetical protein	514	host killer protein, small toxic membrane polypeptide
POG355	hypothetical protein	516	
POG356	hypothetical protein, eaa2 homolog	219	
POG357	hypothetical protein	640	
POG358	hypothetical protein	236	match putative antirepressor protein Ant with short length
POG359	hypothetical protein	227	
POG360	unknown	29	too short
POG361	DNA methylase	202	too short
POG362	unknown		
POG363	replication protein O	547/257	
POG364	replication protein P	548	
POG365	hypothetical protein	235	
POG366	Lipoprotein Rz1 protein	562	

	precursor		
POG367	putative tail fiber assembly protein	425/316	
POG368	hypothetical protein		
POG369	Middle operon regulator, MOR		
POG370	terminase large subunit	901/5	
POG371	Hypothetical protein	983	
POG372	phage_Mu_F phage Mu protein F like protein, virion morphogenesis protein, possiblely a minor head protein.	205/607/580	
POG373	virion morphogenesis protein		
POG374	Phage virion morphogenesis family, Phage_tail_S, involved in tail completion and stable head joining	377/11	
POG375	putative protease (I) and scaffold (Z) protein	300	
POG376	major head subunit gpT		
POG377	Hypothetical protein		
POG378	Hypothetical protein	452	
POG379	tail sheath protein	417	
POG380	tail/DNA circulation protein	420	
POG381	putative tail protein	421/320	
POG382	putative baseplate assembly protein	312	
POG383	putative tail protein	423	
POG384	putative tail protein, baseplate J-like protein	382/802/314	
POG385	putative tail protein	424	
POG386	tail assembly protein	425	
POG387	resolvase	88	one type of site specific recombinase
POG388	virulence-associated E family protein	200	
POG389	holin	233	
POG390	terminase large subunit	240/5/47	
POG391	portal protein	241/401/204/355	
POG392	scaffold protein	242	
POG393	capsid protein	155/357	
POG394	DNA stabilization protein	243	
POG395	DNA stabilization protein	244	
POG396	packaged DNA stabilization protein gp26	245	
POG397	DNA transfer protein	247	
POG398	DNA transfer protein	248	
POG399	DNA transfer protein	249	

POG400	tail fiber assembly protein	425/316	
POG401	Xis, excisionase	450/451/902	
POG402	(putative) helicase	133	SNF2 family N-terminal domain
POG403	(putative) helicase	133	
POG404	antirepressor	113	
POG405	(putative) DNA polymerase	90	
POG406	DNA polymerase	90	
POG407	hypothetical protein	662	
POG408	hypothetical protein	610/666/102	
POG409	S-adenosyl-L-methionine hydrolase		
POG410	protein kinase	339	
POG411	RNA polymerase	340	
POG412	RNA polymerase	340	
POG413	Bacteriophage gene 1.1 protein		
POG414	deoxyguanosine triphospho-hydrolase inhibitor	341	
POG415	DNA ligase, C terminal OB-fold.	342	
POG416	(putative) HNH homing endonuclease	273/715	
POG417	Hypothetical protein	343	
POG418	hypothetical protein	282	
POG419	bacterial RNA polymerase inhibitor	344	
POG420	single-stranded DNA-binding protein	345/662	
POG421	endonuclease	935	
POG422	N-acetylmuramoyl-L-alanine amidase	346	
POG423	Putative phage DNA primase/helicase	932	
POG424	T7 gene 4.2	347	
POG425	gene 4.3 protein	348	
POG426	Phage T7 superfamily gene 4.5 protein	349	
POG427	DNA polymerase	406/90	
POG428	5.3 protein, homing endonuclease	273	
POG429	Phage T7 superfamily gene 5.5 protein	350	
POG430	Phage T7 superfamily gene 5.7 protein	292	
POG431	Phage T7 superfamily gene 6.5 protein	352	
POG432	Phage T7 superfamily gene 6.7 protein	353	
POG433	Phage T7 superfamily gene 7.3 protein	354	

POG434	head-to-tail joining protein, portal protein	355	
POG435	capsid assembly protein	356	
POG436	Capsid assembly protein	356	scaffolding protein
POG437	capsid protein	357	
POG438	tail tubular protein	358/394	
POG439	tail tubular protein B	359	
POG440	internal virion protein A	360	
POG441	internal virion protein B	361	
POG442	internal virion protein C	362	
POG443	transglycosylase	910/363/789	some also match internal virion protein D
POG444	internal virion protein D	363	
POG445	internal virion protein D	363	
POG446	internal virion protein D	363/755	
POG447	tail fiber protein	925	
POG448	lysis protein	364	
POG449	DNA packaging protein A	365	
POG450	endopeptidase	606	77118208
POG451	gene 18.7 protein	366	
POG452	large terminase subunit, DNA packaging protein B, DNA maturase B	47/901/5	
POG453	gene 19.2 protein	367	
POG454	gene 19.5 protein	369	
POG455	putative capsid portal protein	6	
POG456	terminase large subunit	901/5/47/240	
POG457	Putative ATPase subunit of terminase, gpP-like	5/901/47	
POG458	putative capsid scaffolding protein	300	
POG459	phage major capsid protein	301	
POG460	phage small terminase subunit	302	
POG461	phage head completion protein	303	
POG462	phage Tail Protein X	305	
POG463	holin	740	
POG464	Lambda lysozyme	234	
POG465	LysB	308	
POG466	LysC	309	
POG467	phage tail completion protein R, GpR	310/376	
POG468	Protein S of phage P2, phage_tail_S, tail completion and stable head joining	377	link to POG613
POG469	baseplate assembly protein gpV	312/933/422	

POG470	bacteriophage baseplate assembly protein J	314/382	
POG471	gpI Bacteriophage P2-related tail formation protein	315	
POG472	phage tail sheath protein	317/733	
POG473	phage major tail tube protein	318	
POG474	phage tail protein E	899/419	
POG475	phage tail protein, gpE like protein	899	
POG476	putative tail protein, gpU	319	
POG477	bacteriophage late control gene D protein	320	
POG478	putative phage transcriptional activator Ogr/Delta	322	
POG479	repressor protein CI	112/370	
POG480	CII	371/321	
POG481	DNA polymerase alpha subunit, exonuclease domain	665	
POG482	hypothetical protein	746	
POG483	putative PAPS reductase/sulfotransferase	948	
POG484	Tum, DinI-like protein	846	
POG485	hypothetical protein	643	
POG486	hypothetical protein	68	
POG487	minor tail subunit	712	
POG488	peptidoglycan-binding domain	73	Hhsearch found this match, but psiblast doesn't
POG489	putative portal protein	204	
POG490	hypothetical protein	74/580	
POG491	hypothetical protein	714	
POG492	hypothetical protein	77/582	
POG493	unknown	78	
POG494	unknown	79	
POG495	unknown	80	
POG496	major tail subunit	81/428	
POG497	unknown	82	
POG498	unknown	83	
POG499	minor tail subunit	76/928/171	
POG500	minor tail protein	84	psiblast supports minor tail protein, while hhsearch finds major tail protein
POG501	unknown	694	
POG502	unknown	921/928	
POG503	unknown	921	
POG504	unknown	921/713	
POG505	predicted 6.3Kd protein	451/902	
POG506	MazG nucleotide	276	

	pyrophosphohydrolase		
POG507	unknown	89/302	
POG508	hypothetical protein		predicted 9.7Kd protein
POG509	predicted 8.0Kd protein		
POG510	DNA polymerase I	90	
POG511	Thymidylate synthase complementing protein	407	RP thymidylate synthase
POG512	unknown	91	
POG513	unknown	92	
POG514	DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog	93	
POG515	unknown	94	
POG516	unknown	95/636	
POG517	Glutaredoxin-like NRDH-redoxin, thioredoxin	100/785	
POG518	Putative DNA primase	96	
POG519	hypothetical protein	706	
POG520	predicted 5.8Kd protein		one matches phage minor tail protein U
POG521	hypothetical protein	97	
POG522	putative phosphoesterase or phosphohydrolase	847/759	
POG523	hypothetical protein	101	
POG524	Putative RecB family exonuclease	102	
POG525	hypothetical protein		
POG526	hypothetical protein	103	
POG527	hypothetical protein	104	
POG528	predicted 28.2Kd protein		
POG529	hypothetical protein	107	
POG530	predicted 10.1Kd protein		
POG531	predicted 14.2Kd protein	42	
POG532	predicted 7.4Kd protein, gp93		
POG533	hypothetical protein	727	
POG534	hypothetical protein	718	
POG535	predicted 26.5Kd protein, gp85	720	
POG536	hypothetical protein		
POG537	hypothetical protein		
POG538	hypothetical protein	3	
POG539	hypothetical protein		
POG540	structural protein VP1	443/869	
POG541	AAA+ ATPase, DnaA	856	
POG542	putative integrase	299	
POG543	unknown	858	
POG544	hypothetical protein		
POG545	hypothetical protein	864	Hhsearch matches 'zinc finger C2H2 protein' with short length

POG546	hypothetical protein	861	
POG547	hypothetical protein	862	
POG548	zinc finger protein		
POG549	unknown	865	
POG550	unknown	866	SecD Preprotein translocase subunit
POG551	unknown	867	
POG552	structural protein VP, unique to SSV viruses	869/443	
POG553	unknown	870	Unique to SSV viruses
POG554	unknown	854	Unique to SSV viruses
POG555	unknown	855	Unique to SSV viruses
POG556	transcription regulator ArsR family	868	Unique to SSV viruses
POG557	hypothetical protein		
POG558	Bor protein precursor	435	Bor protein or Bor Protein precursor?
POG559	Phage terminase large subunit	646/901/5	
POG560	gpW, head-to-tail protein W	853	
POG561	capsid protein	647/6/41	probable portal protein
POG562	peptidase, Clp protease	26/386	capsid protein?
POG563	capsid protein	386	Still matches 'peptidase'
POG564	Phage major capsid protein E	852/590	
POG565	Prophage minor tail protein Z, gpZ		
POG566	Minor tail protein U		
POG567	phage tail protein	12	also matches 'capsid protein', and surface proteins containing Ig-like domains
POG568	major tail protein (psiblast)		Hhsearch matches bacterial Ig-like domain
POG569	Bacteriophage lambda minor tail protein, gpG		
POG570	Minor tail protein T		
POG571	prophage tail length tape measure protein	903	
POG572	tail component	400	possible phage-related minor tail protein, or tail length tape measure protein
POG573	tail component	400	possible phage-related minor tail protein, or tail length tape measure protein
POG574	phage minor tail protein	16	
POG575	phage minor tail protein L	17	tail component
POG576	NlpC/P60 famlily, tail component	18	C terminal of different bacterial and viral proteins. Weral proteins are described as tail assembly proteins or Gp19.
POG577	tail component	19	bacterial lambda tail assembly protein I
POG578	tail component	20/8	
POG579	Excisionase	893	
POG580	Bacteriophage lambda Kil protein	222	short length
POG581	Bacteriophage lambda Kil protein	222	short length

POG582	restriction alleviation protein	639	
POG583	Cro Lambda repressor	225/937	
POG584	Regulatory protein CII	226	
POG585	NinD	641	
POG586	NinF	230	
POG587	ser/thr protein phosphatease		
POG588	antitermination protein	232	DnaJ-class molecular chaperone with C-terminal Zn finger domain
POG589	Bacteriophage lysis protein	606	PspB Phage shock protein B
POG590	Dna polymerase	52	viral, bacteria, plant organelles, type B DNA polymerase
POG591	early protein gp6	55	
POG592	late gene activator, early protein GP4	54	Phi-29
POG593	Phi-29 DNA terminal protein GP3		
POG594	DNA polymerase	52	
POG595	major head protein (phi29-like phages)	57	
POG596	tail protein	58	
POG597	upper collar protein	59	head-tail connector, portal, SH3-like, helix bundle
POG598	lower collar protein	60/188	
POG599	pre-neck appendage protein	61	
POG600	pre-neck appendage protein	61	
POG601	pre-neck appendage protein	61	
POG602	morphogenesis protein	625/900	lysin?
POG603	morphogenesis protein	900/307/910	zinc metallopeptidase?
POG604	holin	62/566	
POG605	lysin		
POG606	DNA encapsidation protein, Gp16	66/568	
POG607	XerD Site-specific recombinase	299	XerD and XerC integrase
POG608	T7 gene 0.3 product	757	PF08684 DNA mimic ocr
POG609	gp1.8		
POG610	HNH endonuclease	273	DNA binding domain?
POG611	predicted transcriptional regulator	38	homolog to E.Coli TOR inhibitor protein
POG612	hypothetic protein	397	HicB family, involoved in pilus formation
POG613	Protein S of phage P2, phage_tail_S, tail completion and stable head joining	377	link to Pog468
POG614	Phage tail sheath protein	378	
POG615	putative tail tube protein	379	

POG616	hypothetical protein	380	phage protein
POG617	hypothetical protein	381	phage protein
POG618	Baseplate J-like protein	382/314/802	
POG619	phage tail protein	383	
POG620	hypothetical protein	384	
POG621	hypothetical protein	385	
POG622	hypothetical protein	15	phage protein
POG623	hypothetic protein, putative phage gene	15	
POG624	hypothetical protein		
POG625	Single-strand binding protein	119	
POG626	ERF superfamily, the DNA single-strand annealing proteins (SSAPs)	118	
POG627	hypothetic protein	873	
POG628	hypothetic protein	282	
POG629	DNA polymerase subunit	724/650	DNA repair protein, recombinase RadA, RadB
POG630	Holliday junction endonuclease	659	
POG631	minor structural protein	929/909	
POG632	(putative) terminase small subunit	894	
POG633	prohead protease	904/48	
POG634	putative minor structural protein		short length
POG635	major capsid protein	7	can't find match for '9629659-2'
POG636	putative structural protein		
POG637	putaitive phage structural prtein		
POG638	putative structural protein		
POG639	hypothetical protein		
POG640	Lactophage P2 receptor binding protein	215	
POG641	holin		
POG642	amidase	573/272/346	lysin, N-acetylmuramoyl-L-alanine amidase
POG643	hypothetical protein		
POG644	hypothetical protein		sklp23, e32
POG645	hypothetical protein		sklp24,e31
POG646	hypothetical protein		sklp26,e28
POG647	hypothetical protein		sklp30,e22
POG648	hypothetical protein		sklp33,e17
POG649	hypothetical protein	930	sklp34,e15
POG650	hypothetical protein		sklp35,e14
POG651	DNA repair protein FAD52 honolog	157	the homologous-pairing domain of Rad52 recombinase
POG652	hypothetical protein		sklp38
POG653	hypothetical protein		sklp42
POG654	hypothetical protein		sklp43,e7

POG655	hypothetical protein		sklp52
POG656	hypothetical protein		sklp53
POG657	hypothetical protein		sklp54
POG658	hypothetical protein		sklp54
POG659	holin	144	
POG660	peptidase, related to metalloendopeptidase	900/307/910	related to putative tail fiber protein, POG238
POG661	DNA polymerase II small subunit/DNA polymerase delta, subunit B	962	DNA repair exonuclease
POG662	histone-like DNA binding protein		HU protein
POG663	hypothetical protein	203	
POG664	DNA polymerase III alpha subunit	730	
POG665	DNA polymerase III alpha subunit	730	
POG666	DNA methylase	260	C5 cytosine-specific DNA methylase
POG667	NrdI Protein involved in rionucleotide reduction		
POG668	dUPTase	128/258	N-end
POG669	gpH, tail fiber protein	914	
POG670	chitinase	72	
POG671	hypothetical protein		
POG672	hypothetical protein		
POG673	hypothetical protein		exonuclease
POG674	hypothetical protein		
POG675	Trp repressor protein	715/716	short length
POG676	hydrolase	43	Hhsearch supports chloride peroxidase and bromoperoxidase
POG677	unknown	717	
POG678	tail component	18	seems like to be a proteasomal regulaorty subunit
POG679	hypothetic protein, gp22		
POG680	Cor protein		hhsearch links to Bor protein
POG681	Repressor protein CI	112	also matches peptidase_S24, LexA repressor, transcription repressor.
POG682	ParB-like nuclease domain	944	
POG683	Helix Turn Helix DNA binding domain	112	ParB, transcription repressor
POG684	Helix-turn-helix transcription regulaotor	902	
POG685	partitioning protein ParA	569	ATPase? MinD, nitrogenase iron protein
POG686	protelomerase	645/299	Integrase?
POG687	Primase/helicase	649/932	
POG688	DNA replication protein		
POG689	helicase	649/295/294	
POG690	helicase	649/295/294	
POG691	helicase	649	
POG692	helicase	649	

POG693	replication protein	649	
POG694	antitermination protein Q		
POG695	antirepressor, regulatory protein Cro	937/38	
POG696	phage immunity repressor protein		
POG697	hypothetic protein		
POG698	hypothetical protein	446	
POG699	hypothetical protein		
POG700	hypothetical protein		
POG701	hypothetical protein		
POG702	DNA methylase	202	
POG703	hypothetical protein		Hhsearch matches "Nucleoside 2-deoxyribosyltransferase" with short length
POG704	C-5 cytosine-specific DNA methlase	260	
POG705	hypothetical protein	46	
POG706	phage terminase large subunit	240/47/5/901	
POG707	putative head-tail adaptor	10	
POG708	structural protein	11/570/138/183	putative head-tail joining protein, or tail component
POG709	putative structural protein		
POG710	putative structural protein		Nanovirus component 8 (C8) protein
POG711	neck passage structural protein	909/929/50	minor structural protein
POG712	hypothetic protein		
POG713	hypothetic protein		
POG714	hypothetic protein		
POG715	hypothetic protein		
POG716	transcglycosylas	910/494	lysozyme
POG717	endonuclease	273	putative endodeoxyribonuclease; HNH homing endonuclease
POG718	HNH homing endonuclease	273	
POG719	hypothetical protein		
POG720	CHC2 zinc finger	96	primase
POG721	protein of unknown function	33/171/214/186/670	holin-like, structural protein?
POG722	tail protein		Bactriaophage Mu P protein
POG723	protein of unknown function	682/433/102	RecB, exodeoxyribonuclease V beta chain
POG724	unknown	448	
POG725	hypothetical protein	231	
POG726	holiday juction resolvase -- RusA	32	
POG727	KilA-N domain	235	DNA binding
POG728	hypothetical protein	237	
POG729	(putative) phage head-tail joining protein	10/137	
POG730	head-tail adaptor protein	10/182/137	
POG731	anti-RecBCD protein 2	221	

POG732	hypothetical protein	35	
POG733	single-stranded DNA binding protein	119	
POG734	Repressor Protein Cro	38/937	CI?
POG735	Phage major tail protein	12/211/268	
POG736	Putative tail protein		putative major tail protein, Immunoglobulin I-set domain protein
POG737	putative tail length tape measure protein	400/903	methyl-accepting chemotaxis protein (MCP) signaling domain
POG738	phage tail assembly chaperone	13	
POG739	minor tail protein T	14	
POG740	hypothetical protein	545	
POG741	Gp53 or Gp58		
POG742	Replication protein O	547/257	
POG743	DnaD like domain	121	
POG744	DnaC	122	ATPase
POG745	Single-strand binding protein	119	
POG746	unknown	187	ORF37, ORF56
POG747	transcription regulator ArpU family	284/441/338	putative transcriptional activator RinA
POG748	terminase small subunit	4/197	
POG749	phage portal protein, SPP1 Gp6-like	204/286/41	
POG750	phage minor head protein	205/607	phage Mu protein F like protein
POG751	Phage major capsid protein E	590/852	
POG752	unknown	209	
POG753	structural protein	210/571/966/139	tail protein, head taill joining protein
POG754	Phage major tail protein	211	
POG755	hypothetical protein	212	
POG756	hypothetical protein	213	Difference between the two?
POG757	hypothetical protein	213	
POG758	capsid protein, N terminal, related to POG228		
POG759	hypothetical protein	669	
POG760	putaitive head tail joining protein	137/10/182	
POG761	putative head-tail joining protein	571/139/966	short length
POG762	phage major tail protein	140/184	
POG763	Bacterial Ig-like domain		
POG764	hypothetical protein	653	
POG765	structural protein (tail?)	186/670/214/171	
POG766	putative tail fiber protein; mttA/Hcf106 family	922/254/387/915	
POG767	hypothetical protein	671	

POG768	N-acetylmuramoy-L-alanine amidase	64	it matches 'N-acetylmuramoy-L-alanine amidase', which is POG63; but it matches POG64 and doesn't find out POG63
POG769	N-acetylmuramoy-L-alanine amidase	64	
POG770	Leucocidin S component LukS-PV	145/146	Luk protein, hemolysin/cytotoxin
POG771	LukF-PV	145/146	precursor
POG772	putative membrane protein	110	Na+/K+ ATPase alpha subunit, phi PVL ORF 30 analogue
POG773	transcriptional regulator		hypothetical protein, phi PVL orf 35-like protein
POG774	hypothetical protein	667	phi PVL orf 35-like protein
POG775	hypothetical protein		
POG776	hypothetical protein	114	
POG777	hypothetical protein	115	phi PVL orf 38-like protein
POG778	hypothetical protein	116	phi PVL ORF 39 analogue
POG779	hypothetical protein		
POG780	hypothetical protein	686	
POG781	hypothetical protein		phiPVL ORF41-like protein
POG782	RecT Recombinational DNA repair protein	190/539	
POG783	hypothetical protein		PhnP Metal-dependent hydrolases of the beta-lactamase superfamily
POG784	hypothetical protein		
POG785	hypothetical protein	125	PVL ORF-50-like family
POG786	phage conserved open reading frame 51	126	
POG787	hypothetical protein	127	PVL orf52-like protein
POG788	hypothetical protein		
POG789	hypothetical protein		
POG790	hypothetical protein		
POG791	transcriptional activator rinB	131	
POG792	hypothetical protein	132	
POG793	hypothetical protein	333/338	transcription activator, HTH?
POG794	protein containing tetrapyrrole methyltransferase domain and MazG-like	275/276	
POG795	putative head-tail joining protein	570	
POG796	hypothetical protein		
POG797	hypothetical protein		
POG798	hypothetical protein	238	
POG799	GtrC superfamily, O-antigen conversion protein		predicted membrane protein
POG800	bacteriaophage, scaffolding protein	242	
POG801	head assembly protein	246	
POG802	EaF protein	968	

POG803	NinZ		
POG804	terminase small subunit		
POG805	Lipoprotein Rz1 precursor	562	
POG806	Glycosyl transferase, GtrB	453/946	
POG807	Glycosyl transferase, GtrB	453	
POG808	GtrA-like protein	454	putative flippase, GtrA, glycosyl translocase
POG809	unknown	218/695	EaG.
POG810	anti-RecBCD Abc1, transcriptional regulator	572	
POG811	hypothetical protein		
POG812	antirepressor, regulatory protein Cro	937	
POG813	hypothetical protein NinX		
POG814	hypothetical protein NinY		
POG815	coat protein	155/357	
POG816	Mnt repressor		Mnt repressor mutant with C-terminal residues deleted
POG817	tailspike protein	251	
POG818	hypothetical protein	37	
POG819	phage minor tail protein L	17	
POG820	phage-related protein, tail component	8/20	
POG821	unknown		gp38
POG822	putative phage encoded membrane protein	941	
POG823	putative holin	412	
POG824	hypothetical protein	123	phage-like protein
POG825	hypothetical protein		
POG826	hypothetical protein	323	
POG827	hypothetical protein		
POG828	hypothetical protein		
POG829	hypothetical protein	663	
POG830	hypothetical protein		
POG831	putative transcription regulator	324	
POG832	Siphovirus Gp157	255	It is thought that bacteria possessing the gene coding for this protein have an increased resistance to the bacteriophage.
POG833	hypothetical protein	120/228/25	
POG834	hypothetical protein	257/121	putative DNA replication protein
POG835	hypothetical protein	124	
POG836	hypothetical protein	328	
POG837	hypothetical protein	30	
POG838	Deoxyuridine 5'-triphosphate nucleotide hydrolase	331/770	
POG839	hypothetical protein	130	
POG840	hypothetical protein	332	
POG841	hypothetical protein	139/571/966	putative structural protein, tail head joining

			protein
POG842	hypothetical protein	731	
POG843	hypothetical protein	143	
POG844	lytic enzyme		
POG845	Phosphate starvation-inducible protein PhoH, predicted ATPase	294	
POG846	acetyltransferase		
POG847	N-acetylmuramoy-L-alanine amidase	64	endolysin
POG848	putative tail fiber protein	69	
POG849	unknown	70	
POG850	hypothetical protein	69	tail fiber protein?
POG851	hypothetical protein	71	gp7?
POG852	contain peptidoglycan-binding domain	307/72/73	putative lysin
POG853	putative lysin	72/307	
POG854	putaive lysin	72	
POG855	putative minor tail subunit	76	
POG856	unknown		gp25,24?
POG857	unknown		gp30,28?
POG858	unknown		gp28,29?
POG859	unknown	921/928/713	putative structural protein
POG860	beta-lactamase	86?	penicilllin binding protein?SxxK motif
POG861	hypothetical protein		
POG862	hypothetical protein		gp33?
POG863	RNA helicase GLH-2	70	not sure
POG864	unknown	87	
POG865	DNA invertase, serine recombinase, integrase	88	
POG866	unknown		gp37?
POG867	unknown		gp37, 38, 40?
POG868	unknown		gp39?
POG869	unknown		gp39, 40?
POG870	unknown		gp42, 45?
POG871	unknown	91	
POG872	hypothetical protein		
POG873	unknown		
POG874	unknown		
POG875	unknown	99	
POG876	unknown		
POG877	terminase small subunit		short length
POG878	unknown		porcine reproductive and respiratory syndrome virus (PRRSV) 2b protein; invasion gene expression up-regulator
POG879	unknown	105	
POG880	unknown		Seq protein
POG881	unknown		
POG882	unknown	40	

POG883	unknown	106	
POG884	peptidase	111,151	transcription regulator?
POG885	putative antirepressor		
POG886	hypothetical protein		
POG887	hypothetical protein		
POG888	hypothetical protein		
POG889	hypothetical protein		
POG890	Bacteriophage Mu Gam like protein	117/255	
POG891	putative recombination protein ERF	118	DNA sigle-strand annealing proteins (SSAPs)
POG892	hypothetical protein		
POG893	HTH DNA binding motif	121	transcription regulator, replication protein
POG894	sigma factor?	338/170/284	Transcription regulator?
POG895	major head protein	135	
POG896	major capsid protein	7/135/335/9	
POG897	hypothetical protein	136	
POG898	hypothetical protein	138/570/11	putative structural protein, or head-tail joining protein
POG899	phage major tail protein	140/184	
POG900	major tail protein		
POG901	Bacterial Ig-like domain		bacteria and phage surface protein
POG902	hypothetical protein		
POG903	structural protein (tail?)	33/214/186/171 /670	DNA circulation protein?
POG904	structural protein (tail?)	141/900/50/922	
POG905	hypothetical protein		SLT orf 96-like protein
POG906	minor structure protein	924/61	
POG907	structural protein (tail? Neck passage?)	920/172/909/27 0/917	
POG908	hypothetical protein	142	
POG909	hypothetical protein		
POG910	hypothetical protein		
POG911	hypothetical protein	150	
POG912	antirepressor	113	
POG913	antirepressor		
POG914	hypothetical protein	152	
POG915	unknown	153	orf8
POG916	unknown	154	orf9
POG917	unknown	39	orf10
POG918	hypothetical protein	908	orf11
POG919	DNA replication protein	158/257/274	
POG920	putative replisome organiser protein C-terminus	158	
POG921	hypothetical protein		
POG922	protein of unknown function	160	
POG923	protein of unknown function	161/162	
POG924	protein of unknown function	162/167	

POG925	protein of unknown function	163	
POG926	protein of unknown function	164	
POG927	hypothetical protein	168	prophage protein
POG928	unknown	169	
POG929	protein of unknown function	170	
POG930	phage terminase, small subunit	134	
POG931	hypothetical protein	208	
POG932	hypothetical protein	10	head-tail adaptor?
POG933	hypothetical protein	11/570/138	
POG934	hypothetical protein		
POG935	hypothetical protein	173	
POG936	putative holin		
POG937	lysin	307/306	putative peptidoglycan-binding domain-containing protein
POG938	hypothetical protein		
POG939	putative repressor	175	
POG940	bacterial phage repressor		
POG941	hypothetical protein	176	
POG942	DNA recombination protein	118	? Homologous to Rad52? Eukaryotic recombination protein
POG943	DnaC	122	
POG944	hypothetical protein	177	
POG945	unknown	218/159/166	
POG946	hypothetical protein	170/338	putative transcription regulator
POG947	hypothetical protein		
POG948	Repressor		CI?
POG949	putative repressor	189	transcription regulator
POG950	putative repressor	189	transcription regulator
POG951	phage replication protein	121	N-terminal phage replisome organiser
POG952	RusA, endodeoxyribonuclease, holliday junction resolvase	32	
POG953	unknown	192	Dam-replacing family
POG954	unknown	193	
POG955	unknown	147	
POG956	terminase small subunit	197/239	
POG957	GepA uncharacterized phage-associated protein	149	
POG958	excisionase	199	
POG959	unknown	201	
POG960	terminase large subunit	47/5/240/901	
POG961	minor structure protein	206	
POG962	DNA binding, zinc finger		
POG963	major capsid protein	9/7/575	
POG964	putative structural protein	207	short length
POG965	putative protein	208	
POG966	putative structural protein	214/171	
POG967	putative structural protein	172/920/909	

	(tail? Neck passage?)		
POG968	putative neck passage structure	909/50	
POG969	lysin		50S ribosomal protein L24P
POG970	lysin motif		
POG971	unknown	217	
POG972	unknown	448	
POG973	terminase small subunit	239/197	DNA binding domain?
POG974	portal protein	241/401	
POG975	unknown	250	
POG976	hypothetical protein	252	
POG977	hypothetical protein	253	
POG978	minor head protein		
POG979	replication protein	257/121	LexA DNA binding protein
POG980	phage minor capsid protein	262/205	
POG981	putative scaffold protein	263	
POG982	major capsid protein		
POG983	minor capsid protein	265	
POG984	minor capsid protein	266/11	
POG985	minor capsid protein	267	
POG986	major tail shaft protein	268	
POG987	hypothetical protein		
POG988	hypothetical protein	269	
POG989	putative holin 2	271	
POG990	hypothetical protein	285	
POG991	ParB-like nuclease domain	944	
POG992	putative portal protein	286/41/401	
POG993	DNA methyltransferase	202	
POG994	hypothetical protein		
POG995	LexA DNA binding domain	936	
POG996	probable acetyltransferase	277	
POG997	RecA bacterial DNA recombination protein	279/709	
POG998	N-acetylmuramoy-L-alanine amidase	63	
POG999	hypothetical protein		
POG1000	hypothetical protein		
POG1001	hypothetical protein	264	
POG1002	structural protein	900/50/141/922 /577	
POG1003	L-alanyl-D-glutamate peptidase, hydrolase		
POG1004	hypothetical protein		
POG1005	hypothetical protein		
POG1006	hypothetical protein		
POG1007	gp40		
POG1008	nuclease	298/536	match both endonuclease and exonuclease
POG1009	putative methyltransferase	202/399	

POG1010	gp52		
POG1011	gp54		
POG1012	gp55		
POG1013	gp65		
POG1014	hypothetical protein		
POG1015	putative peptidoglycan binding domain	307/306	
POG1016	endolysin, N-acetylmuramoyl-L-alanine amidase	306	transglycosylase SLT domain
POG1017	Ogr_Delta phage transcriptional activator	322	
POG1018	hypothetical protein		
POG1019	hypothetical protein		
POG1020	hypothetical protein		
POG1021	hypothetical protein		
POG1022	transcription regulator	112/109/716	
POG1023	hypothetical protein		
POG1024	hypothetical protein	325	
POG1025	Ssb, single stranded DNA-binding protein	327/119	
POG1026	hypothetical protein		
POG1027	hypothetical protein		ETA orf 28-like protein
POG1028	hypothetical protein	333	
POG1029	unknown	297/207/581	Rho termination factor, N-terminal domain
POG1030	hypothetical protein	208/414	
POG1031	major tail protein		
POG1032	hypothetical protein		
POG1033	tail assembly protein		
POG1034	phi ETA orf 54-like protein		
POG1035	hydrolase?	141	
POG1036	tail fiber protein	920/172	
POG1037	N-acetylmuranoyl-L-alanine amidase	573/588/272	
POG1038	amidase, hydrolase	63	
POG1039	tail fiber protein	917/919/920	
POG1040	hypothetical protein		phi ETA orf 63-like protein
POG1041	amidase	63	
POG1042	exfoliative toxin A		heat shock protease?
POG1043	ORF 098		
POG1044	phage uncharicterized protein		
POG1045	NfeD nodulation efficiency protein D		
POG1046	plypsa, alpha/beta hydrolase, multi-domain	64	
POG1047	conserved hypothetical protein		
POG1048	conserved hypothetical protein		
POG1049	helix-turn-helix	112	CI-like transcription regulator

POG1050	putative transcriptional activator RinA	338/170/284	sigma factor
POG1051	putative major capsid protein		
POG1052	structural protein	966/571/139	tail component? Or head tail joining protein?
POG1053	putative tail length tape measure protein	400	
POG1054	phage-related minor tail protein	400	putative tail length tape measure protein
POG1055	putative tail protein	387/254	
POG1056	hypothetical protein		
POG1057	hypothetical protein	444/445	
POG1058	lysis protein S, holin	559	
POG1059	bacterial phage lysis protein	606	
POG1060	hypothetical protein	972	
POG1061	conserved hypothetical protein	801	
POG1062	VRR_NUC domain		
POG1063	conserved hypothetical protein		
POG1064	hypothetical protein		
POG1065	conserved hypothetical protein	389	
POG1066	phage protein Gp37/Gp68	390	
POG1067	ISBt3 transposase subunit protein	24	
POG1068	ISBt3 transposase subunit protein	24	
POG1069	HflC/HflK family inner membrane lipoprotein	392	
POG1070	hypothetical protein		
POG1071	hypothetical protein		
POG1072	bacteriophage protein		
POG1073	transcription repressor	112/109/937	
POG1074	hypothetical protein		
POG1075	putative chromosome partitioning protein PARB	944	
POG1076	putative chromosome partitioning protein	112	
POG1077	replication protein	257	transcription regulator?
POG1078	bacteriophage replication protein	257	
POG1079	bacteriophage replication protein	257	
POG1080	hypothetical protein		
POG1081	protein of unknown function	231	
POG1082	hypothetical protein		
POG1083	hypothetical protein		
POG1084	putative helix-turn-helix	112	

	transcriptional regulator		
POG1085	hypothetical protein	394	
POG1086	hypothetical protein		
POG1087	Cox, transcription regulator, K139-like phages	742	
POG1088	hypothetical protein	374	
POG1089	P2 phage tail completion protein R, GpR	376	
POG1090	hypothetical protein	197; 239	small terminase subunit; phage protein. Hhrsearch found transposase, recombinase
POG1091	hypothetical protein	405, 693	
POG1092	putative antirepressor	113	
POG1093	replication protein	257,547	
POG1094	hypothetical protein		transcription regulator? Short length match
POG1095	putative integral membrane protein	413	
POG1096	hypothetical protein	414, 181, 136	phage protein DNA packaging protein
POG1097	hypothetical protein	415	conserved hypothetical protein
POG1098	hypothetical protein	416	putative bacterialphage protein
POG1099	tail tube protein	418	
POG1100	putative tail like protein	419, 899	hhr support tail like protein, but psiblast indicates sigma factor domain-containing protein
POG1101	tail/DNA circulation protein	420	
POG1102	conserved hypothetical protein		
POG1103	putative terminase	5, 901, 47, 240	large subunit
POG1104	hypothetical protein	427	
POG1105	major tail subunit	428, 81, 689	
POG1106	hypothetical protein	928, 921	Xylansas Y, carbogydrate-binding module, xylan-binding
POG1107	putative peptidase	72	
POG1108	hypothetical protein	346, 573	peptidogylcan binding domain-containing protein; peptidoglycan recognition protein I-alpha
POG1109	Whib Transcription factor	429	
POG1110	DNA polymerase III, alpha subunit, exonulease	665	
POG1111	hypothetical protein		psiblast found Na (+)-translocating NADH-quinone reductase subunit F
POG1112	hypothetical protein		
POG1113	thioredoxin	100	
POG1114	hypothetical protein		
POG1115	hypothetical protein		
POG1116	acetyltransferase	360	hhrsearch supports
POG1117	hypothetical protein		
POG1118	hypothetical protein	455	metal dependent phosphohydrolase
POG1119	conserved hypothetical protein		

POG1120	hypothetical protein		
POG1121	KilA-N domain family	235	
POG1122	hypothetical protein	120	putative cytoplasimic protein; HHr found endonulease
POG1123	hypothetical protein	475	unique to stx phages?
POG1124	hypothetical protein	550	
POG1125	hypothetical protein	557	
POG1126	HslV family protein		
POG1127	methyltransferase subunit	567	
POG1128	hypothetical protein		
POG1129	minor head protein	205, 607, 580	HHr found Mu portein F like protein
POG1130	Baseplate J-like protien	314	
POG1131	CopG/DNA-binding domain-containing protein	296	
POG1132	dUTPase / dCTPase	128	
POG1133	lysin / amidase	272, 573	
POG1134	hypothetical protein		
POG1135	hypothetical protein	255	
POG1136	hypothetical protein	165	
POG1137	hypothetical protein	574	
POG1138	coat protein	575, 9	
POG1139	phage putative tail component	171, 214	
POG1140	putative phage pre-neck appendage protein	924, 61	
POG1141	putative site specific recombinase	88	
POG1142	repressor or peptidase?	151,111	
POG1143	hypothetical protein	581, 207	
POG1144	phage protein	583	
POG1145	phage protein	584, 966	
POG1146	phage protein	585	
POG1147	hypothetical protein	586	
POG1148	holin	587	
POG1149	hypothetical protein	596	
POG1150	major capsid protein	603	
POG1151	phage head protein; minor head protein-like protein	607, 205, 580	
POG1152	putative portal protein	608, 734, 6, 204	NUDIX hydrolase
POG1153	BcepNY3gp31	617	
POG1154	phage baseplate_J	382, 314	
POG1155	hypothetical protein	630	
POG1156	hypothetical protein	761	
POG1157	phage protein	605	Scaffold protein (Burkholderia phage BcepNY3)
POG1158	putative baseplate protein	312, 422	
POG1159	BcepNY3gp39		
POG1160	pentapeptide repeat-containing protein		

	[Trichodesmium erythraeum IMS101]		
POG1161	hypothetical protein		
POG1162	recombination associtaed protein		Hhr search found putative exonuclease
POG1163	putative secretion activating protein	644	
POG1164	probable pyocin R2_PP, tail fiber protein	914	
POG1165	partition protein ParA	569	
POG1166	putative phage related protein		
POG1167	hypothetical protein	651	
POG1168	hypothetical protein		
POG1169	hypothetical protein	431	
POG1170	excisionase		
POG1171	hypothetical protein		hhr found high affinity transport system protein
POG1172	hypothetical protein		
POG1173	phi PV83 orf 10-like protein	663	
POG1174	DnaD-like domain	121	phi PV83 orf 20-like protein
POG1175	DNAB replicaton fork helicase		
POG1176	phi Mu50B-like protein	merge?	
POG1177	phi Mu50B-like protein		
POG1178	unknown phi protein	213	
POG1179	Amidase_2 N-acetylmuramoyl-L-alanine amidase; cell wall hydrolase	174	
POG1180	domain of unknown function in cell wall hydrolase	174	
POG1181	hypothetical protein	664	
POG1182	hypothetical protein		
POG1183	conserved hypothetical protein		
POG1184	unknown		
POG1185	hypothetical protein		
POG1186	hypothetical protein		
POG1187	Psiblast found 'probable ATP-dependent helicase' while HHsearch matches EXOIII exonuclease.	665	POG 1187 and POG 1188 are different domains in same protein
POG1188	Pol III-like ATP dependent helicase domain C-t		
POG1189	hypothetical protein		
POG1190	hypothetical protein		
POG1191	hypothetical protein		
POG1192	virulence associated protein E	200	VirE N-terminal domain

POG1193	staphylokinase		
POG1194	Staphylococcal complement inhibitor		
POG1195	hypothetical protein		
POG1196	hypothetical protein		
POG1197	chemotaxis-inhibiting protein CHIPS		
POG1198	putative Ribonuclease HI		
POG1199	hypothetical protein	673	only one (66391701) doesn't matches DNA protecting protein in psiblast
POG1200	hypothetical protein		
POG1201	probable acetyltransferase		
POG1202	bacterial phage protein	68	
POG1203	putative portal protein	204, 41	
POG1204	putative head protein	205, 74	
POG1205	bacteriophage protein	426 (scaffolding protein)	
POG1206	bacteriophage protein		
POG1207	bacteriophage protein		
POG1208	hypothetical protein	11	
POG1209	unknown	688, 966	Unknown refers to GP...
POG1210	unknown	689	
POG1211	unknown	690	
POG1212	unknown	691	
POG1213	phage tail tape measure protein	903	
POG1214	putative phage tail	76, 928, 921	
POG1215	putative tail protein	692	
POG1216	aminotransferase class III	928	
POG1217	unknown	70	
POG1218	unknown	693, 405	
POG1219	unknown		
POG1220	unknown		
POG1221	unknown	974	
POG1222	unknown	71	
POG1223	Amidase_2 N-acetylmu	573, 346	
POG1224	unknown		
POG1225	putative DNA polymerase III epsilon subunit	695	
POG1226	unknown		
POG1227	unknown		
POG1228	unknown		Hhsearch found Actin-binding LIM Zn-finger protein Limatin involved in axon guidance
POG1229	unknown		
POG1230	unknown		
POG1231	unknown		
POG1232	unknown	902	might be DNA binding domain -containing protein

POG1233	unknown		
POG1234	unknown		
POG1235	unknown	696	
POG1236	unknown		this one has interesting sequence, low complexity
POG1237	unknown		
POG1238	hypothetical protein	977	
POG1239	transcription factor WhiB	429	
POG1240	transcription factor?		not very comfirmed.
POG1241	unknown		
POG1242	unknown		
POG1243	putative endonulease	697, 725	not very comfirmed.
POG1244	Domain of unknown function		
POG1245	unknown		
POG1246	unknown		
POG1247	unknown		
POG1248	unknown		
POG1249	unknown		
POG1250	hypothetical protein	274	
POG1251	unknown	576	
POG1252	unknown		
POG1253	unknown	698	
POG1254	unknown		
POG1255	unknown		
POG1256	unknown		
POG1257	unknown	973	
POG1258	unknown		
POG1259	unknown		
POG1260	unknown		
POG1261	unknown	700	
POG1262	unknown		
POG1263	unknown		
POG1264	unknown	701	
POG1265	unknown		
POG1266	unknown	975	
POG1267	glycosyltransferase	702	hhsearch found polypeptide N-acetylgalactosaminyltransferase 1
POG1268	methyltransferase	702	
POG1269	glycosyltransferase fmily 25	703	
POG1270	polypeptide N-acetylgalactosaminyltransferase	704, 946	
POG1271	unknown		no hits found in psiblast and hhrsearch
POG1272	unknown		
POG1273	unknown	698	
POG1274	unknown		
POG1275	unknown	431	
POG1276	unknown		

POG1277	unknown	721	
POG1278	unknown	722	
POG1279	cutinase	73	
POG1280	unknown		
POG1281	exodeoxyribonuclease VIII (RecE)		
POG1282	glycosyltransferase-like protein	946	
POG1283	glycosyltransferase		not so sure
POG1284	ExsB protein		
POG1285	GTP cyclohydrolase I		
POG1286	holiday junction resolvases	659	
POG1287	protein of unknown function		
POG1288	phage portal protein	6	
POG1289	unknown		
POG1290	unknown		Not sure: according to structure, hhsearch found Cellobiose phosphorylase, beta-sandwich, (alpha/alpha) 6 barrel
POG1291	Mycobacterium phage protein		
POG1292	unknown		
POG1293	unknown		
POG1294	putative minor tail subunit	76	not sure
POG1295	unknown		
POG1296	unknown		
POG1297	unknown		
POG1298	tail component protein	11, 570	
POG1299	unknown		
POG1300	tail protein	917	?
POG1301	phage protein		
POG1302	tail protein; baseplate,gp44		hhrsearch supports
POG1303	unknown		
POG1304	unknown		
POG1305	unknown		
POG1306	unknown		
POG1307	unknown	577, 69	tail fiber?
POG1308	unknown	723	
POG1309	DNA binding domain	109	matches transcription regulator; SinR repressor, DNA-binding domain
POG1310	hypothetical protein		
POG1311	unknown		
POG1312	unknown	433, 682	
POG1313	unknown		
POG1314	bifunctional DNA primase/polymase N-terminal domain	195	
POG1315	putative DNA primase	724	
POG1316	unknown		
POG1317	unknown		

POG1318	unknown		
POG1319	unknown		may have helics turn helics structure
POG1320	unknown		
POG1321	HNH endonuclease	725, 25	
POG1322	bifunctional DNA primase/polymase	195	
POG1323	unknown		
POG1324	unknown	44	
POG1325	unknown	726	
POG1326	unknown		
POG1327	putative phage/prophage antirepressor	113	
POG1328	unknown		
POG1329	unknown	99	
POG1330	FtsK/SpoIIIE family protein	857	
POG1331	unknown		
POG1332	unknown		
POG1333	helicase	133	Hhsearch found chromatin remodeling complex, chromodomain-helicase DNA-binding protein… it may be DNA- protein interaction domain
POG1334	ATP-dependent nuclease, subunit B	102, 191, 610	RecB family exonuclease
POG1335	RuvC resolvase		?
POG1336	transposase		putative transposase DNA-binding domain
POG1337	unknown		
POG1338	unknown	653	
POG1339	unknown	139	
POG1340	unknown	652	
POG1341	unknown		
POG1342	unknown		
POG1343	unknown		
POG1344	unknown		
POG1345	phage terminase, large subunit	901, 240	
POG1346	unknown	748	
POG1347	putative endoprotease	749	
POG1348	unknown	750	
POG1349	unknown	751	
POG1350	unknown	752	
POG1351	major tail subunit	753	
POG1352	phage protein	86, 359	Maybe tail tubular subunit according to hhsearch
POG1353	putative phage related protein	363	
POG1354	adenine methylase	202	
POG1355	hypothetical protein		
POG1356	putative bacteriophage protein	600	
POG1357	hypothetical protein		hhsearch found Frequency Clock Protein

POG1358	DNA primase / helicase like protein		
POG1359	unknown		
POG1360	unknown		
POG1361	unknown		
POG1362	ATP-dependent DNA ligase-like protein, C-terminal	342	
POG1363	DNA endonuclease-like protein		
POG1364	unknown		
POG1365	scaffolding-like protein		
POG1366	major capsid protein	357	
POG1367	internal virion protein	361	
POG1368	hypothetical protein		
POG1369	phage-related lysozyme	933	not so sure
POG1370	internal virion protein	363	
POG1371	small terminase subunit		
POG1372	hypothetical protein	94	
POG1373	hypothetical protein		
POG1374	unknown		
POG1375	putative head protein		
POG1376	putative head to tail joining protein	10, 137	
POG1377	phage protein	966, 571	
POG1378	putative tail protien	12, 211	
POG1379	unknown		
POG1380	conserved phage protein		
POG1381	tail length tape measure protein	400	
POG1382	tail length tape measure protein	400	
POG1383	conserved phage protein		hhsearch found phage minor tail protein
POG1384	phage minor tail protein		
POG1385	peptidoglycan hydrolase/transglycosylase	18	148727198 does not have good E value
POG1386	hypothetical protein		
POG1387	hypothetical protein		
POG1388	hypothetical protein		
POG1389	hypothetical protein		
POG1390	hypothetical protein		
POG1391	hypothetical protein		
POG1392	putative DNA polymerase	90	
POG1393	putative phage helicase	650	
POG1394	putative DNA primase	932	
POG1395	hypothetical protein		
POG1396	hypothetical protein		
POG1397	putative inhibitor of transcription and anti-terminator protein		

POG1398	hypothetical protein		
POG1399	hypothetical protein		
POG1400	HNH endonuclease family protein / DNA binding domain		this pog has short length
POG1401	hypothetical protein		
POG1402	hypothetical protein		
POG1403	hypothetical protein	762?	
POG1404	hypothetical protein		
POG1405	hypothetical protein	898, 738	major head protein?
POG1406	hypothetical protein		
POG1407	hypothetical protein		
POG1408	hypothetical protein		
POG1409	hypothetical protein	793	
POG1410	hypothetical protein		
POG1411	gp5 baseplate hub subunit and tail lysozyme	933, 311	
POG1412	gp38 phage tail fibre adhesin		
POG1413	hypothetical protein		
POG1414	DEAD/DEAH box helicase	843	
POG1415	hypothetical protein	843	
POG1416	DNA methylase	260	
POG1417	DNA ligase C-terminal domain	342	link to POG29,30
POG1418	hypothetical protein	445	
POG1419	conserved hypothetical protein	777	
POG1420	homing endonuclease	832	
POG1421	hypothetical protein	984	
POG1422	adenylyltransferase	848	
POG1423	adenylyltransferase	608	phosphotase?
POG1424	conserved hypothetical protein		
POG1425	nicotinamide phosphoribosyl transferase	849	
POG1426	hypothetical protein	850	
POG1427	base plate hub assembly catalyst gp51	819	
POG1428	putative baseplate tail tube cap	822	
POG1429	baseplate structural protein GP27	820	
POG1430	hypothetical protein	986	
POG1431	conserved tail assembly protein	20, 8	
POG1432	hypothetical protein		this pog has short length; not good alignment to 'conserved tail assembly protein'
POG1433	conserved tail assembly protein		

POG1434	conserved tail assembly protein	17	
POG1435	primase	195	
POG1436	hypothetical protein		
POG1437	hypothetical protein	984	
POG1438	hypothetical protein	392	
POG1439	adenylate cyclase	985	
POG1440	Srd postualted decoy of host sigma70 or sigmaS	768, 25, 725	
POG1441	hypothetical protein	814	
POG1442	metal dependent phosphohydrolase, HD region		
POG1443	hypothetical protein		
POG1444	hypothetical protein		hhsearch found tail sheath protein
POG1445	hypothetical protein		
POG1446	hypothetical protein		
POG1447	hypothetical protein		
POG1448	hypothetical protein	850	
POG1449	terminase small subunit	4, 197	
POG1450	domain of unknown function	863, 102	hhsearch found exonulease
POG1451	unknown	871	
POG1452	hypothetical protein	872	
POG1453	putative HNH endonuclease	273	
POG1454	hypothetical protein		
POG1455	phage conserved protein		Hhsearch found 'intron-associated endonuclease 1', but not sure.
POG1456	putative ribose-phosphate pyrophosphokinase		
POG1457	hypothetical protein		
POG1458	DNA binding protein	894	
POG1459	phage protein		
POG1460	sheath tail protein	733, 417, 317	
POG1461	putative phage core tail protein		
POG1462	phage XkdN-like protein		
POG1463	tail length tape measure protein	903	
POG1464	putative LysM like protein		
POG1465	phage late control gene D	320	hydrolase?
POG1466	phage protein		
POG1467	phage protein	631	hhsearch found tial lysozyme. ?
POG1468	putative recombination protein		
POG1469	putative main capsid protein	357	
POG1470	hypothetical protein		
POG1471	hypothetical protein	329	

POG1472	unknown		
POG1473	hypothetical protein	876	
POG1474	hypothetical protein	681	
POG1475	hypothetical protein		
POG1476	putative major head subunit precursor	605	
POG1477	hypothetical protein		
POG1478	uncaracterized protein conserved in bacteria	603, 7	
POG1479	hypothetical protein		
POG1480	hypothetical protein		
POG1481	phage-related hypothetical protein	11	
POG1482	hypothetical protein		
POG1483	hypothetical protein		
POG1484	putative tail tape measure protein		
POG1485	zinc finger, principally involved in DNA binding in DNA primases	649, 96	1485-1486: same protein, different domain
POG1486	putative DNA primase		
POG1487	putative DNA primase		
POG1488	putative transcription regulator		
POG1489	putative ATP-dependent helicase		
POG1490	hypothetical protein	28	
POG1491	hypothetical protein		
POG1492	HYS2 Archaeal DNA polymerase II, small subunit/DNA polymerase delta, subunit B	962	
POG1493	hypothetical protein		
POG1494	hypothetical protein		
POG1495	hypothetical protein, RzlA		
POG1496	hypothetical protein, YdbL		
POG1497	hypothetical protein, YfilJ		
POG1498	hypothetical protein		
POG1499	putative phage head morphogenesis protein		
POG1500	putative phage head morphogenesis protein	205, 607	
POG1501	NAD-dependent DNA ligase	813	
POG1502	putative resolvase		
POG1503	hypothetical protein	636, 95	
POG1504	integrase	196	
POG1505	hypothetical protein		
POG1506	ATPase family associated with various cellular activities		

POG1507	ribonuclease H		
POG1508	terminase, large subunit	646	
POG1509	hypothetical protein		
POG1510	hypothetical protein		
POG1511	putative portal protein	6	
POG1512	putative prohead protease	904	
POG1513	putative major capsid protein		
POG1514	hypothetical protein		
POG1515	hypothetical protein		
POG1516	hypothetical protein		
POG1517	hypothetical protein		
POG1518	putative major tail sheath protein	317, 417, 733	
POG1519	hypothetical protein		
POG1520	hypothetical protein		
POG1521	putative tail lysin	18	but hhsearch doesn't agree with psiblast, hhsearch found phage late control gene D protein, phage_GPD
POG1522	putative tail lysin	18	hhsearch found amidas
POG1523	hypothetical protein		
POG1524	hypothetical protein		
POG1525	putative bacteriophage baseplate protein		
POG1526	baseplate structural protein	804	hhsearch
POG1527	putative adsorption associated tail protein		
POG1528	putative Replication protein		
POG1529	putative exonuclease	759	needs more work. May manually join with POG63 and some domains from POG62; link to POG63, gp47 recombination endonuclease subunit, C terminal; may be C terminal of exonuclease
POG1530	hypothetical protein		
POG1531	DNA primase / Zinc finger?	96	
POG1532	DNA primase		
POG1533	putative primase		
POG1534	hypothetical protein		
POG1535	hypothetical protein		hhsearch found predicted membrane protein
POG1536	hypothetical protein		
POG1537	putative sigma factor		
POG1538	putative structural protein		
POG1539	hypothetical protein		
POG1540	hypothetical protein		
POG1541	hypothetical protein		
POG1542	hypothetical protein		
POG1543	hypothetical protein		
POG1544	hypothetical protein		
POG1545	hypothetical protein		

POG1546	hypothetical protein		
POG1547	hypothetical protein		
POG1548	hypothetical protein		
POG1549	hypothetical protein		
POG1550	hypothetical protein		
POG1551	hypothetical protein	276, 275	hhsearch found MazG nucleotide pyrophosphohydrolase domain
POG1552	general secretion pathway protein A, SecA?	762, 294	type II secretory pathway, component ExeA
POG1553	integrase core domain	24	putative transposase A subunit
POG1554	putative transposase A subunit		
POG1555	putative transcriptional regulator, IclR helix-turn-helix domain		
POG1556	putative bacteriophage	38	
POG1557	putative phage-related DNA-binding protein, helix-turn-helix		
POG1558	putative phage encoded membrane protein		
POG1559	putative inner membrane protein		
POG1560	endolysin	910, 363, 789, 494	
POG1561	phage-like protein		putative transcriptional regulator TM1602, N-terminal domain
POG1562	small terminase subunit		
POG1563	protease / phage capsid scaffolding protein	300	?
POG1564	phage major capsid protein	852, 590	
POG1565	hypothetical protein		
POG1566	putative structural protein		
POG1567	putative structural protein		
POG1568	phage DNA polymerase	406, 52, 695	
POG1569	putative structural protein	621, 320, 15, 421	phage late control gene D protein
POG1570	putative holin protein		
POG1571	hypothetical protein		
POG1572	hypothetical protein		
POG1573	hypothetical protein		
POG1574	dUTPase	331, 770	
POG1575	holin	964	
POG1576	probable transposase		
POG1577	transition state regulatory protein?		?
POG1578	hypotheticla protein		
POG1579	20G-Fe (II) oxygenase superfamily		
POG1580	hypothetical protein		
POG1581	baseplate wedge	804	

POG1582	terminase small subunit	810	
POG1583	hypothetical protein		
POG1584	contractile sheath protein gp18	733, 317	
POG1585	contractile sheath protein gp19	733, 317	
POG1586	hypothetical protein		
POG1587	hypothetical protein		
POG1588	CobT Cobalamin biosynthesis protein		
POG1589	hypothetical protein		
POG1590	hypothetical protein		
POG1591	small heat shock protein		
POG1592	hypothetic protein		
POG1593	Nicotinic acid mononucleotide adenylyltransferase	848	
POG1594	hypothetic protein		psiblast found aerobic NDP reductase small subunit NrdB with higher E value
POG1595	hypothetic protein		
POG1596	hypothetic protein		
POG1597	glutaredoxin	785, 782	
POG1598	hypothetic protein		
POG1599	putative high-light inducible protein		merge?
POG1600	putative high-light inducible protein		
POG1601	photosystem II D1 protein		
POG1602	photosystem II D2 protein		
POG1603	hypothetic protein		
POG1604	hypothetic protein		
POG1605	hypothetic protein		
POG1606	hypothetic protein		
POG1607	putative exonuclease, C terminal	102, 682, 191, 610, 863, 433	
POG1608	hypothetical protein		
POG1609	baseplate structural proein GP27	820	
POG1610	hypothetical protein		
POG1611	20G-Fe (II) oxygenase superfamily		connect to POG1579
POG1612	putative helicase	295	
POG1613	hypothetical protein		
POG1614	hypothetical protein		
POG1615	hypothetical protein		
POG1616	putative translaldolase		
POG1617	hypothetical protein		
POG1618	putative adsorption associated tail protein		connect to POG1527
POG1619	possible DNA binding		

	protein		
POG1620	hypothetical protein		
POG1621	endonuclease		
POG1622	hypothetical protein		
POG1623	hypothetical protein		
POG1624	hypothetical protein		
POG1625	hypothetical protein		
POG1626	major structural phage protein		
POG1627	structural phage protein		
POG1628	hypothetical protein		
POG1629	exonuclease; RecD, C terminal	294	check relationship with POG125; ATP-dependent exoDNAse (exonuclease V), alpha subunit - helicase superfamily I member [DNA replication, recombination, and repair]
POG1630	hypothetical protein		
POG1631	putative structural protein	204, 41	
POG1632	phage protein		
POG1633	head morphogenesis protein	205, 580, 607, 262	
POG1634	hypothetical protein		
POG1635	putative helicase	650, 724, 709	
POG1636	conserved hypothetical protein	148	
POG1637	hypothetical protein		
POG1638	ribosomal protein L22	151	
POG1639	hypothetical protein		
POG1640	hypothetical protein		
POG1641	unknown		
POG1642	phage head morphogenesis protein		connect to POG1499 and POG1500
POG1643	unknown		
POG1644	excisionase		
POG1645	hypothetical protein		
POG1646	unknown		
POG1647	ParB-like nuclease domain	944	
POG1648	unknown		
POG1649	hypothetical protein		
POG1650	amidase		
POG1651	unknown		
POG1652	unknown		
POG1653	putative terminase subunit	134	
POG1654	unknown	143	
POG1655	unknown		
POG1656	phage replication protein	547	
POG1657	unknown		
POG1658	unknown		
POG1659	unknown		

POG1660	hypothetic protein	330	
POG1661	conserved phage protein		
POG1662	tail assembly protein		
POG1663	hypothetical protein	334	
POG1664	putative terminase small subunit	38	
POG1665	unknown		
POG1666	phage anti-repressor protein		Hhsearch found 'MCLC Mid-1-related choride channel'
POG1667	unknown		
POG1668	hypothetical protein		
POG1669	hypothetical protein		
POG1670	hypothetical protein		
POG1671	unknown		
POG1672	unknown		
POG1673	unknown		
POG1674	carbohydrate-binding domain, family V/XII	357, 9	hhsearch found major capsid protein
POG1675	hypothetical protein		
POG1676	putative HNH endonuclease, NUM0D4 motif, a putative DNA-binding motif	273	connect to POG1453
POG1677	hypothetical protein		
POG1678	hypothetical protein		
POG1679	putative transposase		
POG1680	hypothetical protein		
POG1681	hypothetical protein		
POG1682	hypothetical protein		
POG1683	unknown		
POG1684	unknown		
POG1685	DNA repair protein RAD52 homolog		
POG1686	unknown		
POG1687	unknown		
POG1688	unknown		
POG1689	putative thymidylate synthase	407, 94	157786444 doesn't match well

Appendix B: parameters for Roche 454 Newbler

Input Parameters

Large or complex genome: selected; large or complex datasets will be assembled successfully and speedily.

Expected depth: 0; the assembler does not use expected depth information in its computation.

Minimum read length: 20

Computation Parameter

Increment *de novo* assembler analysis: selected; New read data will be assembled into any existing assembly that was created during previous computations in this project.

Number of CPUs to use (0=all): 0

Overlap Detection Parameters:

Seed step: 12; the numbers of bases between seed generation locations used in the exact k-mer matching part of the overlap detection.

Seed length: 16; the number of bases used for each seed in the exact k-mer matching part of the overlap detection.

Seed count: 1; the number of seeds required in a window before an extension is made.

Minimum Overlap length: 40

Minimum Overlap Identity: 90%

Alignment identity score: 2; when multiple overlaps are found, the per-overlap-column identity score used to sort the overlaps for use in the progressive alignment.

Alignment difference score: -3; when multiple overlaps are found, the per-overlap-column difference score used to sort the overlaps for use in the progressive multi-alignment.

Output Parameters

Include consensus: selected; the application will generate all the output files, including the output files related to the generation of contigs and consensus sequence information.

Pairwise alignment: none; the 454PireAlign.txt file is not generated.

Ace/Consed: single ACE file for small genome – a single ACE file is generated if fewer than four million reads are input to the assembly.

Ace read mode: Default; when the ACE file is generated, output reads use the trimmed portion of the reads.

All contigs threshold: 100; the minimum number of bases for a contig to be output in the 454AllContigs.fna file.

Large contigs threshold: 500; the minimum number of basesd for a contig to be output in the 454LargeContigs.fna file.

Appendix C: the virus protein orthologs idenitified in Metagenomics data

The table was organized as follows: the first column lists the contig name along with its reading frame and ORF, while the second column gives the GI number (note that in cases where the search against individual sequences was inconclusive, a GI number is not available). Column three lists the protein family that the ORF was matched to, while columns four and five give the virus species and species taxonomy, respectively. Column six is the PSI-BLAST score or HHsearch probability, and column seven is the E-value.

contigID_frame_ORF	GI #	matched protein	virus	taxonomy	score / probability	E-value
contig07130_3_1	1020269	GKS/GKT ATPase / replication protein E1	Human papillomavirus type 55	dsDNA viruses, no RNA stage; Papillomaviridae; Alphapapillomavirus; Human papillomavirus 6; Human papillomavirus type 44; Human papillomavirus type 55	40	0.64
contig07533_6_1	1020181	replication protein E1	Human papillomavirus type 22	dsDNA viruses, no RNA stage; Papillomaviridae; Betapapillomavirus; Human papillomavirus 9	38.9	0.21
contig00382_6_1		Replication protein E1; DNA helicase, AAA+, ATPase, replication, initiator protein;	Bovine papillomavirus type 1	dsDNA viruses, no RNA stage; Papillomaviridae; Deltapapillomavirus; Bovine papillomavirus 1	95.77	0.057
contig53506_4_1		Replication protein E1; DNA helicase, AAA+, ATPase, replication, initiator protein; HET: ADP	Bovine papillomavirus type 1	dsDNA viruses, no RNA stage; Papillomaviridae; Deltapapillomavirus; Bovine papillomavirus 1	95.09	0.041
contig10302_3_1	296495831	replication protein E1	Human papillomavirus type 119	dsDNA viruses, no RNA stage; Papillomaviridae; Gammapapillomavirus; Human papillomavirus 112	35.4	2.3
contig66735_2_7	40556138	CNPV200 Rep-like protein	Canarypox virus	dsDNA viruses, no RNA stage; Poxviridae; Chordopoxvirinae; Avipoxvirus	59.7	6.00E-07

contig42337_1_1	41018595	ORF108 DNA packaging protein/ATPase	Orf virus	dsDNA viruses, no RNA stage; Poxviridae; Chordopoxvirinae; Parapoxvirus	39.3	0.35
contig11512_5_1		Poxvirus D5 protein-like;	Poxvirus	dsDNA viruses, no RNA stage; poxviridae; poxvirus	100	0
contig39131_1_2	45686095	Poxvirus A32 protein	Poxvirus	dsDNA viruses, no RNA stage; poxviridae; poxvirus	100	4.10E-37
contig13259_3_1		helicase-primase, but most likely cellular	Poxvirus D5 protein-like	dsDNA viruses, no RNA stage; poxviridae; poxvirus	100	0
contig06781_1_1	284504397	HNH endonuclease	Marseillevirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses	42.3	0.017
contig09166_4_1	134303403	occlusion-derived virus envelope-56 protein	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	236	2.00E-60
contig12819_3_2	134303461	ribonucleotide reductase small subunit	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	197	9.00E-49
contig12819_5_4	134303460	hypothetical protein GrBNV_gp62	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	44.7	0.008
contig13781_3_1	134303439	hypothetical protein GrBNV_gp41	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	152	3.00E-35
contig31500_1_1	134303432	hypothetical protein GrBNV_gp34	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	70.9	4.00E-11
contig33914_2_3	134303426	hypothetical protein GrBNV_gp28	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	42	0.07
contig41784_1_1	134303443	envelope protein P74	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	52	2.00E-05
contig68088_5_1	134303480	ribonucleotide reductase large subunit	Gryllus bimaculatus nudivirus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	38.1	0.35
contig16319_6_1	22788829	Orf123	Heliothis zea virus 1	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	133	8.00E-30
contig41784_3_1	22788862	p74 protein	Heliothis zea virus 1	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	51.6	3.00E-05
contig11131_1_1	213159320	lef-5	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	93.3	8.00E-18

contig11934_6_3	213159269	dnapol B	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	327	7.00E-88
contig14230_2_1	213159330	hypothetical protein OrNV_gp062	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	36.9	0.87
contig15667_5_1	108515125	DNA polymerase B	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	94.8	3.00E-18
contig16064_4_1	213159283	vp39	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	256	1.00E-66
contig16064_6_2	213159281	hypothetical protein OrNV_gp013	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	176	5.00E-42
contig16319_1_2	213159284	polh/gran	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	46.2	0.006
contig16436_4_2	213159355	38K protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	170	3.00E-42
contig30243_5_1	213159376	dnahel 2	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	49.3	7.00E-04
contig31940_1_1	213159314	GrBNV_gp22-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	77.4	2.00E-12
contig31940_2_2	213159314	GrBNV_gp22-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	70.9	1.00E-10
contig34436_3_2	213159392	hypothetical protein OrNV_gp124	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	98.6	1.00E-18
contig36247_4_1	213159332	lef-8	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	77.4	5.00E-13
contig36796_3_1	213159290	GrBNV_gp72-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	48.1	3.00E-04
contig37013_5_1	213159302	dnahel	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	112	6.00E-23
contig37728_5_1	213159374	vp91	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	67.8	4.00E-10
contig37728_5_3	213159374	vp91	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	51.2	1.00E-04
contig37728_5_4	213159375	pif-3	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	82.4	4.00E-14
contig37	213159374	vp91	Oryctes	dsDNA viruses, no RNA stage;	73.2	3.00E-

728_6_2			rhinoceros virus	unclassified dsDNA viruses; Nudivirus		11
contig39213_3_1	213159302	dnahel	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	73.2	1.00E-11
contig39525_2_1	213159328	pif-1	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	99.8	4.00E-19
contig40626_2_1	213159405	GrBNV_gp17-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	47	7.00E-04
contig40626_3_2	213159405	GrBNV_gp17-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	39.7	0.4
contig40627_1_1	213159301	19kda protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	62.4	2.00E-08
contig40627_6_1	213159397	hypothetical protein OrNV_gp129	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	50.4	7.00E-05
contig41186_3_1	213159376	dnahel 2	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	65.9	5.00E-09
contig43139_4_1	213159312	GrBNV_gp97-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	99	2.00E-19
contig43192_5_1	213159296	hypothetical protein OrNV_gp028	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	44.3	0.006
contig43559_3_1	213159290	GrBNV_gp72-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	73.2	1.00E-11
contig44129_2_1	213159400	GrBNV_gp48-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	79	2.00E-13
contig45395_3_1	213159332	lef-8	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	55.1	2.00E-06
contig45625_1_1	213159354	GrBNV_gp61-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	44.7	0.004
contig45625_2_3	213159348	GrBNV_gp60-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	41.2	0.044
contig45625_3_1	213159354	GrBNV_gp61-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	34.7	5
contig45625_3_2	213159348	GrBNV_gp60-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	54.3	5.00E-06
contig45874_3_4	108515104	late expression factor 4	Oryctes rhinoceros	dsDNA viruses, no RNA stage; unclassified dsDNA viruses;	65.5	2.00E-09

			virus	Nudivirus		
contig45874_6_3	213159307	GrBNV_gp93-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	37.7	0.42
contig47227_4_1	213159387	GrBNV_gp36-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	100	8.00E-20
contig47515_4_1	213159293	GrBNV_gp76-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	91.3	4.00E-17
contig47515_4_2	213159293	GrBNV_gp76-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	71.2	4.00E-11
contig47721_4_2	213159313	GrBNV_gp23-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	70.5	6.00E-11
contig48075_5_1	213159384	GrBNV_gp33-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	54.7	4.00E-06
contig48075_6_2	213159384	GrBNV_gp33-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	130	2.00E-28
contig48447_2_2	213159286	GrBNV_gp67-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	95.1	3.00E-18
contig50581_5_1	213159302	dnahel	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	40	0.15
contig51009_4_1	213159376	dnahel 2	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	39.7	0.65
contig51009_5_1	213159376	dnahel 2	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	46.6	0.001
contig56242_2_1	213159321	GrBNV_gp84-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus	61.2	4.00E-08
contig34436_3_2	213159392	?hypothetical protein OrNV_gp124	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus; Oryctes rhinoceros virus	98.6	1.00E-18
contig12567_2_1	213159382	??GrBNV_gp06-like protein unclassified	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus;Oryctes rhinoceros virus	47	0.01
contig14933_6_2	108515093	hypothetical protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus;Oryctes rhinoceros virus	346	3.00E-93
contig48002_1_1	213159297	GrBNV_gp81-like protein	Oryctes rhinoceros virus	dsDNA viruses, no RNA stage; unclassified dsDNA viruses; Nudivirus;Oryctes rhinoceros virus	62	2.00E-07
contig51009_4_1	213159376	? weak match to a less-conserved	Oryctes rhinoceros	dsDNA viruses, no RNA stage; unclassified dsDNA viruses;	43.2	0.055

		portion of helicase from insect nudivirus - needs further analysis???dnahel 2	virus	Nudivirus;Oryctes rhinoceros virus		
contig34054_3_1	50261337	minor core protein VP6	Palyam virus	dsRNA viruses; Reoviridae; Sedoreovirinae; Orbivirus	90.9	9.00E-17
contig39779_1_1	50261339	non structural protein NS1	Palyam virus	dsRNA viruses; Reoviridae; Sedoreovirinae; Orbivirus	136	3.00E-30
contig08744_3_1	118420305	VP2 protein	Bluetongue virus 1	dsRNA viruses; Reoviridae; Sedoreovirinae; Orbivirus; Bluetongue virus	261	4.00E-68
contig14397_2_1	4959681	nonstructural protein NS3/NS3A	Bluetongue virus 12	dsRNA viruses; Reoviridae; Sedoreovirinae; Orbivirus; Bluetongue virus	163	6.00E-39
contig46070_3_1	61654798	VP5 protein	Bluetongue virus 2	dsRNA viruses; Reoviridae; Sedoreovirinae; Orbivirus; Bluetongue virus	187	1.00E-45
contig03369_6_1	226423322	RNA-directed RNA polymerase	Homalodisca vitripennis reovirus	dsRNA viruses; Reoviridae; Sedoreovirinae; Phytoreovirus; unclassified Phytoreovirus	36.2	1.5
contig04898_6_1		GAG polyprotein; NMR	Human immunodeficiency virus type 1	Retro-transcribing viruses; Retroviridae; Orthoretrovirinae; Lentivirus; Primate lentivirus group	100	0
contig12479_1_1		polyprotein; HIV-1, integrase, domain organization	Human immunodeficiency virus type 1	Retro-transcribing viruses; Retroviridae; Orthoretrovirinae; Lentivirus; Primate lentivirus group	98.59	2.50E-08
contig03247_2_1	49256769	replication associated protein	Nanovirus-like particle	Satellites; Satellite Nucleic Acids; Single stranded DNA satellites; Begomovirus-associated alphasatellites	282	3.00E-74
contig32008_3_1	308275331	replication associated protein	Nanovirus-like particle	Satellites; Satellite Nucleic Acids; Single stranded DNA satellites; Begomovirus-associated alphasatellites	74.7	2.00E-11
contig08002_5_1		TT_ORF1 TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	2.70E-35
contig14322_1_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	99.85	.2.5e-21
contig15523_2_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0

contig15772_2_1		TT viral ORF2	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	99.96	5.70E-29
contig30199_2_2		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig30199_3_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig30379_2_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig30436_3_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	2.10E-42
contig31623_5_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig31635_2_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	1.90E-37
contig32236_5_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	4.20E-45
contig34478_6_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	4.90E-38
contig35321_3_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	94.25	0.1
contig46759_2_1		TT viral ORF2	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	99.88	2.10E-22
contig46980_4_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig51114_3_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig54892_5_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig55082_3_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	0
contig55085_4_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	1.30E-41
contig55325_3_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	3.10E-43
contig55564_1_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	100	1.30E-36
contig64360_6_1		TT viral orf 1	TT virus (TTV)	ssDNA viruses; Anelloviridae; unclassified Anelloviridae	98.6	1.10E-08
contig01729_1_1	225904214	putative capsid protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	224	4.00E-57
contig02720_6_1	46404804	putative replication-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	162	1.00E-38
contig04103_2_1	46404813	putative replication-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	193	9.00E-48
contig06	225904193	putative	Beak and	ssDNA viruses; Circoviridae;	42.4	0.049

131_3_1		replication-associated protein	feather disease virus	Circovirus		
contig17631_6_7	9630730	replication-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	147	4.00E-34
contig18923_3_2	67773351	replicase-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	35	7.6
contig31855_3_2	283777591	putative replication-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	50.1	1.00E-04
contig78068_4_2	46558867	putative replicase-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	50.1	2.00E-04
contig78070_5_1	46404814	putative coat protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	40.8	0.13
contig78122_2_2	46558861	putative replicase-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	40.8	0.14
contig78610_3_2	46558867	putative replicase-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	54.3	1.00E-05
contig78750_5_1	46369931	putative capsid protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	42.4	0.016
contig79175_4_2	46558861	putative replicase-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	52.8	4.00E-05
contig79486_6_2	225904193	putative replication-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus	67.4	5.00E-10
contig00269_4_3	46558867	putative replicase-associated protein	Beak and feather disease virus	ssDNA viruses; Circoviridae; Circovirus; Beak and feather disease virus	3.00E-66	256
contig31002_6_1	18875310	replicase	Canary circovirus	ssDNA viruses; Circoviridae; Circovirus	52.4	2.00E-05
contig00774_5_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	83.51	7
contig03309_6_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	77.54	15
contig05565_1_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	80.67	9.3
contig05		Circovirus ORF-2	Circovirus	ssDNA viruses; Circoviridae;	95.91	0.04

789_3_1		protein		Circovirus		
contig05952_1_1		Viral_Rep Putative viral replication protein. This is a family of viral ORFs from various plant and animal ssDNA circoviruses.	Circovirus	ssDNA viruses; Circoviridae; Circovirus	99.85	7.30E-21
contig06030_2_1		Circo_ORF2 Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	95.71	0.056
contig08584_1_1		Circo_ORF2: Circovirus ORF-2 protein;	Circovirus	ssDNA viruses; Circoviridae; Circovirus	81.35	7.5
contig09589_4_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	94.21	0.1
contig17310_2_1		Circo_ORF2; Circovirus ORF-2 protein;	Circovirus	ssDNA viruses; Circoviridae; Circovirus	88.28	0.95
contig18230_4_1		Circo_ORF2 Circovirus ORF-2 protein.	Circovirus	ssDNA viruses; Circoviridae; Circovirus	81.57	12
contig18572_2_1		Circo_ORF2 Circovirus ORF-2 protein.	Circovirus	ssDNA viruses; Circoviridae; Circovirus	78.64	1.6
contig20036_3_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	95.02	0.031
contig21667_4_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	97.42	0.00032
contig32889_5_1		Circo_ORF2 Circovirus ORF-2 protein.	Circovirus	ssDNA viruses; Circoviridae; Circovirus	89.81	8.2
contig78139_4_2		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	91.77	2.3
contig78274_5_4		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	94.32	0.13
contig78678_5_1		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	91.16	5.3
contig79784_5_2		Circovirus ORF-2 protein	Circovirus	ssDNA viruses; Circoviridae; Circovirus	97.31	0.0017
contig03874_1_1		Circovirus ORF-2 protein;	Circovirus	ssDNA viruses; Circoviridae; Circovirus	85.57	2.8
contig78678_6_3		Viral_Rep Putative viral replication protein.	Circovirus	ssDNA viruses; Circoviridae; Circovirus	90.79	0.032
contig13211_2_1	76009497	Rep	Goose circovirus	ssDNA viruses; Circoviridae; Circovirus	47	0.003

contig17631_6_8	50261991	rep	Porcine circovirus 1	ssDNA viruses; Circoviridae; Circovirus	45	0.002
contig33516_2_1	122891978	replicase	Porcine circovirus 1	ssDNA viruses; Circoviridae; Circovirus	39.7	0.11
contig00098_3_1		hw0_A Replicase; alpha+beta; NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	86.26	2.8
contig03303_4_2		Replicase; alpha+beta;NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	85.39	4.9
contig05127_6_1	73622369	rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	142	8.00E-33
contig06735_4_1	167832568	rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	180	5.00E-44
contig12474_6_1	217323269	capsid protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	221	3.00E-56
contig16953_1_1	57868104	capsid protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	248	5.00E-64
contig16953_4_1	164419583	Rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	256	1.00E-66
contig17769_2_1	121545854	rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	130	6.00E-29
contig17915_6_3	224381732	Rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	158	2.00E-37
contig17937_1_1	63259175	replicase	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	167	8.00E-40
contig18893_5_1	167832568	rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	151	3.00E-35
contig26785_5_1		Replicase; alpha+beta;	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	83.46	0.6
contig30124_3_2	228583494	capsid protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	38.5	0.045
contig33909_6_3		Replicase; alpha+beta; NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	98.07	2.80E-06
contig36857_2_1		Replicase; alpha+beta; NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	73.1	10
contig39461_5_1		Replicase; alpha+beta; NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	92.68	0.036
contig50879_5_1		Replicase; alpha+beta; NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	100	2.00E-36
contig78337_1_5	162950972	rep protein	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	48.1	0.002
contig79103_1_2		Replicase; alpha+beta;	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	90.11	0.71
contig79744_6_1		Replicase; alpha+beta; NMR	Porcine circovirus 2	ssDNA viruses; Circoviridae; Circovirus	98.21	2.20E-07

contig26523_1_1		Putative viral replication protein	This is a family of viral ORFs from various plant and animal ssDNA circoviruses	ssDNA viruses; Circoviridae; Circovirus;	99.77	1.10E-18
contig78884_6_1		Putative viral replication protein	This is a family of viral ORFs from various plant and animal ssDNA circoviruses	ssDNA viruses; Circoviridae; Circovirus;	100	0
contig79161_2_1		Putative viral replication protein	This is a family of viral ORFs from various plant and animal ssDNA circoviruses	ssDNA viruses; Circoviridae; Circovirus;	100	0
contig01618_5_1	15141811	replication associated protein	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	170	6.00E-41
contig02524_2_1	118201949	replication-associated protein	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	159	3.00E-37
contig02797_3_3	74136915	Rep	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	124	3.00E-27
contig03706_1_1	15141804	replication associated protein	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	186	8.00E-46
contig04504_3_2	74136915	Rep	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	188	1.00E-46
contig18314_5_1	15141804	replication associated protein	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	161	2.00E-38
contig18563_2_1	118201941	replication-associated protein	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	190	4.00E-47
contig19401_2_1	15141811	replication associated protein	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	239	7.00E-62
contig20043_4_2	74136915	Rep	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	221	1.00E-55
contig20636_3_1	77735205	putative replication associated protein	Duck circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	57.4	6.00E-07

contig00448_3_1	116630444	replication-associated protein	Finch circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	132	2.00E-29
contig02140_5_3	116630444	replication-associated protein	Finch circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	39.3	0.13
contig03639_6_1	116630444	replication-associated protein	Finch circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	140	5.00E-32
contig78066_2_3	116630444	replication-associated protein	Finch circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	56.6	3.00E-06
contig50348_6_1	116630440	replication-associated protein	Gull circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	73.9	1.00E-11
contig79340_5_1	116630440	replication-associated protein	Gull circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	41.2	0.13
contig24324_5_1	257815101	rep	Muscovy duck circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	71.6	3.00E-11
contig04740_6_2	115334608	rep protein	Raven circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	131	2.00E-29
contig09808_1_1	115334608	rep protein	Raven circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	180	6.00E-44
contig20268_4_2	115334608	rep protein	Raven circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus	38	0.29
contig66735_2_7	74136915	Rep	Columbid circovirus	ssDNA viruses; Circoviridae; Circovirus; unclassified Circovirus; Columbid circovirus	3.00E-06	57.4
contig22820_2_1	297598949	putative Rep	Circoviridae TM-6c	ssDNA viruses; Circoviridae; unclassified Circoviridae	62	2.00E-08
contig26117_5_4	297598949	putative Rep	Circoviridae TM-6c	ssDNA viruses; Circoviridae; unclassified Circoviridae	76.3	2.00E-12
contig46732_6_1	297598949	putative Rep	Circoviridae TM-6c	ssDNA viruses; Circoviridae; unclassified Circoviridae	64.7	4.00E-09
contig47444_4_1	297598949	putative Rep	Circoviridae TM-6c	ssDNA viruses; Circoviridae; unclassified Circoviridae	52.4	2.00E-05
contig53244_6_1	297598949	putative Rep	Circoviridae TM-6c	ssDNA viruses; Circoviridae; unclassified Circoviridae	55.8	1.00E-06
contig00766_3_4	290783641	replication-association protein	Cyclovirus NG12	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	43.1	0.02
contig18911_3_1	290783641	replication-association protein	Cyclovirus NG12	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	52	5.00E-05
contig38714_6_1	290783644	replication-association protein	Cyclovirus NG14	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	63.5	7.00E-09
contig41649_4_1	290783644	replication-association protein	Cyclovirus NG14	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	38.5	0.24

contig42833_1_1	290783644	replication-association protein	Cyclovirus NG14	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	62.4	2.00E-08
contig61767_5_1	290783644	replication-association protein	Cyclovirus NG14	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	65.1	3.00E-09
contig64667_4_1	290783644	replication-association protein	Cyclovirus NG14	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	45.1	0.003
contig79687_6_1	290783644	replication-association protein	Cyclovirus NG14	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	61.6	3.00E-08
contig16142_4_1	290783611	replication-association protein	Cyclovirus PK5006	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	134	2.00E-30
contig03140_6_2	290783623	replication-association protein	Cyclovirus PK6197	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	208	4.00E-52
contig78966_4_1	290783653	replication-association protein	Cyclovirus TN18	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	44.7	0.01
contig74365_5_1	290783650	replication-association protein	Cyclovirus TN25	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	63.2	1.00E-08
contig79244_5_1	290783650	replication-association protein	Cyclovirus TN25	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	62.8	1.00E-08
contig79535_1_1	290783650	replication-association protein	Cyclovirus TN25	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	59.7	1.00E-07
contig79688_3_1	290783650	replication-association protein	Cyclovirus TN25	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	60.1	9.00E-08
contig41263_3_1	295148501	replication-association protein	Cyclovirus TN26	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	47.8	5.00E-04
contig08191_3_1	290783648	capsid protein	Human stool-associated circular virus NG13	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	42	0.025
contig08420_2_1	290783647	replication-association protein	Human stool-associated circular virus NG13	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	44.2	0.05
contig10280_1_1	290783648	capsid protein	Human stool-associated circular virus NG13	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	39.6	0.13
contig16	290783648	capsid protein	Human	ssDNA viruses; Circoviridae;	427	0.014

735_6_1			stool-associated circular virus NG13	unclassified Circoviridae; suggested genus Cyclovirus		
contig18462_6_1	290783648	capsid protein	Human stool-associated circular virus NG13	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	45.1	0.01
contig24764_2_1	290783647	replication-association protein	Human stool-associated circular virus NG13	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	42.4	0.018
contig78166_1_1	290783647	replication-association protein	Human stool-associated circular virus NG13	ssDNA viruses; Circoviridae; unclassified Circoviridae; suggested genus Cyclovirus	45.8	0.002
contig41063_6_1		Galleria mellonella densovirus capsid protein	Galleria mellonella densovirus	ssDNA viruses; Parvoviridae; Densovirinae; Densovirus	96.12	0.0067
contig11482_6_1	23334609	structural protein VP1	Galleria mellonella densovirus	ssDNA viruses; Parvoviridae; Densovirinae; Densovirus	215	1.00E-54
contig33665_2_1		DNA replication protein; AAA+ protein, P-loop atpases, helicase; HET: DNA ADP;	Adeno-associated virus 2	ssDNA viruses; Parvoviridae; Parvovirinae; Dependovirus	90.12	0.09
contig71155_3_1		DNA replication protein; AAA+ protein, P-loop atpases, helicase; HET: DNA ADP	Adeno-associated virus 2	ssDNA viruses; Parvoviridae; Parvovirinae; Dependovirus	95.71	0.026
contig42335_2_1	2766609	nonstructural protein Rep78	adeno-associated virus 3B	ssDNA viruses; Parvoviridae; Parvovirinae; Dependovirus; Adeno-associated virus - 3	37	1.1
contig03253_2_2	283488695	putative replicase protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	43.1	0.011
contig03	283488693	putative capsid	Chimpanzee	ssDNA viruses; unclassified ssDNA	180	6.00E-

253_6_1		protein	stool associated circular ssDNA virus	viruses		44
contig04675_3_1	283488708	putative replicase protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	35.7	1.6
contig17637_4_3	283488716	putative capsid protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	458	1.00E-127
contig33271_4_1	283488692	putative replicase protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	47.7	8.00E-04
contig37567_6_2	283488716	putative capsid protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	70.1	4.00E-10
contig38694_6_1	283488695	putative replicase protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	73.9	6.00E-12
contig39169_6_1	283488703	putative capsid protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	44.7	0.011
contig59642_3_3	283488696	putative capsid protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses	36.2	1.4
contig37567_6_2	283488716	putative capsid protein	Chimpanzee stool associated circular ssDNA virus	ssDNA viruses; unclassified ssDNA viruses;	70	4.00E-10
contig02281_5_1	254688516	hypothetical protein	Circovirus-like genome BBC-A	ssDNA viruses; unclassified ssDNA viruses	40.1	0.089
contig17199_2_1	254688516	hypothetical protein	Circovirus-like genome BBC-A	ssDNA viruses; unclassified ssDNA viruses	38.6	0.29
contig02	254688534	putative Rep	Circovirus-	ssDNA viruses; unclassified ssDNA	90.4	6.00E-

671_1_1		protein	like genome CB-A	viruses		17
contig17773_3_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	50.1	2.00E-04
contig21167_2_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	35	3.1
contig27423_6_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	105	4.00E-21
contig30864_5_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	36.2	1.3
contig35517_6_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	57.8	3.00E-06
contig39591_2_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	55.5	2.00E-06
contig48273_3_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	37.7	0.42
contig64087_2_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	63.5	8.00E-09
contig79158_2_1	254688534	putative Rep protein	Circovirus-like genome CB-A	ssDNA viruses; unclassified ssDNA viruses	37.4	0.51
contig00621_6_2	254729602	hypothetical protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	34.2	4.5
contig05565_2_1	254729603	putative Rep-associated protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	44.7	0.07
contig07808_6_1	254729602	hypothetical protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	265	4.00E-69
contig30885_6_1	254729602	hypothetical protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	42.7	0.013
contig32845_3_1	254729602	hypothetical protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	53.5	4.00E-05
contig33297_2_3	254729603	putative Rep-associated protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	42	0.07
contig78799_2_2	254729603	putative Rep-associated protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses	52	9.00E-05
contig01080_2_1	254688519	hypothetical protein	Circovirus-like genome	ssDNA viruses; unclassified ssDNA viruses	44.7	0.004

			RW-A			
contig07107_6_1	254688519	hypothetical protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	40.4	0.064
contig17125_4_1	254688518	putative Rep protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	123	5.00E-27
contig17469_1_1	254688518	putative Rep protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	146	9.00E-34
contig22824_3_1	254688518	putative Rep protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	82.8	5.00E-14
contig23791_4_1	254688519	hypothetical protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	44.7	0.015
contig39678_5_1	254688518	putative Rep protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	59.3	1.00E-07
contig64662_3_2	254688518	putative Rep protein	Circovirus-like genome RW-A	ssDNA viruses; unclassified ssDNA viruses	64.7	4.00E-09
contig00451_6_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	37.8	0.48
contig01651_4_2	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	39.7	0.14
contig03299_3_2	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	36.9	2.1
contig03639_4_2	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	36.9	0.79
contig03901_5_3	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	41.2	0.15
contig05970_3_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	40.8	0.11
contig06047_1_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	37.4	0.77
contig06229_4_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	175	1.00E-42
contig07933_1_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	43.6	0.021
contig17606_2_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	42.3	0.054

contig18036_5_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	180	3.00E-44
contig22033_4_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	44.7	0.009
contig22712_6_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	45.1	0.005
contig29258_1_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	38.5	0.24
contig30896_4_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	38.5	0.48
contig32567_5_2	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	197	3.00E-49
contig33064_1_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	42	0.091
contig36221_2_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	42	0.023
contig65252_5_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	50.8	5.00E-05
contig78128_1_2	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	39.7	0.11
contig78656_2_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	40.8	0.15
contig78747_1_2	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	74.3	2.00E-11
contig78863_4_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	40	0.11
contig78972_6_1	254688523	hypothetical protein	Circovirus-like genome RW-B	ssDNA viruses; unclassified ssDNA viruses	53.5	2.00E-05
contig00098_1_2	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	42.3	0.041
contig00473_5_2	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	40.8	0.2
contig00527_1_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	36.6	2.4
contig00	254688526	hypothetical	Circovirus-	ssDNA viruses; unclassified ssDNA	35	9.8

767_4_1		protein	like genome RW-C	viruses		
contig01026_3_2	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	43.5	0.015
contig02692_3_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	40.9	0.047
contig03208_6_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	44.3	0.01
contig05148_4_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	45.8	0.004
contig05963_3_3	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	190	4.00E-47
contig11715_4_1	254688525	putative Rep protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	37.7	0.47
contig17220_1_2	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	44.3	0.05
contig30156_4_1	254688525	putative Rep protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	53.5	1.00E-05
contig33909_4_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	44.7	0.01
contig37580_4_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	55.8	2.00E-06
contig78265_3_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	79.7	2.00E-13
contig78755_6_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	38.5	0.52
contig78807_5_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	40	0.24
contig79444_2_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	56.6	2.00E-06
contig79462_5_1	254688526	hypothetical protein	Circovirus-like genome RW-C	ssDNA viruses; unclassified ssDNA viruses	88.6	4.00E-16
contig02914_3_1	254688528	hypothetical protein	Circovirus-like genome RW-D	ssDNA viruses; unclassified ssDNA viruses	42	0.031
contig03140_1_2	254688528	hypothetical protein	Circovirus-like genome	ssDNA viruses; unclassified ssDNA viruses	45.8	0.08

			RW-D			
contig03713_3_1	254688528	hypothetical protein	Circovirus-like genome RW-D	ssDNA viruses; unclassified ssDNA viruses	249	2.00E-64
contig17054_6_1	254688528	hypothetical protein	Circovirus-like genome RW-D	ssDNA viruses; unclassified ssDNA viruses	37.3	0.56
contig17448_5_1	254688532	hypothetical protein	Circovirus-like genome RW-E	ssDNA viruses; unclassified ssDNA viruses	270	4.00E-71
contig23184_1_1	254688531	putative Rep protein	Circovirus-like genome RW-E	ssDNA viruses; unclassified ssDNA viruses	35	3.1
contig26964_1_1	254688531	putative Rep protein	Circovirus-like genome RW-E	ssDNA viruses; unclassified ssDNA viruses	46.6	0.001
contig67760_3_1	254688531	putative Rep protein	Circovirus-like genome RW-E	ssDNA viruses; unclassified ssDNA viruses	115	1.00E-24
contig04394_1_2	254688512	hypothetical protein	Circovirus-like genome SAR-B	ssDNA viruses; unclassified ssDNA viruses	268	1.00E-69
contig05808_4_1	254688512	hypothetical protein	Circovirus-like genome SAR-B	ssDNA viruses; unclassified ssDNA viruses	41.6	0.12
contig08111_4_1	254688512	hypothetical protein	Circovirus-like genome SAR-B	ssDNA viruses; unclassified ssDNA viruses	39.7	0.42
contig78672_1_1	254688512	hypothetical protein	Circovirus-like genome SAR-B	ssDNA viruses; unclassified ssDNA viruses	40.8	0.29
contig03140_1_2	254688528	hypothetical protein from a circovirus RW-D, the part missing in this peptide but found in another one has HxH motif,maybe rolling-circle replication initiator	Circovirus-like genome RW-D	ssDNA viruses; unclassified ssDNA viruses;	315	4.00E-84
contig17773_4_2	254688528	?match to a circovirus protein, but composition problems	Circovirus-like genome RW-D	ssDNA viruses; unclassified ssDNA viruses;	43	0.096
contig04394_1_2	254688512	hypothetical protein from a circovirus RW-D, HxH	Circovirus-like genome SAR-B	ssDNA viruses; unclassified ssDNA viruses;	232	6.00E-59

		motif,maybe rolling-circle replication initiator///hypothetical protein				
contig08165_3_1	254729602	?hypothetical protein	Circovirus-like genome CB-B	ssDNA viruses; unclassified ssDNA viruses;	296	4.00E-78
contig01656_1_1	62956490	Phosphoprotein	Rabies virus	ssRNA viruses; ssRNA negative-strand viruses; Mononegavirales; Rhabdoviridae; Lyssavirus	40.5	0.15
contig13431_3_1	15213612	glycoprotein	rabies virus	ssRNA viruses; ssRNA negative-strand viruses; Mononegavirales; Rhabdoviridae; Lyssavirus	273	4.00E-72
contig06068_5_1	253756677	orf1a polyprotein	Feline coronavirus UU9	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nidovirales; Coronaviridae; Coronavirinae; Alphacoronavirus; Alphacoronavirus 1; Feline coronavirus	42.4	0.02
contig44391_3_1	20260794	protein A	Pariacato virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Alphanodavirus	42.7	0.022
contig29860_5_1	20260794	protein A, RNA-dependent RNA polymerase	Pariacato virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Alphanodavirus; Pariacato virus	77.4	3.00E-12
contig05514_5_1	295314691	coat protein	Alphanodavirus HB-2007/CHN	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Alphanodavirus; unclassified Alphanodavirus	95.9	2.00E-18
contig41979_5_1	295314693	RNA-dependent RNA polymerase	Alphanodavirus HB-2007/CHN	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Alphanodavirus; unclassified Alphanodavirus	47	8.00E-04
contig29434_1_1	225580032	RNA-dependent RNA polymerase	Redspotted grouper nervous necrosis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Betanodavirus	59.3	2.00E-07
contig33983_6_1	56967887	RNA-dependent RNA polymerase	Redspotted grouper nervous necrosis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Betanodavirus	45.4	0.004
contig34	225580032	RNA-dependent	Redspotted	ssRNA viruses; ssRNA positive-	63.2	1.00E-

426_4_1		RNA polymerase	grouper nervous necrosis virus	strand viruses, no DNA stage; Nodaviridae; Betanodavirus		08
contig38476_4_1	81337918	capsid precursor protein	Wuhan nodavirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Betanodavirus; unclassified Betanodavirus	101	2.00E-19
contig04709_3_1	81337918	capsid precursor protein	Wuhan nodavirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Betanodavirus; unclassified Betanodavirus;Wuhan nodavirus	92	1.00E-16
contig16051_6_1	81337918	coat protein	Wuhan nodavirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; Betanodavirus; unclassified Betanodavirus;Wuhan nodavirus	95	2.00E-17
contig02861_6_2	297598937	Viral RNA-directed RNA-polymerase	Bat guano associated nodavirus GF-4n	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; unclassified Nodaviridae;Bat guano associated nodavirus GF-4n	147	6.00E-33
contig31667_3_1	297598937	putative RNA-dependent RNA polymerase	Bat guano associated nodavirus GF-4n	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; unclassified Nodaviridae;Bat guano associated nodavirus GF-4n	79	1.00E-12
contig34923_5_1	297598937	RNA-dependent RNA polymerase	Bat guano associated nodavirus GF-4n	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; unclassified Nodaviridae;Bat guano associated nodavirus GF-4n	70	8.00E-10
contig41274_3_1	297598937	RNA-dependent RNA polymerase	Bat guano associated nodavirus GF-4n	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Nodaviridae; unclassified Nodaviridae;Bat guano associated nodavirus GF-4n	63.5	9.00E-08
contig20074_3_1	162417285	Polyprotein	cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Cardiovirus; unclassified Cardiovirus	170	6.00E-41
contig12177_5_1	21321710	structural polyprotein	Cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	101	3.00E-20

contig37283_5_2	21321709	nonstructural polyprotein	Cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	45.8	0.002
contig41105_2_1	21321709	nonstructural polyprotein	Cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	51.6	1.00E-04
contig41530_3_1	21321710	structural polyprotein	Cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	36.2	1.3
contig50662_3_1	21321709	nonstructural polyprotein	Cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	33.1	10
contig51108_1_1	21321709	nonstructural polyprotein	Cricket paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	72	2.00E-11
contig27668_6_2		Insect picorna-like virus coat proteins	CRPV (Cricket paralysis virus)	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	95.18	0.04
contig42971_6_2		Insect picorna-like virus coat proteins	CRPV (Cricket paralysis virus)	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	96.83	0.0089
contig31601_5_1	9629651	replicase polyprotein	Drosophila C virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	42.4	0.018
contig32622_2_1	9629651	replicase polyprotein	Drosophila C virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	68.9	1.00E-09
contig53173_2_1	9629652	capsid polyprotein	Drosophila C virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	62.4	5.00E-08
contig32622_2_1	9629651	replicase polyprotein	Drosophila C virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus;	68.9	1.00E-09
contig41498_2_1	9629651	replicase polyprotein	Drosophila C virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus;	0.33	41.6
contig32832_3_1	20451019	nonstructural protein	Himetobi P virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage;	47	7.00E-04

		precursor		Picornavirales; Dicistroviridae; Cripavirus		
contig33030_6_2	50251149	nonstructural protein precursor	Himetobi P virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	47	0.001
contig53879_4_1	50251149	nonstructural protein precursor	Himetobi P virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; Cripavirus	40.8	0.065
contig36782_5_1	283979151	non-structural polyprotein	Taura syndrome virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae	35.8	1.9
contig41505_5_1	66775618	non-structural polyprotein	Taura syndrome virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae	51.6	3.00E-05
contig42395_2_2	78057527	capsid protein precursor	Taura syndrome virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae	37.4	1.2
contig17525_6_1	283979151	non-structural polyprotein, RNA dependent RNA polymerase domain	Taura syndrome virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae;Taura syndrome virus	75.5	5.00E-11
contig31532_6_2	119372481	non-structural polyprotein, RNA helicase P-loop domain	Taura syndrome virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae;Taura syndrome virus	63.5	6.00E-08
contig12035_6_1	187234322	polymerase polyprotein	Israel acute paralysis virus of bees	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unclassified Dicistroviridae	41.2	0.11
contig14647_2_1	165906128	structural polyprotein	Israel acute paralysis virus of bees	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unclassified Dicistroviridae	218	5.00E-55
contig31601_6_1	165906127	non-structural polyprotein	Israel acute paralysis virus of bees	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unclassified Dicistroviridae	78.2	7.00E-13
contig42925_1_1	187234328	polymerase polyprotein	Israel acute paralysis virus of bees	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unclassified Dicistroviridae	64.3	1.00E-08

contig26471_6_2	156563985	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	88.6	1.00E-15
contig27682_6_1	156563986	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	99	2.00E-19
contig30776_1_1	156563985	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	95.9	1.00E-18
contig46635_5_1	156563986	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	62	2.00E-08
contig48479_5_1	156563986	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	51.2	4.00E-05
contig61848_4_1	156563985	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	115	8.00E-24
contig79647_2_1	156563985	hypothetical protein	Marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	63.9	6.00E-09
contig10273_4_1	156563988	hypothetical protein	Marine RNA virus JP-B	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	44.2	0.004
contig12174_6_2	296005647	polyprotein	Slow bee paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Iflaviridae; Iflavirus; unclassified Iflavirus	42.3	0.017
contig16622_3_1	297578409	polyprotein	Slow bee paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Iflaviridae; Iflavirus; unclassified Iflavirus	45	0.013
contig22807_3_1	46810913	polyprotein, RNA-helicase domain	Satsuma dwarf virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Secoviridae; Sadwavirus	504	1.00E-140
contig12174_6_2	296005647	polyprotein	Slow bee paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Iflaviridae; Iflavirus; unclassified Iflavirus	42.3	0.017

contig16622_3_1	297578409	polyprotein	Slow bee paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Iflaviridae; Iflavirus; unclassified Iflavirus	45	0.013
contig02778_2_1		Rhinovirus coat proteins	Human rhinovirus A 2 (HRV-2)	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Enterovirus	80.35	3.2
contig03976_4_1	217316369	Polyprotein; ATPase from circovirus!	circovirus	ssDNA viruses; Circoviridae	215	2.00E-54
contig09506_3_1	95102514	Polyprotein	Taura Syndrom Virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae	302	9.00E-81
contig19570_1_1	217316359	Polyprotein	Dicstrovirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales	270	7.00E-71
contig24777_6_1	409464	polyprotein	marine RNA virus JP-A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; environmental samples	299	1.00E-78
contig09271_6_2	9626736	Polyprotein	Taura Syndrom Virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Dicistroviridae; unassigned Dicistroviridae	301	4.00E-80
contig07475_4_1	255682334	Polyprotein	Human rhinovirus C	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Enterovirus; unclassified Rhinovirus	42.7	0.015
contig30096_3_1	27657699	polyprotein, 3C cysteine protease	Dicstrovirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales	294	3.00E-77
contig32693_1_1	56799471	polyprotein	Human enterovirus 76	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Enterovirus; Human enterovirus A	41.6	0.038
contig11831_3_1	121484964	Polyprotein	European brown hare syndrome virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Caliciviridae; Lagovirus	227	4.00E-58
contig09673_4_1	162417285	Polyprotein	Chaetoceros tenuissimus RNA virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales;	234	2.00E-59

				Bacillariornaviridae; Bacillariornavirus		
contig05171_3_1	226876797	nonstructural polyprotein, RNA helicase, P3A, peptidase_C domain	Chaetoceros socialis f. radians RNA virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Bacillariornaviridae; Bacillariornavirus	0	704
contig12042_5_1	33309090	Polyprotein	Chaetoceros socialis f. radians RNA virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Bacillariornaviridae; Bacillariornavirus	263	7.00E-69
contig40824_3_1			Mycovirus FusoV	dsRNA viruses; Partitiviridae; unclassified Partitiviridae	90.84	0.04
contig35517_6_1	115430549	polyprotein, RNA helicase P-loop domain	Human poliovirus 1	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Enterovirus; Human enterovirus C;Human poliovirus 1 / replication assiciated protein, Circovirus-like genome	358	8.00E-97
contig42395_3_2	189170126	polyprotein	Human parechovirus 7	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Parechovirus; Human parechovirus	41.6	0.03
contig21996_4_2	209363551	polyprotein, RNA helicase containing P-loop NTPase damain	Seneca valley virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Senecavirus;Seneca valley virus	64.7	6.00E-08
contig16622_2_1	209363551	polyprotein, RNA helicase P-loop NTPase domain	Seneca valley virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; Picornaviridae; Senecavirus;Seneca valley virus/???Picornavirales; Picornaviridae; Parechovirus; Human parechovirus;Human parechovirus 3	69	1.00E-09
contig29860_5_1	302171582	unknown	Tetnovirus 1	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; unclassified Picornavirales	84	3.00E-14
contig43478_3_1	302171582	unknown	Tetnovirus 1	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Picornavirales; unclassified	38.5	0.26

				Picornavirales		
contig00925_1_6		1ng0_A Coat protein; sobemovirus, virus assembly;	Cocksfoot mottle virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Sobemovirus	93.45	0.1
contig03442_4_1		Sobemovirus coat protein	Cocksfoot mottle virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Sobemovirus	86.42	10
contig34361_2_1	48697149	putative RNA-dependent RNA polymerase	Dendrolimus punctatus tetravirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Tetraviridae; Omegatetravirus; unclassified Omegatetravirus	40	0.36
contig34361_3_1	48697149	putative RNA-dependent RNA polymerase	Dendrolimus punctatus tetravirus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; Tetraviridae; Omegatetravirus; unclassified Omegatetravirus	65.9	2.00E-09
contig05867_6_1	45476497	coat protein	Sclerophthora macrospora virus A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; unclassified ssRNA positive-strand viruses	40.8	0.097
contig31455_6_1	45476497	coat protein	Sclerophthora macrospora virus A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; unclassified ssRNA positive-strand viruses	42	0.022
contig38143_2_1	45476494	putative RNA dependent RNA polymerase	Sclerophthora macrospora virus A	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; unclassified ssRNA positive-strand viruses	48.5	3.00E-04
contig06254_6_1	226821746	RNA-dependent RNA polymerase	Chronic bee paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; unclassified ssRNA positive-strand viruses;Chronic bee paralysis virus	64	6.00E-08
contig11882_6_1	226821754	RNA-dependent RNA polymerase	Chronic bee paralysis virus	ssRNA viruses; ssRNA positive-strand viruses, no DNA stage; unclassified ssRNA positive-strand viruses;Chronic bee paralysis virus	77	6.00E-12
contig00596_3_2	301070440	coat protein	Plasmopara halstedii virus A	unclassified viruses	44.6	0.04
contig04762_3_1	301070442	coat protein	Plasmopara halstedii virus A	unclassified viruses	165	1.00E-39
contig31812_3_1	301070430	RNA-dependent RNA polymerase	Plasmopara halstedii virus A	unclassified viruses	65.1	3.00E-09
contig36	157087440	putative coat	Plasmopara	unclassified viruses	72.4	9.00E-

501_5_2		protein	halstedii virus A			11
contig37408_1_1	157087442	putative coat protein	Plasmopara halstedii virus A	unclassified viruses	42.4	0.016
contig40554_5_1	301070432	RNA-dependent RNA polymerase	Plasmopara halstedii virus A	unclassified viruses	66.2	1.00E-09
contig37493_3_1	295148403	replication-association protein	Virus PK1142	unclassified viruses	80.1	9.00E-14
contig22643_1_2	295148421	replication-association protein	Virus PK1813	unclassified viruses	44.7	0.004