The Assembly and Analysis of 60,000 Bases of the Nuclear Genome of *Trichinella murrelli*

By

Travis Charlton

April, 2012



Allegheny College

Bioinformatics Program

The Assembly and Analysis of 60,000 Bases of the Nuclear Genome of *Trichinella murrelli*

By

Travis Charlton

April, 2012

A Senior Comprehensive Project in Partial Fulfillment of the Requirements for a Bachelor of Science Degree from Allegheny College

I hereby recognize and pledge to fulfill my responsibilities as defined in the Honor Code

and to maintain the integrity of both myself and the College community as a whole.

_____

Pledge

Approval by the Senior Thesis Committee:

_____

Dr. Kristen Webb

_____

Dr. Matthew Jadud

Table of Contents

Abstract……

Introduction……..….

Materials and Methods……...

Results……

Discussion……..

Acknowledgements……...……..

Literature Cited……...

Appendices…….

Abstract:

Analysis of single nucleotide polymorphisms (SNPs) is one of the most popular methods in modern science for determining levels of diversity between two or more organisms. Short reads of the *T. murrelli* nuclear genome, collected via next-generation sequencing, were assembled and aligned to a reference contig of the *T. spiralis* nuclear genome that was approximately 60,000 bases long. No *T. murrelli* nuclear reference genome data was available, so *T. murrelli* short reads were assembled to a reference contig of the *T. spiralis* nuclear genome due to necessity for a reference sequence for the assembly process. This marks the first time that reads from one species were mapped to a reference contig of another species. Approximately 94% coverage of the reference genome was achieved. SNP analysis of the assembly showed a SNP occurrence rate of 1 in 60 bases, and a SNP occurrence rate of 1 in 73 bases was determined over the exons of the 17 genes in the assembly. Also, 87 non-synonymous changes were found in the exons. These results show distinct diversity between the two organisms across the 60,000 bases, and serve as a baseline of estimation of diversity between the two species across the entire genome. This project marks the first successful assembly of *T. murrelli* reads to a *T. spiralis* reference sequence, and the first cross species assembly of any kind.

Introduction:

Trichinella spiralis is a parasitic, microscopic nematode that has been shown to infect humans, rats, and swine, but is best known for its ability to infect humans (Mitreva et al. 2011). Upon infection, humans may develop trichinosis, which can be deadly in severe cases. One potential way that *T. spiralis* can infect its hosts is via the path shown in Figure 1. A pig eats a rat infected with *T. spiralis*. When the infected pig dies of trichinosis, it is then eaten by other pigs. These infected pigs are then slaughtered and consumed by humans, causing the onset of trichinosis.

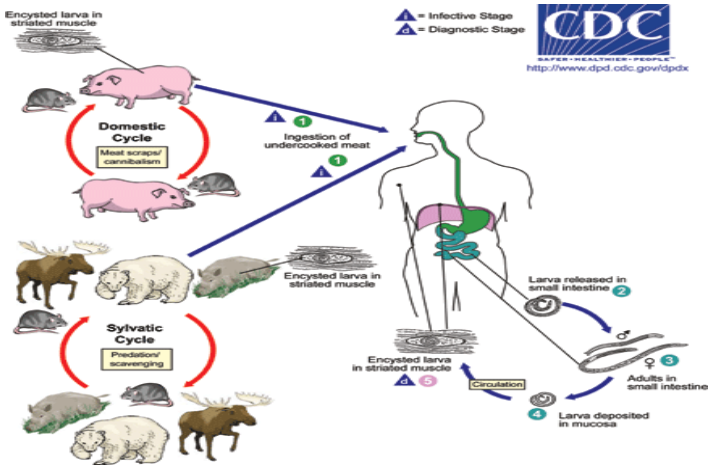


Figure 1: Parasite life-cycle obtained from the CDC website (http://dpd.cdc.gov/dpdx/HTML/Image_Library.htm).

Though trichinosis has been eradicated in the commercial pork supply of the United States with the use of careful agricultural practices, it still remains a serious problem in developing nations around the world (Rosenthal et al. 2008).

Trichinella murrelli, a member of the same genus as *T. spiralis*, is a parasite that is found in North America (Rosenthal et al. 2008). Unlike its counterpart though, *T. murrelli* is not capable of sustaining an infection in swine. However, it is able to sustain infection in a wide variety of mammals including rats, foxes, black bears, coyotes, gray wolves, raccoons, and many others (Hill et al. 2008). In 2006, Dubey et al. performed a study where a *T. murrelli* isolate was taken from a female beagle, grown, and fed to rats, which subsequently became infected (Dubey et al. 2006). Not only was this the first diagnosed case of a companion animal being infected by *T. murrelli*, it demonstrated that a single isolate of *T. murrelli* could potentially be capable of sustaining infections in an incredibly wide range of animals, from wild sylvatic mammals to domesticated companion animals. There is also data that suggests that swine previously infected by *T. murrelli* will not only survive the infection due to *T. murrelli's* inability to sustain infection

in swine, but might potentially gain some form of immunity from further infection from *T. spiralis* (Hill, unpublished).

The first ever draft genome of *T. spiralis* was sequenced and assembled totaling approximately 64 Mb (64 million bases) in length (Mitreva et al. 2011). The majority of the genome was assembled into 9267 contigs, which are assemblies of individual sequences grouped together based on similarity of nucleotide composition in order to form one longer sequence of DNA (Figure 2).

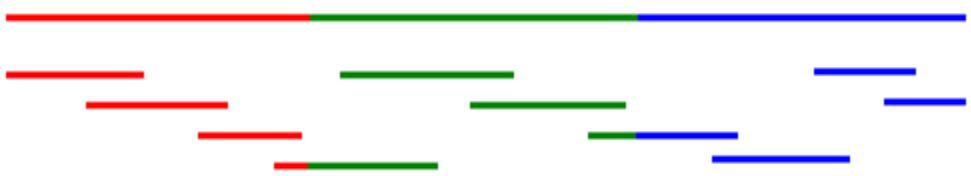


Figure 2: Represents assembly of DNA contig from sequence reads based on sequence similarity. Here, the upper line is being assembled together based on overlap of the shorter DNA sequences due to sequence similarity. Sequence similarity is represented as the different colors.

These contigs were then assembled together to form nine ultracontigs which comprised a large amount of the assembly (Mitreva et al. 2011). Orthologous protein groups were compared for *T. spiralis* as well as a number of other species in the Nematoda phylum showing that host characteristics may have a significant effect on the parasite characteristics that are present, even for species that are not closely related genetically. This is an example of convergent evolution and can be explored more fully in the future as more parasites from this phylum are sequenced and analyzed (Mitreva et al. 2011).

The mitochondrial genome of *T. spiralis* was assembled and analyzed by Webb and Rosenthal. A comparison was drawn between the assembled genome and the second assembled mitochondrial genome from *T. spiralis*. The results showed distinct within species variation in the form of single nucleotide polymorphisms. A SNP is variation between two similar DNA sequences, meaning it is a place in the alignment one sequence contains a base change relative to the other sequences in the alignment. Measuring the number of SNPs, as well as their locations within genomes, can give researchers an idea of the level of variation between the organisms whose genomes are being compared. This study went on to locate all the SNPs in their assembly compared to the earlier isolate, and performed a wide variety of analyses on the assembled genome. These include analysis of synonymous vs. non-synonymous positions and 4-fold degenerate vs. non-degenerate sites. Seven SNPs were located in region in question. Six of the seven SNPs caused non-synonymous changes (Webb and Rosenthal 2010). Non-synonymous changes are defined as SNPs that cause changes in amino acid coded for by the codon where the SNP occurs, whereas synonymous changes are SNPs that do not cause a change in amino acid coded for by the codon where the SNP occurs. Non-synonymous change is a strong indicator of variation as it represents change in the fundamental proteins of which organisms are made. Finding six non-synonymous changes compared to the one synonymous change suggests distinct variation between the two isolates of *T. spiralis* in the mitochondrial genome.

Fast genome sequencing and analysis is now possible because of next-generation sequencing and assembly. Traditional Sanger sequencing methods typically yield DNA sequences of between 300 and 1,000 bases in length. Next-generation sequencing has allowed for the collection of thousands, sometimes millions, of short DNA sequences at once (Martinez and Nelson 2010). Each of these DNA sequences can be between 25-50 bases in length,

classified as short reads, while some DNA sequences can be near or over 100 bases in length, classified as long reads. Though huge numbers of reads can be collected, the location of a given read within the genome cannot be determined. Therefore, it becomes necessary to assemble these reads to form a consensus DNA sequence.

The assembly of DNA can be completed in two forms: *de novo* assembly or assembly by reference. *De novo* assembly is completed with reads being compared to one another for similarity and overlapping regions to form larger chunks, or contigs, of DNA sequence. Assembly by reference is performed by comparing reads to a reference sequence for similarity. Reads then "map" to a specific portion of a reference sequence where they are similar in content to the reference. Reads overlap each other in parts of the reference sequence, and a consensus contig is formed (Martinez and Nelson 2010). The number of reads that overlap at a given position in the assembly are reported as read depth. Low read depth across the length of the assembly is usually an indicator of a failed assembly process, meaning very few reads on average mapped across each position and confidence in the correctness of the assembly should be low. This can also be an indicator of distinct variation between the reference sequence the reads being assembled together. High read depth is an indicator of the opposite, meaning many reads on average mapped across each position in the assembly and confidence in the correctness of the assembly should be high. Incredibly large read depths are usually indicative of repeat regions in the reference genome. Reads can be sequenced in one direction, single-end reads, while paired-end reads are sequenced in both directions. Without the presence of long reads and paired-end reads, *de novo* assembly becomes nearly impossible for eukaryotic genomes and assembly by reference can become more difficult (Martinez and Nelson 2010).

SNP analysis plays a vital role in comparative genomics. Analysis of SNPs from coding and non-coding regions of the mitochondrial genomes of 220 Portugese residents was done to compare different regions of the mitochondrial genome (*tRNA Leu*, *ND1*, *D-loop*, etc.) in an attempt to better understand mutation rates in the mitochondrial genome as a whole (Santos et al. 2008). The assumption was made that mutation occurred randomly throughout the mitochondrial genome, and the results show that non-coding regions acquired change at a faster rate than coding regions. However, there were also many complex factors, including sex of individual and level of heteroplasmy occurring within samples, that determine rate of change in the different regions of mitochondrial genomes (Santos et al. 2008).

Another function of SNP analysis involved determining whether a SNP causes a synonymous or a non-synonymous change in the genetic code of the organism. For example, the NKX2.5 gene in the DNA of humans with congenital heart disease (CHD) was studied for exactly this purpose (Dinesh 2010). The proteins coded for by NKX2.5 serve as early markers during heart formation. Sequences of healthy and diseased individuals were collected and aligned for comparison. It was found that humans with congenital heart disease, a birth defect that encompasses a number of different problems with the heart, can have up to six SNPs in the NKX2.5 exons. Three of the SNPs cause non-synonymous changes while two of the SNPs cause synonymous changes. The sixth SNP was not categorized as causing synonymous or non-synonymous changes. This shows that while SNPs in the NKX2.5 exons are not the cause of CHD, non-synonymous changes in this region of DNA may well be strong indicators of CHD in humans (Dinesh 2010).

SNP analysis can be used to determine how closely related organisms are at the genomic level. This method of analysis can be applied to *T. spiralis* and *T. murrelli*. Analyzing even a

small portion of the nuclear genomes of *T. spiralis* and *T. murrelli* is the first step in determining the amount of genetic diversity that exists between these two organisms. Therefore the main goals of this project are twofold. First, the project will explore the aspect of the assembly of next-generation sequenced reads from one species using the reference genome of another species. More specifically, this project will attempt to determine the feasibility of assembling short, single-end reads from *T. murrelli* to the full genome of *T. spiralis* by first assembling reads of *T. murrelli* to a reference contig of *T. spiralis* of approximately 60,000 bases in length. The project will also characterize the diversity that exists between an isolate of *T. murrelli* and an isolate of *T. spiralis* by analyzing the SNP content present in the assembly of *T. murrelli* throughout the coding and non-coding regions of the 60,000 base *T. spiralis* reference contig. The SNP content of the region as a whole will be examined and compared, and the non-coding region will be analyzed for synonymous and non-synonymous changes to the coded amino acid structure in the exon sequences of the genes throughout the region.

Methods:

T. spiralis Reference Sequence:

T. spiralis genome assembly contig 8.7 (Genbank accession number ABIR02000007) was chosen as the reference sequence for the project. This 60,000 base region of the *T. spiralis* genome was chosen because while only being approximately 1/1000 of the estimated entire nuclear genome length, it contains 17 protein coding regions, most of which contained intron and exon sequences, and 33,893 bases of non-coding sequence.

T. murrelli Novel Sequence Data:

106,858,709 short reads of mitochondrial and nuclear DNA from *T. murrelli* were collected and provided by Dr. Kristen Webb during her work at the USDA. The sequences were

collected using a Solexa Illumina Sequencing System in 2009 (unpublished) resulting in single-ended 25 or 40 base pair reads.

Assembly of *T. murrelli* to *T. spiralis*:

A preliminary assembly was created to check the viability of the 60,000 base *T. spiralis* contig as a potential reference sequence for the assembly of *T. murrelli*. This initial assembly used only 38,781,491 reads (the 25 base reads) and was evaluated based on read depth, position coverage, and repeat content. The assembly was done using Mapping and Assembly with Quality v. 0.7.1 (MAQ) (http://maq.sourceforge.net/maq-manpage.shtml). First, *T. murrelli* short reads were converted from fastq to bfq format using the MAQ command *fastq2bfq*. The fasta formatted *T. spiralis* reference contig was converted to bfa format using the MAQ command *fasta2bfa*. Files containing up to two million reads were mapped to the *T.* spiralis reference sequence using the *map* command under the following parameters: 1) two mismatches allowed for a read to map to the reference contig; 2) maximum outer distance for a correct read pair is 250, 3) maximum outer distance of two RF paired reads is 0; 3) mutation rate between the reference sequences and the reads is 0.001; 4) all unmapped reads and reads containing more than two mismatches were dumped into a separate file; 5) the threshold on the sum of mismatching base qualities is 70 (http://maq.sourceforge.net/maq-manpage.shtml). The mapped reads were merged together to form a consensus sequence using the *mapmerge* command. After MAQ creates an assembly, it generates the *mapcheck.txt*, *assemble.log*, *cns.snp*, *out.filtered.snp*, *out.pileup*, and *out.aln.txt* files as well as other files that can only be read and processed by MAQ. The *mapcheck.txt* file contains general data about the assembly, including assembly and reference sequence lengths, average read depth across the assembly, and base content of the assembly. The *assemble.log* file contains quality scores at each position of the assembly. The

cns.snp file contains all the SNPs reported by MAQ before its initial filtering step. The *out.filtered.snp* file contains all the SNPs reported by MAQ after it filters the SNPs based on quality scores. The *out.pileup* file contains read depth at each position throughout the assembly. The *out.aln.txt* file contains the reads that mapped to the reference sequence. With this information available, the *T. murrelli* assembly was analyzed to determine if the *T. spiralis* reference sequence would be a suitable reference for the assembly. Upon examination, the average read depth across the reference sequence was 24.751 reads and the coverage of the genome was 57,310 of 61853 bases, about 93% of the reference sequence covered. Additionally, there was not a significant amount of repeat content in the region. With these results, further analysis would be possible.

The comprehensive assembly was created using all 106,858,709 the reads (25 base pair reads and 40 base pair reads) using the same parameters and generating the same outfiles as described above. A script, runmaq.sh (Appendix A) written by Dr. Kristen Webb, was adapted and used to automate the process of running MAQ commands to create the assembly and its outputs.

Analysis of Assembly:

The output of the MAQ assembly yielded an assembled region of *T. murrelli* and the outfiles that accompany a completed assembly. Microsoft Excel (http://office.microsoft.com/en-us/excel/) was used to analyze the *out.pileup* file. Using the information provided in *out.pileup*, it was possible to determine a read depth cutoff for the SNPs that MAQ reported in the assembly. SNPs reported in regions with extremely high read depth are likely to not actually be SNPs. Due to MAQ's allowance of two mismatches during read mapping, reads that map to repeat regions in the reference sequence are likely to belong in different parts of the genome, but map to repeat

regions simply due to sequence similarity. This is because the reference contig being used is only approximately 1/1,000th of the full genome and reads belong elsewhere in the genome.

The read depths were sorted in ascending order and graphed to look for the point in which read depth underwent a spike (Figure 3).

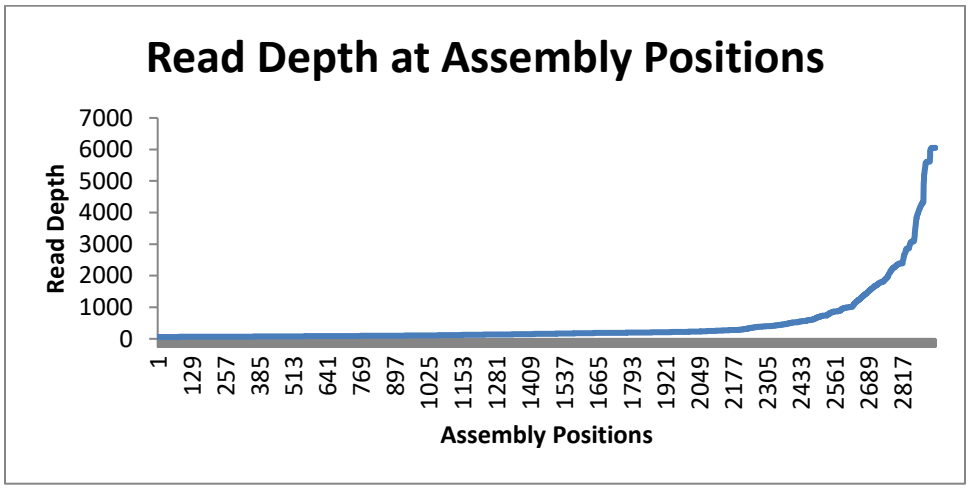


Figure 3: Shows the rapid increase of read depth across the 3,000 positions with the highest read depth in the full assembly. The other 57,000 positions contained read depths that grew at a slow, linear speed. Positions in the assembly (including positions with a read depth of 0) were sorted in ascending order by read depth. Across the final 3,000 positions in the list, a rapid increase takes place beginning near positions with a read depth of 100 reads.

The read depths of the assembly began to show a sharp increase. A read depth cutoff was chosen, meaning no SNPs that MAQ reported which had a read depth greater than the cutoff would be considered to actually be SNPs.

Excel was used to analyze the *out.filtered.snp* file, and exclude SNPs that fit the following criteria: 1) a read depth at the position of the SNP over 100, 2) two SNPs occurred at adjacent positions, 3) the SNP was reported only by reads where it was in the last position of the read, 4) less than 75% of the reads reported a SNP occurring at that position. The exclusion of SNPs occurring in adjacent positions was included for the same reasons as the exclusion of SNPs at positions of read depth greater than 100. Due to sequence similarity and allowance of two

mismatches during mapping, reads are likely mapping in locations where they do not truly belong. SNPs reported in the last position of the reads used in this project are ambiguous because preliminary data reporting average quality scores of base calls at each position in a read drop drastically at the ends of reads (Appendix B). Finally, if the majority of the reads did not report a SNP at a position, then the SNP was not considered to be real. This is for the same reason as described earlier, in that those reads likely belong to a different part of the *T. murrelli* genome but were aligned to the given position due to mismatch allowance.

A script, SNP_filter.py (Appendix A), was written in the programming language Python (http://www.python.org/) to automate the above step. Due to limitations in the assembly output information that MAQ provides, it was impossible to fully automate the process to filter SNPs based on all exclusion criteria. The script was able to account for all exclusion criteria, except it was unable to fully determine exclusion based on position on reads where SNPs were reported. The purpose of the script was to act as an estimate of the number of SNPs to be excluded by the initial SNP filtering step.

The coordinates for each *T. murrelli* gene were found from the *T. spiralis* reference contig on Genbank (ABIR02000007). Seventeen files containing only single fasta formatted genes from the *T. murrelli* assembly were created by trimming the full assembly sequence using Sequencher 5.0 (http://genecodes.com/) to include only the positions inside the gene coordinates found from the *T. spiralis* reference contig. Each newly created gene sequence was aligned to the full reference contig and the reported SNPs were edited down to include only the true SNPs determined from the earlier analysis. The 17 *T. murrelli* gene sequences were aligned to the reference contig of *T. spiralis* and the gene sequences as well as the reference contig were translated to amino acid sequence using Sequencher 5.0. Being sure to account for reverse

complementing of certain genes and intron sequence as reported by Genbank, the amino acids coded for by the SNP containing codons were analyzed for differences relative to the amino acids coded for by the codons in the reference contig. The synonymous and non-synonymous changes found in this comparison were counted for further analysis, and were not compared to possible synonymous or non-synonymous sites.

Results:

Of the 106,858,709 reads that went into the assembly, 92,065 mapped to the reference contig of *T. spiralis*. The reference sequence was 61,853 bases in length, and the mapped reads covered 58,148 positions, or 94% of the reference genome. The average read depth across the regions where reads mapped was 44.69 bases, meaning that on average 44.69 reads mapped to positions where any single read mapped. There were 3,705 positions where no reads mapped to the reference genome. The percent of mapped reads in the exon regions and non-coding regions are all above 90% except for one gene (Table 1).

MAQ reported 1,458 SNPs across the 58,148 mapped positions of the contig. The reported SNPs were subsequently filtered to 978 SNPs, hereby referred to as "true" SNPs. This number can be confirmed as accurately calculated based on the results of SNP_filter.py, which calculated 989 SNPs to be considered true SNPs. The SNP_filter.py script was expected to report an overestimate based on design of its algorithm. The SNP occurrence rate in the portion of genome was determined to be 1 SNP in every 60 mapped positions (Table 1**)**.

Table 1: Data from Regions in the Assembly of *T.murrelli*

The table reports information about the exons of each gene (size includes only the exon regions of each gene). In the non-coding region, intron sequences as well as between gene sequences are included.

	Size (in bases)	Percent of mapped positions	Number of SNPs	Percent SNPs
Hypothetical Protein (1)	590	100	4	<1
Conserved Hypothetical Protein (1)	2437	100	38	2
Hypothetical Protein (2)	211	100	0	0
Conserved Hypothetical Protein (2)	367	90	7	2
Protein UPS1	604	100	8	1
Conserved Hypothetical Protein (3)	987	99.6	24	2
Putative Sperm-Specific Protein Phi-1	281	100	3	1
Putative Lectin C-Type Domain Protein	598	98	4	<1
Receptor Expression-Enhancing Protein 5	497	100	3	<1
Asparagine--tRNA Ligase	1166	100	11	1
Conserved Hypothetical Protein (4)	467	100	5	1
Netrin-1	371	100	4	1
ATP-Dependent RNA Helicase DDX42	2210	100	23	1
Conserved Hypothetical Protein (5)	748	97	17	2
Putative Immunity Protein PlnM	351	49	3	17
Conserved Hypothetical Protein (6)	447	94	4	1
Conserved Hypothetical Protein (7)	803	100	18	2
Non-Coding Regions	45013	92.5	802	2

Of the 17 genes in the contig, 458 SNPs were found in the genes and 520 SNPs were in non-coding regions. Of the 458 SNPs in the gene regions, 177 were in exons. The combined length of all the exons in these genes was 13,135 bases. Using this data, it was determined that the SNP occurrence rate in the exons of the 17 genes was approximately 1 in 73. Additionally, there was a 1:4.5 ratio of SNPs in coding regions, SNPs in exons, versus SNPs in non-coding regions, SNPs in introns and non-coding regions between genes. The percent of SNP content in the exons of the genes and the non-coding regions is under 5% for all except Putative Immunity Protein PlnM (Table 1). Of the 177 SNPs occurring in the exons of the 17 genes, the number of synonymous changes was 90 while the number of non-synonymous changes was 87. This makes the ratio of synonymous versus non-synonymous changes just over 1:1 at 1.03:1.

Discussion:

Assembly Quality:

Considering all the factors surrounding the assembly of *T. murrelli*, the quality of the assembly was high. The assembly can be considered good because of its high average read depth and its high percentage of sequence coverage. Additionally, and perhaps most importantly, all but one of the genes in the region had a very high percentage of coverage (over 90%). This is especially true considering the data being used as well as the use of a different species as a reference during the assembly process. The data used were short reads comprised of sequences that were either 25 or 40 bases in length. Read length plays a major role in determining coverage of the reference sequence (Martinez and Nelson 2010). A longer read will be able to map to a larger part of the genome. If longer reads are present, then a single read will be able to provide coverage over more positions. This leads to a larger average read depth for a given assembly, which bolsters the confidence in correctness of assembly. The average read

depth of the *T. murrelli* assembly was 44.69. That many successfully mapped over a given position in the reference genome due to sequence similarity allows for confidence in the correctness of the assembly. Other studies, such as one performed by Dalloul et al. where the turkey genome was sequenced and assembled using next-generation techniques, was only able to achieve an average read depth of 15.3 reads (Dalloul et al. 2010). Next-generation assembly of a snail, *Radix balthica*, transcriptome yielded read depth of approximately 30 reads across a region of approximately 17.5 million positions (Feldmeyer et al. 2011). The assembly of *T. murrelli* was able to achieve a read depth greater than both of these other studies adding evidence to confirm that the assembly was created correctly. Also, the data were only single-end reads, rather than paired-end reads. Paired-end reads are the result of high quality sequencing of a sequence in both directions. The lack of paired-end reads makes *de novo* assembly, or the assembly of data without a reference genome, extremely difficult as well as increasing the difficulty of reference assembly (Martinez and Nelson 2010). Additionally, the gaps in the assembly may in fact represent true diversity between the two species. This study is the first of its kind to assemble reads from one species to a reference genome of another species. Even when considering that the data was single-end short reads assembled to a different species, the assembly quality was still high, as indicated by the high read depth and low number of gaps.

As further proof of the accuracy of the assembly, the large number of unmapped reads is reasonable for the scope of this project. The collection of reads in next-generation sequencing leads to the collection of all types of DNA from a given sample. In this project, the DNA collected includes genomic DNA as well as mitochondrial DNA. The reads are therefore comprised of DNA from all parts of the nuclear and mitochondrial genomes. Since the region in question is only about 1/1000 of the full nuclear genome, the large number of unmapped reads is

actually a good indicator that the assembly was made correctly. A small number of unmapped reads would imply that many reads that did not belong in the assembly were mapped to the reference contig. This project becomes the first that was able to successfully create an assembly of a partial sequence of genome of one species while using a reference contig of another closely related species as the reference genome.

SNP Occurrence Rates:

By comparing the SNP occurrence rates found in this project to those from other studies, it is possible to determine a basic estimate of the amount of diversity between these two closely related species of parasite. It has been previously discovered that the SNP occurrence rate between the mitochondrial genomes of two isolates of *T. spiralis* was 1 in 3100 (Webb and Rosenthal 2010). Comparing this number to the occurrence rate in the coding region, shows that there is clearly more diversity between the nuclear genomes of the two isolates of *T. murrelli* and *T. spiralis* than exists between the mitochondrial genomes of two isolates of *T. spiralis*. Additionally, SNP occurrences in the 13,720 base long IFN regulatory factor 6 gene were explored. This gene was sequenced and aligned from two pigs, discovering 3 locations of exonic variation, which equates to a SNP occurrence rate of 1 in 4,573 bases (Liu et al. 2012). This is comparable to the previously mentioned SNP occurrence rates, in that the occurrence rate discovered here is less frequent than the *T. spiralis* mitochondrial coding region. This can be attributed to the fact that the mitochondrial genome tends to acquire change at a faster rate than the nuclear genome. And still, this SNP occurrence rate of 1 in 4,573 is less than the occurrence rate of 1 in 73 found in the assembly of *T. murrelli*. These results suggest that distinct variation exists between these two species of parasites.

Variation within Genes:

Of the 17 genes in the assembly, four were found to be moderately conserved while five were found to be moderately variable. The conserved genes, hypothetical proteins 1 and 2, putative lectin C-type domain protein, and receptor expression-enhancing protein 5 were comprised of less than one percent SNPs. The variable genes which were conserved hypothetical proteins 1, 2, 3, 5, and 7, were all comprised of two percent SNPs. Variation is an indicator of positive selection on a gene while conservation is an indicator of negative selection on a gene (Snoeck et al. 2011). Positive selection means that variation is being encouraged in a specific region while negative selection means that the organism functions more successfully if the region is conserved. The hypothetical and conserved hypothetical proteins could not be identified and no notion of possible function could be discerned. However, the putative lectin C-type domain protein plays a major role in the immune system of the eastern oyster, *Crassostrea virginica*, where it helps to recognize particles foreign to the organism (State University of New York 2011). It very likely plays a similar role in *T. spiralis* and *T. murrelli*. Receptor expression-enhancing protein 5 is a member of a family of genes that aid in cell surface expression of various proteins (Park et al. 2010). Both of these genes fulfill important roles in all organisms, and it would be reasonable for them to remain conserved throughout the *Trichinella* genus. The gene putative immunity protein PlnM only achieved just under half coverage at 49%. This lack of coverage could represent true diversity between the two organisms (and perhaps the two species), or it could simply represent an area of the genome where sequencing and assembly were not able to produce high enough quality reads to cover the area. Not much is known about this gene, but further study of it might lead to determining the differences in host/parasite interaction between these two species.

This project helped to determine the feasibility of assembly reads of *T. murrelli* to a *T. spiralis* reference genome. The reference contig used in this project only represents 1/1,000 of the estimated length of the *T. spiralis* genome, but it is good representation of the genetic content likely found within the rest of the genome. This contig contained 17 genes, many containing numerous exons and introns as well as a few that contained only exon sequence, and large amounts of non-coding sequence, which was partly comprised of repeat content. That being said, the result was high confidence in the assembly based on strong coverage and high average read depth. The gaps in the assembly may truly represent genetic difference between the two species or may represent areas where the sequencing and assembly processes were unable to produce assembly sequence, both of which are to be expected during a random assembly project where reads from one species are assembled to reference sequence of another species. It is a strong possibility that the assembly of *T. murrelli* can be accomplished using the *T. spiralis* genome as a reference. Even though the assembly process would be greatly aided by the presence of paired-end, long reads, the assembly of *T. murrelli* short, single-end reads to a reference genome of *T. spiralis* would likely yield high read depth and comprehensive genome coverage based on the assembly quality in this project.

Most research in the comparative genomics field tends to examine different organisms of the same species where this study examined two organisms of different, but closely related species. One study compared the IRF6 gene complexes between two swine and found only three synonymous changes and zero non-synonymous changes. These changes were found over a 13,720bp length (Liu et al. 2012). This exonic length is comparable to that which was examined in this study, differing by only 585 bases. However, the amount of change found in these exonic regions differs drastically. Liu et al. found an exonic SNP occurrence rate of approximately 1 in

4,573 whereas the exonic SNP occurrence rate found in this study was 1 in 73. Not only this, but the only changes found in the IRF6 gene complex were synonymous. This is opposed to the region examined in this study having half of the changes being synonymous and the other half being non-synonymous (Liu et al. 2012). This indicates significant differences between the two species in this region.

Additionally, a second study examined the NKX2.5 gene sequence, which contains two exons of a combined length of approximately 3,120 bases in length, of 220 people. Within these samples, the researchers found six SNPs. These SNPs, as with the data presented in this study, caused approximately half synonymous changes and half non-synonymous changes with two causing synonymous changes and three causing non-synonymous changes (one SNP was uncategorized) (Dinesh et al. 2010). The SNP occurrence rate found in the NKX2.5 exons is 1 in 52. Compared, to the results of this study, that rate is more frequent than the SNP occurrence rate of 1 in 73. However, the study done by Dinesh et al. found these six SNPs cumulatively among the 220 participants. This means that many of the participants had zero or one SNPs occur in their 3,120 base region, and only collectively were the six SNPs able to be identified (Dinesh et al. 2010). When this fact is accounted for, it can once again be seen that the NKX2.5 gene contains less variation on an organism to organism basis than does the assembled region of *T. murrelli* compared to *T. spiralis*.

Overall, the project was a success, and distinct variation was found between these two species of parasite. Though the size of the assembly and number of genes examined were not a large enough portion of the full genome to determine any conclusive information about the diversity between these two species, the data serves as a basis of comparison for future studies regarding these parasites. Future work that compares a much greater amount of the genomes will

be invaluable in determining differences between these parasites. This research will lead scientists closer to determining in what ways these organisms function differently and may present clues as to how and why they do not possess the same abilities of infection.

Acknowledgements:

I would like to acknowledge Dr. Kristen Webb for her continual support throughout the project. I would also like to thank Ben Rosenthal and Dr. Kristen Webb for their work at the USDA and for providing me with the *T. murrelli* sequence data which made this project possible. Thank you to Dr. Matt Jadud for all his programming help and his constant support throughout my years at Allegheny. Finally, I would like to thank all of my fellow comp group students who aided me on numerous occasions with this project and other endeavors over the past year and a half.

Literature Cited

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., et al. (2010). Multi-platform next-generation sequencing of the domestic turkey (meleagris gallopavo): Genome assembly and analysis. *PLoS Biology, 8*

Dinesh, S. M., Kusuma, L., Smitha, R., Savitha, M. R., Krishnamurthy, B., Narayanappa, D., et al. (2010). Single-nucleotide polymorphisms of NKX2.5 found in congenital heart disease patients of mysore, south india. *Genetic Testing and Molecular Biomarkers, 14*, 873+.

Dubey, J. P., Hill, D. E., & Zarlenga, D. S. (2006). A trichinella murrelli infection in a domestic dog in the united states. *Veterinary Parasitology, 137*, 374+.

Feldmeyer, B., Wheat, C. W., Krezdorn, N., Rotter, B., & Pfenninger, M. (2011). Short read illumina data for the de novo assembly of a non-model snail species transcriptome (radix balthica , basommatophora, pulmonata), and a comparison of assembler performance. *BMC Genomics, 12*, 317.

Hill, D. E. USDA - ARS.

Hill, D. E., Samuel, M. D., Nolden, C. A., Sundar, N., Zarlenga, D. S., & Dubey, J. P. (2008). TRICHINELLA MURRELLI IN SCAVENGING MAMMALS FROM SOUTH-CENTRAL WISCONSIN, USA. *Journal of Wildlife Diseases, 44*(3), 629.

Kapel, C. M. O., Webster, P., & Gamble, H. R. (2005). Muscle distribution of sylvatic and domestic trichinella larvae in production animals and wildlife. *Veterinary Parasitology, 132*, 101+.

Liu, Y., Xu, J., Fu, W., Weng, Z., Niu, X., Liu, J., et al. (2012). Tissues expression, polymorphisms of IFN regulatory factor 6 (IRF6) gene and their associated with immune traits in three pig populations. *Asian - Australasian Journal of Animal Sciences, 25*, 163+.

Martinez, D. A., & Nelson, M. A. (2010). The next generation becomes the now generation. *PLoS Genetics, 6*

Mitreva, M., Jasmer, D. P., Zarlenga, D. S., Abubucker, S., Wang, Z., Martin, J., et al. (2011). The draft genome of the parasitic nematode trichinella spiralis. *Nature Genetics, 43*, 228+.

Park, S. H., Zhu, P., Parker, R. L., & Blackstone, C. (2010). Hereditary spastic paraplegia proteins REEP1, spastin, and atlastin-1 coordinate microtubule interactions with the tubular ER network. *Journal of Clinical Investigation, 120*, 1097+.

Rosenthal, B. M., LaRosa, G., Zarlenga, D., Dunams, D., Chunyu, Y., Mingyuan, L., et al. (2008). Human dispersal of *trichinella spiralis* in domesticated pigs. *Infection, Genetics and Evolution, 8*, 799.

Santos, C., Montiel, R., Arruda, A., Alvarez, L., Aluja, M. P., & Lima, M. (2008). Mutation patterns of mtDNA: Empirical inferences for the coding region. *BMC Evolutionary Biology, 8*, 167.

Snoeck, J., Fellay, J., Bartha, I., Douek, D. C., & Telenti, A. (2011). Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology, 8*, 87.

State university of new york publishes research in fish and shellfish research.(2011, 05/03; 2012/4). pp. 2801.

Webb, K. M., & Rosenthal, B. M. (2010). Deep resequencing of trichinella spiralis reveals previously un-described single nucleotide polymorphisms and intra-isolate variation within the mitochondrial genome. *Infection, Genetics and Evolution, 10*, 304+.

Webb, K. M., & Rosenthal, B. M. (2011). Next-generation sequencing of the trichinella murrelli mitochondrial genome allows comprehensive comparison of its divergence from the principal agent of human trichinellosis, trichinella spiralis. *Infection, Genetics and Evolution, 11*, 116+.

Appendix A:

SNP_filter.py:

```
#!/usr/bin/env python

# Travis Charlton
# SNP_filter.py
# This script will take a list of reported SNPs from MAQ and filter through
them to determine
# whether or not they are indeed true SNPs.

# Importing the necessary libraries
import collections
import fileinput
import time
import sys, os
import re


#### Filtering constants ####
# MAXDEPTH - is used as a cap on possible read depth.  No SNPs reported at
positions where
#           read depth exceeds MAXDEPTH will be included as true SNPs
# throwOut25s - is a boolean (True or False) that when set to true, will
exclude reads where
#           the 25th position of a read aligned with position of the SNP.
#           Reads where the 40th positions align are automatically thrown out
because
#           this is the last base in the read and average quality scores show
that the
#           there can be no confidence in the called base.
#           NOTE:  I realize that throwing out 25 aligned reads means that
some reads of
#           of length 40 that happen to have an alignment with a SNP at the
25th
#           position will occur, hence my estimates have this value set to
False.
MAXDEPTH = 100
throwOut25s = False

#### Data Structures ####
# These structures contain the position, base, and read depth of the
reference sequence and
# asembly sequences (SNP containing sequence) respectively.  'snp' also
contains the ref
# seq base as well as the assembly (snp) base
refinfo = collections.namedtuple('refinfo', ['refpos', 'base', 'readdepth'])
snp = collections.namedtuple('snp', ['snppos', 'refbase', 'snp',
'readdepth'])


#### Calculated/Extracted Values ####
```

```
# result_list - contains the pertinent information extracted from the
out.revised.pileup
#           file
# matchpositions - contains the positions on each read where the read mapped
to the SNP
#                 position in the reference genome.
#                 NOTE: This does not account for whether or not a SNP was
actually
#                       reported or not reported by the read, only reports
what
#                       position on the read aligned to the reference
position
# mismatches
result_list = []
matchpositions = {}
mismatches = {}


pileup_file = open('out.revised.pileup', 'rb')


# This loop populates the above variables with the appropriate values
for line in pileup_file:
      try:
            # We want data in the last field, pipe-separated.
            # gi|316979797|gb|ABIR02000007.1|   13    C     4     @.,,,
      34,5,5,1,
            first_split = line.split('|')

            # Grab just the bits to the right of the last pipe.
            # len(first_split) - 1
            # gets us the length of first_split, and then subtracts one...
            # so, we pull the last element from the first_split array.
            interesting_stuff = first_split[len(first_split) - 1]

            # Now, we have data that looks like...
            #   13      C     4     @.,,, 34,5,5,1,
            # We will split on one or more whitespace characters.
            final_split = interesting_stuff.split('\t')

            # We want columns one, two, and three.
            # The split leaves an empty string at the start of the split,
hence it seems
            # like we're off by one.
            # Push a pileup structure onto the list.
            result_list.append(refinfo(final_split[1], final_split[2],
int(final_split[3])))

            # Grab the numbers at the end. Store them.
            # Store a pair, which contains the... posn and then the list of
matchpositions.
            numbers = final_split[len(final_split) - 1].split(',')
            numbers = filter(lambda s: s != '\n', numbers)
            numbers = map(int, numbers)

            # Now a hash. Reference by key (the posn) and value (number list)
            matchpositions[int(final_split[1])] = numbers
```

```
            # Actual mismatches -- the commas and periods and stuff
            #print final_split
            mismatches[int(final_split[1])] = final_split[len(final_split) -
2]


      except:
            print "Something went wrong."

#### SNP Filter ####
# here is where the reported SNPs will be filtered down to only the true SNPs

out_filtered = open('out.filtered.snp', 'rb')

# filtered - contains the pertinent contents of the out.filtered.snp file
filtered = []

# This loop populates filtered with pertinent contents of the
out.filtered.snp file
for line in out_filtered:
      # Data looks like...
      # gi|316979797|gb|ABIR02000007.1|   267   G     A     75    16    1.00
      63    60    N     255   N
      fsplit = line.split('|')
      # Again, last column.
      last = fsplit[len(fsplit) - 1]
      # Now, split on tabs
      tsplit = last.split('\t')
      # We want the list of good snips...
      filtered.append(snp(tsplit[1], tsplit[2], tsplit[3], int(tsplit[5])))

# a function that creates a hash table containing the 'matchpositions' for
each SNP position
# with SNP being the key and 'matchpositions' being the value
def listOfNumbers (key, dict):
      return dict[key]

# a function that determines if 40s (and 25s if 'throwOut25s' is set to True)
are present in 'matchpositions' on SNPs
# and counts the number of 40s (and 25s if included) are present
def countEnds (lon):
      global throwOut25s
      c = 0
      for n in lon:
            if n == 40:
                  c += 1
            if throwOut25s and n == 25:
                  c += 1
      return c

# a function that creates a hash table of the number of reads reporting a SNP
# at each SNP position in out.filtered.snp
def calcMismatches(key, dict):
      s = dict[int(key)]
      s = re.sub('\\.', '', s)
      s = re.sub(',', '', s)
```

```
	s = re.sub('@', '', s)
	print s, " ", len(s)
	return len(s)

# The filtering step:
# list that will contain the true SNPs after the filter
keepers = []

# a reference index to help pass through out.filtered.snp
ndx = 0

#
while (ndx < len(filtered)):
	# Pull out the snp
	me = filtered[ndx]
	# Assume we will keep the read
	keep = True

	# If the read depth is under 100, continue
	if (me.readdepth < MAXDEPTH):
		keep = True
	else:
		keep = False

	# We will never do more checks if we've decided to throw things
	# out. Hence, add the check for 'keep'.

	if keep:
		print "==== ", ndx, " ===="

		# Check to see how many reads end with a mismatch
		lon = listOfNumbers(int(me.snppos), matchpositions)
		mismatchCount = calcMismatches(me.snppos, mismatches)
		count = 0
		print lon
		if lon:
			count = countEnds(lon)
		revisedReadDepth = me.readdepth - count

		if keep and (revisedReadDepth >= 3):
			keep = True
		else:
			keep = False

		if keep:
			# Not subtracting 40s from the mismatchCount...
			# So, mismatches on the end of a read are considered "OK"
			goodnessScore = (float(mismatchCount) /
float(revisedReadDepth))
			print "c: ", count, " mc: ", mismatchCount, " rd: ",
me.readdepth, " rrd: ", revisedReadDepth, " gs: ", goodnessScore

			# goodnessScore is the percentage of reads that must report
a SNP in order for it to be counted as a true SNP
			# 75% was used for this project
			if goodnessScore >= 0.75:
				keep = True
```

```
                  else:
                        keep = False


      # Look forward and backward...
      # start by looking forward.
      if keep and (ndx < len(filtered) - 1):
            next = filtered[ndx + 1]
            # They are not back-to-back if...
            # the next one is one more than me.
            # Subtract from the next; if they're equal,
            # we must be adjacent. That's bad.
            if (int(me.snppos) == int(next.snppos) - 1):
                  keep = False
            else:
                  # Look backward
                  if (ndx > 0):
                        prev = filtered[ndx - 1]
                        if (int(me.snppos) == int(prev.snppos) + 1):
                              keep = False
                        else:
                              keep = True

      # Finally, decide if it is a keeper.
      if keep:
            print "KEEP"
            keepers.append(me)
      else:
            print "KICK"

      # Increment
      ndx += 1

# print keepers
print len(filtered)
print len(keepers)




#### OUTPUT EXAMPLE ####
#
#
#     ====  SNP Position  ====
#     'String of letters,     Number of SNP bases reported
#     bases reported
#     as SNPs'
#     [positions on read where read aligned to reference at each reported SNP
position]
#     c: count = number of 40s (and 25s if included) found  mc: mismatchCount
= number of reads reporting a SNP
#     rd: readdepth = reported read depth  rrd: revisedreaddepth = read depth
minus count (corrected for bad quality base calls)
#     gs: goodnessScore = percentage of reads reporting SNPs compared to
total reads (corrected for count of 40s (and 25s if included)
```

```
#     KICK or KEEP = whether the SNP was kept as a true SNP or removed from
the data set
#
#     (at bottom of list)
#
#     number of SNPs reported in out.filtered.snp
#     number of SNPs remaining after the filtering step




#### My results:
#### SNPS in out.filtered.snp = 1458
#### SNPs remaining after filter = 989

#### This is good!  The total SNPs found in my lab work were 978.  This
program is not able to account for all of the factors
#### that I was able to in my lab work, so it serves as an estimate.  I
reported nearly 99% of the SNPs reported by this filter.
#### This program is able to successfully estimate the number of true SNPs
present in a dataset!
```

runmaq.sh:

```
#!/bin/sh

#path to directory containing 2million read files
dir1=/SeniorComps2011-2012/Travis/T5Outfiles.dir/All.dir/

i=0

#for each outfile in the directory, convert to bfq and map to reference
sequence
for file in `ls $dir1/outfile*`

do
    maq fastq2bfq $file reads$i.bfq
    maq map -u unmapped$i.txt reads$i.map /SeniorComps2011-
2012/Travis/T1ReferenceSequences/contig8.7.bfa  reads$i.bfq
    let i=$i+1
done

mapmerge=""
for file in `ls reads*.map`
do
    mapmerge="$mapmerge $file"
done
echo $mapmerge
maq mapmerge out.map $mapmerge
```

```
maq mapcheck /SeniorComps2011-2012/Travis/T1ReferenceSequences/contig8.7.bfa
out.map > mapcheck.txt

maq assemble consensus.cns /SeniorComps2011-
2012/Travis/T1ReferenceSequences/contig8.7.bfa  out.map 2>assemble.log

maq cns2fq  consensus.cns > cns.fq

maq cns2snp  consensus.cns > cns.snp

maq.pl SNPfilter cns.snp > out.filtered.snp

maq  pileup /SeniorComps2011-2012/Travis/T1ReferenceSequences/contig8.7.bfa
out.map > out.pileup

maq mapview out.map > out.aln.txt

maqindex -l 1024 -c consensus.cns -i out.map

#maqview -c consensus.cns out.map

rm reads*.bfq
```

Appendix B:

Supplemental Data:

Average read qualities:

base	mean_quality
1	33.0
2	32.5
3	31.6666666667
4	31.75
5	31.4
6	31.1666666667
7	30.2857142857
8	30.5
9	30.4444444444
10	30.5
11	30.5454545455
12	30.1666666667
13	30.0769230769
14	29.6428571429
15	29.6666666667
16	29.6875
17	29.2352941176
18	28.7222222222
19	28.7894736842
20	28.65
21	28.4285714286
22	28.2727272727
23	28.2608695652
24	28.0
25	27.84
26	27.7307692308
27	27.5925925926
28	27.3571428571
29	27.2068965517
30	26.9666666667
31	26.7741935484
32	26.53125
33	26.2424242424
34	26.0588235294
35	25.9714285714
36	25.8333333333
37	25.7567567568
38	25.7105263158
39	25.6666666667
40	25.65

T. murrelli Gene Sequences:

>hypothetical_protein1
ATGCGTTTGCCTGTTTGTACAAGGCAAACAGATGCCCCGCAGCTGGATCC
CGCAGTTGTTCACAAATATCCGAGGTCCAGCATTCGAAAGGTAACAAGTG
GGTGGACCGCGTAAGAGTCTGCTTTGCAGAAATTTGTCCAAATAAGAACT
TCTCCAACAGTGTATATTCTTGATATGTAAAACTGTTTACTCAACAGCAT
CCGTCAGCCGGCAATTTGCACGGGGGCCGTTGTGACAGCCGACAACACGT
CATTAAAAGTGACAACCCGATTGGTTAAAGCACAATCAGGCAAAAGTCTA
AAGGTGCTTCGACTGACCGTTTTCGAGTCAGTGTTTTGCCGACCTACCTC
GAAGCTTACATGGATGAACTATCGAAACGCTGAACTGTTCCTGCTGGCCT
TCTTGACGCTTTCACTTTTGGCCGTCTTAGCCTTAGCTACACTTTGCGCC
ACCGCTTCACCAAATCTCATACAGAGGTGAATATAATTATTTAGCTTGTT
GTTTATTTTCATTTGAATAATAATATTCATTCAAATGAGTTCACCATCTC
GTTTAATGAACGCAACATCCCTGTAATATTTTAGGGTGAGAAGACCGAGA
AGTGGAATTAGTGACCAATTGCTTGCCCATCACTCGGGCCACGGCAAACC
GTCTGAAGCGCCGACAAAGAACATCGCCTACGGTTTAGACCTTTGGTAAG
TACATATTTGGCATACTCATATACTTACACTCAAATGCGACTAAATTGCA
TTTGTAAATTTTAAGCATGAAGCAAGTGATATACAAAAACAAAGTCAGTA
ATTATTTTGAACTTTCATATTTTGGGTATTATTAAATTAATTTTAATTAA
AATTTGAAATTACCTTCAAAACACAACCTCATAGCAATAGCAAACAACTG
AGTACAACTTGGTATATACTACATCATTTACTTTAGAATGAATTGTGAAA
AATTCAAAATTATGCACGCAATATTATTATTGAACATTAGACAAAGTGTA
TTAACAAATTGATTGAAATACACCACAAAAAATTTAAACTAGGCAAGAAA
AATAAAGCAAATAATAACAACGGCAATAATTGTAATAAATGAAATTTTCT
ATAATAAATAAATCACCAAATTGAAATTATGAAGTACTTTTTCAATTTTT
ATTTCAACAATAGTCATAATTAATAGAATAACTAACTAAATGTTCATATT
AACGAGCAGTTGTTCATGTTAAATAATAGGTATAGTAAAAAAAATAATAA
ATAATACTATNNNNNNNNNNNNTATTTAAATTTAATTGTACAAATTGCAA
ATGAACAAATTAAATTCTACTTTAAANNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNCAACAAAAAAGCAATTCTGAAAGTG
AAAGTAAGTAATAATTACCGATTAATNNNNNNNNNNNNNNNNTGATTGAT
TTCAAAATCATTAAGACAAAATAATTTCATTACAACGGATGAGAGGAAAT
TGTATTTTAAATAAATAAATATATATATGCATGCCATTTTTCAATCATTT
TGTCCATCTACGATTGNNNNNNNNNNNNNNNTGCTTAATGGAAATATCAC
ATAATCTTGTAAAAAGTAGTTTAACAAAAAAAACCTGTTAAAGTTATACT
TTGTTCTGCTTTTTTGTAGTTGCCTGCATTTTTTTCTTTTAACAACTATA
CTGTATCACATTAAGAATAAATAATATGATGCAGACAGAACAAATTTGCT
AGAATCGGCGTTATCATCATTGAATGAATCAGATCAGATAGATAGATGTT
CGTGTTGTTGATCAGTAAACAACGCAGCTTGTTTTTTTATGCAATATAAA
CATTTTTGATTAAACAAAAAAAAGCATTAATACAAATTGTATTGTGTATA
AATAAATTTCCGCTAACTGGACGTTGCCTAACAATGCAATGCTTCTTTTC
TTGCCTTTTTTATTGTCCTGCGCCAAATTCACTCTATTATTAATTTNNNN
NNNNNNNNNNAAATTCCAAAAAAAGAAATACTCTACTATATTCGACCAAT
CAAGAAGTGCTATTCTGTTGGTTGGATGTTTGTTTTTTTTCTGAAAAAGC
ATCGAAAAAGAATAGTCGAATCATATTCACTTTTTCTGTATTACATGATT
CTTGTTGTTGTTGTTGTTGGTATTTGTTACTGAAATTTTSAAAAAAT
TATTTAAATTTTTAAATTCAAATTCTTCTCATTTGGAAAATATACAAAAG
TGAAGTGTACTGTATTTATATATGAGAAACAATTGCAGGAACGGTACCAC
AACGCCGTACATTTTGAACAAAGCGGAACGTGAAGAACAGCTGCTGAAAC
TTTTGCTGAAAGTGGTACAAAGCTTGCGATGA

>conserved_hypothetical_protein1
TCAGAATGTTGGGTGGCGACTGTCCATATTGTAATAAGTGTCGCCGTAGA
CTGTGTTAGGCGGGTGACCCATCGCTGCCAATTGCTCTGGCAAAGGCATT
TGTCCCGGCGGTGGCGGTGGCGGAGGGGGAGCAAATGGGTGAGGGGGCCC

AGCAGGCGTCCATTGGGTATTTGGTGGGGGTGGGGGTGGCATATCCCAAA
CTATTCCAGGTGGCGGGGGTGGCAATGATCCTGGTGGACCAAATATCTGA
CAGTTCTTTGTCGACTGAAGCAATCGCTCCACTTCATTGTCTTTGGCTAT
GTAAATAACTAACCCTGAAACGTGAACAAAAAAACGATCTAAATAATAAA
ATAAGCTTCGAGATAGATGATGTGAGATGATGAAGTAAAAAGAAAAGAAT
ATACCGTTTGTGTAAACTGAGCAATCCTGTTGCAGGCGAGCGAAAGCTTT
TTCTGCTGAAACGTCTTTGAATTTGACAGCAGCGAATCCAGAACCCGGAG
GAAAAACAAAAGCTTGTACTGGTCCAACTCGATTGAACAGGTATTGCAAT
GTTTGTTCATCCATCTGAAATAAAAAAAAAAATGAATTATTATTACTATT
CAATAAAAAGGTTGGAAAAAGTTTGTGCTCAACGTGGTATTGTTGGAGCT
GAACGCACCGTTTGCTGAGTTAAATGCATAAGCACAATCCATTGCAGTCC
ACAAGATGTCGATGGGATTGATGATGAAGAAGMAGCATATTGWGGAGCTG
TTGTTATTGGTGTTGCTGCTGGTGCAGCAGCAGCAGCAGCAGCAGCAGCA
TACGGCATGACACGAGACGGAGATTTCGCACTGAATGGAGAATAGGCAGC
GGAACTAGTCGTCGTCTGCATTATCGGTCGGAATGATTTTGCGGACCAAC
TATCGGTCGCATAGTTTGGACTGATGCTCGAGATCGTCGGGTCGATACAG
CTTTTGGTCGGTGTTGTAATGGCGGACAGCAAAGAGGTGCTGGGCATGGA
ATAGGCGCCGTGGGCAAAACCGCTGGTAGAAGGACCAGCTTGATAGGTCG
GTTTCAGATCGTGACCAAAACGGCCCAAACCAATTGAACTGGTCTTGACA
CTACCTGGTGTAGCGTTCGGGTCGTCGGCCACCTCTTTCGGATCTCGCCA
TTCCCATTTTTGGCCTGGTCGAAATTCAGGCGGTCCATCGTCAATTCCAA
TAATACCCGTGCCAAAAGTGTGCGCGCCGCTGGTGCCGGGCATTGAACTA
ACCAGGCTCATGTCGAGAGACATATGACGGCCGATTATTCCATTTTTATC
CGATTGACTGACAACGGCCGCAACAGCATCTTCCAAACTTCCGTAGGCAG
CAGCAGCACTGGGAAAAATACTGCTTCTCAGATCTTTCATGTCGCCGGTT
TTCATGAAAGAATCGATTGGCCATTCGAAACCAGACTGTCGCTTGGTATT
CGACCACGGATCACTACTACTACTGCCAGCAGTCATGTCAAAAACGCCAA
ACTGTTTGCTCGTAGATGCGGTGGCGGAAATGGCAGCAGCACAATCGCCG
CCATTAATATAAATCGGTTTACGATCGGTACTCAAATACATATTCARTGG
ATTCAACGGGCCATGTTTCCACTGCTGAAAAATCAAAAAAAAAATATTATT
ATTAATCAAAAAACAATATTCGTCGATTCACGTATANNNNNNNNNNNNNN
NNNAAAAACTGACCATTAATTTTGACTGATTTGGTAAATCTGATGTTGGC
ATGGAACAGCCGAAAGTTACATCGTTGGATGAGGTGTTTTTCCTCAGTGA
AGTTTGTTGCGCGGCGCTGAACTGACTGCGCAAATTTCGCGCGGTCAGTC
GAAGCTGATTCACCTCACCGGCCAGGCGCCGGTATTCTTGCTGCTGAGCG
GCGTTCAACGTCCCGTTATGTTCCAACGTCTGCAAGTGTTGAAGCTCGAT
TTTCGACTGTTCCAATTTCGTTATACACTGTAGCAGCTGTCTCAACTTAA
GTAAAGTTTCTTGCGGCAGCGGTTGATTTAAAATAGTACTGTCTATGTGT
CCGGTCCTCATGGCTATATTGAGCTTCTCCGCTACTTGGTGCGCCAAAGC
CTAAAGAGCAGCAGCGACCATAGCGACCGTACTAAGGGGCAAATTTACCA
CAAACGAGCAAATCAAGTCACACACACACACACACACGTTCCTAACCGAA
CCAATGGGCAAAAAAACCCTTAAGAGACTTGGATTTTTCTCTTACGCATT
CTCCAATTAATTAACTAACTCTCTATCTCAAAGTTTACGAAATAAAAAAA
AAACTTCTCAAGTATCAGTGTCAAATCAATCCACTGCTGCCACAGCAACT
CTCCTCATGCACTATACACCCTCCACCGGTGTTTTTTTAAATAATAAACA
TTTACATTAAAAAACCGTTAAAACTCACCATGCTATTATCACAATTTGTG
CTCGGACCAGGTATCATATGGCTAGCAAATATGTCATCGCCAACGAGATT
CGAATCAACGTTGCGACGATTGCCAACATCTTGCCTACTGGTGCCAGGCT
AAAATGAACAATAAAAATAATACAATATATAAANNNNNNNNNNNNNNNNN
NNNNNNNNNNNNAAAAAAAAAACTTTTCTCCATACTTTCGCAGATGCAGC
ATCATTATTCCAAGCCTGCCCACTGGCAGCGGCTGCGTCGTTCATCCACG
ATGAACTGGCGTCCAGTACACTACAAACAAAACCATTAGTATATNNNNNT
TAGTTTTATTTCAAAAAAAAGAAAAAGAAATTACCCAGTAGCGTTCAAAC
CACCATACGGGCTGTTAATACCAGTAGTACCCCAACCGTTCAGAGGTTTC
ACAGGTTGTGGAGCGATTGGAGAATCTTTCCAAACGCTACCGGCATTCGG
CAGAATAGCCACCTTAGAACTTGTGGATGGAGCGCCGGCAGAAGTAGACG
GAGCTGGATTTGGAGGAGTGCACACCGAATGGGCATCCTCATTACCGTTT

TTCCAGCGCTTAATCTCACCTTGAGCGGCGGGATCGCCCCATATGGCAGT
ACCTTTATCCGGCATTGTATTTTTCTTCGGATATGGATTATTCCATATCA
TGCTGCTATCTGTAAAGTAAACGCAATGGGGTCAATTAATCAATGCAAAA
TATACAACACTTACCTTGTGCCAGTGGCGGGCCATTCCAAATGTTGCTAT
TGTTGTTGTTGTTGTTGTTACCGTTATTGTTAGCGTTGTTGTTATTGTTG
TTGCTGCTGTTGCCATTGTTGTTGTTATTGTTGTTGTTGTTTCCGTTGTT
ATTCGGTGCCCCTTCTTGACTGTTCCAGGCGTTATTTCGTGAAGTACTAG
TACTCTAAAATGTAAAAACAACAATCATCATCTGAATCTGAAAAAGTCAA
ATTGCTTTCTAAAAATATTATTTCACTAACTAAATCTAAAATAAAAAACA
ATTTAAAATTCTNNNNNNNNNCTCCAACAAGAAACGAACACTTACTGGAA
CATCTTCCCATCCACCTGATGTACGTGGTGGTGGTGGTGGCGGTAATGGT
GGAGCAGCCCACTGCTGTTCTTTATCAGTGGGAGGCTTCCAGAACATACT
GTTGTTGGTGTTATTGCTGCTATTATTATTATTATTATTATTGCTGCTGC
TGCTGTTGTTGTTATTATTGCCACCGCTTGAATTGTTAGAAGTATTCCAG
GCCGATTCACGTGCCGCCGGCCTGCTACCCCAACTTGAAGATGCCACCAA
AGATTCGTCATCCGTTCGTCGTCCTCCAGCATTAGCTGAAAAGAACAAAA
AAATTATATTAAAAAAAGTAAAAGAAAATACAGAAAAAAAATAACAAGTA
CCTGGATCATTGCCGCCGCCATTGGTGTTATTGCTGCTACCAGCATTACC
AACATCCCACGGTGTTCCCTGATCTACAGGTTTTCCCCAATTTCCTGTGC
CCATTTCGTAATTGCAGTCTTTCGCAGGGTTTATTACACTACCAATTGGA
TTTGATTGTGCAGGAGGAGCAGTATTGATGCTTTTTCTTGCAATTTCCGA
CCAACTGCTGCTGCACACAGATTGGCTATTTTCGCCCTGAAAAGAAGAAA
ACAACAATACAATAAGAGAGAAAACAAAAATAAAGGGTGAAAGKTATTCT
TTTTTCGCTTACTTGGGTCGTTCGACTTTGAGACCAAGCAGAATTGTTGA
CGTCCGCACTCGCTTTCACGACTCCGCACCAGTCGGAGCTGCCACCAACA
CTACCAGTAGGACCTGAGTTTTTCCAATTAGAAGCGGAGGTAGAACTGCT
AGGGTCCCAAGCGTTTTTGGAAGAATCTGAAACAAAGGGAAAAAAAGAAG
TCAAAAAAAANNNNNNNNNNNNNNNNNNNCTGATGATGGATTACTTTTTA
CACTAGACGATGAATTGCCAGATATCAAAAGATTGTTACTATTATTATTT
GGATTGCTGCTGTTCACGTTACTCCAAACAGAAGAAGAAGAAGAAGAAGA
GGATGAAGAAGAAGAAGTATTTGGATCTAAATTTTCAGTAAACAGAGGAT
TCTCAGTTCCGCTGTTGTTTCCTCCGTCAATATTCCACATTATACTTGAG
CAGGCAAATAGGAATCTGACATTGGACCGGTACTCGCTAAAAAAGAAAAA
AAAATTAGAAGAAAAAAAACAAATTGAAACGCTTTGTGCCAATGGAAAAA
GCAGTGCACAACGATGATGCATGTAGGGGTATGTGATGATGATTCGCGAC
GGCAGCAGCAACAATTTTACTAATTCTTATTCTCTAGATAGTATAAAATG
TGAATAAATAAAATAGCATGCGCGCAAAGTTTGCAGATGAGAAGAGCAAA
AAAATGAAAATGGAAAAAAATCTGCTGTCGAACAACGACCGACTGTGTGA
CATGTGTGAATGCAGAATGAGAGAAAGATGACAAAACAACAGCTCTCTCT
CTCTCTYTCTCTCTGAATGGAATACTGTAGAGATGACCTTCATCAACAGA
TCAGACAAGCTTTAAACGAGGGCAATCGTGTGTGTGTGTGCTGAGAGGTT
GAGATAGAAAGAAAGTACGCGTTGATTGGCTGCTTCTTGGTTGCACTCGC
TTGAACAAACAGAGCAGGAATTACCTCAATCTTTGTAATCAAGGGCAAGA
GAGAAGAGGAAAAACAAAGMAAATGAATGATTGAATGAATACCAGTACAC
TTTTTTTTTGGCATCATTGTTCCGTTTATCAGCAGATAGATAGAAATTTG
ATCCGTTGGCCGCAATTACAATGCAGAATCCAAAGCCATAAGGTCAAAGA
AGTTCACCATAATTATCATGATGATCATCAAGCAAAAGTAGAAAAAGAAA
TGGGCCGCCACCGCCTAGTATTATATTTTTCTCTTCTGCGATTGCAATTC
GATGCTGCTGCTGCTATTATTATTATTATTATTCTATAAAAGTGTTTGTG
CTCGCTTCTGATCAGCCAGCCAGCAAGCCGGAGTAGCGGTTGAAGAAACA
RGCAATGCTCGAGAAAAACTCGTTTACCGAAATGAAATCGCACGCAATCG
ATTGTCAT

>hypothetical_protein2

ATGGAACTTCACTTAGCAAAGTTTAATGGCGAAAAAGTGAAACACGCACG
ACCGTATCGCAACCAAAAGCCCACTAACGCGAAAAATCAGGTTGTTTTGA
TTTTCTCGATTCAATTATACTTGCCAATCAATACTCACAGTATGCATAAA
TTAAGTCGATCAATTGCAAGTGCGGTGGTGATAATTTTCCTGTTTTCGTT
ACTTGTTTTTTACTTTTCAAAAATTGTAAAATTTATTACAATCAAAATGA
AATTAATGTTACTTAATTGCATTACGCCATGCTGTCGCTCTTGTCAGTTA
TTAAATTAATTAGATTTTTTTTTAATTTCAGTTTTTTTTTCAAATTTTTA
ATTTAACATCAAAACCGAAAAGATATCAAAAGCAAAATATATCAATACAA
ACTACTCTTGTTTACATGTCGTTATAAATTTTATCCTTAAGCCTTGATCA
CTTATCCAATCACCCAATGTATAGTACTATTTATACTTGAAAATTTTCTT
TCTTACAATATAATATATGGAACCGGTTGTGGAAAATTTTCTAATTTCGA
TCATACAAANNNTTAGCACTGTACAATACTTGTTTAATTTGAAATGAGAA
CAACTTATTCACCTTTAGATGCAAAACCAAAATTTGTAGAGTTCTGATTT
CAAAATTGTAAATCAATGAAATTTCTTAATCTTTCAAATGATAAAAAAAA
TATCAGTGAATCATCTGGTACATTTGAAATTTATTACAGTTAACAATACG
ACAGAATGAATAGTAATTTTGAAACAAAACAATACCAACACATCACACTT
ACGGAAAACTTGAACTTTAAATTATCATCTACTTCGAATTTTAAGAATAT
CAYTTACTTAACTGAAATTGTTCACTGAAAATAAGATGAAAAAGAAAAAC
CATTATTATCACATAATAAAATAATATAAAGAAAAAATTGCCATTAATGT
TTTGTACTTATGTAAACCATGACGTCACGTGAAATCAGTTATGTGTCATG
TGACAGTTTTTTTCCAAAGGTTGGAAAAATTATAACCTTTACAGTCAAAA
ATGGACAATACGTTTGTCATAGTTACAGAATTAAAACTGAATACATTTAT
CTGAACTGCACTGCTTCNNNNCGATTTAAATTCATTTATTAAATATGTCC
TCAATTCTAGCTATGCATNNNNNNNNNNNNCAGCAATGTAAATTTTGCAT
TTTAATGAACANNNNNNNNNNNNNNNNNAACAGTAGATTGCAATTTTGAA
AATAAATTTCGAAACTAAATCCCATTTCAAAGTGTATGTGTTNNNNNACA
AACAGTGAAAAAAAAAATTTTAATTTAAGCTCTTCTCAGTCAAAAACGAC
AATCTTTCAAAGCTGTGCAGTTTGTTGAAATGTCACAAACTTTATTCGAT
TAA

>conserved_hypothetical_protein2
TCAATTCTCGCTAAATATAATTATGAGTTGAACAATAATGTAAATAAATT
AATTAATCCAAAGCGATTGCAGTGTAGTTTCAATACACACAATTTGAAAA
CGAATTCTCAATTAATCTAATTCAGGAATAGCATATCAAACAAAAGGAAG
CATTAAGTTCTACAAATAAGTACAATGAAAACCCTAAAATGCTGCTTTTT
CATCCAAGTCAAAATTGACCCGGTTTTAGTCATTGGAAGGACAAAAAATA
TTTTAGTCATCCAAAGTTAAACAACGTAAAAATGAACCTGATTTTCATAA
TTTTATATAGCACAACAGTAATTACGCAAAATTGAAATTGTTGTTAAAAG
CGAACATCAAACAAATATGAAATAGTAAATATAAAAATCTAATAAATGAA
CAATTCAAAGCAGTAATTATTGGTGATATGAATCATTCGAATAAAAATAA
TAATATGAAGCCTTTGTAATACTTAAGNNNNNNNNNNNNNNTTATTATTA
ATCACACTACTATTTTACATTTTATAAATTACATTAAGTAATTATTATTG
AATATATTAATAATWATTATTATAGAATTTTGTACATATTTCAGTATGGT
ACATAGTATTACATANNNNNNNNNNGTATTTATTTAGTGTTAGTTCTTCG
ACTTACGAGTTGGAAATGGTTTCAAAATTTTGCAAAATATTACATTTCAG
AATGTATTTTCTTATGTTTTTGTTTGTATTCTTTCACTTTGCTAACTACA
ATTTATATAATTTATAYTTTATAATATTGTAACGCGCCGCCCAAGCCAGC
TTATTCCACAGCTTCTGAGCCGTCAAATCATCCCTGAGCGGCGGCATTCT
CCGTCGACTCTCCGTCAATGTTTNNNTTGTTCCTCTCGATTTATGATATG
CAACGTTTAATGGTAATAACAACACAACGCACACAATCCAACTAAACTCT
ATCTACATATATACCTCGGCTCCGGGTCATGACACACAGCAAGGCCTTCG
CCTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGATCTTGCCG
TGGCCTTCTGACCCTCGAAACGCACCTGGCTTGGCGACGCAATGCGTCCA
T

>protein_USP1
TCATAGATTCAGGCTTCGCCTCTTCTCCAACCCAACCTGGACAAGGATCG
CGTTAATAACTGCTACAATATATATAACCGTTTTGATGATAAACTAATAG
TTTTCAACAAATTTGAAAGCAAAATGTTTATAAATGTCATGGTTCGAATG
AAATCATGAATTTCATAAAGATAACAAAGCCATTTGCAAAAATTATGAAG
CATAATTCAGACGTGTAATTTGATTTAAATGTCRCTTTGCTATTTTATTA
CTTTATTATAAAATTTCTAACTTAAATTCTATATTACAATCTTTTACTAT
TTAAAATTGCACGACAAATTTCTAAACTTATAATTTAAACTTAAATATAA
AACAACAATTCCCAACAAATTTGTGGAATACACAAACTAAAACTGCAACA
ACGTTTGATCAACAACCATGTTAAATTATCAAAAACCCANNNNATCACTT
CCAAGGATGGTCATAGTANNNNNNNNNNNNNAAAATATTTCGCATCTACA
TCTGCTAAATTTATCATCATCAAGAGCATCAAAATTGAACTTTATTATTT
TAAAAAAACATGCGAACATTGTACATTAAAAAATCAACATTCTGACTTGG
AATAAACTTGATTTTGCCTTGCAGAAAGCATCTTTGCAATGACAAATTGC
AATCCTTTATTTGCATTGTTGGCATGATGTTTAAATCGTTCCAAACTTAA
TCGCTGGATAGCTGAAGCAAAGCCTTTCAAATTGCATTTAAACCACGCTT
CACGAGTAAGTATTGTACATTGCGAATTCATTGGATCCGCTTGATAAATG
CACTTTTCTTCTATGGACATCAGACGCGCGTAGCTAATGTTCCGAGTATA
GGTGATCAACTTTTGAGAATCACGATCTACCACAGACTCTTCCAATACTG
GAACTTGTTTGCAGGAAATAAAACGTTCTCCCCATTTTGGAATGCGAGTT
GTCTTTATAAAAATCCGTTTGGTGATCAACTTATTACCAACCACATTACG
AAAAACAGTGTCTTCACAAAGAACATGACTGCTGTACGGATTTGGATAGC
GCTGCCAAAAAGCGCTTACAACTTCATCCCAAGTATTTTTAAAATCGCTC
TTCACACAGTAGTATTTCAT

>conserved_hypothetical_protein3
ATGCTCTTCTTTTGTTCGCTCGATTCTATCCTCCTCTTTGTTCACTGATT
GGCTATTGCATTTTGATTTCTAATTGGATGGTTATTTATACTATTTCATT
AGCTATTTAGCGTTGATCTGTTCTTACGNNNNNNNNNNNNNNNNNNNNN
NNNNNNTTCTACGAATAATTWATCTTATATAGACATTCATTTTCTATTTT
TCAATAAAACCTTTCCTGAGACAAACTTCTGATGTRCAATTATGAAATTT
TTATTTCCTTAAATTGCATTGCATCTCTTTTCCTTAAATGTCAGTTTGTT
TCACAATTTAAATATTTCAGCATTTTCCTTAATTTCTAGAAATTCTGTCT
ATTCATAACAACTAGTTTAAAATTTCCTTTATGTTTAAACAAATTAATTT
AAACGTTTTGTTTAAACTTCAGCATTAAACTTTTAATGTTTTGTTTAAAC
TTCAGAATATTTATAATAATGGAAGAAAATGAAATAATTCAGTTGGTCAG
TTTCAAGCATGGCAAAACGTGCTCGTTTTCGTTCTCGCACTCGTTCTCGA
TCACGTTCCCGAAGCTATTCCAAAGGTAATGTTGATGCATCGCATTCACC
GAGAAGTCGAAGAAGTAGAAGTCGAAGAAGCAGATCAGCAACAAGTCGAC
GTTCACATTCTAGAAACTGGTATCCTAGAGATGGAAGTGAAGTATACCAT
TCGACGAGGGGTTGGAAAAGTGAAAGGCGAAGCAGATCAAGACGTAGAAG
GAGATCTCGTGTAAGTCGACGACGTTCCCGTTCCAGAAGCTATTCCAGAG
GAAGTAGTTTTGATGACCGTTCACCGAAAAGTCGACGAAGCTTAAGAGAA
AGAAATTCAAGAGAAAGAAGCAGATTTCATGTAAATCGGCGACGACGTTC
ACATTCTAGAAACTATTCCAAAGATCGAAGTCACAAATCCCGATTGTCAA
AAGATCGAAGCAGGAGGAAAGCAACAAGGAACAGATATTATACAAGTCGT
TCTCAAAGTTACCCTAGAGATGAATTTGAAGGATTCCGTTTGTCGAAAGA
TTATAAAAGAGNNNNAAGAGAAAAAGAAATAATAAAAAGAAGGCTAATGA
ATATCATTCGTAAACATTATTCTTCCAGAAATGATGGACTGTACACCTTA
GATTCTGTACCGGGGAATTATTGCAAAAATTGTTCAAAAGTTTTGCAACG
TCTGTATGAAAAAAGATCACATCGTCAATCAATAATCAAACCTGAAAATC
TCAGTCCGTGGAACGATCAACCTCAACCATTAACTCAAGCTAGTAGTGAT
GGCAGTTCGTCGCAAATTTCTCTCCATACCTCTGATGCCAAAGATGAGAG
ATCTCCAAAATTTTACACCTTTACTAACAGCGATGATTCAGATCACAAAG

TTAAAAAGAAACATCATTCCAAAAATCATAAGAAAAATAGAAAGCACAAG
AAGAAAAAGACTAAATCTGGACATGGATCTGTTTCATCCAGTCATCAGCA
TGATTAA

>putative_sperm-specific_protein_Phi-1
ATGAAACGGTCAAACAGTCAAGCTTGGCGTCGAACAGCAACGAGGCGCAG
TTCATCGAAAAGATTTTACGTGAAACTCAATTCCCCAGATCGTGGAGCAA
TTCGTTCCAGACCCGCAAGACCCCGAAGACGGTACAGTTCCACAGATACT
GAGGGTAATGATCCCAAAAGTAAAAGGAAGTTTCAGTCGAAGAAGCCGAA
GATTCGTCGTACTTGGAAAGGAAAAAGAAAAGCTAGACTTGGTCGCGGAA
TTGTCCCAGTCTATCACGTGAAACGTCGATAA

>putative_lectin_C-type_domain_protein
ATGTAACAAATTGTTGGAAAAAAATAAGCAATTTTAAATTTTCCTTCTTC
TTTATTGCCATTAATAAAAAATACATTGAAAATTTAAAAAAAACATATAA
TTGAAATAAACAAGTGATGCTGCATTGTAAAATTAGCAAAGCAATAAAAA
ACCATTTGTAAATTTGCAATTATTGATGCAAACAAAATTTTTTTCTTATA
TTAAAATGGAAAATACATATTATTAGCATCTTTTATATAAAGAGTAAAAA
TAAAGTTTAATTTGATTGAAGAGATGTTTTATCGCTTTTACCAAATTACT
TAAATTAGGTTAGATGTAGTGAAATTTCTCATTACTATCATGAAATCTTT
CATTGCAAAACCAAATTTAAAGAGAACAAGAACTGGACAACACTGTTTTT
CATTTCTGCACTGCATTCGGCAGGATTTTTTTTAATACAAAATAAAAGCT
ATAATTTAAAATTTTGTAATTTTAGACTGCGCTAATTGTTCTGCATACAC
AACTTCATAAACAGTGAACAACACTGTCATTGAAAAGTACATTTCTTTTT
GGACAAATAATAATAATAATTAACATTTTATGCAAATTTTGCATACATTG
ATTTAATCAATTTAAGATTAAATAAATTTTTCAGATTGAGGCCAACATGC
ATGCAAAATTGTTGACTTTGCTGCTGACTTTGTCGGCTGCAATTTCAACT
TTGGCCGAAGAAAATTTGCAAAACTGTCCCGATGAAAGCTGGATATATTT
CAACAATTTCTGTTATAAAATTGTTACCCAGAAGAGAACATTCCAGCATG
CAGAACAAGAATGCGCCAAGCTGAATGCTGAGCTATTTTTCCCTCCGTCT
ATTAGTGAATTTGTGAGTTATTAAAAAARAATTTCGAACGAAAAATATGG
AATTTTTTTACAGGGAAAAATTGTTCTTACTATACCTCATTCGAAAGAAT
TTTGGATTGATACCACTTTGAGAAAAAATCCCGTTTGGATAAACAATCCC
AGGCTTGACCCAGTACAGAGGTAAGATTACAAGCAATTTTTGTTTCAGCC
ATTTCAGCTTTGTTTTAGACTTTATTTATGGCCCAAACGATTAGCATTGA
AAGAAGATGGTAAAAATTCTTCAGACATTGAATGCGCAGCTTTTCGTCGG
GTGGATAACATCATTGGCGGATATGTTTGCTTGCACAGTTGTAGAAGTTT
AAAAAATTATATTTGTCAAATAGAAATTGACAATACACTACAAATGTAAT
TTTTATCAACACAGAATTTTAATTATTTAATAAAAGGTAGCTAAGTCTCC
TTTTTAACTTAGCTTAGATAATTTAAACACTAAAAATAAAACTTATATAT
TCATTTGTTCTTGTTTTTTTATATCATAATATGACTTTTTGATCATACAG
TTTAATTTTAAAATTRAAAAACCAAAACTATCAATACGTAATGTTTTCAA
AATGCATAATACAGTCGAATTTGGTTAAATGCACATGGAATTTGCCAAKA
AGCTCCATGCATTTATCGAAAGTTGGAAATTATGCAGAGTAAYTGAAAGG
AGCAATAACTTCAGATTTATGAATATGTTTAATGCAAAAAATATCCATTT
ATTTGAAGTAAAATACTCGGAAATGTAGTTTTAAATTCTATTTTGTGGTA
CCAGCGGTACCAGTNNNNNNNNNNNNNAGTCTAAGTAGTCTGATTTATGC
TGTTTTCGTGAAAATATGA

>receptor_expression-enhancing_protein_5

TTATTCAGTGCTCATTTCAACGGCAGTAGTACAGCCAGATATAAACTGGC
TGTAGCGAGATTGGAAGAACTTCACATACAGTATTTCAACTCCAAAACAG
CTCGGCGTGAAAAGCGAAAGGAAGAAAATGCATTTCGGTAACCAATAGAT
TGCAATAAACAGAAGATTGTAGCCAGCAAATAAATCAAAAATTGTAAACA
AACCAAAAATAACCCAATAAAACATCCACTGCTTGCATCTGCTTTTCTCA
ATGGATTCCATCTCTGCTAAAGTTTTCATTACTGGATAGACAATGCAAAT
CAAGTCGCACAACAATTTTACAGCGGGATTGTATATAGCCAACAGCATAC
AAATCACACCAGCCAGCACAATCGCAGCTTGATAACGACTTCGACCACTT
CTTTGTTCAATTTTACCAAAAAGTTTGCCAACCAATGCGCTGTCATCATT
GATGGCATTGGCAACATCCTTGCAAATACTTCGGACAATTTTTTCCAT

>asparagine_tRNA_ligase
ATGAAAGTCAGCTTCCAACTATTCAACTGCGCCAAAACAATTGCTCAGCT
TGCGAAAGAGAAACCAACTGGCCAAGTAGTTACGGTTAAGGTAAGCTGTA
TTTCAACATCAAATAATTTTCAACAATTCAAAAGAATATTTGAAAATATT
TGTARATTATTAAATTTATCTTCAARTTTTGTATTATCGTATTATTATTA
ACTTTTATAATATTGCATTACCTTTTTTGTTGATTTTGAATTCTATTTCA
CTTAAATGCAGATAATGTGAACGGTTGATCTAAGCAGATTTAACTAATTT
ACTTTAATGAGGAAACTTTCAGTATATATTGTTAATTAGGATCATAGAAA
TTGTAATACCAGAATTCAGTATTTCACGACATGATGAATAATCTGCGATT
CAATGTTAATCATTCTAGGGTTGGGTGAAGCATTTTCGAAGTTATAAAAA
TTTCACCTTCATGCACATCAGTGATGGCTTGTCTGCTGCCCTTCTTCAAG
TGGTTATTCCAGAGCCTTTGAATAACAGGTACGCAAGATGAAATTTAAAT
TAGACGGCTATTCCAATTTTTATGCAGTCACGTCACCGTTGGATCAGCAG
TTGCAGTTGAAGGAGAGTTAAATCCCAGCGTTGGTCCAGAACAGCATCTC
GAGTTGGTTGCTACTCAATTAAATGTCGTAGCTCACAATGACAGAAAAGT
ATAAGAAATATTTTCACGTAAAACAAAACAAATGTTCGTTATTAATAGAA
TTTTCCCTTTTCTGTGAGACAAAGGCAATCACCTGATACATTACGTCAGY
ACGAGCATCTGCGTCCAAAAACTCGCAGTTTCGCAGCGTTACTACGAATC
AGGAGTAATTTAAAGCACTTTATACATCACTACTTAAAGGTGAATGATTA
CACCATAATTTGGAAAAATTATAAGTAAACTTAATTATGCTACTGTTTTA
TAATTATTCTTCAAAGATTGTGARCATGAGTTTAAATTTGTTTATAGTGT
TACCGTAATCATAATAATCAAGTTTTTGTTTTGTGATAATGTTGATTAAA
TTTATTGAAATCATATTCAGACTGTTTAAACGAAAAGTTGTTAATTTACA
TTTTTAGATTAATAATTATGTATATATTGATACTCCAGTTTTGACAGCCA
ACGACTGTGAAGCGGCTGGAGATGTTTTCACCGCTGAAAAAGTAAAAGTT
TTTATTTTAAAACTGTATCCGTTATAAAGCAATTTTAGTTTGAAAAGAAT
TCAGCGTTGGCATTTGCATGTGACAGTATACTAGAAGACAGCACTAATGA
GAAAACTTTTTTTGGGAAAAAGATTTTTTTAACAGTTTCTGGTCAGTTAC
ATTTAGAAGCTGCCGCTTGGTAATTAATTTATTTTAAAATAAATTTTTTT
CATTTTAATTTTCTGGCATTAATATGTAATAACAGTTTCTTTAGTGGGTT
GAATAAGGTTTATACATTTAATCCAGCATTTCGAGCTGAAAGTTCTCTTA
CCAGACAACATTTGGCGGAATTTTACATGCTTGAGGTGGAACAGGCGTTT
TGCACTAGTTTGGAAGACTTAATGAAAGTAATTATTCTGTAATTGTTAAT
TTATAAGGCTAAAANNNNNNNNNTTTCTCTGTAATTGTTTTGCATATAGC
CACATGATTATGATTATGATTAAAATGATTATGATTTTTRATTTTAAATA
TTTAAATAACGAGCTTATCAATATAAATTTATCGATATATTTTATATTGT
GTTGTTTAAATTGTATGGAATTTTGAAATTTATTAACTATAAATTTAACT
ACAATATAATTTATNNNNNNNNNNCGTTTCAACAATGCATAAATATTCTT
AATATTTGCAGTAAATGTTTTGCTGATGAAATTCTCCATTTTTCAGTTAG
TGGAAAACTTAATCTTAAATTGTATCAAAACAGTGATGCATGAATCAGTT
GAAGATTTGAAAATTTGTCAGCAATATCATCAGACGGTTGCTTTTGTAAG
TATTAAATTTGCATTTGTTTTAAATTTTACTGCGAAGACTTTTCATGATG
CAAGTTTGAAATGACGCTTTAAACTTGATTTAATACTTTTATTTTTAGGA

TGATCTGTGTCGTATTGAGAAACTCAGTTCTTTTCCCAGGATTACATTTG
TTGAAGCAACAGAAATTTTGCAAAATTCAAACAAAAACTTTAGCAAGGTA
TAAATTTATAATATATGCAATCGTAATTATATGTATCCTTAATTTAGAAA
ATTAATTATATGTTGTAAAAATATGATTTGTTTTTTTTTTAATGAAAAGA
TATCAATATAAAGTTTGTTAATTGCATTTAGTTTAGAATTTGAAACATCT
TAGAAAGGAAGATGAACGATTTTTGGTGCAACATTTTGGAAATTTGCCTT
TGTTTGTTACAGAATACCCAGCTGAGACGAAACCATTTTATATGAAGCGG
AATCAGGACAATACATGTGTAAGTAACGATTATGCACTCGTTTTCATTTC
TGTATTAAAAAGAACCAGTATATTTATTTATTATATGCATATGTTTCAGG
CACTTTGCTTTGATTTATTGGTCCCTCATTCCGGAGAATTATGTGGAGGC
AGTCTTCGTGAATGTAATCTTGATATTTTGAGTCATCGTATGGAATTATT
GAAAATTGCCGAACAAAGTCAATGGTCAAAGAGCACTTTTTTTTCATAAA
CGTTTCTAATGTGGTTTATGAGAATATGTTTTTTAGGTATAGCGAATTGC
GAAAATTTGGCTATCCTGATTTGGGTGGTTTTGGTTTAGGATTTGAGCGC
TTACTGCAGTTTTTCATTGGAATACGAAATATTAAAGACTGTATCCCGTT
TCCTCGATGGTTCAGACATTGCAGCACTTAGTTTACTGAATGCAATTTTA
TTGAAATTTCTTTTCAAAATTTTTGTTTCATTTTCTAATAATAAAACTCT
GTTTACTTTGTTTGTATATGTATATTTAGTTTAAAAATATTTTTCTAATA
ACCAAAACTTTTGTTTTCTCTTTCCTTACACTGCACTCTTTATTCTTTCA
TCAACTGTTATTATTATTATTATTATCGTAAAAGAGTGCAATAACAGAAA
AGGCTTTGTTGAAATTTCTTTTAATATAACAGTTATTAGTTGATTGTGTT
GGCAACCAACTCTGCTAATAACCGTTGATAATTAGCGTTAATAAATTTGC
TTGTCCTTGTTTCCAAACGAACTTGCCCTTAAAAGCTTTATTTTTTATTG
TGAAGCTAAACGTGCCGGTAAAATTACCCACATGGTAA

>conserved_hypothetical_protein4
ATGTGCAAACAGTGGTTTCTATTTTTTCTCGTCGCTCTTGCTGCAGTTGT
GTTTTGCAACGGTTTGCCGTGCCGTATCAAATGGCATTTGCATGGTCAAA
ATTGCTATCGTATCATGAATCATAAATCAACATTTTCAACCGCTCAAAAA
ATATGCAAATTGTACAATTCAAAATTGCTTGTTCTTCGTTCATTTGAGGA
AATGGTAAGTAAGGCAGTTATATGTTTTTTAATATTTGTAATTTKCTTTA
TTTTTCAGAAAAATGCGTTTGAAATGTTTTCTGATTCACAATTCTTTTGG
GTTACGCTTAATATGACTGAAGAAAAAGATATGGATGTTCAAAACACGCA
GTATTAGAAAACATTTGTTTTTTTAATCGTAAGCATATTGATTTTAAATG
ATTTTAGCGATTTACTTTATGCTATACCTGTGATCAATGGAAGGCAGATG
AGATCAAACGAAGGTGAAAAATGTTTTGTAATAACAAAATCGAATGCCGA
AGATGAACGACATGTTTTACAAAAAAGCTGTTCTTTGCGATATAATTATA
TTTGTCAATCGAACGTCGAAATTGATAACAACAACAGCACTTTTAAATGC
AGAGATGATTAAAGTTTGTCTATATTTGCATCTGTTCCATAAATGGGAAA
TGTGAGGCATTTTATGCATATTTCATGATTAACATTGCAACTGTATCCAT
CCTTTTTAAAATTGTAGCTTATATTTTACGCAGTGAAATGAAAATCACTA
TCAAATACTTTTCTGTTTGTTAATTTGCATATCATTATTTAGTTATACCC
AGTAATANNNNCAAGTGTGARTAAATTTTTATAAATTTATTCTGTTTGCA
TTTAATGTTTTTAAATAATAAAACTGACGACACTCTTGCATTATCAAACT
TAATGCTTGGAATGTTCATTTGAAAATCATACTGTAGCCTTNNNNNNNNN
NNNNNTGCAATGGCGTTGACATGTTAGTTGATGAAAAGAAAAATTGCTAA
TGGGAGTTTCTAGAAACCCTGCCGAACAGCAATCTCAAGTCGAATCAGTC
AAGCGAAACGAGCTGTCTTCACTGCAATTTAG

>netrin-1
TTAATGCCTCTCTGCAGATAAATTGTGTAGAAATGAAATTAAAAAATAGC
ATATTACCTGCATCAAATCCATCTTCTGGAAAAGCTGAAACAACTTTTAC

GCAAGGAGCAATAGGTGATCGACTTTGTTGATATCCTTTAGCGCATTGAT
TGCAAGTGATTCCTGTAACGCCTGGCTTACATGGACACTGAGCCGTAGTT
TGATTACAATATCTGCTCAGCGCACCAACAGGATGGCAATTGCATGCTAA
AATATGCACAACATCTCTTGAATATTTTTTTCACAAAGTGTCAATTTACA
CATTAATAATAAAAACGGTCTAAACCTTTTTAAAATAACTTAATCATCGN
NNNNTAAAGATTAAATTTTTACTACTTTCAAAATAATATCTATGCAGAAC
GTTTACTTTTTTCATTNNNNNNNNNNNNNNNNATATGAAATTAATCTGTT
TTATTTGTCTTCTGTTCTATTTCATATGTATAATACAATAAAATTTGCAA
TTCCTATTTGTATATACAAGCTGCAATTGTGGAAACTGAAATTTAAATGC
TTTCATTCAAATAATAGCTTTGTATAGGCTAAAGCGGTATAATTGACCAC
AAAGCAGACAGACACAACGCTAAATTCAAATYGCCAAACGGAGATAGGAC
GTCACTGACACGAGCCATTGTTTGCCCTTAATGACAGTGCAATTTTGATG
ATTAATGATCATGCTAGCTACATATTGGCAAAACGGCATCTCTCAGATTA
CCCCGGATTTCTGGCAATCCAAATGGTCACATAAACCTGACAACGATTCG
TTTGAATAAGAAAATCAATTTCTTTTAATTGGCTACTCATGGTTTGTATC
CATCCGGATTAGATTTTTTAGCCAATATGATATTTTTTAGAGCTTTATTA
CCTGTATTGTGCAAAATTCCCTTTCCACCTAAGGAAAAAGTAATAAAAAA
AACAAGTGTTACTTACGTTTGCAAGCTCTTCGATGAGTCATTGGCAGTTT
GCGATCACGATAATAACCTTCTTTGCAATAGTCACAGTGAACTCCGGCAG
TGTTGTGGCGACAGTTGATGCACATGCCTCCTGATTTATAGTTACTCATA
CGATACAGTTCCAT

>ATP-dependent_RNA_helicase_DDX42
TTAGTTCCATCTGGATCTTCGTTTAGTAGCCTCCTCATCCGTCGACTGGC
CAGTTGAATCATTATCACMACCACCAGCAGAAGATGAAGCAGCTTCCGCG
TTGGAGGAAAAATTTTCTTCTTCTTCGTTCATTATTAATTKTTTTCCTTC
TTCAACTTGCTGTTCATGAGCCACTTGGAAAGCTTGTAACTTTTTCTTCC
ATTCTGGAATTGGATCCGTAAACGTTGCCTGATTTGTCCCCATCACTGGT
GTACCGGCAGATGGTTTAAATGTCGACTTGTAATGATTCTGTTGAATGGA
AAAAAACAATAATCAAGCTTTCACTCAAAAAATAAAAAAAATATTAATTA
CAGCATAAGCAGCTTTTATTGAATTCAGCCTGTCTGAACTGCAAGATGCT
TCTTTCACGCTCTGTTGCAATTGGGCGTTAGCTTGATCTGAAAACAAAAA
ATGAAAAAAAAYAGACAAAAAACATATACAAAACAACAACAAAAAACAAA
AAAAAACAAAAAACAAACAAATATATAACCTGATGAATTTCCCTTCTTGT
AACCGAAACCAAAACCCCCAATAGAACTGGGCTTTTTACCACCAACTGAA
TTGGAACGAGTTTTGCGAAACCATGCACTCTGCATTGCTAAACTGAGCAA
AGCAGTCGGAACTTCTTGATTAACACTTTCCAAACTTTTCACCAAATGTC
CAGCAAATTCTTTATCCTTGTCCGTGACCAAAGTATACGCGAAACCTTTT
TCTCCTTCAACAAACCATTATTATAGAAATTAGCATAACAACAACAACAA
CAACAAAAACTTGAATATCTGAAAAAAAATTCCACTGACCAGCTCGTCCC
GTACGTCCAATGCGATGGACATGCGTATCCAGATCTTTGGCCACATCGTA
ATTAATAACCGTCTTAATTGTTGGAATATCCAAACCCCGAGCTAGAGAAA
TTTTTTCAATGAACAACACAATATTTTAAAATTTTTATCAACAATGAATA
TACCGGCAACATCGGTTGCGATCAAAATTGGTACATCATCTTTTTTGAAC
GCATGAATGACTTGGTTGCGTTCGTGTTGAAACATGTCACCATGCAAAAT
GCAAACACGGAAATCTTTCGCTTTCAAATTCTCTGCAATTTCTTCGACAT
GCACTTTCTGGGATATAAATATCAGCACCTTTCCCACTGAAAATTCTTCA
TTTCATTAAACATCATTCAAATTTTACAATTATTCATTGTCACTTACTGG
AGCAAAATTTTACCAAATTTCCCAACAGCCAAGTCCATTTTGTCGGAGGT
GAAGCAAAATATTCCACTATCTGCTGAATGTCAGCGTTGGCCTGAAATTT
GAAAATCAATTTTAAAAAATCCAAAACAGCATCAACAAAAACTGTATGTT
AACAGGTGAAAGCGAACCTCTCCAACTTCACCCTGAATAACTTTTACAGG
ATTTGTCAATACATCGCGTGCTAGACGTTCAATTTTCTTTTTAAAAGTCG
CGCTGAACATCAAACACTGACGATCTGGACGAATGTGGTCCGCAATCGAA

CGAACTTGAGCTTCTGAAATATGGCACAATTTCACATTAAAATATAAAAA
TTTTTTAATTCAGAAAAGARAACCAAAGCCGAGGTCGAACATTCGATCCG
CTTCGTCTAAAACAAAATATGTAACGCGCTCGAAATTTGTTGCCTTCGCT
TTAATCAAATCAATAATACGCCCCTGAAACCAAATTTAAAAAAAACTCTT
ATTCACCAATATCACCAGTGAAAGCAAAACAAAATACATACAGGAGTACA
AACAACCACCTCAGCGCCTTCGGCTAACGCTTTCGATTGTTCCCATTTTG
AACCTCCTCCGTAAGCGCAAACTACTCGAATATCAAAATATTTGCCATAT
TTTTTAGCTTCGTTATAAACCTAATTTTCAAAATATTTAAATCTCATAAC
ACTTCTAAAAATTAAAATATACTTGTATGGCCAATTCTCTAGTCGGCACA
ACAATTAAACAAATAGGTCCTTCTTTCTCATCAAGGTGGCGCTAAATGAA
CATTTTAACATTAGCCTAATAATTATTTAAAAATGCAAATAATAATCCAA
CAAACTTGACTAATAATATGATATATTGCAGGCCACAAATAAGCTGCCGT
TTTACCACTGCCAGTCTTGGCAATGCCTATTATATCTCGACCACTCATTA
CAACTGGTATTCCCTAAATAATATATGAATAAATAAATAAATAAATAAAT
GTTGCACATTAACAATACTTGTGCTTGGATAGGAGTTGGCTGGGAAAATT
CATATTTTCGAATCAGGCTCATCAATTTTTCATCAAATCCAAAATGAGCA
AAGCTGGTAACCAATTTTGGTGCTCGGTCGCCGGTGACTTTTAGATCCAT
TTTTTCACGCAATTCATTTACTTGTGGTTGGGTCAATTTTGCAATATCTG
CATGTTCAATATAAAAGTTTTTTTCAAATGGTTCATAATCAATTTGCGAA
TGATCAATCGGTGGAAGAGGGTCAATATACTAAACAAAACAGCAAATGCA
AACTGANNACTCAAAATAAAGAACTCATAAATTACTTTGCATTTTTTATA
GATAGGATTTCCTTCTTCGTCATATTCTATATCATCCTCTTCTTCAACTT
CTGAGATTTTTCCAAGTTTATCTTTGTTTTGTTCAACCCAGCTAAATTTG
ATAATAAACAAATRCAATAAAGTAACATTCAAAATATTCAACATACTTAA
AGTACGATTCTTGTTCGTCTTCATCTTCAATATCAGTTCGACCAAGTTTA
CTGCTAGTCAATTTTTCATCTTTGGAATGTTTAGCATTTTCTATACTAAT
TTCCTTATCTCTTTTCGCCTGTTTCTAAACAAAAACTACTGTCAAAATAT
GCTTAGTAAAAAAAGAAAGATTCCATCACCTCGACATCAGCCATAAAAGC
ATCTAATGGATCCACTTCGTCATCACTGTTAGATCCTCCTTTATGTTGAT
TGGAAAAATTAGATTCTGCTTCATCAGTATCTTCCAAGCTTAAAATACCC
AACAATCATTAAAAATTTCTCAAAACAAATTTATGAGTCTAAACACTTCA
TATTAATTATAATTCTAAACTGAAACTTGTTGGCCTAGAAACTGAAATTT
TACTATTTTGCATCGGCGCATTGCATGGATTATTTTAAACTTACTTATCA
TCGTAATCACGCATTCTTCTTTTACGAATTTGTCGCGAATATTCCATCTT
AGCGCATTCAAATTGCGACTGCTGATAATTCGACAT

>conserved_hypothetical_protein5
TTAAAGAATCTGCTCCGACAATTCTTGCGAACAACTTCTGATGCTCAGGC
AGTGGCCTGTAGGCATGGAAGTTAGCGGCGACTCAAGTTCTAGAAATACC
GACCAGGAAAAACAAAATGTTCTGGATAGGGATTTATTGATCAACTTTAC
AACACTAAAAAGCATGCAAGAGCTCGATTCTCGTTGATGCTTGACCWTGT
AACGAACCGTCTGCCGTGGYTGATGGATCAGCCTGAGTTTCAGAGTAAAC
GCAACTTTCTTCGCGCTTGCAATAACCTTTTGTCCACGCTCGCGTTTATT
TTGCTGACATCGAATAAATACAAGAACTAACAACTGAGCAAGACAGAACG
CCTTAATATAACAAGAGAACAACAACCAATGCACTGAACCACATCACGAC
CGACTAACCAGCGTACCGCCACCAATTCCAGTCATCCATTCTCCACCGTA
ACCCCATTATAGCAATCAAGTGCAAAAGTGCTTTCAACTGTAATTTACAT
GAAACGANNNNGCGTGAGAAGGCAATCAAGTACGCGGTTCCGGCCCTTCT
GCAGCCGTATCTCTCACAAAGCGAGCGTCACCGTCTGGCGAGGACTGTTG
CAACCATTTGGCTCTTCCACTTTGCTCTTGCTCTTCGCTTTACTTCACTT
TACTCTGATTCAAAAATGACAACTGAGTCAATGTTCTGAAACAGAATCAA
TAGCTGCGTTTCAGAATTTTTCTGGTGTAACGAGCCATCCTGGTCTGCCG
GTTCTGCAAATTCCACGCCATTGAGTCTTGCCTCGACGGCCGTTAGTTCT
CTAACACACCCAACACCAATCTTTTACATTTTCCCGCAGTTCAGGMTAGT

AAAACCGCTGATACACTTTTACCAGGGCTTTCGCCAAGCCCAGAAGGACT
TCTTTCTGGTGATTGTGGATGGCTCTCAGAATCTCCGCGATTTTCTCACA
GGGAATAAGTTGCAGCAACCTTCTCTTGTCAGTGTCCAGTATCTCCCATA
GGATCCTTCGGTTCACGGTTATCTGGTCCCGCTGGGCCTGAGGACTTTTT
ATGATGTGACTGCTCTCTGGAGCTTCTAATGGCCAATCCTTCACTATAAT
CAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTTTCTGCCAT

>putative_immunity_protein_PlnM
ATGAAACCGCATCAGCATTCCGATGTTTCTGTCCAGGTCGGTGAATCACT
TCGAAATCATACATACACACATACRTACGTACAGGCATATATATATTATA
TGGCTGCATACAATACATATATATGCGTATATGCTTTTCTTACTGCATTA
TAATATTATACACAGACAACGAATACATTTTATTTGTGCAAACGCAAAGG
CCAAAAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNGTTTGTTGATTTTTTTTATCTTTATGTATTTGTCTTTTTC
TTATTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNTAAGCTTGACTGCAGAATTCAGTTA
ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNAGAAAATTAAACAAAAGAAAAAGTANNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAATGA

>conserved_hypothetical_protein6
NNNNNNNNNNNNNNNNNTATAAGTGATGAAACTATGGTCTTANNNNNNNN
NNAGGACACAGATATTGGCAATGTGGTTTCAGAGACAGATAATTCAGTAG
ATGAAGGCAAAGCAGTAAACTGCCAAATTTGCATATTCTGNNNNNNNNNN
NNNNNNNNNNNNNCATCCCCAAGAAAAATAAAAATGTTTTGGCTCTCAGT
ATGATGCACAATGATAGCAGGNNTAGCGATGGAAAAGGCAACAAGCCTGA
TTTTATTGTACACTACAATAATACTAAAGGTGGCGTGGATAACTTGGATA
AATGATATCAGCGTATTCTTGTCAAAGAATGACWGCTCAATGGCCATTGG
TAATATTGTACAAAATGATAGAGGTATCGGCATACAACGCCTACGTTCTT
TGGACAGATATACACCCAACATGGAATGCAGGACGGTTGCATAAACGACG
AGTATTCCTTGAGAAACTCTGCACGGCTCTAGTGCAACCAGAAATGTGAG
ACGCGAAACACTACCCAGAACAGCAGCAGCAAAAACGGCAGTTGAAAGGT
TGCGACAAGAAGCTGAGGAGCCATCTACAAGCACCAGCACTGATACGGAC
AGAAGTGGTAAAAAGAGCCCGATGTAA

>conserved_hypothetical_protein7
ATGACAAGACGTTATCGTGTTCCCTCTCAATCRCGATTACGTTCTCGAAA
TTATTCTAGATGCAGAAGTAATGCATGGCGTCCATCAAGAAATCAAAGAA
TCGGAAAGGGAAAAATACTCATAACACGCAATACATCTCACGTTAATCGA
CACCGTTCCCGTTCTAGAAGTTATTCTAGAAAAAGTCGCTCTGAATCATG
GTCACGGGAAAATAGAAAAAGCAGAAGAGGAAGAAGACGAACCTCTGCAG
CAAGGCGCAGATATCATGCAAATCGACAACGCTCACATTCAGGACACCTT
TCTACAGACATTCGTGATGCATGTAGTTCTTCAAAAGATATTATGGGAAG
AAGATTCCCAAAAACAAAACGAAGTAGATCCATTTCAAGTGATCAAGGGT
CGTCTTCTGAATTTGATAGATCCTGCAGTTCAAGCATCATTTCGGATTGC

ACGCTTGTAAATTACTGTGAAAGATGTTCTCAAATTTTACGACATTTCTA
TCCAGAATCATATATTCAATCGGCGACCAAATCTGATGTGCTCAGTTTAT
CGAACGATGAATATTTCCCATCAAACCAGGCCAGTAGTAGCAGTGTTTGC
AGTTCGTCAAAAATTTCCCTCTATTCCGCTCATGCCAAAGTTGAGAGATC
TCCAAAATTTTACATCTTTACTGACATTGAGGACCCTGATCATAACGTTA
ACAAGAAAATTCATTCCAAAAATCATCAGAAAAACAAAAAGAAAAAGAAG
AAAAAGATTAAATCTGGACATAAATCTGTTATATCCAGTCATCATTATGC
CTAA