Published online 17 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D793–D800***
doi: 10.1093/nar/gkv1208

Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine

Christine G. Elsik[1,2,3,*], Aditi Tayal[1], Colin M. Diesh[1], Deepak R. Unni[1], Marianne L. Emery[2], Hung N. Nguyen[3] and Darren E. Hagen[1]

[1]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA, [2]Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA and [3]MU Informatics Institute, University of Missouri, Columbia, MO 65211, USA

Received September 22, 2015; Revised October 26, 2015; Accepted October 27, 2015

ABSTRACT

We report an update of the Hymenoptera Genome Database (HGD) (http://HymenopteraGenome.org), a model organism database for insect species of the order Hymenoptera (ants, bees and wasps). HGD maintains genomic data for 9 bee species, 10 ant species and 1 wasp, including the versions of genome and annotation data sets published by the genome sequencing consortiums and those provided by NCBI. A new data-mining warehouse, HymenopteraMine, based on the InterMine data warehousing system, integrates the genome data with data from external sources and facilitates cross-species analyses based on orthology. New genome browsers and annotation tools based on JBrowse/WebApollo provide easy genome navigation, and viewing of high throughput sequence data sets and can be used for collaborative genome annotation. All of the genomes and annotation data sets are combined into a single BLAST server that allows users to select and combine sequence data sets to search.

INTRODUCTION

The insect order Hymenoptera—which includes bees, ants and wasps—is one of the most species-rich insect orders, with around 150 000 described species (1). Their diverse roles, such as pollinators of flowering plants (e.g. *Apis* and *Bombus*), parasitoids of other insects (e.g *Nasonia*) and cultivators of fungi (e.g *Atta* and *Acromyrmex*) make them vital to natural and agricultural ecosystems. For example, the value of crops in the United States attributed to pollination by the European honey bee, *Apis mellifera*, alone has been estimated at $16.4 billion (2). In addition to the economic benefits of crop pollination and natural pest control, hymenopteran insects have provided humans with products such as honey and beeswax for thousands of years. The order Hymenoptera also includes pests, such as fire ants (e.g *Solenopsis invicta*), which attack small mammals, and leaf cutting ants (e.g. *Atta* and *Acromyrmex*), which damage vegetation. Hymenopteran insects have been subjects of study not only because of their beneficial or detrimental impacts on agriculture and human welfare, but also because of their fascinating lifestyles. They have been used to address questions regarding the biology of social insects, symbiosis, parasitoidal relationships and invasive species. Ants, bees and wasps serve as premier models for investigating the evolution of eusociality, which refers to social systems in which many non-reproductive individuals provide support for the colony; eusociality has evolved independently in nine hymenopteran lineages (3). The first hymenopteran genome to be sequenced was that of *A. mellifera* in 2006 (4). Since then, genome sequences of over 20 hymenopteran species have been generated in projects that were aimed at addressing questions most pertinent to the subject species. The Hymenoptera Genome Database (HGD) (http://HymenopteraGenome.org), first reported in Nucleic Acids Research in 2011 (5), strives to make genomic data of hymenopteran insects easily accessible so that researchers can continue to leverage it to understand the biology of each organism, and to perform comparative studies to investigate processes that are common or diverse across the Hymenoptera order.

HGD consists of three divisions, BeeBase, NasoniaBase and the Ant Genomes Portal, and covers a range of hymenopteran species that have diverged from a common ancestor approximately 170 million years ago (6), with the two most closely related species (the ants *Acromyrmex echinatior* and *Atta cephalotes*) having diverged as recently as 10 million years ago (Figure 1). Since the first report (5), the number of species represented in HGD has more than doubled (Table 1). The seven genomes included in the initial report were those of European honey bee (*A. mellifera*) (4), parasitoid jewel wasp (*Nasonia vitripennis*) (7) and five ant species – the leaf-cutter ant (*Atta cephalotes*)

*To whom correspondence should be addressed. Tel: +1 573 884 7422; Fax: +1 573 882 6527; Email: elsikc@missouri.edu

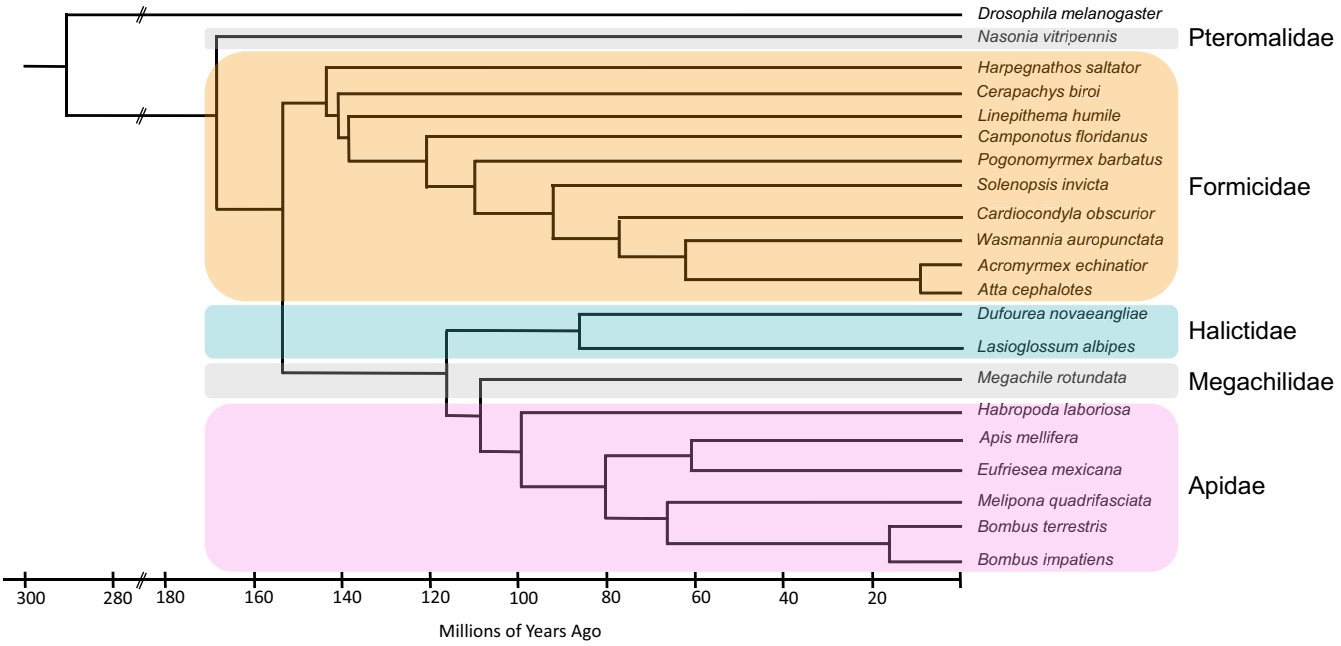

Figure 1. Phylogenetic relationships and approximate divergence times of species included in HGD. Families Apidae, Megachilidae and Halictidae are bees; Formicidae are ants; Pteromalidae are wasps. This figure was redrawn from (6,54–55).

(8), the Florida carpenter ant (*Camponotus floridanus*) (9), the jumping ant (*Harpegnathos saltator*) (9), the Argentine ant (*Linepithema humile*) (10) and the red harvester ant (*Pogonomyrmex barbatus*) (11). In the past five years we have incorporated the genomes of 13 additional hymenopteran species. These are the Panamanian leaf cutter ant (*Acromyrmex echinatior*) (12), the ant *Cardiocondyla obscurior* (13), the clonal raider ant (*Cerapachys biroi*) (14), the red fire ant (*Solenopsis invicta*) (15), the little fire ant (*Wasmannia auropunctata*), the common eastern bumble bee (*Bombus impatiens*) (16), the buff-tailed bumble bee (*Bombus terrestris*) (16), alfalfa leafcutting bee (*Megachile rotundata*) (17), an orchid bee (*Eufriesa mexicana*) (17), a stingless bee (*Melipona quadrifasciata*) (17), the southeastern blueberry bee (*Habropoda labrosia*) (17) and two halictid bees (*Lasioglossum albipes* (18), *Dufourea novaeangliae* (17)). There are varying amounts and types of genomic resources for each species. The minimum set of data for each species is an assembled reference genome and at least one set of gene predictions, which may be a consortium gene set or an annotation release from NCBI. Our policy is to incorporate a genome after it has been published, unless we receive a request from an investigator prior to publication. Some species have additional information submitted by researchers, such as RNAseq alignments and genome resequencing alignments from variation studies. Since the initial publication, tools available at HGD have been used to support reannotation of *A. mellifera* (19), an ant genome comparative analysis (20), and community annotation of the two bumble bee genomes (16), *C. obscurior* (13) and *W. auropunctata*, which is not yet published.

GENOME BROWSERS AND COMMUNITY ANNOTATION

A major focus of HGD has been community genome annotation. Since the previous HGD publication, we supported community annotation of the two *Bombus* genomes by computing homolog and RNAseq alignments, creating consensus gene sets, providing genome browsers (GBrowse (21)) with several gene prediction sets, and providing the Apollo annotation tool (22) with direct connections to Chado databases (23). More recently we replaced the HGD community annotation platform based on Apollo/Chado with WebApollo (24), a browser-based genome annotation tool that leverages the JBrowse (25) genome browser.

WebApollo has provided significant improvements over the previous Apollo/Chado annotation platform, both from data management and user perspectives. From the data management standpoint, loading data tracks, and collecting and exporting manual annotation data is simplified. From the user standpoint, WebApollo's integration with JBrowse allows fast genome navigation and viewing individual read alignments from high-throughput sequencing data sets. The user interface is highly configurable. Manual annotations are saved dynamically, so they are immediately available for viewing by other registered users. Gene model change histories are maintained; users can undo changes and administrators can view histories.

We have deployed WebApollo 2.0 for all species at HGD, providing publicly available JBrowse-based viewing of genome annotations. The WebApollo manual annotation tools and manual annotation data track are available to registered users for the species for which open annotation has been requested. We are actively annotating *A. mellifera* genes, and we provide tracks that allow users to focus on priority genes, such as loci that are in disagreement across

Table 1. Species and data sets in HGD. JBrowse and BLAST is provided for all of the species listed

Species	New	NCBI Assembly Name[a]	HGD or Consortium Assembly Name	NCBI Annot. Release	OGS Name	Hymenop-teraMine	Ortho -DB	Ensembl Metazoa	Uni-Prot	Inter-Pro	GO	SRA	Pub-Med
Acromyrmex echinatior	√	Aech_3.9*	Aech_2.0	100	aech_OGSv3.8	√	√		√	√	√		
Apis mellifera		Amel_4.5	Amel_4.5	102	amel_OGSv3.2	√	√	√	√	√	√	√	√
Atta cephalotes		Attacep1.0*	Acep_1.0	100	acep_OGSv1.2	√	√	√	√	√	√		
Bombus impatiens	√	BIMP_2.0	Bimp_2.0	101	bimp_OGSv1.0	√	√		√	√	√		√
Bombus terrestris	√	Bter_1.0	Bter_1.0	101	NCBI Annotation	√	√		√	√	√		√
Camponotus floridanus		CamFlo_1.0*	Clfo_3.3	100	cflo_OGSv3.3	√	√		√	√	√		
Cardiocondyla obscurior	√		Cobs_1.4		cobs_OGSv1.4	√							
Cerapachys biroi	√	CerBir1.0*	Cbir.assembly.v3.0	100	armyant.OGS.V1.8.6								
Dufourea novaengliae	√		Dufourea_novaeangliae		Dufourea_novaeangliae_v1.1								
Eufriesa mexicana	√		Eufriesea_mexicana.v1.0		Eufriesea_mexicana_v1.1								
Habropoda labrosia	√		Habropoda_laboriosa.v1.0		Habropoda_laboriosa_v1.2								
Harpegnathos saltator		HarSal_1.0*	Hsal_3.3	100	hsal_OGSv3.3	√	√		√	√	√		
Lasioglossum albipes	√		Lalb_v2		Lalb_OGS_v5.42	√	√						
Linepithema humile		Lhum_UMD_V04*	Lhum_1.0	100	lhum_OGSv1.2	√	√		√	√			
Megachile rotundata	√	MROT_1.0	Mrot_1.0	101	Megachile_rotundata_v1.1								
Melipona quadrifasciata	√		v1.0		Melipona_quadrifasciata_v1.1								
Nasonia vitripennis		Nvit_2.1*	Nvit_1.0	101	OGSv2, nvit_OGSv1.2	√	√	√	√	√	√		√
Pogonomyrmex barbatus		Pbar_UMD_V03*	Pbar_1.0	100	pbar_OGSv1.2	√	√		√	√	√		
Solenopsis invicta	√	Si_gnG*	Sinv_1.0	100	sinv_OGS2.2.3	√	√	√	√	√	√		√
Wasmannia auropunctata	√	wasmannia.A_1.0*	Waur_1.0	100		√							

[a]An asterisk (*) next to an NCBI Assembly name indicates that the NCBI reference genome assembly differs from the consortium assembly, and HGD provides JBrowse for both assemblies.

gene prediction sets. For example, tracks include OGSv3.2 genes that are split and merged compared to RefSeq and the older gene set, OGSv1. RNAseq read alignment tracks are especially helpful for evaluating intron splice sites, or determining whether a gene should be split or merged. In addition to viewing tracks showing priority genes for annotation, users may navigate to genes of interest by performing a search with a protein homolog or cDNA sequence using the WebApollo built-in BLAT feature or the HGD BLAST server (described below).

DATA MINING

We have developed a new data warehouse called HymenopteraMine using the InterMine data warehousing system (26). InterMine was originally developed for FlyMine (27), and now has become widely used for other model organism databases (e.g. (28–33)). The InterMOD consortium, a collaboration that includes the development teams for InterMine and five model organism databases, has worked to provide a platform for cross-species analyses through FlyMine, MouseMine, RatMine, ZebrafishMine, YeastMine and WormMine (34). The efforts of the InterMOD Consortium to ease the integration of genomic and functional data for cross-species comparison will prove to be especially valuable for research in non-model organisms, which relies heavily on comparative analyses. These efforts have made InterMine well suited to host the species of HGD. Integrating data across species in HymenopteraMine allows users to leverage cross-species data sets via orthologous relationships. The InterMine platform also eases the incorporation of external data sets, facilitating integration of the genome data in HGD with data from widely used insect genomic resources, such FlyBase (35), FlyMine (27), OrthoDB (36) and EnsemblMetazoa (37).

Another strength of the InterMine platform is its ability to resolve and manage multiple identifiers for the same entity. A common practice in hymenopteran genome projects has been for the research community to annotate the genome prior to annotation at NCBI. Members of the insect genome consortiums often perform their own automated gene prediction, followed by community manual annotation of gene families of particular interest, with analyses culminating in a consortium publication. Since their original publications, NCBI has annotated many of the hymenopteran genomes currently in HGD, resulting in the existence of at least two sources of genome annotations for these species, the consortium's official gene set (OGS) and the NCBI annotation release. HymenopteraMine addresses the common request among HGD users to provide tables with database cross-references or aliases for gene identifiers in different gene sets of the same species.

HymenopteraMine data sources

Data in HymenopteraMine (Table 1) include genome assemblies and official gene sets developed by genome consortiums for the different species (8–13,15–16,18–19), annotations from NCBI RefSeq (38), Gene Ontology (39) and protein annotations from UniProt (40), protein family and domain assignments from InterPro (41), and *A. mellifera* gene expression information computed from RNAseq data downloaded from the NCBI Sequence Read Archive (SRA) (42). HymenopteraMine includes orthologues and paralogues from OrthoDB (36) and EnsemblMetazoa (37), facilitating comparison across the hymenopteran species and with *Drosophila melanogaster*. We have also acquired database cross references from NCBI for *A. mellifera* genes, and have computed database cross references for the other species by using IntersectBed (43) to identify overlapping gene predictions. Gene aliases for some of the species include the original gene prediction pipeline identifiers, such as those assigned by MAKER (44), or new versus old gene set identifiers, such as those for *A. mellifera* OGSv3.2 and OGSv1.0 (4,19). We used the gene2pubmed file downloaded from the NCBI Gene FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz) (45) to provide links to PubMed (46) abstracts, and used database cross references to link those abstracts to consortium OGS identifiers.

We have incorporated RNAseq data from the NCBI SRA into HymenopteraMine to allow users to investigate gene expression patterns. Our initial focus was to incorporate RNAseq data sets for *A. mellifera*, because at the time of download (September 2014) there were many more Illumina RNAseq data sets for *A. mellifera* than any other hymenopteran species in HGD (160 for *A. mellifera* compared to 31 or less for the other insects). We used the SRAInfo table for *Apis mellifera* to identify mRNA sequence data sets with 100 bp reads generated using an Illumina platform. We downloaded FASTQ files for 107 paired-end read runs and 51 single-end runs, trimmed for adaptors using Fastq-MCF (https://code.google.com/p/ea-utils/wiki/FastqMcf); we eliminated 17 paired-end runs and 10 single-end runs in which adaptors were not found. We then trimmed for quality using DynamicTrim (47). We aligned reads to the *A. mellifera* genome assembly Amel_4.5 using TopHat2 (48) and determined FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and normalized read counts for each expression data set for transcripts in *A. mellifera* OGSv3.2 using cuffquant and cuffnorm, which are part the Cufflinks package (49). We also used CoverageBed (43) to determine raw read counts per transcript, and used the raw counts to compute RPKM (Reads Per Kilobase of transcript per Million mapped reads). HymenopteraMine users can retrieve expression levels along with metadata associated with expression experiments by entering OGSv3.2 gene identifiers, or they can retrieve gene lists based on gene expression constraints. More complex queries can use *A. mellifera* gene expression criteria and obtain data for orthologues, allowing researchers of hymenopteran species without comprehensive RNAseq data sets to leverage the *A. mellifera* expression resources.

HymenopteraMine home page and quick search

The HymenopteraMine home page is accessible from the HGD navigation bar. HymenopteraMine has its own navigation bar that is available on all the HymenopteraMine pages. It provides a tab for the HymenopteraMine home page, a 'Help' tab linking to a tutorial, and tabs for the HymenopteraMine tools.

The HymenopteraMine home page provides a quick search tool and a quick list analysis tool, both of which

accept multiple input data types. For example, a user can search for gene identifiers, transcript identifiers, gene symbols, functional annotation terms and species names. The quick search tool performs a full text search and supports wildcards; it returns results from all of the data sets loaded into HymenopteraMine and can be used to explore the data before performing more complex queries. Searching by species name with the quick search tool provides a list of all objects and data sets for that species. A faceted search tool in the quick search result page allows users to filter the results by category. The results of a quick list search can be used in downstream data mining.

In addition to quick search and list analysis forms, the home page provides tabs for the major data set categories in HymenopteraMine: 'Genes', 'Gene Expression', 'Protein', 'Homology' and 'Function'. Each data category tab provides a list of relevant template queries. Clicking 'More queries' leads to a full list of templates on the 'Templates' page (described below), which is also accessible by clicking 'Templates' in the main HymenopteraMine navigation bar.

HymenopteraMine reports

HymenopteraMine provides a report for each entity in the database. Each report is divided into sections including 'Summary', 'Gene', 'Gene Expression', 'Protein', 'Function', 'Homology', 'Others' or a subset of those categories, depending on the identifier searched. For example, searching a protein identifier would provide sections relevant to proteins. Each section of the report provides information in the form of a table that can be customized and downloaded in various formats. The Summary section of a Gene Report provides gene identifiers, symbols, description, organism, chromosome, strand and other identifiers such as aliases (identifiers from previous OGS versions) and database cross references. The Transcripts section provides information about the gene model (transcripts, exons, coding sequences) and gives a visual representation of the gene model highlighting the structure of the gene. Users can download FASTA-formatted sequences for each type of feature provided in the gene model section. Links are provided to JBrowse and to NCBI gene pages, when applicable. Transcript identifiers are linked to Transcript Reports, which include a Gene Expression section that provides raw read counts, normalized read counts, FPKM and RPKM values for RNAseq data with SRA metadata. The Protein section of the Gene Report includes protein name, accession and length. The protein name/accession numbers are linked to Protein Reports, with more information including protein family, GO annotations, InterPro domains, related publications, protein features, and curated notes from UniProt describing tissue specificity, function and developmental stages. The Function section provides GO annotation with evidence codes from the Biological Process, Molecular Function and Cellular Component ontologies. Clicking on the symbols to the right of each GO term leads to the directed acyclic graph showing all parents and children of the term, developed using the BioJS DAG Viewer (50). Clicking on a GO term provides a report for that term, a list of genes annotated with that term and tables showing relationships to other GO terms. The Homologue section lists homologues in other hymenopteran species and *Drosophila melanogaster*. The 'Other' section provides publications and a list of overlapping features.

HymenopteraMine QueryBuilder

The QueryBuilder allows users to explore the underlying data model and construct queries that integrate data sources in HymenopteraMine. QueryBuilder does not require the user to have any programing experience, but it does require some practice. Investigating the predefined query Templates (described below) is a good way to get started. Users can navigate the hierarchical structure of data objects (classes) and subclasses by clicking on 'Browse the Data Model'. Mousing over the symbol next to each class provides a description. The largest class is 'Bio-Entity', which is divided into 'Protein', 'Protein Domain' and 'Sequence Feature'. The 'Sequence Feature' class is further divided into classes, such as 'Gene' and 'Transcript'. Clicking on a class opens the Model Browser and reveals the class attributes, which can be selected to serve as search constraints or as output columns. For example the class 'Transcript' includes attributes 'Description', 'Length', 'Name', 'DB identifier', 'Gene', 'Organism'. References between classes allow them to be combined within queries. Query construction is initiated by clicking the word 'constrain' next to a class. A box appears allowing the user to enter a constraint identifier; if no identifier is provided, all entities of the class will be searched. If the user is logged in and has already used the List tool (described below), an option will allow using the list to constrain multiple searches. Clicking on 'show' next to a class attribute adds the attribute as an output column. The Query Overview in the right panel shows the query construction. Once construction is complete, the 'Fields selected for output' section below the Model Browser can be used to rearrange column order. The query output can be filtered, sorted, reordered and downloaded in various formats, including XML, GFF3, tab-delimited text, JSON and BED. A detailed QueryBuilder example is provided in Supplementary Data.

HymenopteraMine templates

The Templates page provides a list of easy-to-use predefined queries. The templates range in complexity, and can serve as learning tools to make more complex queries with QueryBuilder.

Some of the template queries were adopted from FlyMine (27), and others were custom developed for special use cases in HymenopteraMine. Clicking on a template provides a query interface that may include pull-down menus and may be pre-populated with an example identifier. Users may perform the query using the default identifier or may enter a different value. For some template queries, users are provided with constraint options that may include numerical operations, such as 'less than', 'greater than' and 'not equal to'. Results are obtained by clicking the 'Show Results' button. Alternatively, clicking 'Edit Query' provides access to the QueryBuilder for query modification. As an example, QueryBuilder may be used to remove an identifier constraint, so that a query can be run on an entire data set at

once. Modified templates can be saved for later use and can be exported in XML format to allow sharing among users.

HymenopteraMine lists

The List tool allows users to create and manage lists of identifiers that can be used in further analysis. The user may enter a list of identifiers into the text box, upload a list as a text file or create a list by saving the results of a query. The pull-down menu in List indicates the objects for which identifiers may be entered. After the list is entered, the database performs a lookup of identifiers and prompts the user to disambiguate duplicate or unresolved identifiers. Logged-in users can save lists to refer to at a later date or to use in QueryBuilder. Before clicking 'Save a list', it is recommended that the user enter a name for the list so it can be easily recognized in later analyses. After a list has been saved, the QueryBuilder and Template queries automatically provide the option to use the list in any constraint with the same data type. For example, the template query 'Gene->Alias', which normally accepts a single gene identifier, will automatically include the option to input a saved list of gene identifiers created using the List tool.

HymenopteraMine regions search

The Genomic Region search page allows users to perform an organism specific coordinate based search for genomic features. Users can paste a list of scaffold identifiers with coordinates into the text area, upload the list as a text file, or use the results of a query. For example, the output of the template query 'All Gene → chromosomal location' could be used in the Regions search after setting the sliding bar, to identity features within a specified distance of all genes.

HGD BLAST USING SEQUENCE SERVER

We have deployed SequenceServer (http://sequenceserver.com) to create a unified BLAST interface that allows users to select and query multiple BLAST databases across the hymenopteran species. BLAST hits are linked to JBrowse viewers based on match coordinates when the search database is a genome assembly. When the search database is a CDS or peptide, BLAST hits are linked to JBrowse viewers based on the hit identifiers. The updated BLAST server is an improvement over the previous HGD BLAST interfaces, which used separate BLAST search pages for BeeBase, NasoniaBase and the Ant Genomes Portal, and did not allow the combination of individual search databases. SequenceServer also provides downloadable tab-delimited or BLAST XML reports and graphical overviews of the matches. The HGD BLAST server is similar to the BLAST server of the Fourmidable Ant Genome Database (51), but HGD provides data sets for a larger number of non-ant species. Fourmidable provides transcriptome assemblies for ant species that do not have genome assemblies, while HGD provides only genome assemblies and gene sets.

GENOME ASSEMBLY AND GENE SET VERSIONS

For each species, the use of a single genome assembly and gene set by different researchers is advantageous, because it facilitates data sharing and comparison across studies. The existence of different assemblies for the same species is a common problem because genome assemblies and annotations are often updated when new data or improved assembly methods become available. Several of the genome assemblies in HGD have been updated. We strive to maintain the most recent assembly for each species and provide older assemblies in archive data set pages. For most of the ant genomes, the assemblies were submitted to NCBI after gene prediction and publication by the consortium. Processing and filtering for contaminants at NCBI has introduced changes to the assembly sequences. As a result, the annotation releases from NCBI and the respective consortium OGS (8–12,14–15) do not map to the same coordinate systems. For *Nasonia vitripennis*, the genome assembly has been through two upgrades at NCBI after the consortium publication (7), and there are two consortium official gene sets, neither of which was generated with the newest assembly. The existence of different gene sets and different assemblies is especially problematic when external data providers use different versions. For example, OrthoDB currently provides orthologs for each ant consortium OGS rather than the NCBI annotations on the latest assemblies; UniProt provides protein information for *Nasonia* OGSv1.2, while OrthoDB provides orthologues for *Nasonia* OGSv2.0, and neither of these gene sets is on the same coordinate system as the most recent NCBI annotation release. HGD attempts to address these conflicts in several ways. We provide multiple JBrowse/WebApollo instances for some species, with multiple assemblies and gene sets available for search using BLAST. HymenopteraMine provides coordinate information and links to JBrowse instances based on the more recent NCBI assemblies, but we also include functional information from external sources and database cross references for consortium gene sets. An ongoing effort at HGD is to migrate consortium gene sets to the updated assemblies.

CITING HGD AND DATA SETS

Cite this article for the use HGD or its divisions (BeeBase, NasoniaBase and Ant Genomes Portal), tools such as HymenopteraMine, JBrowse/WebApollo and BLAST, or our code modifications available on GitHub (https://github.com/elsiklab/). In addition, cite the original genome publication for the use of a consortium genome or gene set downloaded from HGD. A list of genome publications may be found by clicking 'Genome Consortium Publications' in the HGD navigation bar.

FUTURE DIRECTIONS AND CONCLUSIONS

The future goals of HGD are to add annotation resources for new hymenopteran genomes, improve connections with other insect bioinformatics resources, and further develop HymenopteraMine to enhance usability and incorporate new data types. We will continue to expand HGD with new species and additional high throughput sequencing data sets. We will strive to deploy WebApollo and BLAST resources for new genomes as they become available, rather than adhering to a predefined update schedule for those tools, and will rely on the NCBI genomes site, the Fourmidable Ant Genomics Database (51), PubMed searches

and direct communication with researchers to identify new genome assemblies. We will continue to work with researchers to support community genome annotation, and encourage leaders of hymenopteran genome projects to contact us if they would like HGD to host community annotation prior to publication.

Our efforts in comparative genome analysis have been focused on either comparing hymenopteran species to each other or comparing hymenopteran species to model organisms, mainly *Drosophila*. We intend to improve connections to other model organism InterMine instances, and to improve connections to non-model insects via the i5k Workspace@NAL (52). Although our goal to support community annotation overlaps that of the i5k Workspace@NAL, the i5k Initiative is focused on 'orphaned' groups that do not have support for genome hosting (52,53). It was agreed early in i5k planning that HGD would continue to host the genomes of hymenopteran species. However, we have recognized the value of comparative insect genome analysis. To that end we are working with developers of the i5k Workspace@NAL, who will mirror HGD genome browsers and BLAST data sets allowing i5k users to compare Hymenoptera with other insects.

We will continue to modify HymenopteraMine to improve usability. Migrating annotations to upgraded assemblies, and computing gene aliases and database cross-references are perpetual efforts necessary for integrating data in HymenopteraMine. Therefore, we anticipate releasing HymenopteraMine updates on an annual basis rather than each time a new genome becomes available. We will continue to make source code modifications for HymenopteraMine and HGD BLAST available on GitHub (https://github.com/elsiklab/).

HymenopteraMine is the first InterMine instance for a group of non-model insects. With limited funding available for the study of non-model organisms, insect researchers need to leverage the knowledgebase and genomic resources of model species like *Drosophila melanogaster*. Similarly, researchers of hymenopteran insects need to leverage the resources created for the better-funded hymenopteran species such as *A. mellifera*. The InterMine platform accommodates these needs by enabling cross-species data mining using orthology. We have shown that InterMine can provide a solution to data integration and mining needs emerging with non-model insect genome sequencing initiatives such as i5k.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Christopher P. Childers, Anna K. Bennett, Justin T. Reese and Monica C. Munoz-Torres for assistance in some of the early JBrowse instances and web pages. We thank Farida Olden for providing feedback on HymenopteraMine. We also thank Christopher Childers for discussing plans to share data with the i5k Workspace@NAL.

FUNDING

USDA National Institute of Food and Agriculture - Agriculture and Food Research Initiative [2010–65106–20634; 2010–65106–21301] and the University of Missouri. Funding for open access charge: University of Missouri.
Conflict of interest statement. None declared.

REFERENCES

1. Grissell,E. (2010) *Bees, Wasps and Ants - The Indispensable Role of Hymenoptera in Gardens*. Timber Press, Portland, OR.
2. Losey,J.E. and Vaughan,M. (2006) The economic value of ecological services provided by insects. *Bioscience*, **56**, 311–323.
3. Hughes,W.O., Oldroyd,B.P., Beekman,M. and Ratnieks,F.L. (2008) Ancestral monogamy shows kin selection is key to the evolution of eusociality. *Science*, **320**, 1213–1216.
4. Honeybee Genome Sequencing Consortium. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, **443**, 931–949.
5. Munoz-Torres,M.C., Reese,J.T., Childers,C.P., Bennett,A.K., Sundaram,J.P., Childs,K.L., Anzola,J.M., Milshina,N. and Elsik,C.G. (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res.*, **39**, D658–D662.
6. Biewer,M., Schlesinger,F. and Hasselmann,M. (2015) The evolutionary dynamics of major regulators for sexual development among Hymenoptera species. *Front. Genet.*, **6**, 124.
7. Werren,J.H., Richards,S., Desjardins,C.A., Niehuis,O., Gadau,J., Colbourne,J.K. and Nasonia Genome Working GroupNasonia Genome Working Group, Werren,J. H., Richards,S., Desjardins,C.A. *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, **327**, 343–348.
8. Suen,G., Teiling,C., Li,L., Holt,C., Abouheif,E., Bornberg-Bauer,E., Bouffard,P., Caldera,E.J., Cash,E., Cavanaugh,A. *et al.* (2011) The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.*, **7**, e1002007.
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. Smith,C.D., Zimin,A., Holt,C., Abouheif,E., Benton,R., Cash,E., Croset,V., Currie,C.R., Elhaik,E., Elsik,C.G. *et al.* (2011) Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5673–5678.
11. Smith,C.R., Smith,C.D., Robertson,H.M., Helmkampf,M., Zimin,A., Yandell,M., Holt,C., Hu,H., Abouheif,E., Benton,R. *et al.* (2011) Draft genome of the red harvester ant Pogonomyrmex barbatus. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5667–5672.
12. Nygaard,S., Zhang,G., Schiott,M., Li,C., Wurm,Y., Hu,H., Zhou,J., Ji,L., Qiu,F., Rasmussen,M. *et al.* (2011) The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. *Genome Res.*, **21**, 1339–1348.
13. Schrader,L., Kim,J.W., Ence,D., Zimin,A., Klein,A., Wyschetzki,K., Weichselgartner,T., Kemena,C., Stokl,J., Schultner,E. *et al.* (2014) Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat. Commun.*, **5**, 5495.
14. Oxley,P.R., Ji,L., Fetter-Pruneda,I., McKenzie,S.K., Li,C., Hu,H., Zhang,G. and Kronauer,D.J. (2014) The genome of the clonal raider ant Cerapachys biroi. *Curr. Biol.*, **24**, 451–458.
15. Wurm,Y., Wang,J., Riba-Grognuz,O., Corona,M., Nygaard,S., Hunt,B.G., Ingram,K.K., Falquet,L., Nipitwattanaphon,M., Gotzek,D. *et al.* (2011) The genome of the fire ant Solenopsis invicta. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5679–5684.
16. Sadd,B.M., Barribeau,S.M., Bloch,G., de Graaf,D.C., Dearden,P., Elsik,C.G., Gadau,J., Grimmelikhuijzen,C.J., Hasselmann,M., Lozier,J.D. *et al.* (2015) The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.*, **16**, 76.
17. Kapheim,K.M., Pan,H., Li,C., Salzberg,S.L., Puiu,D., Magoc,T., Robertson,H.M., Hudson,M.E., Venkat,A., Fischman,B.J. *et al.* (2015) Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. *Science*, **348**, 1139–1143.

18. Kocher,S.D., Li,C., Yang,W., Tan,H., Yi,S.V., Yang,X., Hoekstra,H.E., Zhang,G., Pierce,N.E. and Yu,D.W. (2013) The draft genome of a socially polymorphic halictid bee, Lasioglossum albipes. *Genome Biol.*, **14**, R142.
19. Elsik,C.G., Worley,K.C., Bennett,A.K., Beye,M., Camara,F., Childers,C.P., de Graaf,D.C., Debyser,G., Deng,J., Devreese,B. *et al.* (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics*, **15**, 86.
20. Simola,D.F., Wissler,L., Donahue,G., Waterhouse,R.M., Helmkampf,M., Roux,J., Nygaard,S., Glastad,K.M., Hagen,D.E., Viljakainen,L. *et al.* (2013) Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.*, **23**, 1235–1247.
21. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Harris,T.W., Arva,A., Stajich,J.E. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
22. Lee,E., Harris,N., Gibson,M., Chetty,R. and Lewis,S. (2009) Apollo: a community resource for genome annotation editing. *Bioinformatics*, **25**, 1836–1837.
23. Mungall,C.J., Emmert,D.B. and FlyBase Consortium. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
24. Lee,E., Helt,G.A., Reese,J.T., Munoz-Torres,M.C., Childers,C.P., Buels,R.M., Stein,L., Holmes,I.H., Elsik,C.G. and Lewis,S.E. (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14**, R93.
25. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
26. Smith,R.N., Aleksic,J., Butano,D., Carr,A., Contrino,S., Hu,F., Lyne,M., Lyne,R., Kalderimis,A., Rutherford,K. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, **28**, 3163–3165.
27. Lyne,R., Smith,R., Rutherford,K., Wakeling,M., Varley,A., Guillier,F., Janssens,H., Ji,W., McLaren,P., North,P. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, **8**, R129.
28. Motenko,H., Neuhauser,S.B., O'Keefe,M. and Richardson,J.E. (2015) MouseMine: a new data warehouse for MGI. *Mamm. Genome*, **26**, 325–330.
29. Harris,T.W., Baran,J., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., Done,J., Grove,C., Howe,K. *et al.* (2014) WormBase 2014: new views of curated biology. *Nucleic Acids Res.*, **42**, D789–D793.
30. Krishnakumar,V., Hanlon,M.R., Contrino,S., Ferlanti,E.S., Karamycheva,S., Kim,M., Rosen,B.D., Cheng,C.Y., Moreira,W., Mock,S.A. *et al.* (2015) Araport: the Arabidopsis information portal. *Nucleic Acids Res.*, **43**, D1003–D1009.
31. Krishnakumar,V., Kim,M., Rosen,B.D., Karamycheva,S., Bidwell,S.L., Tang,H. and Town,C.D. (2015) MTGD: The Medicago truncatula genome database. *Plant Cell Physiol.*, **56**, e1.
32. Balakrishnan,R., Park,J., Karra,K., Hitz,B.C., Binkley,G., Hong,E.L., Sullivan,J., Micklem,G. and Cherry,J.M. (2012) YeastMine–an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database*, bar062.
33. Contrino,S., Smith,R.N., Butano,D., Carr,A., Hu,F., Lyne,R., Rutherford,K., Kalderimis,A., Sullivan,J., Carbon,S. *et al.* (2012) modMine: flexible access to modENCODE data. *Nucleic Acids Res.*, **40**, D1082–D1088.
34. Sullivan,J., Karra,K., Moxon,S.A., Vallejos,A., Motenko,H., Wong,J.D., Aleksic,J., Balakrishnan,R., Binkley,G., Harris,T. *et al.* (2013) InterMOD: integrated data and tools for the unification of model organism research. *Sci. Rep.*, **3**, 1802.
35. dos Santos,G., Schroeder,A.J., Goodman,J.L., Strelets,V.B., Crosby,M.A., Thurmond,J., Emmert,D.B., Gelbart,W.M. and FlyBase,C. (2015) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.*, **43**, D690–D697.
36. Kriventseva,E.V., Tegenfeldt,F., Petty,T.J., Waterhouse,R.M., Simao,F.A., Pozdnyakov,I.A., Ioannidis,P. and Zdobnov,E.M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, **43**, D250–D256.
37. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
38. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
39. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
40. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
41. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
42. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database Collaboration. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
43. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
44. Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.*, **12**, 491.
45. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
46. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
47. Cox,M.P., Peterson,D.A. and Biggs,P.J. (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinform.*, **11**, 485.
48. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
49. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
50. Kalderimis,A., Stepan,R., Sullivan,J., Lyne,R., Lyne,M. and Micklem,G. (2014) BioJS DAGViewer: A reusable JavaScript component for displaying directed graphs. *F1000Research*, **3**, 51.
51. Wurm,Y., Uva,P., Ricci,F., Wang,J., Jemielity,S., Iseli,C., Falquet,L. and Keller,L. (2009) Fourmidable: a database for ant genomics. *BMC Genomics*, **10**, 5.
52. Poelchau,M., Childers,C., Moore,G., Tsavatapalli,V., Evans,J., Lee,C.Y., Lin,H., Lin,J.W. and Hackett,K. (2015) The i5k Workspace@NAL–enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.*, **43**, D714–D719.
53. i5K Consortium. (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.*, **104**, 595–600.
54. Brady,S.G., Fisher,B.L., Schultz,T.R. and Ward,P.S. (2014) The rise of army ants and their relatives: diversification of specialized predatory doryline ants. *BMC Evol. Biol.*, **14**, 93.
55. Ward,P.S., Brady,S.G., Fisher,B.L. and Schultz,T.R. (2015) The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). *Syst. Entomol.*, **40**, 61–81.