HymenopteraMine: An Integrated Annotation Data Warehouse Hymenoptera Genome Database

Aditi Tayal[1], Deepak Unni[1], Colin M. Diesh[1], Christine G. Elsik[1,2,3] and Darren E. Hagen[1]
Divisions of [1]Animal Sciences and [2]Plant Sciences and [3]MU Informatics Institute, University of Missouri, Columbia, MO 65211, USA

United States Department of Agriculture
National Institute of Food and Agriculture

Abstract

The Hymenoptera Genome Database (HGD http://HymenopteraGenome.org) is a genome informatics resource for hymenopteran insect species. HGD includes genomic data of three honey bee species (*Apis mellifera*, *A. florea, A. dorsata*), two bumble bee species (*Bombus terrestris* and *B. impatiens*), nine ant species (*Acromyrmex echinatior*, *Atta cephalotes*, *Camponotus floridanus*, *Cardicondyla obscurior*, *Harpegnathos saltator*, *Linepithema humile*, *Pogonomyrmex barbatus*, *Solenopsis invicta, Wasmannia auropunctata*), a parasitoid wasp (*Nasonia vitripennis*) and a sweat bee (*Lasioglossum albipes*). We have used InterMine to deploy a new data warehouse called HymenopteraMine to allow fast and flexible queries of annotation and ortholog data across species. HymenopteraMine integrates information from many data sources including RefSeq, UniProt, InterPro, OrthoDB, Pubmed, Gene Ontology. Users may perform a "Quick Search" or use the "Query Builder" for specialized searches. The "Genome Region Search" and "Overlapping Feature Search" allow users to download annotations with a specified genomic context. Links between genes and publications facilitate access to relevant scientific literature. Users may download query results in various formats, such as tab-delimited files, GFF, Fasta, BED, JSON and XML. In addition to HymenopteraMine, users may leverage the HGD genome browsers (GBrowse and JBrowse), BLAST searching and data download pages to access genome assemblies, computed and manually-annotated genes, protein homologs, cDNA sequences, non-coding RNA sequences, RNAseq- based expression data and genetic markers.

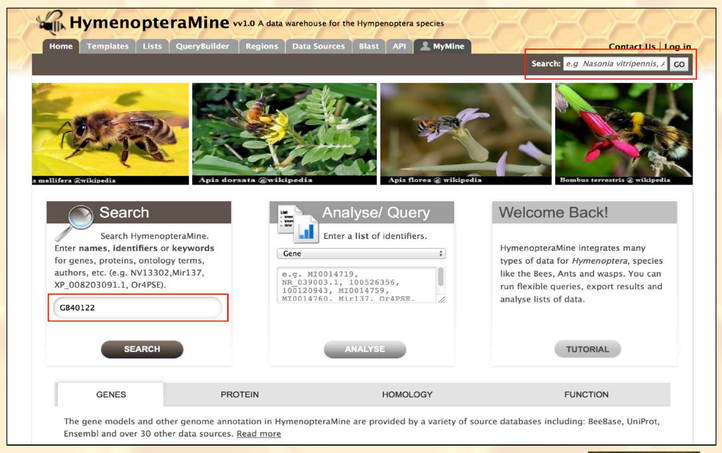


Figure 1: Web interface for HymenopterMine

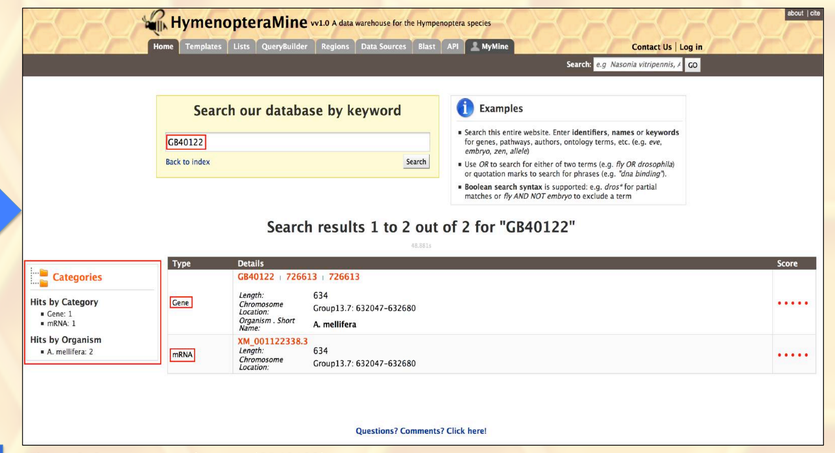


Figure 2: Search results for gene 'GB40122'.

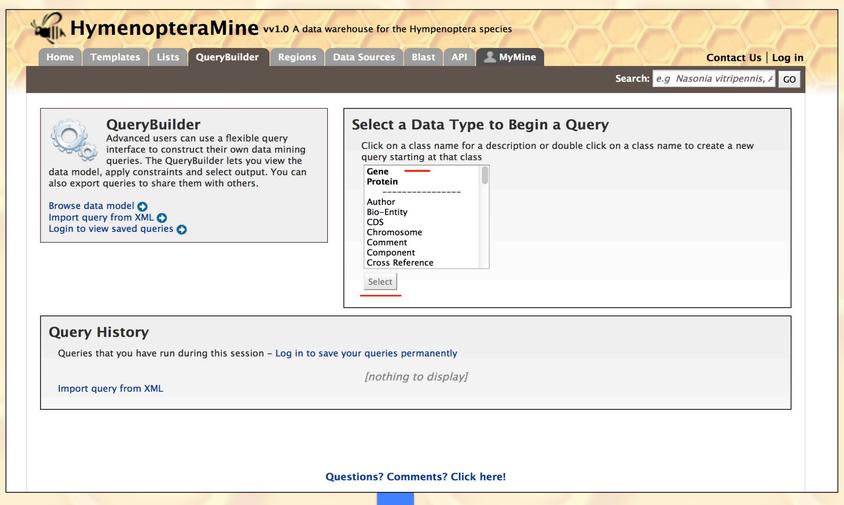


Figure 9: Query Builder tool with data types

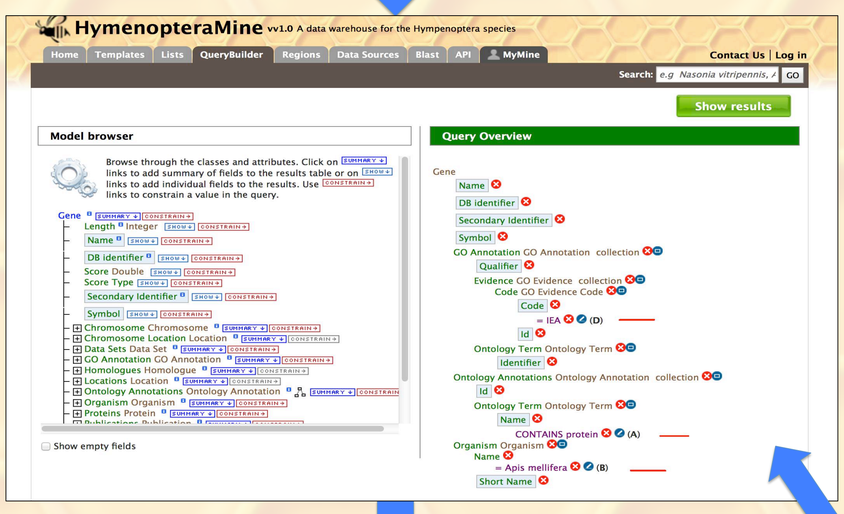


Figure 10: Query Builder tool querying the data model for information

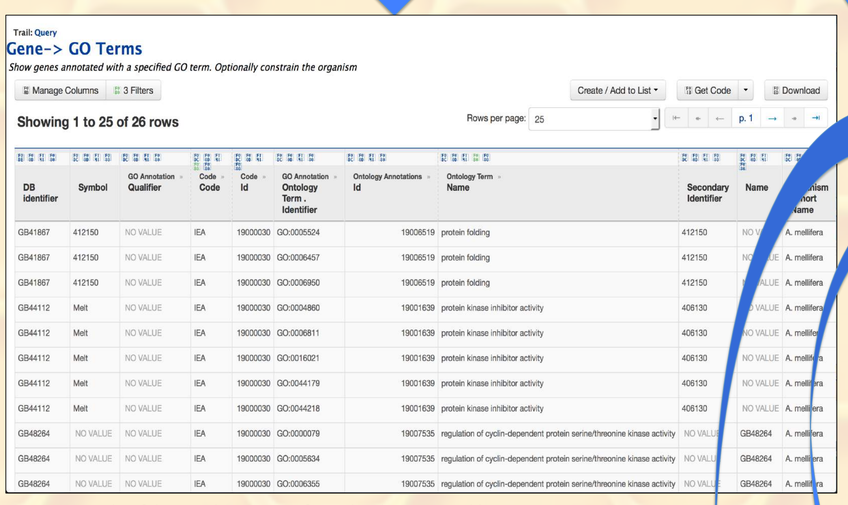


Figure 11: Results of the complex query performed in the Query Builder

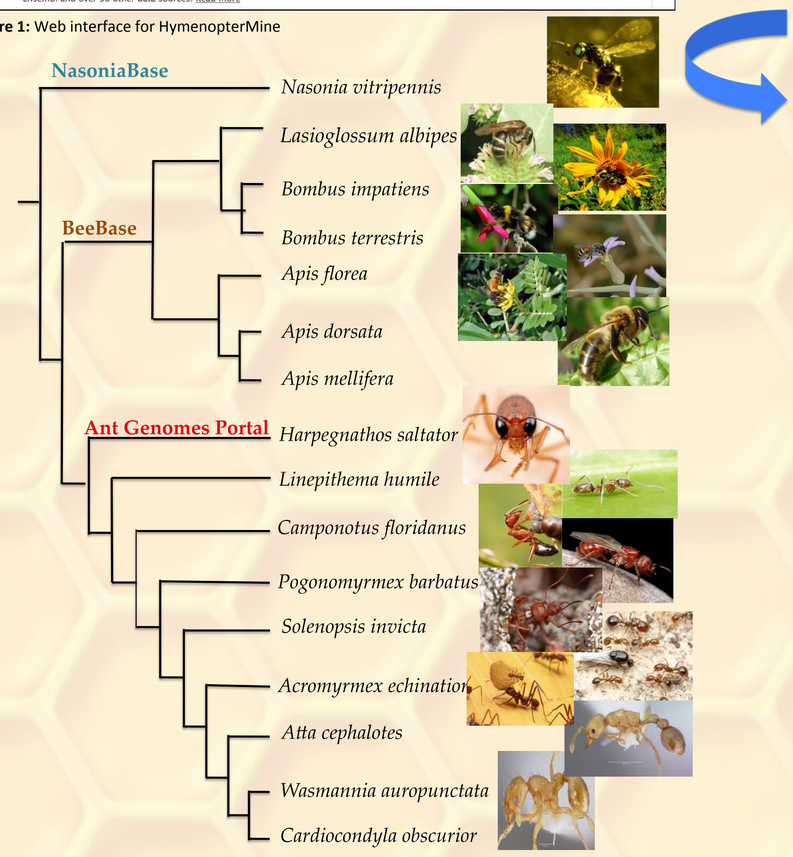


Figure 17: Various species in the HymenopteraMine database

Exciting highlights

- Various tools like Jbrowse and Gbrowse make data more interesting and visually appealing.
- Queries can also be exported and imported as XML and thus can be shared between users. Query results can be saved and exported from the InterMine web application in a range of common formats, for example as tab-delimited, comma-delimited, GFF3 or BED files and the users can customize the data fields to export.
- Homolog comparison across species like Bees, Wasps, ants and flies.
- Web interface and analysis widgets for visualizing chromosome distribution, Publication enrichment, Ortholog count, enrichment of Gene Ontology Terms for facilitating rapid exploration and discovery.
- There is a sophisticated QueryBuilder (Figure 9) for constructing advanced custom queries.
- Template queries can be created for commonly run searches
- Complex data integration from various sources like NCBI, Uniprot, OrthoDB, Pubmed and many more for a comprehensive source of data
- Developed set of APIs and web tools
- Blast searches sequence comparisons
- Fast, flexible querying with functionality for querying features which overlap specific genome ranges. Specific distances upstream and downstream of genes are represented to enable querying for genes that are near other features.

Figure 18: Exciting highlights of HymenopteraMine

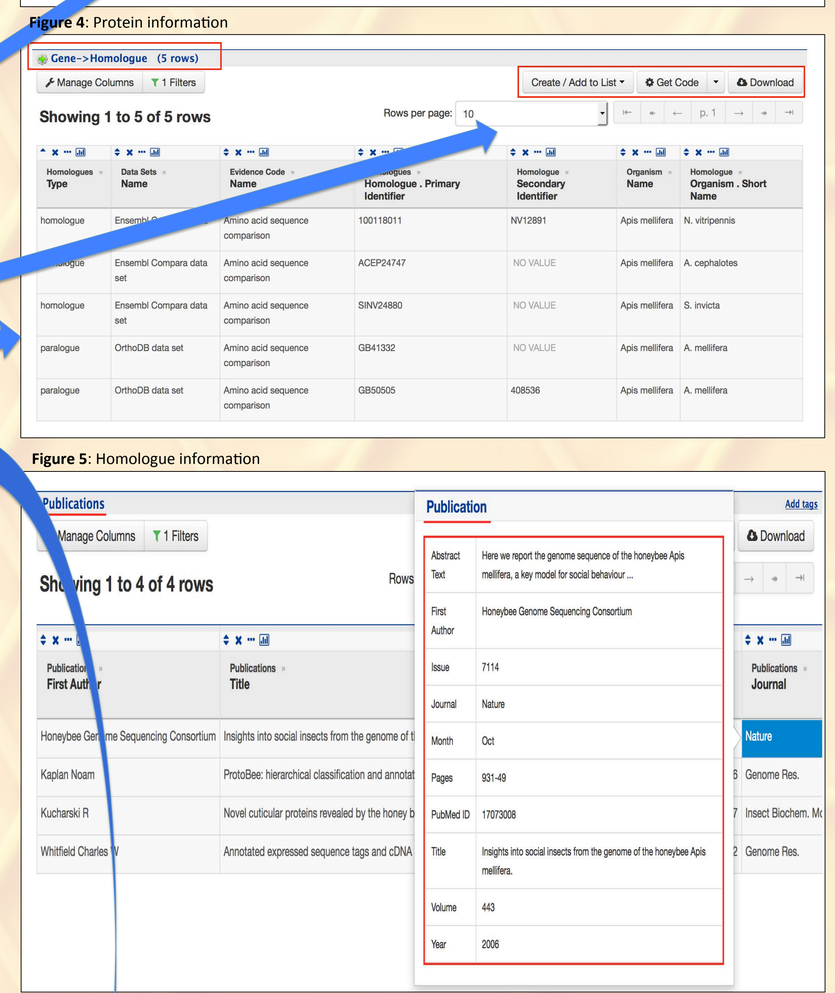
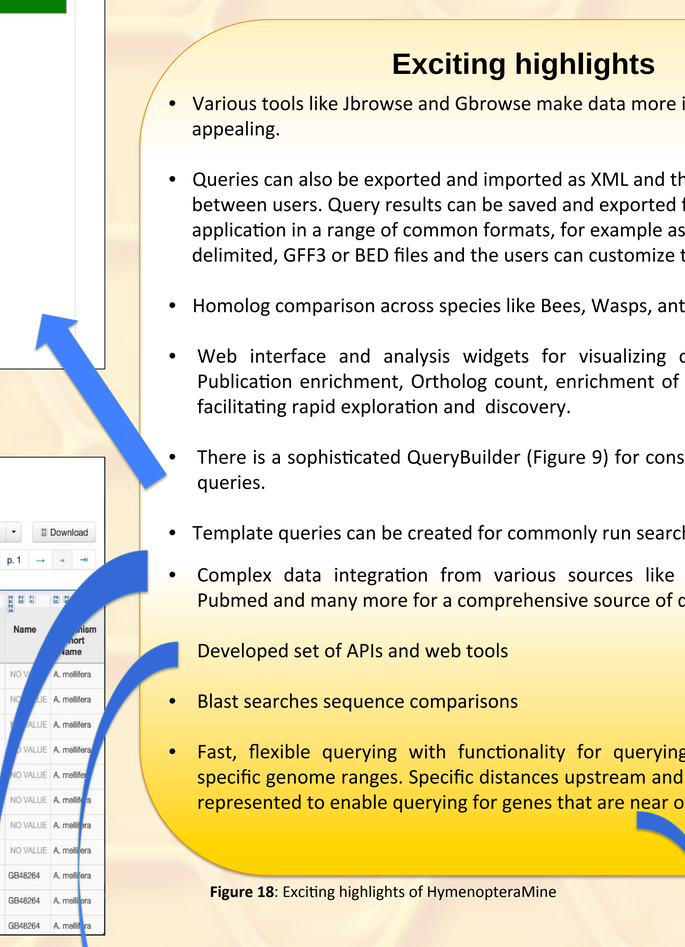


Figure 5: Homologue information

Data Warehouses

System	Disadvantages
BIOMART	-Queries limited to only two datasets at once Not possible to edit or create new filters
Open Genome Resource (OGeR)	-No advanced data mining tools beyond sequence alignment
PROFESS	-Query system is based on a non- customizable set of filters -Does not support similarity functions between standard BLAST searches
BioXRT	-No advanced data mining tools
Intermine	-No tools for classifying or clustering data

Table 1: A concise comparison of various data warehouse tools.
Reference: Triplet, T. and Butler, G. (2011). Systems biology warehousing: Challenges and strategies to- ward effective data integration, *DBKDA 2011, The Third International Conference on Advances in Databases, Knowledge, and Data Applications*, pp. 34–40



Figures 12-16 : Screenshot of various tools provided by intermine. 12) RegionSearch 13) API and web tools 14)Data sources 15) Pre-computed templates 16) lists and widgets

Figure 4: Protein information

Figure 6: Publication information

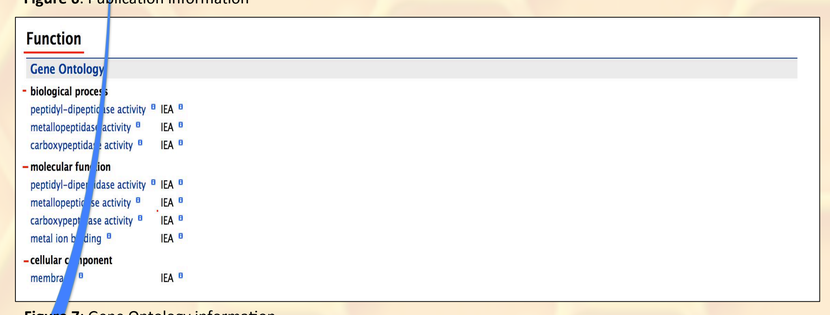


Figure 7: Gene Ontology information

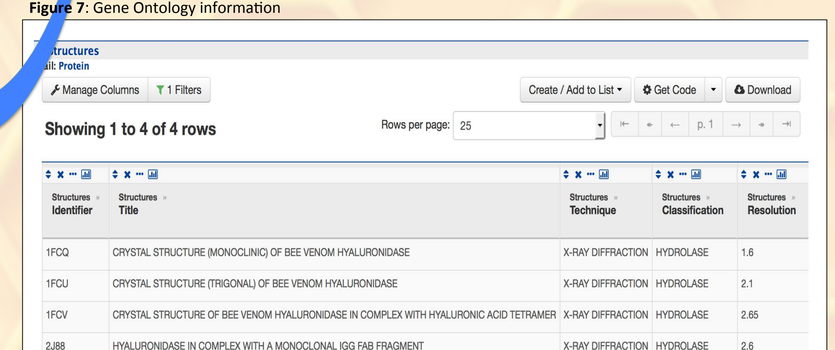


Figure 8: Protein structure information

Acknowledgements

HymenopteraMine is part of the Hymenoptera Genome Database project and is supported by Agriculture and Food Research Initiative Competitive grant no. 2010-65106-21301 from the USDA National Institute of Food Agriculture. We would also like to acknowledge Alex Wild Photography (http://www.alexanderwild.com) for pictures of Hymenoptera.