Upgrading the honey bee genome sequence

The Honey Bee Genome Sequencing Consortium[A]

[A]Represented by the authors of this document: J.D. Evans[*], M. Beye, C. Elsik, R. Maleszka, H.M. Robertson, G.E. Robinson, D.B. Weaver, C.W. Whitfield

[*]Contact information: Ph. 301-504-5143; Fax 301-504-8736 Email: evansj@ba.ars.usda.gov

Summary

Honey bees (*Apis mellifera, A.m.*), insects endowed with great cognitive and social abilities and amenable to molecular, genetic, neural, and ecological manipulation, provide an important model for understanding and improving human health. The Honey Bee Genome Project (HBGP) has successfully organized a large and diverse research community around the bee model. With 7.5x genomic sequence coverage and a robust assembly carried out at the Baylor College of Medicine NHGRI Sequencing Center, and gene-prediction strategies based on orthology, transcript evidence and de novo models, *A.m.* is able to fill a central role in research on diverse issues related to behavior, development, reproduction, and immunity. More importantly, HBGP has united a broad range of scientists, from leaders in human genomics and bioinformatics to ecologists. This inter-disciplinary group has already started to generate unexpected insights into human health and genetics. Analysis of the bee genome has just begun with publications planned for late '05. Our analyses have identified pressing needs, and we propose additional sequencing to improve the *A.m.* genome in the following three ways. 1) Targeted sequencing of critical genome regions using libraries biased toward specific isochores. 2) To fill knowledge gaps for the transcriptome and better exploit honey bee sequence variation, EST screening of libraries from several important developmental stages and tissues, from both *A.m.* and the Africanized "Killer" bee, *A.m. scutellata* along with minimal (0.2X) coverage of the Africanized bee genome. 3) 2X sequence coverage for three additional bee species, chosen to form a ladder of genetic distance between *A.m.* and other insect genomes: the Asian honey bee *A. dorsata*, the bumble bee *Bombus terrestris* and the alfalfa leafcutter bee, *Megachile rotundata*. These species differ genetically from *A.m.* at levels well suited for comparative inference techniques for identifying and validating predicted *A.m.* genes and regulatory elements and micro- and macrosynteny analysis. They also represent the full range of sociality, from solitary to highly social. A key need is to explore the validity of thousands of *A.m.* gene models, most commonly in areas of GC-richness, that appear to have no sequence similarity to genes from other eukaryotes. These genes may well reflect the impact of sociality and/or haplodiploidy on the genome. Together with *A.m.*, this suite of species will provide a strong foundation for understanding how haplodiploidy and social evolution affect genome structure, function, and organization. The results of these efforts, and the complementary proposal to sequence the genome of the fire ant, promise to dramatically increase the value of the HBGP for comparative genomics and biomedical research.

Table of Contents

A. Specific biological rationale for utility of new sequence data

1. Suitability of bees as a model for human health. The White Paper for a Honey Bee Genome Project submitted in 2002 (http://www.genome.gov/11008252) discussed how sequencing the bee genome will benefit human health and medicine in diverse areas, including venom toxicology, allergic disease, mental illness, infectious disease, parasitology and gerontology; improve human nutrition by enabling enhanced pollination of food plants and accelerated delivery of hymenopteran parasitoids for biological control of pests; and improve sentinel function for detection and location of chemical and biological agents of harm. We do not repeat all of these arguments here, but instead highlight developments that have occurred since the original submission and how they relate to this proposal.

1a. Instincts and mental health. The original White Paper (HBGPWP) outlined the attributes of bee society that make it a compelling model for understanding sociality in general. "The societies of honey bees and other social insects occupy Wilson's[3] second "pinnacle of social evolution," with complexity that rivals our own." Recent studies—enhanced by genome sequence--are starting to implicate genes in complex social instincts, using microarray, QTL, RNAi, and pharmacological analysis[4-8]. We predict that interest and productivity in this area are poised for an explosive increase, but are dependent on having a better annotated genome than we now have. Better annotation is necessary to facilitate going from QTL to gene, an especially important issue for getting beyond well known candidate genes from *Drosophila* to truly novel findings, including those expected from QTLs for several complex social behaviors[9-12]. Understanding the regulation of behavior by elucidating how nature/nurture interactions act at the molecular level is a pressing question in human biology[13]; with a sequenced and annotated genome, *A.m.* will deliver important answers.

1b. Cognition. Bees display "vertebrate-like" cognitive abilities, with a brain with only 4X > neurons than *Drosophila*. They are excellent at associative learning, based on the need to associate a color, shape, scent, or location of a flower with a food reward. Recent studies have shown that *A.m.* also can learn abstract concepts such as "similar" and "dissimilar," and are able to negotiate complex mazes by using visual stimuli as abstract "signposts" or by recognizing path irregularities[14-16]. A set of candidate genes for behaviors representing diverse signaling pathways underlying these impressive abilities has recently been identified[17]. Finding these highly conserved molecules suggests that insights from *A.m.* will play an important part in bridging the chasm between genotype and behavior[8]. These studies and others suggest that *A.m.* can be used to identify drug-sensitive genes and networks critical for disease phenotypes, a major challenge in pharmacogenomics. Such questions require an excellent annotated genome and cDNA reagents to identify both known and novel neurotransmitters. These issues are important for understanding human cognition and the treatment of neurodegenerative disease.

1c. Immunity and disease. Innate immune responses are major players in human disease as both a first-line defense and against secondary infections in injured or immuno-compromised individuals[18]. These responses are remarkably conserved across the eukaryotes[19-23]; insect models play a prominent role in understanding immunity. An understanding of insect immune responses also helps in the design of controls against undesirable insects, and in their mitigation as vectors of disease[24-26]. Honey bees are an excellent model for elucidating immune system function; the natural pathogens and parasites of honey bees are well known because they cause substantial economic loss[27]. Research in this area has started to take off[28-31]; there are currently over 20 collaborators annotating the *A.m.* genome for genes involved with immunity but they have found orthologs for only ca. 60% of the *Drosophila* and *Anopheles* genes implicated in the two primary immune-response pathways. In addition, there appear to be fewer paralogs for gene families

related to immune recognition, signal transduction, and effectors in the bee genome than in *Drosophila* or *Anopheles*. We believe that many important immunity genes, and arguably novel gene families, are absent from the current annotation. All three of the proposed project goals will help improve this situation.

1d. Developmental biology. Social insects are known for their striking developmental polymorphisms, which will help answer fundamental questions related to reproduction, nutrition and growth rates, and aging, and provide insight into how genes and the environment interact during development. The best studied case is the honey bee queen/worker polymorphism. Female larvae develop into queens or workers on the basis of larval nutrition and endocrine signaling. Earlier studies [32-34] suggest that finding the most important genes cannot be achieved by searches for orthologous genes in *Drosophila* and other models. Genetic approaches, such as those underway for the pea aphid (a non-social insect with a comparably striking polymorphism[35]) will be needed, supported by improved annotation. One genome-enabled pathway that is receiving great interest involves honey bee sex determination, the centerpiece of which is the *complementary sex determination (Csd)* gene. In a landmark study recently published in *Cell*, a team of honey bee genome consortium members led by White Paper co-author Martin Beye implicated this gene in the ability of bees to determine sex based solely on ploidy level[36]. Female bees are typically diploid while males derive from unfertilized eggs and begin development with a haploid genome, and *csd* is the first gene identified in a eukaryote involved in regulating haplodiploidy. On top of having interesting implications for dosage compensation, imprinting, and other fundamental questions of developmental biology, haplodiploid-induced asymmetries in relatedness between social insect offspring and sisters have been crucial in the development of one of the most prominent theories of social evolution, kin selection[37,38]. *csd* was identified by positional cloning, which was aided immensely by the extremely high *A.m.* recombination rate[9,39], then was validated by RNA interference. Current efforts to understand *csd* and downstream pathway members will require more complete coverage of the area around *csd* (which ranges from 1X to 4X coverage) and around other candidate pathway members.

1e. Gene regulation. *A.m.* offers two distinct avenues for better understanding gene regulation in humans. First, *A.m.* is an intriguing model for dosage compensation, an important feature of development in humans as well as in understanding disease-causing chromosomal ploidy mutations. Since *A.m.* is haplodiploid, each chromosome is effectively an X-chromosome, i.e., one copy in the male and two copies in the female. During development, many but not all cells in both male and female bees undergo chromosomal endoduplication[40]. These studies will complement similar analyses in other (haplodiploid) Hymenoptera: *Nasonia vitripennis* (accorded High Priority for sequencing), the three additional species in this proposal and the fire ant (White Paper submitted). Second, *A.m.* is an excellent model to study large-scale coordinated changes in gene expression; the type long suspected to regulate complex phenotypes. Recent microarray studies have demonstrated just such large-scale coordinated changes in gene expression[41-43]; for example 40% of the genes on a cDNA microarray show differential regulation in the bee brain when comparing two behavioral states: brood care and foraging[41]. An improved genome assembly coupled with DNA-DNA alignments with the proposed species should help identify members of two key components of gene regulation, conserved microRNAs[44-47], and cis regulatory elements. These analyses will be both "blind" with respect to known regulatory sequences, in order to discover statistically over-presented short DNA stretches in the *A.m.* genome associated with coregulated genes, and directed by known DNA regulatory elements, an approach recently used with the current genome data to find predicted Nf-kappaB binding sites upstream from *A.m.* genes encoding antimicrobial peptides[48].

1f. Gerontology. Honey bee queens and their workers have identical genotypes but queens live two orders of magnitude longer[49]. Moreover, this difference is natural; while it has been relatively easy to select for extended longevity in *Drosophila* and other laboratory models, long-lived strains have not been observed in the wild[50]. Identification of genes responsible for these striking lifespan differences has important implications for human longevity and aging; these issues are beginning to be addressed by Consortium members. Preliminary results implicate genes related to respiration efficiency[51] and the insulin pathway[52]. These studies will be enhanced with sequence information from two of the proposed additional species; their queen-worker differences in longevity are either reduced (*B. terrestris*) or nonexistent (*M. rotundata*). *A.m.* also offers lessons with respect to sperm longevity; sperm remains viable and active in queens for up

to several years[53]. An additional goal of uniting more genes with their upstream regions (through an improved assembly) will be to find regulatory sequences associated with genes that mitigate the effects of aging.

1g. Comparative genomics. Insects provide outstanding material for comparative genomics because they are so old and so diverse. This great potential has been recognized by NHGRI and elsewhere, resulting so far in genome sequencing projects for multiple *Drosophila* and mosquito species, the beetle *Tribolium castaneum*, the silkmoth *Bombyx mori, Nasonia*, and the pea aphid (*Acyrthosiphon pisum*). Insect genomes promise to provide insights into the genomic mechanisms that generate diversity of lifeforms and that maintain deep conservation of biological processes. This proposal, and the fire ant White Paper, provide a powerful set of species to address issues related to sociogenomics[17]; how genes affect the diverse set of biological processes that are subject to social influence and how sociality in turn affects all aspects of genome structure, function, and organization. Fire ants show advanced eusociality, like *A.m.* and *A. dorsata*, but they represent a completely independent evolution of eusociality whereas *A. dorsata* and *A.m.* are part of a monophyletic clade[1]. *A. dorsata* differs from *A.m.* by nesting outside of cavities, a fact that affects social organization[54]. These two species also differ in susceptibility to pests and pathogens[55-57]. *B. terrestris* is primitively eusocial but it is in the same family as *A.m.*. *M. rotundata*, in the same superfamily as honey bees and bumble bees, is completely non-social. No other animal taxon besides the bees shows the complete sweep of sociality –from solitary to highly eusocial. This proposal thus not only promises to upgrade the genome of an important model species but to provide an initial framework for powerful comparative genomic analyses of social evolution. Both *terrestris* and *rotundata* also have other attributes that make them compelling choice for economical low-coverage sequencing, as follows.

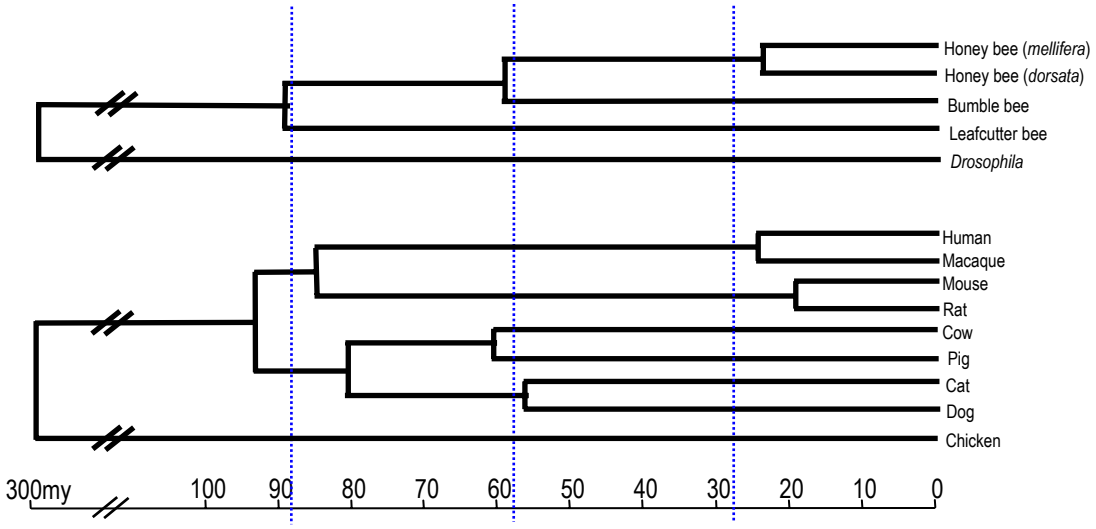

Figure 1. Phylogenetic relationships and divergence dates for a) proposed bee species and *Drosophila* outgroup and b) human and other vertebrate species. Data from [1,2]

Both *terrestris* and *rotundata* have annual lifecycles (unlike *Apis*), and a facultative diapause. The comparable "dauer" stage in *C. elegans* has proven an excellent model for studies of aging[58]. Both *terrestris* and *rotundata* are sold commercially because of their pollination value, so methods of controlled rearing are well established. *terrestris* has the queen/worker polymorphism described above (as does *Apis*) but *rotundata* does not, providing an important "outgroup" for studies of developmental plasticity. Both *terrestris* and *rotundata* have well studied pests and pathogens. Bumble bees comprise one of the best studied social insect groups genetically, next to the honey bee and have been used as a model species with respect not only to social behavior, but also for population and community ecology, parasitism, immunology, life history, physiology, reproduction, sex determination or pollination ecology. *B. terrestris* has been used to describe several key elements of insect immunity, tolerance, and resistance to pathogens and

parasites[59-61], including novel findings on transitive immune defense passed from queen to offspring[62] and parasite-induced changes in behavior and life history[63,64]. *Bombus* also shares some pests and pathogens with *Apis*, and thus provides excellent comparative material. Finally, genetic maps in *B. terrestris* indicate a substantially slower recombination rate than that of *A.m.*, a find which could help test hypotheses for why *A.m.* is the leading eukaryote in this trait[65]. Solitary bees such as *M. rotundata* comprise the vast majority of the 3750 recognized bee species worldwide, and play major roles as pollinators in both natural and agricultural plant communities[66]. Bumble bees are increasingly important for pollination of greenhouse crops, an agricultural need for which *A.m.* is poorly suited[67,68] while *rotundata* is a primary pollinator of many seed crops[69]. *B. rotundata* can provide a better understanding of foraging behavior and learning without the complications of sociality.

The proposed species represent estimated evolutionary distances that roughly match some of the important splits in mammalian lineages that have been exploited for comparative genomics purposes (Figure 1). *Mellifera-dorsata* is close to mouse-rat (18 my) and human-old world primate (25 my), *Apis-Bombus* is shorter than human-rodent (60 vs 75 my) and similar to cat-dog, and *Apis-Megachile* farther than human-rodent but much less than human-other vertebrates. While *Drosophila's* molecular clock apparently is faster than for vertebrates[70], it is not obvious that this is the case for *A.m.*. Alignments (cDNA) for two genes (*Ef-1a* and *rhodopsin2*) show sequence-level (non-indel) differences of 4-6% for *mellifera-dorsata* and 11-20% for *mellifera-Bombus* and *mellifera-Megachile.* These species thus provide an optimal ladder with which to better annotate and understand the *A.m.* genome using both searches for conserved non-coding DNA elements[71-73] and gene-centric searches tuned to identify novel genes based on mutational variation such as TWINSCAN[74-76] and SGP2[77].

B. Strategic issues in acquiring new sequence data

1. Current resources in honey bees. Table 1 provides an overview, highlighted by great advances generated by HBGP, primarily through genomic sequencing and assembly at the Baylor College of Medicine NHGRI Genome Sequencing Center. There is an assembled genome from 7.5X sequence coverage (23X clone coverage) consisting of 231 MB, 71% in mapped scaffolds. Scaffold N50 = 362 Kb, contig N50 = 26 kbp. 98% of the known honey bee cDNA's and EST's are in the assembly. However, close scrutiny of many of these genes (as conducted by co-author C. Elsik and members of the annotation list, Appendix 1) suggests that >10% of these genes have exons in intra or inter-scaffold gaps and many more are missing nearby upstream regions. Moreover, > 40% of the predicted genes (4200 out of ca. 10,000 orthology-based predictions from Ensembl and other major pipelines, consolidated by "GLEAN," A. Mackey, Univ. Penn.) are not assigned to a chromosome, but instead are in several thousand ungrouped contigs (median size 2.7 kbp). Detailed analyses by Elsik, to be described below, reveal one problem: the bee genome is unexpectedly complex, with a large AT-rich region that is also gene rich, and there is a relative paucity of sequence reads in the AT-rich region. *These limitations will severely limit research on the above topics.*

Database development proceeds aggressively to make maximum use of sequence information. BeeBase is a dedicated analysis and display environment for the honey bee genome (C. Elsik, Texas A&M Univ., PI), which will be closely tied to FlyBase and will be a "spoke" in the planned InsectBase (W. Gelbart, Harvard Univ.). Other databases include: NCBI Honey Bee Genomic Resource, ENSEMBL, EBI-Heidelberg, UC Santa Cruz, US-DOE, and the central site at BCM-HGSC, which also offers sequence data and assemblies for two key *A.m.* pathogens, *Paenibacillus larvae* and *Ascosphaera apis*. BeeSpace (Univ. Illinois) is a 5-year, $5M project (NSF Frontiers in Biological Research Program) for information scientists and biologists to leverage the bee genome to create a new information environment for the study of social behavior (http://www.beespace.uiuc.edu/). The HBGP has united a broad range of scientists, from leaders in human genomics and bioinformatics at BCM and elsewhere to members of diverse disciplinary and organism-based communities, including those studying mammals and humans; 112 individuals in 63 institutions around the world have signed on to analyze the newly available bee genome sequence. New genomic resources are being created in collaboration with industry leaders, government labs, and academia. For example, a leading genomics lab at NASA-Ames (Viktor Stolc, PI) is performing a genome-tiling experiment using state-of-the-art technology to explore the extent and location of transcriptional units

in the bee genome. This was prompted by the intriguing *in silico* prediction (FGENESH, Softberry, Inc.) of thousands of genes that show no orthology with any other known genes. An important goal of this proposal is to generate resources to make it possible to determine whether they are orthologous, novel, or artifacts.

Table 1. Available Resources	Prior to HBGP (2002)	Since HBGP
Sequenced fraction of genome	< 3%	~ 95%
Complete cDNA's in GenBank	9	120
Predicted genes	< 20	11,000 (NCBI, ENSEMBL predictions)
Estimated gene count	14,000 (inferred from flies)	16,000+
Gene expression	Microarray from bee brain (6,200 genes)	Whole-genome oligo array in production (already funded), several microarray studies
Tiled genome array	-	In production at NASA-Ames for gene validation
BAC libraries	TAMU-Baylor (DeJong) 25x clone coverage, 3x sequence coverage, Purdue (Clemson[78])- - 15x, end-sequences	-
Microsatellite markers	120	2100+
Transgenesis	Sperm-mediated	-
Gene inactivation	Single RNAi example	Systemic RNAi, multiple life stages
Mutants	Many morphological	-
Germplasm storage	Short-term sperm, nuclei transfer	-
Sex determination	Linkage maps	Identified locus, proposed pathway
Genomes of bee pathogens	RNA virus (1)	RNA viruses (7); bacterial pathogen *Paenibacillus larvae**, fungal pathogen *Ascosphaera apis** (*7X drafts by BCM-HGSC)
Public web resources	Gene-expression data, EST resource (UIUC) linkage maps, (UC-Davis, Purdue)	BCM-HGSC, Beebase (TAMU), BeeSpace (Univ. Ill), NCBI Genome Resources, EBI
Genome Browsers	-	Beebase, UCSC, ENSEMBL, LBL

2. Demand for honey bee genome sequence. The honey bee community numbers over 150 laboratories worldwide. There are 4369 peer-reviewed articles on *A.m.* in the past 10 years (Scopus index), with 168 authors publishing ≥ 10. About 15% of this literature has an explicit genetic basis, although much *A.m.* research is strongly cross-disciplinary and addressed jointly by molecular and non-molecular strategies. A special feature of the bee community is its explicit organismal focus; consequently a great deal is known about bee learning, nutrition, development, reproduction, and disease resistance. This information promises to add context and meaning to molecular information from this next stage of the HBGP. Achieving this synergy is an explicit goal of the NSF BeeSpace project described above. There are ca. 50 laboratories currently focused on molecular analyses of honey bees, and ca. 112 individuals in 63 institutions currently engaged in annotation of the bee genome (Appendix 1). Of these, nearly half are new to bees and were attracted specifically by the potential of the new genome-enabled biology now possible. Laboratories that study *A.m.* are joined by a much larger group of researchers interested in other social insects, parasitoid wasps, insect genetics, and the broad community of comparative genomics, especially those studying *Drosophila* and mosquitoes. HBGP has enabled the bee community to forge close ties with the *Drosophila* community (Appendix 1).

3. Rationale for improving upon the current sequence. Most recognized *A.m.* genes are included in the current assembly (version 3.0, 6/1/05) but they are found in 442 mapped scaffolds (71% of the genome) as well as in ca. 9432 unmapped, smaller, scaffolds (N50 = 33 kb, median = 2.7 kb). The assembly is

excellent, given the total number of reads (3M wgs) and the complexity of the genome. This unexpected complexity has negatively affected the assembly statistics; and a substantial fraction of homology-based gene models (4184/9157) have not been mapped to chromosomes. These problems cannot be fully resolved computationally. While an extensive effort to use superscaffolding based on gene-centric information (led by Richa Agarawal, US-NIH-NCBI) has proposed joining of >1000 scaffolds, additional sequencing and new genome and map resources will be critical to join unmapped scaffolds and their associated genes to mapped regions of the genome. The complexity of the bee genome is vividly seen in the following analysis. Contigs in the smaller, unmapped (GroupUn) scaffolds of Assembly 3.0 show approximately 75% AT content, while those in mapped (generally longer and with higher average fold coverage) scaffolds approach 65% AT. A global analysis of GC-content patterns (using[79]) in the current assembly reveals an AT-rich shoulder below 30% GC (Figure 2A). The distribution of homology-based predicted genes with respect to segment GC content (Figure 2B) shows a substantial fraction of genes in this AT-rich component. Separating the mapped (Figure 2C) from unmapped scaffolds (Figure 2D) shows that the AT-rich shoulder is missing in the mapped scaffolds but enriched in the unmapped scaffolds. Repeating the segmentation analysis with thresholds that vary granularity has indicated that the AT-rich shoulder is not due to artificially short segments in the unmapped scaffolds (data not shown). The data suggest that the difficulty in anchoring the unmapped scaffolds arises from the lack of suitable markers in AT-rich regions as well as the small size of unmapped scaffolds (Figure 2F, compare to mapped scaffolds Figure 2E). The data also suggest that one effective strategy for improving the bee genome will be targeted AT sequencing, as described below.

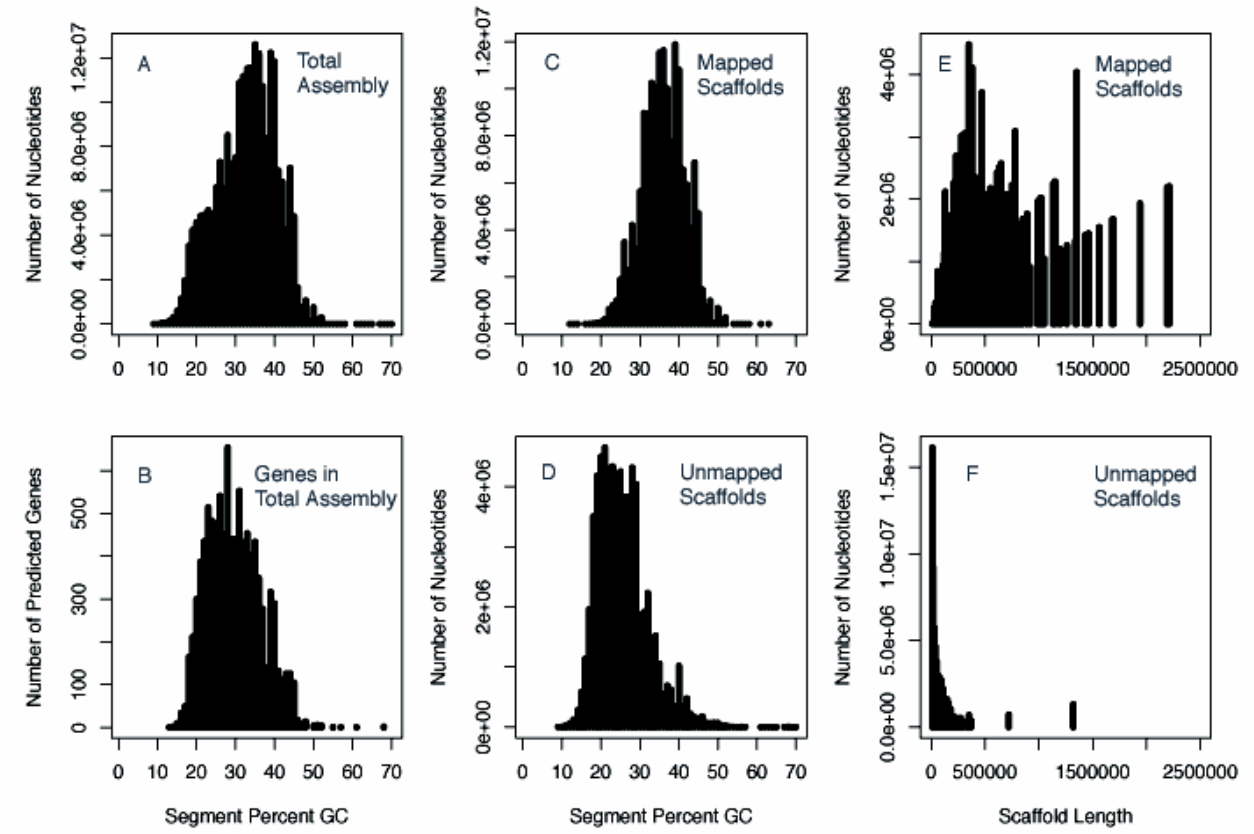

Figure 2. A, C, D. Number of nucleotides in GC segments of the total assembly, mapped scaffolds and unmapped scaffolds, respectively. B. Number of predicted genes found in GC segments. E, F. Number of nucleotides distributed by length of mapped and unmapped scaffolds, respectively. Scaffold lengths grouped into bins of 10,000. (C. Elsik, unpub.)

We propose a cost-effective strategy to improve significantly upon the current assembly. We also propose additional sequencing activities involving ESTs and outgroup informant bee species that will help critically with gene annotation, an important roadblock for a species 300 my from the nearest neighboring model species. Sequence from these species is expected to add a significant number of currently uncalled genes to the *A.m.* gene list, and should extend up to half of the current gene prediction models. Most importantly, informant species will test the evolutionary conservation of a huge set of unique gene models, found especially often in GC-rich isochores of the bee genome (e.g.,[11]). These gene models, which have no significant sequence-level homology to known eukaryotic genes yet seem to contribute transcripts to honey bee RNA pools, would make up a substantial fraction of the bee transcriptome if validated, providing a new frontier for eukaryotic genetics. The added sequence information will help improve upon the current assembly by uniting scaffolds into gene-based super-assemblies, a strategy that is becoming routine in genome sequencing projects and that is well suited to building on the coverage level anticipated here.

4. Overview of the three project goals

4a. *A. mellifera genomic sequencing.* We propose to use additional genomic sequencing to better characterize thinly covered genome regions in the current assembly. Specifically, we propose continuation of a strategy used for the last stages of sequencing for HBGP, whereby short-insert genomic libraries were established from DNA that had been fractionated so as to be biased toward higher AT%. During HBGP, 1M reads from this library effectively doubled the contig N50, and greatly improved the evenness of assembly coverage (Fig. 3).

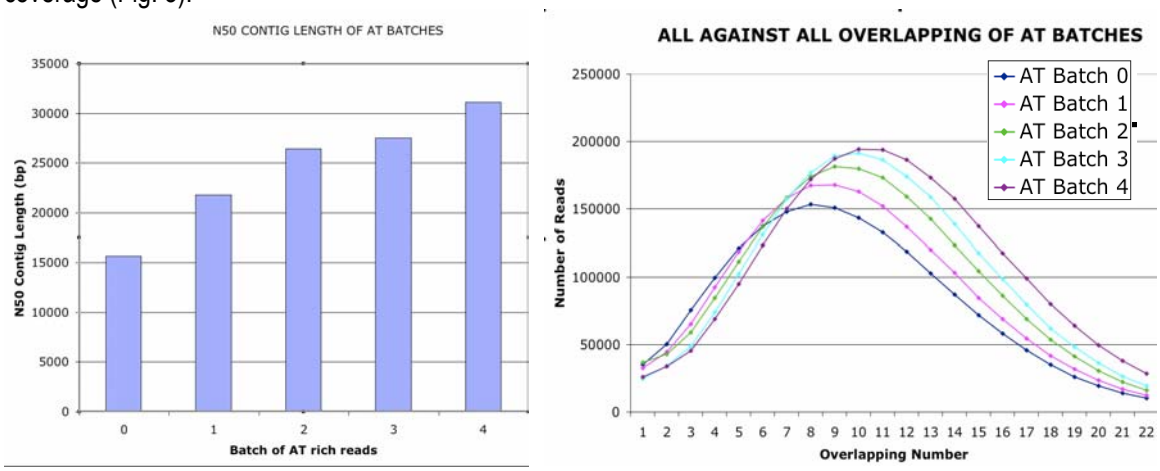

Figure 3. Changes in a) contig N50 and b) overlap density with successive addition of 200K reads from an AT-biased library comprising part of Honey Bee assembly 2.0. Analysis by Worley, K., Weinstock, G, Gibbs, R, (BCM-HGSC).

We propose to generate and sequence similar short-insert libraries from the same genetic stock as in HBGP, with DNA fractionated as before into AT-biased contingents by M. Beye. Since diminished returns are inevitable at some point, we propose an empirical approach whereby new assemblies are built after each 200K sequence reads, and are reviewed by the Sequencing Center to determine when this approach is losing effectiveness. We anticipate considerable improvement for the assembly after up to 600-800k of new reads by this method.

4b. Characterization of A. mellifera EST's and sequence variation. We propose an additional 100K EST sequencing for *A. mellifera ligustica* (European honey bee, EHB) along with 120K EST sequencing for *A. mellifera scutellata,* the subspecies from which "Africanized" honey bees (AHB) derive. For both AHB and EHB, we propose two-directional sequencing of 10,000 clones each from 5 cDNA libraries collected to reflect a diversity of developmental stages and tissues, including pooled embryos (4-72 h), early-instar larvae, pupae, adult males, and adult queens. For AHB, an additional 10,000 clones will be derived from a brain library, to complement excellent existing brain ESTs from EHB (Table 1). Normalized, directional

libraries will be constructed from cDNA derived from whole-body RNA extracts (or dissected brains). The libraries will be provided and tested by members of the HBGP Consortium.

We also propose minimal (0.2X, 70K reads) survey sequencing of the AHB genome. This will provide ample material for SNP-based mapping between and within these subspecies. An incredibly long genetic map for honey bees (>4K cM[80], analogous to that of humans and other organisms with vastly larger genome sizes) allows for very precise positional mapping of markers, and having honey bee SNP's at this high coverage will allow the characterization of recombinant neighborhoods that occur at the level of 10-50kbp. Pilot studies are currently being conducted by Consortium members (using SNPs identified from genome traces and ca. 2100 high quality AHB genomic sequences generated at BCM-HGSC, each of which contained at least one useful SNP) at a scale of 1536 SNP markers to be genotyped in ca. 1000 individual bees. Results from these studies will be used to guide execution of this project.

Along with providing tools for mapping traits of interest, SNPs identified between and within subspecies will be used immediately to associate unmapped genes (ca. 4200, section B1) to chromosome region. The mapping approach will take advantage of bee haplodiploidy, in which drones (males) are meiotic haploid progeny of a diploid queen (female), and thus serve as an F2 mapping population for linkage analysis of markers in the single parental queen. Because the 7.5x genome sequence was obtained entirely from drone progeny of a single queen (called "DH4"), additional DH4-derived drones (banked) will serve as F2 mapping population for SNPs present in the two haplotypes present in genome trace sequence. SNPs derived from AHB - EHB comparison likewise will be mapped using extant drone progeny of an AHB/EHB hybrid queen. In parallel, targeted AHB sequencing will be performed (single traces from PCR derived genomic DNA) by Consortium members, to identify SNPs linked to the remaining unmapped genes (not successfully mapped above). An additional benefit of generating large numbers of novel SNP's will be the development of fast and economical diagnoses of honey bee "Africanization" in the U.S., an objective of considerable and direct economic impact (see original HBGP white paper).

4c. Genomic survey sequencing of outgroup bee taxa. We propose low-level draft (2X sequence coverage) genomic sequencing of short-insert plasmid libraries from three “outgroup” species in the bee superfamily Apoidea: *Apis dorsata*, *Bombus terrestris* and *Megachile rotundata*. For each species, library clones (5-7 kbp) will be sequenced from both directions, and reads will be assembled using the ATLAS assembly pipeline at BCM-HGSC, then annotated using pipelines (based on ENSEMBL along with other, *de novo,* tools) developed during genome projects of the honey bee, parasitoid wasp *Nasonia vitripennis*, and other insect species. Genome size estimates by flow cytometry are 274 MBp[81] and 330 MBp[82], respectively, for *B. terrestris* and *M. rotundata*. While there is no concrete genome size estimate for *A. dorsata*, nearly identical estimates for congeners *A. mellifera* and *A. cerana* suggest this species will similarly be ~ 280Mbp. For *B. terrestris*, several genomic linkage maps exist [83] and QTLs have been identified for immune defense phenotypes and other traits. These efforts include the development of several hundred genetic markers for this species which should help in assembly of the *B. terrestris* low-level draft. It is also anticipated that within one year a BAC-library of the *B. terrestris* genome (12X clone coverage) will be available from the lab of collaborator P.Schmid-Hempel.

Each outside species will then be used separately and in combination as informants for the gene inference prediction program TWINSCAN[74,75]. This program is well suited for gene prediction using evolutionary distances within the span of informant species we propose. Indeed, TWINSCAN has been used successfully for gene predictions in both substantially longer[84] and shorter[85-87] distances. One example comes from the chicken. Like honey bees, the chicken lacks both extensive transcript-based gene evidence (EST’s and cDNA’s) and close neighbors for which genomic data are available. Here, TWINSCAN correctly predicted hundreds of validated novel genes using the human genome sequence as an informant (300 my distant)[88]. TWINSCAN and the related program SGP2[77] also extended thousands of extant, orthology- or EST-based, gene predictions to better represent true gene structure for the chicken. While this distance reflects an extreme for TWINSCAN usefulness[86], the results suggest that all three informant species are well within range for these methods. Accordingly, we anticipate synergism between these taxa as TWINSCAN informants for honey bee, and a further refinement of estimates of the effective distance of this method for gene inference[86]. Dr. Michael Brent (Washington Univ., developer of TWINSCAN) has

offered to assist the bee community and Sequencing Center with respect to the effective implication of TWINSCAN and interpretation of results. We will also use sequences from the three outgroup species, and from ongoing sequencing projects from *Nasonia vitripennis* and other insects, to screen the honey bee genome for conserved noncoding elements[72,89] and miRNA's[73,90].

5. Costs and readiness

5a. Predicted costs. The direct sequencing cost for an NIH-NHGRI Sequencing Center to perform this project is $3M ($0.90/read x 3.34M reads), along with costs associated with genome assembly and analyses. Predicted sequencing throughput includes ≤800K genomic sequence reads from *A.m.*, 100K EST reads from *A.m.,* 120K EST reads from AHB, 70K genomic survey sequences from AHB, and a total of 2.25M reads (2X coverage each) from *A. dorsata, B. terrestris* and *M. rotundata*.

5b. Biological material. Males derived from the exact same queen used for HBGP have been banked in sufficient numbers in several laboratories (BCM, UI, USDA-Beltsville, and Univ. Halle) to generate ample amounts of AT-biased DNA for new libraries. These samples will be provided for whichever sequencing strategy is chosen in consultation with the Sequencing Center. Libraries for generation of EST reads from *A.m.* will be funded by ongoing research projects at Australia (R. Maleska), Illinois (G. Robinson) and Maryland (J. Evans). Africanized bee samples will be derived from populations in Brazil or Mexico for which admixture with other bee subspecies is believed to be minimal. All cDNA libraries will be normalized and screened prior to delivery to the sequencing center. Specimens of *dorsata* will be provided from a single native colony by Dr. Kiyoshi Kimura (Tsukuba, Japan). *B. terrestris* will be obtained from a well-studied population in Switzerland[91-95] by collaborator P. Schmid-Hempel (Swiss Federal Institute of Technology, Zurich). This is the same population used for the described BAC-sequencing and QTL projects. *B. terrestris* colonies are singly mated and sufficiently large as to allow collection of ample (male) material from a single colony for several short-insert genomic DNA libraries (as was done for *A.m.*). As for the *A.m.* samples described here and the HBGP, progeny from one queen will ensure that precisely two haplotypes are present in the sequenced libraries, at equal ratios. We plan to choose a single North American commercial lineage of *M. rotundata*, then will select individuals (male and female) from several matrilines in order to pool sufficient amounts of genomic DNA. Collaborators R. James and T. Pitts-Singer (USDA-ARS, Logan, Utah) have volunteered samples of *M. rotundata* from their cultivated populations.

6. Other (partial) sources of funding. All biological material will be collected by the authors of this document and collaborators. Funding to establish the cDNA libraries will be obtained from Texas A&M University (TAMU), the Texas Beekeepers Association and researchers in the bee community. BeeBase is located and supported by TAMU, BeeSpace (Univ. of Illinois) and the pursuit of other NIH or NSF funding. BeeSpace (Univ. of Illinois) is funded by a $5M NSF FIBR grant to the UI Institute for Genomic Biology (White Paper Co-Author is a Co-PI). Additional informatics support has been provided *gratis* by NCBI, US-DOE, Univ. California, and European Bioinformatics Institute. Dr. Michael Brent (Washington Univ.) has offered to share expertise with his annotation tool TWINSCAN.

7. Letters of support roster. We have not repeated support letters from HBGP in 2002, but instead offer letters from collaborators involved with collecting new material and insights (P. Schmid-Hempel and R. James), those for whom resources involving non-*Apis* bees will be extremely useful (K. Winter, B. Danforth), and a leader of a proposed collaborating NHGRI Human Genome Sequencing Center (R. Gibbs). Collaborators now supporting honey bee annotation are listed as Appendix I.

8. Institutional Affiliations of White Paper Authors and Acknowledgements. J.D. Evans, USDA Bee Research Lab., Beltsville, MD; M. Beye, Univ. Halle, Germany, C. Elsik, Dept. Animal Sci. Texas A and M Univ., R. Maleszka, Dept. Biology, Australian Natl. Univ.; H.M. Robertson, Dept. Entomology, Univ. Illinois at Urbana-Champaign; G.E. Robinson, Neuroscience Program and Institute of Genomic Biology, Univ. Illinois at Urbana-Champaign, D.B. Weaver, B Weaver Apiaries, Inc., Navasota, TX, C.W. Whitfield, Dept. of Entomology, Univ. Illinois. Additional assistance and advice provided by D. Inouye, Univ. Maryland; P. Schmid-Hempel, ETH-Zurich; T. Pitts-Singer and R. James, USDA-ARS, Logan, Utah; W. Kemp, USDA-ARS, Fargo N.D, M. Brent, Washington Univ., B. Danforth, Cornell Univ., S. Cameron, Univ. Illinois.

9. References

1. Willis L. G., Winston M. L., Honda B. M. (1992) Phylogenetic relationships in the honeybee (genus *Apis*) as determined by the sequence of the cytochrome oxidase II region of mitochondrial DNA. *Mol Phylogenet Evol*, 1:169-178.
2. Springer M. S., Murphy W. J., Eizirik E., O'Brien S. J. (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A*, 100:1056-1061.
3. Wilson E. O. (1975) *Sociobiology:The New Synthesis*. Cambridge, MA: Harvard Univ. Press.
4. Maleszka R., Helliwell P., Kucharski R. (2000) Pharmacological interference with glutamate re-uptake impairs long-term memory in the honeybee, *Apis mellifera*. *Behavioural Brain Research*, 115:49-53.
5. Dacher M., Lagarrigue A., Gauthier M. (2005) Antennal tactile learning in the honeybee: Effect of nicotinic antagonists on memory dynamics. *Neuroscience*, 130:37-50.
6. Wüstenberg D. G., Grünewald B. (2004) Pharmacology of the neuronal nicotinic acetylcholine receptor of cultured Kenyon cells of the honeybee, *Apis mellifera*. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 190:807-821.
7. Si A., Helliwell P., Maleszka R. (2004) Effects of NMDA receptor antagonists on olfactory learning and memory in the honeybee (*Apis mellifera*). *Pharmacology Biochemistry and Behavior*, 77:191-197.
8. Kucharski R., Maleszka R. (2005) cDNA microarray and real-time analyses of gene expression in the honey bee brain following caffeine treatment. *J. Mol. Neurosci.*
9. Rueppell O., Pankiw T., Nielsen D. I., Fondrk M. K., Page Jr. R. E., Beye M. (2004) The genetic architecture of the behavioral ontogeny of foraging in honeybee workers. *Genetics*, 167:1767-1779.
10. Spivak M., Lapidge K. L., Oldroyd B. P. (2002) Seven suggestive quantitative trait loci influence hygienic behavior of honey bees. *Naturwissenschaften*, 89:565-568.
11. Lobo N. F., Hill C. A., Romero-Severson J., Hunt G. J., Collins F. H., Ton L. Q., Emore C. (2003) Genomic analysis in the sting-2 quantitative trait locus for defensive behavior in the honey bee, *Apis mellifera*. *Genome Research*, 13:2588-2593.
12. Arechavaleta-Velasco M. E., Hunt G. J., Emore C. (2003) Quantitative trait loci that influence the expression of guarding and stinging behaviors of individual honey bees. *Behavior Genetics*, 33:357-364.
13. Robinson G. E. (2004) Genomics. Beyond nature and nurture. *Science*, 304:397-399.
14. Giurfa M. (2004) Conditioning procedure and color discrimination in the honeybee *Apis mellifera*. *Naturwissenschaften*, 91:228-231.
15. Guerrieri F., Schubert M., Sandoz J.-C., Giurfa M. (2005) Perceptual and neural olfactory similarity in honeybees. *PLoS Biology*, 3:0718-0732.
16. Giurfa M., Zhang S., Jenett A., Menzel R., Srinivasan M. V. (2001) The concepts of 'sameness' and 'difference' in an insect. *Nature*, 410:930-933.
17. Robinson G. E., Grozinger C. M., Whitfield C. W. (2005) Sociogenomics: Social life in molecular terms. *Nature Reviews Genetics*, 6:257-270.
18. Dommett R., Zilbauer M., George J. T., Bajaj-Elliott M. (2005) Innate immune defence in the human gastrointestinal tract. *Molecular Immunology*, 42:903-912.
19. Hoffmann J. A. (2003) The immune response of *Drosophila*. *Nature*, 426:33-38.
20. Imler J.-L., Zheng L. (2004) Biology of Toll receptors: Lessons from insects and mammals. *Journal of Leukocyte Biology*, 75:18-26.

21. Schulenburg H., Kurz C. L., Ewbank J. J. (2004) Evolution of the innate immune system: The worm perspective. *Immunological Reviews*, 198:36-58.
22. Williams M. J. (2001) Regulation of antibacterial and antifungal innate immunity in fruitflies and humans. *Advances in Immunology*, 79:225-259.
23. Royet J. (2004) Infectious non-self recognition in invertebrates: Lessons from *Drosophila* and other insect models. *Molecular Immunology*, 41:1063-1075.
24. Christophides G. K., Vlachou D., Kafatos F. C. (2004) Comparative and functional genomics of the innate immune system in the malaria vector *Anopheles gambiae*. *Immunological Reviews*, 198:127-148.
25. Nappi A. J., Christensen B. M. (2005) Melanogenesis and associated cytotoxic reactions: Applications to insect innate immunity. *Insect Biochemistry and Molecular Biology*, 35:443-459.
26. Meister S., Koutsos A. C., Christophides G. K. (2004) The *Plasmodium* parasite - A 'new' challenge for insect innate immunity. *International Journal for Parasitology*, 34:1473-1482.
27. Morse R. A., Flottum K. (Ed): (1997) *Honey Bee Pests Predators and Diseases* Medina, Ohio: A.I. Root Co.;
28. Evans J. D. (2004) Transcriptional immune responses by honey bee larvae during invasion by the bacterial pathogen, *Paenibacillus larvae*. *Journal of Invertebrate Pathology*, 85:105-111.
29. Aronstein K., Saldivar E. (2005) Characterization of a honey bee Toll related receptor gene Am18w and its potential involvement in antimicrobial immune defense. *Apidologie*, 36:3-14.
30. Gregory P. G., Evans J. D., Rinderer T., De Guzman L. (2005) Conditional immune-gene suppression of honeybees parasitized by *Varroa* mites. *Journal of Insect Science*, 5:1-5.
31. Yang X., Cox-Foster D. L. (2005) Impact of an ectoparasite on the immunity and pathology of an invertebrate: Evidence for host immunosuppression and viral amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 102:7470-7475.
32. Corona M., Estrada E., Zurita M. (1999) Differential expression of mitochondrial genes between queens and workers during caste determination in the honeybee *Apis mellifera*. *Journal of Experimental Biology*, 202:929-938.
33. Evans J. D., Wheeler D. E. (1999) Differential gene expression between developing queens and workers in the honey bee, *Apis mellifera*. *Proceedings of the National Academy of Sciences of the United States of America*, 96:5575-5580.
34. Evans J. D., Wheeler D. E. (2001) Expression profiles during honeybee caste determination. *Genome biology*, 2.
35. Braendle C., Caillaud M. C., Stern D. L. (2005) Genetic mapping of aphicarus - A sex-linked locus controlling a wing polymorphism in the pea aphid (*Acyrthosiphon pisum*). *Heredity*, 94:435-442.
36. Beye M., Hasselmann M., Fondrk M. K., Page R. E., Omholt S. W. (2003) The gene csd is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell*, 114:419-429.
37. Crozier R. H., Pamilo P. (1996) Oxford Series in Ecology and evolution: Evolution of social insect colonies: Sex allocation and kin selection.
38. West S. A., Murray M. G., Machado C. A., Griffin A. S., Herre E. A. (2001) Testing Hamilton's rule with competition between relatives. *Nature*, 409:510-513.

39. Beye M., Moritz R. F. A., Hunt G. J., Page R. E., Kim Fondrk M., Grohmann L. (1999) Unusually high recombination rate detected in the sex locus region of the honey bee (*Apis mellifera*). *Genetics*, 153:1701-1708.
40. Aron S., De Menten L., Roisin Y., Van Bockstaele D. R., Blank S. M. (2005) When hymenopteran males reinvented diploidy. *Current Biology*, 15:824-827.
41. Whitfield C. W., Cziko A.-M., Robinson G. E. (2003) Gene expression profiles in the brain predict behavior in individual honey bees. *Science*, 302:296-299.
42. Kucharski R., Maleszka R. (2002) Evaluation of differential gene expression during behavioral development in the honeybee using microarrays and northern blots. *Genome biology*, 3.
43. Robinson G. E. (2002) Genomics and integrative analyses of division of labor in honeybee colonies. *American Naturalist*, 160.
44. Coffman C. J., Wayne M. L., Nuzhdin S. V., Higgins L. A., McIntyre L. M. (2005) Identification of co-regulated transcripts affecting male body size in *Drosophila*. *Genome Biol*, 6:R53.
45. Burgler C., Macdonald P. M. (2005) Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics*, 6:88.
46. Lai E. C., Tam B., Rubin G. M. (2005) Pervasive regulation of *Drosophila* Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes Dev*, 19:1067-1080.
47. Zhang Y. Q., Broadie K. (2005) Fathoming fragile X in fruit flies. *Trends Genet*, 21:37-45.
48. Klaudiny J., Bachanová K., Simuth J., Albert S., Kopernický J. (2005) Two structurally different defensin genes, one of them encoding a novel defensin isoform, are expressed in honeybee *Apis mellifera*. *Insect Biochemistry and Molecular Biology*, 35:11-22.
49. Page R.E. J., Peng C. Y.-S. (2001) Aging and development in social insects with emphasis on the honey bee, *Apis mellifera* L. *Experimental Gerontology*, 36:695-711.
50. Vettraino J, Buck S, R A. (2001) Direct selection for paraquat resistance in *Drosophila* results in a different extended longevity phenotype. *J Gerontol A*, 56:B415-425.
51. Corona M., Hughes K. A., Weaver D. B., G.E. R. (accepted pending revision) Gene expression patterns associated with queen honey bee longevity. *Mechanisms of aging and development.*
52. Corona M., Hughes K. A., Weaver D. B., G.E. R. (unpublished data).
53. Collins A. M., Williams V., Evans J. D. (2004) Sperm storage and antioxidative enzyme expression in the honey bee, *Apis mellifera*. *Insect Molecular Biology*, 13:141-146.
54. Woyke J., Wilde J., Reddy C. C. (2004) Open-air-nesting honey bees *Apis dorsata* and Apis laboriosa differ from the cavity-nesting *Apis mellifera* and *Apis cerana* in brood hygiene behaviour. *Journal of Invertebrate Pathology*, 86:1-6.
55. Kavinseksan B., Wongsiri S., De Guzman L. I., Rinderer T. E. (2003) Absence of Tropilaelaps infestation from recent swarms of *Apis dorsata* in Thailand. *Journal of Apicultural Research*, 42:49-50.
56. Wongsiri S., Rinderer T. E., Sylvester H. A., Crozier R. H., Oldroyd B. P., Clifton M. J. (1997) Polyandry in the genus Apis, particularly *Apis andreniformis*. *Behavioral Ecology and Sociobiology*, 40:17-26.
57. Wattanachaiyingcharoen W., Wongsiri S., Oldroyd B. P., Palmer K., Paar J. (2003) A scientific note on the mating frequency of *Apis dorsata*. *Apidologie*, 34:85-86.
58. Melendez A., Talloczy Z., Seaman M., Eskelinen E. L., Hall D. H., Levine B. (2003) Autophagy genes are essential for dauer development and life-span extension in C. elegans. *Science*, 301:1387-1391.
59. Schmid-Hempel P. (2005) Evolutionary ecology of insect immune defenses. *Annual Review of Entomology*, 50:529-551.

60. Brown M. J. F., Moret Y., Schmid-Hempel P. (2003) Activation of host constitutive immune defence by an intestinal trypanosome parasite of bumble bees. *Parasitology*, 126:253-260.
61. Lord G. M., Matarese G., Howard J. K., Lechler R. I., Moret Y., Schmid-Hempel P. (2001) The bioenergetics of the immune system [3]. *Science*, 292:856.
62. Moret Y., Schmid-Hempel P. (2001) Immune defence in bumble-bee offspring. *Nature*, 414:506.
63. Schmid-Hempel P., Schmid-Hempel R. (1990) Endoparasitic larvae of conopid flies alter pollination behavior of bumblebees. *Naturwissenschaften*, 77:450-452.
64. Shykoff J. A., Schmid-Hempel P. (1991) Parasites delay worker reproduction in bumblebees: consequences for eusociality. *Behavioral Ecology*, 2:242-248.
65. Gadau J., Gerloff C. U., Krüger N., Schmid-Hempel H. C. P., Wille A., Page R.E. J. (2001) A linkage analysis of sex determination in Bombus terrestris (L.) (Hymenoptera: Apidae). *Heredity*, 87:234-242.
66. Allen-Wardell G., Bernhardt P., Bitner R., Burquez A., Buchmann S., Cane J., Cox P. A., Dalton V., Feinsinger P., Ingram M., et al. (1998) The potential consequences of pollinator declines on the conservation of biodiversity and stability of food crop yields. *Conservation Biology*, 12:8-17.
67. Morandin L. A., Laverty T. M., Kevan P. G. (2001) Bumble bee (Hymenoptera: Apidae) activity and pollination levels in commercial tomato greenhouses. *Journal of Economic Entomology*, 94:462-467.
68. Higo H. A., Rice N. D., Winston M. L., Lewis B. (2004) Honey bee (Hymenoptera: Apidae) distribution and potential for supplementary pollination in commercial tomato greenhouses during winter. *Journal of Economic Entomology*, 97:163-170.
69. Cane J. H. (2002) Pollinating bees (Hymenoptera: Apiformes) of U.S. alfalfa compared for rates of pod and seed set. *Journal of Economic Entomology*, 95:22-27.
70. Holt R. A., Subramanian G. M., Halpern A., Sutton G. G., Charlab R., Nusskern D. R., Wincker P., Clark A. G., Ribeiro J. M. C., Wides R., et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298:129-149.
71. Boffelli D., Nobrega M. A., Rubin E. M. (2004) Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics*, 5:456-465.
72. Numata K., Kanai A., Saito R., Tomita M., Kondo S., Adachi J., Hayashizaki Y., Arakawa T., Carninci P., Kuwai J., et al. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Research*, 13:1301-1306.
73. Lagos-Quintana M., Rauhut R., Meyer J., Tuschl T., Borkhardt A. (2003) New microRNAs from mouse and human. *RNA*, 9:175-179.
74. Chuang T.-J., Che F.-C., Chou M.-Y. (2004) A comparative method for identification of gene structures and alternatively spliced variants. *Bioinformatics*, 20:3064-3079.
75. Wu J. Q., Gibbs R. A., Shteynberg D., Arumugam M., Bren M. R. (2004) Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Research*, 14:665-671.
76. Flicek P., Keibler E., Hu P., Korf I., Brent M. R. (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome research*, 13:46-54.
77. Parra G., Agarwal P., Abril J. F., Wiehe T., Fickett J. W., Guigo R. (2003) Comparative gene prediction in human and mouse. *Genome Res*, 13:108-117.
78. Tomkins J. P., Fang G. C., Main D., Atkins M., Luo M., Goicoechea J. L., Yu Y., Wing R. A., Frisch D. A., Page R. E., et al. (2002) New genomic resources for the honey bee (*Apis*

mellifera L.): Development of a deep-coverage BAC library and a preliminary STC database. *Genetics and Molecular Research*, 1:306-316.
79. Li W., Bernaola-Galvan P., Haghighi F., Grosse I. (2002) Applications of recursive segmentation to the analysis of DNA sequences. *Computers and Chemistry*, 26:491-510.
80. Solignac M., Vautrin D., Baudry E., Mougel F., Loiseau A., Cornuet J. M. (2004) A microsatellite-based linkage map of the honeybee, *Apis mellifera* L. *Genetics*, 167:253-262.
81. Gadau J., Page Jr. R. E., Werren J. H., Schmid-Hempel P. (2000) Genome organization and social evolution in hymenoptera. *Naturwissenschaften*, 87:87-89.
82. Petitpierre E. (1996) Molecular cytogenetics and taxonomy of insects, with particular reference to the Coleoptera. . *International Journal of Insect Morphology & Embryology*, 25:115-133.
83. Gadau J., Gerloff C. U., Krueger N., Chan H., Schmid-Hempel P., Wille A., Page R. E. (2001) A linkage analysis of sex determination in Bombus terrestris (L.) (Hymenoptera: Apidae). *Heredity [Heredity]*, 87:234-242.
84. Consortium C. G. S. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. 432:695-716.
85. Wei C., Lamesch P., Arumugam M., Rosenberg J., Hu P., Vidal M., Brent M. R. (2005) Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions. *Genome research*, 15:577-582.
86. Wang M., Buhler J., Brent M. R. (2003) The effects of evolutionary distance on TWINSCAN, an algorithm for pair-wise comparative gene prediction. *Cold Spring Harbor Symposia on Quantitative Biology*, 68:125-130.
87. Tenney A. E., Brown R. H., Vaske C., Brent M. R., Lodge J. K., Doering T. L. (2004) Gene prediction and verification in a compact genome with numerous small introns. *Genome Research*, 14:2330-2335.
88. Eyras E., Reymond A., Castelo R., Bye J. M., Camara F., Flicek P., Huckle E. J., Parra G., Shteynberg D. D., Wyss C., et al. (2005) Gene finding in the chicken genome. *BMC Bioinformatics*, 6:131.
89. Postlethwait J., Amores A., Cresko W., Singer A., Yan Y.-L. (2004) Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends in Genetics*, 20:481-490.
90. Ambros V., Bartel B., Bartel D. P., Burge C. B., Carrington J. C., Chen X., Dreyfuss G., Eddy S. R., Griffiths-Jones S., Marshall M., et al. (2003) A uniform system for microRNA annotation. *RNA*, 9:277-279.
91. Muller C. B., Schmid-Hempel P. (1993) Exploitation of cold temperature as defence against parasitoids in bumblebees. *Nature*, 362:65-67.
92. Baer B., Schmid-Hempel P. (1999) Experimental variation in polyandry affects parasite loads and fitness in a bumble-bee. *Nature*, 397:151-154.
93. Allander K., Schmid-Hempel P. (2000) Immune defence reaction in bumble-bee workers after a previous challenge and parasitic coinfection. *Functional Ecology*, 14:711-717.
94. Mallon E. B., Loosli R., Schmid-Hempel P. (2003) Specific versus nonspecific immune defense in the bumblebee, Bombus terrestris L. *Evolution*, 57:1444-1447.
95. Baer B., Schmid-Hempel P. (2005) Sperm influences female hibernation success, survival and fitness in the bumble-bee Bombus terrestris. *Proc Biol Sci*, 272:319-323.

Appendix 1. Honey Bee Genome Analysis Plan Draft 5/15/05

Agriculture and Medicine (Jay Evans)

Innate immunity (Jay Evans)

Fly/moth functional inferences

Charles Hetru – Univ. Stausbourg	
antimicrobial peptides	10+
GNBPs (*b*-glucan)	40?
TEPs	10?
Serpins	30?
Phenoloxidases	3?
Insights from fly arrays wrt bacterial, fungal, viral infection	
Dan Hultmark – Univ. Umea, Sweden	
PGRP's	5
Dorsal, DIF, Cactus	3
Cellular and humoral immunity	?
Jean-Luc Imler – Univ. Strausbourg	
JAK-STAT pathway	?
Cellular immunity	?
M. Kanost (UKansas State Univ) H. Jiang (Univ. Oklahoma)	
Serine proteases, CLIP, etc.	?
C-type lectins	?
Galectins	?

Bee Disease subgroup

Katherine Aronstein - USDA Kika de la Garza	
Toll/tlrs	6
Imd/Rel and associates	2+
RNAi of potential toll pathway members	
Judy Chen - USDA Bee Lab	
viral immunity genes, transcriptional response	?
Jay Evans - USDA Beltsville	
IG superfamily genes/fibrinogen	80
AMP's:	10
Scavenger receptors	6
Transcriptional response to disease	

Sequence-level and paralogy stories

Andy Clark – Cornell
Global/SNP analyses

Brian Lazzaro – Cornell
Global/SNP analyses

Graham Thompson – Univ. Sydney

Global/familial selection analyses (KS/Ka)

Major Royal Jelly Proteins (Stefan Albert)

Josef Simuth/Katarina Bilikova – Slovak Acad Sciences/Max Planck Berlin
major proteins of larval food (royal jelly) ?
exogenous and endogenous defense proteins and peptides secreted into honey bee products ?

Stefan Albert – U. Wuerzberg
Major proteins of royal jelly and other yellow-related proteins

Mark Drapeau -NIH
MRJP's

Ryszard Maleska
MRJPs

Pesticide and stress resistance (Charles Claudianos)

Katherine Aronstein - USDA Kika de la Garza
pesticide resistance (GABA subunits, sodium channel, metabolic enzymes..) ?

Reed Johnson and May Berenbaum - UIUC
p450s 100

Charles Claudianos, Rene Feyereisen, Hilary Ranson- ANU, INRA-Antibes, Univ. Liverpool
p450s, pesticide resistance
COE

Pamela Gregory - USDA-Weslaco
Stress-related proteins/mite responses ?

Population genetics/migration insights (Charlie Whitfield)

Steve Sheppard - Washington State Univ.
Population genetics (honey bee diaspora from Eurasia)

David Wheeler- BCM
SNP detection across races

Charlie Whitfield- Univ. Illinois
SNP detection in North American populations

Michel Solignac- CNRS, Gif-sur-Yvette
microsatellite polymorphism and mapping)

Deb Smith- Univ. Kansas
Phylogenetics within the genus *Apis*

Jay Evans (USDA-Beltsville)
Msat development/trace polymorphism

Neurobiology & Behavior (Ryszard Maleszka)
Arnd Baumann - Institut fuer Biologische Informationsverarbeitung, Juelich, Germany

Biogenic amine receptors	30
adenylyl and guanylyl cylases	10
NOS system	5

Yehuda Ben-Shahar – Univ. Iowa

Serine/threonine and serine/tyrosine kinases	160
DEG/ENaC Na channel family	30

Kyle Beggs and Alison Mercer – Univ. Otago New Zealand

Biosynthetic enzymes for biogenic amines (eg. Tyrosine hydroxylase)	5
Identification of G protein components	9

David Bernard - NHGRI

inositol phosphatases and kinases	?
Neuronal Ceroid Lipfuscinosis (NCL)	?
synucleins	?

Wolfgang Blenau - Universität Potsdam

biogenic amine and acetylcholine receptors	20
Other G protein-coupled receptors, e.g. peptide receptors	25

Guy Bloch and Michal Linial - Hebrew University Jerusalem

Circadian rhythm genes	10
Small GTP binding proteins	80
JH binding proteins (takeout and related genes)	30
Tubulins and actins	25
SNARE (secretory proteins) and trafficking	50

Charles Claudianos and Ryzard Maleszka - Australian National University

neuroligins, neurotactins, gliotactins, glutactins, etc. (proteins containing non-catalytic choline esterase domains)	45

Paul Ebert coordinate a topic on longevity, stress resistance and respiration

Paul Ebert - University of Queensland

Biogenic amine receptors - ready to go	20
Glutathione-S-transferases	30
UDP-glucurosyltransferases	60
ABC transporters	30

Dorothea Eisenhardt - Freie Universität Berlin

CREB/CREM family of transcription factors and other bZIP proteins	?

Dorothea Eisenhardt, Gerard Leboulle - Freie Universität Berlin

Protein kinase A genes (catalytic and regulatory subunit)	?
RAC1 protein	

Susan Fahrbach, Rodrigo Velarde, Klaus Hartfelder - Wake Forest, UIUC, University of Sao PaoloD

Nuclear hormone receptors	20

Brenton Graveley - Univ. Connecticut Health Center

DSCAM	1

Frank Horodyski - Ohio University

neuropeptides, particularly allatotropins, allatostatins, eclosion hormone	?

Tatsuhiko Kadowaki - Nagoya University

Wnt signalling pathway	?
TRP channel gene family	?
Mechanosensory and auditory pathway	?

Greg Hunt – coordinate the topic

Candidate genes for social behavioral traits mapped as QTLs

Ryszard Maleszka - Australian National University

Glutamate transporters, receptors, etc:	30
(also other receptors; adenosine, serotonin, aetylcholine, etc unless somebody else is keen to do these?)	100
IP3 system	?
Structural organization of the synapse, adhesion, vesicular proteins, etc.)	?
OBPs (done)	

Jonathan Sweedler, Amanda Hummon, Gene Robinson, Sandra Rodriguez-Zas Timothy Richmond– UIUC
Peter Verleyen and Lilian Schoofs – Katholieke Univ Belgium

Neuropeptide genes	100

Cornelius Grimmelikhuijzen –Denmark

Molecular endocrinology and neurohormone GPCRs (done)	60

Hugh Robertson lab- UIUC

Odorant and gustatory chemoreceptors (done)	150

Sylvain Foret, Ryszard Maleszka

Odorant binding proteins

Zachary Huang and Ke Dong - Michigan State University

voltage- and ligand-gated ion channels	70

Catherine Hill - Purdue University

QFC algorithm for finding GPCRs	?

Stefan Albert – U. Wuerzberg

small G-protein - rab proteins in particular
regulators of G-proteins, GAPs, GDIs, GEFs
Raf kinases and signal transduction

Takeo Kubo lab - University of Tokyo

carbohydrate metabolizing enzymes	50

Evolution (Hugh Robertson)

Stewart Berlocher - UIUC

Glycolytic pathway	10
Other commonly used allozymes	20

Greg Hunt

Genes relate to the high meiotic recombination rate in honey bees

Hugh Robertson lab - UIUC

Tetraspanins (done)	40
Opsins (done)	10
Methuselahs	10
Transposons	

bee/human genes missing from *Drosophila*

Peer Bork, Evgeny Zdobnov - EMBL
- Orthology finding (requires annotated genes)
- Global comparative analyses
- Synteny to diptera (hard due to distance)
- Family expansion (pfam + inparalog identification –> more recent expansion of genes)
- Gene family gains and losses over longer time scales
- Intron evolution, alternative splicing evolution (exploratory, might not work)

Chris Elsik lab – BeeBase, Texas A&M
- Global comparative analyses

Miguel Corona and Gene Robinson - UIUC
- Toxins and venoms 30

Rob Cutler – Bard College
- Non-coding RNA

Gene Regulation (Charles Whitfield)

Christina Grozinger – NC State
- Chromatin remodeling proteins (HDATs, HATs, etc.) 40

Charlie Whitfield - UIUC
- MicroRNAs 20?

Karl Gordon - CSIRO, Canberra
- microRNAs

Rob Reenan
- RNA editing

Ryszard Maleszka, Hugh Robertson, Yu Ling, Ying Wang and Gene Robinson
- methylation
- DNA methyltransferases (done) 3

Chris Elsik, Danny Weaver, Juan Anzola – Texas A&M
- miRNAs

Jeff Shen – Nevada
- Transcription factors from rice BHLH and MYB gene families
- Plant – insect interactions

Michael Linial
- Alternative splicing levels in bee vs. other insects

Development & Metabolism (Martin Beye)

Martin Beye, Martin Hasselmann, Tanja Gempe, Morten Schioett- Universitat Halle
- Haplodiploidy 120
- sex determining genes 20

Sydney Cameron - UIUC
- pigment patterning genes ?

Zachary Huang, Ke Dong and Klaus Hartfelder group - Michigan State, Sao PaoloD
- JH and melatonin enzyme pathways 30

Shu-ning Hsu, Hugh Robertson and Akira Chiba – UIUC (Chiba lab)	
Cadherins	?
Anita Collins and Jay Evans - USDA Beltsville	
Seminal and other sperm storage proteins	30
Craig Coates and Danny Weaver - Texas A&M	
DEAD-box family of proteins	30
Danny Weaver - Texas A&M	
Integrins	
Miguel Corona and Gene Robinson – UIUC	
Antioxidant proteins (done)	20
Nuclear-encoded mitochondrial proteins (done)	50
Other proteins implicated in aging and longevity	?
Michelle Elekonich – Univ. Nevada	
Heatshock proteins and chaperones	50
Tugrul Giray, Pedro Alvarez, Felipe Soto-Adames and Jim Vigoreaux - University of Puerto Rico and University of Vermont	
Muscle contractile proteins	
Titin-like	8
Myosin & associated proteins	25
Actin & associated proteins	40
Karl Gordon - CSIRO, Canberra	
Peritrophins/mucins	
Proteinases	
Transporters	
Innexins	
Apoptosis	
Signalling e.g. wnt	
Florence Mougel - CNRS Gif-sur-Yvette	
ribosomal proteins	100
Michel Solignac - University Paris Sud	
aminoacyl-tRNA-synthetases	20
Randi Aamodt and Stig Omholt – Norwegian Univ.	
Regulatory aspects of aging	
DNA repair and aging	
Immunological processes and aging	
Diana Wheeler - University of Arizona	
hexamerins	12
Michael Eisen – UC Berkeley	
Early development	
Bill Gelbart, Tatsuhido Kadowaki – Harvard, Nagoya University	
Embryogenesis and imaginal disc development	
(gap genes, pair-rule genes, segment polarity genes, selector genes, Hox genes, etc.)	

Ross Overbeek, Amy Toth, Seth Ament and Gene Robinson - Fellowship for Interpretation of Genomes and UIUC
metabolism

Judy Willis -Univ. Georgia
Cuticular proteins

Reproduction (Klaus Hartfelder)

Klaus Hartfelder - University of Sao Paolo
Caste determination

Klaus Hartfelder group - University of Sao PaoloD

oogenesis-regulating genes	60
melanization-regulating enzymes	5

Robin Moritz and Michael Lattorff - University Halle

reproductive behavior	30
female fecundity	20
foraging behavior	10
allatotropins/statins	10

Ben Oldroyd and Ryszard Maleszka - Sydney University and Australia National University (ANU)

ovary action, egg development, vitellogenesis etc.	?

Genome Assembly and features (George Weinstock)

Lan Zhang – Baylor College of Medicine
Sequencing data
Methods for the assembly
Statistics on the assembly
Anchoring to chromosomes
Quality assessment

Gene predictions
Vivek Iyer - Ensembl
Evgeny Zdobnov - Homology
Victor Solovyev - FgenesH
Barbara Ruef – NCBI Gnomon
Mike Eisen-LBL
Aaron Mackey Univ. Penn – GLEAN for consolidation of sets
Chris Elsik – Texas A and M – collect and display gene sets in BeeBase
Gene prediction validation
Gene Robinson – Nimblegen high density arrays, qRT-PCR, Northern methods on GC-rich region *ab initio* predictions without orthology or EST evidence.
Gene prediction discussion (including above)
Ryszard Maleszka, Australian National University, MALESZKA@rsbs.anu.edu.au
Tom Newman, Univ. Illinois newmant5550@life.uiuc.edu
Manoj Samanta, NASA, Ames Center, manoj-samanta@yahoo.com
Viktor Stolc, NASA, Ames Center, vstolc@mail.arc.nasa.gov
Kevin White, Yale University kevin.white@yale.edu
Charlie Whitfield, Univ. Illinois charlie@life.uiuc.edu