7-25-2012

Annotation of Cytochrome P450 Genes In "Harmonia axyridis' And a Comparative Study of CYP Genes in "Harmonia axyridis' and "Tribolium castaneum'

Supriya Priyadarshini Pati

ANNOTATION OF

CYTOCHROME P450 GENES IN *Harmonia axyridis*

AND

A COMPARATIVE STUDY OF CYP GENES IN *Harmonia axyridis*

AND

Tribolium castaneum

By

Supriya Priyadarshinini Pati

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Bioinformatics

The University of Memphis

[July 2012]

This Thesis is dedicated to my Father, Mr. Lakshmidhar Pati, and Mother the role model of my Life, Ms. Suprava Pati, for always being a source of inspiration and support in my life.

Acknowledgement

The completion of the daunting task of obtaining my Master of Science degree was only possible because of the continuous help, support and guidance of many people. First, I would like to express my gratitude and heartfelt appreciation to my mentor Dr. Duane D. McKenna and my committee members for their support and guidance. Words cannot express my gratitude and respect for all of my committee members. Dr. McKenna, my major advisor, has made this project possible, and he has provided his time and guidance throughout the study. I also want to thank Dr. David R. Nelson for his kindness, excellent guidance and patience while teaching me fundamental concepts. It was a blessing to work and learned from such a renowned scientist who is extremely dedicated to science.

My heart is filled with deep respect and honor for the benevolent help of Dr. Ramin for providing me the opportunity to pursue an M.S. in Bioinformatics and for guiding me in various fields of science, both in and out of the classroom. I was always encouraged by his passion for research and science. His support and encouragement motivated me to work hard. I really respect the Man of Honor in him. Lastly, I want to thank Dr. Sutter. In my short interaction with him, I was inspired by his dedication to science and research. His encouragement guided me in the pursuit of my goals. In addition, he made a huge contribution in reviewing and editing my thesis.

I also want to acknowledge my previous major advisor at the University of Memphis, Dr. I. Khan, for his initial help and guidance. I am grateful very grateful to all of my Bioinformatics professors and colleagues, Dr. Vasile Rus**,** Dr. Vinhthuy T**.** Phan,

Ms. Tallulah B Campbell, Dr. Melvin L. Beck in the Department of Biological Sciences, Ms. Donna Haskins and my friends Dr. MRK Subbarao (St. Jude), Dr. Sailaja Krishna (UT Health Science Center), Dr. Vinay Jain (UT Health Science Center), Ms. Jamie Brown, Tina Chou, Behrouz Madahian, Nam Vo, Jingnan Zhao, Fazle Faisal, Amin Ahsan Ali, Paul Kim (Young June Kim), Matthew Ryan Krull, Alex Aitken, Laasya Vadlamudi, Pragya Sharma, Bhargavi Manda and John K. Anderson.

Finally, I would like to thank my parents, siblings and husband; words are not enough to thank all of them for their support and patience during the completion of this thesis. I would also like to thank all of my teachers from kindergarten through graduate school for igniting in me the light of knowledge. Finally, I would like to thank the Almighty for everything that I have today.

Abstract

Pati, Supriya Priyadarshini. MS. The University of Memphis. 07/2012. Annotation of cytochrome P450 genes in *Harmonia axyridis* and a comparative study of CYP genes in *Harmonia axyridis* and *Tribolium castaneum.* Major Professor: Dr. Duane D. McKenna.

Our knowledge of beetle cytochrome P450 (CYP) genes was primarily obtained from studies of the model beetle and grain pest *Tribolium castaneum*. To gain additional insight into beetle CYPs and ultimately to inform our understanding of beetle CYP evolution, we identified and annotated all of the CYP genes present in a new draft genome of *Harmonia axyridis* by using traditional and automated methods for gene annotation. Overall, we identified somewhat fewer CYPs in *H. axyridis* (at least 94 genes and 3 pseudo genes representing 17 families and 42 subfamilies) compared to the number of known CYPs in *T. castaneum* (137 plus 2 slight variants and 10 pseudogenes). The *H. axyridis* CYPs could be divided into 4 distinct clans: Mito, CYP2, CYP3 and CYP4. The phylogeny of the CYPs from these species reveals that the Mito, CYP2, CYP3, and CYP4 clans are major (monophyletic) groups with strong support for most relationships and illustrates the presence of CYP "blooms" in *T. castaneum* that are lacking in *H. axyridis*. Several additional CYPs that are present in *H. axyridis* are missing in *T. castaneum*. The Mito clan of *H. axyridis* contains 6 genes in 5 families and 6 subfamilies. We found 7 genes in the CYP2 clan with 5 families and 6 sub-families. We found 2 distinct families (4 and 349) and a minimum of 22 genes in the CYP4 clan in *H. axyridis*. Interestingly, both *H. axyridis* and *T. castaneum* carry CYP4G genes, which are candidate resistance genes for insecticides, including permethrins. The function of CYP4G was associated with pesticide resistance. The CYP3 clan has 59 genes in 5

families in *H. axyridis*: CYP6, CYP9, CYP345, CYP435 and CYP436. These 5 families in CYP3 are classified into at least 21 subfamilies. Our work focused on the automated annotation of CYP genes involved several software programs, the most efficient and sensitive of which were Augustus, GenScan and Fgenesh. Although it is likely that a few CYP genes remain to be identified in the *H. axyridis* genome, our ongoing work suggests that the vast majority of CYPs have been identified.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CYP	Cytochrome P450
BLAST	Basic Local Alignment Search Tool
cDNA	Complimentary DNA
DDBJ	DNA Data Bank of Japan
EcRs	Ecdysone Receptors
ESTs	Expressed Sequence Tags
GA	Genome Analyzer
HMM	Hidden Markov Model
PCR	Polymerase Chain Reaction
PET	Paired end Tag
Next-gen	Next-generation
Spook, Spo	CYP307A1
Spookier	CYP307A2
Spookiest	CYP307B1
Phantom, Phm	CYP306A1
Disembodied, Dib	CYP302A1
Shadow, Sad	CYP315A1
Shade, Shd	CYP314A1
McKenna_xxxx	McKenna_1kb_R1.PF_(paired)_contig_xxxx
gDNA	genomic DNA
LGX	Linkage Group X

Introduction

The identification and annotation of cytochrome P450 (CYP) genes in the beetle *Harmonia axyridis* (Coleoptera: Coccinellidae), the Asian multicolored ladybug, using traditional methods and automated gene annotation tools were the primary objectives of this thesis. The CYPs identified were also compared and contrasted with those in the genome of *Tribolium castaneum* (Coleoptera: Tenebrionidae), the red flour beetle, the only annotated beetle genome available to date (*T. castaneum* Genome Sequencing Consortium 2008). An assembly of the *H. axyridis* genome, based on next-generation (next-gen) DNA sequence data, served as the basis for an annotation and comparative study. The process of gene annotation has undergone several notable modifications over the last 30 years. For example, with the development of new time-efficient and less memory-intensive algorithms, interest in automated gene annotation has grown. Traditional gene cDNA mapping is no longer the primary means employed to annotate genomes (Brent et al. 2005; Carvajal et al. 2010; Nielsen et al. 2010; Petty et al. 2010).

However, challenges associated with simulated annotation include the accuracy and sensitivity of the protocol followed and the sensitivity of the computational tools (software) used. Therefore, it is helpful to use as many different protocols as possible to compare results and thereby take advantage of the strengths of each approach while ameliorating their weaknesses. The method of gene annotation we used is semiautomated and has a higher confidence limit for achieving optimal accuracy than fully automated annotation protocols. Therefore, in the current genome project, we focused on obtaining

the most accurate annotation of CYP genes, and then we compared the output of traditional annotation with that from automated annotation tools (Singh et al. 2000).

The CYP genes of insects play numerous and important physiological roles. The CYPs that regulate an essential biochemical pathway related to development and metamorphosis are known as Halloween genes in insects. They function specifically at the cellular level in different developmental stages of insects such as embryonic development, larval development, pupal development and reproduction (Parish et al. 2002; Petryk et al. 2003; Inoue et al. 2004; Namiki et al. 2005; Parvy et al. 2005; Ono et al. 2006). CYPs encoded by the Halloween genes are members of the CYP 2 and Mito clans (Nelson et al. 2006). The biosynthesis of 20E (an important insect growth hormone) from cholesterol occurs through a series of oxidation reactions involving Halloween genes. CYP307A1 (*Spook*) and CYP307A2 (*Spookier*) are expressed at different developmental stages. However, they are alleged to mediate the same enzymatic reaction (Ono et al. 2006). CYP307A1 and CYP307A2 play vital roles in the formation of 7-dehydrocholesterol (7dC) by directing an uncharacterized series of oxidations known as the Black Box (Namiki et al. 2005; Ono et al. 2006). CYP307B1 (*Spookiest*) is a member of the same family. The Halloween genes in *Drosophila erecta*, *D. virilis* and *D. willistoni*, CYP307A1 (*Spook*), CYP307A2 (*Spookier*), CYP307B1 (*Spookiest*), CYP306A1 (*Phantom*), CYP302A1 (*Disembodied*), CYP315A1 (*Shadow*), and CYP314A1 (*Shade*), are observed to have consensus regions. In studies of the intron-exon boundaries of different insects, it was observed that specific conserved introns were located at the same position and in the same phase (CYP302A1 (*Disembodied*), CYP306A1 (*Phantom*), CYP314A1 (*Shade*) and CYP315A1 (*Shadow*) as single-copy

orthologs of different insect species (Savard et al. 2006). Ecdysone receptors (EcRs) comprise an interesting and important class of nuclear receptors because they serve as receptors for insect molting hormones. Studies have revealed that the nonsteroidal ecdysteroid synthetic agonists for EcR could disrupt molting and function as safe pesticides. Hence, they have potential applications in pest management and medicine (Fahrbach et al. 2012).

Automated Gene Annotation

Gene annotation involves processing the functional elements of a genetic region including exons and functional segments of introns (Brent et al. 2005). In the past, the gene annotation process was more experimental, but with increasing genomic research and advances in technology, the annotation process was gradually automated. After sequencing the human genome and the introduction of large-scale high-throughput sequencing techniques, most genomic research protocols were automated, including gene annotation. Detecting regions such as the pre-exonic region (e.g., transcription regulatory region, transcription start site) remains challenging for *de novo* genome annotation projects (Nielsen et al. 2010).

For gene annotation using next-gen DNA sequence data, the gene annotation process can be broadly classified into 2 categories: (i) Structural annotation of the gene, including exons, introns, untranslated regions and splice forms; and (ii) Functional annotation, detailing the involvement of specific genes in different biological, molecular and cellular processes and the detection of specific expression sites. Further, the

structural annotation processes can be broadly classified as (i) *ab initio* gene predictions, (ii) expressed sequence tag (EST) matching and (iii) protein homology.

The problems of gene identification and the prediction of functional gene regions in genomic DNA sequences by computational means have attracted considerable attention (Mathe et al. 2002). With access to reliable computational tools to locate functionally significant regions within numerous genes, automated characterization of genomic sequences is of the utmost importance for large-scale sequence annotation, e.g., as part of a genome project. Automated annotation focuses on specific features; for example, protein coding genomic sequences exhibit characteristic features that distinguish them from noncoding sequences. The uneven usage of both amino acids and synonymous codons in existing proteins imposes constraints on the interpretation of nucleotide coding sequences. The asymmetrical usage of oligonucleotides, redundancy and short-range correlations constrains gene-coding sequences. Global computational searches using content methods were performed to exploit the existence of such constraints and statistically measure the coding potential of DNA sequences (Fickett et al. 1992; Gelfand et al. 1995). It has also long been known that DNA encodes sequence signals that instruct the cellular machinery in the pathway, leading from DNA to amino acid sequences. Among others, promoter motifs regulate the transcription of the genomic DNA regions encoding individual proteins, and splice sites direct the removal of introns in the primary RNA transcript to produce the mature mRNA sequences. Start and stop codons delimit the portion of the mRNA sequence that will finally be translated into amino acid sequences. Different advanced computational techniques known globally as search-by-signal methods have been developed to define and locate the signals in the

DNA sequences involved in gene specification (Gelfand et al. 1995). Neither seeking by signal nor by content can elucidate the complex and variable genomic structure of the genes of higher eukaryotic organisms encoded primarily by a number (sometimes large) of small coding portions (exons) separated by larger intervening noncoding sequences (introns). With statistical coding measures, their value cannot always be unequivocally interpreted. Often, noncoding sequences exhibit features typical of coding sequences and *vice versa*. In addition, in specific cases in which a large fraction of higher eukaryotic coding exons are incomplete (in short fragment form), statistical measures will provide less meaningfully computed output. Additionally, the DNA sequence signals involved in gene prediction are ill defined in some groups of organisms, and they appears to be less accurately interpreted with unspecific models. Further, the current algorithms executed in gene detection software use heuristic algorithms to distinguish the signals present. In reality, this makes the output challenging to interpret. Gelfand (1990) and Fields and Soderlund (1990) pioneered computer programs designed to integrate search-by-signal and search-by-content methods. Such programs have the objectives of both gathering potential coding regions in anonymous DNA sequences and producing (in principle) full predictions of the (exonic) structure of the genes potentially encoded by the DNA sequences. After these programs were introduced, a number of programs that use these techniques were developed (Kel et al. 1995; Fickett et al. 1996). Recently developed programs employ information from sequence similarity database searches. These programs are routinely used in genome centers to study processed DNA sequences, and they have contributed in numerous cases to the discovery and characterization of new genes. However, their success in processing large, uncharacterized genomic sequences of

higher eukaryotic organisms is indefinite or moderate (Burset et al. 1996). Thus, the predictions of such programs are not usually certain unless they are supported by additional evidence derived from similarly known EST amino acid sequences or experimental results. In prokaryotic organisms and yeast, computational simulations have substantially contributed to annotation through automated genome analysis (Borodovsky et al. 1993). Automated sequence annotation or automated identification of protein coding genes and elucidation of their structure remained impractical and inconsistent for the genomes of high eukaryotic organisms. However, reliable computational tools or software could substantially reduce the cost required to interpret the vast amount of data produced by genome projects. The tools or software could also reduce the cost of the projects themselves. Considerable effort is being devoted to developing increasingly accurate computational techniques for the detection of genes and their structure. Consequently, gene structure prediction programs are evolving into complex and sophisticated heterogeneous systems for the automated annotation of genomic sequences, and these programs utilize sequence statistics, signal identification and similarity to known database sequences. Such systems could be useful for both the analysis of existing genomes and gene and genome engineering. Although emphasis in computational gene identification is directed toward specific applications, e.g. developing increasingly powerful tools for automated sequence analysis and annotation, computational gene identification also addresses a fundamental biological problem: the problem of elucidating the “computation” occurring in the cell that outputs protein sequences from inputted DNA sequences. From this perspective, rather than attempting to locate likely coding exons by discovering sequence composition biases or database homologues, the

problems in attempting to elucidate gene structure are addressed by understanding and replicating it "*in silico*." The rules governing how the DNA sequence signals are involved in the specification of the genes are recognized and processed during such a "cellular computation." In this thesis, the accuracy of a number of currently available gene structure prediction programs will be reviewed. Then, the DNA sequence signals involved in gene specification are demonstrated to carry substantial amounts of information, and it will be revealed that even simple methods of processing such information results in the prediction of gene structures that are comparable, to some extent, with those obtained using other, more sophisticated methods.

Metagenomic Studies of Gene Annotation

The process of gene annotation has undergone many changes with the development and growth of modern genomic research, starting in the 1970s with Sanger sequencing (Sanger et al. 1975, 1977) and Maxam-Gilbert dye terminator sequencing (Donis-Keller et al. 1977; Perler et al. 1980; Maxam et al. 1986). This was followed by capillary sequencing (Swerdlow et al. 1991) and most recently by high-throughput silicon chip sequencing. With each of these advancements, sequencing has become less expensive and more efficient (Miller et al. 2008). The accuracy of traditional dye terminator sequencing is typically the benchmark against which other (and often newer) approaches are measured. High-throughput sequencers tend to have tradeoffs between the time interval for sequencing and read length (accuracy and quality). With the advancement of high-throughput DNA sequencing, we are approaching a platinum era of sequencing in which sequencing a human genome for $1000 will soon be possible (Locali-Fabris et al. 2006). In

Table 1.1, we present the different genome sequencers and the basic principles on which they are based. For this project, we used data from an Illumina Hi-Seq 2000 sequencer.

Table 1.1. The development of next-gen high-throughput DNA sequencing techniques over time. The first column from the left shows the high-throughput sequencing techniques, the second column shows the year of their introduction, the third column shows the methodology and approach followed and the fourth column includes references to the relevant literature.

High-throughput Sequencing	Genome Era	Methodology	Reference(s)
Massively Parallel Signature Sequencing	1990	Adapter ligation followed by adapter decoding	Brenner et al. 2000
Polony sequencing	2005	Paired-end tag (PET) library, emulsion PCR, cost is $1/10^{th}$ that of Sanger sequencing, > 99.9999% accuracy.	Shendure et al. 2005
454 pyrosequencing	2005	Emulsion PCR primer-coated bead	Margulies et al. 2005; Schuster et al. 2008
Illumina (Solexa)	2008	Reversible dye terminators bridge amplification	Mardis et al. 2008
SOLiD	2008	Sequencing by ligation	Schuster et al. 2008
Helioscope	2008	True single-molecule sequencing	Harris et al. 2008
SMRT	2008	Zero-mode waveguides technology	Foque et al. 2008; Korlach et al. 2008
Nanopore	2009	Electrical signal-based alpha-hemolysin pore binding technology	Clarke et al. 2009
VisiGen	2009	Fluorescent dye-based (fluorescent resonant energy transfer technology)	http://www.bio-itworld.com/news/01/13/10/Visigen-founder-sues-Life-Technologies.html

Table 1.1. The development of next-gen high-throughput DNA sequencing techniques over time

High-throughput Sequencing	Genome Era	Methodology	Reference(s)
DNA nanoball sequencing	2010	Rolling circle replication Amplification of small fragments of genomic DNA into DNA nanoballs	Drmanac et al. 2010; Porreca et al. 2010
Ion semiconductor	2011	Detection of hydrogen ions	Rusk et al. 2011
Single-molecule real-time sequencing	2011	RNA polymerase, polystyrene bead technology	Pareek et al. 2011

The rapidly advancing technology of next-gen DNA sequencers facilitates the acquisition of large quantities of data. In addition, this technology enables us to sequence an entire genome in a more cost-effective and time-efficient manner (Ridley et al. 2008). This facilitates the scientific study of genomes. In the last 2–3 years, several high-throughput DNA sequencers have gained widespread use. These include The Roche 454, Illumina, Life Tech SOLiD and Pacific Biosciences (PacBio) sequencers. These DNA sequencers have their advantages and disadvantages (Kircher et al. 2010) with regard to their computational burden and support, economic aspects (e.g., cost/base) and experimental limitations. These factors influence the final choice of genome researchers (Metzker et al. 2010). Next-gen sequencers such as Illumina can be used to generate paired-end DNA sequences by exploiting PET sequencing and the reversible dye terminator technique (Fullwood et al. 2009).

The Illumina technology begins with adapter ligation followed by fixation of the genome sequence to the substrate, library amplification by local *in situ* PCR and

sequence amplification using fluorophore-labeled chain terminators (Bennett et al. 2005). Sequences obtained by Illumina sequencing are usually 180–200 nucleotides long, but technology improvements are expected to make longer read lengths possible and allow more data to be generated in a shorter timeframe. Advantages of the Illumina technique include the large amount of data produced (600 gigabase pairs total per run) with greater than 99% sequencing accuracy (http://www.illumina.com/truseq/quality_101/quality_scores.ilmn) in approximately 11 days (runtime is scaled in days). However, the read length can be increased by increasing the runtime. The sequencing practice is modified to have an optimal balance between read length and run time per cost with consideration of low-complexity regions and sequence repeats. There is an advantage in generating large amounts of data, but along with these data, there is also a drawback due to the requirements for storage space (Hsi-Yang et al. 2011).

Brief Overview of Next-gen Assemblers

The next-gene massively parallel sequencing platforms such as the Illumina Genome Analyzer (Bentley et al. 2006), Applied Biosystems SOLiD System and Helicos BioSciences HeliScope (Harris et al. 2008) initiated a new genomic era by generating high-quality data, i.e., >6 kilobase (kB) of data. However, they all have the drawback of producing short reads (shorter than the Sanger reads). The Illumina system generates short read lengths of 180–200 bp, which is substantially shorter than Sanger sequencer reads (500–1000 bp; Shendure et al. 2004). Moreover, their accuracy is suboptimal if the

quality of the genomic sequence is not high, which in turn has raised concerns about their accuracy in producing large assembled contigs of genomes. However, the generation of large quantities of data facilitates many lines of research, such as SNP detection and reference genome studies (Sachidanandam et al. 2001; Wang et al. 2011). Studies of insertions, deletions and mutations leading to structural variations of proteins and the *de novo* assembly of individual genomes benefit from the longer read lengths of different sequencers such as the 454 system. Sanger sequencing technology is still the preferred method for generating reference genomes because of its read length and accuracy. Sanger sequencing is not economical for large genome projects.

The development of *de novo* assembly methods for short reads facilitated the building of reference sequences for unexplored genomes in a more cost-effective manner than traditional Sanger sequencing, helping researchers to conduct different *de novo* analyses. Software such as ARACHNE (Batzoglou et al. 2002), Phrap (http://www.phrap.org), Atlas (Havlak et al. 2004), Celera assembler (Meyer et al. 1997), RePS (Wang et al. 2002), PCAP (Warren et al. 2006) and Phusion (Mullikin et al. 2003), are well-established *de novo* assemblers for whole-genome shotgun sequencing projects. The genome assemblies are based on an overlapping layout for all short reads. However, recording the sequence overlaps requires a substantial amount of computer memory.

The de Bruijn graph data structure was first implemented in the program EULER (Pevzner et al. 2001), a genome assembler that is based on assembly, by using a suitable overlap of short reads in a suitable data structure. In the de Bruijn graph data structure, the K-mers are considered vertices, and the read paths along the K-mers are considered the edges of the graph. The graph size provides the genome size and repeat content of the

sequenced sample. This graph works efficiently with high-redundancy and deep read coverage. Various short-read assemblers, including ALLPATHS (Butler et al. 2008), SOAP (Li et al. 2009), EULER-SR (Chaisson et al. 2008) and Velvet (Zerbino et al. 2008), also implement this algorithm. However, other short-read assemblers based on overlap and extension are VCAKE (Jeck et al. 2007), SSAKE (Warren et al. 2007), SHARCGS (Dohm et al. 2007) and Edena (Farrer et al. 2009), which can handle sequencing errors efficiently. However, these assemblers were designed to handle small genomes, and they are not suitable for the assembly of large genomes. A few programs have been designed for the *de novo* assembly of large genomes, such as ALLPATHS-LG and SOAP.

The Traditional Annotation Process

The traditional gene annotation process is based on the relatedness of specific genes to previously annotated genes.

Sequence Similarity-based Analysis

Pairwise sequence similarity is a measure of how closely related 2 proteins or DNA sequences are, defined in terms of the percent identity or similarity. This is accomplished using different pairwise sequence alignment tools such as the Basic Local Alignment Search Tool (BLAST). This software is used to determine the percent identity between amino acids or nucleotide residues. The alignment between closely related species is used to quantify evolutionary changes based on variation in exonic regions or

conserved regions in sequences. Some of the structure-based algorithms are used to detect the structure and function of proteins. The percent identity alone does not always indicate the best and most reliable match; however, other parameters, such as the expectation value (E-value) cutoff level and other substitution parameters, are also considered to score both matching residues and residue substitutions, insertions and deletions. The score for a particular substitution is empirically measured by considering substitution frequencies. The most frequently used scoring matrices for proteins are PAM (Point Accepted Mutation) (Mount et al. 2008) and BLOSUM (BLOcks Substitution Matrices) (Halperin et al. 2003). The similarity score of 2 sequences is then obtained by aligning the 2 sequences using the substitution matrices. Alignments using different scoring matrices may lead to different similarity scores. Based on the algorithm and the substitution matrices (in the case of protein sequences) or the gap penalty parameters (in the case of nucleotide sequences), the optimal alignment is detected as the alignment that leads to the maximum similarity score. Dynamic programming algorithms have been widely implemented to solve this problem, but there are many other heuristic algorithms preferred to enhance the speed of analysis. Optimal alignment search algorithms can be broadly categorized into 2 types. The first one is global alignment, in which the entire sequences are considered to obtain the global optima. An example is the Needleman-Wunsch algorithm (Needleman et al. 1970) that uses dynamic programming. The other type of alignment is known as local alignment. Local alignments compute the local optima by generating the score of subsequences of the 2 query sequences. However, the local optima are not guaranteed to be the global optima. An example of a dynamic programming algorithm implementing the aforementioned principle is the Smith-

Waterman algorithm (Smith et al. 1981). BLAST, the most widely used technique for calculating sequence similarity, implements global and local alignment. BLAST uses a heuristic algorithm to calculate the optimal local alignment (Altschul et al. 1990). The output of BLAST against a database of sequences returns the top hits of the query sequence, reporting for each the score, E-value and the local alignments themselves. The E-value provides a statistical measure of the significance of the alignment and score (S). The E-value gives the probability of hits having a score of S or more by chance. Low E-values imply biological significance, whereas high values imply the potential for false positives (Karlin et al. 1997). The simple BLAST algorithm can be further subdivided and refined to compare nucleotide query sequences to a protein database through BLASTx or compare protein queries to nucleotide databases through tBLASTn. In each of these cases, the nucleotide sequences are translated into 6 different reading frames during alignment. The usual nucleotide-to-nucleotide BLAST is BLASTn, and the typical protein-to-protein BLAST is BLASTp.

Gene Annotation by the BLAST Search Method

In the traditional gene annotation approach, the BLAST search is done against the publicly available nucleotide datasets in GenBank, DNA Data Bank of Japan (DDBJ), EBI-EMBL, Flybase or open access and annotated sequence databases. In some cases, the BLAST search is also done against the protein databases Swiss-Prot and Pfam. For specific CYP gene analysis, we used Dr. Nelson's publicly available CYP dataset (http://drnelson.uthsc.edu/CytochromeP450.html).

A gene annotator can manually infer the origins of a fragment by using the top hits of the BLAST query (even with a low E-value to gather different functional regions of a specific gene from the sequence). During the process of gene annotation by the traditional approach, the query genome is studied on the basis of the annotation of a reference genome (Nebert et al. 1989, 1991). During the process of annotation, care is taken to find all of the functional regions of a new gene. If it is not possible to curate the entire gene, then one can curate fragments of the gene, and the missing parts should be noted. Automated methods are very poor at *de novo* annotating incomplete gene sequences, such of those often found in next-gen sequence data. Additionally, some of the unclassified sequences may actually be similar to known genes in the database. These may be simply missed because of the presence of partial genes (incomplete sequences) or limitations of the alignment algorithm.

Considering that a metagenomics project can currently produce a half a million or more fragments and those projects will produce much more data as sequencing costs decrease, it may soon be impossible to annotate without automation. If the protein sequence information of closely associated taxonomic groups and gene families is available, then better results may be obtained.

The Hidden Markov Model (HMM) is a mathematical model that is widely implemented in biology to efficiently resolve many complex problems, including gene prediction and phylogeny reconstruction. It linearly assigns a sequence to different states from a finite set of states (that have biologically significant meaning). In the case of gene annotation, the finite states are mutually exclusive classes: ‘introns’ and ‘exons.’ Similarly, one could assign other biological sets of classes based on specific criteria. The

finite sets of states are connected by directional transitions. Each transition is associated with a former and latter state, and it has a transition probability associated with it (Howard et al. 1971). While reading each character of the sequence or walking along the state path in a Markov Model, the input sequences are labeled with state labels (Eddy et al. 1995; Sonnhammer et al. 1998).

The emission probability in HMMs facilitates the definition of states, which closely represent biologically significant classes (Bird et al. 1987). For example, in exon detection, the character inside an exon sequence has a higher emission probability than the character inside an intron (an exon compared to an intron). After determining the transition and emission probabilities, a probability is assigned to the given input sequence and chosen path. As there could be multiple state paths, this is not always an easy task. There are different biologically relevant rationalizations associated with HMM. These facilitate the selection of the correct path to annotate a specific region in a genome. When protein sequences are used instead of nucleotide sequences, the specific approach is known as a profile HMM model (Krogh et al. 1994). To use the protein sequences in annotation, one needs to translate all DNA fragments into their 6 possible frames of translation. HMMs also have relevance in position-specific score models in HMMER. HMMs have been found to be widely useful in gene finding, protein secondary structure prediction (Eddy et al. 1998) and genetic linkage mapping.

HMMs can be used together with multiple sequence alignment tools to achieve more precision in gene annotation. Multiple sequence alignment assists with selecting the optimal state path in a specific alignment. Compared to BLAST, multiple sequence alignment provides more position-specific information that is conserved across genes.

This information improves the performance, thus providing a more precise annotation. This can aid in detecting a distantly related member of a gene family and improving the results of database searches for homologous sequences (Gribskov et al. 1987).

For gathering genetic information from different related taxa, GenBank, EBI-EMBL, DDBJ and other publicly available datasets are the best sources. Similarly, for protein datasets, the information could be gathered from different publicly available protein resources, e.g., Swiss-Prot and Pfam. For specific research on CYP genes, we collected protein data from Dr. Nelson's publicly available website (http://drnelson.uthsc.edu/CytochromeP450.html).

Our Implementation

In our gene annotation process, we compared the annotation obtained using HMM with traditional gene annotation, with a focus on the accuracy of each methodology. Augustus, Fgenes, GenScan, GeneID, GeneMark, RepeatMasker, NCBI ORF finder, UCSC Blat and the specific algorithms used in designing the automated annotation are discussed in Table.1.2. We used a previously annotated large piece of the *T. castaneum* genome: the LGX chromosome (bp 33,080–236,581).

Table 1.2 Gene annotation software and the algorithms implemented in them. The first column from the left indicates the software, and the second column indicates the algorithms or approaches implemented.

Software	Algorithms
Augustus	Generalized HMM
Fgenes	HMM-based gene structure prediction
GenScan	Fourier Transformation algorithm + HMM
GeneID	A dynamic programming algorithm
Snap	Exact and inexact string matching and HMM
GeneMark	Specific inhomogeneous and homogeneous Markov chain models of noncoding DNA
RepeatMasker	Efficient implementation: Smith-Waterman-Gotoh algorithm: By Phil Green
NCBI ORF finder	Sequence alignment in 6 reading frames and own algorithm
UCSC Blat	BLAT

Classification of *Harmonia axyridis* and *Tribolium castaneum*

The insects are currently classified into 34 different orders. Eleven orders are classified under the supra-ordinal group Endopterygota (=Holometabola), including Coleoptera (beetles), Diptera (flies), Lepidoptera (butterflies and moths) and Hymenoptera (bees and wasps). This thesis focused on the evolution of CYPs in insects belonging to the order Coleoptera. *Harmonia axyridis* belongs to the cerylonid series of families in the superfamily Cucujoidea. Cucujoidea is believed to have originated approximately 202.9 ± 11.44 million years (Ma) ago (McKenna et al. 2011). *H. axyridis* has 16 chromosomes and an estimated genome size of 322 megabase pairs (Mb; Gregory et al. 2003).

Brief General Background on Insect Evolution and *Harmonia axyridis*

Insects are an important component of ecosystems, serving as pollinators and decomposers among other functions (Grimaldi et al. 2005). Coevolution between herbivorous insects and their host plants, including codiversification, resource tracking and sequential evolution, has been proposed to play an important role in insect diversification (Farrell et al. 1993; McKenna et al. 2006, 2009). For this reason, insects are considered ideal models for ecological and taxonomic diversification. Insects have extended mutual interactions with their hosts (plants, animals and other insects). In addition, insects are expected to exhibit specializations for their habits. Numerous studies have focused on the diversification of insects in relation to the diversification of genes associated with host association/host usage in different insect species (Janz et al. 2006). These studies have revealed that host association is not a spontaneous event; rather, it is the result of the accumulation of specializations in numerous genes over millions of years (McKenna et al. 2011). *Harmonia axyridis* and *Tribolium castaneum* belong to different superfamilies of beetles. *H. axyridis* belongs to the super family Cucujoidea, whereas *T. castaneum* belongs to the super family Tenebrionoidea. They last shared a common ancestor approximately 225 Ma ago (McKenna et al. 2009).

Interspecies interaction between insects is a complicated event. However, when the interaction is allied with humans or resources exploited by humans (e.g., timber, crop plants), the insect species are considered economically important. Originally native to Asia, *H. axyridis* was introduced into North America several times as a biocontrol agent

for aphids beginning in 1916 (Krafsur et al. 1994) in addition to its later introduction into Europe in 1982 (Ongagna et al. 1993) and 1986 in South America (Poutsma et al. 2008).

Even with multiple introductions of ladybugs from different source populations, the species was not established in the United States for decades. However, they recently and suddenly became invasive on several continents. Invasive populations were first observed in western Oregon, USA, and eastern Louisiana, USA in 1988 and 1991, respectively (Chapin et al. 1991; LaMana et al. 1996). They were then recorded in Europe (Belgium et al. 2003) and South America (Argentina et al. 2004) in 2001 and Africa (South Africa; Stals et al. 2007) in 2004. The species have been observed to form large populations in these areas, where it has become a predator of many other insects and noninsect arthropods, a household invader and a pest of fruit cultivation (Koch et al. 2003).

Harmonia axyridis produces secondary metabolites, particularly alkaloids, in response to perceived threats. It emits toxins from its tibiofemoral joints, a phenomenon known as reflex bleeding. This is a well-characterized defense mechanism in the beetle family Coccinellidae. The secondary toxins Har a1 and Har a2 are known to cause allergic reactions in people with secondary immune responses (Nakazawa et al. 2007). *Harmonia axyridis* has a toxic diet; it feeds on toxic insects such as aphids and other ladybugs. Hence, it is expected to have a highly evolved detoxification system of CYPs, which may allow *H. axyridis* to tolerate its own toxins. The different secondary metabolites found in *H. axyridis* are listed in Table 2.

Table 2. Secondary metabolites of *Harmonia axyridis*. The table presents the secondary metabolites found in a survey of the literature.

Secondary metabolites of *H. axyridis*
1) 2-alkyl-3-methoxypyrazines
2) hydropyrido [2, 1, 6-de] quinolizine
3) 5-arly-4-hydroxy-3-(2H)-isothiazolone-1, 1-dioxide derivatives
4) coccinelline
5) 2-methoxy-3-alkylpyrazines
6) 3-alkyl-2-methoxypyrazines
7) adaline
8) precoccineline
9) C2-symmetric 2, 6-diallypiperidine carboxylic acid methyl ester
10) 1, 3-dioxol-2-one
11) 3-isobutyl-2-methoxypyrazine
12) *N*-cyanomethyloxazolidines
13) porantherine
14) azabicyclo[3.3.1] alkenes
15) poranthericine

For gene identification and evolutionary diversification studies of *H. axyridis*, the most preferred genome is the *T. castaneum* genome. *T. castaneum* is a well-established stored grain pest, and it is a source of interest to many researchers because of its extraordinary pesticide resistance. The *T. castaneum* genome project was completed in 2008 by researchers based at the Baylor College of Medicine (Richards et al. 2008).

Introduction to CYP Enzymes

CYP enzymes have broad-spectrum gene regulatory activity. These genes are mostly activated or induced by a chemical signal. There are 2 different types of electron

transfer chains for CYPs (Guengerich et al. 2008). These are again classified depending on their location in the cell. Some CYPs are present in the mitochondrial inner membrane, whereas others are present in the endoplasmic reticulum (ER). Both classes of CYPs are classified as membrane-bound proteins. In the ER, the protein that donates electrons to the CYP is NADPH CYP reductase (Omura et al. 1980). It is also bound to the membrane by an N-terminal tail that crosses the ER membrane. The bulk domain of this protein is on the cytosolic side of the ER membrane. The CYP reductase protein has 2 domains, each containing a flavin. In the catalytic process, the 2 electrons are accepted subsequently from NADPH, after which they migrate from FAD to FMN and then to the CYP heme iron via CYP reductase (Sligar et al. 1979).

For CYPs to function, they need a source of electrons. The heme ring transfers 2 electrons to oxygen, which undergoes reduction, eventually breaking the fairly stable oxygen-oxygen bond. The electrons are donated by the reductase protein that binds to the CYP and passes the electrons from 2 flavin prosthetic groups in the molecule. This exchange of electrons between proteins is also termed an electron transfer chain, and it is similar to the electron transfer in complexes I to IV found in mitochondria (Hannemann et al. 2007).

Brief Overview of Insect CYPs

CYPs have a signature heme group and display a very strong light absorption band at 450 nm. In spectrometric studies, when microsomes are treated with dithionite (reduces microsomes) and carbon monoxide gas in a single cuvette, a very strong

absorption band observed at 450 nm, hence the name "P450" (P is for pigment). This is also referred as a reduced CO difference spectrum. As CO binds tightly to the ferrous heme iron, it is responsible for the difference in absorbance. This spectrum was first observed in 1958 (Mason et al. 1957; Garfinkel et al. 1958; Klingenberg et al. 1958; Sato et al. 1964). CYPs display absorption in this range due to ligation to the heme iron. Four electron pairs are provided by nitrogens on the heme ring, and the fifth ligand is a thiolate anion from the conserved cysteine (Prosite: PDOC00081 Cytochrome P450 cysteine heme-iron ligand signature (November et al. 1997) (http://www.faqs.org/patents/app/20090023173#ixzz1v8Z6NzlK).

CYP enzymes are observed to be present in different forms of life, i.e., in plants, animals, fungi, bacteria, Archaea, protists and even viruses (Danielson et al. 2002; Roland et al. 2007). More than 18,000 distinct CYP proteins are known (http://drnelson.uthsc.edu/CytochromeP450.html). The CYP proteins are classified into families and subfamilies according to their sequence similarity. Protein sequences with greater than 40% identity at the amino acid level are grouped into the same family. Sequences that are greater than 55% identical are placed into the same subfamily. Sequences with greater than 95% similarity can be annotated as alleles (http://drnelson.uthsc.edu/CytochromeP450.html; Lindberg et al. 1989; Nebert et al. 1989; Nelson et al. 1993).

Evolution of CYPs

The existence of the CYP superfamily of enzymes in widely divergent groups of organisms ranging from microbes to plants and animals validates the contention that these groups of enzyme systems are very ancient and all of the current CYPs are derived from a single ancestor (Nebert et al. 1987; Nelson et al. 1987). Gene duplication and adaptive diversification play vital roles in divergence during reproduction (Gotoh et al. 1993). The rapid increase in the number of new CYPs during the past 400 Ma may be due to the increased interaction between animals and plants (Nebert et al. 1988, 1989; Gonzalez et al. 1990). Molecular drive or DNA turnover also plays a vital role in the diversification of CYPs (Dover et al. 1986; Gonzalez et al. 1990; Jeffry et al. 2001). Molecular drive also includes gene transposition, slippage replication, unequal crossover, gene duplication and RNA-mediated transfer (Gonzalez et al. 1990; Dover et al. 1986; Fogleman et al. 1997). Over time, molecular evolution occasionally results in the production of pseudogenes (Nebert et al. 1988). Pseudogenes are detected by their aberrant coding region, transcriptional silence or both (Dwyer et al. 2011).

Insect CYPs Elaborated

The first sequenced insect genome was that of the fruit fly *Drosophila melanogaster*, which carries 86 CYP genes and 4 pseudogenes (Tijet et al. 2001; http://P450.antibes.inra.fr). The number of fully sequenced insect genomes is growing, and in addition to *D. melanogaster* (180-Mb genome), a dozen closely related species of *Drosophila* (which all diverged within 40 Ma) have been sequenced. Moreover, many

insect disease vector genomes are available, such as those of the mosquitoes *Anopheles gambiae* (220 Mb), the vector of malaria, *Aedes aegypti* (800 Mb), the vector of yellow fever and dengue, and *Culex pipiens* (540 Mb), the vector of West Nile virus. All of these species belong to the order Diptera. Additionally, the genomic sequences for economically important insects such as the honeybee *Apis mellifera* (200 Mb) (order Hymenoptera), the silkworm *Bombyx mori* (530 Mb) (order Lepidoptera) and the red flour beetle *T. castaneum* (200 Mb) (order Coleoptera) are available. Other recently completed insect genome projects include *Bicyclus anynana* (490 Mb), a tropical butterfly, and *Nasonia vitripennis* (330 Mb), a parasitoid wasp. The genomes of the following insects belonging to the supra-ordinal group Holometabola (=Endopterygota) have been sequenced: the body louse *Pediculus humanus*, an exopterygote, order Phthiraptera, the cotton bollworm *Helicoverpa armigera* (400 Mb), the tsetse fly *Glossina morsitans* (200 Mb), the Asian tiger mosquito *Aedes albopictus*, the Eastern tree hole mosquito *Ochlerotatus* (*Aedes*) *triseriatus* and the medfly *Ceratitis capitata*. These insects undergo complete metamorphosis, and they have well-characterized larval, pupal and adult stages. The fossil record indicates that holometabolous insects appeared 300 Ma ago. In addition to the holometabolous insect genomes, the sequences of a few hemimetabolous insects are also available, such as the hemipterans, including the pea aphid *Acyrthosiphon pisum* (300 Mb) and the blood-sucking bug *Rhodnius prolixus* (670 Mb), the vector of *Trypanosoma cruzi*. A few noninsect arthropod genomes have been studied, including those of the crustacean *Daphnia pulex*, a water flea, and 2 chelicerate arthropods, the tick *Ixodes scapularis* and the spider mite *Tetranychus urticae*.

Because of the genome projects of the aforementioned insect and noninsect arthropods, there is a tremendous annotation challenge for the ancient superfamily of CYP genes because it is one of the most abundant families of genes found in eukaryotic genomes. Annotation of the fruit fly (Tijet et al. 2001), mosquito (Ranson et al. 2002) and honeybee (Claudianos et al. 2006) CYP sequences each brought specific problems. These included (i) nomenclature, i.e., lumping new sequences into existing CYP families or splitting them into new CYP families; (ii) sequence and assembly quality, i.e., how to assemble genes that are apparently spread onto 2 or more sequence contigs; e. g., will a pseudogene in release 1.0 become a 'real' gene in release 2.0; and (iii) homozygosity of the genomic DNA that was sequenced (in *Anopheles gambiae*, a large CYP cluster was present on contigs that originated from the DNA of 2 different cytotypes). Annotation work is therefore a continuous effort, and new experimental data, particularly those for the PCR amplification and resequencing of problematic CYPs, are often needed.

Four Clans of Insect CYP Genes

Analysis of the available sequences indicates that insect CYP genes fall into 4 major clans. These clades correspond to branches above the family level described as subclasses by Gotoh et al. (1993) and as clans by Nelson et al. (1998). Each one of these 4 clans includes some CYP families from vertebrate species, and for clarity, these were previously named (Claudianos et al. 2006) the CYP2, CYP3, CYP4 and Mito clans. Implicit in this designation is the fact that these 4 groups of genes were represented by at least 1 member in the last common ancestor of vertebrates and insects. Insects do not produce sterols and have lost the CYP51 gene. Interestingly, insects appear to lack CYPs

related to CYP26, CYP7 and CYP8, and these families have probably diverged from CYP51 in more ancient animals, such as cnidarians or *Trichoplax* species, but were lost in insects.

Moreover, indicated are the CYP genes for which presumed orthologs are found in almost all insect species. In the mosquito *A. gambiae*, only 10 CYP orthologs were found compared with the predicted number of ~40 (Ranson et al. 2002). The 46honeybee *A. mellifera* CYP genes (Claudianos et al. 2006) also include 10 genes with orthologs in *Diptera*. However, there are more CYPs with known physiological functions, and this indicates that specialized physiological functions in different insect orders recruit new CYPs after gene duplication events. The evolution of new functions following gene duplication is therefore not restricted to 'detoxification' or 'environmental response.' In the *Drosophila* lineage, a CYP gene duplication event was estimated to occur on average every 5 Ma (Feyereisen et al. 2005). The sequences of 12 closely related *Drosophila* species will provide a fertile ground for the study of CYP gene duplication and gene loss as well as intron loss and gain (Hahn et al. 2007). A brief list of the different CYPs and their functional annotation is provided in Table 3.

Table 3. Insect CYPs and their functions. The CYP column presents the different CYPs, and the infer column provides the functional annotation according to different case studies presented in the Reference column.

CYP	Infer	Reference
CYP345	CYP345 group of enzymes exhibits insecticide tolerance properties	Jiang et al. 2008
CYP9	CYP9 group of enzymes exhibits xenobiotic tolerance that is proportional to the duration of exposure to and concentration of xenobiotics	Stevens et al. 2000
CYP6F/CYP6	Insecticide resistance and the metabolism of exogenous compounds	Zhang et al. 2011b
CYP6BS/CYP6	Allelochemical-metabolizing	Zhang et al. 2011a
CYP4AA1	Xenobiotic and chemosensory responses: 20-hydroxyecdysone (20-HE) biosynthesis, pheromone metabolism and pyrethroid insecticide resistance	Oakeshott et al. 2010
CYP4Q	Xenobiotic detoxification	http://pubs.aic.ca/doi/pdfplus/10.4141/CJPS07001
CYP4G	Xenobiotic activity differs for different chemicals	Pedro et al. 2012
CYP49A1	Anticipated to participate in the steroid-mediated regulation of the insect early larval stage	http://flybase.org/reports/FBgn0033524.html
CYP314A1 /*Shade*	Hydroxylation of ecdysone to the corresponding steroid	Srinivasan et al. 2003
CYP307B1	Involved in the initial conversion of 7-dehydrochoresterol to ketodiol	http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2583.2009.00957.x/pdf
CYP307A1 /Spook	Regulator of ecdysone synthesis in insects	Namiki et al. 2005
CYP306A1	Ecdysteroid biosynthesis in the prothoracic glands	Niwa et al. 2004
CYP303A1	Required for the structure and function of *D. melanogaster* sensory organs	http://agris.fao.org/openagris/search.do?recordID=CN2010001174

Table 3. Insect CYPs and their functions.

CYP	Infer	Reference
CYP305B1V1	Metabolism of xenobiotics and drugs	http://agris.fao.org/openagris/search.do?recordID=CN2010001174
CYP301A1	Involved in cuticle formation	Willingham et al. 2004
CYP18A1	Steroid hormone inactivation is essential for metamorphosis.	Guittard et al. 2011
CYP15A1	CYP15A1 (methyl farnesoate epoxidase) has a specific role during embryogenesis	Dwyer et al. 2011b

Materials and Methods (Protocols)

Gene Identification Protocol

Despite the rapid development in gene prediction techniques, the present understanding of the gene repertoire is still incomplete (Brent et al. 2004; Eyras et al. 2005). Even the definition of a gene remains blurry in some situations (in terms of specific regions that define a gene; Snyder et al. 2003). Nonetheless, well-conserved genes are now relatively easy to annotate in different species. To study differences in the numbers and types of CYP genes present in the genomes of the beetles *H. axyridis* and *T. castaneum*, we sought to identify and annotate the CYPs present in the genome of *H. axyridis*. The CYPs of *T. castaneum* have been identified and annotated (Richards et al. 2008). The analytical workflow for traditional gene annotation and comparative classification of CYP genes in *Harmonia axyridis* considering *Tribolium castaneum* as reference is provided in Fig.1.1.

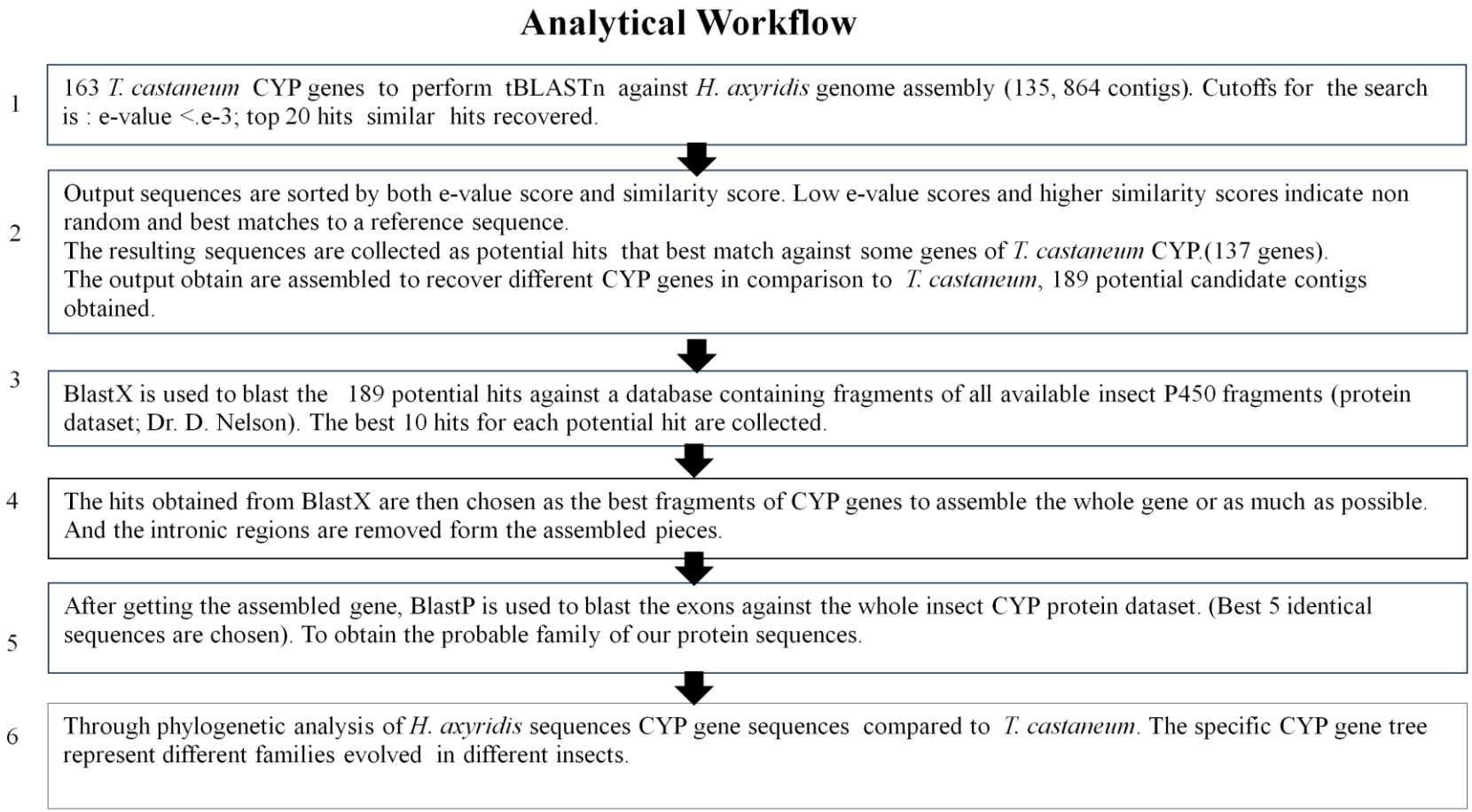

Fig. 1.1. Overview of traditional gene annotation and comparative classification of CYP genes in *Harmonia axyridis* vs. *Tribolium castaneum.*

The Traditional Method of Gene Annotation

Harmonia axyridis Genome Sequences and de novo Genome Assembly

Genomic DNA extraction, DNA sequencing and genome assembly were performed by Dr. Duane McKenna using the following methods. Genomic DNA from one individual of *H. axyridis* was sheared to produce 1 library of 1 kb genomic DNA fragments and another library of 500 bp fragments. These were sequenced on separate lanes (100 bp paired-end sequencing) on an Illumina Hi-Seq 2000 DNA sequencer (San Diego, CA, USA). A total of 592,992,560 paired-reads of 100 bp were produced. The paired Illumina reads were assembled into contigs using the CLCbio Genomics Workbench version 4.9 (http://www.clcbio.com/) with conflict resolution by voting (A, C, G, T). Nonspecific matches were given a random value. The resulting assembly contained 135,864 contigs larger than 1 kb with an average length of 2,219 bp. The contig N50 was 2,457 bp and average coverage of the assembled contigs was 142.8×.

In the Process of Traditional Method of Gene Annotation

The *H. axyridis* assembly of 135,864 contigs was searched by BLAST using standalone BLAST software (http://www.ncbi.nlm.nih.gov/books/NBK1763/) against the 163 CYP sequences of *T. castaneum* representing different genes. The standalone BLAST software is multiplatform compatible. It was locally installed on the McKenna

lab Linux server (aka Tesla1). Tesla1 is a Supermicro server with 48 core AMD Opteron processors and 192 GB of RAM (24×8 GB).

Initially, the assembly was subjected to BLAST against the CYP sequences of most closely related well-annotated *Tribolium* CYP amino acid sequences. There were 163 *Tribolium* CYP sequences that were subjected to BLAST against the *H. axyridis* assembly (135,864 contigs) by tBLASTn batch BLAST. In this manner, we obtained the BLAST output of potential contigs or hit sequences (approximately 198) based on their lower E-value and higher rank of identity. A low E-value indicates that the hit is not accidental, and a higher identity signifies sequence similarity. Therefore, the low E-value and high identity score contigs provide us with the most statistically relevant output.

The E-value of an alignment depends on 3 criteria: the alignment itself, the query sequence length and sequence composition and the total length of database sequences and their composition. The first step is computation of scoring of the alignment to generate the raw score S. This score is based on a nonspecific scoring system and must be normalized (Karlin et al. 1990; Dembo et al. 1994) to give a score S′ as follows:

$$S' = (\lambda S - \ln K)/\ln 2,$$

where λ and K are parameters that illustrate the expected distribution of S for the scoring system used. The normalized score S′ (has units in bits) provides the calculation of actual probabilities. The E-value for the alignment is as follows:

$$E = m\, n\, 2^{S'},$$

where m and n are the lengths of the database and query sequences, respectively, and S′ is the normalized score described previously in this section. (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). The potential contigs are

extracted from the assembly and subjected to BLAST against the protein dataset of *T. castaneum* CYPs. However, in the aforementioned case, the selection of *T. castaneum* sequences influenced results toward the existence of *T. castaneum* CYPs. In addition, the assembled CYP exonic fragments did not completely recover entire CYP sequences. Hence, to allow a greater possibility of fragment recovery and better assembly of the fragments, they were subjected to BLASTx against a mixture of different groups of well-annotated insect CYP sequences (Dr. Nelson's CYP database). This permitted the retrieval of all possible CYP fragments of *H. axyridis*.

The 189 hit sequences of *H. axyridis* were again subjected to BLASTx and ranked on the basis of their similarity to all possible CYP fragments in Dr. Nelson's CYP database for insects (the insect data set contains ~3000 sequence CYP genes fragments; http://drnelson.uthsc.edu/CytochromeP450.html). Based on the similarity score, the sequences were subjected to a second BLAST search. However, in the second BLAST, the E-value cutoff was relaxed (1.00 because the insect CYP consensus regions could have been present in smaller fragments of sequences, providing them with lower E-value scores). The BLASTx similarity search provided the best possible CYP fragments of *H. axyridis* CYPs, which were manually assembled with each other after removing the intronic regions.

On final assembly of the exons, all possible CYP fragments were gathered together. However, we were not able to recover sufficient numbers of fragments to assemble entire CYP sequences (the assembled sequences of several CYPs were not complete). There are various possibilities explaining the absence of missing fragments of CYPs, such as areas of low coverage in the *H. axyridis* assembly or low similarity to

database fragments. Finally, the assembled CYPs sequences were subjected to BLASTp to obtain the best-fit CYP family among insects. After obtaining the best insect CYP match, the sequences were annotated to their respective family, subfamily or allele.

Via automated gene annotation, we were able to recover 94 CYP genes, and by repeating the BLAST search (BLASTx) using the final *H. axyridis* contigs against the *H. axyridis* genome assembly, we recovered from some missing exonic fragments. Ultimately, through traditional gene annotation (assembling and BLASTp), we were able to recover and identify 94 *H. axyridis* CYP genes and 3 pseudogenes (from CYP301A1, CYP9BD1P and CYP9BE7P), in contrast to *T. castaneum*, which has 137 genes and 10 pseudogenes.

Nucleotide Sequence BLAST to Recover the Exons of the CYP Genes of Harmonia axyridis

We began to hunt for genes among 135,864 assembled contiges of *H. axyridis* using the well-annotated CYP protein sequences of *T. castaneum*. We performed a BLAST search for the CYP genes of *T. castaneum* in the form of a protein query against a database of *H. axyridis* contig sequences via a tBLASTn search. The output obtained from this search is given in Fig 1.2.

```
Query= CYP351A3_seq_98_CYP4_clan

Length=488
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  McKenna_1kb_R1.PF_(paired)_contig_69083 Average coverage: 96.49      119    5e-26
  McKenna_1kb_R1.PF_(paired)_contig_73025 Average coverage: 700.30     112    5e-24
  McKenna_1kb_R1.PF_(paired)_contig_56469 Average coverage: 93.07      111    9e-24
  McKenna_1kb_R1.PF_(paired)_contig_97615 Average coverage: 122.47     110    2e-23
  McKenna_1kb_R1.PF_(paired)_contig_65850 Average coverage: 206.28     107    1e-22
  McKenna_1kb_R1.PF_(paired)_contig_93511 Average coverage: 128.07     104    1e-21
  McKenna_1kb_R1.PF_(paired)_contig_41406 Average coverage: 134.59     102    6e-21
  McKenna_1kb_R1.PF_(paired)_contig_73302 Average coverage: 120.40     100    1e-20
  McKenna_1kb_R1.PF_(paired)_contig_48771 Average coverage: 219.24    99.4    4e-20
  McKenna_1kb_R1.PF_(paired)_contig_110020 Average coverage: 86.57    99.0    5e-20
  McKenna_1kb_R1.PF_(paired)_contig_68213 Average coverage: 79.54     88.2    1e-16
  McKenna_1kb_R1.PF_(paired)_contig_38109 Average coverage: 119.87    85.9    5e-16
  McKenna_1kb_R1.PF_(paired)_contig_44905 Average coverage: 120.15    85.5    7e-16
  McKenna_1kb_R1.PF_(paired)_contig_97615 Average coverage: 122.47    84.0    2e-15
  McKenna_1kb_R1.PF_(paired)_contig_34157 Average coverage: 101.84    83.2    3e-15
  McKenna_1kb_R1.PF_(paired)_contig_100427 Average coverage: 152.34   83.2    4e-15
  McKenna_1kb_R1.PF_(paired)_contig_87030 Average coverage: 62.26     82.8    4e-15
  McKenna_1kb_R1.PF_(paired)_contig_9307 Average coverage: 87.47      82.0    6e-15
  McKenna_1kb_R1.PF_(paired)_contig_30031 Average coverage: 99.18     82.0    7e-15

> McKenna_1kb_R1.PF_(paired)_contig_69083 Average coverage: 96.49
Length=1285

 Score =  119 bits (297),  Expect = 5e-26, Method: Compositional matrix adjust.
 Identities = 56/119 (48%), Positives = 81/119 (69%), Gaps = 1/119 (0%)
 Frame = +3

Query  318   QKKIGKELDVIFGKDDRVPTLEDINRMEYLERVIKETLRFLTPVPFMLRTNNQDITLDSN  377
             Q++I +EL+ I   ++R PT ED+ +M+ LER IKE+LR    V  + R  ++D TL S
Sbjct  726   QEQIVEELNSILEGEERQPTYEDLQKMDLLERCIKESLRLYPSVHLISREADEDTTLHSG  905

Query  378   TIPA-GSCIMIPIFHIHKKPEYWKNPNEFDPDRFLPENSSKRPRCAFIPFSSGPRNCIG  435
              + A G+ ++IPI  +H+ PE + +P +FDPDRFLPEN   R   A++PFS+GPRNCIG
Sbjct  906   CVVAKGATVLIPIMSVHRNPEIYPHPEKFDPDRFLPENCIGRHPFAYLPFSAGPRNCIG  1082
```

Fig. 1.2. tBLASTn output of *Tribolium castaneum* and *Harmonia axyridis:* This figure presents an example of the tBLASTn output of *T. castaneum* CYP gene sequences against the assembled contigs of *H. axyridis*. This example shows the tBLASTn output from the CYP351A3 gene and its first hit (identity value = 48%; E-value = e−26); both of these parameters indicate that this is a good match. The 'Query' line shows the gene as a protein sequence, and the subject line shows the best match from the database of *T. castaneum* sequences.

The tBLASTn contigs, or output contigs with low E-values and high similarity scores, are considered the best candidate contigs for further study in the CYP gene recovery process. In the tBLASTn process, the nuclotide sequences are translated into 6

different reading frames before being compared to the protein sequences. The cutoff value of tBLASTn in the first BLAST search was kept at an E-value of e−3. At this cutoff, we retrieved 189 candidate contigs. However, in the second retrieval of genes from the assembled contigs of *H. axyridis*, the cuttoff E-value was set to 1, resulting in the addition of 46 new candidate contigs. The resulting candidate contigs were subjected to BLAST using BLASTx against all available insect CYP gene fragments from Dr. Nelson's publicly available CYP dataset (http://drnelson.uthsc.edu/CytochromeP450.html). In this BLASTx search, the sequences were used for a batch BLAST search against the protein database after being translated in 6 different reading frames (see the example BLASTx output in Fig 1.3). Fig 1.3 shows the BLAST similarity search output of the query sequence of *H. axyridis* contig sequence (candidate genes) hits matched with the subject sequence of protein sequences of CYP genes (the figure shows the match with CYP345B1 from *T. castaneum*) from different insects incuding *T. castaneum*. To perform the batch BLAST, we used batch BLAST standalone BLAST software.

The aligned fragments from *H. axyridis* fragments in comparison to different insects are used to assemble the genes. While assembling the exon pieces, parameters such as reading frame and phase are carefully considered. The aligned regions provide evidence about the location of inton-exon GT and AG boundaries. These boundaries lie at or near the edges of the aligned fragments. The phase information (if the GT boundry is exactly at the starting position of the codon, then it is in phase (0); if the GT boundry is 1 nucleotide into the codon, then it is in phase (1), if the GT boundry is 2 nucleotides into the codon, then it is in phase (2). The AG boundry is complementary to the GT boundary.

The GT and AG boundry have to be in the same phase to correctly assemble the protein sequence (Long et al. 1998). The phase parameters facilitate the efficientremoval of intronic regions.

```
> CYP345D2Triboliumcastaneum111
Length=492

 Score =  124 bits (310),  Expect = 3e-29
 Identities = 60/124 (49%), Positives = 86/124 (70%), Gaps = 3/124 (2%)
 Frame = +2

Query  1400  YTYR-SFKYWETRNVYYEKPVPIFGNFYDVAIRKRHMGDVLKEIHLKLDDNVPYFGVYIF  1576
             Y YR ++KYWE++ V  EKP  IFG+FYDVA+R++H+   ++EI+ K   + PY G+YIF
Sbjct  15    YLYRKNYKYWESKGVLTEKPFFIFGSFYDVALRRKHLFHKVREIYDKF--STPYVGIYIF  72

Query  1577  HAPNLVVRTKEMIKEVLIKKFSSFPNRMDYTNEVVDPLSSYDLFSMKEDLWKFTRTKLSP  1756
             + P LV+R+ E++K+VL+K F  F NR    NE VDP+  + LFS K D W+  R K+SP
Sbjct  73    NQPTLVIRSPELLKKVLVKDFDKFINRKVAANESVDPVFFHTLFSAKNDNWRNLRAKISP  132

Query  1757  AFSS  1768
              F+S
Sbjct  133   VFTS  136


> CYP345D1Triboliumcastaneum111
Length=494

 Score =  117 bits (294),  Expect = 2e-27
 Identities = 54/124 (44%), Positives = 86/124 (70%), Gaps = 3/124 (2%)
 Frame = +2

Query  1400  YTY-RSFKYWETRNVYYEKPVPIFGNFYDVAIRKRHMGDVLKEIHLKLDDNVPYFGVYIF  1576
             Y Y +++ YW++RNV  +KP   FG+FY++A+RK H+ + ++ I+ +   + PY G+YIF
Sbjct  15    YIYAKNYNYWQSRNVPTDKPFLFFGSFYNIAVRKEHIFERIRTIYNQF--SAPYVGIYIF  72

Query  1577  HAPNLVVRTKEMIKEVLIKKFSSFPNRMDYTNEVVDPLSSYDLFSMKEDLWKFTRTKLSP  1756
             H P L++R+ E++++VL+K F  F NR   TNE VDPL+ + LF  K+ +W+  R KLSP
Sbjct  73    HQPVLLIRSPEILRKVLVKDFDKFTNRNIATNEAVDPLAFHTLFISKDAVWRNLRAKLSP  132

Query  1757  AFSS  1768
              F+S
Sbjct  133   VFTS  136
```

Fig. 1.3. A snapshot of the BLASTx output of candidate contigs. The snapshort of the BLASTx output of candidate contig sequences against the CYP gene fragment protein database. During the BLAST search, the genomic sequences are translated into 6 different reading frames to obtain the optimal hit in the BLAST search.

After the removal of intronic regions, the sequences are assembled together to allow the maximum possible assembly of exons. Intronic sequences are important when they contain regulatory regions, but these sequences are irrelevant for assembling a protein sequence.

To ensure the completeness and correctness of assembly, the assembled sequences are again subjected to BLASTp. The assembled sequences are further studied for their nomenclature based on their sequence similarity to known CYPs proteins. The sequences with >55% sequence identity are assigned to the same subfamily, and sequenses with >95% identity are assigned to the same allelic group. The motifs located in a specific sequence fragment also provide evidence for selecting the correct exon and quantifying the number of genes. The specific information based on the similarity search of our study is presented in Table A1.3 (Appendix). The presence of motif regions, frequently in the C-terminal half, with high similarity scores can help to assign 2 sequences to the same subfamily. After assembling a major portion of a gene (at least 350–400 amino acids long from a total of ~500 amino acids in length), one could use the fragment for comparative studies.

An Example of the Process of Gene detection, Starting from the BLAST Search Using Reference Sequences of Tribolium castaneum till the assembly of Harmonia axyridis CYP genes

As an example of the entire BLAST search process, we chose the CYP345A2 gene, as it was one of the most complete CYP genes recovered in our study (small

fragments of exons missing; coverage range: amino acids 19–487).

Firstly, the reference genes from *T. castaneum* were taken and subjected to a tBLASTn search to obtain output contigs, which served as candidates for further study. The most promising potential candidate contigs were those with both low E-values and high identity scores. An example showing a search involving the CYP345A2 gene. For the output of the tBLASTn search is presented in Fig. 1.4, the very first contig sequence with an E-value of 2e−56 was chosen as the best hit with an identity of 40%.

```
Query= CYP345A2_seq_37_CYP3_clan

Length=505
                                                                                 E
Sequences producing significant alignments:                                    Value
  McKenna_1kb_R1.PF_(paired)_contig_63104 Average coverage: 216.51              2e-56
  McKenna_1kb_R1.PF_(paired)_contig_93915 Average coverage: 68.76               2e-56
  McKenna_1kb_R1.PF_(paired)_contig_34157 Average coverage: 101.84              8e-55
  McKenna_1kb_R1.PF_(paired)_contig_38109 Average coverage: 119.87              3e-51
  McKenna_1kb_R1.PF_(paired)_contig_73770 Average coverage: 137.57              8e-47
  McKenna_1kb_R1.PF_(paired)_contig_100427 Average coverage: 152.34             9e-40
  McKenna_1kb_R1.PF_(paired)_contig_123923 Average coverage: 92.06              9e-35
  McKenna_1kb_R1.PF_(paired)_contig_11981 Average coverage: 122.71              2e-33
  McKenna_1kb_R1.PF_(paired)_contig_93915 Average coverage: 68.76               3e-33
  McKenna_1kb_R1.PF_(paired)_contig_12024 Average coverage: 148.44              3e-31
  McKenna_1kb_R1.PF_(paired)_contig_125712 Average coverage: 125.18             3e-31
  McKenna_1kb_R1.PF_(paired)_contig_97615 Average coverage: 122.47              3e-28
  McKenna_1kb_R1.PF_(paired)_contig_20898 Average coverage: 139.63              7e-28
  McKenna_1kb_R1.PF_(paired)_contig_101506 Average coverage: 67.17              3e-27
  McKenna_1kb_R1.PF_(paired)_contig_44663 Average coverage: 129.82              4e-27
  McKenna_1kb_R1.PF_(paired)_contig_100712 Average coverage: 100.12             5e-27
  McKenna_1kb_R1.PF_(paired)_contig_29860 Average coverage: 165.14              6e-27
  McKenna_1kb_R1.PF_(paired)_contig_9307 Average coverage: 87.47                7e-27
  McKenna_1kb_R1.PF_(paired)_contig_100427 Average coverage: 152.34             2e-26
  McKenna_1kb_R1.PF_(paired)_contig_24576 Average coverage: 80.97               5e-26

> McKenna_1kb_R1.PF_(paired)_contig_63104 Average coverage: 216.51
Length=4164

 Score   = 152 bits (385)  , Expect (2) = 2e-56, Method: Compositional matrix
adjust.
 Identities = 72/180 (40%), Positives = 109/180 (61%), Gaps = 0/180 (0%)
 Frame = +1

Query   112   SPHHDPLVKNMLFLNKNPEWKNVRVKMTPVFTTGKLKGMIPLINDVGETMTKYIAQKIPN 171
           +  +DP+  +MLF+ K  EWK +R K++P F+  KLK M   I+++G ++ ++I
Sbjct  1498  AAEYDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNR  1677
Query  172   FSLEAKEICAKFSTDVIAKCAFGINANSFKNEDAEFRKIGRRIFDFRWSTAIQQTSYFFL  231
             L+ KE+ +KFS DVIAKC FGI+A S + ED EF +I  +IFD R  T+ +   YFF
Sbjct  1678  SGLDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFF  1857
Query  232   PGLVNLLKFRMLDKDASDFLRETFWHTIKLREEKNLKANDLIDAIIALKDNQEFCKNMNF  291
                + + ++ D D   FLR  FW  I+LRE+ N+K NDLID I+ L+ + E  + + F
Sbjct  1858  HSFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNELSERIKF  2037

 Score = 89.0 bits (219),  Expect(2) = 2e-56, Method: Compositional matrix adjust.
```

```
 Identities = 42/87 (49%), Positives = 59/87 (68%), Gaps = 0/87 (0%)
 Frame = +3
Query  25    YFSRNFDHWEKKNVFYFKPIPFFGNFVDISLFRTTIGEHLAKLYNQTTEPFFGIFVFDKP  84
           Y SRN+D+W+K+NV + KP PF GN  +I L +  +    KLYN    PFFGIFVF KP
Sbjct  1239  YISRNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKP  1418
Query  85    HLIIKSPELVKTILVRDFNNFDDRCIA  111
             LI+K  +L+K I V+DF++F D+ ++
Sbjct  1419  ALIVKDVKLLKNIFVKDFDHFMDQHVS  1499

 Score = 97.1 bits (240),  Expect = 2e-19, Method: Compositional matrix adjust.
 Identities = 48/95 (51%), Positives = 63/95 (67%), Gaps = 8/95 (8%)
 Frame = +3
Query  355   KYLNMCVCETLRKYPVLPFLDRTCKEDYKLPNSNVVIEKGTPVFIPMFGLHYDPQYFPNP  414
           +YL +   ETLRKY  +P LDR C +DYK+P +++VIEKG    +P +G   DP+YF NP
Sbjct  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
Query  415   QKYDPERFSD--ENMQNITPFSYIPFGEGPRNCIG  447
             ++Y PERF    E+M      Y+PFG GPRNCIG
Sbjct  3582  EEYIPERFESIKEDM------FYMPFGHGPRNCIG  3668

 Score = 62.4 bits (150),  Expect = 7e-09, Method: Compositional matrix adjust.
 Identities = 39/77 (51%), Positives = 54/77 (71%), Gaps = 1/77 (1%)
 Frame = +1
Query  291   FEGDKvvaqaaqffvaGFETTSSTMAFTLYELCLQPQFQRRVRAEIATCLKEHNG-LTYE  349
             +GDKV+AQA  FFVAGFETT ST+AFTL+ LCL    QR++R  I   +K+H G LT E
Sbjct  2710  LDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLTME  2889
Query  350   ALQSMKYLNMCVCETLR  366
           ++++M YL+  +   L+
Sbjct  2890  SIENMDYLDNVIKGILK  2940

 Score = 52.8 bits (125),  Expect = 5e-06, Method: Compositional matrix adjust.
 Identities = 28/71 (40%), Positives = 40/71 (57%), Gaps = 1/71 (1%)
 Frame = +2
Query  427   MQNITPFSYIPFGEGPRNCIGERFGLIGTKLGLIHILSEFEVEKSSDTPVPLEFEPKSFV  486
           + NI P S+  +   P++ +G RFGLI  KL ++ IL EFE+  + +TPV LEF   S +
Sbjct  3929  LNNILPISF-KWKLKPQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLI  4105
Query  487   LASKVGLPMKF  497
                L M F
Sbjct  4106  PQPIQQLKMSF  4138
```

Fig. 1.4. tBLASTn output showing the BLAST output obtained by taking the *Tribolium castaneum* query against the database of assembled *Harmonia axyridis* contigs (135,864 contigs). The figure presents the BLAST search output for the first hit (best hit).

Despite the low identity score, the contig was still considered a good hit for further study due to its low E-value. In this example, we considered only 1 contig, as we recovered most of the genes from this contig. In many other cases, we had to evaluate more than 1 contig to assemble the entire gene.

In the second step, the best hit contigs associated with this specific gene were subjected to a BLASTx search against various fragments of CYP genes (from different insects, CYP genes collected from Dr. Nelson's CYP protein dataset). The BLASTx

output gives all of the translated regions of contigs that are similar to fragments of a CYP gene. The BLASTx output was further analyzed to differentiate exons from introns based on the AG-GT boundary. Detecting the exact intronic boundary is a tedious task, occasionally involving ambiguous options. During the process of detecting the phase boundary, the specific contigs were translated into 6 reading frame using translators such as ExPASy translator (http://web.expasy.org/translate/). Details of the detection of AG-GT phase boundaries are provided in text elaborating of fig.1.3. The BLASTx output obtained when contig 'McKenna_63104' subjected to a BLAST search against the entire database of CYP genes (different fragments of CYPs from different organisms) given in Fig. 1.5. The output was further scanned to obtain exon-intron boundaries after translating the entire contig.

```
Query= McKenna_1kb_R1.PF_(paired)_contig_63104 Average coverage: 216.51
Length=4164
                                                                      Score    E
Sequences producing significant alignments:                          (Bits)  Value
  CYP345A2Triboliumcastaneum111                                       150    3e-57
  CYP345A1Triboliumcastaneum111                                       142    5e-56
  CYP345C1Triboliumcastaneum111                                       127    4e-49
  CYP345H1Leptinotarsadecemlineata50                                  120    2e-47
  CYP6AQ13LinepithemahumileargentineantLh6                            133    7e-47
  CYP345B1Triboliumcastaneum111                                       118    1e-43
  CYP6K1BlatellagermanicaGermancockroachJeffScottAF281328             120    2e-43
  CYP6AQ14LinepithemahumileargentineantLh6                            122    5e-43
  CYP6J1AF281325BlattellagermanicaGermancockroachJeffScott12099       108    2e-42
  CYP6AQ18PogonomyrmexbarbatusPb6AQseq4i291                           110    2e-42

> CYP345A2Triboliumcastaneum111
Length=505

 Score =  150 bits (380),  Expect(2) = 3e-57
 Identities = 72/177 (41%), Positives = 108/177 (62%), Gaps = 0/177 (0%)
 Frame = +1
Query  1507  YDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNRSGL  1686
             +DP+  +MLF+ K  EWK +R K++P F+  KLK M   I+++G ++ ++I        L
Sbjct  115   HDPLVKNMLFLNKNPEWKNVRVKMTPVFTTGKLKGMIPLINDVGETMTKYIAQKIPNFSL  174
Query  1687  DVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFFHSF  1866
             + KE+ +KFS DVIAKC FGI+A S + ED EF +I  +IFD R  T+ +   YFF
Sbjct  175   EAKEICAKFSTDVIAKCAFGINANSFKNEDAEFRKIGRRIFDFRWSTAIQQTSYFFLPGL  234
Query  1867  AKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNELSERIKF  2037
               + + ++ D D   FLR  FW  I+LRE+ N+K NDLID I+ L+ + E  + + F
Sbjct  235   VNLLKFRMLDKDASDFLRETFWHTIKLREEKNLKANDLIDAIIALKDNQEFCKNMNF  291

 Score = 89.0 bits (219),  Expect(2) = 3e-57
 Identities = 40/84 (48%), Positives = 57/84 (68%), Gaps = 0/84 (0%)
```

```
 Frame = +3
Query  1248  RNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKPALI  1427
             RN+D+W+K+NV + KP PF GN  +I L +  +     KLYN    PFFGIFVF KP LI
Sbjct  28    RNFDHWEKKNVFYFKPIPFFGNFVDISLFRTTIGEHLAKLYNQTTEPFFGIFVFDKPHLI  87
Query  1428  VKDVKLLKNIFVKDFDHFMDQHVS  1499
             +K  +L+K I V+DF++F D+ ++
Sbjct  88    IKSPELVKTILVRDFNNFDDRCIA  111

 Score =  101 bits (251),  Expect = 4e-22
 Identities = 47/93 (51%), Positives = 64/93 (69%), Gaps = 4/93 (4%)
 Frame = +3
Query  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
             +YL +   ETLRKY  +P LDR C +DYK+P +++VIEKG    +P +G   DP+YF NP
Sbjct  355   KYLNMCVCETLRKYPVLPFLDRTCKEDYKLPNSNVVIEKGTPVFIPMFGLHYDPQYFPNP  414
Query  3582  EEYIPERF--ESIKE--DMFYMPFGHGPRNCIG  3668
             ++Y PERF  E+++      Y+PFG GPRNCIG
Sbjct  415   QKYDPERFSDENMQNITPFSYIPFGEGPRNCIG  447

 Score = 77.8 bits (190),  Expect = 5e-15
 Identities = 38/68 (56%), Positives = 51/68 (75%), Gaps = 1/68 (1%)
 Frame = +1
Query  2713  DGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLTMES  2892
             +GDKV+AQA  FFVAGFETT ST+AFTL+ LCL    QR++R  I   +K+H G LT E+
Sbjct  292   EGDKVVAQAAQFFVAGFETTSSTMAFTLYELCLQPQFQRRVRAEIATCLKEHNG-LTYEA  350
Query  2893  IENMDYLD  2916
             +++M YL+
Sbjct  351   LQSMKYLN  358

 Score = 52.0 bits (123),  Expect = 3e-07
 Identities = 28/71 (40%), Positives = 40/71 (57%), Gaps = 1/71 (1%)
 Frame = +2
Query  3929  LNNILPISF-KWKLKPQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLI  4105
             + NI P S+  +   P++ +G RFGLI  KL ++ IL EFE+  + +TPV LEF   S +
Sbjct  427   MQNITPFSYIPFGEGPRNCIGERFGLIGTKLGLIHILSEFEVEKSSDTPVPLEFEPKSFV  486
Query  4106  PQPIQQLKMSF  4138
                   L M F
Sbjct  487   LASKVGLPMKF  497

> CYP345A1Triboliumcastaneum111
Length=505

 Score =  142 bits (358),  Expect(2) = 5e-56
 Identities = 66/177 (38%), Positives = 106/177 (60%), Gaps = 0/177 (0%)
 Frame = +1
Query  1507  YDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNRSGL  1686
             +DP+  +MLF  K  EWK +R K++P F+  KLK M   I+++G +L ++I    +   L
Sbjct  115   HDPLVKNMLFFNKNPEWKNVRVKMTPVFTTGKLKGMIPLINDIGETLTKYIAQKTSNLSL  174
Query  1687  DVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFFHSF  1866
             + KE+S+K++ DVIAKC FGI+A SL+ ED EF  +  + FD R   + +   YF
Sbjct  175   EAKEISAKYTTDVIAKCAFGINANSLKNEDAEFRNLGRRFFDFRWSNAIQQTSYFLLPGL  234
Query  1867  AKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNELSERIKF  2037
               + ++++ D     FLR  FW+ I+LR++NN K  DLID I+ ++++ E  +   F
Sbjct  235   VNVLKLRVMDKKDSNFLRETFWQTIKLRQENNSKAKDLIDAIIAMKENKEFCKNFNF  291

 Score = 93.2 bits (230),  Expect(2) = 5e-56
 Identities = 43/84 (52%), Positives = 58/84 (70%), Gaps = 0/84 (0%)
 Frame = +3
Query  1248  RNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKPALI  1427
             RNYD+W+K+NV F KP PF GN+ +I L +  +     KLYN    PFFGIFVF KP LI
Sbjct  28    RNYDHWEKKNVFFFKPTPFFGNILDISLFRTTIGEHLAKLYNQTTEPFFGIFVFDKPHLI  87
Query  1428  VKDVKLLKNIFVKDFDHFMDQHVS  1499
             +K  +L+K I V+DF++F D+ V+
Sbjct  88    IKSPELVKTILVRDFNNFDDRGVA  111

 Score =  101 bits (251),  Expect = 4e-22
 Identities = 47/93 (51%), Positives = 64/93 (69%), Gaps = 4/93 (4%)
 Frame = +3
Query  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
             +YL +   ETLRKY  +P LDR C +DYK+P +++VIEKG    +P +G   DP+YF NP
Sbjct  355   KYLNMCVCETLRKYPVLPFLDRTCKEDYKLPNSNVVIEKGTPVFIPMFGLHYDPQYFPNP  414
Query  3582  EEYIPERF--ESIKE--DMFYMPFGHGPRNCIG  3668
```

```
             ++Y PERF  E+++      Y+PFG GPRNCIG
Sbjct  415   QKYDPERFSDENMQNITPFSYIPFGEGPRNCIG  447

 Score = 78.2 bits (191),  Expect = 4e-15
 Identities = 38/71 (54%), Positives = 52/71 (74%), Gaps = 1/71 (1%)
 Frame = +1
Query  2704  FVLDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLT  2883
             F  +GDKV+AQA  FF+AGFETT +T+AFTL+ LCL   IQ K+R  I   +K+H G LT
Sbjct  289   FNFEGDKVVAQAAQFFIAGFETTSATMAFTLYELCLQPQIQSKVRTEIMTCVKEHNG-LT  347
Query  2884  MESIENMDYLD  2916
              E++++M YL+
Sbjct  348   YEALQDMKYLN  358

 Score = 49.7 bits (117),  Expect = 2e-06
 Identities = 26/71 (37%), Positives = 40/71 (57%), Gaps = 1/71 (1%)
 Frame = +2
Query  3929  LNNILPISF-KWKLKPQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLI  4105
             + NI  P S+  +   P++ +G RFGLI+ KL ++ +L  FE+  + +TPV LEF   S +
Sbjct  427   MQNITPFSYIPFGEGPRNCIGERFGLISTKLGLIHVLSNFEVERSSDTPVPLEFEPKSFV  486
Query  4106  PQPIQQLKMSF  4138
                   L M F
Sbjct  487   LASKVGLPMKF  497

> CYP345C1Triboliumcastaneum111
Length=494

 Score =  127 bits (318),  Expect(2) = 4e-49
 Identities = 69/205 (34%), Positives = 116/205 (57%), Gaps = 10/205 (4%)
 Frame = +1
Query  1510  DPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDAS-PNRSGL  1686
             DP++ H+LFI K  +W+ +R+K++P F+ SK+K M   I+N    +  +++    + + +
Sbjct  108   DPISTHILFILKNPDWRELRTKMTPVFTSSKIKIMSELIENASHEMTNYLNNHIKDYNSV  167
Query  1687  DVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPI-TSFRFLCYFFFHS  1863
             +++++  KF+VDVI   +FG+ A S + E+ +F  +A ++ D   I T+FRF CY
Sbjct  168   EMRDVCLKFTVDVIGSTIFGVQANSFKDENSQFSSVAKRLIDWDDIVTAFRFRCYLLAPL  227
Query  1864  FAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNE--------L  2019
             F  +FRMKLF  D V FL+  F + ++ R  +N   NDLIDI++ ++ DN         +
Sbjct  228   FVNLFRMKLFPPDCVNFLKNTFLDIMDKRSVSNKSRNDLIDILLQMKNDNRNFIEGDILV  287
Query  2020  SERIKFGECFFCFCFVIMHFDLYYF  2094
             S+ + F    F      M F LY F
Sbjct  288   SQALMFFVAGFETTSSTMGFALYEF  312

 Score = 85.5 bits (210),  Expect(2) = 4e-49
 Identities = 35/82 (43%), Positives = 55/82 (68%), Gaps = 0/82 (0%)
 Frame = +3
Query  1248  RNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKPALI  1427
             RN+ +W+K+NVPFIKP PF G++ + +L ++++   F  LY     PF G F+  KP L+
Sbjct  19    RNFKHWEKKNVPFIKPLPFFGSIYDGVLMRHSIGEVFYDLYYKSTKPFVGFFILDKPCLL  78
Query  1428  VKDVKLLKNIFVKDFDHFMDQH  1493
             ++D KL+K I V DF +F D++
Sbjct  79    IRDPKLIKKILVNDFQYFYDRN  100

 Score = 86.7 bits (213),  Expect = 1e-17
 Identities = 44/93 (48%), Positives = 55/93 (60%), Gaps = 4/93 (4%)
 Frame = +3
Query  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
             EYL +   E LRKY  VP LDR C   Y IP+T++ I+K     +P      DP+YF N
Sbjct  345   EYLDMCVKEVLRKYPVVPFLDRKCNTTYTIPDTNVTIDKDTPIFIPSLALHYDPQYFPNA  404
Query  3582  EEYIPERFESIKE---DMF-YMPFGHGPRNCIG  3668
             + + PERF S  +   D F Y+PFG GPRNCIG
Sbjct  405   DIFDPERFSSNNKTGIDSFAYLPFGEGPRNCIG  437


 Score = 75.9 bits (185),  Expect = 2e-14
 Identities = 36/77 (47%), Positives = 55/77 (72%), Gaps = 1/77 (1%)
 Frame = +1
Query  2710  LDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLTME  2889
             ++GD +++QAL+FFVAGFETT ST+ F L+    N DIQ K+R  I+DI  K+ G +  +
Sbjct  281   IEGDILVSQALMFFVAGFETTSSTMGFALYEFARNPDIQDKIRNEIKDISDKY-GDIKYD  339
Query  2890  SIENMDYLDNVIKGILK  2940
             S++ M+YLD  +K +L+
```

```
Sbjct  340   SLKEMEYLDMCVKEVLR  356

 Score = 36.6 bits (83),  Expect = 0.014
 Identities = 14/45 (32%), Positives = 29/45 (65%), Gaps = 0/45 (0%)
 Frame = +2
Query  3971  PQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLI  4105
              P++ +G RFGL+  KL ++ ILKEF +   +++   ++F+   ++
Sbjct  432   PRNCIGARFGLLTAKLGLVHILKEFVVSCNEKSNEKIKFNPKGMV  476


> CYP345H1Leptinotarsadecemlineata50
Length=503

 Score =  120 bits (300),  Expect(2) = 2e-47
 Identities = 55/178 (31%), Positives = 104/178 (59%), Gaps = 2/178 (1%)
 Frame = +1
Query  1498  AAEYDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNR  1677
              A   D + ++ +F +K   WK  R+K++P F+  KLKAMF  I N+   + ++++   N
Sbjct  110   AEPQDEIMSNFMFFQKSPRWKSDRTKLTPVFTSGKLKAMFSLIYNVAEEMVKYLE--DNV  167
Query  1678  SGLDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFF  1857
                ++ K++S+++S DVIAKC FGIDA   + ++ +F +    + +     +F  + YFF
Sbjct  168   GKIEAKDISARYSTDVIAKCAFGIDAYCFDDQESDFRKYGRLMLEFSLRNAFSQMSYFFI  227
Query  1858  HSFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNELSERI  2031
               ++ KIF + LF  +V  +  + F + ++ RE +  + ND +D+++DL+  N+L E +
Sbjct  228   QTWVKIFHINLFSEEVRNYFSQAFTQTMKSRELSKTRVNDFVDLLIDLKNSNQLPEEL  285


 Score = 87.0 bits (214),  Expect(2) = 2e-47
 Identities = 44/105 (42%), Positives = 71/105 (68%), Gaps = 0/105 (0%)
 Frame = +3
Query  1176  IMHWFIEVlllisfllyllhlyisRNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSF  1355
              I  W +++++ I FL+  L +Y  RN+DYW+KR V   KP PF+GN+GE++  K  +S +
Sbjct  3     IPSWILQLIIFIVFLICFLWMYSVRNFDYWKKRGVYSPKPTPFLGNIGELVFLKKCLSEW  62
Query  1356  FEKLYNDMDAPFFGIFVFSKPALIVKDVKLLKNIFVKDFDHFMDQ  1490
                 LY   D  FFG+F+F +P+L++KD KL++ + +KD D+F D+
Sbjct  63    LSSLYFSTDERFFGVFMFDEPSLVLKDPKLIQLVMMKDADYFPDR  107

 Score = 95.9 bits (237),  Expect = 2e-20
 Identities = 46/93 (50%), Positives = 58/93 (63%), Gaps = 3/93 (3%)
 Frame = +3
Query  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
              +YL     ETLRKY  +  LDR C  DYK+P TD+VIEKG+   +P  G   D KYF+ P
Sbjct  349   KYLKNCIYETLRKYPVLAFLDRSCIADYKLPGTDLVIEKGMRVYIPLAGLHLDEKYFEEP  408
Query  3582  EEYIPERFES---IKEDMFYMPFGHGPRNCIGG  3671
               +Y P+RF      +  +FYMPFG GPR C+GG
Sbjct  409   RKYNPDRFSEKMYNQNGLFYMPFGEGPRKCLGG  441

 Score = 82.0 bits (201),  Expect = 3e-16
 Identities = 41/73 (57%), Positives = 52/73 (72%), Gaps = 1/73 (1%)
 Frame = +1
Query  2722  KVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLTMESIEN  2901
              K   QAL+FF AGFETT S+I+FTLH LCLN ++Q K+R  I + IK HGG +T ESI++
Sbjct  289   KACGQALMFFAAGFETTSSSISFTLHELCLNREVQNKVRAEILETIKNHGG-ITYESIQD  347
Query  2902  MDYLDNVIKGILK  2940
              M YL N I   L+
Sbjct  348   MKYLKNCIYETLR  360

 Score = 43.9 bits (102),  Expect = 8e-05
 Identities = 20/45 (45%), Positives = 30/45 (67%), Gaps = 0/45 (0%)
 Frame = +2
Query  3971  PQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLI  4105
              P+  LG RFGLI+ +LA++ IL +FE+    ETP  +EF   S++
Sbjct  435   PRKCLGGRFGLISTQLALIHILSKFEVQKCAETPDPIEFEPKSIL  479

> CYP6AQ13LinepithemahumileargentineantLh6
Length=516

 Score =  133 bits (334),  Expect(2) = 7e-47
 Identities = 69/176 (40%), Positives = 103/176 (59%), Gaps = 3/176 (1%)
 Frame = +1
Query  1498  AAEYDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASP--  1671
```

```
            A E D +    LF+ K   WKL+R+K++PFF+  K++ MF  I   G +L  ++D S
Sbjct  115  ADENDRLGCANLFLIKNPAWKLLRTKLTPFFTSGKMRKMFDLILQCGENLDAYLDKSEFE  174
Query  1672 -NRSGLDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCY  1848
             N S +DVKEL++KF+ DV+    FG++  S +  D EF +    IFD   + +F  L
Sbjct  175  DNGSIVDVKELTAKFTTDVVGSTAFGLEVNSFKYPDAEFSKCRRMIFDYSTVRAFELLMV  234
Query  1849 FFFHSFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNE  2016
             FF  S   +F +++F  +   FLR+VFWE    R  + VK NDLIDI+V L+++NE
Sbjct  235  FFIPSIVSLFSIRIFGKEPTIFLRKVFWETFTQRINSGVKRNDLIDILVKLKENNE  290

 Score = 72.0 bits (175),  Expect(2) = 7e-47
 Identities = 34/85 (40%), Positives = 53/85 (63%), Gaps = 2/85 (2%)
 Frame = +3
Query  1248 RNYDYWQKRN-VPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDA-PFFGIFVFSKPA  1421
             RN++YW+KR  V    P PF+GN  + L  K   + F ++LY+     P+ G +V  KP
Sbjct  29   RNFNYWKKRGIVEMTPPTPFLGNFSDCLRFKKAPADFLKELYDQAKGLPYIGFYVLDKPF  88
Query  1422 LIVKDVKLLKNIFVKDFDHFMDQHV  1496
             L++ D +L+K I VKDF+HF D+++
Sbjct  89   LLICDRELVKQILVKDFNHFSDRYI  113

 Score = 80.9 bits (198),  Expect = 6e-16
 Identities = 39/92 (43%), Positives = 55/92 (60%), Gaps = 4/92 (4%)
 Frame = +3
Query  3405 YLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNPE  3584
             YL +   ETLR Y P+  L+RV  K YK+ ++D+V+EK I   +   G   D +YF NPE
Sbjct  365  YLDMVLSETLRMYPPLGYLNRVANKTYKMSDSDLVLEKNIPVYISALGLHYDAEYFPNPE  424
Query  3585 EYIPERFES----IKEDMFYMPFGHGPRNCIG  3668
             ++ PERF+      +    Y+PFG GP +CIG
Sbjct  425  QFDPERFDEKNRHNRPSCVYLPFGDGPHSCIG  456

 Score = 68.6 bits (166),  Expect = 3e-12
 Identities = 34/83 (41%), Positives = 52/83 (63%), Gaps = 0/83 (0%)
 Frame = +1
Query  2692 NFFTFVLDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHG  2871
             N   F  DGD ++AQA  FF AG +T+ +TIAF L+ L +  +IQ +LRE I   + +
Sbjct  293  NTENFTYDGDDLMAQAASFFSAGSDTSATTIAFALYELAVKPEIQNRLREEILHALDQSN  352
Query  2872 GKLTMESIENMDYLDNVIKGILK  2940
             GK+T + I+++ YLD V+   L+
Sbjct  353  GKITYDMIQSLPYLDMVLSETLR  375

> CYP345B1Triboliumcastaneum111
Length=506

 Score =  118 bits (295),  Expect(2) = 1e-43
 Identities = 58/182 (32%), Positives = 107/182 (59%), Gaps = 8/182 (4%)
 Frame = +1
Query  1504 EYDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNRSG  1683
             E DP+ +H+LF+ K  +W+ MR KI+P F+  K+K M+  I   G  + +H+    ++S
Sbjct  116  ENDPMGSHLLFLLKTPDWRDMRRKITPVFTSGKMKNMYSLISEAGNDMIQHMRKEVSKSD  175
Query  1684 -LDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFD----TRPITSFRFLCY  1848
              L+++E++++++ D I    FGI+A   + E  EF  ++ ++F+     R I++    CY
Sbjct  176  QLEMREVAARYTTDAITSTSFGINANCFKNEKAEFREVSRRVFNWAIWERSIST---TCY  232
Query  1849 FFFHSFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNELSER  2028
             F   +  K+F++K  D+   TFLR  FW  +  RE+     NDL+DI++D++K  ++++
Sbjct  233  FIAPNLVKLFKLKFIDSASATFLREAFWRTMTDREEKKFVRNDLLDILIDIKKQEDINDP  292
Query  2029 IK  2034
              K
Sbjct  293  YK  294

 Score = 76.3 bits (186),  Expect(2) = 1e-43
 Identities = 31/84 (37%), Positives = 53/84 (64%), Gaps = 0/84 (0%)
 Frame = +3
Query  1248 RNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKPALI  1427
             R +++W+ +NVP + P PF GN  E+   + N+  F  ++YN    PF G F+  +P L+
Sbjct  29   RKFNHWKSKNVPQVAPIPFFGNAFEVFTWRKNIGEFARQIYNSTTKPFIGFFICDEPYLL  88
Query  1428 VKDVKLLKNIFVKDFDHFMDQHVS  1499
             ++D +L+K+I VKDF  F ++ +S
Sbjct  89   IRDPELVKSILVKDFAVFSNRSIS  112

 Score = 93.6 bits (231),  Expect = 9e-20
 Identities = 47/93 (51%), Positives = 60/93 (65%), Gaps = 4/93 (4%)
 Frame = +3
```

```
Query  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
             EYL +   ETLRKY  +P LDR C  DY+IP +D+V+EKG    +   G   DP+YF +P
Sbjct  359   EYLDMCIKETLRKYPVLPFLDRKCDTDYRIPGSDVVLEKGSPVFISVSGLHYDPQYFPDP  418
Query  3582  EEYIPERF--ESIKE--DMFYMPFGHGPRNCIG  3668
             ++Y P RF  E+IK      Y+PFG GPRNCIG
Sbjct  419   DKYDPLRFTEENIKSRPQFTYLPFGEGPRNCIG  451

 Score = 91.7 bits (226),  Expect = 4e-19
 Identities = 43/79 (55%), Positives = 61/79 (78%), Gaps = 1/79 (1%)
 Frame = +1
Query  2704  FVLDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLT  2883
             + LDGDK++AQA  FFVAGFETT STI FTL+ L +N D+Q KL+  IRD+++KH G+++
Sbjct  293   YKLDGDKLVAQATQFFVAGFETTSSTICFTLYELAINKDLQNKLKSEIRDVVRKH-GEIS  351
Query  2884  MESIENMDYLDNVIKGILK  2940
               S+++M+YLD  IK  L+
Sbjct  352   YNSLKDMEYLDMCIKETLR  370

 Score = 35.8 bits (81),  Expect = 0.023
 Identities = 15/60 (25%), Positives = 31/60 (52%), Gaps = 0/60 (0%)
 Frame = +2
Query  3971  PQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLIPQPIQQLKMSFINTD  4150
             P++ +G RFG ++ K  + +I+ EFE+   ++T   ++      +  P+  L + F   D
Sbjct  446   PRNCIGARFGSVSSKSGVAKIISEFEVDLCEKTQHPIQIDPKGFLMAPVSDLVLKFKRLD  505

> CYP6K1BlatellagermanicaGermancockroachJeffScottAF281328
Length=524

 Score =  120 bits (300),  Expect(2) = 2e-43
 Identities = 57/185 (31%), Positives = 101/185 (55%), Gaps = 7/185 (3%)
 Frame = +1
Query  1498  AAEYDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNR  1677
             AAE D + +  LF   G  WK +R K+SP F+  ++K M+  +++    L+ ++  + N
Sbjct  113   AAESDTLGSQNLFTLNGAPWKYLRVKLSPTFTSGRMKKMYPLVESCAKQLQDYLKENCNT  172
Query  1678  SGLDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFF  1857
               ++VKE ++K++ DVI+ C FGI++ SL+  + EF     KIF+     +F  +  FF
Sbjct  173   KAIEVKETTAKYATDVISTCAFGIESNSLKDPNAEFREFGRKIFEFTRYRTFEVMALFFS  232
Query  1858  HSFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLR-------KDNE  2016
                 K      F  +   FLR+VFW+ I  RE N +  +D +D+++ L+       +D E
Sbjct  233   PGLVKFLNGNFFTKETTEFLRKVFWDTINFRESNKISRDDFMDLLIQLKNKGTIDNEDGE  292
Query  2017  LSERI  2031
             ++E++
Sbjct  293   VTEKV  297

 Score = 73.2 bits (178),  Expect(2) = 2e-43
 Identities = 30/82 (37%), Positives = 51/82 (63%), Gaps = 0/82 (0%)
 Frame = +3
Query  1254  YDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKPALIVK  1433
             + YW+++ V   KP P  GN    +LQK +       +Y   +APF G ++F++PA+++K
Sbjct  31    FTYWKRKGVVNPKPLPVFGNFLPSVLQKRSPGQILWDIYKAAEAPFVGFYIFARPAILIK  90
Query  1434  DVKLLKNIFVKDFDHFMDQHVS  1499
             D  ++K++ VKDF+ F D+H S
Sbjct  91    DPNIIKHVLVKDFNAFSDRHAS  112

 Score =  100 bits (250),  Expect = 6e-22
 Identities = 45/92 (49%), Positives = 61/92 (67%), Gaps = 4/92 (4%)
 Frame = +3
Query  3405  YLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNPE  3584
             YL +   ETLRKY P+P+LDRVC +DYK+P TD++IE+     +   G  +DP+Y+ NPE
Sbjct  375   YLHMVVSETLRKYPPLPLLDRVCLQDYKVPGTDLIIERDTPVFIALLGLHRDPQYYPNPE  434
Query  3585  EYIPERF----ESIKEDMFYMPFGHGPRNCIG  3668
              Y PERF    +  ++   Y+PFG GP NCIG
Sbjct  435   RYDPERFSEENKRQRKAYTYLPFGEGPHNCIG  466

 Score = 68.6 bits (166),  Expect = 3e-12
 Identities = 33/79 (42%), Positives = 47/79 (60%), Gaps = 0/79 (0%)
 Frame = +1
Query  2704  FVLDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLT  2883
             F   GD +++Q  LFF AGFET  +T++FTL+ L L  D+Q +LR  I  ++K   GK T
Sbjct  307   FEFTGDNLVSQPALFFTAGFETNATTLSFTLYELSLQPDLQNRLRSEIAGVMKTSNGKPT  366
Query  2884  MESIENMDYLDNVIKGILK  2940
              E +  M YL  V+   L+
```

```
Sbjct  367   YEDVFGMPYLHMVVSETLR  385

 Score = 46.2 bits (108),  Expect = 2e-05
 Identities = 21/59 (36%), Positives = 36/59 (62%), Gaps = 0/59 (0%)
 Frame = +2
Query  3971  PQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLIPQPIQQLKMSFINT  4147
              P + +G RFG +AVK A++ +L EFE+    +TP+ LE ST S +      + ++F+ +
Sbjct  461   PHNCIGLRFGYMAVKTALVHMLAEFEVKPCKDTPIPLELSTRSSVLATTSGIPLTFVKS  519

> CYP6AQ14LinepithemahumileargentineantLh6
Length=516

 Score =  122 bits (305),  Expect(2) = 5e-43
 Identities = 64/176 (37%), Positives = 100/176 (57%), Gaps = 5/176 (2%)
 Frame = +1
Query  1495  LAAEYDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPN  1674
              LA   D +    LF  + + WK++R+K++PFF+  K+K MF  +   G  L  H++ SP
Sbjct  114   LADPKDRLGYATLFFLRNQAWKIIRTKMTPFFTSGKMKKMFELMIQCGKHLDEHLN-SPE  172
Query  1675  RSG----LDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFL  1842
                G    +++KEL++KF+ DVI    FG++  S +  D EF +    IF    I  F  L
Sbjct  173   FEGKGKTIEIKELTAKFTTDVIGSTAFGLEVNSFKDPDAEFRKYGKMIFHYNAIRGFEML  232
Query  1843  CYFFFHSFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKD  2010
                FF     ++ ++K+F  +   FLR+VFWE I  R K+  K NDLIDI+V+L+++
Sbjct  233   AIFFLPEIVRLAKVKMFGKEPTEFLRKVFWETINQRMKSGAKRNDLIDILVELKQN  288

 Score = 70.1 bits (170),  Expect(2) = 5e-43
 Identities = 33/85 (39%), Positives = 53/85 (63%), Gaps = 2/85 (2%)
 Frame = +3
Query  1248  RNYDYWQKRNVP-FIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDA-PFFGIFVFSKPA  1421
              R + YW+KR +     P PF GN  + LL K   + F ++LY+     P+ G +V  KP
Sbjct  29    RKFKYWKKRGISETAPPTPFFGNFADCLLFKKAPADFLKELYDQAKGLPYIGFYVLDKPI  88
Query  1422  LIVKDVKLLKNIFVKDFDHFMDQHV  1496
              L+++D +L+KNI VKDF++F +++V
Sbjct  89    LLIRDRELVKNILVKDFNYFSNRYV  113

 Score = 77.8 bits (190),  Expect = 5e-15
 Identities = 37/92 (41%), Positives = 54/92 (59%), Gaps = 4/92 (4%)
 Frame = +3
Query  3405  YLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNPE  3584
              YL +   E+LR Y P+  L+R+ T+ YK+P +++V+EK     +   G   DP+YF NPE
Sbjct  366   YLDMVVSESLRMYPPLGYLNRITTEPYKLPNSNLVLEKDTPVYISMLGMHYDPEYFPNPE  425
Query  3585  EYIPERF----ESIKEDMFYMPFGHGPRNCIG  3668
              ++ PERF    +  +    Y PFG GP  CIG
Sbjct  426   KFDPERFNEENKRNRPSTVYFPFGEGPHACIG  457

 Score = 68.2 bits (165),  Expect = 4e-12
 Identities = 32/79 (41%), Positives = 53/79 (68%), Gaps = 0/79 (0%)
 Frame = +1
Query  2704  FVLDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLT  2883
              F  DGD ++AQA  FF AGFET+ +TI+F L+ L L  ++Q +LR+ I + +++  GK+T
Sbjct  298   FTYDGDDLMAQAASFFSAGFETSATTISFALYELALCPEMQNRLRKEILEALEQSNGKIT  357
Query  2884  MESIENMDYLDNVIKGILK  2940
               + + ++ YLD V+   L+
Sbjct  358   YDLVMSLPYLDMVVSESLR  376

 Score = 37.0 bits (84),  Expect = 0.010
 Identities = 22/63 (35%), Positives = 38/63 (61%), Gaps = 3/63 (4%)
 Frame = +2
Query  3971  PQSFLGRRFGLIAVKLAILQILKEFELHSTDET--PVNLEFSTASLIP-QPIQQLKMSFI  4141
              P + +G RFGL+  KL I++ILK+ E+  +++T  PV ++   A L P   +  L +  +
Sbjct  452   PHACIGNRFGLLQSKLGIMEILKKCEVTPSEKTTIPVQIDPRGAMLAPLNGVLYLNIRKL  511
Query  4142  NTD  4150
              NT+
Sbjct  512   NTN  514

> CYP6J1AF281325BlattellagermanicaGermancockroachJeffScott12099
Length=501

 Score =  108 bits (270),  Expect(2) = 2e-42
 Identities = 55/165 (34%), Positives = 93/165 (57%), Gaps = 1/165 (0%)
 Frame = +1
```

```
Query  1513  PVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDA-SPNRSGLD  1689
             P+ A+ +F  +G++WK +R+ ++P F+  K+K MF  +D  G  L   I+  +   + +
Sbjct  119   PLNANAIFALRGQKWKHVRTSLTPTFTTGKMKNMFYLVDKCGQQLVLFIEKFAKAENPVA  178
Query  1690  VKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFFHSFA  1869
             VK+   +F++DV A C FGI+  SL+    EF  + H+IF     ++   L  FF
Sbjct  179   VKDAVERFTMDVTAMCAFGIECNSLQDPKAEFSNLLHRIFQLSFTSAVANLATFFAPWVQ  238
Query  1870  KIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLR  2004
               FR+KL D+++   +R + W  + LREK   K NDL+D +++LR
Sbjct  239   NFFRLKLMDSEIEDRIRDIVWRAVHLREKTGEKRNDLLDYLMELR  283


 Score = 82.0 bits (201),  Expect(2) = 2e-42
 Identities = 34/84 (41%), Positives = 62/84 (74%), Gaps = 1/84 (1%)
 Frame = +3
Query  1248  RNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYND-MDAPFFGIFVFSKPAL  1424
             R++++W+KR V +++P PF GN+ ++LLQK  +  + + +Y + ++ P+ GIF F +PAL
Sbjct  28    RHFNFWKKRGVIYVRPLPFFGNLKDVLLQKKYIGYYLKDIYEENINKPYVGIFAFDQPAL  87
Query  1425  IVKDVKLLKNIFVKDFDHFMDQHV  1496
             +V D++++KNI VKD  +F+D+ V
Sbjct  88    LVNDLEIVKNILVKDSRNFIDRMV  111

 Score = 82.0 bits (201),  Expect = 3e-16
 Identities = 42/93 (46%), Positives = 58/93 (63%), Gaps = 5/93 (5%)
 Frame = +3
Query  3402  EYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
             +YL +   ETLRKY  +P LDR C +DY + + D+++  G    +P Y    D KYF +P
Sbjct  355   KYLDMVVNETLRKYPAIPFLDRRCQEDYPLTQ-DLMLPAGTGVYIPVYALHHDSKYFPSP  413
Query  3582  EEYIPERF-ESIKEDM---FYMPFGHGPRNCIG  3668
              ++ PERF E  K+++    YMPFG GPRNCIG
Sbjct  414   AKFDPERFSEKNKQNIPHFAYMPFGEGPRNCIG  446

 Score = 69.7 bits (169),  Expect = 1e-12
 Identities = 35/80 (44%), Positives = 50/80 (63%), Gaps = 0/80 (0%)
 Frame = +1
Query  2701  TFVLDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKL  2880
             T  LDGD  +AQA  F VAGF T+  T+ F L+ L ++ DIQ   R  I+D++ H  K+
Sbjct  287   TSKLDGDDFVAQAFGFLVAGFHTSSMTLTFALYELSVHQDIQTTARTEIKDVLEHHKKKV  346
Query  2881  TMESIENMDYLDNVIKGILK  2940
             T  SI++M YLD V+   L+
Sbjct  347   TYYSIKDMKYLDMVVNETLR  366

 Score = 33.1 bits (74),  Expect = 0.15
 Identities = 15/37 (41%), Positives = 23/37 (63%), Gaps = 0/37 (0%)
 Frame = +2
Query  3971  PQSFLGRRFGLIAVKLAILQILKEFELHSTDETPVNL  4081
             P++ +G RFG + VK A++ IL  FE+    ET + L
Sbjct  441   PRNCIGMRFGSMQVKAALIHILSNFEVSPCKETRIPL  477

> CYP6AQ18PogonomyrmexbarbatusPb6AQseq4i291
Length=291

 Score =  110 bits (275),  Expect(2) = 2e-42
 Identities = 59/162 (37%), Positives = 95/162 (59%), Gaps = 3/162 (1%)
 Frame = +1
Query  1531  LFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNRSG-LDVKELSS  1707
             LF  K   WK++R K++PFF+  KLK MFG +     +L  ++D+   +   +DVK+LS+
Sbjct  125   LFSIKNPAWKIIRMKLTPFFTSGKLKKMFGLMLECTKNLEDYLDSLKLKGKVIDVKDLSA  184
Query  1708  KFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFFHSFAKIFRMK  1887
             K +++VI    +G+D    +  D  F +    IF+   I S+  L  FF  +  +  R K
Sbjct  185   KLTMNVINSTAYGLDTNPFK--DPNFYKYGRMIFNNSYIRSWEILALFFLPNVVRRTRFK  242
Query  1888  LFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDN  2013
             LF  +   FLR++FWE I  R +++ K NDLIDI+V+L+K++
Sbjct  243   LFGKETTIFLRKIFWETITKRMESDTKRNDLIDILVELKKNS  284
```

Fig. 1.5. BLASTx output obtained when contig ‘McKenna_63104’ was subjected to a BLAST search against the entire database of CYP genes (different fragments of CYPs from different organisms). The output was further scanned to obtain exon-intron boundaries after translating the entire contig. The fragments were then assembled

together with consideration of the ‘nucleotide number.’ When the count is on the positive reading frame, then the ‘nucleotide number’ on the query sequence is considered when assembling the pieces. However, if the count is on negative reading frame, then the ‘nucleotide number’ of the subject sequence is used.

During the phase boundary detection, the consensus regions are generally considered part of the exon. After detecting the boundary, intronic regions were eliminated from the final query sequence output.

The resulting sequences were assembled on the basis of the ‘nucleotide number’ on the query sequence if the sequence was in the positive reading frame. Otherwise (if the query was in the negative reading frame’), the ‘nucleotide number’ of the subject sequence was considered. The output obtained after the phase detection for contig’McKenna_63104’ is given in Fig. 1.6.

```
Query= McKenna_1kb_R1.PF_(paired)_contig_63104 Average coverage: 216.51

Query  1176  (2)WFIEVlllisfllyllhlyisRNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSF     1355
Query  1356  FEKLYNDMDAPFFGIFVFSKPALIVKDVKLLKNIFVKDFDHFMDQHVS(2)              1499
Query  1507  (2)YDPVTAHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNRSGL  1686
Query  1687  DVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFFHSF     1866
Query  1867  AKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNE(2)
Query  2713  (1)GDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGGKLTMES    2892
Query  2893  IENMDYLDNVIKG(1)                                                   2916
Query  3402  (1)SEYLFVQTVETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNP  3581
Query  3582  EEYIPERFESIKEDMFYMPFGHGPRNCI(1)                                   3668
              (1)GRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLI                      4105
Query  4106  PQPIQQLKMSFINT(1)                                                  4147
```

Fig. 1.6. BLASTx output when contig 63104 was assembled on the basis of phase information. In the output, the number on the query and subject sequence is the ‘nucleotide number,’ and the number within the brackets indicates the phase of the sequence.

The final assembled exon sequence for specific genes or contigs are checked through BLASTp to assign them to specific families and to check the coverage for the gene. In the following example, the previously assembled sequence subjected to a BLASTp search was revealed to provide good coverage for the CYP345A2 gene (presenting 19–487 amino acids in contig 'McKenna_ 63104'). If the gene fragments are present on multiple contigs, then the fragments are assembled by observing their family and the region of specific gene coverage. In the latter case, after assembling the specific fragment of the gene, the assembly was subjected to a BLASTp search (given in Fig. 1.7) to detect their family and gene coverage.

```
Query=McKenna_1kb_R1.PF_paired_contig_63104Averagecoverage:216.51
Length=490
                                                                    Score    E
Sequences producing significant alignments:                        (Bits)  Value

  CYP345A2Triboliumcastaneum111                                     426    1e-120
  CYP345A1Triboliumcastaneum111                                     421    7e-119
  CYP345H1Leptinotarsadecemlineata50                                389    3e-109
  CYP345C1Triboliumcastaneum111                                     365    5e-102
  CYP345B1Triboliumcastaneum111                                     363    1e-101

> CYP345A2Triboliumcastaneum111
Length=505

 Score =  426 bits (1096),  Expect = 1e-120, Method: Compositional matrix adjust.
 Identities = 222/487 (46%), Positives = 303/487 (63%), Gaps = 32/487 (6%)

Query  19   YISRNYDYWQKRNVPFIKPRPFVGNMGEILLQKYNMSSFFEKLYNDMDAPFFGIFVFSKP  78
            Y SRN+D+W+K+NV + KP PF GN  +I L +  +     KLYN    PFFGIFVF KP
Sbjct  25   YFSRNFDHWEKKNVFYFKPIPFFGNFVDISLFRTTIGEHLAKLYNQTTEPFFGIFVFDKP  84
Query  79   ALIVKDVKLLKNIFVKDFDHFMDQHVS---YDPVTAHMLFIEKGEEWKLMRSKISPFFSP  135
             LI+K  +L+K I V+DF++F D+ ++   +DP+  +MLF+ K  EWK +R K++P F+
Sbjct  85   HLIIKSPELVKTILVRDFNNFDDRCIASPHHDPLVKNMLFLNKNPEWKNVRVKMTPVFTT  144
Query  136  SKLKAMFGAIDNLGVSLRRHIDASPNRSGLDVKELSSKFSVDVIAKCVFGIDAKSLEIED  195
             KLK M   I+++G ++ ++I        L+ KE+ +KFS DVIAKC FGI+A S + ED
Sbjct  145  GKLKGMIPLINDVGETMTKYIAQKIPNFSLEAKEICAKFSTDVIAKCAFGINANSFKNED  204
Query  196  GEFLRIAHKIFDTRPITSFRFLCYFFFHSFAKIFRMKLFDADVVTFLRRVFWECIELREK  255
             EF +I  +IFD R  T+ +   YFF      + + ++ D D   FLR  FW  I+LRE+
Sbjct  205  AEFRKIGRRIFDFRWSTAIQQTSYFFLPGLVNLLKFRMLDKDASDFLRETFWHTIKLREE  264
Query  256  NNVKGNDLIDIIVDLRKDN---------EGDKVIAQALLFFVAGFETTGSTIAFTLHALC  306
             N+K NDLID I+ L KDN         EGDKV+AQA  FFVAGFETT ST+AFTL+ LC
Sbjct  265  KNLKANDLIDAIIAL-KDNQEFCKNMNFEGDKVVAQAAQFFVAGFETTSSTMAFTLYELC  323
Query  307  LNLDIQRKLRENIRDIIKKHGGKLTMESIENMDYLDNVIKGSEYLFVQTVETLRKYSPVP  366
             L    QR++R  I   +K+H G LT E++++M YL+  +           ETLRKY  +P
Sbjct  324  LQPQFQRRVRAEIATCLKEHNG-LTYEALQSMKYLNMCV----------CETLRKYPVLP  372
Query  367  ILDRVCTKDYKIPETDIVIEKGIITLVPPYGFQKDPKYFDNPEEYIPERFESIKEDM---  423
             LDR C +DYK+P +++VIEKG    +P +G   DP+YF NP++Y PERF    E+M
Sbjct  373  FLDRTCKEDYKLPNSNVVIEKGTPVFIPMFGLHYDPQYFPNPQKYDPERFSD--ENMQNI  430
Query  424  ---FYMPFGHGPRNCIGRRFGLIAVKLAILQILKEFELHSTDETPVNLEFSTASLIPQPI  480
```

```
               Y+PFG GPRNCIG RFGLI  KL ++ IL EFE+  + +TPV LEF   S +
Sbjct  431  TPFSYIPFGEGPRNCIGERFGLIGTKLGLIHILSEFEVEKSSDTPVPLEFEPKSFVLASK  490
Query  481  QQLKMSF  487
             L M F
Sbjct  491  VGLPMKF  497
```

Fig. 1.7. BLASTp output obtained after assembling contig 'McKenna_63104' (for which the identity value for the sequence search is 48%, the E-value is e−120 for the search and the coverage is amino acids 19−487).

Annotation of CYPs though Global Protein Database Searches

The potential 189 *H. axyridis* genomic sequence pieces were further studied by BLAST search (BLASTx) of the 189 candidate contigs in different publicly available protein databases such as TrEMBL, Flybase and Beetlebase. The output obtained are studied on the basis of their identity (proteins) using BLASTx by search against all possible genome sequences in publicly available global protein database. The search output infer that the sequence matches obtained from Flybase were of low identity in comparison to the original annotation in the case of Flybase. One possible reason for the low similarity was through Flybase (http://flybase.org/BLAST/), the best possible similarity matches obtained were mostly *Drosophila* sequences, which are not the best matches for a beetle gene. According to phylogeny, though *Tribolium* is not closely associated with *H. axyridis* (diverged more than 200 Ma ago), but is also a Coleopteran genome (the only well annotated beetle genome available), hence preferred as a reference genome. Forwhich, the contigs were searched against Beetlebase (http://beetlebase.org/BLAST/BLAST.html) to locate a suitable match. However, the BLASTx output obtained represents various chromosomal regions of *T. castaneum* that

are difficult to associate with specific genes; hence, annotation using Beetlebase was not preferred.

Nevertheless, the BLASTx output from Swiss-Prot/TrEMBL (http://www.ebi.ac.uk/Tools/sss/ncbiBLAST/) resulted in a few very good annotations that are comparable to our original CYP annotations (obtained from the traditional gene annotation process). During the BLASTx search using TrEMBL, all possible protein sequences in the several databases were considered, as listed in Table 4.1.

Table 4.1. Databases searched for the annotation of the CYPs of *H. axyridis* by NCBI BLAST searches using Swiss-Prot/TrEMBL.

Primary database	Subsets of the primary database
UniProt Knowledgebase	
UniProtKB/Swiss-Prot	
UniProtKB/Swiss-Prot isoforms	
UniProtKB/TrEMBL	
UniProtKB Taxonomic Subsets	UniProtKB Archaea UniProtKB Arthropoda UniProtKB Bacteria UniProtKB Complete Microbial Proteomes UniProtKB Eukaryota UniProtKB Fungi UniProtKB Human UniProtKB Mammals UniProtKB Nematoda UniProtKB PDB UniProtKB Rodents UniProtKB Vertebrates UniProtKB Viridiplantae UniProtKB Viruses
UniProt Clusters	UniProt Clusters 100% UniProt Clusters 100% (SEG-filtered) UniProt Clusters 90% UniProt Clusters 50%

Table 4.1. Databases searched for the annotation of the CYPs of *H. axyridis* by NCBI BLAST searches using Swiss-Prot/TrEMBL.

Primary database	Subsets of the primary database
Patents	EPO Patent Protein Sequences JPO Patent Protein Sequences KIPO Patent Protein Sequences USPTO Patent Protein Sequences NR Patent Proteins Level 1 NR Patent Proteins Level 2
Structure	
Other Protein Databases	UniProt Archive International Protein Index IntAct IMGT/HLA IPD-KIR IPD-MHC

During the BLAST search using the individual web browsers in the aforementioned 3 websites, only 1 contig could be used for BLAST search at a time. We were able to recover 75 good matches with a good percentage identity (ranging from 45 to 90%) with preferred E-values (<e−5). However, the top 10 BLASTp hits from these database (publicly available protein database) searches were observed to have lower identity scores; although the BLAST output in the web TrEMBL gave us a much better output, the total number of best hits obtain from entire 189 candidate contigs was very low. Hence we conclude that the traditional semiautomated gene annotation is precise and accurate.

Protocol for Automated Gene Annotation

The traditional semiautomated gene annotation process was found to be the most accurate. However, when the time allotted for complete gene annotation is limited, automated gene annotation is preferred. The accuracy of a single automated annotation algorithm is questionable. Different software are preferred for specific organisms, as their algorithms can fit best to estimate specific genomes based on the nature of the genomic elements (e.g., intron size, codon usage, the intron-exon boundary) (Carle-Urioste et al. 1997). One such example is Glimmer, which is a preferred tool for microbial automated gene annotation (Kelley et al. 2012). Automated gene annotation procedures (based on HMM) are gradually adopting different parameters to estimate (and therefore annotate) the genome more accurately (Pedersen et al. 2003).

Before using different available automated gene annotation software on the *H. axyridis* genome, we determined the most useful software for high-sensitivity gene detection. In this study, we used the LGX chromosome of *T. castaneum* (33,080–236,581 bp). This region of the *T. castaneum* genome is well annotated with 111 exons and 24 genes (http://beetlebase.org/cgibin/gbrowse/BeetleBase3.gff3/#search). We attempted to annotate this genomic region using different gene prediction software and then contrasted it with the original (traditional) annotation to assign a sensitivity estimate to each software.

In this procedure, the DNA sequence of the *T. castaneum* LGX chromosome was collected from GenBank. The single large contig was used for gene prediction, after

which the results were compared from various software programs. The output of the gene estimation sensitivity from software is given in appendix Table A.1.2, and the summary table from the sensitivity study is shown in Table 4.2 and software vs sensitivity plot for is provided in Fig.2.

Table 4.2. Estimate of the total number of exonic regions in the LGX genomic region of *Tribolium castaneum* predicted by each software. The gene prediction software used included Augustus, Fgenesh, GeneID, GeneMark, NCBI ORF Finder, SNAP and GenScan [sensitivity = (number of true positive exons obtained/total number of exons)].

Estimate	Augustus	Fgenesh	Gene ID	GeneMark	ORF Finder NCBI	SNAP	GenScan
Sum	83	68	27	11	50	34	80
Sensitivity	0.741	0.607	0.241	0.098	0.446	0.304	0.714

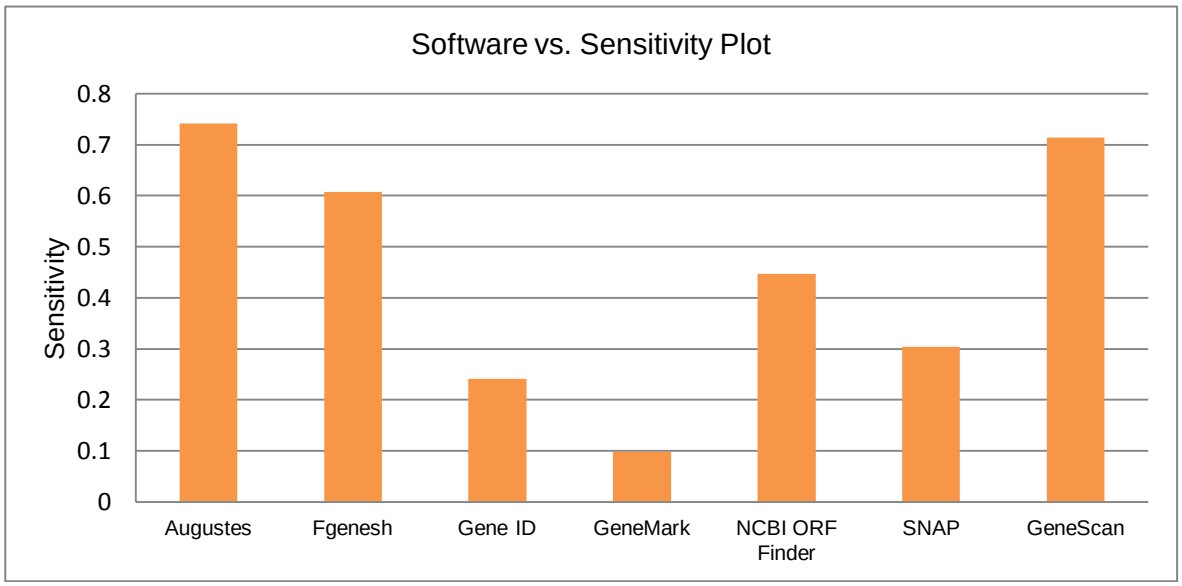

Fig. 2. Plot of the sensitivity of different software used in gene prediction analysis [sensitivity = (number of true positive exons obtained/total number of exons)].

The annotation tools and software used in the analysis above employ HMM-based algorithms. However, they use different criteria to detect the coding regions. In the aforementioned analysis, the sensitivity study was performed based on the sensitivity of software to estimate the 5′ coding region. However a cutoff of ±20 was allowed to obtain optimal sensitivity. In addition, in specific cases, we expanded the cutoff to ±30, e.g., when we found that none of the above software were able to detect specific exonic sequences. The various algorithms implemented in the aforementioned software are given in Table 1.2.

In summary, we found that Augustus, GenScan and Fgenesh have highest sensitivity scores, whereas estimation by GeneMark had the lowest sensitivity. In addition, we considered the software with the highest and lowest sensitivity scores for automated CYP gene annotation.

During the automated gene annotation process, nucleotide sequences were provided as input to the software (Augustus, Fgenesh (pipeline), Fgenes (pipeline), GenScan, GeneMark and Swiss-Prot/TrEMBL NCBI BLASTx). The resulting exonic sequences from Fgenesh and Fgenes were subjected to BLASTp to find the similarity of their output. Based on the resulting output (using a cutoff < e−3), we obtained the best matching hit from each software. We screened out the false positives from the actual output using the best E-value cutoff and similarity score.

Gene Prediction Using Augustus, GenScan and GeneMark

Gene prediction Using Augustus

Augustus uses species-specific parameters together with estimation using the Markov chain transition probability of coding and noncoding regions. The annotation parameters could also be trained on training sets in the GenBank format for specific annotated genes. The aforementioned parameters are stored in the config directory in different files containing the parameters for the exon, intron and intergenic region-related parameters. In addition, various species-specific information was included, such as the optimal order of the Markov chain or the length of the window used in the splice site model (otherwise known as metaparameters). However, these parameters are stored in a species-specific file that is separate from the aforementioned files, e.g. Tribolium_parameters.cfg. These parameters are optimized to give the best estimates for a working set based on different species along with the training set. In particular, the specific information on the training set includes extremely crucial parameters that are prestandardized and optimized for a specific group of organisms to provide the best estimation. In sum, the estimates with the meta-parameters of another species of another training set would result in a less than optimal prediction.

The species with nondocumented meta-parameters also could be documented using the program's 'entraining' reads. This helps to store the meta-parameters in the form of .cfg and GenBank files with annotated information and also in writing the species specific parameters into 3 .pbl files.

For our specific gene annotation process, we used the readily available web-based tool, a snapshot of web interface is provided in Fig.3. However, one could download the Augustus pipeline onto a local server and use the command line options (elaborated in the program manual: http://augustus.gobics.de/binaries/README.TXT).

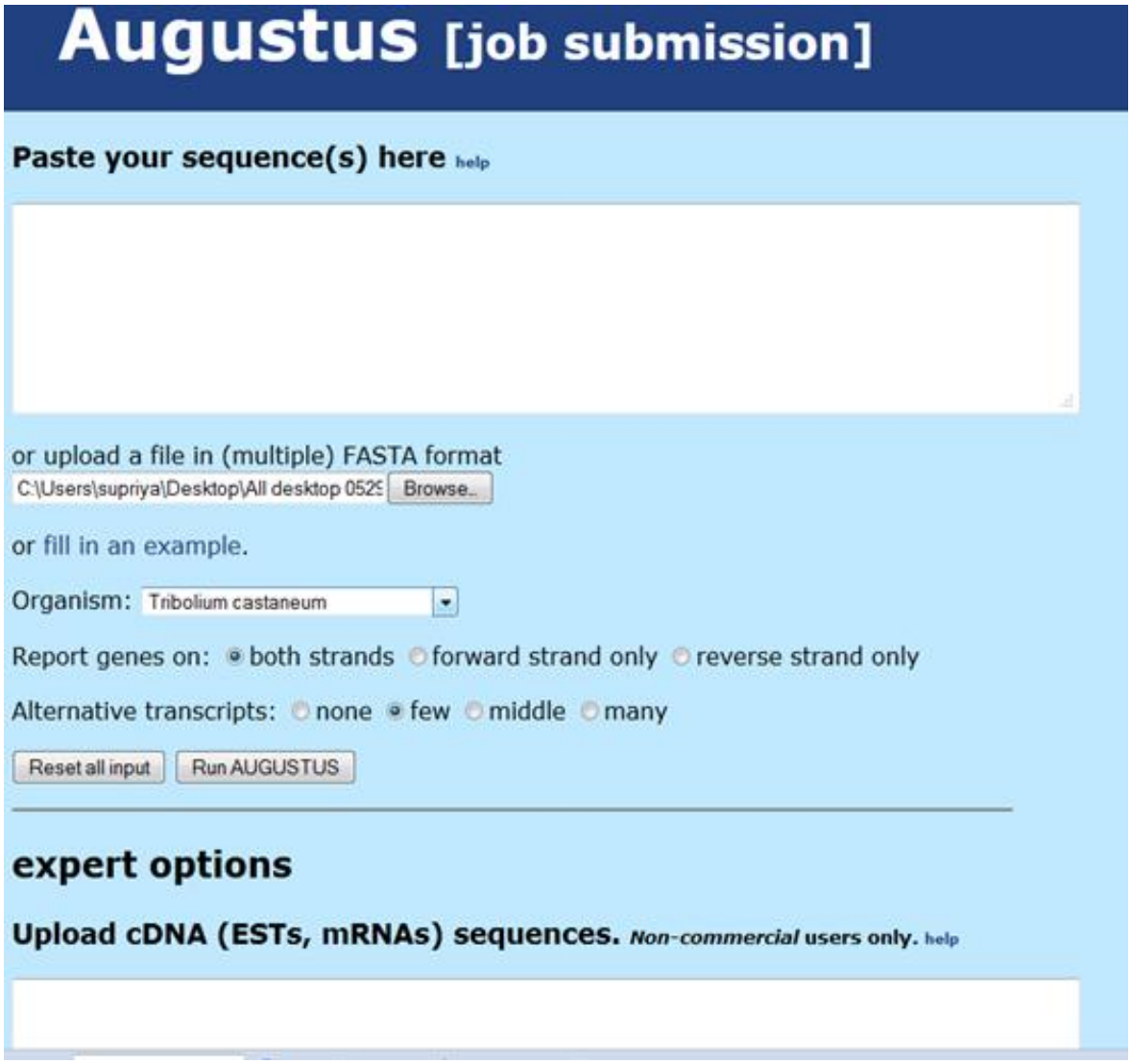

Fig. 3. Web interface for the Augustus gene prediction server. In the "paste sequence(s) here" tab, the user can paste the sequences or upload a file to provide a path to a file in the local system for analysis. In the "Organism" tab, the user can select various taxonomic groups of organisms; specifically, the most closely related groups are preferred for the most accurate gene prediction." The "Report gene on" tab provides the

option to search genes on both strands, and the “alternative transcript” option allows different probable transcripts.

In the process of gene prediction for sensitivity testing using a small number of contigs, we used the web server. One can simply provide the path to the query file and name the model set organism. One could also specify options for the location of genes in specific strands and the degree of alternative splicing to allow during the prediction of the expected exons. An example of Augustus gene prediction output is shown in Fig. 4.

Augustus [result display]

Your job id is AUG-91312805.
The graphical and text results are here

Calculating ...

```
# This output was generated with AUGUSTUS (version 2.6).
# AUGUSTUS is a gene prediction tool for eukaryotes written by Mario Stanke (mario.stanke@uni-greifswald.de)
# and Oliver Keller (keller@cs.uni-goettingen.de).
# Please cite: Mario Stanke, Mark Diekhans, Robert Baertsch, David Haussler (2008),
# Using native and syntenically mapped cDNA alignments to improve de novo gene finding
# Bioinformatics 24: 637-644, doi 10.1093/bioinformatics/btn013
# No extrinsic information on sequences given.
# Initialising the parameters ...
# tribolium version. Use default transition matrix.
# Looks like /var/tmp/augustus/AUG-91312805/input.fa is in fasta format.
# We have hints for 0 sequences and for 0 of the sequences in the input set.
#
# ----- prediction on sequence number 1 (length = 3323, name = McKenna_1kb_R1PF_paired_contig_945) -----
#
# Constraints/Hints:
# (none)
# Predicted genes for sequence number 1 on both strands
# start gene g1
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        gene    1       1260    0.97    -       .       g1
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        transcript      1       1260    0.97    -       .       g1.t1
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        internal        14      176     1       -       0       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        internal        237     625     0.98    -       2       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        initial 894     1260    0.99    -       0       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        intron  1       13      1       -       .       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        intron  177     236     0.99    -       .       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        intron  626     893     1       -       .       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        CDS     14      176     1       -       0       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        CDS     237     625     0.98    -       2       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        CDS     894     1260    0.99    -       0       transcript_id "g1.t1"; gene_id "g1";
McKenna_1kb_R1PF_paired_contig_945      AUGUSTUS        start_codon     1258    1260    .       -       0       transcript_id "g1.t1"; gene_id "g1";
# coding sequence = [atgttgatattaaattgtgctaaaagcgtgtttcacgacttacatcagggtgacttacaaaccatcctcttcgtttttct
# tcggtgtgctagtgcttgttcgattgttccaaaaattgagagaactgtacatcttacccccgggtccatggggttatcccatcgtcggcagtatatgc
# tccctcaagaaagaccttcacgtccacttcaacgacctagctgctgaatatggttccctgttctccacgaggttcggcaaccagctgattgttgtgct
# gagtgactacaagatgatcagggatatattccgcagagaagagttcacaggaaggcccaacaacgaattcaccaagatcctagacggctacggtctga
# tcaacattgccgggaagttgtggaaggaacagaggaaattcgtacacgaaggtttacgtcacttcggcatgtcttacttgggaaccaagaaggctcag
# atggaaaccaggatcacaaaagaagttgaggaattcctcagaatgcttaagtccaagaggggccaaaaggtagatctcaacccttactttgccgtgtc
# catctcaaacgttatctgcgagatcttgatgagtgtaagattctctttcgacgacaaaaggtttaggaggttcatgcaattgatagatgagggattca
# gactcttcggctctctggataaggctgtcttcattcccatcatgagatacctacctggtaaatggcagacgttgaacaaaatcagacagaaccgtaac
# gaaatgggattgttcttgcaagaaaccatcgatgaacatagaaggaccttcaacagagacaacataagagacatcttggatacgtatctgcttgagat
# tgccaaggcttccgatgagggtactgaagactgtttgttccaaggaaaagaccacg]
# protein sequence = [MLILNCAKSVFHDLHQGDLQTILFVFFGVLVLVRLFQKLRELYILPPGPWGYPIVGSICSLKKDLHVHFNDLAAEYGS
# LFSTRFGNQLIVVLSDYKMIRDIFRREEFTGRPNNEFTKILDGYGLINIAGKLWKEQRKFVHEGLRHFGMSYLGTKKAQMETRITKEVEEFLRMLKSK
# RGQKVDLNPYFAVSISNVICEILMSVRFSFDDKRFRRFMQLIDEGFRLFGSLDKAVFIPIMRYLPGKWQTLNKIRQNRNEMGLFLQETIDEHRRTFNR
# DNIRDILDTYLLEIAKASDEGTEDCLFQGKDH]
# end gene g1
###
```

Fig. 4. Output format of Augustus gene prediction. The output format of Augustus gene prediction proceeds as follows. The first few lines provide the title heading information, and the subsequent lines provide information about the predicted gene from the sequences (details provided in the text). The first few lines are divided into specific columns that provide details about the predicted region. The next few lines provide information about the open reading frame (ORF) and the most probable translated protein.

In the aforementioned output of Augustus, the first column specifies the sequence name, and the second column provides the name of the program that generated the output. The third column specifies the type of hit generated, the fourth and fifth columns specify the starting and ending positions of the hit, respectively, the sixth column gives scores for the prediction, the seventh column indicates the strand on which the gene was found and the eighth column gives the reading frame defined in the GFF standard. The ninth column gives other arbitrary information and information about the string identifier. Following the former information, the sequence of the ORF after alternative splicing is given, and information about the exonic regions is provided in the form of amino acid sequences (alternate output could be specified before running Augustus). However, while running Augustus for whole-genome assembly, installation of the software locally by establishing the pipeline is recommended. While running in the pipeline mode, one could provide the input and then specify different parameters at the command prompt.

Gene Prediction Using GenScan

The GenScan gene prediction algorithm is based on probabilistic models for sequence composition and the basic structure of structural elements in genes. The algorithm computes a meaningful probability (for an event E, i.e., the probability that a particular exon is correct). This probability, P (E), is the sum of all of the probabilities in the model of all possible structural region of gene (termed as parses) that restrain the exon in the matching (accurate) reading frame. And the sum is computed in a reasonable amount of time by a "forward-backward" procedure (Rabiner et al. 1989). The aforementioned probability gives the quantitative estimate of the likelihood that an exon is true (Burset et al. 1996; Fernandez-Carvajal et al. 2006).

During gene estimation using GenScan, the query file was provided in the webpage in a manner similar to that of Augustus. The web interface of GenScan is shown in Fig.5. Other parameters, including the model organism name, suboptimal cutoff score and the format of output, were chosen during gene prediction. However, it is worth mentioning that although the model organism options are limited and only distantly related organism information related to specific higher taxonomic groups (e.g., vertebrates, *Arabidopsis*, maize) are available, the algorithm still provides very reasonable output (Issac et al. 2002).

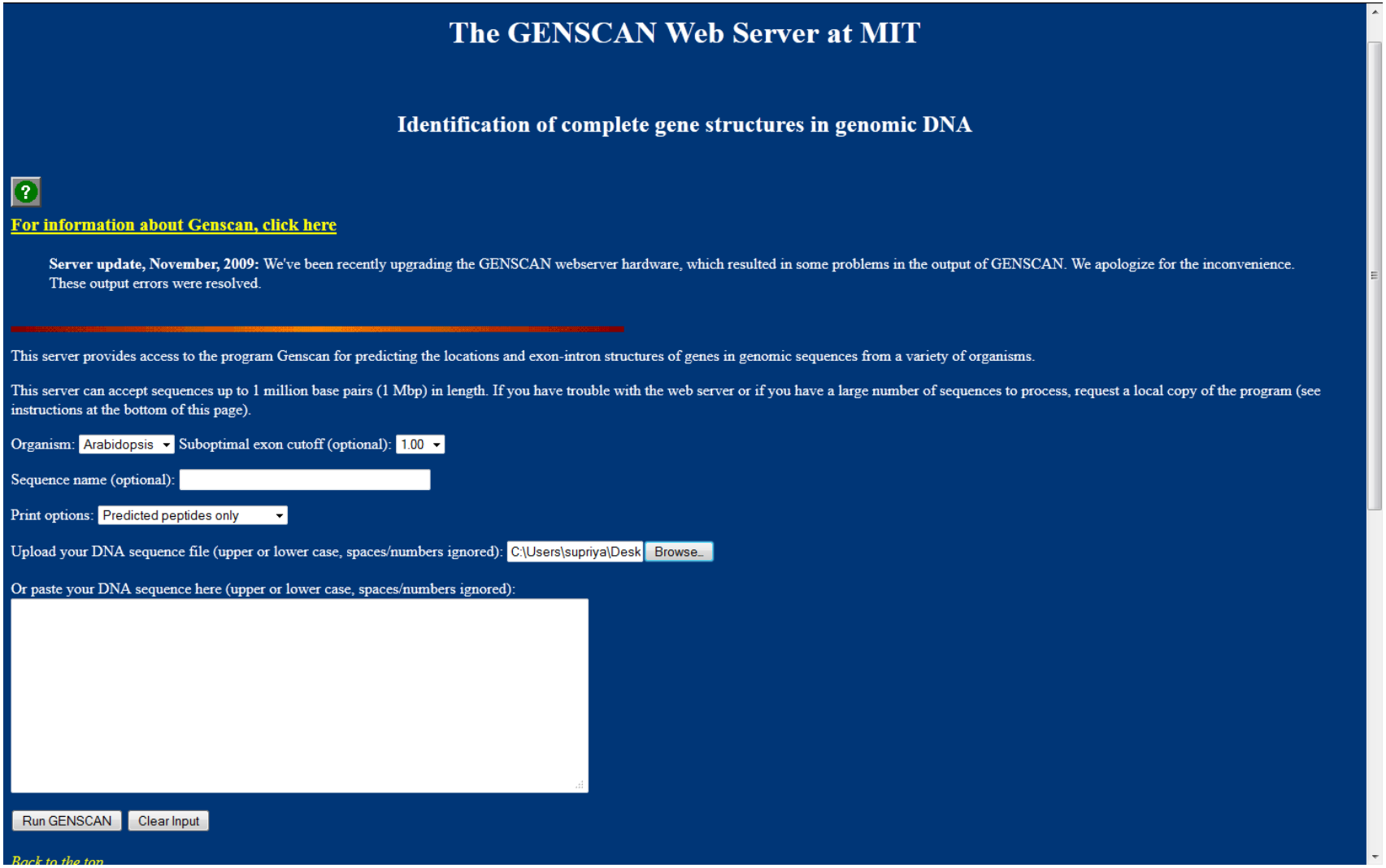
The GENSCAN Web Server at MIT
Identification of complete gene structures in genomic DNA
For information about Genscan, click here
Server update, November, 2009: We've been recently upgrading the GENSCAN webserver hardware, which resulted in some problems in the output of GENSCAN. We apologize for the inconvenience. These output errors were resolved.
This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.
This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page).
Organism: Arabidopsis Suboptimal exon cutoff (optional): 1.00
Sequence name (optional):
Print options: Predicted peptides only
Upload your DNA sequence file (upper or lower case, spaces/numbers ignored): C:\Users\supriya\Desk Browse...
Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):
Run GENSCAN
Clear Input
Back to the top

Fig. 5. Web interface of GenScan. The web interface of GenScan provides options of parameterizing the input. The organism option permits the selection of a specific group of organisms, the related parameters of which are considered for final evaluation of the sequences, with a cutoff value providing the relatedness cutoff of the query sequence to database sequences. The web interface also provides options for specific parameters for more precise output. The DNA sequence can either be uploaded as a file or pasted into the provided space.

The output of GenScan, similarly as Augustus, contains the following details,(a snapshot of the output obtain in web interface is provided in Fig.6).: the first column specifies the exon number, the second column specifies the exon type, the third column provides the strand information, the fourth and fifth columns specify the beginning and end positions of exonic sequences, respectively, the sixth column specifies the size of the exon, the seventh and eighth columns specify the reading frame and the remaining columns assign different probability scores to the sequence. After interpreting the estimate for all possible sequences, the amino acid regions are predicted for the optimal coding regions.

```
64.03 PlyA + 500846 500851    6                                    1.05


65.07 PlyA - 501348 501343    6                                    1.05
65.06 Term - 501576 501566   11  1  2   98   47     5 0.621  -0.22
65.05 Intr - 502469 501892  578  1  2   72   -4   400 0.020  25.63
65.04 Intr - 503287 503247   41  1  2  100   18    53 0.019   0.60
65.03 Intr - 505184 504909  276  2  0   77   84   117 0.480  11.99
65.02 Intr - 507018 506175  844  2  1   43   94   640 0.632  55.55
65.01 Intr - 508734 508532  203  0  2   72   98    92 0.753  10.76

Suboptimal exons with probability > 1.000

Exnum Type S .Begin ...End .Len Fr Ph B/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------



NO EXONS FOUND AT GIVEN PROBABILITY CUTOFF



Predicted peptide sequence(s):



>/tmp/06_02_12-19:30:13.fasta|GENSCAN_predicted_peptide_1|329_aa
MEKLEKCTDVKSETFEDAFIRHHINLLRVSKALQFLHRQCTEDLERIVSTSAVLVRLFQK
LRELYILPPGPWGYPIVGSICSLKKDLHVHFNDLAAEYGSLFSTRFGNQLIVVLSDYKMI
RDIFRREEFTGRPNNEFTKILDGYGLINIAGKLWKEQRKFVHEGLRHFGMSYLGTKKAQM
ETRITKEVEEFLRMLKSKRGQKVDLNPYFAVSISNVICEILMSVRFSFDDKRFRRFMQLI
DEGFRLFGSLDKAVFIPIMRYLPGKWQTLNKIRQNRNEMGLFLQETIDEHRRTFNRDNIR
DILDTYLLEIAKASDEGTEDCLFQGKDHX
```

Fig. 6. Output format of GenScan. The output format for gene prediction by GenScan is slightly different from that of Augustus. In GenScan, the output amino acid sequences obtained from the entire batch of sequences are represented together rather than separately with specific ORF and contig annotation.

Gene prediction using GeneMark is performed following the same protocol as Augustus and GenScan. The output page also displays the starting and ending positions of the coding sequence in a similar manner. In the process of gene prediction using Augustus, GenScan and GeneMark, an online web server is used; however, these pipelines could be installed on a local server.

In the process of gene prediction by automated annotation, the query sequences are provided in a web server, and the exonic sequences of proteins are collected. Based on the number of true exons generated in contrast to the number of false positives, one can compute the sensitivity of the software. For gene prediction using Fgenesh, Fgenes and NetBLASTx, the pipelines were established using MolQuest. Fgenesh and Fgenes are available from Softberry Inc. (Mount Kisco, NY) and operate using different algorithms. NetBLASTx works on a simple BLAST algorithm (Altschul et al. 1990).

Gene Prediction Using Fgenesh and Fgenes

Fgenesh is the fastest (50–100-fold faster than GenScan) and most accurate gene finder available. Fgenesh was well established for gene annotation by its use in the rice and *Drosophila* genome projects, and it has been described as "the most successful (gene finding) program" (Yu et al. 2002). For example, it was used to detect 87% of all high-evidence predicted genes in the genome of rice (Goff et al. 2002).

Fgenesh is an HMM-based algorithm with the parameters trained on the set of organism-specific genes annotated in GenBank (Benson et al. 1999). For the accuracy of

gene prediction on different levels, several schemes are designed to predict an unambiguous set of genes; for example, in the *Drosophila melanogaster* genome, several gene sets are generated, such as annotation CGG1, again based on a set of reliable exons, annotation CGG2 and finally the most complete set of exons, annotation CGG3.

Fgenesh+ is a variant of Fgenesh that considers some information about similar proteins. Fgenes is based on discriminate functions trained to predict human genes, but it also efficiently predicts other genes. It utilizes pattern recognition to detect different types of exons, promoters and polyA signals. The Fgenesh and Fgenes output structures are similar and have the following elements in the output.

The output of these programs uses the following abbreviations: G, predicted gene number across the entire length of the gene; Str, DNA strand (+ or −); Feature, type of coding sequence; CDSf, first (start codon); CDSi, internal (internal coding region); CDSl, last coding segment; TSS: position of transcription start (TATA box); TSS, position of transcription start site; TATA, TATA box; wTATA – discriminate function score for the TATA box; and ORF, start/end positions of the ORF where the first complete codon starts and the last codon ends. The output consists of columns for gene number (G), stand, the feature of the structure, start and ending positions of a gene, weight score computed though specific algorithm and the ORF start and ending positions of the specific gene. Following the aforementioned columns, information about the predicted proteins is given. An example of Fgenesh output is given in Fig. 7.

```
FGENESH 2.6 Prediction of potential genes in Tribolium_castaneum genomic DNA
 Time     :   Sat May 05 10:14:16 2012
 Seq name: McKenna_1kb_R1PF_paired_contig_945
 Length of sequence: 3323
 Number of predicted genes 1: in +chain 0, in -chain 1.
 Number of predicted exons 3: in +chain 0, in -chain 3.
 Positions of predicted genes and exons: Variant   1 from   1, Score:86.128058
   G Str   Feature   Start        End    Score           ORF           Len

   1 -    1 CDSi        14 -       176   23.33        15 -       176    162
   1 -    2 CDSi       237 -       625   31.83       237 -       623    387
   1 -    3 CDSf       894 -      1260   42.01       895 -      1260    366
   1 -      TSS       1325              -3.14

Predicted protein(s):
>FGENESH:   1   3 exon (s)     14  -   1260   306 aa, chain -
MLILNCAKSVFHDLHQGDLQTILFVFFGVLVLVRLFQKLRELYILPPGPWGYPIVGSICS
LKKDLHVHFNDLAAEYGSLFSTRFGNQLIVVLSDYKMIRDIFRREEFTGRPNNEFTKILD
GYGLINIAGKLWKEQRKFVHEGLRHFGMSYLGTKKAQMETRITKEVEEFLRMLKSKRGQK
VDLNPYFAVSISNVICEILMSVRFSFDDKRFRRFMQLIDEGFRLFGSLDKAVFIPIMRYL
PGKWQTLNKIRQNRNEMGLFLQETIDEHRRTFNRDNIRDILDTYLLEIAKASDEGTEDCL
FQGKDH
//
```

Fig. 7. Output of Fgenesh gene prediction. The output, similar to Augustus and GenScan, represents different predicted structural regions of a specific gene from different chains of proteins followed by the complete predicted exonic protein sequence.

Gene Prediction Using NetBLAST through MolQuest

NetBLASTx is an extension of the BLASTx program that helps in sequence search in different publicly available databases. It performs BLAST searches of the nucleotide query sequence against all possible matches in the online NCBI database. The NetBLASTx feature is provided in the Softberry MolQuest package (a snapshot of softberry interface is provided in Fig.8.) A nucleotide or protein sequence sent to the BLAST server is compared against databases at the NCBI, and a summary of matches is returned as output. The original web server for BLAST is available through the home page of NCBI (www.ncbi.nlm.nih.gov).

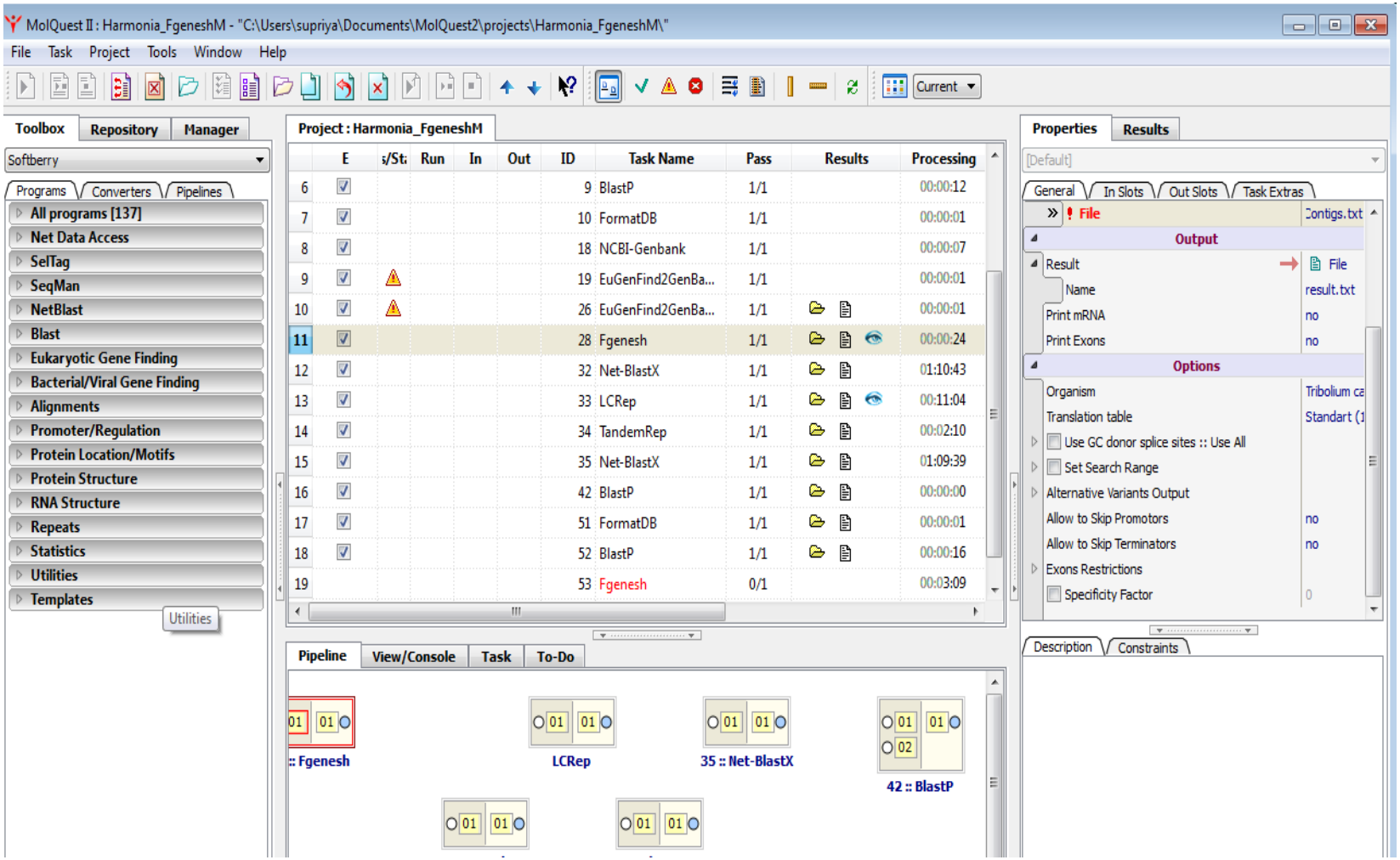

Fig. 8. Softberry interface. Softberry provides the MolQuest interface for eukaryotic gene prediction by generating a pipeline for different software (such as Fgenes, Fgenesh and Fgenesh+).

In the earlier protocol for detecting the 5′-end region exonic sequences in the *T. castaneum* LGX chromosome, the major drawback was the large number of false-positive sequences. In contrast to the former study, the output obtained by CYP gene detection through BLASTp provided a very nice estimation to avoid false positives. The exonic sequences with high similarity and E-values less than 0.001 compared to manually annotated sequences are found to give a good estimate of exact exonic sequences, generally avoiding false positives.

Visualization of the Predicted Genes

Softberry provides a visualization package through MolQuest that has graphic features to adjust the scale and range for visualization (a snapshot of web interface is provided in Fig.9.) It also provides 2 petition displays in which the left side window provides the summary of the structural information and right side window provides the complete display of the exonic region of gene predicted in the specific range. The estimated output from each program is presented in the Appendix, and a summary of the output is provided in Table A.1.2.

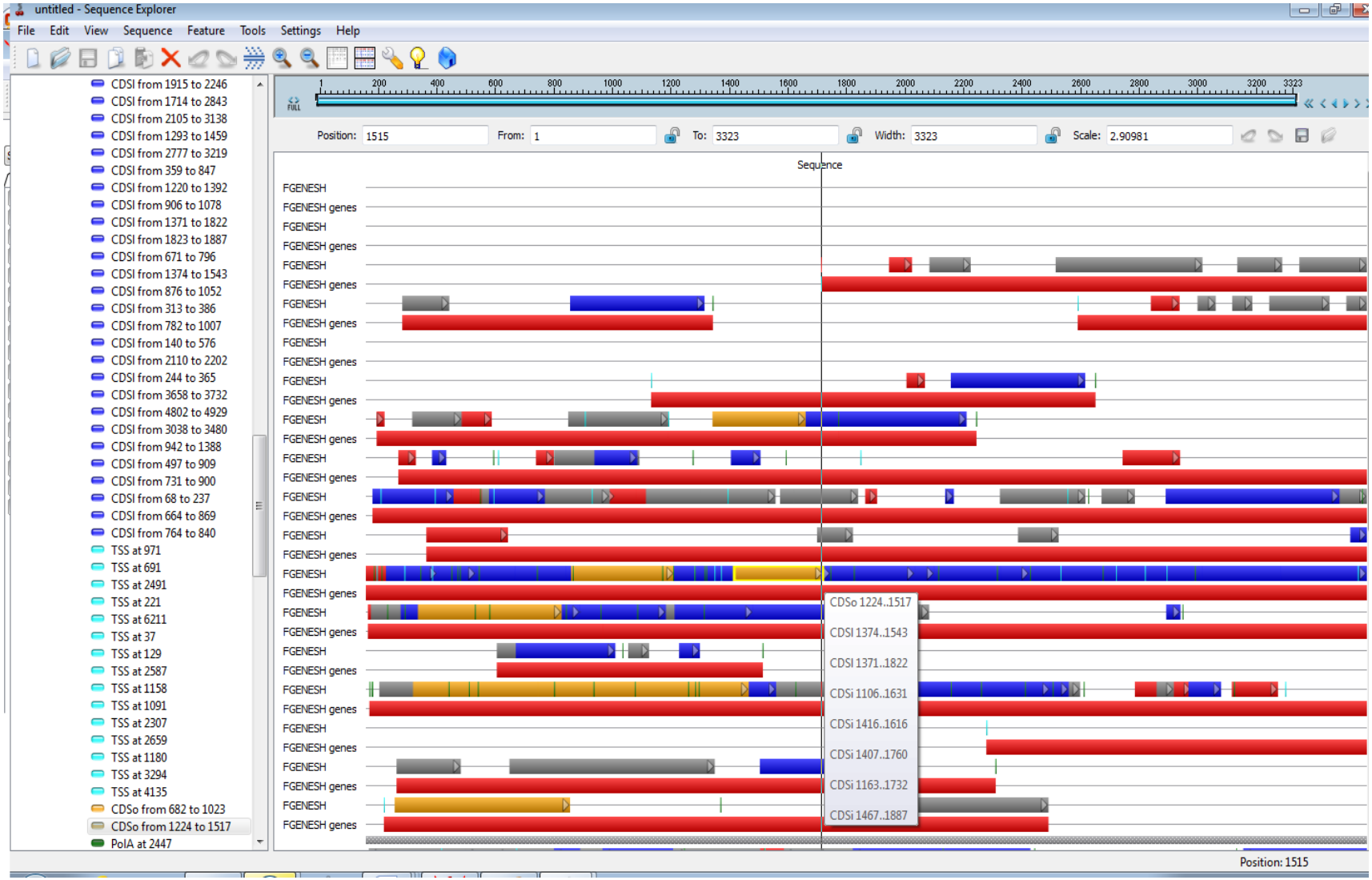
untitled - Sequence Explorer
File Edit View Sequence Feature Tools Settings Help
CDSI from 1915 to 2246
CDSI from 1714 to 2843
CDSI from 2105 to 3138
CDSI from 1293 to 1459
CDSI from 2777 to 3219
CDSI from 359 to 847
CDSI from 1220 to 1392
CDSI from 906 to 1078
CDSI from 1371 to 1822
CDSI from 1823 to 1887
CDSI from 671 to 796
CDSI from 1374 to 1543
CDSI from 876 to 1052
CDSI from 313 to 386
CDSI from 782 to 1007
CDSI from 140 to 576
CDSI from 2110 to 2202
CDSI from 244 to 365
CDSI from 3658 to 3732
CDSI from 4802 to 4929
CDSI from 3038 to 3480
CDSI from 942 to 1388
CDSI from 497 to 909
CDSI from 731 to 900
CDSI from 68 to 237
CDSI from 664 to 869
CDSI from 764 to 840
TSS at 971
TSS at 691
TSS at 2491
TSS at 221
TSS at 6211
TSS at 37
TSS at 129
TSS at 2587
TSS at 1158
TSS at 1091
TSS at 2307
TSS at 2659
TSS at 1180
TSS at 3294
TSS at 4135
CDSo from 682 to 1023
CDSo from 1224 to 1517
PolA at 2447
1 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3323
Position: 1515 From: 1 To: 3323 Width: 3323 Scale: 2.90981
Sequence
FGENESH
FGENESH genes
CDSo 1224..1517
CDSi 1374..1543
CDSi 1371..1822
CDSi 1106..1631
CDSi 1416..1616
CDSi 1407..1760
CDSi 1163..1732
CDSi 1467..1887
Position: 1515

Fig. 9. Visualization of different structural regions in specific predicted sequences. The left side partition window displays the name of the structural elements that is displayed in detail in the right side window. Both the right side partition window and left side partition window contain interactive objects that display the number of structural elements associated with each contig, helping one to interpret the protein structure in detail.

CYP Sequence Estimation Using the Aforementioned Software and Web Search

Analysis of *T. castaneum* LGX gene prediction was performed using several gene prediction tools. These tools considered nucleotide sequences to predict the suitable coding region and exonic fragments. We selected tools with high sensitivity for further analysis in *de novo* studies. In our CYP gene annotation study using different highly efficient tools, we used the protein sequence as a reference instead of the nucleotide sequence. The manually annotated sequences are used as a reference to estimate the efficiency of the automated annotation tools. The sequence identity is compared at the level of proteins, opposed to genomic sequences, to make the process a bit more focused on comparative analysis for gene finding (in contrast to sequence identity and comparison of genomic sequence). In addition, many ambiguities (such as the wobble codon) could be avoided by using amino acid sequences. Hence, this is a widely used approach. Table 4.3 shows the summary information for gene prediction using different software and web searches and the software vs. sensitivity plot presented in Fig.10, in detecting exonic regions and their fragments of CYP genes.

Table 4.3 Estimate of the total number of exonic regions from the 198 assembled potential CYP contig sequences from *Harmonia axyridis* predicted by each software and NCBI BLAST on different publicly available protein databases. The gene prediction software used included Augustus, Fgenesh (pipeline), Fgenes (pipeline), GeneMark, NCBI BLAST at Swiss-Prot/TrEMBL, the combined output of sensitivity predicted by Fgenesh and Fgenes and the combined output of Fgenesh, Fgenes and BLAST search output from Swiss-Prot/TrEMBL [sensitivity = (number of true positive exons obtained/total number of exons)].

Swiss-Prot/TrEMBL output estimate	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	NCBI BLAST at Swiss-Prot/TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
Sum	60	123	135	60	30	75	160	172
Sensitivity	0.337	0.691	0.758	0.337	0.169	0.421	0.899	0.966292

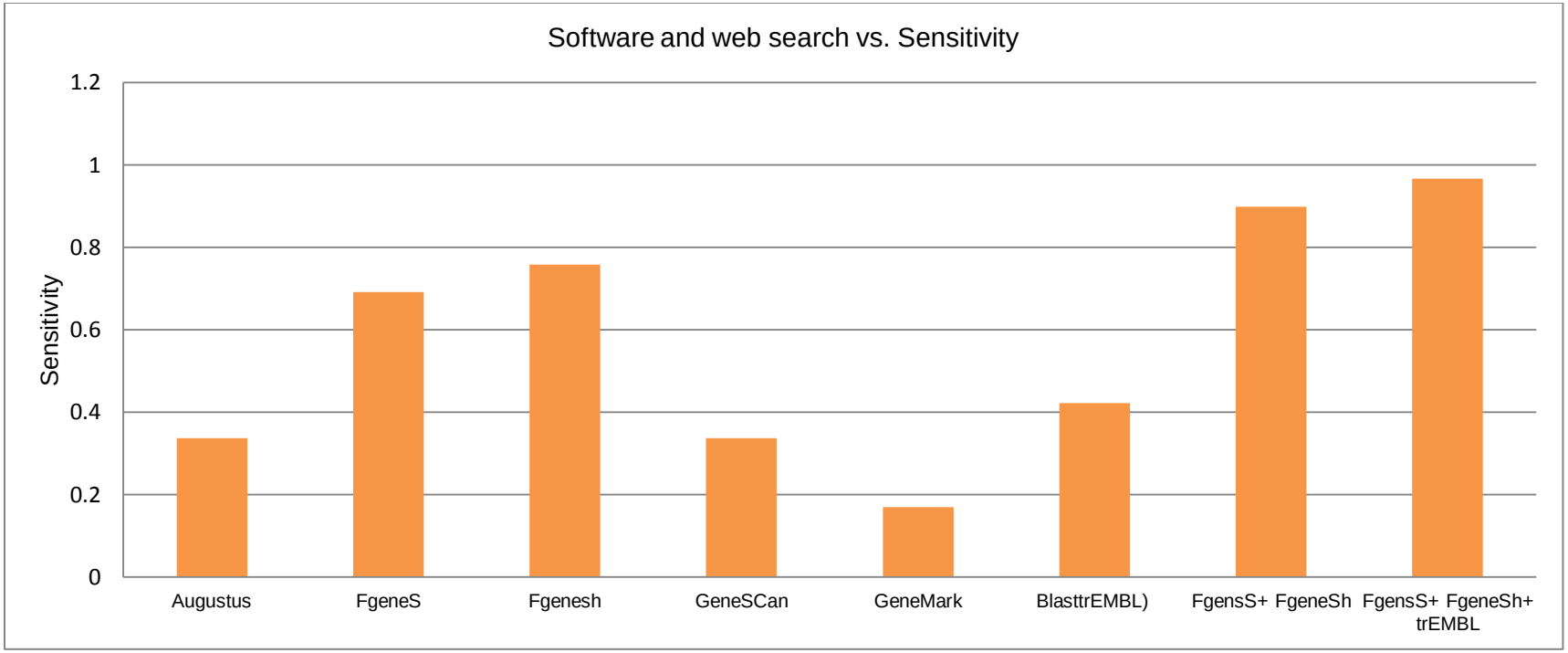

Fig. 10. Bar graph showing the summary of the software vs. sensitivity study for gene prediction. The gene prediction software included in this study were Augustus, Fgenesh (pipeline), Fgenes (pipeline), GeneMark, NCBI BLAST at Swiss-Prot/TrEMBL, the combined output of sensitivity predicted by Fgenesh and Fgenes and the combined output of Fgenesh, Fgenes and the BLAST search output from Swiss-Prot/TrEMBL. The final combined output of Fgenesh, Fgenes and Swiss-Prot/TrEMBL exhibited the highest sensitivity [sensitivity = (number of true positive exons obtained/total number of exons)].

Repeat Masking

In genome annotation, some of the most pronounced ambiguities arise due to the presence of repeated regions such as Long Interspersed nuclear Elements (LINEs), and Short Interspersed nuclear Elements (SINEs) are often predicted as genes by different software tools. To avoid these problem of associated with repetitive elements, software such as RepeatMasker masks repeats by using specific symbols. Even BLAST programs and gene prediction algorithms accept sequences containing Ns or specific symbols. To analyze large amounts of data analysis in different metagenomic projects, screening repeats using BLASTn, tBLASTn or tBLASTx is also a good choice (Claverie et al. 1994). Additionally, repeats can be screened after gene prediction by contrasting them against a database of repeat sequences (translated in all 6 frames for protein comparison). Programs such as RepeatMasker may also delete simple sequence repeats (microsatellites) such as AT repeats and CAG repeats, the latter of which codes for poly-glutamine in certain triplet repeat disease genes. Hence, in some cases, repeat masking is avoided.

While analyzing the data of 189 potential contigs from *H. axyridis* using RepeatMasker, we found that it masked contig 'McKenna_9346.' The contig is a fragment of a real gene found in the estimate of gene prediction by the traditional gene annotation process. Hence, we decided not to use gene masking before gene prediction. When we screened the entire genome using Fgenesh without repeat masking, we obtained 97,704 gene fragments (until this point), which we are still studying to remove false

positives and to find the true exon fragments of genes and their exact number.

Phylogenetic Analysis of CYP Sequences

The assembled CYP genes with sequence lengths exceeding ~300–350 amino acids long were further considered for the comparative analysis study against all CYP sequences of *T. castaneum*. Pairwise comparisons of single-copy orthologs in insect genomes exhibited an average protein identity conservation of 45–95%. Although the most conserved regions of the CYP proteomes have been identified, the identity distributions in each pair or class of organisms are relatively different and vary between various functional gene classes. In this study, the CYP genes from both species were collected and aligned using MAFFT multiple sequence alignment software. The resulting alignments were then manually adjusted.

Sequence Alignment and Molecular Phylogenetic Analyses

The *H. axyridis* and *T. castaneum* CYP amino acid sequences were aligned using the E-INS-i algorithm in MAFFT version 6.8 (Katoh et al. 2008). The alignments were manually adjusted in Mesquite 2.75+ (Maddison et al. 2008). The aligned amino acid sequence matrix was analyzed under Bayesian inference using the fixed rate WAG substitution model (Wheland et al. 2001) in MrBayes version 3.1.2 (Huelsenbeck et al. 2001; Ronquist et al. 2003). Two independent runs were executed in MrBayes, each with 4 chains run for 3×10^6 generations. Trees were sampled every 100 generations. The runs

converged (standard deviation of split frequencies below 0.05) by 2×10^6 generations. All post-convergence trees (1000 trees from each of the 2 independent runs) were used to construct a 50% majority rule consensus tree in Mesquite showing Bayesian posterior probabilities for each node. This tree was exported from Mesquite as a circle tree and midpoint-rooted in Figtree version v1.3 (http://groups.google.com/group/figtree-announce/browse_thread/thread/dea093d866dddbe8).

Results and Discussion

Traditional Gene Annotation

Traditional gene annotation or gene discovery and the curation of the CYPs in the *H. axyridis* v1.1 draft genome sequence assembly (February 2012) produced 94 CYP genes and 3 apparent pseudogenes. *T. castaneum* has 137 CYP genes and 10 pseudogenes. Other insect genomes have different number of CYPs, e.g., *A. mellifera* (46 CYPs), *D. melanogaster* (86 CYPs), *Anopheles gambiae* (105 CYPs), *Aides aegypti* (158 CYPs) and *Acyrthosiphon pisum* (67 CYPs).

The *H. axyridis* CYPs can be divided into 4 distinct clans that are present in all other insects that have been studied: Mito, CYP2, CYP3 and CYP4 (Feyereisen et al. 2006). The 4 clans can be further subdivided into 18 families, at least 43 subfamilies and 3 possible pseudo-genes (CYP9BD1P, CYP9BE7P and CYP301A N-terminal fragment). The sequences are mostly in fragments, and some pseudogenes may not be accounted for among these fragments. The phylogenetic tree resulting from a Bayesian analysis of the *H. axyridis* and *T. castaneum* CYP genes (Fig. 11.1) was used in combination with the previously described rules for naming CYPs (40% identity for family, 55% identity for subfamily) to name the *H. axyridis* CYPs (Nelson et al. 2006).

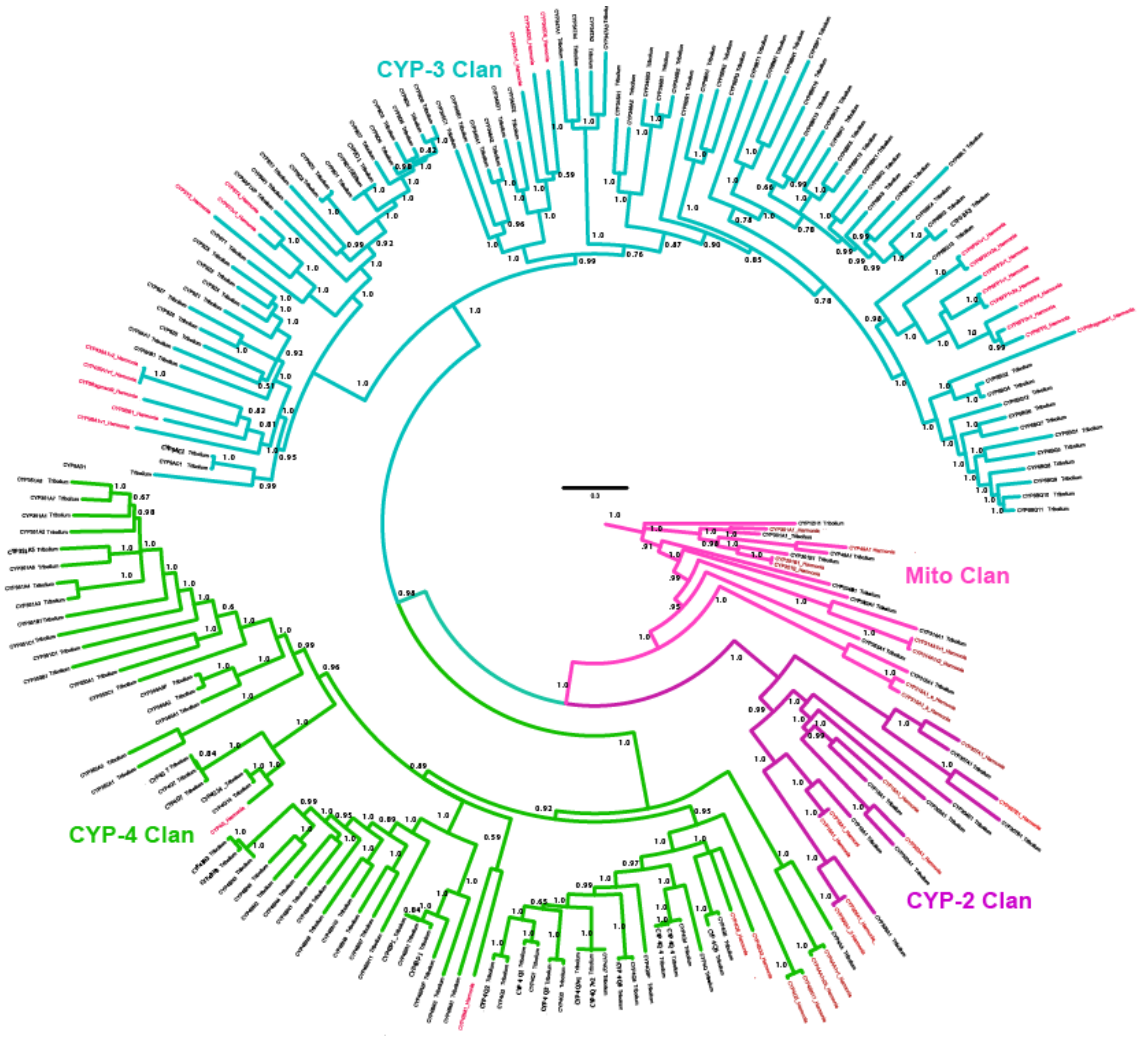

Fig. 11.1 Fifty percent majority rule consensus tree with midpoint rooting resulting from the Bayesian analysis of complete or nearly complete CYP genes (amino acid sequences) from *Harmonia axyridis* and *Tribolium castaneum*.

Numbers adjacent to the nodes are posterior probabilities, and they indicate the proportion of trees sampled in the Bayesian analysis that included the node. Branch lengths are proportional to the number of substitutions per site, and colors indicate the 4 CYP clans. *H. axyridis* sequences are indicated by red sequence names. Black sequence names indicate CYPs from *T. castaneum.* We reconstructed a phylogenetic tree for *H. axyridis* and *T. castaneum* that included all *H. axyridis* sequences that are at least 300–

350 amino acids long (out of ~500 amino acids) together with all *T. castaneum* sequences. This tree included 42 (less than 50% of the 94 CYP sequences) CYP sequences (Fig. 11.2). This tree is also available as a PDF (Appendix), and it is divided into 3 parts, each of which has been individually enlarged (Figs. 11.2, 11.3 and 11.4). A summary of the annotation obtained by annotating CYP genes to different CYP families is presented in Fig. 12.

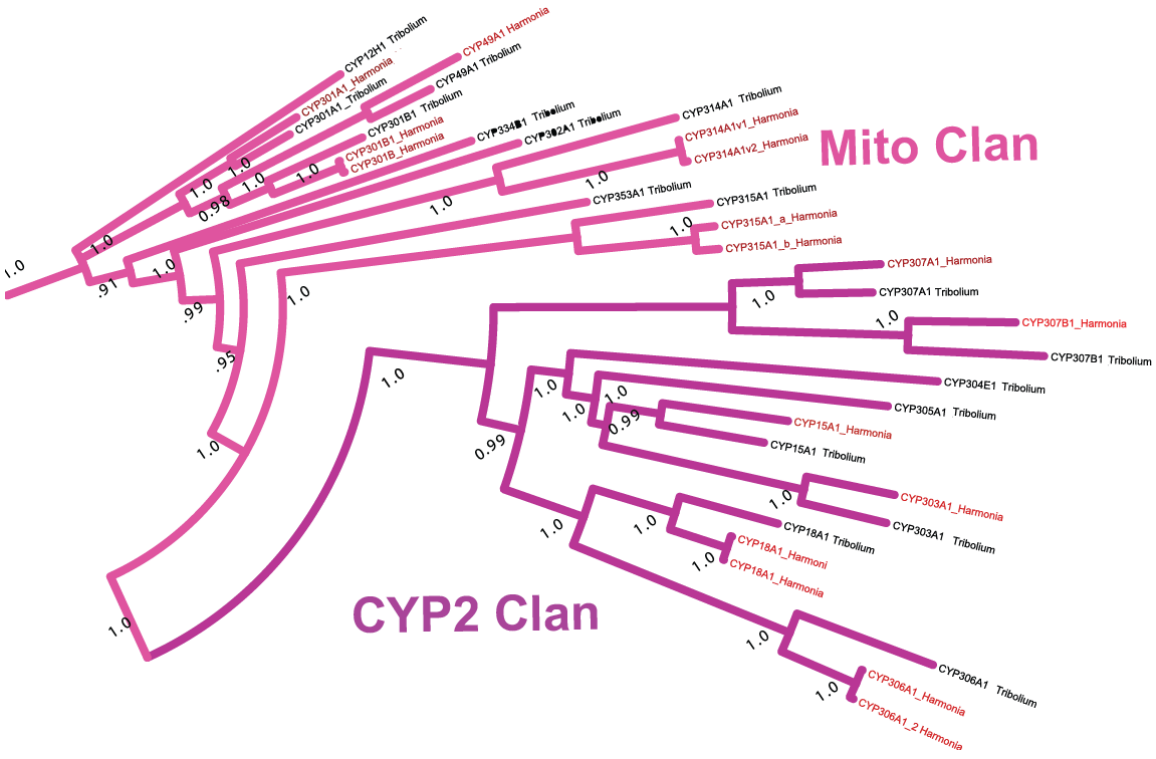

Fig. 11.2. Enlarged region of the 50% majority consensus tree from the Bayesian analysis of complete and nearly complete *Harmonia axyridis* and *Tribolium castaneum* CYPs showing members of the Mito and CYP 2 clans as leaves of the tree, and the nodal support obtained from Bayesian posterior probability is represented in the CYP tree as numbers with *H. axyridis* CYP genes presented in red and *T. castaneum* CYP genes presented in black.

Each of the 4 CYP clans is monophyletic in this tree (posterior probabilities of 1.0 in Fig. 11.2.) excluding the Mito clan. The Mito clan is the sister clan of the CYP2 clan, and the CYP3 clan is the sister clan of the CYP4 clan, consistent with other studies (Baldwin et al. 2009).

Typically, the Mito clan is monophyletic. In this tree, it is not, but that may be caused by the use of partial sequences from *H. axyridis* mixed with full-length sequences of *T. castaneum*. The tree did have some earlier versions showing the monophyletic to the Mito clan. Within each of the 4 clans, relationships are generally well resolved and consistent with accepted views on the inter-relationships between various CYP genes. It is apparent from this tree that several CYPs have undergone "blooms" in *T. castaneum* that are lacking in *H. axyridis* (that is specifically discussed in detail for each clan later in this section). However, because only complete or nearly complete *H. axyridis* sequences were included in this analysis, the absence of individual CYP genes from *H. axyridis* in the phylogeny does not necessarily mean that *H. axyridis* does not carry these genes. The major branches include the Mito, CYP2, CYP3 and CYP4 clans. These 4 clans include 18 families and a minimum of 43 subfamilies (See Fig.12.) In *H. axyridis*, most of the CYP genes recovered from the *H. axyridis* genome were incomplete and smaller than our 300-amino acid cutoff (see above) for inclusion in these analyses.

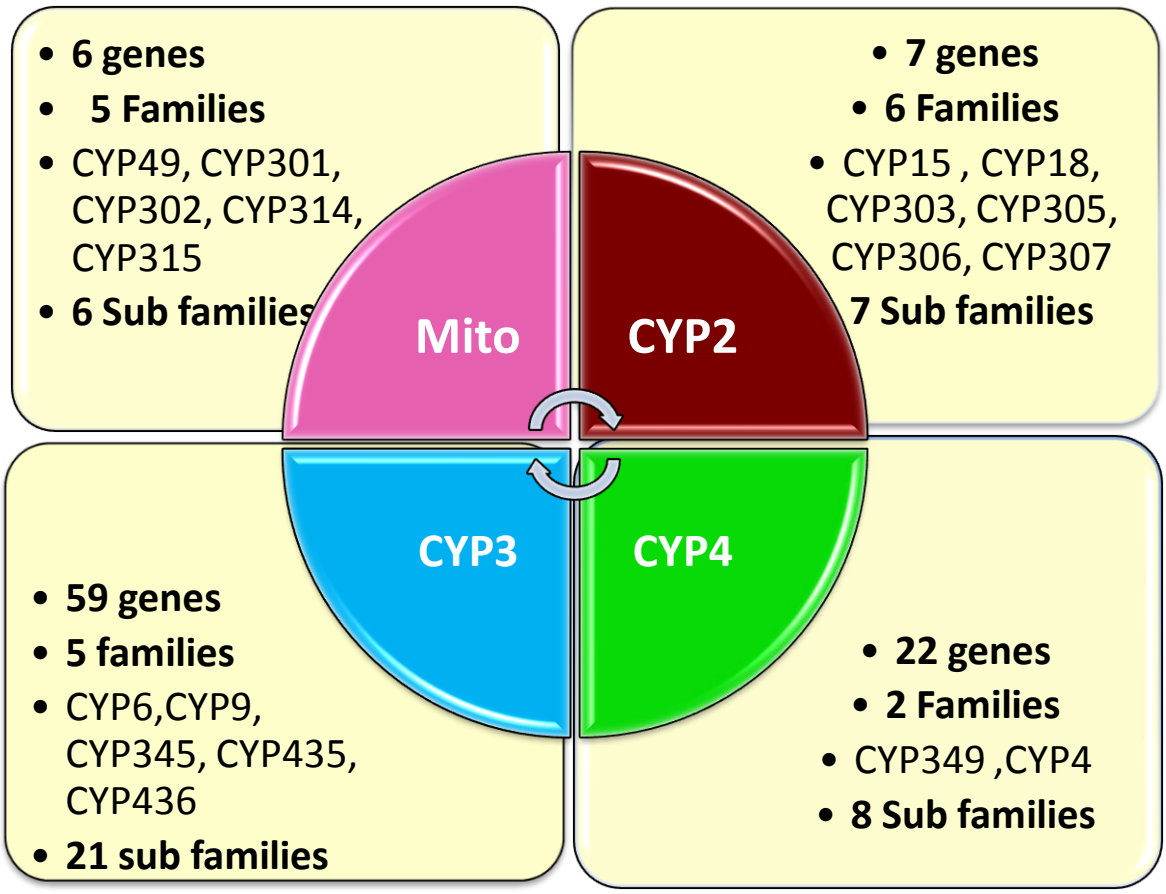

Fig. 12. Summary of the output obtained by annotating different CYP genes. CYP clans are presented as linked parts of a quarter-circle, and the text adjacent to the central circular clan region indicates the number of family members in each clan followed by the name of individual family members and the number of genes and subfamilies predicted in a specific clan in the *H. axyridis* genome (a detailed table is provided in appendix Table A.1.3 with all of the predicted and assembled contigs).

The Mito clan of *H. axyridis* contains 6 members in 5 families and 6 (see Fig. 12.) subfamilies. Three are highly conserved Halloween genes involved in ecdysone synthesis (Rewitz et al. 2006, 2008). Genes such as *Disembodied* (CYP302A1, *Dib*; *H. axyridis* CYP302, not shown in the tree), *shade* (CYP314A1, *Shd*), and *shadow* (CYP315A1, *Sad*) are the Mito CYPs involved in the biosynthesis of 20-hydroxyecdysone (20-HE). CYP314A1 is the CYP that mediates the conversion of ecdysone to 20-HE (Rewitz et al.

2006, 2008). The remaining 3 CYPs are divided into 2 new families CYP301 (CYP301A1, CYP301B1) and CYP49A1 have specific pathway regulators and new functions. Most of the genes are in fragments, and thus, the number of pseudo genes in *H. axyridis* remains uncertain. CYP301A1 is involved in cuticle formation (Willingham et al. 2004). CYP49 has specific functions during the early larval stage (see http://flybase.org/reports/FBgn0033524.html). All Mito clan members have 1:1 orthologs in *T. castaneum* that are also members of the Mito clan. However, we recovered CYP302A1 in *H. axyridis* only when we rescanned the database for all possible CYP matches using a more relaxed E- value.

The genes found in the CYP2 clan include 7 gene members, with some of the genes belonging to the Halloween family of genes. The members of the CYP2 clan are divided into 6 different families (CYP15, CYP18, CYP303, CYP305, CYP306 and CYP307). Three genes present in this clan such as CYP306 (*Phantom*) and CYP307 (*Spook*), including both CYP307A1 and CYP307B1; (Rewitz et al. 2006, 2007) are Halloween genes. The functions of the CYP307 and CYP306 genes families are not exactly known, but these genes are known to participate in regulating the biosynthesis of ecdysone from cholesterol. They are also members of the Black Box regulating early oxidation reactions of 20E. CYP305 (http://agris.fao.org/openagris/search.do?recordID=CN2010001174) members play a role in the detoxification of xenobiotics. CYP15A1 (Dwyer et al. 2011) has a specific role during embryogenesis, and CYP18A1 (Guittard et al. 2011) is a negative regulator of metamorphosis. The CYP304 gene (present in *T. castaneum*) is absent in *H. axyridis*. All

other members of the CYP2 clan in *H. axyridis* and *T. castaneum* exhibit 1:1 correspondence with each other, even at the subfamily level.

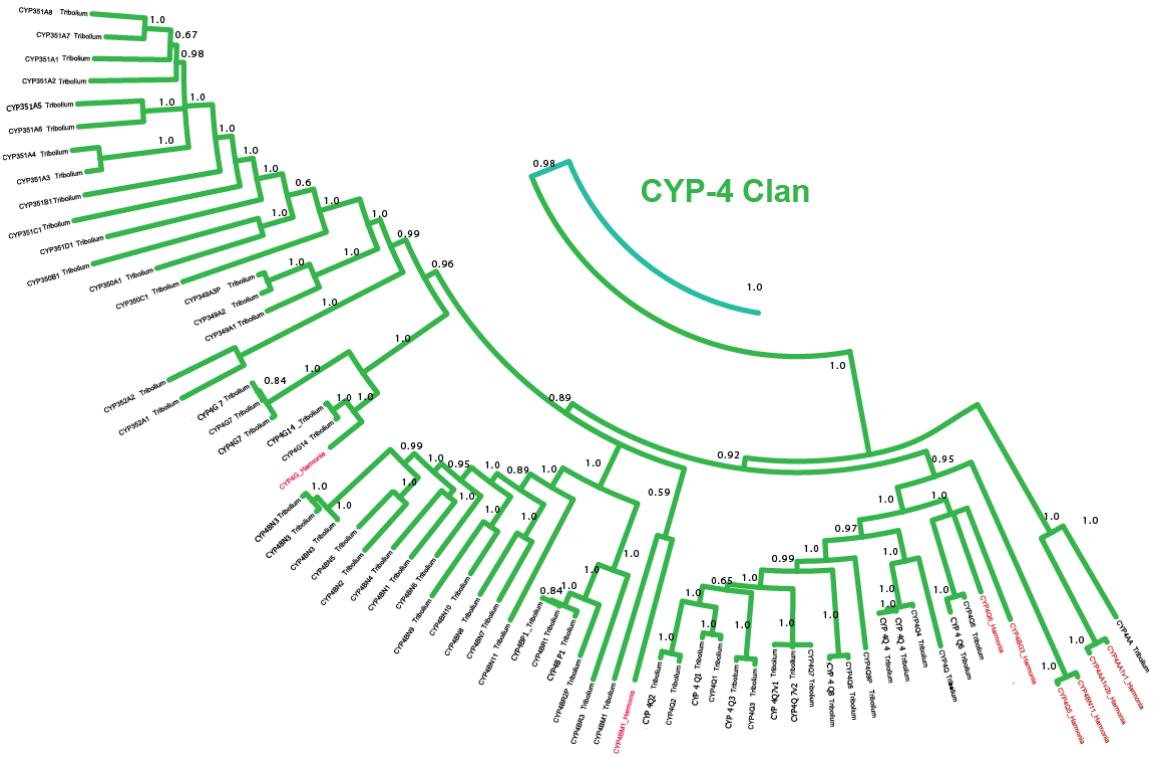

Fig. 11.3. Enlarged region of the 50% majority consensus tree from the Bayesian analysis of complete and nearly complete *Harmonia axyridis* and *Tribolium castaneum* CYPs showing members of the CYP4 clan as leaves of the tree, and the nodal support obtained from Bayesian posterior probabilities is represented in the CYP trees as numbers, with *H. axyridis* CYP genes presented in red and *T. castaneum* CYP genes presented in black.

The CYP4 and CYP3 clans members are monophyletic, and these clans are sister clans. The CYP4 clan in *H. axyridis* includes 2 distinct families (CYP349, CYP4) with 22 genes and 8 subfamilies see Fig. 11.3 and Fig.12. (sequence fragments are not assigned to specific families; further information is needed to assemble them). Specific members of the CYP4AA1 subfamily have known functions in xenobiotic resistance,

chemosensory function, ecdysone biosynthesis and pheromone metabolism (Oakeshott et al. 2010). CYPAA1 was expanded in *H. axyridis* in comparison to *T. castaneum*, and it may have some expanded functions. CYP4Q is involved in xenobiotic detoxification (http://pubs.aic.ca/doi/pdfplus/10.4141/CJPS07001), and members of CYP4G have been associated with different allelochemical detoxification phenotypes (Pedro et al. 2012). However, no literature evidence is available to assign function to CYP349 members.

The presence of CYP4G in *H. axyridis* indicates that it has a natural resistance to xenobiotics such as pyrethroids (Pridgeon et al. 2003; Yang et al. 2006; Guo et al. 2010; Karatolos et al. 2012; Martinez-Paz et al. 2012). When compared with *T. castaneum*, the CYP4 clan family CYP348A1 is absent in *H. axyridis*; there is no literature evidence of the function of these genes. Both *H. axyridis* and *T. castaneum* have CYP349A1 and CYP349A2 in common, but *T. castaneum* has an extra gene, CYP349A3 (in addition, *H. axyridis* has many small fragments of CYP349 that are not assigned to any family). Additionally, a few subfamily members of CYP350, CYP351, CYP352, CYP4BN and CYP4BR are present in *T. castaneum* and absent from the current *H. axyridis* assembly. However, *H. axyridis* carries the CYP4AW and CYP4DW subfamilies, which are absent in *T. castaneum*. Other than the aforementioned families, all other subfamilies have 1:1 correspondence to the same genes in *T. castaneum*.

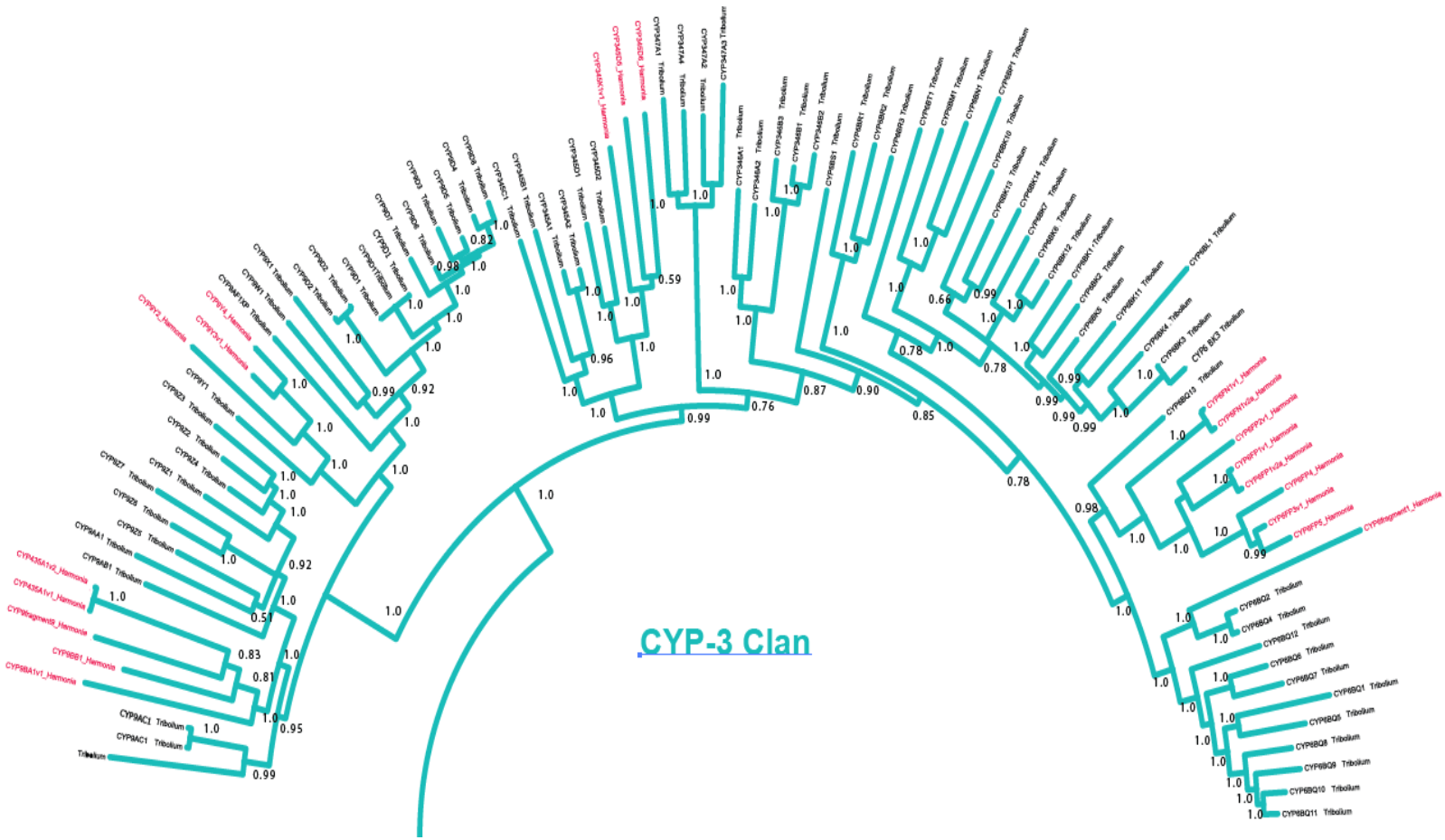

Fig. 11.4 Enlarged region of 50% majority consensus tree from the Bayesian analysis of complete and nearly complete *Harmonia axyridis* and *Tribolium castaneum* CYPs showing members of the CYP3 clan as leaves of the tree, and the nodal support obtained from Bayesian posterior probabilities are represented in the CYP tree as numbers, with *H. axyridis* CYP genes presented in red and *T. castaneum* CYP genes presented in black.

The monophyletic CYP3 clan has 5 major family members in *H. axyridis*: CYP345, CYP435, CYP436, CYP6 and CYP9 (see Fig. 11.4. and Fig.12.). The members of CYP3 consist mostly of genes responsible for the detoxification of xenobiotics and other endobiotics (Strode et al. 2008). Some CYP3 members are activated by hormones such as ecdysone (Le Goff et al. 2006) and regulated by different detoxification pathways (Baldwin et al. 1994). The 4 major families in CYP3 are classified into approximately 59 genes.

CYP345 is involved in insecticide resistance (Jiang et al. 2008). In *H. axyridis*, we found CYP345D4, CYP345D5, CYP345D6, CYP345D7, CYP345K1v1, CYP345K1v2, CYP 345K2, CYP 345K3v1, CYP 345K3v2, CYP 345K4v1, CYP 345K4v2, CYP 345 new, CYP345 fragment 1, CYP435A1v1 and CYP435A1v2. To date there is no specific phenotype correlation assigned to CYP435 and CYP436 members. CYP9 members display xenobiotic tolerance proportional to the time and concentration of xenobiotics (Stevens et al. 2000). The members of CYP9 found in *H. axyridis* are CYP9Y2, CYP9Y3v1, CYP9Y3v2, CYP9Y4, CYP9BA1v1, CYP9BA2v1, CYP9BA2v2, CYP9BA1v2, CYP9BA3, CYP9BA4, CYP9BA5, CYP9BB1, CYP9BC1, CYP9BD1P, CYP9BE1, CYP9BE2, CYP9BE3, CYP9BE4v1, CYP9BE4v2, CYP9BE5v1, CYP9BE5v2, CYP9BE6v1, CYP9BE6v2, CYP9BE7P, CYP9BF1, CYP9BG1v1, CYP9BG1v2 and CYP9 fragment 1. Members of the CYP6 and CYP9 subfamilies, participate in tolerance to different xenobiotics. We found the following CYP genes (partially complete and in fragments up to this point): CYP6BS3av1, CYP6BS3av2, CYP6BS3bv1, CYP6BS3bv2, CYP6BS3c, CYP6CR fragment, CYP6FN1v1, CYP6FN1v2a, CYP6FN1v2b, CYP6FN2, CYP6FN fragment 1, CYP6FP1v1, CYP6FP1v2, CYP6FP2v1, CYP6FP2v2, CYP6FP3v1, CYP6FP3v2, CYP6FP4, CYP6FP5, CYP6FP6, CYP6FP fragment 1, CYP6FP fragment 4v1,

CYP6FP fragment 4v2, CYP6FP fragment 5v1, CYP6FP fragment 5v2, CYP6FP fragment 6, CYP6FQ1 CYP6 fragment 8v1 and CYP6 fragment 8v2. Specifically, CYP6BS metabolizes allelochemicals (Zhang et al. 2011a) and also helps to induce insecticide resistance. *H. axyridis* lacks CYP334, CYP346, CYP347, CYP348, CYP6BK1, CYP6BL1, CYP6BQ9, CYP6BR1, CYP9D1 and CYP9Y1, which are present in *T. castaneum*. In addition, many subfamilies that are expanded in *H. axyridis* are absent in *T. castaneum*, such as CYP435, CYP436, CYP6BQ1, CYP6BR2, CYP6BT1 and CYP9Z2. Other subfamilies have 1:1 orthologs in *T. castaneum*. Several of the aforementioned families are missing in the phylogeny (Figure 11.2) because they did not meet our criteria for inclusion. The missing subfamilies may be present as orthologs that have diverged too far to be included in the same subfamily.

Comparison of Traditional Gene Annotation vs. Automated Gene Annotation

After completing the traditional gene annotation, the sensitivity of different automated annotation software was tested. Specifically, we tested the sensitivity of software by taking the query as a large insert genomic DNA sequence and also as multiple queries of short amino acid sequences. One hundred eighty-eight sequences of amino acids were run individually or simultaneously in a batch to obtain their sensitivity (see "Materials and Methods" for more information).

Study Using Genomic Sequences for Gene Prediction

In this process, the previously annotated *T. castaneum* LGX sequence (33,080–236,581 bp region) with 111 exons and 24 genes (http://beetlebase.org/cgi-bin/gbrowse/BeetleBase3.gff3/#search) was analyzed using different gene prediction tools and other software. The obtained results were tested for their sensitivity to predict the 5′-end regions of specific exonic pieces on the chromosome within a specified window of cutoff (±20). This cutoff was adjusted to obtain the optimal output of sensitivity from most of the software.

The output of the specific study (Table 4.2) indicates that Augustus, GenScan and Fgenesh provided the best output. These results imply that these tools are very sensitive regarding their site-specific accuracy for detecting specific exonic sequences. However, we obtained many false-positive sequences (Fig.2. provides sensitivity of each software for comparative study.)

Table 4.2. Estimate of the total number of exonic regions in the LGX genomic region of *T. castaneum* predicted with each software. The gene prediction software included Augustus, Fgenesh, GeneID, GeneMark, NCBI ORF Finder, SNAP and GenScan [sensitivity = (number of true positive exons obtained/total number of exons)].

Estimate	Augustus	Fgenesh	Gene ID	GeneMark	ORF Finder NCBI	SNAP	GenScan
Sum	83	68	27	11	50	34	80
Sensitivity	0.741	0.607	0.241	0.098	0.446	0.304	0.714

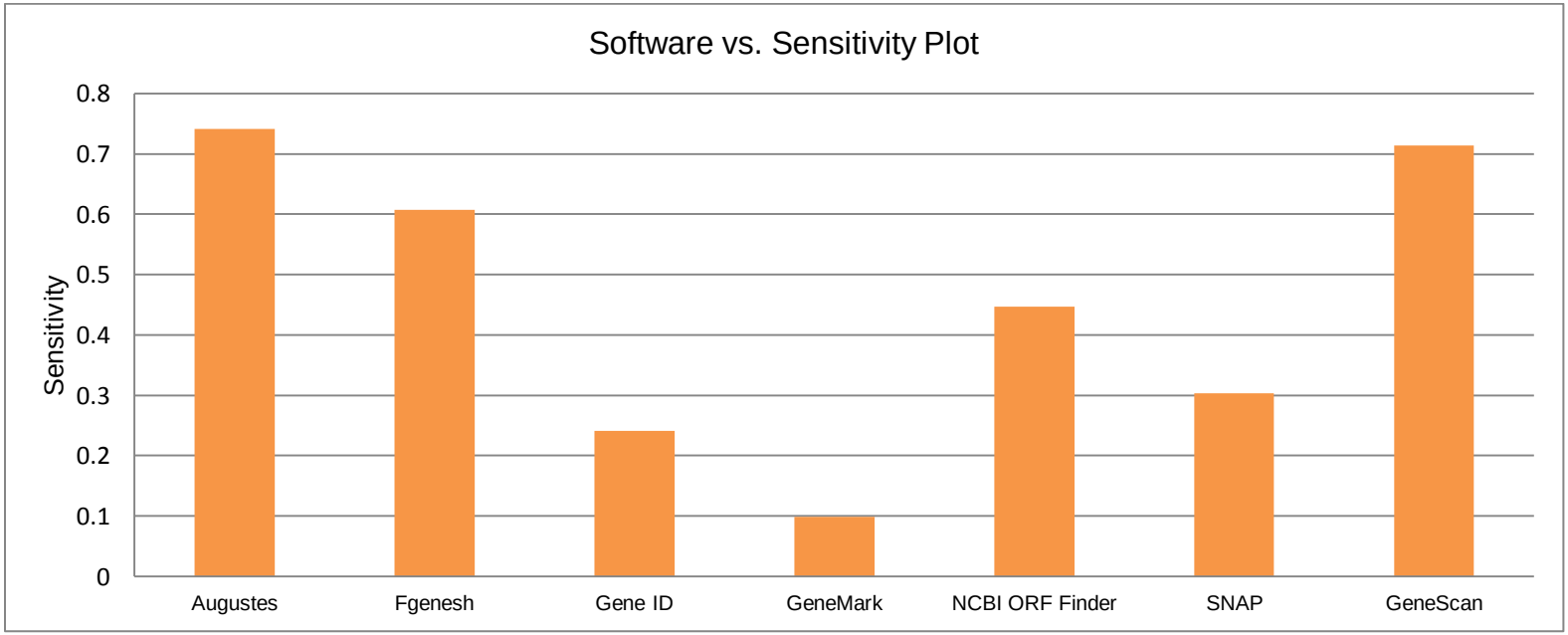

Fig.2. Plot of the sensitivity of the different software used in gene prediction [sensitivity = (number of true positive exons obtained/total number of exons)].

Studying Genes with the Amino Acid Fragments of CYP Sequences

In this process, the results obtained from automated gene annotation were gathered for comparison with the high-quality results obtained by traditional annotation. The procedure followed is elaborated in the protocol. The sensitivity output obtained from searches through Swiss-Prot/TrEMBL databases and automated gene annotation were tested through querying in a web server and querying using pipelines. The output of sensitivity was then studied individually for each software, and BLAST searches of Swiss-Prot/TrEMBL were performed. The output obtained from this process is expanded in Table 4.3. (Fig.10. provides sensitivity of each software for comparative study.)

Table 4.3. Estimate of the total number of exonic regions from the 198 assembled potential contig sequences of *H. axyridis* predicted by each software and NCBI BLAST searches on different publicly available protein databases. The gene prediction software used included Augustus, Fgenesh (pipeline), Fgenes (pipeline), GeneMark, NCBI BLAST at Swiss-Prot/TrEMBL, the combined output of sensitivity predicted by Fgenesh and Fgenes and the combined output of Fgenesh, Fgenes and BLAST search output from Swiss-Prot/TrEMBL [sensitivity = (number of true positive exons obtained/total number of exons)].

Estimate	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	NCBI BLAST at Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
Sum	60	123	135	60	30	75	160	172
Sensitivity	0.337	0.691	0.758	0.337	0.169	0.421	0.899	0.966292

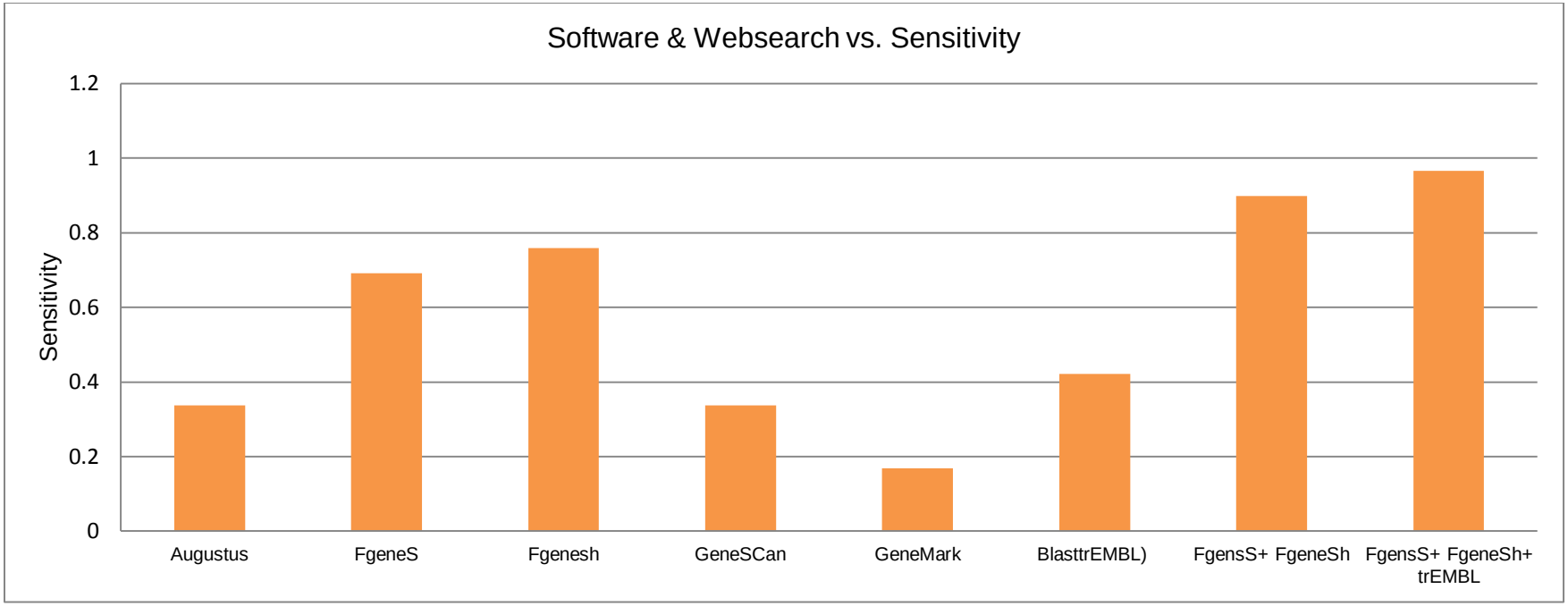

Fig. 10. Bar graph showing the summary of the software vs. sensitivity for gene prediction. The software studied included Augustus, Fgenesh (pipeline), Fgenes (pipeline), GeneMark, NCBI BLAST at Swiss-Prot/TrEMBL, the combined output of sensitivity predicted by Fgenesh and Fgenes and the combined output of Fgenesh, Fgenes and BLAST search output from Swiss-Prot/TrEMBL. The graph indicates that the final combined output of Fgenesh, Fgenes and Swiss-Prot/TrEMBL displays the highest sensitivity [sensitivity = (number of true positive exons obtained/total number of exons)].

The specific advantage of this approach was that we were able to avoid false-positive output completely by considering the similarity and E-value scores obtained from the BLASTp output. The output summary table indicates that Fgenesh and Fgenes pipeline estimation of sensitivity outperformed Augustus and GenScan. However, the table of detailed results in Appendix A1.2 gives a clear overview that the regions detected by Fgenesh and Fgenes complement each other in many cases, and the other few regions remaining are complemented by BLAST searches of publicly available databases.

In the aforementioned study, the output was obtained from the 188 contigs, but none of the software could detect 10 contigs that were detected by traditional gene annotation. These contigs are Mckenna_82098, Mckenna_79748, Mckenna_79212, Mckenna_77757, McKenna_59612, McKenna_59278, McKenna_54994, McKenna_35973, McKenna_29860 and McKenna_11720. When we reviewed the gene annotation and looked for these contigs, we found 2 of the novel contigs that were not detected by any of the aforementioned automated gene annotation method. These included McKenna_82098 (fragment of CYP4 and 42% similar to CYP4S4 of *Mamestra brassicae*, Lepidoptera) and McKenna_29860 (most similar to McKenna_32062), and they were assigned to the subfamily CYP6FP3v2. In the latter case, automated annotation missed one of these fragments but detected the other; this reveals that the automated annotation with less alternative splicing allowed some false-negative errors. We also found that 2 of the selected contigs had ambiguities at specific regions, suggesting that they are slight variants. McKenna_59612 and McKenna_43307 were slight variants, although both were revealed to belong to the same subfamily (CYP349A1). Similarly, McKenna_ 79748 and McKenna_81771 were slight variants assigned to subfamily

CYP6CR. In addition, the other 6 contigs not detected by automated gene annotation were redundant fragments. In both aforementioned cases, the automated annotation detected only one of them. Our final calculation obtained a maximum sensitivity 0.956 (until this point) by considering the false-negative output. The maximum sensitivity that could be obtained through the entire process was computed to be 0.97 when the allowed alternative splicing was maximized to avoid all false positives. This automated gene annotation output displayed 90–100% identity when subjected to a BLASTp search against the output obtained through traditional gene annotation.

The entire process was designed to optimize gene prediction using automated gene annotation. However, we achieved our result partially by obtaining a good sensitivity and avoiding false-positive results completely. The next challenge is to work with large-scale datasets, and we are looking forward to automating the entire process as much as possible. We performed gene prediction using Fgenesh, and we recovered 97,704 gene pieces that are present in *H. axyridis* without using masking before interpretation. We are still working to obtain the sensitivity of the output by detecting the number of false positives. We have already identified the combination of 3 software tools (including 1 web search) for obtaining optimal output, and we are working on that output to annotate the entire genome.

Application and Future Work

The applications of this study include (i) a foundation for finding and establishing the roles of various CYPs in relation to specific phenotypes and (ii) the use of the

automated gene annotation protocols to annotate the entire *H. axyridis* genome (or other genomes) in a more time-efficient manner (for large-scale metagenomic studies).

Potential future work includes (i) obtaining transcriptome information for the previously discussed fragments that will help in further joining fragments to obtain the best possible output, (ii) obtaining the mate-pair information to arrange the contigs in a large scaffold to obtain best possible output for specific genes and (iii) further studying the expression of the specific CYP genes obtained in the aforementioned process. For example, CYP4G could be studied for xenobiotic tolerance by clarifying its gene expression through quantitative PCR. Specific future objectives for automated gene annotation include using the information gathered in this small-scale project to implement large-scale annotation by automating many of the steps.

Overall Summary

Through traditional gene annotation procedures, we identified CYPs in *H. axyridis* (at least 94 genes) via comparison with CYPs in the *T. castaneum* genome (137 genes plus 2 slight variants and 10 pseudogenes). The *H. axyridis* CYPs were divided into 4 distinct clans: Mito, CYP2, CYP3 and CYP4. The 4 clans include 17 families, 42 subfamilies and most probably 3 pseudogene (thus far). The Mito clan of *H. axyridis* contains 6 genes in 5 families and 6 subfamilies. We found 7 genes in CYP2 clan with 6 families (CYP 15, CYP18, CYP303, CYP305, CYP306 and CYP307) and 7 subfamilies. CYP4 clan has 2 distinct families (4 and 349), 8 subfamilies and a minimum of 22 genes in *H. axyridis*. The CYP3 clan has 59 genes in 5 major family members in *H. axyridis*:

CYP6, CYP9, CYP345, CYP435 and CYP436. These 5 major families in CYP3 are classified into at least 21 subfamilies.

Through a comparative annotation study of the *T. castaneum* LGX chromosome, we determined that Fgenesh, Fgenes and Augustus provide the best estimates of sensitivity, with Augustus providing the best sensitivity (0.74). Through a comprehensive automated annotation study of CYP genes followed by comparisons of the estimate to the traditional gene annotation of *H. axyridis*, the best sensitivity estimate was obtained using Fgenesh (0.78). The combined output of sensitivity from multiple software and web searches was 0.96. The exons obtained are 90–100% similar to the exons obtained by traditional methods of gene annotation. The advantage of this approach of BLASTp over the detection of accuracy of specific regions is that we could optimize the output to avoid false positives. In addition, we could avoid false negatives by allowing more alternative splicing. The overall sensitivity we obtained without any false negatives was 0.97.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Siden-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). "The genome sequence of *Drosophila melanogaster*." *Science* 287 (5461): 2185-2195.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). "Basic local alignment search tool." *J. Mol. Biol.* 215 (3): 403-410.

Baldwin, W. S., Marko, P. B., and Nelson, D. R. (2009). "The cytochrome P450 (CYP) gene superfamily in *Daphnia pulex*." *BMC Genomics* 10:169.

Batzoglou, S. (2005). "The many faces of sequence alignment." *Brief. Bioinform.* 6 (1): 6-22.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. (2002). "ARACHNE: a whole-genome shotgun assembler." *Genome Res.* 12 (1): 177-189.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. A., and Jones, S. J. (2009). "De novo transcriptome assembly with ABySS." *Bioinformatics* 25 (21): 2872-2877.

Bolt, J. R, Lombart, E. (2010). "*Deltaherpeton hiemstrae*, a New Colosteid Tetrapod from the Mississippian of Iowa." *J. Paleontol.* 84 (6): 1135-1151.

Bradfield, J. Y., Lee, Y. H., and Keeley, L. L. (1991). "Cytochrome P450 family 4 in a cockroach: molecular cloning and regulation by regulation by hypertrehalosemic hormone." *Proc. Natl. Acad. Sci. U.S.A.* 88 (10): 4558-4562.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." *Nat. Biotechnol.* 18 (6): 630-634.

Brent, M. R. (2005). "Genome annotation past, present, and future: how to define an ORF at each locus." *Genome Res.* 15 (12): 1777-1786.

Brent, M. R. and Guigo, R. (2004). "Recent advances in gene structure prediction." *Curr. Opin. Struct. Biol.* 14 (3): 264-272.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads." *Genome Res.* 18 (5): 810-820.

Carvajal-Rodriguez, A. (2010). "Simulation of genes and genomes forward in time." *Curr.Genomics* 11 (1): 58-61.

Carle-Urioste, J.C., Brendel, V., and Walbot, V. (1997). "A combinatorial role for exon, intron and splice site sequences in splicing in maize." *Plant J.* 11 (6): 1253-1263.

Chaisson, M. J., and Pevzner, P. A. (2008). "Short read fragment assembly of bacterial genomes." *Genome Res.* 18 (2): 324-330.

Ciccarelli, F. D., von Mering, C., Suyama, M., Harrington, E. D., Izaurralde, E., and Bork, P. (2005). "Complex genomic rearrangements lead to novel primate gene function." *Genome Res.* 15 (3): 343-351.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). "Continuous base identification for single-molecule nanopore DNA sequencing." *Nat. Nanotechnol.* 4 (4): 265-270.

Claudianos, C., Ranson, H., Johnson, R. M., Biswas, S., Schuler, M. A., Berenbaum, M. R., Feyereisen, R., and Oakeshott, J. G. (2006). "A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee." *Insect Mol. Biol.* 15 (5): 615-636.

Claverie, J.-M. (1994) In: "Automated DNA Sequencing and Analysis Techniques," M. D. Adams, C. Fields, and J. C. Venter, eds.: pp. 267-279.

Crampton, A. L., Baxter, G. D., and Barker, S. C. (1999). "A new family of cytochrome P450 genes (CYP41) from the cattle tick, *Boophilus microplus*." *Insect Biochem. Mol. Biol.* 29 (9): 829-834.

Danielson, P. B. (2002). "The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans." *Curr. Drug Metab.* 3 (6): 561-597.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2007). "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing." *Genome Res.* 17 (11): 1697-1706.

Donis-Keller, H., Maxam, A. M., and Gilbert, W. (1977). "Mapping adenines, guanines, and pyrimidines in RNA." *Nucl. Acids Res.* 4 (8): 2527-2538.

Dover, G. A. and Tautz, D. (1986). "Conservation and divergence in multigene families: alternatives to selection and drift." *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 312 (1154): 275-289.

Dwyer, D. (2011). "Experiences of registered nurses as managers and leaders in residential aged care facilities: a systematic review." *Int. J. Evid1 Based Healthc.* 9 (4): 388-402.

Eddy, S. R. (1998). "Profile hidden Markov models." *Bioinformatics* 14 (9): 755-763.

Eddy, S. R., Mitchison, G., and Durbin, R. (1995). "Maximum discrimination hidden Markov models of sequence consensus." *J. Comput. Biol.* 2 (1): 9-23.

Eyras, E., Reymond, A., Castelo, R., Bye, J. M., Camara, F., Flicek, P., Huckle, E. J., Parra, G., Shteynberg, D. D., Wyss, C., Rogers, J., Antonarakis, S. E., Birney, E., Guigo, R., and Brent, M. R. (2005). "Gene finding in the chicken genome." *BMC Bioinformatics* 6: 131.

Fahrbach, S. E., Smagghe, G., and Velarde, R. A. (2012). "Insect nuclear receptors." *Annu. Rev. Entomol.* 57: 83-106.

Farrer, R. A., Kemen, E., Jones, J. D., and Studholme, D. J. (2009). "De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads." *FEMS Microbiol. Lett.* 291 (1): 103-111.

Fernandez-Carvajal, A. M., Encinar, J. A., Poveda, J. A., de Juan, E., Martinez-Pinna, J., Ivorra, I., Ferragut, J. A., Morales, A., and Gonzalez-Ros, J. M. (2006). "Structural and functional changes induced in the nicotinic acetylcholine receptor by membrane phospholipids." *J. Mol. Neurosci.* 30 (1-2): 121-124.

Feyereisen, R. (2006). "Evolution of insect P450." *Biochem. Soc. Trans.* 34(6): 1252-1255.

Fickett, J. W. (1996). "The gene identification problem: an overview for developers." *Comput. Chem.* 20 (1): 103-118.

Finlay, B. J., Thomas, J. A., McGavin, G. C., Fenchel, T., and Clarke, R. T. (2006). "Self-similar patterns of nature: insect diversity at local to global scales." *Proc. Biol. Sci.* 273 (1596): 1935-1941.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Ratsch, G., and Mott, R. (2011). "Multiple reference genomes and transcriptomes for Arabidopsis thaliana." *Nature* 477 (7365): 419-423.

Garfinkel, D. (1958). "Studies on pig liver microsomes. I. Enzymic and pigment composition of different microsomal fractions." *Arch. Biochem. Biophys.* 77 (2): 493-509.

Gelfand, M. S. (1995). "FANS-REF: a bibliography on statistics and functional analysis of nucleotide sequences." *Comput. Appl. Biosci.* 11(5): 541.

Gilbert, L. I. and Warren, J. T. (2005). "A molecular genetic approach to the biosynthesis of the insect steroid molting hormone." *Vitam. Horm.* 73: 31-57.

Gonzalez, F. J. and Nebert, D. W. (1990). "Evolution of the P450 gene superfamily: animal-plant 'warfare', molecular drive and human genetic differences in drug oxidation." *Trends Genet.* 6 (6): 182-186.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). "Profile analysis: detection of distantly related proteins." *Proc. Natl. Acad. Sci. U.S.A.* 84 (13): 4355-4358.

Grimaldi, D., Engel, M. S. (2005). "Evolution of the Insects." Cambridge University Press

Gregory, T.R., Nedved, O., and Adamowicz, S.J. (2003). "C-value estimates for 31 species of ladybird beetles (Coleoptera: Coccinellidae)." *Hereditas* 139: 121-127

Guengerich, F. P. (2008). "Cytochrome p450 and chemical toxicology." *Chem. Res. Toxicol.* 21 (1): 70-83.

Guittard, E., Blais, C., Maria, A., Parvy, J., Pasricha, S., Lumb, C., Lafont, R., Daborn, P. J., Dauphin-Villemant, C. (2011). "CYP18A1, a key enzyme of *Drosophila* steroid hormone inactivation, is essential for metamorphosis." *Dev. Biol.* 349: 35-45.

Guo, G. Z., Geng, Y. J., Huang, D. N., Xue, C. F., and Zhang, R. L. (2010). "Level of CYP4G19 Expression Is Associated with Pyrethroid Resistance in *Blattella germanica*." *J. Parasitol. Res.* 2010: 517-534.

Guzov, V. M., Unnithan, G. C., Chernogolov, A. A., and Feyereisen, R. (1998). "CYP12A1, a mitochondrial cytochrome P450 from the house fly." *Arch. Biochem. Biophys.* 359 (2): 231-240.

Hahn, M. W., Han, M. V., and Han, S. G. (2007). "Gene family evolution across 12 *Drosophila* genomes." *PLoS Genet.* 3 (11): e197.

Halperin, E., Buhler, J., Karp, R., Krauthgamer, R., and Westover, B. (2003). "Detecting protein sequence conservation via metric embeddings." *Bioinformatics* 19: 122-129.

Hannemann, F., Bichet, A., Ewen, K. M., and Bernhardt, R. (2007). "Cytochrome P450 systems--biological variations of electron transport chains." *Biochim. Biophys. Acta* 1770 (3): 330-344.

Havlak, P., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Weinstock, G. M., and Gibbs, R. A. (2004). "The Atlas genome assembly system." *Genome Res.* 14 (4): 721-732.

Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G., and Birney, E. (2011). "Efficient storage of high throughput DNA sequencing data using reference-based compression." *Genome Res.* 21 (5): 734-740.

http://agris.fao.org/openagris/search.do?recordID=CN2010001174

http://www.faqs.org/patents/app/20090023173#ixzz1v8Z6NzlK

Inoue, Y., Miyawaki, K., Terasawa, T., Matsushima, K., Shinmyo, Y., Niwa, N., Mito, T., Ohuchi, H., and Noji, S. (2004). "Expression patterns of dachshund during head development of *Gryllus bimaculatus* (cricket)." *Gene Expr. Patterns* 4 (6): 725-731.

"Insects." http://crazydaz.com/insects.pdf. pp. 4. Retrieved 2009-05-17.

Issac, B., Singh, H., Kaur, H., and Raghava, G. P. (2002). "Locating probable genes using Fourier transform approach." *Bioinformatics* 18 (1): 196-197.

Janz, N., Nylin, S., and Wahlberg, N. (2006). "Diversity begets diversity: host expansions and the diversification of plant-feeding insects." *BMC Evol. Biol.* 6: 4.

Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., and Jones, C. D. (2007). "Extending assembly of short DNA sequences to handle error." *Bioinformatics* 23 (21): 2942-2944.

Jiang, H. B., Wang, J. J., Liu, G. Y., and Dou, W. (2008). "Molecular cloning and sequence analysis of a novel P450 gene encoding CYP345D3 from the red flour beetle, *Tribolium castaneum*." *J. Insect Sci.* 8: 1-7.

Karatolos, N., Williamson, M. S., Denholm, I., Gorman, K., Ffrench-Constant, R. H., and Bass, C. (2012). "Over-expression of a cytochrome P450 is associated with resistance to pyriproxyfen in the greenhouse whitefly *Trialeurodes vaporariorum*." *PLoS One* 7 (2): e31077.

Karlin, S.,and Mrazek, J. (1997). "Compositional differences within and between eukaryotic genomes." *Proc. Natl. Acad. Sci. U.S.A.* 94 (19): 10227-10232.

Kel, A. E., Kondrakhin, Y. V., Kolpakov, P., Kel, O. V., Romashenko, A. G., Wingender, E., Milanesi, L., and Kolchanov, N. A. (1995). "Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences." *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3: 197-205.

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). "Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering." *Nucl. Acids Res.* 40 (1): e9.

Kircher, M. and Kelso, J. (2010). "High-throughput DNA sequencing--concepts and limitations." *Bioessays* 32 (6): 524-536.

Klingenberg, M. (1958). "Pigments of rat liver microsomes." *Arch. Biochem. Biophys.* 75 (2), 376-386.

Koblizek, M., Shih, J. D., Breitbart, S. I., Ratcliffe, E. C., Kolber, Z. S., Hunter, C. N., and Niederman, R. A. (2005). "Sequential assembly of photosynthetic units in *Rhodobacter sphaeroides* as revealed by fast repetition rate analysis of variable bacteriochlorophyll a fluorescence." *Biochim. Biophys. Acta* 1706 (3): 220-231.

Koch R.L. (2003). "The multicolored Asian lady beetle, *Harmonia axyridis*: A review of its biology, uses in biological control, and non-target impacts." *J. Insect Sci.* 3:32.

Krafsur, E. S., Kring, T. J., Miller, J. C., Nariboli, P., Obrycki, J. J., Ruberson, J. R., Schaefer, P. W. (1997) Gene Flow in the Exotic Colonizing Ladybeetle *Harmonia axyridis* in *North America*. 8: 207-214.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). "Hidden Markov models in computational biology. Applications to protein modeling." *J. Mol. Biol.* 235(5): 1501-1531.

Kuwahara, S., Omura, T. (1980). "Different requirement for cytochrome b5 in NADPH-supported O-deethylation of p-nitrophenetole catalyzed by two types of microsomal cytochrome P-450." *Biochem. Biophys. Res. Comm.* 96 (4):1562-1568.

LaMana ML, Miller JC. (1996). Field observations on *Harmonia axyridis* Pallas (Coleoptera: Coccinellidae) in Oregon. *Biol. Control* 6: 232-237.

Le Goff, G., Hilliou, F., Siegfried, B. D., Boundy, S., Wajnberg, E., Sofer, L., Audant, P., ffrench-Constant, R. H., and Feyereisen, R. (2006). "Xenobiotic response in *Drosophila melanogaster*: sex dependence of P450 and GST gene induction." *Insect Biochem. Mol. Biol.* 36 (8): 674-682.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., and Wang, J. (2009). "SOAP2: an improved ultrafast tool for short read alignment." *Bioinformatics.* 25 (15): 1966-1967.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). "De novo assembly of human genomes with massively parallel short read sequencing." *Genome Res.* 20 (2): 265-272.

Lindberg, R. L., and Negishi, M. (1989). "Alteration of mouse cytochrome P450 coh substrate specificity by mutation of a single amino-acid residue." *Nature* 339 (6226): 632-634.

Locali-Fabris, E. C., Freitas-Astua, J., Souza, A. A., Takita, M. A., Astua-Monge, G., Antonioli-Luizon, R., Rodrigues, V., Targon, M. L., and Machado, M. A. (2006). "Complete nucleotide sequence, genomic organization and phylogenetic analysis of *Citrus leprosis* virus cytoplasmic type." *J. Gen. Virol.* 87 (9): 2721-2729.

Lombaert, E., Guillemaud, T., Cornuet, J. M., Malausa, T., Facon, B., and Estoup, A. (2010). "Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird." *PLoS One* 5 (3): e9743.

Long, M., de Souza, S. J., Rosenberg, C., and Gilbert, W. (1998). "Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis." *Proc. Natl. Acad. Sci. U.S.A.* 95(1): 219-223.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." *Nature* 437 (7057): 376-380.

Martínez-Paz, P., Morales, M., Martínez-Guitarte, J. L., Gloria Morcillo. (2012). "Characterization of a cytochrome P450 gene (CYP4G) and modulation under different exposures to xenobiotics (tributyltin, nonylphenol, bisphenol A) in *Chironomus riparius* aquatic larvae." *Comp. Biochem. Physiol. C. Toxicol. Pharmacol*. 155 (2): 333-343.

Mathe, C., Sagot, M., Schiex, T., and Rouze, P. (2002) "Current methods of gene prediction, their strengths and weaknesses." *Nucl. Acids Res.* 30 (19): 4103-4117.

Mason, H.S. (1957). Mechanisms of oxygen metabolism. *Adv. Enzymol. Relat. Subj. Biochem.* 19: 79-233.

Maxam, A. M. and Gilbert, W. (1986). "A method for determining DNA sequence by labeling the end of the molecule and cleaving at the base. Isolation of DNA fragments, end-labeling, cleavage, electrophoresis in polyacrylamide gel and analysis of results." *Mol. Biol. (Mosk)* 20(3): 581-638.

Mccouch, S. R., Mcnally, K. L., Wang, W., and Hamilton, R. S. (2012). "Genomics of gene banks: A case study in rice." *Am. J. Bot.* 99: 407-423.

McKenna, D. D. (2006). "Towards a Temporal Framework for “Inordinate Fondness”: Reconstructing the Macroevolutionary History of Beetles (Coleoptera)." *Entomologica Americana*. 117 (1 & 2): 28-36.

McKenna, D. D. and Farrell, B. D. (2006). "Tropical forests are both evolutionary cradles and museums of leaf beetle diversity." *Proc. Natl. Acad. Sci. U.S.A.* 103 (29): 10947-10951.

McKenna, D. D., Sequeira, A. S., Marvaldi, A. E., and Farrell, B. D. (2009). "Temporal lags and overlap in the diversification of weevils and flowering plants." *Proc. Natl. Acad. Sci. U.S.A.* 106 (17): 7083-7088.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." *Nat. Rev. Genet.* 11 (1): 31-46.

Meunier, B., de Visser, S. P., and Shaik, S. (2004). "Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes." *Chem. Rev.* 104 (9): 3947-3980.

Meyer, U. A. and Zanger, U. M. (1997). "Molecular mechanisms of genetic polymorphisms of drug metabolism." *Annu. Rev. Pharmacol. Toxicol.* 37: 269-296.

Miller, N., Estoup, A., Toepfer, S., Bourguet, D., Lapchin, L., Derridj, S., Kim, K. S., Reynaud, P., Furlan, L., and Guillemaud, T. (2005). "Multiple transatlantic introductions of the western corn rootworm." *Science* 310 (5750): 992.

Miller, W., Drautz, D. I., Ratan, A., Pusey, B., Qi, J., Lesk, A. M., Tomsho, L. P., Packard, M. D., Zhao, F., Sher, A., Tikhonov, A., Raney, B., Patterson, N., Lindblad-Toh, K., Lander, E. S., Knight, J. R., Irzyk, G. P., Fredrikson, K. M., Harkins, T. T., Sheridan, S., Pringle, T., and Schuster, S. C. (2008). "Sequencing the nuclear genome of the extinct woolly mammoth." *Nature* 456 (7220): 387-390.

Mount, D. W. (2008). "A test of the markov model of evolution in proteins." *CSH Protoc.* doi:10.1101/pdb.ip58.

Mullikin, J. C. and Ning, Z. (2003). "The phusion assembler." *Genome Res.* 13 (1): 81-90.

Myers, E. W. (2005). "The fragment assembly string graph." *Bioinformatics* 21 (2): ii79-ii85.

Nakazawa, T., Satinover, S. M., Naccara, L., Goddard, L., Dragulev, B. P., Peters, E., and Platts-Mills, T. A. (2007). "Asian ladybugs (Harmonia axyridis): a new seasonal indoor allergen." *J. Allergy Clin. Immunol.* 119 (2): 421-427.

Namiki, T., Niwa, R., Sakudoh, T., Shirai, K., Takeuchi, H., and Kataoka, H. (2005). "Cytochrome P450 CYP307A1/Spook: a regulator for ecdysone synthesis in insects." *Biochem. Biophys. Res. Commun.* 337 (1): 367-374.

Nebert, D. W. and Gonzalez, F. J. (1987). "P450 genes: structure, evolution, and regulation." *Annu. Rev. Biochem.* 56: 945-993.

Nebert, D. W., Jones, J. E., Owens, J., and Puga, A. (1988). "Evolution of the P450 gene superfamily." *Prog. Clin. Biol. Res.* 274: 557-576.

Nebert, D. W., Nelson, D. R., Adesnik, M., Coon, M. J., Estabrook, R. W., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., Johnson, E. F., and Kemper, B. (1989). "The P450 superfamily: updated listing of all genes and recommended nomenclature for the chromosomal loci." *DNA* 8 (1): 1-13.

Nebert, D. W., Nelson, D. R., Coon, M. J., Estabrook, R. W., Feyereisen, R., Fujii-Kuriyama, Y., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., and Johnson, E. F.

(1991). "The P450 superfamily: update on new sequences, gene mapping, and recommended nomenclature." *DNA Cell Biol.* 10(1): 1-14.

Nebert, D. W., Nelson, D. R., and Feyereisen, R. (1989). "Evolution of the cytochrome P450 genes." *Xenobiotica* 19 (10): 1149-1160.

Nelson, D. R. (2006). "Cytochrome P450 nomenclature, 2004." *Methods Mol. Biol.* 320: 1-10.

Nelson, D. R. (1998). "Metazoan cytochrome P450 evolution." *Comp. Biochem. Physiol. C. Pharmacol. Toxicol. Endocrinol.* 121 (1-3): 15-22.

Nelson, D. R. and Strobel, H. W. (1987). "Evolution of cytochrome P-450 proteins." *Mol. Biol. Evol.* 4(6): 572-593.

Nelson, D. R., Kamataki, T., Waxman, D. J., Guengerich, F. P., Estabrook, R. W., Feyereisen, R., Gonzalez, F. J., Coon, M. J., Gunsalus, I. C., and Gotoh, O. (1993). "The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature." *DNA Cell Biol.* 12 (1): 1-51.

Nelson, D. R., Zeldin, D. C., Hoffman, S. M., Maltais, L. J., Wain, H. M., and Nebert, D. W. (2004). "Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants." *Pharmacogenetics* 14 (1): 1-18.

Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D., and Wang, T. (2010). "Visualizing genomes: techniques and challenges." *Nat. Methods* 7 (3): S5-S15.

Niwa, R., Matsuda, T., Yoshiyama, T., Namiki, T., Mita, K., Fujimoto, Y., and Kataoka, H. (2004). "CYP306A1, a cytochrome P450 enzyme, is essential for ecdysteroid biosynthesis in the prothoracic glands of *Bombyx* and *Drosophila*." *J. Biol. Chem.* 279 (34): 35942-35949.

Oakeshott, J. G., Johnson, R. M., Berenbaum, M. R., Ranson, H., Cristino, A. S., and Claudianos, C. (2010). "Metabolic enzymes associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*." *Insect Mol. Biol.* 19 (1): 147-163.

Omura, T., Sato, R. (1964). "The Carbon Monoxide-Binding Pigment of Liver Microsomes. Ii. Solubilization, Purification, and Properties." *J. Biol. Chem.* 239: 2379-2385.

Ongagna P., Giuge L., Iperti G., Ferran A. (1993): "Life cycle of *Harmonia axyridis* (Col.: Coccinellidae) in its area of introduction: south-eastern France." *Entomophaga* 38: 125-128.

Ono, H., Rewitz, K. F., Shinoda, T., Itoyama, K., Petryk, A., Rybczynski, R., Jarcho, M., Warren, J. T., Marques, G., Shimell, M. J., Gilbert, L. I., and O'Connor, M. B. (2006). "Spook and Spookier code for stage-specific components of the ecdysone biosynthetic pathway in *Diptera*." *Dev. Biol.* 298 (2): 555-570.

Pareek, C. S., Smoczynski, R., Tretyn, A., (2011). "Sequencing technologies and genome sequencing." *J. Appl. Genet*. 52 (4): 413-435.

Parish, C. R. and Warren, H. S. (2002). "Use of the intracellular fluorescent dye CFSE to monitor lymphocyte migration and proliferation." *Curr. Protoc. Immunol.* Chapter 4 Unit 4.9.

Parvy, J. P., Blais, C., Bernard, F., Warren, J. T., Petryk, A., Gilbert, L. I., O'Connor, M. B., and Dauphin-Villemant, C. (2005). "A role for betaFTZ-F1 in regulating ecdysteroid titers during post-embryonic development in *Drosophila melanogaster*." *Dev. Biol.* 282 (1): 84-94.

Pedersen, J. S., Hein, J. (2003). "Gene finding with a hidden Markov model of genome structure and evolution." *Bioinformatics* 19 (2): 219-227.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980). "The evolution of genes: the chicken preproinsulin gene." *Cell* 20 (2): 555-566.

Petryk, A., Warren, J. T., Marques, G., Jarcho, M. P., Gilbert, L. I., Kahler, J., Parvy, J. P., Li, Y., Dauphin-Villemant, C., and O'Connor, M. B. (2003). "Shade is the *Drosophila* P450 enzyme that mediates the hydroxylation of ecdysone to the steroid insect molting hormone 20-hydroxyecdysone." *Proc. Natl. Acad. Sci. U.S.A.* 100 (24): 13773-13778.

Petty, N. K. (2010). "Genome annotation: man versus machine." *Nat. Rev. Microbiol.* 8 (11): 762.

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

Poutsma, J., Loomans, A. J. M., Aukema, B., and Heijerman, T. (2008). "Predicting the potential geographic distribution of the harlequin ladybird, *Harmonia axyridis*, using the CLIMEX odel." *BioControl* 53: 103-125.

Pridgeon, J. W., Zhang, L., and Liu, N. (2003). "Overexpression of CYP4G19 associated with a pyrethroid-resistant strain of the German cockroach, *Blattella germanica* (L.)." *Gene* 314: 157-163.

Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." *Proc. IEEE* 77(2): 257-285.

Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M. V., Unger, M. F., Collins, F. H., and Feyereisen, R. (2002). "Evolution of supergene families associated with insecticide resistance." *Science* 298 (5591): 179-181.

Rewitz, K. F. and Gilbert, L. I. (2008). "*Daphnia* Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone: evolutionary implications." *BMC Evol. Biol.* 8: 60.

Rewitz, K. F., O'Connor, M. B. and Gilbert, L. I. (2007). "Molecular evolution of the insect Halloween family of cytochrome P450s: phylogeny, gene organization and functional conservation." *Insect Biochem. Mol. Biol.* 37 (8): 741-753.

Rewitz, K. F., Rybczynski, R., Warren, J. T. and Gilbert, L. I. (2006). "The Halloween genes code for cytochrome P450 enzymes mediating synthesis of the insect moulting hormone." *Biochem. Soc. Trans.* 34 (6): 1256-1260.

Richards, S. (2008). "The genome of the model beetle and pest *Tribolium castaneum*." *Nature* 452: 949-955.

Ridley, A. M., Allen, V. M., Sharma, M., Harris, J. A., and Newell, D. G. (2008). "Real-time PCR approach for detection of environmental sources of Campylobacter strains colonizing broiler flocks." *Appl. Environ. Microbiol.* 74 (8): 2492-2504.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., Altshuler, D., and International SNP Map Working Group. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." *Nature* 409 (6822): 928-933.

Sanger, F. (1975). "The Croonian Lecture, 1975. Nucleotide sequences in DNA." *Proc. R. Soc. Lond. B. Biol. Sci.* 191 (1104): 317-333.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." *Proc. Natl. Acad. Sci. U.S.A.* 74 (12): 5463-5467.

Saini, E.D. (2004). "Presencia de *Harmonia axyridis* (Pallas) (Coleoptera: Coccinellidae) en la provincia de Buenos Aires. Aspectos biológicos y morfológicos." *RIA.* 33: 151-160.

Savard, J., Tautz, D., Richards, S., Weinstock, G. M., Gibbs, R. A., Werren, J. H., Tettelin, H., and Lercher, M. J. (2006). "Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects." *Genome Res.* 16 (11): 1334-1338.

Scott, J. G., Wen, Z. (2001). "Cytochromes P450 of insects: the tip of the iceberg." *Pest Manag. Sci.* 57: 958-967.

Secnik, J., Gelfand, C. A., and Jentoft, J. E. (1992). "Retroviral nucleocapsid protein specifically recognizes the base and the ribose of mononucleotides and mononucleotide components." *Biochemistry* 31 (11): 2982-2988.

Shendure, J., Mitra, R. D., Varma, C., and Church, G. M. (2004). "Advanced sequencing technologies: methods and goals." *Nat. Rev. Genet.* 5 (5): 335-344.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." *Science* 309 (5741): 1728-1732.

Sigel, H., Sigel, A., and Sigel, R. K. O. (2007). "Metal Ions in Life Sciences. The Ubiquitous Roles of Cytochrome P450 Proteins Volume 3." John Wiley and Sons.

Singh, G. B. (2000). "Computational approaches for gene identification." *Methods Mol. Biol.* 132: 351-364.

Sligar, S. G., Cinti, D. L., Gibson, G. G., and Schenkman, J. B. (1979). "Spin state control of the hepatic cytochrome P450 redox potential." *Biochem. Biophys. Res. Commun.* 90 (3): 925-932.

Sloggett, J. J. (2010). Predation of ladybird beetles by the orb-web spider *Araneus diadematus*. *BioControl* 55: 631-638.

Smith, T. F., Waterman, M. S., and Fitch, W. M. (1981). "Comparative biosequence metrics." *J. Mol. Evol.* 18 (1): 38-46.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). "Pfam: multiple sequence alignments and HMM-profiles of protein domains." *Nucl. Acids Res.* 26(1): 320-322.

Stals, R. and Prinsloo, G. (2007). "Discovery of an alien invasive, predatory insect in South Africa: the multicoloured Asian ladybird beetle, *Harmonia axyridis* (Pallas; Coleoptera: Coccinellidae)." *S. Afr. J. Sci.* 103 (3-4): 123-126.

Stevens, J. L., Snyder, M. J., Koener, J. F., and Feyereisen, R. (2000). "Inducible P450s of the CYP9 family from larval *Manduca sexta* midgut." *Insect Biochem. Mol. Biol.* 30 (7): 559-568.

Strode, C., Wondji, C. S., David, J. P., Hawkes, N. J., Lumjuan, N., Nelson, D. R., Drane, D. R., Karunaratne, S. H., Hemingway, J., Black, W. C.,4th, and Ranson, H. (2008).

"Genomic analysis of detoxification genes in the mosquito *Aedes aegypti*." *Insect Biochem. Mol. Biol.* 38 (1): 113-123.

Swerdlow, H., Zhang, J. Z., Chen, D. Y., Harke, H. R., Grey, R., Wu, S. L., Dovichi, N. J., and Fuller, C. (1991). "Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence." *Anal. Chem.* 63 (24): 2835-2841.

Tijet, N., Helvig, C., and Feyereisen, R. (2001). "The cytochrome P450 gene superfamily in *Drosophila melanogaster*: annotation, intron-exon organization and phylogeny." *Gene* 262 (1-2): 189-198.

Tompkins, L. M. and Wallace, A. D. (2007). "Mechanisms of cytochrome P450 induction." *J. Biochem. Mol. Toxicol.* 21 (4): 176-181.

Wang, J., Wong, G. K., Ni, P., Han, Y., Huang, X., Zhang, J., Ye, C., Zhang, Y., Hu, J., Zhang, K., Xu, X., Cong, L., Lu, H., Ren, X., Ren, X., He, J., Tao, L., Passey, D. A., Wang, J., Yang, H., Yu, J., and Li, S. (2002). "RePS: a sequence assembler that masks exact repeats identified from the shotgun data." *Genome Res.* 12 (5): 824-831.

Wang, R., Qiu, Z., Chen, J., Warren, A., and Song, W. (2007). "Morphogenesis of the freshwater ciliate *Neokeronopsis spectabilis* (Kahl 1932) Warren et al., 2002, based on a China population (Ciliophora: Urostylidae)." *J. Eukaryot. Microbiol.* 54 (2): 184-190.

Wang, Y., Sun, S., Liu, B., Wang, H., Deng, J., Liao, Y., Wang, Q., Cheng, F., Wang, X., and Wu, J. (2011). "A sequence-based genetic linkage map as a reference for *Brassica rapa* pseudochromosome assembly." *BMC Genomics* 12: 239.

Warren, R. L., Varabei, D., Platt, D., Huang, X., Messina, D., Yang, S. P., Kronstad, J. W., Krzywinski, M., Warren, W. C., Wallis, J. W., Hillier, L. W., Chinwalla, A. T., Schein, J. E., Siddiqui, A. S., Marra, M. A., Wilson, R. K., and Jones, S. J. (2006). "Physical map-assisted whole-genome shotgun sequence assemblies." *Genome Res.* 16 (6): 768-775.

Wiegmann, B. M., Mitter, B., and Farrell, B. (1993). "Diversification of parasitic insects: extraordinary radiation or specialized dead end?"*Am. Nat.* 142 (5): 737-754.

Wilkinson, G. R. (2005). "Drug metabolism and variability among patients in drug response." *N. Engl. J. Med.* 352 (21): 2211-2221.

Willingham, A. T. and Keil, T. (2004). "A tissue specific cytochrome P450 required for the structure and function of Drosophila sensory organs." *Mech. Dev.* 121 (10): 1289-1297.

Yang, Y., Chen, S., Wu, S., Yue, L., and Wu, Y. (2006). "Constitutive overexpression of multiple cytochrome P450 genes associated with pyrethroid resistance in *Helicoverpa armigera*." *J. Econ. Entomol.* 99 (5): 1784-1789.

Zerbino, D. R. and Birney, E. (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." *Genome Res.* 18 (5): 821-829.

Zhang, Y. L., Kulye, M., Yang, F. S., Xiao, L., Zhang, Y. T., Zeng, H., Wang, J. H., and Liu, Z. X. (2011). "Identification, characterization, and expression of a novel P450 gene encoding CYP6AE25 from the Asian corn borer, *Ostrinia furnacalis*." *J. Insect Sci.* 11: 37.

Zhu, F., Li, T., Zhang, L., and Liu, N. (2008). "Co-up-regulation of three P450 genes in response to permethrin exposure in permethrin resistant house flies, *Musca domestica*." *BMC Physiol.* 8:18.

Appendix

Table A.1.1. *Tribolium castaneum* genomic sequence annotation

This table provides an estimate of the total number of exonic regions on the LGX genomic region of *T. castaneum* predicted by each software. The utilized gene prediction software included Augustus (Aug), Fgenesh, GeneID, GeneMark, NCBI ORF Finder, SNAP and GenScan. In this table, "Exon" represents the number of exonic sequences, and the subsequent columns present the estimates from all of the software tools with their different exonic estimates. The "Freq" column provides the number of software tools that could detect the exon positively. The "Putative exons" column shows the starting and ending codons and their respective bp positions. The estimates of the total number of true exons and sensitivity of each software tool are summarized at the bottom of table.

Exon	Aug	Fgenesh	Gene ID	GeneMark	NCBI ORF Finder	SNAP	GenScan	Freq	Putative exons
1	1	1			1		1	3	CAA 33,080–33,186 TTT
2	1	1			1		1	4	GCG 34,038–34,099 GCG
3	1	1		1			1	3	AGA 34,174–34,370 AAT
4	1	1		1	1		1	5	GTT 34,422–35,313 CTT
5	1	1			1		1	4	ATG 49,582–49,965 TAA
6					1			1	TTT 67,681–67,954 CAG
7	1	1		1				3	GAT 67,999–68,232 TTA
8	1	1			1	1	1	5	ATG 68,673–68,916 GAG

Exon	Aug	Fgenesh	Gene ID	GeneMark	NCBI ORF Finder	SNAP	GenScan	Freq	Putative exons
9	1					1	1	3	ATT 68,964–69,151 CGA
10					1			1	GTG 69,691–69,870 GAG
11		1						1	ATG 69,913–70,173 TAA
12							1	1	ATT 85,003–85,355 AAG
13	1		1			1	1	4	GTG 85,399–85,831 GAG
14		1	1			1		3	ACA 88,087–88,263 CCG
15	1	1	1			1		4	CGC 88,314–88,679 TAA
16	1	1			1	1	1	5	ATG 106,518–106,661 AAG
17	1	1		1	1	1	1	6	ATG 106,706–107,012 ATG
18		1			1	1	1	4	CTG 107,066–107,289 ATG
19	1	1	1		1	1	1	6	CCT 108,018–108,238 ATC
20	1				1		1	3	ATT 108,285–108,524 CAA
21	1	1			1		1	4	GTT 108,574–108,760 GCG
22	1	1			1			3	GCC 108,811–108,903 AAA
23	1	1				1	1	4	ACT 108,955–109,017 TAA
24			1					1	ATG 127,429–127,591 ATG
25							1	1	CAT 132,124–132,209 TGA
26	1	1					1	3	ATG 139,056–139,073 AAG
27	1	1					1	3	AAC 139,124–139,240 CAT
28	1							2	GTT 139,291–139,387 GAA
29		1					1	2	AGG 139,437–139,626 ACA
30					1			1	GAT 140,239–140,298 GGA
31								1	NAA 140,556–141,196 AAA

Exon	Aug	Fgenesh	Gene ID	GeneMark	NCBI ORF Finder	SNAP	GenScan	Freq	Putative exons
32			1		1	1	1	4	ACG 141,813–141,924 CAC
33	1		1		1	1	1	5	GCA 141,971–142,849 GAG
34		1		1	1		1	4	AAG 143,149–143,385 GCT
35	1	1					1	3	GTC 143,556–143,721 CGT
36	1							1	ATC 143,775–143,849 CAG
37			1		1	1		2	GTT 144,402–144,839 CTT
38	1							1	GAG 144,900–144,989 GAG
39	1							1	GTG 145,023–145,121 GAG
40	1				1			2	GGG 145,215–145,376 TAA
41		1			1		1	3	ATG 150,523–150,687 CGT
42	1	1			1			3	ATG 156,442–156,720 CAG
43	1	1	1	1			1	5	GTG 156,795–158,059 GAT
44		1						1	ATC 158,841–158,872 TAT
45							1	1	ATG 159,489–159,717 CTC
46	1							1	ACT 159,766–159,856 TCA
47			1					1	GTG 159,902–161,017 CAG
48	1	1			1		1	4	GTT 162,003–162,155 AAG
49	1	1					1	3	AGT 162,200–162,346 ACC
50	1				1		1	2	GGC 162,395–162,580 CGC
51	1				1		1	2	AGG 162,630–163,354 CAA
52	1	1			1			4	GCC 163,409–163,601 CCT
53	1							1	AGT 164,357–164,483 TAG
54	1			1		1	1	4	GTA 164,535–164,638 TGA

Exon	Aug	Fgenesh	Gene ID	GeneMark	NCBI ORF Finder	SNAP	GenScan	Freq	Putative exons
55	1	1	1		1	1	1	6	ATG 171,858–172,109 AAT
56	1	1	1			1	1	5	ATA 172,638–173,544 TAG
57	1	1					1	3	TTT 173,595–173,796 AAG
58	1					1	1	3	ACC 175,601–175,767 CAG
59	1	1			1		1	4	GAA 175,823–175,990 AAG
60	1	1	1		1	1	1	6	TTG 176,127–176,587 AAG
61	1				1		1	3	AAT 176,635–176,855 ATG
62	1	1	1			1	1	5	GTG 177,462–178,324 TGG
63	1	1			1	1	1	5	CTC 178,742–178,898 TAA
64		1					1	2	GGT 178,949–179,208 GAG
65		1	1		1	1	1	5	AGA 179,497–179,834 TAG
66	1	1					1	3	ATG 186,166–186,251 TAG
67	1	1					1	3	CCG 186,311–186,377 AAG
68	1	1		1	1		1	5	GCC 186,429–186,597 AGG
69							1	2	CCC 186,650–186,761 GTG
70		1			1		1	3	GGC 195,906–196,081 TCA
71	1						1	2	TCA 199,928–200,056 TAG
72	1	1						2	GTT 200,103–200,328 AGG
73	1	1			1		1	4	AGT 200,380–200,697 CCA
74	1	1		1			1	4	GCG 202,518–203,302 GTG
75		1						1	AAA 204,579–204,798 CAA
76	1	1			1		1	4	AGG 204,852–205,293 CGG
77	1		1			1	1	4	GAA 205,334–205,470 AAA

Exon	Aug	Fgenesh	Gene ID	GeneMark	NCBI ORF Finder	SNAP	GenScan	Freq	Putative exons
78	1		1			1	1	4	GTT 205,890–206,052 AAG
79		1	1		1	1	1	5	GTA 206,752–207,218 TAA
80	1	1				1	1	4	GAA 207,270–207,396 GAG
81	1	1				1	1	4	ATC 207,443–207,649 TAG
82	1	1					1	3	ATG 216,340–216,420 GAG
83	1							1	CAC 216,468–216,641 TAA
84		1					1	2	ATG 216,785–217,216 CAG
85		1					1	2	AAT 219,157–219,300 AAA
86			1				1	2	GAT 219,346–219,526 CCA
87							1	1	ACC 219,571–219,696 AAA
88	1	1					1	2	AGG 220,229–220,365 GCC
89	1				1		1	1	GAA 220,411–220,512 GAG
90	1	1					1	3	GCG 220,987–221,347 AAC
91	1	1			1		1	2	ACG 221,401–221,594 CAG
92	1							2	AGT 221,640–221,801 GAG
93	1	1			1			1	ATT 222,839–223,066 GGG
94	1	1			1			2	TGC 223,127–223,144 TGA
95	1	1						2	GGG 223,146–223,238 AGG
96	1	1	1	1	1	1	1	4	GTG 224,097–224,359 GAG
97	1		1				1	3	GAC 224,406–224,441 GAG
98	1	1	1			1	1	4	GAG 224,489–225,799 AAA
99	1					1	1	2	GTC 225,846–226,037 CAG

Exon	Aug	Fgenesh	Gene ID	GeneMark	NCBI ORF Finder	SNAP	GenScan	Freq	Putative exons
100	1	1			1	1	1	3	ACA 226,082–226,345 GCG
101	1		1		1	1	1	4	GTT 226,392–227,239 TTC
102	1	1		1	1		1	5	ATG 230,664–230,828 CAG
103	1						1	3	ACC 230,874–231,019 TGC
104	1	1			1			2	TAC 231,057–231,238 CAA
105	1	1	1		1		1	4	GTC 231,294–231,596 GTT
106	1	1	1		1		1	5	CGA 231,643–231,886 CGG
107	1		1		1	1	1	5	GCG 233,510–233,706 CAG
108	1		1			1	1	5	GCA 233,751–234,939 GAA
109	1					1	1	4	GCC 234,984–235,218 AAG
110	1	1						3	TAT 235,960–236,204 TAA
111	1	1			1		1	2	GTT 236,251–236,581 TGA
Sum	83	68	27	11	50	34	80		
Sensitivity	0.741071	0.607143	0.241071	0.098214	0.446429	0.303571	0.714286		

Table A.1.2. *Harmonia axyridis* gene annotation using automated gene prediction software

The table provides the estimates of total number of exonic regions from the 198 assembled potential contig sequences of *H. axyridis* predicted by different software and an NCBI BLAST on different publicly available protein databases. The utilized gene prediction software included Augustus, Fgenesh (pipeline), Fgenes (pipeline), GeneMark, NCBI BLAST at Swiss-Prot/TrEMBL, the combined output of sensitivity predicted by Fgenesh and Fgenes and the combined output of Fgenesh, Fgenes and BLAST search output from Swiss-Prot/TrEMBL. The contig name provides the abbreviated name of the contigs, e.g., 'McKenna_945' represents 'McKenna_1kb_R1PF_paired_contig_945,' followed by the software that positively detected the exonic region. The estimates of the total number of true exons and the sensitivity of each software tool are given at the bottom of table.

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_945	1	1	1	1	1	1	1	1
McKenna_2304				1	1	1		1
McKenna_5400			1	1		1	1	1
McKenna_6103		1	1	1		1	1	1
McKenna_9307			1	1		1	1	1
McKenna_10941				1	1	1		1
McKenna_11720						1		1
McKenna_11981	1	1		1		1	1	1
McKenna_12024	1	1	1			1	1	1

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_12807	1	1	1	1		1	1	1
McKenna_15733			1		1	1	1	1
McKenna_17948	1		1	1		1	1	1
McKenna_18780	1	1	1	1		1	1	1
McKenna_18893		1	1	1		1	1	1
McKenna_19826	1	1	1		1		1	1
McKenna_20011	1	1	1	1		1	1	1
McKenna_20430								
McKenna_20898	1	1	1				1	1
McKenna_21703	1	1	1	1		1	1	1
McKenna_22059			1			1	1	1
McKenna_22233						1		1
McKenna_22342	1	1	1		1		1	1
McKenna_22609			1				1	1
McKenna_23880		1	1				1	1
McKenna_24087			1				1	1
McKenna_24576	1	1	1	1			1	1
McKenna_24719		1	1				1	1
McKenna_25162		1	1				1	1
McKenna_25931		1	1	1			1	1
McKenna_26312		1	1	1			1	1
McKenna_27165	1		1	1	1		1	1
McKenna_27215		1		1	1		1	1
McKenna_27981	1	1	1				1	1
McKenna_28250		1	1				1	1
McKenna_29860								
McKenna_30031		1	1	1			1	1

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_30862		1	1	1			1	1
McKenna_32062		1	1				1	1
McKenna_32234			1				1	1
McKenna_32781	1	1	1	1			1	1
McKenna_33418			1				1	1
McKenna_34157	1		1	1	1		1	1
McKenna_35031		1					1	1
McKenna_35973								
McKenna_36044			1				1	1
McKenna_37958			1				1	1
McKenna_38109	1	1	1	1			1	1
McKenna_38292		1	1	1			1	1
McKenna_39779		1		1			1	1
McKenna_40698	1	1	1	1	1		1	1
McKenna_41406		1	1				1	1
McKenna_43307		1		1			1	1
McKenna_43434	1	1	1				1	1
McKenna_44095			1				1	1
McKenna_44239		1					1	1
McKenna_44351	1	1	1	1			1	1
McKenna_44663	1		1	1	1		1	1
McKenna_44691			1				1	1
McKenna_44706		1	1				1	1
McKenna_44905	1	1	1	1			1	1
McKenna_45495	1	1	1	1	1		1	1
McKenna_45581				1				
McKenna_46662								

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_46973		1		1			1	1
McKenna_47539	1	1	1	1			1	1
McKenna_48071		1	1				1	1
McKenna_48771		1					1	1
McKenna_49810	1	1	1	1	1		1	1
McKenna_50287	1	1	1				1	1
McKenna_51177	1	1	1		1		1	1
McKenna_51726	1	1	1	1	1		1	1
McKenna_52730	1		1				1	1
McKenna_53591		1	1	1			1	1
McKenna_53778				1			1	1
McKenna_53784	1	1	1		1		1	1
McKenna_54774				1				
McKenna_54994								
McKenna_55760	1	1	1	1	1		1	1
McKenna_56446	1	1	1				1	1
McKenna_56469		1	1				1	1
McKenna_57285		1	1	1			1	1
McKenna_58185		1	1				1	1
McKenna_59278								
McKenna_59440	1	1	1	1			1	1
McKenna_59612								
McKenna_59660			1				1	1
McKenna_59965	1	1	1				1	1
McKenna_59987		1					1	1
McKenna_61856	1	1	1				1	1
McKenna_62152	1	1	1				1	1

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_64461		1					1	1
McKenna_64985	1	1	1				1	1
McKenna_65594		1					1	1
McKenna_65850		1	1				1	1
McKenna_67139	1		1				1	1
McKenna_67576		1					1	1
McKenna_67641		1	1				1	1
McKenna_68213	1	1	1	1	1		1	1
McKenna_69083		1	1	1			1	1
McKenna_69116		1	1				1	1
McKenna_70825			1	1			1	1
McKenna_70895		1	1				1	1
McKenna_70904	1	1	1				1	1
McKenna_71125		1					1	1
McKenna_71298		1	1				1	1
McKenna_71905		1	1	1			1	1
McKenna_72064		1	1				1	1
McKenna_72206			1				1	1
McKenna_72384			1		1		1	1
McKenna_72692				1				
McKenna_73025		1	1				1	1
McKenna_73302		1	1				1	1
McKenna_73464		1					1	1
McKenna_73770			1				1	1
McKenna_77757								
McKenna_79212								
McKenna_79748								

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_79978				1				
McKenna_80432	1			1				
McKenna_81616		1					1	1
McKenna_81771			1				1	1
McKenna_82098								
McKenna_82275	1	1	1	1	1		1	1
McKenna_83602		1	1				1	1
McKenna_83750		1	1			1	1	1
McKenna_84359						1		1
McKenna_84618						1		1
McKenna_87030			1			1	1	1
McKenna_87274		1	1			1	1	1
McKenna_89270		1	1			1	1	1
McKenna_89867						1		1
McKenna_90090		1	1			1	1	1
McKenna_90133		1				1	1	1
McKenna_91842		1	1			1	1	1
McKenna_92383			1			1	1	1
McKenna_92563		1				1	1	1
McKenna_93461		1	1			1	1	1
McKenna_93511	1	1	1			1	1	1
McKenna_93915		1	1			1	1	1
McKenna_94145	1		1	1	1	1	1	1
McKenna_97505		1				1	1	1
McKenna_97615	1	1	1	1	1	1	1	1
McKenna_97724						1		1
McKenna_100427		1	1				1	1

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_100712		1	1			1	1	1
McKenna_101287	1	1	1			1	1	1
McKenna_101506	1	1	1		1		1	1
McKenna_101568	1	1	1	1	1	1	1	1
McKenna_101612		1	1			1	1	1
McKenna_104261		1	1			1	1	1
McKenna_104318						1		1
McKenna_104580	1	1	1	1		1	1	1
McKenna_104719			1			1	1	1
McKenna_107371			1			1	1	1
McKenna_108449	1	1	1			1	1	1
McKenna_109359		1	1	1	1	1	1	1
McKenna_109885	1	1	1			1	1	1
McKenna_109920	1	1	1	1	1	1	1	1
McKenna_110020	1		1		1		1	1
McKenna_110460		1		1			1	1
McKenna_113570		1	1			1	1	1
McKenna_114066			1	1		1	1	1
McKenna_114665			1			1	1	1
McKenna_116773						1		1
McKenna_117319		1	1			1	1	1
McKenna_118324		1	1			1	1	1
McKenna_118374		1	1			1	1	1
McKenna_118825	1	1	1			1	1	1
McKenna_120940		1	1			1	1	1
McKenna_121348						1		1
McKenna_121359		1				1	1	1

Contig name	Augustus	Fgenes	Fgenesh	GenScan	GeneMark	BLAST Swiss-Prot/ TrEMBL	Fgenes + Fgenesh	Fgenes + Fgenesh + Swiss-Prot/TrEMBL
McKenna_121511						1		1
McKenna_123923			1	1		1	1	1
McKenna_124192	1		1		1	1	1	1
McKenna_125545		1	1			1	1	1
McKenna_125712	1	1	1			1	1	1
McKenna_125815	1				1	1	1	1
McKenna_126849	1	1	1			1	1	1
McKenna_127722		1		1			1	1
McKenna_128699		1	1			1	1	1
McKenna_129906		1	1		1	1	1	1
McKenna_130670	1	1	1			1	1	1
McKenna_131727		1	1			1	1	1
McKenna_132274			1			1	1	1
McKenna_134640		1	1			1	1	1
McKenna_71811	1	1	1				1	1
McKenna_128566		1				1	1	1
Sum	60	123	135	60	30	75	160	172
Sensitivity	0.337079	0.691011	0.758427	0.337079	0.168539	0.421348	0.898876	0.966292

Table A.1.3. Traditional gene annotation in the estimation of different genes from *Harmonia axyridis*

The "Clan" column indicates the clan to which the new CYP genes detected in *H. axyridis* are assigned. The "Family" column indicates the family to which the specific sequence was assigned. "CYP name" indicates the nomenclature assigned to specific sequences. "Remarks" shows the exon fragment or the specific region recovered from the entire contig. "Best insect hit" shows the specific CYP gene from the best insect match that best matches our query sequence. "Percent ID" gives the percentage identity of the previously described genes with *H. axyridis* genes. "Region" shows the amino acid regions of CYP genes detected in *H. axyridis* genes. "Contig_name" indicates the name of the contig in which the specific fragment was found. In the table, "aa" abbreviates amino acids.

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
2	306	CYP306A1		CYP306A1	53%	1–352	McKenna_12807
2	306	CYP306A1	adjacent to CYP18A1	CYP306A1	53%	5:163	McKenna_101568
2	306	CYP306A1		CYP306A1	59%	1–110	McKenna_97505
2	307	CYP307A1		CYP307A1	71%	1:367	McKenna_44905
2	307	CYP307B1	exon 1	CYP307B1	51%	1:75	McKenna_44706

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
2	307	CYP307B1	exons 2–3	CYP307B1	60%	270–641	McKenna_18893
Mito	49	CYP49A1	exon 1	CYP49A1	47%	1:65	McKenna_33418
Mito	49	CYP49A1	exon 1	CYP49A1	36%	7:54	McKenna_15733
Mito	49	CYP49A1	aa 196–291	CYP49A1	83%	4:75	McKenna_53778
Mito	49	CYP49A1	I-helix	CYP49A1	65%	1:30	McKenna_118324
Mito	49	CYP49A1	I-helix	CYP49A1	65%	1:57	McKenna_35973
Mito	49	CYP49A1	HEME	CYP49A1D	76%	1:99	McKenna_70895
Mito	49	CYP49A1	last exon	CYP49A1	81%	6–40	McKenna_121511
Mito	49	CYP49A1					McKenna_54335
Mito	301	CYP301A1	exon 1	CYP301A1	52%	5:57	McKenna_84359
Mito	301	CYP301A1	exons 2–3	CYP301A1	60%	7:161	McKenna_104580
Mito	301	CYP301A1	exons 3–4	CYP301A1	78%	213–347	McKenna_59440
Mito	301	CYP301A1	last exon	CYP301A1	76%	1:102	McKenna_37958
Mito	301	CYP301A1	last exon	CYP301A1	79%	1:102	McKenna_59965
Mito	301	CYP301A1					McKenna_32237
Mito	301						McKenna_80357
Mito	301	CYP301B1	exon 1	CYP301B1	45%	1:80	McKenna_28250
Mito	301	CYP301B1	exon 1	CYP301B1	44%	18:78	McKenna_38292
Mito	301	CYP301B1	exons 2–3	CYP301B1	50%	1:123	McKenna_117319
Mito	301	CYP301B1	exons 2–3	CYP301B1	48%	1:116	McKenna_25931

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
Mito	301	CYP301B1	exons 4–9	CYP301B1	62%	1:325	McKenna_30862
Mito	302	CYP302A1		CYP302A1			McKenna_59802
Mito	314	CYP314A1	complete	CYP314A1v1	48%	31–515	McKenna_124192
Mito	314	CYP314A1	complete	CYP314A1v2	46%	2:459	McKenna_45495
Mito	315	CYP315A1		CYP315A1	45%	27–315	McKenna_71811
Mito	315	CYP315A1		CYP315A1	45%	7:300	McKenna_40698
Mito	315	CYP315A1		CYP315A1	40%	14:160	McKenna_55760
Mito	315	CYP315A1		CYP315A1	52%	1:59	McKenna_2304
Mito	315	CYP315A1		CYP315A1	50%	1:65	McKenna_80432
4	4	CYP4G79		CYP4G14	78%	31–509	McKenna_97615
4	4	CYP4Q14		CYP4Q6	53%	20–413	McKenna_93511
4	4	CYP4Q15		CYP4Q6	65%	335–453	McKenna_110020
4	4	CYP4Q16		CYP4Q7v1	67%	22–150	McKenna_93461
4	4	CYP4Q17		CYP4Q7v1	65%	1:120	McKenna_73025
4	4	CYP4Q18		CYP4Q7v1	65%	11:135	McKenna_69083
4	4	CYP4Q19		CYP4Q7v1	63%	10:137	McKenna_65850
4	4	CYP4Q fragment 1		CYP4Q5	59%	1:55	McKenna_72206
4	4	CYP4Q fragment 2		CYP4Q9P	50%	4:58	McKenna_77757

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
4	4	CYP4Q fragment 3		CYP4Q3	52%	1:56	McKenna_11720
4	4	CYP4Q fragment 4		CYP4M7	52%	3:96	McKenna_54774
4	4	CYP4Q fragment 5		CYP4Q2	60%	4:63	McKenna_72692
4	4						McKenna_60607
4	4						McKenna_15994
4	4						McKenna_88883
4	4						McKenna_93842
4	4						McKenna_61221
4	4	CYP4AA1v1		CYP4AA1	48%	8–461	McKenna_18780
4	4	CYP4AA1v2a		CYP4AA1	38%	1:126	McKenna_22609
4	4	CYP4AA1v2b		CYP4AA1	54%	205–482	McKenna_56446
4	4	CYP4AW fragment 1		CYP4AW1	57%	5:72	McKenna_58185
4	4	CYP4BM fragment 1		CYP4C62	57%	9–173	McKenna_25162
4	4	CYP4BM fragment 2v1		CYP4BM1	59%	1:99	McKenna_22342
4	4	CYP4BM fragment 2v2		CYP4BM1	58%	24:86	McKenna_126849
4	4	CYP4BM fragment 3v1		CYP4BM1	53%	4–116	McKenna_27981
4	4	CYP4BM fragment 3v2		CYP4BM1	55%	1:54	McKenna_57285
4	4	CYP4BM fragment 4		CYP4J20	75%	5:78	McKenna_51726
4	4	CYP4BM fragment 5		CYP4BM1	53%	1:52	McKenna_104318
4	4	CYP4DW1v1		CYP4Q5	35%	25:305	McKenna_56469

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
4	4	CYP4DW1v2		CYP4BN11	29%	56–321	McKenna_59987
4	4	CYP4DW2		CYP4Q5	46%	4:107	McKenna_41406
4	4	CYP4DW3		CYP4Q5	46%	3:122	McKenna_73302
4	4	CYP4DW4v1		CYP4AB6	28%	1:181	McKenna_35031
4	4	CYP4DW4v2		CYP4AB6	25%	1:69	McKenna_65594
4	4	CYP4DW fragment 1		CYP4AX1	44%	1:93	McKenna_90133
4	4	CYP4 fragment 1		CYP4S4	42%	12:93	McKenna_82098
4	4	CYP4 fragment 2		CYP4Q1	41%	30:61	McKenna_44239
4	4	CYP4 fragment 3		CYP4BN11	44%	7:171	McKenna_48071
4	4	CYP4 fragment 4		CYP4AY2	39%	1:137	McKenna_128699
4	4	CYP4 fragment 5		CYP4Q12	43%	10:148	McKenna_89270
4	4	CYP4 fragment 6		CYP4BQ1	45%	8:131	McKenna_92563
4	4	CYP4 fragment 7		CYP4M14	40%	1–117	McKenna_113570
4	4	CYP4 fragment 8		CYP4BN4	41%	6:129	McKenna_70904
4	4	CYP4 fragment 9		CYP4BN4	53%	10:134	McKenna_97724
4	4	CYP4 fragment 10		CYP4BN12v2	51%	446–497	McKenna_89867
4	4						McKenna_63047
4	4						McKenna_41802
4	4						McKenna_58192
4	349	CYP349 fragment 1		CYP349A1	39%	10:81	McKenna_59612

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
4	349	CYP349 fragment 2		CYP349A1	36%	3:73	McKenna_43307
4	349	CYP349 fragment 3		CYP349B1	42%	14–92	McKenna_36044
4	349	CYP349 fragment 4v1		CYP349A2	63%	34–95	McKenna_104719
4	349						McKenna_106919
4	349	CYP349 fragment 5		CYP349A2	55%	51–102	McKenna_121359
4	349	CYP349 fragment 6		CYP349A2	52%	1:134	McKenna_48771
4	349						McKenna_51341
4	349						McKenna_114597
4	349						McKenna_74465
3	6	CYP6BS3av1		CYP6BS1	54%	8–167	McKenna_19826
3	6	CYP6BS3av2		CYP6BS1	54%	5:157	McKenna_72384
3	6	CYP6BS3bv1		CYP6BS1	54%	7:192	McKenna_82275
3	6	CYP6BS3bv2		CYP6BS1	53%	1:165	McKenna_125815
3	6	CYP6BS3c		CYP6BS1	66%	365–460	McKenna_107371
3	6	CYP6CR fragment		CYP6BK5	72%	1:66	McKenna_79748
3	6	CYP6FN1v1		CYP6BQ13v1	39%	5:501	McKenna_52730
3	6	CYP6FN1v2a		CYP6BQ13v1	40%	10:465	McKenna_47539
3	6	CYP6FN1v2b		CYP347A1	44%	61:97	McKenna_46662
3	6	CYP6FN2		CYP6BQ13v2	58%	4:93	McKenna_90090
3	6	CYP6FN fragment 1		CYP9A	60%	13:68	McKenna_84618

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
3	6	CYP6FN fragment 2		CYP347A1	52%	444–498	McKenna_81616
3	6	CYP6FP1v1		CYP6BQ13v1	53%	9:326	McKenna_49810
3	6	CYP6FP1v2a		CYP6BQ13v1	53%	1:284	McKenna_61856
3	6	CYP6FP1v2b		CYP6BL1	58%	1:49	McKenna_116773
3	6	CYP6FP2v1		CYP6BQ13v1	52%	2–363	McKenna_21703
3	6	CYP6FP2v2		CYP6BQ15	45%	1:179	McKenna_67641
3	6	CYP6FP3v1		CYP6BQ13v1	60%	11:373	McKenna_24576
3	6	CYP6FP3v2		CYP6BQ13v1	71%	12:185	McKenna_44691
3	6	CYP6FP4		CYP6BQ13v1	56%	1–356	McKenna_11981
3	6	CYP6FP5		CYP6BQ11	52%	1–403	McKenna_12024
3	6	CYP6FP6		CYP6BQ5	54%	14:202	McKenna_53784
3	6	CYP6FP fragment 1		CYP6BQ13v1	63%	8:191	McKenna_32062
3	6	CYP6FP fragment 2		CYP6BQ13v1	72%	1:55	McKenna_29860
3	6	CYP6FP fragment 3		CYP6BK11	52%	1–107	McKenna_5400
3	6	CYP6FP fragment 4v1		CYP6BQ13v1	60%	5:60	McKenna_59278
3	6	CYP6FP fragment 4v2		CYP6BQ13v1	60%	1:51	McKenna_71125
3	6	CYP6FP fragment 5v1		CYP6BK5	56%	1:25	McKenna_79212
3	6	CYP6FP fragment 5v2		CYP6BK5	56%	1:36	McKenna_79978
3	6	CYP6FP fragment 6		CYP6BQ15	43%	16–135	McKenna_24719
3	6						McKenna_33613

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
3	6	CYP6FQ1		CYP6BQ13v1	42%	11:173	McKenna_125712
3	6	CYP6 fragment 1		CYP6BQ4	45%	5–326	McKenna_17948
3	6	CYP6 fragment 2		CYP6BQ1	43%	2:78	McKenna_45581
3	6	CYP6 fragment 3		CYP6CE2	53%	1:61	McKenna_129906
3	6	CYP6 fragment 4		CYP6BQ13	42%	10–167	McKenna_20011
3	6	CYP6 fragment 5		CYP6BQ4	35%	1:177/154:199	McKenna_81771
3	6	CYP6 fragment 6		CYP6BQ15	42%	8:172	McKenna_130670
3	6	CYP6 fragment 7		CYP6BQ21	36%	1:124	McKenna_69116
3	6	CYP6 fragment 8v1		CYP6BQ13v1	46%	16:177	McKenna_109885
3	6	CYP6 fragment 8v2		CYP6BQ13v1	47%	25:190	McKenna_62152
3	6	CYP6 fragment 9v1		CYP6BS2	40%		McKenna_104261
3	6	CYP6 fragment 9v2		CYP6BS2	40%	55–106	McKenna_22059
3	6						McKenna_17693
3	9	CYP9Y2		CYP9Y1	47%	9:487	McKenna_44663
3	9	CYP9Y3v1		CYP9Y1	54%	1:467	McKenna_34157
3	9	CYP9Y3v2		CYP9Y1	54%	1:196	McKenna_67139
3	9	CYP9Y4		CYP9Y1	56%	5:458	McKenna_38109
3	9	CYP9BA1v1		CYP9Z2	44%	1:584	McKenna_68213
3	9	CYP9BA2v1		CYP9Z4	61%	10–159	McKenna_131727
3	9	CYP9BA2v2		CYP9Z4	61%	6:143	McKenna_30031a

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
3	9	CYP9BA1v2		CYP9Z6	41%	1:181	McKenna_91842
3	9	CYP9BA3		CYP9Z4	44%	6:143	McKenna_30031b
3	9	CYP9BA4		CYP9Z2	41%	12:356	McKenna_32781
3	9	CYP9BA5		CYP9Z1	42%	1:108	McKenna_44095
3	9						McKenna_60215
3	9	CYP9BB1		CYP9Z4	48%	2:389	McKenna_100427
3	9	CYP9BC1		CYP9Z5	41%	5:251	McKenna_114665
3	9	CYP9BD1P		CYP9Z4	47%	797–1227	McKenna_94145
3	9	CYP9BE1		CYP9Z6	55%	1:283	McKenna_50287
3	9	CYP9BE2		CYP9Z4	57%	1–154	McKenna_20898
3	9	CYP9BE3		CYP9Z2	40%	3–168	McKenna_109359
3	9	CYP9BE4v1		CYP9Z6	63%	1–129	McKenna_9307
3	9	CYP9BE4v2		CYP9Z5	65%	102–155	McKenna_121348
3	9	CYP9BE5v1		CYP9Z6	59%	2–120	McKenna_120940
3	9	CYP9BE5v2		CYP9Z5	57%	11:129	McKenna_118374
3	9	CYP9BE6v1		CYP9Z4	52%	2:134	McKenna_24087
3	9	CYP9BE6v2		CYP9Z4	52%	2:134	McKenna_114066
3	9	CYP9BE7P		CYP9Z4	42%	1:215	McKenna_108449
3	9	CYP9BF1		CYP9Z4	48%	1:198	McKenna_83602
3	9	CYP9BG1v1		CYP9Z4	41%	24:169	McKenna_87030

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
3	9	CYP9BG1v2		CYP9Z1	42%	23–117	McKenna_59660
3	9	CYP9 fragment 1		CYP9Z4	43%	7:272	McKenna_43434
3	9	CYP9 fragment 2		CYP9V1	45%	1–111	McKenna_128566
3	9	CYP9 fragment 3		CYP9Z2	45%	16–134/165–216	McKenna_101612
3	9	CYP9 fragment 4		CYP9Z5	38%	8:73	McKenna_73464
3	9	CYP9 fragment 5		CYP9V1	47%	5:54	McKenna_54994
3	9	CYP9 fragment 6v1		CYP9V1	44%	2:136	McKenna_27215
3	9	CYP9 fragment 6v2		CYP9Z1	55%	2:72	McKenna_39779
3	9	CYP9 fragment 7		CYP9V1	41%	2:87	McKenna_71298
3	9	CYP9 fragment 8		CYP9Z6	57%	10:69	McKenna_22233
3	9	CYP9 fragment 9		CYP9Z1	42%	1–223	McKenna_6103
3	9	CYP9 fragment 10		CYP9AC1	44%	11:139	McKenna_87274
3	9	CYP9 fragment 11v1		CYP9Z4	54%	2:111	McKenna_72064
3	9	CYP9 fragment 11v2		CYP9Z4	54%	2:111	McKenna_23880
3	9	CYP9 fragment 12		CYP9AC1	51%	1–148	McKenna_26312
3	9	CYP9 fragment 13		CYP9AS3	42%	290–588	McKenna_134640
3	9	CYP9 fragment 14v1		CYP9Z2	37%	1:254	McKenna_125545
3	9	CYP9 fragment 14v2		CYP9A20	35%	3:134	McKenna_92383
3	9						McKenna_44011

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
3	9						McKenna_58513
3	9						McKenna_133538
3	9						McKenna_62451
3	9						McKenna_111718
3	345	CYP345D4		CYP345D2	48%	4–312	McKenna_132274
3	345	CYP345D5		CYP345D3	40%	19–297	McKenna_73770
3	345	CYP345D6		CYP345D3	43%	58–321	McKenna_123923
3	345	CYP345D7					McKenna_84502
3	345	345K1v1		CYP345D3	44%	1:516	McKenna_44351
3	345	345K1v2		CYP345D1	48%	27:246	McKenna_101506
3	345	345K2		CYP345D2	48%	1:106	McKenna_32234
3	345	345K3v1		CYP345D3	47%	1:195	McKenna_71905
3	345	345K3v2		CYP345D2	49%	283–407	McKenna_110460
3	345	345K4v1		CYP345D3	44%	5:140	McKenna_53591
3	345	345K4v2	EXXR to HEME	CYP345D1	51%	2–270	McKenna_93915
3	345	345 new					McKenna_63104
3	345	CYP345 fragment 1		CYP345A2	43%	1:109	McKenna_70825
3	345	CYP345 fragment 2		CYP345A2	47%	1:51	McKenna_20430
3	435	CYP435A1v1		CYP9D8	31%	20–130	McKenna_109920

Clan	Family	CYP name	Remarks	Best insect hit	Percent ID	Region	Contig name
2	15	CYP15A1	exon 2	CYP15A1	60%	30:115	McKenna_83750
2	15	CYP15A1	exons 3–4	CYP15A1	53%	3:138	McKenna_118825
2	15	CYP15A1	exon 5	CYP15A1	67%	4:64	McKenna_67576
2	15	CYP15A1	exons 5–6	CYP15A1	68%	284–433	McKenna_100712
2	15	CYP15A1	exon 7	CYP15A1	67%	2:64	McKenna_64461
2	15	CYP15A1	exon 7, partial	CYP15A1	83%	1:26	McKenna_10941
2	18	CYP18A1	last exon	CYP18A1	68%	5:163	McKenna_101568
2	18	CYP18A1	exons 1–3	CYP18A1	64%	2–270	McKenna_945
2	303	CYP303A1		CYP303A1	63%	1:55	McKenna_27165
2	303	CYP303A1		CYP303A1	68%	1:306	McKenna_64985
2	305	CYP305G1	aa 127–346	CYP305A1	40%	1:221	McKenna_101287
3	435	CYP435A1v2		CYP9D8	31%	195:326	McKenna_51177
3	436	CYP436A1		CYP9X1	35%	1:245	McKenna_46973
3	436	CYP436A2		CYP9V1	32%	8:151	McKenna_127722

List A.1.4. List of Annotated *Harmonia axyridis* CYP Sequences

At least 94 different genes, based on unique exon counts

Sequenced by D. Nelson on April 13, 2012

Added 27 new sequences including 2 full genes on June 11.

CYP2 clan (16 sequences, probably 7 different sequences)

The annotated CYP sequences presents the detailed information on annotation. For interpreting the annotation, we presented the first few annotations with subheadings (such as for CYP genes); other annotations without subheadings could be interpreted similarly.

CYP Name: >CYP15A1 assembled sequence missing the first 28 amino acids from 6 contigs
Sequence Identity: 61% to CYP15A1 *T. castaneum*
Obtained by assembling contig(s): McKenna_83750
McKenna_118825
McKenna_67576
McKenna_100712
McKenna_10941
McKenna_64461
Remarks on protein sequence: Search for exon 1
Assembled CYP Sequence:
GPLWLPIVGCLFQFKKLKEKFGYQHLAWQELYNRYGEVVGTKLGRKYVVVVFG
EEAVAQILTREEFEGRPNGFIFRMRTFGRRLGLVFSEGDFWKKQRKFSIQRLKKF
GYGGKQMIEKVEEETSYLMEAFMKKCAEPILMHDAFDVSVINVLWTMIAGERF
DLEDAKLKELMGIIHDAFRMLDMSGGLLNQIPLLRYLAPRSSGYRQIMKVMQGL
WNFLEDTVKDHKNNFSTDNPKDLIEAFLQEMNTHNDDTFSDEQLIASCLDLLMA
GSETTSNTLSFAVVYMLENPEVMRRVQLEQDKVVGRNRCPSLQDSTRLPYTNAV
LKEVQRIANIPPVGIAHRAVRDAELFGYTIPEDTIVLTSLYSVHMDPKTWKDPHVF
RPERFLGEEGKTIRERAFLPFGYGKRQCLGISLAKANYFLFFTSLLHRFNLEKTPG
ELQPDINGYDGITLSPKPFKVKLVPR*

Special information: Note: *T. castaneum* exon 1 =
MLFFVTLVISLVLLFLILDTIKPRRYPP

Obtained by assembling contig: McKenna_83750
revised DRN 4/3/12
Sequence Identity: 60% to CYP15A1 *T. castaneum*
Assembled CYP Sequence:
GPLWLPIVGCLFQFKKLKEKFGYQHLAWQELYNRYGEVVGTKLGRKYVVVVF

GEEAVAQILTREEFEGRPNGFIFRMRTFGRRL

Obtained by assembling contig: McKenna_118825
revised DRN 4/3/12
Sequence Identity: 53% to CYP15A1 *T. castaneum*, exons 3–4
Assembled CYP Sequence:
GLVFSEGDFWKKQRKFSIQRLKFGYGGKQMIEKVEEETSYLMEAFMKKCAEPIL
MHDAFDVSVINVLWTMIAGERFDLEDAKLKELMGIIHDAFRMLDMSGGLLNQIP
LLRYLAPRSSGYRQIMKVMQGLWNFLEDTVKDHKNNFSTDNPKDLIEAFLQEM
NTHNDDTFS
Obtained by assembling contig: McKenna_64461
McKenna_10941 same as McKenna_64461
revised DRN 4/2/12
Sequence Identity: 67% to CYP15A1 *T. castaneum*
Assembled CYP Sequence:
GKRQCLGISLAKANYFLFFTSLLHRFNLEKTPGELQPDINGYDGITLSPKPFKVKL
VPR*

CYP Name: >CYP15A1
Obtained by assembling contig: McKenna_100712
McKenna_67576 same as McKenna_100712 exon 3
revised DRN 4/2/12
Sequence Identity: 68% to CYP15A1 *T. castaneum*, amino acids 284–431
Assembled CYP Sequence:
DEQLIASCLDLLMAGSETTSNTLSFAVVYMLENPEVMRRVQLEQDKVVGRNRCP
SLQDSVRLPYTNAVLKEVQRIANIPPVGIAHRAVRDAELFGYTIPEDTIVLTSLYS
VHMDPKTWKDPHVFRPERFLGEEGKTIRERAFLPFGY

CYP Name: CYP18A1 Query
Obtained by assembling contig: McKenna_945
revised DRN 4/2/12
Assembled CYP Sequence:
MLILNCAKSVFHDLHQGDLQTILFVFFGVLVLVRLFQKLRELYILPPGPWGYPIV
GSICS
LKKDLHVHFNDLAAEYGSLFSTRFGNQLIVVLSDYKMIRDIFRREEFTGRPNNEF
TKILDGYGLINIAGKLWKEQRKFVHEGLRHFGMSYLGTKKAQMETRITKEVEEF
LRMLKSKRGQKVDLNPYFAVSISNVICEILMSVRFSFDDKRFRRFMQLIDEGFRLF
GSLDKAVFIPIMRYLPGKWQTLNKIRQNRNEMGLFLQETIDEHRRTFNRDNIRDIL
DTYLLEIAKASDEGTEDCLFQGKDHGKFYS

CYP Name: CYP18A1
Obtained by assembling contig: McKenna_101568
Remarks: C-terminal minus strand
Sequence Identity: 68% to CYP18A1 C-terminal exon
Assembled CYP Sequence:

DVKLNGFHIPQDVQVVPLLHAVHMDPALWDEPEKFNPSRFVNAEGKVVKPEFFL
PFGVGRRVCLGEIMARMELFLFFSNFIHSFDVSVPEDEPLPSLTGIAGITISPKHYN
VVLEPRPNGWDIPSVSIRSTGSH*

CYP Name: CYP18A1 *H. axyridis* assembled sequence
Obtained from McKenna_945
Sequence Identity: 64% to CYP18A1 *T. castaneum* amino acids 64–311
Remarks:
Proline-rich region in the N-terminus is approximately 30 amino acids from the starting methionine
Adjacent to CYP306A1 as observed in *Drosophila* and *Daphnia*
Assembled CYP Sequence:
MLILNCAKSVFHDLHQGDLQTILFVFFGVLVLVRLFQKLRELYILPPGPWGYPIV
GSICSLKKDLHVHFNDLAAEYGSLFSTRFGNQLIVVLSDYKMIRDIFRREEFTGRP
NNEFTKILDGYGLINIAGKLWKEQRKFVHEGLRHFGMSYLGTKKAQMETRITKE
VEEFLRMLKSKR GQKVDLNPYFAVSISNVICEILMSVRFSFDDKRFRRFMQLIDE
GFRLFGSLDKAVFIPIMRYLPGKWQTLNKIRQNRNEMGLFLQETIDEHRRTFNRD
NIRDILDTYLLEIAKASDEGTEDCLFQGKDHGKFYS
(85-amino acid gap)
DVKLNGFHIPQDVQVVPLLHAVHMDPALWDEPEKFNPSRFVNAEGKVVKPEFFL
PFGVGRRVCLGEIMARMELFLFFSNFIHSFDVSVPEDEPLPSLTGIAGITISPKHYN
VVLEPRPNGWDIPSVSIRSTGSH*

>CYP303A1 *H. axyridis* assembled sequence
67% identical to CYP303A1 *T. castaneum*
McKenna_27165 exon 1
McKenna_64985 exons 2–3
MWLVVGLFIVILGLLVFLDTKKPKNYPPGPKWLPLLGSALEVIKERKRTGYLYV
ATAEMSKTYGPVVGLKVGKDLLVIVYGGTALKEFLTSEDLAGRPTGIFFEKRTW
GKRLGIMLTDSDFWQEQRRFVLRQLREFGFGRKNMSLMIEEESDHMVHDIKDL
MNNSGHLIADMESIFNIHVLNTLWTMLAGVRYSSQDKGLKELQGILGELFAHID
MVGAPFSQFPVLRFIAPELSGYKRYVKTHIHVWDFINRELKRHKESHNPAEPRDF
MDVYINILNSPDRKSSFTEDQLLAICMDMFMAGSETTSKTLGFMFLYLILNTDVQ
KKAQEEIDRVVGKNRFPSLEDRPKMPYMESCVLEALRMFAGRAFSVPHRAMRD
THLQGYFIPK

>CYP305G1
McKenna_101287
revised DRN 4/2/12
40% identical to CYP305A1 *T. castaneum* amino acids 127–346
RQFVVKHLKTLGLGKEIMSNLVKYEVANLLKILPNNNTAIKTLLAPCVINIFWSLT
TGSRFDVNDCRIEKLIAILGARSKVFDLAGGIMNALPWLRFIAPKKIGYEMICKLN
KELYTLFMKTIEEHYKTYDEDNDEDLINMYIKEIKKGNAHFTGIFNIQFFFSDEKL
VIVLLDLFIAGSQTTSTTLDFALMMMILRPELQEKLQKQLDDAFDKNVPIDYSQK
WR

>CYP306A1
McKenna_97505
revised DRN 4/3/12
59% identical to CYP306A1 *T. castaneum*
MFNFSLQMTSTFLIGLILVVLFLSWWRKRNQFLPPGPWNIPIIGSLLWLDPKKPYK
TLHDFAKRYGPIYGLYMGGIYTVVLSDAKLIKKVLSKEATTGRAPLYITHGLMK
GC

>CYP306A1
McKenna_12807
amino acids 396–531
revised DRN 4/3/12
53% identical to CYP306A1 *Leptinotarsa decemlineata*
GIICAEGPLWKDQRKLLFNFLRNIGAAKVSTKRTAMENLILKHVDGFVDYIESVG
ESPNFSPLEHLRHTIGSFMNEIVFGESWSKDDKTWIYLQHLQEEGIKYIGIAGPVN
FLPILR
FLPIFRKNLKFLQDGLETTHKLYDEIIANHRKTLKQQLDENPDFEPQNLLDTFLLE
KKKKENTPDRLFYNDQQLRYLLADVFGAGLDTTLTTLSWYLLFMSLHEDLQKL
VREELASVLQGRRPTMVDFQELPLFEASIAEVQRIRSTVPVGIPHGTIAPIEIEGFQI
PKETMIIPLQWSVHMDEKKWRDPEIFDPKRFLDENGSFCRSENFIPFQA

>CYP306A1
McKenna_101568
revised DRN 4/3/12
plus strand
53% identical to CYP306A1 *Leptinotarsa decemlineata*
adjacent to CYP18A1 tail-to-tail
GKRICIGDELAKMLLYLFASTILQKYRLSIDSEENVNLDGHCGITLSPESFKITFKK
VQA*

>CYP306A1 assembled sequence
McKenna_97505
McKenna_12807
McKenna_101568
52% identical to CYP306A1 Leptinotarsa decemlineata
MFNFSLQMTSTFLIGLILVVLFLSWWRKRNQFLPPGPWNIPIIGSLLWLDPKKPYK
TLHDFAKRYGPIYGLYMGGIYTVVLSDAKLIKKVLSKEATTGRAPLYITHGLMK
GCGIICAEGPLWKDQRKLLFNFLRNIGAAKVSTKRTAMENLILKHVDGFVDYIES
VGESPNFSPLEHLRHTIGSFMNEIVFGESWSKDDKTWIYLQHLQEEGIKYIGIAGP
VNFLPILRFLPIFRKNLKFLQDGLETTHKLYDEIIANHRKTLKQQLDENPDFEPQN
LLDTFLLEKKKKENTPDRLFYNDQQLRYLLADVFGAGLDTTLTTLSWYLLFMSL
HEDLQKLVREELASVLQGRRPTMVDFQELPLFEASIAEVQRIRSTVPVGIPHGTIA
PIEIEGFQIPKETMIIPLQWSVHMDEKKWRDPEIFDPKRFLDENGSFCRSENFIPFQ
AGKRICIGDELAKMLLYLFASTILQKYRLSIDSEENVNLDGHCGITLSPESFKITFK
KVQA*

>CYP307A1
McKenna_44905
71% identical to CYP307A1 *T. castaneum*
ALALCDWSSLQKTRRNIARAYCSPKLTSLQCDKVKAVALEELKVFLTELNKLPL
EEPVDIKPIVLKACANMFTKYMCSTSFAYDDKEFSLVVRYFDEIFWEINQGYAVD
FMPWLLPMYTGHMNKISNWATVIRSFILERIINEHEASLDYDGIPRDFTDALLMH
LAEDPNMNWQHIIFQLEDFIGGHSAVGNLIMVTLSSVIKYPEVAAKIQKEVDSVT
GGNRCPNLFDKEAMPYTEATIMEALRLASSPIVPHVATVDSNIGGYTVNKGTMIF
INNYELNIGDDYWSEPNKFKPERFISAGGHISKPAHFIPFSTGKRTCIGQKLVQYFC
FVILATLVQHFDLSVPVPPKLKPGCVAVHPDCFKFILKSRGHNNV*

>CYP307B1
McKenna_18893
McKenna_44706 exon 1
revised DRN 4/3/12
60% identical to CYP307B1 *T. castaneum*
MISLLLLSQFTMVLIVILMVIVLLVIYENKRSNSKLPTLDNGLLLPPVPFQLPLIGH
LHLMGGYDVPYKALTEIGRKYGNVVKLQLGNVKCVVVNDQRNIREAIINKSHHF
DARPNFARYEHLFSGNKQNSLAFCDWSEVQRTRRDMLKFHTFPRVSTGIFNSLES
IINKNTRGIIAKVDTGKPVNLKPIIIQHCANIFTKHFCTKQFSYNDKEFLNMVEDFD
EIFYEVNQGYAADFLPFLMPFHKGNLKRISSLTHNIRNFVLNRIIEDRYEKFNKEN
EPKDYVDDLLRHVKTEGEFDWETALFALEDIIGGHSAVGNFFMKLLGFLVKETE
VQKKIQKEIDNIISSEGRDILIKDRNRMPYTEATIYEAIRLIASPIVPRVANQNSSIDG
YLIEKDTLLLLNNHDLSMSNKLWLHPEKFMPERFIKDGRIVKPDHYLPFGGGKRS
CMGYKMVQLISLGILGGLLQHYTIHPVDKEEYKVPIGSLALPKNTFHFKFIKRFAP
SC*

CYP3 clan (104 sequences, at least 47 different sequences)
Note: Thirty-eight different CYP3 clan genes have been named; however, some are incomplete, and some of these incomplete genes may be joined to form whole genes.

There are 5 N-terminal CYP6 fragments and 5 CYP6 C-helix fragments
Thus, there are at least 5 genes in this set of fragments

>CYP6BS3av1
McKenna_19826
starting methionine not found
revised DRN 4/9/12
98% identical to McKenna_72384 PROBABLE ALLELE
54% identical to CYP6BS1 *T. castaneum*
TLFTACLLLLYIYSRITYTYWKSRGVPQLDPSFPFGDVTSVIFRLKNMGERTKEMF
DWARLRGHRYIGVYSFFRKGILLADPVLIREVMGRFNDRGIHYDEKNDPISAHLF
SLAGPKWKNLRTKLTPAYSPKQLRNLFNVVMDCGMKLIEFVEENV

>CYP6BS3av2
McKenna_72384

starting methionine not found
revised DRN 4/9/12
54% identical to CYP6BS1 *T. castaneum*
98% identical to McKenna_19826 PROBABLE ALLELE
TLFTACLLLLYIYSRITYNYWKSRGVPQLDPSFPFGDVTSVIFRRKNMGERTKEM
FDWARLRGHRYIGVYSFFRKGILLADPVLIREVMGRFNDRGIHYDEKNDPISAHL
FSLAGPKWKNLRTKLTPAYSPKQLRNLFNVVMDCGMKLIEFVEENV

>CYP6BS3bv1
McKenna_82275
revised DRN 4/9/12
98% identical to McKenna_125815 probable allele
54% identical to CYP6BS1 *T. castaneum*
45% identical to CYP6FP4 McKenna_11981

GRPIEIEIKETFARYNTDVIGSTALGLECNSLKDPNAEFRQIGKRAFTQNYVDIFRI
SIIRNMPRLAEFFGLGIFTPQVTEFFRRVVRETVEYREKNNVRREDFLQILIELKNN
ATSNALTMDEIAAQVFIFFLAGFETTSTTTCFAVYELAKNKEIQDRARLEIRKGLE
KHGGVLDYEILQEFTYLDMIVH

>CYP6BS3bv2
McKenna_125815
revised DRN 4/9/12
98% identical to McKenna_82275
53% identical to CYP6BS1 *T. castaneum*
DNTDVIGSTALGLECNSLKDPNAEFRQIGKRAFTQNYVDIFRISIIRNIPRLAEFFGL
GIFTPQVTEFFRRVVRETVEYREKNNVRREDFLQILIELKNSATSNALTMDEIAAQ
VFIFFLAGFETTSTTTCFAVYELAMNKEIQDRARLEIRKGLEKHGGVLDYEILQEF
TYLDMIVH (1)

>CYP6BS3c
McKenna_107371
revised DRN 4/2/12
66% identical to CYP6BS1 *T. castaneum*
Significantly better than any other *T. castaneum* match
EAMRKYPAAPFYLRRCPERYQLPDSNVVIEEGFNFVVSSLGLHRDPEYFPDPDRF
DPDRFSEENRSKIWDYTYMPFGSGPRICI

>CYP6CR fragment
McKenna_79748
revised DRN 4/9/12
53% identical to CYP6M6 *Aedes aegypti*
63% identical to CYP6CR3 *Dendroctonus*
both ends run off the contig

VETSATTATFALYEIALNEDIQNKLREEIRTVLKKHNGMTYEAFMEMNYLEQVIK
ETLRKYPPVPLAP

>CYP6 fragment 5
McKenna_81771
revised DRN 4/9/12
runs off the contig
35% identical to CYP6BQ4 *T. castaneum* N-term
40% identical to CYP6CR3 *Dendroctonus*
39% identical to McKenna_17948 CYP6 fragment 5
MDNSFSPIQHAIVPVVICILISILIFFKWSFCYWSRQGIPCVPNPIIPFGNLITVVQQK
ETLFELTKKIYLDAKKIKLKHLGMYIFAKPIWIPVDSDLIRNILIKDSSHFMNHGM
YINERDDPLTGNIFLLEDEKWKHTRAVTTKLFSPAKMKTILEFMLKSVQDIEEVL
NRA

>CYP6FN1v1 Query
McKenna_52730
revised DRN 4/8/12 complete gene
39% identical to CYP6BQ13v1 *T. castaneum*
MTVATTISIIIILLGLLVTLVKRKYTFWKNRNVQTPPFKFGWGNFQDPILQRRCFS
DTLKDFYRFFKSRQLQHGGLYSFIFPVYVPIDLDIIKSILLVDFDHFAHRGFYMNE
VDDPLSATIFALNEDKWRWVRAKFSPAFSPAKLKVMFDSIMRCGEEFADTIGEV
CNGAVDITELNGGYTMNIIGQCAFGIECNSLKEPKNEYIVNGTAFFERSLLENIKIQ
LTEIVPELMKKLHISITRPSVTDFYKRLTLQAMEYREKSKIVKKDLLDLLIKMKN
MNESNDDVDNLDKTAKKLPFGEIVGLAYSMFLASFETTSIAISFTLYLLALHRDIQ
EKARDEVIRVMKKYNGKLTYESMSEFLYLDMVIDESMRLYPPVHHLYRMCTKD
YKVPGTDLILRKGQKVFVPVIGIHHDEEYYPDPEKFDPERFSTENSQDRNPITFMP
FGIGPKYCLAQRFGSLQVKIALIMVLQRFEMSTNEKTKNPLEFSTGNHITLFPKDG
IWLNMKKIIE*

>CYP6FN1v2a Query
McKenna_47539
revised DRN 4/8/12
40% identical to CYP6BQ13v1 *T. castaneum*
97% identical to McKenna_52730
MTVATTISIIIILLGLLVTLMKRKYTFWKNRNVPTPPFKFGWGNFQDPILQRRCFS
DTLKDFYRFFKSRQLQHGGLYSFIFPVYVPIDLDIIKSILLVDFDHFAHRGFYMNEI
DDPLSATIFALNEDKWRWVRAKFSPAFSPAKLKVMFDSIMRCGEEFADTIGEVC
NEAVDITELNGGYTMNIIGQCAFGIECNSLKEPKNEYIVNGTAFFERSLLENIKIQL
TEIVPELMKRLHISITRPSVTDFYKRLTLQAMEYREKNKIVNKDLLDLLIKMKNM
NESNDDVDNIDKTCAKKLPFGEIVGLAYSMFLASFETTSIAISFALYLLALHGDIQ
EKARDEVIRVMKKYNGKLTYESMSEFVYLDMVIDESMRLYPPVHHLYRMCTKD
YKVPGTDLILRKGQKVFVPVIGIHHDEEYYPDPEKFDPERFSIENSQDRNPITFMPF
GIGPKYCL

>CYP6FN1v2b
McKenna_46662
revised DRN 4/9/12
98% identical to McKenna_52730
44% identical to CYP347A1 *T. castaneum* C-term
This sequence appears to be part of the CYP6FN1v2 allele
AQRFGSLQVKIALIMVLQKFEMSTNEKTKNPLEFSTGNHITLFPKDGIWLNMKKII
E*

>CYP6FN2
McKenna_90090
revised DRN 4/9/12
70% identical to McKenna_52730
58% identical to CYP6BQ13v2 *T. castaneum*
ESMRYYPPVHVLNRVCTKDYQVPGSDLLLRKGQKVLIPVLGIQRDEEYYPDPDT
FNPERFSDENSTGRNPATFLPFGIGPRNCI

>CYP6FN fragment 1
McKenna_84618
revised DRN 4/9/12
90% identical to McKenna_90090
60% identical to CYP9A *Manduca sexta*
possible pseudogene because the exon ends are missing
VLGSDLLLRKGQKVLIPVLGIQRDEEYYPDPDTFNPEKFSDKIS

>CYP6FN fragment 2
McKenna_81616
revised DRN 4/9/12
52% identical to CYP347A1 *T. castaneum* C-term
56% identical to CYP6FN1v2b McKenna_46662
56% identical to CYP6FN1v1 McKenna_52730
GLRFGMVQVKTTIVSIIDNYEISLNKKTKYPLEFDTGHPLTLSPKGGIWLNLKKIE
K*

>CYP6FP1v1
McKenna_49810
revised DRN 4/8/12
53% identical to CYP6BQ13v1 *T. castaneum*
GRFTTDVIGTCAFGIECNSLENPDNEFRLKGKAIFERPKDFWMIMYERFFIYLPNL
MQFLNLKYIDKEVTNFFVGITKKTIEYREKNGVRRKDIMDLLIQLKNNMKLADN
DQTPVDKDLPEEESGISVDEIAGQAFLFFEAGFETSSTAMTFCLYELASNKDVQD
KLRQEINEVLAKYDNKITYDAIMDMSYLEMVIQESLRKYPPIPTFRRVCTKSYRV
PGTEIVLQKGANVLIPVYGIHYDPLYYPEPEKFIPERFSDENKKSRHPFAFLPFGEG
PRMCIGIRFGLMEAKVGIITLLKRYEFEVSSKTKQPLEWNPKNFVLSAKGEIWLK
HKKIH*

>CYP6FP1v2a
McKenna_61856
revised DRN 4/8/12
97% identical to McKenna_49810
56% identical to McKenna_11981
53% identical to CYP6BQ13v1 *T. castaneum*
GRFTTDVIGSCAFGIECNSLENPDNEFRLKGKAIFERPKDFWMIMYERFFIYLPNL
MQFLNLKYIDKEVTNFFVGITKKTIEYREKNGVRRKDIMDLLIQLKNNVKLADN
DQAPVDKDLPEEESGISVDEIAGQAFLFFEAGFETSSTAMTFCLYELASNKDVQD
KLRQEINEVLEKYDNKITYDAIMDMPYLEMVIQESLRKYPPIPTFRRVCTKSYRV
PGTEIVLQKGANVLIPVYGIHYDPLYYPEPEKFIPERFSDENRKSRHPFAFLPFGEG
PRMCI

>CYP6FP1v2b
McKenna_116773
revised DRN 4/9/12
96% identical to McKenna_49810
58% identical to CYP6BL1 *T. castaneum*
Probable last exon of McKenna_61856
GIRFGLMEAKVGIITLLKRYEFEVSSKTKQPLEWNPKNFVLSAKGEIWLKHKRIN*

>CYP6FP2v1
McKenna_21703
revised DRN 4/7/12
Missing the last 2 exons
62% identical to McKenna_11981
52% identical to CYP6BQ13v1 *T. castaneum*
57% identical to CYP6FP1v2a
96% identical to McKenna_67641
86% identical to McKenna_24719
MELFNTFTLLVCTTFLLIWFWLKNNYNFWKKRNIENSHYEYFWGSMKEVFLQK
DTFSDSTRKIYREFKDKGVRHGGLFFLWVPLYMPIDIELVKSIIQVDFQHFVDRGI
YVNEKADPLSAHLFSLQGKKWKLLRNKLTPTFTSGKIRMMFETLVDCTKGLEKL
MDKEMGKSVDIKDILGRFTTDVIGSCAFGLNCNSLEDPKSEFRVRGKSLFEKELG
RSIKDAILFLLPNLMKKLNMKMIPEDISNFFMTVVKDTVEYREKNNISRKDFMDL
LIQLKNKGKLVDDDQIGTEQITEAENYITLDEICAQAFIFFEAGFETSSTAMTFCLY
ELAKNKEIQKKAREEMRIVLSRHEGKLTYDAAMEMHYVEQVIN

>CYP6FP2v2 Query
McKenna_67641
revised DRN 4/9/12
45% identical to CYP6BQ15 *Leptinotarsa decemlineata*
96% identical to McKenna_21703
85% identical to McKenna_24719
57% identical to McKenna_12024
52% identical to McKenna_11981

MELFNTSTLLICTTFLLIWFWLKNNYNFWKKRNIESSHYEYFWGSMKEVFLQKD
TFSDSTRKIYREFKDKGARHGGLFFLWVPLYMPIDIEIVKSIIQVDFQHFVDRGIYL
NEKADPLSAHLFSLQGKKWKLLRNKLTPTFTSGKIRMMFETLVDCTKGLEKLMD
KEMGKSVDIKDIL

>CYP6FP3v1
McKenna_24576
revised DRN 4/7/12
missing first exon
79% identical to McKenna_11981
60% identical to CYP6BQ13v1 *T. castaneum*
GRFTTDIIGSCAFGIECNSLEDPKSIFRQKGREFFDADLKQNMKNFSSLFVPQIMR
AFKITLIPRDISSFFINLVKDTVEYREKNNVIRKDFMQLLIQLKNEGKLVDDEKME
AEKINEDENRITMDEIAAQAFVFFQAGFETSSTTMTFCLYELARNKDIQEKLRKEII
QTLDRHGGKITYENVHEMVYLDQVINETLRKYPPLPNLNRVCTKEYKVPGTDLV
LEKGMQVIIPVFGIHRDPEYYPDPDKFDPERFTEENKRNRHQFSFLPFGEGPRICIG
LRFGLMQTKVGLIALLSKYEFSVNKKTIEPLQFKGTSFIMTTKGDIWLDFKKISESI
*

>CYP6FP3v2
McKenna_44691
revised DRN 4/9/12
98% identical to McKenna_24576 possible allele
80% identical to McKenna_32062
71% identical to CYP6BQ13v1 *T. castaneum*
ETLRKYPPLPNLNRVCTKEYKVPGTDLVLEKGMQVIIPVFGIHRDPEYYPDPDKF
DPERFTEENKRNRHHFSFLPFGEGPRICIGLRFGLMQTKVGLIALLSKYEFSVNKK
TIEPLQFKGTSFIMTTKGDIWLDFEKISESRQ*

>CYP6FP4 Query
McKenna_11981
revised DRN 4/7/12
runs off the end of the contig
56% identical to CYP6BQ13v1 *T. castaneum*
56% identical to CYP6FP2v1 McKenna_21703
71% identical to McKenna_12024
78% identical to CYP6FP3v1
80% identical to McKenna_32062
82% identical to McKenna_44691
LLTAIFLIYKWVNYKLTYWKRRDLLVPDKNYLKDTLDAVLSRNAFADRVLALY
RDFKSKGVHHGGNYVMLMPQYVPMDLDIIKSIMQVDFQHFVDRGVYMDEKND
PISAHLFSLTGAKWRILRHKLTPTFTSGKMKMMFGTLLDCTKGLHKVLEKTNGK
EVDIKDILARFTTDIIGSCAFGIECNSLDDPNSIFREKGVKIFDVDSKQNVIQFVSLV
MPQILKIFNIPMVPSDVSSFFINLVKDTVDYREKNNITRKDFMDLLIQLKNKGKLV
DDQEVEINENENRITIDEIAAQAFIFFEAGFETSSTTMTFCLYELARNKDVQEKLR

KEILETLDRHGGKITYENIHSMVYLDQVIKETLRKYPPLPQLTRLCTKEYKVPGTD
LILEKGMQVMIPVMGIHRDADYYPDPDKFDP
ERFSEENKDSRHQYSYLPFGEGPRVCI

>CYP6FP5 Query
McKenna_12024
revised DRN 4/7/12
70% identical to CYP6FP4 McKenna_11981
52% identical to CY6BQ11 *T. castaneum*
63% identical to CYP6FP2v1 McKenna_21703
85% identical to CYP6FP3v1 McKenna_24576
Missing the last 2 exons
MEILNTFNLLVLTGVFLIYKWVNNKLNYWKSRGLQVSDISPLENLLNIIFARNSFA
DRGLQLYTDFKSKGIEHGGFYILLKPFYMPINLDLVKSILQVDFQHFVDRGVYMD
EKNDPLSAHLFSLTGKKWKNLRYKLTPTFTSGKMKMMFETVVDCTKGLHKLM
DKTNGKEVDIKDVLGRFTTDIIGSCAFGIECNSLEDPKSIFRQKGKEFFEADLIRNA
KFFMAFSFPQFMKTFNIAFIPRDISSFFMNLVKDTVEYREKNNVKRKDFMQLLIQ
LKNEGKLVDDEKMEAEKINEDENRITMNEITAQAFVFFQAGFETSSTTMTFCLYE
LARNKEIQKKLRTEIIQTLDRHGGKITYENVHEMTYLDQVIN

>CYP6FP6
McKenna_53784
revised DRN 4/9/12
52% identical to CYP6BQ5 *T. castaneum*
53% identical to CYP6FP5
54% identical to CYP6FP3v1 McKenna_24576
ARFTTDIIGSVAFGVKCNSLENPKCEFRERAQAVFRTDILAAVKQGALYAFPELF
QSLNVKTLDPKSSEFFMDVIKDTVNYREKNNVFRRDFMHLLIQLKNSNKLMDTG
NKTVEDTSITVEEIAAQAFIFFEAGYESSSNTMTFALYELARHMDIQDRVREEIFQ
VLDKHGGVLNYDAVEEMKYLDQVLC

>CYP6FP fragment 1
McKenna_32062
revised DRN 4/9/12
63% identical to CYP6BQ13v1 *T. castaneum*
80% identical to CYP6FP3v2
80% identical to CYP6FP4 McKenna_11981
ETLRKYPPLPNLNRICTKEYKIPGTDVILEKGMQVMIPVVGIHRDPEYYPDPDKFD
PDRFTEENKKSRHHFSFIPFGEGPRICIALRFGVLQSKVGLIALLSKYEFSINKKTV
EPIQFKASSIVTAAEGEIWLDFSKIPEGN*

>CYP6FP fragment 2
McKenna_29860
revised DRN 4/9/12
79% identical to CYP6FP3v1
79% identical to CYP6FP3v2 McKenna_44691

72% identical to CYP6BQ13v1 *T. castaneum*
ETLRKYPPGPILNRICTKEYKVPNTDIVLEKGIRVMIPVMGIHRDPEHYPDPEKFD
PERFSEENKRNIKPFTFLPFGEGPRVCI

>CYP6FP fragment 3
McKenna_5400
revised DRN 4/4/12
Last 2 exons (runs off the contig)
52% identical to CYP6BK11 *T. castaneum*
53% identical to McKenna_44691
54% identical to McKenna_24576
59% identical to McKenna_11981
YPPVPLAPRLCTKDYKVPGCNTVIEKDTLVMVPITGVQRDADIYPDPDKFDPERF
GEGCS
IPSMAFLSFGEGPRLCIGKRFGTLQTKVALATVLKNYQVTMNREKTEAPLQFAPK
SLITTPKGDVWLNVKRIQ*

>CYP6FP fragment
McKenna_33613
49% identical to CYP6FP fragment 3 McKenna_5400
This part is immediately downstream of the heme signature
GMRVGITQVKVILSTILKNHRVTFDRQKTPYPLQYSRKSMIATPQGNIWFHIEKA*

>CYP6FP fragment 4v1
McKenna_59278
revised DRN 4/3/12
98% identical to McKenna_71125 (1-amino acid difference) probable allele
60% identical to CYP6BQ13v1 *T. castaneum*
GTRFGMMQTKVGLTALLLNYDFDVSEDTKEPIEFDPKSFILLTKGDIWLKYRRID
TTETEL*

>CYP6FP fragment 4v2
McKenna_71125
revised DRN 4/9/12
1-amino acid difference from McKenna_59278
60% identical to CYP6BQ13v1 *T. castaneum* C-term
GTRFGMMQTKVGLTALLLNYDFDMSEDTKEPIEFDPKSFILLTKGDIWLKYRRID
TTETEL *

>CYP6FP fragment 5v1
McKenna_79212
revised DRN 4/9/12
79% identical to 44691 C-terminus
56% identical to CYP6BK5 *T. castaneum*
GLRFGMMQTKVGLVTLLSNYEFSVNKKTIEPIEFKETSFIITTKGN

>CYP6FP fragment 5v2
McKenna_79978
56% identical to CYP6BK6 *T. castaneum*
79% identical to McKenna_44691
100% identical to McKenna_79212
GLRFGMMQTKVGLVTLLSNYEFSVNKKTIEPIEFKETSFIITTKGN

>CYP6FP fragment 6 Query
McKenna_24719
revised DRN 4/9/12
43% identical to CYP6BQ15 *Leptinotarsa decemlineata* N-terminus
86% identical to McKenna_21703
85% identical to McKenna_67641
MELFNTFTILVCTTILLIWLWLKNNYDFWKKRNIENAHYAYFWGSLKEVFLQKD
TFSNSTRKIYREFKEKGARHVGIFLLWVPLYMPIDIEIVKSIMQVDFQHFVDRGVY
VNEKADPLSAHLFSLEGKRWKLLRNKLSPTFTSGKMKMMFETLVDCTKGLKKV
MDEEMGKSVDIKDIL

>CYP6FQ1
McKenna_125712
revised DRN 4/9/12
42% identical to CYP6BQ13v1 *T. castaneum*
43% identical to CYP6FP3v1
49% identical to CYP6FP4
DICLTMDEVVAQAYTYLFTGSETASSTLIMLFYECSLNKNIQHKLQKDIDDALAA
SGGEITYEAILEMKYLNMVISETLRKYPTLQFLMRKAVEDYRIEKINLQIKKGTRV
YVSIQGMHRDPKYYPNPEVFDPERFSKENQAERHPLTYIPFGEGPRNCIGKRLGY
FMIAVGAIHMFRKFSISPYKDSGKQLKLHPYSYTIRPEGSLMLNVTPR*

>CYP6 fragment 1
McKenna_17948
revised DRN 4/8/12
45% identical to CYP6BQ4 *T. castaneum*
runs off the contig end
40% identical to CYP6FP4 McKenna_11981 N-terminus to I-helix
MALSLIIVEILVALLVVSAAMVFALYKYNFTYWRRKKVLHIPHPTIPFGNINDLVN
QSESLAELVINTYNFIKSHNAKHGGLYFFGKPIWMPVDLDLLNHVMIKDFTHFVN
HGFYVNAEHDPLSAHLFSLEDDNWRRMRAALTHTFTTGKIKMMFPIMIQNAQN
LEKIIKKAEAAGEAVNVKDLISRFTIDVITSTAFGLEINSIENPDADFRKAGNRFFV
DSGYEAFRNLLSFIIPRKFLDAIKFKLIKPDITEYFVNIVKKTIEYRENNNIERPDFIQ
LLIQLKNSGKVMNTGEETVKSKSVADKQMDMHLSVEEMAAQIFVFFLAGFETSA
TTA

>CYP6 fragment 2
McKenna_45581
revised DRN 4/9/12

43% identical to CYP6BQ1 T. *castaneum*
42% identical to CYP6FN1v1 McKenna_52730 N-term
MLSLLTLGFLILTCLYLINKWLRGKFSFWSERGVITPPVNLGWGNYQEVILQKKS
FSQALHEFYKLFKAKGVLHGGLYSFTSPIYMPVDINI

>CYP6 fragment
McKenna_17693
32% identical to CYP6BR2 sequence56, XP_969813.1 N-terminus CYP 6 family
36% identical to CYP6AZ1 *Mayetiola destructor*
MAFLIFLRIYIHFKKTYQYWQERGIKVVQSAVFPLGNFWTLIKRDIGVGHMLAE
WYNEVDAEAVGLYAINEPFLLVR

>CYP6 fragment 3
McKenna_129906
revised DRN 4/9/12
45% identical to CYP6BQ4 *T. castaneum*
may join with McKenna_45581
DPDIIKDILVKDFNSFMDRGIFMDKNDVLTGQLWKLPGYKWKPLRAKMSTCFTI
GKLKMMFSNLLKSGSEMEIFLNKAAEAESVDVSISL

>CYP6 fragment 4
McKenna_20011
revised DRN 4/9/12
42% identical to CYP6BQ13v1 *T. castaneum* N-terminus to C-helix
46% identical to McKenna_67641
46% identical to McKenna_21703
48% identical to CYP6FP4 McKenna_11981
MADPLSLALIATCVVLFCTWLRYKFTHWSRKNIPTPSLQLFLKLFKMGILQQNAY
ADRDLYLYNNFKQQNFLHGGTYFYHLPIYIPIDLKIIKHILQIDHQHFTDRGLYVN
EKADPLSAHLFSLGGERWRSLRKKLTPTFTSGKLRMMFDTMQDCTCDLQRVMT
ENIGTDFDVRDVM

>CYP6 fragment 6 Query
McKenna_130670
revised DRN 4/9/12
42% identical to CYP6BQ15 *Leptinotarsa decemlineata*
42% identical to CYP6FP2v1 McKenna_21703
MEFYTATLTLLTATLILFWIWTRKVFSFWSNCGVETTPFSYLWGHLRTPFFHGPA
LGDRIKLIYNHLKSKNLKHGGFYLLFDPIYVPMDLDICKAILQTDFQHFVDRGGR
VFNGDPLTAHLLNLKGKKWKKMRSKLTPAFSSGKMKMMFETLLACTNSLDKIM
DELLSSDIDIKDVL

>CYP6 fragment 7
McKenna_69116
revised DRN 4/9/12
32% identical to CYP6FB1 Colorado potato beetle

36% identical to CYP6BQ21 *Leptinotarsa decemlineata* before the I-helix
VDDLCRRFSADIIGSSLFGLEVKSFKDRDDDFLRMSTCLLSTFNRKNSIKAALQV
IAPNLIDVFSFFKIETVNKYALAFVYNMVEKIIDFREKTGNVGKDMMQLLIQLKK
FGKVDGDNGEK

>CYP6 fragment 8v1
McKenna_109885
revised DRN 4/9/12
98% identical to McKenna_62152 PROBABLE ALLELE
46% identical to CYP6BQ13v1 *T. castaneum*
48% identical to CYP6FP3v1 McKenna_24576
GSTFFNRTFFENVKLYFSDVFPGIMKKLHVRVSRPSTIKFFERLTRETIEYREENNII
RKDFMHLLIQLKNSGKLIDSEEVGNIGKSENGETISFADLVGQAYVFFLAGYETPS
NTISFTLYFLSVFKEVQEKCRKEIKTVLDKYGGNLTYQAVMELHYLESVIN

>CYP6 fragment 8v2
McKenna_62152
revised DRN 4/9/12
98% identical to McKenna_109885 PROBABLE ALLELE
47% identical to CYP6BQ13v1 *T. castaneum*
GSTFFNRTFFENVKLYFSDVFPGIMKKLHVRVSRPSTIKFFERLTRETIEYREQNNI
IRKDFMHLLIQLKNSGKLIDSEEVGNIGKSENGETISFADLVGQAYVFFLAGYETP
SNTISFTLYFLSVFKEAQEKCRKEIKTVLDKYGGKLTYQAVMELHYLESVIN

>CYP6 fragment 9v1
McKenna_104261
same as McKenna_22059
40% identical to CYP6BS2 *Dendroctonus ponderosae*
last exon
GMRFAILEIKITLVNLLNNYVLRINSNTKLPASLAKQGVLLAPANPIWVDFEPARL
AVRINRYPNDQFE*

>CYP6 fragment 9v2
McKenna_22059
same as McKenna_104261
40% identical to CYP6BS2 *Dendroctonus ponderosae*
last exon
GMRFAILEIKITLVNLLNNYVLRINSNTKLPASLAKQGVLLAPANPIWVDFEPARS
AVHINRFPNDQFE*

>CYP9Y2
McKenna_44663
revised DRN 4/2/12
42% identical to CYP9AB1 *T. castaneum*
47% identical to CYP9Y1 *T. castaneum*
53% identical to CYP9Y McKenna_34157

54% identical to McKenna_38109
MMLLLLVILLILIIIGHLLTTLVPAYYYWKIRGVPHVSLLTVLKRMIMKDKPYFDL
VISDHNKFAGKRYYGSYMSVMPALFIRDLDLIKKILIKDFEHFTNRLDILNASSDPI
FRNNLLVAKGKKWKALRSTISPVFTTAKMKAMYVLIAEEAKKFVVYYNSMNQD
VIEVEMKEMYSKFTNDVIASCSFGVQIDSLKDPKNDMFLAGKVITQTTPFNVAWI
LISMLFPKLLKHLKVNIFPKQITQYFKNMISATIRSREEGKITRPDLIQLLVEAKKG
KLKRENTPQAEETGFATVRESTDLQVDSSLEITDEIIVAQAIIFFLAGFEASSSLLSC
LSYELALNPKIQQKLIEEIDEHLASSDIITYEMITKMKYLDQVVSEALRKWPTAFL
QLRICTKDYVIKPEEELEVPLKVSKGSLVVIPIIGTHYDERYFENPDAFIPERFSDGE
NIVPQSFAPFGIGPRNCVGSRFALLEMKCILANILSKFEMHITKKTNIPFKKLNTIPF
AIDGGITLALKKRKNTSR*

>CYP9Y3v1
McKenna_34157
sequence revised 4/5/12
54% identical to CYP9Y1 *T. castaneum*
49% identical to CYP9D5 *T. castaneum*
78% identical to McKenna_38109
MIFTVFVLILIVYVANRVIPPIYYWKKRNIIYVNPLFRVYQVFFGIKSFAEIVQEAY
NEYPDKRYYGSYQFLKPSLFVRDLDLIKQITIKDFDHFTDHVDILNSNNDPIFSKN
LFSLKGREWRELRSTLSPAFTSSKMKAMFVLISEASKKFVEHFEAKKEEIIEVEMK
ETYSKFTNDVIATCAFGINCDTLENPENEFFTMGKAVTRSTFLRMMRALVRMLFP
TIFEVLKIPTFPKDVTNFFKKIITDTLTTREKEGTIRPDLIHLLMEARKGKLQQETSQ
VTEDTGFATALEVKDTKIKSDLEISDDLITAQALIFFLAGLETSSALLSFLSYELAK
NPDIQQKLISEIDDNLSSSEASISYEKLAKMKYLDQVVSEALRKWTPGFGLNRICT
KEYTIPPIKEGEIPLTLSKGCFITILVIGIHYDPQFFENPEVFDPERFNDENKKKIVPG
SFIPFGSGPRNCI

>CYP9Y3v2
McKenna_67139
sequence revised 4/5/12
1-amino acid difference from McKenna_34157
YYGSYQFLKPSLFVRDLDLIKQITIKDFDHFTDHVDILNSNNDPIFSKNLFSLKGKE
WRELRSTLSPAFTSSKMKAMFVLISEASKKFVEHFEAKKEEIIEVEMKETYSKFTN
DVIATCAFGINCDTLENPENEFFTMGKAVTRSTFLRMMRALVRMLFPTIFEVLKIP
TFPKDVTNFFKKIITDTLTTREKEGTIRPDLIHLLMEARKGKLQQETSQVTEDTGF

>CYP9Y4
McKenna_38109
sequence revised 4/5/12
78% identical to McKenna_34157
56% identical to CYP9Y1 *T. castaneum*
YFGSYQLFNPCLVVKDVDLIKQITIKDFDHFTDHVDLMNTDHDSLFSKNLFFLKG
KQWR
EMRNTLSPAFTSSKMKAMFVLISEASKKFVQHFEAKNDEIIEVEMKDAYSKFTTD
VIATCAFGINCDTLENPENEFFTMGKAVIRSNFWMIIKAFLRTIFPRVFGFFRIPILP

VEVTNFFKMIISDTIKSREKEGTIRPDLIHLLMEARKGKLQQETSSQSTEDTGFATA
QEIRDSKPNSDLEITDDLITAQAILFFLAGLDTSSSLLSLLSYELAKNPDIQQKLISE
VDDNLSSSEDFISYEKLAKMKYLDQVVSEALRKWTPGFVLDRLCTKDYSIPPVK
EDEVPLTISKGCYVHIPVIGLHYNPQFFENPDVFDPERFNDENKKKIAPGTYIPFGS
GPRNCI

>CYP9BA1v1
McKenna_68213
44% identical to CYP9Z2 *T. castaneum*
MLLVLIVIGTIFFYFFIIKPGNYWKERHVPTGKIIPIFGEHYLNILGKDCSTEFAQRI
YNNVPDARYLGIYLFQTPVLVLRSPDLIKDICVKNFNNVLDRRNIPPDCDELLVSK
NLMGQQWKEFRHLLSPSFSSSKMKAIYVLLCDNASKIVQYLMDKHENLIEEEVK
DTFTRYTNDAIANTIYGLEINSFTERKNEFFMMGKKVTDFSAPWKIFVVLLYTLA
PKIAKVRIKFKVALYGKKTADFYTDLCKRIIELREEKNIERPDVLGLLIEARNKFE
KSKSCNEESEGVTYYSTSETPNEKVLEKYLSYEEIAAHIFLFKLGGYDTSSSAMCY
MAYELAIHPEIQKKLIAEIDQIKSEIATPSYEAIMNMTYLDMVVSETLRMWPSLAL
TDRLVTTSFTIAAEQPGEKPLQMKEKDLIIIPIFGIHRDPKYYENPDRFDPERFSPEN
RKNITPYTYMPFGVGPRNCIGMRLALLEIKVLFFYLLSHFEIVATERTKVPLRLKR
TMVTSTADDDFPLAFKKRGSKI*

>CYP9BA fragment
McKenna_60215
83% identical to CYP9BA1v1 McKenna_68213 N-terminus
MIAIVIAIGTILFYFFIIKPRNYWKERHIPTGKITPIFGEHYRNILGKDCSTEFVQRIY
NKVPEGR

>CYP9BA1v2 or end of CYP9BA3
McKenna_91842
96% identical to McKenna_68213
41% identical to CYP9Z6 *T. castaneum*
exons 4–5 runs off the contig end
KFKVALYGKKTADFYTDLCKRIIELREEKNIERPDVLGLLIEARKKFEKSKTCNEE
SEGVTYYSTSETPNEKVLEKYLSYEEIAAHLFLFKLGGYDTSSSAMCYMAYELA
KHPEIQKKLIAEIDQIKSEITTPSYEAIMNMTYLDMVVSETLRMWPSLALTDRLVT
APFTIAAEQPGEKPLQMK

>CYP9BA2v1
McKenna_30031a
Length = 4101 (2 sequences)
revised DRN 4/7/12
2 genes
end of 1 gene
98% identical to McKenna_131727
61% identical to CYP9Z4 *T. castaneum*

ETLRKWPSIAHTDRVVSTPFTIETELPEEKALHMREKSKIMIPIFGIHRDPKYYEEP
DRFDPQRFSPENRKNINPYTYMPFGVGPRNCIGMRLALLEVKVLFFYLLSHFEIV
KTEKTEIPLKLKRTVVTLTAENGFPLAFRRRHFVKK*

>CYP9BA3
McKenna_30031b
Length = 4101 (2 sequences)
revised DRN 4/7/12
2 genes
beginning of a second gene
97% identical to McKenna_68213
44% identical to CYP9Z4 *T. castaneum*
MLLVLIVIGTIFFYFFIIRPGKYWKERHVPTGKIIPIFGEHYLNILGKDCSTEFAQRIY
NNVPDARYLGIYLFQTPVLVLRSPDLIKDICVKNFNNVLDRRNIPPDCDELLVSKN
LMGQQWKEFRHLLSPSFSSSKMKAIYVLLCDNASKIVQYLMDKDENLIEEEVKD
TFTRYTNDAIANTIYGLEINSFTERKNEFFVMGKKVTDFSAPWKMFVVLLYTLAP
KIAKVRI

>CYP9BA2v2 or end of CYP9BA3
McKenna_131727
revised DRN 4/9/12
98% identical to McKenna_30031a probable allele
61% identical to CYP9Z4 *T. castaneum*
ETLRKWPSIAHTDRVVSTPFTIEAELPEEKALHMREKSKIMIPIFGIHRDPKYYEEP
DRFDPERFSPENRKNINPYTYMPFGVGPRNCIGMRLALLEVKVLFFYLLSHFEIVK
TEKTEIPLKLKRTVVTLTAENGFPLAFRRRHFVKK*

>CYP9BA4 Query
McKenna_32781
revised DRN 4/7/12
94% identical to McKenna_44095
62% identical to CYP9BA1 McKenna_68213
41% identical to CYP9Z2 *T. castaneum*
YAGIYLFQTPVLVLKSPELIKEICVKNFNNVLNRQALTPDCSEPLMSKNLLALKG
QHWKDIRHLLSPSFTISKIKAIHVLLCDNASKTMQYFHDKDEELIEVEVKDTFTRF
TNDVLANTIFGLEINSFKDNQNEFYMMGKDASDFSKPWKIFVILLHHISSKLARTL
KVELYGQETTDFYTNIVKQTIKIREEKNIQRNDVLGNMMEERKKLRTDTNCNENI
NEKSDEKSVETSLSDEDIAAHLFLYMLGGYDTTSTAICFMAYELAINPDIQKKLIE
EIDGVGTENGMPSYEDISNMVYMDMVLS

>CYP9BA5
McKenna_44095
revised DRN 4/7/12
94% identical to McKenna_32781
42% identical to CYP9Z1 *T. castaneum*

TLKIELYGQQTTDFYTNIVKQTIKIREEKNIQRNDVLGNMMEERKKLRTDTNCNE
NINEKSDEKCVETSLSDEDIAAHLFLYMLGGYDTTSTAICFMAYELAINPHIQKKL
IEEIDGISTENGTPSYEDISNMMYMDMVLS

>CYP9BB1
McKenna_100427
revised DRN 4/2/12
48% identical to CYP9Z4 *T. castaneum*
58% identical to C-term part of McKenna_20898
LMHVCSKNFVKYLKEKPEELIELELKDVFTRYTNDIIASTAFGIQCDSLKERGNEF
YMNGRRVTNLSGVIRNLKFLIVFIFPKLSKLVKATFFDEEVASFFRGVVRDTLKY
RTNNKIDRPDLLQLLMQAKKNIQKEADGKNIERTDSSVVEDFTVSKRGKLDLSL
DDITSQALIFFFAGFETVAAVMCFVAYELAVNPNIQTRLIEELDEFRASNEKFSYD
SLTKLPYLDMILSETLRKWPVMISTDRKCNKPYMIKAELDEETSLQLNGGEIISIPI
YALHRDEKYWENPDNFDPERFSTENKHKIDPYTYIPFGTGPRNCIGSRFAISEVKV
IFFELLNQFEIVPTKKSCIPLVLDRKSFSLNSSTGFWFGLKKRRC*

>CYP9BC1
McKenna_114665
revised DRN 4/9/12
41% identical to CYP9Z5 *T. castaneum*, most similar to CYP9Zs
IVSESLEKENINKHRSSLMDSIHKLKQEEQSENLDGDIKRAKIDLNEDEIVSQLCM
FLFSSIDGLLPTLIFITYELAINPNVQQKLRNEMDQIRVEGELPEFHTIMGLQYLNM
VISETMRKWPSVTMTDRLCTKPYTIEPVLPEEKPLHLDIGDCVAVPLYALHHDPQ
YFPEPDTFDPERFADGNRHKIKPFTYLPWGIGPRNCNAQKLSLLILKVFFFQLLEH
LEVIPIDETVIPIELQKGVIKVEPKCDIRLGLRKREDL*

>CYP9BD1P
McKenna_94145
DRN 4/1/12 (−) strand
revised DRN 4/1/12
pseudogene, 47% identical to CYP9Z4 *T. castaneum*
46% identical to McKenna_20898
ETLRKWPPLLNMDRMCVEPYTIKPEKPGEETVHLEESTLIWIPIWAIHHDPEHWP
NPEVFDPEKFRDNTRTPNGFIPFGLGNRGCLGFRFATMEIKLLFIHLLTSFKIERSS
KSQVPF VPVVHSFSLTSEDGFCFKLVKRQKPN

>CYP9BE1 Query
McKenna_50287
revised DRN 4/4/12
54% identical to CYP9Z6 *T. castaneum*
92% identical to McKenna_9307
77% identical to McKenna_120940
77% identical to McKenna_118374
ILKIPIISKKVQDFFINLVKDTLRMREENNIKRPDVLGLLLDARKGQLDIKEDQEEE
DGG FAVVKEQLEMKTLNCELTDLDIAAQAFIFFLAGFEGVATLICQTVYELAIN

PDVQKKLIDEIDENWPEDDKPSYNKVMNMTYLDMVVSEALRKWPNGIQTDRVV
TKKYTIEPELPGEKPLTLEEGSILLIPILGMHHDPKYFPNPEKFDPERFSAENKNNI
DPYTYMPFGIGPRNCIGSRFALMEVKALLFYLFRHFKVVPIEKTEIPIKFNKNAFG
FIPENGYQLGLKKRTMEK*

>CYP9BE2
McKenna_20898
revised DRN 4/7/12
runs off the end of the contig
57% identical to CYP9Z4 *T. castaneum*
59% identical to McKenna_50287
56% identical to CYP9BA1v1 McKenna_68213
54% identical to CYP9BB1 McKenna_100427

EAGFAVVEEHLEVNEIKPQDLTDTDIASQVFIFFFGGFETVSTAMCFMAHELASSP
DVQAKLIEEIDESVRKNGEPTYESIANMIYLDMVVCETLRKWPVNIATDRMVTKP
YTINPEEPGEKPVHLEVGDVAIIPIIALHYDPKYYENPEKFDPERFSPENRKNIDPY
AYIPFGVGPRNCIGSRFALLELKAVFFHILRHFEIVPVEKTNIPIKLSTKSINLSSEN
DYPLGLKKREISKN*

>CYP9BE3 Query
McKenna_109359
revised DRN 4/9/12
40% identical to CYP9Z2 *T. castaneum*
64% identical to CYP9BE1 McKenna_50287
possible pseudogene since heme region exon does not end in phase 1
ILKIPIIAKYPQDIFINLVKDTIKTREKQNIKLPDFLGLLLDARKEKVRAKEQHGRN
DSGAVIKEESEQKNSNCQLTDVDITAQVFVYFLAGFEAVATQICLTILELAIHSDV
QNRLIDEIYRNWPEDREPNYNEVINMPYLDMVVS

>CYP9BE4v1
McKenna_9307
revised DRN 4/4/12
59% identical to CYP9Z4 *T. castaneum*
96% identical to McKenna_121348
93% identical to McKenna_50287
81% identical to McKenna_120940
81% identical to McKenna_118374
63% identical to CYP9Z6 *T. castaneum*
EALRKWPNAILTDRVVTKEYTIEPELPGEKPLTLKEGSILVIPILGMHYDPKYFPNP
EKFDPERFSPENKHNIDPYTYMPFGIGPRNCIGSRFALMEVKALLFYLFRHFEV
VPIEKTEIPIKFNKNAFGFIPENGYHLGLKKRAMEK*

>CYP9BE4v2
McKenna_121348
revised DRN 4/4/12

65% identical to CYP9Z5 *T. castaneum* runs off the end of the contig
96% identical to McKenna_9307
EALRKWPNGILTDRVVTKEYTIEPELPGEKPLTLKEGSILLIPILGMHFDPKYFPNP
EKFDPERFSPENKHNIDPYTY

>CYP9BE5v1
McKenna_120940
revised DRN 4/4/12
59% identical to CYP9Z6 *T. castaneum*
ESLRKWPNGILTDRIVTKEYTIQPELPGEKPVTLKEGTILMISILGLHNDPKYFPNP
EKFDPERFSPENKSNIHPYTYIPFGIGPRNCIGSRFALLEIKALLFYFFRHFEVVPIEK
TDIPPKFNKNTAGFIPINGFHVGFKRRRLN*

>CYP9BE5v2
McKenna_118374
revised DRN 4/4/12
57% identical to CYP9Z5 *T. castaneum*
63% identical to CYP9Z18 *Dendroctonus valens*
96% identical to McKenna_120940
ESLRKWPNGILTDRIVTKEYTIHPELPGEKPVTLKEGTVLMISILGLHNDPKYFPNP
EKFDPERFSPENKSNIHPYTYIPFGIGPRNCIGSRFALLEIKALLFYLFRHFEVVPIET
TDIPPKFNKNAAGFIPINGFHVGFKRRRLN*

>CYP9BE6v1
McKenna_24087
revised DRN 4/9/12
97% identical to McKenna_114066
62% identical to McKenna_9307
52% identical to CYP9Z4 *T. castaneum*
EGLRMWPFLTIIDRVSTKKYTIQPEKPDELPLTLEPGSLILIPIIGLHYDPKNYEHPE
KFDPERFSAENKKTINPYAYLPFGVGPRLCLGNRFALMEMKIILFKMLCNFEIVPI
QKTEVCLKLNKQALGLTPINGYHLGLKKRNKHTY*

>CYP9BE6v2
McKenna_114066
revised DRN 4/9/12
97% identical to McKenna_24087
EGLRMWPFLTIIDRVSTKKYTIQPEEPDEQPLTLEPGSLILIPIIGLHYDPKNYEHPE
KFDPERFSAENKKTINPYAYLPFGVGPRLCLGNRFALMEMKIILFKMLCNFEIVPI
QRTEVCLKLNKQALGLTPINGYHLGLKKRNKHTY*

>CYP9BE7P Query
McKenna_108449
revised DRN 4/9/12
42% identical to CYP9Z4 *T. castaneum*
41% identical to McKenna_50287

53% identical to CYP9BE2 McKenna_20898
43% identical to CYP9BE1
possible pseudogene because the heme region exon does not end in phase 1
TLGLSLFPPRAREFFLNIVKDTMKKRKEEHLKRDDMIGLLIELRNAQLKAREEAG
PKAKPQKLMTVQEMASYIFVMYFGVIDSVTTVLAFLAWELAMAPEIQERLRREC
DSLASSTNEATHQDIQGLKYLDMVLCEVLRKWPGAIATDRLVTKSYTIEPELPHE
KPVHLKEGDNIMIPIYALHRDAKYFPEPEKFDPERFSPQRKHEMNSNAYIPFGIGP
R*CL

>CYP9BF1
McKenna_83602
revised DRN 4/9/12
48% identical to CYP9Z4 *T. castaneum*
CFMAYELAVNPDMQKKLLEEIDTLRRKTEQINYEDLIELSYLHMIISETMRKYPPF
CRNDRRCTKSLTIKNEEYPEKSFTIEAGHDIWFPIYALHHDSQYWPEPEKFIPERFS
PENKKNIKPGTYLPFGIGPRGCMGYRFVLQTVKILFYELLSEFEIIPVAKTAIPCQL
SKNTSHVIPGKGFWFGFKRREQSI*

>CYP9BG1v1
McKenna_87030
revised DRN 4/9/12
41% identical to CYP9Z4 *T. castaneum*
ETLRKWPPCYFTERCCNETWIIKKRWANERNAWLLDGDAVWIPIWAIHRDPKH
WWQPEVFDPERFSPNSHKTIEPGTFIPFGVGPRTCLGDKYSLSQIKIVVSEILALFE
VVPTNKTPKTIELNHRTLNLSPKDGLWLGLKKRFM*

>CYP9BG1v2
McKenna_59660
revised DRN 4/9/12
42% identical to CYP9Z1 *T. castaneum*
44% identical to McKenna_100427
98% identical to McKenna_87030
ETLRKWPPCYFTERCCNETWIIKKRWANERNAWLLDGDAVWIPIWAIHRDPKH
WWQPEVFDPERFSPNSRKTIEPGTFIPFGVGPRTCLGDKYSLSQIKIVVSEILALFE
VVPTNRTPKTIELNHRTLNLSPKDGLWLGLKKRFM*

There are at least 6 C-helix regions in this set and 6 exon 1 N-termini
There are no C-termini here, and thus, they must be identified as some other sequence group.

>CYP9 fragment 1
McKenna_43434
revised DRN 4/8/12
43% identical to CYP9Z4 *T. castaneum* N-terminus exons 1–3
43% identical to CYP9BA3 McKenna_30031b

MLLVYTVLCLLLYILVIKPYLYWSNKNVPFRFSIPLFGEGIYMIMGKENMSDTIKR
MYNKIDVVRYLGVFQFMQPVLIVKSTELIKQICVKDFDHFLNHKVLLPDGVEELL
SKNLLQLKDQTWKNMRATLSPSFTTSKMKSMFTLISQNADAFAKYFKEKNDGIV
EVEFKDAFTRYTNDVIASAAFGLQVDSLTDRENDFYVLGREMSDFSTFWKKFVF
FFFQISPTIAR

>CYP9 fragment
McKenna_44011
95% identical to CYP9 fragment 1 McKenna_43434
MLLVYTVLCLLLYILVIKPYTYWSNKNVPFRFSIPLFGEGIYMIMGKENMSDSIKR
MYNKISDVRYLGVFQFMQPILIVKSTELIKQICVKDFDHFLNHKVLLPDGVEELLS
KNLLQLK

>CYP9 fragment
McKenna_58513
43% identical to CYP9Z6 sequence34, XP_972794.1 N-terminus
37% identical to CYP9 fragment 1 McKenna_43434
MSNVPVAVSIIFSTLILYYYNYTYWRRNRVKQRFPQVPIFGDNYRHFFGKSPYHK
VLEALYWKFPTARYTGFYQYMQPGLMIRDMKLVRKVCVEDFDTFKDRKSFLPR
SADPLWNKNLFALQGDKWRQMRNTLSYSFTANNMGVMRSSLPLWS runs off the
contig

>CYP9 fragment 2 Query
McKenna_128566
revised DRN 4/5/12
63% identical to McKenna_27215
45% identical to CYP9V1 *Leptinotarsa decemlineata* N-term
MFVVFLLLSVLMMLVYFMVIKPFQYWKKKNVKTGVIIPLFGDNFRVVFGLEGIID
QARRIYNSFPNERYQFSRPVLVIRNPDLIKKFCVKDFEYFFNRRFATE

>CYP9 fragment 3
McKenna_101612
45% identical to CYP9Z2 *T. castaneum*, exons 1–2
43% identical to McKenna_27215
MFWLFSILILLILWYYYERSTFNFWSEKGVKQKIPRSILSDILGSVFQTTAVCDMIK
GIYDSFDDVRYVGFYQFLQPILMIKDPKLIKQICVKDFEYFLDRRPFVAEDVDPL
WGKNLVALS

>CYP9 fragment 4
McKenna_73464
revised DRN 4/9/12
38% identical to CYP9Z5 *T. castaneum* N-terminus
MFWLLFCAVLLVVFVFLKKKYTYWKEKGVVQDLAVPVFGLNWKLLTHQASNI
EIFQYLYDKHKKER

>CYP9 fragment 5
McKenna_54994
revised DRN 4/9/12
47% identical to CYP9V1 *Leptinotarsa decemlineata* N-terminus
85% identical to McKenna_27215
MIAFILLSILITLVYFLAIKPFNYWKERNVKTGRVIPVFGDNFGVIFGLESFIDQTRR
MYNTLPNER

>CYP9 fragment 6v1
McKenna_27215
revised DRN 4/5/12
98% identical to McKenna_39779
63% identical to McKenna_132274
44% identical to CYP9V1 *Leptinotarsa decemlineata*, exons 1–2
MIAFILLLSILIILVYFLAIKPFNYWKERGVKTERVIPVFGDNFGVMFGLESFIDQT
KRMYNTLPNERYYGLYQFSLPTLVIRTPDLIKKLCVKDFEYFLNRRNLTPEGCDV
LFSKNLINLK

>CYP9 fragment 6v2
McKenna_39779
revised DRN 4/5/12
exon 2
100% identical to McKenna_27215
55% identical to CYP9Z1, exon 2
YYGLYQFSLPTLVIRTPDLIKKLCVKDFEYFLNRRNLTPEGCDVLFSKNLINLK

>CYP9 fragment 7
McKenna_71298
revised DRN 4/9/12
41% identical to CYP9V1 *Leptinotarsa decemlineata*
50% identical to McKenna_27215, exons 1–2
runs off the contig end
ILFFGENFWVLLSLESLTEQVHRIYNLFSNERYFGVYFFTTPLLVVKSPDLIKELCV
KNFDHFVNRMTLVPDHADDLFTRNSVSIK

>CYP9 fragment 8
McKenna_22233
revised DRN 4/9/12
57% identical to CYP9Z6 *T. castaneum*, exon 2
77% identical to McKenna_27215
77% identical to McKenna_39779
YYGLYQFFNPALVIRTPDIIKKCCVKDFDHFLNRRNFAAEGADVLFSKNLLNLK

>CYP9 fragment
McKenna_133538
90% identical to CYP9 fragment 8 McKenna_22233

YYGLYQFFMPALVIRTPDIIKKLCVKDFDYFLNRRNFAAEGSDVLFSKNLINLK

>CYP9 fragment
McKenna_62451
88% identical to CYP9 fragment 8 McKenna_22233
94% identical to contig_133538
exon 2
YYGLYQFFLPALVIRTPDIIKKLCVKDFEYFLNRRNFAAEGTDVLFSKNLINLK

>CYP9 fragment 9
McKenna_6103
revised DRN 4/7/12
42% identical to CYP9Z1 *T. castaneum* amino acids 68–288
41% identical to CYP9Y3v1
47% identical to McKenna_23880
runs off the end of the contig, exons 2–4
YIGVYDFLTPIVIIKDPDLLQEVLIKKFEHFTDHRPTIPSDVDPLWNKNVANLKGN
KWKEMRSSLSPSFTSNKMRVMFVLMEQCAKNFVQYFLLKNEEIVETEIRESFSRY
ISDVIASTCFGYQCNSMTNPDNEFYVKGLQASYDMSFWRHVKSSLLRVFPWMSK
VIKVTRFGDDCTRFFRDIIKENVEFRDKNKVKRPDMITLLMEARNEPTSNEDKDN
NDSSFLIYELNASASKKKIDLTVDDIAANAMVFFFAGFSTTTSLLCFMAYELAQ

>CYP9 fragment 10
McKenna_87274
revised DRN 4/9/12
58% identical to McKenna_23880
53% identical to CYP9Y3v2
44% identical to CYP9AC1 *T. castaneum*, exon 3
GQIWRDTRSTLSPTFTSSKMKAMFMLICYNAKLFTDYFLEKEEDNIEVELKGILH
RYTCGVMTSTLLGIQVNFMKDINHFFFELGADFGSLRMKISSFLYQISPRIAS

>CYP9 fragment 11v1
McKenna_72064
revised DRN 4/9/12
95% identical to McKenna_23880
54% identical to CYP9Z4 *T. castaneum*, exon 3
GQSWKDMRSTLSPTFTSSKMKSMFVLICQNANLFTEHFLEKKEDLIEVEFKDSIC
RFTSDVIASCAFGLQVDSLKDRNNTFFKMGRESLDFTSIRLKIAFCLYQVSPKLAS

>CYP9 fragment 11v2
McKenna_23880
revised DRN 4/9/12
95% identical to McKenna_72064
75% identical to McKenna_26312
54% identical to CYP9Z4 *T. castaneum* C-helix region, exon 3

GQSWKDMRSTLSPTFTSSKMKSMFVLICHNAKLFTEHFLEKKEDLIEVEFKDSIC
RFTSDVIASSAFGLQVDSLKDRNNTFFKMGRESLDFTSIRLKIAFCLFQLSPKLAS

>CYP9 fragment
McKenna_111718
88% identical to CYP9 fragment 11v1 = McKenna_72064
C-helix (2-amino acid difference)
GQSWKNMRTTLSPTFTS runs off the contig

>CYP9 fragment 12
McKenna_26312
revised DRN 4/9/12
49% identical to CYP9Z1 *T. castaneum*
51% identical to CYP9AC1 *T. castaneum*
74% identical to McKenna_23880
55% identical to McKenna_43434, exon 3
GQSWKNMRSTLSPTFTGSKMKAMFELVCQNAKLFTDYFLEKNEDITEIELKDSV
QR FTSDVIANAAFGIQVDSLKDRKNTFFMMGRDGLNFNNIRIKIAFILYQLSPKL
AS

>CYP9 fragment 13
McKenna_134640
42% identical to CYP9AS3 *Linepithema humile* I-helix up to EXXR
52% identical to CYP9AZ1 *Dendroctonus*, exon 4
IFKAHIFDKEVSRYFRSLIWGIIDRREENDIERADMLQLLVRRIPETHFDNDDITSQ
ALVFFLGGFETVSSALCFMAYELALNPDIQERLYKEINDFRETHKREITYEDTKG
MRYLEMVFN

>CYP9 fragment 14v1
McKenna_125545
revised DRN 4/9/12
37% identical to CYP9Z2 *T. castaneum*
36% identical to CYP9BA4 McKenna_32781
95% identical to McKenna_92383, exons 3–4
GEEWKNVRSTMSPVYTSSKMKVFFNQISHNADKLVDYIMKNDQIILEVELKDLL
SRFSNDVIARNIYGVEIDSLKDRNNIFYEMAKNGTHICGYRKKYSILFYQLAPRIS
AFLRLPLAEKEVQSFFVKLIEDTIAIRKTQNITQPDLVGKLIQSTEANEVKTITTKD
GSVVKEKNWRTEEIKNDLSVTDMAALTFFAFNAGFEAISNVLCFVAHQLALNPDI
QKRLIDEIDENFADNEMPSYGKLLNMSYLDMVIS

>CYP9 fragment 14v2
McKenna_92383
revised DRN 4/9/12
95% identical to McKenna_125545 possible allele
42% identical to McKenna_50287
35% identical to CYP9A20 *Bombyx mori*, exon 4

FFRLPLAEKEVQSFFVKLIKDTISIRKAQNITQPDLVGKLIQSTEANEVKTITTKDG
SVVKEKNLRTEEIKNDLSVTDMAALTFFAFNAGFEAISNVLCFVAHQLALNPDIQ
KRLIDEINENFTDNEMPSYGKLLNMSYLDMVIS

>CYP345D4 Query
McKenna_132274
revised DRN 4/9/12
49% identical to McKenna_123923
48% identical to CYP345D2 *T. castaneum* N-term (identity will increase with full length)
47% identical to McKenna_73770
MIWLILLVTIFMLFYIYTYRSFKYWEIRNVYYEKPVPIFGNFYDVAVRKKHMGD
VLKEIHLKLDDNVPYFGVYIFHAPNLVVRTKEMIKEVLIKNFTSFPNRMDYTNEV
VDPLSSYDLFSMKEDLWKFTRTKLSPAFSSGKIKMMGPLMKEVTDQLENLLESS
NGQQVDVRDLAKRYLVDIISTCAFGINAESLKDQNSKIKVMANQMLDQRGFVRS
FAVFAWFFCPLLVDIFRLPFVEK

>CYP345D fragment
McKenna_84502
94% identical to CYP345D4 McKenna_132274
runs off the contig
MIWLILVVTILILFYIYTYRSFKYWETRNVYYEKPVPIFGNFYDVAIRKRHMGDVL
KEIHLKLDDNVPYFGVYIFHAPNLVVRTKEMIKEVLIKKFSSFPNRMDYTNEVVD
PLSSYDLFSMKEDLWKFTRTKLSPAFSS

>CYP345D5
McKenna_73770
revised DRN 4/8/12
47% identical to CYP345D4 McKenna_132274
40% identical to CYP345D3 *T. castaneum*
42% identical to CYP345K1 McKenna_44351
41% identical to McKenna_123923
MIWIVLISLCLLLFYYRTVRTFSYWKLKKVPFRRPLPIFGNIMKVALGQQQVGTAI
REIYDSFNENVPYFGMYILHKPFLVIRSKELAKKILIKDFNTFQNRPLYHNKEIDP
MASNALFIMRGEEWKGLRNKLSPIYTSGKMKKMMPLITKIADQMELYVDTLTD
GQEIDVRDLANRYSLDVICSCAFGINSNSLIQKDNEIKNTANKMLDLRSIKRSFAI
GSYFFCPLMVELFKLTFLDKESSDYFVDIFKTSFEQRKNSKEKRNDLIDLMYKIKL
SESEEDTFKFGKKYL

>CYP345D6
McKenna_123923
revised DRN 4/9/12
43% identical to CYP345D3 *T. castaneum*
49% identical to CYP345D4
MMLSLILFIVAASLIYLYVKRSLSYWEKKNVYAEQPYPLFGNLLNVAFRGESMM
NIIKRINNGLSDKHKYFGFYLFHKPVLIVTDKELIKEVLIKDFNTFANRANYTNEK
VDPIASNNLFSLRDQVWRTVRNKLSPVFTSGKMKLMMPLMTEISDNLEDVLNNT

HNTNIDVKDITKRYAIDITTSCAFGFDSESLKDASSEVKLMSDKLLNPTSFISRFG
MFCWVLCPFLVDIFRVPFVDREASRYFINLFKQSEDERIKKNIQRNDFVDLMIEM
TKQE

>CYP345K1v1
McKenna_44351
Sequence revised 4/5/12
44% identical to CYP345D3 *T. castaneum*
MIWLVFVFLFLVLLYIYSIPNLWKNKNIDFVPPFPVVGNFLNTILRTEGIGEIIKKN
YDGYDSSTPYFGMYFFRTPFLVIKSRELVKTVLVKDFSHFPNRPTYADEYIDPWS
TKTLLTLRNEEWRSLRNKLSTVFTSGKMKMMIHLMKEVSDQMEEYLEERKSED
VDVRELSRRFFINIITKCVFGIKPNNLKDDNSYIRNLASRLIDIENYKWIFSLLLYFT
FPIIVKLCGLQFVDKEAADYLVKVFEDSYKERKKTKMVTNDLVDLLHNLKEQEK
EDDTFKFDDMKLAAQALVFFQAGSEPASSTLSFCLYELALNKEVQDTLREEIHQN
IDADGFLSYEVLMGLTYLDLVLKEALRKYPLVQVLVRLVEKEYTFEKTGLKVEK
GVSVVIPMLALHYDPKYYPDPEKFDPERFRGEKFRKLDYVYLPFGDGPRKCIGY
RFAVMSLKIALAMFIKSFEVLPCTKTEIPLEFNKTSFFISPKSSSIVLRVNAVDKYK
KD*

>CYP345K1v2
McKenna_101506
revised DRN 4/2/12
48% identical to CYP345D1 *T. castaneum*
1-amino acid difference from McKenna_44351
DDMKLAAQALVFFQAGSEPASSTLSFCLYELALNKEVQDTLREEIHQNIDADGFL
SYEVLMGLTYLDLVLKEALRKYPLVQVLVRLVEKEYTFEETGLKVEKGVSVVIP
MLALHYDPKYYPDPEKFDPERFRGEKFRKLDYVYLPFGDGPRKCIGYRFAVMSL
KIALAMFIKSFEVLPCTKTEIPLEFNKTSFFISPKSSSIVLRVNAVDKYKKD*

>CYP345K2
McKenna_32234
revised DRN 4/9/12
48% identical to CYP345D2 *T. castaneum*
59% identical to 345K1v2 McKenna_101506
58% identical to CYP345K1v1
NMDSEGNITYEALFEMTYLDLVIKETLRKYPLTHLLMRYPETDYTFEETGLHLEK
GTPVLLSMTSLHFDQRYFPDPDKFDPERFRGEKAKDLQYVYMPFGEGPRKCIGD
RFALLSMKIAVFMFLKKFEITPCNKTEVPVKQSKTAFFIIPDSGSIFLNVRKIE*

>CYP345K3v1
McKenna_71905
revised DRN 4/9/12
54% identical to CYP345K1v1 McKenna_44351
53% identical to CYP345K2
47% identical to CYP345D3 *T. castaneum*
98% identical to McKenna_11046

59% identical to McKenna_53591
DEIKMSAQAMAFFNAGTDTTSILLTLCLYEMAMNLHIQEKLRNEIRDSIDPTGSIS
YETLLSMEYLDLVTKETLRKYPFLQLLQRYCVKDHTFETGLTIKKGQRVLIPTLA
LHHDPKYYPNPEEFNPERFRGDRIKDLQYVFLPFGEGPRKCIGEKFGLISFKLGLS
MILKNFEVLPGPQTEIPLKFEKAAFFTTPESAEIMLTIRRVD*

>CYP345K3v2
McKenna_110460
revised DRN 4/9/12
55% identical to McKenna_101506
49% identical to CYP345D2 *T. castaneum*
DEIKMSAQAMAFFNAGTDTTSILLTLCLYEMANLHIQEKLRNEIRDSIDPTGSISY
ETLLSMEYLDLVTKETLRKYPFLQLLQRYCVKDHTFETGLTVKKGQRVLIPTLAL
HHDPKYYPNPEE

>CYP345K4v1
McKenna_53591
revised DRN 4/9/12
44% identical to CYP345D3 *T. castaneum*
57% identical to CYP345K1v1 McKenna_44351
53% identical to CYP345K2
98% identical to McKenna_93915
ETLRKYPFAQFISRYAEKDHIFESTGLKIEEGTQIIIPTLAIHENPKHYPDPSKFDPE
RFRGDKMKDLQYVYLPFGEGPRKCIGERFAMMAMKLVIVTFLKKFELVPGEKT
DIPLKFVKNGFFTSTKDGCIYLKIKERSQ*

>CYP345K4v2
McKenna_93915
revised DRN 4/3/12
51% identical to CYP345D1 *T. castaneum* (CYP3 clan)
98% identical to McKenna_53591
ETLRKYPFAQFISRYAEKDHIFESTGLKIEEGTQIIIPTLAIHENPKHYPDPSKFDP
ERFRGDKMKDLQYAYLPFGEGPRKCI

>CYP345 fragment 1
McKenna_70825
revised DRN 4/9/12
43% identical to CYP345A2 *T. castaneum* amino acids 172-291
SGLDVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIYDTRPITSFRFLCYF
FFHSFARIFKMKLFDADVVTFLRRVFWECIELREKNNVRGNDLIDIIVDLRKDNEL
SERIKF

>CYP345 fragment 2
McKenna_20430
revised DRN 4/9/12
47% identical to CYP345A2 *T. castaneum*

GRRFGLIAAKLAILQILKEFELHSTDETPVNLEFSPASTIPQPIQQLKMSFINTDPLF
*

>CYP345 new sequence
McKenna_63104
46% identical to CYP345A2 *T. castaneum* sequence37, XP_970633.1
MHWFIEVLLLISFLLYLLHLYISRNYDYWQKRNVPFIKPRPFVGNMGEILLQKYN
MSSFFEKLYNDMDAPFFGIFVFSKPALIVKDVKLLKNIFVKDFDHFMDQHYDPVT
AHMLFIEKGEEWKLMRSKISPFFSPSKLKAMFGAIDNLGVSLRRHIDASPNRSGL
DVKELSSKFSVDVIAKCVFGIDAKSLEIEDGEFLRIAHKIFDTRPITSFRFLCYFFFH
SFAKIFRMKLFDADVVTFLRRVFWECIELREKNNVKGNDLIDIIVDLRKDNELSER
IKFDGDKVIAQALLFFVAGFETTGSTIAFTLHALCLNLDIQRKLRENIRDIIKKHGG
KLTMESIENMDYLDNVIKETLRKYSPVPILDRVCTKDYKIPETDIVIEKGIITLVPP
YGFQKDPKYFDNPEEYIPERFESIKEDMFYMPFGHGPRNCIGRRFGLIAVKLAILQ
ILKEFELHSTDETPVNLEFSTASLIPQPIQQLKMSFINTDPLF*

>CYP435A1v1 Query
McKenna_109920
revised DRN 4/4/12
31% identical to CYP9D8 *T. castaneum*
VLTFFLTASLMIDQYYRFWNRRGVKQKYTFLFTDDMTMRLLKKQSFGDMLTDI
YYRFSGTRYIGFYQFHKPILLVKDPHLIRNLCITNYRIFRDHTRILPFRCDPFWNKT
IFALKGKKWKKARQQLMPLFNGRNMRNMYETTRSYAYDFTNSLVPIQTKVLEA
DFKELFSKFSTELTARVVYDIHNKSFKNVSTSFYTKLRDSNENFGARRFIRIFLSQ
MCPCIGKLLNVSIFSTDLSIFFSKVTHNIMRHRERFGIEKMDLISLLMNTRNREGR
RDFMRFIGETDPEEVEEDMTEADEDLTEETIVAHAMTFFYAGFDVISTVLSFLFYE
LALNPDIQRRLLKDLDAWRSADNIDPYKSLFGIQYLSMVIS

>CYP435A1v2 Query
McKenna_51177
revised DRN 4/4/12
2-amino acid difference from contig_109920
VLTFFLTASLMIDQYYRFWNRRGVKQKYTFLFTDDMTMRLLKKQSFGDMLTDI
YYRFSGTRYIGFYQFHKPILLVKDPHLIRNLCITNYRIFRDHTRILPFRCDPFWNKT
IFALKGKKWKKARQQLMPLFNGRNMRNMYETTRSYAYDFTNSLVPIQTKVLEA
DFKELFSKFSTELTARVVYDIHNKSFKNVSTSFYTKLRDSNENFGARRFIRIFLSQ
MCPCIGKLLNVSIFSTDLSIFFSKVTHNIMRHRERFGIEKMDLISLLMNTRNREGR
RDFMRFIGETDPEEVEEDMTEADEDLTEETIVAHAMTFFYAGFDVISTVLSFLFYE
LALNPDIQRRLLKDLDAWRTAGNIDPYKSLFGIQYLSMVIS

>CYP436A1
McKenna_46973
revised DRN 4/3/12
35% identical to CYP9X1 *T. castaneum* N-term
MNVFGVSSGFSTSFLLWLVLFVFIYLCCTVVLKPYKYWSSKGVRTAKILPVIGDA
IFAICFSKKPYPKLIQDVYDSVGNSRYIGYSELRTPILMIKSPELIGKLFVRDAENFQ

DRRVVFLENFDTLMSKTLVHLRGQKWKNIRGTWNMVLTTSKVERMYPLIYKMT
ESFVTSFNNVEENEVKTFNVKTVYRKYICDLSASIYFGIDKDFDIFREKLSNFVNF
NGNSIWTLLSFCVACLAPELSI

>CYP436A2 Query
McKenna_127722
revised DRN 4/9/12
84% identical to McKenna_46973
32% identical to CYP9V1 *Leptinotarsa decemlineata*
MNIFDVLSGFFFLLLLLLSVFIYICCTVDLKPFYYWSSKGVRTANILPVIGDAIFGIC
FSKKPYPKLIQDVYDSAGNSRYIGYSELRTPILMIKSPELIGKLLVRDAENFQDRR
VVFLGNSDTVMSKTLVQLRGQEWKNIRGIWNMVLTTS

CYP4 clan (45 sequences, at least 18 different sequences)

>CYP4G79 Query
McKenna_97615
revised DRN 4/1/12
78% identical to CYP4G14 *T. castaneum*
Missing first 73 amino acids
There appears to be only 1 CYP4G gene
VIFNHMIEKSKEFGKIIRMWIGHKLFVFLMHPDDVELILGSHEHIDKAPEYRFFKP
WLGDGLLISTGPKWRAHRKLIAPAFHLNVLKSFIDLFNTNSLDVVNRLKKTMGK
EIDCHDYMSEATVEILLETAMGVSKKTQDQSGYDYAMAVMKMCDILHLRHTK
MWLYPDVLFNLSKYKSYQDKLINTIHSLTRKVIKSKKAAFAKGIRGSIAEVPENL
KSKSVDNTEVKTVVEGLSFGQSAGLKDDLDVDDDIGEKKRMAFLDLMIEASQSG
VVINDEEIKEQVDTIMFEGHDTTAAGSSFFLCMMGLHTDIQEKVYEELNEIFKGS
DRPATFADTLEMKYLERCLMETLRMYPPVPLIARQLRRDLKLTSEDLTVPAGCT
VIIPTFKIHRDPVTYPNPEKFDPDNFLPERTANRHYYSFIPFSAGPRSCVGRKYAM
LKLKILLSHLLRNFRIVSDVPEAEYKLQADIILKREDGFRIKLEPRKKVAA*

>CYP4Q14
McKenna_93511
revised DRN 4/1/12
53% identical to CYP4Q6 *T. castaneum*
48% identical to CYP4BG3 Dendroctonus ponderosae
GEKWHARRKLLTPTFHFKILQDFLIAFNEETNKFVEKINGCLDDAGVDISKFIDNL
TLQAVGETAMGLELVDEGIMIDYRQNLRKMGKIIIERITKFFYRFDFIYRWTDLA
HEEKVVTQSLHRFSETVIRRRKVMRRGSTIERKRLAMLDLLLQYKEEGADIDDLG
IREEVDTFLFEGHDTTSIALSMLILLLANHKEVQERVIEEIDQVLGKKAKNPTHED
LPKLEYLECCIKETLRLYPSVPSIGRIAGEDFTTTTGYRIPKGTIMLIQIYDLHRDAS
VFPEPEKFNPDRFKPENTMNRNSFAYIPFSAGPRNCIGQKFAMWEIKAVMCGLLQ
HFKLAPIDTPMGIKFYTDLILRTNCPIKVKFIKRQ*

>CYP4Q15
McKenna_110020
revised DRN 4/4/12
65% identical to CYP4Q6 *T. castaneum*
EKAVAEILEAIGPDSDLAYSDTQKFPYLERCIKEGLRLYPSVPMISRTAGCDYVTS
TGYRIPAGTTLHLHIFDLHRKPYIYPDPDLFDPDRFLKENCSLRHPFAYIPFSAGPR
NCI

>CYP4Q16
McKenna_93461
revised DRN 4/5/12
87% identical to McKenna_69083
87% identical to McKenna_73025
67% identical to CYP4Q7v1 *T. castaneum*
NQIVEEINSVLEGEDRHPTYDDLQKMDLLERCIKESLRLYPSVHLISREVEEDTTL
HSGCVVAKGAMVVISIQFVHRNPEIYPDPEKFDPDRFLPENCIRRHPFAYLPFSAG
PRNCI

>CYP4Q17
McKenna_73025
revised DRN 4/5/12
94% identical to McKenna_69083
87% identical to McKenna_93461
65% identical to CYP4Q7v1 *T. castaneum*
EQIVEELNSILEGEERQPTYEDLQKMDLLERCIKESLRLYPSVHLISRKADEDTTL
HSGCVVAKGATVVIPIIFVHRNPEIYPEPEKFDPDRFLLENCIGRHPFAYLPFSAGP
RNCI

>CYP4Q18
McKenna_69083
revised DRN 4/5/12
94% identical to McKenna_73025
87% identical to McKenna_93461
65% identical to CYP4Q7v1 *T. castaneum*
EQIVEELNSILEGEERQPTYEDLQKMDLLERCIKESLRLYPSVHLISREADEDTTLH
SGCVVAKGATVLIPIMSVHRNPEIYPHPEKFDPDRFLPENCIGRHPFAYLPFSAGP
RNCI

>CYP4Q19
McKenna_65850
revised DRN 4/5/12
62% identical to McKenna_93461
65% identical to McKenna_69083
63% identical to CYP4Q7v1 *T. castaneum*

DNIVDEMMKVMDGRTTSPTYDDLQKMEYMERCIKETLRLYPSVYFISRVADEDT
VLNSGMLVPKGTHVHIHIYDVMRNPEIFPSPDTFDPDRFLPENCQHRHPFAYIPFS
GGPRNCI

>CYP4Q fragment 1
McKenna_72206
revised DRN 4/5/12
last exon
90% identical to McKenna_77757
76% identical to McKenna_11720
59% identical to CYP4Q5 *T. castaneum*
GQRFAMMELKVVLSAIIRKFQLVAVDTPDDIELKNESVLRANGIRVKFIPRV*

>CYP4Q fragment 2
McKenna_77757
revised DRN 4/5/12
last exon
50% identical to CYP4Q9P *T. castaneum*
GQRFAMMELKVVLSAIIRKFQLVAVDTPDNLELRIESILRANGIRVKFIPRV*

>CYP4Q fragment 3
McKenna_11720
revised DRN 4/5/12
77% identical to McKenna_77757
76% identical to McKenna_72206
52% identical to CYP4Q3 *T. castaneum*
GQRFAMMELKAVLSGIIRRFQLVAVSTPDDLEIKLETILRTNGIKMKFIPRV*

>CYP4Q fragment 4
McKenna_54774
revised DRN 4/5/12
78% identical to McKenna_72692
52% identical to CYP4M7 *Helicoverpa armigera*
IFLSTPKHTTKSLLYLIFGRWLQDGLLLSEGEKWQQRRKLLTPAFHFNILREFISVF
NEKSDDLSKKLSETAGKEIDVMPLFHECSLQMIL

>CYP4Q fragment
McKenna_60607
57% identical to CYP4Q fragment 5 McKenna_72692
before C-helix
IFMSTKKHSTKSEVYDNLKNWLKEGLLLSS

>CYP4Q fragment
McKenna_15994
81% identical to CYP4Q fragment 4 McKenna_54774

GEKWQQRRKLLTPAFHFNILREFICVFNEKSDELMEKLRETAGKEIDVMPLFHEC
SLQMII

>CYP4Q fragment 5 Query= McKenna_72692
revised DRN 4/3/12
60% identical to CYP4Q2 *T. castaneum* C-helix region, runs off the end of the contig
IFLSTPKHSSKSVLYQIFGRWLKEGLLLSKGEKWQQRRKILTPAFHFSILKEFIDAF
NEKSDELMEKLRETAGT

>CYP4Q fragment
McKenna_88883
73% identical to CYP4Q7 *T. castaneum* sequence7v1, AAF70496.1
GHDTTAAAIGFCLMLIANHPEVQ

>CYP4Q fragment
McKenna_93842
52% identical to CYP4 fragment 6 McKenna_92563
I-helix
GHDTVATALSFCVMNIANHPEIQ(0)

>CYP4Q fragment
McKenna_61221
52% identical to CYP4Q4 *T. castaneum* just past the heme signature
KVAMLGIKTTICGILKKFRLEPIGSLEDIIFVPHLLLKSKHPLEVKFIPRVSL*

>CYP4AA1v2a Query
McKenna_22609
MVLKLHEYILTTFYFLFLFYCLWKLRNYFRAVLLALHLPGPFAYPIIGNALTLFFF
TELEYLGNNSYRLFGPIFRCWISIVPFFFVTDPAHLQTLLSNGRLTKKNMFYSLLH
NFIGEGLITNNGEKWKLHRKLIQPYFHINVLENFIPIFAETSRDLATNFKDITEVNIT
TFINDWVLDTLH

>CYP4AA1v2b
McKenna_56446
revised DRN 4/3/12
54% identical to CYP4AA1 *T. castaneum*
97% identical to CYP4AA1v1 McKenna_18780
ELIVPYRISRPWMLIDFIFNLTKSADQEQKQRSNLHEFTKKLLYSRRNQEISSIKFV
SLMEIFMNLSETNTDFSEQDVIDETCTFMLAGQDSVGAATAFTLHFLAKHEEIQR
KVYEEQVRIFENDTRLVTTNDLNEMRYLEQCIKETMRLCPSVPIVCRKLNEDLRL
GNYVLPGGTNIFISPFITHRLEHLYPEPQKYDPDRFSAENSHKIHPYGYIPFSAGPR
NCIGYKFAMLEMKAAISAIMREYQLSLVPGKEETIFSYRITLRCKGGIWIRLTRRD
KDL*

>CYP4AA1v1
McKenna_18780

revised DRN 4/3/12
identical to 22609, almost identical to 56446
48% identical to CYP4AA1 *T. castaneum*
MVLKLHEYILTTFYFLFLFYCLWKLRNYFRAVLLALHLPGPFAYPIIGNALTLFFF
FTELEYLGNNSYRLFGPIFRCWISIVPFFFVTDPAHLQTLLSNGRLTKKNMFYSLL
HNFIGEGLITNNGEKWKLHRKLIQPYFHINVLENFIPIFAETSRDLATNFKDITEVNI
TTFINDWVLDTLHMSSFYKDQNSRKCVSVLYSFRGELIVPYRISRPWMLIDFIFNL
TKSADQEQKQRSNLHEFTKKLLYTRRNQEISSSIKFVSLMEIFMNLSEANTDFSEQ
DVIDETCTFMLAGQDSVGAATAFTLHFLAKHEDIQRKVYEEQVRIFENDTRLVTT
NDLNEMRYLEQCIKETMRLCPSVPIVCRKLNEDLRLGNYVLPVGTNIFISPFITHR
LEHLYPEPQKYDPDRFSAENCHKIHPYGYIPFSAGPRNCIGYKFAMLEMKAAISAI
MREYKLSLVPGKEETIFSYRITLRCKGGIWIRLTRRDKDL*

>CYP4AW fragment
McKenna_58185
revised DRN 4/4/12
57% identical to CYP4AW1 *Phyllopertha diversa* scarab beetle
EKVFNEQSEIYSKKTDAQITYADLVEMKYLEMVIKESLRIHTPIPFFARKLEEDTY
Y

There appears to be only 1 CYP4BM gene with 2 alleles represented

>CYP4BM fragment 1
McKenna_25162
revised DRN 4/4/12
57% identical to CYP4C62 *Laodelphax striatellus*
52% identical to CYP4BM1 *T. castaneum*, best BLAST hit to 4BM1
probable N-terminal region for 4BM1-like gene, exons 1–3
MSPLSYFNIFFLTELYQSIYDRTEKFGPIFRSWVGFIPQIHLTRARHAEIILRSSVNIT
KGMNYTFVKHWLGDGLITGTGSYWQRHRKLITPTFHFKILDSFQEVFSEKAHLLL
DELKPLADGKFFDISTLVTHCALDIIC

missing exon 4 (24 amino acids)

>CYP4BM fragment 2v1
McKenna_22342
revised DRN 4/4/12
59% identical to CYP4BM1 *T. castaneum* up to I-helix exons 5–6
ILEIFIYRWFRPWLHSDFIFALTSKGREQKKVLEILHGFSNKVIADRKKIIQNSKGV
EELSEEDMLLGKKRRLAFLDLLLQQNMEKNEWTDTELREEVDTFMFA

>CYP4BM fragment 2v2
McKenna_126849
revised DRN 4/4/12
58% identical to CYP4BM1 *T. castaneum*
100% identical to McKenna_22342, exon 6

VIADRKKIIQNSKGVEELSEEDMLLGKKRRLAFLDLLLQQNMEKNEWTDTELRE
EVDTFMFA

>CYP4BM fragment 3v1
McKenna_27981
revised DRN 4/4/12
53% identical to CYP4BM1 *T. castaneum*, I-helix to EXXR region
exons 7–8
GHDTTTSSVLWNLFVLGNSPKFRVYEEIDSVFHGEERPIMPEDIAKMQYMERVM
KETLRIYSVVPYIMRRLEEDTEI

>CYP4BM fragment 3v2
McKenna_57285
revised DRN 4/4/12
100% identical to McKenna_27981 exon 8
VYEEIDSVFHGEERPIMPEDIAKMQYMERVMKETLRIYSVVPYIMRRLEEDTEI

>CYP4BM fragment 4
McKenna_51726
revised DRN 4/5/12
75% identical to CYP4J20 *Culex quinquefasciatus*
71% identical to CYP4BM1 *T. castaneum* (probably a missing exon in the CYP4BM gene)
exon 9
EGTIIPAGVCVAIHITNVHKDPEQFPDPFRFDPDRFLPENVAKRNPYAHIPFSAGPR
NCI

>CYP4BM fragment 5
McKenna_104318
revised DRN 4/2/12
53% identical to CYP4BM1 *T. castaneum*, exon 10
GQKFAIRNTKTMLTAILRKYKIKSKLKPEDMKFYGDIILKPQEGIFISLEPRN*

assembled hypothetical gene

MSPLSYFNIFFLTELYQSIYDRTEKFGPIFRSWVGFIPQIHLTRARHAEIILRSSVNIT
KGMNYTFVKHWLGDGLITGTGSYWQRHRKLITPTFHFKILDSFQEVFSEKAHLLL
DELKPLADGKFFDISTLVTHCALDIIC
missing exon 4 (24 amino acids)
ILEIFIYRWFRPWLHSDFIFALTSKGREQKKVLEILHGFSNKVIADRKKIIQNSKGV
EELSEEDMLLGKKRRLAFLDLLLQQNMEKNEWTDTELREEVDTFMFAGHDTTT
SSVLWNLFVLGNSPKFRVYEEIDSVFHGEERPIMPEDIAKMQYMERVMKETLRIY
SVVPYIMRRLEEDTEIEGTIIPAGVCVAIHITNVHKDPEQFPDPFRFDPDRFLPENV
AKRNPYAHIPFSAGPRNCIGQKFAIRNTKTMLTAILRKYKIKSKLKPEDMKFYGDI
ILKPQEGIFISLEPRN*

>CYP4DW1v1
McKenna_56469
revised DRN 4/3/12
35% identical to CYP4Q5 *T. castaneum*
98% identical to McKenna_59987
VILGDPKFTSKGRLYEPSKFWLGEGLLVSAGEKWRKGRKLLTKTFHFGVLKNYM
HIFNEQLEILNKSFESKHGEPAQLTSLLQFHSLRIICATTVGGQVEISKKSGESLLNS
LETLTKIVGIKMTIPLMNFLYKFTYLLREEREALQDYNDFAYRLIEKNDDCKDMD
EDADHPRLLKVLLQNCDSEDVKDHMKNFLFAGQDTMTTTLTFYLYVLANKPEI
QDEILNEILAISVNENPTYGEITQMGLLDRFIKECLRLYSPAPFIGRTVEENVSLPSG
YTIPAGTSVFIDIFDIHRHPKLYPNPENFDPSRFLPENCGKRHPYSFIPFSAGPRNCI

>CYP4DW1v2
McKenna_59987
revised DRN 4/3/12
29% identical to CYP4BN11 *T. castaneum* amino acids 116–344
64% identical to McKenna_90133
69% identical to McKenna_35031
69% identical to McKenna_65594
97% identical to McKenna_56469
VILGDPKFTSKGRLYEPSKFWLGEGLLVSAGEKWRKGRKLLTKTFHFGVLKNYM
HIFNEQLEILNESFESKHGEPAQLTSLLQFHSLRIICSTTVGGQVEISKKSGESLLNS
LETLTKIVGIKMTIPLMNFLYKFTYLLREEREALQDYNDFAFRLIEKNDDCKDMD
EDTDHPRLLKVLLQNCDSEDVKDHMKNFLFAGQDTMTTTLTFYLYVLANKPEIQ

>CYP4DW2
McKenna_41406
revised DRN 4/3/12
82% identical to McKenna_56469
46% identical to CYP4Q5
DEILNEIMTTSANENPTYGEITQMGLLDRFIKECLRLYPPAPFIGRTIDENVGLPSG
YTIPAGTFIFMSIFDIHRHPKLYTNPENFDLNRFSPENCRKRHPFSFIPFSAGPRNCI

>CYP4DW3
McKenna_73302
revised DRN 4/3/12
93% identical to McKenna_41406
NEILNEIMTTSANENPTYGEITQMGLLDRFIKECLRMYPPAPFIGRTIDEDVGLPSG
YTIPAGTFIFMNIFDIHRHPKLYPNPEIFDLNRFSPEHCRKRHPFSFIPFSAGPRNCI

>CYP4DW4v1
McKenna_35031
revised DRN 4/3/12
27% identical to CYP4AB6 *Nasonia*

GEKWKKNRRLLTKSFHFGVLKNYMHIFNEQLENLKSSLNSNNGEPTELISVMKY
HSLNIICTTLAGDQFGIEKTSKKFLSSLETLTKMVYIKMTVPLMNFLYKFTYLLKE
ETEAMQSYKDFARALIAKNCDSNNVDDDTDLPRLLKILSRNCDSQGVEEQMNTF
LFAGQDTTTTSLTFFLYVLANKPDLQ

>CYP4DW4v2
McKenna_65594
revised DRN 4/3/12
25% identical to CYP4AB6 *Nasonia*
63% identical to CYP4DW1v2 McKenna_59987
95% identical to McKenna_35031
GEKWKKNRRLLTKSFHFGVLKNYMHIFNEQLENLKSSLYSNNGEPTELISVMKY
HSLNIICTTLAGDQFGIEKTSKKFLSSLETLTRMVYIKMTVPLMNFLYKFTYLLKE
ETEAMKSYKDFARALIEKNRDSNDVDDDTDLPRLLKILSRNFDSQGVEEQMNTF
LFA

>CYP4DW fragment 1
McKenna_90133
revised DRN 4/3/12
44% identical to CYP4AX1 *Bombyx mori*
91% identical to McKenna_35031
90% identical to McKenna_65594
VILSNPKYISKGIFYEPMRYWLGDGLLVSA
GEKWKKNRRLLTKSFHFGVLKNYMHIFNELSENLKSCLNSNNGEPTELISVMEY
HSLNIICA

>CYP4 fragment 1
McKenna_82098
revised DRN 4/9/12
42% identical to CYP4S4 *Mamestra brassicae* N-terminus
MLPSVKLGIWMAFFTLLLLIGKMLYRHLRDILVLLEIPAPPAKPIIGLGTEFFGVSQ
EEIFRKFREYSKQFRPVYRLPLFHIQAINLINSGDIEQVLSSYNHLKKSMTYDFLNK
WLGTGLLTSS

>CYP4 fragment 2
McKenna_44239
revised DRN 4/5/12
exon 2 near N-terminus
41% identical to CYP4Q1 *T. castaneum*
37% identical to McKenna_82098
DVIHQSLRNYAKEFYPHYRLRILHSISVNIVSPEDCE

>CYP4 fragment 3
McKenna_48071
revised DRN 4/5/12
C-helix region

44% identical to CYP4BN11 *T. castaneum*
53% identical to CYP4AW1
GMLRAFNDYTKKYGDIVYTRIGPLYHGLLFTDADLAKEVFQANINLTKGSAYEF
CRSWLGHGLLTTDEKRWKKQRKIVTPAFNTQLLIEFIPVFDKQSNILIEKLDNAPS
KDSLNIHRLIGLCSLDITC

>CYP4 fragment
McKenna_63047
91% identical to CYP4 fragment 3 McKenna_48071
immediately before the ETAM exon, runs off the contig
KDSLNIHRIIGLCSLDITC

>CYP4 fragment 4
McKenna_128699
revised DRN 4/5/12
C-helix region
39% identical to CYP4AY2 *Ips paraconfusus*
45% identical to McKenna_54774
ELLPYLHEQLDKFDGLMQIHIGTEVMLTASNYKFVQWLMTSTTIITKSLQYSFFN
EWLGHALLITSGSRWKSYRKILTPAFHFKRIENLISVCQKASDDLIKKMSNNLDK
DCFDVYPIISNFSLDVLC

>CYP4 fragment 5
McKenna_89270
revised DRN 4/5/12
ETAM exon, 43% identical to CYP4Q12 *Leptinotarsa decemlineata*
40% identical to McKenna_113570
ETAMGYKLNTQEEHSKDYIKALHEIELLFNYRMMRPWLHIPLIYWFNSTSRREA
QLLKILHGFTKAIIKERMDSFECDVLSLQSDEKSEGKLQNKTRRRRLVLLDVLLQ
TRAIDGSIDYEGICEEVDTFMFE

>CYP4 fragment
McKenna_41802
ETAM exon 82% identical to CYP4 fragment 5 McKenna_89270
ETAMGYKLNTQEEHSRNYIKAVHEIEKLFLHRMMRPWLQIPLIYWFSPISRIEAQ
KLKVLHGFTRAIIKKRMDSFECDVISSQSDERSEDKLQNKTKRRRLAFLDVLLQA
RAIDGSIDYEGICEEVDTFMFE

>CYP4 fragment 6
McKenna_92563
revised DRN 4/5/12
ETAM exon and I-helix
45% identical to CYP4BQ1 *Dendroctonus ponderosae*
ESSMATSINAQEEESEYRSCIRTLCQIILERAMDPLYMHDLLFYFHPQYQNFQKSI
KTIHEFDQTVIEKRRKLLQNVHEKNDDDLENVYGKKKTPFLDILLKARDEDGNP
LSNKDIRDQVDGIMFAGHDTTASAISFILYNLSVHPDVQ

>CYP4 fragment
McKenna_58192
65% identical to CYP4Q8 *T. castaneum* sequence19, XP_970987.1 I-helix
44% identical to CYP4 fragment 6 McKenna_92563
GHDTVSTALSFCAMNIANHPEIQ

>CYP4 fragment 7
McKenna_113570
revised DRN 4/5/12
ETAM exon up to I-helix
40% identical to CYP4M14 *Spodoptera frugiperda*
40% identical to McKenna_89270
ETAMGYKTKTNDNAMKKYLKAVEGLEKVINLFIVRPWLRYFGVFYRFTSYGKE
EGRHLDVLHNYTKTIIKEKMEISKNNPSESSEIEKDDIYFGKKRRRMAMLDILLEA
HRNGNQIDFNGICEEVDTFTFE

>CYP4 fragment 8
McKenna_70904
41% identical to CYP4BN4 *T. castaneum*
revised DRN 4/5/12
ETAM exon up to I-helix
41% identical to CYP4BR4 *Dendroctonus*
41% identical to CYP4BN4 *T. castaneum*
ETSMGVKLGAQNEQNTEYVLAVKEMCRIIATRSYSMMKAMKLTYPFTEDYSIEK
KSVKILHGFTDRVVQKKREQRKEQTESNLDQDGRSRRSNLLDILLDYSDKEKLLS
EKEIRDEIHTFMFA

>CYP4 fragment 9
McKenna_97724
revised DRN 4/1/12
53% identical to CYP4BN4 *T. castaneum*
RRDFNFSDGKLIPKNTTilillhhlHRNAEVFPDPEKFDPTRFDDNSKIPNFAFLPFSAG
LRNCIGQRFAMLEIKVAVAELVRNFEFLPAPNYKPNVISEISLKSSNGICIRLKKRQ
*

>CYP4 fragment 10
McKenna_89867
revised DRN 4/4/12
last exon
43% identical to McKenna_97724
51% identical to CYP4BN12v2 *Leptinotarsa decemlineata*
GQKFANNEMKVILSKLIRTFEFRPAEPDHELDLRAEVVLTSKNGINVKIIRRQQ*

>CYP349 fragment 1

McKenna_59612
revised DRN 4/4/12
N-terminus
80% identical to McKenna_43307
39% identical to CYP349A1 *T. castaneum*
DRLMFINISEIFETFLNYFRNSPDIFKLWFGHMLIIGVSKPEHMEIVLTNPNTMNKS
HLVDFTKPYMGDGLFSAS

>CYP349 fragment 2
McKenna_43307
revised DRN 4/4/12
N-terminus
36% identical to CYP349A1 *T. castaneum*
82% identical to McKenna_59612
DRLMFINISDILKILLSYYRNSPDIFKFWLGHILFIGVSKPEHMEIVLTNPNTMNKS
HLVAFTKPYMGDGLFSAT

>CYP349 fragment
McKenna_51341
47% identical to CYP4BM1 *T. castaneum* sequence89, XP_966563.1 from KYG motif near N-terminus
44% identical to CYP349 fragment 2 McKenna_43307
YGPNFAVWILDHIFIVVGKPDDVEKILQSPSCLTKNELYRFTHGIVGTGLFTAP

>CYP349 fragment 3
McKenna_36044
revised DRN 4/4/12
C-helix region and ETAM exon
42% identical to CYP349B1 *Dendroctonus ponderosae*
AEKWRQHRKIISPTFNAKILEGYLPIFCKTGNIFVDEILPNNVDKKDADWYTLFTA
VNLDVIM QTAMGIDKDVQRKDVPFGQWLEK

>CYP349 fragment 4a
McKenna_104719
revised DRN 4/4/12
63% identical to CYP349A2 *T. castaneum* EXXR region
QKVYEEAIEVLGQDRYPTKADIPKLQFTEMFIKETLRLFPIAPLFLRVASDNFRM

>CYP349 fragment
McKenna_114597
90% identical to CYP349 fragment 4 McKenna_104719
QKVYEEAIEVLGQDRYPTKADIPKLKFTEMFIKETLRLFPIGPFLLRVASDDFRM

>CYP349 fragment 4b
McKenna_106919
Matches CYP349 fragment 4 McKenna_104719

1-amino acid difference
QKVYEEAIEVLGQDRYPTKADIPKLQFTEMFIKETLRLFPIAPLFLRVASDDFRM

>CYP349A fragment
McKenna_74465
60% identical to CYP4G14 *T. castaneum* sequence9, gi|91080899|ref|XP_973423.1
60% identical to CYP4G79 McKenna_97615
60% identical to CYP349A3P *T. castaneum* sequence64, XP_968399.1
GTDTSAVTACFFLTMMGMHQDIQ

>CYP349 fragment 5
McKenna_121359
revised DRN 4/4/12
50% identical to CYP4C12 *Mastotermes darwiniensis* AF067632, EXXR region
55% identical to CYP349A2 *T. castaneum*
EKVYEEVMGVLGPEGQPTVKDLNEMHFFERCLKETMRLFTPAPIILRKNTGGDL
KI

>CYP349 fragment 6
McKenna_48771
revised DRN 4/4/12
last 2 exons
50% identical to CYP349B2 *Dendroctonus ponderosae*
52% identical to CYP349A2 *T. castaneum*
DDLTIPKDSIILIGIMHLHRSPKYWEDPLKFDPNRFLPEKLSKMHPYSYLPFSGGPR
ICFGYRYAMIVMKLILSKIIRKFELKTEYKSIEEIQLKINLMMRPSNGFKVTLQPRE
*

Mito clan (23 sequences, possibly 6 different genes)

>CYP49
McKenna_33418
revised DRN 4/3/12
N-terminus
MKKKIIEISKSVQNFAVRRSYATDRNYSTVMKLDIFEDDIEVISKEANVEEKEYIK
PYSDVPGPQQLPIIGNAWRFAPFIGQYKIHEL

>CYP49
McKenna_15733
identical to contig_33418
MKKKIIEISKSVQNFAVRRSYATDRNYSTVMKLDIFEDDIEVISKEANVEEKEYIK
PYSDVPGPQQLPIIGNAWRFAPFIGQYKIHEL(gap)

>CYP49

McKenna_53778
83% identical to CYP49A1 *T. castaneum* amino acids 196–291
EESLDNNQELPPHFLSEIYKWALESVVRVSLDTRLGCLEPNLPKDSEQQKIIDSINT
FFWNVAEVELKMPVWRIYHNNAFKKYIGALENFRI

>CYP49 Query
McKenna_118324
revised DRN 4/3/12 amino acids 326–381 up to I-helix
65% identical to CYP49A1 *T. castaneum*
ALCSKYIQQTMANMDLKDFKNMKKENISIVEKILLQTGNPKLATVLAIDLLLVGV
DT

>CYP49Q
McKenna_35973
revised DRN 4/3/12 amino acids 326–381 up to I-helix
1-amino acids difference from 118324
65% identical to CYP49A1 *T. castaneum*
LCSKYIQQTMANMALKDFKNMKKENISIVEKILLQTGNPKLATVLAIDLLLVGVD
T(gap)

>CYP49
McKenna_54335
66% identical to CYP49A1 *T. castaneum* sequence87, gi|91086895|ref|XP_970738.1
I-helix to EXXR
TSIAVASTLYQLSQNPDKQEKLYQELKTVLPTANSKFTAETQENIPYLKACIKETL
R

>CYP49 Query
McKenna_70895
revised DRN 4/3/12
Heme signature
MYPVIIGNGRSLQSDTEIGGYHIPKGTHVIFPHLVVSNTEEYFSEPQRFEPERWLK
KKDESTKCPIKQDKIHPFVSLPFGYGRRSCLGRRFAETELQILLSK

>CYP49
McKenna_121511
revised DRN 4/3/12
81% identical to C-term of CYP49A1 DPO024_I03 *Dendroctonus ponderosae*
IFRKYRVDYNYEALTYKITPTYVPEQPLKFKLTERMS*

>CYP49A1 assembled sequence
74% identical to CYP49A1 *T. castaneum* sequence87, gi|91086895|ref|XP_970738.1
MKKKIIEISKSVQNFAVRRSYATDRNYSTVMKLDIFEDDIEVISKEANVEEKEYIK
PYSDVPGPQQLPIIGNAWRFAPFIGQYKIHEL
(gap)

EESLDNNQELPPHFLSEIYKWALESVVRVSLDTRLGCLEPNLPKDSEQQKIIDSINT
FFWNVAEVELKMPVWRIYHNNAFKKYIGALENFRILCSKYIQQTMANMALKDF
KNMKKENISIVEKILLQTGNPKLATVLAIDLLLVGVDTTSIAVASTLYQLSQNPDK
QEKLYQELKTVLPTANSKFTAETQENIPYLKACIKETLRMYPVIIGNGRSLQSDTE
IGGYHIPKGTHVIFPHLVVSNTEEYFSEPQRFEPERWLKKKDESTKCPIKQDKIHPF
VSLPFGYGRRSCLGRRFAETELQILLSKIFRKYRVDYNYEALTYKITPTYVPEQPL
KFKLTERMS

>CYP301A1 *H. axyridis*
revised DRN 6/11/12, complete
76% identical to CYP301A1 *T. castaneum*
McKenna_84359 exon 1
McKenna_104580 exons 2–3
McKenna_59440 exon 4–5
CYP301A1 fragment McKenna_32237 2 exons
McKenna_37958 last exon
McKenna_59965 last exon
MSKRIAAKLKKNSLRDWMQYSTSTIARAEGISCPHIQDVNIKPYSEIPGPKPLPFL
GNTWRLLPVIGQYDISDVGKLSKRFHEQYGKIVKLSGLVGRPDLLFVYDADEIQK
VYSNEGPTPFRPSMPCLVKYKSEVRKEFFGDLAGVVGVHGEPWKTFRTKVQRPI
LQLKTVKKYITPIEEVTNYFIERILEMKDDKDEMPGDFDNEIHKWSLECKVSLDT
RLGCLDPNLPSDSEPQKIINAAKYALRNVAILELKFPFWRYFPTTVWTNYVKNM
DYFIEICMKHIDAAMERLKSKSITNENELSLIERILASEPDPKTAYILALDLILVGID
TISMAVCSILYQLATRPNEQEKLYQELRRVLPDPKTPLTASLLDEMVYLKAFVKE
VLRMYSTVIGNGRTLQEDTVIQGYHVPKGVVVFPTLVTGNMPEFVSEPSKFIPER
WMKDSGLDYKLHPYASLPYGHGARMCLGRRFADLEIQVLLAKLVRSFKLEYNH
EPLEYKVTFMYAPEGELKFKMTPRHETKE*

>CYP301A fragment
McKenna_80357
83% identical to CYP301A1 *H. axyridis* N-terminus
MSKRIAAELKKNGLRDWKQYSTSTITRAEGVSCPHIQEVNIKPYSEIPGPKPLPFL

>CYP301B.a Query
McKenna_28250
revised DRN 4/3/12
N-terminus, exon 1
MLVKNVRFSRKLISRPFSQTLSPGVIDGVSSEWDNASPYSSIPGPKALPLVGNTWR
FLPYI

>CYP301B.b Query
McKenna_38292
86% identical to contig_28250
exon 1
MLVKNVRFSRKLITRSFSQTLNPGVIDGLSSEWDSALPYCSIPGPKALPLVGNTW
RFLPYV

>CYP301B.a
McKenna_117319
revised DRN 4/3/12
exons 2–3
65% identical to CYP301B1 *T. castaneum*
GGFQIEHIDKLCKQLHQKYGKIVKMEGLLGRPDMLFLFDPDLIQRVFKQEDNLPY
RPSMPSLTYYKHKHKKEIFGEDGGVIAVHGEEWQKFRSKVNQIMLHASAAHQYI
DTINEAGNNFIER

>CYP301B.b
McKenna_25931
only 4-amino acid difference from 117319
GGFQIEHIDQLCKQLHQKYGKIVKMEGLLGRPDMLFLFDPDLIQRVIKQEDNLPY
RPSMPSLTYYKHKHKKEIFGEDGGVIAVHGEEWQKFRSKVNQIMLQATAAHQYI
DTINEAGNNFIER

>CYP301B Query
McKenna_30862
revised DRN 4/3/12
62% identical to CYP301B1 *T. castaneum*
IEFLKDDKEEVPGDFLNEIHKWSLESLARVALDIKLNCLAENPKKHTQELIDAVN
TFFMGVPILELKNPMWRIISTPLFKKYIKALDTIYELCNIHIEDAMRNPSSSTDENC
SVLQKVLRNNNPKTAKHLALDLFLVGIDTTSNAVASILYQLAQHQDAQEKLYKH
LVSSKITANSDITIDFLKNSGYLKFCIKETMRMFPVVIGNGRCTTNNTVIGGYQVP
KGVQVIFQHYVISNLDEYFPRSSEFLPERWSDPKGKTHNFASVPFGHGRRMCLGK
RFADMEMQVVISK

>CYP301B1 assembled sequence (note: there are two exon 1s and two exon 2, 3s)
revised DRN 4/3/12
McKenna_28250 exon 1
McKenna_117319 exons 2–3
McKenna_30862 exons 4–9
60% identical to CYP301B1 *T. castaneum*
missing the last exon
MLVKNVRFSRKLISRPFSQTLSPGVIDGVSSEWDNASPYSSIPGPKALPLVGNTWR
FLPYIGGFQIEHIDKLCKQLHQKYGKIVKMEGLLGRPDMLFLFDPDLIQRVFKQE
DNLPYRPSMPSLTYYKHKHKKEIFGEDGGVIAVHGEEWQKFRSKVNQIMLHASA
AHQYIDTINEAGNNFIERIEFLKDDKEEVPGDFLNEIHKWSLESLARVALDIKLNC
LAENPKKHTQELIDAVNTFFMGVPILELKNPMWRIISTPLFKKYIKALDTIYELCNI
HIEDAMRNPSSSTDENCSVLQKVLRNNNPKTAKHLALDLFLVGIDTTSNAVASIL
YQLAQHQDAQEKLYKHLVSSKITANSDITIDFLKNSGYLKFCIKETMRMFPVVIG
NGRCTTNNTVIGGYQVPKGVQVIFQHYVISNLDEYFPRSSEFLPERWSDPKGKTH
NFASVPFGHGRRMCLGKRFADMEMQVVISK

>CYP302A1

McKenna_59802
58% identical to CYP302A1 *T. castaneum* sequence107, gi|91083869|ref|XP_974252.1
MCLKRKILSQLTSNKRELSTDIRKPFTSIPGPISLPVIGTLYQYIPLI
GKYKFTKLETTGLKKYKKFGSIVKEEIAPGVNTVWLYDPNDIEHLFRNEGIYPRR
RSHLALEKYRLDKPHVYNTGGLLPTNGENWWKLRQVFQKGLSSPMAVHNFISG
SNEIIDEFLDRINYIKEYSNVDYLPEISRLFLECMYVFDVRMDSFSEIELRKNSRST
KLLKSALTTNSCILKLDNGFHLWKHFATPLYWRMKRAQTYMEDVAIDLVGIKM
HTYEELPYNHPKCLMDIYLRSKDLDFKDIIGISCDFLLAGMDTTSYTTSFLLYYLA
LNKEVQNRLWEEVKRLLPNSNSPVTKEVLAEAHYAKACLKETHRLRPISVGVGR
ILDRPTVFSGYEVPEDTIIVTQNQVSCRLEDYFPEANKFLPERWLKNDPLYRKIHP
FLVLPFGHGKRSCIARRLAEQNMLILLLKISRCYEVGWIGYKLDTISSLINKPDGPI
LLNFQAR*

>CYP314A1v1 Query
McKenna_124192
revised DRN 4/3/12
48% identical to CYP314A1 *T. castaneum*
MISQIVSLTPDKLAIVLIILLIYYLDYRPPWWNRTEKKRQTIPGPKPIPLFGTNWLF
YLGCYNIHKIRDFYMAMYKKYGPIVKQETYFNFPIYSVFERKDIEKVLKVPSKYP
LRPASEAVIAYRASRPDRYASAGITNTQGETWHYLRTTLTPPLISPKTMSSFVTEV
NSLAEDWVSYLGRIRGPDNQIKDLAEIVVPLCIETTLDLVLGRRLGFLSPNPISSKS
RQLVEALEGHFVGLRDTQFNFPWWKFFPTKAYKTLTTCENYIYETVLEMVAEYA
DEEGEETVYKSLLKADIDQREKTGAIIDYISAGIHTLKNSLIFLFHLVAEHPELQGK
IGADLSYAKACMHEAFRLLPTATPLSRILDQDMELGGHQVKAGSIVLCHGDIASR
NEENFERPDEFIPERWLGDDKHKNISAGTYIMIPFGAGRRICPGKRFIELVLPIFLQ
QTVKKFELTTEHKMEVEFQFLTAPKGAVSIKLTDRD*

>CYP314A1v2
McKenna_45495
revised DRN 4/3/12
3-amino acid difference from McKenna_124192 allele
MISQIVSLTPDKIAIVLIILLIYYLDYRPPWWNRTEKKRQTIPGPKPIPLFGTNWLFY
LGCYNIHKIRDFYIAMYKKYGPIVKQETYFNFPIYSVFERKDIEKVLKVPSKYPLR
PASEAVIAYRASRPDRYASAGITNTQGETWHYLRTTLTPPLISPKTMSSFVTEVNS
LSEDWVSYLGRIRGPDNQIKDLAEIVVPLCIETTLDLVLGRRLGFLSPNPISSKSRQ
LVEALEGHFVGLRDTQFNFPWWKFFPTKAYKTLTTCENYIYETVLEMVAEYADE
EGEETVYKSLLKADIDQREKTGAIIDYISAGIHTLKNSLIFLFHLVAEHPELQGKIG
ADLSYAKACMHEAFRLLPTATPLSRILDQDMELGGHQVKAGSIVLCHGDIASRN
EENFERPDEFIPERWLGDDKHKNISAGTYIMIPFGAGRRICPGKRFIELVLPIFLQQ
TVKKFELTTEHKMEVEFQFLTAPKGAVSIKLTDRD*

Note: There are 3 sequences for the N-terminus of CYP315A and 2 sequences for the C-terminus. They are labeled a, b and c, but it is not clear which ones join together.

>CYP315A1.a
McKenna_71811

NOTE: this was labeled McKenna_24087
CYP315A 45% identical to CYP315A1 *T. castaneum* amino acids 1–315
MSMRRELCVRLSKITIQSKRDCSKIARDFRDVPSPRGLPIIGTRLAILAAGSSKYLH
EYVDKRHRELGPIYRERLGPVTGVFVSDPDSIRSIFAQEGKYPIHVIPEAWTLFNQ
KHNYDRGLFFMNGQEWLDTRRIMNNLLLKGDMSWMEEAADAVITGFLEDLEK
NGSLRTSLDSEFYQVFLEVIVSVLLGADTYRKHRDVIQQRVNHLAQSVQLVFETT
MKLEMIKAEWAEKLGLKRWKNFESSMLNALSGTARMLERLTEECGRGNGLME
MLRAKNLSEDRINAIVTDLFLAAADTTAYTMQWMVYMVAKNQNIQ
(86 amino acids gap)

>CYP315A1.a
McKenna_80432
54% identical to CYP315A1*T. castaneum*
amino acids 400–465 C-terminus
ASLPFAMGSRSCVGKKLAEYELQTMLSQLVKNFKIELLNTKEIRMVMKMIAAPS
EPIRLNLKKIG*

>CYP315A1.b
McKenna_40698
46% identical to CYP315A1 Colorado potato beetle
amino acids 25–316
91% identical to contig_71811
MSLTRELFVRFNIISILSKRDCSKVVRDFRDVPSPRGFPILGTRLAILAAGSSKYLH
EYVDKRHRELGPIYRERLGPVTGVFVSDPDSIRSIFSQEGKYPIHVIPEAWSLFNQ
KHNCARGLFFMNGQEWLDTRRIMNNLLLKGDMSWMEEAADAVITGFLEDLEN
NGSLRTSLDLEFYQVFLEVIVSVLLGADTYRKHRDVIQQRVDHLAQSVQLVFETT
MKLDMLKVKWADTLGLKRWKHFESSMLNALSGTTRMLERLTEECGRGNGLME
MLRAKNLSEDRINAIVTDLFLAAADTTAYTMQWMVYMVAQNQNIQ

>CYP315A1.b
McKenna_2304
amino acids 417–471 C-terminus, 4-amino acid difference from contig_80432 93%
XXXXXXMGSRSCVGKKLAEYELQSMLSQLVKNFKIELLNTKEIRMVLKMIAAPS
EPLRLSLKKI*

>CYP315A1.c
McKenna_55760
amino acids 169–324 92% identical to contig_40698, 91% identical to McKenna_71811
40% identical to CYP315A1 Colorado potato beetle
TGFLEDLDKNGFLGTSLDSEFYQVFLEVIVSVLLGADTYRKHRDVIQKRVNQLAE
SVQLVFETTMKLDMLKVKWAEKLGLKRWKHFESSMLNALSGTARMLERLTEE
CGHGNGLMEMLRAKNLSEDRINAIVTDLFLAAADTTAYTMQWVVYMVAKNQN
IQ