2003 BELTWIDE COTTON CONFERENCES, NASHVILLE, TN – JANUARY 6-10

GENOMICS OF DIGESTION OF THE CABBAGE LOOPER, *TRICHOPLUSIA NI*

Deborah M. Thompson, Joanna R. Carlson, Sayed M. Khalil, Matthew B. Vanderherchen, Douglas D. Anspaugh, and R. Michael Roe
Department of Entomology, North Carolina State University
Raleigh, NC

Abstract

Among insect pests, some systems have been used repeatedly as pesticide targets. The field of genomics however, gives us a new tool in the development of previously untapped targets for pesticide action. The digestive system has been proposed as a likely organ of choice for identification of new pesticide targets (Wang and Granados, 2001). This is partly because the food source can be genetically engineered to produce the pesticide and therefore to deliver it to the insect through feeding. In order to take advantage of the pesticide development tools genomics can provide and to effectively target a species, knowledge of gene expression in the digestive tract of pest organisms is needed. In this project, 957 new expressed sequence tags (ESTs) were randomly sequenced from a cDNA library isolated from the digestive system of the cabbage looper, *Trichoplusia ni* (Lepidoptera: Noctuidae). Bioinformatics tools were used to assign putative functions to these sequences. Twelve percent of the ESTs were digestive proteins, 9% were ribosomal proteins, 7% matched previously sequenced genes of unknown function, and interestingly, 29% were novel proteins that did not match any published sequence. Over 29% of the sequences matched lepidopteran sequences with 21% matching noctuid genera and 2% specifically matching *T. ni.* The elongation factor 1 alpha (EF1α) gene, which is necessary for protein translation, was studied further. Bioinformatics analysis of the EF1α-like ESTs indicated that this group actually consists of two genes; the well-studied EF1α and an translation releasing factor, RF3, not previously sequenced in this species. The *T. ni* sequence database offers researchers the chance to explore specific genes present within this insect and examine new potential insecticide targets.

Introduction

Much has been made recently of the power of genomics to create "personalized" or "designer" pharmaceuticals (see for example Johnson, 2001). Similarly, research into the genomics of crop pests may allow us to better understand the molecular steps in plant infestation, and therefore to develop targeted approaches to preventing or reducing crop damage. These "designer pesticides" based on the genomics of insect crop pests should be highly specific for the insect targeted and will therefore have greatly improved toxicological and environmental profiles. Heckel (2003) provides a good review of the burgeoning field.

Our target for the study of genomics is the insect digestive system. When a pest reaches a food source, its first contact is through touch. Because this is the first exposure to a crop plant, many pesticides act through contact with the cuticle of the insect. The pest next eats a portion of the plant, resulting in contact with the organs of the digestive system. Many pesticides, including genetically engineered pesticides like Bt, are designed to act at the point of digestion. However, our ability to target the digestive system has been hampered by a relative lack of information about the gene products that are present in the insect gut.

The intersection of plant recombinant biology and gut genomics research provides an opportunity for the discovery of new targets for insect control. The mRNAs of these genes (and others that will be found as a result of this research) or their gene products may become targets for pesticide action. Novel receptors may provide leads for new chemistries for pest control. Knowing the sequence of the metabolic enzymes (digestive and those with other actions) may provide us with new targets for pesticide action. Having the ability to specifically "turn off" expression of gut-specific RNAs in Lepidoptera would be a strong tool in the management of these crop pests.

This is the first preliminary report of the results of random sequencing of a cDNA library isolated from the digestive system of *Trichoplusia ni*. There was redundancy in the genes sequenced with a high proportion of digestive proteins and other enzymes as well as a large number of novel sequences. Therefore, a final report of results awaits development of a Unigene database, where each allele is represented only once. We identified sequences that matched genes published in the NCBI GenBank database for Lepidoptera with a small percentage matching *T. ni* specifically. Nucleotide and amino acid sequence alignments allowed us to separate one group of related sequences, the elongation factor 1α-like sequences into two more appropriate groups; EF1α and releasing factor 3 (RF3). The EF1α gene is the first gene that we examined in detail from our sequence database, therefore other genes could also produce interesting results. The sequence database allows us to mine for new information and look for new possible pesticide targets of interest.

Materials and Methods

Insect Rearing and cDNA Library Construction

Larvae of the cabbage looper, *Trichoplusia ni* (Lepidoptera: Noctuidae), were reared on artificial diet at 27 ± 1°C, and a 14 h light:10 h dark cycle (Roe *et al*., 1982). The ages of gate I, last (fifth) stadium larvae were determined as previously described (Kallapur *et al*., 1996). Total RNA was isolated (Chomczynski and Sacchi, 1987) from guts of an equal proportion of day 1 (L5D1) and day 2 (L5D2) fifth stadium larvae. Poly (A^+) RNA was purified using a Poly(A) Quik mRNA Isolation kit (Stratagene, La Jolla, CA). The cDNA library was constructed using Zap Express cDNA Synthesis and Gigapack II Gold Cloning kits (Stratagene) according to the manufacturer's instructions.

Plating, Plasmid Isolation, and Sequencing

Phagemids from the cDNA library were excised from the Lambda vector and used to transfect XLOLR strain *Escherichia coli* cells (Stratagene) following supplier's protocol. Cells were plated on LB Kanamyacin (50µg/mL) agar plates coated with X-gal and IPTG (Sambrook et al., 1989) for selection of clones with cDNA inserts. White colonies (those containing inserts), were selected, picked, and transferred to 384-well plates containing LB Kanamyacin (50µg/mL) media with 15% glycerol by the Genetix Q-Bot (Genetix USA, Inc., Boston, MA). These plates were allowed to grow overnight at 37°C. The 384-well plates were used to inoculate 96-well growth blocks containing LB Kanamyacin media. The growth blocks were allowed to grow 16 to 20 h at 32°C with shaking at 260 rpm. The blocks were then centrifuged for 20 minutes at 4000 xg and the media was removed by decanting. Plasmid DNA was isolated using the R.E.A.L. Prep 96 BioRobot Kit (Qiagen, Valencia, CA) with the Qiagen BioRobot 9600 following Qiagen's specifications. DNA sequencing reactions were then performed in 96-well plates using the BigDye Terminator Cycle Kit (Applied Biosystems, Foster City, CA). Each 10 µL reaction was composed of 4 µL of BigDye Terminator Matrix, 250-500 ng of purified plasmid DNA and 2 pmol of the T3 primer (kindly provided by Ms. Winnell Newman, Nucleic Acids Facility, NCSU, Raleigh, NC). Reactions were run on a GeneAmp 9700 thermocycler (Applied Biosystems) using the cycling parameters provided by the BigDye Terminator Cycle Kit. Samples were run on an ABI PRISM 370 DNA Analyzer (Applied Biosystems) by the Genome Research Lab at North Carolina State University.

Bioinformatics Analysis

Chromatograms were viewed in Chromas (Technelysium Pty Ltd., Southport, Queensland, Australia) and only those with high quality sequences (957 ESTs) were included in subsequent analyses. BLASTx searches were conducted with these sequences using the NCBI GenBank database (www.ncbi.nlm.nih.gov) to determine putative protein function for each EST. A cutoff expect value of 0.1 was used and results with a BLAST score higher than this number were not considered. Putative functions were identified based on function listed for the closest BLASTx match.

Elongation Factor 1 Alpha (EF1α) Gene Family

Nucleotide and amino acid alignments of elongation factor 1 alpha (EF1α) gene sequences from our database and previously published *T. ni* and *Drosophila* sequences were performed using SeqLab (Accelrys, Madison, WI). Pairwise percent identity between sequences (Table 1) was determined using the Gap function in SeqLab.

Results and Discussion

We have initiated a search for new targets for pesticide action. In the crop pests, the gut is an early point of contact for pesticides. As an organ, the gut is therefore an excellent target. However, we know very little about the gut system at the molecular level. Our lab has undertaken a random sequencing approach designed to identify some of the major genes that are transcribed into mRNA in the gut of feeding *Trichoplusia ni*. These are expected, by analogy, to be expressed in the gut system of other lepidopteran insects, including those known to be pests of cotton, like the cabbage looper (*Trichoplusia ni*), tobacco budworm (*Heliothis virescens*) and cotton earworm (*Helicoverpa zea*). Nearly one thousand cDNAs from the gut of late stage *T. ni* were partially sequenced as expressed sequence tags (ESTs). High quality sequences were analyzed further while sequences of low quality were removed from the database. High quality sequences were studied to identify their closest relative on the GenBank database. ESTs were then grouped according to proposed function.

High Abundance Putative Functions

BLASTx analysis was used to identify putative functions for each EST (Figure 1). Putative function was assigned only if the expect value of a match was better than 0.1 ($E \leq 0.1$). In cases where there were matches with $E \leq 0.1$ the top matches were compared. If within the best matches there was general agreement as to gene function, and the expect value was less than or equal to 0.1, that putative function was assigned to the EST in question. If there was not a general agreement, the closest match with $E \leq 0.1$ was used to assign putative function. In the case where expect values of $E \leq 0.1$ matched GenBank database accessions of unknown function, ESTs were assigned function "unknown". In the case where the expect value of the nearest match was 0.1 or greater, or where there was no match on the database, ESTs were declared "novel". Because the library used for random sequencing was not normalized, the relative abundance of ESTs of a particular function is representative of the relative abundance of its cognate message. ESTs were therefore grouped according to function in order to better understand the relative expression of these genes in the gut of feeding insects.

Figure 1 shows the results of this analysis. Only functions that are represented by one percent or more of the ESTs are shown. The largest groupings are the novel and unknown groups. However, these may be artificially large in that each is most likely made up of more than one putative function. (That is, the novel and unknown groups probably result from more than one gene or gene family. When these functions are separated, each group; e.g.: unknown 1, unknown 2, etc. will be smaller.) As predicted due to their abundance in all tissues, ribosomal proteins are a highly expressed group of ESTs. Because we generated the cDNA library from the digestive tissue of *T. ni*, we expected that the digestive enzymes lipase, trypsin and chymotrypsin would be abundantly expressed. This expectation was realized, as we see in the graph. The third grouping, serine proteases, will be split as appropriate to the chymotrypsin or trypsin group using bioinformatics tools. While of interest to the basic study of gene expression, highly expressed mRNAs and proteins like these are not as likely to make good targets for molecular based pesticides as genes that are expressed at low level. A notable exception may be the trypsins, which have been identified as a target for the hormone trypsin modulating oostatic factor (TMOF) in mosquito (Borovsky et al. 1990).

ESTs Have a High Level of Relatedness to Previously Sequenced DNAs from the Insecta

ESTs were tested to determine their relatedness to cDNA and genomic DNA sequences already on the GenBank database. The organism with BLASTx result E≤0.1 that was most closely related to *T. ni* was identified (Figure 2). Thus, if the closest relative also derived from the genus *Trichoplusia* the sequence was listed as matching *Trichoplusia*. If the relative was one of the Lepidoptera but not of the Noctuidae group, the sequence was grouped with the Lepidoptera. Groupings are therefore additive.

Of the high quality sequences obtained, only 2% had a high degree of similarity to sequences already obtained from *Trichoplusia*. This is most likely a reflection of the scant amount of sequencing work that has been done in this insect. Twenty one percent of the sequences were similar to a previously sequenced gene from the Noctuidae. Once Insecta and its three subgroups (Lepidoptera, Noctuidae, *Trichoplusia*) are added together, 50% of the newly sequenced ESTs have a closest match in the Insecta class. Expanding the grouping to the Arthropoda does not increase the sequence matches, most likely a reflection that the majority of the sequencing that has been done in the Arthropoda has been from the insects. Only 3.4% of the over 794,000 published Arthropoda sequences are non-insects (DMT, unpublished). The majority of the sequences had homologies to published sequences from organisms within the Animalia (a total of 57%). Interestingly, 14% of the high quality sequences had as their closest match an organism that was outside the Animal kingdom. The majority of these matches, although they had expect values that met the relaxed criterion of E<0.1, had values that were not very good, ranging from 0.1 to 0.0001. Several of these matches with low expect values were to ubiquitin protein from *Entodinium caudatum*, a bacterium. Other low stringency matches were to other bacteria or viruses. One interesting exception is an EST with high similarity to a protein of unknown function from the pea (*Pisum sativum*). The expect value of this match is quite good, E=9x10-33. Other surprising matches were to an endopeptidase from *Escherichia coli* (10-24), a proline-rich cell wall protein from *Caenorhabditis elegans* (10-7), large T antigen from Simian virus 40 (10-7) and a biotin biosynthetic gene from *E. coli* (10-78). Finally, a high percentage of the ESTs sequenced (29%) had no match on the GenBank database, even at the relatively low stringency of E<0.1.

Elongation Factor 1 Alpha is a Model for Data Mining in the New Database

In any large sequencing project, data mining becomes important. The next step for our *T. ni* sequences will be the creation of a Unigene database, where each gene is represented only once. After creation of the Unigene database, we will be able to identify the number of different alleles of each gene product and possibly the number of gene loci they represent. As a step towards that analysis, the ESTs that were assigned a putative function elongation factor 1α (EF1α) were analyzed. EF1α is a robust marker of molecular evolution and divergence in insects (Friedlander et al., 1992). In *Drosophila* (Hoveman et al., 1988) and the honeybee, *Apis mellifera* (Danforth and Ji, 1998) there are two genes encoding EF1α. Of the 957 *T. ni* ESTs sequenced five were assigned the putative function EF1α or EF1α-like. An additional two ESTs with the predicted function EF1α were sequenced previously by this laboratory (GenBank nos. BG354606 and BG354592) and were included in the current analysis. Preliminary alignments of *T. ni* EF1α and EF1α-like ESTs and the previously sequenced EF1α cDNAs from Drosophila (GenBank X06869, X06870 and U88868) and *T. ni* (GenBank U20140) were made using the Pileup command in the program SeqLab (data not shown). These alignments provided strong evidence for two alleles of EF1α-like genes in the cabbage looper. An evolutionary tree constructed using PAUP* (data not shown) confirmed the assignment of *T. ni* EF1α-like nucleotide sequences to two separate groups that represent at least two different genes. These genes were presumed similar to the EF1α F1 and F2 genes of *Drosophila* and honeybee. However, this would represent a departure from the published literature, which provides evidence for only a single EF1α gene in *T. ni* (Friedlander et al., 1992, Cho et al., 1995 and Mitchell et al., 1997).

Further bioinformatics analysis of the sequences began to introduce doubt as to the assignment of function. For example, two honeybee EF1α genes (F1 and F2) are 25.0% different from each other at the nucleotide level (Danforth and Ji, 1998). Drosophila EF1α F1 and F2 differ by 18.6%. However, comparison of the regions of overlap between the *T. ni* EF1α and EF1α-like genes found only 34.1-41.5% identity between the two forms (Table 1). If, as indicated by Danforth and Ji (1998) and Friedlander et al. (1992), the split between the two EF1α genes occurred in parallel, we would not anticipate such a low level of identity between EF1α genes in the same organism. A more careful review of the BLASTx results from the ESTs showed that three of the sequences were most closely related to an *Anopheles gambiae* clone of unknown function (Genbank acces-

sion number: AAAB01008984). The next closest match was to a *Drosophila* EF1α-like gene (Genbank accession number: U88868). This gene had a low level of identity to both EF1α F1 and F2 from *Drosophila* (49.2% and 49.9%, respectively, see Table 1). This low level of identity suggested that, although the putative function listed for U88868 was EF1α-like, it was a different gene from EF1α F1 and F2. Furthermore, U88868 is found on chromosome 2. EF1α F1 and F2 are found on chromosome 3 of the *Drosophila* genome. Originally sequenced by Basu et al. (1998), U88868 is closely related to *Drosophila* EF1α. However, it is more closely related to the human and yeast translational release factor, and is now designated RF3 (Lasko, 2000). Like *Drosophila* RF3, the *T. ni* EF1α-like ESTs are more similar to their homologue from other species (e.g.: 74.2-76.1% identical to *Drosophila* RF3, see Table 1) than to EF1α from *T. ni* (34.1-41.5%). Therefore, the *T. ni* ESTs originally assigned the putative function EF1α-like have been, upon further bioinformatics analysis, assigned the function RF3. (For an alignment of a portion of the EF1α and EF1α-like genes of *Drosophila* and *T. ni*, see Figure 3.) In fact, the *A. gambiae* EST (AAAB01008984) that most closely matched these three *T. ni* ESTs is also, most likely, a RF3 as well. Interestingly, mutation of the *Drosophila* sup35P disrupts male mitosis, leading to male sterility. Therefore, the *T. ni* RF3 could become a novel target for pesticide development.

Summary

For this work, 957 new ESTs were generated from a cDNA library to the gut system of feeding *T. ni* larvae. Taken together with the 33 ESTs previously sequenced from the same library and released on the GenBank database (accession numbers:BG354582-BG354614) we have a database of 990 total EST sequences from this stage. Digestive enzymes and structural enzymes are highly represented in this library. ESTs with homology to genes of unknown function and ESTs with no detectable homology to any sequence already on the GenBank database were found. Homologies were found to genes from organisms as close evolutionarily as the *Trichoplusia* genus, and as far as bacteria and viruses. The translational elongation factor, EF1α, was studied further. Bioinformatics analysis of the *T. ni* translational elongation factors helped us identify not only the previously published EF1α but also a translational release factor, RF3, not previously sequenced in *Trichoplusia*. We are in the process of reducing the 990 ESTs, which represent a redundant database, to a Unigene database which will have each allele represented only one time. This database will contain a wealth of information that may lead to the discovery of novel targets for pesticide action, in addition to being a valuable resource for questions of basic insect biology and genetics.

Acknowledgements

The authors wish to thank the staff of the NCSU Genome Research Laboratory for use of the facility and for technical assistance. In particular, we wish to thank Ms. Aimée Salstead for assistance above and beyond the call of duty. We thank Dr. Brian Wiegmann for helpful discussions on analysis of the EF1∝ genes. We wish to thank Dr. Long Long Yang for technical assistance. This work was funded in part by the NCSU Center for Integrated Pest Management in grant to DMT and by Cotton Incorporated in a grant to RMR.

Bibliography

Borovsky D., D.A. Carlson, P.R. Griffin, J. Shabanowitz and D.F. Hunt. 1990. Mosquito oostatic factor: a novel decapeptide modulating trypsin-like enzyme biosynthesis in the midgut. FASEB J 4:3015-3020.

Cho, S., A. Mitchell, J.C. Regier, C. Mitter, R.W. Poole, T.P. Friedlander and S. Zhao. 1995. A highly conserved nuclear gene for low-level phylogenetics: elongation factor 1-alpha recovers morphology-based tree for heliothine moths. Mol. Biol. Evol. 12:650-656.

Chomczynski, P. and N. Sacchi. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analyt. Biochem. 162: 156-159.

Friedlander, T.P., J.C. Regier and C. Mitter. 1992. Nuclear Gene Sequences for Higher Level Phylogenetic Analysis: 14 Promising Candidates. Syst. Biol. 41:483-490.

Hoveman, B., S. Richter, U. Walldorf and C. Cziepluch. 1988. Two genes encode related cytoplasmic elongation factors 1a (EF-1a) in *Drosophila melanogaster* with continuous and stage specific expression. Nucleic Acids Res. 16:3175-3194.

Johnson, J.A. 2001. Drug target pharmacogenomics: an overview. Am. J. Pharmacogenomics 1:271-81

Kallapur, V.L., C. Majumder and R.M. Roe. 1996. *In vivo* and *in vitro*-tissue specific metabolism of juvenile hormone during the last stadium of the cabbage looper, *Trichoplusia ni*. J. Insect Physiol. 42: 181-190.

Mitchell, A., S. Cho, J.C. Regier, C. Mitter, R.W. Poole and M. Mathews. 1997. Phylogenetic utility of elongation factor-1a in Noctuoidea (Insecta: Lepidoptera): the limits of synonymous substitution. Mol. Biol. Evol. 14:381-390.

Roe, R.M., A.M. Hammond, Jr. and T.C. Sparks. 1982. Growth of larval *Diatraea saccharalis* (Lepidoptera: Pyralidae) on artificial diet and synchronization of the last larval stadium. Ann. Entomol. Soc. Am. 75: 421-429.

Sambrook, J., E.F. Fritsch and T. Maniatis. 1989. Molecular Cloning: A Laboratory Manual, second edition. Cold Spring Harbor Press, Cold Spring Harbor, NY.

Wang, P. and R.R. Granados. 2001. Molecular structure of the peritrophic membrane (PM): identification of potential PM target sites for insect control. Arch. Insect Biochem. Physiol. 47:110-118.

Table 1. Pairwise Percent Identity of *Drosophila* and *T. ni* EF1α-like genes. *Tn* is used to designate *T. ni* and *Dm* is used for *D. melanogaster.* Sequences listed as "EF1α *Tn* EST" and "EF1α-like *Tn* EST" were obtained from our *T. ni* database. All other sequences were obtained from NCBI GenBank database (www.ncbi.nlm.nih.gov) with accession numbers listed in parentheses. Multiple EF1α and EF1α-like ESTs were sequenced from the cDNA library therefore we displayed the range for these values. The sample size for the EF1α ESTs was four and EF1α-like was three. EF1α *Tn* ESTs compared to each were 98.2% percent identical. EF1α-like *Tn* ESTs were 94.8-100% identical.

	EF1α F1 *Dm* (X06869)	**EF1α F2 *Dm* (X06870)**	**EF1α *Tn* (U20140)**	**EF1α-like *Dm* (U88868)**	**EF1α *Tn* EST**
EF1α F2 *Dm* (X06870)	81.1				
EF1α *Tn* (U20140)	85.6	79.8			
EF1α-like *Dm* (U88868)	49.2	49.9	48.0		
EF1α *Tn* EST (n=4)	73.3-84.9	40.1-73.4	89.3-99.8	39.8-52.8	
EF1α-like *Tn* EST (n=3)	40.7-41.7	35.7-39.9	39.3-40.0	74.2-76.1	34.1-41.5

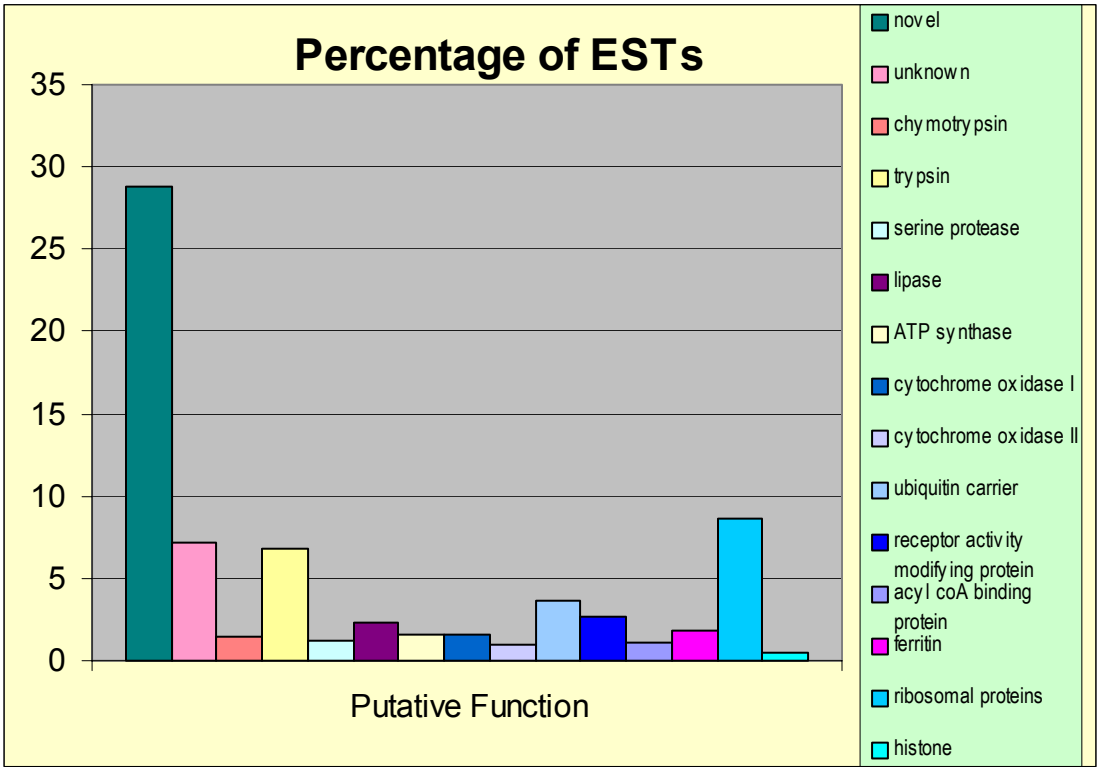

Figure 1. Percentage of ESTs matching putative functions listed. Only the categories consisting of one percent or more of the total ESTs were included in this figure. The category "Novel" contains sequences that did not match any sequences listed in the GenBank database and therefore no function can be assigned presently. Novel sequences will be explored further with additional searches to identify interesting motifs and possibly determine function. The category "Function Unknown" contains sequences that did have matches in the database, however no function was listed for the sequence in the database.

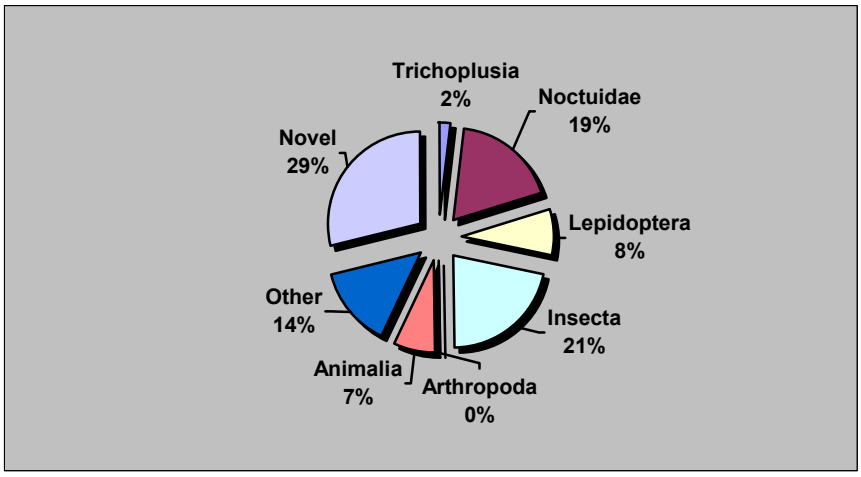

Figure 2. Percentage of ESTs that matched organisms in the GenBank database and the closest match to *Trichoplusia ni* (Lepidoptera: Noctuidae). "Novel" sequences did not match any published sequences in the database therefore the closest matching organism could not be identified. The category "Animalia" is all non-arthropod animals. The category "Insecta" is all insect orders excluding Lepidoptera. The category "Lepidoptera" is all Lepidopteran families excluding Noctuidae. The category "Noctuidae" is all noctuid genera excluding *Trichoplusia*.

```
     Tn EF1α-like EST    CGCTCAAGTT GTGATCCTTG AACACAAGTC CATTATCTGT GCGGGTTACT
Dm EF1α-like (U88868)    TGCTCAGGTC GTAATTTTAG AACACAAATC AATTATCTGT GCGGGTTACT
          Tn EF1α EST    AGGCAGAGTC GAAACTGGTA TCCTCAAGCC TGGTACCATC GTCGTCTTCG
     Tn EF1α (U20140)    AGGCAGAGTC GAAACTGGTA TCCTCAAGCC TGGTACCATC GTCGTCTTCG
   Dm EF1α F1 (X06869)   GGGTCGTGTG GAGACTGGTG TGCTGAAGCC CGGTACCGTT GTGGTCTTCG
   Dm EF1α F2 (X06870)   AGGTCGTGTG GAGACTGGTC TCCTCAAGCC AGGCATGGTC GTCAACTTTG
```

Figure 3. Portion of the nucleotide sequence alignment of elongation factor 1 alpha (EF1α) and EF1α-like fragments found in *Trichoplusia ni* and *Drosophila melanogaster*. *Tn* is used for *T. ni* and *Dm* is used for *D. melanogaster*. Sequences listed as "*Tn* EF1α EST" and "*Tn* EF1α-like EST" are representative samples from our *T. ni* cDNA database. All other sequences were obtained from the NCBI GenBank database (www.ncbi.nlm.nih.gov) with accession numbers listed in parentheses.