Combined assembly of metagenomic libraries from the stool samples of IBD and PSC patients allowed the identification of African swine fever virus-like sequences

by

Md Salman Reza

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science
in
Experimental Medicine

Department of Medicine

University of Alberta

ABSTRACT

Metagenomics is an emerging discipline to explore microbial diversity in clinical samples, independent of the limitations of cell cultures. This method is widely accepted as a modern technique to detect novel viruses in clinical and environmental samples and has the potential to contribute to the identification of unknown etiological agents causing diseases. This thesis utilizes the combined assembly approach on deep sequencing data in search for novel viral populations in clinical samples of patients diagnosed with inflammatory bowel disease (IBD) and primary sclerosing cholangitis (PSC), which is an extraintestinal manifestation of IBD. In this study, twelve mammalian viral sequences and other plant, bird and insect viral sequences were found. Among the viral sequences, we have detected many sequences of African swine fever like virus (ASFLV) in the IBD and PSC patients after the combined assembly approach but not in the initial metagenomics data sets of the assembly from the individual patient. Interestingly, the ASFLV sequences showed similarity to thirty-nine genes along the ASFV genome, but only 38-62% identity at the amino acid level, suggesting that they are related by distinct sequences. Phylogenetic analyses positioned the ASFLV sequences in a clade different from those clustering ASFV and that was consistent for the topoisomerase, capsid and helicase genes. Furthermore, nested PCR confirmed the presence of ASFLV sequences in one ulcerative colitis and PSC patients for multiple loci including the capsid, helicase and origin binding proteins. Thus, we report for the first time the presence of ASFLV sequences in human clinical samples in North America. This study also investigates the suitability of several viral enrichment methods to isolate viruses from clinical samples before performing metagenomics. The detection of ASFLV sequences was possible only after adopting the combined assembly approach, which enabled the identification of a novel virus sequences and improved the overall identification of viruses in the metagenomic libraries.

Dedication

To my beloved parents

Acknowledgements

All praises due to Allah, the Creator of the Universes, in whom I believe and towards Him I show my utmost gratitude for making me able to complete this thesis work. I always seek refuge and support in Him during all my good times and the bad.

I would like to thank my supervisor, Dr. Gane Ka-Shu Wong for his constant guidance and support. Dr. Wong is an enthusiastic and dedicated scientist and a great human being. Working under his supervision, has been a great boost in my research career. His critics and discussions have been really helpful throughout my research work.

I would like to thank my co-supervisor, Dr. Andrew L Mason for helping me immensely throughout my work. I really appreciate his motivation, enthusiasm and intelligence. His invaluable suggestions for progress and troubleshooting provided me a great support. I am truly grateful to him.

I would like to thank Dr. David Marchant for his valuable comments and the time being on my supervisory committee.

I acknowledge Dr. Weiwei Wang and Dr. Juan Jovel for teaching me different techniques and helping me carrying out experiments, proof-reading and troubleshooting. Their faith in my abilities gave me confidence and motivation when sometimes things went wrong.

M.Sc. Jordan Patterson in the Wong's lab has developed the metagenomics pipeline and helped me in many steps of bioinformatics. I would like to thank my colleagues in Dr. Wong's and Mason's labs for their long hours of discussions, planning, helping in my experiments and encouragement and advice throughout the years.

Finally, I would like to thank my family and loved ones for their constant support and belief in my abilities. I would also like to give a special thanks to Simrika Thapa for being willing to listen to the gripes and rants while offering me her sound advice.

CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1. VIRAL METAGENOMICS TO DISCOVER NOVEL VIRUS 1
1.2. ASSEMBLY APPROACH FOR VIRAL METAGENOMICS 3
1.3. INFLAMMATORY BOWEL DISEASE (IBD) 5
1.4. PRIMARY SCLEROSING CHOLANGITIS (PSC) 6
1.5. VIRUSES IN AUTOIMMUNE DISEASES 7
1.6. HYPOTHESIS AND SPECIFIC AIMS 8
1.6.1. HYPOTHESIS 8

CHAPTER 2: MATERIALS AND METHODS 10

2.1. ETHICS APPROVAL 10
2.2. DESCRIPTION OF PATIENTS 10
2.3. VIRAL PURIFICATION FROM FAECAL SAMPLES 11
2.3.1. TANGENTIAL FLOW FILTRATION (TFF) METHOD 11
2.3.2. ULTRAFREE MC (MICROCENTRIFUGATION) METHOD 12
2.3.3. GLASS MILK/ SILICON DIOXIDE METHOD 12
2.4. CONSTRUCTION OF LIBRARY USING TRUSEQ KIT 13
2.4.1. PRE-PROCESSING OF RNA: STEPS PRIOR TO THE CONSTRUCTION OF METAGENOMIC LIBRARIES FROM RNA 13
2.4.2. PREPROCESSING OF DNA: STEPS PRIOR TO THE CONSTRUCTION OF METAGENOMIC LIBRARIES FROM DNA 14
2.4.3. COMMON STEPS FOR METAGENOMIC LIBRARIES CONSTRUCTION 16
2.5. DATA ANALYSIS OF INDIVIDUALLY SEQUENCED METAGENOMIC LIBRARIES 17
2.6. USING COMBINED ASSEMBLY FOR THE ANALYSIS OF ALL METAGENOMIC LIBRARIES FROM DIFFERENT SAMPLES 20
2.7. APPLICATION OF OTHER SOFTWARE 21
2.8. VERIFICATION OF VIRUS PRESENCE: NESTED PCR PROCEDURE 22

CHAPTER 3: ANALYSIS OF INDIVIDUAL METAGENOMICS LIBRARIES 23

3.1. INTRODUCTION 23
3.1.1. TAXONOMIC IDENTIFICATION OF VIRAL SEQUENCES IN THE METAGENOMICS LIBRARY FROM CD, UC AND PSC STOOL SAMPLES 24
3.2. RESULTS 25
3.2.1. DETECTION OF VIRAL SEQUENCES BY BLASTN SEARCH 25
3.2.2. DETECTION OF VIRAL SEQUENCES BY BLASTX SEARCH 28
3.3. KEY FINDINGS 30

CHAPTER 4: COMBINED ASSEMBLY IN THE IDENTIFICATION OF VIRUSES IN IBD AND PSC PATIENTS 31

4.1. INTRODUCTION 31
4.2. COMBINED ASSEMBLY GENERATES LONGER SCAFFOLDS 32
4.3. TAXONOMIC CLASSIFICATION OF SEQUENCED DATA 33
4.4. DETECTED SEQUENCES OF VIRUS IN THE METAGENOMICS LIBRARIES AFTER COMBINED ASSEMBLY 36
4.5. DISTRIBUTION OF VIRAL READS IN IBD AND PSC PATIENTS 39
4.6. EXAMPLES OF THE FULL OR PARTIAL ALIGNMENT OF CONTIGS/SCAFFOLDS WITH VIRAL REFERENCE GENOMES 42
4.7. PRESENCE OF MULTIPLE GENOTYPES OF A VIRUS IN IBD AND PSC SAMPLES 44
4.7.1. MULTIPLE STRAINS OF PICOBIRNAVIRUS 44
4.7.2. MULTIPLE STRAINS OF HUMAN PAPILLOMA VIRUS 47
4.7.3. PHYLOGENETIC ANALYSIS OF ASFLV SEQUENCE SUGGESTS DISTANT RELATION FROM ASFV GENES AND THE PRESENCE OF MULTIPLE STRAINS IN THE METAGENOMIC LIBRARIES 48
4.8. SIMILARITY BETWEEN THE ASSEMBLED ASFLV CONTIGS/SCAFFOLDS AND THE SEQUENCES OF ASFLV FROM HUMAN AND SEWAGE SAMPLES AT THE NCBI DATABASE 51
4.9. ASFLV SEQUENCE VARIANCE AMONG IBD AND PSC PATIENTS 53
4.10. VALIDATION OF COMBINED ASSEMBLY RESULTS USING NESTED PCR 53
4.11. KEY FINDINGS 56

CHAPTER 5: COMPARISON OF DIFFERENT VIRAL ENRICHMENT METHODS 58

5.1. INTRODUCTION 58
5.1.1. ADVANTAGES AND DISADVANTAGES OF EACH APPROACH FOR VIRAL PREPARATION OF STOOL SAMPLES FOR NEXT GENERATION SEQUENCING 60
5.2. COMPARISON OF DIFFERENT METHODS FOR VIRUS PARTICLE PURIFICATION FOR THE CONSTRUCTION OF METAGENOMIC LIBRARY 62
5.2.1. BASED ON THE TAXONOMIC CLASSIFICATION 62
5.2.2. BASED ON THE DISTRIBUTION OF VIRUSES FOUND AMONG THE METAGENOMES 64
5.2.3. BASED ON NUMBER OF VIRAL SEQUENCES OBTAINED FROM DIFFERENT VIRAL PREPARATIONS METHODS IN VARIOUS PATIENTS 66
5.3. KEY FINDINGS 67

CHAPTER 6: DISCUSSION 68

6.1. OVERVIEW 68
6.2. COMBINED ASSEMBLY IMPROVES THE IDENTIFICATION OF VIRUS IN NGS DATA THAN THE VIRUS IDENTIFIED BEFOREHAND USING ASSEMBLY OF INDIVIDUAL METAGENOMIC LIBRARY 69
6.3. DETECTION OF VIRAL SEQUENCES IN THE CLINICAL SAMPLES 70
6.4. COMPARISON OF VIRAL ENRICHMENT METHODS 73
6.5. FUTURE DIRECTION 74

LIMITATIONS OF THE STUDY 75

BIBLIOGRAPHY 77

APPENDICES 85

1.1. THE PATIENTS' INFORMATION AND CODE 85
1.2. COLON BIOPSIES OF UC AND CD PATIENTS FOR PREVALENCE OF ASFLV STUDY 87
2. DISTRIBUTION OF TOTAL VIRAL READ COUNTS SIMILAR TO THE SEQUENCES OF NCBI VIRUS DATABASE AND LIST OF CORRESPONDING VIRUS DETECTED ACROSS THE METAGENOMICS LIBRARIES 88
3.1. ALIGNMENT OF CONTIGS/SCAFFOLDS SIMILAR TO HERPESVIRUS SEQUENCES WITH SEQUENCES OF NCBI DATABASE 90
3.2. ALIGNMENT OF CONTIGS/SCAFFOLDS SIMILAR TO POXVIRUS SEQUENCES WITH SEQUENCES OF NCBI DATABASE 133
3.3. ALIGNMENT OF CONTIGS/SCAFFOLDS SIMILAR TO RETROVIRUS SEQUENCES WITH SEQUENCES OF NCBI DATABASE 158
4.1. HIERARCHICAL CLUSTERING FOR VIRUSES DETECTED IN THE RNA LIBRARIES OF INDIVIDUAL PATIENTS. 170
4.2. HIERARCHICAL CLUSTERING FOR VIRUSES DETECTED IN THE DNA LIBRARIES OF INDIVIDUAL PATIENTS 172
5. HUMAN PICOBIRNAVIRUS RNA DEPENDENT RNA POLYMERASE USED IN THE PHYLOGENETIC ANALYSIS 173
6.1 HUMAN PAPILLOMAVIRUS CAPSID AND TRANSCRIPTION REGULATORY PROTEIN USED IN THE PHYLOGENETIC ANALYSIS 174
6.2. HUMAN PAPILLOMAVIRUS TRANSCRIPTION REGULATORY PROTEIN USED IN THE PHYLOGENETIC ANALYSIS 176
7. AFRICAN SWINE FEVER VIRUS USED IN THE PHYLOGENETIC ANALYSIS 177
8. DIFFERENT PROTEIN CODING CONTIGS/SCAFFOLDS OF ASFLV ARE QUITE DIVERGENT FROM THE PROTEINS OF ASFV AT THE NCBI DATABASE (FROM PIG). SCAFFOLDS WERE SEARCHED AGAINST PROTEIN DATABASE USING BLASTX, AND THE LEVEL OF IDENTITY OF SEQUENCES WITH MATCHED DATABASE PROTEINS WAS DETERMINED. 178
9. CONFIRMATION OF PCR PRODUCT (ASFLV SEQUENCES) AMPLIFICATION BY BIOANLYZER 179
10. SUMMARY OF THE PCR EXPERIMENT SEEKING TO DETECT ASFV-LIKE VIRUS IN SAMPLES OF OTHER BODY PARTS OF THE TWO POSITIVE PATIENTS. 180
11. TEST FOR ASFLV-LIKE SEQUENCES IN OTHER PATIENTS 181
12. PRIMER SETS USED FOR PCR EXPERIMENTS 182
13. DISTRIBUTION OF THE MAJOR TAXONOMIC GROUPS OF SEQUENCES USING ENTIRE DATASETS 183

LIST OF TABLES

Table 3.1: The distribution of the total viral read counts based on the sequence similarity search with NCBI virus database using BLASTn (nt/nr) in metagenomics libraries of faecal samples from different patients. 27

Table 3.2: List of the name of the detected viruses, the number of the virus sequences (total read counts) similar to the virus sequences in NCBI database, the genome coverage of the corresponding virus by the reads from patient CD1 RNA metagenomics library 28

Table 3.3: List of sequences from patient CD1 that had similarity with the protein sequences of NCBI virus database after BLASTx search 29

Table 4.1: List of sequences similar to the sequence of mammalian viruses identified across the metagenomics libraries after BLASTn and BLASTx search on the assembled sequences from combined assembly approach 36

Table 4.2: List of sequences similar to the sequence of plant and other (insect, bird and fish) viruses identified across the metagenomics libraries after BLASTn and BLASTx search on the assembled sequences from combined assembly approach 37

Table 4.3: The distribution of the total viral read counts based on the sequence similarity search with NCBI virus database using BLASTn and BLASTx on the combined assembled sequences in metagenomics libraries of faecal samples from different patients 41

Table 4.4: Assembled sequences (contigs/scaffolds) of ASFLV protein coding genes are distantly related with previously reported ASFV-like sequences found in human and sewage samples 52

Table 4.5: Alignment of African swine fever virus like capsid protein with capsid protein of ASFV from the reference genome 53

Table 4.6: A summary of the PCR experiment for the confirmation of the presence of ASFV sequences in patients' stools. (NC= negative control) 54

Table 5.1: The distribution of sequences (total reads) that showed similarity to sequences at the NCBI virus database and distribution of virus species identified based on the sequence similarity search using combined assembly datasets 66

LIST OF FIGURES

Figure 1: Possible etiology of inflammatory bowel disease (IBD) 6

Figure 2.1: Schematics of the metagenomic library construction process and data analysis 15

Figure 2.2: Schematics of the bioinformatics pipeline for analysis of individual metagenomic libraries 19

Figure 2.3: Schematics of the bioinformatics pipeline for the combined assembly of different metagenomics libraries 21

Figure 4.1: Combined assembly generates longer sequences when compared to the assembly of reads from individual libraries 32

Figure 4.2: Bioinformatics approach to perform the combined assembly of metagenomics datasets and sequence annotation of the assembled contigs/scaffolds. 34

Figure 4.3: Classification of contigs/scaffolds based on the taxonomy of the assembled sequences (contigs/scaffolds) after running all the reads from metagenomics libraries through the SOAPdenovo-trans package 35

Figure 4.4: The alignment of sequences similar to the sequences of pepper mild mottle virus (PMMV) with PMMV reference genome 42

Figure 4.5: The alignment of assembled sequences (contigs/scaffolds) similar to the sequences of human picobirnavirus against the reference genome of human picobirnavirus from NCBI database 43

Figure 4.6: Alignment of ASFLV assembled sequences (contigs/scaffolds) against the reference genome of ASFV, BA71V 44

Figure 4.7: Phylogenetic analysis of assembled sequences of picobirnavirus RNA dependent RNA polymerase found in CD patients 46

Figure 4.8: Phylogenetic analysis of assembled sequences of HPV capsid protein and transcription regulatory protein (E1) found in CD and PSC patients 47

Figure 4.9: Phylogenetic analysis of assembled sequences of ASFLV capsid, topoisomerase and helicase protein found in IBD and PSC patients 50

Figure 4.10: Alignment of PCR amplified sequences with assembled sequences from the combined assembly of metagenomics datasets. 55

Figure 5.1: Comparison of assembled sequences matched as bacteria, human, phage and virus from faecal samples of three CD, one UC and one PSC patients using the three viral enrichment methods 63

Figure 5.2: Comparison of three viral preparation methods on the basis of the viral sequences detected among the metagenomic libraries of the patients. 65

LIST OF ABBREVIATIONS

ASFLV: African swine fever like virus

ASFV: African swine fever virus

BLAST: Basic Local Alignment Search Tool

BP: Base pair

CD: Crohn’s Disease

CDD: Conserved Domains Database

cDNA: complementary Deoxyribonucleic acid

CMV: Cytomegalovirus

DBG: De Bruijn Graph

DNA: Deoxyribonucleic acid

dsDNA: Double stranded deoxynucleic acid

GBP: Giga Base Pair

HMM: Hidden Markov model

HPV: Human papillomavirus

HSV: Herpesvirus

IBD: Inflammatory Bowel disease

MEGA6: Molecular Evolutionary Genetics Analysis

NCBI: National Center for Biotechnology Information

NGS: Next generation sequencing

NR: Non redundant

NT: Nucleotides

ORF: Open reading frame

PMMV: Pepper mild mottle virus

PSC: Primary Sclerosing Cholangitis

PTFE: Polytetrafluoroethylene

RdRp: RNA dependent RNA polymerase

RNA: Ribonucleic acid

RPM: Rotation per minutes

rRNA: Ribosomal ribonucleic acid

SOAP: Short Oligonucleotide Analysis Package

ssDNA: Single stranded deoxynucleic acid

TFF: Tangential Flow Filtration

TPM: Total reads per million

UC: Ulcerative Colitis

CHAPTER 1: INTRODUCTION

1.1. VIRAL METAGENOMICS TO DISCOVER NOVEL VIRUS

The method incorporating virus culture *in vitro* using a cell culture system and virus neutralization assay were traditionally used for virus discovery [1]. This option has the limitations of lacking the appropriate cell culture system for some viruses and unavailability of antibodies for the neutralization assay that present a great challenge for viral discovery and further viral research [2]. New molecular identification methods such as PCR have been applied to overcome these obstacles to the study of unculturable and nonisolated viruses. However, the application of PCR is not without potential flaws; the sequence information of a targeted new virus genome has to be known or inferred in order for viral sequences to be amplified. Moreover, to identify a virus as a pathogen, an individual analytical test for each pathogen must be performed [3]. These limitations associated with such techniques make it challenging to identify unknown viruses. In addition, sequencing of the 16S fragment of the small subunit of the ribosomal RNA (rRNA) is a proven technique for the detection of novel or known microbes [4-10]. However, this technique cannot be applied to viral discovery as viruses do not possess this gene and do not share any common genes that is so highly conserved [10].

To overcome all hindrances involved in virus discovery, metagenomics offers an alternative culture-independent and sequence-independent technique. Unlike sequencing of 16S fragments for bacterial discovery, in viral metagenomics, researchers need not to have prior knowledge of a common particular gene of the target virus. This technique was first used to analyze the genetic content of environmental samples, to determine its metabolic traits, to characterize it, and to identify new enzymes and antibodies [11-15]. Later the scope of metagenomics was expanded to the field related to marine environmental research, plant and agricultural biotechnology, human genetics and the diagnosis of human diseases [16].

One of the promising areas of metagenomics is its use in detecting viral pathogens in clinical samples. The viral metagenomics technique allows the identifications of co-infections, which some diagnostic tests may overlook after identifying the initial infectious agent [17]. The application of viral metagenomics to virus discovery has been explored by many researchers [18, 19]. Viral pathogen outbreaks can also be detected using viral metagenomics [20-23]. One of a successful use of viral metagenomics was during influenza outbreaks in which a rapid determination of the viral subtype was achieved [24]. On a different note, a gene can be inserted into a virus (transposons), and this could lead to the virus becoming a more infectious agent; it could also help the virus to escape detection by traditional diagnostic tests [25]. The novel wild type ebola virus, for example, has not been detected in any traditional tests, but the metagenomic approach identified it in all clinical cases [26]. Metagenomic studies of diarrhoea of unknown etiology helped researchers to find novel and potentially pathogenic human-associated viruses like cosavirus [27, 28] and klassevirus [29]. The list of new types of human viruses being discovered through the use of viral metagenomics is still expanding; it includes rhinovirus [30], a novel bocavirus [31], a novel arenavirus [32] and a novel parechovirus [33] etc.

Human pathogens are very diverse and approximately two hundreds viral pathogens have been identified to date, with a rate of two being discovered each year [34]. Most of these viruses discovered are found in acute diseases. For instance, parvovirus and coronavirus in lower respiratory tract infection [35], polyomavirus in human merkel cell carcinoma [36], astrovirus, torque teno virus, norovirus, picobirnavirus, enterovirus and nodavirus in diarrhea patients [37], arenavirus in hemorrhagic fever [38], bocavirus, picornaviruses, circovirus, nodavirus and dicistroviruses in acute flaccid paralysis [39], astrovirus in encephalitis [40] and circovirus in tropical febrile illness[41] were discovered by using metagenomics approach. Mostly, if not all, these new viruses were discovered in acute diseases. Chronic diseases, on the other hand, were not studied much using viral metagenomics for viral screening often which leads low viral load except in immune suppressed condition due to the induction of drugs.

For viruses, the mutation rate is higher in comparison to other microbes, and this dynamic ability renders them able to jump between species, as is the case in avian [42] and swine [43] influenza epidemics, and to overcome the effectiveness of drugs. The viral species are unique in their infection type, mode of transport and characteristics. It is therefore important to identify them rapidly as pathogens and to know and identify the novel mutant subtypes for the purpose of treatment to control viral diseases. Given all the advantages of metagenomics, this technique have been used in this study to detect the viruses present in inflammatory bowel disease (IBD) and primary sclerosing cholangitis (PSC) faecal samples with the aim to investigate the presence of novel virus in human clinical samples of patients with chronic diseases.

1.2. ASSEMBLY APPROACH FOR VIRAL METAGENOMICS

Assembly is a process by which overlapping short sequences are joined together to form longer sequences. Since most high throughput sequencing platforms generate short sequences, alignment of these latter to reference databases frequently leads to ambiguous results; therefore, assembly of read (short sequences of ~150 bp length) into contigs/scaffolds (longer than reads) is often implemented prior to alignment. There are two strategies employed for assembling the metagenomic data; namely, reference based assembly and *de novo* assembly. To date, the *de novo* assembly approach has been used in most studies. The advantage of *de novo* is that it is suitable for the analysis of next generation sequencing data where the reference genome of targeted species could be unavailable and still using *de novo* approach in NGS data can recover the whole genome or lager sequence fragments to recognise the origin of these sequences.

The SOAPdenovo-trans package [44] was used in this study to assemble short reads into longer sequences in order to recover the genome of uncultured virus or at least some of its coding sequences for subsequent characterization. SOAPdenovo-trans utilizes *de novo* assembly based on the De Bruijn graph (DBG) approach, which was specifically created to handle very large amounts of data [45, 46]. During the assembly process, the similar sequences are joined using the DBG approach, and longer sequences, called contigs. Reads are decomposed into *k*-mers (fragments of *k* nucleotides); these fragments become nodes of the De Bruijn graph. Overlapping *k* - 1 bases are then used to elongate the consecutive sequences to find contigs. The end pairs of two related contigs are used to link to a longer non-contiguous DNA sequence to form a scaffold [47].

Sequence assembly is an important computational aspect of the metagenomics process because with the assembly approach more data are used for the analysis whereas without the assembly many single-sequenced reads would remain unused and unidentified in the metagenomic data. However, the assembly process is still in its early stage, and typically there are no references against which we can compare results. Another advantageous aspect of assembly is the generation of sequences that are generally longer than the unassembled reads, which in turn leads to more accurate taxonomic classification. Taxonomic classification of unassembled sequences is often ambiguous. Ideally, assembly provides full-length coding sequences for subsequent analyses and produces less data than unassembled sequences, reducing the processing time as well. With all these advantages of assembly approach a virus with higher titre in a patient sample would be detected but the non-abundance viruses may remain undetected from NGS data.

For this study we also used combined assembly, a process whereby reads from different samples were run together in a single assembly to generate longer sequences through contigs and scaffold formation. During the process, sequences were tagged with a short unique sequence for each metagenomics library so that after sequencing the origin of the fragment could be traced back after assembly had completed and the relative contribution of each sample to the resulting scaffolds could be determined. Combined assembly was performed to increase the use of additional data that were found to be unknown in similarity searches of the metagenomes of individual patients.

1.3. INFLAMMATORY BOWEL DISEASE (IBD)

Idiopathic inflammatory bowel disease (IBD) is an illness including a range of chronic, non-communicable and diarrheal diseases such as ulcerative colitis (UC) and crohn's disease (CD) [48]. IBD is characterized mainly by ulcerations in the gut wall which can result in excessive bleeding, anaemia, perforation with abscesses and consequent fibrosis in the intestine. Although IBD affects the intestines, it sometimes leads to extraintestinal manifestations such as disorders of joint, biliary tract, lungs, skin and eye [49, 50]. UC and CD are distinguishable based on clinical and pathological outcome; however, because many similarities are shared, diagnosis is difficult in 10% of cases. Due to this difficulty, it remains elusive whether UC and CD share a common or different etiology [51].

The etiology of IBD is not yet clear; however, it is thought to be caused by a combination of genetic composition of individual, different immunological factors, environmental factors and gut microflora [52-55]. The most widely accepted account of the pathogenesis of IBD considers it a condition in which T-cell responses occur against gut microflora in genetically and environmentally susceptible hosts. The increasing prevalence and the complexity in the etiology have made IBD a disease of concern. Figure 1 shows the complexity of the pathogenesis of IBD and the factors involved.

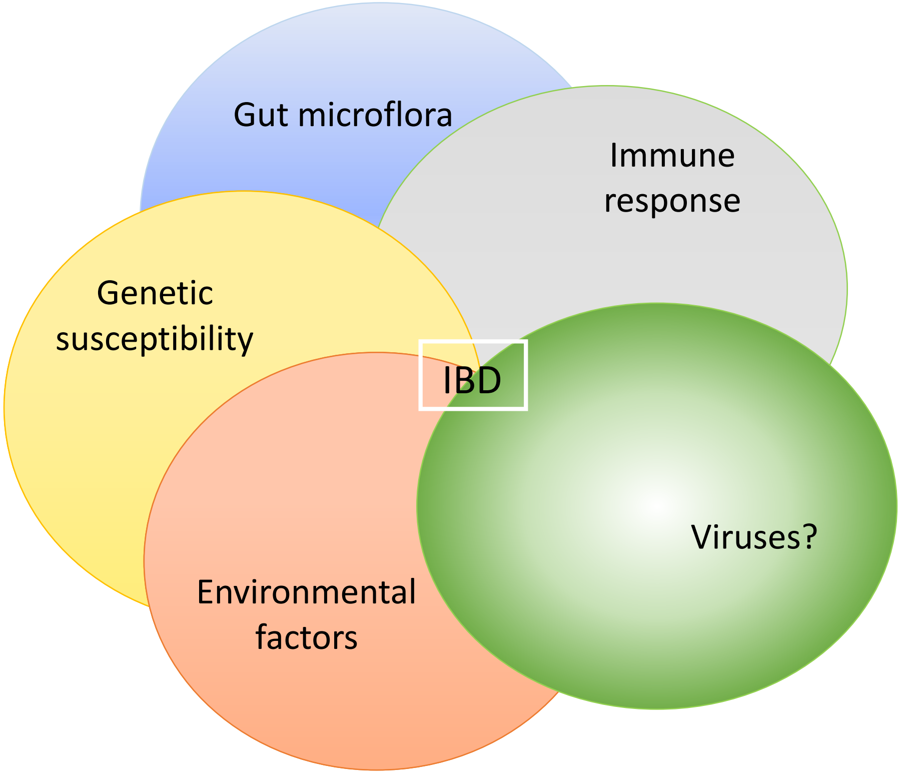


Figure 1: Possible etiology of inflammatory bowel disease (IBD). IBD is thought to be caused by a complex interaction between the genetics and immunity of the host and environmental factors

1.4. PRIMARY SCLEROSING CHOLANGITIS (PSC)

PSC is a rare liver disorder which is characterized by inflammation and fibrosis of the bile ducts present in liver; both the intrahepatic and extrahepatic bile ducts are affected [56, 57]. As in the case of IBD, the pathogenesis of PSC is not yet known, but is thought to be associated with complex interactions between genetic composition, environmental factors, and an imbalance in the gut microbiome, resulting in auto immunological responses in the host [58]. The susceptible genes thought to be involved in PSC pathogenesis are HLA B8 and HLA DR3 [59, 60]. PSC is a progressive disorder which leads to other diseases such as liver cirrhosis and portal hypertension [61]. PSC is considered to be one of the extraintestinal manifestations of IBD; approximately 70% to 80% of PSC patients are found to have IBD [62-64]. In addition, studies have found that PSC develops in 2% to 7.5 % of UC patients [65, 66].

1.5. VIRUSES IN AUTOIMMUNE DISEASES

Role of viruses in autoimmune diseases are not known. Attempts have been made to detect a particular virus contributing to IBD using the traditional technique of tissue culture based detection [67-69] and more recent molecular-based methods of virus screening [70, 71], but none have so far been successful or conclusive. Evidence of the presence of bacteriophages in the mucosa of CD patients has been reported by microscopy [72]. In addition, the potential role of viruses like herpes, measles and rubella as causative agents of IBD has interested some researchers; they found that the viruses were present in many occasions in the serum and tissues of a host, but did not find clear evidence of their relationship with IBD [73, 74]. IBD has also been associated with the presence of cytomegalovirus, which was found to play a role in causing colitis and superimposed disease in the context of immunosuppressive therapy [75, 76].

Further strong evidence of a relationship between viruses and IBD has been reported in studies using a mouse model [77, 78]. Cadwell et al. showed that CD-like pathologies were observed in norovirus-infected mice in which a CD-susceptible gene, Atg16L1, had been previously mutated [77]. This study demonstrated that a viral infection within mice carrying a mutated CD-susceptible gene, in combination with environmental factors and commensal bacterial, can cause an IBD-like phenotype. Hubbard and Cadwell later reported on the host-pathogen interaction that can develop

IBD by studying the three-way relationship between viruses, autophagy genes and IBD [78]. Since the presence of viruses was reported in IBD patients from many researchers, but the viral metagenomics have not been yet performed on IBD or PSC samples. We chose faecal samples from IBD and PSC patients for performing viral metagenomics to know the viruses harboured in the gut of the patients with these diseases.

1.6. HYPOTHESIS AND SPECIFIC AIMS

We assumed that samples from patients affected by similar diseases contain similar types of viral species. Thus, if all libraries from patients with related pathologies are considered as one for the purpose of sequence assembly, sequences from similar viruses from different patients could be joined together to form longer sequences and consequently the chances of achieving viral sequence recognition would be much higher. This strategy was utilized in this study for virus discovery in clinical samples from chronic disease patients. Since the patients with chronic diseases have an active immune system which may diminish viral burden, the combined assembly would be the ideal technique to detect those viruses residing in the gut by enabling getting longer sequences through joining the short reads together. This study aims at discovering novel viruses in stool samples from patients suffering from several similar types of autoimmune diseases.

1.6.1. Hypothesis

Combined assembly will increase the sensitivity of virus detection in stool metagenomic samples derived from patients affected by IBD or PSC

To address this hypothesis, this thesis has three specific aims.

1.6.1.1. To profile virus populations in fecal samples from IBD or PSC patients using viral metagenomics

1.6.1.2. To assess the efficiency of several virus enrichment protocols in clinical fecal samples

1.6.1.3. To implement the combined assembly approach for virus identification in metagenomics libraries

CHAPTER 2: MATERIALS AND METHODS

2.1. ETHICS APPROVAL

The Human Research Ethics Board at the University of Alberta have reviewed and approved the sample collection and handling procedures. All fecal samples were collected after obtaining a patient's consent form that had been read and signed by the patients.

2.2. DESCRIPTION OF PATIENTS

Faecal samples were obtained from a total of nine IBD (five UC and four CD) patients, one with PSC, two controls (who suffered from diarrhoea; designated as CON1 and CON2), and one with an unknown diagnosis (designated as UNK). All patients were coded (see the Appendix 1.1). For the comparisons of viral enrichment methods, three CD patients (CD1, CD2 and CD3), one UC patient (UC1) and the sole PSC patient's faecal samples were processed by Tangential flow filtration (TFF), Glass milk and the Ultrafree MC microfiltration method. DNA and RNA were extracted from all purified viral preparations, and metagenomic libraries were constructed. Libraries were also constructed for the rest of the samples either viruses enriched using TFF, glass milk or ultrafree mc microfiltration method (CD4, CD5, UC2, UC3, UC4, CON1, CON2 and UNK). Data analysis for individual metagenomes was conducted using the bioinformatics pipeline described in Section 2.5 and Section 2.6. All the reads from individual libraries were used for the combined assembly analysis.

2.3. VIRAL PURIFICATION FROM FAECAL SAMPLES

2.3.1. Tangential flow filtration (TFF) Method

First, the amount of starting material from either patients or control subjects was determined. Liquid samples of about 5 to 10 ml were mixed with five volumes of TN Buffer (0.1 M Tris, pH 7.6; 0.1 M NaCl). The solution was transferred into the appropriate MidJet containers that had been washed and UV treated before use. For solid stools, 10 g of stool were mixed in 100 ml of TN. The solution was homogenized by being stirred briefly in TN buffer until the faeces fully dissolved under the fume hood. The solution was transferred to 50 ml tubes and was centrifuged at 300 x g for 5 min at room temperature (RT). Then the mixed solution was transferred into the appropriate MidJet container that had been washed and UV treated before use.

Next, the solution was filtered through a 0.45 μm MidJet column (GE Healthcare) at 4^0C, and then concentrated through a 300 KDa MidJet column at 4^0C. This filtration step is intended to remove bacterial and human cells. The final volume of the filtrated sample was about 4-6 ml. Half of the concentrated sample was treated with 50 U/ml or 100 U/ml of Benzonase (Sigma) in 2 mM $MgCl_2$ at 21^0C for 1.5 hours. The reminder of the final concentrated sample was kept at -80^0C.

The DNA was extracted from 2 ml of nuclease-treated filtrate using multiple columns from the DNeasy Blood & Tissue Kit (Qiagen), as described in the Purification from Animal and Blood or Cells Protocol. RNAse A (100 mg/ml) was used to get rid of the RNA present in the solution. DNA was eluted by sequentially passing 70μl of AE buffer through all columns used for each preparation (the final elution volume was ~ 40-50 μl). RNA was extracted using 2 ml of nuclease-treated filtrate with the QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer's protocol. RNA was eluted in AVE buffers described for DNA. Subsequently DNAse I was used in the Qiagen RNeasy Mini Kit as recommended in the RNA clean up protocol to eliminate the DNA.

2.3.2. Ultrafree MC (microcentrifugation) method

Faecal samples were suspended in TN Buffer as described above. A microcentrifugal tube (Ultrafree-MC 0.45μm non-sterile filter, Millipore, UFC3 0HV 25) was first sterilized with ethanol by adding 500 μl ethanol into the tube and spinning it at 12,000 x g for 5 min. Then 500 μl of suspended solution was loaded into the microcentrifugal tube and centrifuged at 12,000 x g for five min. Once samples were passed through these filters, they were concentrated using 30 KDa Centriprep Filters. For 1 ml of suspended solution, 20 μl Turbo DNAse, 40 μl 10X Turbo buffer, 20 μl Baseline Zero DNAse, 20 μl Benzonase, and 20 μl RNAse A were added and incubated at 37^0C for 1.5 hours. The DNA and RNA were extracted following the same procedure employed in the TFF method of virus purification (see above).

2.3.3. Glass milk/ Silicon dioxide method

For the preparation of the glass milk solution, 5 g of Silicon dioxide (Sigma-Aldrich S-5631) were washed in 45 ml of PBS overnight. The solution was left overnight to settle and the wash repeated twice. The mixture was transferred to a glass bottle and was sterilized by autoclaving. The glass milk was cleaned up by adding 525 μl of 95 mM NaAc (pH 4.5) onto 725 μl of glass milk in a microcentrifuge tube. After mixing by inversion, the tubes were incubated at RT for 10 min (the tubes were inverted to mix properly every few minutes). The tubes were centrifuged at 5000 x g for 5 min. Supernatant was removed and 500 μl of water was added, and the solution was heated at 55˚C for 5 min. This process of adding water and centrifugation step was repeated for two or three more times. Finally the glass milk pellet was resuspended with water to obtain 1 ml of final "clean" glass milk. To purify the virus particles from the sample, each ml of sample was supplemented with 50 µl 2M NaAc (pH 4.5) and mixed properly. 1.45 ml of “clean” glass milk solution was added and mixed. The solution was incubated at RT for 10 min (the solution tubes were inverted to mix properly every few minutes). The sample tube was centrifuged at 5000 x g for 5 min. The supernatant from the centrifugation steps were transferred to a clean sterile tube without disturbing the glass milk pellet at the bottom. The DNA and RNA were extracted following the procedure described above for DNA and RNA extraction in the TFF method of virus purification.

2.4. CONSTRUCTION OF LIBRARY USING TRUSEQ KIT

The extracted RNA and DNA from the virus enrichment procedures needed to be preprocessed prior to construction of metagenomic libraries. RNA was reverse transcribed into cDNA. For DNA the genomes of viruses were too long to start the library construction. Hence, these DNAs were fragmented by shearing. After those steps, all prepared double stranded cDNA and fragmented DNA were used and a common procedure was followed to construct metagenomics library.

2.4.1. Pre-processing of RNA: steps prior to the construction of metagenomic libraries from RNA

The sequencer cannot sequence single stranded nucleic acids like RNA. Therefore, it is needed to make cDNA from RNA prior to the construction of a metagenomic library. cDNA synthesis was performed using a two-step procedure. Six samples were processed in each batch. In 0.5 µl PCR tubes, 1µl random primers (Invitrogen) were added to 11.2 µl RNA for first-strand synthesis. Then, the tubes were incubated at 65˚C for 5 min in a thermal cycler, and then quickly chilled on ice. Next, 4 µl of 5X first strand buffer, 2 µl 100 mM DDT, 0.4 µl 25 mM dNTPs mix and 0.5µl RNAseOUT (Invitrogen) were added to each tube, and the tubes were incubated in a thermal cycler at 25˚C for 2 min. 1 µl of SuperScript-II (Invitrogen) was added in each tube and incubated under the following conditions: 25˚C for 10 min, 42˚C for 50 min, 70˚C for 2 min, and 4˚C on hold; the tubes were put on ice afterwards. For the second strand cDNA synthesis, the following components were added to the first strand cDNA synthesis mix: 75 µl second strand buffer (56.8 µl H_2O, 10 µl 500 mM Tris, pH 7.8, 5 µl 100 mM $MgCl_2$, 1.2 µl 25 mM dNTPs, 1 µl 100 mM DTT, and 1 µl RNaseH [2 µ/µl]) (Invitrogen) and 5 µl DNA Pol I (Invitrogen). After mixing well those tubes were incubated at 16°C for 2.5 h. The synthesized cDNA was purified using the QIAquick PCR purification kit (Qiagen) and eluted in 45 µl elution buffer (Qiagen).

2.4.2. Preprocessing of DNA: steps prior to the construction of metagenomic libraries from DNA

The viral genomic DNA extracted after the virus enrichment procedures needed to be fragmented into small pieces of DNA. DNA was sheared in a Covaris S2 instrument in 6x16 mm AFA fiber snap-cap microtubes (Covaris) under recommended conditions to generate the target pick of 300 bp (duty cycle: 10%, intensity: 4; cycles per bust: 200; treatment time: 80 sec; water bath temperature: 7°C).

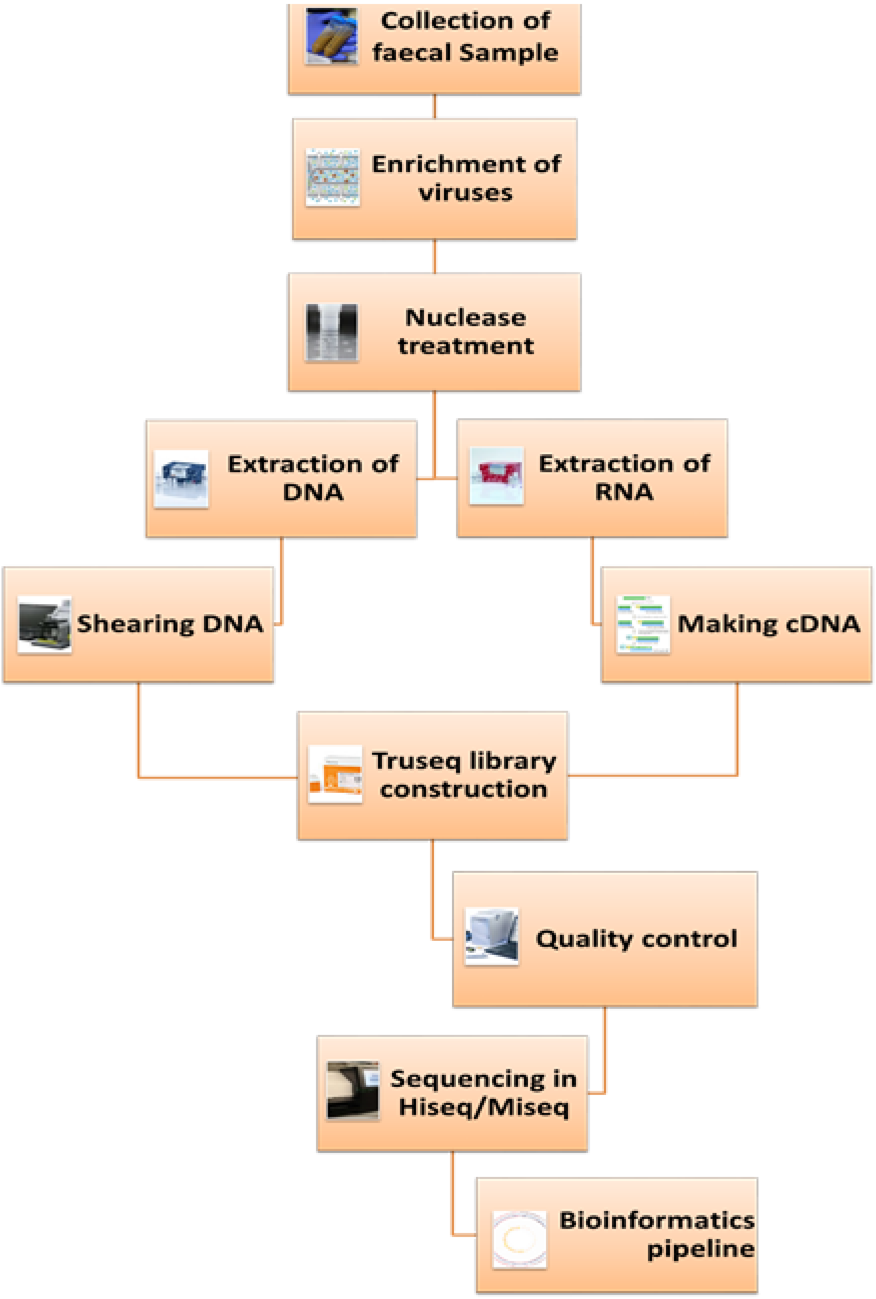


Figure 2.1: Schematics of the metagenomic library construction process and data analysis

2.4.3. Common steps for metagenomic libraries construction

Twelve samples per batch were processed for library construction (six cDNA samples and six sheared DNA samples). Samples were end repaired to make both ends of DNA molecules blunt. 42.5 µl of cDNA or fragmented genomic DNA were supplemented with 2.5 µl of End Repair Enzyme Mix (NEB) and 5 µl of 10x NEBNext End Repair Reaction Buffer (NEB), mixed and incubated in a thermal cycler at 20˚C for 30 min. The reaction was purified using the QIAquick PCR purification kit (Qiagen) and eluted in 45 µl elution buffer. Next, a dA-tailing step that creates a 3' dA overhang at the ends of the DNA molecules was performed. To perform this step 42 µl of end-repaired DNA, 5µl of 10X NEBNext dA-Tailing Reaction Buffer (NEB) and 3 µl of Klenow Fragment (NEB) (3' → 5' exo) were mixed and incubated in a thermal cycler at 37˚C for 30 min. The reactions were purified using the QIAquick PCR purification kit (Qiagen) and eluted in 27 µl elution buffer. The next step in library construction is the ligation of pair-end (PE) adapters (which have a 3' dT overhang). To conduct this step, 25µl of dA-tailed DNA, 10µl of NEBNext Quick Ligation Reaction Buffer (NEB) (5X), 1µl of PE adapters (Illumina) and 5 µl of Quick T4 DNA ligase (NEB) were mixed in the tubes and were incubated in a thermal cycler at 20˚C for 15 min. After the incubation period, 1.5 µl of USER enzyme (NEB) was added to each of the reaction tubes, and again all the tubes were incubated in a thermal cycler at 37˚C for 30 min. The reactions were purified using the QIAquick PCR purification kit (Qiagen) and eluted in 27 µl elution buffer. In the next step, the size selection of ligated product was performed by running ligated DNA molecules on an E-Gel® EX Agarose Gel (2% agarose gel). The samples were loaded and were run according to the manufacturer's protocols. The gel area between 250 and 400 bp was cut and the DNA was extracted using a MinElute Gel Extraction Kit (Qiagen) and eluted in 40 µl of elution buffer. Libraries were then indexed and amplified by PCR using *Pfx* DNA polymerase. The PCR reaction was mixed under the hood: 10.4 µl Ultra-pure water, 5 µl HF buffer [10X], 2 µl Mg^{2+} (50 mM), 1 µl InPE 1.0 [25 µM], 1 µl InPE 2.0 [0.5 µM], 0.8 µl dNTP [25 mM], 0.8 µl Plat *Pfx* DNA polymerase [50 U/µl], 1 µl Primer Index [25 µM] and 28 µl of DNA from the previous step were added on the bench, and the PCR was run for 19 cycles using the *Pfx* recommended reaction conditions.

1 µl of resuspended PCR product from each library was run on the BioAnalyzer (Agilent) for titration of the amplification using a DNA-specific chip (Agilent DNA-1000) following the manufacturer's instructions. Additional PCR cycles were added as deemed necessary. The formation of amplified adaptor dimers were inevitable for some of the samples, these were removed using Agencourt AMPure magnetic beads (Beckmann Coulter). The libraries were size selected using 2% agarose gel on an E-Gel® EX Agarose Gel. Except for one sample that was sequenced at BGI-Shenzhen (China) on a Genome Analyzer II instrument, the rest of the libraries were sequenced at The Applied Genomic Centre at the University of Alberta on an Illumina Miseq instrument. The schematics of the metagenomic library construction process outlined in the Figure 2.1.

2.5. DATA ANALYSIS OF INDIVIDUALLY SEQUENCED METAGENOMIC LIBRARIES

Individual metagenomic data were analyzed according to the procedure described elsewhere [79], and summarized in Figure 2.2 (figure modified from Law et al. [79]). In short, the metagenomes were analyzed using the developed pipeline, which consisted of a sequential and multitier algorithm performed using a combination of the Basic Local Alignment Search Tool (BLAST) [80] and Short Oligonucleotides Analysis Program (SOAP) [81]. In the first step, low quality sequences, primer and adapter sequences were eliminated. The simple repeat sequences, ribosomal, mitochondrial and sequences were removed using the SOAPaligner [81] and RepeatMasker [82] programs, using the metagenome data to align against the NCBI nt/nr and RepBase (version 16.04) (http://www.girinst.org/ [date last accessed 26 Sep 2011]) databases respectively. The taxonomic classification or binning of sequences according to taxon were performed in a two-step process. Since the BLAST algorithm is a time-consuming step, it was important to feed this algorithm with as few sequences as possible. Therefore, SOAPaligner was used to align high-quality reads against the human, bacteria and virus database and to bin the matched sequences according to the taxon they aligned against. SOAPaligner is a stringent program which consumes less memory than the BLAST program. The unaligned remaining read ends were aligned with the NCBI nt/nr database using a BLASTn [80] search (using E value 1e-05 as the cut off). For both programs, the human database sequences were obtained from GRCh37/hg19 (Feb, 2009) and all human transcript were obtained from refseq at genbank database. The bacteria and virus database were constructed using taxon ID deposited in the NCBI nt/nr. A sequence was regarded as ambiguous when a sequence could not be defined and differentiated from human to bacteria or virus to with three base pairs variations existed within the matched region of reference sequence. The unknown sequences category contained those which did not match any of the human, bacteria or virus datasets.

De novo assembly was performed with SOAPdenovo-Trans [44] using default parameters. The SOAPdenovo-Trans assembler in contrast to the SOAPdenovo [47] considers transcript sequences and gene expression levels of allelic variants and gene fusion for sequence analysis on NGS data by reconstructing the full length transcript and alternative splicing forms of mRNA from very short reads. To assemble sequences: virus, phage, and HERV, a subset of ambiguous sequences which had a viral end at one end and an unknown at the other end and unknown sequences (read ends) from previous steps were submitted to the program and subsequently all reads were broken into *k*-mers to construct De Bruijn *k*-mer graphs. The software uses this De Bruijn graph (DBG) to build reads into contigs. Then the program discarded the DBG graph and used the presence of read ends on different contigs to join them together to form scaffolds. It uses overlapping reads to reduce or close the gaps between two sequence regions of a scaffold. The sequences were then aligned against the NCBI nr database, which includes archaea, bacteria, HERV, fungi, plant, human, invertebrate, mammal, phage, vertebrate, protista, and virus sequences, using the BLASTx program (the E value cut off was 1e-05 and the top hit was deemed as the similar taxon of the assembled sequences) (Figure 2.2).

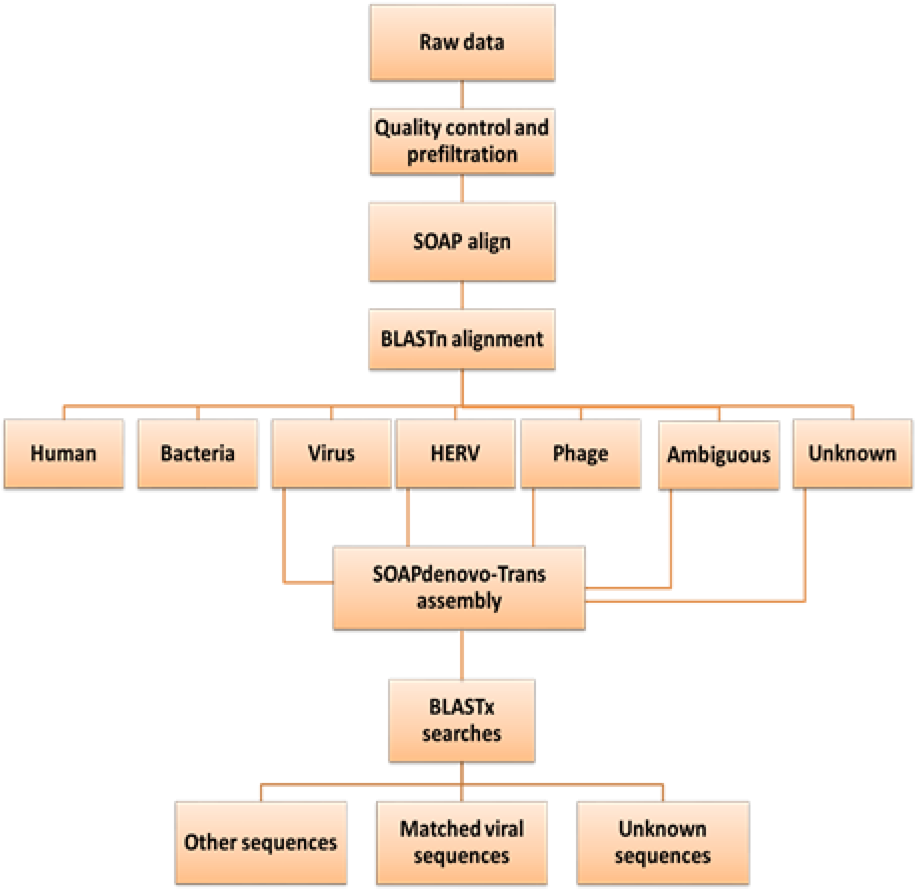


Figure 2.2: Schematics of the bioinformatics pipeline for analysis of individual metagenomic libraries

2.6. USING COMBINED ASSEMBLY FOR THE ANALYSIS OF ALL METAGENOMIC LIBRARIES FROM DIFFERENT SAMPLES

Reads from RNA and DNA metagenomic libraries from different samples were first trimmed using a quality score of 15 and a window size of 5, which will trim bases off an end if the average quality of the last 5 bases is below 15. Primers and adapters sequences were trimmed off in the same step. The high quality read ends were then run through the assembly software, SOAPdenovo-Trans [44]. This software generated contigs and scaffolds in a process described in Section 2.5. The reads that take part in the formation of a contig or scaffold were traceable as each reads had unique short-sequences tags for each of the metagenomics library. The assembled sequences were then searched against the NCBI nt/nr database using the BLASTn algorithm. The cut-off value of E was determined as 1e-05. The sequences were then binned into different taxa according to their E value. The remaining unmatched sequences from BLASTn were then searched against the NCBI nr database which includes archaea, bacteria, HERV, fungi, plant, human, invertebrate, mammal, phage, vertebrate, protista, and virus sequences using the BLASTx program. The E value of 1e-05 was taken as cut off and all hits below this threshold were binned as significant matches. The left over sequences which had no sequence similarity to the database sequences were binned as unknown sequences category. The bioinformatics pipeline summarizing the process is given in Figure 2.3.

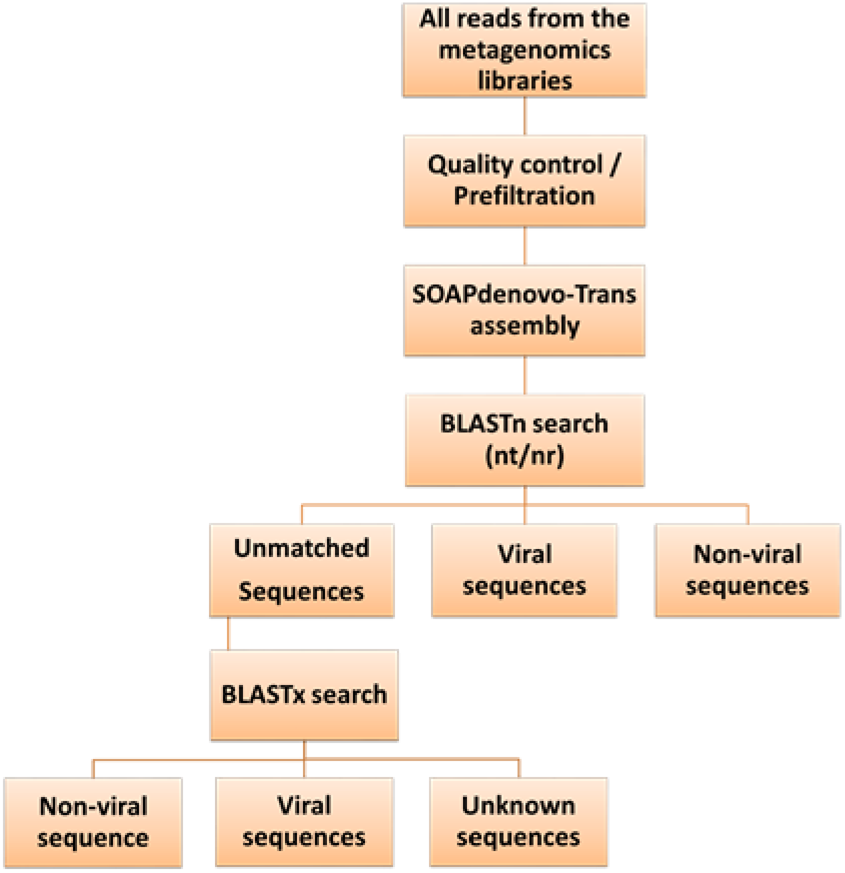


Figure 2.3: Schematics of the bioinformatics pipeline for the combined assembly of different metagenomics libraries

2.7. APPLICATION OF OTHER SOFTWARE

The phylogenetic analysis was performed with the Molecular Evolutionary Genetic Analysis software version 6.0. (MEGA6) [83]. MeV 4.8.1 was used to generate hierarchical clustering for the viruses detected in the RNA and DNA metagenomics libraries of individual patients.

2.8.VERIFICATION OF VIRUS PRESENCE: NESTED PCR PROCEDURE

DNA from virus enrichments (Section 2.3) was used as template for PCR. Platinum Taq High Fidelity polymerase (Invitrogen) and the buffer that came along with the box were used. The primers were added at a concentration of 10 µM each. The primer sets used for the amplification of capsid, helicase and origin binding protein of ASFLV sequences are given in Appendix 12. For each round of PCR, 35X cycles were applied in a thermal cycler and the PCR conditions were followed according to the polymerase manufacturer's instructions.

CHAPTER 3: ANALYSIS OF INDIVIDUAL METAGENOMICS LIBRARIES

3.1. INTRODUCTION

Sequencing technology is an extensively studied area due to its potential to bring together the interests of scientists in the fields of comparative genomics, evolution, diagnostics, metagenomics and epidemiology. New sequencing technologies are being developed widely to overcome the expensive, labor intensive, and time consuming nature of the present ones. Among them, the Illumina/Solexa sequencing platform is the most successful and widely accepted next-generation sequencing technology for genome sequencing.

The Illumina/Solexa sequencing platform works on the principle of 'sequencing by synthesis' and is capable of generating hundreds of megabases of sequence data in a single run. It detects a single base when an individual base is incorporated to complement the DNA. DNA fragments that are intended to be sequenced are first immobilized on a surface. Then these fragments are PCR amplified so that they can be copied. Sequencing is then executed through a synthesis process: a mixture of four fluorescently labelled reversible chain terminators and DNA polymerase is used to sequence the clustered DNA [84]. The detection of the fluorescent signal is done for each template. The chain terminators and enzyme mix are added to start the next cycle, and the process is continued until the run ends [85].

The low sequencing cost and the amount of data generated are two prime advantages of this method. Because all four chain terminators are present in the reaction, the risk of misincorporation is lower in the Illumina/Solexa platform, which gives it greater accuracy and limits systematic errors [85]. However, the datasets from this process have been shown to have high error rates at the tail end of the reads [86]. To eliminate the errors in bad datasets, clipping the reads is one proven solution. Also, using higher sequence quality values could be a good strategy to detect bad sequences.

Among the instruments using the Illumina/Solexa platform, the 16 channel Illumina/Solexa HiSeq 2000 sequencer generates 60 Gbp on each channel and requires approximately 10 days of instrument time. In this study, faster runtime was achieved using the Illumina MiSeq instrument. However, it has higher costs per Gbp of sequencing data. Most of the data in this study were generated using the Illumina MiSeq platform.

3.1.1. Taxonomic identification of viral sequences in the metagenomics library from CD, UC and PSC stool samples

Taxonomic identification, i.e. assigning each sequence to the corresponding genome with the highest similarity, is one of the main goals of metagenomics. Due to the short length of the reads, the annotation of sequences is often a very challenging and tedious work, as the short sequences are less informative.

After the removal of low-quality data, the remaining high-quality sequence data were annotated using an approach called the sequence similarity-based method. This is the approach most commonly used to classify the taxonomy profile of metagenomics data; it is usually based on sorting the sequences using the BLAST search algorithm [80]. There is a number of challenges to deal with when BLAST is used to perform the annotation step, as the public database does not contain all the sequences of all species and it is not a true representation of the diversity among the living organisms [87]. This is particularly true in case of viruses, so the sequences generated in viral metagenomics remain unknown; and as a result, the classification of the sequence by sequence similarity method is elusive [88]. Moreover, novel viral sequences of unknown origin which have high genetic diversity from known virus sequences are difficult to classify. Therefore, finding sequence similarity using a BLASTx search increase the sensitivity of the sequence annotation step than solely using BLASTn for the classification of viral metagenomics data [89]. The advantage of BLASTx is that it can recognize and classify more related sequences based on similarities in amino acid coding codons, despite the existence of the differences in the synonymous codons in the sequences nucleotide fragments. However, one should keep in consideration that achieving more recognition of NGS data using rigorous and aggressive BLASTx search takes longer processing-time than BLASTn search.

3.2. RESULTS

3.2.1. Detection of viral sequences by BLASTn search

The reads were aligned against the sequences from the virus in the NCBI database using a BLASTn (nt/nr) search. These reads were recognized from individual metagenomics libraries were classified according to their similarity with the top hit of viral species. The majority of viral sequences (reads) were similar to the sequences from plant viruses that mainly infect edible fruits or crops and most likely ingested by vegetables consumption. Table 3.1 summarizes the list of major viruses detected and the distribution of the viral reads across the individual metagenomic libraries.

All the virus names detected similar to the reads in metagenomics libraries across the samples are listed in the Appendix 2. Since all of the libraries of CD, UC and PSC patients typically consisted of similar types of viruses, the virus sequence distribution of one typical patient (patient CD1) is discussed here as a representative of the rest. Patient CD1's RNA library consisted of a mixture of viral populations including eight viral families. Out of the eight families, the majority of sequences belonged to the viral family *Virgaviridae*. This family includes positive sense single-stranded RNA viruses like the pepper mild mottle virus, paprika mild mottle virus, tobacco mild green mosaic virus, cucumber green mottle mosaic virus, etc. Moreover, the major mammalian viral family included *Adenoviridae, Herpesviridae* and *Polyomaviridae*. The number of the virus sequences (read counts) similar to the virus sequences in NCBI database and the name of the corresponding viruses from patient CD1 RNA libraries are depicted in the Table 3.2.

Table 3.1: The distribution of the total viral read counts based on the sequence similarity search with NCBI virus database using BLASTn (nt/nr) in metagenomics libraries of faecal samples from different patients. The value in each boxes represents the number of total read counts for a virus and the value in the parentheses represents the number of positive patient(s) for the corresponding detected virus. The cut-off value was E=1e-5. CD: Crohn's Disease, UC: Ulcerative Colitis and PSC: Primary Sclerosing Cholangitis. n= Number of patients.

Viruses sequences identified by BLASTn	**Genome**	**CD (n=5)**	**UC (n=4)**	**PSC (n=1)**	**Control (n=2)**
Autographa californica nucleopolyhedrovirus	DNA	190 (3)	13 (2)	34 (1)	0
Pepper mild mottle virus	RNA	184,143 (3)	272 (2)	44 (1)	0
Paprika mild mottle virus	RNA	19,755 (4)	11 (1)	0	0
Paramecium bursaria chlorella virus	DNA	2014 (3)	75 (2)	154 (1)	10 (1)
Tobacco mild green mosaic virus	RNA	1660 (1)	79 (1)	0	0
Cucumber green mottle mosaic virus	RNA	814 (1)	0	4 (1)	0
Tomato mosaic virus	RNA	352 (3)	1 (1)	44 (1)	0
Bell pepper mottle virus	RNA	296 (1)	17 (1)	0	0
Plutella xylostella multiple nucleopolyhedrovirus	DNA	167 (2)	10 (1)	26 (1)	0
Cafeteria roenbergensis virus	DNA	15 (3)	40 (2)	104 (1)	0
Emiliania huxleyi virus	DNA	0	27 (2)	14 (1)	1 (1)
Tobacco mosaic virus	RNA	102 (2)	26 (1)	0	0
Herpesvirus	DNA	766 (4)	695 (3)	45 (1)	438 (1)
Poxvirus	DNA	6 (2)	66 (2)	44 (1)	9 (1)
Wiseana iridescent virus	DNA	14 (2)	47 (2)	24 (1)	8 (1)
human papillomavirus	DNA	4 (1)	0	37 (1)	0
Torque teno virus	DNA	0	0	22 (1)	0
Lumpy skin disease virus	DNA	5 (1)	2 (1)	2 (1)	0

Table 3.2: List of the name of the detected viruses, the number of the virus sequences (total read counts) similar to the virus sequences in NCBI database, the genome coverage of the corresponding virus by the reads from patient CD1 RNA metagenomics library

Viral sequences identified by BLASTn	Virus family	Read counts	Genome coverage (%)
Pepper mild mottle virus	*Virgaviridae*	172,184	99.69%
Paprika mild mottle virus	*Virgaviridae*	19,599	97.95%
Tobacco mild green mosaic virus	*Virgaviridae*	1,558	94.48%
Cucumber green mottle mosaic virus	*Virgaviridae*	616	69.72%
Tomato mosaic virus	*Virgaviridae*	335	65.38%
Bell pepper mottle tobamovirus	*Virgaviridae*	294	27.34%
Tobacco mosaic virus	*Virgaviridae*	92	9.76%
Simian virus 40	*Polyomaviridae*	91	7.23%
Rehmannia mosaic virus	*Virgaviridae*	83	9.44%
Tobacco vein clearing virus	*Caulimoviridae*	68	2.77%
Broad bean wilt virus 2 RNA1	*Secoviridae*	31	12.13%
Cucumber mosaic virus RNA 3	*Virgaviridae*	10	13.63%
Autographa californica nucleopolyhedrovirus	*Baculoviridae*	6	0.41%
Plutella xylostella multiple nucleopolyhedrovirus	*Baculoviridae*	5	0.30%
Human adenovirus type 1	*Adenoviridae*	2	0.42%
Human adenovirus C	*Adenoviridae*	2	0.42%
Human adenovirus 2	*Adenoviridae*	2	0.42%
Garlic latent virus	*Carlavirus*	2	3.18%
Human adenovirus 5	*Adenoviridae*	2	0.42%
Broad bean wilt virus 2 RNA2	*Secoviridae*	1	4.19%
Human herpesvirus 5	*Herpesviridae*	1	0.06%

3.2.2. Detection of viral sequences by BLASTx search

After the SOAPdenovo assembly was run to generate longer sequences in individual metagenomics libraries, a BLASTx search at a cut-off value of 1e-5 was executed on the assembled longer sequences for the accurate more annotation of NGS data. A typical list of virus species, identified from the annotated sequences of patient CD1, are presented in Table 3.3. The longest viral sequence assembled using this method was longer than 6000 bp and it aligned best to the paprika mild mottle virus genome. The longest mammalian virus sequence, a capsid coding gene of picobirnavirus, was 2040 bp long. Along with many plant virus sequences, various mammalian virus sequences, including picobirnavirus sequences, were assembled through BLASTx search.

Table 3.3: List of sequences from patient CD1 that had similarity with the protein sequences of NCBI virus database after BLASTx search

Assembled Sequences	Length (Bp)	Matched bases	Similarity (BP)	Protein	Virus
scaffold9	1,602	519	32.20%	Replicase protein	Pepper mild mottle virus
scaffold7	997	268	16.66%	p183 KDa protein	Tobacco mild green mosaic virus
scaffold30	6,237	899	55.63%	183kDa protein	Paprika mild mottle virus)
C15175	347	115	43.56%	30 KDa movement protein	Tomato mosaic virus
C15167	318	105	0.397727	Transport protein	Cucumber green mottle mosaic virus
scaffold13	1,033	201	78.21%	Transport protein	Pepper mild mottle virus
scaffold28	615	63	22.66%	183 kDa protein	Pepper mild mottle virus
scaffold49	2,040	345	65.71%	Capsid	Picobirnavirus
scaffold2	330	119	7.22%	RNA replicase	Cucumber green mottle mosaic virus
scaffold5	362	120	7.43%	183 kD protein	Tomato mosaic virus
scaffold17	363	78	7.02%	p126 kDa protein	Tobacco mild green mosaic virus
C15193	402	133	11.63%	129K replicase protein	Cucumber green mottle mosaic virus
scaffold21	941	274	17.03%	RNA-directed RNA polymerase	Tobacco mild green mosaic virus
scaffold40	457	62	38.99%	Coat protein	Tobacco mild green mosaic virus
C15255	1,646	534	100.00%	RNA dependent RNA polymerase	Human picobirnavirus

3.3. KEY FINDINGS

- After the alignment of short reads to the sequence of NCBI nt/nr database using BLASTn program, most of the reads due to its short sequence length remained ambiguous.
- Short contigs/scaffold using *de novo* assembly from individual samples remained unrecognized even after using more aggressive protein level searches.
- Therefore, there was a need for adopting a more efficient technique(s) to increase the sequence length to improve the identification of the NGS data.

CHAPTER 4: COMBINED ASSEMBLY IN THE IDENTIFICATION OF VIRUSES IN IBD AND PSC PATIENTS

4.1. INTRODUCTION

Metagenomics is a cost-effective, reliable and high throughput method to obtain the sequence information of collective genomes from all the microbes in a particular environment [90]. However, researchers have encountered the problem of recognizing the sequences because of the generation of short reads from NGS data which makes the interpretation of metagenomic data more challenging. Moreover, up to 60 to 99 % of the sequences from metagenomics samples may not be recognized [16, 22, 91, 92] due to the incomplete reference genomes in the databases and due to the lack of similarities between protein coding sequences from NGS data and the database sequences that caused by high relative mutation rates [93].

In this study, we obtained the dominant distribution of phage and plant viral sequences in the metagenomic data of the patients' stool samples; however, the major proportion of the short read sequences were not recognized. Thus, the strategy of combined assembly was adopted to maximize the coverage of the data and improve the overall sequence recognition. We assumed if a novel virus is present in multiple autoimmune diseases of similar kinds regardless of abundance of this unknown virus, the virus sequences from different samples will be joined by combined assembly to form longer sequences. In result the recognition of these sequences will then be improved so does the identification of viral species in the NGS data [94, 95].

4.2. COMBINED ASSEMBLY GENERATES LONGER SCAFFOLDS

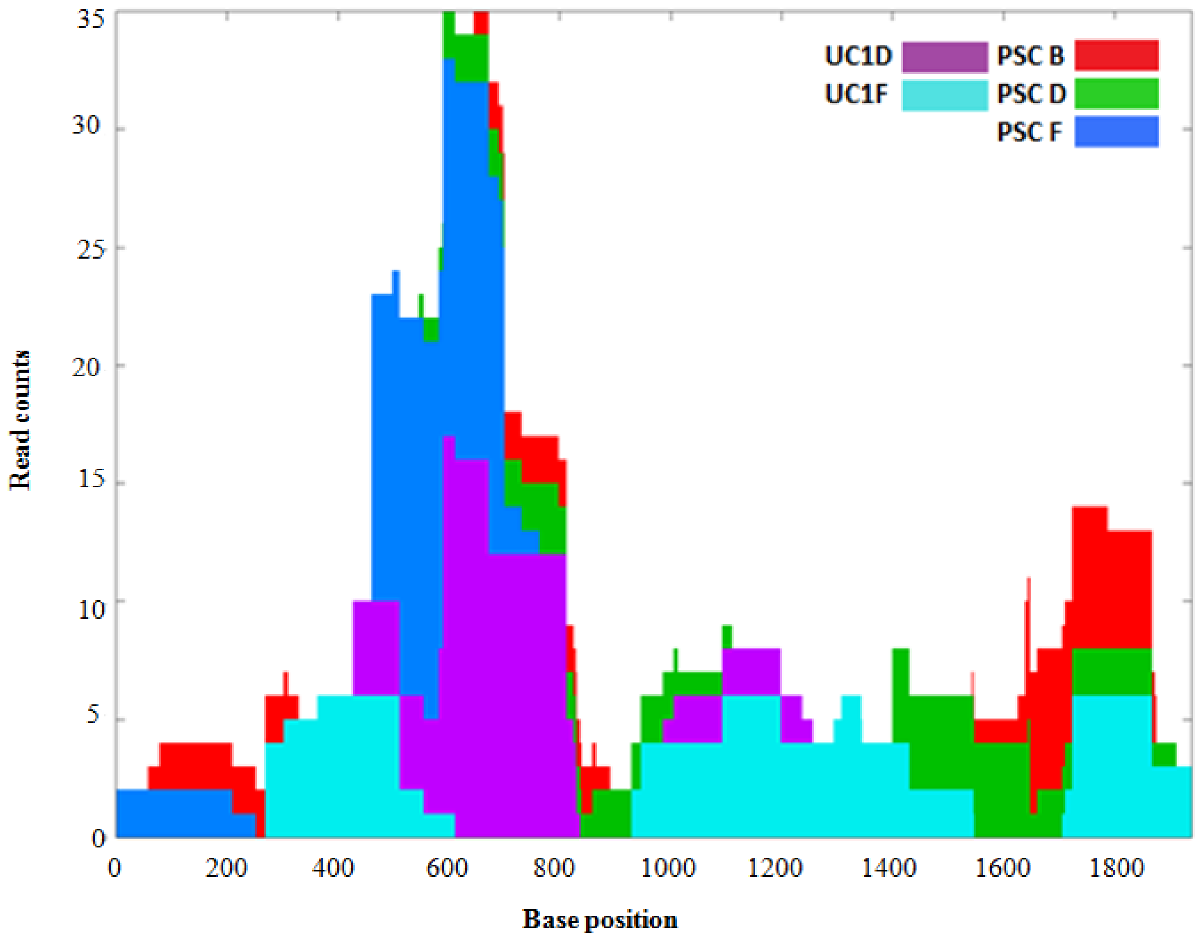


Figure 4.1: Combined assembly generates longer sequences when compared to the assembly of reads from individual libraries. An assembled sequence after combined assembly which was similar to the origin binding protein of an African swine fever like virus sequence of 1939 BP is taken as an example. Illustrated in the figure is the assembled sequence from multiple metagenomics libraries after performing the combined assembly approach to all reads from the metagenomics libraries and overlapping sequences joined in to form a long sequence. The longest assembled sequence of ASFLV were chosen to show how combined assembly benefits annotation of sequence after providing longer sequences than reads which otherwise might remained unrecognized. BP: base pairs

Combined assembly can generate sequences longer than the reads in a process described in Section 2.6. This ensures the use of more metagenome data from cross samples to increase the length of sequences and thus, more interpretation of sequencing data is achieved [96]. Figure 4.1 is an example from the real NGS data that shows, how many sequences (reads) from metagenomics libraries can join to form a longer sequence. When reads from individual libraries were aligned with BLAST, the short reads remained unrecognizable due to the length of sequences. But when these reads from different samples were assembled using the combined assembly approach, the reads appeared to originated from a virus that existed in samples of several patients and joined to form a longer sequence (contig/scaffold). This phenomenon is exactly illustrated in Figure 4.1 showing the overlapping reads joined into one long sequence after the combined assembly. Whereas without combined assembly these reads would remain as short fragments of DNA. In this example, reads from two different patients, UC1 and PSC, were obtained using the three viral enrichment methods. Reads of UC1D and UC1F, for instance, were originated from the nucleic acids, enriched by using two different methods of virus enrichment in a patient's sample (patient UC1). Likewise, reads of PSC B, PSC D and PSC F were obtained from the nucleic acids of a PSC patient's sample using three different methods. Each color represents the reads were originated from the same metagenomics library. Upon combined assembly these short reads were stitched into a sequence of 1939 bp in length.

4.3. TAXONOMIC CLASSIFICATION OF SEQUENCED DATA

A total of 159,342,088 reads were assembled using SOAPdenovo-Trans [144] assembly to give rise to 2,721,912 long sequences (contigs/scaffolds). The assembled contigs or scaffolds were searched for sequence similarity search using BLASTn algorithm [97]. This search of non-redundant nucleotide databases using assembled sequences as query sequences led to the recognition of a total of 877,821 contigs/scaffolds. These were annotated, and the sequences were binned as human, animal, algal, fungal, protozoan, bacterial, viral and phage sequences. However, 1,844,091 contigs/scaffolds were not recognized by the BLASTn search, which indicated that these sequences did not have any sequence similarity at the nucleotide-based alignment with the database sequences. Hence, these were tagged as un-annotated scaffolds. Furthermore, these un-annotated scaffolds were searched against the protein database using a translated nucleotide query (BLASTx) which led to the further annotation of 503,508 scaffolds that matched the protein coding ORFs from human and non-human sources. After BLASTx, total of 1,341,033 (49%) assembled contigs/scaffolds had no sequence similarity even after the protein coding ORF search of the assembled sequences. These sequences were tagged as unknown contigs/scaffolds. The bioinformatics pipeline used for combined assembly is shown in Figure 4.2.

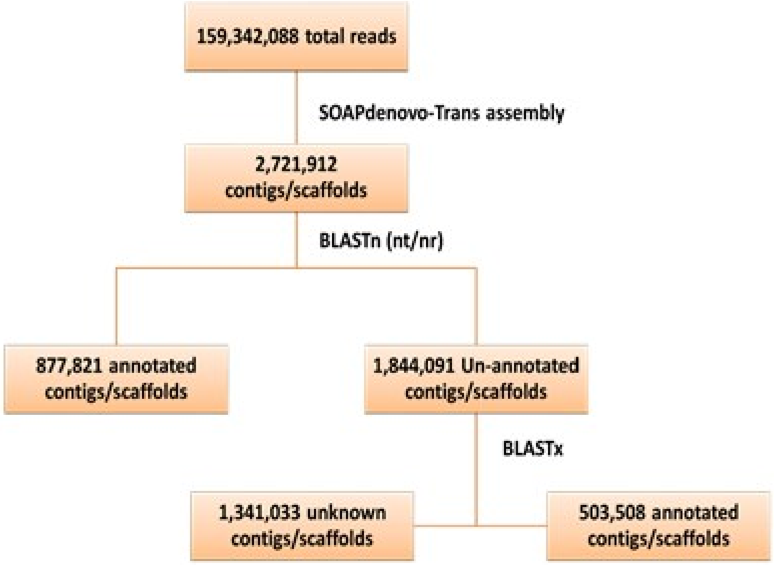

159,342,088 total reads
SOAPdenovo-Trans assembly
2,721,912 contigs/scaffolds
BLASTn (nt/nr)
877,821 annotated contigs/scaffolds
1,844,091 Un-annotated contigs/scaffolds
BLASTx
1,341,033 unknown contigs/scaffolds
503,508 annotated contigs/scaffolds

Figure 4.2: Bioinformatics approach to perform the combined assembly of metagenomics datasets and sequence annotation of the assembled contigs/scaffolds

After the search for similarities in the nucleotide sequences and protein homologies was completed, the sequences were classified into different taxa based on the search results (figure 4.3 and Appendix 13). Of these contigs/scaffolds, over a million sequences (37.65 % to total contigs/scaffolds) were categorized as bacterial sequences. The longest of assembled bacterial sequences was 145,465 bp in length. The second most abundant sequences belonged to the human taxon; it constituted 9.27% of the total number of sequences. 1.6 million reads were formed into 2,492 assembled viral sequences. The proportion of the viral assembled sequences was 0.1%. The longest virus sequence length was 6397 bp. 49% of contigs/scaffolds were remained unknown.

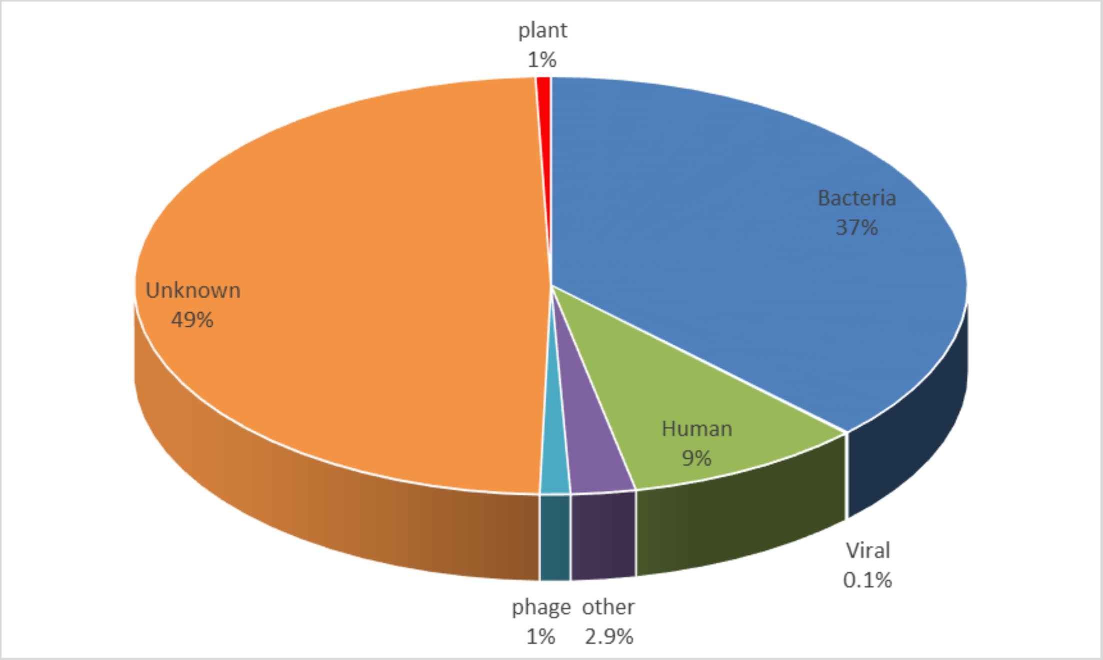


Figure 4.3: Classification of contigs/scaffolds based on the taxonomy of the assembled sequences (contigs/scaffolds) after running all the reads from metagenomics libraries through the SOAPdenovo-trans package [44]

4.4. DETECTED SEQUENCES OF VIRUS IN THE METAGENOMICS LIBRARIES AFTER COMBINED ASSEMBLY

About 2500 virus contigs/scaffolds were detected, and each one was classified into different viral species according to their top sequence match with sequences in database for the both nucleotides (BLASTn) and protein homology (BLASTx) searches. A total of twelve different types of viral sequences were identified using BLAST algorithms. They included: picobirnavirus, African swine fever like virus (ASFLV), herpesvirus, poxvirus, human papillomavirus (HPV), retrovirus and circovirus (Table 4.1). In result of combined assembly short viral reads were formed into longer sequences (contigs/scaffolds) that were used for further downstream analyses. Besides these mammalian virus sequences identified, some other assembled sequences was similar to the sequence from plant, fish, bird and insect viral sequences (Table 4.2).

Table 4.1: List of sequences similar to the sequence of mammalian viruses identified across the metagenomics libraries after BLASTn and BLASTx search on the assembled sequences from combined assembly approach. Name of the mammalian viruses with which assembled sequences showed similarity, the number of contigs/ scaffolds and the number of reads assembled into the contigs/scaffolds corresponding to each type of viruses, are shown

Mammalian virus sequences identified by BLASTn and BLASTx	Contigs/ Scaffolds	Number of reads
Human picobirnavirus	8	10215
African swine fever virus-like	96	576
Herpesviruses	12	134
Poxviruses	6	62
Human papillomavirus	11	68
Retroviruses	3	12
Torque teno virus	10	61
Circovirus like viruses	8	72
Hepatitis E virus	1	16
Rodent stool-associated circular genome virus	7	112
Cotia virus	1	26
Lumpy skin disease virus	1	7

Table 4.2: List of sequences similar to the sequence of plant and other (insect, bird and fish) viruses identified across the metagenomics libraries after BLASTn and BLASTx search on the assembled sequences from combined assembly approach. Name of the major plant and other viruses with which assembled sequences showed similarity, the number of contigs/ scaffolds and the number of reads assembled the contigs/scaffolds corresponding to each type of viruses, are shown

Plant and other virus sequences identified by BLASTn and BLASTx	Contigs/ Scaffolds	Number of reads
Pepper mild mottle virus	20	925,162
Paprika mild mottle virus	4	189,982
Tobacco mild green mosaic virus	15	15,099
Cucumber green mottle mosaic virus	10	4,045
Tomato mosaic virus	14	3,286
Bell pepper mottle virus	11	975
Broad bean wilt virus	3	520
Garlic common latent virus	5	416
Wheat rosette stunt virus	5	891
Tropical soda apple mosaic virus	1	154
Infectious pancreatic necrosis virus	1	106
Lymphocystis disease virus 1	12	65
Kadipiro virus	1	24
Rotifer birnavirus strain Palavas	2	20
Beak and feather disease virus	1	4
Tiger frog virus	1	2

Those virus like assembled sequences, identified after BLASTn and BLASTx, were searched once again by BLASTn and BLASTx one by one to gain more insights about those contigs/scaffolds. Following are the summary of the finding.

The picobirnavirus sequences were detected in mainly in patient CD1 RNA library among the metagenomics libraries. The picobirnavirus is commonly found in patient with diarrhoea. The capsid coding gene was aligned with the reference sequence of this virus and its identity was 40% at the amino acid-level. This distant relation implies picobirnavirus capsid protein present in the patient CD1 may be a new genotype of picobirnavirus.

ASFLV sequences were not reported to be found in patients with chronic diseases like IBD and PSC before. We also have not seen these viral sequences in individual metagenomics libraries even before adopting combined assembly approach. Therefore, we have analyzed further about this virus sequences which will be discussed later sections in this chapter.

A total of twelve assembled sequences in IBD and PSC patients were similar to the sequences from different types of herpesviruses of NCBI database. These sequences were ambiguous in nature as all of these sequences were similar not only to different types of herpesviruses but also were matched with human, bacteria and marine species (e.g., species belongs to Cnidaria and Chordata phylum). The results of BLASTn and BLASTx search of these sequences are listed in the Appendix 3.1.

A total of six assembled sequences in IBD and PSC patients were similar to the sequences from poxvirus of NCBI database. Except two of these assembled sequences (C4880850 and C7235113), rest of the sequences were similar to other species and rather these sequences fall under ambiguous category, having showed less similarity to the sequences from poxvirus after checking those sequences manually. The contig C4880850 was similar to the sequence of ribonucleotide reductase protein of variola virus with identity of 57% at the amino acid-level (E value was 5e-10) and the contig C7235113 was similar to the sequence of structural protein of *Molluscum contagiosum* (a DNA pox virus) with identity of 37% at the amino acid level, the E value was 4e-09 (see appendix 3.2).

There were three assembled sequences (contig C7407449 in PSC, C4274868 in CD2 and C6591025 in UC1) that showed similarity to the sequences from retroviruses after BLASTx search. Only the contig C7407449 was similar to the sequence of conserved reverse transcriptase-like protein of retrovirus with the identity of 40% at the amino acid-level, the E value was 8e-09 (see Appendix 3.3). The remainder two contigs, C4274868, and C6591025 were aligned with trimeric dUTP diphosphatase protein superfamily the E-values were 6e-12 and 4e-10 respectively. However, trimeric dUTP diphosphatases are the most common family of dUTPase that is found in bacteria, eukaryotes, and archaea.

4.5. DISTRIBUTION OF VIRAL READS IN IBD AND PSC PATIENTS

After combined assembly, the assembled sequences that showed similarity with some virus sequences from the NCBI database were reported to be associated with autoimmune diseases like IBD and also there were some assembled sequences that matched with other virus sequences from the NCBI database that were not found in IBD or PSC patients before. However, among the total read counts, the proportion of viral read counts after combined assembly ranged from 0 to 25.53% among the metagenomics libraries and the proportion of average viral read counts among the fifty-two metagenomics libraries was 0.6% of total read counts. The lower number of viral reads forming these assembled sequences among the metagenomics libraries suggested that the matched viruses were not abundant in any of the disease type and the distribution of viral reads did not show higher viral load to any particular disease types. Also in this study, we had a small sample size to perform a meaningful comparison of patients with different diagnosis. Hence we have not been able to perform any statistical analysis to associate the presence of any of these viruses with IBD or PSC.

A number of virus sequences were common in the samples from all of the patients. The human picobirnavirus sequences were detected in three CD patients, one of the four UC patients and the sole PSC patient. However, Majority of this virus sequences (98.48%) were detected in patient

CD1 (using TFF method). African swine fever virus-like sequences were detected in two of the CD (patient CD1 and CD2) and UC (patient UC1 and UC3) patients and in the PSC patient as well. Herpesviruses, which were previously reported as being associated with IBD [76, 98-102], were also detected different types of herpesvirus-like sequences in at least one of each types of IBD (CD and UC) and PSC patients. Assembled sequence similar to human papillomavirus type 8 was found in one CD patient (patient CD2) and assembled sequences similar to the sequences from HPV type 4, HPV type 24, HPV type 36, HPV type 123 and HPV type 155 were identified in the PSC patient. The retroviral sequences were observed in one of each the CD (feline immunodeficiency virus in patient CD2), UC (Feline immunodeficiency virus in patient UC1) and PSC (simian immunodeficiency virus) patients as well. Table 4.3 shows the virus distribution across the metagenomics libraries of differently diagnosed patients. In addition to mammalian and other viruses, plant viruses were also widely distributed among metagenomic libraries. Pepper mild mottle virus is the predominant plant virus found in the CD and UC patients. Wheat rosette stunt virus is the most common virus and is ubiquitously found in all the patients including in control libraries. Although we have seen a handful sequences from mammalian virus we have found the presence of african swine fever like virus (ASFLV) interesting and performed additional analyses on the sequences of ASFLV. The detection of the sequences of ASFLV in IBD and PSC was interesting because, this virus sequences were not detected from assembly of individual patients and upon combined assembly ninety six of the contigs/scaffolds were detected in multiple patients of IBD and PSC.

Table 4.3: The distribution of the total viral read counts based on the sequence similarity search with NCBI virus database using BLASTn and BLASTx on the combined assembled sequences in metagenomics libraries of faecal samples from different patients. The value in each boxes represents the number of total read counts for a virus and the value in the parentheses represents the number of positive patient(s) for the corresponding detected virus. The cut-off value was E=1e-5 for both BLASTn and BLASTx search. CD: Crohn's Disease, UC: Ulcerative Colitis and PSC: Primary Sclerosing Cholangitis. n= Number of patients for a particular diagnosis

Virus name	CD (n=5)	UC (n=4)	PSC (n=1)	Control (n=2)	References
Human picobirnavirus	10,207 (3)	6 (1)	2 (1)	0	Not reported
African swine fever virus	27 (2)	294 (2)	235 (1)	20 (2)	Not reported
Herpesviruses	7 (1)	38 (2)	70 (1)	19 (2)	Kandiel et al. [76] Kim et al. [103], Lawlor et al. [104] Spieker et al. [101]
Poxviruses	4 (1)	44 (1)	12 (1)	2 (1)	Not reported
Human papillomavirus	6 (1)	0	62 (1)	0	Greenberg et al. [105] Kong et al. [106]
Retroviruses	6 (1)	2 (1)	4 (1)	0	Vicente et al. [107]
Hepatitis E virus	0	0	6 (1)	0	Not reported
Lumpy skin disease virus	7 (1)	0	0	0	Not reported
Circovirus like virus	4 (1)	33 (1)	74 (1)	32 (1)	Not reported
Rodent stool-associated circular genome virus	2 (1)	26 (1)	73 (1)	11 (2)	Not reported
Cotia virus	22 (1)	4 (1)	0	0	Not reported
Infectious pancreatic necrosis virus	0	0	106 (1)	0	Not reported
Kadipiro virus	0	24 (1)	0	0	Not reported
Rotifer birnavirus strain Palavas	0	0	20 (1)	0	Nor reported
Lymphocystis disease virus 1	8 (2)	31 (2)	30 (1)	2 (1)	Not reported

4.6. EXAMPLES OF THE FULL OR PARTIAL ALIGNMENT OF CONTIGS/SCAFFOLDS WITH VIRAL REFERENCE GENOMES

The assembled contigs/scaffolds were utilized to align them against the reference genome as to observe if combined assembly help improving the genome coverage. Three examples of plant and mammalian virus genome coverage using the assembled sequences are given in Figure 4.4, 4.5 and 4.6.

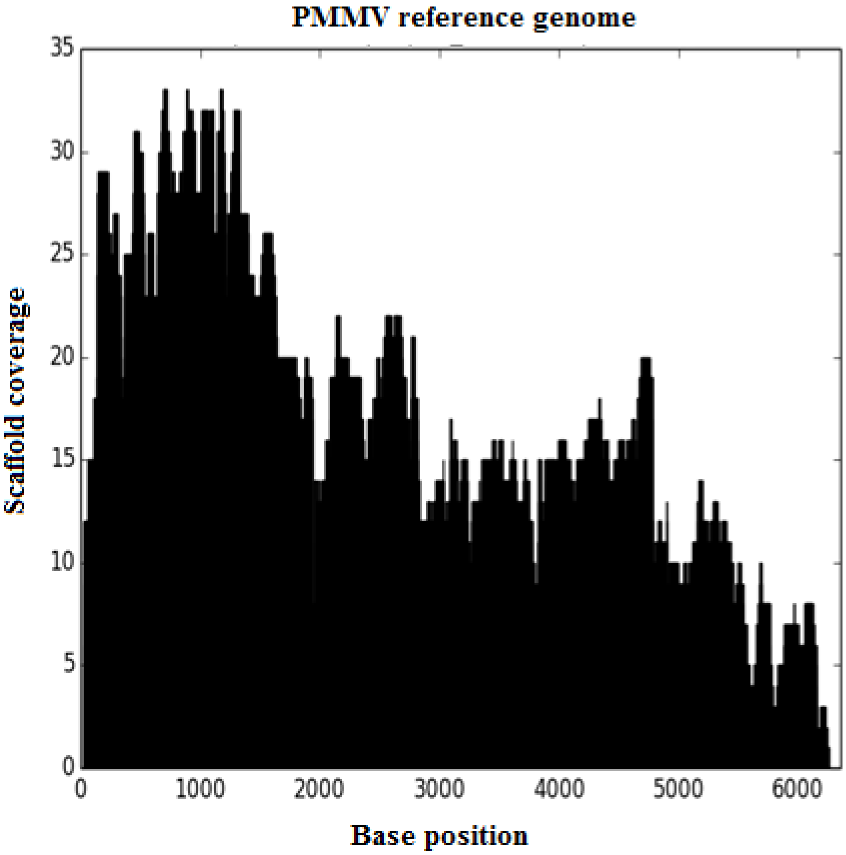


Figure 4.4: The alignment of sequences similar to the sequences of pepper mild mottle virus (PMMV) with PMMV reference genome (NCBI GenBank ID, gi: 20177424)

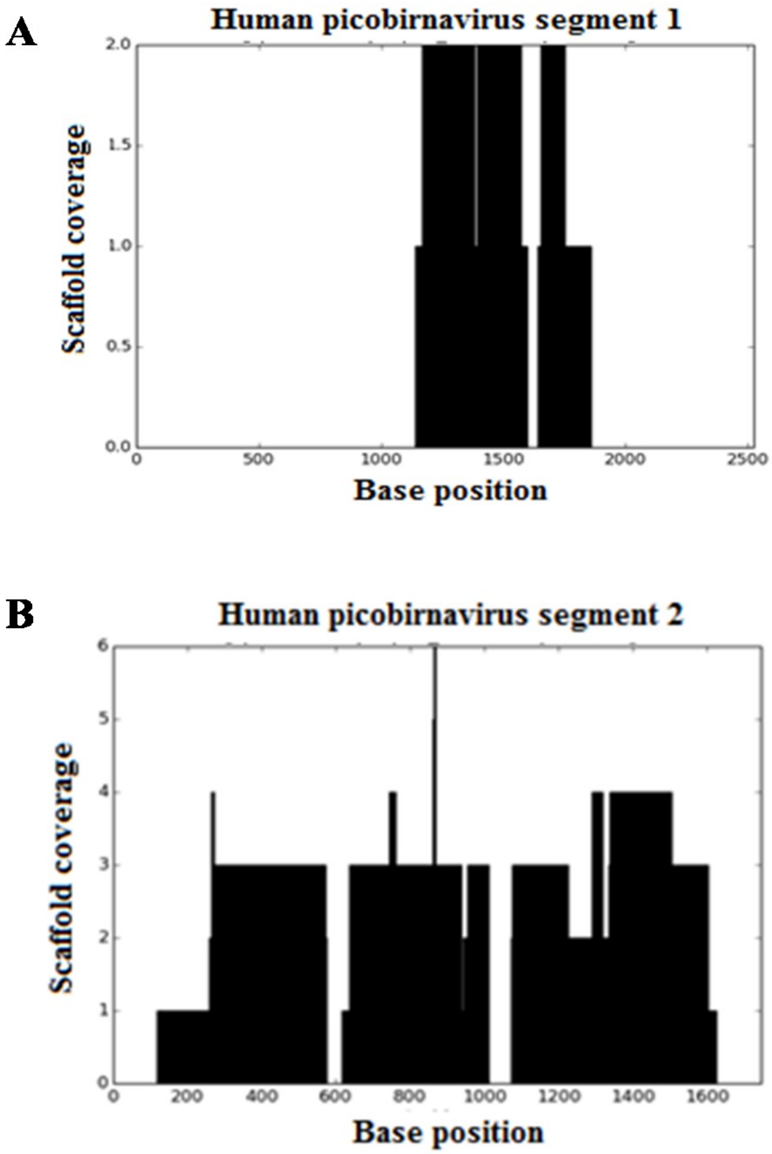


Figure 4.5: The alignment of assembled sequences (contigs/scaffolds) similar to the sequences of human picobirnavirus against the reference genome of human picobirnavirus from NCBI database. A) 'Segment 1' of the reference genome for human picobirnavirus (GI: 66391744) was used to align the contigs/scaffolds matched as human picobirnavirus capsid. Due to the sequence variability of the virus, the contigs/scaffolds were matched partially with its reference genome. The other region was highly variable and remained unmatched. B) 'Segment 2' of the reference genome for human picobirnavirus (gi:66391747) was used to align the contigs/scaffolds matched as human picobirnavirus, RNA-dependent RNA polymerase, and those sequences had coverage of 95% of the sequence of 'segment 2' the virus genome

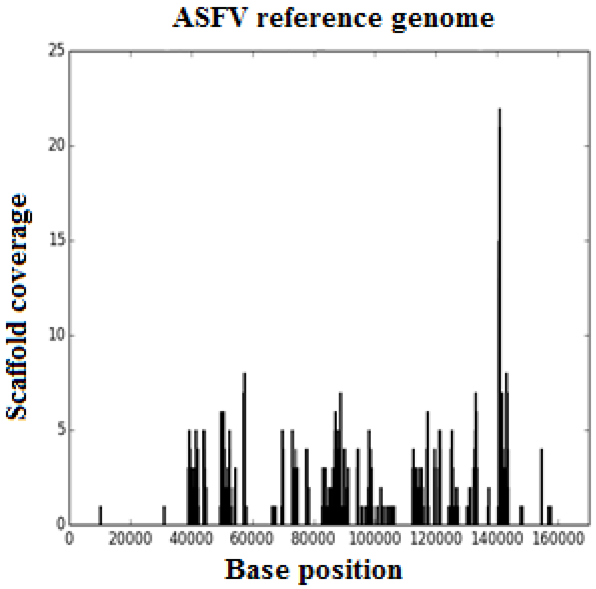


Figure 4.6: Alignment of ASFLV assembled sequences (contigs/scaffolds) against the reference genome of ASFV, BA71V (GI: 9628113). Sequences of the ASFLV were aligned against a total of thirty-nine protein-coding genes of reference genome of ASFV with coverage 11.2% of ASFV reference genome

4.7. PRESENCE OF MULTIPLE GENOTYPES OF A VIRUS IN IBD AND PSC SAMPLES

4.7.1. Multiple strains of picobirnavirus

To investigate the genetic diversity of assembled sequences that matched as picobirnavirus sequences, overlapping assembled sequences of RNA dependent RNA polymerase (RdRp) gene were taken for the phylogenetic analysis. the code name for a contig were presented here as start with an identifier C that is preceded by an integer number which was followed by the code name of the origin of the sequence among the metagenomics libraries, and for the assembled scaffolds that are represented here as with the identifier word "scaffold" preceded by an integer number and followed by the code name of the origin of the sequence among the metagenomics libraries. These sequences, along with the RdRp gene coding sequences from the NCBI database, were used to construct the tree. The constructed phylogenetic tree showed that the contigs or scaffolds were not clustered in a same clade; rather these assembled sequences were clustered together in different clades with different species according to their evolutionary relationship. The contig, C8033813 sequence was more related to mouse picobirnavirus RdRp than to human picobirnavirus RdRp. The topology also suggested that C3610560 and C6969679 were closer to the origin of porcine and feline picobirnavirus RdRp than to that of other human picobirnavirus RdRp. Scaffold172055, scaffold172056 and scaffold82127 remained in the clade with other RdRp of human picobirnavirus originating in the United States of America (USA) (Figure 4.7).

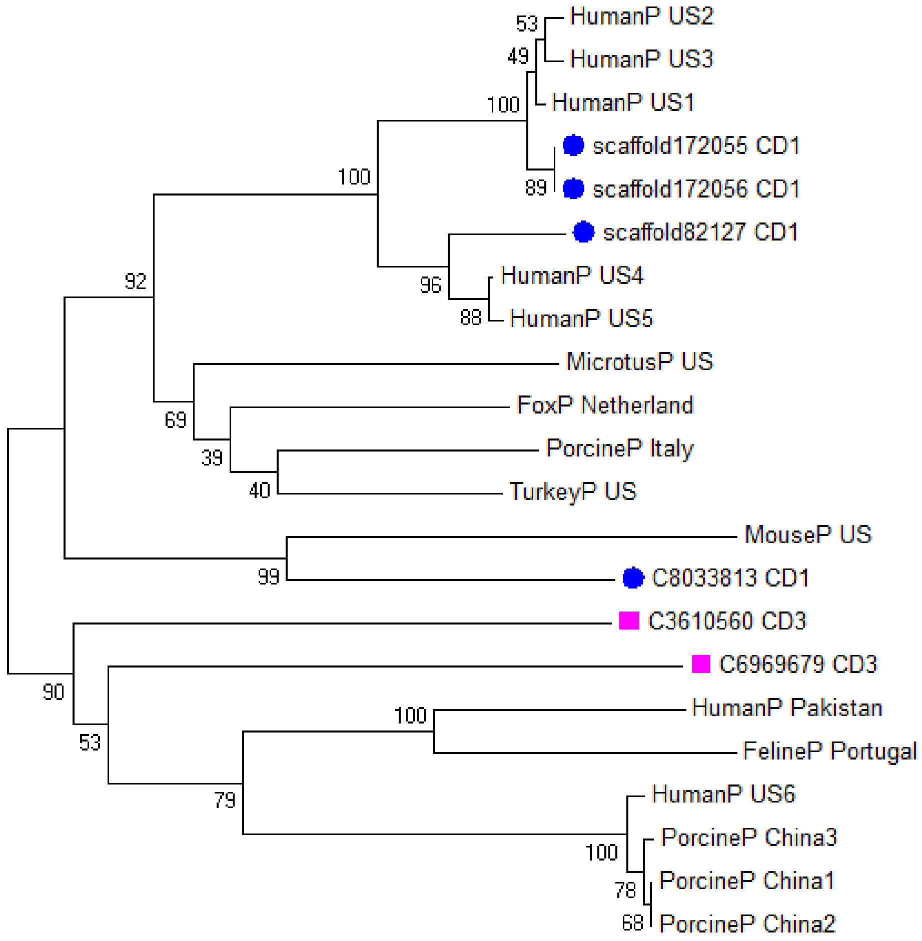


Figure 4.7: Phylogenetic analysis of assembled sequences of picobirnavirus RNA dependent RNA polymerase found in CD patients. Overlapping assembled sequences of RdRp from combined assembly of metagenomics libraries and the corresponding protein coding sequences of related picobirnavirus strains (human picobirnavirus RdRp US sequence 1-6, human picobirnavirus Pakistan, mouse picobirnavirus US, fox picobirnavirus Netherland, microtus picobirnavirus US, porcine picobirnavirus Italy, turkey picobirnavirus US sequence, porcine picobirnavirus China 1-3, and feline picobirnavirus Portugal) from NCBI were aligned for multiple alignment using ClustalW. MEGA 6 was used to generate a phylogenetic tree using a neighbor-joining (NJ) method with 1000 replication. Similarly colored, shaped circles and squares represents same origin of assembled sequences. Sequence names and respective Genebank IDs are given in the appendix 5.

4.7.2. Multiple strains of human papilloma virus

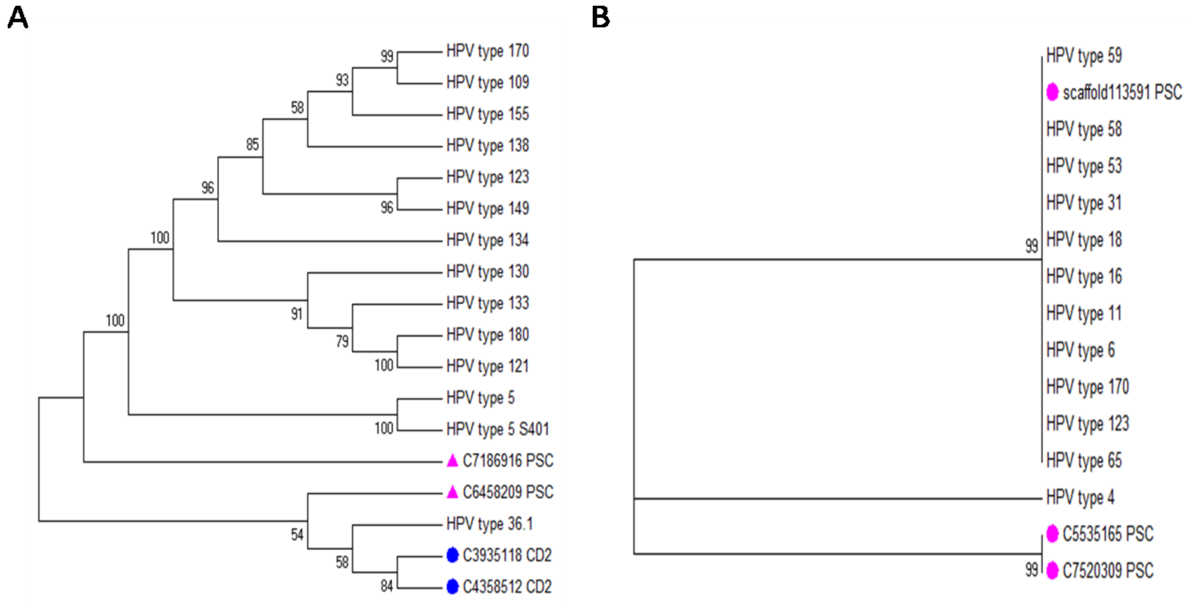


Figure 4.8: Phylogenetic analysis of assembled sequences of HPV capsid protein and transcription regulatory protein (E1) found in CD and PSC patients. Assembled sequences of capsid protein and transcription regulatory protein (E1) coding genes from two patients (patient CD2 and Patient PSC) and the equivalent protein coding nucleotide sequences of different HPV types were aligned in ClustalW, and a tree was constructed with MEGA 6 software using the NJ method with 1000 replication. Similarly colored, shaped circles and triangles represent same origin of assembled sequences. Sequence names and respective Genebank IDs are given in the appendix 6.1 and 6.2.

Overlapping assembled sequences similar to the sequence of capsid protein of HPV from patient PSC and patient CD2 were chosen to perform phylogenetic study of human papilloma virus (HPV) (Figure 4.8 A). The topology of the generated tree suggested that two contigs from a CD patient (patient CD2) were closely related to HPV type 36 and that the contig from patient PSC (C7186916) was more closely related to HPV type 5. The same procedure was followed to study the diversity of HPV transcription regulatory protein (E1) gene (Figure 4.8 B). Two sequences from the PSC samples were closer to sequence of HPV type 4 than to other types of HPV sequences. Scaffold 113591 from the patient PSC showed relatedness to HPV types 53, 58 and 59, and was distantly related with HPV type 4 virus.

4.7.3. Phylogenetic analysis of ASFLV sequence suggests distant relation from ASFV genes and the presence of multiple strains in the metagenomic libraries

To determine the diversity of the of ASFLV sequences, genes that code similar to ASFV-like proteins, namely- capsid, topoisomerase and helicase were used for phylogenetic analysis. Capsid protein coding (both the nucleotide and amino acid) sequences from 13 ASFV from reference genome, along with overlapping assembled sequences similar to the sequences of ASFV capsid protein from metagenomics libraries of patients, were used to construct phylogenetic trees. The two trees shared an almost identical topology, in which all of the ASFV-like capsid coding sequences resided in a cluster separate from the reference capsid protein cluster. The contigs, C6773767 and C8031519 clustered together in a clade, and the scaffold, *Scaffold68529* and the contig, C5671736 were clustered together, in separate clade. These two clades merged with the ASFV reference sequences (Figure 4.9 A). The amino acid-based tree showed the same topology (Figure 4.9 B).

Trees that were based on topoisomerase coding sequences, whether nucleotide or amino acid-based, showed a topology similar to that observed in the case of capsid protein—i.e., the sequenced gene of the ASFV-like virus topoisomerase clustered in a clade separate from that of the reference genes of the ASFV. As was expected, the two topoisomerase coding assembled sequences, C7483237 and scaffold152037, were in the same branch, merging with that of the contig C5979116 (Figure 4.9 C and D).

The phylogenetic analysis of helicase gene of ASFLV assembled sequences and ASFV reference gene sequences showed similar topology as observed in the previous two genes phylogenetic tree analyses. In this case, two contigs C6430553 and C6522131, from patient PSC and patient UC1 respectively, showed relatedness with the contig C7223994 from patient PSC, which remained distantly related to the reference ASFV gene sequences. Both the amino acid and nucleotide based sequences showed the same topology in the phylogenetic tree (Figure 4.9 E and F).

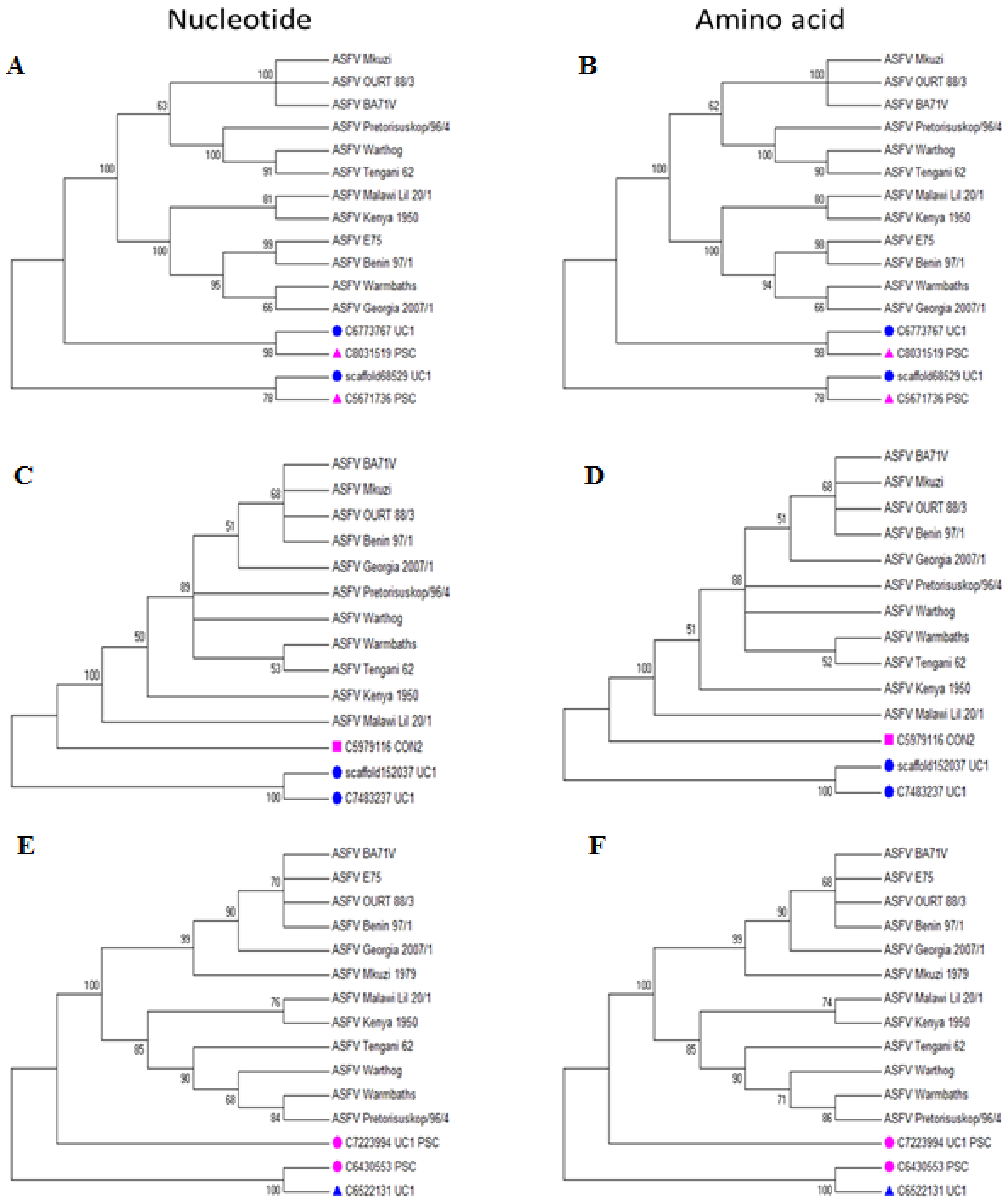


Figure 4.9: Phylogenetic analysis of assembled sequences of ASFLV capsid, topoisomerase and helicase protein found in IBD and PSC patients. The left panel shows the nucleotide-based phylogenetic tree, whereas the right panel depicts the amino acid-based trees which were constructed using the assembled sequences from the datasets of all the metagenomics libraries after performing the combined assembly. Overlapping assembled sequences and translated amino acid of respective genes were used for the phylogenetic analysis along with the ASVF capsid,

topoisomerase and helicase genes from the NCBI database for multiple alignments using ClustalW. MEGA 6 [83] was used to generate a phylogenetic tree using NJ with 1000 replications. (A) and (B) indicate the nucleotide and amino acid-based capsid protein phylogenetic trees respectively. (C) and (D) represent the nucleotide and amino acid-based topoisomerase phylogenetic trees respectively. (E) and (F) represent the nucleotide and amino acid-based helicase protein phylogenetic trees respectively. Similar colored, shaped squares, circles and triangles represent same origin of assembled sequences. Sequence names and respective Genebank IDs are given in the appendix 7.

4.8. SIMILARITY BETWEEN THE ASSEMBLED ASFLV CONTIGS/SCAFFOLDS AND THE SEQUENCES OF ASFLV FROM HUMAN AND SEWAGE SAMPLES AT THE NCBI DATABASE

A total of ninety six assembled sequences showed similar to thirty-nine genes of ASFV from NCBI database. Hence, these sequences were considered and annotated as ASFLV sequences. These contigs and scaffolds of varying length matched with different proteins of ASFV with identity level were ranging between 38% and 62% at amino acid-based levels (Appendix 8).

ASFLV sequences were reported to be found in human serum and sewage samples before. To determine the relationship between the assembled metagenomic sequence data and reported ASFV like sequences [108], both sets of sequences were aligned using the BLASTx algorithm. The results are presented in Table 4.4. The identity varied among different protein coding genes. The identity level ranges between 25.56 and 70.83 %, with most of the sequences identity ranged from 40 to 55%.

Table 4.4: Assembled sequences (contigs/scaffolds) of ASFLV protein coding genes are distantly related with previously reported ASFV-like sequences found in human and sewage samples

Contigs/scaffolds	Database	Length of Sequence (BP)	Match (AA)	Identity at amino acid level (%)	Protein
scaffold127139	gi270342034	488	54	57.41	Putative DNA primase
C5695234	gi270342037	219	64	26.56	DNA-directed RNA polymerase
C5744036	gi270342033	221	70	38.57	Putative helicase
C6200256	gi270342041	237	28	60.71	Transcription factor
C6428863	gi270342031	244	61	45.9	RNA helicase
C6430553	gi270342033	244	74	40.54	putative helicase
C6522131	gi270342030	246	63	46.03	genomic DNA
C6601827	gi270342017	249	62	41.94	RNApol2
C6653524	gi270342041	251	19	63.16	ATP- or GTP-binding motif
C6762348	gi270342020	258	24	70.83	origin binding protein
C6786441	gi270342041	259	57	40.35	Transcription factor
C6929668	gi270342017	269	67	49.25	RNApol2
C6995270	gi270342042	273	49	51.02	Alpha-NAC binding
C7163504	gi270342030	285	33	51.52	RNA helicase
C7318908	gi270342010	301	53	49.06	RNApol1
C7456564	gi270342030	318	28	53.57	RNA helicase
C7483237	gi270342012	322	64	40.62	topoisomerase homolog
C7735476	gi270342028	367	39	43.59	polymerase
C8031519	gi270342029	489	26	53.85	major structural protein p72

4.9. ASFLV SEQUENCE VARIANCE AMONG IBD AND PSC PATIENTS

Major structural protein p72 (gi: 210647), which encodes the capsid protein of ASFV, is a 2416 bp long sequence. This sequence was chosen since multiple assembled sequences (contigs) from metagenomics libraries showed best aligned with this capsid coding sequences. In the next step, all the sequences of ASFLV coding capsid protein from metagenomics libraries were aligned with the database capsid coding sequence. This alignment suggested that despite their similar functional prediction at amino acid level, at the nucleotide level they possessed a significant variation. All of the scaffolds matched only partially with the reference one. These variations in the sequences suggests that there may be more than one ASFLV genotypes present among the IBD and PSC libraries (Table 4.5).

Table 4.5: Alignment of African swine fever virus like capsid protein with capsid protein of ASFV from the reference genome

Sample code of the source	Contigs ID	Identity (AA)	Start	End
PSC B	C8031519	96	1	201
PSC D	C6721688	46	593	646
PSC F	C4156102	38	414	464
UC1C	C7230138	36	600	646
UC1E	C6098262	41	412	475
UC3B	C4155466	32	67	113

4.10. VALIDATION OF COMBINED ASSEMBLY RESULTS USING NESTED PCR

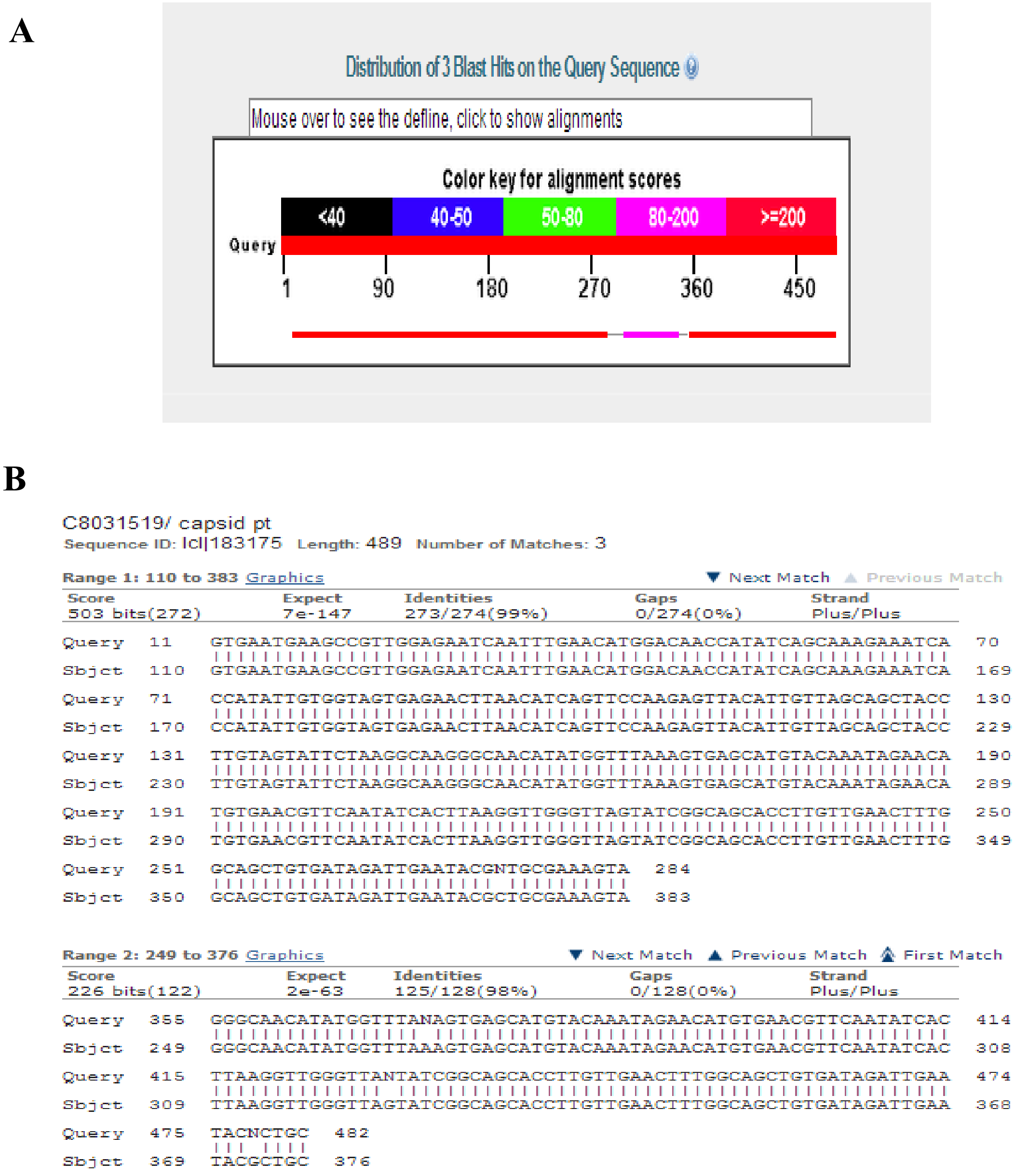
The combined assembly result was validated using nested PCR. Assembled sequences of four genes of ASFV having ten reads per million (TPM) for these genes present in at least one patient were chosen for the nested PCR amplification. Concentrated DNA from virus enriched stool samples from each of two CD patients (patient CD1 and CD2), UC patients (patient UC1 and

UC3) and the patient PSC were tested by nested PCR experiment for the validation of our findings. Since the ASFV-like virus was not highly abundant in our samples, nested PCR was used to amplify the target. The amplification was confirmed by running chips in Bioanlyzer or, QIAxcel run (Appendix 9). Three out of four test genes were positive in two of our five test samples (Table 4.6). Based on this result, we concluded that samples that have undergone viral enrichment can be detected by nested PCR, and that the threshold for such detection level is >= 30 reads/million. Also, the presence of multiple genotypes of this virus (which was evidenced from phylogenetic analysis) might be the reason that this virus was not detectable across all the positive samples from NGS datasets. The amplified fragments were Sanger sequenced. The sequenced fragments were then aligned with the assembled sequences of metagenomics data (Figure 4.10).

Table 4.6: A summary of the PCR experiment for the confirmation of the presence of ASFV sequences in patients' stools. (NC= negative control)

Patients		Origin Binding Protein	Capsid	Helicase	Reads/million
UC1	Test Sample	+	+	+	51.64
PSC		+	+	-	29.01
CD2		-	-	-	10.56
UC3		-	-	-	10.22
CD1		-	-	-	3.33
Water	NC*	-	-	-	0

A


Distribution of 3 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
90
180
270
360
450

B

```
C8031519/ capsid pt
Sequence ID: lcl|183175  Length: 489  Number of Matches: 3

Range 1: 110 to 383 Graphics                          ▼ Next Match  ▲ Previous Match
 Score              Expect     Identities        Gaps          Strand
 503 bits(272)      7e-147     273/274(99%)      0/274(0%)     Plus/Plus

Query  11    GTGAATGAAGCCGTTGGAGAATCAATTTGAACATGGACAACCATATCAGCAAAGAAATCA  70
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  110   GTGAATGAAGCCGTTGGAGAATCAATTTGAACATGGACAACCATATCAGCAAAGAAATCA  169

Query  71    CCATATTGTGGTAGTGAGAACTTAACATCAGTTCCAAGAGTTACATTGTTAGCAGCTACC  130
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  170   CCATATTGTGGTAGTGAGAACTTAACATCAGTTCCAAGAGTTACATTGTTAGCAGCTACC  229

Query  131   TTGTAGTATTCTAAGGCAAGGGCAACATATGGTTTAAAGTGAGCATGTACAAATAGAACA  190
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  230   TTGTAGTATTCTAAGGCAAGGGCAACATATGGTTTAAAGTGAGCATGTACAAATAGAACA  289

Query  191   TGTGAACGTTCAATATCACTTAAGGTTGGGTTAGTATCGGCAGCACCTTGTTGAACTTTG  250
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  290   TGTGAACGTTCAATATCACTTAAGGTTGGGTTAGTATCGGCAGCACCTTGTTGAACTTTG  349

Query  251   GCAGCTGTGATAGATTGAATACGNTGCGAAAGTA  284
             ||||||||||||||||||||||| ||||||||||
Sbjct  350   GCAGCTGTGATAGATTGAATACGCTGCGAAAGTA  383

Range 2: 249 to 376 Graphics         ▼ Next Match  ▲ Previous Match  ▲ First Match
 Score              Expect     Identities        Gaps          Strand
 226 bits(122)      2e-63      125/128(98%)      0/128(0%)     Plus/Plus

Query  355   GGGCAACATATGGTTTANAGTGAGCATGTACAAATAGAACATGTGAACGTTCAATATCAC  414
             ||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  249   GGGCAACATATGGTTTAAAGTGAGCATGTACAAATAGAACATGTGAACGTTCAATATCAC  308

Query  415   TTAAGGTTGGGTTANTATCGGCAGCACCTTGTTGAACTTTGGCAGCTGTGATAGATTGAA  474
             |||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  309   TTAAGGTTGGGTTAGTATCGGCAGCACCTTGTTGAACTTTGGCAGCTGTGATAGATTGAA  368

Query  475   TACNCTGC  482
             ||| ||||
Sbjct  369   TACGCTGC  376
```

Figure 4.10: Alignment of PCR amplified sequences with assembled sequences from the combined assembly of metagenomics datasets. Amplified segments of capsid protein from nested PCR were sequenced by Sanger sequencing, and the verification of the amplified segment was done by aligning the sequence data with previously sequenced Miseq data. A) Sequenced PCR product was matched with assembled sequences. The red color key indicates highly matched alignment. B) Pairwise alignment between PCR product and combined assembled sequenced data.

The nested PCR experiments were also performed using a pair of capsid primers to detect this virus in the different samples (plasma, white blood cells, lymph nodes, liver and colon tissue) of patients UC1 and PSC (Appendix 10). Only plasma from patient PSC was positive for this virus. The PCR result was positive when 25 ng/µL of extracted DNA from the plasma sample was used to detect capsid protein. Besides, test for the ASFV-like sequences in other patients was performed using DNA from colon samples, extracted from each of five Crohn's disease and ulcerative colitis patients to perform nested PCR. The PCR experiment was negative for all the samples (see Appendix 11). In this study, evidence for the presence of multiple genotypes of ASFLV in a viral population in some of the patients was observed after the data analysis and we have also observed that the primers in the PCR experiment only be able to amplify the targeted sequence in a sample but failed to amplify the sequence from other patient. This observation suggests the presence of multiple genotypes of ASFLV which has distant relation among its genotypes. Therefore, we concluded that the primers are specific for each genotypes of this putative virus like sequences. Conserved sequences are yet to be identified to detect the presence of this virus like sequences regardless of the genotypes present in host.

4.11. KEY FINDINGS

- Combined assembly is a promising approach to improve the identification of viruses with longer contigs/scaffolds.
- Different genotypes of mammalian viruses were identified from clinical samples based on phylogenetic analysis. The same protein coding sequences were distantly related to each other.

- Such sequences were amplified and verified by Sanger's sequencing. Using concentrated DNA from five positives patients for ASFLV sequences, two patients (Patient UC1 and Patient PSC) were positive for ASFLV by nested PCR.
- As it was clear from phylogenetic analysis that there is a presence of distantly related multiple genotypes of ASFLV in a sample, specific PCR primers were required to amplify for the targeted viral sequences to be amplified.
- We also have found none of the samples from different body-part of same patient except plasma from PSC were positive for ASFLV capsid by nested PCR which suggested that this virus may have specificity to a certain body part where it harboured more than others.
- It was observed that virus sequence with >30 reads per million in a sample from next generation sequencing data were required to amplify and verify by nested PCR and Sanger sequencing

CHAPTER 5: COMPARISON OF DIFFERENT VIRAL ENRICHMENT METHODS

5.1. INTRODUCTION

Success in viral metagenomics depends on molecular methods like PCR and on sequencing techniques, for both of which viral nucleic acid is the target. Therefore, focus on the methods of extraction and purification of viral nucleic acid is the priority. Depending on the nature of starting material, need for purity and quality of DNA or RNA from the sample and that of the viral nucleic acid collected after sample preparation, the enrichment procedure of viral nucleic acid is determined [109].

Viruses are difficult to study due to its small size, their nature to evolve rapidly, and the genomic flexibility of the viral nucleic acids [110]. The majority of viruses are difficult to culture in laboratory conditions due to the lack of a proper host system. In addition, viruses lack a single phylogenetic marker (e.g., conserved 16sRNA sequence of bacteria) to be used for studying their diversity and evolutionary pathways. Thus, it has become a necessity to adopt alternative technique(s) to study the virome [111, 112]. Viral metagenomics in corporation with bioinformatics and genome assembly are techniques that could be applied to study viruses in a manner in which a viral culture method is avoidable. However, for applying the techniques of high-throughput sequencing and analysis, enrichment of viruses from the samples needs to be done. There is a notable fact about our study sample i.e., clinical samples contain low viral loads and have high bacterial concentration as observed by the previous bacterial metagenomic studies on faecal samples. Thus, one of the principle aim for this study was to get rid of this unwanted background noise so that low abundant virus sequences can be detectable after the NGS data analysis. As the viruses have viral capsids providing them with the stiffness required to go through the concentration and purification steps, and given the small genome size of viruses [113], it is important to determine the potential method for viral enrichment prior to metagenomic library construction.

The aim of viral enrichment is to concentrate the isolated viral nucleic acid from the clinical samples and to get rid of any remaining host cells, nuclei and free nucleic acids in the resultant sample preparation. In the clinical sample, particularly in the stool samples, it is possible that a larger size and quantity of bacterial and eukaryotic genome than that of the target viral genome present. These unwanted sequences would show up as contaminants upon high-throughput sequencing, which would be time consuming and costly and could ruin the purpose of the experiment (to identify viruses). Thus, attempts should be made to overcome this masking of viral sequences by contaminant microbial and eukaryotic nucleic acids. Different viral preparation methods are effective in the enrichment of viral nucleic acids and the clearance of unwanted sequences. Among the viral preparation methods, tangential flow filtration (TFF), glass milk (silicon dioxide) particles and ultrafree MC microfiltration methods are most commonly adopted, depending on the nature of samples [31, 39, 114-117]. These three methods have their own advantages and disadvantages; as a result, it is important to compare their use in viral preparation from clinical faecal samples in order to determine the best method for viral enrichment in case of our sample before high-throughput next generation sequencing.

5.1.1. Advantages and disadvantages of each approach for viral preparation of stool samples for next generation sequencing

5.1.1.1. Tangential flow filtration (TFF)

TFF is an ultrafiltration membrane based method which is widely used to harvest viruses from natural water samples [114, 115, 118, 119]. TFF allows researchers to process a wide range of sample volumes, ranging from tens of millilitres to thousands of litres, by scaling the system components. This filtration approach is advantageous in many respects. For example, in the TFF method there is no preconditioning required; hence it is possible to recover a wide range of viruses including bacteriophages [120]. The large surface area of the filter allows filtrate to pass through rapidly in large volume. Moreover, tangential flow helps to prevent the retentate from clogging up the system. However there are some drawbacks to this method. Firstly, the cost and complexity of the equipment permits only one sample to be processed at a time, making the process costly and tedious. Secondly, significant losses of viruses occur due to the adsorption of the virus into the membrane. Particularly, the sensitivity of enveloped viruses to modifications occurring during the different TFF steps may cause this type of virus to be lost. Thirdly, although the majority of the viruses extracted from the filters are small, the large viruses (~720 nm particle) along with some filamentous viruses (2nm in length) cannot be extracted through the filters and will be lost in the TFF method [121, 122]. Finally in TFF method, a large volume of starting material is preferred to concentrate viruses than small volume of sample to start with the process of virus enrichment.

5.1.1.2. Glass milk (silicon dioxide) particles method

Several studies have adopted the glass milk (silicon dioxide) particles method for the extraction of nucleic acids [116, 117, 123, 124] where glass milk beads are used to which bind virus particles and pull down the viruses from the sample to settle at the bottom of the microcentrifuge tube. It has been reported to be particularly efficient in extracting viral RNA by RT-PCR from fecal materials [125, 126]. The major advantage of this method is that it is less labour intensive and a more rapid procedure than the TFF method. The method is simple, requiring only a single reaction tube and not involving any filtration membrane. Thus, the chance of losing virus particles is low in this method. However, due to requirement of careful pipetting and handling, there is up to a 50% increase in the risk of losing the RNA material [117].

5.1.1.3. Ultrafree MC microfiltration method

Ultrafree MC microfiltration method relies mainly on the hydrophobic polytetrafluoroethylene (PTFE) membrane and is an easy and fast microfiltration technique for the removal of prokaryotic and eukaryotic cells, leaving behind the virus particles in aqueous filtrate solution [31, 127-130]. This method is faster and less likely to be contaminated, since only a single tube is used to process per sample. There can be clogging of the PTFE membrane as the turbid fecal specimen passes through it, but this is avoidable if it is centrifuged at 12000 rpm for 2 minutes before the supernatant is filtered through the centrifugal filter units [31]. This method is much more efficient than the other methods, and it reduces the loss of virus particles due to its simple procedure requiring only one step. However, a number of enzymes need to be used to chew up nucleic acids, which may have an inhibitory effect in the process of metagenomic library construction.

5.2. COMPARISON OF DIFFERENT METHODS FOR VIRUS PARTICLE PURIFICATION FOR THE CONSTRUCTION OF METAGENOMIC LIBRARY

The above three methods were adopted for the enrichment of viral particle in the faecal samples of three CD, each of one UC and PSC patient to determine the best method to be applied in metagenomes. After that, the nucleic acids from these preparations were used to construct a metagenomic library. Sequencing was done, and high quality sequences were searched for sequence similarity to public databases using BLASTn and BLASTx searches and binned according to taxon (see Section 2.6 for the method of analysing metagenomic datasets). Because the focus was on viruses only, the sequences generated from different methods classified as viruses were then compared at the virus species level to determine the distribution of viral species in the metagenomes.

5.2.1. Based on the taxonomic classification

The number of reads, similar to the sequences of bacteria, human and phage taxa generated from different virus enrichment methods, did not greatly varied among the metagenomic libraries. The generation of sequences similar to virus from different methods also varied from sample to sample both in DNA and RNA metagenomics libraries.

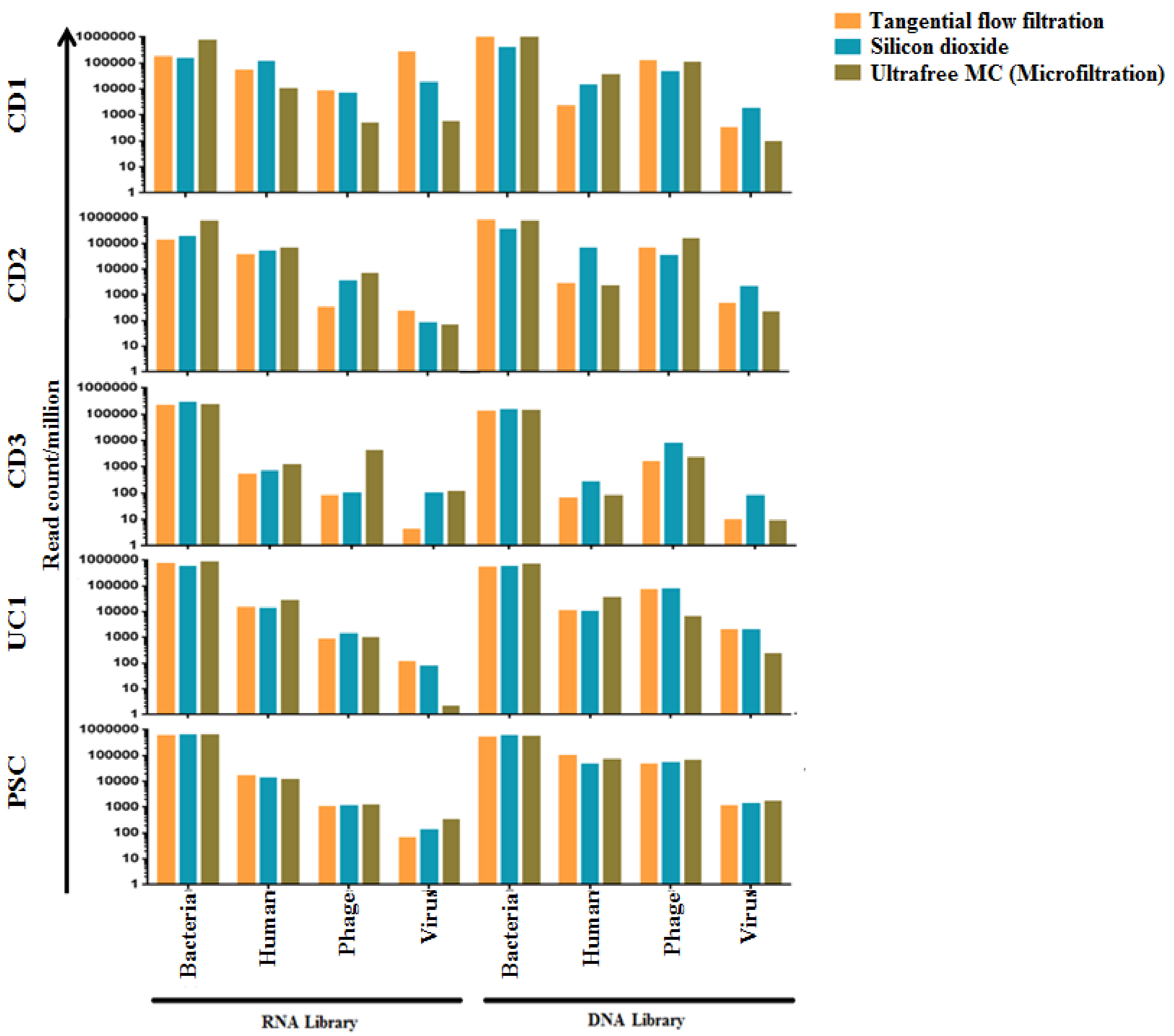


Figure 5.1: Comparison of assembled sequences matched as bacteria, human, phage and virus from faecal samples of three CD, one UC and one PSC patients using the three viral enrichment methods. The isolation and purification of virus from faecal samples of three CD patients (CD1, CD2 and CD3), one UC patient (UC1) and one PSC patient were done using tangential flow filtration, glass milk and ultrafree MC (microfiltration) methods simultaneously. DNA and RNA were extracted from each viral preparation and subjected to sequencing. The sequences were assembled using combined assembly and BLASTn and BLASTx were used to annotate those assembled sequences to determine the presence of sequences belonging to different taxa, and the three methods are compared for their ability to isolate the variety of taxa from samples. The read counts per million are shown for RNA and DNA library in a log scale for each of the taxa from sequences from different patients. CD1: First Crohn's Disease patient, CD2: Second Crohn's Disease patient, CD3: Third Crohn's Disease patient, UC: Ulcerative Colitis, PSC: Primary Sclerosis Cholangitis.

5.2.2. Based on the distribution of viruses found among the metagenomes

Again, not a single type of virus enrichment method outperformed the other two methods. The number of viral sequences detected by different methods varied greatly. One virus sequence found using one method was not necessarily found in the two libraries generated by the other two methods for the same patient. Considering the variability in virus species identification by different methods, in all CD and PSC DNA libraries, silicon di-oxide viral preparation method performed better than TFF or ultrafree MC (Microfiltration). However, for the UC patient, it was TFF that worked better to detect more viral species (Figure 5.2).

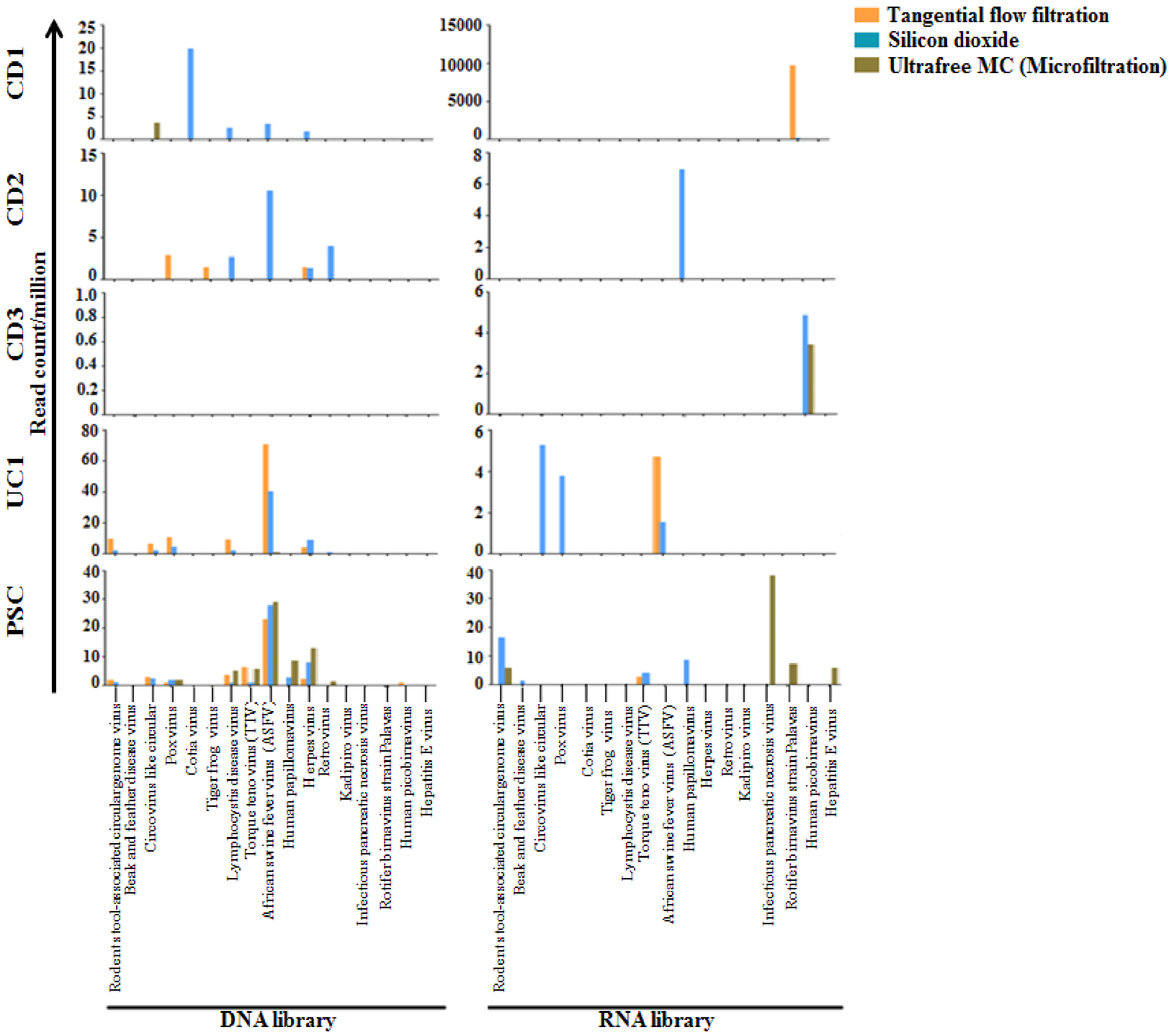


Figure 5.2: Comparison of three viral preparation methods on the basis of the viral sequences detected among the metagenomic libraries of the patients. The isolation and purification of viruses from stool samples of three CD patients (CD1,CD2 and CD3), one UC patient (UC1) and one PSC patient were done using tangential flow filtration, silicon dioxide and ultrafree MC (microfiltration) methods simultaneously. DNA and RNA were extracted from each viral preparation and subjected to sequencing. The sequences were assembled to determine the presence of viral sequences, and the three methods are compared for their ability to isolate the variety of viruses in the samples. The read counts per million are shown for the RNA and DNA library for each virus detected in the samples from different patients. CD1: First Crohn's Disease patient, CD2: Second Crohn's Disease patient, CD3: Third Crohn's Disease patient, UC: Ulcerative Colitis patient, PSC: Primary Sclerosis Cholangitis patient.

5.2.3. Based on number of viral sequences obtained from different viral preparations methods in various patients

When the total numbers of all viral sequences found in three CD, UC and PSC patients using the three methods of viral preparation were calculated, the results suggested that the TFF and silicon dioxide methods performed almost similarly in CD and UC patients' libraries, whereas the ultrafree MC (Microfiltration) method shows a higher viral sequence number with lower viral variation at the species level (Table 5.1).

Table 5.1: The distribution of sequences (total reads) that showed similarity to sequences at the NCBI virus database and distribution of virus species identified based on the sequence similarity search using combined assembly datasets. The values are shown as total viral read counts/number of viral species. #: number

Diagnosis	Tangential flow filtration (# read counts/ # viral species)	Silicon di-oxide (# read counts/ # viral species)	Ultrafree MC (microfiltration) (# read counts/ # viral species)
CD (3 patients)	10069/5	284/5	22/4
UC (1 patient)	264/6	189/6	2/1
PSC (1 patient)	95/7	172/8	206/6

Based on the sequencing datasets we found that there were cases when TFF method was better than other two methods (glass milk and Ultrafree MC microfiltration) for virus enrichment process prior to the construction of metagenomics library and there were cases where glass milk method was better choice over other two methods. The small sample size was a limitation for making a concrete choice of virus enrichment procedure. Roughly if the aim is to pick the viruses of various kind (diversity) then our result showed glass milk will be the appropriate choice. Whereas when the aim is to generate more viral sequences regardless of the number of different virus types then TFF method should be chosen. We also found TFF may be a good option for detecting RNA viruses. In our data, one RNA virus (human picobirnavirus) was detected in the RNA libraries in abundance whereas other two methods did not help concentrating this virus from the same patient or in a similar abundance from other patients.

5.3. KEY FINDINGS

- Not a single virus enrichment method was constantly stand-alone choice for isolating viral sequences and providing greater variability of viruses from the clinical samples
- In addition, there was variation on the virus detection quantity on the individual patient. The best method for one type of patient was not as good for other patients.
- The small sample size prevented us from making a conclusive statement regarding the appropriate method for viral enrichment.
- In conclusion, the viral preparation method has to be chosen carefully considering factors that include the diagnosis of patients from whom samples are collected, the origin of viral preparation (DNA or RNA), and the nature of samples.

CHAPTER 6: DISCUSSION

6.1. OVERVIEW

We have constructed the metagenomics libraries, sequenced them, and adopted the combined assembly approach to facilitate the use of more data than that available in individual metagenomic libraries in a cross sample analysis for the identification of viruses in faecal samples of IBD and PSC patients. The combined assembly approach, adopted in this study to improve the overall usage of metagenomics data, did indeed help to improve annotation and sequence recognition, and ultimately improved virus identification, in the clinical samples. Before obtaining raw sequencing data from samples for combined assembly in order to identify the viral content, the samples were enriched with virus particles through the use of various virus isolation and purification techniques. The comparison of viral data generated using different viral preparations methods was made afterwards to see the correlation of viral enrichment methods and the identification of viral content in the sample. After being sequenced, NGS data were run through the BLAST algorithms to analyze them for similarity to sequences in the known databases. Viral matches were compared so that known or unknown virus-like sequences were sorted. Finally, to utilize the unknown data, all of the reads from different samples were combined through *de novo* assembly process. The sequences were sorted according to the taxa. Overall virus identification was improved after adopting the combined assembly approach. Moreover, new virus-like sequences, including many genes having some relatedness to the ASFV genes, were detected in the stool samples of IBD and PSC patients. The longer sequences generated from the combined assembly approach were used in further downstream analysis for the interpretation and characterization of metagenomes in faecal samples, which showed that the combined assembly approach can be efficient in identifying virus sequences which otherwise might remain undetected. The sequences of virus detected in this thesis have opened a gateway to study the relation of this putative virus (if any) with autoimmune diseases like IBD and PSC. PCR experiments also verified the presence of ASFLV sequences in the clinical samples.

6.2. COMBINED ASSEMBLY IMPROVES THE IDENTIFICATION OF VIRUS IN NGS DATA THAN THE VIRUS IDENTIFIED BEFOREHAND USING ASSEMBLY OF INDIVIDUAL METAGENOMIC LIBRARY

Before adopting the combined assembly approach many of the short reads remained ambiguous owing to the short length of sequences generated from individual metagenomics libraries. Even after the assembly approach on individual libraries, the detection of notable mammalian viruses was not possible, rather a handful sequences of plant virus were obtained. Still many reads did not match with known sequences of NCBI database and hence, these reads remained un-annotated; and were categorized as unknown. Other groups also have performed viral metagenomics to explore and characterize the role of virome in a non-pathological state in humans and reported to detect most of them as phage and plant viruses [91, 131-136]. In 2003, Breitbart et al. found that in the faeces of a healthy adult, 59% of their sequences were unknown, with a majority of recognized sequences being identified as phages [135]. Reyes et al. and Minot et al. found 81% and 98% of sequences to be unknown in a virome dominated by phages [133, 136].

Since through the assembly of individual metagenomes, very few sequences similar to the sequences of mammalian viruses were found in this study, a comparative metagenomics study based on combined assembly approach was performed to increase the virus identification. Since the incomplete nature of the reference virus database creates a biased conclusion about unknown sequences of individual libraries after a sequence homology search, combined assembly was adopted so that more data could be utilized. Combined assembly uses the *de novo* assembly approach; thus, no reference genome sequence of a virus is required [44]. The approach takes the advantage of the presence of unique identifiers for each of the reads from all the metagenomics libraries to be combined [96]. In combined assembly, it was expected that the sequences derived from the same biological source would link up together with the assembly process owing to an existing sequence relationship within the genome.

In this study, the proportion of unknown sequences in the individual metagenomics DNA libraries ranged from 47 to 86 % whereas the unknown proportion for the RNA libraries ranged from 16 to 55 %. After performing the combined assembly, overall, identification of viral sequences in metagenomics libraries was seen to be much improved and only 49% of total assembled sequences remained as unknown categories. The assembled sequences from combined assembly approach showed similarity to the sequences of twelve mammalian virus sequences and six virus sequences from avian, insects and fishes were detected. A new virus-like sequence containing thirty-nine genes closest to the virus family *Asfarviridae* was found in UC, CD and PSC patients. The existence of variant sequences from the same virus particles suggested the presence of viral quasi species. Moreover, a nested PCR experiment validated the result of the metagenomics virus detection in the stool samples of the IBD and PSC patients. All in all, combined assembly has shown to be better for assembly of the reads from individual patients' data, and more viruses were effectively identified.

6.3. DETECTION OF VIRAL SEQUENCES IN THE CLINICAL SAMPLES

In the case of autoimmune diseases like IBD, viruses may possibly act directly on the host epithelium and immune system to induce inflammation, or they may induce dysbiosis in balanced commensal microbiota [137]. The presence of Epstein Barr virus (EBV) has been reported in a case-control-based IBD study [100-102]. This virus was found to be virulent to lymphocyte in intestines of UC patients and in a lesser level in CD patients [98, 99]. Lawlor et al. found that 79% of his IBD patients had cytomegalovirus (CMV) [104]. He found that upon inactivation of CMV it goes into a latent state, and under an immunosuppressed or stressed condition, it is reactivated, inducing colitis disease with few symptoms of IBD. We also found assembled sequences similar to herpesvirus sequences in faecal samples of IBD patients, with sequence variation from the existing public database. Similarly, some assembled sequences were similar to the gene coding sequences of human papillomavirus (HPV) in the metagenomics libraries of CD and PSC patients. This virus has been previously reported to be associated with squamous cell carcinoma in IBD patients [105, 106, 138]. However, further investigation is required to confirm that HPV may have an effect in autoimmune diseases such as IBD.

Detection of ASFLV sequences had been reported in human serum and sewage samples [108]. We have detected ninty-six contigs/scaffolds (including 39 encoding proteins) of the ASFLV in two UC, two CD and one PSC patients. Sequences of ASFLV were aligned with 11.2% (~19,041 BP) of the ASFV reference genome. The african swine fever virus (ASFV), sole member to the *Asfarviridae* family, causes hemorrhagic virema leading to death in domesticated and wild pigs; it is characterized by fever, hyperaemia of the skin, and haemorrhages of the internal organs, and carries a high mortality rate [139]. Although the structure or full genome information of this putative ASFLV is yet to be revealed, no threat to humans has been reported. However, in one metagenome-based study, ASFLV sequences were also detected in serum of patients who had febrile illness cases (with dengue-like symptoms) but tested negative for the dengue virus [140]. It is a possibility that the patients included in our study might have consumed ASFV infected pork due to which ASFV like sequences were found in the stool samples. Further investigation is needed to get insight into this possibility. Although no risk to humans of this putative virus infection has been reported so far, a virus may become pathogenic by gene mutation or insertion of gene (transposons) on its genome and may lead to significant health, economic and environmental problems. Therefore, further investigation of this putative ASFLV is required to understand its transmission and role of this virus (if any) in human.

The protein coding sequences of ASFLV in IBD and PSC patients were 38 to 62% identical with ASFV genes from NCBI that was isolated from pig at the amino acid levels in our study, whereas Loh et al. showed that identity levels between African swine fever-like virus (ASFLV) sequences isolated from human serum and sewage samples and ASFV sequences isolated from pig ranged from 27 to 64% [108]. ASFLV is most likely a new member in the family *Asfarviridae*, as the sequences of the virus found in this study have 25.56 to 70.83 % sequence similarities with ASFLV sequences detected in human serum and sewage samples [108]. The phylogenetic analysis of different protein from ASFLV suggested not only that these sequences are distantly related with ASFV virus sequences, but also that multiple genotypes of ASFLV might be present in patients with IBD and PSC.

Finally, the PCR amplification of detected nucleic acids is a widely accepted way to validate the annotated data of metagenomes [141-145]. In our sample, we have been able to amplify capsid, helicase and origin binding protein regions in the fecal DNA extracts from UC and PSC patients using nested PCR. The PCR detection of genes is correlated with the total numbers of reads per million (TPM) for individual samples. So far it is only possible to detect if a sequence has a representation of >30 reads/million in a sample. Part of the capsid gene of ASFLV has also been found in the plasma sample of the same PSC patient. This was an indication that the virus was transported to other parts of the body (via the bloodstream) from the gut. Moreover, the presence of multiple ASFLV genotypes evidenced from the NGS data. Therefore, we did not attempt to recover the unknown middle sequence region of the two genes through a PCR experiment. Also Wan et al., suggested that the gene order of ASFLV is not similar to the ASFV [144]. Hence, their attempt for amplifying the linking sequences between two genes was failed.

6.4. COMPARISON OF VIRAL ENRICHMENT METHODS

The comparison of different virus enrichment methods was performed to evaluate the efficiency of virus enrichment method on faecal samples prior to the construction of metagenomic libraries. Enrichment of virus particles in the sample prior to sequencing is a very crucial step for viral detection in samples; it leads to efficient viral nucleic acid extraction and removes the contamination of non-viral cells, resulting in the maximum number of viral sequences [23, 146, 147]. In this study, we found that the method of viral enrichment of samples is dependent on the nature of nucleic acid to be isolated. The glass milk method was found to be effective in generating more viral reads in the case of DNA preparation, whereas the tangential flow filtration (TFF) method was better for the enrichment of virus particles for RNA preparation from samples. It has been reported that the efficiency of the TFF method can be increased by the process described by Wommack et al. by using a large volume of 10 litres as the starting material. However, for most clinical samples, this is nearly impossible, as the volume of starting material is far less than required. Hale et al., 1996 found the glass milk method to be sensitive enough to enrich RNA viruses to a detectable level [124, 148]. In addition, Boom et al. found this method to work well in purifying RNA from human serum and urine samples [116].

Although ultrafree MC (Millipore) microfiltration method is a widely practised method of virus enrichment prior to the construction of metagenomes [131, 146, 149-151], its use on the libraries of the stool samples from all of the patients in this study, except that of the PSC patients, showed low efficiency in detecting virus sequences. This might be due to the detrimental effects of freezing and thawing sample, and harsh treatment with a mix of nuclease enzymes involved in the method, which might have led to physical damage to some of the viruses [152]. A study has been performed in which different methods of virus enrichment, either based on centrifugation, syringe-based filtration, nuclease base treatment or a combination of these were used. However, only the combination strategy led to an effective increase in the abundance of targeted viral particles and no substantial increase in the relative abundance of viruses in the metagenomics datasets with varying number of reads from bacteria and human host cells using either of the methods had been found [153]. However, the proportion of viral sequences they gained was only 1% of total metagenomics datasets, which is consistent with the results obtained in this study. Based on this study, it can be suggested that when viral enrichment methods need to be chosen to perform a viral metagenomics project, the method should not be adopted arbitrarily based on the previously published results; rather a method should be chosen based on the nature of the starting material and required end point purity of the nucleic acids. Another consideration to take before choosing virus enrichment methods is the genome of the targeted virus (i.e., DNA or RNA). However this consideration may not be applicable in viral metagenomics study aiming to discover a new virus as prior knowledge of what sort of virus are to be detected is lacking.

6.5. FUTURE DIRECTION

A major proportion of sequences from most of the clinical and the environmental metagenomics in the previous studies was remained unrecognized [16, 22, 88, 91, 92]. In our study after the combined assembly approach, 49% of total assembled sequences remained as unknown categories. Thus, an amino acid sequence similarity based approach (i.e., a conserved protein domain search) can be adopted to characterize the unknown sequences. Evolutionally related proteins from the same family member have more conserved sequences than their primary sequences [154, 155]. Many bioinformatics tools, including Pfam, CDD, SMART, and TIGRFAM have been developed to find these remote sequence similarities among the proteins and to identify them from their raw sequences [156-159]. More *in silico* tools are needed to be developed in order to characterize the unknown sequences in metagenomes. Secondly, we have detected some sequences of virus in the clinical samples with autoimmune diseases. We can perform, in future, case-control study based on the viral species detected in this thesis. Furthermore, to confirm the presence of ASFLV in the IBD patients and to investigate whether their entry into patient's body is via infected pork or the virus is associated with IBD, antibody detection specific to ASFLV in blood serum could be done using enzyme-linked immunosorbent assay (ELISA).

LIMITATIONS OF THE STUDY

One of the limitations of the study was that the sequence data from Illumina/Solexa technology generated were limited-length reads which could not be annotated using sequence similarity from the public database. As a result, the metagenomics data from individual samples remained unassembled given that the short sequences could not be related to each other [95]. The assembly approach was adopted to eliminate this limitation.

Secondly, availability of starting materials for performing viral metagenomics was limited; this was especially true of the material required to analyze the metagenomes of RNA libraries. For the isolation of 50 to 100 ng of yield viral RNA, 500 g of feces is required, whereas for isolation of the same amount of DNA only 1 g of human feces is sufficient [132, 160]. Given the lack of sufficient materials, we may not have obtained enough representative viruses in the gut virome of these patients. In addition, the number of subjects in the study from whom the samples were collected were only patients with nine IBD (five CD and four UC patients) and one PSC and two were normal controls. Due to the low sample size we have not been able to come to a conclusive evidence regarding virus involvement with IBD or PSC.

Another important limitation in this study was the bias created in the amplification step of library construction. Since the extracted nucleic acid material was not adequate to start the construction of metagenomes, there was a need for a nucleic acid amplification step before sequencing. Amplification steps introduce many biases in sequencing results: an adapter will ligate only to the dsDNA and will not bind to the ssDNA, and some ssDNA viruses will not be amplified and will be left out from a metagenomic library [161]. In addition, chimeras (reads from more than one species) can be generated and the process may introduce quantitative biases [162].

Contaminating nucleic acid material exists even after the nuclease treatment is applied to digest all of the unprotected DNA/RNA; nucleic acid remains protected inside the viral capsid. Most of the viral nucleic acids extracted in this study were apparently cleared from contaminant ribosomal and human mitochondrial DNA tested by PCR using 16S and 18S and human mitochondrial primers. However, a sample treated with DNase cannot completely eliminate free DNA [127]. Therefore, it is impossible to get rid of all the host genetic material from a viral concentration. Also, the introduction of nucleic acids can happen during the steps of the construction of metagenomic library or from the reagents and kits. However, we took enough precautions to keep the contamination level minimal.

BIBLIOGRAPHY

1. Leland, D.S. and C.C. Ginocchio, *Role of cell culture for virus detection in the age of technology.* Clin Microbiol Rev, 2007. **20**(1): p. 49-78.
2. Wang, D., et al., *Microarray-based detection and genotyping of viral pathogens.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(24): p. 15687-15692.
3. Bibby, K., *Metagenomic identification of viral pathogens.* Trends in Biotechnology, 2013. **31**(5): p. 275-279.
4. Schmidt, T.M., E.F. Delong, and N.R. Pace, *ANALYSIS OF A MARINE PICOPLANKTON COMMUNITY BY 16S RIBOSOMAL-RNA GENE CLONING AND SEQUENCING.* Journal of Bacteriology, 1991. **173**(14): p. 4371-4378.
5. Gross, E.L., et al., *Bacterial 16S Sequence Analysis of Severe Caries in Young Permanent Teeth.* Journal of Clinical Microbiology, 2010. **48**(11): p. 4121-4128.
6. Tringe, S.G. and P. Hugenholtz, *A renaissance for the pioneering 16S rRNA gene.* Current Opinion in Microbiology, 2008. **11**(5): p. 442-446.
7. Manichanh, C., et al., *A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library.* Nucleic Acids Research, 2008. **36**(16): p. 5180-5188.
8. Streit, W.R. and R.A. Schmitz, *Metagenomics - the key to the uncultured microbes.* Current Opinion in Microbiology, 2004. **7**(5): p. 492-498.
9. Liles, M.R., et al., *A census of rRNA genes and linked genomic sequences within a soil metagenomic library.* Applied and Environmental Microbiology, 2003. **69**(5): p. 2684-2691.
10. Rohwer, F. and R. Edwards, *The Phage Proteomic Tree: a genome-based taxonomy for phage.* Journal of Bacteriology, 2002. **184**(16): p. 4529-4535.
11. Riesenfeld, C.S., P.D. Schloss, and J. Handelsman, *Metagenomics: Genomic analysis of microbial communities.* Annual Review of Genetics, 2004. **38**: p. 525-552.
12. Schloss, P.D. and J. Handelsman, *Biotechnological prospects from metagenomics.* Current Opinion in Biotechnology, 2003. **14**(3): p. 303-310.
13. Krause, D.O., et al., *Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics.* Fems Microbiology Reviews, 2003. **27**(5): p. 663-693.
14. Rondon, M.R., et al., *Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms.* Applied and Environmental Microbiology, 2000. **66**(6): p. 2541-2547.
15. Brady, S.F. and J. Clardy, *Palmitoylputrescine, an antibiotic isolated from the heterologous expression of DNA extracted from bromeliad tank water.* Journal of Natural Products, 2004. **67**(8): p. 1283-1286.
16. Mokili, J.L., F. Rohwer, and B.E. Dutilh, *Metagenomics and future perspectives in virus discovery.* Curr Opin Virol, 2012. **2**(1): p. 63-77.
17. Yang, J., et al., *Unbiased Parallel Detection of Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach.* Journal of Clinical Microbiology, 2011. **49**(10): p. 3463-3469.
18. Tang, P. and C. Chiu, *Metagenomics for the discovery of novel human viruses.* Future Microbiology, 2010. **5**(2): p. 177-189.
19. Bexfield, N. and P. Kellam, *Metagenomics and the molecular identification of novel viruses.* Veterinary Journal, 2011. **190**(2): p. 191-198.

20. Rosario, K. and M. Breitbart, *Exploring the viral world through metagenomics.* Current Opinion in Virology, 2011. **1**(4): p. 289-297.

21. Patowary, A., et al., *De novo identification of viral pathogens from cell culture hologenomes.* BMC research notes, 2012. **5**: p. 11-11.

22. Nakamura, S., et al., *Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach.* PLoS One, 2009. **4**(1): p. e4219.

23. Delwart, E.L., *Viral metagenomics.* Rev Med Virol, 2007. **17**(2): p. 115-31.

24. Yongfeng, H., et al., *Direct pathogen detection from swab samples using a new high-throughput sequencing technology.* Clinical Microbiology and Infection, 2011. **17**(2): p. 241-244.

25. Clem, A.L., et al., *Virus detection and identification using random multiplex (RT)-PCR with 3 '-locked random primers.* Virology Journal, 2007. **4**.

26. Towner, J.S., et al., *Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in Uganda.* Plos Pathogens, 2008. **4**(11).

27. Kapoor, A., et al., *A highly prevalent and genetically diversified Picornaviridae genus in South Asian children.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(51): p. 20482-20487.

28. Holtz, L.R., et al., *Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea.* Virology Journal, 2008. **5**.

29. Greninger, A.L., et al., *The complete genome of klassevirus - a novel picornavirus in pediatric stool.* Virology Journal, 2009. **6**.

30. Lysholm, F., et al., *Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing.* Plos One, 2012. **7**(2).

31. Kapoor, A., et al., *A newly identified bocavirus species in human stool.* J Infect Dis, 2009. **199**(2): p. 196-200.

32. Palacios, G., et al., *A new arenavirus in a cluster of fatal transplant-associated diseases.* New England Journal of Medicine, 2008. **358**(10): p. 991-998.

33. Li, L., et al., *Genomic characterization of novel human parechovirus type.* Emerg Infect Dis, 2009. **15**(2): p. 288-91.

34. Woolhouse, M.E.J., et al., *Temporal trends in the discovery of human viruses.* Proceedings of the Royal Society B-Biological Sciences, 2008. **275**(1647): p. 2111-2115.

35. Allander, T., et al., *Cloning of a human parvovirus by molecular screening of respiratory tract samples.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(36): p. 12891-12896.

36. Feng, H., et al., *Clonal integration of a polyomavirus in human Merkel cell carcinoma.* Science, 2008. **319**(5866): p. 1096-1100.

37. Finkbeiner, S.R., A.F. Allred, and P.I. Tarr, *Metagenomic analysis of human diarrhea: viral detection and discovery.* PLoS Pathog, 2008. **4**: p. e1000011.

38. Briese, T., et al., *Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever-Associated Arenavirus from Southern Africa.* Plos Pathogens, 2009. **5**(5).

39. Victoria, J.G., et al., *Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis.* J Virol, 2009. **83**(9): p. 4642-51.

40. Quan, P.-L., et al., *Astrovirus Encephalitis in Boy with X-linked Agammaglobulinemia.* Emerging Infectious Diseases, 2010. **16**(6): p. 918-925.

41. Yozwiak, N.L., et al., *Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing.* Plos Neglected Tropical Diseases, 2012. **6**(2).

42. Ungchusak, K., et al., *Probable person-to-person transmission of avian influenza A (H5N1).* New England Journal of Medicine, 2005. **352**(4): p. 333-340.

43. Smith, G.J.D., et al., *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.* Nature, 2009. **459**(7250): p. 1122-U107.

44. Xie, Y., et al., *SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads.* Bioinformatics, 2014. **30**(12): p. 1660-6.

45. Miller, J., S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**(6): p. 315 - 327.

46. Pevzner, P., H. Tang, and M. Waterman, *An Eulerian path approach to DNA fragment assembly.* Proc Natl Acad Sci USA, 2001. **98**(17): p. 9748 - 9753.

47. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.

48. Xavier, R.J. and D.K. Podolsky, *Unravelling the pathogenesis of inflammatory bowel disease.* Nature, 2007. **448**(7152): p. 427-434.

49. Stange, E.F., et al., *European evidence based consensus on the diagnosis and management of Crohn's disease: definitions and diagnosis.* Gut, 2006. **55 Suppl 1**: p. i1-15.

50. Rothfuss, K.S., E.F. Stange, and K.R. Herrlinger, *Extraintestinal manifestations and complications in inflammatory bowel diseases.* World J Gastroenterol, 2006. **12**(30): p. 4819-31.

51. Stenson, W.F., S. B. Hanauer, et al. , *Inflammatory bowel disease* Textbook of Gastroenterology, ed. D.H.A. T. Yamada, A. N. Kallooet al. Hoboken. Vol. 1. 2009, NJ: Wiley-Blackwell. 1386-1472.

52. Kugathasan, S. and C. Fiocchi, *Progress in basic inflammatory bowel disease research.* Seminars in pediatric surgery, 2007. **16**(3): p. 146-53.

53. Danese, S. and C. Fiocchi, *Etiopathogenesis of inflammatory bowel diseases.* World Journal of Gastroenterology, 2006. **12**(30): p. 4807-4812.

54. Podolsky, D.K., *Inflammatory bowel disease.* New England Journal of Medicine, 2002. **347**(6): p. 417-429.

55. Frank, D.N., et al., *Disease Phenotype and Genotype Are Associated with Shifts in Intestinal-associated Microbiota in Inflammatory Bowel Diseases.* Inflammatory Bowel Diseases, 2011. **17**(1): p. 179-184.

56. Smith, M.P. and R.H. Loe, *Sclerosing Cholangitis; Review of Recent Case Reports and Associated Diseases and Four New Cases.* Am J Surg, 1965. **110**: p. 239-46.

57. Lee, Y.M. and M.M. Kaplan, *Primary sclerosing cholangitis.* N Engl J Med, 1995. **332**(14): p. 924-33.

58. Mitchell, S.A., et al., *Cigarette smoking, appendectomy, and tonsillectomy as risk factors for the development of primary sclerosing cholangitis: a case control study.* Gut, 2002. **51**(4): p. 567-73.

59. Chapman, R.W., et al., *Association of primary sclerosing cholangitis with HLA-B8.* Gut, 1983. **24**(1): p. 38-41.

60. Farrant, J.M., et al., *Amino acid substitutions at position 38 of the DR beta polypeptide confer susceptibility to and protection from primary sclerosing cholangitis.* Hepatology, 1992. **16**(2): p. 390-5.

61. Tischendorf, J.J., et al., *Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: A single center study.* Am J Gastroenterol, 2007. **102**(1): p. 107-14.

62. Wiesner, R.H., et al., *Primary sclerosing cholangitis: natural history, prognostic factors and survival analysis.* Hepatology, 1989. **10**(4): p. 430-6.

63. Broome, U., et al., *Natural history and prognostic factors in 305 Swedish patients with primary sclerosing cholangitis.* Gut, 1996. **38**(4): p. 610-5.

64. Loftus, E.V., Jr., et al., *PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis.* Gut, 2005. **54**(1): p. 91-6.

65. Olsson, R., et al., *Prevalence of primary sclerosing cholangitis in patients with ulcerative colitis.* Gastroenterology, 1991. **100**(5 Pt 1): p. 1319-23.

66. Saich, R. and R. Chapman, *Primary sclerosing cholangitis, autoimmune hepatitis and overlap syndromes in inflammatory bowel disease.* World J Gastroenterol, 2008. **14**(3): p. 331-7.

67. Yoshimura, H.H., M.K. Estes, and D.Y. Graham, *Search for evidence of a viral aetiology for inflammatory bowel disease.* Gut, 1984. **25**: p. 347-355.

68. Phillpotts, R.J., J. Hermon-Taylor, and B.N. Brooke, *Virus isolation studies in Crohn/'s disease: a negative report.* Gut, 1979. **20**: p. 1057-1062.

69. Phillpotts, R.J., J. Hermon-Taylor, and N.M. Teich, *A search for persistent virus infection in Crohn/'s disease.* Gut, 1980. **21**: p. 202-207.

70. Van Kruiningen, H.J., M. Poulin, and A.E. Garmendia, *Search for evidence of recurring or persistent viruses in Crohn/'s disease.* APMIS, 2007. **115**: p. 962-968.

71. Sura, R., B. Gavrilov, and L. Flamand, *Human herpesvirus-6 in patients with Crohn/'s disease.* APMIS, 2010. **118**: p. 394-400.

72. Lepage, P., J. Colombet, and P. Marteau, *Dysbiosis in inflammatory bowel disease: a role for bacteriophages?* Gut, 2008. **57**: p. 424-425.

73. Bernstein, C.N., P. Rawsthorne, and J.F. Blanchard, *Population-based case-control study of measles, mumps, and rubella and inflammatory bowel disease.* Inflamm Bowel Dis, 2007. **13**(6): p. 759-62.

74. Bernstein, C.N. and J.F. Blanchard, *Viruses and inflammatory bowel disease: is there evidence for a causal association?* Inflamm Bowel Dis, 2000. **6**(1): p. 34-9.

75. Hussein, K., et al., *Acute cytomegalovirus infection associated with the onset of inflammatory bowel disease.* Am J Med Sci, 2006. **331**(1): p. 40-3.

76. Kandiel, A. and B. Lashner, *Cytomegalovirus colitis complicating inflammatory bowel disease.* Am J Gastroenterol, 2006. **101**(12): p. 2857-65.

77. Cadwell, K., et al., *Virus-Plus-Susceptibility Gene Interaction Determines Crohn's Disease Gene Atg16L1 Phenotypes in Intestine.* Cell, 2010. **141**(7): p. 1135-U64.

78. Hubbard, V.M. and K. Cadwell, *Viruses, autophagy genes, and Crohn/'s disease.* Viruses, 2011. **3**: p. 1281-1311.

79. Law, J., et al., *Identification of Hepatotropic Viruses from Plasma Using Deep Sequencing: A Next Generation Diagnostic Tool.* PLoS ONE, 2013. **8**(4): p. e60595.

80. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

81. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-714.

82. Tarailo-Graovac, M. and N. Chen, *Using RepeatMasker to identify repetitive elements in genomic sequences.* Curr Protoc Bioinformatics, 2009. **Chapter 4**: p. Unit 4 10.

83. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.

84. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

85. Bentley, D.R., *Whole-genome re-sequencing.* Curr Opin Genet Dev, 2006. **16**(6): p. 545-52.

86. Nakamura, K., et al., *Sequence-specific error profile of Illumina sequencers.* Nucleic Acids Res, 2011. **39**(13): p. e90.

87. McHardy, A.C. and I. Rigoutsos, *What's in the mix: phylogenetic classification of metagenome sequence samples.* Curr Opin Microbiol, 2007. **10**(5): p. 499-503.

88. Edwards, R.A. and F. Rohwer, *Viral metagenomics.* Nat Rev Microbiol, 2005. **3**(6): p. 504-10.

89. Kunin, V., et al., *A bioinformatician's guide to metagenomics.* Microbiol Mol Biol Rev, 2008. **72**(4): p. 557-78, Table of Contents.

90. Hugenholtz, P. and G.W. Tyson, *Microbiology: metagenomics.* Nature, 2008. **455**(7212): p. 481-3.

91. Minot, S., et al., *Hypervariable loci in the human gut virome.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(10): p. 3962-3966.
92. Reyes, A., et al., *Viruses in the faecal microbiota of monozygotic twins and their mothers.* Nature, 2010. **466**(7304): p. 334-U81.
93. Bonhoeffer, S. and P. Sniegowski, *Virus evolution: the importance of being erroneous.* Nature, 2002. **420**(6914): p. 367, 369.
94. Charuvaka, A. and H. Rangwala, *Evaluation of short read metagenomic assembly.* BMC Genomics, 2011. **12**(Suppl 2): p. S8.
95. Wommack, K., J. Bhavsar, and J. Ravel, *Metagenomics: read length matters.* Appl Environ Microbiol, 2008. **74**(5): p. 1453 - 1463.
96. Dutilh, B.E., et al., *Reference-independent comparative metagenomics using cross-assembly: crAss.* Bioinformatics, 2012. **28**(24): p. 3225-3231.
97. Altschul, S.F., et al., *BASIC LOCAL ALIGNMENT SEARCH TOOL.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
98. Lidar, M., P. Langevitz, and Y. Shoenfeld, *The role of infection in inflammatory bowel disease: initiation, exacerbation and protection.* Isr Med Assoc J, 2009. **11**(9): p. 558-63.
99. Yanai, H., et al., *Epstein-Barr virus infection of the colon with inflammatory bowel disease.* Am J Gastroenterol, 1999. **94**(6): p. 1582-6.
100. Sankaran-Walters, S., et al., *Epstein-Barr virus replication linked to B cell proliferation in inflamed areas of colonic mucosa of patients with inflammatory bowel disease.* J Clin Virol, 2011. **50**(1): p. 31-6.
101. Spieker, T. and H. Herbst, *Distribution and phenotype of Epstein-Barr virus-infected cells in inflammatory bowel disease.* Am J Pathol, 2000. **157**(1): p. 51-7.
102. Wakefield, A.J., et al., *Detection of herpesvirus DNA in the large intestine of patients with ulcerative colitis and Crohn's disease using the nested polymerase chain reaction.* J Med Virol, 1992. **38**(3): p. 183-90.
103. Kim, J.J., et al., *Cytomegalovirus infection in patients with active inflammatory bowel disease.* Dig Dis Sci, 2010. **55**(4): p. 1059-65.
104. Lawlor, G. and A.C. Moss, *Cytomegalovirus in inflammatory bowel disease: Pathogen or innocent bystander?* Inflammatory Bowel Diseases, 2010. **16**(9): p. 1620-1627.
105. Greenberg, R., et al., *Squamous dysplasia of the rectum in a patient with ulcerative colitis treated with 6-mercaptopurine.* Dig Dis Sci, 2008. **53**(3): p. 760-4.
106. Kong, C.S., M.L. Welton, and T.A. Longacre, *Role of human papillomavirus in squamous cell metaplasia-dysplasia-carcinoma of the rectum.* Am J Surg Pathol, 2007. **31**(6): p. 919-25.
107. Perez-Brocal, V., et al., *Study of the Viral and Microbial Communities Associated With Crohn/'s Disease: A Metagenomic Approach.* Clin Trans Gastroenterol, 2013. **4**: p. e36.
108. Loh, J., et al., *Detection of novel sequences related to african Swine Fever virus in human serum and sewage.* J Virol, 2009. **83**(24): p. 13019-25.
109. Steward, G.F. and A.I. Culley, *Extraction and purification of nucleic acids from viruses.* Manual of Aquatic Viral Ecology, 2010(In S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.]): p. 154-165.
110. Angly, F.E., et al., *The marine viromes of four oceanic regions.* PLoS Biol, 2006. **4**(11): p. e368.
111. Schloss, P.D. and J. Handelsman, *Metagenomics for studying unculturable microorganisms: cutting the Gordian knot.* Genome Biol, 2005. **6**(8): p. 229.
112. Tringe, S.G. and E.M. Rubin, *Metagenomics: DNA sequencing of environmental samples.* Nat Rev Genet, 2005. **6**(11): p. 805-14.
113. Angly, F., et al., *PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information.* BMC Bioinformatics, 2005. **6**: p. 41.

114. Proctor, L.M. and J.A. Fuhrman, *Viral mortality of marine bacteria and cyanobacteria.* Nature, 1990. **343**(6253): p. 60-62.

115. Suttle, C.A., A.M. Chan, and M.T. Cottrell, *Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton.* Appl Environ Microbiol, 1991. **57**(3): p. 721-6.

116. Boom, R., et al., *Rapid and simple method for purification of nucleic acids.* J Clin Microbiol, 1990. **28**(3): p. 495-503.

117. de Paula, V.S., L.M. Villar, and A.M. Coimbra Gaspar, *Comparison of four extraction methods to detect hepatitis A virus RNA in serum and stool samples.* Braz J Infect Dis, 2003. **7**(2): p. 135-41.

118. Wommack, K.E., et al., *Filtration-based methods for the collection of viral concentrates from large water samples.* Manual of Aquatic Viral Ecology, 2010. In S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.]: p. 110-117.

119. Paul, J.H., S.C. Jiang, and J.B. Rose, *Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration.* Appl Environ Microbiol, 1991. **57**(8): p. 2197-204.

120. Urase, T., K. Yamamoto, and S. Ohgaki, *Evaluation of virus removal in membrane separation process using coliphage Q beta.* Water science and technology 1993. **28**: p. 9-15.

121. Koonin, E.V., *Virology: Gulliver among the Lilliputians.* Curr Biol, 2005. **15**(5): p. R167-9.

122. Raoult, D. and P. Forterre, *Redefining viruses: lessons from Mimivirus.* Nat Rev Microbiol, 2008. **6**(4): p. 315-9.

123. Green, J., et al., *Norwalk-like viruses: demonstration of genomic diversity by polymerase chain reaction.* J Clin Microbiol, 1993. **31**(11): p. 3007-12.

124. Hale, A.D., J. Green, and D.W. Brown, *Comparison of four RNA extraction methods for the detection of small round structured viruses in faecal specimens.* J Virol Methods, 1996. **57**(2): p. 195-201.

125. Koopmans, M., A. , Herrewegh, and M.C. Horzinek, *Diagnosis of torovirus infection.* Lancet, 1991. **337**: p. 859.

126. Lambden, P.R., et al., *Sequence and genome organization of a human small round-structured (Norwalk-like) virus.* Science, 1993. **259**(5094): p. 516-9.

127. Allander, T., et al., *A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11609-14.

128. Kapoor, A., et al., *A highly divergent picornavirus in a marine mammal.* J Virol, 2008. **82**(1): p. 311-20.

129. Svraka, S., et al., *Metagenomic sequencing for virus identification in a public-health setting.* J Gen Virol, 2010. **91**(Pt 11): p. 2846-56.

130. Victoria, J.G., A. Kapoor, and L. Li, *Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis.* J Virol, 2009. **83**: p. 4642-4651.

131. Kim, M.S., E.J. Park, and S.W. Roh, *Diversity and abundance of single-stranded DNA viruses in human feces.* Appl Environ Microbiol, 2011. **77**: p. 8062-8070.

132. Zhang, T., M. Breitbart, and W.H. Lee, *RNA viral community in human feces: prevalence of plant pathogenic viruses.* PLoS Biol, 2006. **4**: p. 0108-0118.

133. Minot, S., R. Sinha, and J. Chen, *The human gut virome: inter-individual variation and dynamic response to diet.* Genome Res, 2011. **21**: p. 1616-1625.

134. Breitbart, M., et al., *Viral diversity and dynamics in an infant gut.* Research in Microbiology, 2008. **159**(5): p. 367-373.

135. Breitbart, M., I. Hewson, and B. Felts, *Metagenomic analyses of an uncultured viral community from human feces.* J Bacteriol, 2003. **185**: p. 6220-6223.

136. Reyes, A., M. Haynes, and N. Hanson, *Viruses in the faecal microbiota of monozygotic twins and their mothers.* Nature, 2010. **466**: p. 334-338.

137. Sun, L., G.M. Nava, and T.S. Stappenbeck, *Host genetic susceptibility, dysbiosis, and viral triggers in inflammatory bowel disease.* Curr Opin Gastroenterol, 2011. **27**(4): p. 321-7.

138. Saul, S.H., *Inflammatory cloacogenic polyp: relationship to solitary rectal ulcer syndrome/mucosal prolapse and other bowel disorders.* Hum Pathol, 1987. **18**(11): p. 1120-5.

139. Penrith, M.-L., G.R. Thomson, and A.D.S. Bastos, *African swine fever*. Infectious diseases of livestock (2nd edn), ed. J.A.W. Coetzer and R. Tustin. Vol. 2. 2004: Oxford University Press, Cape Town: 1087–1119.

140. Yozwiak, N.L., et al., *Virus identification in unknown tropical febrile illness cases using deep sequencing.* PLoS Negl Trop Dis, 2012. **6**(2): p. e1485.

141. Daly, G.M., et al., *A Viral Discovery Methodology for Clinical Biopsy Samples Utilising Massively Parallel Next Generation Sequencing.* PLoS ONE, 2011. **6**(12): p. e28879.

142. Sachsenröder, J., et al., *Simultaneous Identification of DNA and RNA Viruses Present in Pig Faeces Using Process-Controlled Deep Sequencing.* PLoS ONE, 2012. **7**(4): p. e34631.

143. Bibby, K. and J. Peccia, *Identification of viral pathogen diversity in sewage sludge by metagenome analysis.* Environ Sci Technol, 2013. **47**(4): p. 1945-51.

144. Wan, X.F., et al., *Detection of African swine fever virus-like sequences in ponds in the Mississippi Delta through metagenomic sequencing.* Virus Genes, 2013. **46**(3): p. 441-6.

145. Stenglein, M., et al., *Complete genome sequence of an astrovirus identified in a domestic rabbit (Oryctolagus cuniculus) with gastroenteritis.* Virology Journal, 2012. **9**(1): p. 216.

146. Thurber, R.V., et al., *Laboratory procedures to generate viral metagenomes.* Nat. Protocols, 2009. **4**(4): p. 470-483.

147. Daly, G.M., et al., *A Viral Discovery Methodology for Clinical Biopsy Samples Utilising Massively Parallel Next Generation Sequencing.* Plos One, 2011. **6**(12).

148. Vogelstein, B. and D. Gillespie, *Preparative and analytical purification of DNA from agarose.* Proc Natl Acad Sci U S A, 1979. **76**(2): p. 615-9.

149. Breitbart, M., et al., *Metagenomic analyses of an uncultured viral community from human feces.* Journal of Bacteriology, 2003. **185**(20): p. 6220-6223.

150. Djikeng, A., et al., *Viral genome sequencing by random priming methods.* Bmc Genomics, 2008. **9**.

151. Kapoor, A., et al., *A highly divergent picornavirus in a marine mammal.* Journal of Virology, 2008. **82**(1): p. 311-320.

152. Sambrook, J. and D.W. Russell, *Molecular cloning: A laboratory manual* 3rd ed. 2001: Cold Spring Harbor Laboratory Press.

153. Hall, R.J., et al., *Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery.* J Virol Methods, 2014. **195**: p. 194-204.

154. Weimbs, T., et al., *A conserved domain is present in different families of vesicular fusion proteins: a new superfamily.* Proc Natl Acad Sci U S A, 1997. **94**(7): p. 3046-51.

155. Sitbon, E. and S. Pietrokovski, *Occurrence of protein structure elements in conserved sequence regions.* BMC Struct Biol, 2007. **7**: p. 3.

156. Marchler-Bauer, A., et al., *CDD: a Conserved Domain Database for the functional annotation of proteins.* Nucleic Acids Research, 2011. **39**: p. D225-D229.

157. Punta, M., et al., *The Pfam protein families database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D290-301.

158. Letunic, I., T. Doerks, and P. Bork, *SMART 7: recent updates to the protein domain annotation resource.* Nucleic Acids Res, 2012. **40**(Database issue): p. D302-5.

159. Haft, D.H., J.D. Selengut, and O. White, *The TIGRFAMs database of protein families.* Nucleic Acids Research, 2003. **31**(1): p. 371-373.

160. Breitbart, M., et al., *Metagenomic analyses of an uncultured viral community from human feces.* J Bacteriol, 2003. **185**(20): p. 6220-3.

161. Kim, K.H. and J.W. Bae, *Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses.* Appl Environ Microbiol, 2011. **77**(21): p. 7663-8.

162. Yilmaz, P., et al., *Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.* Nat Biotechnol, 2011. **29**(5): p. 415 - 420.

Appendices

1.1. The patients' information and code

Sample	Sample Name	Lab. ID	Code	DNA/RNA	Viral enrichment method
RV630 CD Stool	CD1	CDS01	CD1A	RNA	Tangential flow filtration (June 15, 2010)
		CDS01D	CD1B	DNA	Tangential flow filtration (June 15, 2010)
		CDS02	CD1C	RNA	Tangential flow filtration (April 30, 2012)
		CDS02D	CD1D	DNA	Tangential flow filtration (April 30, 2012)
		CDS03	CD1E	RNA	Silicon di-oxide
		CD03D	CD1F	DNA	Silicon di-oxide
		CDS07	CD1G	RNA	Ultrafree MC (microfiltration)
		CDS07D	CD1H	DNA	Ultrafree MC (microfiltration)
RV37 CD Stool	CD2	CDS04	CD2A	RNA	Tangential flow filtration
		CDS04D	CD2B	DNA	Tangential flow filtration
		CDS05	CD2C	RNA	Silicon di-oxide
		CDS05D	CD2D	DNA	Silicon di-oxide
		CDS06	CD2E	RNA	Ultrafree MC (microfiltration)
		CDS06D	CD2F	DNA	Ultrafree MC (microfiltration)
S.R. CD Stool sample	CD3	CDS10	CD3A	RNA	Tangential flow filtration
		CDS10D	CD3B	DNA	Tangential flow filtration
		CDS11	CD3C	RNA	Ultrafree MC (microfiltration)
		CDS11D	CD3D	DNA	Ultrafree MC (microfiltration)
		CDS12	CD3E	RNA	Silicon di-oxide
		CDS12D	CD3F	DNA	Silicon di-oxide
RV 616 CD stool	CD4	CDS08	CD4A	RNA	Ultrafree MC (microfiltration)
		CDS08D	CD4B	DNA	Ultrafree MC (microfiltration)
RV 624 CD stool	CD5	CDS09	CD5A	RNA	Ultrafree MC (microfiltration)
		CDS09D	CD5B	DNA	Ultrafree MC (microfiltration)
UC 12 stool	UC1	UCS06	UC1A	RNA	Ultrafree MC (microfiltration)
		UCS06D	UC1B	DNA	Ultrafree MC (microfiltration)

Sample	**Sample code**	**ID**	**Code**	**DNA/RNA**	**viral prep**
		UCS07	UC1C	RNA	Tangential flow filtration
		UCS07D	UC1D	DNA	Tangential flow filtration
		UCS08	UC1E	RNA	Silicon di-oxide
		UCS08D	UC1F	DNA	Silicon di-oxide
RV575 UC stool	UC2	UCS01	UC2A	RNA	Tangential flow filtration
		UCS01D	UC2B	DNA	Tangential flow filtration
		UCS04	UC2C	RNA	Ultrafree MC (microfiltration)
		UCS04D	UC2D	DNA	Ultrafree MC (microfiltration)
RV712 UC Stool	UC3	UCS02	UC3A	RNA	Tangential flow filtration
		UCS02D	UC3B	DNA	Tangential flow filtration
		UCS03	UC3C	RNA	Silicon di-oxide
		UCS03D	UC3D	DNA	Silicon di-oxide
RV641 UC stool	UC4	UCS05	UC4A	RNA	Ultrafree MC (microfiltration)
		UCS05D	UC4B	DNA	Ultrafree MC (microfiltration)
RV914 PSC stool	PSC	PSCS01	PSC A	RNA	Tangential flow filtration
		PSCS01D	PSC B	DNA	Tangential flow filtration
		PSCS02	PSC C	RNA	Silicon di-oxide
		PSCS02D	PSC D	DNA	Silicon di-oxide
		PSCS03	PSC E	RNA	Ultrafree MC (microfiltration)
		PSCS03D	PSC F	DNA	Ultrafree MC (microfiltration)
RV543 control stool	CON1	CONS01	CON1A	RNA	Tangential flow filtration
		CONS01D	CON1B	DNA	Tangential flow filtration
RV620 control stool	CON2	CONS02	CON2A	RNA	Ultrafree MC (microfiltration)
		CONS02D	CON2B	DNA	Ultrafree MC (microfiltration)

1.2. Colon biopsies of UC and CD patients for prevalence of ASFLV study

ID	**Code**
CD06	CDC06
CD10	CDC10
CD12	CDC12
CD19	CDC19
UC01	UCC01
UC04	UCC04
UC06	UCC06
UC07	UCC07
UC13	UCC13

2. Distribution of total viral read counts similar to the sequences of NCBI virus database and list of corresponding virus detected across the metagenomics libraries

Viruses	CD1A	CD1B	CD1C	CD1D	CD1E	CD1F	CD1G	CD1H	CD2A	CD2B
Autographa californica nucleopolyhedrovirus	128	0	6	0	126	0	0	0	0	0
Pepper mild mottle virus	514318	0	172184	34	11668	17	0	0	140	20
Paprika mild mottle virus	48430	0	19599	2	144	4	0	0	0	2
Paramecium bursaria chlorella virus	27	1	0	0	6	7	0	0	0	1
Tobacco mild green mosaic virus	6810	0	1558	0	102	0	0	0	0	0
Cucumber green mottle mosaic virus	2130		616	0	198	0	0	0	0	0
Tomato mosaic virus	2015		335	0	0	0	0	0	2	0
Bell pepper mottle virus	0	0	294	0	0	0	0	0	0	0
Plutella xylostella multiple nucleopolyhedrovirus	128	0	5	0	110	0	0	0	0	0
Cafeteria roenbergensis virus	0	0	0	0	0	2	0	6	0	0
Emiliania huxleyi virus	0	1	0	0	0	0	0	0	0	0
Tobacco mosaic virus	109	0	92	0	0	0	0	0	0	0
Herpesvirus	0	20	1	7	22	138	12	0	1	10
Poxvirus	0	0	0	0	5	0	0	0	0	0
Wiseana iridescent virus	0	0	0	0	2	0	0	0	0	0
human papillomavirus	0	0	0	0	0	0	0	0	0	0
Torque teno virus	0	0	0	0	0	0	0	0	0	0
Lumpy skin disease virus	0	1	0	0	0	5	0	0	0	0

Viruses	CD2C	CD2D	CD2E	CD2F	CD3A	CD3B	CD3C	CD3D	CD3E	CD3F
Autographa californica nucleopolyhedrovirus	8	0	44	0	0	0	2	0	0	0
Pepper mild mottle virus	32	18	0	0	0	0	0	0	30	0
Paprika mild mottle virus	2	0	0	0	0	0	0	0	2	0
Paramecium bursaria chlorella virus	0	10	0	0	0	0	0	0	0	0
Tobacco mild green mosaic virus	0	0	0	0	0	0	0	0	0	0
Cucumber green mottle mosaic virus	0	0	0	0	0	0	0	0	0	0
Tomato mosaic virus	0	0	0	0	7	0	8	0	0	0
Bell pepper mottle virus	0	0	0	0	0	0	0	0	2	0
Plutella xylostella multiple nucleopolyhedrovirus	8	0	44	0	0	0	0	0	0	0
Cafeteria roenbergensis virus	1	0	0	0	0	0	0	0	0	0
Emiliania huxleyi virus	0	0	0	0	0	0	0	0	0	0
Tobacco mosaic virus	0	0	0	0	0	0	0	0	10	0
Herpesvirus	0	570	0	0	0	1	0	0	2	0
Poxvirus	0	1	0	0	0	0	0	0	0	0
Wiseana iridescent virus	12	0	0	0	0	0	0	0	0	0
human papillomavirus	0	0	0	0	0	0	0	0	0	0
Torque teno virus	0	0	0	0	0	0	0	0	0	0
Lumpy skin disease virus	0	0	0	0	0	0	0	0	0	0

Viruses	CD4A	CD4B	CD5A	CD5B	UC1A	UC1B	UC1C	UC1D	UC1E	UC1F
Autographa californica nucleopolyhedrovirus	0	0	4	0	0	0	10	0	2	
Pepper mild mottle virus	0	0	0	0	0	0	0	0	0	0
Paprika mild mottle virus	0	0	0	0	0	0	0	0	0	0
Paramecium bursaria chlorella virus	0	0	0	1990	0	0	0	26	0	26
Tobacco mild green mosaic virus	0	0	0	0	0	0	0	0	0	0
Cucumber green mottle mosaic virus	0	0	0	0	0	0	0	0	0	0
Tomato mosaic virus	0	0	0	0	0	0	0	0	0	0
Bell pepper mottle virus	0	0	0	0	0	0	0	0	0	0
Plutella xylostella multiple nucleopolyhedrovirus	0	0	0	0	0	0	10	0	0	0
Cafeteria roenbergensis virus	0	6	0	0	0	0	2	8	2	13
Emiliania huxleyi virus	0	0	0	0	0	0	0	22	0	4
Tobacco mosaic virus	0	0	0	0	0	0	0	0	0	0
Herpesvirus	0	0	1	1	2	0	1	7	8	3
Poxvirus	0	0	0	0	0	0	32	0	0	34
Wiseana iridescent virus	0	0	0	0	27	0	0	4	5	2
human papillomavirus	0	0	0	0	0	0	0	0	0	0
Torque teno virus	0	0	0	0	0	0	0	0	0	0
Lumpy skin disease virus	0	0	0	0	0	0	0	2	0	0

Viruses	UC2A	UC2B	UC2C	UC2D	UC3A	UC3B	UC3C	UC3D	UC4A	UC4B
Autographa californica nucleopolyhedrovirus	0	0	0	0	0	0	1	0	0	0
Pepper mild mottle virus	8	0	0	0	130	9	95	30	0	0
Paprika mild mottle virus	0	0	0	0	0	0	4	7	0	0
Paramecium bursaria chlorella virus	0	0	0	0	0	11	0	2	0	10
Tobacco mild green mosaic virus	0	0	0	0	77	0	2	0	0	0
Cucumber green mottle mosaic virus	0	0	0	0	0	0	0	0	0	0
Tomato mosaic virus	0	0	0	0	1	0	0	0	0	0
Bell pepper mottle virus	0	0	0	0	13	0	4	0	0	0
Plutella xylostella multiple nucleopolyhedrovirus	0	0	0	0	0	0	0	0	0	0
Cafeteria roenbergensis virus	0	0	0	0	15	0	0	0	0	0
Emiliania huxleyi virus	1	0	0	0	0	0	0	0	0	0
Tobacco mosaic virus	0	0	0	0	6	0	20	0	0	0
Herpesvirus	667	0	0	0	0	5	0	2	0	0
Poxvirus	0	0	0	0	0	0	0	0	0	0
Wiseana iridescent virus	0	0	0	0	0	9	0	0	0	0
human papillomavirus	0	0	0	0	0	0	0	0	0	0
Torque teno virus	0	0	0	0	0	0	0	0	0	0
Lumpy skin disease virus	0	0	0	0	0	0	0	0	0	0

Viruses	PSC A	PSC B	PSC C	PSC D	PSC E	PSC F	CON1A	CON1B	CON2A	CON2B
Autographa californica nucleopolyhedrovirus	28	0	4	0	2	0	0	0	0	0
Pepper mild mottle virus	44	0	0	0	0	0	0	0	0	0
Paprika mild mottle virus	0	0	0	0	0	0	0	0	0	0
Paramecium bursaria chlorella virus	57	53	0	22	0	22	0	2	8	0
Tobacco mild green mosaic virus	0	0	0	0	0	0	0	0	0	0
Cucumber green mottle mosaic virus	0	0	0	0	4	0	0	0	0	0
Tomato mosaic virus	44	0	0	0	0	0	0	0	0	0
Bell pepper mottle virus	0	0	0	0	0	0	0	0	0	0
Plutella xylostella multiple nucleopolyhedrovirus	26	0	0	0	0	0	0	0	0	0
Cafeteria roenbergensis virus	10	10	8	14	42	20	0	3	0	0
Emiliania huxleyi virus	2	2	0	5	0	5	0	1	0	0
Tobacco mosaic virus	0	0	0	0	0	0	0	0	0	0
Herpesvirus	14	14	7	0	1	9	95	343	0	0
Poxvirus	12	12	0	4	0	16	3	0	6	0
Wiseana iridescent virus	2	2	4	12	0	4	0	0	0	8
human papillomavirus	4	4	8	6	0	15	0	0	0	0
Torque teno virus	6	5	10	0	0	0	0	0	0	0
Lumpy skin disease virus	0	0	0	2	0	0	0	0	0	0

3.1. Alignment of contigs/scaffolds similar to herpesvirus sequences with sequences of NCBI database

3.1.1. BLAST alignment for contig C2872238 (116 bp)

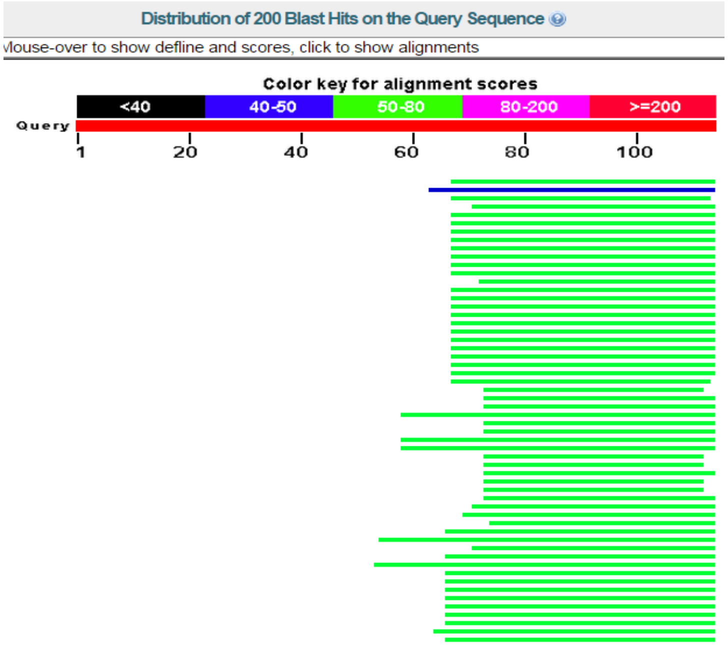
Nucleotide sequence in FASTA format

```
>C2872238
ATGTCCTTTAGACACGGGTCAGTACACTTACTAGGTGTTCCTGTTCCTCCGGTTCCTC
CTCCTCCCCCAGTGTCAGGATTAGGATTAGGACTAGGGTTAGGGCTAGGGTTCGGG
TT
```

BLASTn alignment for the contig C2872238

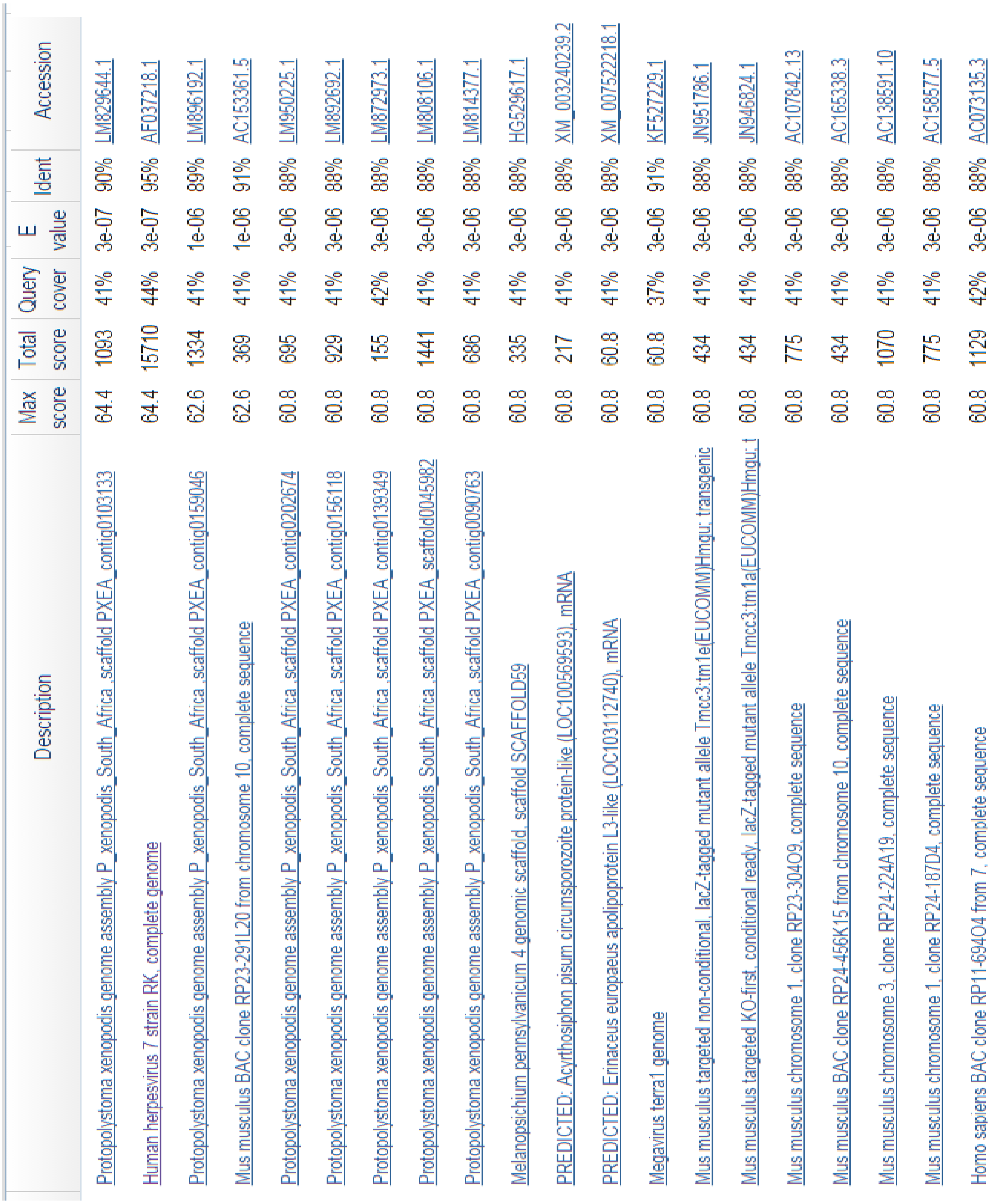
BLASTn search result showed that the assembled query sequence was similar to a species belonging to the phylum Platyhelminthes along with repeat regions of the herpesvirus 7. There was about 44% of total coverage (40 bp match) when aligned with the sequence of herpesvirus 7 with E-value of 3e-07 and the identity of 95% for the matched region. The first sixty base pair fragment of this contig was taken for a separate BLASTn search. The query sequence did not match with any sequence in the database. The BLASTx search of that same fragment showed similarity to the sequence of roundworm and different types of birds with insignificant E-value (1.1-3.7)

A) Graphic display



Distribution of 200 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
20
40
60
80
100

B) Hit list of aligned NCBI sequences for the contig C2872238 after BLASTn search



Description	Max score	Total score	Query cover	E value	Ident	Accession
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0103133	64.4	1093	41%	3e-07	90%	LM829644.1
Human herpesvirus 7 strain RK, complete genome	64.4	15710	44%	3e-07	95%	AF037218.1
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0159046	62.6	1334	41%	1e-06	89%	LM896192.1
Mus musculus BAC clone RP23-291L20 from chromosome 10, complete sequence	62.6	369	41%	1e-06	91%	AC153361.5
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0202674	60.8	695	41%	3e-06	88%	LM950225.1
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0156118	60.8	929	41%	3e-06	88%	LM892692.1
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0139349	60.8	155	42%	3e-06	88%	LM872973.1
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_scaffold0045982	60.8	1441	41%	3e-06	88%	LM808106.1
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0090763	60.8	686	41%	3e-06	88%	LM814377.1
Melanopsichium pennsylvanicum 4 genomic scaffold, scaffold SCAFFOLD59	60.8	335	41%	3e-06	88%	HG529617.1
PREDICTED: Acyrthosiphon pisum circumsporozoite protein-like (LOC100569593), mRNA	60.8	217	41%	3e-06	88%	XM_003240239.2
PREDICTED: Erinaceus europaeus apolipoprotein L3-like (LOC103112740), mRNA	60.8	60.8	41%	3e-06	88%	XM_007522218.1
Megavirus terra1 genome	60.8	60.8	37%	3e-06	91%	KF527229.1
Mus musculus targeted non-conditional, lacZ-tagged mutant allele Tmcc3:tm1e(EUCOMM)Hmgu; transgenic	60.8	434	41%	3e-06	88%	JN951786.1
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Tmcc3:tm1a(EUCOMM)Hmgu; t	60.8	434	41%	3e-06	88%	JN946824.1
Mus musculus chromosome 1, clone RP23-304O9, complete sequence	60.8	775	41%	3e-06	88%	AC107842.13
Mus musculus BAC clone RP24-456K15 from chromosome 10, complete sequence	60.8	434	41%	3e-06	88%	AC165338.3
Mus musculus chromosome 3, clone RP24-224A19, complete sequence	60.8	1070	41%	3e-06	88%	AC138591.10
Mus musculus chromosome 1, clone RP24-187D4, complete sequence	60.8	775	41%	3e-06	88%	AC158577.5
Homo sapiens BAC clone RP11-694O4 from 7, complete sequence	60.8	1129	42%	3e-06	88%	AC073135.3

BLASTx alignment for the contig C2872238

After BLASTx search, query sequence had been found to have no significant similarity with nucleotide sequences in the NCBI database.

3.1.2. BLAST alignment for the contig C7151139 (284 bp)

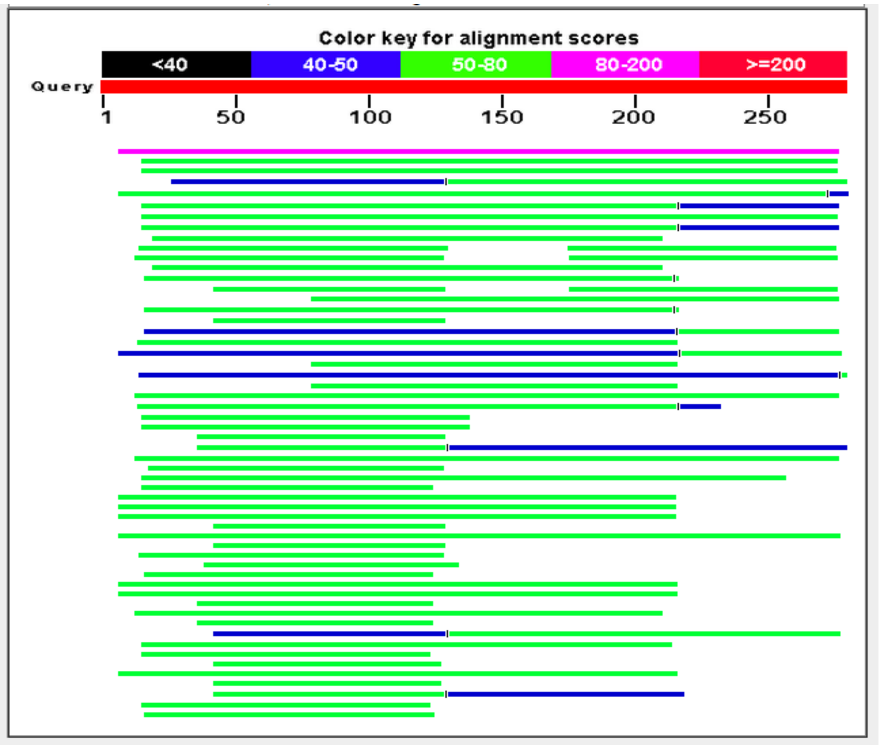
Nucleotide sequence in FASTA format

```
>C7151139
TTGCCACTGCATTCACGTGTGTGTTTGAGCATGCACATGTTCTTGTATGTATGTGTGT
GTGTGCTTGCCACTGCATGCACGTGTGTGTTTGTGCATGTGCGTGTGAACGTATGCA
TGGGTATGTGTGCTTGCCACCGCATGCACCCGCGTATTTGTGCATGTCCCTGCATAC
TTTGTTACATACATGTGTTCAGCTGCATGTGCATGCATGTGTCTGTGTACCACCACA
TGCATGTGCGATTTTGCGCATGTAGGTGCATGGATTTGCTTGTGTGTGCATGGAT
```

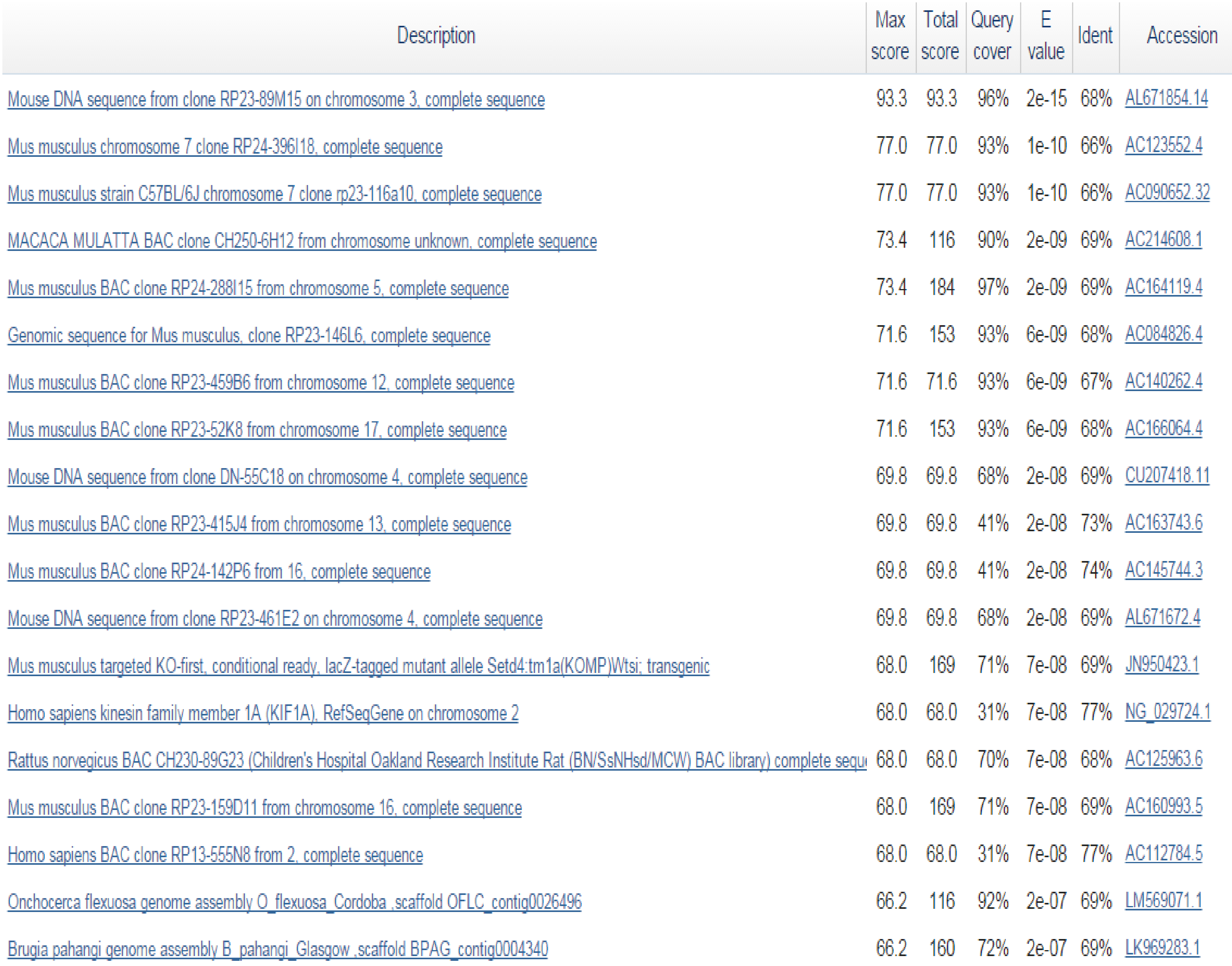
BLASTn alignment for the contig C7151139

BLASTn search of the contig result showed that the sequence was similar to the mouse DNA sequence with E-value of 2e-15. The alignment showed the total coverage of 96% (279 bp match) with the identity of 68% within the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C7151139 after BLASTn search

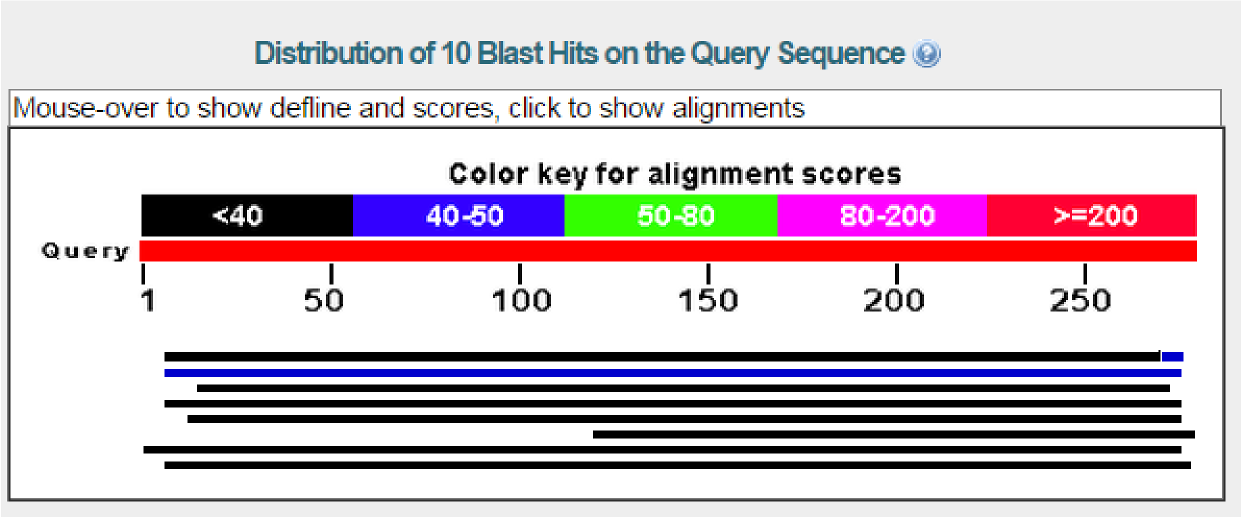


Description	Max score	Total score	Query cover	E value	Ident	Accession
Mouse DNA sequence from clone RP23-89M15 on chromosome 3, complete sequence	93.3	93.3	96%	2e-15	68%	AL671854.14
Mus musculus chromosome 7 clone RP24-396I18, complete sequence	77.0	77.0	93%	1e-10	66%	AC123552.4
Mus musculus strain C57BL/6J chromosome 7 clone rp23-116a10, complete sequence	77.0	77.0	93%	1e-10	66%	AC090652.32
MACACA MULATTA BAC clone CH250-6H12 from chromosome unknown, complete sequence	73.4	116	90%	2e-09	69%	AC214608.1
Mus musculus BAC clone RP24-288I15 from chromosome 5, complete sequence	73.4	184	97%	2e-09	69%	AC164119.4
Genomic sequence for Mus musculus, clone RP23-146L6, complete sequence	71.6	153	93%	6e-09	68%	AC084826.4
Mus musculus BAC clone RP23-459B6 from chromosome 12, complete sequence	71.6	71.6	93%	6e-09	67%	AC140262.4
Mus musculus BAC clone RP23-52K8 from chromosome 17, complete sequence	71.6	153	93%	6e-09	68%	AC166064.4
Mouse DNA sequence from clone DN-55C18 on chromosome 4, complete sequence	69.8	69.8	68%	2e-08	69%	CU207418.11
Mus musculus BAC clone RP23-415J4 from chromosome 13, complete sequence	69.8	69.8	41%	2e-08	73%	AC163743.6
Mus musculus BAC clone RP24-142P6 from 16, complete sequence	69.8	69.8	41%	2e-08	74%	AC145744.3
Mouse DNA sequence from clone RP23-461E2 on chromosome 4, complete sequence	69.8	69.8	68%	2e-08	69%	AL671672.4
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Setd4:tm1a(KOMP)Wtsi; transgenic	68.0	169	71%	7e-08	69%	JN950423.1
Homo sapiens kinesin family member 1A (KIF1A), RefSeqGene on chromosome 2	68.0	68.0	31%	7e-08	77%	NG_029724.1
Rattus norvegicus BAC CH230-89G23 (Children's Hospital Oakland Research Institute Rat (BN/SsNHsd/MCW) BAC library) complete sequ	68.0	68.0	70%	7e-08	68%	AC125963.6
Mus musculus BAC clone RP23-159D11 from chromosome 16, complete sequence	68.0	169	71%	7e-08	69%	AC160993.5
Homo sapiens BAC clone RP13-555N8 from 2, complete sequence	68.0	68.0	31%	7e-08	77%	AC112784.5
Onchocerca flexuosa genome assembly O_flexuosa_Cordoba ,scaffold OFLC_contig0026496	66.2	116	92%	2e-07	69%	LM569071.1
Brugia pahangi genome assembly B_pahangi_Glasgow ,scaffold BPAG_contig0004340	66.2	160	72%	2e-07	69%	LK969283.1

BLASTx alignment for the contig C7151139

After BLASTx search, query sequence had been found to align with the sequences from different birds with higher E value of 0.011 suggesting the alignment to be insignificant.

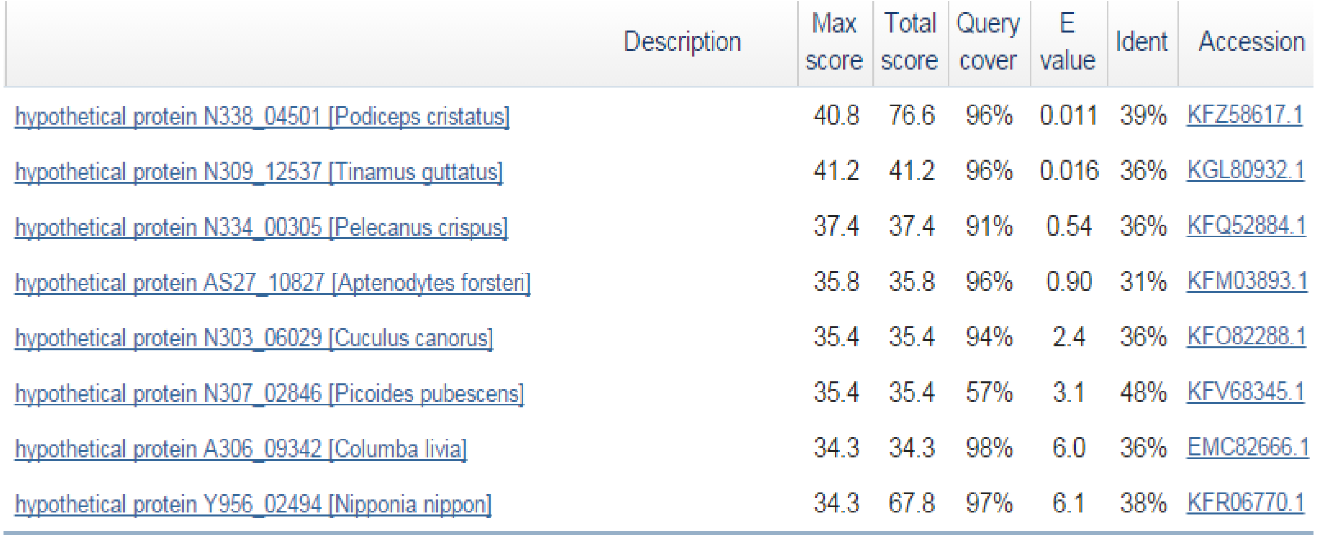
A) Graphic display



Distribution of 10 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C7151139 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
hypothetical protein N338_04501 [Podiceps cristatus]	40.8	76.6	96%	0.011	39%	KFZ58617.1
hypothetical protein N309_12537 [Tinamus guttatus]	41.2	41.2	96%	0.016	36%	KGL80932.1
hypothetical protein N334_00305 [Pelecanus crispus]	37.4	37.4	91%	0.54	36%	KFQ52884.1
hypothetical protein AS27_10827 [Aptenodytes forsteri]	35.8	35.8	96%	0.90	31%	KFM03893.1
hypothetical protein N303_06029 [Cuculus canorus]	35.4	35.4	94%	2.4	36%	KFO82288.1
hypothetical protein N307_02846 [Picoides pubescens]	35.4	35.4	57%	3.1	48%	KFV68345.1
hypothetical protein A306_09342 [Columba livia]	34.3	34.3	98%	6.0	36%	EMC82666.1
hypothetical protein Y956_02494 [Nipponia nippon]	34.3	67.8	97%	6.1	38%	KFR06770.1

3.1.3. BLAST alignment for the contig C5219408 (192 BP)

Nucleotide sequence in FASTA format

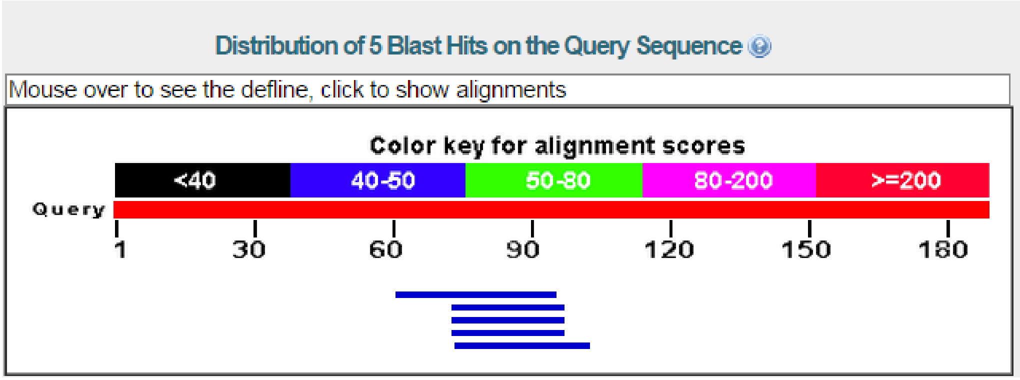
>C5219408
CCCGTAACCAGTGTTGGCTGCTGTGGATGGACACTCTCGATCAGGCTGGCAAAGCG
AAGGGACGAGCGGAAGGCCGAGCCACTGGTGCAACTGGTGCAAGCGTGGGCCTCTT
TGTTGGACTCTTTGTTGGTCTCCTTGTTGGTCTTCTTGTTGGACGCTTTGTTGGACGC
TTTGTTGGTCTCCTTGTTGGTC

BLASTn alignment for the contig C5219408

BLASTn search result showed that assembled query sequence was similar to the sequence of *Agrobacterium*, the alignment was not significant as the E value was 0.14 (higher than 1e-5) with only 18% of total coverage (35 bp match). The identity was found to be 89% within the matched region.

The separate BLASTn for the unmatched first sixty base pairs showed no alignment with any of the reference sequences in the NCBI database and BLASTx result showed query sequence being similar to the sequence of *Fasciola hepatica* (phylum Platyhelminthes) with higher E-value of 3.4 (insignificant). Whereas the sequence fragment (91-192 base pairs position) of contig C5219408 was similar to the sequence of Armadillo species with significant E-value of 1e-05.

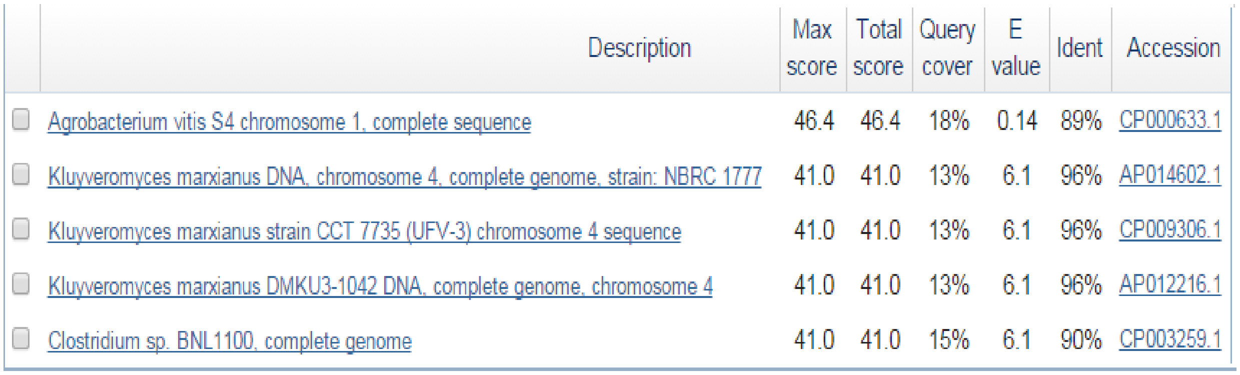
A) Graphic display



Distribution of 5 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C5219408 after BLASTn search



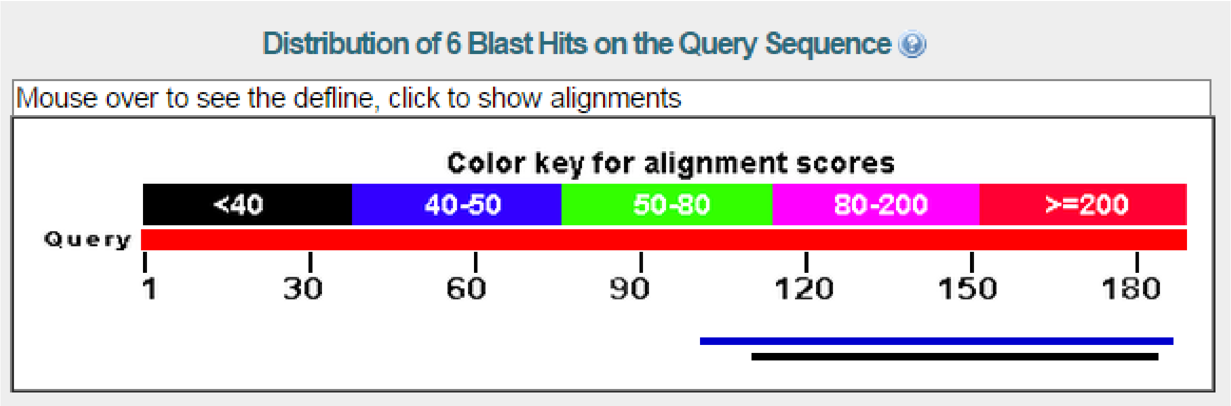
	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Agrobacterium vitis S4 chromosome 1, complete sequence	46.4	46.4	18%	0.14	89%	CP000633.1
□	Kluyveromyces marxianus DNA, chromosome 4, complete genome, strain: NBRC 1777	41.0	41.0	13%	6.1	96%	AP014602.1
□	Kluyveromyces marxianus strain CCT 7735 (UFV-3) chromosome 4 sequence	41.0	41.0	13%	6.1	96%	CP009306.1
□	Kluyveromyces marxianus DMKU3-1042 DNA, complete genome, chromosome 4	41.0	41.0	13%	6.1	96%	AP012216.1
□	Clostridium sp. BNL1100, complete genome	41.0	41.0	15%	6.1	90%	CP003259.1

BLASTx alignment for the contig C5219408

After BLASTx search, query sequence aligned with a mammalian species (*Armadillo*) and with herpesvirus 6. The E value was 2.0 for the herpesvirus sequence alignment which means that this is not a significant hit. Sequence identity with HSV-6 was 64 % with matched length of 16 amino acid out of total 25 amino acids, and query coverage was 39%.

This contig could belong to an unknown species as after BLASTn and BLASTx search the contig showed only partial alignment and less similarity to the sequences of NCBI database

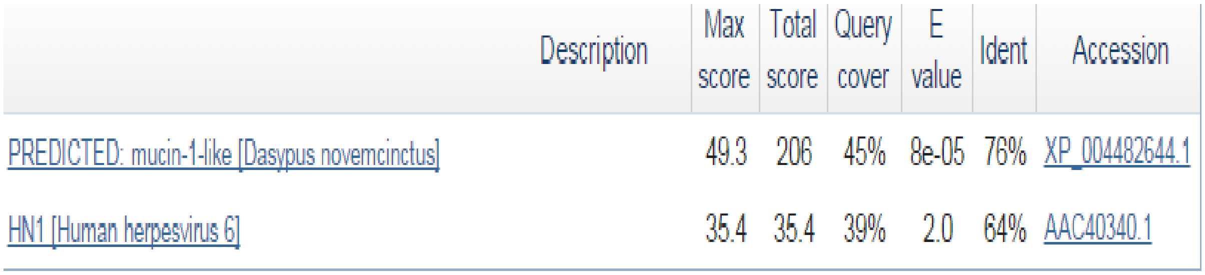
A) Graphic display



Distribution of 6 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C5219408 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: mucin-1-like [Dasypus novemcinctus]	49.3	206	45%	8e-05	76%	XP_004482644.1
HN1 [Human herpesvirus 6]	35.4	35.4	39%	2.0	64%	AAC40340.1

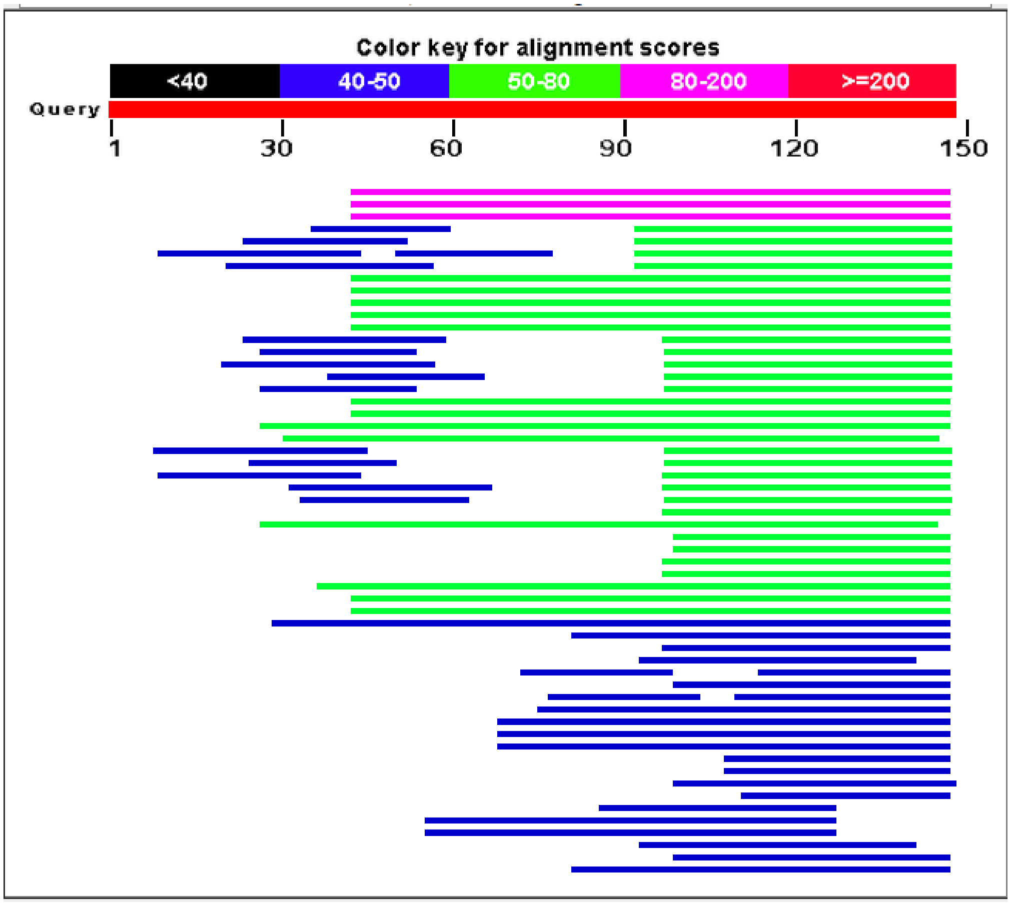
3.1.4. BLAST alignment for the contig C3733744 (150 BP)

>C3733744
TTGTATTCCAAGTTTTACCCCAAGAAATGAAGATAAATTTGTTGGTGCAACAGTTTT
TGAACCAAAAGTAGGGATTTATTTAACACCTGTTACAGTTGTTGATTTCAAAAGTTT
GTATCCAAGCATTATGAGAGCACATAATTTGTGTTT

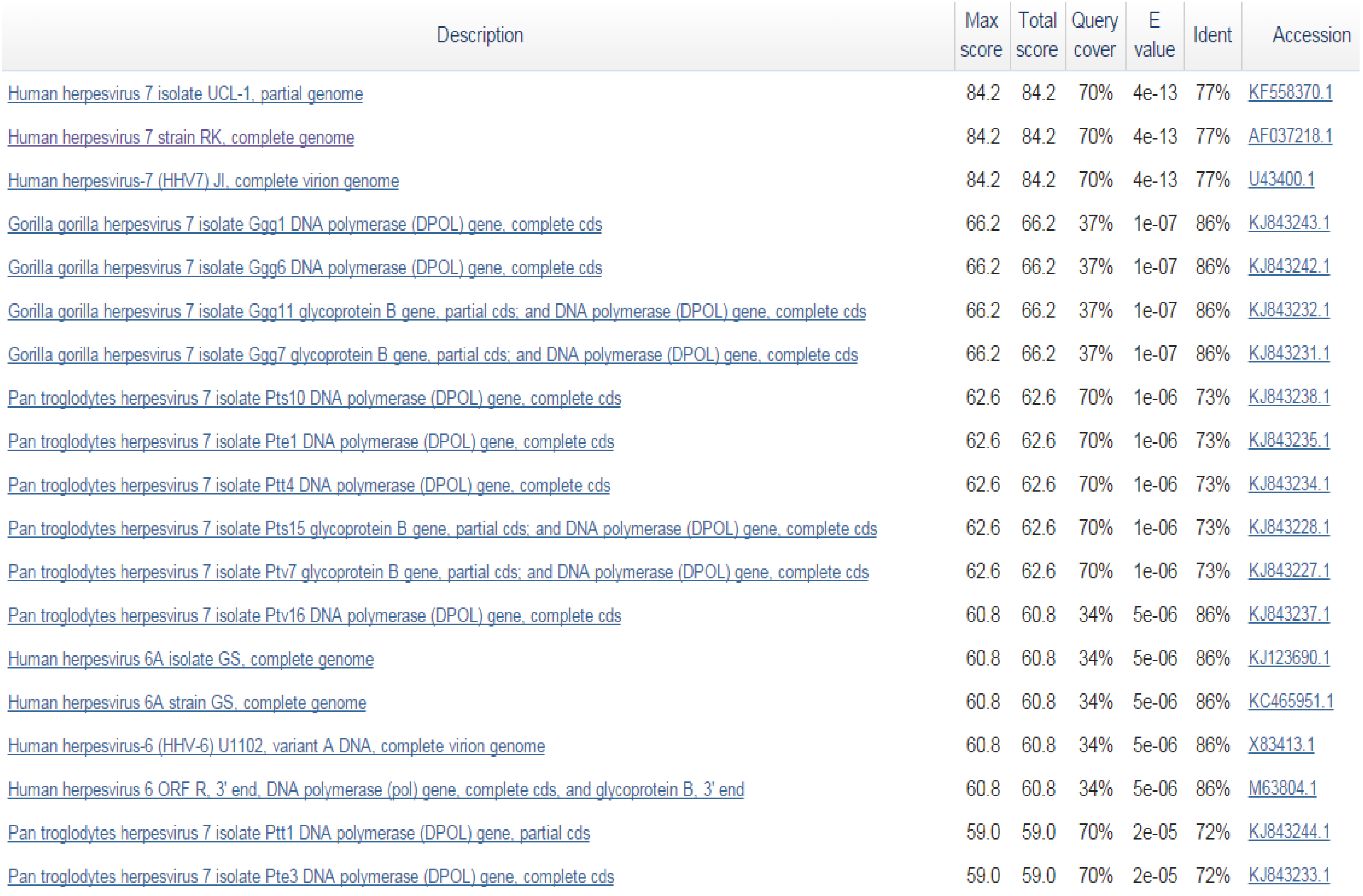
BLASTn alignment for the contig C3733744

BLASTn search result showed query sequence was similar to the sequence of herpesvirus 7 DNA polymerase gene. For top hit (HSV-7) the match length was 70% of total coverage (106 bp match), E value was 4e-13 and the identity of 77% was found with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C3733744 after BLASTn search

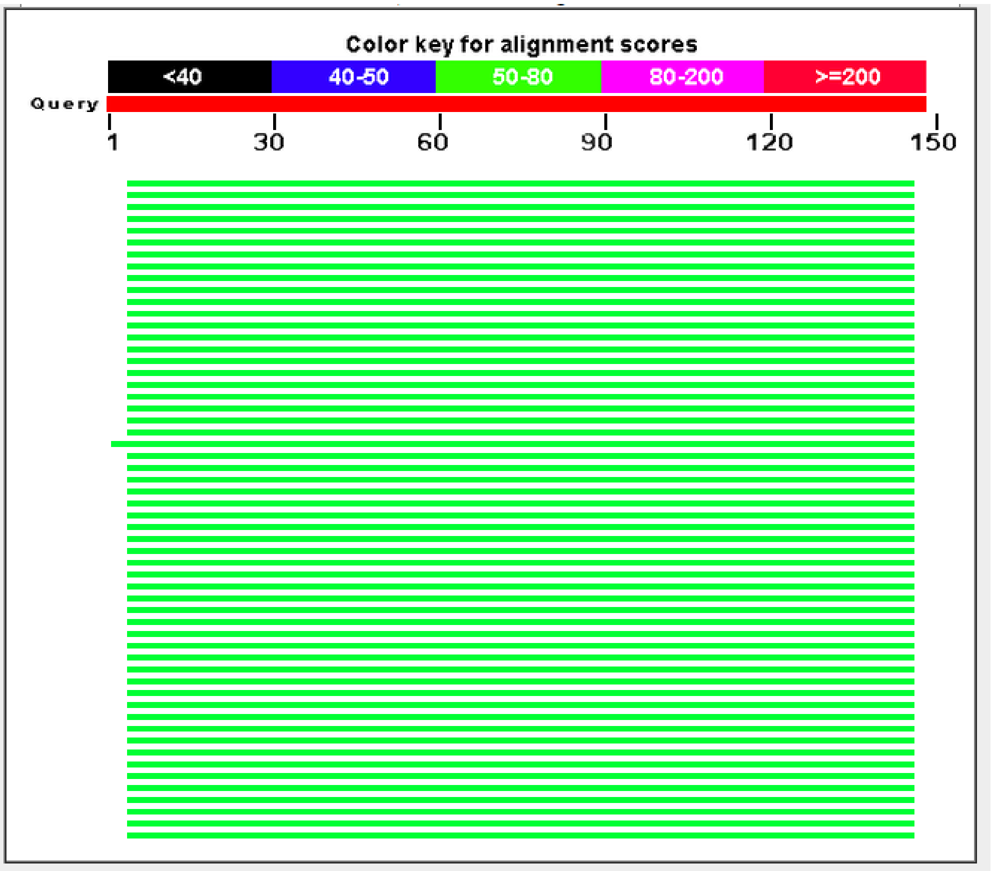


Description	Max score	Total score	Query cover	E value	Ident	Accession
Human herpesvirus 7 isolate UCL-1, partial genome	84.2	84.2	70%	4e-13	77%	KF558370.1
Human herpesvirus 7 strain RK, complete genome	84.2	84.2	70%	4e-13	77%	AF037218.1
Human herpesvirus-7 (HHV7) JI, complete virion genome	84.2	84.2	70%	4e-13	77%	U43400.1
Gorilla gorilla herpesvirus 7 isolate Ggg1 DNA polymerase (DPOL) gene, complete cds	66.2	66.2	37%	1e-07	86%	KJ843243.1
Gorilla gorilla herpesvirus 7 isolate Ggg6 DNA polymerase (DPOL) gene, complete cds	66.2	66.2	37%	1e-07	86%	KJ843242.1
Gorilla gorilla herpesvirus 7 isolate Ggg11 glycoprotein B gene, partial cds; and DNA polymerase (DPOL) gene, complete cds	66.2	66.2	37%	1e-07	86%	KJ843232.1
Gorilla gorilla herpesvirus 7 isolate Ggg7 glycoprotein B gene, partial cds; and DNA polymerase (DPOL) gene, complete cds	66.2	66.2	37%	1e-07	86%	KJ843231.1
Pan troglodytes herpesvirus 7 isolate Pts10 DNA polymerase (DPOL) gene, complete cds	62.6	62.6	70%	1e-06	73%	KJ843238.1
Pan troglodytes herpesvirus 7 isolate Pte1 DNA polymerase (DPOL) gene, complete cds	62.6	62.6	70%	1e-06	73%	KJ843235.1
Pan troglodytes herpesvirus 7 isolate Ptt4 DNA polymerase (DPOL) gene, complete cds	62.6	62.6	70%	1e-06	73%	KJ843234.1
Pan troglodytes herpesvirus 7 isolate Pts15 glycoprotein B gene, partial cds; and DNA polymerase (DPOL) gene, complete cds	62.6	62.6	70%	1e-06	73%	KJ843228.1
Pan troglodytes herpesvirus 7 isolate Ptv7 glycoprotein B gene, partial cds; and DNA polymerase (DPOL) gene, complete cds	62.6	62.6	70%	1e-06	73%	KJ843227.1
Pan troglodytes herpesvirus 7 isolate Ptv16 DNA polymerase (DPOL) gene, complete cds	60.8	60.8	34%	5e-06	86%	KJ843237.1
Human herpesvirus 6A isolate GS, complete genome	60.8	60.8	34%	5e-06	86%	KJ123690.1
Human herpesvirus 6A strain GS, complete genome	60.8	60.8	34%	5e-06	86%	KC465951.1
Human herpesvirus-6 (HHV-6) U1102, variant A DNA, complete virion genome	60.8	60.8	34%	5e-06	86%	X83413.1
Human herpesvirus 6 ORF R, 3' end, DNA polymerase (pol) gene, complete cds, and glycoprotein B, 3' end	60.8	60.8	34%	5e-06	86%	M63804.1
Pan troglodytes herpesvirus 7 isolate Ptt1 DNA polymerase (DPOL) gene, partial cds	59.0	59.0	70%	2e-05	72%	KJ843244.1
Pan troglodytes herpesvirus 7 isolate Pte3 DNA polymerase (DPOL) gene, complete cds	59.0	59.0	70%	2e-05	72%	KJ843233.1

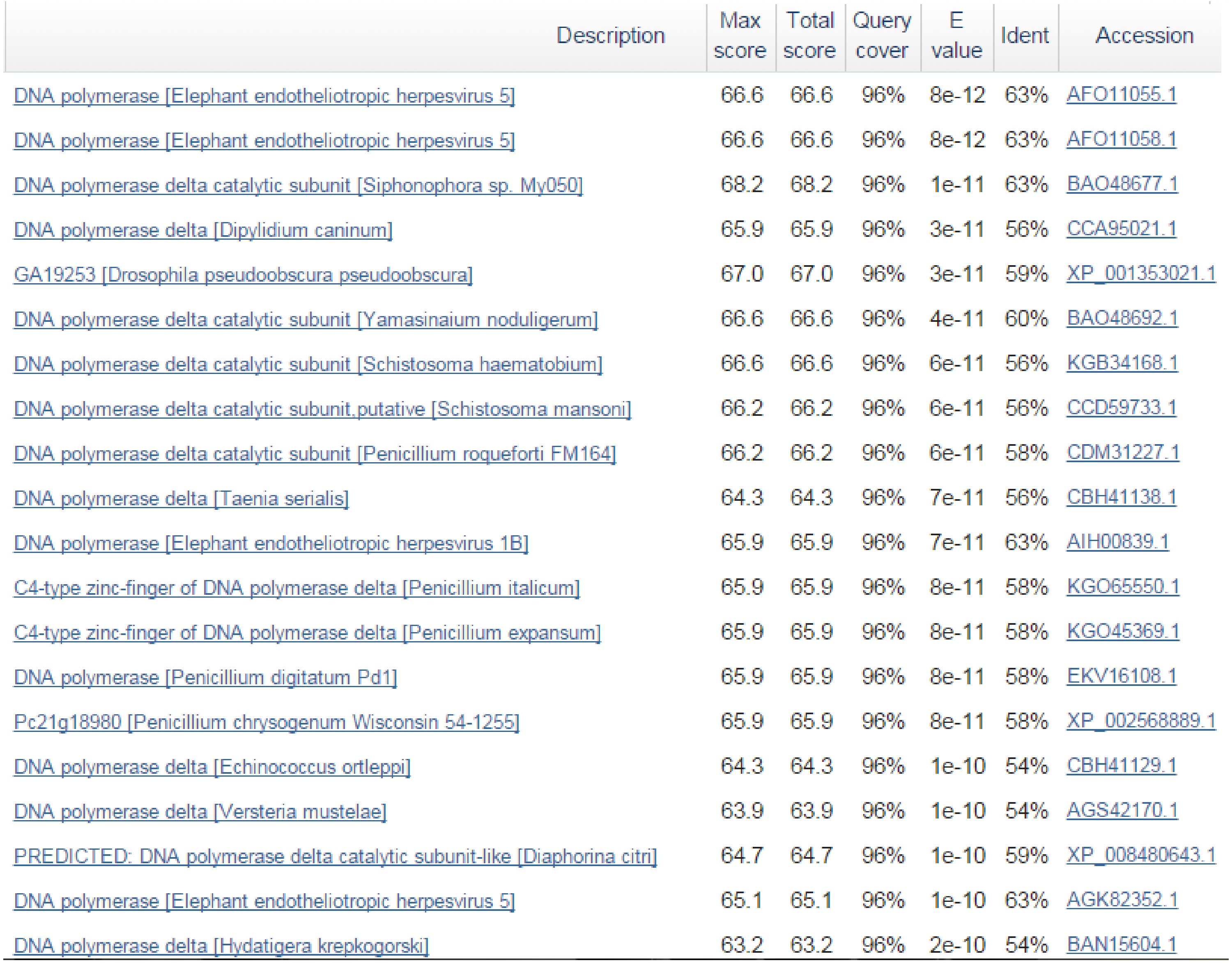
BLASTx alignment for the contig C3733744

BLASTx search showed that the assembled query sequence was similar to conserved DNA polymerase of type-B delta sub-family catalytic domain. Sequence identity to top hit (Elephant endotheliotropic herpesvirus 5) was 63%, matched length was 30 amino acid of total 48 amino acids, query coverage was 96% and E value was 8e-12. However, the query sequence was also similar to *Siphonophora* species (Cnidaria phylum) E-value of 1e-11, for *Dipylidium caninum* (double-pore tapeworm), and *Drosophila* E-value was 3e-11.

A) Graphic display



B) Hit list of aligned NCBI sequences for C3733744 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
DNA polymerase [Elephant endotheliotropic herpesvirus 5]	66.6	66.6	96%	8e-12	63%	AFO11055.1
DNA polymerase [Elephant endotheliotropic herpesvirus 5]	66.6	66.6	96%	8e-12	63%	AFO11058.1
DNA polymerase delta catalytic subunit [Siphonophora sp. My050]	68.2	68.2	96%	1e-11	63%	BAO48677.1
DNA polymerase delta [Dipylidium caninum]	65.9	65.9	96%	3e-11	56%	CCA95021.1
GA19253 [Drosophila pseudoobscura pseudoobscura]	67.0	67.0	96%	3e-11	59%	XP_001353021.1
DNA polymerase delta catalytic subunit [Yamasinaium noduligerum]	66.6	66.6	96%	4e-11	60%	BAO48692.1
DNA polymerase delta catalytic subunit [Schistosoma haematobium]	66.6	66.6	96%	6e-11	56%	KGB34168.1
DNA polymerase delta catalytic subunit,putative [Schistosoma mansoni]	66.2	66.2	96%	6e-11	56%	CCD59733.1
DNA polymerase delta catalytic subunit [Penicillium roqueforti FM164]	66.2	66.2	96%	6e-11	58%	CDM31227.1
DNA polymerase delta [Taenia serialis]	64.3	64.3	96%	7e-11	56%	CBH41138.1
DNA polymerase [Elephant endotheliotropic herpesvirus 1B]	65.9	65.9	96%	7e-11	63%	AIH00839.1
C4-type zinc-finger of DNA polymerase delta [Penicillium italicum]	65.9	65.9	96%	8e-11	58%	KGO65550.1
C4-type zinc-finger of DNA polymerase delta [Penicillium expansum]	65.9	65.9	96%	8e-11	58%	KGO45369.1
DNA polymerase [Penicillium digitatum Pd1]	65.9	65.9	96%	8e-11	58%	EKV16108.1
Pc21g18980 [Penicillium chrysogenum Wisconsin 54-1255]	65.9	65.9	96%	8e-11	58%	XP_002568889.1
DNA polymerase delta [Echinococcus ortleppi]	64.3	64.3	96%	1e-10	54%	CBH41129.1
DNA polymerase delta [Versteria mustelae]	63.9	63.9	96%	1e-10	54%	AGS42170.1
PREDICTED: DNA polymerase delta catalytic subunit-like [Diaphorina citri]	64.7	64.7	96%	1e-10	59%	XP_008480643.1
DNA polymerase [Elephant endotheliotropic herpesvirus 5]	65.1	65.1	96%	1e-10	63%	AGK82352.1
DNA polymerase delta [Hydatigera krepkogorski]	63.2	63.2	96%	2e-10	54%	BAN15604.1

3.1.5. BLAST alignment for the contig C3526692 (144 BP)

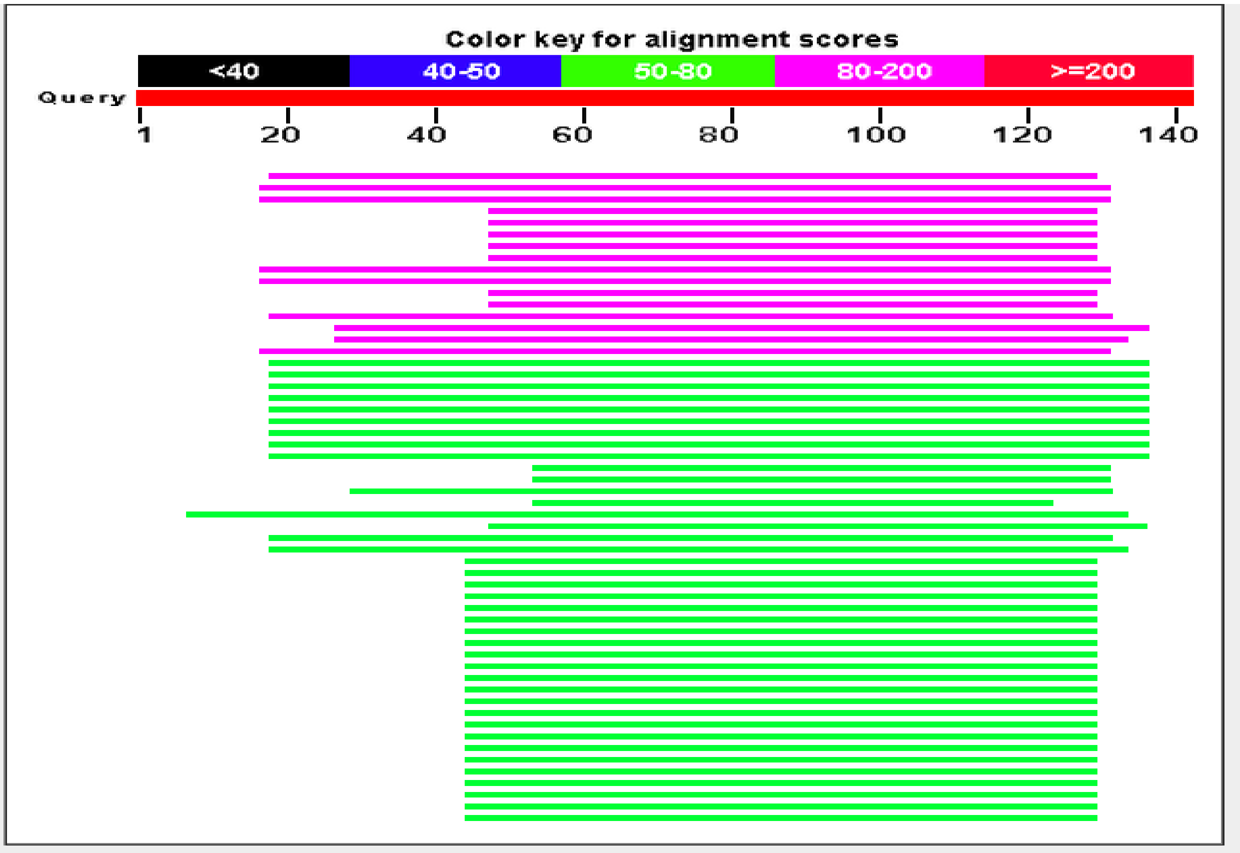
Nucleotide sequence in FASTA format

>C3526692
GAAGATTCTTGTTCTTATTTCGGTACTCAGATGAGATTTGACATACGCAACGAATTT
CCGCTATTAACTACAAAGAAGGTATATTGGAGAGGAGTAGTGGAAGAACTGTTATG
GTTTATCAGAGGTTCTACAAGTTCTTTAGAA

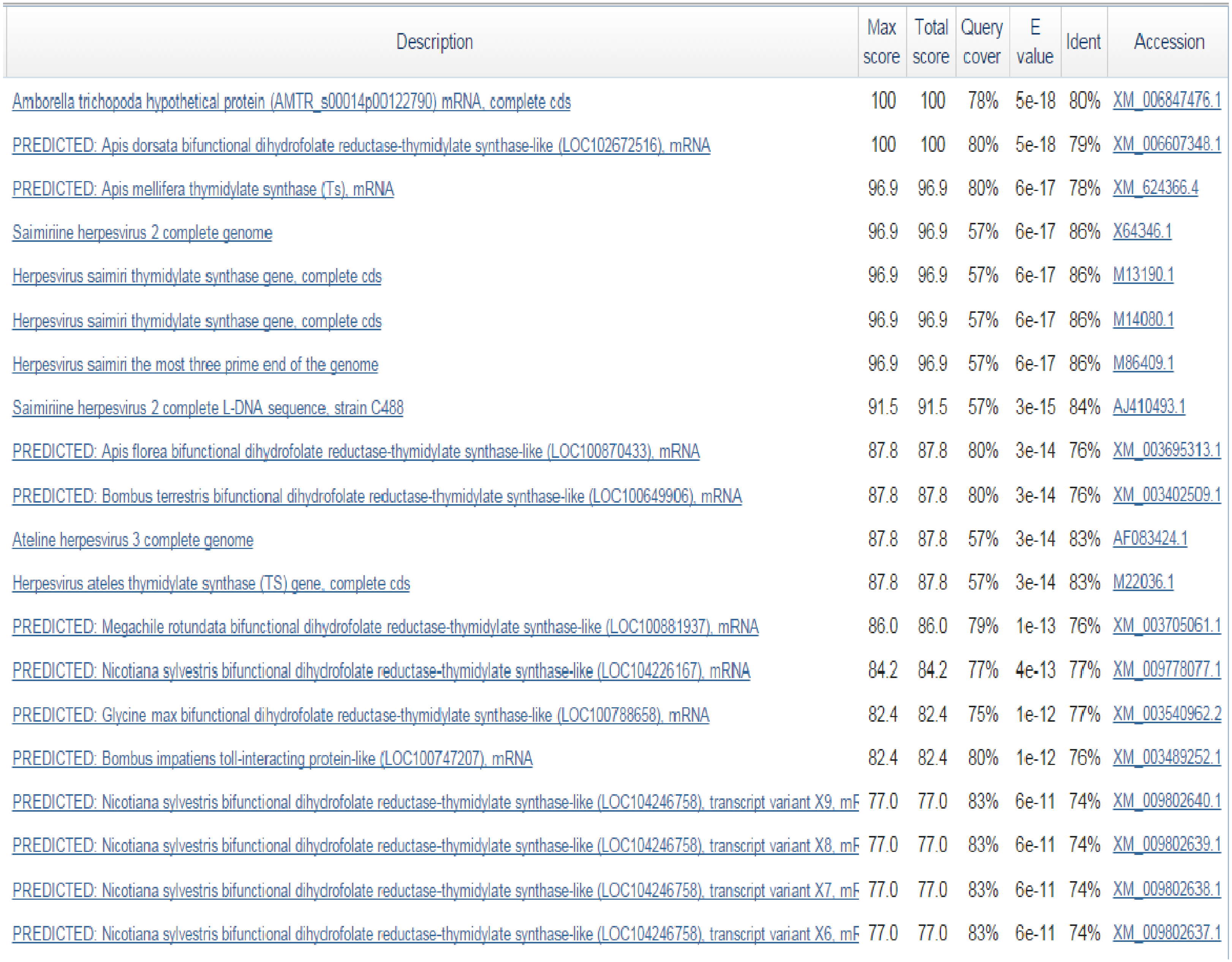
BLASTn alignment for the contig C3526692

BLASTn search result showed query sequence to be similar to a plant sequence *(Amborella)* and with Saimiriine herpesvirus 2. The best match sequence (plant sequence) had 78% of query coverage (113 bp match), E value was 5e-18 and the identity was of 80% with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C3526692 after BLASTn search

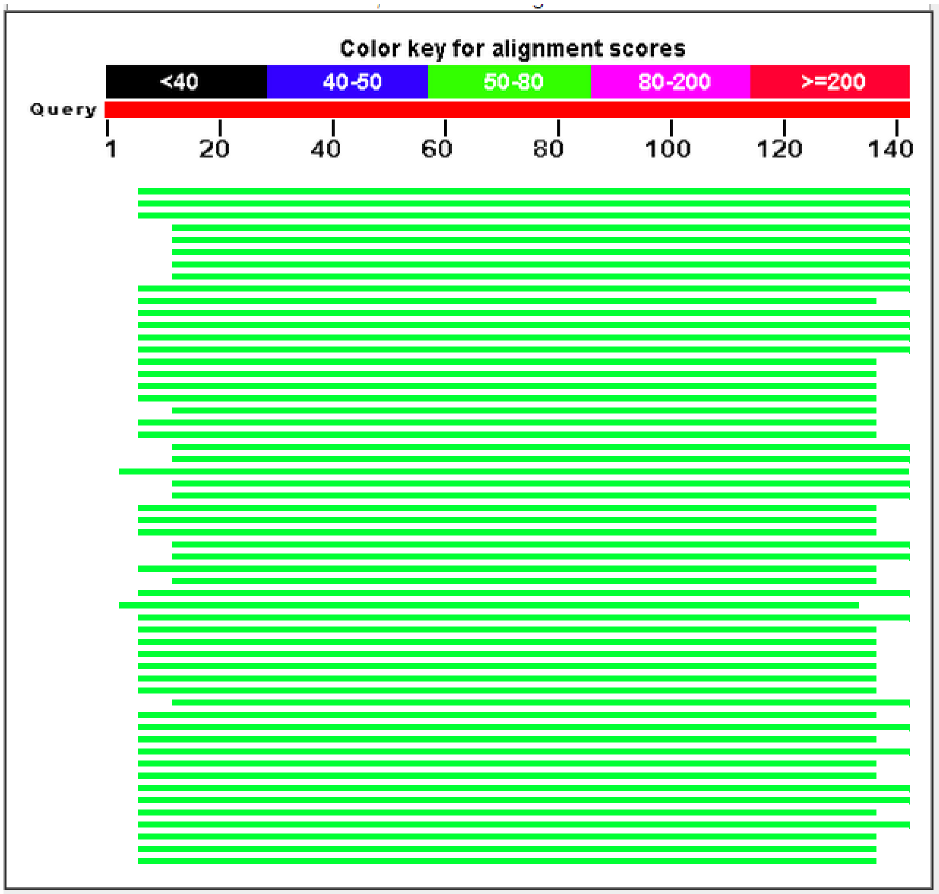


Description	Max score	Total score	Query cover	E value	Ident	Accession
Amborella trichopoda hypothetical protein (AMTR_s00014p00122790) mRNA, complete cds	100	100	78%	5e-18	80%	XM_006847476.1
PREDICTED: Apis dorsata bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC102672516), mRNA	100	100	80%	5e-18	79%	XM_006607348.1
PREDICTED: Apis mellifera thymidylate synthase (Ts), mRNA	96.9	96.9	80%	6e-17	78%	XM_624366.4
Saimiriine herpesvirus 2 complete genome	96.9	96.9	57%	6e-17	86%	X64346.1
Herpesvirus saimiri thymidylate synthase gene, complete cds	96.9	96.9	57%	6e-17	86%	M13190.1
Herpesvirus saimiri thymidylate synthase gene, complete cds	96.9	96.9	57%	6e-17	86%	M14080.1
Herpesvirus saimiri the most three prime end of the genome	96.9	96.9	57%	6e-17	86%	M86409.1
Saimiriine herpesvirus 2 complete L-DNA sequence, strain C488	91.5	91.5	57%	3e-15	84%	AJ410493.1
PREDICTED: Apis florea bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC100870433), mRNA	87.8	87.8	80%	3e-14	76%	XM_003695313.1
PREDICTED: Bombus terrestris bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC100649906), mRNA	87.8	87.8	80%	3e-14	76%	XM_003402509.1
Ateline herpesvirus 3 complete genome	87.8	87.8	57%	3e-14	83%	AF083424.1
Herpesvirus ateles thymidylate synthase (TS) gene, complete cds	87.8	87.8	57%	3e-14	83%	M22036.1
PREDICTED: Megachile rotundata bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC100881937), mRNA	86.0	86.0	79%	1e-13	76%	XM_003705061.1
PREDICTED: Nicotiana sylvestris bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC104226167), mRNA	84.2	84.2	77%	4e-13	77%	XM_009778077.1
PREDICTED: Glycine max bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC100788658), mRNA	82.4	82.4	75%	1e-12	77%	XM_003540962.2
PREDICTED: Bombus impatiens toll-interacting protein-like (LOC100747207), mRNA	82.4	82.4	80%	1e-12	76%	XM_003489252.1
PREDICTED: Nicotiana sylvestris bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC104246758), transcript variant X9, mF	77.0	77.0	83%	6e-11	74%	XM_009802640.1
PREDICTED: Nicotiana sylvestris bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC104246758), transcript variant X8, mF	77.0	77.0	83%	6e-11	74%	XM_009802639.1
PREDICTED: Nicotiana sylvestris bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC104246758), transcript variant X7, mF	77.0	77.0	83%	6e-11	74%	XM_009802638.1
PREDICTED: Nicotiana sylvestris bifunctional dihydrofolate reductase-thymidylate synthase-like (LOC104246758), transcript variant X6, mF	77.0	77.0	83%	6e-11	74%	XM_009802637.1

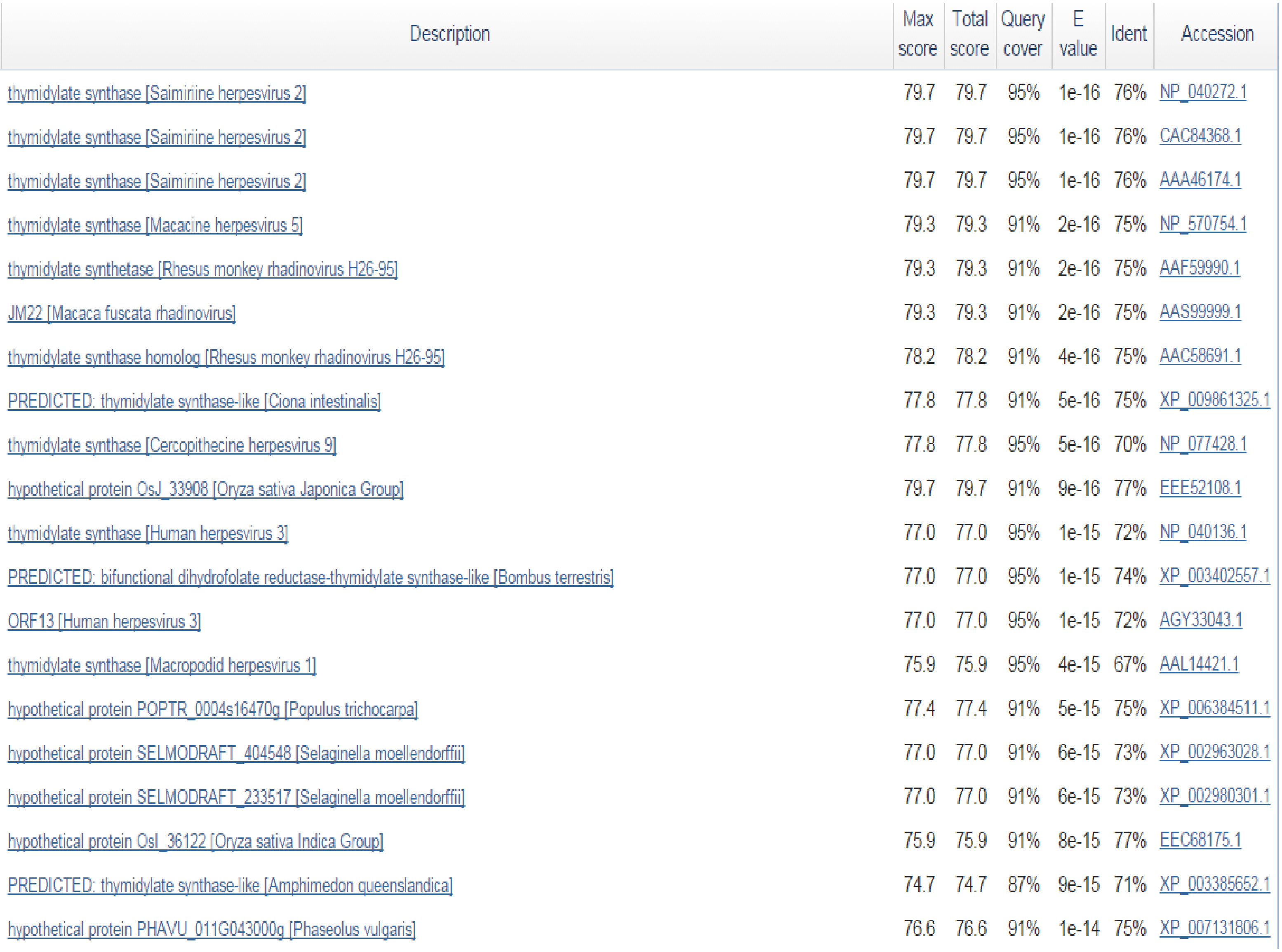
BLASTx alignment for the contig C3526692

After BLASTx search, query sequence was aligned with the conserved thymidylate synthase superfamily. Sequence identity to top hit Saimiriine herpesvirus 2 was 76%, matched length was 35 amino acid out of total 46 amino acids, query coverage was 95% and E value was 1e-16. Within the top five hits, the query sequence was also similar to the conserved thymidylate synthase of rhesus monkey with E-value of 2e-16, 91% of query coverage and 75 % identity at the amino acid level.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C3526692 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
thymidylate synthase [Saimiriine herpesvirus 2]	79.7	79.7	95%	1e-16	76%	NP_040272.1
thymidylate synthase [Saimiriine herpesvirus 2]	79.7	79.7	95%	1e-16	76%	CAC84368.1
thymidylate synthase [Saimiriine herpesvirus 2]	79.7	79.7	95%	1e-16	76%	AAA46174.1
thymidylate synthase [Macacine herpesvirus 5]	79.3	79.3	91%	2e-16	75%	NP_570754.1
thymidylate synthetase [Rhesus monkey rhadinovirus H26-95]	79.3	79.3	91%	2e-16	75%	AAF59990.1
JM22 [Macaca fuscata rhadinovirus]	79.3	79.3	91%	2e-16	75%	AAS99999.1
thymidylate synthase homolog [Rhesus monkey rhadinovirus H26-95]	78.2	78.2	91%	4e-16	75%	AAC58691.1
PREDICTED: thymidylate synthase-like [Ciona intestinalis]	77.8	77.8	91%	5e-16	75%	XP_009861325.1
thymidylate synthase [Cercopithecine herpesvirus 9]	77.8	77.8	95%	5e-16	70%	NP_077428.1
hypothetical protein OsJ_33908 [Oryza sativa Japonica Group]	79.7	79.7	91%	9e-16	77%	EEE52108.1
thymidylate synthase [Human herpesvirus 3]	77.0	77.0	95%	1e-15	72%	NP_040136.1
PREDICTED: bifunctional dihydrofolate reductase-thymidylate synthase-like [Bombus terrestris]	77.0	77.0	95%	1e-15	74%	XP_003402557.1
ORF13 [Human herpesvirus 3]	77.0	77.0	95%	1e-15	72%	AGY33043.1
thymidylate synthase [Macropodid herpesvirus 1]	75.9	75.9	95%	4e-15	67%	AAL14421.1
hypothetical protein POPTR_0004s16470g [Populus trichocarpa]	77.4	77.4	91%	5e-15	75%	XP_006384511.1
hypothetical protein SELMODRAFT_404548 [Selaginella moellendorffii]	77.0	77.0	91%	6e-15	73%	XP_002963028.1
hypothetical protein SELMODRAFT_233517 [Selaginella moellendorffii]	77.0	77.0	91%	6e-15	73%	XP_002980301.1
hypothetical protein OsI_36122 [Oryza sativa Indica Group]	75.9	75.9	91%	8e-15	77%	EEC68175.1
PREDICTED: thymidylate synthase-like [Amphimedon queenslandica]	74.7	74.7	87%	9e-15	71%	XP_003385652.1
hypothetical protein PHAVU_011G043000g [Phaseolus vulgaris]	76.6	76.6	91%	1e-14	75%	XP_007131806.1

3.1.6. BLAST alignment for the contig C8217291 (730 BP)

Nucleotide sequence in FASTA format

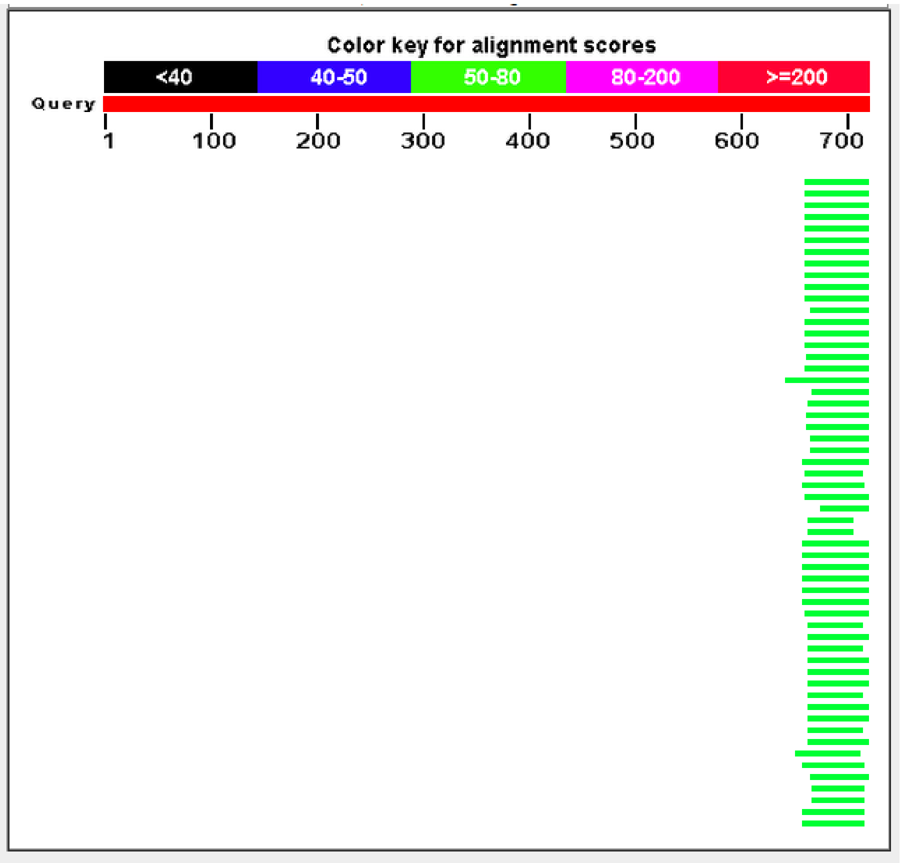
```
>C8217291
TGCTGCTATTCTTGCTGGAGAAAAAGCTGCTGGTGCTGCTGCTCTTGAGACAGCAAG
ACTTGGCACTGCTGTAGCAATTGGTCAGGCACAACTCAGCAAAGAGATTGCTGAAT
CTAAGTATGAAATTAGCAGACATGTTTCTGCCGAAAGCGATGCAACCCGTGGACTG
ATTAACAGTTTAAAGACTGATGAGCTCAATCGTATGCTGATTGAGCGCAATACCGA
TTGCAATCACTTTCGTCACGGATATTGGGATGCTGTAGGTGGTGCTAACAATGCACA
GTTCGCTTCAGTTGCTTCCCAGCTTAATGCTTTCCAAAGCCAATTACAAGAAACCCG
TCAGGGAATGGTAAACTTTGGAACAATGGCTGGAGTTCGTCAGGATTCAACCAGCA
ACAATGTTCGCTGATCTAGTTCAGTAGTTATACTGGGGAGAAGTACTTCTCCCCCTT
TTTAAAAGGAGAATAACTATGGACTCAGCAGAAAGGAAACTTATTGATCTATACAA
TCTTCTTGCTCAGTATCAAAGAAGTAATGATCCAGGTCTGATTAGTAATATTCAGGC
AATTCGGACTGAAATTTCACAAAACCTAAATGACCGCATTGGTGGTAGTGGTAATG
ACAATATCAACATCAATATAGATGCAGATAATTGTCCTGATGATTGTCCTCCGGGAC
CTCCAGGGCCTCCTGGAGAGCCAGGACCACCAGGGCCTCCAGGACCACCAGG
```

BLASTn alignment for the contig C8217291

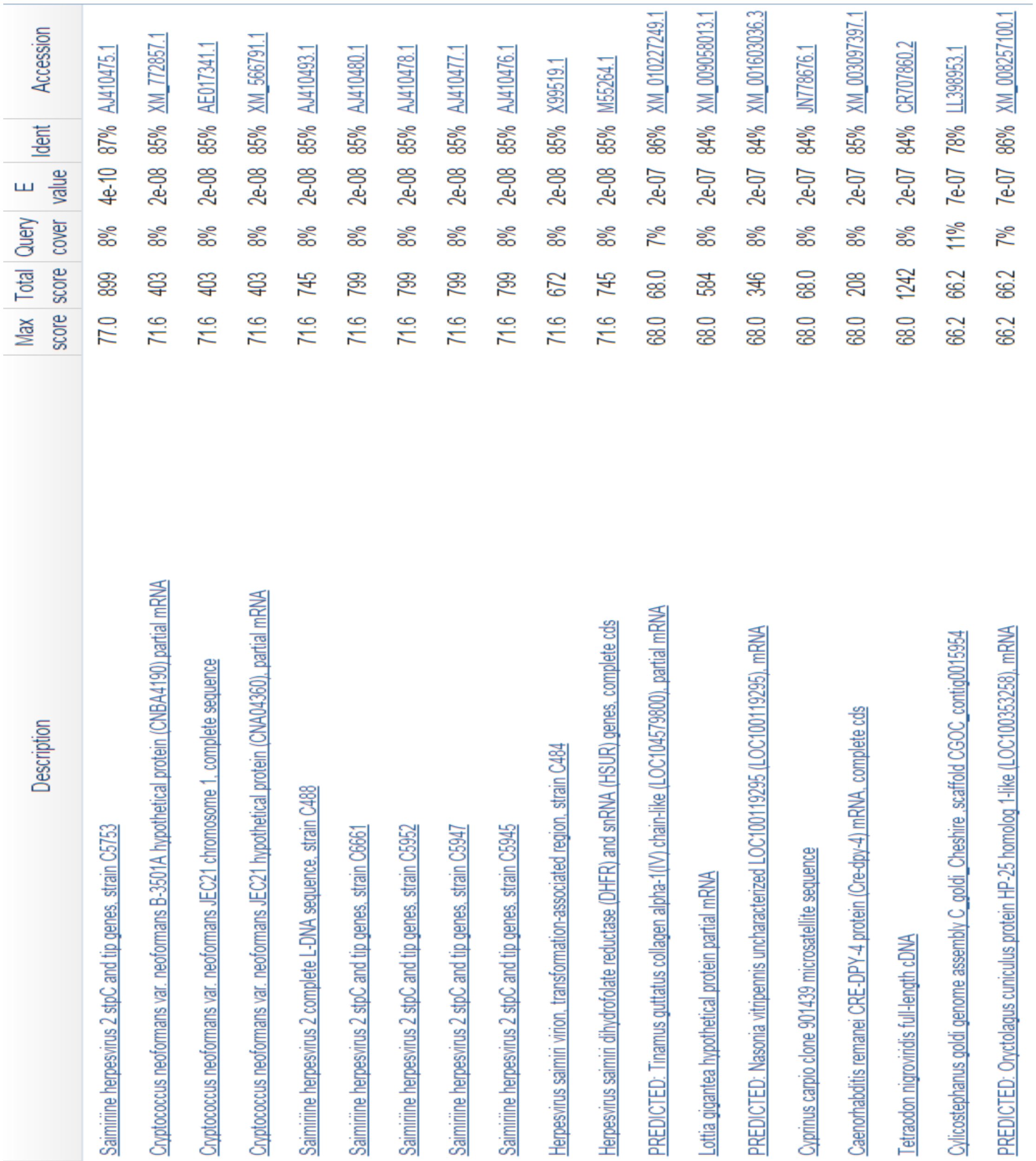
BLASTn search result showed that the assembled query sequence was similar to Saimiriine herpesvirus 2 and to other bacterial sequences. The alignment with the top hit (Saimiriine herpesvirus 2) was only 8% of total coverage (62 bp matched) with E value of 4e-10 and the identity of 87% for the matched region.

The unmatched 650 base pairs fragment from contigs C8217291 was again subjected to BLASTn and BLASTx search separately, which showed that this fragment had no match with reference sequences in the NCBI database which indicates that this sequence could belong to a completely unknown organism.

A) Graphic display



B) Hit list of aligned NCBI sequences for C8217291 after BLASTn search



Description	Max score	Total score	Query cover	E value	Ident	Accession
Saimiriine herpesvirus 2 stpC and tip genes, strain C5753	77.0	899	8%	4e-10	87%	AJ410475.1
Cryptococcus neoformans var. neoformans B-3501A hypothetical protein (CNBA4190) partial mRNA	71.6	403	8%	2e-08	85%	XM_772857.1
Cryptococcus neoformans var. neoformans JEC21 chromosome 1, complete sequence	71.6	403	8%	2e-08	85%	AE017341.1
Cryptococcus neoformans var. neoformans JEC21 hypothetical protein (CNA04360), partial mRNA	71.6	403	8%	2e-08	85%	XM_566791.1
Saimiriine herpesvirus 2 complete L-DNA sequence, strain C488	71.6	745	8%	2e-08	85%	AJ410493.1
Saimiriine herpesvirus 2 stpC and tip genes, strain C6661	71.6	799	8%	2e-08	85%	AJ410480.1
Saimiriine herpesvirus 2 stpC and tip genes, strain C5952	71.6	799	8%	2e-08	85%	AJ410478.1
Saimiriine herpesvirus 2 stpC and tip genes, strain C5947	71.6	799	8%	2e-08	85%	AJ410477.1
Saimiriine herpesvirus 2 stpC and tip genes, strain C5945	71.6	799	8%	2e-08	85%	AJ410476.1
Herpesvirus saimiri virion, transformation-associated region, strain C484	71.6	672	8%	2e-08	85%	X99519.1
Herpesvirus saimiri dihydrofolate reductase (DHFR) and snRNA (HSUR) genes, complete cds	71.6	745	8%	2e-08	85%	M55264.1
PREDICTED: Tinamus guttatus collagen alpha-1(IV) chain-like (LOC104579800), partial mRNA	68.0	68.0	7%	2e-07	86%	XM_010227249.1
Lottia gigantea hypothetical protein partial mRNA	68.0	584	8%	2e-07	84%	XM_009058013.1
PREDICTED: Nasonia vitripennis uncharacterized LOC100119295 (LOC100119295), mRNA	68.0	346	8%	2e-07	84%	XM_001603036.3
Cyprinus carpio clone 901439 microsatellite sequence	68.0	68.0	8%	2e-07	84%	JN778676.1
Caenorhabditis remanei CRE-DPY-4 protein (Cre-dpy-4) mRNA, complete cds	68.0	208	8%	2e-07	85%	XM_003097397.1
Tetraodon nigroviridis full-length cDNA	68.0	1242	8%	2e-07	84%	CR707860.2
Cylicostephanus goldi genome assembly C_goldi_Cheshire ,scaffold CGOC_contig0015954	66.2	66.2	11%	7e-07	78%	LL398953.1
PREDICTED: Oryctolagus cuniculus protein HP-25 homolog 1-like (LOC100353258), mRNA	66.2	66.2	7%	7e-07	86%	XM_008257100.1

BLASTx alignment for C8217291

After BLASTx search, query sequence had been found to have no significant similarity with nucleotides in the NCBI database. This contig could belong to an unknown species as after BLASTn and BLASTx search the contig showed partial match to the sequences of NCBI database.

3.1.7. BLAST alignment for the scaffold *scaffold123145 (*845 BP)

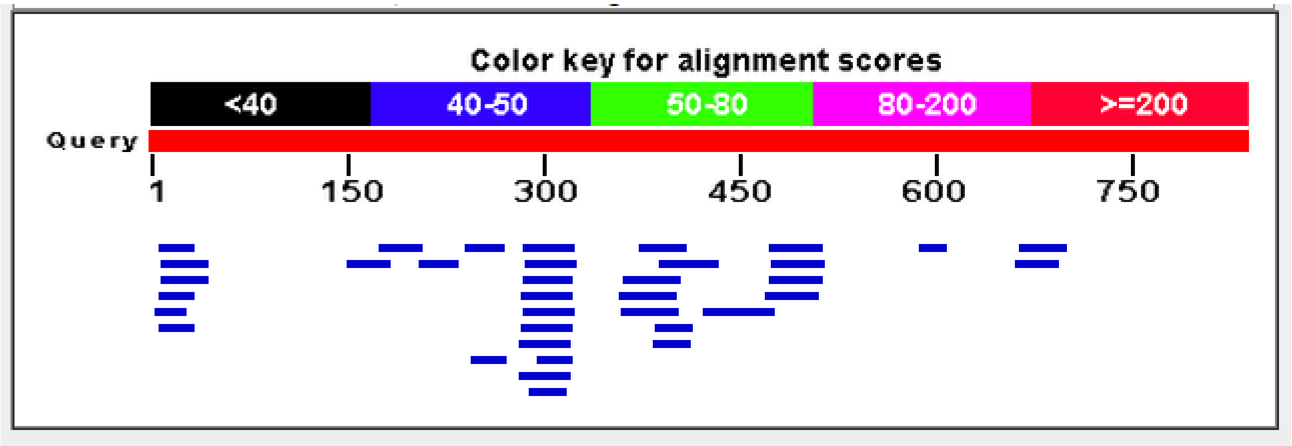
Nucleotide sequence in FASTA format

```
>scaffold123145
TTCGCCTTGAAGGACAAGGACAAGGAACTGGCAAAAGAACGCGGCAAAAGGTCCA
GACTGCAATCCGATGTGATTTTCAAGAATCAGGAGCTGAAGTCTTTCGAAGACAGA
ATCGCTGAACTTGTTCACACCCTCTCCTGCCGCGACATCGAAATCGAACGTCTCACC
TTCGCCTTGAAGAACAAGGAGAACGAAAGGGACTCTCTGCTTGACACGCAAGAGCA
AGAGATCGACGACATTACCCAAAAACTGCAGGAGAAAACCGAGCAACTGGAGCAT
GCCAATCTCGAATTGGAAGCCAAGGGTGAGCAGATCGACGAACTCATGAAGGAACT
CAACGACGTGAAAGAACTGTATTACAACAGGTTGAAAATCGAGGAGCAGCAGGAC
GAAGAAATCGAACGCCTGACGCGCAAGAACAAGAAGCCCTTCAAGGCGGCTGAGT
TGCGGTGGCTGATCGACGGGCACACAGACGTGTTGGAAGAGCATCGCACCGCCATC
CTGGAAGTGCGCGAGGCCATTGACCAGAACAAAGAAACGTTGAGAGCTGCTCAAC
GGGAGATGCTGGCAACCCTTGATCGCCAAGATGGGTTGATCAGCAAGGCGGTTGCC
TGCATCGACAACCACTCGGACTCGATCTACAACCAACAGCAGTCCATCGACGCCTT
GTGGGACTATGTCAAGAGCGTTGCGGCGCACATCAACACAGGTAGCAGTCAGGAGA
ACCCGTGATGAACCCGAACAAATGTGAAACCTGTGACTACTCGAAGATGGGTAGCA
CAGGAGGCTGGTGCTACGTGTGGCGCTCAGAACCAACGCATGTGTGTTATCAAGAT
CGGAAGAGC
```

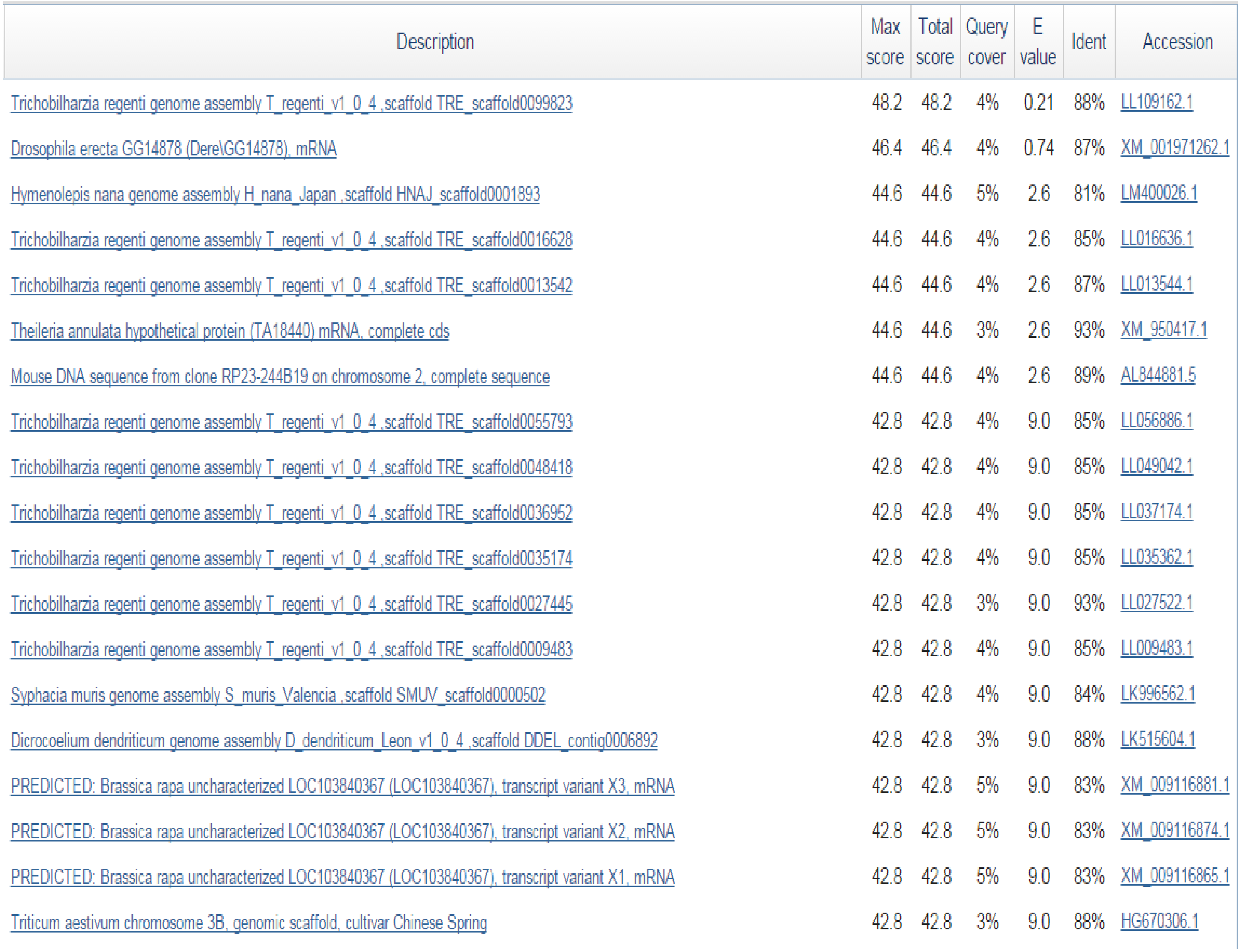
BLASTn alignment for the scaffold *scaffold123145*

BLASTn search result showed the assembled query sequence was similar to a sequence of nasal parasite of bird (the parasite belongs to platyhelminthes phylum), E value was 0.21 and 4% of total coverage (40bp matched), and the identity was with 88% with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the scaffold *scaffold123145* after BLASTn search



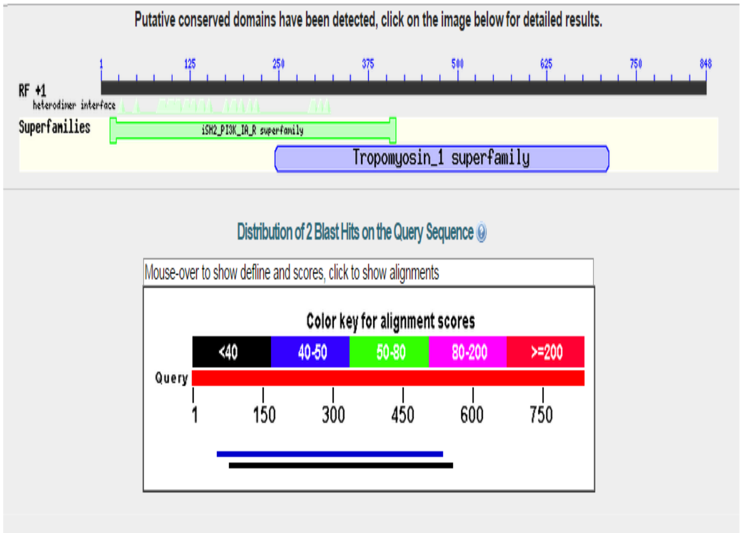
Description	Max score	Total score	Query cover	E value	Ident	Accession
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0099823	48.2	48.2	4%	0.21	88%	LL109162.1
Drosophila erecta GG14878 (Dere\GG14878), mRNA	46.4	46.4	4%	0.74	87%	XM_001971262.1
Hymenolepis nana genome assembly H_nana_Japan ,scaffold HNAJ_scaffold0001893	44.6	44.6	5%	2.6	81%	LM400026.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0016628	44.6	44.6	4%	2.6	85%	LL016636.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0013542	44.6	44.6	4%	2.6	87%	LL013544.1
Theileria annulata hypothetical protein (TA18440) mRNA, complete cds	44.6	44.6	3%	2.6	93%	XM_950417.1
Mouse DNA sequence from clone RP23-244B19 on chromosome 2, complete sequence	44.6	44.6	4%	2.6	89%	AL844881.5
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0055793	42.8	42.8	4%	9.0	85%	LL056886.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0048418	42.8	42.8	4%	9.0	85%	LL049042.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0036952	42.8	42.8	4%	9.0	85%	LL037174.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0035174	42.8	42.8	4%	9.0	85%	LL035362.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0027445	42.8	42.8	3%	9.0	93%	LL027522.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0009483	42.8	42.8	4%	9.0	85%	LL009483.1
Syphacia muris genome assembly S_muris_Valencia ,scaffold SMUV_scaffold0000502	42.8	42.8	4%	9.0	84%	LK996562.1
Dicrocoelium dendriticum genome assembly D_dendriticum_Leon_v1_0_4 ,scaffold DDEL_contig0006892	42.8	42.8	3%	9.0	88%	LK515604.1
PREDICTED: Brassica rapa uncharacterized LOC103840367 (LOC103840367), transcript variant X3, mRNA	42.8	42.8	5%	9.0	83%	XM_009116881.1
PREDICTED: Brassica rapa uncharacterized LOC103840367 (LOC103840367), transcript variant X2, mRNA	42.8	42.8	5%	9.0	83%	XM_009116874.1
PREDICTED: Brassica rapa uncharacterized LOC103840367 (LOC103840367), transcript variant X1, mRNA	42.8	42.8	5%	9.0	83%	XM_009116865.1
Triticum aestivum chromosome 3B, genomic scaffold, cultivar Chinese Spring	42.8	42.8	3%	9.0	88%	HG670306.1

BLASTx alignment for the scaffold *scaffold123145*

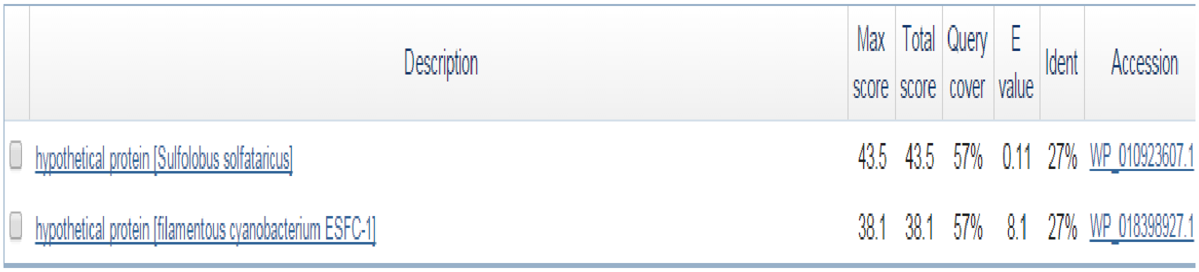
After BLASTx search, query sequence was aligned with a member of archaea and cyanobacteria. The E value was 0.11 for the archaea sequence which implied less significant hit. The fragment of *scaffold123145* (fragment from 550-845 bp position) was searched using BLASTn and BLASTx which also showed no similarity of this fragment with sequences in the NCBI database.

This contig could belong to an unknown species as after BLASTn and BLASTx search the contig showed less similarity with insignificant E-value to sequences of NCBI database

A) Graphic display


Putative conserved domains have been detected, click on the image below for detailed results.
1
125
250
375
500
625
750
848
RF +1
heterodimer interface
Superfamilies
iSH2_PI3K_IA_R superfamily
Tropomyosin_1 superfamily
Distribution of 2 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
150
300
450
600
750

B) Hit list of aligned NCBI sequences for scaffold *scaffold123145* after BLASTx search



	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	hypothetical protein [Sulfolobus solfataricus]	43.5	43.5	57%	0.11	27%	WP_010923607.1
□	hypothetical protein [filamentous cyanobacterium ESFC-1]	38.1	38.1	57%	8.1	27%	WP_018398927.1

3.1.8. BLAST alignment for the contig C3668266 (149 BP)

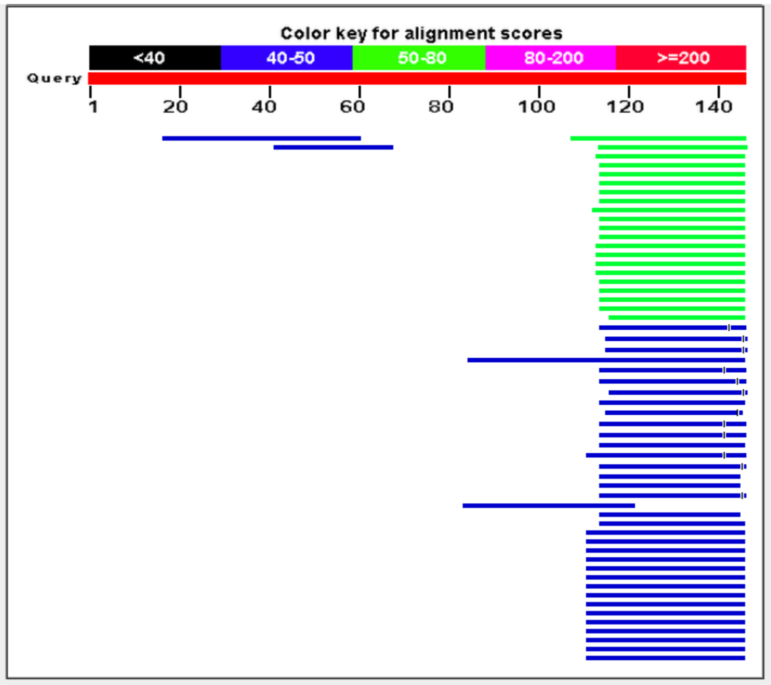
Nucleotide sequence in FASTA format

>C3668266
CCTCTTGGAAGTTCGGTGCGTCCACAGGGTCAAGGCGGATTACCTTGTTGCCCTTCT
TCACCTCAACCGACCAGCGTTTCATCTTGGGTGTGGAAGGAGACGCGCCAGCCGCC
CCAGTGGGTGCTTTGGGTGCTGTGGGTGCTGTGGGT

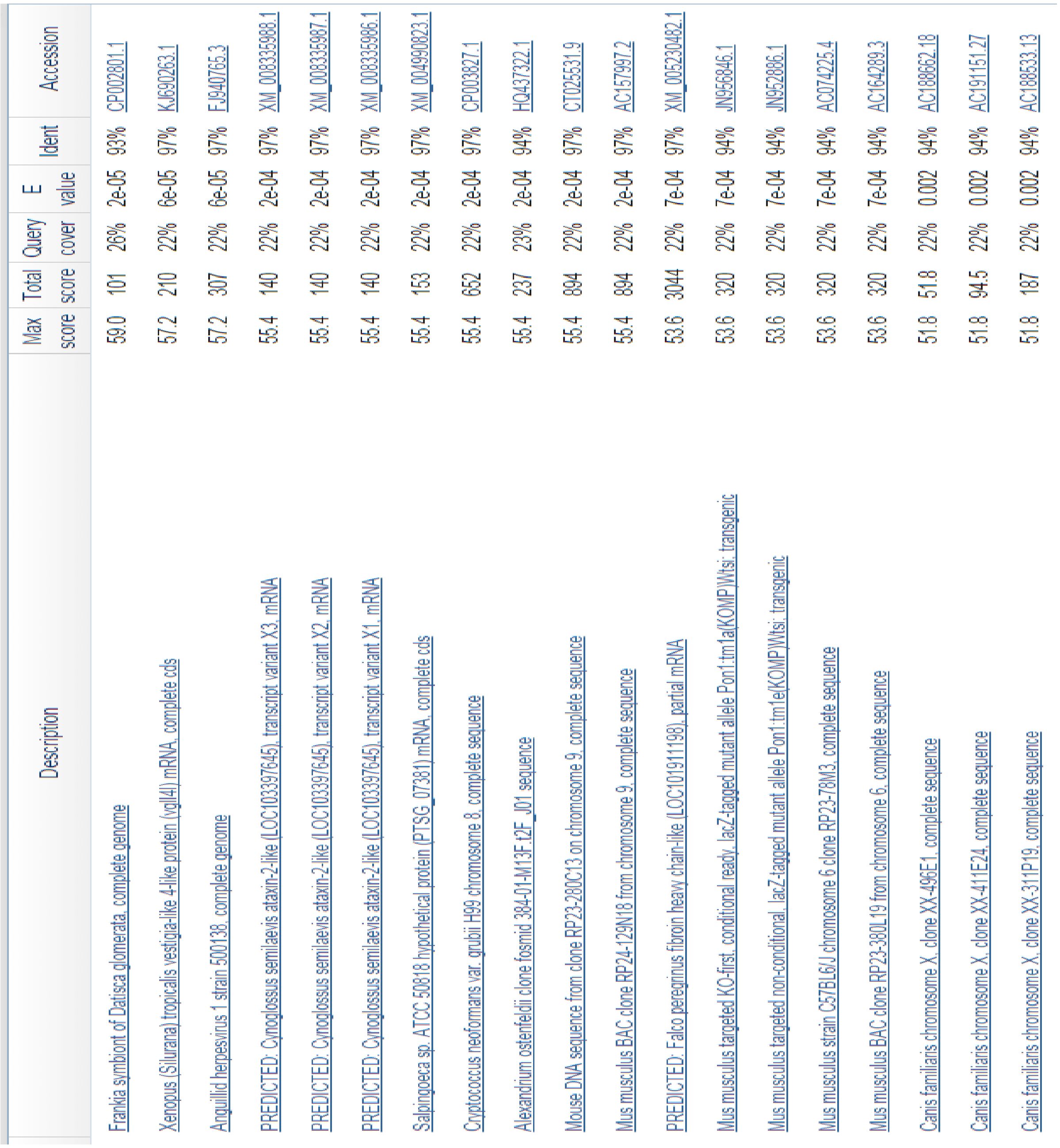
BLASTn alignment for the contig C3668266

BLASTn search result showed that the query sequence was similar to the sequence from bacteria, *Mus musculus* BAC clone (mouse genome) or Anguillid herpesvirus sequence. The E value was 2e-05 for the top hit (bacteria). Since the query sequence was hitting with different kind of sequences with E value ranged from 7e-04 to 2e-05, this sequence is ambiguous in nature. BLASTn search of the first hundred base pairs of this contigs C3668266 showed the query sequence was similar to the sequence of *Streptococcus* species with E-value of 0.061 and BLASTx search showed this sequence was matched with *Campylobacter* with insignificant E-value of 5.9 (higher than 1e-5).

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C3668266 after BLASTn search



Description	Max score	Total score	Query cover	E value	Ident	Accession
Frankia symbiont of Datisca glomerata, complete genome	59.0	101	26%	2e-05	93%	CP002801.1
Xenopus (Silurana) tropicalis vestigia-like 4-like protein (vgll4l) mRNA, complete cds	57.2	210	22%	6e-05	97%	KJ690263.1
Anguillid herpesvirus 1 strain 500138, complete genome	57.2	307	22%	6e-05	97%	FJ940765.3
PREDICTED: Cynoglossus semilaevis ataxin-2-like (LOC103397645), transcript variant X3, mRNA	55.4	140	22%	2e-04	97%	XM_008335988.1
PREDICTED: Cynoglossus semilaevis ataxin-2-like (LOC103397645), transcript variant X2, mRNA	55.4	140	22%	2e-04	97%	XM_008335987.1
PREDICTED: Cynoglossus semilaevis ataxin-2-like (LOC103397645), transcript variant X1, mRNA	55.4	140	22%	2e-04	97%	XM_008335986.1
Salpingoeca sp. ATCC 50818 hypothetical protein (PTSG_07381) mRNA, complete cds	55.4	153	22%	2e-04	97%	XM_004990823.1
Cryptococcus neoformans var. grubii H99 chromosome 8, complete sequence	55.4	652	22%	2e-04	97%	CP003827.1
Alexandrium ostenfeldii clone fosmid 384-01-M13F.t2F_J01 sequence	55.4	237	23%	2e-04	94%	HQ437322.1
Mouse DNA sequence from clone RP23-280C13 on chromosome 9, complete sequence	55.4	894	22%	2e-04	97%	CT025531.9
Mus musculus BAC clone RP24-129N18 from chromosome 9, complete sequence	55.4	894	22%	2e-04	97%	AC157997.2
PREDICTED: Falco peregrinus fibroin heavy chain-like (LOC101911198), partial mRNA	53.6	3044	22%	7e-04	97%	XM_005230482.1
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Pon1:tm1a(KOMP)Wtsi; transgenic	53.6	320	22%	7e-04	94%	JN956846.1
Mus musculus targeted non-conditional, lacZ-tagged mutant allele Pon1:tm1e(KOMP)Wtsi; transgenic	53.6	320	22%	7e-04	94%	JN952886.1
Mus musculus strain C57BL6/J chromosome 6 clone RP23-78M3, complete sequence	53.6	320	22%	7e-04	94%	AC074225.4
Mus musculus BAC clone RP23-380L19 from chromosome 6, complete sequence	53.6	320	22%	7e-04	94%	AC164289.3
Canis familiaris chromosome X, clone XX-496E1, complete sequence	51.8	51.8	22%	0.002	94%	AC188662.18
Canis familiaris chromosome X, clone XX-411E24, complete sequence	51.8	94.5	22%	0.002	94%	AC191151.27
Canis familiaris chromosome X, clone XX-311P19, complete sequence	51.8	187	22%	0.002	94%	AC188533.13

BLASTx alignment for C3668266

After BLASTx search, query sequence had been found to have no significant similarity with nucleotide nr NCBI database. This contig could belong to an unknown species as after BLASTn and BLASTx search the contig showed only partial alignment and less similarity to the sequences of NCBI database.

3.1.9. BLAST alignment for the contig C5052190 (178 BP)

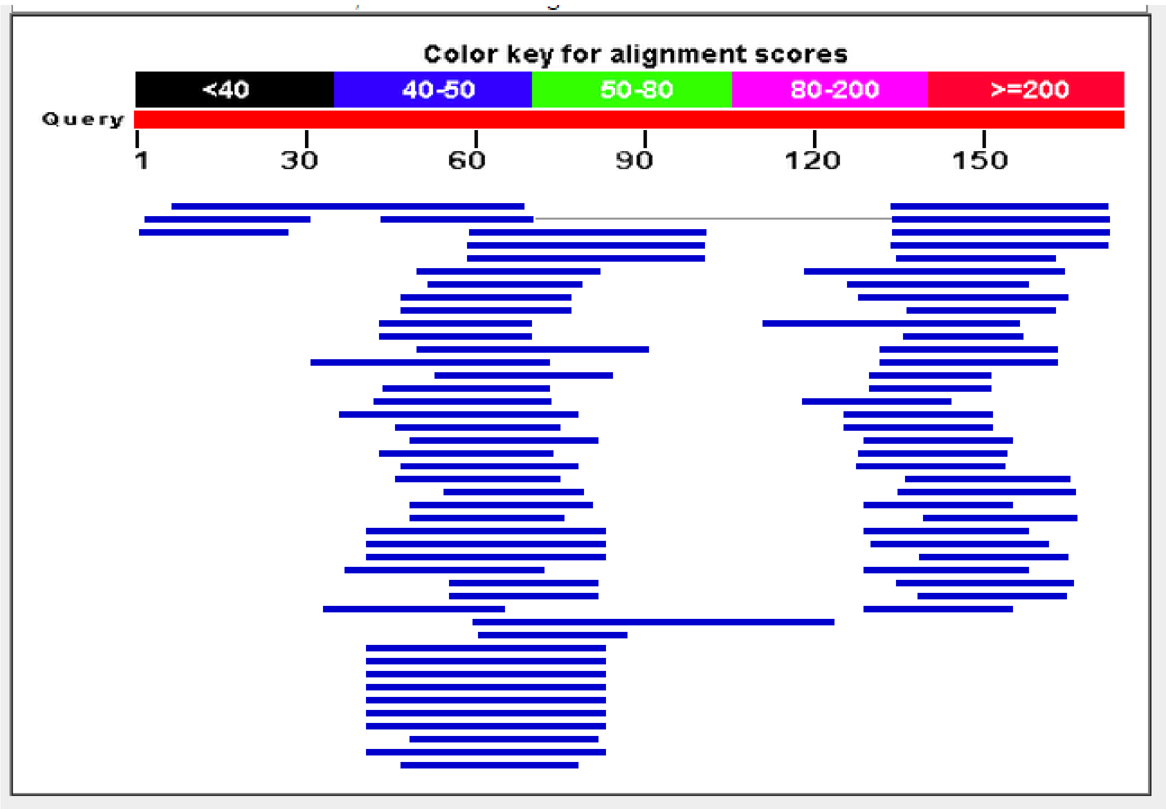
Nucleotide sequence in FASTA format

>C5052190
TGTTTCGATTGTTTATGAACCACTTGATCAATGGCAGTCTGTTAATGGTCAAAGTTT
ACTTGATAAATTTTATCAAGATACAAAACGGTGGGATATACCTTTCAATCGTATGC
ATTTGTAACTCGGGTAATGGAGCAAGAAAAATATAAAAAGCTTTATCCTTTATTGCC
CCAAGTA

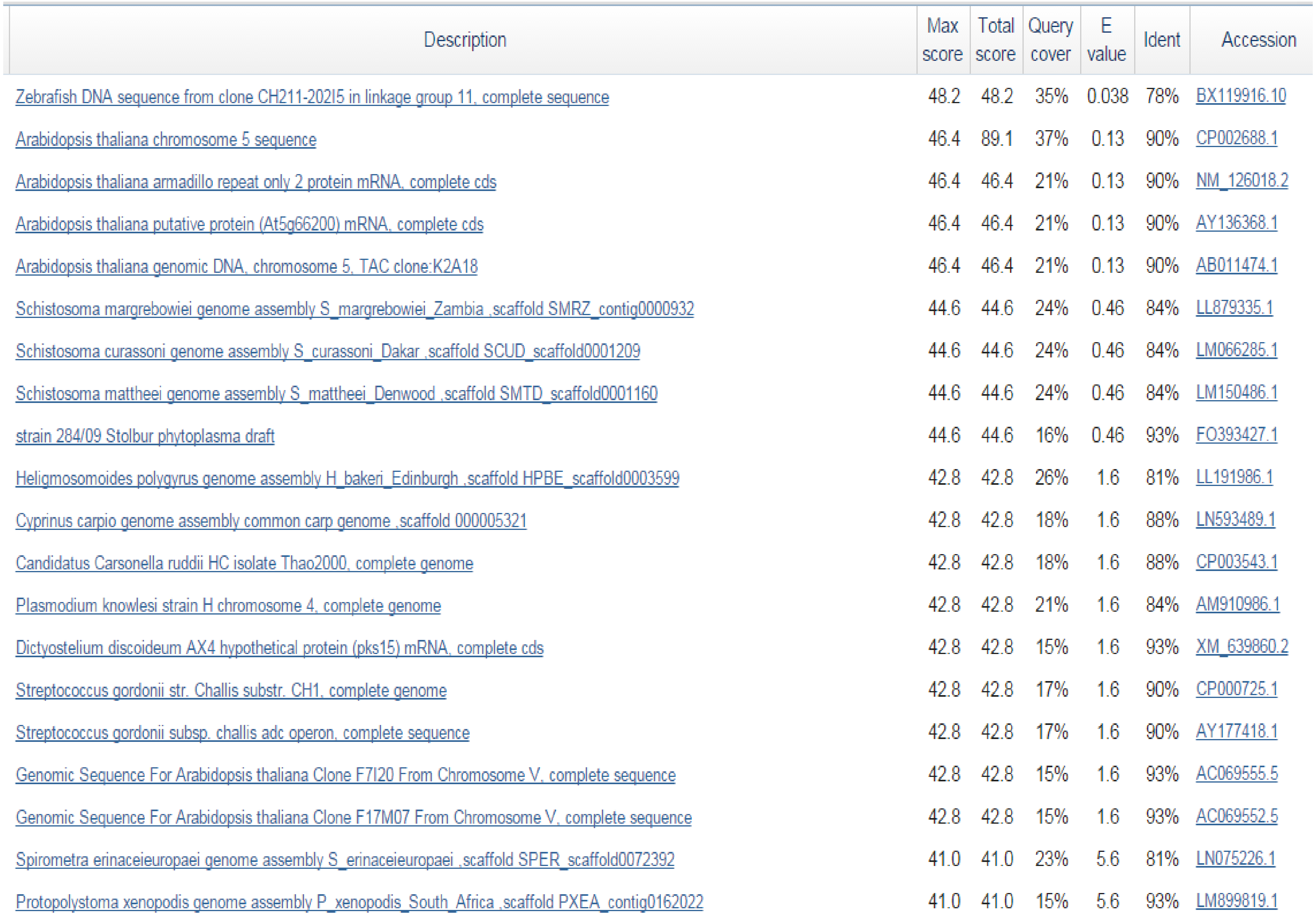
BLASTn alignment for the contig C5052190

BLASTn search result showed that the assembled query sequence had similarity to sequences from different species, whereas for top hit (Zebrafish) E value was 0.038. The E value higher than 0 implied the insignificant alignment with reference sequences.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C5052190 after BLASTn search

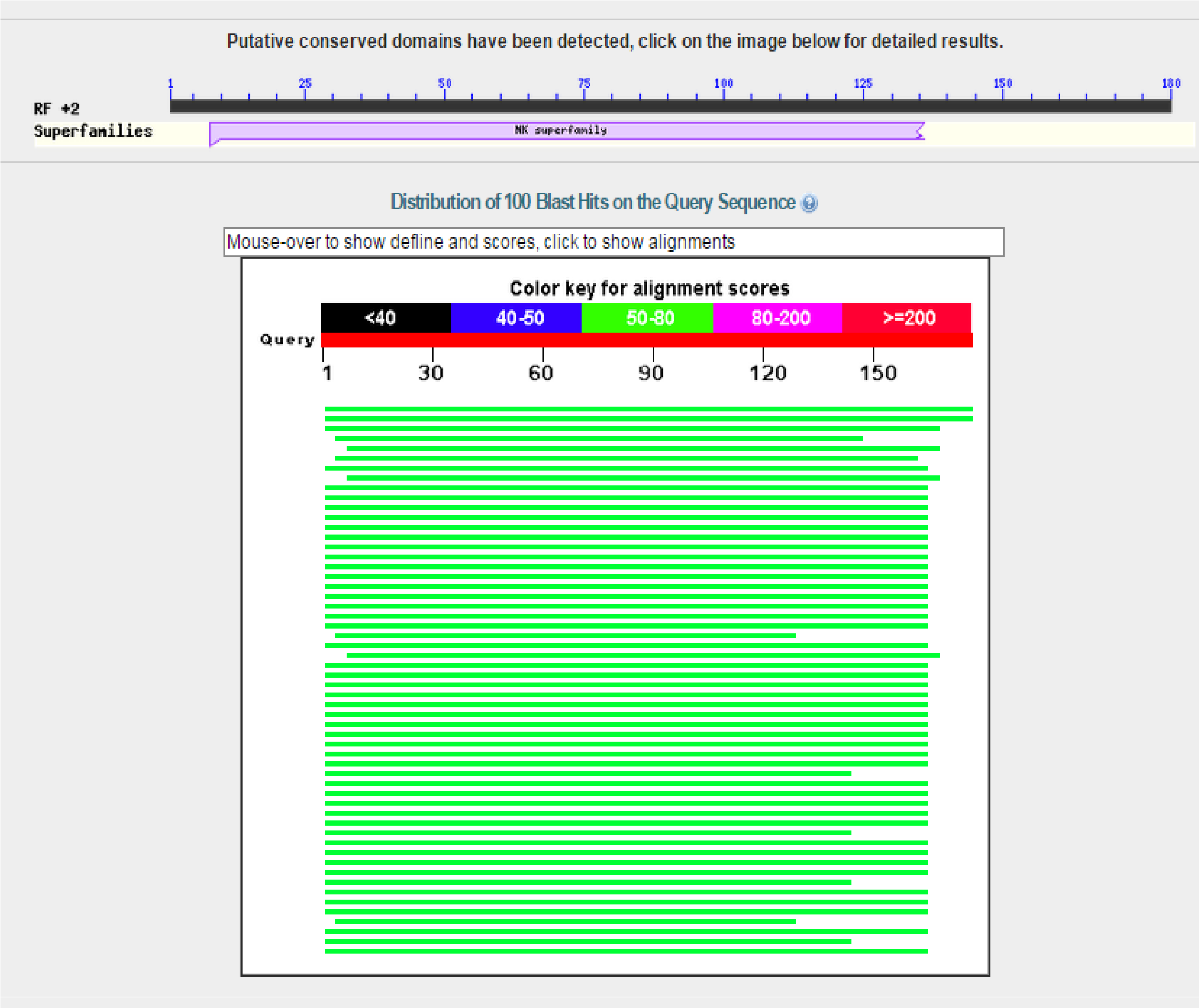


Description	Max score	Total score	Query cover	E value	Ident	Accession
Zebrafish DNA sequence from clone CH211-202I5 in linkage group 11, complete sequence	48.2	48.2	35%	0.038	78%	BX119916.10
Arabidopsis thaliana chromosome 5 sequence	46.4	89.1	37%	0.13	90%	CP002688.1
Arabidopsis thaliana armadillo repeat only 2 protein mRNA, complete cds	46.4	46.4	21%	0.13	90%	NM_126018.2
Arabidopsis thaliana putative protein (At5g66200) mRNA, complete cds	46.4	46.4	21%	0.13	90%	AY136368.1
Arabidopsis thaliana genomic DNA, chromosome 5, TAC clone:K2A18	46.4	46.4	21%	0.13	90%	AB011474.1
Schistosoma margrebowiei genome assembly S_margrebowiei_Zambia ,scaffold SMRZ_contig0000932	44.6	44.6	24%	0.46	84%	LL879335.1
Schistosoma curassoni genome assembly S_curassoni_Dakar ,scaffold SCUD_scaffold0001209	44.6	44.6	24%	0.46	84%	LM066285.1
Schistosoma mattheei genome assembly S_mattheei_Denwood ,scaffold SMTD_scaffold0001160	44.6	44.6	24%	0.46	84%	LM150486.1
strain 284/09 Stolbur phytoplasma draft	44.6	44.6	16%	0.46	93%	FO393427.1
Heligmosomoides polygyrus genome assembly H_bakeri_Edinburgh ,scaffold HPBE_scaffold0003599	42.8	42.8	26%	1.6	81%	LL191986.1
Cyprinus carpio genome assembly common carp genome ,scaffold 000005321	42.8	42.8	18%	1.6	88%	LN593489.1
Candidatus Carsonella ruddii HC isolate Thao2000, complete genome	42.8	42.8	18%	1.6	88%	CP003543.1
Plasmodium knowlesi strain H chromosome 4, complete genome	42.8	42.8	21%	1.6	84%	AM910986.1
Dictyostelium discoideum AX4 hypothetical protein (pks15) mRNA, complete cds	42.8	42.8	15%	1.6	93%	XM_639860.2
Streptococcus gordonii str. Challis substr. CH1, complete genome	42.8	42.8	17%	1.6	90%	CP000725.1
Streptococcus gordonii subsp. challis adc operon, complete sequence	42.8	42.8	17%	1.6	90%	AY177418.1
Genomic Sequence For Arabidopsis thaliana Clone F7I20 From Chromosome V, complete sequence	42.8	42.8	15%	1.6	93%	AC069555.5
Genomic Sequence For Arabidopsis thaliana Clone F17M07 From Chromosome V, complete sequence	42.8	42.8	15%	1.6	93%	AC069552.5
Spirometra erinaceieuropaei genome assembly S_erinaceieuropaei ,scaffold SPER_scaffold0072392	41.0	41.0	23%	5.6	81%	LN075226.1
Protopolystoma xenopodis genome assembly P_xenopodis_South_Africa ,scaffold PXEA_contig0162022	41.0	41.0	15%	5.6	93%	LM899819.1

BLASTx alignment for the contig C5052190

BLASTx search showed that the assembled query sequence was aligned with the hypothetical protein of uncultured bacteria. The E value was 3e-13 for that protein sequence alignment. The identity was 52% for the matched amino acid of 60 and the query coverage was 99%.

A) Graphic display

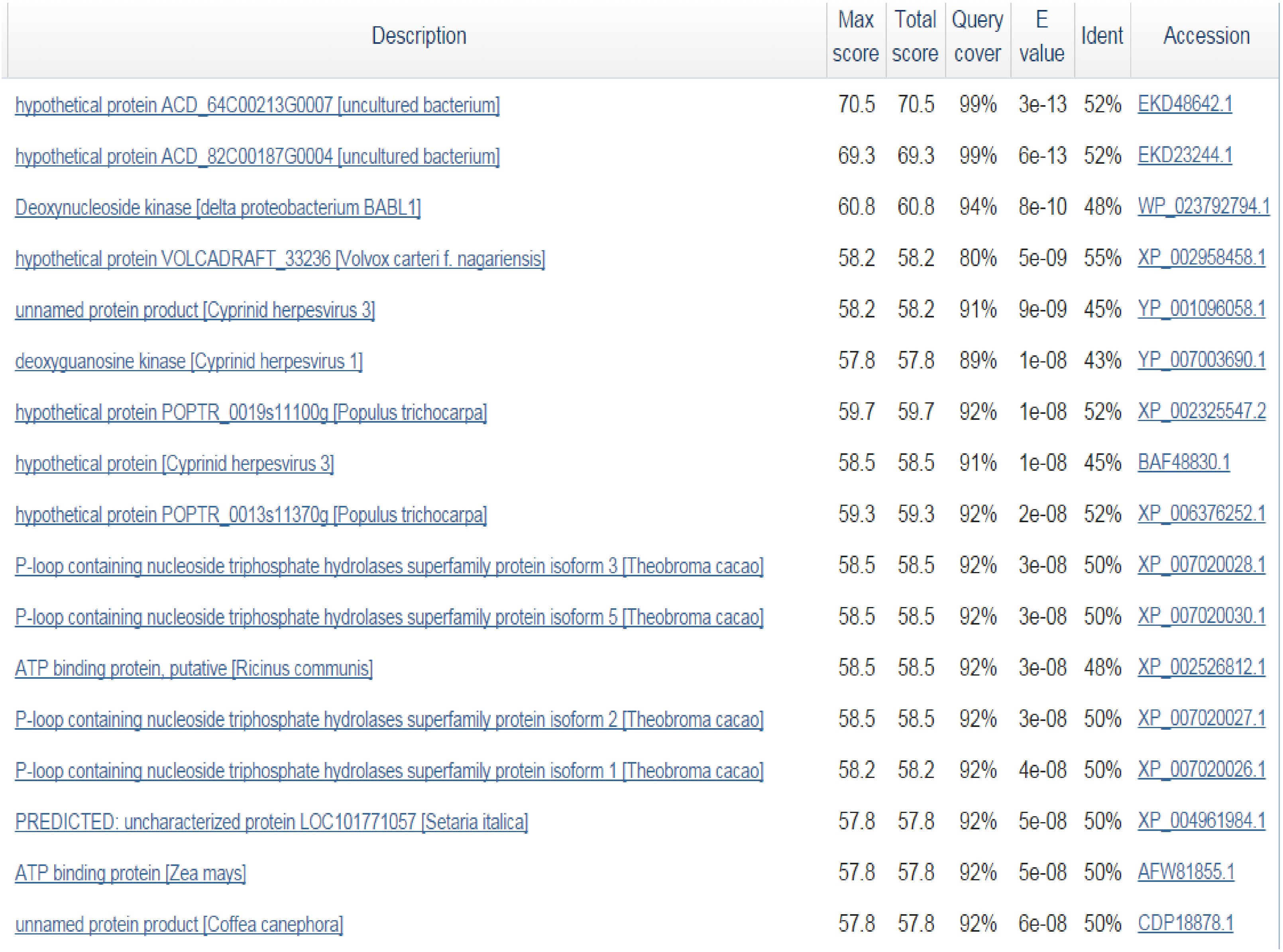


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C5052190 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
hypothetical protein ACD_64C00213G0007 [uncultured bacterium]	70.5	70.5	99%	3e-13	52%	EKD48642.1
hypothetical protein ACD_82C00187G0004 [uncultured bacterium]	69.3	69.3	99%	6e-13	52%	EKD23244.1
Deoxynucleoside kinase [delta proteobacterium BABL1]	60.8	60.8	94%	8e-10	48%	WP_023792794.1
hypothetical protein VOLCADRAFT_33236 [Volvox carteri f. nagariensis]	58.2	58.2	80%	5e-09	55%	XP_002958458.1
unnamed protein product [Cyprinid herpesvirus 3]	58.2	58.2	91%	9e-09	45%	YP_001096058.1
deoxyguanosine kinase [Cyprinid herpesvirus 1]	57.8	57.8	89%	1e-08	43%	YP_007003690.1
hypothetical protein POPTR_0019s11100g [Populus trichocarpa]	59.7	59.7	92%	1e-08	52%	XP_002325547.2
hypothetical protein [Cyprinid herpesvirus 3]	58.5	58.5	91%	1e-08	45%	BAF48830.1
hypothetical protein POPTR_0013s11370g [Populus trichocarpa]	59.3	59.3	92%	2e-08	52%	XP_006376252.1
P-loop containing nucleoside triphosphate hydrolases superfamily protein isoform 3 [Theobroma cacao]	58.5	58.5	92%	3e-08	50%	XP_007020028.1
P-loop containing nucleoside triphosphate hydrolases superfamily protein isoform 5 [Theobroma cacao]	58.5	58.5	92%	3e-08	50%	XP_007020030.1
ATP binding protein, putative [Ricinus communis]	58.5	58.5	92%	3e-08	48%	XP_002526812.1
P-loop containing nucleoside triphosphate hydrolases superfamily protein isoform 2 [Theobroma cacao]	58.5	58.5	92%	3e-08	50%	XP_007020027.1
P-loop containing nucleoside triphosphate hydrolases superfamily protein isoform 1 [Theobroma cacao]	58.2	58.2	92%	4e-08	50%	XP_007020026.1
PREDICTED: uncharacterized protein LOC101771057 [Setaria italica]	57.8	57.8	92%	5e-08	50%	XP_004961984.1
ATP binding protein [Zea mays]	57.8	57.8	92%	5e-08	50%	AFW81855.1
unnamed protein product [Coffea canephora]	57.8	57.8	92%	6e-08	50%	CDP18878.1

3.1.10. BLASTn alignment for contig C8227460 (756 BP)

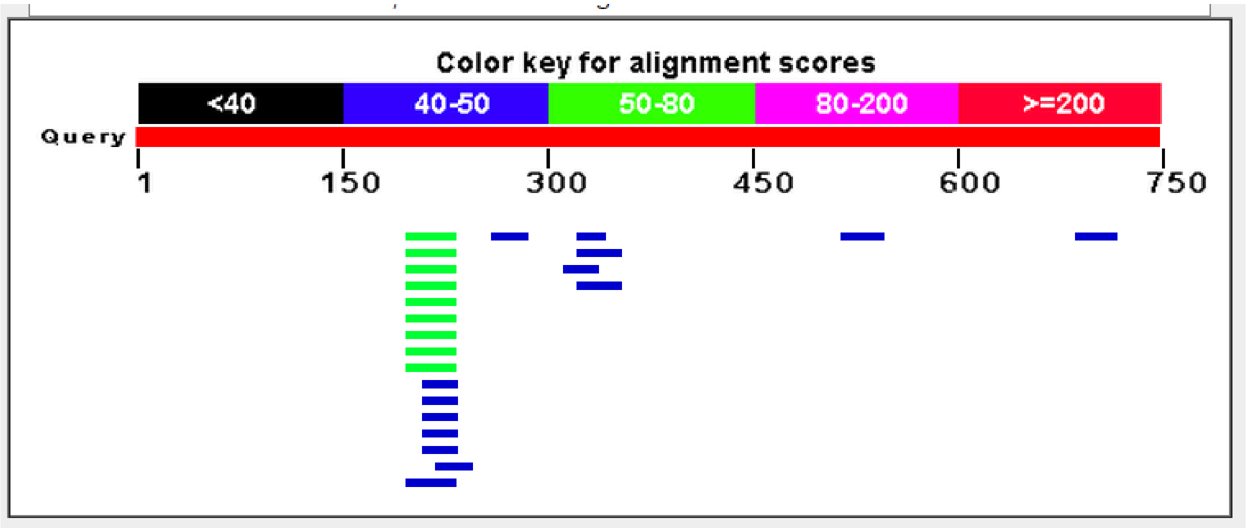
Nucleotide sequence in FASTA format

```
>C8227460
GGGTTCTAGCCAGTGTCAGAGCGCCGCTTGTGGCAATTGTGCCGTATGCCGTATTCG
TTAGCGCGCGATTTAACGCTACGTTCTCCGCGCCTACTTGACTGCCGACTTCAATAA
TCGCCTTCACTACTGCCAAAGCAATTGAATCGGTGGTGGCCTGATTACCAAGTGCCG
ACAATGCCTTCGTTACGGCTAGTATCAGCGCTTCATTGGTTGCTTGATTGCCTGCTTC
GCTGATTGCCTTAACAATTGCCAAAGCCAGCGTTTCATTGCCTGCCTGAGTGCTTCC
TGCTGATGCACCGCGTATCATACCTAGCGTCGCATTAGCCAGCGCCAGCAGCGTGT
AGGCGTTGGCTACACCTAACGTTCGTCCTAGTGTCGCATTGCCTAGCATGGCGGCTG
TATTGCCTACCGTAGCCGACCGTGTAAGCGCTATGGTTGTGCCCTCTGATCCGGCAA
GCGCGTTACCTTGTGTAATTCCTAGCGTTTTCGCTAGTGTTGTATTGCCAGGAACAG
TCGCAATGTTCGCAGCGGTTGCTCCCTCGGTTCTGGCTAGTGTGTTGGCGTCTGTGG
TTGTTAATGTGCTATTTGGTGTAGCGGCGACTGTACGCGCAATAGAAACAGCCCCCA
CACCCGCAAGTATTTCGACTTGGGTTACTCCTGATGTTCTCGTTAACGTCAGGCTAC
CTACCGCGTTTGCAATTTGACTGGATACAACACCTTCGGTGCGGGCTAAAATCTGAC
TCTCGGCAAAAGTTG
```

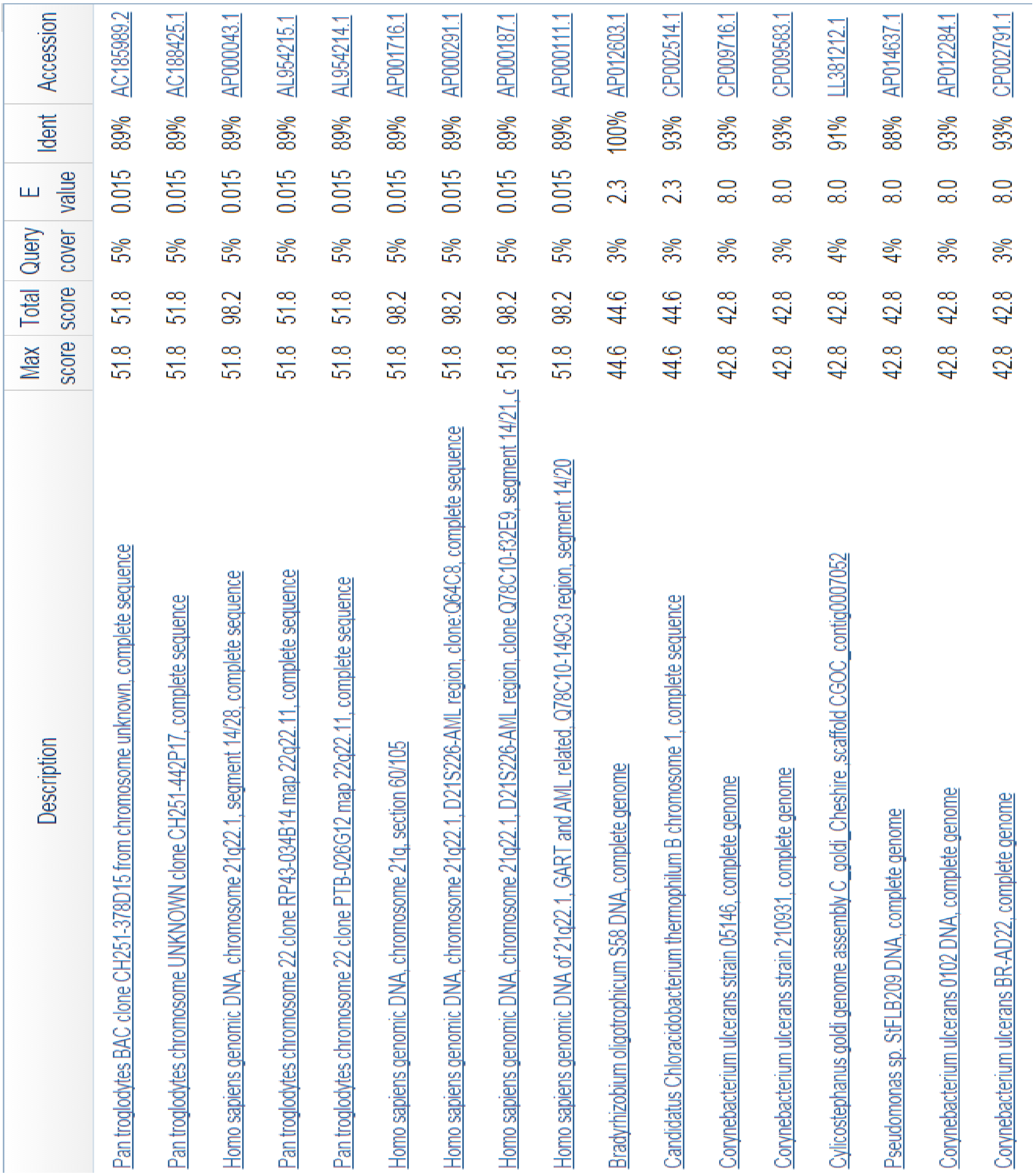
BLASTn alignment for the contig C8227460

BLASTn search result showed that the assembled query sequence was ambiguous and the query sequence was similar to the sequences from human, bacteria or chimpanzee. E value of the best hit was 0.015 for *Pan troglodytes* (chimpanzee) BAC clone sequence. The E value higher than zero implied the insignificant alignment with reference sequences.

A) Graphic display



B) Hit list of aligned NCBI sequences for C8227460 after BLASTn search

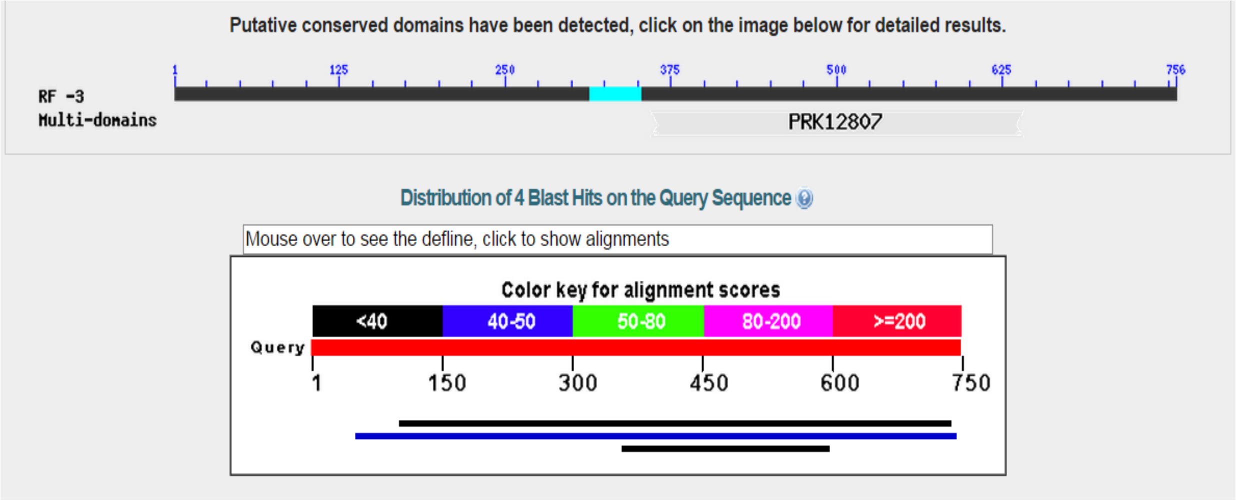


Description	Max score	Total score	Query cover	E value	Ident	Accession
Pan troglodytes BAC clone CH251-378D15 from chromosome unknown, complete sequence	51.8	51.8	5%	0.015	89%	AC185989.2
Pan troglodytes chromosome UNKNOWN clone CH251-442P17, complete sequence	51.8	51.8	5%	0.015	89%	AC188425.1
Homo sapiens genomic DNA, chromosome 21q22.1, segment 14/28, complete sequence	51.8	98.2	5%	0.015	89%	AP000043.1
Pan troglodytes chromosome 22 clone RP43-034B14 map 22q22.11, complete sequence	51.8	51.8	5%	0.015	89%	AL954215.1
Pan troglodytes chromosome 22 clone PTB-026G12 map 22q22.11, complete sequence	51.8	51.8	5%	0.015	89%	AL954214.1
Homo sapiens genomic DNA, chromosome 21q, section 60/105	51.8	98.2	5%	0.015	89%	AP001716.1
Homo sapiens genomic DNA, chromosome 21q22.1, D21S226-AML region, clone:Q64C8, complete sequence	51.8	98.2	5%	0.015	89%	AP000291.1
Homo sapiens genomic DNA, chromosome 21q22.1, D21S226-AML region, clone Q78C10-f32E9, segment 14/21, c	51.8	98.2	5%	0.015	89%	AP000187.1
Homo sapiens genomic DNA of 21q22.1, GART and AML related, Q78C10-149C3 region, segment 14/20	51.8	98.2	5%	0.015	89%	AP000111.1
Bradyrhizobium oligotrophicum S58 DNA, complete genome	44.6	44.6	3%	2.3	100%	AP012603.1
Candidatus Chloracidobacterium thermophilum B chromosome 1, complete sequence	44.6	44.6	3%	2.3	93%	CP002514.1
Corynebacterium ulcerans strain 05146, complete genome	42.8	42.8	3%	8.0	93%	CP009716.1
Corynebacterium ulcerans strain 210931, complete genome	42.8	42.8	3%	8.0	93%	CP009583.1
Cylicostephanus goldi genome assembly C_goldi_Cheshire ,scaffold CGOC_contig0007052	42.8	42.8	4%	8.0	91%	LL381212.1
Pseudomonas sp. StFLB209 DNA, complete genome	42.8	42.8	4%	8.0	88%	AP014637.1
Corynebacterium ulcerans 0102 DNA, complete genome	42.8	42.8	3%	8.0	93%	AP012284.1
Corynebacterium ulcerans BR-AD22, complete genome	42.8	42.8	3%	8.0	93%	CP002791.1

BLASTx alignment for the contig C8227460

BLASTx search showed the assembled query sequence was similar to the sequence of S*treptococcus* species with the E value of 0.003. This contig could belong to an unknown species as after BLASTn and BLASTx search the contig showed only partial alignment and less similarity with insignificant E-value to the aligned sequences of NCBI database

A) Graphic display

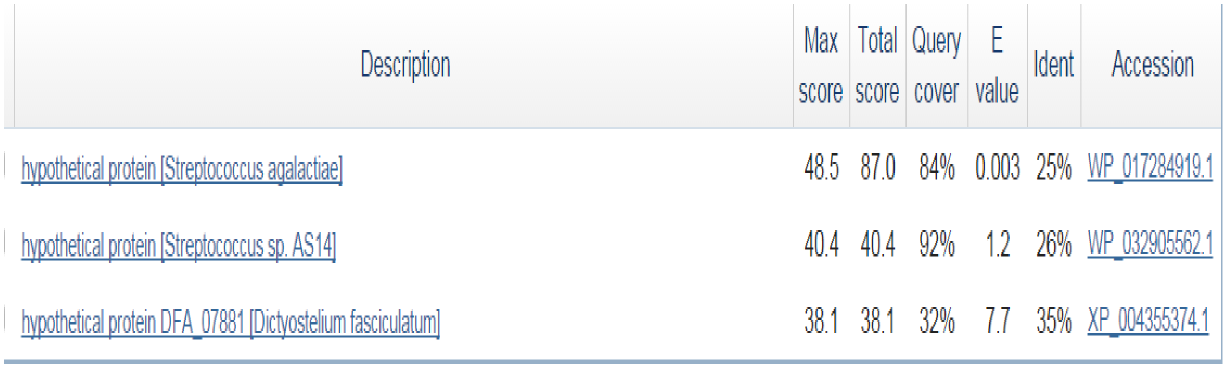


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 4 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C8227460 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
hypothetical protein [Streptococcus agalactiae]	48.5	87.0	84%	0.003	25%	WP_017284919.1
hypothetical protein [Streptococcus sp. AS14]	40.4	40.4	92%	1.2	26%	WP_032905562.1
hypothetical protein DFA_07881 [Dictyostelium fasciculatum]	38.1	38.1	32%	7.7	35%	XP_004355374.1

3.1.11. BLAST alignment for the contig C7779111 (378 BP)

Nucleotide sequence in FASTA format

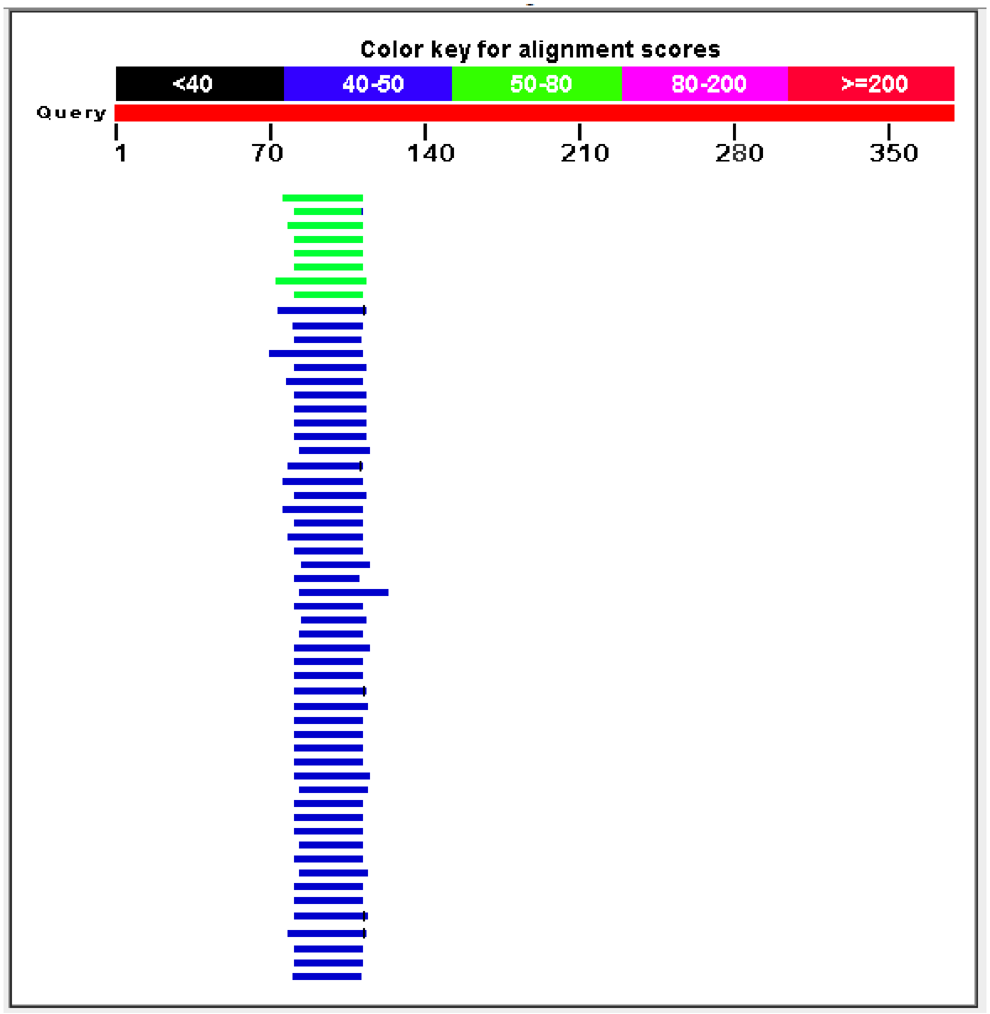
```
>C7779111
CGAAGCTGGCGTGTCTTACAGCAAGCTGAACAATGCTGGAATGCGCGCTGCACTGG
TTGCGCATTACGCCAAGTCCGAAGCGGTTGAAGCTGAAGTTGAAGCTGAAGTTGAA
GAGACCCCGACGTCCAACGGCATGTCCTTCGCTCAGATGCTTGGCCTGACACCTGTT
CCCGCTCCTGCCAATGTTGGTAACGCGACCAGCGTTGTTGATGGTAAGCGGGTTGA
AGCTAAGGCATCGAAAGGGAAGGGTGAAGCTCGCACCCGTTCTGAGAAGCCTGCTG
CCCCTGCTGTGCCCCGCGTCTCGCGCAAGGGTTACACGATCCAGAAGGAACGTGAA
GAGCGCAACGGTGTGAAGCGTCCGTCCGAAGGCACTATCTG
```

BLASTn alignment for the contig C7779111

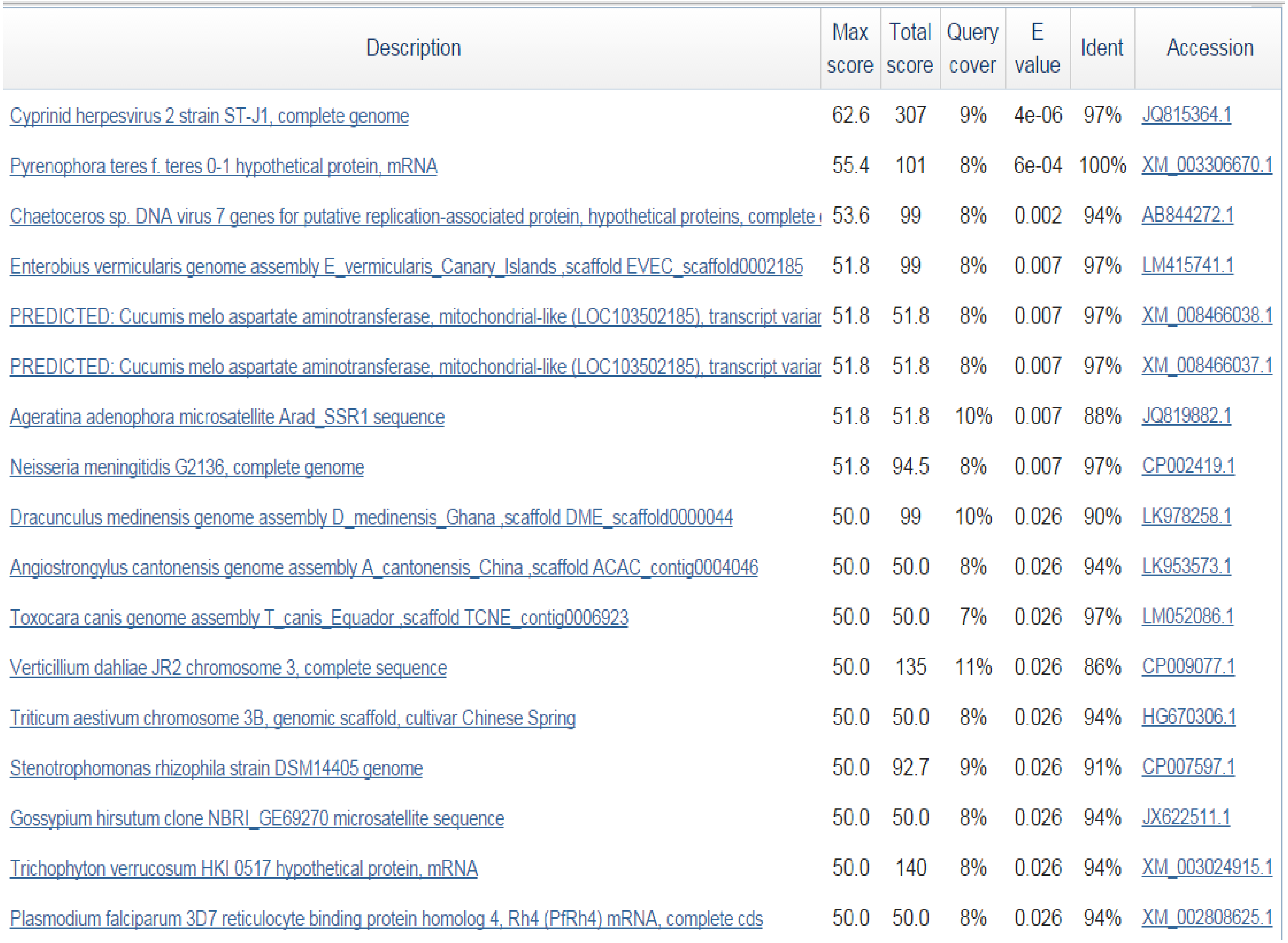
BLASTn search indicated that the query sequence was aligned with Cyprinid herpesvirus 2 as top hit. E value of the top hit was 4e-06 with 9% of total coverage (27 bp matched), the identity of 97% was found with the matched region.

Fragments of contig C7779111 (1-70 base pairs position), showed similarity with sequence of *Agrobacterium tumefaciens* with E-value of 1.5 after BLASTn and other fragment of the same contig (120-378 base pairs position) showed similarity to a sequence from fungi (E-value of 0.72) after BLASTn search. BLASTx search showed the sequence was similar to the sequence from bacteria with E-value of 0.58

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C7779111 after BLASTn search



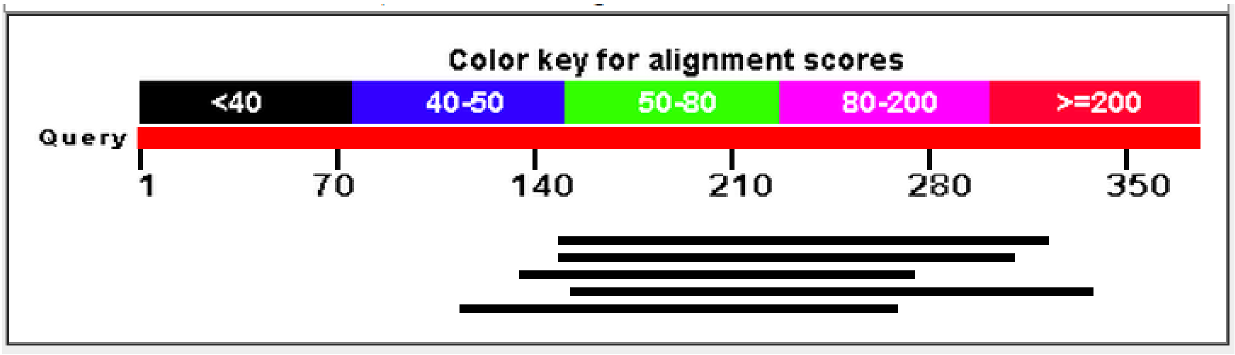
Description	Max score	Total score	Query cover	E value	Ident	Accession
Cyprinid herpesvirus 2 strain ST-J1, complete genome	62.6	307	9%	4e-06	97%	JQ815364.1
Pyrenophora teres f. teres 0-1 hypothetical protein, mRNA	55.4	101	8%	6e-04	100%	XM_003306670.1
Chaetoceros sp. DNA virus 7 genes for putative replication-associated protein, hypothetical proteins, complete (	53.6	99	8%	0.002	94%	AB844272.1
Enterobius vermicularis genome assembly E_vermicularis_Canary_Islands ,scaffold EVEC_scaffold0002185	51.8	99	8%	0.007	97%	LM415741.1
PREDICTED: Cucumis melo aspartate aminotransferase, mitochondrial-like (LOC103502185), transcript variar	51.8	51.8	8%	0.007	97%	XM_008466038.1
PREDICTED: Cucumis melo aspartate aminotransferase, mitochondrial-like (LOC103502185), transcript variar	51.8	51.8	8%	0.007	97%	XM_008466037.1
Ageratina adenophora microsatellite Arad_SSR1 sequence	51.8	51.8	10%	0.007	88%	JQ819882.1
Neisseria meningitidis G2136, complete genome	51.8	94.5	8%	0.007	97%	CP002419.1
Dracunculus medinensis genome assembly D_medinensis_Ghana ,scaffold DME_scaffold0000044	50.0	99	10%	0.026	90%	LK978258.1
Angiostrongylus cantonensis genome assembly A_cantonensis_China ,scaffold ACAC_contig0004046	50.0	50.0	8%	0.026	94%	LK953573.1
Toxocara canis genome assembly T_canis_Equador ,scaffold TCNE_contig0006923	50.0	50.0	7%	0.026	97%	LM052086.1
Verticillium dahliae JR2 chromosome 3, complete sequence	50.0	135	11%	0.026	86%	CP009077.1
Triticum aestivum chromosome 3B, genomic scaffold, cultivar Chinese Spring	50.0	50.0	8%	0.026	94%	HG670306.1
Stenotrophomonas rhizophila strain DSM14405 genome	50.0	92.7	9%	0.026	91%	CP007597.1
Gossypium hirsutum clone NBRI_GE69270 microsatellite sequence	50.0	50.0	8%	0.026	94%	JX622511.1
Trichophyton verrucosum HKI 0517 hypothetical protein, mRNA	50.0	140	8%	0.026	94%	XM_003024915.1
Plasmodium falciparum 3D7 reticulocyte binding protein homolog 4, Rh4 (PfRh4) mRNA, complete cds	50.0	50.0	8%	0.026	94%	XM_002808625.1

BLASTx alignment for the contig C7779111

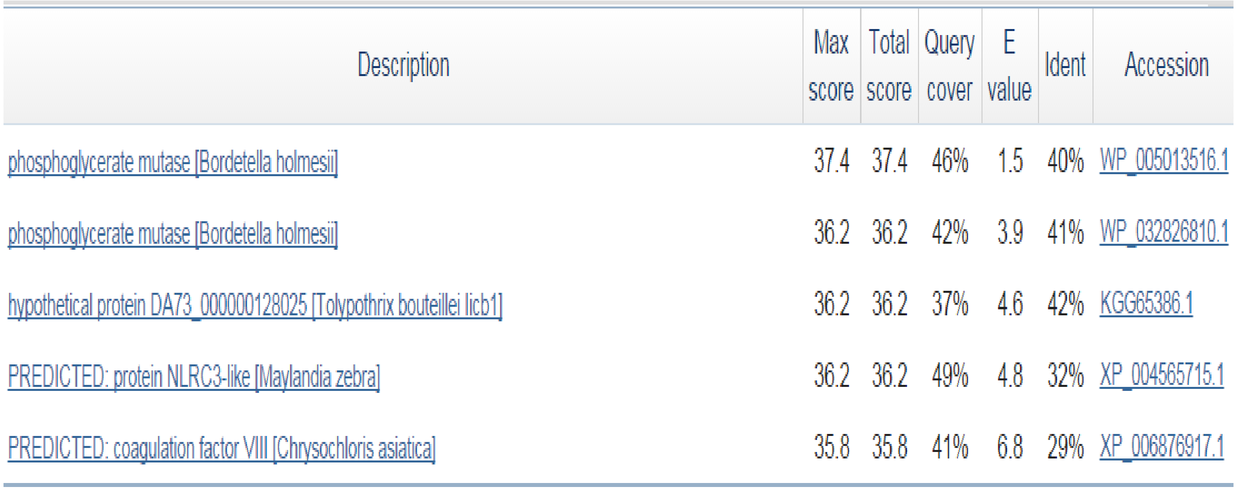
After BLASTx search, assembled query sequence was aligned with sequence from bacteria and algae with the E value ranging from 1.5-6.8. High E-value refers that the matched sequences from NCBI database were not similar enough to the query sequence.

This contig could belong to an unknown species as after BLASTn and BLASTx search the contig showed only partial alignment and less similarity with insignificant E-value to the aligned sequences of NCBI database.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C7779111 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
phosphoglycerate mutase [Bordetella holmesii]	37.4	37.4	46%	1.5	40%	WP_005013516.1
phosphoglycerate mutase [Bordetella holmesii]	36.2	36.2	42%	3.9	41%	WP_032826810.1
hypothetical protein DA73_000000128025 [Tolypothrix bouteillei licb1]	36.2	36.2	37%	4.6	42%	KGG65386.1
PREDICTED: protein NLRC3-like [Maylandia zebra]	36.2	36.2	49%	4.8	32%	XP_004565715.1
PREDICTED: coagulation factor VIII [Chrysochloris asiatica]	35.8	35.8	41%	6.8	29%	XP_006876917.1

3.1.12. BLAST alignment for the contig C7843866 (396 BP)

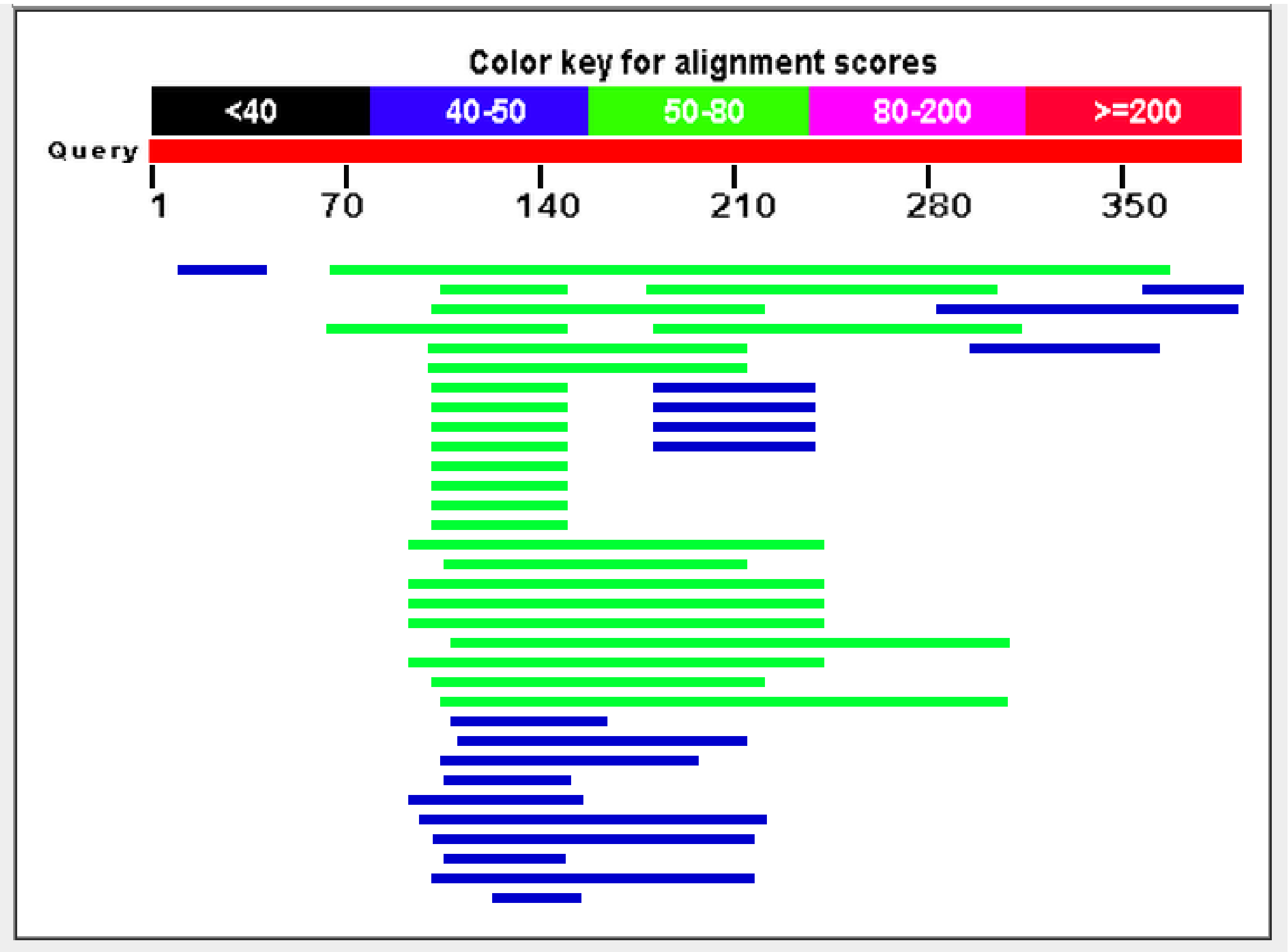
Nucleotide sequence in FASTA format

```
>C7843866
ACGCCTGCTGCACGCTGGCCAGCATCAACTTGCGCAAGTTCCTAGTACCCACTGCAA
CGGGCTATGAGATCGACCACGAGAAACTGCATGACTGCGTTCGCATGATCACCCGC
AATCTGGACATGATCATTGACGTTAATCACTACCCGGTGCCCGAATGCGCGCAGAA
CTCATATGACTACCGCCCAATCGGCATTGGAATACAAGCCCTGGCCGATGTATTTGC
AATCATGCGAATCCCATTCCTGTCCCCGGAGGCCGCGCGGATTGACATAGAAATCG
CAGAGACAATCTATCACGCAGCCATCACCGAGTCGGCCGCGCGAGCGCAAGTTCAT
GGCGCATACAAGGGCTTTGAGGGCTCGCCCGCCAGCCGCGGGTTGTTCCAGTTCGA
CC
```

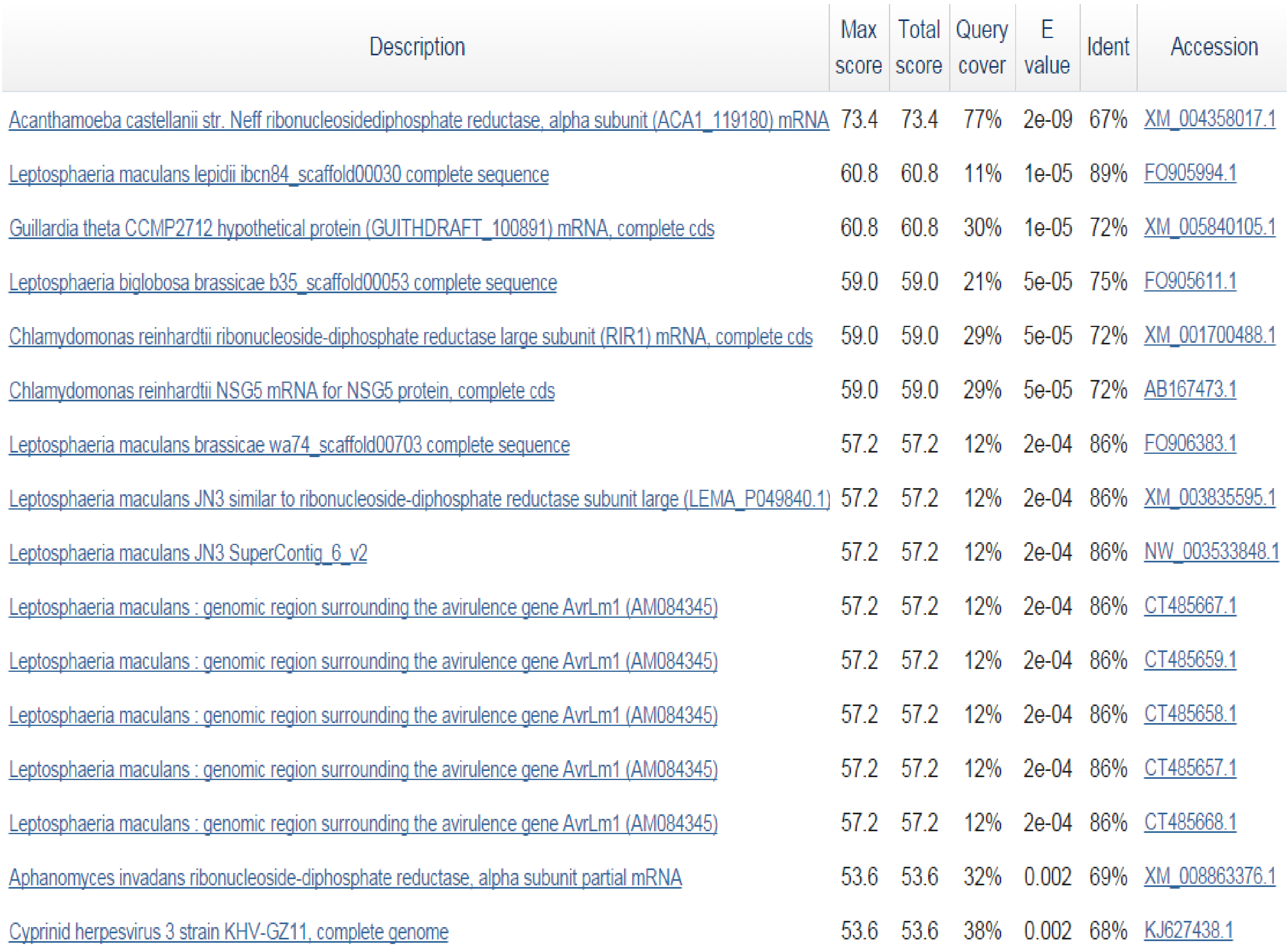
BLASTn alignment for the contig C7843866

BLASTn search result showed the query sequence was ambiguous in nature as the assembled query sequence was similar to the sequence from amoeba, fungi or algae. E value of best matched hit was 2e-09 for amoeba sequence. The query coverage was 77% and the identity of the aligned sequences was 67%

A) Graphic display



B) Hit list of aligned NCBI sequences for C7843866 after BLASTn search

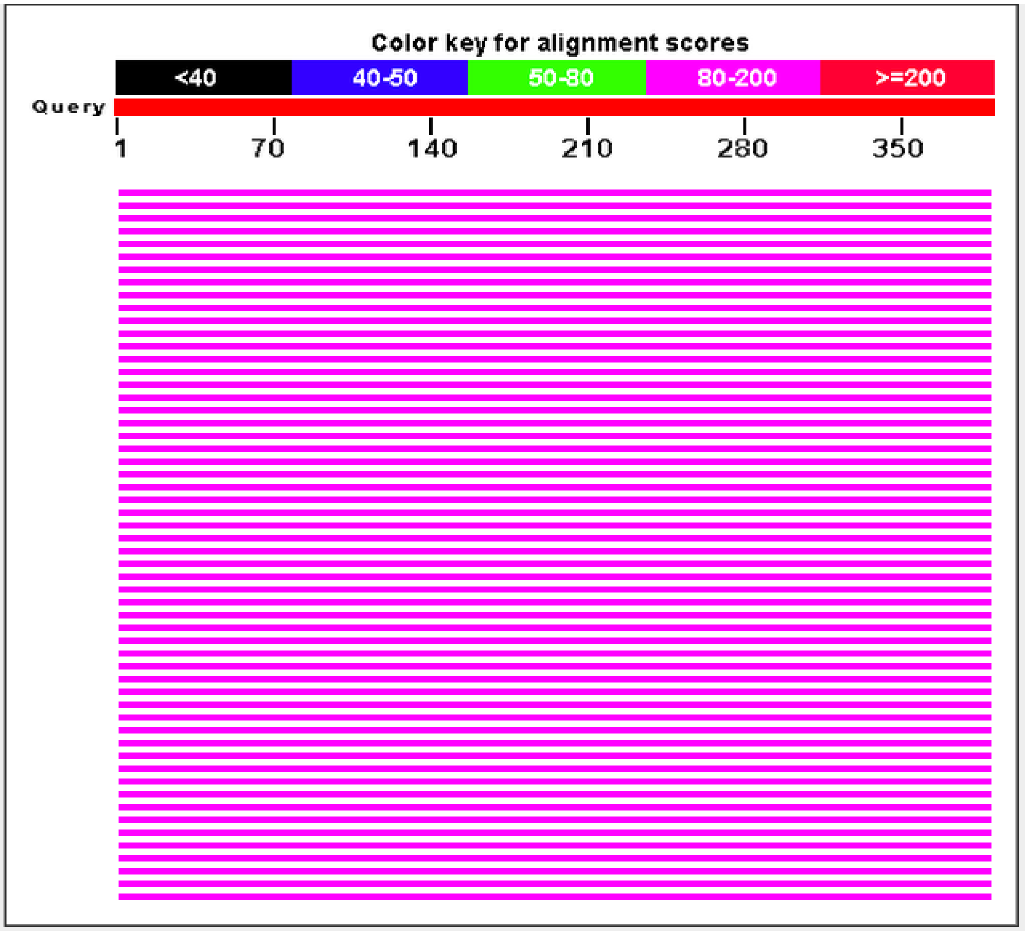


Description	Max score	Total score	Query cover	E value	Ident	Accession
Acanthamoeba castellanii str. Neff ribonucleosidediphosphate reductase, alpha subunit (ACA1_119180) mRNA	73.4	73.4	77%	2e-09	67%	XM_004358017.1
Leptosphaeria maculans lepidii ibcn84_scaffold00030 complete sequence	60.8	60.8	11%	1e-05	89%	FO905994.1
Guillardia theta CCMP2712 hypothetical protein (GUITHDRAFT_100891) mRNA, complete cds	60.8	60.8	30%	1e-05	72%	XM_005840105.1
Leptosphaeria biglobosa brassicae b35_scaffold00053 complete sequence	59.0	59.0	21%	5e-05	75%	FO905611.1
Chlamydomonas reinhardtii ribonucleoside-diphosphate reductase large subunit (RIR1) mRNA, complete cds	59.0	59.0	29%	5e-05	72%	XM_001700488.1
Chlamydomonas reinhardtii NSG5 mRNA for NSG5 protein, complete cds	59.0	59.0	29%	5e-05	72%	AB167473.1
Leptosphaeria maculans brassicae wa74_scaffold00703 complete sequence	57.2	57.2	12%	2e-04	86%	FO906383.1
Leptosphaeria maculans JN3 similar to ribonucleoside-diphosphate reductase subunit large (LEMA_P049840.1)	57.2	57.2	12%	2e-04	86%	XM_003835595.1
Leptosphaeria maculans JN3 SuperContig_6_v2	57.2	57.2	12%	2e-04	86%	NW_003533848.1
Leptosphaeria maculans : genomic region surrounding the avirulence gene AvrLm1 (AM084345)	57.2	57.2	12%	2e-04	86%	CT485667.1
Leptosphaeria maculans : genomic region surrounding the avirulence gene AvrLm1 (AM084345)	57.2	57.2	12%	2e-04	86%	CT485659.1
Leptosphaeria maculans : genomic region surrounding the avirulence gene AvrLm1 (AM084345)	57.2	57.2	12%	2e-04	86%	CT485658.1
Leptosphaeria maculans : genomic region surrounding the avirulence gene AvrLm1 (AM084345)	57.2	57.2	12%	2e-04	86%	CT485657.1
Leptosphaeria maculans : genomic region surrounding the avirulence gene AvrLm1 (AM084345)	57.2	57.2	12%	2e-04	86%	CT485668.1
Aphanomyces invadans ribonucleoside-diphosphate reductase, alpha subunit partial mRNA	53.6	53.6	32%	0.002	69%	XM_008863376.1
Cyprinid herpesvirus 3 strain KHV-GZ11, complete genome	53.6	53.6	38%	0.002	68%	KJ627438.1

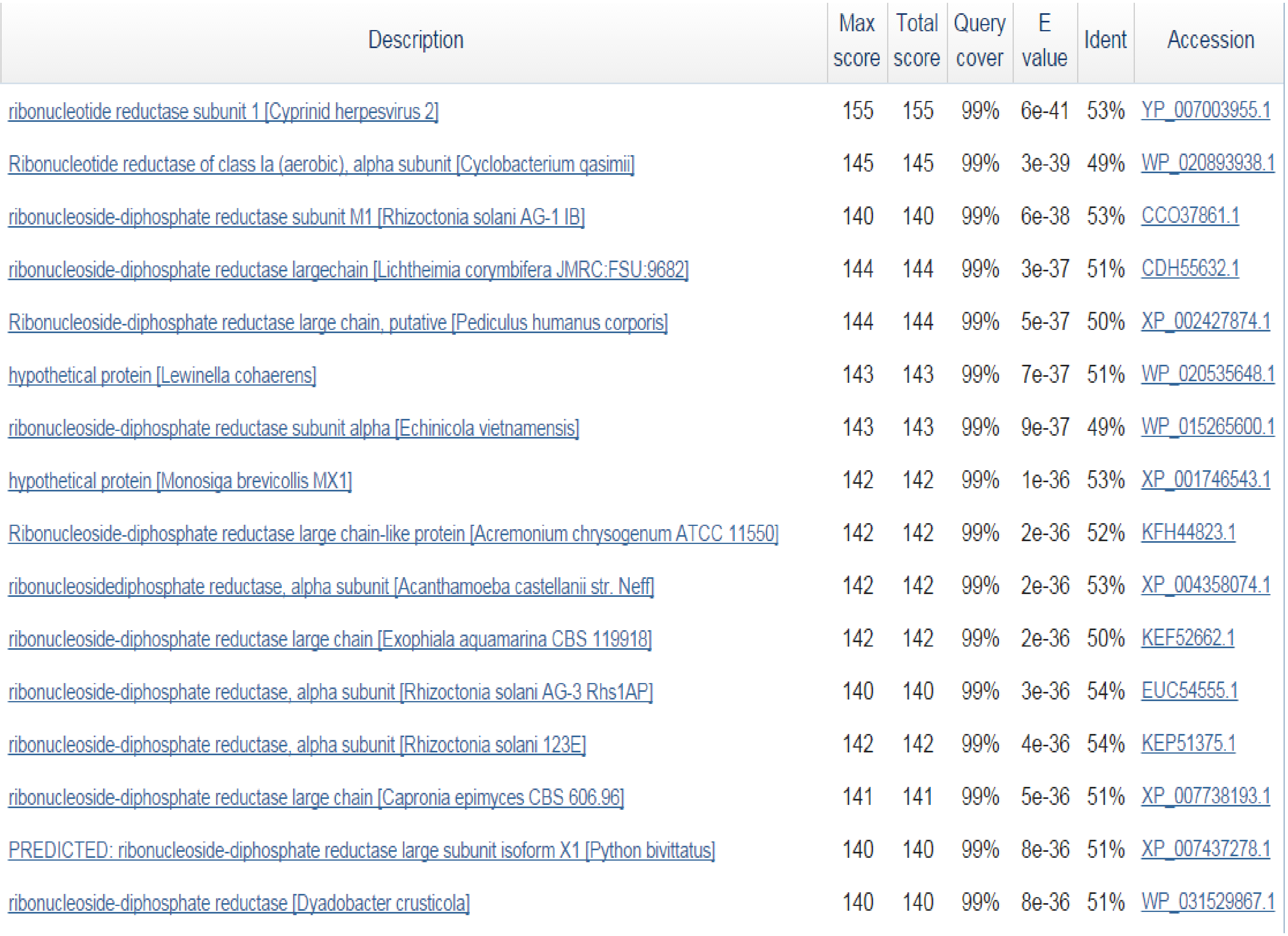
BLASTx alignment for the contig C7843866

After BLASTx search, assembled query sequence was aligned with the conserved ribonucleotide reductase superfamily. The query sequence identity to top hit cyprinid herpesvirus 2 (fish virus) was 53%, matched length was 71 amino acid out of total 133 amino acids, query coverage was 99% at E value of 6e-41.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C7843866 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
ribonucleotide reductase subunit 1 [Cyprinid herpesvirus 2]	155	155	99%	6e-41	53%	YP_007003955.1
Ribonucleotide reductase of class Ia (aerobic), alpha subunit [Cyclobacterium qasimii]	145	145	99%	3e-39	49%	WP_020893938.1
ribonucleoside-diphosphate reductase subunit M1 [Rhizoctonia solani AG-1 IB]	140	140	99%	6e-38	53%	CCO37861.1
ribonucleoside-diphosphate reductase largechain [Lichtheimia corymbifera JMRC:FSU:9682]	144	144	99%	3e-37	51%	CDH55632.1
Ribonucleoside-diphosphate reductase large chain, putative [Pediculus humanus corporis]	144	144	99%	5e-37	50%	XP_002427874.1
hypothetical protein [Lewinella cohaerens]	143	143	99%	7e-37	51%	WP_020535648.1
ribonucleoside-diphosphate reductase subunit alpha [Echinicola vietnamensis]	143	143	99%	9e-37	49%	WP_015265600.1
hypothetical protein [Monosiga brevicollis MX1]	142	142	99%	1e-36	53%	XP_001746543.1
Ribonucleoside-diphosphate reductase large chain-like protein [Acremonium chrysogenum ATCC 11550]	142	142	99%	2e-36	52%	KFH44823.1
ribonucleosidediphosphate reductase, alpha subunit [Acanthamoeba castellanii str. Neff]	142	142	99%	2e-36	53%	XP_004358074.1
ribonucleoside-diphosphate reductase large chain [Exophiala aquamarina CBS 119918]	142	142	99%	2e-36	50%	KEF52662.1
ribonucleoside-diphosphate reductase, alpha subunit [Rhizoctonia solani AG-3 Rhs1AP]	140	140	99%	3e-36	54%	EUC54555.1
ribonucleoside-diphosphate reductase, alpha subunit [Rhizoctonia solani 123E]	142	142	99%	4e-36	54%	KEP51375.1
ribonucleoside-diphosphate reductase large chain [Capronia epimyces CBS 606.96]	141	141	99%	5e-36	51%	XP_007738193.1
PREDICTED: ribonucleoside-diphosphate reductase large subunit isoform X1 [Python bivittatus]	140	140	99%	8e-36	51%	XP_007437278.1
ribonucleoside-diphosphate reductase [Dyadobacter crusticola]	140	140	99%	8e-36	51%	WP_031529867.1

3.2. Alignment of contigs/scaffolds similar to poxvirus sequences with sequences of NCBI database

3.2.1. BLAST alignment for the contig C4880850 (164 BP)

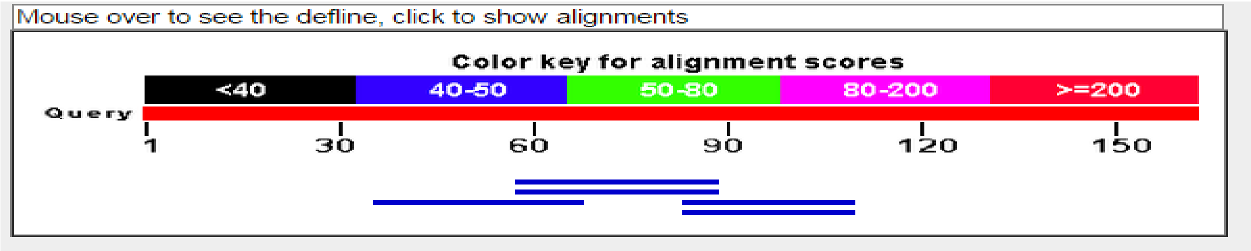
Nucleotide sequence in FASTA format

```
>C4880850
GTTACTACATTGTTAGCCATTGGGGTCGTTGCCGGCCTTGTTCGTATATAGTAAGAC
CCAGTCTTAAGTCCTAACTCCCAGCCCTTTTTCATTACCCCGCGTAAATACTTACTTG
AGTTGTCGCGGAGATATATATTTAGGCTTTGTGATTGGTCAACAAAGGC
```

BLASTn alignment for the contig C4880850

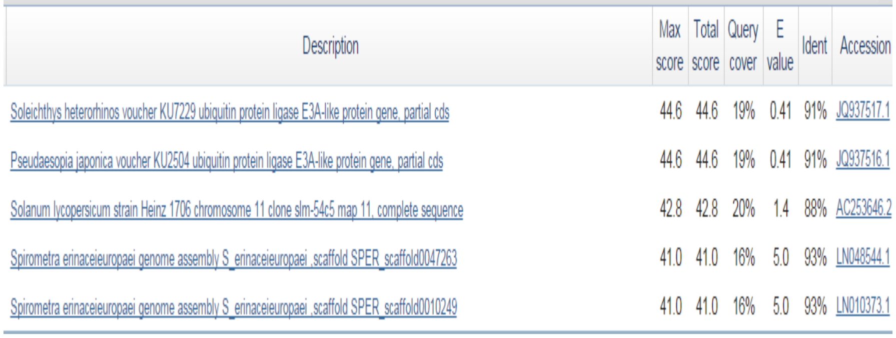
BLASTn search result showed that assembled query sequence was similar to sequences from fish and other species that belonged to phylum Chordata. For the top hit (fish), total coverage was 19% (32 bp matched), E value was 0.41 and the identity was 91% with the matched region.

A) Graphic display



Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
30
60
90
120
150

B) Hit list of aligned NCBI sequences for C4880850 after BLASTn search

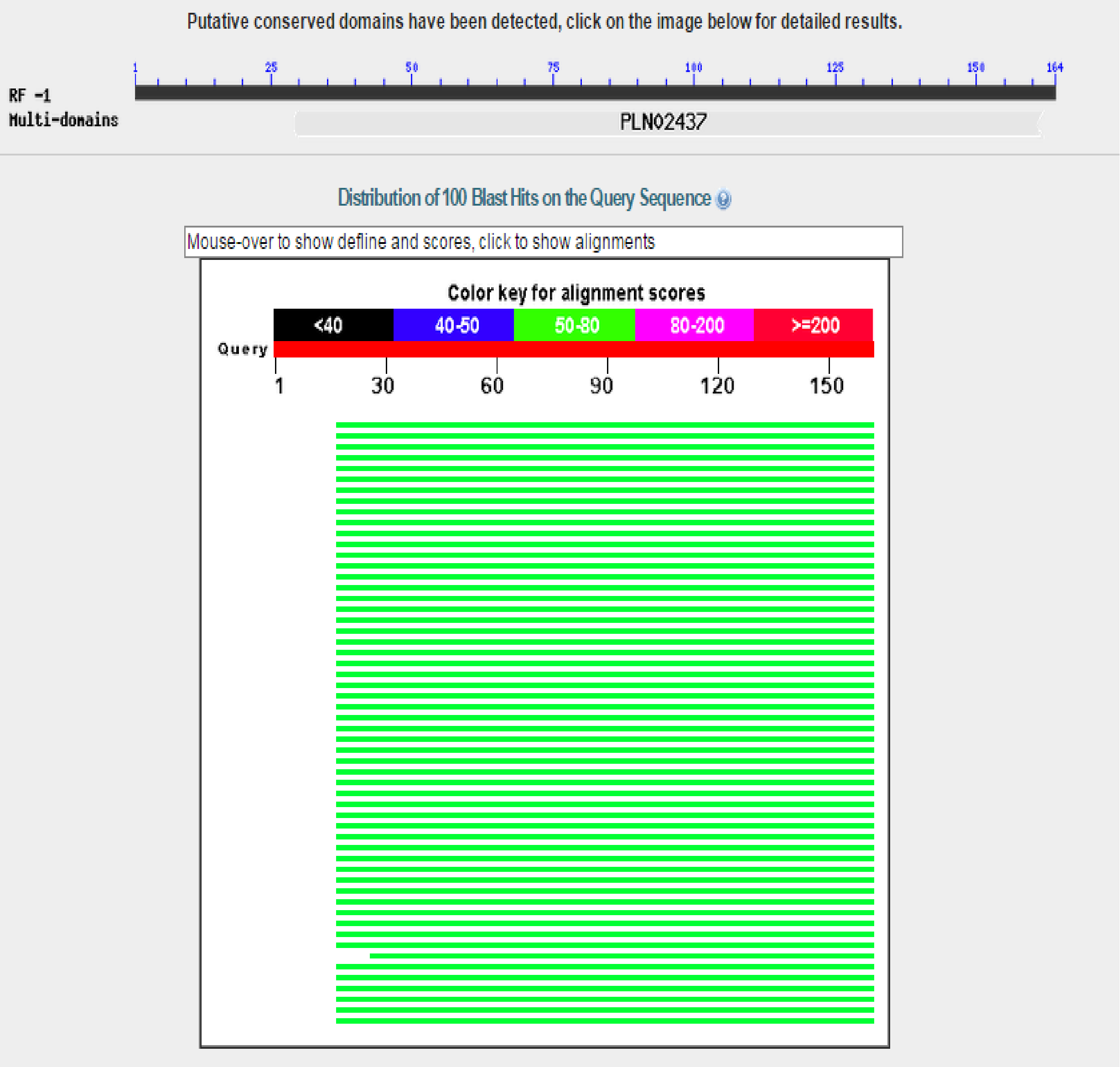


Description	Max score	Total score	Query cover	E value	Ident	Accession
Soleichthys heterorhinos voucher KU7229 ubiquitin protein ligase E3A-like protein gene, partial cds	44.6	44.6	19%	0.41	91%	JQ937517.1
Pseudaesopia japonica voucher KU2504 ubiquitin protein ligase E3A-like protein gene, partial cds	44.6	44.6	19%	0.41	91%	JQ937516.1
Solanum lycopersicum strain Heinz 1706 chromosome 11 clone slm-54c5 map 11, complete sequence	42.8	42.8	20%	1.4	88%	AC253646.2
Spirometra erinaceieuropaei genome assembly S_erinaceieuropaei ,scaffold SPER_scaffold0047263	41.0	41.0	16%	5.0	93%	LN048544.1
Spirometra erinaceieuropaei genome assembly S_erinaceieuropaei ,scaffold SPER_scaffold0010249	41.0	41.0	16%	5.0	93%	LN010373.1

BLASTx alignment for the contig C4880850

BLASTx search result showed that the assembled query sequence was aligned with the conserved ribonucleoside-diphosphate reductase large subunit superfamily. For the query sequence, identity to top hit variola virus was 57%, matched length was 28 amino acid out of total 46 amino acids, and query coverage was 89% at E value of 5e-10.

A) Graphic display

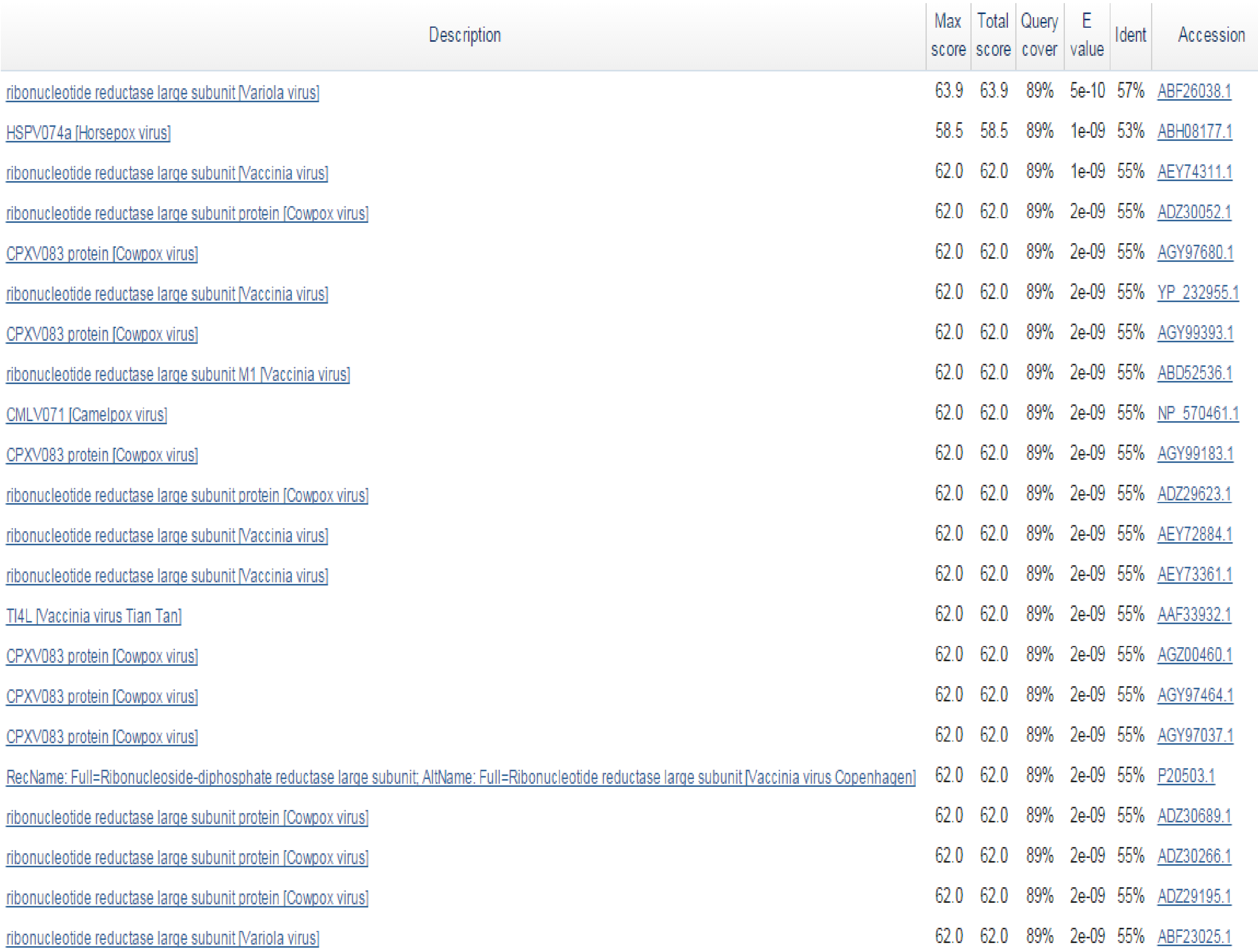


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C4880850 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
ribonucleotide reductase large subunit [Variola virus]	63.9	63.9	89%	5e-10	57%	ABF26038.1
HSPV074a [Horsepox virus]	58.5	58.5	89%	1e-09	53%	ABH08177.1
ribonucleotide reductase large subunit [Vaccinia virus]	62.0	62.0	89%	1e-09	55%	AEY74311.1
ribonucleotide reductase large subunit protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	ADZ30052.1
CPXV083 protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	AGY97680.1
ribonucleotide reductase large subunit [Vaccinia virus]	62.0	62.0	89%	2e-09	55%	YP_232955.1
CPXV083 protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	AGY99393.1
ribonucleotide reductase large subunit M1 [Vaccinia virus]	62.0	62.0	89%	2e-09	55%	ABD52536.1
CMLV071 [Camelpox virus]	62.0	62.0	89%	2e-09	55%	NP_570461.1
CPXV083 protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	AGY99183.1
ribonucleotide reductase large subunit protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	ADZ29623.1
ribonucleotide reductase large subunit [Vaccinia virus]	62.0	62.0	89%	2e-09	55%	AEY72884.1
ribonucleotide reductase large subunit [Vaccinia virus]	62.0	62.0	89%	2e-09	55%	AEY73361.1
TI4L [Vaccinia virus Tian Tan]	62.0	62.0	89%	2e-09	55%	AAF33932.1
CPXV083 protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	AGZ00460.1
CPXV083 protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	AGY97464.1
CPXV083 protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	AGY97037.1
RecName: Full=Ribonucleoside-diphosphate reductase large subunit; AltName: Full=Ribonucleotide reductase large subunit [Vaccinia virus Copenhagen]	62.0	62.0	89%	2e-09	55%	P20503.1
ribonucleotide reductase large subunit protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	ADZ30689.1
ribonucleotide reductase large subunit protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	ADZ30266.1
ribonucleotide reductase large subunit protein [Cowpox virus]	62.0	62.0	89%	2e-09	55%	ADZ29195.1
ribonucleotide reductase large subunit [Variola virus]	62.0	62.0	89%	2e-09	55%	ABF23025.1

3.2.2. BLAST alignment for the contig C8246727 (812 BP)

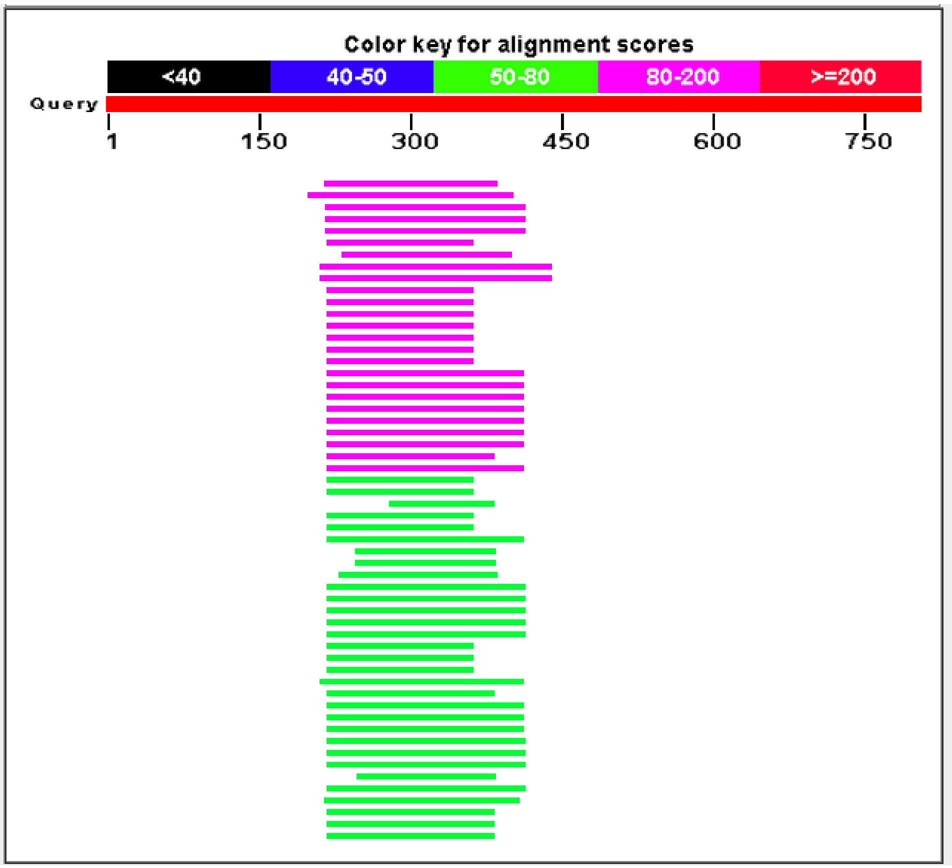
Nucleotide sequence in FASTA format

```
>C8246727
TGTTTGATCTTTATAATGATTAAATTCTGACATTTATATAATTTACCACTCAGATCAT
CAAATTTCAAGTATATTAGTCTTTCTTGTAGAGTTTCACCTCAGTGATGATACTATCA
ATATTAGCCATGGCTCCGTTACCAACGTAACCACAGAATAGTTCTTTCTCGATAGGA
TCAATCACACGGATTCCCCATGACCTGATCTTATCGATGTGCTCCTTGGTGATTGGA
TGCTCCCACATGCGAGTATTCATGGATGGTGCCACGATAAGAGGCTTGTCGAAGTTC
CATGCACGGGCACAGCATGTCAGAAGGTTATCACAGATCCCATTGGCTAATTTAGC
AAGGGTGTTGGCAGAAAGAGGAGCAATGACAAATATATCCGCCCATTTGATCAAGT
CGATATGGAGAACTGTGTCGTCATTTTTATATGCAGACCATTCATCAGAGTCCTCAA
AATCATGCGGATGTTCATAAGGAAAATTATACGAAACGCTTGATCCAAAATGCTTG
GAACTTTCAGTGTGAACACACTTAGTCTTCACCCAACCCATATCACTATACGCTTCT
CCAAATTTATATGCCAGTCGGGAGGCGACACTTCCTGTAAATCCATGTAATATTTTC
ATAAGTTTTTCATATCGTTCGGGGTCCATGCCAGAGGCAGTTTCCTATATTCGACCT
CGGGCATGTCCACATACACACCCACAAGTCCTTCATTCGTCAATTCCTCAGAGGTTC
GTGATTCCGTAGCCTTAGGTGGTCCCTCGACATATACTACACCGATGCGTCTAGGCC
ACAATCCTTTGATGG
```

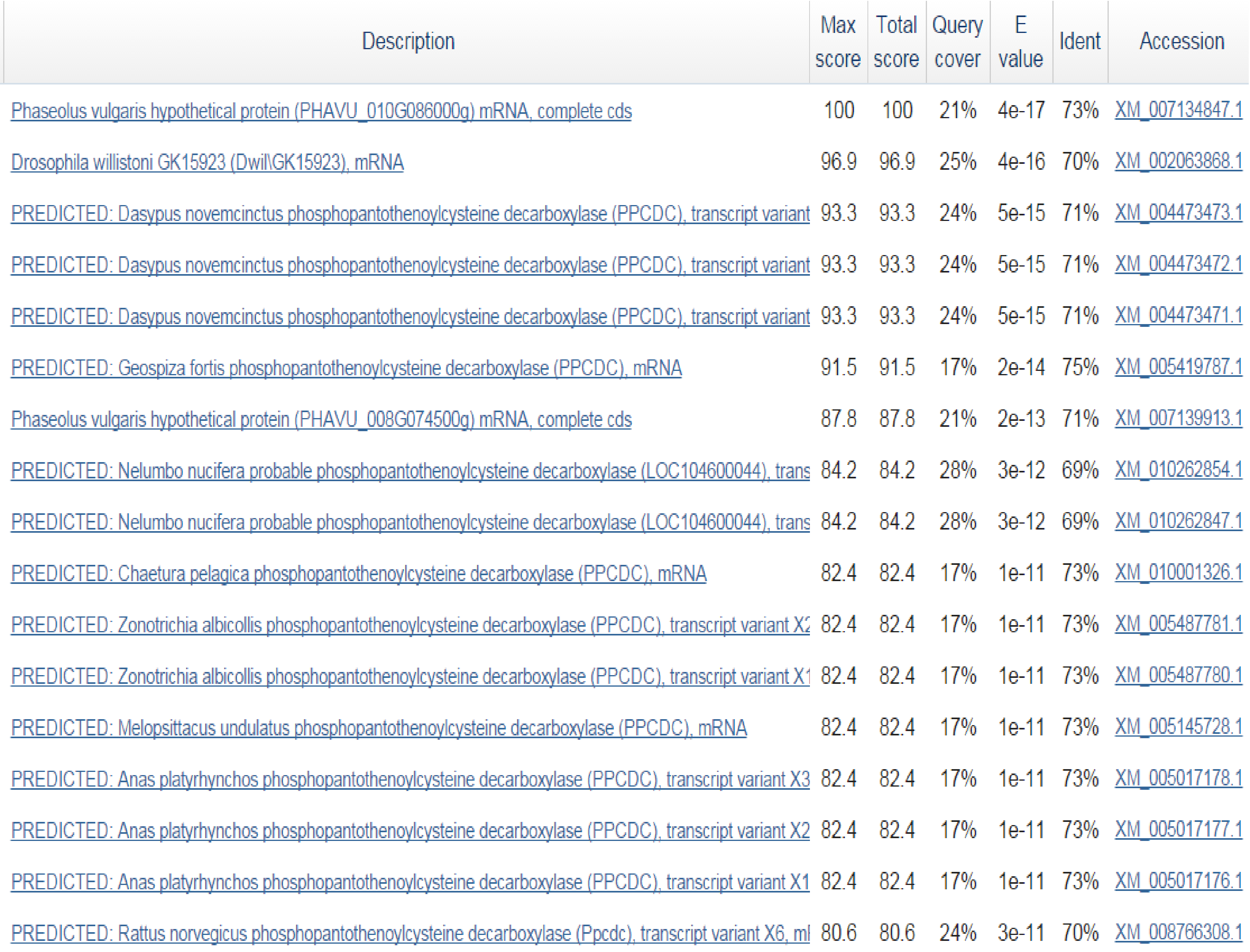
BLASTn alignment for the contig C8246727

BLASTn search result showed that the assembled query sequence was similar to sequences from fish and other species that belong to phylum Chordata. For the top hit, total coverage was 21% (175 bp matched), E value was 4e-17 and the identity was of 73% with the matched region. Low e-value implied that the query sequence may be originated from a fish.

A) Graphic display



B) Hit list of aligned NCBI sequences for C8246727 after BLASTn search

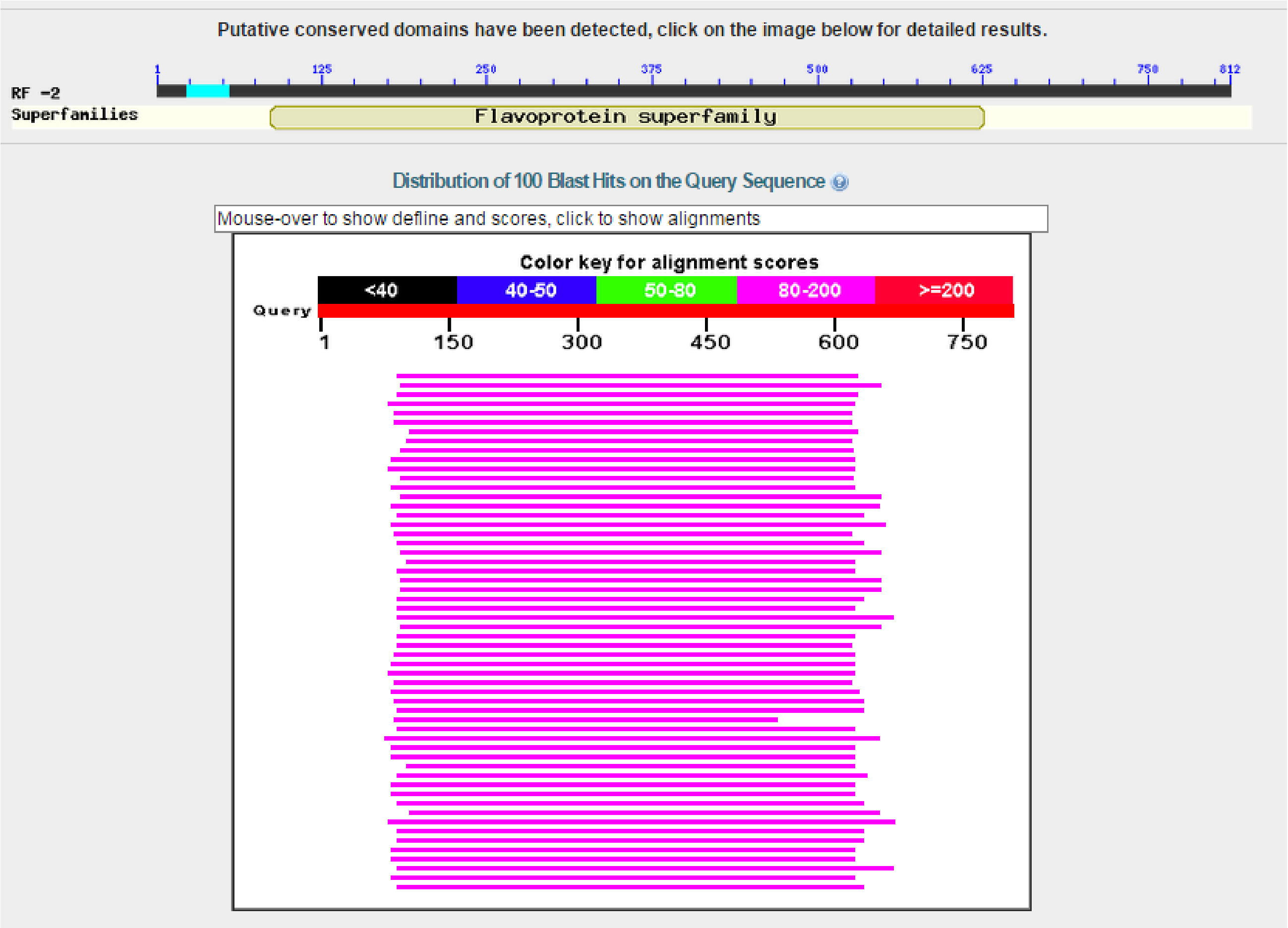


Description	Max score	Total score	Query cover	E value	Ident	Accession
Phaseolus vulgaris hypothetical protein (PHAVU_010G086000g) mRNA, complete cds	100	100	21%	4e-17	73%	XM_007134847.1
Drosophila willistoni GK15923 (Dwil\GK15923), mRNA	96.9	96.9	25%	4e-16	70%	XM_002063868.1
PREDICTED: Dasypus novemcinctus phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant	93.3	93.3	24%	5e-15	71%	XM_004473473.1
PREDICTED: Dasypus novemcinctus phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant	93.3	93.3	24%	5e-15	71%	XM_004473472.1
PREDICTED: Dasypus novemcinctus phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant	93.3	93.3	24%	5e-15	71%	XM_004473471.1
PREDICTED: Geospiza fortis phosphopantothenoylcysteine decarboxylase (PPCDC), mRNA	91.5	91.5	17%	2e-14	75%	XM_005419787.1
Phaseolus vulgaris hypothetical protein (PHAVU_008G074500g) mRNA, complete cds	87.8	87.8	21%	2e-13	71%	XM_007139913.1
PREDICTED: Nelumbo nucifera probable phosphopantothenoylcysteine decarboxylase (LOC104600044), trans	84.2	84.2	28%	3e-12	69%	XM_010262854.1
PREDICTED: Nelumbo nucifera probable phosphopantothenoylcysteine decarboxylase (LOC104600044), trans	84.2	84.2	28%	3e-12	69%	XM_010262847.1
PREDICTED: Chaetura pelagica phosphopantothenoylcysteine decarboxylase (PPCDC), mRNA	82.4	82.4	17%	1e-11	73%	XM_010001326.1
PREDICTED: Zonotrichia albicollis phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant X2	82.4	82.4	17%	1e-11	73%	XM_005487781.1
PREDICTED: Zonotrichia albicollis phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant X1	82.4	82.4	17%	1e-11	73%	XM_005487780.1
PREDICTED: Melopsittacus undulatus phosphopantothenoylcysteine decarboxylase (PPCDC), mRNA	82.4	82.4	17%	1e-11	73%	XM_005145728.1
PREDICTED: Anas platyrhynchos phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant X3	82.4	82.4	17%	1e-11	73%	XM_005017178.1
PREDICTED: Anas platyrhynchos phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant X2	82.4	82.4	17%	1e-11	73%	XM_005017177.1
PREDICTED: Anas platyrhynchos phosphopantothenoylcysteine decarboxylase (PPCDC), transcript variant X1	82.4	82.4	17%	1e-11	73%	XM_005017176.1
PREDICTED: Rattus norvegicus phosphopantothenoylcysteine decarboxylase (Ppcdc), transcript variant X6, ml	80.6	80.6	24%	3e-11	70%	XM_008766308.1

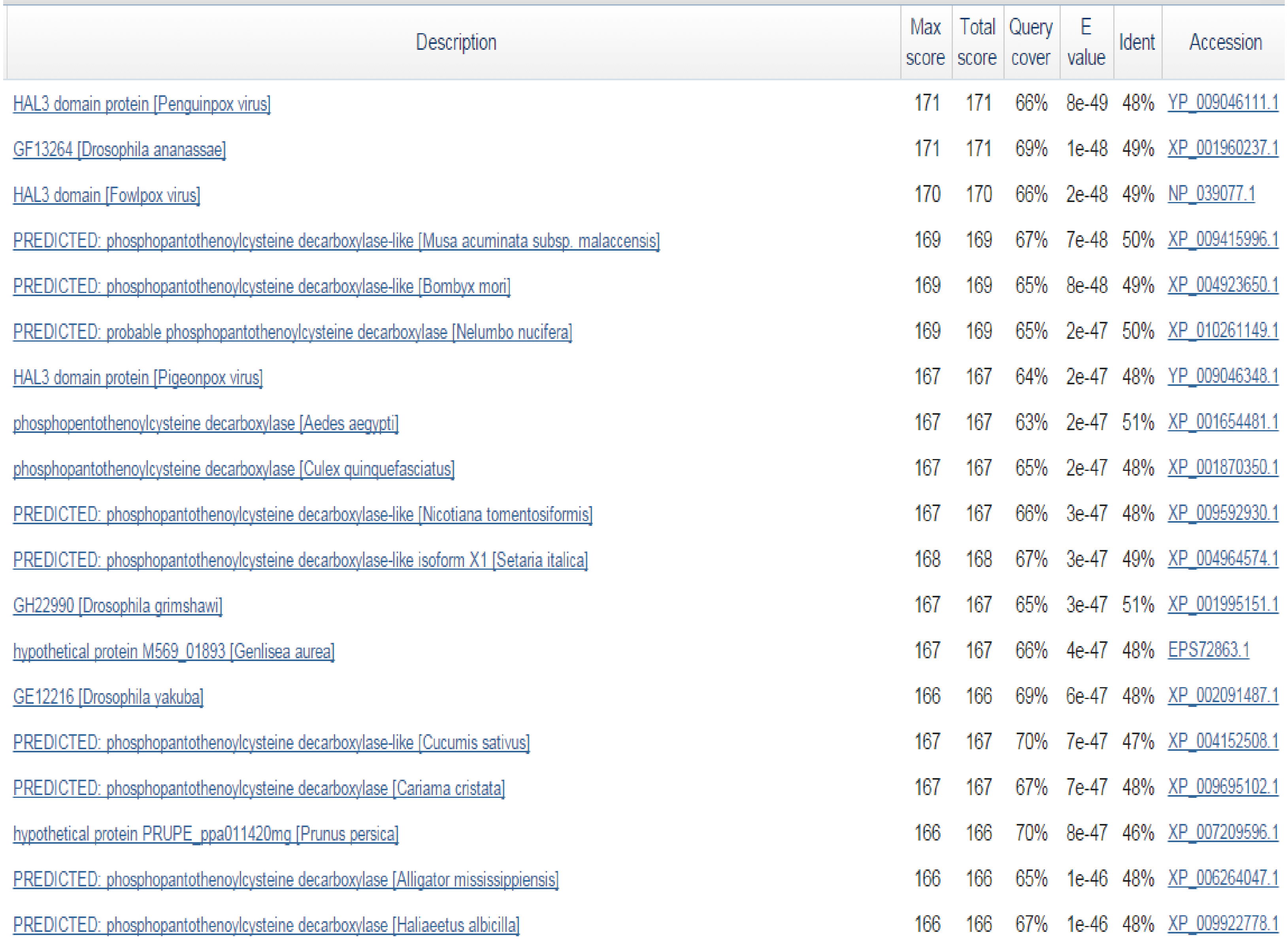
BLASTx alignment for C8246727

BLASTx search result showed that query sequence was aligned with the conserved flavoprotein superfamily. For the query sequence, identity to top hit penguinpox virus was 48%, matched length was 87 amino acid out of total 180 amino acids, and query coverage was 66% at E value of 8e-49.

A) Graphic display


Putative conserved domains have been detected, click on the image below for detailed results.
RF -2
Superfamilies
1
125
250
375
500
625
750
812
Flavoprotein superfamily
Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
150
300
450
600
750

B) Hit list of aligned NCBI sequences for C8246727 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
HAL3 domain protein [Penguinpox virus]	171	171	66%	8e-49	48%	YP_009046111.1
GF13264 [Drosophila ananassae]	171	171	69%	1e-48	49%	XP_001960237.1
HAL3 domain [Fowlpox virus]	170	170	66%	2e-48	49%	NP_039077.1
PREDICTED: phosphopantothenoylcysteine decarboxylase-like [Musa acuminata subsp. malaccensis]	169	169	67%	7e-48	50%	XP_009415996.1
PREDICTED: phosphopantothenoylcysteine decarboxylase-like [Bombyx mori]	169	169	65%	8e-48	49%	XP_004923650.1
PREDICTED: probable phosphopantothenoylcysteine decarboxylase [Nelumbo nucifera]	169	169	65%	2e-47	50%	XP_010261149.1
HAL3 domain protein [Pigeonpox virus]	167	167	64%	2e-47	48%	YP_009046348.1
phosphopentothenoylcysteine decarboxylase [Aedes aegypti]	167	167	63%	2e-47	51%	XP_001654481.1
phosphopantothenoylcysteine decarboxylase [Culex quinquefasciatus]	167	167	65%	2e-47	48%	XP_001870350.1
PREDICTED: phosphopantothenoylcysteine decarboxylase-like [Nicotiana tomentosiformis]	167	167	66%	3e-47	48%	XP_009592930.1
PREDICTED: phosphopantothenoylcysteine decarboxylase-like isoform X1 [Setaria italica]	168	168	67%	3e-47	49%	XP_004964574.1
GH22990 [Drosophila grimshawi]	167	167	65%	3e-47	51%	XP_001995151.1
hypothetical protein M569_01893 [Genlisea aurea]	167	167	66%	4e-47	48%	EPS72863.1
GE12216 [Drosophila yakuba]	166	166	69%	6e-47	48%	XP_002091487.1
PREDICTED: phosphopantothenoylcysteine decarboxylase-like [Cucumis sativus]	167	167	70%	7e-47	47%	XP_004152508.1
PREDICTED: phosphopantothenoylcysteine decarboxylase [Cariama cristata]	167	167	67%	7e-47	48%	XP_009695102.1
hypothetical protein PRUPE_ppa011420mg [Prunus persica]	166	166	70%	8e-47	46%	XP_007209596.1
PREDICTED: phosphopantothenoylcysteine decarboxylase [Alligator mississippiensis]	166	166	65%	1e-46	48%	XP_006264047.1
PREDICTED: phosphopantothenoylcysteine decarboxylase [Haliaeetus albicilla]	166	166	67%	1e-46	48%	XP_009922778.1

3.2.3. BLAST alignment for the contig C6240514 (238 BP)

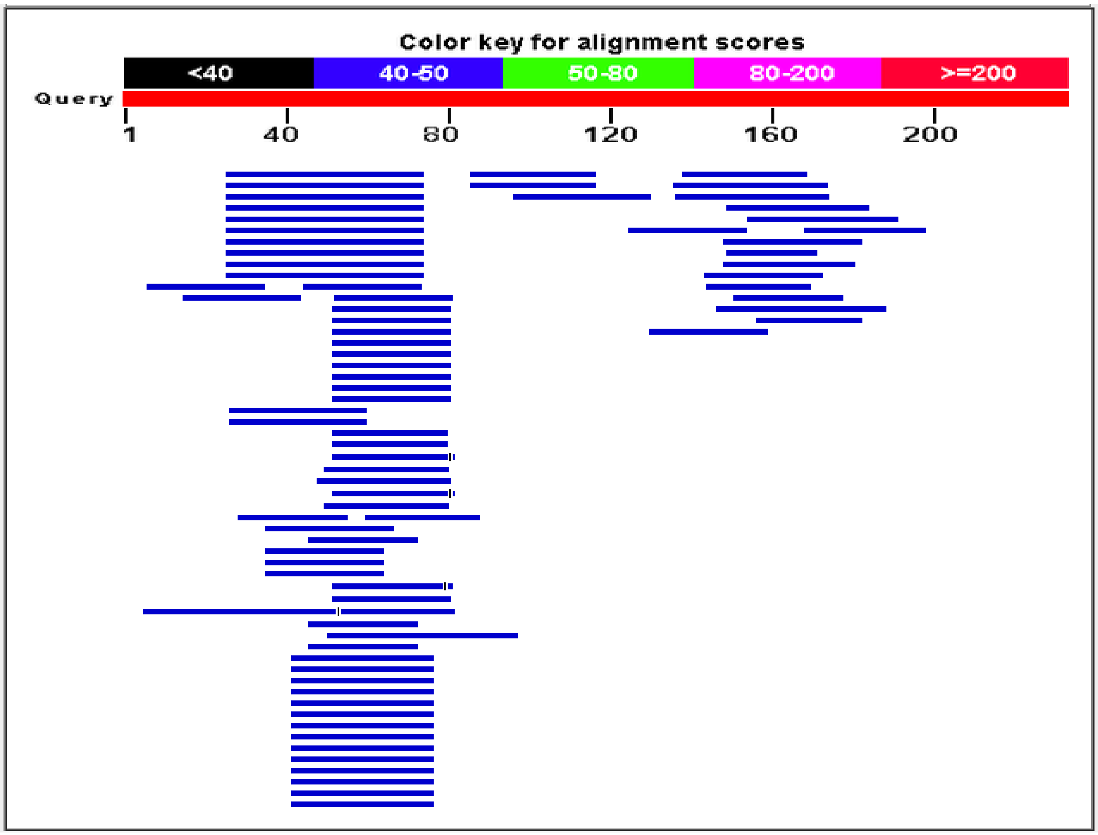
Nucleotide sequence in FASTA format

```
>C6240514
CAAGTCTAAATTAATATTTGAGTCACTAAGTTTTGTATAGTTTTCGTCTTTTTTCTCA
AAGAAGTTTGTTCTTCCTGATATAGCGATTGATTGCATAAAGGAAAGTGGTTGAGAT
ACATTGAATTCCCTTTTTAACCCAAAATCTGTTAGTATGCAATCTGTTGTGTATTTTA
CATAATCTAACATATCTTGTTTATTCATACCCTTCACCCCATCCTCGAAGTAATCATT
TACATAT
```

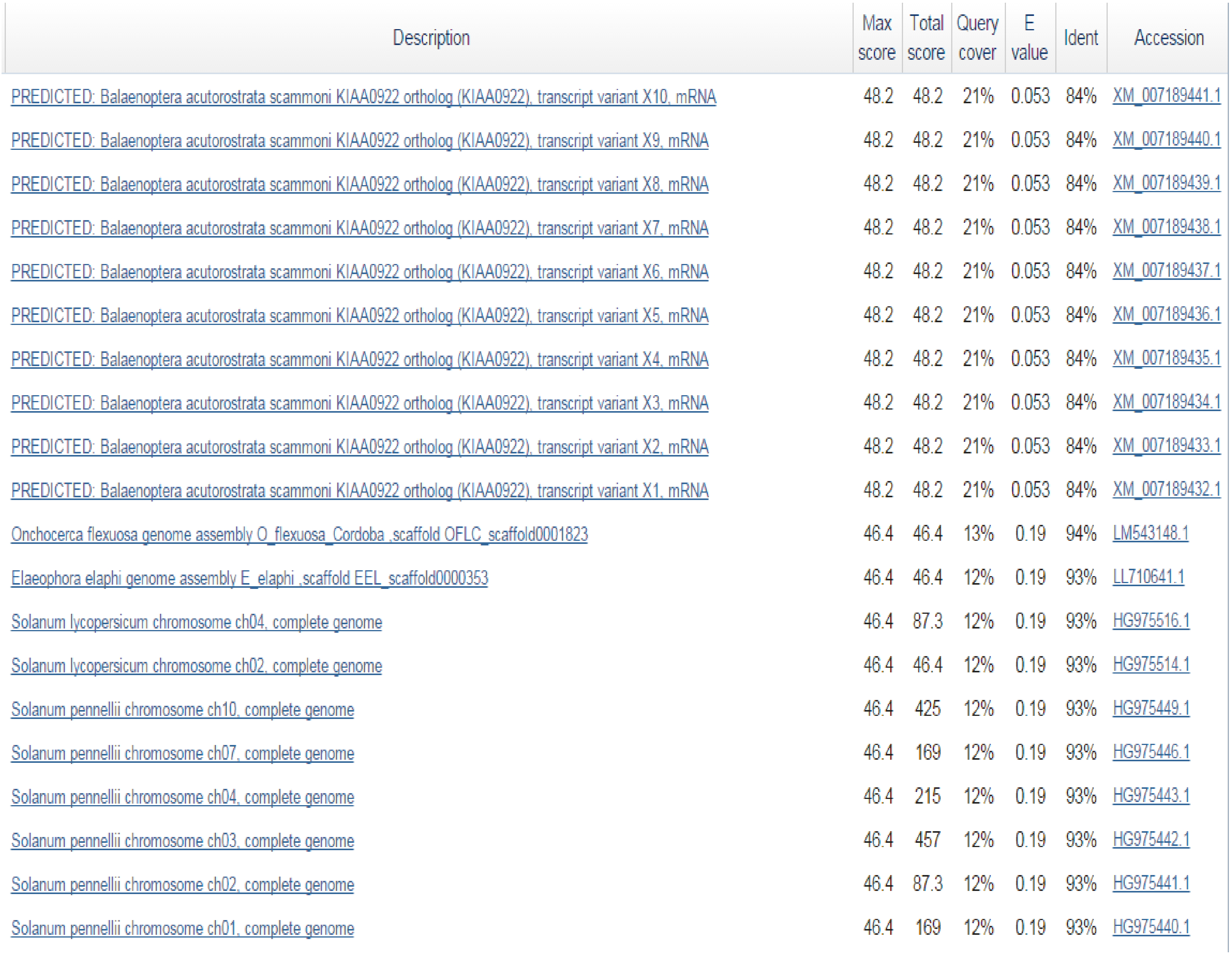
BLASTn alignment for C6240514

BLASTn search result showed that assembled query sequence was similar to a sequence from Minke whale (*Balaenoptera acutorostrata*). For the top hit, total coverage was 21% (50 bp match), E value was 0.053 and the identity was of 84% with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C6240514 after BLASTn search

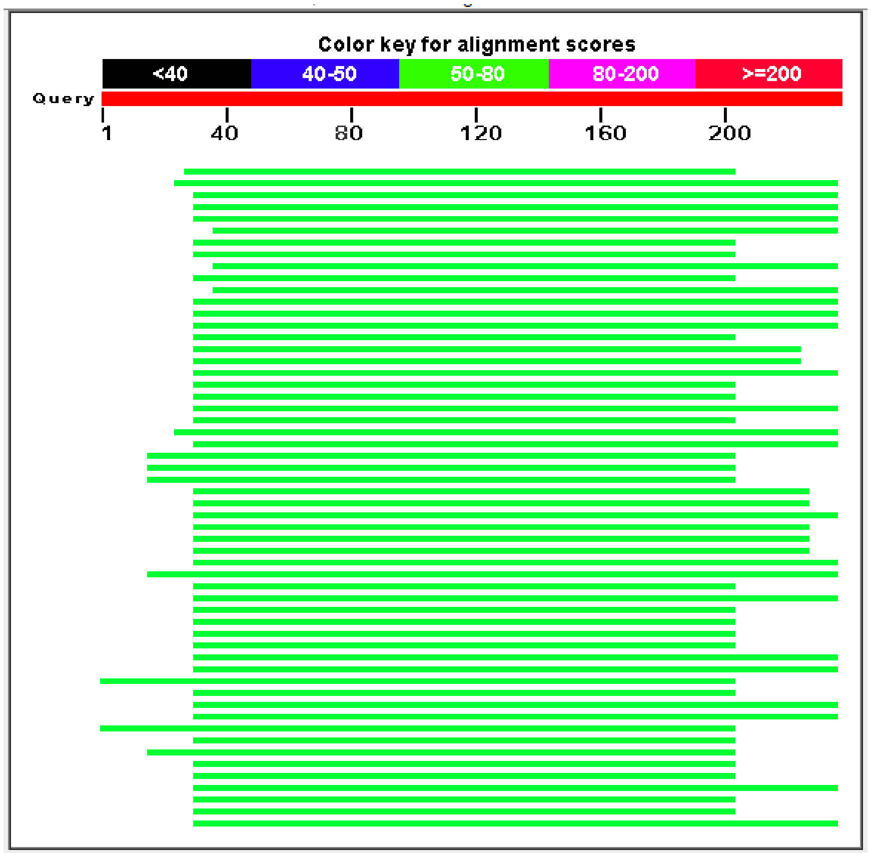


Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X10, mRNA	48.2	48.2	21%	0.053	84%	XM_007189441.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X9, mRNA	48.2	48.2	21%	0.053	84%	XM_007189440.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X8, mRNA	48.2	48.2	21%	0.053	84%	XM_007189439.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X7, mRNA	48.2	48.2	21%	0.053	84%	XM_007189438.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X6, mRNA	48.2	48.2	21%	0.053	84%	XM_007189437.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X5, mRNA	48.2	48.2	21%	0.053	84%	XM_007189436.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X4, mRNA	48.2	48.2	21%	0.053	84%	XM_007189435.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X3, mRNA	48.2	48.2	21%	0.053	84%	XM_007189434.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X2, mRNA	48.2	48.2	21%	0.053	84%	XM_007189433.1
PREDICTED: Balaenoptera acutorostrata scammoni KIAA0922 ortholog (KIAA0922), transcript variant X1, mRNA	48.2	48.2	21%	0.053	84%	XM_007189432.1
Onchocerca flexuosa genome assembly O_flexuosa_Cordoba ,scaffold OFLC_scaffold0001823	46.4	46.4	13%	0.19	94%	LM543148.1
Elaeophora elaphi genome assembly E_elaphi ,scaffold EEL_scaffold0000353	46.4	46.4	12%	0.19	93%	LL710641.1
Solanum lycopersicum chromosome ch04, complete genome	46.4	87.3	12%	0.19	93%	HG975516.1
Solanum lycopersicum chromosome ch02, complete genome	46.4	46.4	12%	0.19	93%	HG975514.1
Solanum pennellii chromosome ch10, complete genome	46.4	425	12%	0.19	93%	HG975449.1
Solanum pennellii chromosome ch07, complete genome	46.4	169	12%	0.19	93%	HG975446.1
Solanum pennellii chromosome ch04, complete genome	46.4	215	12%	0.19	93%	HG975443.1
Solanum pennellii chromosome ch03, complete genome	46.4	457	12%	0.19	93%	HG975442.1
Solanum pennellii chromosome ch02, complete genome	46.4	87.3	12%	0.19	93%	HG975441.1
Solanum pennellii chromosome ch01, complete genome	46.4	169	12%	0.19	93%	HG975440.1

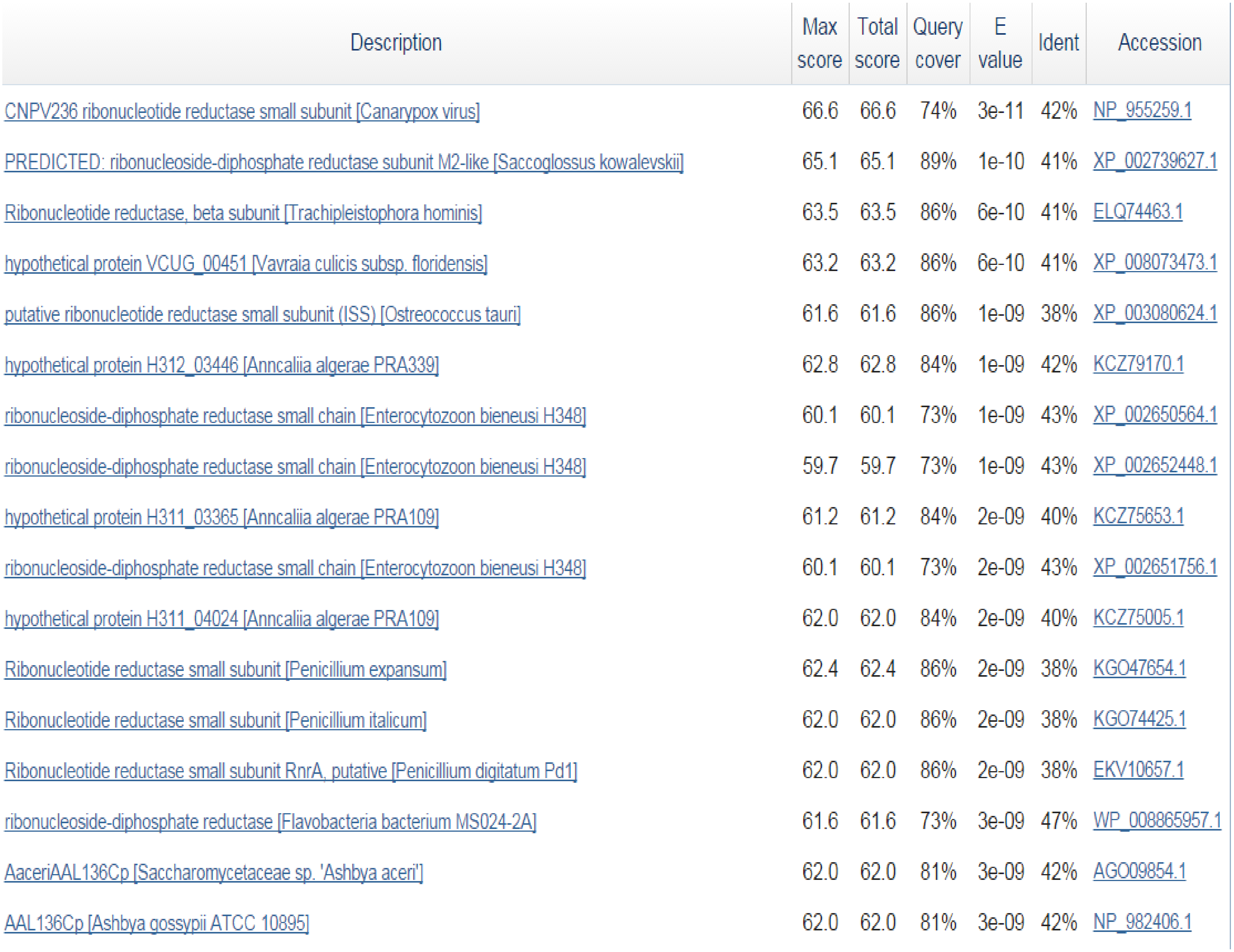
BLASTx alignment for the contig C6240514

After BLASTx search, query sequence was aligned with the conserved Ferritin-like superfamily. For the query sequence, identity to top hit (Canarypox virus) was 42%, matched length was 25 amino acid out of total 59 amino acids, and query coverage was 74% at E value of 3e-11.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C6240514 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
CNPV236 ribonucleotide reductase small subunit [Canarypox virus]	66.6	66.6	74%	3e-11	42%	NP_955259.1
PREDICTED: ribonucleoside-diphosphate reductase subunit M2-like [Saccoglossus kowalevskii]	65.1	65.1	89%	1e-10	41%	XP_002739627.1
Ribonucleotide reductase, beta subunit [Trachipleistophora hominis]	63.5	63.5	86%	6e-10	41%	ELQ74463.1
hypothetical protein VCUG_00451 [Vavraia culicis subsp. floridensis]	63.2	63.2	86%	6e-10	41%	XP_008073473.1
putative ribonucleotide reductase small subunit (ISS) [Ostreococcus tauri]	61.6	61.6	86%	1e-09	38%	XP_003080624.1
hypothetical protein H312_03446 [Anncaliia algerae PRA339]	62.8	62.8	84%	1e-09	42%	KCZ79170.1
ribonucleoside-diphosphate reductase small chain [Enterocytozoon bieneusi H348]	60.1	60.1	73%	1e-09	43%	XP_002650564.1
ribonucleoside-diphosphate reductase small chain [Enterocytozoon bieneusi H348]	59.7	59.7	73%	1e-09	43%	XP_002652448.1
hypothetical protein H311_03365 [Anncaliia algerae PRA109]	61.2	61.2	84%	2e-09	40%	KCZ75653.1
ribonucleoside-diphosphate reductase small chain [Enterocytozoon bieneusi H348]	60.1	60.1	73%	2e-09	43%	XP_002651756.1
hypothetical protein H311_04024 [Anncaliia algerae PRA109]	62.0	62.0	84%	2e-09	40%	KCZ75005.1
Ribonucleotide reductase small subunit [Penicillium expansum]	62.4	62.4	86%	2e-09	38%	KGO47654.1
Ribonucleotide reductase small subunit [Penicillium italicum]	62.0	62.0	86%	2e-09	38%	KGO74425.1
Ribonucleotide reductase small subunit RnrA, putative [Penicillium digitatum Pd1]	62.0	62.0	86%	2e-09	38%	EKV10657.1
ribonucleoside-diphosphate reductase [Flavobacteria bacterium MS024-2A]	61.6	61.6	73%	3e-09	47%	WP_008865957.1
AaceriAAL136Cp [Saccharomycetaceae sp. 'Ashbya aceri']	62.0	62.0	81%	3e-09	42%	AGO09854.1
AAL136Cp [Ashbya gossypii ATCC 10895]	62.0	62.0	81%	3e-09	42%	NP_982406.1

3.2.4. BLAST alignment for C7753049 (371 BP)

Nucleotide sequence in FASTA format

>C7753049

CTCTTCGCTTTCTGAGGACTTACGGATATACAAGACGTACTTTTCCATATAGTCGCA

ATTATGCCCTAGTTGGGTAATTCCTCAAATCAACAGTAGAAATATCATCGACTATGT

CTAGATCAATCTCAATAATACGCCGCTTTAATTGCTCCAAATTCTCCCCCTTACGAA

CCCATGCAAAAGAGTCATCTATAGCCTGTGGAGAGGTGAAGATAATGAACGGCGAG

TTGAGGGGTATCGAGCCACCTTTGAAAAAGAGATTGCATGGATAGCGGTCAGTGAT

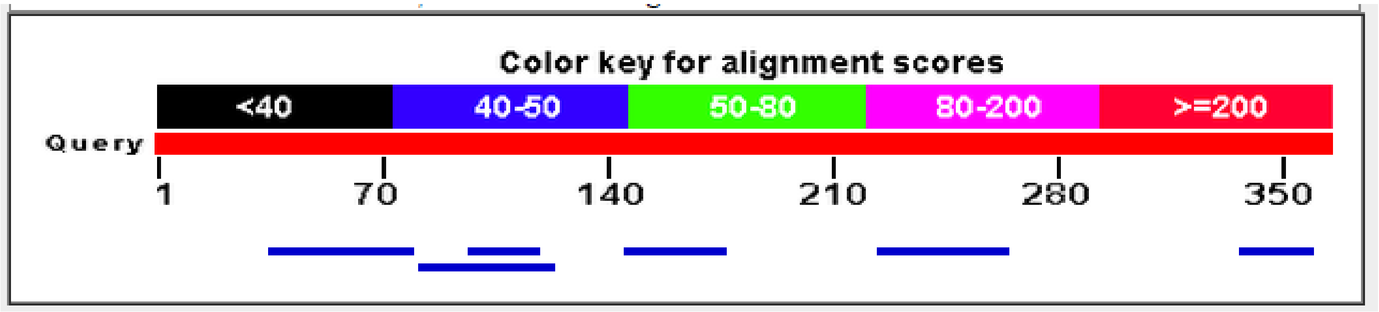
CTTGAGGAGCTGCTCAAAAGGCACGTTCCCTTCACGGAAGTCGTCGAACAATAAAC

ACTCCTGTTGTGAGTATCCCGTGCCAATCCAA

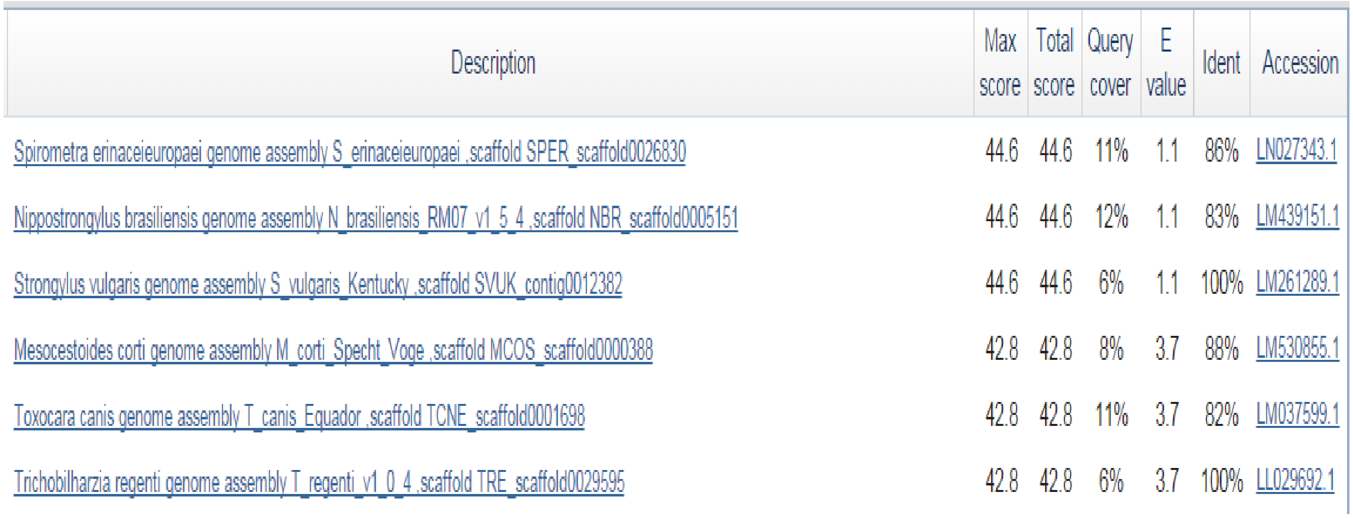
BLASTn alignment for C7753049

BLASTn search result showed that assembled query sequence was similar to a sequence from a tape-worm (S*pirometra erinaceieuropaei*). For the top hit, total coverage was 11% (43 bp matched), E value was 0.11 and the identity was of 86% with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for C7753049 after BLASTn search

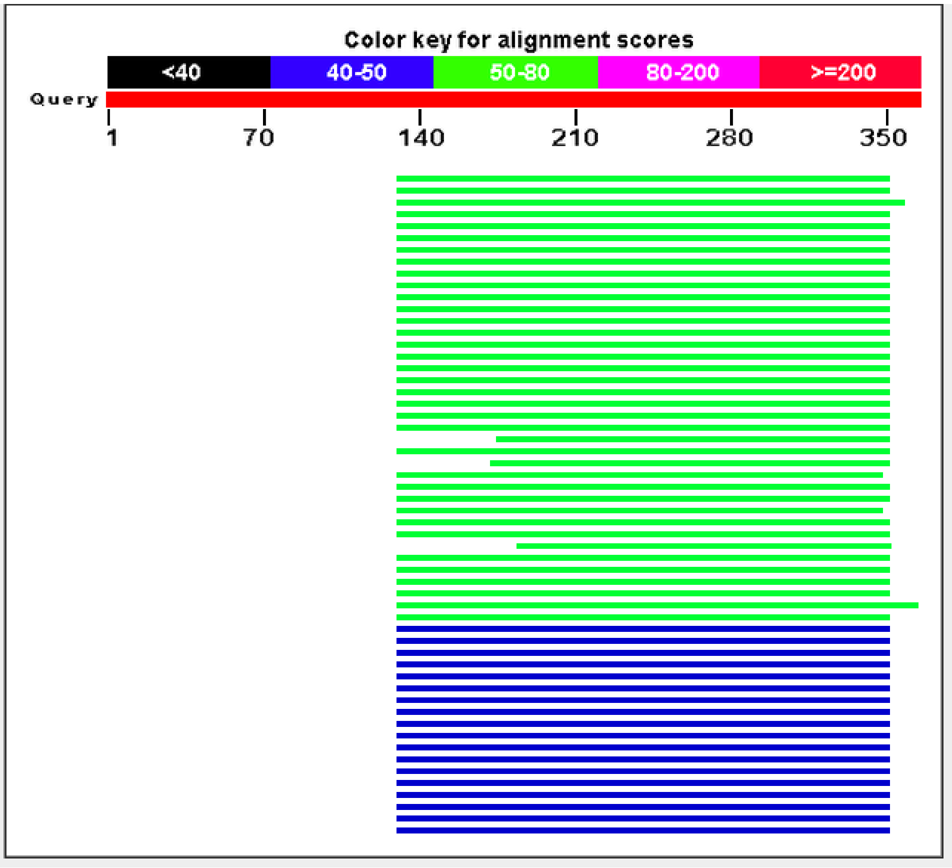


Description	Max score	Total score	Query cover	E value	Ident	Accession
Spirometra erinaceieuropaei genome assembly S_erinaceieuropaei ,scaffold SPER_scaffold0026830	44.6	44.6	11%	1.1	86%	LN027343.1
Nippostrongylus brasiliensis genome assembly N_brasiliensis_RM07_v1_5_4 ,scaffold NBR_scaffold0005151	44.6	44.6	12%	1.1	83%	LM439151.1
Strongylus vulgaris genome assembly S_vulgaris_Kentucky ,scaffold SVUK_contig0012382	44.6	44.6	6%	1.1	100%	LM261289.1
Mesocestoides corti genome assembly M_corti_Specht_Voge ,scaffold MCOS_scaffold0000388	42.8	42.8	8%	3.7	88%	LM530855.1
Toxocara canis genome assembly T_canis_Equador ,scaffold TCNE_scaffold0001698	42.8	42.8	11%	3.7	82%	LM037599.1
Trichobilharzia regenti genome assembly T_regenti_v1_0_4 ,scaffold TRE_scaffold0029595	42.8	42.8	6%	3.7	100%	LL029692.1

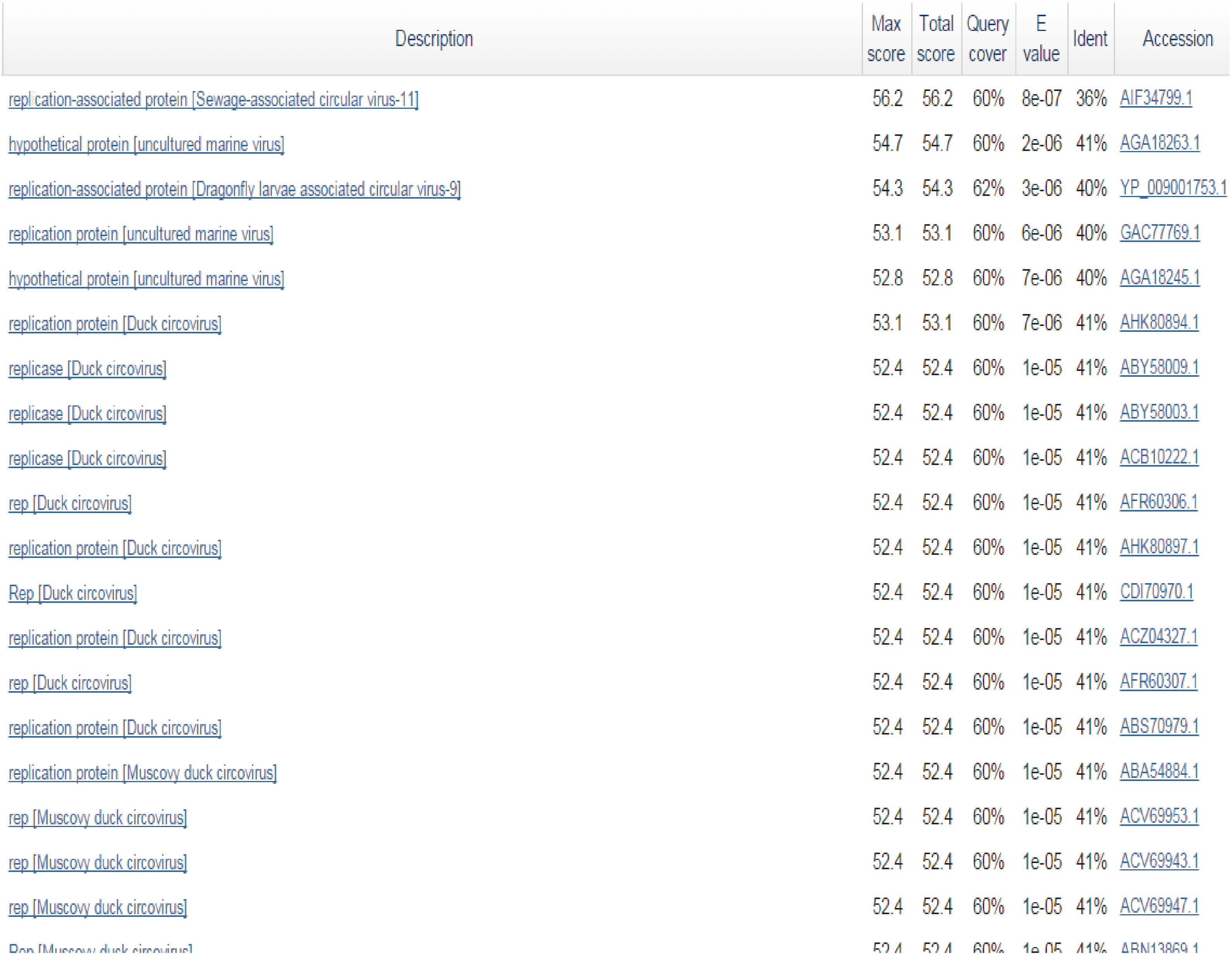
BLASTx alignment for the contig C7753049

BLASTx search result showed that query sequence was aligned with the sequence of sewage-associated circular DNA virus-11 as top match. The query sequence was 42% identical to the top hit sequence, matched length was 27 amino acid out of total 75 amino acids, and query coverage was 60% at E value of 8e-07. The BLASTn and BLASTx search of first 1-135 base pairs position showed no similarity to the sequence of NCBI database.

A) Graphic display



B) Hit list of aligned NCBI sequences for C7753049 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
replication-associated protein [Sewage-associated circular virus-11]	56.2	56.2	60%	8e-07	36%	AIF34799.1
hypothetical protein [uncultured marine virus]	54.7	54.7	60%	2e-06	41%	AGA18263.1
replication-associated protein [Dragonfly larvae associated circular virus-9]	54.3	54.3	62%	3e-06	40%	YP_009001753.1
replication protein [uncultured marine virus]	53.1	53.1	60%	6e-06	40%	GAC77769.1
hypothetical protein [uncultured marine virus]	52.8	52.8	60%	7e-06	40%	AGA18245.1
replication protein [Duck circovirus]	53.1	53.1	60%	7e-06	41%	AHK80894.1
replicase [Duck circovirus]	52.4	52.4	60%	1e-05	41%	ABY58009.1
replicase [Duck circovirus]	52.4	52.4	60%	1e-05	41%	ABY58003.1
replicase [Duck circovirus]	52.4	52.4	60%	1e-05	41%	ACB10222.1
rep [Duck circovirus]	52.4	52.4	60%	1e-05	41%	AFR60306.1
replication protein [Duck circovirus]	52.4	52.4	60%	1e-05	41%	AHK80897.1
Rep [Duck circovirus]	52.4	52.4	60%	1e-05	41%	CDI70970.1
replication protein [Duck circovirus]	52.4	52.4	60%	1e-05	41%	ACZ04327.1
rep [Duck circovirus]	52.4	52.4	60%	1e-05	41%	AFR60307.1
replication protein [Duck circovirus]	52.4	52.4	60%	1e-05	41%	ABS70979.1
replication protein [Muscovy duck circovirus]	52.4	52.4	60%	1e-05	41%	ABA54884.1
rep [Muscovy duck circovirus]	52.4	52.4	60%	1e-05	41%	ACV69953.1
rep [Muscovy duck circovirus]	52.4	52.4	60%	1e-05	41%	ACV69943.1
rep [Muscovy duck circovirus]	52.4	52.4	60%	1e-05	41%	ACV69947.1
Rep [Muscovy duck circovirus]	52.4	52.4	60%	1e-05	41%	ABN13869.1

3.2.5. BLAST alignment for the contig C6842206 (263 BP)

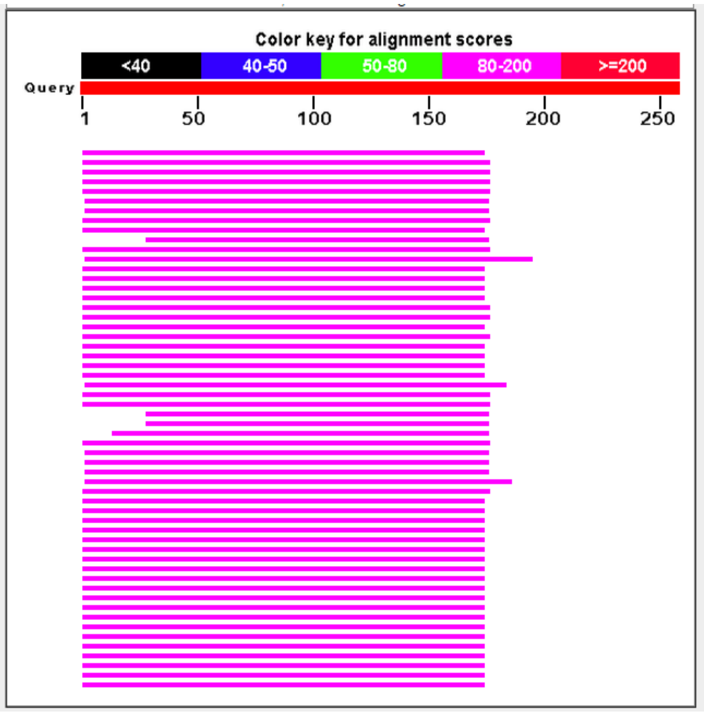
Nucleotide sequence in FASTA format

```
>C6842206
CGCAGACTGCGTTGAGCTTGGAGACCTTTTCGGCAAGAGGGATGAGGTTGATGATA
TTCCCGAAGGGCTTGCGCTCAAAGGTTCCATCCAAAGCAGCCACAACCACTATTTTG
TTCATGTTAGCAAGCTCCTCACAGAAATCAACGATGTCGGAGAAGAACTGTCCCTC
ATCGATTCCGACAACATCATAGTTTTAAGCATGGGACACCCATTTCATAAGGGTCGA
AGTTTTTAGAGCTTTTCTGGTGATTCTGAATAGAAGA
```

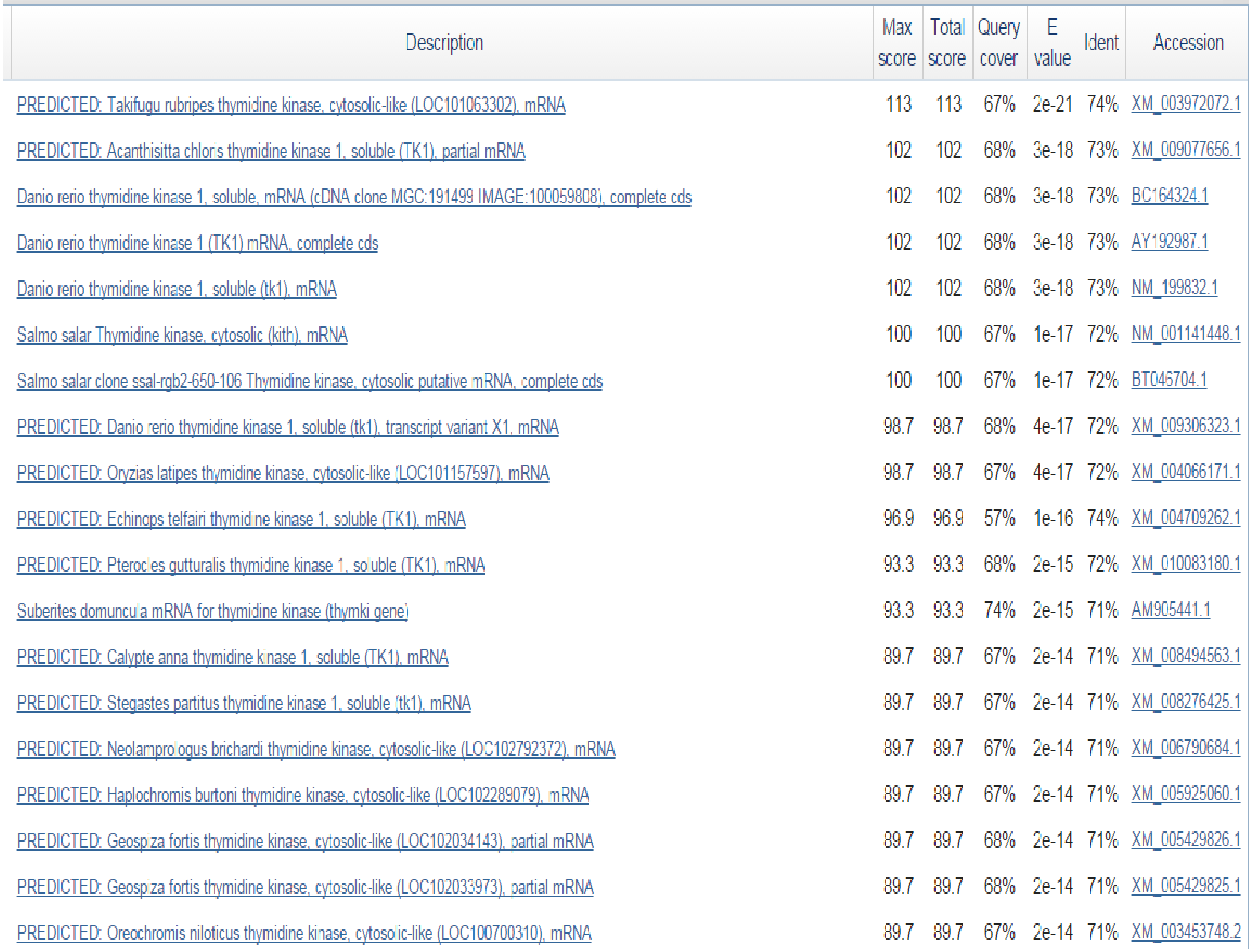
BLASTn alignment for the contig C6842206

BLASTn search result showed that assembled query sequence was similar to the sequences of fish and birds genome. For the top hit (sequence from a fish) total coverage was 67% (177 bp matched), E value was 2e-21 and the identity was of 74% with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C6842206 after BLASTn search

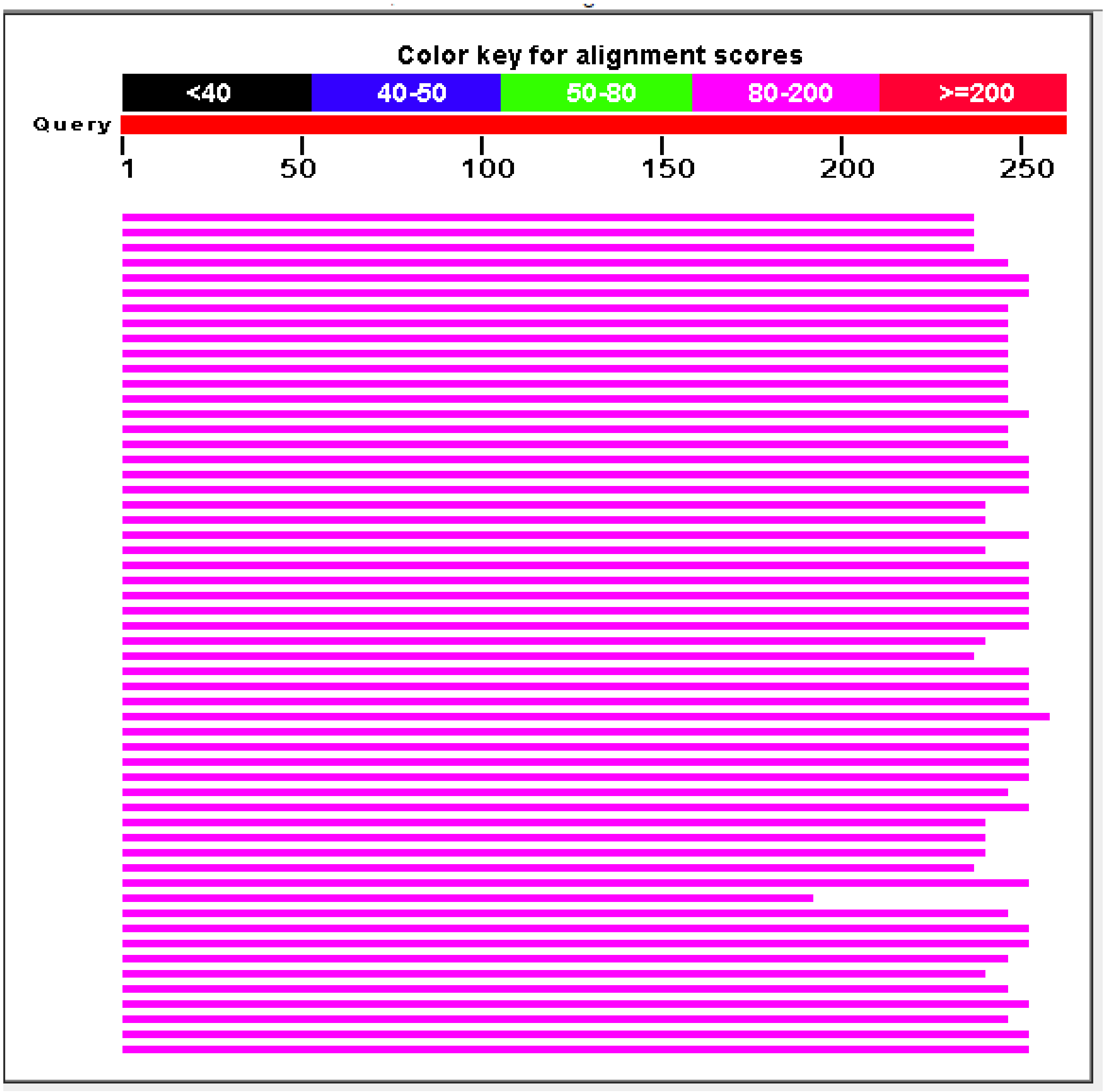


Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Takifugu rubripes thymidine kinase, cytosolic-like (LOC101063302), mRNA	113	113	67%	2e-21	74%	XM_003972072.1
PREDICTED: Acanthisitta chloris thymidine kinase 1, soluble (TK1), partial mRNA	102	102	68%	3e-18	73%	XM_009077656.1
Danio rerio thymidine kinase 1, soluble, mRNA (cDNA clone MGC:191499 IMAGE:100059808), complete cds	102	102	68%	3e-18	73%	BC164324.1
Danio rerio thymidine kinase 1 (TK1) mRNA, complete cds	102	102	68%	3e-18	73%	AY192987.1
Danio rerio thymidine kinase 1, soluble (tk1), mRNA	102	102	68%	3e-18	73%	NM_199832.1
Salmo salar Thymidine kinase, cytosolic (kith), mRNA	100	100	67%	1e-17	72%	NM_001141448.1
Salmo salar clone ssal-rgb2-650-106 Thymidine kinase, cytosolic putative mRNA, complete cds	100	100	67%	1e-17	72%	BT046704.1
PREDICTED: Danio rerio thymidine kinase 1, soluble (tk1), transcript variant X1, mRNA	98.7	98.7	68%	4e-17	72%	XM_009306323.1
PREDICTED: Oryzias latipes thymidine kinase, cytosolic-like (LOC101157597), mRNA	98.7	98.7	67%	4e-17	72%	XM_004066171.1
PREDICTED: Echinops telfairi thymidine kinase 1, soluble (TK1), mRNA	96.9	96.9	57%	1e-16	74%	XM_004709262.1
PREDICTED: Pterocles gutturalis thymidine kinase 1, soluble (TK1), mRNA	93.3	93.3	68%	2e-15	72%	XM_010083180.1
Suberites domuncula mRNA for thymidine kinase (thymki gene)	93.3	93.3	74%	2e-15	71%	AM905441.1
PREDICTED: Calypte anna thymidine kinase 1, soluble (TK1), mRNA	89.7	89.7	67%	2e-14	71%	XM_008494563.1
PREDICTED: Stegastes partitus thymidine kinase 1, soluble (tk1), mRNA	89.7	89.7	67%	2e-14	71%	XM_008276425.1
PREDICTED: Neolamprologus brichardi thymidine kinase, cytosolic-like (LOC102792372), mRNA	89.7	89.7	67%	2e-14	71%	XM_006790684.1
PREDICTED: Haplochromis burtoni thymidine kinase, cytosolic-like (LOC102289079), mRNA	89.7	89.7	67%	2e-14	71%	XM_005925060.1
PREDICTED: Geospiza fortis thymidine kinase, cytosolic-like (LOC102034143), partial mRNA	89.7	89.7	68%	2e-14	71%	XM_005429826.1
PREDICTED: Geospiza fortis thymidine kinase, cytosolic-like (LOC102033973), partial mRNA	89.7	89.7	68%	2e-14	71%	XM_005429825.1
PREDICTED: Oreochromis niloticus thymidine kinase, cytosolic-like (LOC100700310), mRNA	89.7	89.7	67%	2e-14	71%	XM_003453748.2

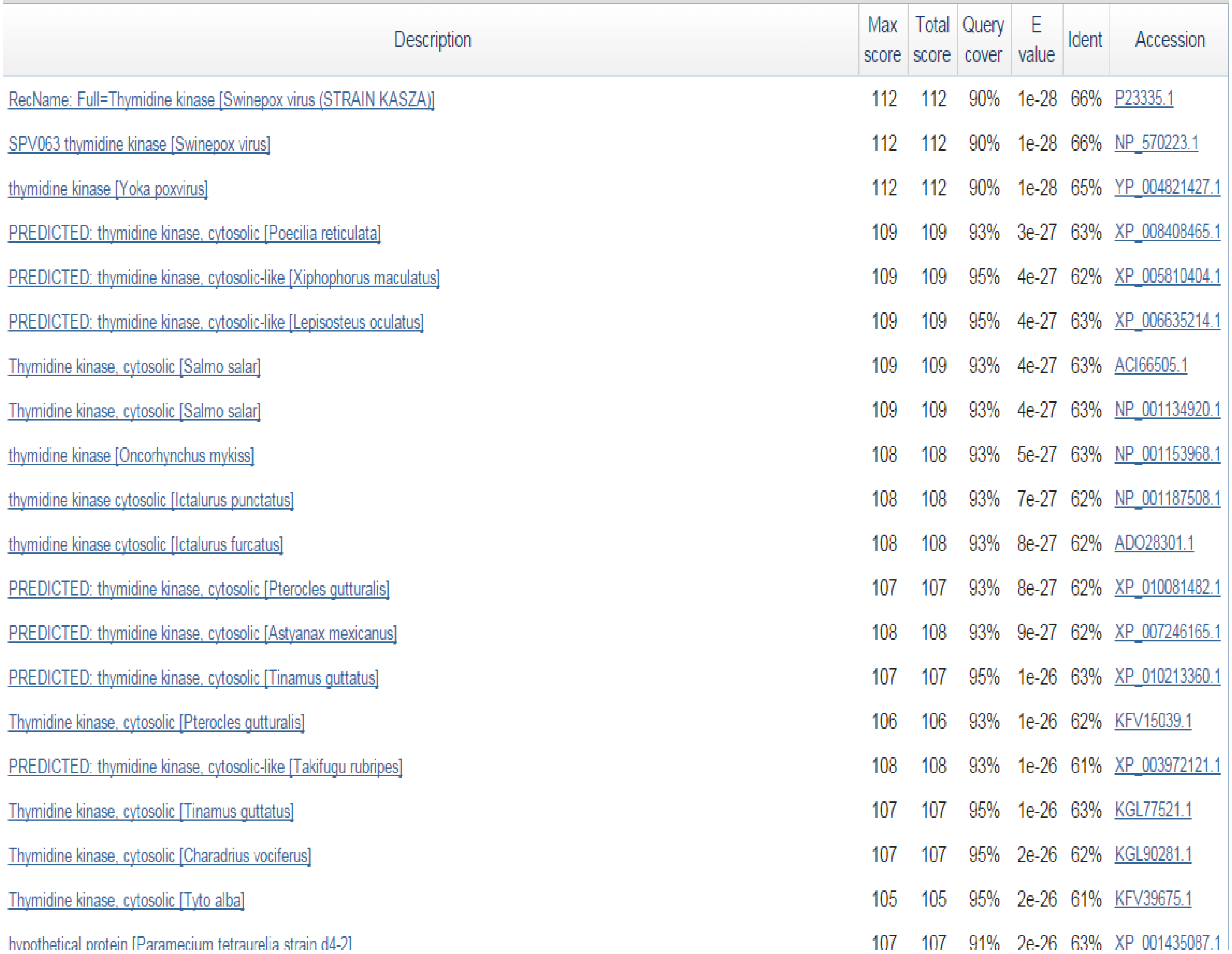
BLASTx alignment for the contig C6842206

BLASTx search showed that assembled query sequence was aligned with the sequence of swinepox virus as the top match and with sequences of fish. For the top hit, identity to the matched length was 66%, matched length was 52 amino acid out of total 79 amino acids, and query coverage was 90% at E value of 1e-28.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C6842206 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Thymidine kinase [Swinepox virus (STRAIN KASZA)]	112	112	90%	1e-28	66%	P23335.1
SPV063 thymidine kinase [Swinepox virus]	112	112	90%	1e-28	66%	NP_570223.1
thymidine kinase [Yoka poxvirus]	112	112	90%	1e-28	65%	YP_004821427.1
PREDICTED: thymidine kinase, cytosolic [Poecilia reticulata]	109	109	93%	3e-27	63%	XP_008408465.1
PREDICTED: thymidine kinase, cytosolic-like [Xiphophorus maculatus]	109	109	95%	4e-27	62%	XP_005810404.1
PREDICTED: thymidine kinase, cytosolic-like [Lepisosteus oculatus]	109	109	95%	4e-27	63%	XP_006635214.1
Thymidine kinase, cytosolic [Salmo salar]	109	109	93%	4e-27	63%	ACI66505.1
Thymidine kinase, cytosolic [Salmo salar]	109	109	93%	4e-27	63%	NP_001134920.1
thymidine kinase [Oncorhynchus mykiss]	108	108	93%	5e-27	63%	NP_001153968.1
thymidine kinase cytosolic [Ictalurus punctatus]	108	108	93%	7e-27	62%	NP_001187508.1
thymidine kinase cytosolic [Ictalurus furcatus]	108	108	93%	8e-27	62%	ADO28301.1
PREDICTED: thymidine kinase, cytosolic [Pterocles gutturalis]	107	107	93%	8e-27	62%	XP_010081482.1
PREDICTED: thymidine kinase, cytosolic [Astyanax mexicanus]	108	108	93%	9e-27	62%	XP_007246165.1
PREDICTED: thymidine kinase, cytosolic [Tinamus guttatus]	107	107	95%	1e-26	63%	XP_010213360.1
Thymidine kinase, cytosolic [Pterocles gutturalis]	106	106	93%	1e-26	62%	KFV15039.1
PREDICTED: thymidine kinase, cytosolic-like [Takifugu rubripes]	108	108	93%	1e-26	61%	XP_003972121.1
Thymidine kinase, cytosolic [Tinamus guttatus]	107	107	95%	1e-26	63%	KGL77521.1
Thymidine kinase, cytosolic [Charadrius vociferus]	107	107	95%	2e-26	62%	KGL90281.1
Thymidine kinase, cytosolic [Tyto alba]	105	105	95%	2e-26	61%	KFV39675.1
hypothetical protein [Paramecium tetraurelia strain d4-2]	107	107	91%	2e-26	63%	XP_001435087.1

3.2.6. BLAST alignment for contig C7235113 (292 BP)

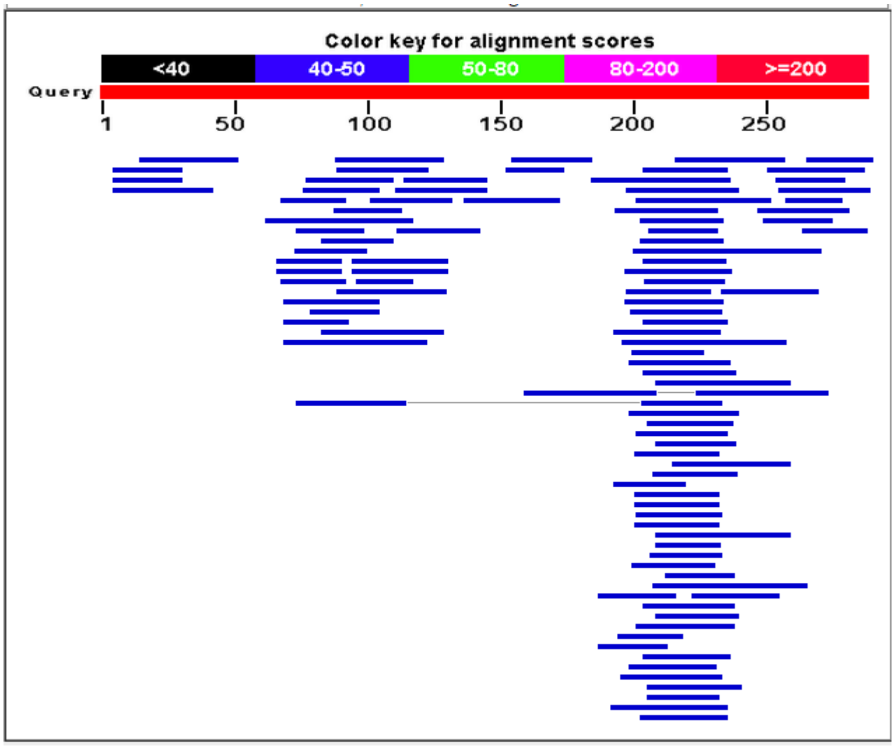
Nucleotide sequence in FASTA format

>C7235113
ATAATCCTGCTACTATAGCTGCAGCAATTGCTATATACATTACCATAGATTCCCAGT
CTACTCCAACTGATTGTGTGATATTAACATTATTAACTGCTTCTAAAGCTACATTCAT
AACTACTTTTACAACACAATTTGCTGCACTTTGTCCCACATTAAGAATTTGAAAAGT
AACAGGAGTTCCAGTAGCATTACAAGTTCCAAAATTATAATTGTTAACTTTTATCTC
ATTATTTACTAAAGCGTTTGCTTTACATTCATTAGTTAAAAATGTTTGAACATCATTA
GTAAA

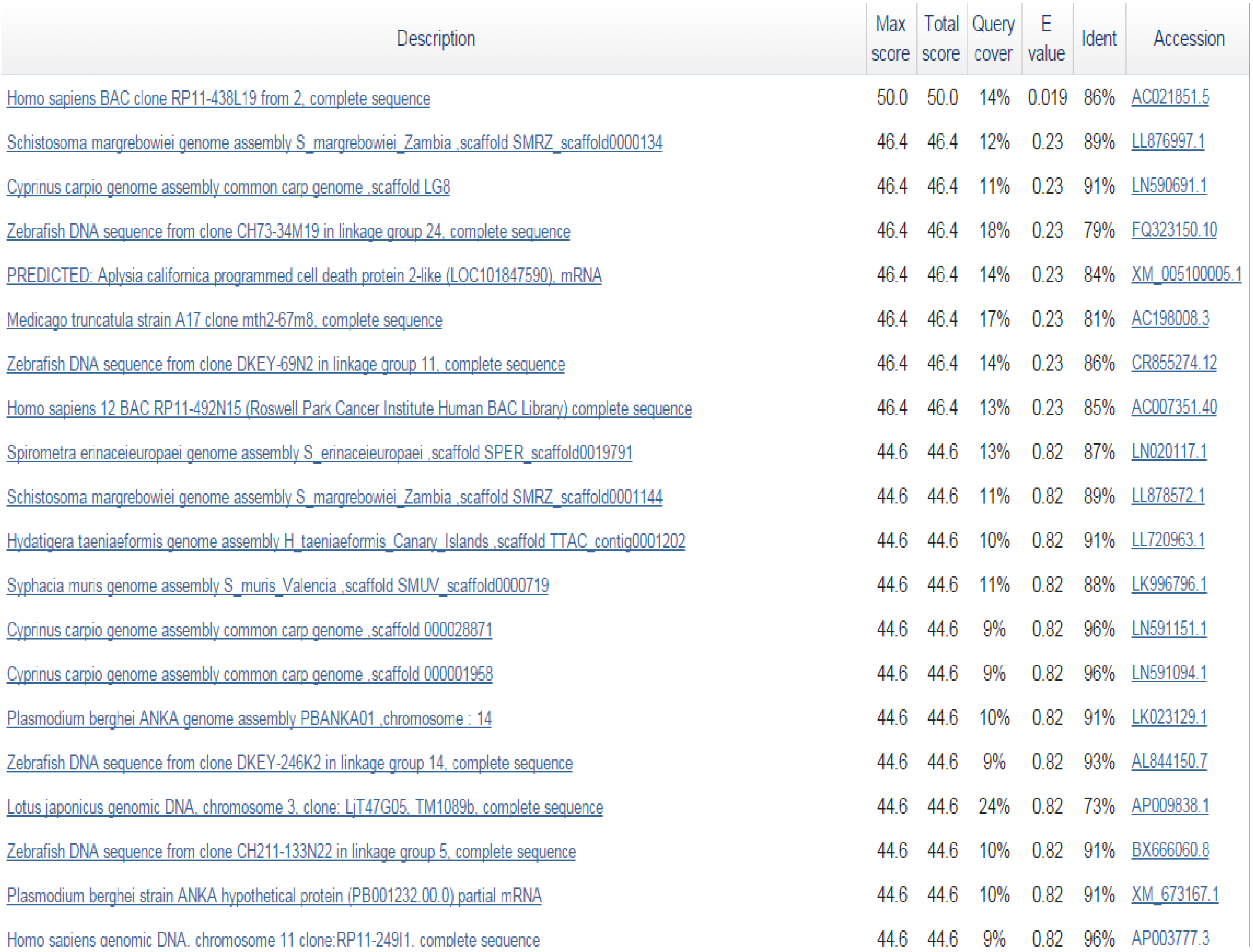
BLASTn alignment for the contig C7235113

BLASTn search result showed that query sequence was similar to the sequence of human BAC clone. For the top hit, total coverage was 14% (42 bp matched), E value was 0.019 and the identity was 86% with the matched region.

A) Graphic display



B) Hit list of aligned NCBI sequences for the contig C7235113 after BLASTn search

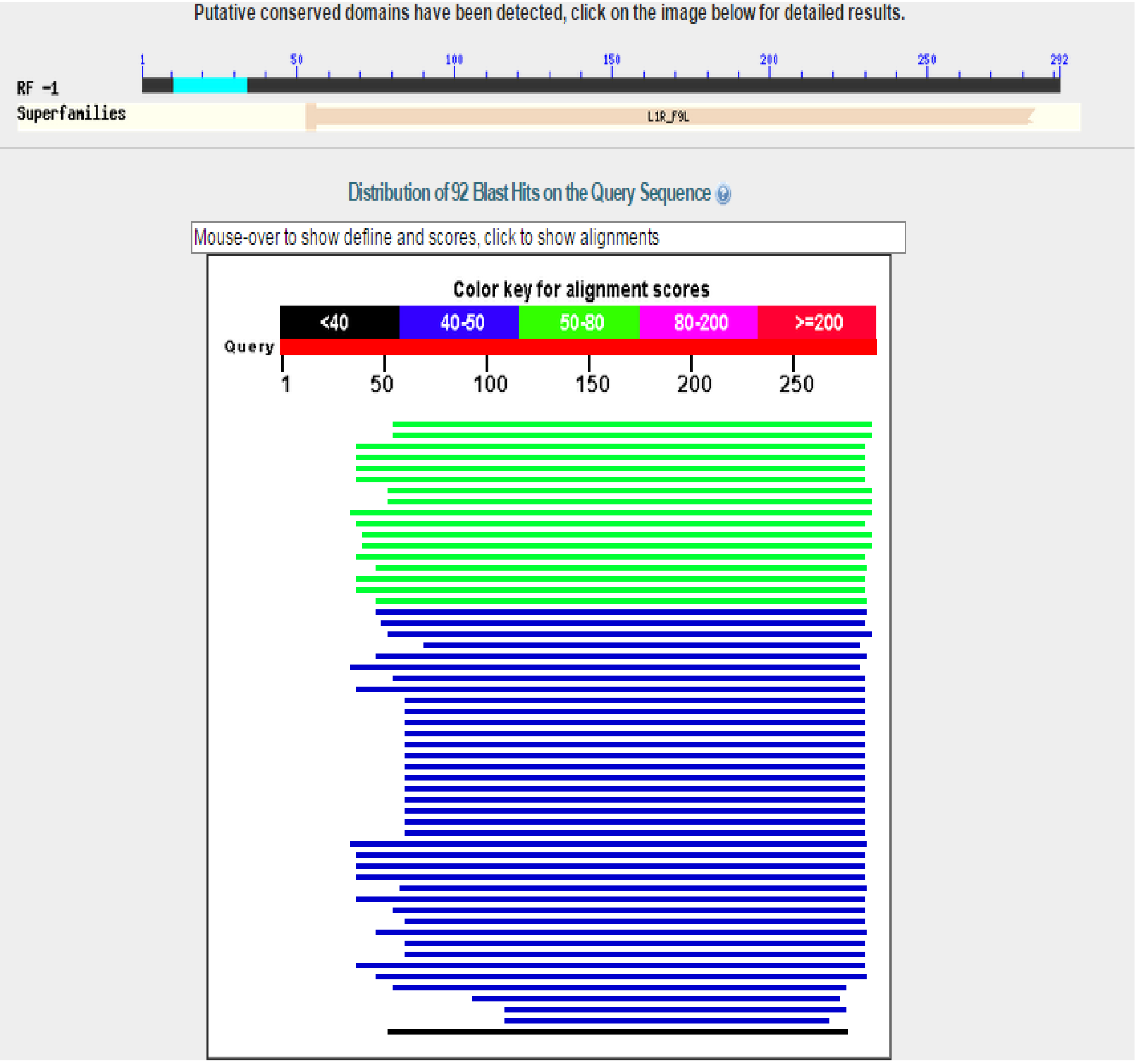


Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens BAC clone RP11-438L19 from 2, complete sequence	50.0	50.0	14%	0.019	86%	AC021851.5
Schistosoma margrebowiei genome assembly S_margrebowiei_Zambia ,scaffold SMRZ_scaffold0000134	46.4	46.4	12%	0.23	89%	LL876997.1
Cyprinus carpio genome assembly common carp genome ,scaffold LG8	46.4	46.4	11%	0.23	91%	LN590691.1
Zebrafish DNA sequence from clone CH73-34M19 in linkage group 24, complete sequence	46.4	46.4	18%	0.23	79%	FQ323150.10
PREDICTED: Aplysia californica programmed cell death protein 2-like (LOC101847590), mRNA	46.4	46.4	14%	0.23	84%	XM_005100005.1
Medicago truncatula strain A17 clone mth2-67m8, complete sequence	46.4	46.4	17%	0.23	81%	AC198008.3
Zebrafish DNA sequence from clone DKEY-69N2 in linkage group 11, complete sequence	46.4	46.4	14%	0.23	86%	CR855274.12
Homo sapiens 12 BAC RP11-492N15 (Roswell Park Cancer Institute Human BAC Library) complete sequence	46.4	46.4	13%	0.23	85%	AC007351.40
Spirometra erinaceieuropaei genome assembly S_erinaceieuropaei ,scaffold SPER_scaffold0019791	44.6	44.6	13%	0.82	87%	LN020117.1
Schistosoma margrebowiei genome assembly S_margrebowiei_Zambia ,scaffold SMRZ_scaffold0001144	44.6	44.6	11%	0.82	89%	LL878572.1
Hydatigera taeniaeformis genome assembly H_taeniaeformis_Canary_Islands ,scaffold TTAC_contig0001202	44.6	44.6	10%	0.82	91%	LL720963.1
Syphacia muris genome assembly S_muris_Valencia ,scaffold SMUV_scaffold0000719	44.6	44.6	11%	0.82	88%	LK996796.1
Cyprinus carpio genome assembly common carp genome ,scaffold 000028871	44.6	44.6	9%	0.82	96%	LN591151.1
Cyprinus carpio genome assembly common carp genome ,scaffold 000001958	44.6	44.6	9%	0.82	96%	LN591094.1
Plasmodium berghei ANKA genome assembly PBANKA01 ,chromosome : 14	44.6	44.6	10%	0.82	91%	LK023129.1
Zebrafish DNA sequence from clone DKEY-246K2 in linkage group 14, complete sequence	44.6	44.6	9%	0.82	93%	AL844150.7
Lotus japonicus genomic DNA, chromosome 3, clone: LjT47G05, TM1089b, complete sequence	44.6	44.6	24%	0.82	73%	AP009838.1
Zebrafish DNA sequence from clone CH211-133N22 in linkage group 5, complete sequence	44.6	44.6	10%	0.82	91%	BX666060.8
Plasmodium berghei strain ANKA hypothetical protein (PB001232.00.0) partial mRNA	44.6	44.6	10%	0.82	91%	XM_673167.1
Homo sapiens genomic DNA, chromosome 11 clone:RP11-249I1, complete sequence	44.6	44.6	9%	0.82	96%	AP003777.3

BLASTx alignment for the contig C7235113

BLASTx search showed that assembled query sequence was aligned as the top match with the sequence of *Molluscum contagiosum* virus (a poxvirus that produces skin infection) and with the sequence of other poxviruses. Identity of the matched length was 37% for the top hit sequence, matched length was 29 amino acid out of total 79 amino acids, and query coverage was 80% at E value of 4e-09.

A) Graphic display

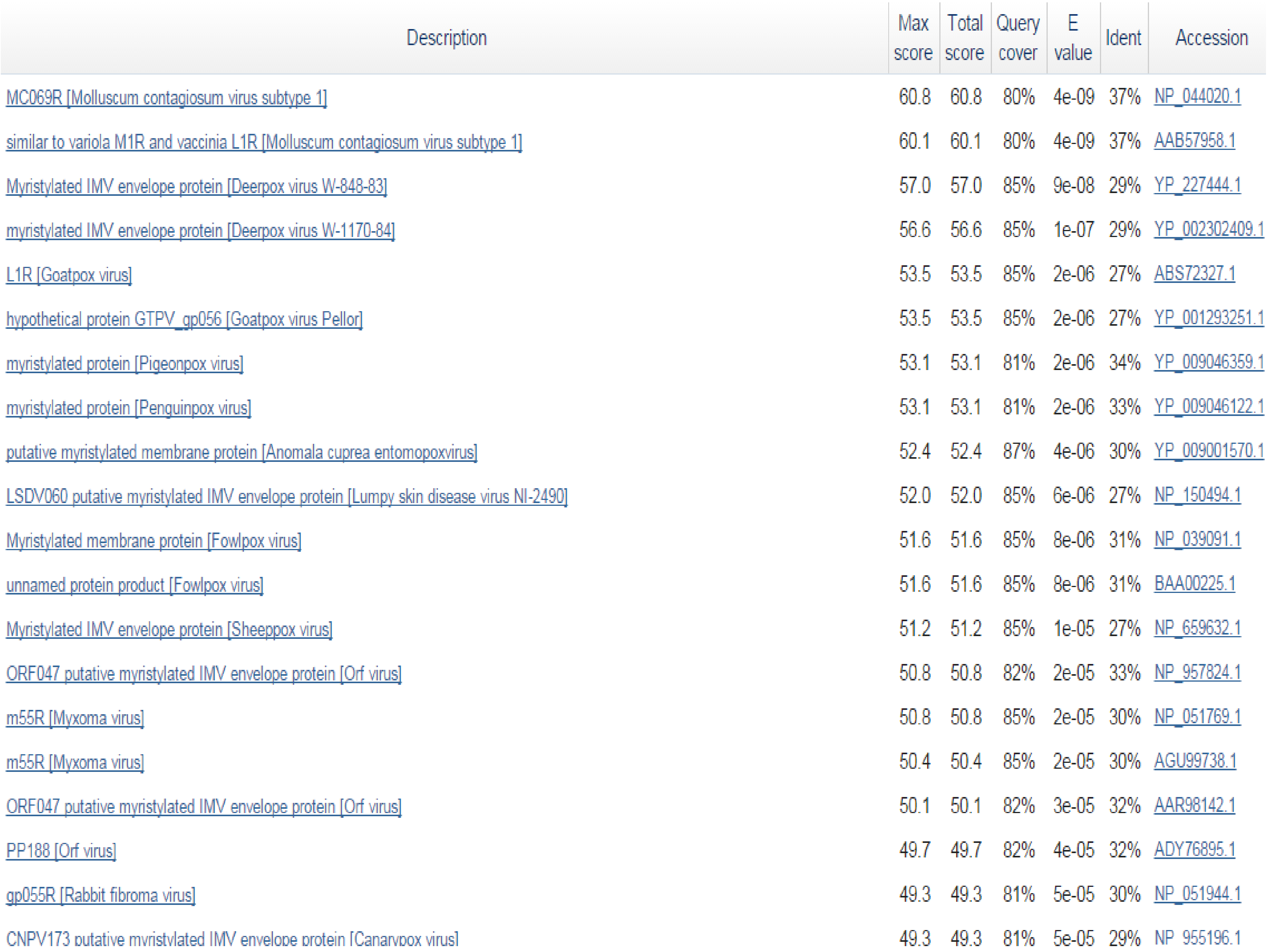


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 92 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C7235113 after BLASTx search



Description	Max score	Total score	Query cover	E value	Ident	Accession
MC069R [Molluscum contagiosum virus subtype 1]	60.8	60.8	80%	4e-09	37%	NP_044020.1
similar to variola M1R and vaccinia L1R [Molluscum contagiosum virus subtype 1]	60.1	60.1	80%	4e-09	37%	AAB57958.1
Myristylated IMV envelope protein [Deerpox virus W-848-83]	57.0	57.0	85%	9e-08	29%	YP_227444.1
myristylated IMV envelope protein [Deerpox virus W-1170-84]	56.6	56.6	85%	1e-07	29%	YP_002302409.1
L1R [Goatpox virus]	53.5	53.5	85%	2e-06	27%	ABS72327.1
hypothetical protein GTPV_gp056 [Goatpox virus Pellor]	53.5	53.5	85%	2e-06	27%	YP_001293251.1
myristylated protein [Pigeonpox virus]	53.1	53.1	81%	2e-06	34%	YP_009046359.1
myristylated protein [Penguinpox virus]	53.1	53.1	81%	2e-06	33%	YP_009046122.1
putative myristylated membrane protein [Anomala cuprea entomopoxvirus]	52.4	52.4	87%	4e-06	30%	YP_009001570.1
LSDV060 putative myristylated IMV envelope protein [Lumpy skin disease virus NI-2490]	52.0	52.0	85%	6e-06	27%	NP_150494.1
Myristylated membrane protein [Fowlpox virus]	51.6	51.6	85%	8e-06	31%	NP_039091.1
unnamed protein product [Fowlpox virus]	51.6	51.6	85%	8e-06	31%	BAA00225.1
Myristylated IMV envelope protein [Sheeppox virus]	51.2	51.2	85%	1e-05	27%	NP_659632.1
ORF047 putative myristylated IMV envelope protein [Orf virus]	50.8	50.8	82%	2e-05	33%	NP_957824.1
m55R [Myxoma virus]	50.8	50.8	85%	2e-05	30%	NP_051769.1
m55R [Myxoma virus]	50.4	50.4	85%	2e-05	30%	AGU99738.1
ORF047 putative myristylated IMV envelope protein [Orf virus]	50.1	50.1	82%	3e-05	32%	AAR98142.1
PP188 [Orf virus]	49.7	49.7	82%	4e-05	32%	ADY76895.1
gp055R [Rabbit fibroma virus]	49.3	49.3	81%	5e-05	30%	NP_051944.1
CNPV173 putative myristylated IMV envelope protein [Canarypox virus]	49.3	49.3	81%	5e-05	29%	NP_955196.1

3.3. Alignment of contigs/scaffolds similar to retrovirus sequences with sequences of NCBI database

3.3.1. BLAST alignment for the contig C7407449 (312 BP)

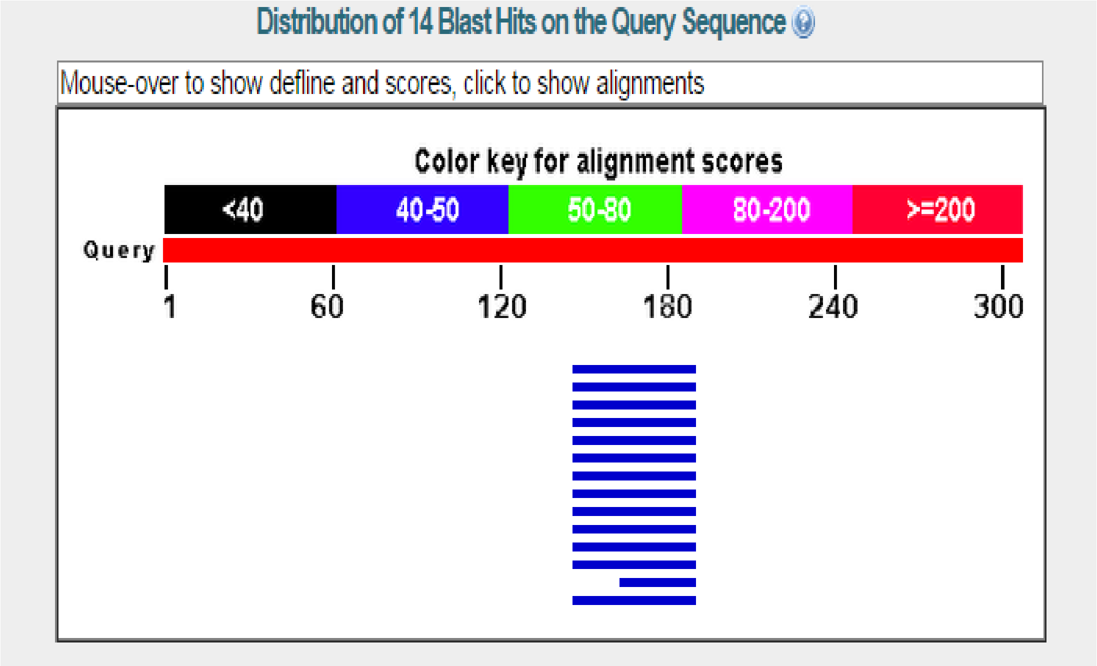
Nucleotide sequence in FASTA format

>C7407449
ATATGCGACTGCCCAACTCGCGAGCCTTCGCTGCCCCGCAGTCTTTCACCTTACCCT
CATTTCGAGATGCTCTTGTAAGCATGAGGGGCCCGACGTACTTCGGGAAAATTGAT
CTGACGGACGCCTTCTATAGCCTGCCGCTGCACCCTGATCTCCAACCGTACTTCGCG
GTGTGGTCCGGTCGGCGACGCCGGATGACATACAGAGTCATGCCGCAGGGGTGGTC
ATGGTCCCCGTACATCTTTCAGACCTCGCTGGCTCCAATTCAAGATCTCGTGCATCG
CCTGCACCCCGCAGTTCGAATGATTCGCT

BLASTn alignment for the contig C7407449

BLASTn search result showed that assembled query sequence was similar to sequence of bacteria. However, the alignment was not significant as the alignment had higher E-value of 3.0, only 14% of total coverage (45 bp matched) and the identity of 82% with the matched region.

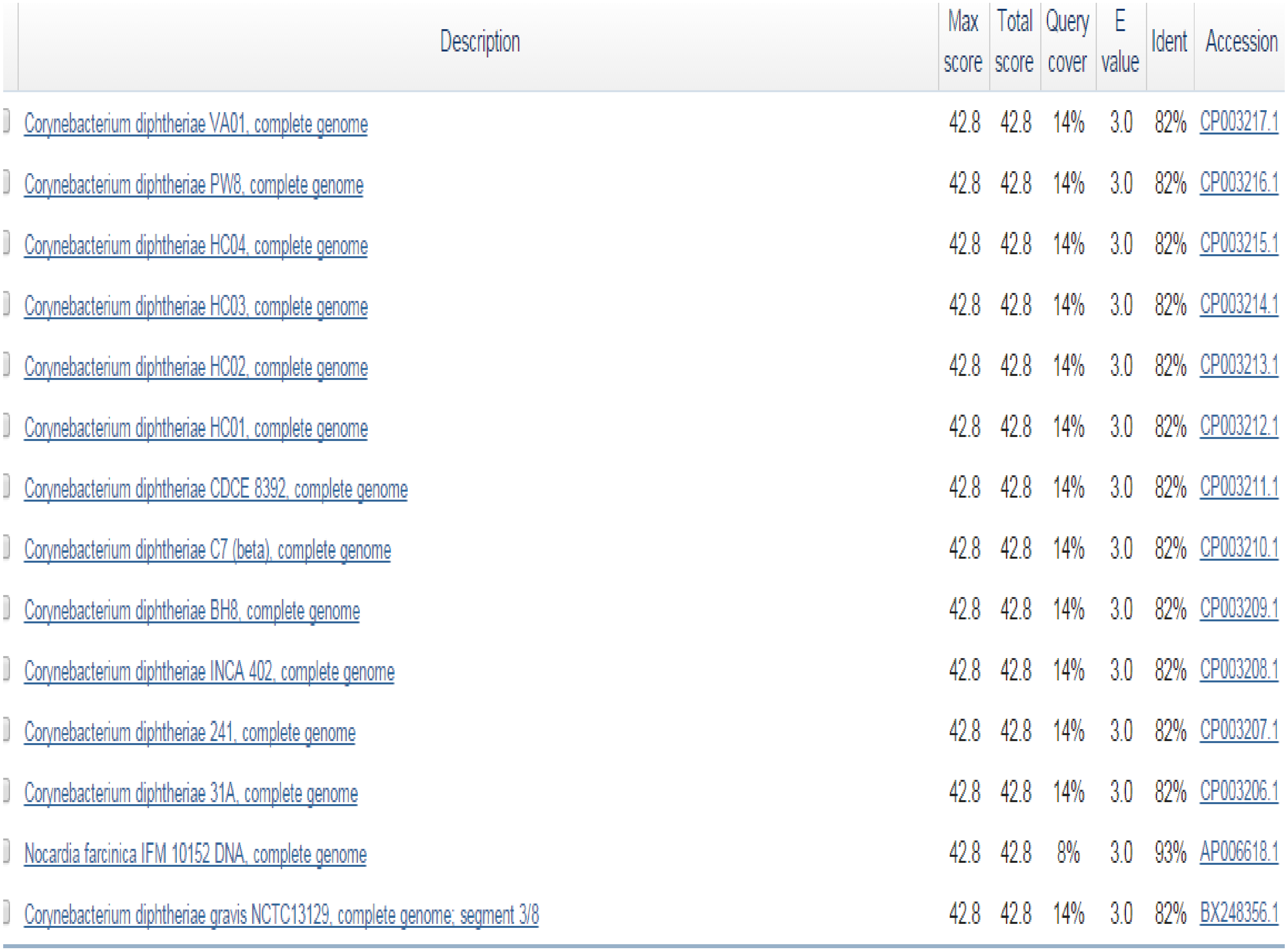
A) Graphic display



Distribution of 14 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C7407449 after BLASTn search

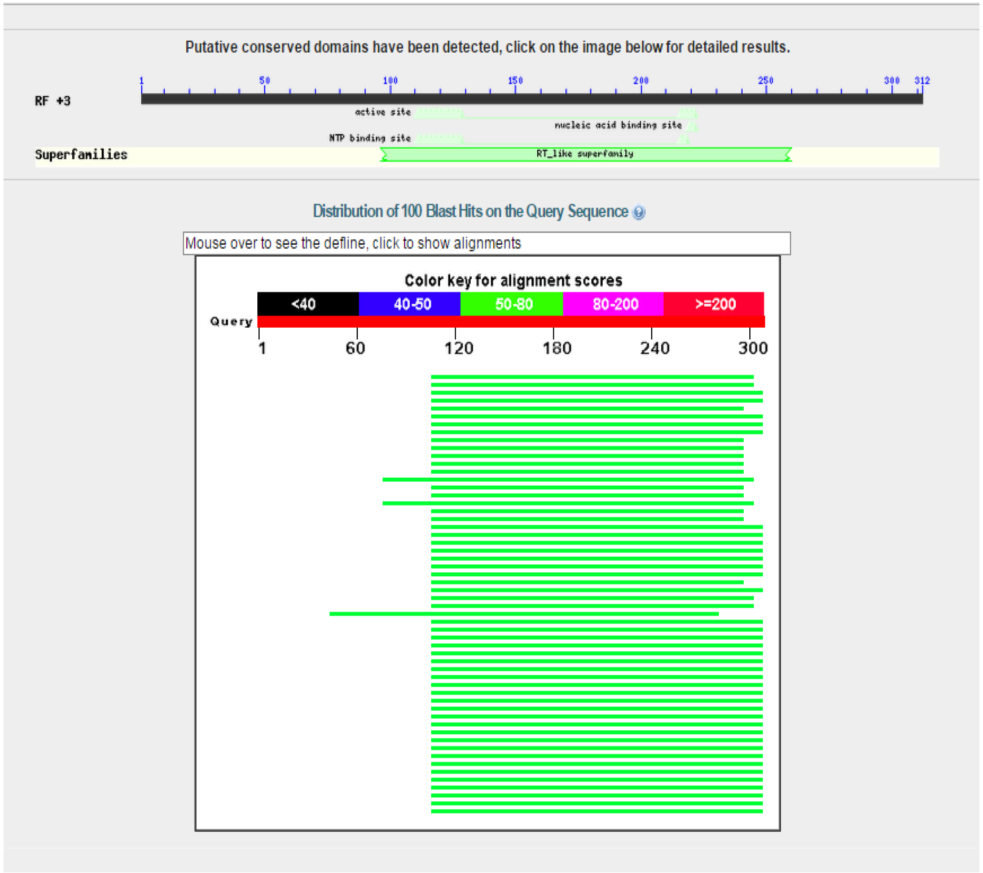


Description	Max score	Total score	Query cover	E value	Ident	Accession
□ Corynebacterium diphtheriae VA01, complete genome	42.8	42.8	14%	3.0	82%	CP003217.1
□ Corynebacterium diphtheriae PW8, complete genome	42.8	42.8	14%	3.0	82%	CP003216.1
□ Corynebacterium diphtheriae HC04, complete genome	42.8	42.8	14%	3.0	82%	CP003215.1
□ Corynebacterium diphtheriae HC03, complete genome	42.8	42.8	14%	3.0	82%	CP003214.1
□ Corynebacterium diphtheriae HC02, complete genome	42.8	42.8	14%	3.0	82%	CP003213.1
□ Corynebacterium diphtheriae HC01, complete genome	42.8	42.8	14%	3.0	82%	CP003212.1
□ Corynebacterium diphtheriae CDCE 8392, complete genome	42.8	42.8	14%	3.0	82%	CP003211.1
□ Corynebacterium diphtheriae C7 (beta), complete genome	42.8	42.8	14%	3.0	82%	CP003210.1
□ Corynebacterium diphtheriae BH8, complete genome	42.8	42.8	14%	3.0	82%	CP003209.1
□ Corynebacterium diphtheriae INCA 402, complete genome	42.8	42.8	14%	3.0	82%	CP003208.1
□ Corynebacterium diphtheriae 241, complete genome	42.8	42.8	14%	3.0	82%	CP003207.1
□ Corynebacterium diphtheriae 31A, complete genome	42.8	42.8	14%	3.0	82%	CP003206.1
□ Nocardia farcinica IFM 10152 DNA, complete genome	42.8	42.8	8%	3.0	93%	AP006618.1
□ Corynebacterium diphtheriae gravis NCTC13129, complete genome; segment 3/8	42.8	42.8	14%	3.0	82%	BX248356.1

BLASTx alignment for the contig C7407449

BLASTx search result showed that the assembled query sequence was aligned with the sequence of conserved reverse transcriptase-like protein superfamily of Simian immunodeficiency virus. For the query sequence, identity to the aligned sequences ranged from 36 – 40%, matched length was 29 amino acid out of total 72 amino acids, query coverage was 63% and E value was 8e-09.

A) Graphic display

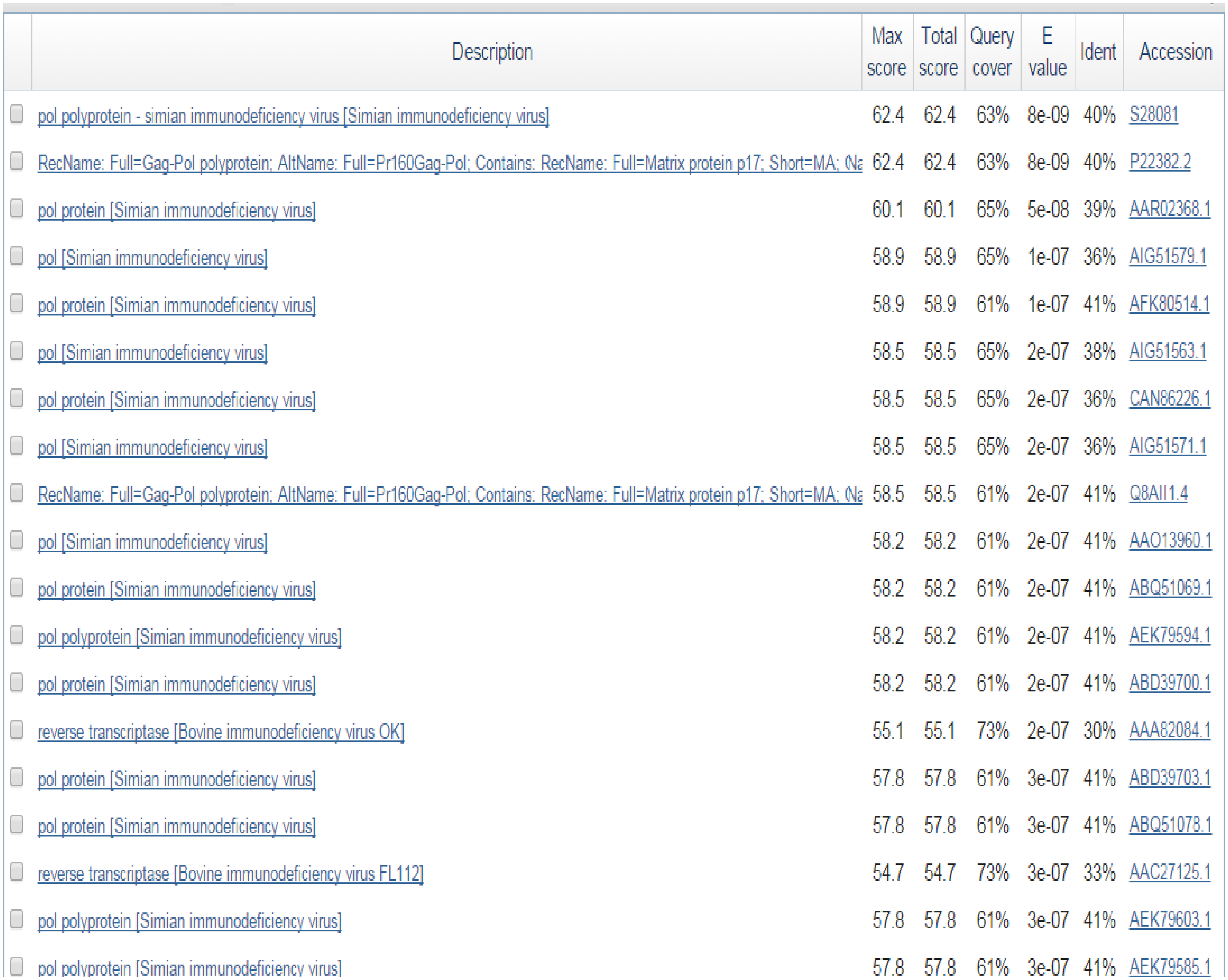


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C7407449 after BLASTx search



	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	pol polyprotein - simian immunodeficiency virus [Simian immunodeficiency virus]	62.4	62.4	63%	8e-09	40%	S28081
□	RecName: Full=Gag-Pol polyprotein; AltName: Full=Pr160Gag-Pol; Contains: RecName: Full=Matrix protein p17; Short=MA; (Na	62.4	62.4	63%	8e-09	40%	P22382.2
□	pol protein [Simian immunodeficiency virus]	60.1	60.1	65%	5e-08	39%	AAR02368.1
□	pol [Simian immunodeficiency virus]	58.9	58.9	65%	1e-07	36%	AIG51579.1
□	pol protein [Simian immunodeficiency virus]	58.9	58.9	61%	1e-07	41%	AFK80514.1
□	pol [Simian immunodeficiency virus]	58.5	58.5	65%	2e-07	38%	AIG51563.1
□	pol protein [Simian immunodeficiency virus]	58.5	58.5	65%	2e-07	36%	CAN86226.1
□	pol [Simian immunodeficiency virus]	58.5	58.5	65%	2e-07	36%	AIG51571.1
□	RecName: Full=Gag-Pol polyprotein; AltName: Full=Pr160Gag-Pol; Contains: RecName: Full=Matrix protein p17; Short=MA; (Na	58.5	58.5	61%	2e-07	41%	Q8AII1.4
□	pol [Simian immunodeficiency virus]	58.2	58.2	61%	2e-07	41%	AAO13960.1
□	pol protein [Simian immunodeficiency virus]	58.2	58.2	61%	2e-07	41%	ABQ51069.1
□	pol polyprotein [Simian immunodeficiency virus]	58.2	58.2	61%	2e-07	41%	AEK79594.1
□	pol protein [Simian immunodeficiency virus]	58.2	58.2	61%	2e-07	41%	ABD39700.1
□	reverse transcriptase [Bovine immunodeficiency virus OK]	55.1	55.1	73%	2e-07	30%	AAA82084.1
□	pol protein [Simian immunodeficiency virus]	57.8	57.8	61%	3e-07	41%	ABD39703.1
□	pol protein [Simian immunodeficiency virus]	57.8	57.8	61%	3e-07	41%	ABQ51078.1
□	reverse transcriptase [Bovine immunodeficiency virus FL112]	54.7	54.7	73%	3e-07	33%	AAC27125.1
□	pol polyprotein [Simian immunodeficiency virus]	57.8	57.8	61%	3e-07	41%	AEK79603.1
□	pol polyprotein [Simian immunodeficiency virus]	57.8	57.8	61%	3e-07	41%	AEK79585.1

3.3.2. BLAST alignment for the contig C4274868 (151 BP)

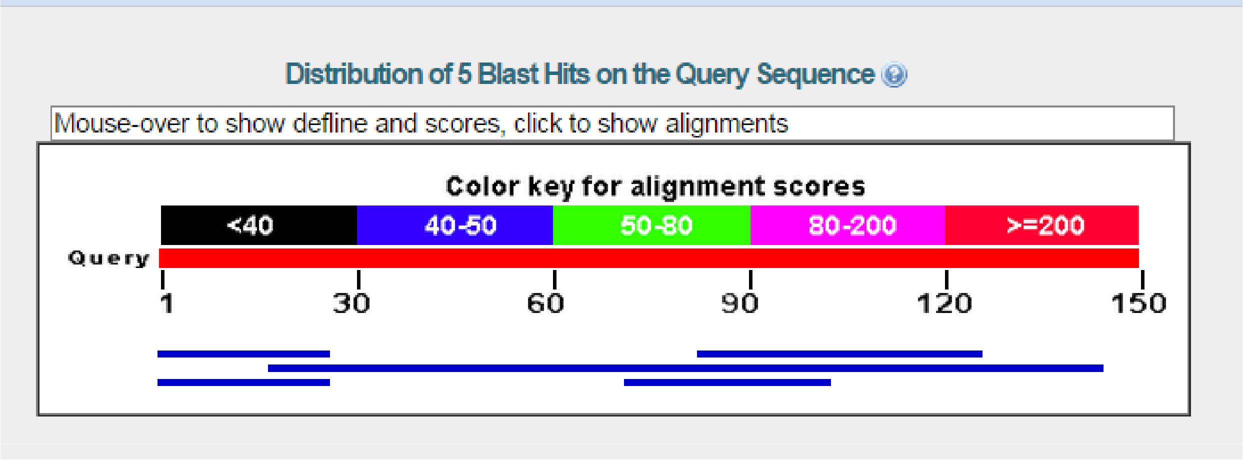
Nucleotide sequence in FASTA format

```
>C4274868
TCACATCAGACAGCAATGATACCAATTGGCGTTGCATTTGAGATAGAAAAGGGTG
TGTCGGTCTTATTTGGGATAAGTCGTCGATTGGTTCAAAAAGCTTAAAGACACTAGG
TGGGGTGATAGATGCTGGGTACCGTGGTGAAGTGTCAG
```

BLASTn alignment for the contig C4274868

BLASTn search result showed that the assembled query sequence was similar to sequence of Sumatran orangutan (*Pongo abelii*). However, that alignment was not significant as the alignment showed higher E-value. For the top hit, E-value was 0.36, matched sequence was 46 bp long, and identity was 83% with 29% of query coverage.

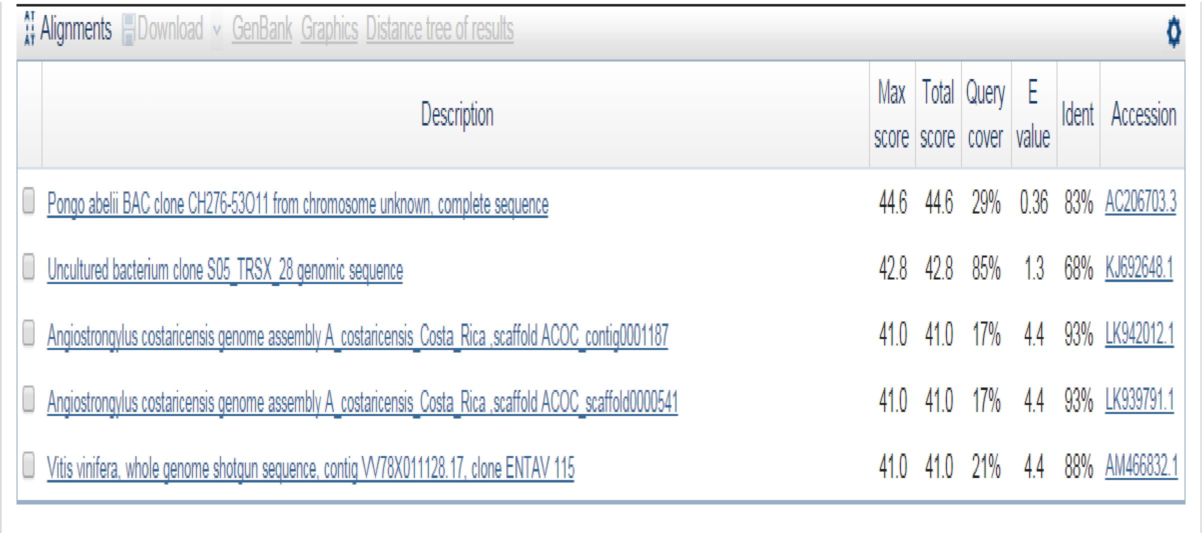
A) Graphic display



Distribution of 5 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C4274868 after BLASTn search

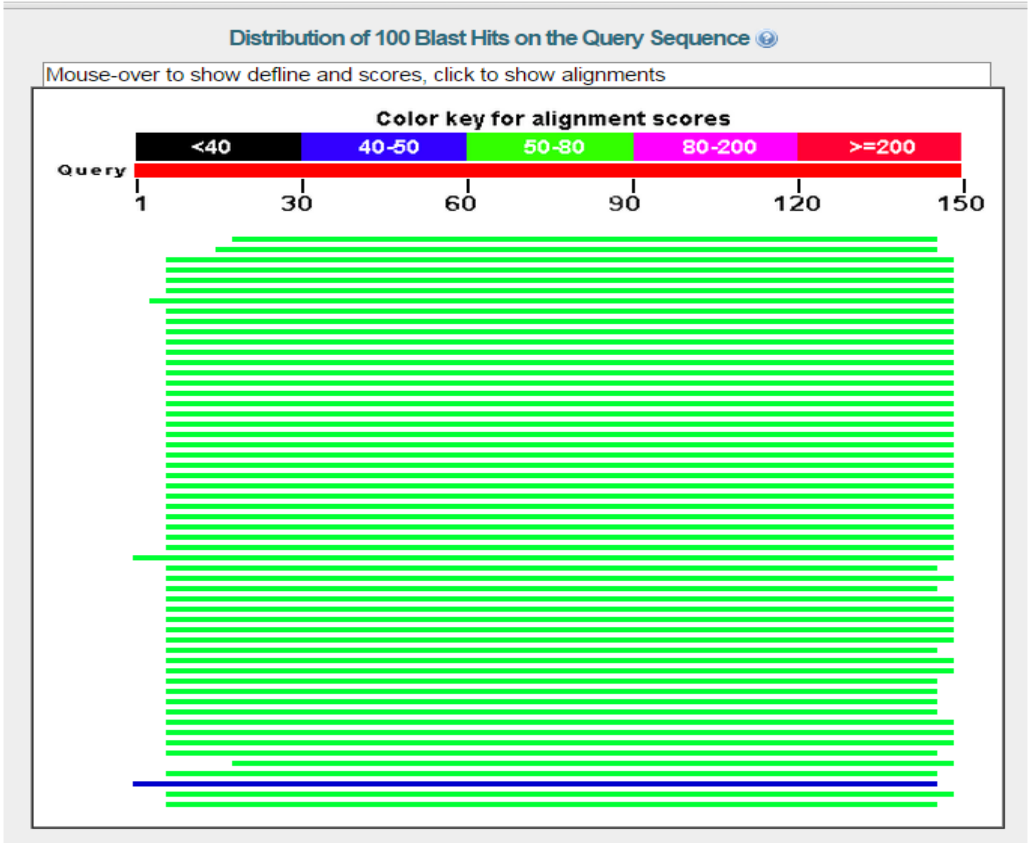


	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Pongo abelii BAC clone CH276-53O11 from chromosome unknown, complete sequence	44.6	44.6	29%	0.36	83%	AC206703.3
□	Uncultured bacterium clone S05_TRSX_28 genomic sequence	42.8	42.8	85%	1.3	68%	KJ692648.1
□	Angiostrongylus costaricensis genome assembly A_costaricensis_Costa_Rica ,scaffold ACOC_contig0001187	41.0	41.0	17%	4.4	93%	LK942012.1
□	Angiostrongylus costaricensis genome assembly A_costaricensis_Costa_Rica ,scaffold ACOC_scaffold0000541	41.0	41.0	17%	4.4	93%	LK939791.1
□	Vitis vinifera, whole genome shotgun sequence, contig VV78X011128.17, clone ENTAV 115	41.0	41.0	21%	4.4	88%	AM466832.1

BLASTx alignment for the contig C4274868

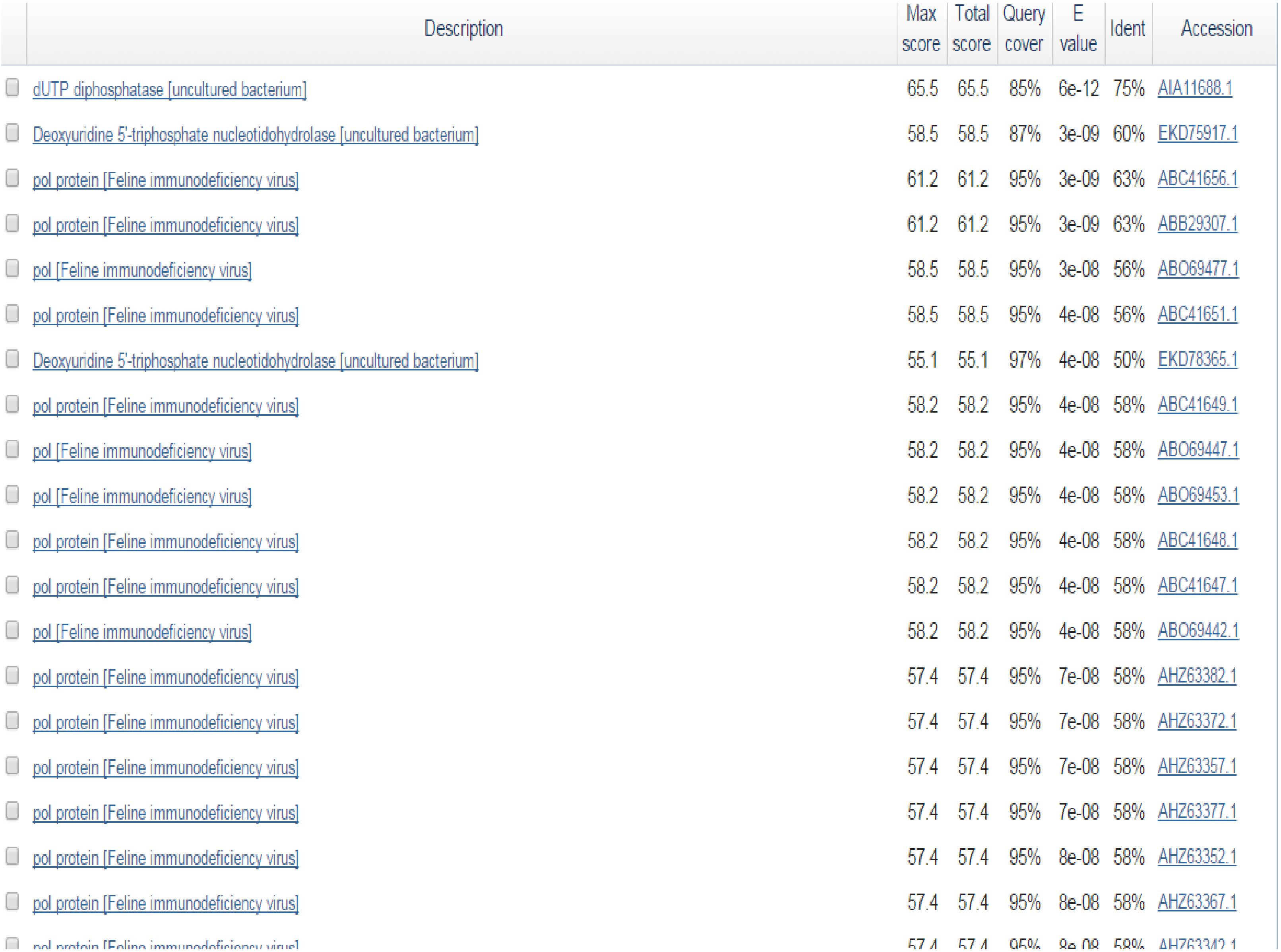
BLASTx search result showed assembled query sequence was similar to the protein coding sequence of trimeric dUTP diphosphatase protein superfamily of feline immunodeficiency virus (Trimeric dUTP diphosphatases, or dUTPases, are the most common family of dUTPase, found in bacteria, eukaryotes, and archaea). For the best aligned sequence against the query sequence, identity was 75%, matched length was 30 amino acid out of total 48 amino acids, query coverage was 85%, and E value was 6e-12. The sequence was ambiguous due to short length with best match as bacterial sequence (75%) alignment. It suggests that this sequence was most likely originated from a bacteria.

A) Graphic display



Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
30
60
90
120
150

B) Hit list of aligned NCBI sequences for the contig C4274868 after BLASTx search



	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	dUTP diphosphatase [uncultured bacterium]	65.5	65.5	85%	6e-12	75%	AIA11688.1
□	Deoxyuridine 5'-triphosphate nucleotidohydrolase [uncultured bacterium]	58.5	58.5	87%	3e-09	60%	EKD75917.1
□	pol protein [Feline immunodeficiency virus]	61.2	61.2	95%	3e-09	63%	ABC41656.1
□	pol protein [Feline immunodeficiency virus]	61.2	61.2	95%	3e-09	63%	ABB29307.1
□	pol [Feline immunodeficiency virus]	58.5	58.5	95%	3e-08	56%	ABO69477.1
□	pol protein [Feline immunodeficiency virus]	58.5	58.5	95%	4e-08	56%	ABC41651.1
□	Deoxyuridine 5'-triphosphate nucleotidohydrolase [uncultured bacterium]	55.1	55.1	97%	4e-08	50%	EKD78365.1
□	pol protein [Feline immunodeficiency virus]	58.2	58.2	95%	4e-08	58%	ABC41649.1
□	pol [Feline immunodeficiency virus]	58.2	58.2	95%	4e-08	58%	ABO69447.1
□	pol [Feline immunodeficiency virus]	58.2	58.2	95%	4e-08	58%	ABO69453.1
□	pol protein [Feline immunodeficiency virus]	58.2	58.2	95%	4e-08	58%	ABC41648.1
□	pol protein [Feline immunodeficiency virus]	58.2	58.2	95%	4e-08	58%	ABC41647.1
□	pol [Feline immunodeficiency virus]	58.2	58.2	95%	4e-08	58%	ABO69442.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	7e-08	58%	AHZ63382.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	7e-08	58%	AHZ63372.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	7e-08	58%	AHZ63357.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	7e-08	58%	AHZ63377.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	8e-08	58%	AHZ63352.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	8e-08	58%	AHZ63367.1
□	pol protein [Feline immunodeficiency virus]	57.4	57.4	95%	8e-08	58%	AHZ63342.1

3.3.3. BLAST alignment for the contig C6591025 (249 BP)

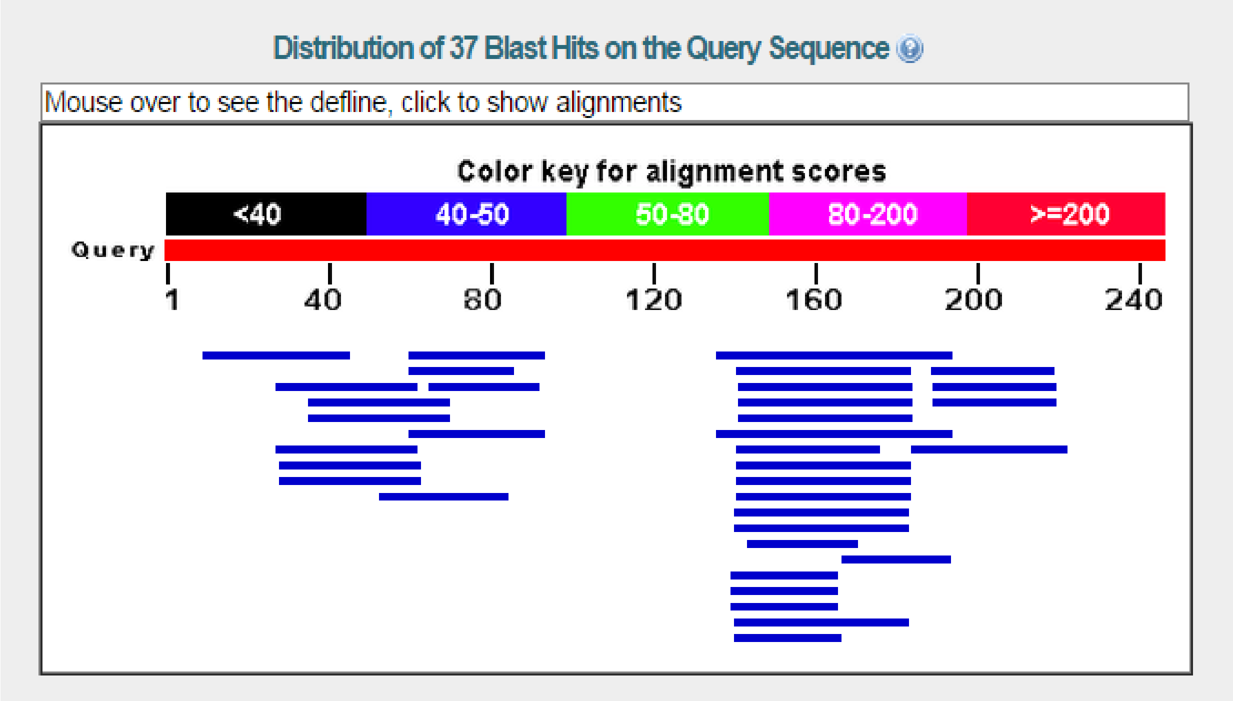
Nucleotide sequence in FASTA format

```
>C6591025
GTTGGCTTACTACGCGAACGTTACTCTATTGGCAAAAAGGGACTTAAGGTAGCTGG
CGGAGTTATTGATGTTGGCTATTCTGGAGATATTTCAGTACTCCTAATGAATGTTAG
CGCAGACAACTTAGATATGCGAGGATACGAAATTAAACCAGGGGATAGGATTGCTC
AAATTCTTATTGTACCTGCAGATAGTTATCCAATTGTAGTAGTAGATACTCTTTGGA
CTAGTGAACGCGGAAGTAAATCT
```

BLASTn alignment for the contig C6591025

BLASTn search result showed assembled query sequence was similar to the sequence of nematode (*Wolbachia endosymbionts*). However, these alignment was not significant as the alignment had higher E-value of 0.055, 23% of total coverage (65 bp matched) and the identity was 77% with the matched region.

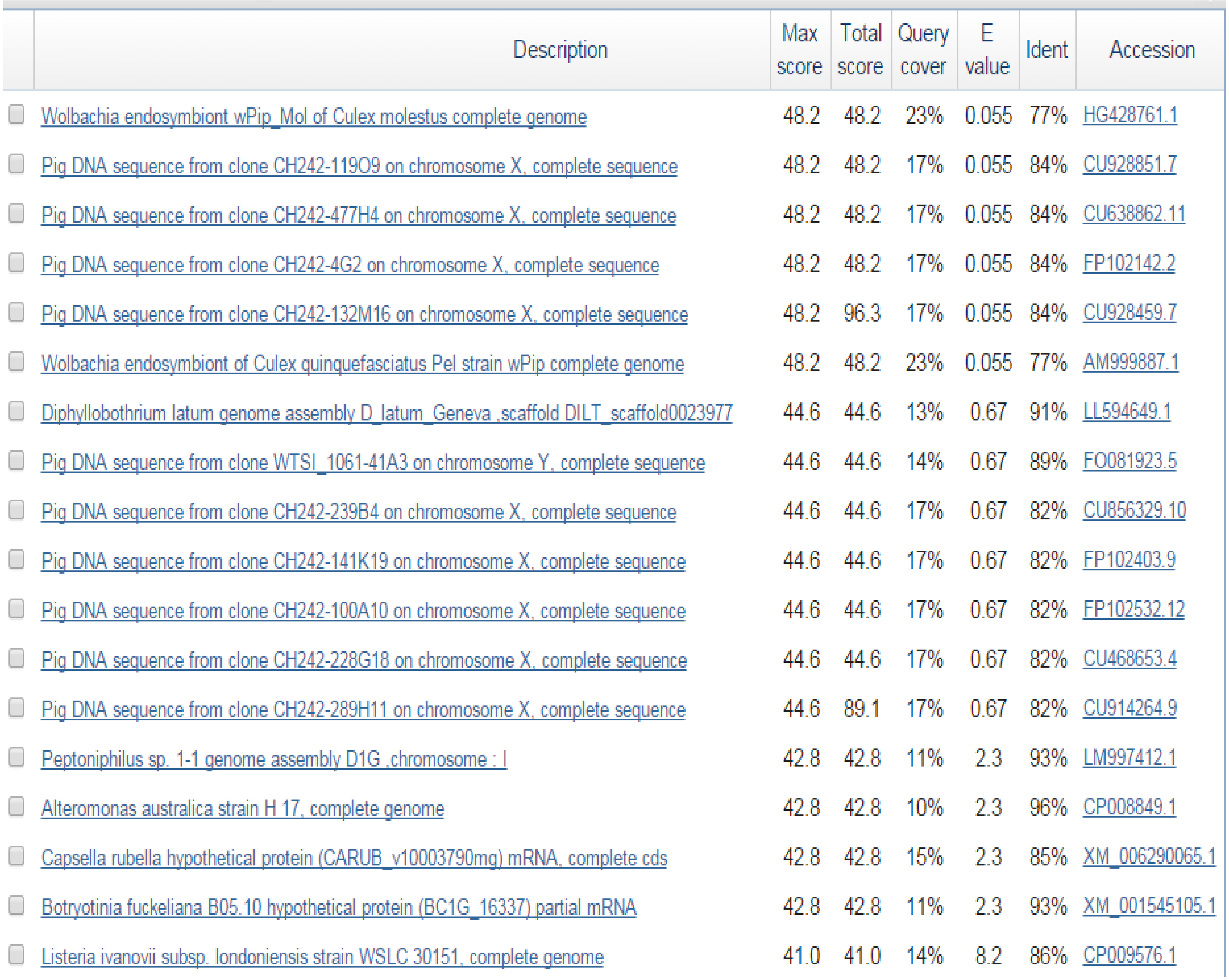
A) Graphic display



Distribution of 37 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C6591025 after BLASTn search

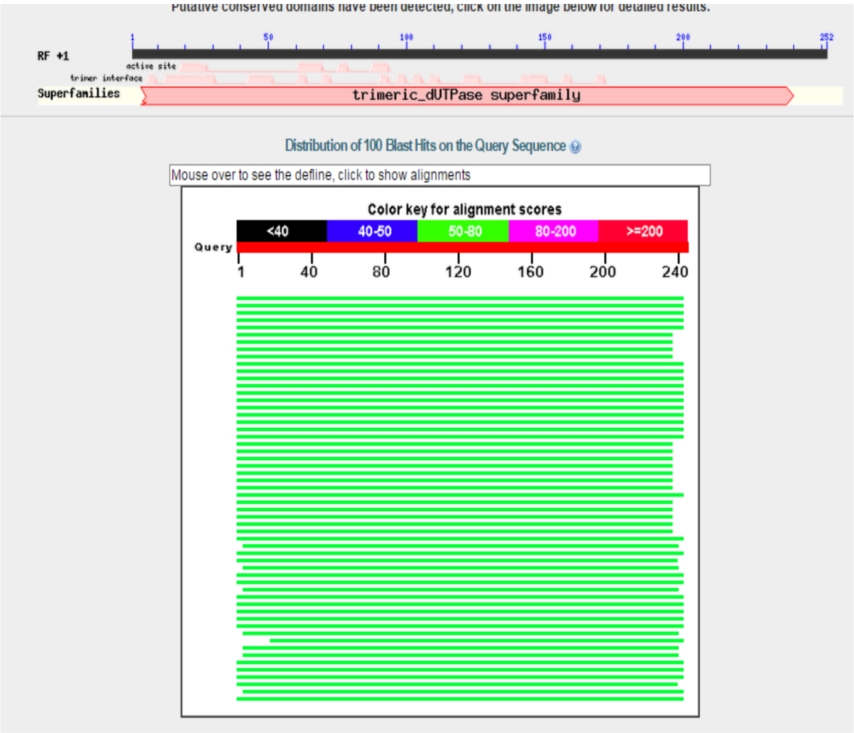


	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Wolbachia endosymbiont wPip_Mol of Culex molestus complete genome	48.2	48.2	23%	0.055	77%	HG428761.1
☐	Pig DNA sequence from clone CH242-119O9 on chromosome X, complete sequence	48.2	48.2	17%	0.055	84%	CU928851.7
☐	Pig DNA sequence from clone CH242-477H4 on chromosome X, complete sequence	48.2	48.2	17%	0.055	84%	CU638862.11
☐	Pig DNA sequence from clone CH242-4G2 on chromosome X, complete sequence	48.2	48.2	17%	0.055	84%	FP102142.2
☐	Pig DNA sequence from clone CH242-132M16 on chromosome X, complete sequence	48.2	96.3	17%	0.055	84%	CU928459.7
☐	Wolbachia endosymbiont of Culex quinquefasciatus Pel strain wPip complete genome	48.2	48.2	23%	0.055	77%	AM999887.1
☐	Diphyllobothrium latum genome assembly D_latum_Geneva ,scaffold DILT_scaffold0023977	44.6	44.6	13%	0.67	91%	LL594649.1
☐	Pig DNA sequence from clone WTSI_1061-41A3 on chromosome Y, complete sequence	44.6	44.6	14%	0.67	89%	FO081923.5
☐	Pig DNA sequence from clone CH242-239B4 on chromosome X, complete sequence	44.6	44.6	17%	0.67	82%	CU856329.10
☐	Pig DNA sequence from clone CH242-141K19 on chromosome X, complete sequence	44.6	44.6	17%	0.67	82%	FP102403.9
☐	Pig DNA sequence from clone CH242-100A10 on chromosome X, complete sequence	44.6	44.6	17%	0.67	82%	FP102532.12
☐	Pig DNA sequence from clone CH242-228G18 on chromosome X, complete sequence	44.6	44.6	17%	0.67	82%	CU468653.4
☐	Pig DNA sequence from clone CH242-289H11 on chromosome X, complete sequence	44.6	89.1	17%	0.67	82%	CU914264.9
☐	Peptoniphilus sp. 1-1 genome assembly D1G ,chromosome : I	42.8	42.8	11%	2.3	93%	LM997412.1
☐	Alteromonas australica strain H 17, complete genome	42.8	42.8	10%	2.3	96%	CP008849.1
☐	Capsella rubella hypothetical protein (CARUB_v10003790mg) mRNA, complete cds	42.8	42.8	15%	2.3	85%	XM_006290065.1
☐	Botryotinia fuckeliana B05.10 hypothetical protein (BC1G_16337) partial mRNA	42.8	42.8	11%	2.3	93%	XM_001545105.1
☐	Listeria ivanovii subsp. londoniensis strain WSLC 30151, complete genome	41.0	41.0	14%	8.2	86%	CP009576.1

BLASTx alignment for the contig C6591025

BLASTx search result showed assembled query sequence was similar to the protein coding sequence of trimeric dUTP diphosphatase protein superfamily of feline immunodeficiency virus (Trimeric dUTP diphosphatases, or dUTPases, are the most common family of dUTPase, found in bacteria, eukaryotes, and archaea). For the best aligned sequence against the query sequence, identity was from 45%, matched length was 37 amino acid out of total 82 amino acids, query coverage was 98% at E value of 4e-10. Within 5 top hits, one sequence was from *Marinitoga piezophila*, a bacteria. This sequence probably was also originated from a bacterial sequence that had some sequence similarity to *Retroviridae* family.

A) Graphic display

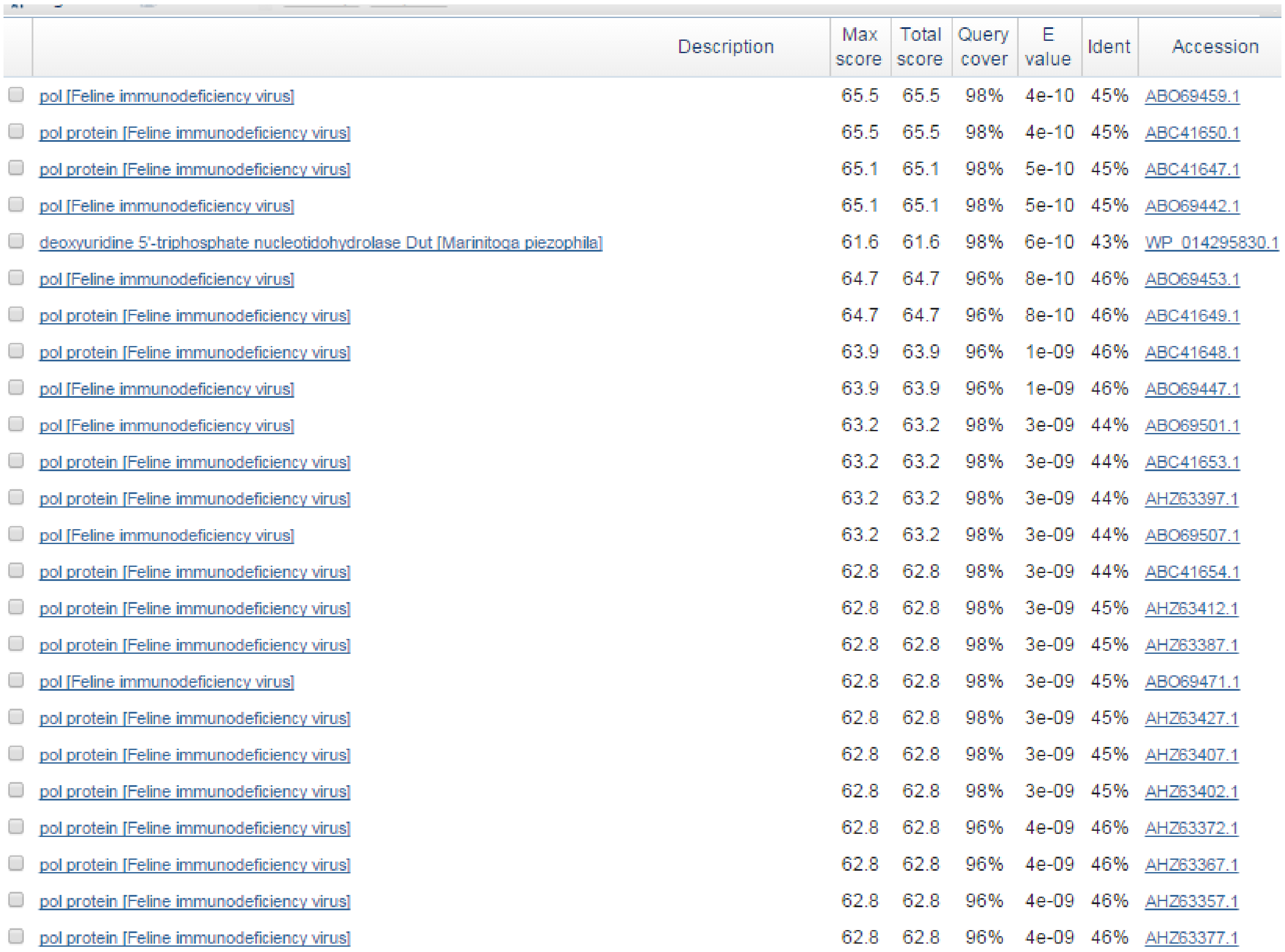


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

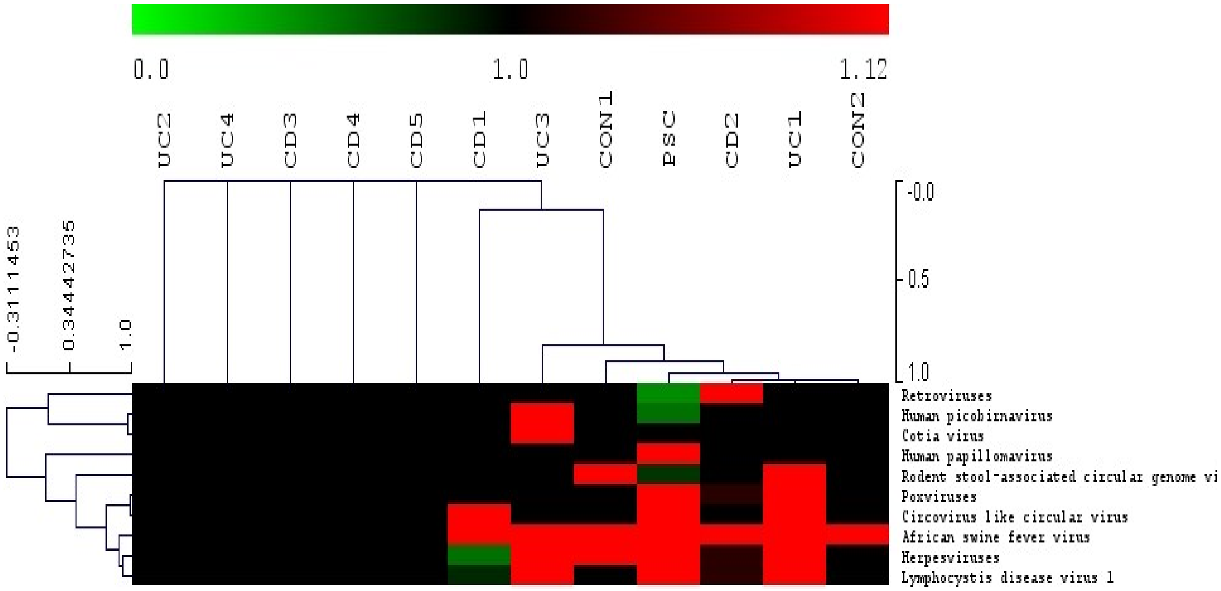
Mouse over to see the defline, click to show alignments

B) Hit list of aligned NCBI sequences for the contig C6591025 after BLASTx search

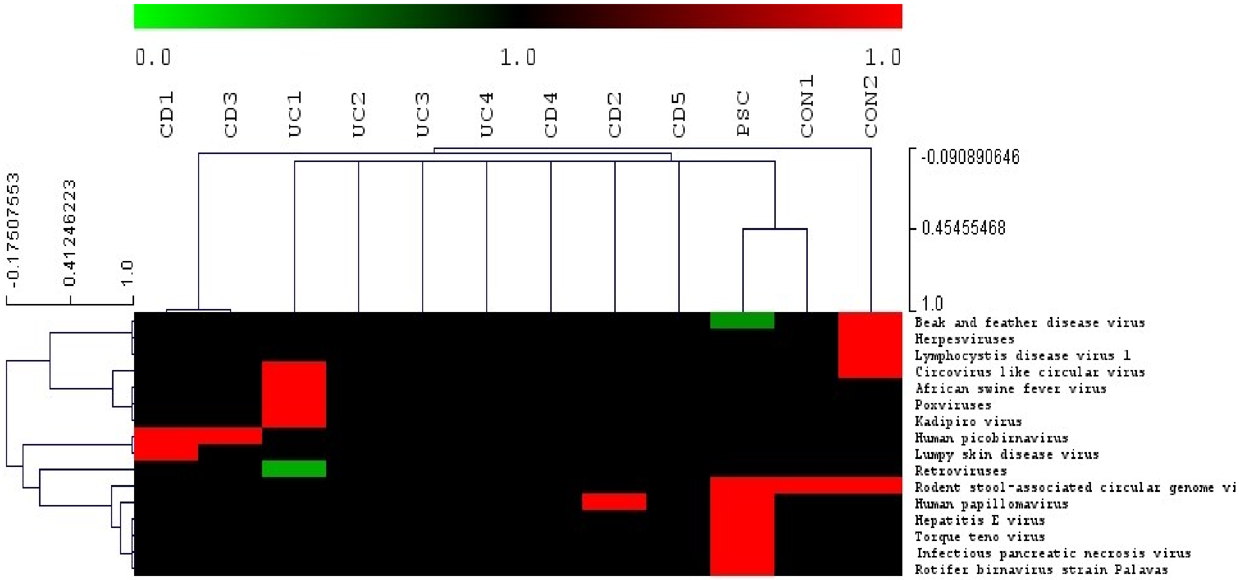


	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	pol [Feline immunodeficiency virus]	65.5	65.5	98%	4e-10	45%	ABO69459.1
□	pol protein [Feline immunodeficiency virus]	65.5	65.5	98%	4e-10	45%	ABC41650.1
□	pol protein [Feline immunodeficiency virus]	65.1	65.1	98%	5e-10	45%	ABC41647.1
□	pol [Feline immunodeficiency virus]	65.1	65.1	98%	5e-10	45%	ABO69442.1
□	deoxyuridine 5'-triphosphate nucleotidohydrolase Dut [Marinitoga piezophila]	61.6	61.6	98%	6e-10	43%	WP_014295830.1
□	pol [Feline immunodeficiency virus]	64.7	64.7	96%	8e-10	46%	ABO69453.1
□	pol protein [Feline immunodeficiency virus]	64.7	64.7	96%	8e-10	46%	ABC41649.1
□	pol protein [Feline immunodeficiency virus]	63.9	63.9	96%	1e-09	46%	ABC41648.1
□	pol [Feline immunodeficiency virus]	63.9	63.9	96%	1e-09	46%	ABO69447.1
□	pol [Feline immunodeficiency virus]	63.2	63.2	98%	3e-09	44%	ABO69501.1
□	pol protein [Feline immunodeficiency virus]	63.2	63.2	98%	3e-09	44%	ABC41653.1
□	pol protein [Feline immunodeficiency virus]	63.2	63.2	98%	3e-09	44%	AHZ63397.1
□	pol [Feline immunodeficiency virus]	63.2	63.2	98%	3e-09	44%	ABO69507.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	44%	ABC41654.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	45%	AHZ63412.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	45%	AHZ63387.1
□	pol [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	45%	ABO69471.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	45%	AHZ63427.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	45%	AHZ63407.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	98%	3e-09	45%	AHZ63402.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	96%	4e-09	46%	AHZ63372.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	96%	4e-09	46%	AHZ63367.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	96%	4e-09	46%	AHZ63357.1
□	pol protein [Feline immunodeficiency virus]	62.8	62.8	96%	4e-09	46%	AHZ63377.1

4.1. Hierarchical clustering for viruses detected in the RNA libraries of individual patients



4.2. Hierarchical clustering for viruses detected in the DNA libraries of individual patients



5. Human picobirnavirus RNA dependent RNA polymerase used in the phylogenetic analysis

Name of RdRp	Genbank ID
HumanP_US1	GI:261865298
HumanP_US2	GI:261865296
HumanP_US3	GI:261865306
HumanP_US4	GI:261865292
HumanP_US5	GI:261865290
HumanP_US6	GI:12407606
MouseP_US	GI:343196972
FoxP_Netherland	GI:645393520
MicrotusP_US	GI:343196974
PorcineP_Italy	GI:583925938
TurkeyP_US	GI:629511266
HumanP_US8	GI:12407606
PorcineP_China1	GI:594499180
PorcineP_China2	GI:594499183
PorcineP_China3	GI:594499174
HumanP_Pakistan	GI:261865226
FelineP_Portugal	GI:557361107

6.1 Human papillomavirus capsid and transcription regulatory protein used in the phylogenetic analysis

Name of HPV capsid protein	Genbank ID
HPV_type_123	GI:296495866
Human papillomavirus type 139	GI:343411551
Human papillomavirus type 170	GI: 409183147
Human papillomavirus type 155	GI:353441726
Human papillomavirus type 138	GI:343411543
Human papillomavirus type 109	GI:225927566
Human papillomavirus type 134	GI:319962667
papillomavirus type 149	GI:312451788
Human papillomavirus type 180	GI:443498384
Human papillomavirus type 121	GI:297342362
Human papillomavirus type 173	GI:564732537
Human papillomavirus type 133	GI:312451820
Human papillomavirus type 142	GI:343411575
Human papillomavirus type 130	GI:312451796
Human papillomavirus type 8	GI:333074
Human papillomavirus type 99	GI:238623435
Human papillomavirus type 105	GI:238623458
Human papillomavirus type 5	GI:68159730
Human papillomavirus type 5	GI:484221
Human papillomavirus type 5b	GI:222402
Human papillomavirus type 47	GI:333062
HPV_type_36.1	GI: 896393
HPV_type_36.2	GI: 623455

6.2. Human papillomavirus transcription regulatory protein used in the phylogenetic analysis

Name of HPV transcription regulatory protein (E1)	Genbank ID
HPV_type_4	GI:312084
Human papillomavirus type 65	GI:312100
Human papillomavirus type 95	GI:40804520
Human papillomavirus type 173	GI:564732532
Human papillomavirus type 163	GI:409183121
Human papillomavirus type 156	GI:410443770
Human papillomavirus isolate 915	GI:371486204
Human papillomavirus type 6	GI:6002612
Human papillomavirus type 11 complete genome	GI:335334258
Human papillomavirus - 18	GI:9626069
Human papillomavirus type 31 isolate QV12357	GI:337238071
Human papillomavirus type 53	GI:9627377
Human papillomavirus type 58 complete genome	GI:222386
Human papilloma virus type 59	GI:557236

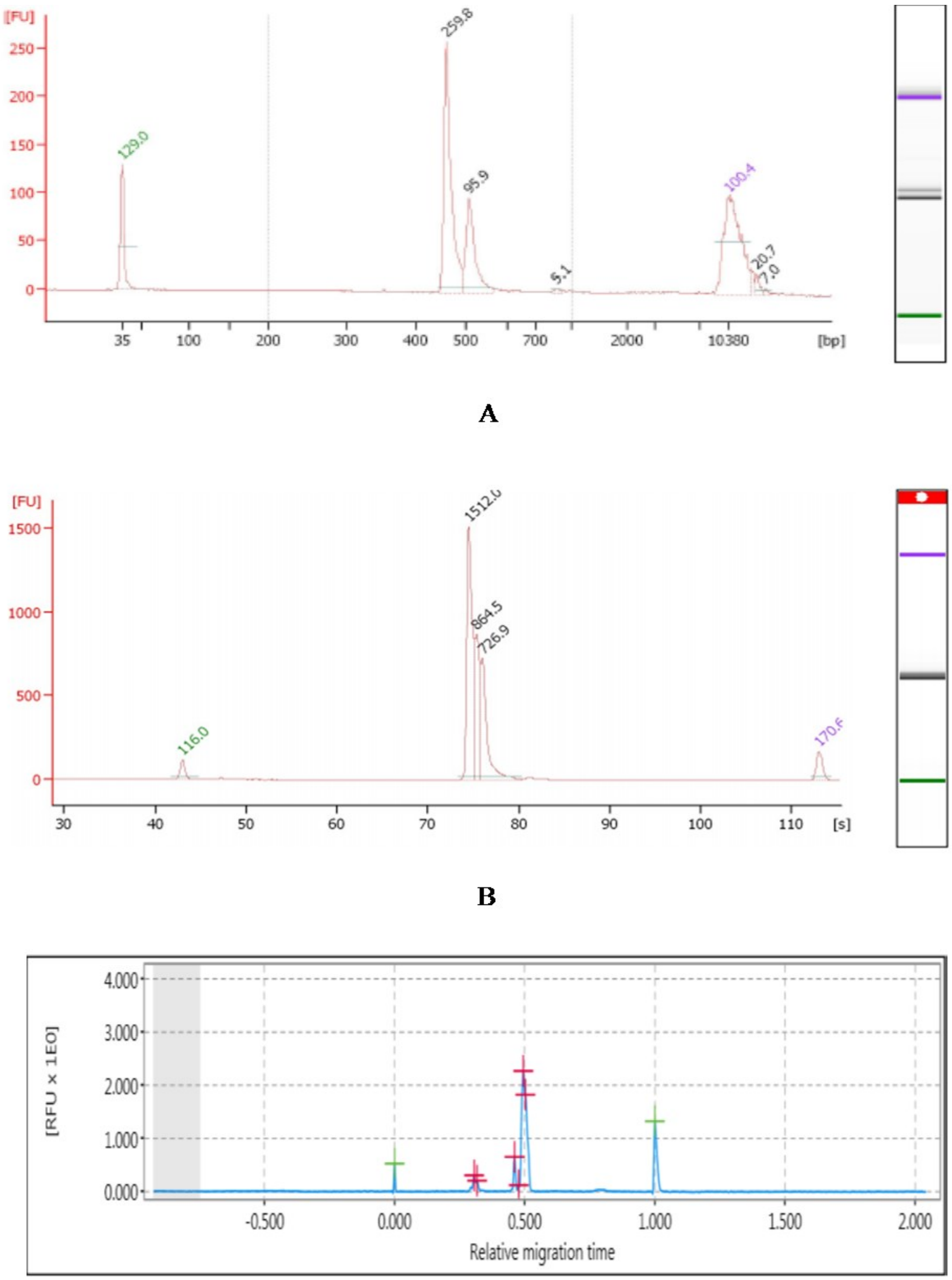
7. African swine fever virus used in the phylogenetic analysis

Name of ASF virus	Genbank ID
BA71V	GI: 9628113
Benin 97/1	GI: 162849209
Kenya 1950	GI 33772320
ASFV_Malawi_Lil-20/1	GI :33772321
ASFV_Mkuzi_1979	GI: 33772322
Pretorisuskop/96/4	GI: 33772323
Tengani_62	GI: 33772324
ASFV_Warmbaths	GI: 33772325
Warthog	GI: 33772326
ASFV_ E75	GI:291289440
OURT 88/3	GI: 162849383
ASFV_Georgia_2007/1	GI: 303398661

8. Different protein coding contigs/scaffolds of ASFLV are quite divergent from the proteins of ASFV at the NCBI database (from pig). Scaffolds were searched against protein database using BLASTx, and the level of identity of sequences with matched database proteins was determined.

Sequence ID	Length of Sequence (BP)	Match (AA)	Identity at amino acid level (%)	Matched Protein
scaffold36309	257	56	57.14	Cysteine protease S273R;
scaffold68529	285	44	56.82	Capsid protein
scaffold84041	274	41	53.66	Topoisomerase
scaffold152037	456	63	38.1	Topoisomerase
C5493134	210	63	47.62	Putative DNA primase
C7573228	335	41	51.22	220kDa Polyprotein
C8357811	1939	59	42.37	Origin binding protein
C8080867	530	99	42	RNApol1
C8031519	489	83	46.99	p72/capsid protein
C7902053	418	56	38.71	DNA ligase
C7753971	371	39	53.85	DNA polymerase
C7730295	366	20	60	capping enzyme large subunit
C7735476	367	71	45.07	Poly (A) polymerase-large subunit
C7456564	318	50	62.07	RNA helicase
C7457362	318	103	42.72	Proliferating cell nuclear antigen
C7317142	301	59	55.93	ATP- or GTP-binding motif
C7100259	279	37	40.54	8-Hydroxy-dGTpase
C6995270	273	45	62.5	Alpha-NAC binding/serine proteinase inhibitor

9. Confirmation of PCR products (sequences similar to ASFLV origin binding protein, helicase and capsid coding genes) amplification by Bioanlyzer (A and B) or QIAxcel A) Origin binding protein region, partially amplified by nested PCR, as shown in the electropherogram. B) Partial amplification of helicase protein by nested PCR, as shown in the electropherogram. C) Partial amplification of capsid protein by nested PCR, as shown in the electropherogram.



A

B

C

10. Summary of the PCR experiment seeking to detect ASFV-like virus in samples of other body parts of the two positive patients.

Patient	Types	Capsid	Concentration of DNA (ng/µL)	Amount of DNA used as template(ng in 25 µL PCR reaction mix)
PSC	Plasma (4/3/2011)	-	6.5	30
	Plasma (14/2/2012)	+	5.5	25
	White Blood cells (14/2/2012)	-	465	50/100
	Liver (1/2/2014)	-	2655	50/100
UC1	Colon tissue	-	1221	50/100
	Lymph node (colon)	-	627	50/100

11. Test for ASFLV-like sequences in other patients

Patient	Concentration of DNA (ng/µl)	Amount of DNA used as template(ng in 25 µL PCR reaction mix)	Result
CDC05	19.9	50/100	-
CDC06	550	50/100	-
CDC10	538	50/100	-
CDC12	1433	50/100	-
CDC19	1642	50/100	-
UCC01	584	50/100	-
UCC04	1554	50/100	-
UCC06	1818	50/100	-
UCC07	73.9	50/100	-
UCC13	1788	50/100	-

12. Primer sets used for PCR experiments

Protein	**Types**	**Primers**	**Round of amplification**
Capsid	Forward	CCGATCCAATCACAATAACGG	First
	Reverse	ACGAAATTACTTTCGCAGCG	First
	Forward	CACAATAACGGTAAAGAGCGC	Second
	Reverse	ACTTTCGCAGCGTATTCAATC	Second
Helicase protein	Forward	GGAAATTACGAATCTTCTCGAGTG	First
	Reverse	GTAACAAATACCTCTGCGGTTG	First
	Forward	CGGACAGGATACGAACGAAC	Second
	Reverse	CTGCGGTTGCTCGAGGTATAG	Second
Origin binding	Forward	ACGGTGTAGTGATTCAACCTG	First
	Reverse	CTGAGACAGATCTCGCGATTTAC	First
	Forward	CGATAATACTTGGGAAATCGTCG	Second
	Reverse	CATCATAGAGCACATGATCGC	Second

13. Distribution of the major taxonomic groups of sequences using entire datasets. BLASTn and BLASTx were used to recognize the assembled longer sequences (contigs/scaffolds). Sequences are grouped based on their sequence similarity (BLASTn) with and protein homology (BLASTx) to bacteria, humans, plants, phages and viruses

Taxonomy	Reads count	Reads count (% by count)	Contigs/ Scaffolds count	Contigs/ Scaffold count (% by count)	Contigs/ Scaffold length (bp)	Contigs/ Scaffold length (% by length)	Size range (bp)
Bacteria	92,766,644	58.29	1,024,798	37.65	423,342,611	52.2	145,465-100
Human	2,725,951	1.71	252,300	9.27	54,634,574	6.74	26,975-100
Plant	1,904,643	1.19	16175	0.6	4,335,416	0.53	35,575-100
Phage	8,173,600	5.13	32,412	1.19	17,711,247	2.18	172,009-100
Virus	1,674,168	1.05	2492	0.1	996,354	0.12	6,397-100