Sweetness and light: illuminating the honey bee genome

G. E. Robinson, J. D. Evans,* R. Maleszka,† H. M. Robertson, D. B. Weaver,‡ K. Worley,§ R. A. Gibbs§ and G. M. Weinstock§

Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL, USA;* Bee Research Laboratory, USDA-ARS, BARC-E, Beltsville, MD, USA;† Visual Sciences & ARC Centre for Molecular Genetics of Development, Research School of Biological Sciences, The Australian National University, Canberra, Australia;‡ Bee Power, L.P., Navasota, TX, USA;§ Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

Instead of dirt and poison we have rather chosen to fill our hives with honey and wax; thus furnishing mankind with two of the noblest things, which are sweetness and light.
Jonathan Swift

The honey bee *Apis mellifera* is the first hymenopteran and the fifth insect genome to be sequenced (Honey Bee Genome Sequencing Consortium, 2006) in what promises to be a swarm of insect genome sequences expected to appear over the next few years (Table 1). The Honey Bee Genome Sequencing Project (HBGSP) was conceptualized over a period from 1998 to 2001 by the community at courses, conferences and workshops (Robinson, 1999; Maleszka, 2000; Pennisi, 2001). In addition, initial efforts were directed at physical and genetic maps of the genome (Estoup et al., 1995; Hunt & Page, 1995), collections of expressed sequence tags (Evans & Wheeler, 2000; Whitfield et al., 2002), and studies using microarrays (Kucharski & Maleszka, 2002; Takeuchi et al., 2002; Whitfield et al., 2003).

At the end of 2001 members of the honey bee community, led by Gene Robinson and Daniel Weaver, and the United States Department of Agriculture, represented by Kevin Hackett, met at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) to discuss a full genome sequencing project. (Representatives of the bovine community were also at this meeting to discuss their genome project, a gathering warmly remembered as the milk and honey workshop.) A White Paper to the National Human Genome Research Institute of the NIH ensued (Honey Bee Genome Sequencing Consortium, 2002), which led to the HBGSP receiving a high priority ranking in the comparative genomics program at the NHGRI. With this support from NHGRI, and additional contributions from the USDA resulting from the efforts of Under Secretary Joseph Jen, the project began in December 2002 at BCM-HGSC.

All genome projects have their challenges as each genome and organism has its own idiosyncrasies. The honey bee was no different. A principal complication was under-representation of AT-rich regions of the genome among the small insert shotgun libraries constructed in *Escherichia coli* for the bulk of the sequencing. Possibly AT-rich DNA was degraded during the preparation of libraries or the clone inserts were not maintained in *E. coli*. To overcome this, Martin Beye supplied AT-rich DNA isolated from dye-CsCl gradients, and this was used to make more shotgun libraries to build up coverage of the AT-rich regions. It was also found that the genome was not fully represented in the large insert BAC clone library, which again could reflect either loss of some regions during clone preparation or in *E. coli*. The BAC problem was never solved and so these clones were used sparingly in the project. A potential problem, polymorphism making it difficult to assemble shotgun sequences, was managed using a partially inbred queen from Daniel Weaver. The DNA for sequencing came from a large number of drones. Although polymorphism was not insignificant, several polymorphic alleles per kilobase, this was a boon for identifying SNPs and quite manageable in genome assembly.

The lack of BAC clones meant that the HBGSP became a pure Whole Genome Shotgun project. In all, the project produced over three million DNA sequences for assembly, mainly from small insert clones, but including a few fosmid and BAC clones. The genome assembly used over 80% of these data. The reads were assembled into the genome with the Atlas assembly software, developed at the BCM-HGSC (Havlak et al., 2004). All overlaps between reads were first found by an alignment process and highly repeated sequences were identified because of their large number of overlapping reads. These were set aside, and
Table 1. Genome projects of insects and other arthropods

| Organism             | Common name | Order          | Size (Mb) | Status   | Sequencing centres* | Reference/source                                                                 |
|----------------------|-------------|----------------|-----------|----------|----------------------|---------------------------------------------------------------------------------|
| Acyrthosiphon pisum  | Pea aphid   | Hemiptera      | 525       | Ongoing  | BCM-HGSC             | www.hgsc.bcm.tmc.edu                                                           |
| Aedes aegypti        | Mosquito    | Diptera        | 1310      | Complete | BI, TIGR             | msc.tigr.org/aedes/aedes.shtml                                                   |
| Anopheles gambiae    | Mosquito    | Diptera        | 264       | Complete | Celera Genomics,     | Holt et al. (2002); Mongin et al. (2004)                                        |
|                      |             |                |           |          | Genoscope, the       |                                                                                |
|                      |             |                |           |          | University of Notre   |                                                                                |
|                      |             |                |           |          | Dame, EBI/Sanger     |                                                                                |
|                      |             |                |           |          | Institute, EMBL, Insti |                                                                                |
|                      |             |                |           |          | tut, IMBB and TIGR   |                                                                                |
| Apis mellifera       | Honey bee   | Hymenoptera    | 262       | Complete | BCM-HGSC             | Honey Bee Genome Sequencing Consortium (2006)                                   |
| Bicyclus anynana     | Butterfly   | Lepidoptera    | 500       | Ongoing  | JGI                  | msc.tigr.org/c_pipiens/index.shtml                                               |
| Bombbyx mori         | Silkworm    | Lepidoptera    | 530       | Complete | International Lepidopteran Genome Project |                                                                                |
| Culex pipiens        | Mosquito    | Diptera        | 540       | Ongoing  | BI, TIGR             | msc.tigr.org/c_pipiens/index.shtml                                               |
| Daphnia pulex        | Water flea  | Siphonaptera   | 200       | Ongoing  | JGI                  | wfileabase.org                                                                 |
| Drosophila melanogaster | Fruit fly | Diptera        | 132       | Complete | CGI; BDGP; BCM-HGSC   | Adams et al. (2000); Celniker et al. (2002)                                     |
| Drosophila pseudobscura | Fruit fly | Diptera        | 139       | Complete | BCM-HGSC             | Richards et al. (2005)                                                          |
| Drosophila species†  | Fruit fly   | Diptera        | –135      | Ongoing  | Multicentre2         | flybase.bio.indiana.edu                                                        |
| Glossina morsitans   | Tsetse fly  | Diptera        | 590       | Ongoing  | WTSI                 | www.sanger.ac.uk/Projects/G_morsitans/                                          |
| ixodes scapularis    | Tick        | Acarina        | 2100      | Ongoing  | BI, TIGR             | www.entm.purdue.edu/igp/default.html                                             |
| Nasonia sp.          | Wasp        | Hymenoptera    | 345       | Ongoing  | BCM-HGSC             | www.hgsc.bcm.tmc.edu                                                            |
| Pediculus humanus    | Body louse  | Phthiraptera   | 107       | Ongoing  | JCVI                 | www.entm.purdue.edu/pittundrigh_lab/default.html                                |
| Rhodnius prolixus    | Chagas' disease vector | Hymenoptera | 670       | Ongoing  | WUGSC                | www.genome.wustl.edu?GENOME=Rhodnius%20prolixus                                  |
| Sand flies‡          | Sand fly    | Diptera        | 170–300   | Ongoing  | BCM-HGSC, WUGSC      | www.genome.wustl.edu                                                            |
| Tribolium castaneum  | Red flour beetle | Coleoptera     | 158       | Ongoing  | BCM-HGSC             | www.hgsc.bcm.tmc.edu                                                            |

*ABC, Agencourt Bioscience Corp.; BCM-HGSC, Baylor College of Medicine Human Genome Sequencing Center; BDGP, Berkeley Drosophila Genome Project; BI, Broad Institute; CGI, Celera Genomics Inc.; JCVI, J. Craig Venter Institute; JGI, Department of Energy Joint Genome Institute; TIGR, The Institute for Genomic Research; WTSI, Wellcome Trust Sanger Institute; WUGSC, Washington University Genome Sequencing Center.

†Drosophila species being analysed and the centres performing this work are virilis, ananassae, mojavensis, erecta, grimshawi (ABC); willistoni (JCVI); persimilis and sechellia (BI); yakuba and simulans (WUGSC).

‡Lutzomyia longipalpis and Phlebotomus papatasi, EST sequencing.
then a series of steps were performed to create a layout of the reads based on their overlapping sequences. This resulted in clusters of overlapping reads (bins of reads), which end in gaps where the repeated sequences have been removed. Each bin of reads was then assembled into a consensus sequence using Phrap (Ewing & Green, 1998; Ewing et al., 1998), generally producing a single contig (a continuous stretch of sequence). Contigs were linked together into scaffolds using the read pairing information (each clone is sequenced from both ends, producing a pair of reads). The highly repeated sequences were now added back to the assembly, using the read pair information for their placement. The scaffolds were used to build chromosomes, by aligning them to the markers of the linkage map (Solignac et al., 2003, 2004, 2006), called superscaffolding. Manual superscaffolding was also performed by placing reads that were not used by these automated procedures.

The product of these activities was a draft assembly, a consensus sequence good enough to represent nearly all reads that were not used by these automated procedures. Whitfield et al. (2006) performed similar SNP discovery efforts with these Africanized sequences as well as ESTs. In addition, DNA was prepared and sequenced from Africanized honey bees and these individual sequences were compared with the assembled honey bee sequence to identify more SNPs. Both of these efforts with these Africanized sequences as well as ESTs. These groups manually analysed over 3000 gene models and identified changes in gene family numbers or in the genetic composition of pathways, by comparison with other insect genomes as well as other genomes, particularly the human genome. In addition there was considerable effort to confirm missing genes: these may be truly absent or they may be present but not recognized if they have a rapidly evolving sequence.

A principal focus was on the honey bee complex social life-style and how it differs from other solitary life-style insects. This large community effort is presented in a special issue of Nature (Honey Bee Genome Sequencing Consortium, 2006) and in more detail in a large number of companion papers forming this issue as well as in other journals. Papers appearing in this volume of Insect Molecular Biology provide new insights into diverse topics in honey bee biology, including neurobiology (Eisenhardt & Leboulle,
and large-scale collections of SNPs for European and Africanized honey bees (above).

The HBGSP has produced an excellent draft honey bee genome sequence, enhanced by coordinating the assembly of the genome at BCM-HGSC and the mapping of the genome by Michel Solignac and colleagues at INRA, France (Solignac et al., 2003, 2004, 2006). To further increase the value of the honey bee genome sequence to researchers, a White Paper to obtain additional sequence information was submitted to NHGRI in July 2005 (Honey Bee Genome Sequencing Consortium, 2005). The project was accorded ‘High Priority’ in August 2005, and this work will begin late in 2006. The HBGSP is expected to usher in a bright era of bee research, for the benefit of agriculture, biological research and human health.

Acknowledgements

We gratefully recognize the financial and administrative support from the NHGRI-NIH throughout this project, as well as additional support from the USDA. Other support for the project has come from the Texas Agricultural Experiment Station, University of Illinois Sociogenomics Initiative, Texas Beekeepers Association, various private donors from the bee industry including Dutch Gold Honey, Golden Heritage Honey, Burleson’s Honey, and Bee Weaver Apiaries, Inc. This project has been a highly socially coordinated effort by the Honey Bee Genome Sequencing Consortium.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., et al. (2000) The genome sequence of Drosophila melanogaster. Science 287: 2185–2195.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., et al. (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol 3, RESEARCH0079.

Claudianos, C., Ranson, H., Johnson, R.M., Biswas, S., Schuler, M.A., Berenbaum, M.R., et al. (2006) A deficit of detoxification...
enzymes: Pesticide sensitivity and environmental response in the honeybee. *Insect Mol Biol* 15: 615–636.

Collins, A.M., Caperna, T.J., Williams, V., Garrett, W.M. and Evans, J.D. (2006) Proteomic analyses of male contributions to honey bee sperm storage and mating. *Insect Mol Biol* 15: 541–549.

Correna, M. and Robinson, G.E. (2006) Genes of the antioxidant system of the honey bee. *Insect Mol Biol* 15: 687–701.

Cristino, A.S., Nunes, F.M., Lobo, C.H., Bitondi, M.M., Simoes, Z.L., Costa, L.F., et al. (2006) Caste development and reproduction – a genome-wide analysis of hallmarks of insect eusociality. *Insect Mol Biol* 15: 703–714.

De Graaf, D.C. (2006) Genomic and transcriptional analysis of protein heterogeneity of the honeybee venom allergen Api m 6. *Insect Mol Biol* 15: 577–581.

Eisenhardt, D. and Leboulle, G. (2006) The PKA-CREB system in the honeybee. *Insect Mol Biol* 15: 551–561.

Elsik, C., MacKey, A.J., Reese, J., Mishina, N. and Weinstock, G.M. (2006) Creating a honey bee consensus gene set. *Genome Biol* (in press).

Estoup, A., Garnery, L., Solignac, M. and Cornuet, J.M. (1995) Microsatellite variation in honey bee (*Apis mellifera L.*) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* 140: 679–695.

Evans, J.D. and Wheeler, D.E. (2000) Expression profiles during honeybee caste determination. *Genome Biol* 2, RESEARCH0001.

Evans, J.D., Aronstein, K., Chen, Y.P., Hetru, C., Imler, J.-L., Jiang, H., et al. (2006) Immune-related genes and honey bee disease responses. *Insect Mol Biol* 15: 645–656.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194.

Ewing, B., Hillier, L., Wendel, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175–185.

Havlík, P., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., et al. (2004) The Atlas genome assembly system. *Genome Res* 14: 721–732.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129–149.

Honey Bee Genome Sequencing Consortium (2002) Proposal for the Sequencing of a New Target Genome: White Paper for a Honey Bee Genome Project www.genome.gov/Pages/Research/Sequencing/SeqProposals/HoneyBee_Genome.pdf.

Honey Bee Genome Sequencing Consortium (2005) Upgrading the honey bee genome sequence white paper. white.papers. genome.gov/Pages/Research/Sequencing/SeqProposals/HoneyBeeSeq-AddCoverage.pdf.

Honey Bee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* (in press).

Hunt, G.J. and Page, R.E., Jr (1995) Linkage map of the honey bee, *Apis mellifera*, based on RAPD markers. *Genetics* 139: 1371–1382.

Kucharski, R. and Maleszka, R. (2002) Evaluation of differential gene expression during behavioral development in the honeybee using microarrays and northern blots. *Genome Biol* 3, RESEARCH0007.

Kunieda, T., Fujiyuki, T., Kucharski, R., Foret, S., Ohashi, K., Takeuchi, H., et al. (2006) Unique characteristics of the honeybee genes for carbohydrate-metabolizing enzymes as revealed by the genome annotation. *Insect Mol Biol* 15: 563–576.

Liu, Q., MacKey, A.J., Roos, D.S. and Pereira, F.C.N. (2006) Consensus eukaryotic gene prediction by latent variable modeling. * Genome Res* (in press).

Maleszka, R. (2000) Molecules to behaviour in the honeybee – the emergence of comparative neurogenomics. *Trends Neurosci* 23: 513–514.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., et al. (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res* 11: 27–35.

Mengin, O., Lewis, C., Holt, R.A., Birney, E. and Collins, F.H. (2004) The *Anopheles gambiae* genome: an update. *Trends Parasitol* 20: 49–52.

Pennisi, E. (2001) Insects rank low among genome priorities. *Science* 294: 1261–1262.

Qin, X., Evans, J.D., Aronstein, K., Murray, D. and Weinstock, G.M. (2006) Genome sequences of the honey bee pathogens *Paenibacillus larvae* and *Ascosphaera apis*. *Insect Mol Biol* 15: 715–718.

Richards, S., Liu, Y., Bettencourt, B.R., Hradyceky, P., Letovsky, S., Nielsen, R., et al. (2005) Comparative genome sequencing of *Drosophila* pseudobscura: chromosomal, gene, and cis-element evolution. *Genome Res* 15: 1–18.

Robinson, G.E. (1999) Integrative animal behaviour and sociogenomics. *Trends Ecol Evol* 14: 203–205.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10: 516–522.

Solignac, M., Vautrin, D., Mougel, E., Baudry, E., Garnery, L., Loiseau, A., et al. (2003) Five hundred and fifty microsatellite markers for the study of the honeybee (*Apis mellifera*). *Genome Biol Ecol Notes* 3: 307–311.

Solignac, M., Vautrin, D., Baudry, E., Mougel, F., Loiseau, A. and Cornuet, J.M. (2004) A microsatellite-based linkage map of the honeybee, *Apis mellifera L.*. *Genetics* 167: 253–262.

Solignac, M., Zhang, L., Mougel, F., Li, B., Vautrin, D., Monnerot, M., et al. (2006) The genome of *Apis mellifera*: dialog between mapping and sequencing. *Genome Biol* (in press).

Takeuchi, H., Fujiyuki, T., Shirai, K., Matsuo, Y., Kamikouchi, A., Fujinawa, Y., Kato, A., Tsujimoto, A. and Kubo, T. (2002) Identification of genes expressed preferentially in the honeybee mushroom bodies by combination of differential display and cDNA microarray. *FEBS Lett* 513: 230–234.

Wheeler, D.E., Buck, N. and Evans, J.D. (2006) Expression of insulin pathway genes during the period of caste determination in *Apis mellifera*. *Insect Mol Biol* 15: 597–602.

Whitfield, C.W., Band, M.R., Bonaldo, M.F., Kumar, C.G., Liu, L., Pardinhas, J.R., et al. (2002) Annotated expression sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res* 12: 555–566.

Whitfield, C.W., Cziko, A.M. and Robinson, G.E. (2003) Gene expression profiles in the brain predict behavior in individual honey bees. *Science* 302: 296–299.

Whitfield, C.W., Behura, S., Berlocher, S., Clark, A.G., Johnston, S., Sheppard, W.S., et al. (2006) Thrice out of Africa: nature and human-facilitated expansions of *Apis mellifera*. *Science* (in press).

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., et al. (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306: 1937–1940.

Zou, Z., Lopez, D., Kanost, M., Evans, J.D. and Jiang, H. (2006) Comparative analysis of serine proteases-related genes in the honey bee genome: possible involvement in embryonic development and innate immunity. *Insect Mol Biol* 15: 603–614.