

EXPLORING THE EVOLUTION OF SOCIAL BEHAVIOUR USING GENOMIC DATA

BY

AARTI VENKAT

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Crop Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Associate Professor Matthew Hudson

ABSTRACT

Sociality is at the root of tremendous ecological success of several taxa, including humans, ants, bees, wasps and termites. The degree and type of sociality varies greatly across taxa. The evolution of complex social behaviour can be studied by performing comparative analyses of organisms across a phylogeny showing diverse social lifestyles. We chose bees as model systems for this study because a wide range of social behaviour patterns, ranging from highly eusocial to solitary can be found in extant bees. Our aim is to identify adaptive changes in the protein coding regions of brain expressed genes. To this end, we used 454 GS FLX sequencing technology to generate the brain Expressed Sequence Tags (ESTs) of twelve socially diverse bees. The ESTs were assembled into species-specific non-redundant contigs and singletons, which were loaded into a MySQL database using custom scripts. The Honey Bee Homolog Blast website was designed to help users access the database. Users can now download these datasets or BLAST against multiple bee and wasp databases to find the homologues. The results are then sorted by e-value and displayed. The ESTs accessed through the website (<http://bee12.cropsci.uiuc.edu>) can be used as a primary tool for gene discovery, genome annotation, and comparative genomic analysis. Since the Honeybee *Apis mellifera* had its genome recently sequenced, we designed an ortholog identification pipeline that generates multiple sequence alignments of putative orthologous genes across the twelve bees, using the gene models of *Apis mellifera* as the reference. The evolutionary changes associated with these alignments were then statistically inferred using maximum likelihood methods that make use of sophisticated codon-substitution models to detect non-neutral evolution in the protein coding genes. The rapidly evolving genes were then annotated using gene ontology to find over representation of associated GO terms. We also recently ventured into whole genome sequencing where we generated both single end and paired end whole genome sequence data for two of the bees, *Bombus impatiens* and *Megachile rotundata* using Illumina sequencing technology. The reads generated were assembled using a de

Bruijn graph based assembly algorithm into scaffolds having a N50 of 1.12 Mb and 31 Kb respectively.

To My Adviser, Dr. Matthew Hudson

ACKNOWLEDGEMENTS

This project is a collaborative effort and would certainly not have been possible without the help of many people. First and foremost, I would like to thank my advisor, Dr. Matthew Hudson, for having given me this wonderful opportunity to work on such an interesting project that shaped my interest in evolutionary genomics. He has played the role of a wonderful advisor and a friend. Thanks to Kranthi Varala, my senior lab mate, who helped me make sense of a lot of things initially when I just started working on the project. Thanks to my collaborators in Dr. Gene Robinson's lab, Hollis Woodard and Brielle Fischman who taught me some of the exciting concepts in evolution and got me a lot interested in bee biology. Thanks to Brandon Smith, a former programmer in our lab, who laid the foundation of the bee BLAST website which I optimized further. Many thanks to my committee members, Dr. Jian Ma and Dr. Nathan Price for their readiness to serve on my committee and thoughtful discussions. Last but not the least, I want to thank my family for always being there for me, and giving me lot of mental courage and strength, so the physical distance of them being miles away did not matter.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: METHODOLOGY.....	14
CHAPTER 3: RESULTS.....	23
CHAPTER 4: DISCUSSION.....	29
CHAPTER 5: CONCLUSIONS.....	36
REFERENCES.....	37
APPENDIX: FIGURES AND TABLES.....	45

CHAPTER 1

INTRODUCTION

Sociogenomics: An Integrative Discipline

Life has evolved starting from single cells, to multicellular organisms, to multicellular organisms forming societies to live in. A lot of research has gone into elucidating the molecular basis of cellular function and development, and the same needs to be done today for social life (sociogenomics) (Robinson et al. 2005). Sociogenomics needs to be understood in terms of how societies evolved, what are the genes influencing them, how are they regulated, how do organisms differ in their social behaviour patterns and so on. We need to understand how behavior influences different aspects of genome structure, genome activity and organismal function (Robinson et al. 1997; Robinson et al. 2009). The conceptual foundation of sociobiology is in Darwinian theory in which emphasis has been laid to group life that is based on mutualism, kin selection and altruism.

The nascent field of sociogenomics is predicated on two of the most significant ideas in biology to emerge from the latter half of the twentieth century (Robinson 2002). First, many aspects of social life, including social behaviour have a biological basis and are thus influenced by genes and the forces of evolution to a large extent (E. O. Wilson 1975). Second, the molecular functions of many genes are highly conserved across species for complex traits (Carroll et al. 2001). One of the challenges in behavioural sciences is to understand at the molecular level, how genes influence social behaviour patterns. There are lots of reasons why we need to use diverse non-model systems for this study. First, traditionally, other forms of behaviour at the molecular level have been studied in model organisms amenable to genetic analysis like *Drosophila melanogaster* in which learning (Dubnau J. and Tully, T. 2001) and circadian rhythms (Panda, S & Kay, S .A, 2002) have been explored. While traditional model systems like the fruit fly have been used to study mating behaviour (which involves structured interactions with conspecifics), mating does not distinguish social animals from most others (Greenspan, R .J. & Ferveur, J. F, 2000). Second, while powerful studies of social behaviour can be performed in the lab (Pfaff, 1999), there is a keen interest in elucidating the molecular

machinery of social behaviour in natural contexts (Jarvis E.D. et al. 1997). We are in need of a broad integrative framework that uses mechanistic and evolutionary perspectives to understand how social behaviour evolved. The mechanistic analysis of social behaviour encompasses the traditional fields of behavioural genetics, neuroscience, cell biology and molecular biology. On the other hand, evolutionary analysis of social behaviour includes fields like phylogeny, population biology, behavioural ecology and sociobiology. While molecular biology helps to target candidate genes of interest, behavioural ecologists study the adaptations occurring on selected genes on interest in a phylogenetic context. Genomics helps to integrate the two perspectives (Robinson et al.2005). Third, social behaviour itself has different levels to consider. Species can be solitary, in which they only interact with conspecifics during mating, or they can live in highly structured colonies in which they interact with conspecifics all the time, in which case they are eusocial. Such a diverse system having multiple levels of sociality gives experimental access to a process involved in all forms of social behaviour and gene regulation. Studying diverse animal societies also allows us to understand if different evolutionary events can have the same end (convergence), and the roles of conservation of genes across species (Robinson G.E. & Ben Shahar, 2002). Also, it is important to use model systems that can be studied under naturalistic conditions, as studies done in natural environments/ecologically relevant conditions will make it easier to interpret the data. This line of study is called evolutionary and ecological functional genomics (Feder ME & Mitchell-Olds, 2003) where there are no other forces that obscure the results as commonly found in laboratory manipulations where lot of extrinsic factors can affect the results (Vignal C Mathevon & Mottin, 2004).

The goal of sociogenomics is to gain a comprehensive understanding of behaviour at the molecular level. This can help us understand how complex behaviour evolved from a simpler ancestral behaviour. Starting with a broad array of models showing diversity in social behaviour, given that many genes and pathways are conserved across species, enables us to compare across diverse taxa. This in turn can help us probe deeper into the evolutionary mechanisms.

Some of the basic activities that need to be performed for survival of a given species include finding food (foraging), mating (which involves identification of mates receptive to reproduction), construction of a nest or shelter to rear young (parental care) and defend the nest from intruders. Such activities need to be performed by both solitary and social animals (Alcock, J, 1998). Social animals perform these activities cooperatively where there is lot of coordination accomplished by structured interactions with other members of the same species. This involves intense communication among individuals, dominance hierarchies and division of labour (Robinson et al. 2005). Genes involved in solitary behaviour are also involved in social behaviour indicating that genes involved in simpler behaviours can be used to identify candidate genes involved in a more complex behaviour. Analysis of certain behaviours shown by solitary animals (e.g model genetic organisms) can be built upon to enhance our understanding of social life.

Behavioural Plasticity: From Highly Eusocial to Solitary

Ants, bees, wasps and termites are the best-known eusocial species (Wilson, E.O.1971; Duffy, J.E., 1996; Sherman, P. et al. 1991; Choe J. C & Crespi B, 1997). In some tropical habitats, ants and termites are dominant terrestrial life forms (Holldobler & Wilson, 1990). One of the significant ideas sociogenomics is built upon is that of conservation of genes across taxa. The insect order Hymenoptera is distinguished by species showing a range of sociality, from solitary to highly eusocial allowing us to exploit three fundamentals of sociogenomics; 1) diversity in social behaviour to understand conservation of genes; 2) the role of selective pressures on these genes that play adaptive roles that can eventually be tied to behavioural differences, and 3) if multiple independent evolutionary events converge. Insights from the integration of evolutionary biology with developmental biology (hybrid evo-devo studies, Toth A.L. & Robinson, G. E. 2007) elucidate the concept of a shared genetic toolkit that is conserved at the molecular level across diverse taxa. For example, the homeobox genes (Hox genes), body form (Gellon, G. et al. 1998), and eye development (Pichaud, F. et al. 2002). The conserved genetic toolkit for development is thought to consist of a set of genes having specialized functions, like transcription factors (Caroll, 2001).

Similarly there are several cases of genes involved in conserved pathways and networks across diverse taxa in Hymenoptera. This makes Hymenopterans excellent model systems to study the evolution of social behaviour.

Eusocial species are those that show extreme form of social organization in which individuals specialize in certain tasks. This behavioural specialization is often linked to differences in age, anatomy and morphology (Robinson, G.E. et al. 2005). Eusociality is rare, but highly successful. In highly eusocial colonies of Honeybees, queens monopolize the reproductive tasks in the colony, while workers are involved in foraging and brood care, or in other words, tasks related to colony growth. Thus, there are three defining characteristics of highly eusocial colonies. There is 1) reproductive division of labour, 2) cooperative brood care and 3) an overlap of generations in which queens and young workers stay in the same colony. In advanced eusocial species, the fate of an individual, queen or worker is determined long before adulthood, and depends on the nutrition fed to the larvae. This sets the stage for colony level selection creating systems of division of labour. In primitively eusocial bees (halictid bees), there are loosely morphologically defined queen and worker classes, so the caste differentiation is more of a behavioural phenomenon. In the solitary bees, every female is fertile and manages all the tasks.

Given all these factors that go into choosing good model systems, in this work, we have focused on using bees (Hymenopterans) to probe into social behaviour evolution. We perform our studies in a naturalistic context, and use good phylogenetic background (species tree) for our downstream evolutionary analyses. We use Genomics to integrate the mechanistic and evolutionary perspectives. The broad goal of our project is to identify the candidate genes that may be involved in the evolution of social behaviour using Genomics, Phylogeny and Behavioural Plasticity in extant bees (see Methods).

Transcriptomics and Social Behaviour

Two approaches can be taken to answer our question, that is, identify genes and pathways involved in social behaviour in Hymenoptera. First, given that the honeybee has recently had its genome sequenced (Robinson, G.E. et al. 2006), genomic resources need to be developed for other species that show advanced forms of eusociality, such as the ones that exist for fire ants (Krieger, M.J. & Ross, K.G. 2002) and leaf cutting ants (Holldobler, B & Wilson, E.O.1990). Second, genomic resources can be developed for selected species of bees that differ in levels of sociality. For example, in Hymenoptera, within the Apini tribe alone, there are species that are solitary, primitively eusocial, and highly eusocial; that is, the euglossines, bumble bees, and honey bees and stingless bees respectively (Lockhart, P.J. & Cameron, S.A. 2001.) While it is realistic to obtain whole genome sequences for many of these species, genomics can provide a wealth of sequence data at an economical cost too, accomplished through Expressed Sequence Tags (ESTs), microarrays and BAC libraries. A lot of progress has been made in using these techniques effectively to discover genes and genomic regions of interest to social behaviour (White et al. 2002; Band et al. 2000; Summers et al. 2001). Sequence information from EST collections and other sources eliminate the need to tediously clone genes on a gene-by-gene basis before experiments with candidate genes can even begin. Microarrays, too, allow for open-ended gene discovery (Fitzpatrick, M. et al. 2005). Cloning each gene is obviously highly inefficient, especially for our purposes, since social behaviour is known to be regulated by a vast repertoire of genes. The two traditional forward genetic models used to discover candidate genes, that is, Seymour Benzer's single gene mutations approach, and Jerry Hirsch's approach of identifying behavioral variants (Tully, T.,1996) are difficult to adapt for this problem because bees are hard to manipulate genetically, though limited success has been obtained in making transgenic bees (Robinson, G.E. et al. 2000). Instead, using transcriptomics one can measure the abundance of genes expressed in brains of social/solitary insects and sequence them. ESTs are single DNA sequencing reads obtained from complementary DNA (cDNA) clone libraries constructed from a known tissue source. Sequencing a large number of these clones

from such a library allows one to get a decent sample of the set of expressed genes, or transcripts, in that particular tissue and experimental state. This provides a snapshot of the tissue's active genes under those defined conditions. ESTs provide a short cut to the transcribed portions of the genome, and this information can be used as key evidence for genome annotations, gene discovery and comparative genomic analysis by bioinformaticists.

EST Sequencing

There are different approaches that can be taken to sequence the ESTs. Next generation sequencing (NGS) technologies are producing tremendous amount of data in a relatively short time as opposed to the traditional sequencing methods like Maxam Gilbert's chemical modification of DNA and cleavage, or the Sanger's di-deoxy chain termination method. The high demand for low-cost sequencing techniques has driven the development of new NGS technologies. Today, the main commercially available technologies are from Roche/454, Illumina/Solexa, Life/APG and Helicos Biosciences. There are a core set of steps that are common to all these technologies, namely template preparation, sequencing and imaging and data analysis. Roche's 454 pyrosequencing method amplifies DNA inside water droplets in an oil emulsion, hence also called emulsion PCR (<http://www.454.com>). Solexa/Illumina uses a cyclic reversible terminator (CRT) system, which is based on reversible dye-terminators. DNA molecules are attached to primers and amplified using bridge amplification to produce clonal copies of a single DNA molecule. The single DNA template is then clonally amplified and sequenced using luciferase that generates light when a nucleotide is added to the nascent growing DNA molecule. The key lies in adding one nucleotide, growing the DNA chain, terminating it and imaging which nucleotide is added using a four dye color system, one for each base, and then adding another nucleotide. A camera takes images of fluorescently labeled nucleotides (<http://illumina.com>). SOLID is similar to Solexa, but uses sequencing by ligation, and makes use of oligonucleotides (www.appliedbiosystems.com). The output of each technology is different, though all of

them provide quality scores for each base sequenced, giving an estimate on what is the probability that the base that is read off by the machine is erroneous.

In this work, we have used ESTs to collect information on the genes that are expressed in our Hymenoptera phylogeny, and sequenced them using Roche's GS-FLX technology. This gave us about 240 bp read length on an average, for all the twelve species. Ability to generate longer read lengths is the main advantage of 454 sequencing.

Transcriptome Assembly

In any transcriptomics or genomics project, the sequenced reads have to first be assembled together so as to get a putative reconstruction of the target. This process is called assembly of the reads. There are lots of software applications available that aid in the assembly process. A transcriptome assembly is very different from a whole genome assembly; In whole genome assembly, a more or less uniform distribution of reads across the genome, or fluctuation arising due to repeat contents is obtained; whereas in a transcriptome assembly, biases in sampling due to the presence of highly expressed genes are largely expected. In a genome assembly, the extension of reads into contigs is ambiguated by the presence of repeats, whereas in transcriptome assembly, the presence of variants/isoforms and gene families confounds the assembly process (Birol et al. 2001). Analysis of the isoforms can help elucidate alternate splicing events. The newer assembly algorithm Abyss, a commonly used transcriptome assembly tool, uses a distributed de Bruijn graph data structure that splits a sequence into K mers and assembles the unique K mers. The distribution of the graph over several nodes of the cluster relaxes the memory/computational requirements for the assembly. This is important because in a de Bruijn graph, the memory requirement scales linearly with the sequence. Abyss falls in the category of Eulerian graph assemblers. Recently, the de Bruijn graph based assembler, Velvet (Zerbino et al. 2007), has been extended into Oases tool for transcriptome assembly, where the uneven sampling bias and alternate transcript information is used to refine the output of Velvet (Zerbino et al., Unpublished).

On the other hand, there are greedy assemblers that just rely on pair-wise alignments, and any two reads having the maximum pair wise overlap similarities will be joined together. The Phrap assembly program is one such example. It was written by Phil Green in 1996 at the University of Washington to provide rapid comparison, alignment, and assembly of large sets of DNA sequences. The PHRAP assembler uses a banded version of the Smith-Waterman-Gotoh algorithm (Smith, T. F. & Waterman, M.S. 1981; Gotoh, O. 1982) to compute pair wise comparisons of the input sequences (de la Bastide, M. 2007). PHRAP is similar to BLAST, in that, it first searches for a “seed” match, and once it finds a perfect word match, it tries to extend the alignment. Since we used 454 for the sequencing, we used Phrap for the assembly of the ESTs into non redundant contigs and singletons. Overlap based greedy assemblers perform well with longer reads.

Database Design

In order to better organize the species specific assembled ESTs, we designed a central MySQL database holding the transcriptome assembly information. We then developed a front end BLAST website which users can use to query the database, and blast against honeybee genes (See Methods, bee12 BLAST server).

Ortholog Detection

Diverse bioinformatics tools have been developed to analyze sequence data from evolutionary and functional perspectives (Ouzounis, C. A. et al. 2003). Evolutionary projects that generate sequence data from closely related species require the concept of phylogenies and orthology, which are crucial to inferring gene content, conserved gene order, gene expression, regulatory networks, metabolic pathways and functional genome annotations, to name a few (Kuzinar et al. 2008; Eisen, J. 1998; Jeffroy et al. 2006; Delsuc, F. 2005; Tatusov et al. 1997; Tatusov et al. 2003; Goodstadt & Ponting, 2006; Bandopadhyaya et al. 2006; Mazurie et al. 2005; Grigoryev et al. 2004; Mao et al. 2006; Hulsen et al. 2006).Walter Fitch in 1970s, divided homology into orthology and paralogy based on speciation and duplication events respectively. Orthology is strictly

an evolutionary concept and can be defined as homologous genes that relate through speciation from a single ancestral gene present in their last common ancestor, whereas paralogs are homologs that arose through gene duplication (Fitch, W.M. 1970; Fitch, W.M. 2000). There are many computational tools developed so far that detect orthologs from sequence data belonging to closely related species. They all have their own set of advantages and disadvantages. The major algorithms developed can be classified into tree based approaches, graph based approaches and those that utilize both / hybrid methods (Kuzinar et al. 2008). A brief review of some of the key algorithms from each of the categories along with their pros and cons is presented here.

The Tree-based methods are used to infer orthologs in entire genes in 2 or more species (Kuzinar et al. 2008). Some popular algorithms include Correlation Coefficient-based Clustering (COCO-CL) and OrthoStrapper. The COCO-CL (Jothi et al. 2006) uses a Pearson's correlation matrix and infers duplication/speciation events. But this method does not implement a tree reconciliation algorithm, and does not require a species tree as input. The Orthostrapper (Storm, C.E.V. & Sonnhammer, E.L.L., 2002) uses a hierarchical grouping of orthologous and paralogous sequences, and requires a set of gene trees from which it calculates bootstrapped values/confidence. However, having a gene tree for every gene can be cumbersome, and moreover, the program is not freely available for download.

The Graph based methods on the other hand use precomputed homologs to infer orthologs and paralogs (Kuzinar et al. 2008). Examples include algorithms based on Nearest Neighbour and Clusters of Orthologous Groups (COGs) of proteins. The Nearest Neighbour methods employ best pair wise sequence alignments of two or more genes as an operational definition of orthology, and can be used as first pass approximations to finding putative orthologs, These methods include best hit (BeT), reciprocal best hit (RBH), bi-directional best hit (BBH), symmetrical best hit (SymBeT) and reciprocal smallest distance (RSD) (Kuzinar et al. 2008; Hirsh & Fraser, 2001; Overbeek et al. 1999; Wall et al. 2003; Lee, Y. et al. 2002; Remm et al. 2001). These methods also identify many-many, one-many orthologous relationships, based on how it

is implemented. Each of them may result in overlapping sets of orthologs. The reciprocal smallest distance does local and global sequence alignments, and uses maximum likelihood estimates of evolutionary distances to predict orthologous proteins (Wall, D.P. et al. 2003). However this method does not permit more than two genomes to be compared at once, and does not allow an outgroup species.

The Hybrid methods are a fusion of both tree based and graph based methods. One can guide the algorithm to refine the results based on species tree input (Hubbard, T.J.P. et al. 2007; Wheeler, D.L, 2007; Cannon, S.B. & Young, N.D., 2003; Dehal, P.S. & Boore, J.L.,2006; Merkeev, I.V.et al. 2006; Li, H. et al. 2006) and do not use information like conserved gene neighbourhood (CGN).A popular hybrid algorithm is OrthoParaMap package (Cannon, S.B. & Young, N.D.,2003) in which a BLAST of two genes along with the gene phylogenies are used to infer events in gene families in two species. But this is limited to two species comparisons. There are other databases like HomoloGene (Wheeler, D.L.2007) and TreeFam (Li, H. et al. 2006) that use clustering methods and store information on precomputed homologs. Here the results may be harder to interpret because the details of the clustering procedure are not clearly described in the literature.

Taking into account the pros and cons of the aforementioned ortholog detection tools, the purpose of our project, availability of ESTs, the phylogenetic background and the computational complexity/scalability issues, we developed our own method, an extension of the reciprocal BLAST method to assign putative orthology (See Methods).

Inferring Selection from Orthologs

Phylogenetic methods that make use of robust statistical models have been widely used of late, to detect natural selection (Yang, Z. 2005). Rapidly evolving regions in genes/genomes occur as a result of positive/Darwanian selection, or evolutionary conservation of the genes occurring as a result of purifying selection. Both these scenarios can be inferred from sequence data (Thomas et al. 2003; Nielsen et al. 2005;

Sawyer et al. 2005). For the purposes of our project, we are mainly interested in statistically inferring signatures of selection/rapidly evolving genes across our orthologs. Analysis of orthologs can help distinguish between synonymous(nucleotide substitutions that do not change the encoded amino acid) and non synonymous substitutions (those that change the underlying amino acid encoded).Since natural selection acts on the protein level, synonymous and non synonymous mutations are under different selective pressures and are fixed at different rates (Yang, Z.2007). Hence we compare the rates of these substitutions to reveal the direction and strength of natural selection acting on the protein (Kimura, M. 1977; Miyata, T. & Yasunaga, T. 1980).

In this work, we use a program called PAML (Yang,Z. 2007) (Phylogenetic Analysis by Maximum Likelihood) that fixes different selective pressures across our phylogeny, and estimates parameters using the maximum likelihood function under a phylogenetic framework. The strength of PAML lies in its collection of sophisticated codon substitution models that use a Markov Model of codon evolution, and make reasonable assumptions about biological processes. The equilibrium frequency of each codon and the transition / transversion rate ratio are taken into account in the Markov Model while computing the log likelihood of the data. This minimizes bias in the dataset resulting from unequal codon usage frequencies, a common problem in most phylogenetic analyses. We automated the whole process computationally, to run PAML across our entire dataset, and use branch models (See Methods) to detect adaptive molecular evolution from the same.

Whole Genome Assembly

We recently embarked on a project to generate whole genome data for the twelve bee species. The availability of the genomes will no doubt advance our knowledge on the genome architecture, provide deeper insights on molecular evolution and enhance our knowledge on social behaviour to test different hypotheses using comparative genomics analyses. The pros and cons of different sequencing techniques have been discussed in the EST Sequencing section of the Introduction chapter. We used Illumina sequencing to sequence the whole genomes (See Methods). A major challenge

following any genome sequencing is the assembly of the reads. We briefly reviewed this in the “Transcriptome Assembly” section. There are two major approaches to genome assembly; 1) de novo assembly, which reassembles the reads purely based on overlaps; 2) mapping, which assembles based on a template reference genome of a closely related species. Since most of the non-model organisms do not have an already available genome sequence of a closely related reference species, it becomes mandatory to choose de novo assembly for putative reconstruction of the target genome. The Classical hierarchical assembly method using Bacterial Artificial Chromosomes / BAC method, which was employed in the human genome project (Lander et al. 2001) consists of building BAC libraries and tracing the contigs using a minimal tiling path approach. This makes assembly within each BAC easier as there are no polymorphisms, but the high cost associated with the BAC library construction necessitated the need for rapid, cost effective methods. The traditional Sanger sequencing method, the low throughput sequencing method, is today being replaced by Ultra High Throughput methods (UHT); Next Generation Sequencing technologies that make use of different chemistries for sequencing and imaging. These UHT methods make use of whole genome shotgun sequencing, where the genome is randomly sheared into a number of fragments, and the ends of each fragment are sequenced. When the distance between two reads and their orientations are known, such “linked reads” help to disambiguate repeats (Edwards, A. & Caskey, T. 1991). Our data makes use of such linked reads or paired-end/mate pair reads (See Methods).

The current genome assemblers can be grouped into major categories based on the approach taken. The greedy assemblers (Phrap (37), Cap3 (Huang, X. & Madan, A. 1999), TIGR Assembler (Sutton et al. 1995) greedily join together the input reads based on local sequence similarity into contigs. But since only the local information is used at each step, this can lead to mis-assemblies caused by repeats, since repeats overlap perfectly. The overlap-layout-consensus based assemblers (Celera (Myers, E. W. et al. 2000), Phusion (Mulikin et al. 2003), MIRA3 (Chevreux et al. 1999), Edena (Hernandez et al. 2008), Arachne (Batzoglou, S. et al. 2002) make use of graph theory. Here any two reads are stored as nodes in the graph, and an edge connects the two

nodes if there is overlap between the corresponding reads. The overlap stage is computationally expensive, since the overlaps across all the reads are calculated, and the graph structure is computed. Following that, in the layout stage, the graph is simplified by removing redundant nodes, and then contigs are created by traversing the Hamiltonian path in the graph. The Eulerian path approaches (Euler-SR (Chaisson, M.J. & Pevzner, P.A., 2008), Velvet (Zerbino et al. 2008), VCAKE (Jeck, W.R. et al. 2007)) make use of graph theoretical models that break up reads into Kmers, and store the Kmers in the edges. Each k-mer is represented in the graph as an edge connecting two nodes, corresponding to its k-1 bp prefix and suffix respectively. The solution to the assembly problem is now traversing all the edges of the graph, an Eulerian path. The repeats are identified very easily using this approach. De Bruijn Graph based assemblers, such as SOAPdenovo (Li, R et al. 2010) and Velvet (Zerbino et al. 2008), were first conceived by Pevzner (Pevzner, P. et al. 2001). They make use of both Hamiltonian and Eulerian paths in the graphs. The Align-layout-consensus based assemblers (Projector2 (Sacha et al. 2005), Mosaik (Smith, D.R. et al. 2008), ELAND (Cox, unpublished software), Mummer (Salzberg et al. 2002)) are similar to the overlap layout consensus assemblers, but the overlap step is replaced by the align step, which means these required a template reference genome to align the reads to. This makes the graph lot simpler. There is lot of demand for these graph based short read assemblers over the conventional assemblers like Atlas (Havlak et al. 2004), which assemble reads from a BAC-by-BAC strategy.

In this work, we have explored algorithms that use de Bruijn graph approaches to putatively reconstruct the draft of the whole genome assemblies for two of our bees.

CHAPTER 2

METHODOLOGY

Bee Collection and RNA Extractions from Brains

Fifty adult females of twelve bees belonging to the insect order Hymenoptera, (*Apis florea*, *Bombus impatiens*, *Bombus terrestris*, *Euglossa cordata*, *Eulaema nigrita*, *Exoneura robusta*, *Megalopta genalis*, *Melipona quadrifasciata*, *Bombus insularis*, *Centris flavifrons*, *Megachile rotundata*, *Frieseomellita varia*) were sampled from Utah and Illinois bee research lab (Robinson lab, Cameron lab, UIUC). The brains were flash frozen in liquid nitrogen to preserve the mRNA. The brains were then dissected, and RNA was extracted and amplified by (Robinson lab, UIUC).

EST Sequencing

454 Genome Sequencer (GS) FLX sequencing technology was used to sequence the cDNAs in a straightforward manner to obtain the Expressed Sequence Tags (EST sequences). The long reads of approximately 300-400 bp produced by the technology enabled coverage of more exons and splice junctions, allowing more positive linkage of variants and longer extension into UTR regions

(<http://www.454.com/applications/transcriptome-sequencing.asp>)

EST Assembly

The ESTs obtained from the sequencing step above were assembled de novo using Phrap algorithm (Green, P.1996) into non-redundant contigs and singletons.

Standardized Species Names and Database Structure

The assembled EST sequences (non-redundant contigs and singletons) were assigned appropriate species-specific FASTA format headers, with standardized species abbreviations that were agreed upon in our group (Table 2). Custom PHP scripts were

written to load the assembled ESTs into a MySQL database (Server version: 5.0.77) named BlastData. In addition to storing sequence information for each species, BlastData records information on markers corresponding to each linkage group of the *Apis mellifera* genome. A total of 643 markers for the 16 linkage groups (Solignac, M. 2007) were loaded in the Markers and Linkage groups tables respectively. The honeybee homologs were computed by blasting the *Apis mellifera* gene models against each of the species specific databases of bees. This information was used to draw the location of the *Apis mellifera* gene for its corresponding bee homolog on the respective chromosome of *Apis mellifera*. The scaffolds table stores about 9870 scaffolds (*A.mellifera*, genome assembly 4.0).A Btree index on selected attributes was created to speed up the BLAST searches. In addition, the database contents are password protected.

Data Statistics

The statistics of the non-redundant contigs and singletons compiled in BlastData Projects are as shown in Figure 2.1. The total number of assembled sequences in the Sequence table is 1,176,683. The BeeHomologs table has precomputed honey bee homologs for 1112178 bee genes.

Honey Bee Homolog Blast Website Design

Website Homepage

In order for users to access the contents of BlastData, we designed the Honey Bee Homolog BLAST website that helps users BLAST against the database/download the datasets. The frontend is designed using HTML, CSS, Javascript and AJAX. PHP connects the front end and the backend MySQL database. Each Project holding the species-specific non-redundant contigs and singletons information is formatted into a BLAST database. Users can run blastn, or tblastn jobs against multiple BLAST databases after logging in to the database.

Website Features

Set optional BLAST parameters

Users can set optional parameters for the BLAST job such as setting an e-value cut off, filtering low complexity regions, customizing how many alignments to display after the BLAST run, selecting from a range of matrices like BLOSUM 45/62/80 and PAM 1/120/250 for the BLAST job.

Graphical Display

Once the BLAST job is completed, a graphical display of the alignment is presented to the user, similar to the NCBI BLAST website graphical display.

Sequence Retrieval

One of the unique features of the website allows users to select multiple high scoring contigs/singletons from the graphical display of the alignments (by holding the z key from the keyboard), and use sliders to define a region of the alignment. The desired Sequence Retrieval Method can then be used, and ClustalW multiple sequence alignment tool can be run. If only the region of the multiple sequence alignment as defined by the sliders is required, the Use Sliders (Compact) option can be used. The "full" option will allow the user to view the complete alignment across the entire length of all the sequences selected. However, this can be slow depending on the length of the sequences, and whether they are well alignable. On the other hand, if some bases/amino acids extension on either sides of the defined region are desired, the extensions (100-5000) can be chosen, and if the sequence extension is within the total length of the contig, the extension can be displayed. In order to take a look at the ClustalW multiple sequence alignment and retrieve the sequences, users must 'allow pop-ups for bee12.cropsci.uiuc.edu' in their browser preferences, and the sequences and the alignment result will open in two different tabs/windows based on how the browser preferences are adjusted by the users. Note that this feature works well only

on Mozilla Firefox/Internet Explorer and Google Chrome web browsers across platforms.

Miscellaneous

Users can also zoom in/out on honey bee chromosomes. This can be done by holding down shift to zoom in and out and requires JavaScript to be enabled. We also make use of our pre-computed homologs to draw the picture or show the location of the *Apis mellifera* gene on its chromosome.

Ortholog Assignment Pipeline

In order to pick out candidate genes that may be involved in social behaviour, a sequence alignment pipeline based on the method of reciprocal BLAST was designed and implemented in PHP (hyper text pre-processor, server side scripting language). *Apis mellifera* gene models from beebase were used as the reference (<http://genomes.arc.georgetown.edu/drupal/beebase>). The newest release of the honeybee gene models in beebase that is the pre-release2 version has approximately 11,062 gene models. The pipeline starts by picking out a honeybee gene model at a time, and blasts the gene model against each of the 12 species specific nonredundant BLAST databases using blastn, $E < 1e-6$. All the hits that are within 10% identity of the top hit are then blasted back to the honey bee gene models database using the same E value cut offs to make sure we get the same gene model A as the top hit. If yes, then it is considered as a putative ortholog according to our operational definition of orthology. We keep track of the coordinate of the gene hits wrt *Apis mellifera*. Each of the gene hits is reciprocally tested to check if it satisfies the condition of orthology. The best reciprocal gene hits are then concatenated together in the order in which the gene fragments occur on the honey bee gene model, a step that involves trimming of overlapping genes, (overlap > 25 bases) and removing the redundant hits. The aforementioned steps are repeated for every database the gene model is blasted against. Care is taken to make sure every gene that goes into the alignment is a

reciprocal best hit. The concatenated orthologous hits are called Gapped Ortholog-reference-based Transcript Assembly (GOTA). A schematic of the pipeline is as shown in Figure 2.2.

Multiple Sequence Alignment of the Putative Orthologous Genes

The Multiple Sequence Alignment by Fast Fourier Transform (MAFFT) (Kato et al. 2002) algorithm was used to align the orthologs obtained from the computational pipeline to the reference *Apis mellifera* gene. We used the E-INS-i alignment strategy of MAFFT, which combines both weighted sum of pairs and consistency scores to generate a multiple sequence alignment. About 1000 maximum iterations for the MAFFT EINS-i run were used to iteratively refine the alignment with each run. This was followed by rigorous manual editing of alignments that contained putative orthologs from all the bee species using Geneious software that enables easy editing of alignments. Here the ambiguous codons were deleted (Robinson lab). An example multiple sequence alignment as viewed in Geneious is shown in Figure 2.3.

Alignment Ranking System

The set of alignments that had putative orthologs from all the species ($n \sim 3647$) were used to generate gapless alignments for inferring the species phylogeny. Among the 3647 genes, alternate transcript alignments were also present. Since alternate transcripts do not help to add any new information to the alignment, a scoring system was developed to rank all the alignments based on depth and coverage. The ranking system works by weighting every site in the alignment based on the coverage, and penalizing the gaps. The total score of the alignment is the sum of the weights across all the sites. The scores range from 1 to N , where N is the maximum number of species covered. Hence an alignment that scores N is very well covered very with no gaps, whereas an alignment that scores 1 is has no hits but the reference *Apis mellifera* sequence, or is widely composed of gaps. The scoring system was very useful in terms of identifying good alignments and the best scoring alignments for the alternative

transcripts. We also used the ranking to test if there is any bias between the alignment score and the rapidly evolving gene lists.

Species Phylogeny

Using the alignment ranking system, alignment scores were obtained for each of the 3647 alignments, including only the best scoring alternative transcript alignments. Two sets of gapless alignments were then generated, the first set only included those sites that are present in all the species in the alignment, while the second set operated on a relaxed criteria, where the sites that had information from all but one species were included. The individual gapless alignments based on both the relaxed and non-relaxed criteria were then concatenated separately, preserving the reading frames, giving two gigantic concatenated alignments, which were subsequently used for codon level analysis. The alignments were analyzed using MrBayes (Huelsenbeck et al. 2001) to infer the species phylogeny (Dr. Sydney Cameron, unpublished). The species tree obtained as a result of these analyses had a high consensus support on each node (Figure 2.4), and was used as the background for tests of selection. Three species out of the twelve species gave ambiguous results in the species phylogeny and hence were removed from further analyses. This finally led to a comparative analyses across nine bee species.

Inferring Selection from the Alignments

The codeml program in the PAML package was used to infer selection signatures from the alignments. PAML implements a maximum likelihood method to estimate parameter values in a phylogenetic framework using an appropriate species phylogeny. Under the codon substitution model of Goldman and Yang, the ω ratio is a measure of natural selection acting on the protein. It is defined as the ratio of the rate of non-synonymous substitutions to the synonymous substitution rate. Simplistically, values of omega, $\omega < 1$, $\omega = 1$, and $\omega > 1$ indicate negative purifying selection, neutral evolution, and positive selection. However, ω averaged over all sites and all lineages is almost never > 1 , since positive selection is unlikely to affect all sites over prolonged time. Thus interest has

been focused on detecting positive selection that affects only some lineages or some sites. (Yang, Z, 2007). For this purpose, Branch Models were used in this study to estimate lineage specific differences from the alignments. The branch models allow the ω ratio to vary among branches in the phylogeny and are useful for detecting positive selection acting on particular lineages (Yang 1998; Yang and Nielsen 1998). They are specified using the variable model in the PAML control file model = 2, which allows several ω ratios across the phylogeny, was preferred over the free-ratios model, which is very parameter rich. Tree files and control files were prepared for each hypothesis and branch node labels were used to specify different rates of evolution in the tree file. The whole process of running branch models on the alignments that have orthologs in all the bees was automated to run over a cluster using a batch submission script. To put it very briefly, the PAML script worked by creating several hypothesis-specific directories inside a main gene model directory. Inside each of the hypothesis-specific directories, the control file, tree file, and the multiple sequence alignment file in the Phylip format were placed and the program codeml was run. The number of jobs submitted to the cluster were tracked, and checked if it is lesser than a threshold count of jobs estimated based on the number of nodes of the cluster. If yes, then the next branch model job was submitted to the cluster. Otherwise, jobs were only submitted as and when they finished.

Hypotheses and Statistical Design

Each Branch model is a specific hypothesis, which is tested through PAML. Three different branch model hypothesis were tested, one null model, and two alternative models that look for lineage specific rapidly evolving genes, and a shared set of genes; the first hypothesis tested if the genes in the eusocial lineages are evolving more rapidly than the non-eusocial lineages (shared set of genes across the entire eusocial clade). The second one tested if the genes in the primitively eusocial lineages evolved more rapidly than the other lineages (lineage specific), while the third one tested if the genes in the highly eusocial lineages are evolving more rapidly than the other lineages (lineage specific). For each hypothesis (model) a log likelihood value was obtained, and

the likelihood ratio test (LRT) was used to compare how well the alternative model fits the data compared to the null model. Subsequently, a decision was made to reject or fail to reject null hypothesis. The test statistic in the LRT is twice the difference in the log likelihoods of the null and the alternative models. The probability distribution of the test statistic was approximated using a chi-square distribution with $(df1-df2)$ degrees of freedom, where $df1$ and $df2$ are the degrees of freedom for model1 and model2 respectively, which are the null and the alternative models respectively. The alternative model in this case being more parameter rich has higher degrees of freedom compared to the null model. It may seem that having more parameters in a model will make the model explain the data better, but this may not be true at all times. We ran our tests of selection at 5% level of significance. The overlap of the results across several hypotheses was also computed to get the rapidly evolving lineage specific gene lists.

Whole Genome Assemblies

We generated whole genome data for two of our bees, one primitively eusocial, *Bombus impatiens*, and another solitary, *Megachile rotundata* using illumina, single end and paired end sequencing technology. About 49 GB of whole genome data was generated for each of the bees. Each run was paired-end (2×124) cycles. The error rate of the Phix control was very low, about (1-1.5%) for each read, each run. The Quake program (Kelley et al, Manuscript in preparation) was used to correct the reads. Quake is used to correct errors in experiments with deep coverage ($>20X$), like those generated using Illumina. It uses a mixture model of genuine and erroneous k-mer distributions, and uses read quality values to learn the nucleotide-to-nucleotide error rates to determine the most likely errors. Following error correction, custom scripts were written to order the mate pairs, and trim the adaptors. The whole genome assembly was then done using the SOAPdenovo software. SOAPdenovo is a de Bruijn graph based algorithm that runs in four distinct steps, pregraph, contig, mapping and scaffolding. The graph construction is computationally most expensive, where each node is a k-mer, and two nodes that overlap by $k-1$ bp are connected by an edge. Once the graph structure is computed, the errors caused by sequencing that appear as bubbles (error in the middle

of the read) or tips (errors at the end of the read), need to be removed. The tips are corrected by trimming the ends of the reads, and the bubble correction is similar to Velvet's Tour bus method based on Dijkstra's algorithm. The repeat sequences that are shorter than the read lengths are resolved using equal N incoming and outgoing edges. The next step is to traverse the edges of the graph to construct the contigs (contiging). The reads are then mapped back to the contig sequences (mapping), and paired-end relationship between the reads is mapped to linkage between contigs, which is then used to construct scaffolds (scaffolding). Once the scaffolds are constructed, a gap closure algorithm was run to close the gaps in the assembly. The memory use for the gap closure is mainly related to the read number and the number of unique k-mers in the reads. The time taken for gap closure depends on the read number, gap number and gap size. Gap closure works by iteratively mapping the reads back to the contigs, and checking for pairs where one read maps to the end of the contig, while the other is in a gap, and then realigning the reads back to the contigs, to extend the contigs locally. SOAPdenovo requires a configuration file describing the insert sizes of the libraries, and allows users to set several parameters. For our purposes, we used several insert sizes for the two bee genomes. For *Bombus impatiens*, we used 500 bp shotgun, 3 kb and 8 kb inserts, while for *Megachile rotundata*, in addition to the inserts as that of *Bombus impatiens*, we used an additional 5kb insert. SOAPdenovo can be run either step-by-step or all at once. The step-by-step requires a user to wait until each step finishes, and tune some parameters before running the second step, and so on. This has to be done for four steps of the algorithm, in total: 1) the pregraph, 2) contig, 3) map and 4) scaffold stages. These steps can be run all at once, in which internally each step is run in turn, and it terminates at the final (scaffolding stage), following which the gap closure can be run separately. After testing several k-mers, k mer of 31 was found to be the most appropriate. It is also the maximum k-mer length that SOAPdenovo can handle.

CHAPTER 3

RESULTS

EST Assembly

About 1G bases were sequenced, and assembled into about 250 M of non-redundant contigs and singlets per bee species using Phrap, version 1.080721 (Table 3.1). The assembled ESTs were analyzed per species for GC bias. We found comparable GC content across the bees, which simplified codon level comparisons (Figure 3.1)

Database Schema

The BlastData database designed to store the species specific contigs and singletons information is implemented in MySQL, server version 5.0.77. There are seven tables in BlastData. The Project table keeps track of any new project referring to any new species that has had its EST sequenced. Every project is given an ID, implemented using the autoincrement field in MySQL, using which every new project inserted into the table gets a unique ID (integer data type). Each project is associated with its ID and its name (varchar data type). A custom PHP script was written to connect to BlastData, and load the assembled EST sequences for each project. The Sequence table has four attributes; 1) Sequence ID, 2) Sequence name, 3) FASTA Sequence, and the 4) Project to which this sequence belongs is referenced by the corresponding Project ID, which is the foreign key linking the Sequence and Project tables (Project ID being the primary key for the Project table). The Scaffold table has the Group number, sequence length and linkage group attributes, which keep track of the honey bee assembly 4.0 scaffolds. The LinkageGroup table has the linkage group number and the length of the chromosome attributes, while the Markers table stores the marker IDs for the corresponding linkage groups (Solignac et al. 2007, Genome Biology). The BeeHomologs table has five attributes; 1) Sequence name (var char data type), stores the name of the honey bee homolog; 2) the E-value attribute, records the E value after the BLAST; 3) mapping information on the scaffold; 4) the length of the sequence. There is also a Users table (not shown here) that is used to validate the user names

and passwords to connect to the database. Figure 3.2 shows the final database schema.

Website Layout and Design

The Honey Bee Homolog BLAST website was designed to provide an intuitive interface that users can use to BLAST against multiple bee and wasp databases. The website can be freely accessed at <http://bee12.cropsci.uiuc.edu>. The front end was designed using HTML, CSS, JavaScript and AJAX. A first glimpse of the website home page is as shown in Figure 3.3. Since the data is password protected, users can log in using the left panel, and if the login is successful, corresponding databases can be chosen for the BLASTs. These databases are blast formatted assembled EST datasets, as described in the Methodology. Users can select the appropriate program to use, blastn or tblastn. If the wrong combination of program and sequence is chosen, an error will be thrown on the screen. For example, choosing blastn, and entering a protein sequence will throw an error. Additionally, advanced parameters for the BLAST can be adjusted (choice of matrix, E-value, Number of alignments to display etc).

A PHP-MySQL script connects to the databases selected, and blasts the query against the databases. Figure 3.4 shows the results of the blast job run with the Frames option selected. The left panel shows the name of the program run (blastn or tblastn), and the databases that were used for the blast corresponds to what the user has chosen before submitting the job. This is followed by the number of hits found for each database. The right panel shows the alignment of the hits to the query gene.

The sequence retrieval feature of the website allows the users to select multiple gene hits, and run a multiple sequence alignment (CLUSTALW) on the selected hits. Multiple hits can be selected and the appropriate sequence retrieval method can be chosen. Figure 3.5 shows the selection of genes, with the Use Sliders (compact) option. This allows the user to define the area of the alignment to be retrieved using sliders, which can be moved on top of the graphical interface of the alignment display. Figures 3.6 and

3.7 show the result of choosing the Use Compact method to retrieve the sequences to obtain the corresponding CLUSTALW alignment. This is displayed in two tabs/windows depending on the browser preferences. Figure 3.6 shows the sequences retrieved using this method, while Figure 3.7 shows the CLUSTALW output. Figure 3.8 shows the location of an *Apis mellifera* gene on the corresponding linkage group for a given homolog query.

Putative Ortholog Detection and Alignment

The concept of reciprocal BLAST was used to define putative orthologs. Each *Apis mellifera* gene model was blasted against each of the bee databases, and all the hits, including the top hit that are within 10% identity from the top hit were blasted back to the gene model database to make sure we got the same gene model that we started with as the top hit. Using this operational definition of orthology, we got approximately 3647 gene models for which we found orthologs from all nine bee species of interest. This gave us a decent gene search space to run the selection analysis. This also reinforces a fundamental concept on which sociogenomics is built; that is, Genes that are conserved evolutionarily across diverse taxa can be used to probe deeper into the evolution of genes. About 1200 gene models did not have any BLAST hits from any of the bee species. This sheds some light into the per species gene gain/gene loss events wrt the reference genes. Each of the 3647 alignments containing orthologs to the reference were aligned to the reference gene using the E-INS-i strategy of the MAFFT algorithm. Manual editing of the alignments (Robinson lab) helped correct for the alignment errors that were hard to solve informatically (Figure 3.9)

PAML Results

Branch models were used to pick our lineage specific differences in the ω ratios. We tested our hypotheses (See Methodology) using a Perl script that automates the PAML Branch Model jobs over a cluster. The codeml program was run, using a batch submission approach on a cluster with 96 nodes. A maximum of 49 codeml jobs were submitted at a time (Figure 3.10) The log likelihoods obtained for each model were

tested using the likelihood ratio test that compares how well the alternative model fits the data compared to the null model. The null model here is the neutral model of evolution, which assumes that a given gene is evolving at the same rate across all the branches of the phylogeny, while the alternative model considers the genes in certain species to be evolving at a given rate, while the gene in the rest of the species across the phylogeny is evolving at a different rate. The p-value of the test statistic, is twice the difference in the log likelihoods of the null and the alternative models, and is estimated using a chi-square distribution. Using a p-value cut off of 0.05, gene lists were made for each hypothesis, indicating rapidly evolving genes. Figure 3.11 shows the results of the three different hypothesis tested, and the number of rapidly evolving genes obtained for each of the hypothesis. Note that we detected significant overlap across the tests (Figure 3.12). The rapidly evolving genes were annotated by gene ontology to obtain the over represented GO terms specific to each lineage. The results of each hypothesis tested, along with the GO annotation results are shown in Tables 3.2, 3.3 and 3.4 respectively. We found genes for gland development, signal transduction, and glycolysis evolving more rapidly in the eusocial lineages, which includes the primitively eusocial and highly eusocial lineages (Table 3.2), while the contrast between highly eusocial versus other lineages gave a lot of metabolic genes, especially glycolysis genes, and some genes involved in biosynthesis. (Table 3.3) The rapidly evolving genes in the primitively eusocial lineages mainly involve developmental processes related genes, like neuron differentiation and embryonic development (Table 3.4).

Whole Genome Assemblies

The de Bruijn graph based assembly algorithm, SOAPdenovo was used to generate a draft genome assembly of two of the bees, *Bombus impatiens* and *Megachile rotundata*. The assembly was run on a large memory cluster in which the compute node had 16 CPUs of 2.34 GHz each, and a total memory of 254.04 GB RAM. The raw reads were error corrected using Quake (Kelley et al. Unpublished). Custom Perl scripts were written to order the mates in the corrected reads into paired reads and singletons. SOAPdenovo was run using the following parameters for both the bee assemblies:


```
/path to SOAPdenovo/ all -s /path to configuration file/ -K 31 -R -o <output file name>
```

The 'all' parameter runs pregraph (construction of Kmer graph), contig (elimination of errors and output contigs), map (map reads to contigs) and scaff (scaffolding), each one in turn, while the -R parameter helps to resolve tiny repeats in reads. The configuration file specifies the path to the error corrected reads for three libraries in this case, 500 bp shotgun, 3kb and 8kb paired end reads. Figure 3.13 shows the statistics on the number of raw reads and the number of corrected reads organized into pairs and singlets used for the *B. impatiens* assembly. The scaffolds obtained from were input into the Gap Closer and run using the following parameters:

```
/path to GapCloser/ -b /path to SOAPdenovo configuration file/ -a /path to scaffold file/ -o <output file name> -p 31
```

The -p parameter specifies the number of threads to run Gapcloser on (default 8).

The *Bombus impatiens* reads could be assembled fairly easily, giving a contig N50 and scaffold N50 of 7.8 Kb (Table 3.5) and 1.2 Mb (Table 3.6) respectively. The sum of the scaffolds and the singletons from the assembly alone is about 260 Mb, which could be estimated to be the genome size of *Bombus impatiens*. Gap Closer was then run on the scaffolds giving a final assembly with 2450 scaffolds and 16% of the total gaps were closed. We also validated the putative Gapped Ortholog Reference Transcript Assembly (GOTA) that we obtained from the computational pipeline run on the EST data to our scaffold assembly. About 95% of the putative orthologs had a >95% match to the genome assembly.

For *Megachile rotundata*, error correction was done only for the 500 bp shotgun library (library was constructed from a haploid male), while the 3kb and 8kb paired end libraries (library was constructed from a pool of individuals) were trimmed for the central 42 bp linker from the reads. Since these reads represent properly circularized DNA molecules, one can be assured about their paired end insert sizes. These trimmed reads were then

checked for their lengths > 31 , (k mer size). However trimming the linkers and filtering them by the lengths reduced the total read number that was used for the assembly considerably (~7.5% of the total reads from 3kb library, and 1.2% of the total reads from the 8kb library). Assembly of these reads gave a contig N50 of 102 bp, while the scaffold N50 was 1.2 Kb. To make a better assembly, we then sequenced outward facing reads from another 5 kb library, which did not have any linkers to trim. The 500 bp reads were then used for the initial contiging step since these were not polymorphic, while the 5kb reads along with the trimmed and filtered 3kb and 8kb mate pairs were used for the later scaffolding step. The assembly slightly improved, with a contig N50 and scaffold N50 of 3.6 Kb (Table 3.7) and 31 Kb (Table 3.8) respectively. Figure 3.14 shows the statistics on the raw and the trimmed/filtered reads used for the *M. rotundata* assembly.

CHAPTER 4

DISCUSSION

Ortholog Identification

In this study, we have attempted to gain an understanding of the evolution of social life in molecular terms. We explored sociogenomics as an integrative discipline by combining our knowledge of phylogeny and wide diversity in social behaviour patterns with genomics. In order to address the question of what are the genomic changes associated with the evolution of social behaviour, we ran tests of selection on the putative orthologous genes identified computationally through the method of reciprocal best hit. This method is a greedy approach, in which it greedily stitches together the contigs to construct the Gapped Ortholog-reference-guided Transcript Assembly (GOTA). While this method solely relies on overlaps to concatenate the contigs, efforts were taken to make sure the overlap length was sufficiently long enough, so that the probability that any two contigs would have the same overlap at random and hence put together is minimized. Care was taken to make sure every contig which goes into the assembly is a reciprocal best hit, and stringent E-value cut offs were used. However, this still does not completely rule out the possibility of paralogs ending up in the assembly for a single contig. Since it is difficult to computationally distinguish the orthologs from paralogs using EST sequence data alone, it is quite possible for our assemblies to contain paralog contamination. A good way to validate the orthologs would be to use whole genome data, where we can look at Conserved Gene Neighbourhood (CGN) and synteny information, which can give more confidence on the orthologous genes, since the gene order across related species should be somewhat conserved. We could have used graph based approaches that have been used for genome assemblies by considering different K mers for each node, and extending unique K mers based on depth information, but this method based on de Bruijn graphs works well for short read sequence data, and for ESTs, this might not be an optimal approach because it will take a lot of memory to store long reads and hash them, which in turn might be heavy on the memory needed to perform these computations. Also, it

works poorly on 454 data as homopolymer errors have a much larger impact on this type of assembly. Overlap based methods are known to work fairly well for EST data, and the reciprocal best hit method is a good first pass approximation to identifying putative orthologs. Using this approach, we were able to identify putative orthologous genes from all species for about 33% of the *Apis mellifera* gene models, while about 10% of the gene models did not have any hits. This could mean that the genes are either missing in the specific lineages, or missing genes yet to be annotated and included the *Apis mellifera* official gene set. The use of the official gene set provides a tunnel vision for the analyses of rapidly evolving genes.

Tests of Selection

Although several methods have been developed that utilize the concept of the rate of synonymous (dS) and non synonymous substitutions (dN) to infer selection, most of these approaches may lack power due to the model assumptions, choice of outgroup species and the number of taxa considered for the analyses (Messier and Stewart 1997; Zhang and Kumar 1997; Yang 1998). We used the branch model tests (codeml program of PAML package) which average ω (dN / dS) over the entire branch and check if its greater than 1, in which case it is inferred as positive selection affecting that branch. However, since positive selection is unlikely to act on all the sites on a branch over evolutionary time, the average ω is very rarely greater than 1. Hence we applied the branch models to pick out genes that are “rapidly evolving” where the ω for a given lineage may not be greater than 1, but will be still greater than that of the background branches put together. The background branches here refer to those where we hypothesize the selective pressure to be absent. The rapidly evolving branches may then be tested by the sites model, which performs a naïve empirical Bayes estimation to identify specific sites under selection on the branch (Yang 2007). The branch-site models on the other hand are designed to detect episodic bouts of positive selection affecting only a few amino acid residues on a few lineages. However, since this test has too many free parameters, and is known to generate many false positives when the model assumptions are violated, we refrained from using these models. Yang et al.

recently developed a modified version of this test found to have a reasonable power under a variety of selection schemes, and this has been used extensively (Crespi et al. 2007; Vamathevan et al. 2008). While the codon substitution model of Goldman and Yang incorporates a lot of parameters that describe biological sequence evolution well, for instance the transition/transversion rate ratio and equilibrium frequency of codons, it assumes a Markov model of codon evolution and does not incorporate codon insertion/deletion effects into the model. Moreover, since one cannot set directional hypothesis in codeml, the rejection of the null hypothesis in these tests just implies that the gene of the foreground branch is evolving at a different rate from the background branches, and gives no information on the direction of the change since the test is two tailed. Some alternate packages like HyPhy (Pond et al. 2004) that allow one to test directional hypothesis in a maximum likelihood framework could be explored in such cases. Also developing a better codon substitution model incorporating indels may be designed. Adding more dimensions to the data can help get closer to true values.

Alignment Accuracy and Selection Estimates

While the maximum likelihood estimation of codeml followed by the likelihood ratio test to detect the rapidly evolving genes is a fairly conservative approach, a more careful consideration of the alignment accuracy on the final outcome is required. Previous studies have noted that different alignment methods lead to different conclusions regarding the detection of positively selected sites (Schneider et al. 2009; Mallick et al. 2010). The effect of insertions, deletions and alignment errors on the branch-site test of positive selection has been systematically studied by Fletcher and Yang (Fletcher and Yang 2010). Here simulations were performed using the program INDELible which generates different data sets under codon models incorporating indels (Fletcher and Yang 2009), followed by the generation of multiple sequence alignments using different algorithms like PRANK (Loytynoja and Goldman 2005), MAFFT (Kato et al. 2002), MUSCLE (Edgar 2004) and ClustalW (Thompson et al. 2004). The results showed that PRANK (codon alignments) consistently performed best having the lowest false positive rates followed by MAFFT, MUSCLE and ClustalW performing the worst. It seems like

the latter algorithms have a tendency to place the non-homologous codons into one column, giving the false impression of selection affecting those sites (Fletcher and Yang, 2010). In this study, all the alignments were manually inspected, and edited for errors by deleting ambiguous codons completely. While this painstaking correction would have reduced the false positive rate to a considerable extent, it still does not rule out the possibility of alignment errors leading to false positives in the selection tests completely. It may have been worthwhile to realign the codons using PRANK and compare the proportion of rapidly evolving genes to the ones obtained after manually editing the alignments. Also, since it is harder to simulate the effects of alignment uncertainties on the branch model test, as compared to the branch-site models, it may be difficult to estimate the error range for our analysis (Fletcher and Yang, 2010).

Lineage Specific Rapidly Evolving Genes

Using our approaches, we identified several rapidly evolving lineage specific genes, and the over representation of their corresponding GO terms. Some of the results have clear implications to social phenotypes, while some are new insights. In the contrast between eusocial lineages versus the rest, the GO enrichment test picked out the genes involved in gland development and signal transduction as rapidly evolving. Gland development genes may well fit into the eusocial evolution scenario for several reasons. Three gland systems, that is, hypopharyngeal, mandibular, and salivary glands are present in the heads of social and solitary bees (Cruz Landim, 1967). The glandular food discharged from the mouth of the workers is eaten by the larvae and the queen of the colony. With some exceptions in solitary bees, the highly eusocial (*Apis mellifera*) bees and the stingless bees (Meliponini) appear to be the only ones in which glandular feeding habit is prominent (Michener, 1974). According to a study by Webster and Ping, the association of glandular feeding habit with sociality is due to four adaptive features of exocrine glands: 1) Glandular food is easily digestible, hence the bulk of faeces accumulation in the hive is minimized, which in turn reduces the load on hive cleaning. 2) Queen's fecundity is increased. 3) Nutrient recovery via cannibalism is facilitated. 4) Rearing of replacement queens is expedited (Webster and Ping, 1988). Since it has

been extensively shown that glandular food is a large part of diet in principally eusocial bees, it can be well hypothesized that the genes for exocrine gland development are constantly evolving in the eusocial species as opposed to solitary bees where the principal diet consists of pollen and nectar, and even though they do have the genes for gland development, there is no selective pressure on these genes to evolve based on colony costs. In addition to this, it is a well-known fact that in eusocial bees, the ovaries of the queens are very well developed. The queen is born with a larger complement of ovarioles than the worker. Several million sperms are deposited in the queen's oviduct, and few of those are stored in the spermatheca, which are then used to fertilize the eggs. These fertilized eggs can then develop into workers and new queens, while the unfertilized eggs develop into drones. We also found over representation of GO terms for genes involved in signal transduction pathways, notably insulin. Toth et al. analyzed the wasp gene expression dataset using next generation sequencing approaches and found insulin related genes among the differentially regulated genes, suggesting that evolution of eusociality involved major nutritional and reproductive pathways (Toth et al. 2007). The insulin pathway has been implicated in honey bee queen worker caste determination (Wheeler et al. 2006; Corona et al. 2007; Patel et al. 2007) and worker foraging behaviour (Ament et al. 2006; Hunt et al. 2007). The honey bee genome project also shed light on some notable differences between honeybees, *C. elegans* and *D. mel* for components of insulin/insulin like growth factor signaling pathways, suggesting that honeybees have evolved a different regulation of this complex pathway (Robinson et al. 2006). Besides these, we also found genes from highly conserved cell signaling pathways that are responsible for most developmental cell-cell interactions in metazoans like thick veins (hedgehog signaling) and costa (epidermal growth factor signaling).

In the contrast between highly eusocial versus the rest of the phylogeny, we identified enrichment for genes involved in odor perception (Tachykinin, no receptor potential A), and metabolism, mainly glycolysis (13 genes, $P < 5.18E-11$). It has been shown that in the highly eusocial *Apis mellifera*, there is a remarkable expansion of insect odorant receptor family (about 170 odorant genes, out of which 7 are pseudogenes) relative to

D.mel and *A.gambiae*. One can hypothesize that these genes are evolving rapidly because the odorant receptor expansion presumably mediates the range of odorant abilities in highly eusocial species, which includes perception of pheromones, kin recognition, and perception of floral colors (Robinson et al. 2006). Also since the eusocial have to perform several energetically demanding tasks like nest thermoregulation and increased foraging activity, the glycolysis genes might be under different selective pressures compared to the non-eusocial.

In the contrast between primitively eusocial bees versus rest of the phylogeny, we found genes implicated in cAMP signaling and learning and memory (*dunce*), development, histone modification and chromatin remodeling factors. It has been shown that *dunce* mutants fail to learn as larvae and to retain memory in adulthood. Such learning tasks are believed to be involved in circuits in mushroom bodies, and may have some restructuring of the brain region related to these tasks (Heisenberg 1989; Laurent and Davidowitz 1994). The effect of visual experience is known to increase the volume of the calyx brain region and is cAMP dependent. This effect is found to be absent in *dunce* mutants.

Whole Genome Assemblies

We used SOAPdenovo assembly algorithm based on de Bruijn graphs to generate draft whole genome assemblies for our bees. SOAPdenovo is designed to assemble large, repetitive genomes from short read sequence data like those generated from Illumina. Unlike other greedy assemblers, SOAPdenovo splits the whole assembly process into distinct phases, with separate processing of repetitive sequences. Unambiguous stretches of sequence form non-branching paths in the de Bruijn graph which makes it easy to read off the contigs (Salzberg et al. 2010). Sequencing error creates tips and bubbles in the de Bruijn graph, which are solved by correcting for dead-end nodes and the Tour bus algorithm as used in Velvet. Though SOAPdenovo's methods are largely derived from Velvet, the graph is more space efficient. However, the de Bruijn graph approach also has some drawbacks. Decomposition of reads into K mers can lead to

loss of information. The graph may not be read coherent (Myers et al. 2005). However these are the only assemblers have shown promising results for human size dataset. SOAPdenovo produced a decent draft assembly for *Bombus impatiens* data, but for *Megachile rotundata*, the assembler seems to be performing poorly largely due to the presence of AT rich repeats, which we analyzed by mapping the EST data onto the scaffold assembly for *M. rotundata* (data not shown). In such cases, it might help to generate a hybrid assembly using both 454 and Illumina sequencing data (Nagarajan et al. 2010) where the complementary nature of assembly algorithms can be used to significantly improve the quality of de novo scaffolds. Also, it might help to have larger insert sizes, which can be used to merge scaffolds by aligning the mate pairs to the contigs. Having a good gap closer algorithm can also significantly improve the quality of the scaffolds. Newer algorithms have been developed that iteratively align sequences against contig ends and perform local assemblies to produce gap-spanning contigs. Such improved iterative mapping methods can be explored to improve the continuity of a draft genome without the need to generate new data (Tsai et al. 2010). The importance of having good error correction algorithms cannot be underestimated. Current algorithms fail to distinguish between true errors and polymorphisms, and hence can be applied only to reads from haploid genomes. A probabilistic and machine learning framework is required to distinguish between the SNPs and the overall error rates for any sequencing technology to improve the current algorithms. Last, but not the least, new methods need to be developed that can help assess the correctness of an assembly.

CHAPTER 5

CONCLUSIONS

We have integrated species phylogeny, differences in social behaviour patterns across a broad array of extant Hymenopterans and Genomics to explore putative molecular signatures of selection that may be involved in complex eusocial behaviour. Eusociality has evolved multiple independent times, and it is unknown if these independent events eventually converged for complex traits. Using our approaches, we identified a shared set of genes, common to species showing different levels of sociality, as well as lineage specific genes. Further research needs to be done to understand how these adaptive changes are actually advantageous to the lineages. About 1GB transcriptome sequence analyzed for this study was assembled into non-redundant contigs and singletons. Users can now access this rich source of assembled EST data, and BLAST against multiple bee databases through our webserver, accessible at <http://bee12.cropsci.uiuc.edu>. A draft genome of a primitively eusocial and a solitary bee was generated using SOAPdenovo algorithm. The whole genome projects will advance our knowledge on the bee genome architecture and provide deeper insights on molecular evolution for sociogenomics studies and other comparative genomic analyses.

REFERENCES

1. Gene E. Robinson, Christina M. Grozinger and Charles W. Whitfield. 2005. Sociogenomics: Social life in molecular terms.
2. Robinson G.E., Fahrbach, S.E. & Winston, M.L. Insect societies and the molecular biology of social behaviour. 1997. *Bioessays* 19, 1099-1108.
3. Robinson, G.E., integrative animal behaviour and sociogenomics. 1009. *Trends Ecol. Evol.* 14, 202-205.
4. Gene E Robinson. 2002. Sociogenomics takes flight. 2002. *Science* 297, 204-205.
5. E.O. Wilson Sociobiology. 1975. The New Synthesis, belkna, Cambridge, MA.
6. S.B. Carroll, J.K. Grenier, S.D. Weatherbee, 2001. From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design, Blackwell Science, Malden MA.
7. Dubnau J. and Tully, T. 2001. Functional anatomy: from molecule to memory. *Curr Biol* 11, R240-R243.
8. Panda, S., Hogenesch, J.B. & Kay, S.A. 2002. Circadian rhythms from flies to human. *Nature* 417, 329-335.
9. Greenspan, R.J. & Ferveur, J.F, 2000. Courtship in *Drosophila*. *Annu Rev Genet* 34, 205-232.
10. Pfaff, D.W., 1999. Drive: Neurobiological and Molecular Mechanisms of Sexual Motivations. MIT Press, Cambridge, MA.
11. Jarvis, E.D., Schwabl, H., Ribeiro, S. & Mello, C.V. 1997. Brain gene regulation by territorial singing behaviour in freely ranging songbirds, *Neuroreport* 8, 2073-2077.
12. Robinson G.E. and Ben Shahr. 2002. Social behaviour and comparative genomics: new genes or new gene regulation ? *Genes, Brain and Behaviour* 1 197-203.
13. Feder ME & Mitchell-Olds T. 2003. Evolutionary and ecological functional genomics, *Nature Rev. Genet.* 4, 651-657.
14. Vignal C., Mathevon, N & Mottin, S. 2004. Audience drives male songbird response to partner's voice. *Nature* 430, 448-451.
15. Alcock, J. 1998. Animal Behaviour: an Evolutionary Approach, Sinauer, Sunderland.
16. Wilson, E.O. 1971. The insect societies, Belknap, Harvard Univ Press, Cambridge MA.
17. Duffy, J.E, 1996. Eusociality in a coral-reef shrimp, *Nature* 381, 512-514.

18. Sherman, P, Jarvis J et al. 1991. The biology of the naked mole rat. Monographs in behaviour and ecology, Princeton Univ Press, MA.
19. Choe J. C. & Crespi, B.J. 1997. The evolution of social behaviour in insects and arachnids, Cambridge Univ Press, Cambridge.
20. Holldobler, B. & Wilson, E.O. 1990. The Ants, Belknap, Harvard Univ Press, Cambridge MA.
21. Amy, L. Toth, Gene E. Robinson. 2007. Evo-devo and the evolution of social behaviour, Trends in Genetics, Vol 23, No. 7.
22. Gellon, G. et al. 1998. Shaping animal body plans in development and evolution by modulation of Hox expression patterns. Bioessays 20, 116-125.
23. Pichaud, F. et al. 2002. Pax genes and eye organogenesis. Current Opin. Genet. Dev. 12, 430-434.
24. Robinson et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443, 931-949.
25. Krieger, M.J. & Ross, K.G. 2002. Identification of a major gene regulating complex social behaviour. Science. 295, 328-332.
26. Lockhart, P.J. & Cameron, S.A. 2001. Trees for bees. Trends Ecol Evol 16, 84-88.
27. White et al. 2002. Social regulation of gonadotropin-releasing hormone, J. Exp Biol, 205, 2567-2581.
28. Band et al. 2000. An ordered and comparative map of the cattle and human genomes. Genome Res 10, 1359-1368.
29. Summers et al. 2001. Comparative physical mapping of targeted regions of the rat genome. Mamm Genome 12, 508-512.
30. Fitzpatrick, M et al. 2005. Candidate genes for behavioural ecology. Trends Ecol Evol. 20, 96-104.
31. Tully, T. 1996. Discovery of genes involved in learning and memory :an experimental synthesis of Hirschian and Benzerian perspectives. Proc Natl Acad. Sci. USA 93, 13460-13467.
32. Robinson et al. 2000. Sperm-mediated transformation of the honey bee, *Apis mellifera*. Insect Mol Biol 9, 625-634.

33. Michael L Metzker.2010.Sequencing technologies, the next generation, Nature Reviews, Genetics.
34. Birol et al. 2001. De Novo transcriptome assembly with Abyss.Bioinformatics.
35. Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147:195-197.
36. Gotoh, O. 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162:705- 708.
37. Green, P. 1999. Phrap, SWAT, Crossmatch. Available from author. University of Washington.
38. Ouzounis,C.A. et al.2003.Classification schemes for protein structure and function.Nat. Rev. Genet.4,508-519.
39. Kuzinar et al.2008.The quest for orthologs: finding the corresponding genes across genomes.Cell, 539-551.
40. Eisen, J.A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8, 163–167.
41. Jeffroy, O. et al. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231.
42. Delsuc, F. 2005.Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.
43. Tatusov, R.L. et al. 1997.A genomic perspective on protein families. Science 278, 631–637.
44. Tatusov, R.L. et al. 2003.The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.
45. Goodstadt, L. and Ponting, C.P. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLOS Comput. Biol. 2, e133.
46. Bandyopadhyay, S. et al. 2006. Systematic identification of functional orthologs based on protein network comparison. Genome Res. 16, 428– 435.
47. Mazurie, A. et al. 2005. An evolutionary and functional assessment of regulatory network motifs. Genome Biol. 6, R35.
48. Grigoryev, D.N. et al. 2004.Orthologous gene-expression profiling in multi-species models: search for candidate genes. Genome Biol. 5, R34.

49. Mao, F. et al. 2006. Mapping of orthologous genes in the context of biological pathways: An application of integer programming. *Proc. Natl. Acad. Sci. U. S. A.* 103, 129–134.
50. Hulsen, T. et al. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* 7, R31.
51. Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113.
52. Fitch, W.M. 2000. Homology a personal view on some of the problems. *Trends Genet.* 16, 227–231.
53. Jothi, R. et al. 2006. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* 22, 779–788.
54. Storm, C.E.V. and Sonnhammer, E.L.L. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18, 92–99.
55. Hirsh, A.E. and Fraser, H.B. 2001. Protein dispensability and rate of evolution. *Nature* 411, 1046–1049.
56. Overbeek, R. et al. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901.
57. Wall, D.P. et al. 2003. Detecting putative orthologs. *Bioinformatics* 19, 1710–1711.
58. Lee, Y. et al. 2002. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* 12, 493–502.
59. Remm, M. et al. 2001. Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052.
60. Hubbard, T.J.P. et al. 2007. Ensembl 2007. *Nucleic Acids Res.* 35, D610–D661.
61. Wheeler, D.L. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36, D13–D21.
62. Cannon, S.B. and Young, N.D. 2003. OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics* 4, 35.
63. Dehal, P.S. and Boore, J.L. 2006. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 7, 201.

64. Merkeev, I.V. et al. 2006.PHOG: a database of supergenomes built from proteome complements. *BMC Evol. Biol.* 6, 52.
65. Li, H. et al. 2006.TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 34, D572–D580.
66. Yang Z.,2005.The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci USA.* 102;3179-3180.
67. Thomas et al. 2003.Comparative analyses of multi-species sequences from targeted genomic regions.*Nature.*424;788-793.
68. Nielsen et al. 2005.A scan for positively selected amino acid sites and applications to the HIV-1 envelope gene.*Genetics.* 148;929-936.
69. Sawyer et al. 2005.Positive selection of primate TRIM5-alpha identifies a critical species-specific retroviral restriction domain.*Proc Natl Acad Sci,USA.* 102;2832-2837.
70. Ziheng Yang.2007.PAML 4:Phylogenetic Analysis by Maximum Likelihood.*Mol Bio and Evolution.*24(8):1586-1591.
71. Kimura, M.1977.Preponderance of synonymous changes as evidence for the neutral theory molecular evolution.*Nature.*267:275-276.
72. Miyata,T. and Yasunaga,T.1980.Molecular evolution of mRNA:a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications.*J Mol Evol.*16:23-36.
73. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczký J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al.*: Initial sequencing and analysis of the human genome. *Nature* 2001, 409(6822):860-921.
74. Edwards, A; Caskey, T.1991. Closure strategies for random DNA sequencing. *Methods: A Companion to Methods in Enzymology* 3 (1): 41–47. doi:10.1016/ S1046-2023(05)80162-8.
75. Huang X., Madan A.(1999) CAP3: A DNA sequence assembly program. *Genome Res.* 9:868–877.

76. Sutton et al. 1995. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *Genome Science and Technology*. 1(1): 9-19.
doi:10.1089/gst.1995.1.9.
77. Myers E.W., et al. (2000) A whole-genome assembly of *Drosophila*. *Science* 287:2196–2204.
78. Mulikin et al. 2003. The Phusion assembler. *Genome Research*.
79. Chevreux et al. (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99*, pp. 45-56.
80. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 18: 802–809.
81. Batzoglu, S. et al. 2002. ARACHNE: A whole genome shotgun assembler. *Genome Research*. 12: 177-189.
82. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Res* 2008;18:324-330.
83. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;18:821-829.
84. Jeck, W.R., et al (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics* 23:2942–2944.
85. Pevzner, P., Tang, H., and Waterman, M.S. 2001. A new approach to fragment assembly in DNA sequencing. *Proceedings of the Fifth Annual International Conference in Computational Molecular Biology (RECOMB)*, April 22–25, 2001, Montreal, pp. 256–267. ACM Press, New York..
86. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20: 265–272.
87. Sacha A. F. T. van Hijum, Aldert L. Zomer, Oscar P. Kuipers, and Jan Kok. 2005. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Res.* July 1; 33(Web Server issue): W560–W566.

88. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* 2008;18:1638-1642.
89. Salzberg et al. 2002. Fast Algorithms for large scale genome alignment and comparison. *Nucl. Acids Res*; 30 (11): 2478-2483.
90. Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, et al. 2004 The Atlas Genome Assembly System. *Genome Res*. 14: 721–732.
91. Loytynoja A., Goldman N.2005.An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA* ;102:10557-10562.
92. Edgar R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*;32:1792-1797.
93. CLUSTALW Thompson J. D., Higgins D. G., Gibson T. J. 1994.CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* ; 22:4673-4680.
94. Katoh et al. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucl. Acids Res.*, 30 (14): 3059-3066.
95. Heisenberg M.1989.Genetic approach to learning and memory in *Drosophila melanogaster*. *Fundamentals of Memory formation: Neuronal Plasticity and Brain Function*, Gustav Fischer; New York.
96. Laurent G.,Davidowitz, H. 1994. Encoding of olfactory information with oscillating neural assemblies. *Science*.265;1872-1875.
97. Farris, SM, Robinson GE, Fahrback SE.J. *Neurosci* 2001.Experience and age related outgrowth of intrinsic neurons in the mushroom bodies of the adult worker honeybee.21;6395-6404.
98. Cruz Landim,C.1967.Estudo comparativo de algumas glandulas das abelhas (Hymenoptera,Apoidea) respectivas implicacoes evolutivas.*Arq. Zool. Sao Paulo*.15.177-290.
99. Michener C.D.1974. *The Social Behaviour of the Bees*. Harvard Univ Press. Cambridge, MA.

100. Pond et al. 2004. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*. 5:676-679.
101. Salzberg et al. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research*, Vol 20, No 9, pp. 1165-1173.
102. Huelsenbeck et al. 2001. MRBAYES: Bayesian Inference of phylogeny. *Bioinformatics*, 17:754-755.

APPENDIX
FIGURES AND TABLES

FIGURES

Figure 2.1 Shows the number of nonredundant contigs and singletons loaded in BlastData.

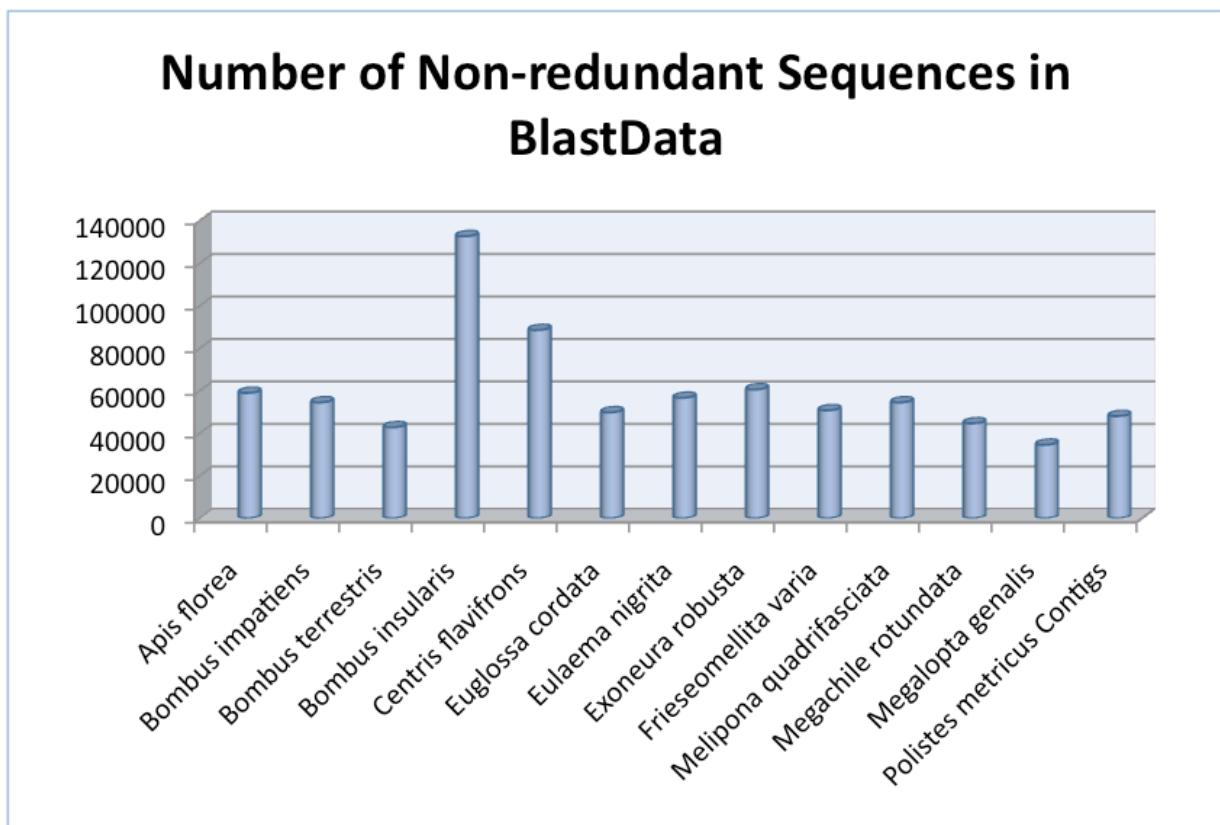


Figure 2.2 Putative Ortholog Assignment Pipeline. Orthology was assigned through the method of reciprocal BLAST. Each *A.mellifera* gene model is blasted against each of the species-specific non-redundant databases, and the top hits are blasted back to ascertain reciprocity. The best reciprocal hits are concatenated, after trimming of the overlaps to generate a Gapped Ortholog-reference-based Transcript Assembly (GOTA).

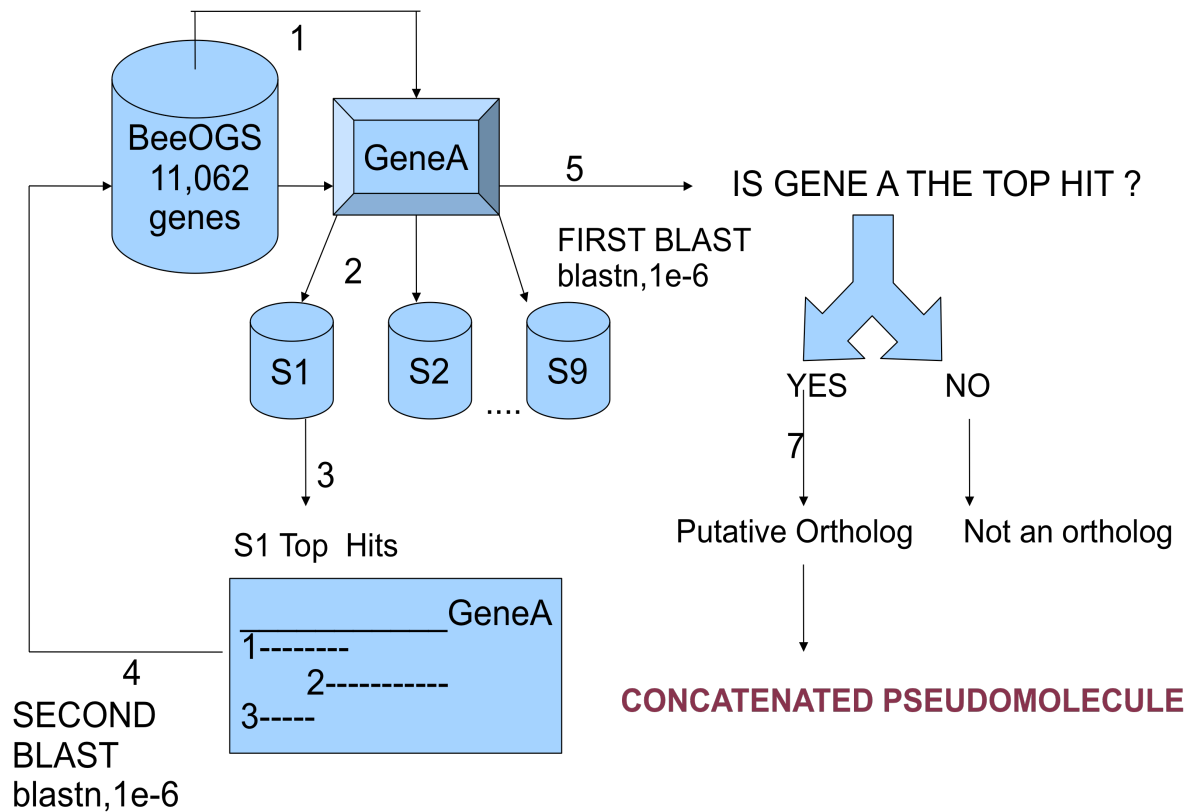


Figure 2.3 An Example Multiple Sequence Alignment as viewed in Geneious. The GOTAs obtained from the reciprocal BLAST pipeline were aligned to the *A. mellifera* reference gene models using the E-INS-i alignment strategy of MAFFT, which uses iterative refinement based on the method of weighted sum of pairs and consistency scores. A max iterations of 1000 was used for the MAFFT runs. The alignments were then manually edited to remove ambiguities using Geneious alignment editing software.



Figure 2.4 Species Phylogeny. Alignments were degapped and concatenated based on different criteria (See Methods) and analyzed by Bayesian Analysis to get the species tree represented below. Each node of the above tree had a posterior probability close to 1 (Dr. Sydney Cameron, UIUC, Unpublished).

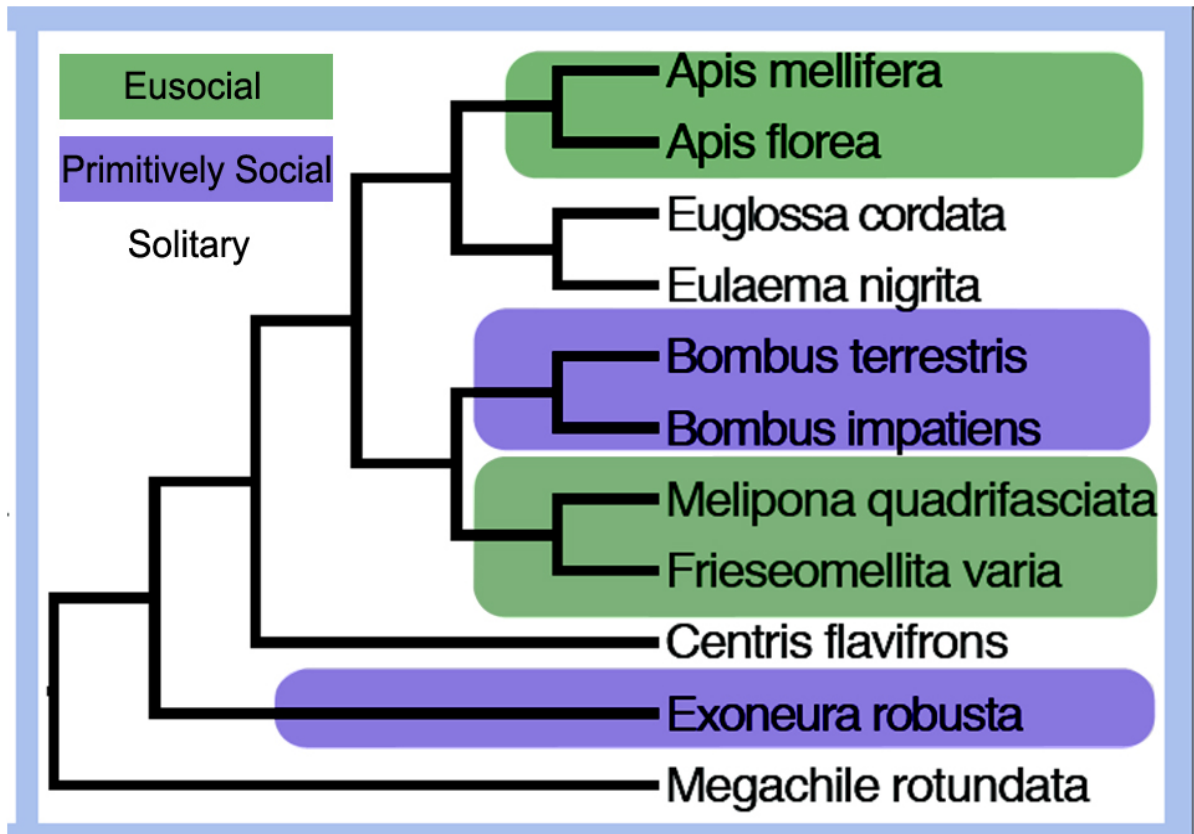


Figure 3.1 %GC of non-redundant contigs and singletons across nine bees.

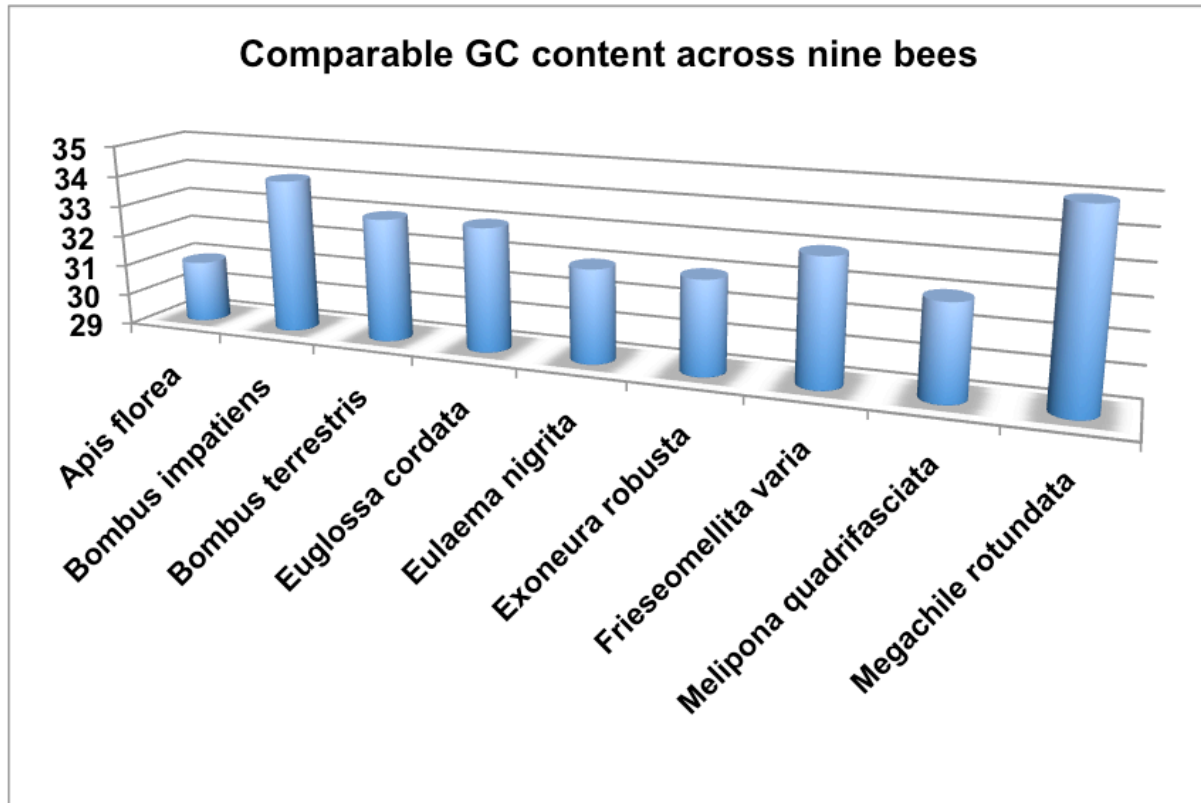


Figure 3.2 Database Schema. The six tables comprising BlastData are shown in Squares, (names in bold) around the central MySQL database. The attributes for each table are shown in brackets below the Table names. In addition, a Users table (not shown here), validates the usernames and passwords to log into BlastData.

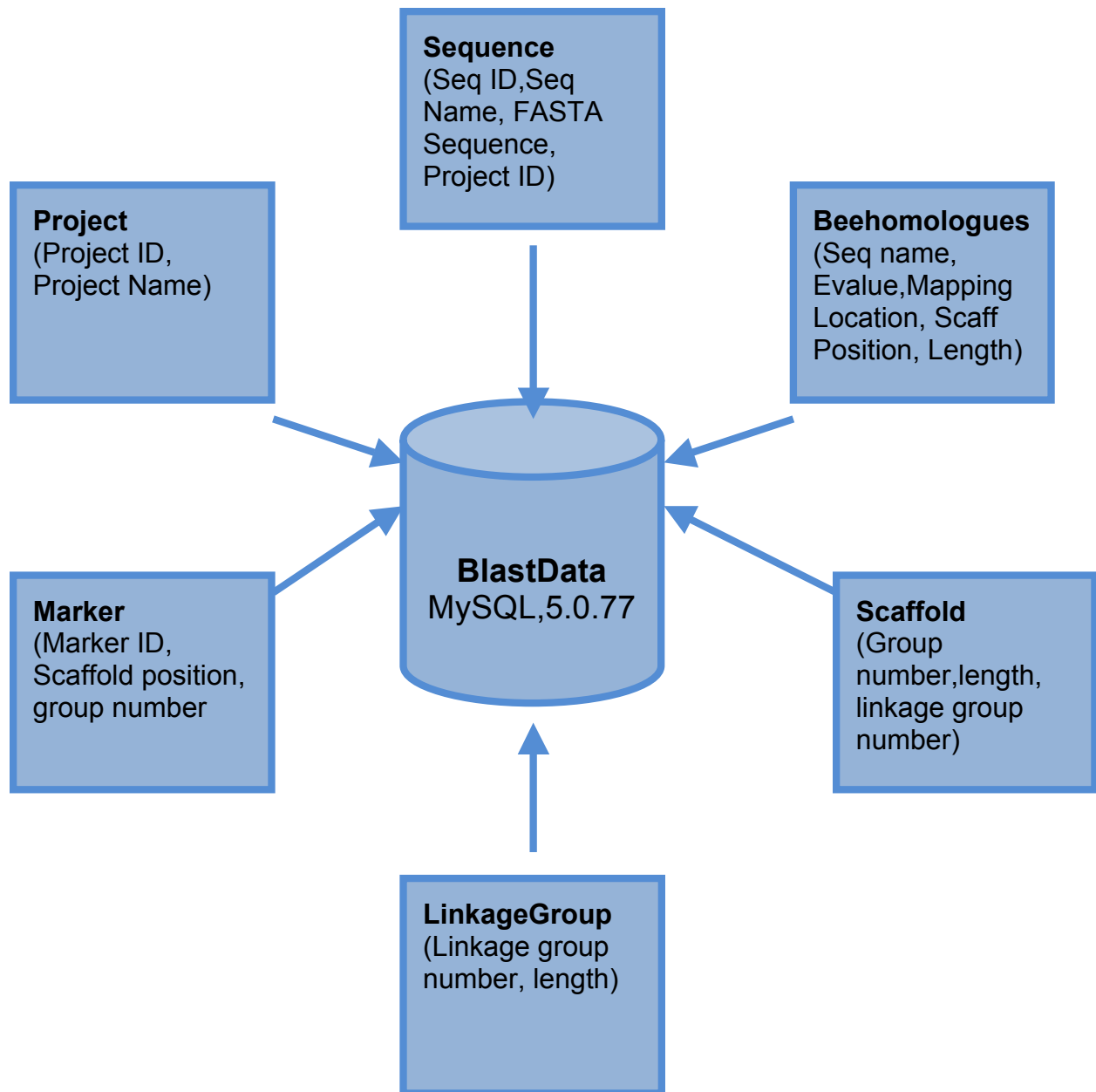
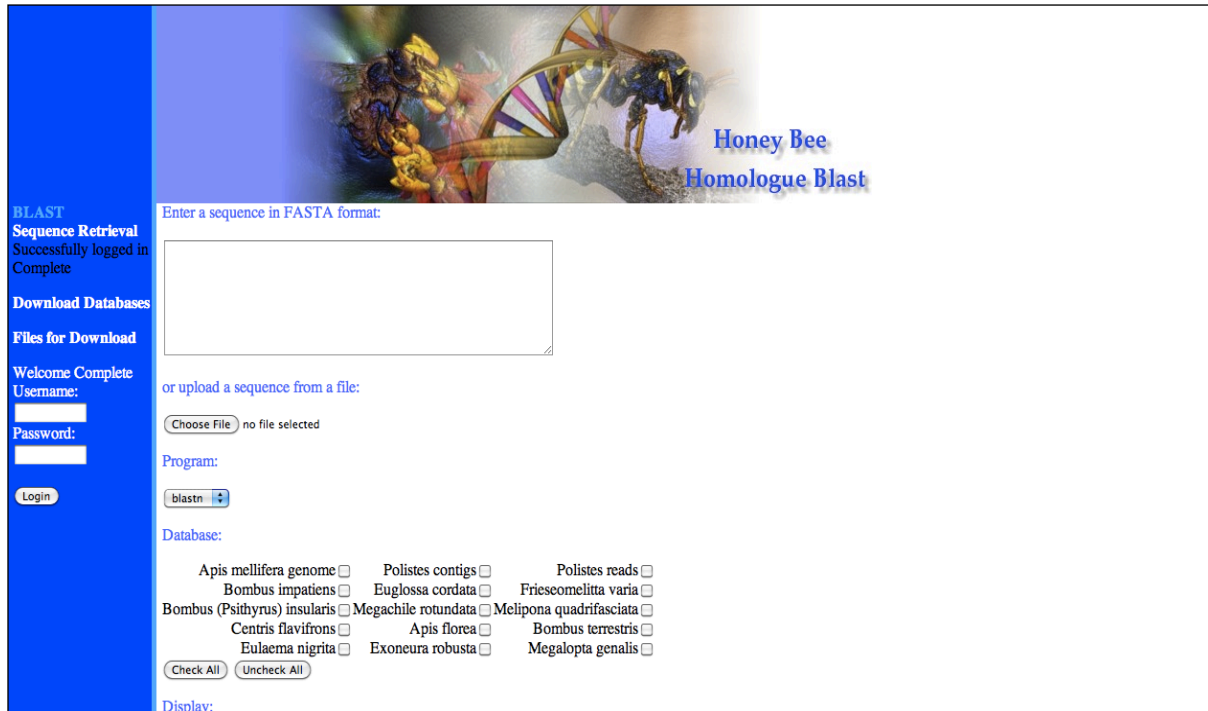


Figure 3.3 Honey Bee Homolog Blast website home page.



The image shows the home page of the Honey Bee Homologue Blast website. The page features a blue sidebar on the left with navigation links and a main content area on the right. The main content area includes a header with a honey bee and a DNA double helix, followed by a text input field for FASTA format sequences, a file upload option, a program selection dropdown (set to 'blastn'), and a database selection grid with checkboxes for various species. A 'Login' button is located in the sidebar.

Honey Bee Homologue Blast

Enter a sequence in FASTA format:

or upload a sequence from a file:

Choose File no file selected

Program:

blastn

Database:

<input type="checkbox"/> Apis mellifera genome	<input type="checkbox"/> Polistes contigs	<input type="checkbox"/> Polistes reads
<input type="checkbox"/> Bombus impatiens	<input type="checkbox"/> Euglossa cordata	<input type="checkbox"/> Frieseomelitta varia
<input type="checkbox"/> Bombus (Psithyrus) insularis	<input type="checkbox"/> Megachile rotundata	<input type="checkbox"/> Melipona quadrifasciata
<input type="checkbox"/> Centris flavifrons	<input type="checkbox"/> Apis florea	<input type="checkbox"/> Bombus terrestris
<input type="checkbox"/> Eulaema nigrita	<input type="checkbox"/> Exoneura robusta	<input type="checkbox"/> Megalopta genalis

Check All Uncheck All

Display:

BLAST
Sequence Retrieval
Successfully logged in
Complete

Download Databases

Files for Download

Welcome Complete
Username:
Password:
Login

Figure 3.4 A blastn run of hsp90 *Apis mellifera* gene against all the databases. The number of contigs/singleton hit per database is displayed on the left panel, the alignment against the query gene is displayed on the right. The blastn program was run using the frames option.

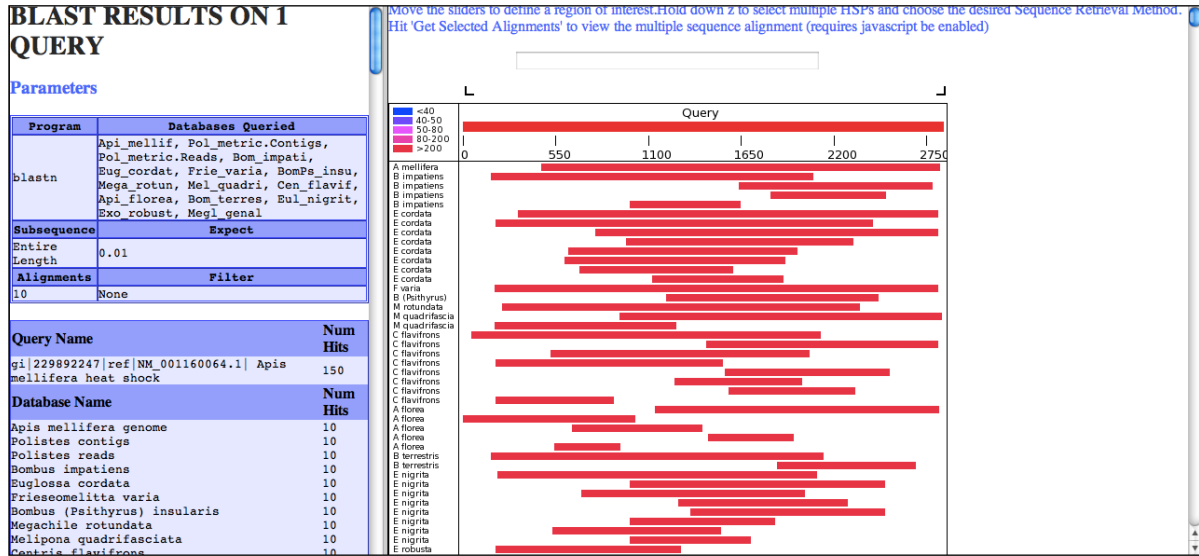


Figure 3.5 Use the sliders to define a region of the multiple sequence alignment. Hold z to select multiple genes, and align them using ClustalW after selecting the appropriate Sequence Retrieval Method.

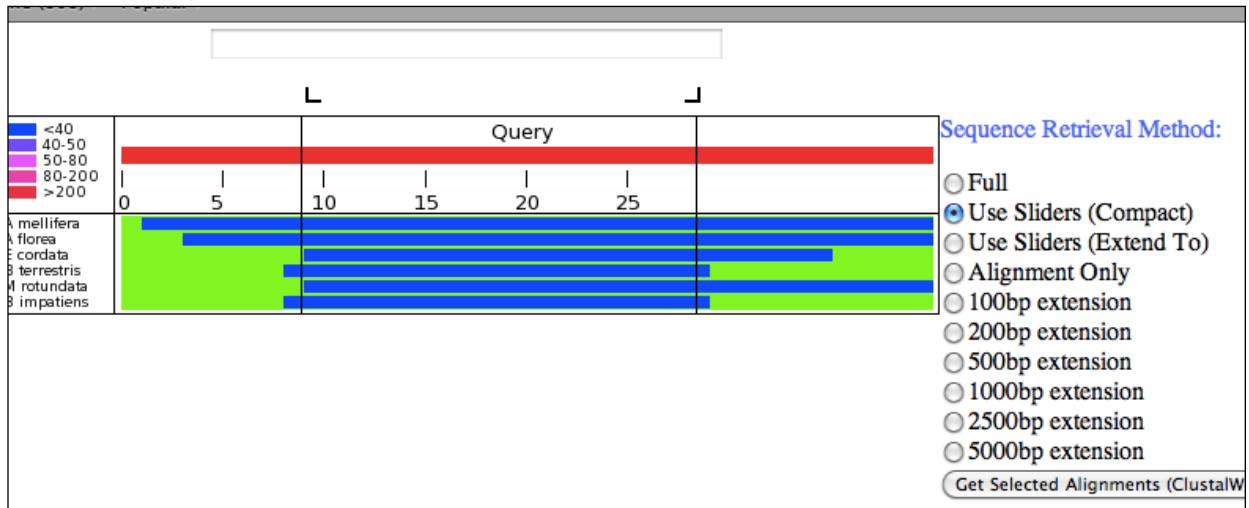


Figure 3.6 Sequence Retrieval Feature allows user to retrieve the selected multiple alignment sequences.

```
>Query
tactctgaaaagtgacg
>Api_mellif.Group7.33
ATACTCTGAAAAGTGACG
>Api_florea.Contig27622
TATGAGACTTTTCACTGC
>Eug_cordat.Contig26359
ATACTCTGAACAGTGACG
>Bom_terres.Contig19747
ATACTCTGAAAAGTGACG
>Mega_rotun.Contig13518
TATGAGACTTTTCATTGA
>Bom_impati.Contig30539
TATGAGGCTTTTCACTGC
```

Figure 3.7 Result of the ClustalW alignment.

```
Api_mellif.Group7.33      ATACTCTGAAAAGTGACG-----
Bom_terres.Contig19747  ATACTCTGAAAAGTGACG-----
Query                    -TACTCTGAAAAGTGACG-----
Eug_cordat.Contig26359  ATACTCTGAACAGTGACG-----
Api_florea.Contig27622  ----TATGAGACTTTTCACTGC
Bom_impati.Contig30539  ----TATGAGGCTTTTCACTGC
Mega_rotun.Contig13518  ----TATGAGACTTTTCATTGA
```

Figure 3.8 The location of a honey bee gene is shown in red on linkage group 7. Zoom levels can be adjusted to one that is convenient.

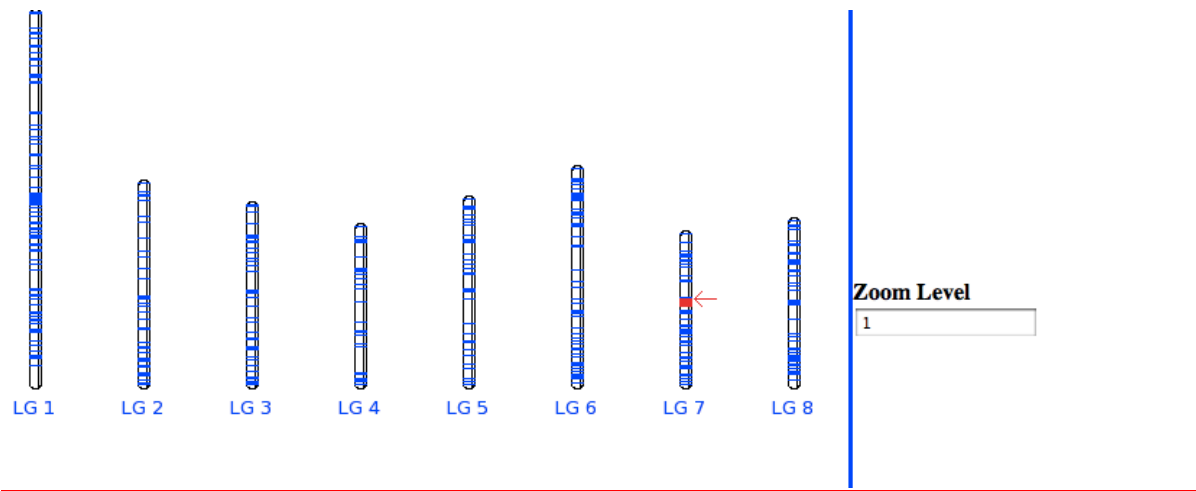


Figure 3.9 Shows the Distribution of the number of orthologs found for each *Apis mellifera* gene model. The computational pipeline based on the method of reciprocal BLAST gave about 33% of the *A. mellifera* gene models had orthologous genes in all nine bee species, while about 10% of the gene models had no hits in any of the species.

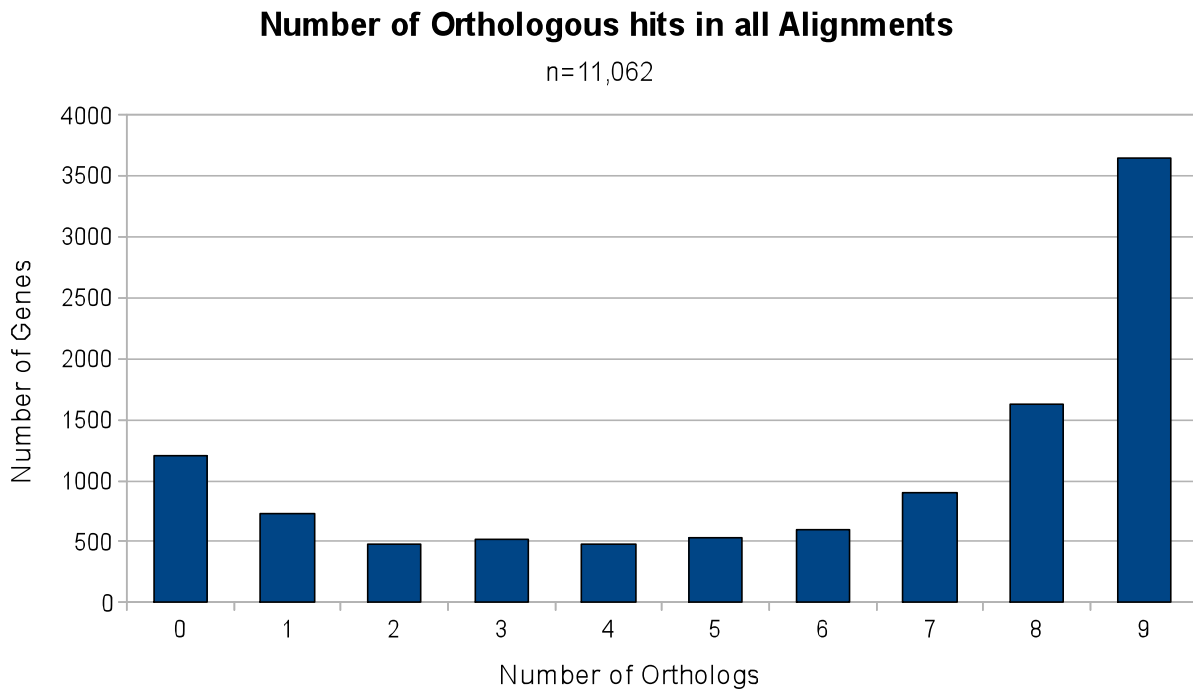


Figure 3.10 PAML Pipeline coded in Perl (dir=directory; chdir=change directory).

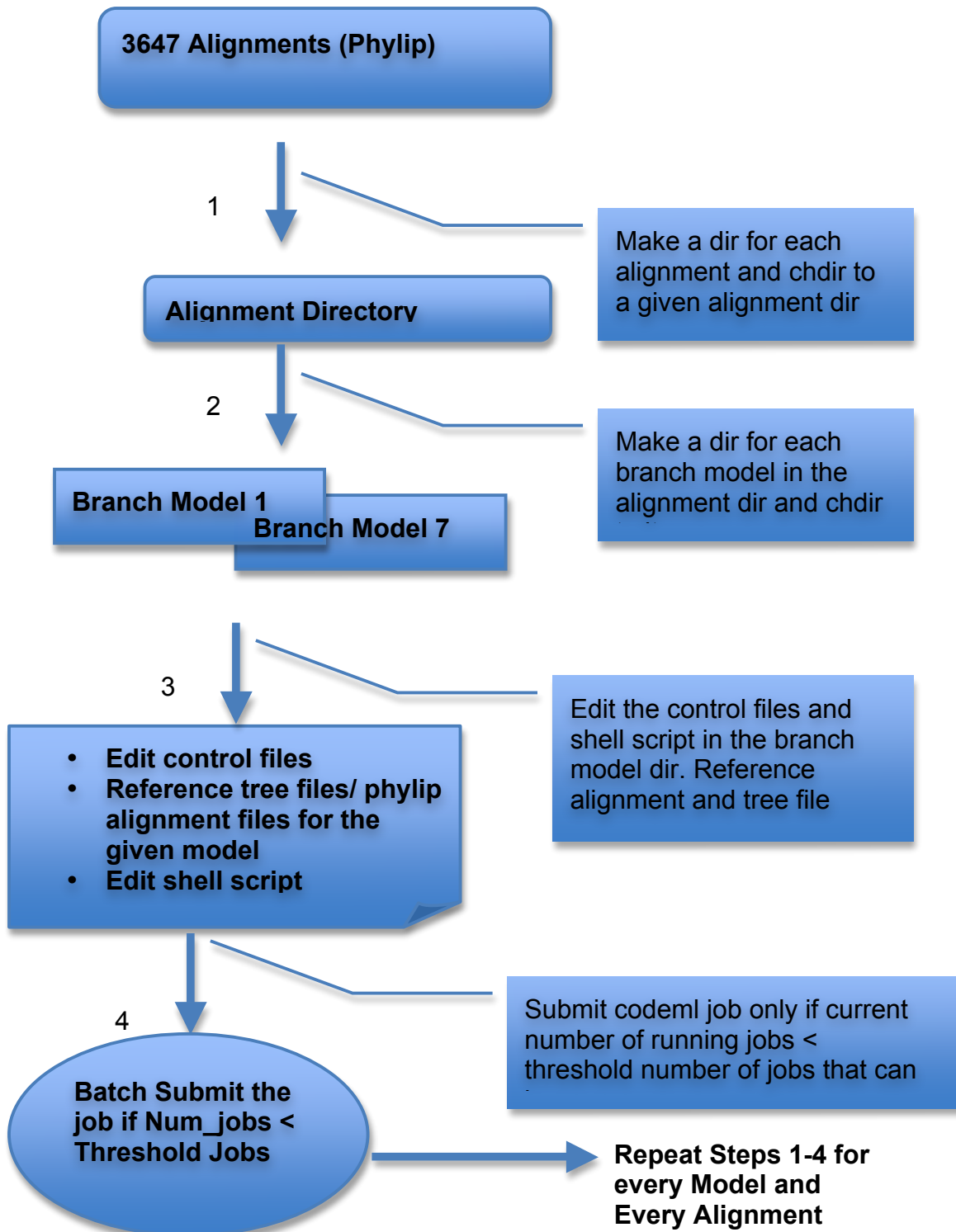


Figure 3.11 Branch model results showing number of rapidly evolving genes for each of the three different hypothesis tested. Each hypothesis contrasts the lineage versus other lineages across the phylogeny. The null model assumes one rate of evolution for all the branches of the phylogeny, while the alternative model assumes different rates of evolution for specific branches across the phylogeny. The Likelihood ratio test followed by a chi-square analysis is used to pick out rapidly evolving genes.

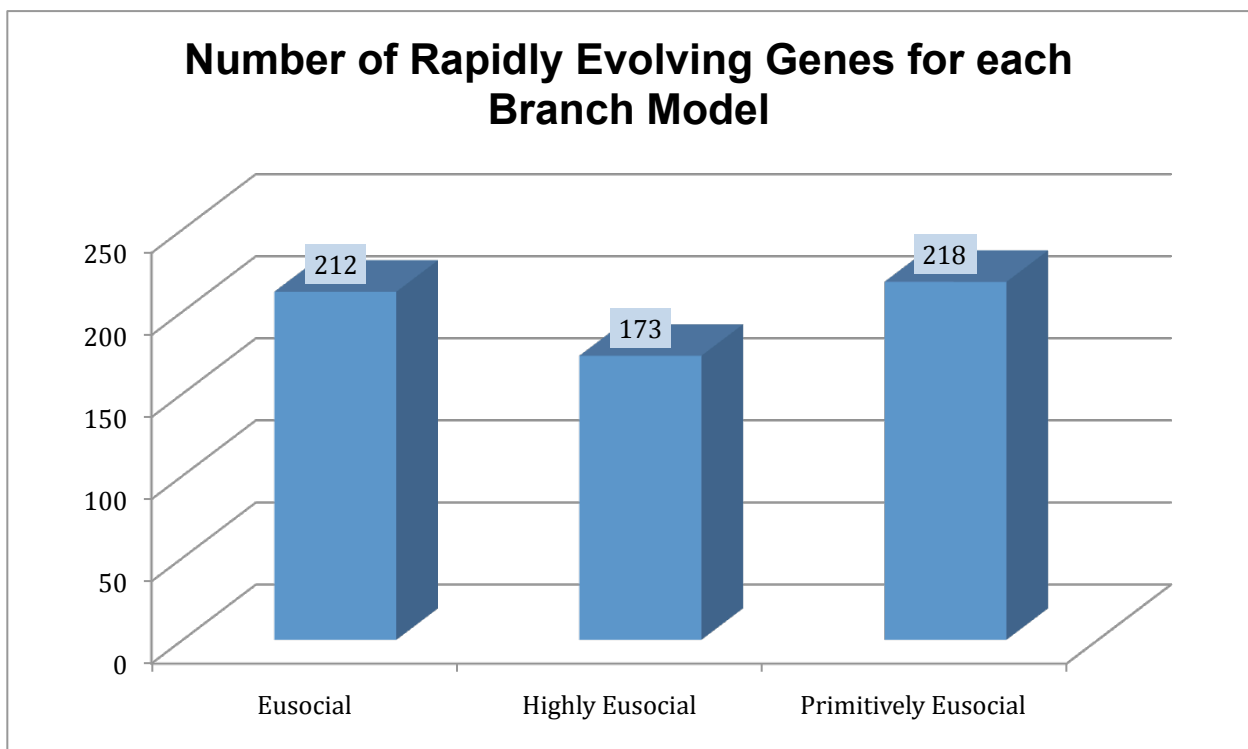


Figure 3.12 The gene lists obtained from each hypothesis were compared to remove the overlaps, and pick out the lineage specific genes that are rapidly evolving for that given hypothesis (Robinson lab).

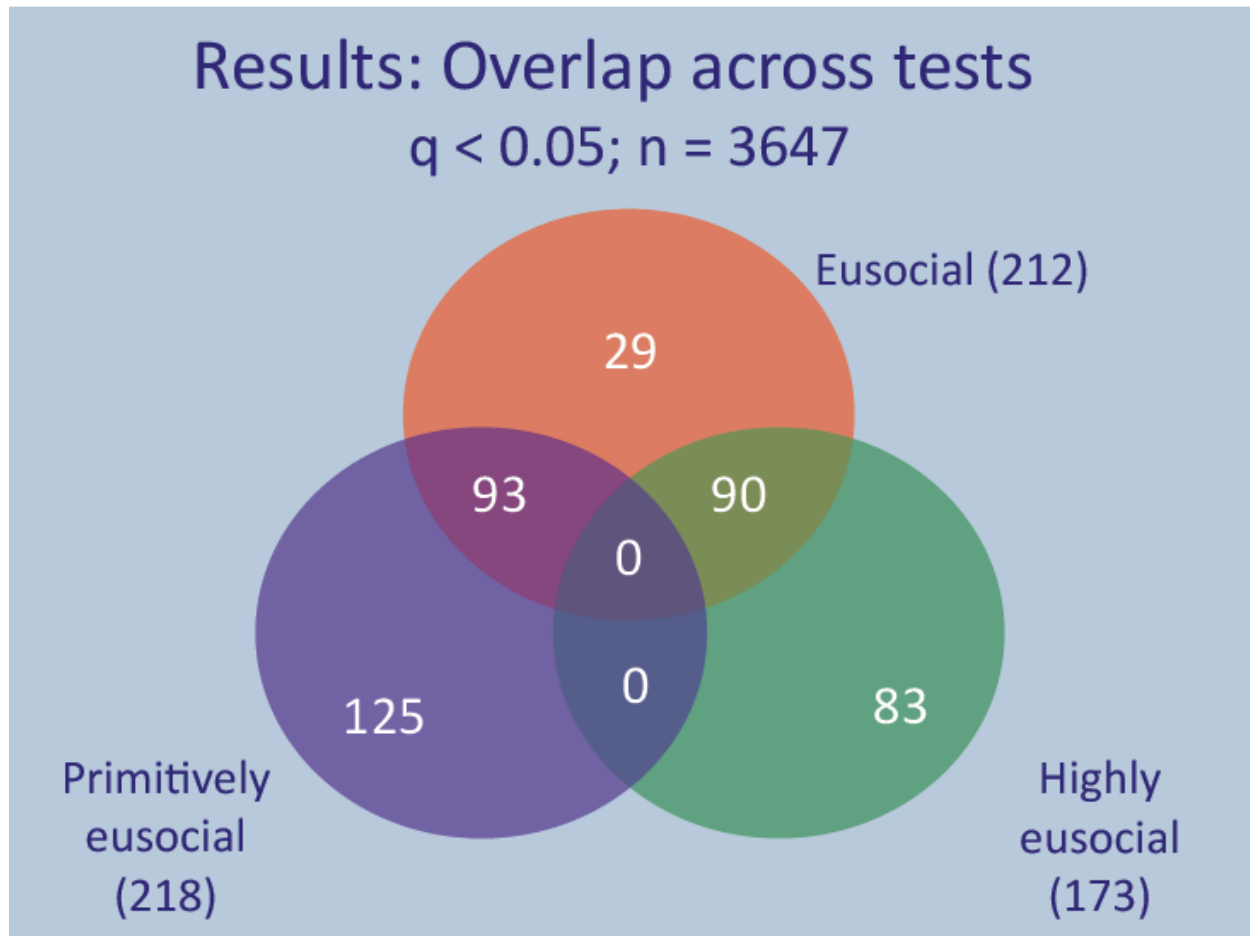


Figure 3.13 Shows the comparison between the raw and corrected read counts used for *Bombus impatiens* assembly. Reads from three libraries (500 bp shotgun, 3kb and 8kb mate pairs) were error corrected using Quake. Custom Perl script was written to order the mates into pairs and singletons. The *pairedCorrected and *singletons were then used for the assembly. Y-axis shows the number of reads.

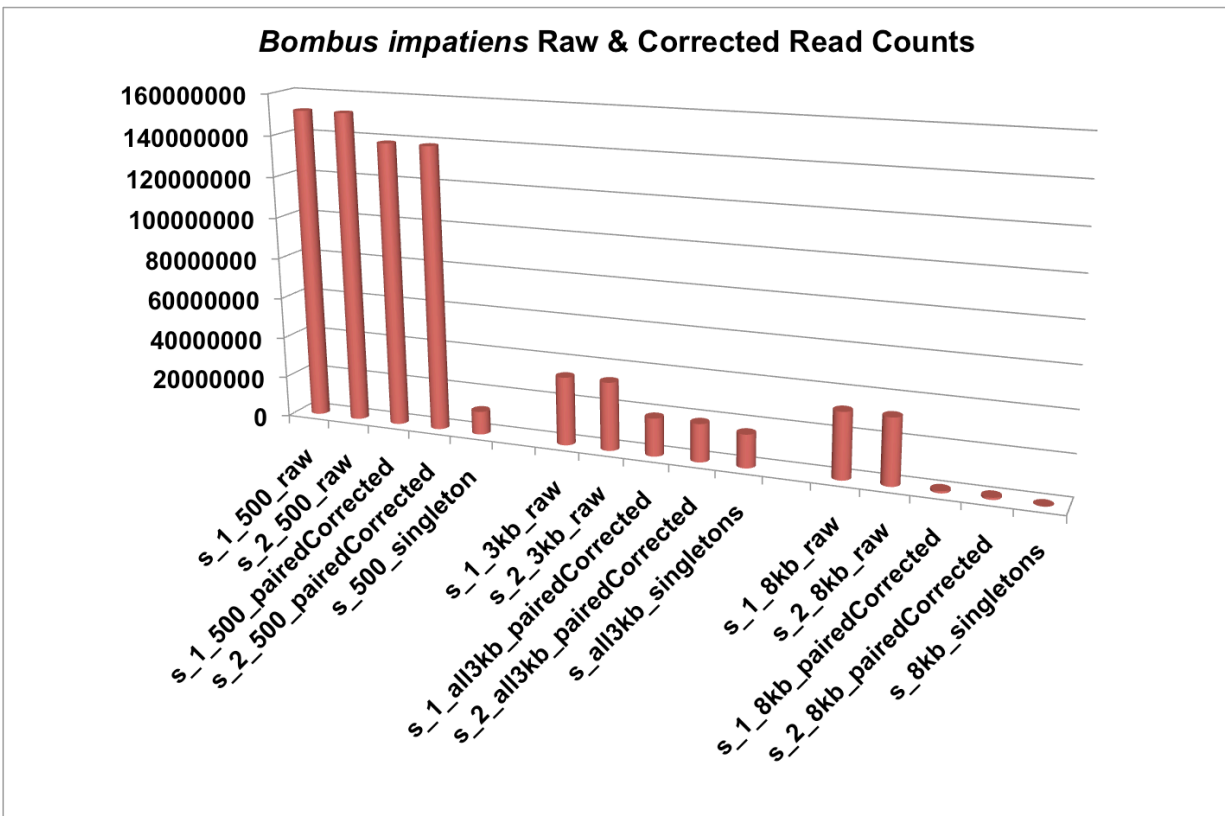
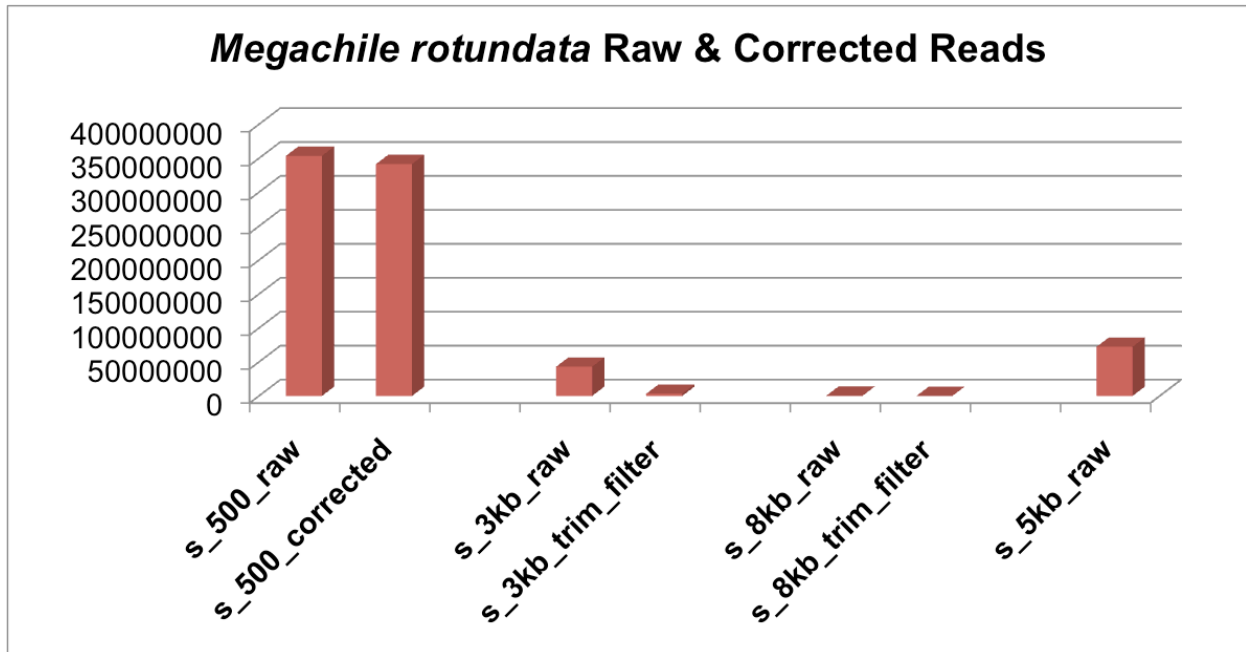


Figure 3.14 Shows the raw and corrected read counts for the *Megachile rotundata* dataset. Reads from four libraries were sequenced: 500 bp shotgun, 3kb, 8kb and 5kb mate pairs. The 500 bp library prepared from a haploid male (s_500_raw) was corrected using Quake (s_500_corrected). The 3kb and 8kb libraries were prepared from a pool of bees, hence were not error corrected. However these were trimmed for linkers and filtered to retain reads \geq Kmer length using custom Perl scrips. The 5kb library was also prepared from a pool of individuals and the raw reads were used for the final assembly. Y-axis shows the number of reads.



TABLES

Table 2.1 Standardized species names. All sequences in the BlastData database are represented by their corresponding standardized abbreviations.

Species	Standardized species name
<i>Apis florea</i>	<i>Api_florea</i>
<i>Bombus impatiens</i>	<i>Bom_impati</i>
<i>Bombus terrestris</i>	<i>Bom_terres</i>
<i>Eulaema nigrita</i>	<i>Eul_nigrit</i>
<i>Exoneura robusta</i>	<i>Exo_robust</i>
<i>Megachile rotundata</i>	<i>Mega_rotun</i>
<i>Euglossa cordata</i>	<i>Eug_cordat</i>
<i>Frieseomellita varia</i>	<i>Frie_varia</i>
<i>Melipona quadrifasciata</i>	<i>Mel_quadri</i>

Table 3.1 Bee EST Assembly using Phrap. Deduplication of clonal reads was done to reduce time taken for the assembly. Many genes are likely represented by multiple contigs/singlets. Average read length was around 240 bp (Varala, K, Hudson Lab).

	Total Bases (~Mb)	No. Of Reads (~ Kb)	No. of contigs	Avg. contig length	#Singlets	#Non redundant bases (~Kb)	Total bp in NR (~Mb)
<i>Bombus impatiens</i>	98	406	30722	556.5	23820	54	22
<i>Megachile rotundata</i>	48	559	13725	592.2	30945	45	15
<i>Euglossa cordata</i>	77	317	26376	560	23454	50	19
<i>Frieseomellita varia</i>	74	307	21052	476.4	29757	51	16
<i>Melipona quadrifasciata</i>	77	317	24797	530	29728	54	19
<i>Apis florea</i>	72	331	28418	464	30592	59	19
<i>Bombus terrestris</i>	76	319	19938	528.6	22878	42	15
<i>Eulaema nigrita</i>	89	376	29509	539.4	27180	57	21
<i>Exoneura robusta</i>	117	421	37791	531.8	22856	61	26

Table 3.2 Shows the GO annotations (over represented terms) specific to the eusocial lineages. Hypothesis tested: Genes in the eusocial lineages are evolving more rapidly than the non-eusocial lineages.

BIOLOGICAL PROCESS	FOLD ENRICHMENT
Cell surface receptor linked signal transduction	1.89
Gland Development	3.00
Protein phosphorylation	2.07
Glycolysis	6.66
RNA Processing	1.73
Transcription	1.70

Table 3.3 Shows the GO annotations (over represented terms) specific to the highly eusocial lineages. Hypothesis tested: Genes in the highly eusocial lineages are evolving more rapidly than the other lineages.

BIOLOGICAL PROCESS	FOLD ENRICHMENT
Glycolysis	14.81
Oxidation Reduction	2.1
Protein phosphorylation	2.6
Carboxylic acid biosynthesis	4.27

Table 3.4 Shows the GO annotations (over represented terms) specific to the primitively eusocial lineages. Hypothesis tested: Genes in the primitively eusocial lineages are evolving more rapidly than the other lineages.

BIOLOGICAL PROCESS	FOLD ENRICHMENT
Histone modification	3.48
Motor activity	4.00
Neuron differentiation	2.50
Post embryonic development	1.86
Response to hormone stimulus	5.00
Transcription	2.27

Table 3.5 *Bombus impatiens* whole genome assembly, Contig Statistics.

NUMBER OF CONTIGS (length > 100)	SUM	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	N50	N90
97971	232 Mb	100 bp	106 kb	2376 bp	7.8 Kb	1.5 Mb

Table 3.6 *Bombus impatiens* whole genome assembly, Scaffold Statistics.

NUMBER OF SCAFFOLDS	NUMBER OF SCAFFOLDS AND SINGLETONS	SUM OF SCAFFOLDS AND SINGLETONS	MAX LENGTH	AVERAGE LENGTH	N50	N90
2450	9359	260 Mb	4.9 Mb	27998 bp	1.2 Mb	148 Kb

Table 3.7 *Megachile rotundata* whole genome assembly, Contig Statistics.

NUMBER OF CONTIGS (length > 100)	SUM	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	N50	N90
207045	239 Mb	100 bp	101 kb	1158 bp	3.6 Kb	810 bp

Table 3.8 *Megachile rotundata* whole genome assembly, Scaffold Statistics.

NUMBER OF SCAFFOLDS	NUMBER OF SCAFFOLDS AND SINGLETONS	SUM OF SCAFFOLDS AND SINGLETONS	MAX LENGTH	AVERAGE LENGTH	N50	N90
21843	60414	274 Mb	1.1 Mb	4641 bp	31 Kb	5 Kb