

Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera

Monica C. Munoz-Torres¹, Justin T. Reese², Christopher P. Childers¹,
Anna K. Bennett¹, Jaideep P. Sundaram¹, Kevin L. Childs², Juan M. Anzola²,
Natalia Milshina² and Christine G. Elsik^{1,2,*}

¹Department of Biology, Georgetown University, Washington, DC 20057 and ²Department of Animal Science, Texas A&M University, College Station, TX 77843, USA

Received August 16, 2010; Revised October 22, 2010; Accepted October 25, 2010

ABSTRACT

The Hymenoptera Genome Database (HGD) is a comprehensive model organism database that caters to the needs of scientists working on insect species of the order Hymenoptera. This system implements open-source software and relational databases providing access to curated data contributed by an extensive, active research community. HGD contains data from 9 different species across ~200 million years in the phylogeny of Hymenoptera, allowing researchers to leverage genetic, genome sequence and gene expression data, as well as the biological knowledge of related model organisms. The availability of resources across an order greatly facilitates comparative genomics and enhances our understanding of the biology of agriculturally important Hymenoptera species through genomics. Curated data at HGD includes predicted and annotated gene sets supported with evidence tracks such as ESTs/cDNAs, small RNA sequences and GC composition domains. Data at HGD can be queried using genome browsers and/or BLAST/PSI-BLAST servers, and it may also be downloaded to perform local searches. We encourage the public to access and contribute data to HGD at: <http://HymenopteraGenome.org>.

INTRODUCTION

The insects of the order Hymenoptera form an extremely diverse group with over 115 000 described species comprising ~10% of the species diversity on the earth. The largest described family in the order (Ichneumonidae) contains more species than all species of birds and mammals combined (1). Some of these species feed solely on plants, some are entomophagous, and some feed on both. Entomophagous species can be parasitic, predatory or a number of stages in between, and their lifestyles range from solitary to complex social communities. This group is essential to sustaining plant diversity through pollination, seed dispersion and protection from predators, as well as essential to regulating population size of other arthropods (2). The order Hymenoptera includes sawflies, bees, ants and wasps, and together they directly affect human health and agriculture through diverse roles such as pollinators, pests and parasitoids. In recent years a large amount of genomic information has become available for several species in the order, including honey bees, parasitoid wasps and ants. The Hymenoptera Genome Database (HGD) strives to effectively integrate these data to convert all resources into knowledge that will benefit both biological and agricultural interests within the research community.

HGD currently hosts data and browsing tools for the genomes of social hymenopterans like honey bees (BeeBase, <http://BeeBase.org>) and ants, as well solitary parasitoid jewel wasps (NasoniaBase, <http://Nasonia>

*To whom correspondence should be addressed. Tel: +1 202 687 4485; Fax: +1 202 687 4662; Email: ce75@georgetown.edu
Present addresses:

Justin T. Reese, Department of Biology, Georgetown University, Washington, DC 20057, USA.

Kevin L. Childs, Department of Plant Biology, Michigan State University, 166 Plant Biology Building, East Lansing, MI 48824, USA.

Juan M. Anzola, IntelligentMDx, 19 Blackstone Street, Cambridge, MA 02139, USA.

Natalia Milshina, Department of Genetics, Southwest Foundation for Biomedical Research, 7620 NW Loop 410, San Antonio, TX 78227, USA.

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Base.org). Honey bees (*Apis mellifera*) are a model species for social behavior and are essential to global ecology as pollinators. Uncovering their genome (3) provided us with the first genome-wide insights into the social life of insects (4). The United States Department of Agriculture (USDA) estimates that honey bees accomplish nearly 80% of insect pollination of crops, contributing over \$14 billion to US agriculture (5). The Nasonia Genome Working Group published the draft sequences of three parasitoid wasps of the genus *Nasonia*, namely *Nasonia vitripennis*, *N. giraulti* and *N. longicornis*. (6). Parasitic wasps have been subjects of ecological, evolutionary, genetic and developmental research for nearly 100 years and are one of the largest and most important groups of Hymenoptera, with more beneficial insects to humans than any other group. They regulate insect pest populations through close and specific associations with their hosts, helping to maintain ecological balance in terrestrial ecosystems (7). Ants are found in most ecosystems on the earth, from the Arctic Circle to the southern tip of South America. They can establish relationships as mutualists with plants and fungi, become resource species as soil turners and providers of food for vertebrates and, in some cases, modify arthropod diversity and ecosystem architecture through mixed communities of two or more interacting ant species (2). Their sophisticated social organization is almost unique amongst living creatures and has often been compared with human societies (8). These and many other aspects of their natural history make them ideal systems to study foraging, caste determination, evolution of social behavior and mutualism, among other topics. HGD includes the genomes of five species of ants, the leaf cutter ant *Atta cephalotes* (Suen, G., *et al.*, manuscript in review), the Florida carpenter ant *Camponotus floridanus* (9), the jumper ant *Harpegnathos saltator* (9), the Argentine ant *Linepithema humile* (Smith, C.D., *et al.*, manuscript in review) and the red harvester ant *Pogonomyrmex barbatus* (Smith, C.R., *et al.*, manuscript in review). The combination of all of these genomes into a single web-accessible resource is a cost effective approach that allows researchers to leverage the genome information. The resources available at HGD facilitate cross-species comparisons and are the starting point to enhancing our understanding of the biology of agriculturally important Hymenoptera species through genomics.

HGD is an order-wide database that brings BeeBase, NasoniaBase and the Ant Genomes Portal under a single umbrella. We first developed BeeBase as a comprehensive data resource for the honey bee genome assembly. BeeBase allowed a dispersed research community to perform manual gene model annotations (10). Genome browsers developed using GBrowse (11), a GMOD tool, provide views of genome databases in an interactive web-based interface that allows manipulating the display of annotations. BLAST and custom PSI-BLAST (12) servers allow researchers to search through scaffolds, contigs, unscaffolded small contigs, and several gene prediction sets. Sequencing of three *Nasonia* genomes generated the need for a similar set of tools, which were made available at NasoniaBase. We improved the biologist-friendly annotation submission web pages built

for BeeBase by developing software that enabled biologists to use the Apollo Annotation Editor (13). Users now connect remotely to our databases through an Apollo client on their personal computers, which grants simultaneous access to all gene evidence for a particular region facilitating annotation of gene models (6). As it was done for BeeBase and NasoniaBase, we have built browsers for the genomes of five species of ants using GBrowse. Genomic resources are essential to evolutionary, ecological and genetic studies that range from genome-wide surveys of an organism's response to environmental cues to identifying markers for breeding of traits of agronomic importance (14). HGD synergistically integrates these genomic resources with the help of more than 400 scientists in nearly 15 countries who peruse our databases to both retrieve information and contribute their own data through manual annotation submissions and interactive commentary.

DATA ORGANIZATION AND MINING TOOLS

HGD utilizes the Chado schema (15), implemented within the PostgreSQL database management system. Each species and assembly version is maintained in a separate Chado database. Users may access the assembled genomes and all available biological evidence supporting gene models, chromosomal organization, etc. Below are the details of how HGD is organized and the tools available to the research community.

Content management

HGD uses Drupal (<http://drupal.org>) for content management. Drupal, distributed under the GNU General Public License, offers a simple and flexible way to design, manage and organize a wide variety of content, which is maintained in a database instead of stored as static pages. The modular nature of Drupal allows HGD curators to precisely determine where and how content is displayed. Additionally, pages generated through external CGI scripts (e.g. CMAP, BLAST) can be made to seamlessly look like Drupal generated pages through the use of HTML iFrames. This management system allows HGD administrators to delegate organization of certain portions of the database to selected users because preconfigured tools facilitate non-programmers to access and edit content.

Genome Browsers

Genome browsers implemented with GBrowse (11) allow users to access the assembled genome and all available biological evidence supporting gene models, chromosomal organization, etc. BeeBase has several genome browsers for different assemblies, and different coordinate systems that allow users to compare data from different sources. We have developed Perl scripts to automatically port features between all coordinate systems. Sequence scaffolds were assembled into chromosomes at Baylor College of Medicine-Human Genome Sequencing Center (BCM-HGSC) using markers and physical maps. We have provided viewers with chromosome coordinates to be

compatible with other genome browsers [NCBI (16), UCSC Genome Browser (17)], as well as a scaffold browser. Scaffolds and gaps within and between scaffolds can be viewed on the chromosome-based genome browser. To improve chromosome assemblies, assigned scaffolds of the BCM-HGSC honey bee assembly, unassigned scaffolds and unscaffolded contigs were manually joined to form superscaffolds (18), and these are also available on genome browsers. We mapped the official honey bee gene set to the latest assembly (version 4) using automated and manual methods, and the assembly 4 genome browser shows release 2 of the official gene set, which incorporates manual annotations performed by members of the sequencing consortium (10). Through these genome browsers researchers can access tracks with information containing honey bee ESTs, homologs, GC compositional domains and genetic map markers. NasoniaBase genome browsers allow access to scaffolds in version 1 of the *N. vitripennis* assembly (Nvit_1.0) and include tracks with the official and all other predicted gene sets, alignments of the *N. vitripennis* genome to *Nasonia* ESTs, small RNA sequences and protein homologs.

BLAST and PSI-BLAST servers

DNA data sets on the BeeBase BLAST website include scaffolds, contigs, unscaffolded contigs, unassembled reads and repeat reads. Protein data sets included the following gene predictions: Ensembl (19), RefSeq (20), Fgenesh, Fgenesh++ (21,22), an Evolutionary Conserved Core set, a *Drosophila* ortholog set, and the Honey Bee Official Gene Set (3), a combination of manual annotations and consensus gene predictions created using GLEAN (23). BeeBase also supports PSI-BLAST searches of special protein data sets that combine the GenBank non-redundant protein set with honey bee predicted genes. The PSI-BLAST site has enabled researchers to identify divergent paralogs that may not be easily identified using other BLAST programs. NasoniaBase users may query the genome using BLAST against scaffolds in Nvit_1.0, RefSeq or *ab initio* predictions in the official gene set, and RNA and protein prediction data sets generated with GLEAN, Augustus (24), Fgenesh, GENEID (25) and SGP2 (26). Functionality on the BLAST sites includes retrieving sequences from HGD or GenBank by clicking on identifiers in the BLAST results. In addition, links are included to allow viewing BLAST hits on the genome browsers. Tracks for the BLAST hits on the genome browsers are maintained using cookies, so that the user may view multiple BLAST hits on the browser at the same time, and or review the hits in a later session.

Comparative map viewer

We have generated a honey bee comparative map viewer (CMAP) using CMAP (<http://gmod.org/wiki/Cmap>) which displays updated versions of the Hunt (27) and Solignac (28) genetic linkage maps. The CMAP viewer connects both linkage maps directly to each other and to the genome assembly. Clicking on map markers brings users to GBrowse.

Apis mellifera pest and pathogens Database

Diseases and pests are contributing factors to the disappearance of pollinator insects on a global scale. Better understanding the relationship between honey bees and their pathogens through genomics will lead to improved management practices. We have computed gene predictions, developed genome browsers and built BLAST databases for three sequenced honey bee pathogens: the fungal parasites *Ascospaera apis* (29) and *Nosema cerana* (30) and the bacterium *Paenibacillus larvae* (29). Homolog analysis and computed functional annotation will be completed for these species and resources for the genome of the mite *Varroa destructor* will soon be added to the database.

COMMUNITY GENE ANNOTATION

The role of a Model Organism Database (MOD) is to maintain a single reference gene set, map gene features to new assemblies and make these data easily accessible for the research community. A crucial step to fulfilling this goal is to accumulate and curate data generated by this community, which in turn is only possible if there is an active group of scientists involved. HGD has supported research contributions from an extensive community from almost 70 institutions in 13 countries, constituting what is perhaps the largest dispersed manual annotation effort reported. To date, members of the honey bee and *Nasonia* research communities have annotated over 4000 gene models (3,6). Researchers used the tools available at HGD to edit gene models of the official gene sets, identify genes missing from the consensus gene sets, and annotate functions using homology and orthology. The HGD team has developed scripts to process this community annotation data and submit it to GenBank (31) as features on the assemblies.

Initial community annotation efforts involved pasting modified sequences onto websites. Later our team deployed a Chado database and configured Apollo to allow remote access to gene evidence from the Chado database from the user's computer. Users are provided with Apollo configuration files that include host specific information to connect to the HGD server. For example, researchers can start their Apollo client using a RefSeq accession number or a scaffold identifier and connect directly to NasoniaBase. Users will access the selected scaffold and retrieve tracks of evidence such as gene predictions obtained with NCBI's GNOMON pipeline, Fgenesh++ and GLEAN, as well alignments of EST/cDNA from *N. vitripennis*, *N. giraulti* and fire ants, and protein homologs from all Metazoans in Swissprot (32), FlyBase (33) and additional insect RefSeq collections available at NCBI. Completed annotations are saved as Chado-XML files using Apollo, then uploaded to NasoniaBase through a submission website. We are in the process of implementing a write-back connection, so that modified annotations may be transferred directly from individual Apollo clients to our Chado database.

ADDITIONAL RESOURCES

Wiki pages

HGD offers interactive resources available to the Hymenoptera research community at large to facilitate the distribution of information and ideas. Individual Wiki sites developed for BeeBase and NasoniaBase can be accessed at: <http://beebase.org/?q=wiki> and <http://nasoniabase.org/?q=wiki>, respectively. We will soon make Wiki sites available for individual species of ants in the Ant Genomes Portal. Users are expected to interact as they would on message boards, discussing subjects directly related to the genome sequencing projects and creating a new wiki page per discussion topic. Wikis can be used to post topics on 'Discussion Board' pages, create or modify annotation information data on gene(s) of interest, communicate news to the research community, and share protocols, images, job postings, etc. Wiki sites are populated with pages that allow for subscribed researchers to add or edit the content. Subscription is obtained after verification of identity and nature of the association to the Hymenoptera research community, in an attempt to prevent spam or electronic attacks. The information does not need to be error-proof nor in its final stages of revision in order to be posted. In fact the idea is to offer the possibility to dynamically update all content, allowing the community to keep the information as up to date as possible. Wiki pages also offer the community an opportunity to post questions, suggestions for obtaining community resources and ideas to develop for future funding. A comprehensive list of names and affiliations of contributing researchers is also available. Other than basic punctuation marks, formatting from text editing software becomes plain-text in a Wiki editing window. Throughout the site, users will find it easier to add and/or modify content when they first become familiar with WikiMarkup, the language of Wiki pages. Simplified tutorials are available in the 'Add Content to This Wiki' sections.

Supplementary data from genome consortia

A comprehensive list of all data sets generated by the community is available for each species in HGD. Every version of the official gene sets, all versions of the genome assemblies, and all text, tables and figures submitted as supporting data for publication of each genome project are found under the 'Genome Consortium data sets' section.

FUTURE DIRECTIONS

To make HGD an even more comprehensive genome database with graphical search tools and viewers, we are currently working on developing additional tools such as cross-species comparisons in the form of synteny browsers, ortholog and homolog families among Hymenoptera, and pre-computed ultraconserved sequences at various evolutionary distances within the Hymenoptera. We also plan to incorporate web accessible search and graphical interfaces for QTL, SNP, haplotype

and expression data, when available, for each species. We are generating searchable gene model pages using a novel Ruby on Rails application that we developed to provide sequences and graphical views of gene intron/exon boundaries, splice variants, and protein domains and motifs, functional annotations, database cross references and links to additional related resources. Finally, we plan to continue helping in the development of new controlled vocabularies required to describe Hymenoptera genomics data computationally, contributing to the Gene Ontology project (34).

FUNDING

Work on BeeBase was supported by: (i) Special Cooperative Agreements between the Texas Agricultural Experiment Station and two United States Department of Agriculture, Agriculture Research Service Laboratories: the Baton Rouge Honey Bee Breeding, Genetics and Physiology Laboratory and the Weslaco Bee Research Laboratory; (ii) a supplement to National Institutes of Health grant (5-P41-HG000739-13); (iii) the Texas Agricultural Experiment Station; (iv) gifts from Golden Heritage Foods and Sioux Honey Association. Addition of NasoniaBase and further development of HGD was funded by the USDA National Institute of Food and Agriculture grant (2008-35302-18804); start-up funds to CGE from Georgetown University. Funding for open access charge: USDA National Institute of Food and Agriculture grant (2008-35302-18804).

Conflict of interest statement. None declared.

REFERENCES

1. Sharkey, M.J. (2007) Phylogeny and classification of Hymenoptera. *Zootaxa*, **1668**, 521–548.
2. LaSalle, J. and Gauld, I.D. (1993) In LaSalle, J. and Gauld, I.D. (eds), *Hymenoptera and Biodiversity*. Wallingford, UK, CAB International, p. 348.
3. Honey Bee Genome Sequencing Consortium. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, **443**, 931–949.
4. Wilson, E.O. (2006) Genomics: how to make a social insect. *Nature*, **443**, 919–920.
5. Hackett, K.J. (2004) Bee benefits to agriculture. *Agr. Res. Mag.*, **52**, 2.
6. Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Beukeboom, L.W., Desplan, C., Elsik, C.G., Gimmelikhuijzen, C.J. *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*, **327**, 343–348.
7. LaSalle, J. (1993) In LaSalle, J. and Gauld, I.D. (eds), *Hymenoptera and Biodiversity*. Wallingford, UK, CAB International, p. 348.
8. Keller, L. and Gordon, E. (2009) *The Lives of Ants*. Oxford University Press Inc., New York, USA.
9. Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C. *et al.* (2010) Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*, **329**, 1068–1071.
10. Elsik, C.G., Worley, K.C., Zhang, L., Milshina, N.V., Jiang, H., Reese, J.T., Childs, K.L., Venkatraman, A., Dickens, C.M., Weinstock, G.M. *et al.* (2006) Community annotation: procedures, protocols, and supporting tools. *Genome Res.*, **16**, 1329–1333.
11. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. *et al.* (2002)

- The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
 13. Lewis,S.E., Searle,S.M., Harris,N., Gibson,M., Lyer,V., Richter,J., Wiel,C., Bayraktaroglu,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
 14. Munoz-Torres,M.C., Sasaki,C., Blackmon,B., Romero-Severson,J. and Werren,J.H. (2010) Development of bacterial artificial chromosome library resources for parasitoid Hymenoptera (*Nasonia vitripennis* and *Nasonia giraulti*: Pteromalidae). *Insect Mol. Biol.*, **19**(Suppl. 1), 181–187.
 15. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
 16. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.
 17. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
 18. Robertson,H.M., Reese,J.T., Milshina,N.V., Agarwala,R., Solignac,M., Walden,K.K. and Elsik,C.G. (2007) Manual superscaffolding of honey bee (*Apis mellifera*) chromosomes 12–16: implications for the draft genome assembly version 4, gene annotation, and chromosome structure. *Insect Mol. Biol.*, **16**, 401–410.
 19. Curwen,V., Eyraes,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
 20. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, Vol. 37, pp. D32–D36.
 21. Salamov,A.A. and Solovyev,V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516–522.
 22. Solovyev,V. (2007) In Balding,D.J., Bishop,M. and Cannings,C. (eds), *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester, England; Hoboken, NJ, pp. 97–159.
 23. Elsik,C.G., Mackey,A.J., Reese,J.T., Milshina,N.V., Roos,D.S. and Weinstock,G.M. (2007) Creating a honey bee consensus gene set. *Genome Biol.*, **8**, R13.
 24. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl. 2), ii215–225.
 25. Blanco,E., Parra,G. and Guigo,R. (2007) Using geneid to identify genes. *Curr. Protoc. Bioinformatics*, Chapter 4, Unit 4.3.
 26. Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigo,R. (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.
 27. 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.
 28. Solignac,M., Mougel,F., Vautrin,D., Monnerot,M. and Cornuet,J.M. (2007) A third-generation microsatellite-based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. *Genome Biol.*, **8**, R66.
 29. Qin,X., Evans,J.D., Aronstein,K.A., Murray,K.D. and Weinstock,G.M. (2006) Genome sequences of the honey bee pathogens *Paenibacillus* larvae and *Ascosphaera apis*. *Insect Mol. Biol.*, **15**, 715–718.
 30. Cornman,R.S., Chen,Y.P., Schatz,M.C., Street,C., Zhao,Y., Desany,B., Egholm,M., Hutchison,S., Pettis,J.S., Lipkin,W.I. *et al.* (2009) Genomic analyses of the microsporidian *Nosema ceranae*, an emergent pathogen of honey bees. *PLoS Pathog.*, **5**, e1000466.
 31. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–51.
 32. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–148.
 33. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–559.
 34. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.