

Preprint

Elsik C.G., Tayal A., Unni D.R., Burns G.W., Hagen D.E. (2018) Hymenoptera Genome Database: Using HymenopteraMine to Enhance Genomic Studies of Hymenopteran Insects. In: Kollmar M. (eds) Eukaryotic Genomic Databases. Methods in Molecular Biology, vol 1757. Humana Press, New York, NY. doi: 10.1007/978-1-4939-7737-6_17 (https://link.springer.com/protocol/10.1007%2F978-1-4939-7737-6_17)

Hymenoptera Genome Database: Using HymenopteraMine to Enhance Genomic Studies of Hymenopteran Insects

Christine G. Elsik^{1,2,3*}, Aditi Tayal¹, Deepak R. Unni¹, Gregory W. Burns¹, Darren E. Hagen¹

Affiliation:

Division of Animal Sciences, University of Missouri, Columbia, MO USA
Division of Plant Sciences, University of Missouri, Columbia, MO, USA
MU Informatics Institute, University of Missouri, Columbia, MO, USA

Running Title: Hymenoptera Genome Database

Corresponding Author:

Christine G. Elsik
Division of Animal Sciences
University of Missouri
920 East Campus Drive
Columbia, Missouri 65211 USA
elsikc@missouri.edu

Abstract

The Hymenoptera Genome Database (HGD; <http://hymenopteragenome.org>) is a genome informatics resource for insects of the order Hymenoptera, which includes bees, ants and wasps. HGD provides genome browsers with manual annotation tools (JBrowse/Apollo), BLAST, bulk data download and a data mining warehouse (HymenopteraMine). This chapter will focus on the use of HymenopteraMine to create annotation data sets that can be exported for use in downstream analyses. HymenopteraMine leverages the InterMine platform to combine genome assemblies and official gene sets with data from OrthoDB, RefSeq, FlyBase, Gene Ontology, UniProt, InterPro, KEGG, Reactome, dbSNP, PubMed and BioGrid, as well as pre-computed gene expression information based on publicly available RNAseq. Built-in template queries provide starting points for data exploration, while the QueryBuilder tool supports construction of complex custom queries. The List Analysis and Genomic Regions search tools execute queries based on uploaded lists of identifiers and genome coordinates, respectively. HymenopteraMine facilitates cross-species data mining based on orthology and supports meta-analyses by tracking identifiers across gene sets and genome assemblies.

Key Words

Hymenoptera, *Apis mellifera*, genome, database, data mining, orthology, pathway, gene expression, single nucleotide polymorphism, InterMine

1. Introduction

The Hymenoptera Genome Database (HGD; <http://hymenopteragenome.org>) is an informatics resource for data associated with sequenced genomes of hymenopteran insects [1]. HGD currently includes genomes of eleven bee species, ten ant species, and the parasitoid jewel wasp (Table 1). Goals of HGD have been to 1) support species genome consortia with genome annotation tools, 2) provide access to data via genome browsers, BLAST and data download, 3) add value to the genome data by integrating it with external data sources in a data mining warehouse (HymenopteraMine) and 4) maintain the value of the genome consortia's published work by porting gene annotations to upgraded genome assemblies and providing identifier cross-references for updated gene sets.

Most hymenopteran insect genome sequencing projects have been carried out by small research consortia. HGD's initial contributions have focused on supporting hymenopteran insect genome consortia in the genome annotation and analysis process. More recently, with the availability of easy-to-deploy annotation pipelines, such as Maker2 [17] and the web-based Apollo annotation platform [18], the need for annotation support from HGD has decreased. Although we still provide Apollo annotation tools, we have shifted our efforts to support the use of the genomics data in downstream analyses. To make HGD more effective for post-genome-sequencing analyses, we have integrated the genome assemblies with other sources of biological data and developed data mining tools that support complex queries across species.

In addition to providing data mining and browsing tools, an important role of HGD is mapping identifiers across alternate datasets of the same species. Following publications of several genome projects, many of the genome assemblies and gene sets available at NCBI have been upgraded. To preserve the information from the original consortium publications and provide users with the most-up-to-date information, we provide resources for both the original consortium gene sets and the updated RefSeq gene sets and assemblies, along with references between gene identifiers across the gene sets. For entry into HGD, we require either a published consortium gene set or availability of a RefSeq gene set.

2. Methods

2.1 Website Navigation

HGD is divided into three major divisions (BeeBase, NasoniaBase, and the Ant Genomes Portal) to facilitate species-specific data download and genome browsing (Fig. 1).

However, data mining (HymenopteraMine) and sequence search (BLAST) tools are unified across HGD. The navigation bar of the HGD home page includes links to the individual divisions and to the pages that are common across divisions (Hymenoptera Home, HymenopteraMine, BLAST, Genome Consortium Publications, Data Usage Policy, How to Cite). Within each division, the navigation bars include the links common to all divisions, as well as links that expand to species-specific pages (JBrowse and Data Sets).

2.2 JBrowse and Apollo

Genome browsing is provided for each species using JBrowse [19] as implemented by Apollo [18]. From a user's perspective, the main differences between JBrowse and Apollo are the gene editing functions and the user annotation pane that are available only when logged into Apollo. The evidence tracks are identical across the browsers. All HGD users can access JBrowse, while only users registered for annotation can access Apollo. An Apollo registration link is provided in the navigation bar of a division only when there is an active annotation project for a species. Currently, active annotation is available for the three Apis species.

2.3 BLAST search

The HGD BLAST search interface leverages the SequenceServer platform [20], modified to make dataset selection easier. Datasets for all species are provided in a single interface. The name of each dataset indicates whether it is a consortium or NCBI dataset. You can select any combination of either protein or nucleotide datasets for a single search. When the search database is a genome assembly, BLAST hits are linked to JBrowse viewers based on matched coordinates. When the search database is coding sequence, transcript or peptide, BLAST hits are linked to a JBrowse location based on the hit identifier. SequenceServer also provides downloadable tab-delimited or BLAST XML reports and graphical overviews of the matches.

2.4 HymenopteraMine

HymenopteraMine is the data mining resource for HGD. It leverages the InterMine platform [21] to integrate biological data from a variety of sources (Table 2). Combining genomes for multiple hymenopteran species in a single data mining warehouse allows users to leverage cross-species information using orthologue relationships from OrthoDB [29]. The complexity of data in HymenopteraMine makes query construction challenging, so the search tools support a range of user skills, such as simple keyword search and predefined template queries for new users and the QueryBuilder for power users.

2.4.1 HymenopteraMine Navigation, Tutorials and MyMine

The HymenopteraMine home page is accessible from the navigation bars of HGD, BeeBase, NasoniaBase and the Ant Genomes Portal. HymenopteraMine has its own navigation bar with HymenopteraMine-specific links (Home, MyMine, Templates, Lists, QueryBuilder, Regions, Data Sources, Data Model, Help, API), as well as a link to HGD BLAST (Fig. 2). The HGD home page is accessible by clicking “Hymenoptera Genome Database” in the header. The Help Tab leads to a HymenopteraMine tutorial that includes a link to a YouTube channel with HymenopteraMine videos. The Data Model tab provides helpful information about the interconnection of data types in HymenopteraMine; it opens a new browser tab showing data network diagrams and links to tables indicating which types of identifiers are needed for specific data sets.

HymenopteraMine maintains user accounts allowing you to save your work after ending a session. Clicking “Log in” to the right of the navigation bar leads to an option to create an account. The MyMine tab in the navigation bar leads to a history of queries performed

during the current session. If you are logged in, lists created, queries performed and template queries are saved for future use.

2.4.2 Quick Search and Report Page

The HymenopteraMine home page provides basic search tools (Fig. 2). The Quick Search tool is used to perform a full text search of all datasets loaded in HymenopteraMine, and supports the use of wild cards. Data input types for Quick Search include gene identifiers, transcript identifiers, protein identifiers, gene symbols, gene names, functional annotation terms and species names. Quick Search is a good place to start to explore the data before performing more complex queries. For example, searching a species name will provide a list of all datasets for that species. A faceted search tool in the search result page allows users to filter the results by category before selecting an entity to access a report page (Fig. 3).

A report is provided for each entity in HymenopteraMine. Each Report is divided into sections appropriate for the data class. Users may find the most familiar reports to be those for genes, transcripts and proteins. These contain information similar to that found in gene pages of other model organism databases. However, most of the information in a HymenopteraMine report is provided in the form of tables that can be customized and downloaded in various formats, or saved as lists for further HymenopteraMine analyses. Sequences can be downloaded from reports in fasta format. The Function section of a Gene Report provides GO annotations and may also include pathways. The Transcripts section of a Gene Report gives a visual representation of the transcripts highlighting gene

structure with links to JBrowse. Transcript identifiers are linked to Transcript Reports. For *A. mellifera*, Transcript Reports include a Gene Expression section that provides various forms of expression values (raw read counts, normalized read counts, FPKM and RPKM) for RNAseq data with metadata from the Sequence Read Archive. The Protein section of the Gene Report lists protein identifiers that link to Protein Reports with more information including protein domains, UniProt keywords, and curated notes from UniProt.

2.4.3 List Analysis

The ability to upload and analyze lists of identifiers is one of the most important HymenopteraMine features because it allows you to gather a variety of functional annotation information associated with your own data. The Quick List tool provided on the home page is a slimmed-down version of the List Tool (Fig. 2). As opposed to Quick Search, which performs a full text search of keywords and supports wildcards, Quick List searches only gene and protein datasets based on gene identifiers (ids and symbols), transcript identifiers and protein identifiers.

The full List Tool is available by clicking the “Lists” tab in the navigation bar or the word “advanced” in the Quick List box. Clicking the “Lists” tab brings you to either an upload interface or a view of existing lists. You can move from one to the other by clicking “Upload” or “View” in the brown bar below the main navigation bar.

HymenopteraMine provides users with premade gene lists, for entire gene sets, that may be used as background populations in enrichment widgets (described in Subheading

2.4.7). The List upload menu of the List tool allows users to select from a pull-down menu of many data classes, to limit the search to a particular species, and to upload a list of identifiers.

Both Quick List and the full List tool perform a database lookup to validate the identifiers and may prompt the user to select from duplicates due to their presence in multiple datasets. For example, entering the gene symbol “Nmdar1” returns results for both *A. mellifera* and *D. melanogaster*. A green button saying “Save a list of 0 Genes” indicates that you must click the “Add” button to the right of a gene to save it to the list. The “Add” button does not appear if there are no duplicate identifiers. Before saving the final list, you may wish to enter a name for it. The list is saved by clicking the green “Save a List of X Genes” (where “X” is a number), and the result is a table with preset column output of associated information, such as gene identifier, gene secondary identifier, gene name, gene symbol, gene source (i.e. the gene set), gene status (i.e. the type of gene), chromosome and coordinate location, and organism. As with all table outputs in HymenopteraMine, the columns can be rearranged or deleted; column management tools (described in Subheading 2.4.8) can be used to add additional information; and the table can be exported. Lists of the additional data types, such as organisms or chromosomes and coordinates, can be saved using the “Save as List” pull-down menu above the table.

Saved lists can be retrieved by clicking “View” in the brown bar that is displayed under the navigation bar when the Lists tab is selected. Here you can perform set operations (union, intersection, subtraction and asymmetric difference) to create new lists. Once you

have saved a list, the predefined template queries and the QueryBuilder (described in Subheading 2.4.5) automatically provide the option to use it as long as the query is based on the appropriate data class. Lists are automatically deleted upon ending a session or accidental server disconnection unless you are logged in to MyMine, so working while logged in is recommended.

2.4.4 Template Queries

The complicated network of data in HymenopteraMine and the presence of alternative identifiers (Tables 3 and 4, and discussed in Subheading 2.4.10) can make it difficult to construct a query. HymenopteraMine provides predefined template queries to serve as starting points for data exploration. The complete list of templates is available via the Templates tab in the main navigation bar. Templates are also divided into categories in the home page template menu, which has tabs for GENES, GENE EXPRESSION, PROTEINS, HOMOLOGY, FUNCTION and VARIATION. In addition, the ALIAS AND DBXREF tab provides templates that convert identifiers between gene sets for an organism, and the ENTIRE GENE SET tab lists templates for queries that output all genes or proteins for an organism.

If you click a template name, you will access a query form that may already be pre-populated with example identifiers, and may include pull-down menus. Some templates include options for numerical operations. An example of a simple template is Gene ID → Alias ID, listed under the ALIAS AND DBXREF category. Clicking on the template name opens the template interface where you can enter a gene id. An example of a

complex template query is *A. mellifera* Transcript → Expression and MetaData, under the GENE EXPRESSION category, in which you enter a transcript id and you have options to constrain the output by expression levels and metadata values. If you have already saved a list with the appropriate data class, an option is provided to constrain the search to the list of identifiers rather than a single identifier. You obtain results by clicking “Show Results”. Alternatively, you can click “Edit Query” to access the QueryBuilder (described in Subheading 2.4.5) if you wish to modify the query.

2.4.5 QueryBuilder

The QueryBuilder is the most flexible and sophisticated search tool of HymenopteraMine. However, use of the QueryBuilder is not intuitive; becoming a power user requires practice. On the other hand, for users without scripting skills, mastering the use of QueryBuilder to create a large complex data set would likely be more efficient than learning a scripting language to compile the data from the original sources. Before trying to build a query with QueryBuilder, it may be helpful to investigate the structure of some of the template queries using the “Edit Query” button available in the template query menus. After clicking “Edit Query” you will see the constructed template query in the same interface that is used to build the query from scratch.

A detailed example using the QueryBuilder is provided below. However, first we will provide an overview of the features. Clicking QueryBuilder in the navigation bar leads to the entry page for construction. The “QueryBuilder” box on the left includes options to browse the HymenopteraMine data model, import a query from XML, and login to view

saved queries. Selecting “Browse the Data Model” leads to a tree-like model of the data classes (Fig. 4). Clicking a plus sign in the tree reveals subclasses, while clicking the name of a data class opens the Model Builder (described below) at that class. The tree is useful for seeing the numbers of data objects within different data classes. Mousing over the “i” symbol next to a class provides a description. The classes that will be of most interest to HymenopteraMine users are listed under the “Sequence Feature” class, which is directly under “BioEntity”. Although exploring the data with this tree is useful for developers, it is not intuitive for most users, and the structure is subject to change.

Instead of selecting “Browse the Data Model”, selecting a data class from the pull-down menu under “Select a Data Type to Begin a Query” on the right side of the main QueryBuilder page leads to the same Model Builder mentioned above; the advantage of starting with the pull-down menu is you do not need to find the correct class in the data model tree. The Model Browser shows the hierarchical data structure, similar to the data tree described above, but is zoomed into the selected data class. The Model Browser also shows relationships between data classes. To the right of each data class is the word **CONSTRAIN**, and either the word **SUMMARY** or **SHOW**. Clicking any of these words initiates query construction. You add constraints to the query using “CONSTRAIN”, and you select output data classes with **SUMMARY** or **SHOW**. **SUMMARY** is provided for any class with a “+” sign next to it, indicating that it can be divided into more than one subclass or attribute; **SUMMARY** is used to select a collection of the attributes as output. **SHOW** is provided for individual attributes that cannot be further subdivided (e.g. an identifier).

As you construct a query using the Model Browser, query building blocks are shown in the Query Overview Panel on the right side of the page. If you initiate query construction by clicking the word “CONSTRAIN” in the Model Browser next to the data class on which the search will be performed, a box appears allowing you to enter a constraint identifier. Once you have made selections in the constraint box, the constraint appears in the Query Overview. You add query output by clicking “SHOW” next to an attribute or “SUMMARY” next to a class. When the word “collection” appears next to a data class name in the Query Overview, it indicates that there is a collection of attributes related to this class. Clicking on the data class name next to “collection” in the Query Overview allows you to easily navigate to that data class in the Model Browser tree for further modifications. In an iterative fashion, you can add additional constraints and outputs to build a complex query. You can remove constraints and outputs using the red “X”, and you can edit constraints by clicking the pencil symbol. Once you have completed query construction, you can view and rearrange the output column order in the “Fields selected for output” section below the Model Browser. You can export the query using “Export XML” at the bottom of the page to save the query locally. If you are logged into MyMine, the “Start building a template query” button at the bottom of the page allows you to make the query into a template that will be saved in your MyMine account. Finally, to run the query, you would click the green “Show results” button. After running the query, it is automatically saved in your MyMine query history. From there you can rerun it or edit it. Clicking edit returns you to the QueryBuilder interface, which provides another opportunity to export or create a template query if you did not already do so.

QueryBuilder Example: Say you would like more information about a gene, “BIMP17180” from the *B. impatiens* official gene set (bimp_OGSv1.0). In this example, you will build a query to retrieve the *A. mellifera* homologue(s) of the *B. impatiens* gene, and pathways of those homologues. You will also retrieve the gene symbols for both the *B. impatiens* and *A. mellifera* genes. As you follow the instructions to build the query, note how the Query Viewer displays each step of query construction. The last few steps in this example provide instructions for turning your constructed query into a template query if you are logged into MyMine. An advantage of creating a template query is it makes it convenient to rerun the query with different constraint values or with a premade list of identifiers.

1. It is advisable to log into MyMine before you start building a query so that you will be able to save your work.
2. Click the QueryBuilder tab in the HymenopteraMine navigation bar.
3. Select “Gene” in the “Select a Data Type to Begin a Query” box. This brings you to the Model Browser, starting with the “Gene” data class.
4. The first query building block will be to constrain the gene id to “BIMP17180”. Click CONSTRAIN next to “DB identifier” under “Gene” in the model browser, enter the gene id “BIMP17180” in the box, and click “Add to Query” (Fig. 5A). Notice that the constraint has been added to the Query Viewer on the right. The pencil symbol allows you to edit the constraint in case you decide to use a

- different id. Since this is a unique identifier that is only found in the bimp_OGSv1.0 gene set, there is no need to constrain the species or dataset.
5. We would like the output to include our entered gene id, so click “Show” next to “DB identifier” under “Gene” in the Model Browser. Notice that in the Query Overview, “DB identifier” is now shown in a light blue box to indicate it will be included in the output.
 6. The next step is determine what kind of identifier is needed in order to output a gene symbol and homologues using Table 4 in this chapter, or the “Identifier Relationship Table” available by clicking the HymenopteraMine “Data Model” tab. Notice that for *B. impatiens*, Table 4 indicates “R” (RefSeq) for gene symbol and “O” (OGS) for OrthoDB.
 7. We will first take steps to retrieve the gene symbol. We will need to use a database cross reference relationship to retrieve the RefSeq id for the gene, so the next query building block to add is the database cross reference id. Use the scroll bar to the right of the Model Browser to scroll down until you see “Db Cross References x Ref”. Click the “+” sign next to “Db Cross References” to show subclasses. Look down a couple lines for “Cross Reference Gene” and click the “+” sign to see its attributes. Click SHOW next to “DB identifier” to output the database cross reference identifier. Notice that several lines have been added to the Query Viewer, and these are indented to indicate that this information is a subclass of “Gene”. The line with “Db Cross References x Ref collection” was added because you selected an attribute that is part of the collection of attributes for the relationship between the Db Cross Reference dataset and the Gene dataset.

- More specifically, you selected the attribute “DB identifier” of the “Cross Reference Gene”.
8. The next query building block is to add the gene symbol for the database cross reference as output. You will notice that the Model Browser view has automatically been reset to the top of the tree. In order to output the gene symbol for the database cross reference gene, rather than the original gene, you must be sure to navigate back to the correct part of the tree. To easily do so, click the word “Gene” within “Cross Reference Gene” in the Query Overview. This causes the Model browser to adjust so that “Cross Reference Gene” and its attributes show in the central area of the view (Fig. 5B). Click SHOW next to “Symbol” to add it as output to the query.
 9. The objective of the next few steps is to add the homologous *A. mellifera* gene (or genes) to the output. Remember that Table 4 indicated that the *B. impatiens* OGS id is required to retrieve relationships to OrthoDB data. Therefore you will go back to the original “Gene” as a starting point when looking at the Model Browser, because you have already constrained the gene based on an OGS id.
 10. Use the Model Browser scroll bar to scroll down. You should click the “-” sign next to “Db Cross References x Ref” to close that part of the tree to avoid confusion. You will see “Homologues Homologue” a few lines down. Click the plus sign to see the subclasses. Many data classes include a subclass that refers back to the parent data class (in this case the initial “Gene”). These are recognized with a red arrow pointing up. Under Homologues, the first subclass listed is “Gene” with the red up arrow (Fig. 5C). This is not the homologous gene, but is a

- reference to the parent “Gene” class. Do not select this. Instead, look further down the list of subclasses until you see “Homologue Gene”. This is where you find attributes for the homologous genes. Click the “+” sign next to “Homologue Gene”.
11. Under “Homologue Gene” select SHOW next to “DB identifier” so the gene id of the homologue will be included in the output (Fig. 5C). Notice that several lines have been added to the Query Viewer. The line with “Homologues Homologue Collection” was added because you selected an attribute that is part of the collection of attributes for homologue relationships to genes. More specifically, you selected the attribute “DB identifier” that is part of the “Homologue Gene”, i.e. the homologous gene itself.
 12. The next building block to add is a constraint specifying that you want only *A. mellifera* homologues. Click the word “Gene” within “Homologue Gene” in the Query Viewer to jump back to the appropriate part of the Model Browser tree.
 13. Scroll down using the Model Browser scroll bar to find the word “Organism”, keeping an eye on the inner line on the left side to ensure you stay next to the line that descends from “Homologue Gene” (i.e. make sure you stay in the “Homologue Gene” subtree).
 14. Click the “+” next to “Organism” to show its attributes.
 15. Click CONSTRAIN next to “Short Name”, and then use the pull-down menu to select “A. mellifera” and click “Add to Query”. The homologue organism constraint will show up in the Query Viewer.

16. The next step is to determine which kind of identifier is used for *A. mellifera* homologues, and also which kind of identifier is needed for *A. mellifera* gene symbols and pathways. Notice in Table 4 that for *A. mellifera* there is an “O” (OGS) in the OrthoDB cell, and “R” (RefSeq) for both Gene Symbol and KEGG. This means that the *A. mellifera* homologues that will be retrieved will have OGS ids, so a database cross reference is needed to connect the OGS ids to the RefSeq ids before KEGG pathways can be retrieved.
17. To add the building block for the database cross reference id of the *A. mellifera* homologue, click “Gene” in “Homologue Gene” to return to the correct subtree in the Model Browser.
18. Follow the line that descends from “Homologue Gene” until you find “Db Cross References x Ref” (i.e. make sure you select the “Db Cross References x Ref” that is a subclass of “Homologue Gene” rather than the one that is a subclass of the root “Gene”) (Fig. 5D).
19. Similar to what you did in step 7, click the “+” sign next to “Db Cross References” to show subclasses under “Db Cross References”. Look down a couple lines for “Cross Reference Gene” and click the “+” sign to see its attributes. Click SHOW next to “DB identifier” to output the database cross reference identifier.
20. The next building block to add is to output the gene symbol for the *A. mellifera* database cross reference (RefSeq) gene. Click the word “Gene” within “Cross Reference Gene” in the Query Viewer, making sure you are looking at the “Cross

- Reference Gene” under “Homologue Gene” to jump to the correct subtree of the Model Browser.
21. Under “Cross Reference Gene” in the Model Browser, click SHOW next to “Symbol” to add the gene symbol for the *A. mellifera* RefSeq gene to the output. A line with the word “Symbol” will appear under “DB identifier” under “Cross Reference Gene”.
 22. The next step is to output pathway information for the database cross reference of the *A. mellifera* homologous gene. To make sure you select the correct pathway data class, again click the word “Gene” within “Cross Reference Gene” in the Query Viewer, making sure you are looking at the “Cross Reference Gene” under “Homologue Gene” to jump to the correct subtree of the Model Browser.
 23. Scroll down using the Model Browser scroll bar to find the word “Pathways”, again keeping an eye on the inner line on the left side to ensure you stay in the correct subtree.
 24. Click the “+” next to “Pathways” to open the subtree.
 25. Click SHOW next to “Identifier” to include the pathway identifier in the output. Notice that several more lines have been added to the Query Viewer, because you have added information from another data collection.
 26. The last building block to add to the query is to output the name of the pathway. To navigate back to the correct Pathway subtree in the Model Browser, click the word “Pathway” in the line “Pathways Pathway Collection” in the query viewer.
 27. Under “Pathways”, click SHOW next to “Name” to output the name of the homologue pathway.

28. At this point, query construction is complete (Fig. 5E). Scroll down to the “Fields Selected for Output” area of the page. You will see blocks representing the output columns (Fig. 5F). The blocks appear in the order you added them to the query, not necessarily the order they are shown in the Query Overview. You may rearrange the columns by dragging the blocks.
29. If you are not logged in to MyMine, and you wish to save this query, you can click “Export XML” to save it locally. You will be able to import it in a future HymenopteraMine session.
30. You can either click “Show Results” at this point to run the query, or if you are logged into MyMine, you can click “Start building a template query” to create a template query saved in your MyMine Template collection. The only additional steps required to create the template are to name the template and click “Save Template”. Before saving, you also have the options of providing a title, description and comments. You will have the opportunity to run the query after you complete the template.
31. Whether or not you have created a template, if you are logged into MyMine the query will automatically be saved in your query history after you run it. From your query history list in MyMine, you can rerun the query or edit it. However, the query history is temporary, and its preservation across sessions is not guaranteed. Therefore, it is advisable to name your query and click “Save query” at the bottom of the QueryBuilder page before you end the session to save the query in your MyMine Queries list. If you have already run the query before saving, simply find it in your query history to save it.

2.4.6 Regions Search

The Genomic Region search tool allows you to perform a coordinate based search for genomic features. The list of available features depends on the species selected. Gene, mRNA, coding regions (CDS), exons, and polypeptides are available for all species. Some species include additional features, such as miRNA, tRNA, indels and SNPs. You perform a search by selecting desired output features, providing a list of regions, and optionally entering a distance in bases to extend the regions. You can either paste lists of locations that include the scaffold or chromosome identifier and the start and end coordinates, or upload the locations as a text file. The search result page provides options to download data for individual regions or all regions at once in tab, csv, gff3, fasta or bed formats. A pull-down menu allows selection of a feature type for creating a list that you can further augment using template queries or “Manage Columns” (described in Subheading 2.4.8). The example in Subheading 2.4.11 includes a Regions search.

2.4.7 Enrichment Widgets

After saving or viewing a list of genes you will automatically be presented with widgets showing results of enrichment analyses for gene ontology terms, pathways and publications. Each widget provides options for test correction (Holm-Bonferroni, Benjamini Hochberg, and Bonferroni) and p-value cutoff. The default background population is all genes in the organism that have annotations of the type being calculated. Therefore, since HymenopteraMine contains multiple gene sets with annotations per organism, it is recommended that you do not use the default background population.

Rather, you can click “Change” to select a background population from among your saved lists. Premade lists of gene ids for each organism’s gene sets are provided. The List Tool makes it easy to create more refined background populations for specific questions. For example, you can create lists of all expressed genes to use as the background to test for enrichment in differentially expressed genes. The example in Subheading 2.4.11 includes the Gene Ontology Enrichment Widget.

2.4.8 Manipulating Outputs and Augmenting Queries through Column

Management

HymenopteraMine table outputs can be manipulated in many ways, from minor changes like sorting rows or deleting columns, to extensive changes, like adding new columns and filters. The example in 2.4.11 includes column management.

Rearranging Columns: You can rearrange column order in two ways. First, you can reorder the columns when building a query with QueryBuilder, as described above. Second, you can use the “Manage Columns” button appearing above the output table. With “Manage Columns”, you can use up and down arrows next to the column list to rearrange the order. The red circle next to each column name lets you delete the column.

Using Column headers: Each column header in a table contains symbols for table management. Arrows are for sorting in descending or ascending order; the “X” is for column deletion, the “...” symbol is to hide a column in order to make other columns more visible; the funnel-shaped symbol is to filter the rows based on values; the

histogram symbol is to show counts of individual values, and to provide an alternate filter interface. An example use of the filter in the histogram interface is to determine which gene source has the largest number of output genes, and apply a filter to select only that gene source. This assumes that gene source was included as output in the original query. If it was not include in the original query, it can be added using Manage Columns as described below.

Manage Columns to Alter Query Output: The Manage Columns interface provides yet another mechanism to build a query. It does not allow the incorporation of new query constraints, but allows the addition of new output columns. The starting point for building a query with Manage Columns is any table; the table may have originated from the List Tool, a Regions Search, a table within a report page, or a query output. You may find that using a simple template query followed by Manage Columns is easier than using the QueryBuilder to construct a complex query. Although you cannot add new query constraints with Manage Columns, you can use the column filters described above or “Manage Filters” described below to further constrain the final output. Upon clicking “Manage Columns”, the “Selected Columns” interface automatically appears, listing the columns; the “+ Add a Column” button opens an interactive data model tree organized hierarchically, similar to the tree in the Model Browser. The tree is automatically rooted with the appropriate data class. Tag symbols within the tree delineate columns that you can add to the table. Plus symbols indicate data subclasses that can be opened to reveal additional tags or subclasses. You select a column to add as output by clicking to highlight in blue any attribute with a tag symbol. Once you have selected the desired

columns, clicking “Apply Changes” brings you back to the column list, where you can use the arrows to rearrange the column order. Clicking “Apply Changes” in this window provides the modified table.

Manage Filters to Alter Query Constraints: Manage Filters provides an interface complementary to the Manage Columns interface. While Manage Columns can be used to alter query outputs, Manage Filters can be used to alter query constraints. Although the word “filter” is used in this interface, you can relate each filter to a constraint from the original query. After clicking Manage Filters, an interface is provided for editing current filters or adding new filters. For example, if the original query was constrained using a particular gene identifier, clicking on the corresponding filter in Manage Filters allows you to change that identifier. To create a new filter, click “Define a New Filter”. An interactive data model tree, similar to the one described under “Manage Columns” will appear. From the tree you can select attributes to use as filters. An important difference between adding a filter with the Manage Filters interface versus incorporating the constraint in the original query is that multiple values for the selected attribute must exist in the current table in order for that attribute to be selected as a filter. If the current table has only one value for a particular attribute, you will receive a message “There is only one possible value... You might want to remove this constraint.”

Sorting with Manage Columns: In addition to the simple ascending or descending sort you can perform on a single column using the sort symbol in the header, you can create a prioritized list of sort attributes using the Sort Order tab in Manage Columns. In the Sort

Order interface, the primary output column on which you can sort is listed on the left side. The panel on the right lists all data classes that are available as sort criteria, and can be selected by clicking the plus sign. In addition to the existing columns, all attributes directly connected to each data class can be used as sort criterion without requiring the attribute itself to be present in the table. For example, gene ids can be sorted based on gene length without adding a gene length column to the output.

Manage Relationships: The default result for a complex query is to show only rows for which values exist in all related data classes. For example, if you wish to generate a table containing *A. mellifera* gene ids, *D. melanogaster* homologues to the *A. mellifera* genes and pathways for the *D. melanogaster* homologues, the default output will not include any *A. mellifera* gene for which there is a *D. melanogaster* homologue but no pathway. You can use “Manage Relationships” to change the behavior such that all *A. mellifera* genes that have *D. melanogaster* homologues are shown regardless of whether pathway information exists. The Manage Relationships interface lists all data class relationships present in your original query, with buttons to switch any of them from “Required” to “Optional”. Selecting “Optional” not only causes rows lacking results to be included in the output, but also modifies the format for the affected column, such that the output is shown as nested sub-tables within the main results set. However, when the table is exported as tab or comma separated values, the format reverts to the inline layout.

2.4.9 Exporting Query Outputs and Queries

The “Export” button at the top of each table provides many download options. Tables can be downloaded as tab separated values (TSV), comma separated values (CSV), JSON or XML; sequences associated with identifiers in the tables can be downloaded in fasta format; coordinate information associated with the identifiers can be downloaded in GFF3 or BED format.

The Export form includes a text box to enter a file name. To the right of the text box the default .tsv file extension opens a pull-down menu allowing you to modify the file type. The tabs listed on the left side of the Export form depend on the file type selected, and allow you to modify default settings. For TSV, CSV, JSON and XML, the “All Columns” tab allows you to select which columns to download, the “All Rows” tab allows you to decrease the number of rows for export. For TSV and CSV, the “No Column Headers” tab allows you to add column headers. By default, none of the file formats are compressed, but the “No Compression” tab allows you to set the option to gzip or zip compression. A preview of the first three rows is provided for TSV, CSV, JSON and XML formats.

The queries themselves can be downloaded as XML, allowing them to be shared with other users or imported back into your private MyMine account. An advantage of saving your query after you are satisfied with the table content and arrangement is that manipulations on the table after running a query, such as sorting or filtering, will be maintained in the XML code.

2.4.10 Gene Aliases and Database Cross References in HymenopteraMine

A HymenopteraMine feature that is particularly important for the species in HGD is the maintenance of gene and transcript alias identifiers and database cross reference identifiers. Table 3 lists the data sources of alias identifiers and database cross references, and Table 4 indicates which type of identifier is used for each data set.

We define a “Database cross reference” as an identifier for an equivalent gene locus in an alternative primary gene set. For example, for most HymenopteraMine organisms, the consortium OGS (with HGD identifiers) contains gene identifiers that cross reference gene identifiers in the RefSeq gene set and vice versa. Any database cross reference id is also the primary database id of another primary gene set, and therefore may be connected to other data classes, such as homologues and pathways (Table 4). We use database cross references for gene sets that have been generated using different methods, and sometimes using different assembly releases. As a result, the cross-referencing gene sets are not equivalent to each other, so any gene in one gene set may not have a database cross reference in another gene set, and some genes may have multiple cross references due to disagreement in gene boundaries. Rather than eliminating database cross references for genes with ambiguous relationships, we keep the full list of cross references for all genes so you will be aware when there is a disagreement between gene sets (e.g. you will see a gene in one gene set that maps to two different genes in the other). Several species have database cross references between the OGS and RefSeq gene sets. For *Nasonia vitripennis*, there are cross references between the original published consortium OGS (nvit_OGSv1.2) [7], the Evidential Gene Set [8], and RefSeq.

When working with a primary gene set identifier, you may need to use a database cross reference in order to retrieve a specific data set. For example, to retrieve KEGG pathways for *A. mellifera* genes, you would need RefSeq gene ids rather than OGS gene ids. Table 4, also available via the HymenopteraMine Data Model tab, indicates the type of identifier required to retrieve information from each data set for each species. If necessary, you can convert identifiers using the Gene ID → Database Cross Reference ID template query.

An alias id is an alternative identifier for a gene or transcript, and is not connected directly to any data class other than the primary gene. Alias identifiers occur due to 1) HGD identifiers being assigned to an OGS, 2) the existence of an upgraded OGS (e.g. *A. mellifera* OGSv1 and OGSv3.2), or 3) an alternative identifier used by OrthoDB. The assignment of HGD identifiers is the major source of aliases. We assign gene and transcript identifiers to official gene sets provided by consortia because many of the original gene sets provide only transcript identifiers that do not indicate which transcripts belong to the same gene locus. The OGS are loaded into HymenopteraMine using the HGD id as the primary id and the consortium id as the alias. We also assign aliases between old and new consortium gene sets in a similar way that we compute database cross references. Finally, we provide aliases when the OrthoDB website displays an alternate identifier for a species, so that you can use the OrthoDB website in downstream analyses. Table 3 indicates the names of alias identifier sources for each species, and also indicates whether the alias identifiers are assigned to transcripts or genes. Whether an alias is associated with a transcript or gene determines which type of template query to

use for conversion to primary ids. Each template query that involves an alias includes a description that indicates which species it is suited for.

If you are unsure whether an identifier is an alias or an id from a primary gene set, you can quickly check the identifier using the List Tool. Use the pull-down menu to select “Gene” or “Alias Name”. If the identifier is verified in the database when “Gene” is selected, it is an identifier from a primary gene set. If it is verified when “Alias Name” is selected, it is an alias. If the identifier is not found with either choice, it is not a gene or transcript identifier in HymenopteraMine. Once you have determined that your identifier is an alias, you can use a template to convert the alias to a primary identifier.

2.4.11 Step-By-Step Meta-Analysis Example including Regions Search, List Tool, Enrichment Widget, Template Query and Column Management

This example will combine two studies on honey bee scouting behavior. The first study used microarray analysis to identify genes that were differentially expressed between foragers that scout for food sources and foragers called recruits, which do not scout for food sources [35]. The second study used genome-wide association analysis (GWAS) to identify genome variants associated with behavioral differences between scouts and recruits [36]. We would like to identify differentially expressed genes from the first study within 50kb of single nucleotide polymorphisms (SNPs) identified in the second study.

1. The first step is to create a list of SNP coordinates to use in a regions search. You have the choice of following instructions for downloading the data from the

original journal source and using Excel or OpenOffice-Calc to format them or using preformatted SNP coordinates provided in Table 5 of this chapter.

Method using original journal source plus Excel: For this example we will use the supplemental data file from the GWAS study [36]. Go to the publicly available journal article webpage

(<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146430>). Scroll

down to the supporting information section. Click “S2 Table”. After downloading the file, change the file extension to “.txt”. Open the file using Excel/OOCalc.

Although the file was originally labeled “CSV”, check both tab and space as column delimiter. Like many supplemental datasets, this file is not optimally formatted for use in meta-analyses. The first step is to delete all unnecessary columns. We need only the chromosome identifier and SNP coordinate, so delete all columns except A and C. Now delete any extra rows, including the first two rows and the rows at the end with “Unplaced” scaffolds. We will ignore any scaffold that was not assigned to a chromosome. We need to create a list of chromosome locations using the original *A. mellifera* genome assembly “Group” identifiers, rather than identifiers like “LG1”. Use the Find and Replace Function to replace “LG” with “Group.” The next step is to format the chromosome location for pasting into the HymenopteraMine Regions text input box, which requires both a start and an end coordinate. For SNPs, you use the same coordinate for each. So in your Excel/OOCalc spreadsheet, copy the column of

coordinates to the next column. Now you should have three columns:
chromosome identifier, start and end.

Alternative method: If you do not have access to the publication or Excel/OOCalc, you should follow the alternative instructions in step 2d below to use identifiers provided in Table 5 of this chapter.

2. Once you have properly formatted SNP locations, the next step is to perform a Regions Search to obtain genes within 50kb of the SNP coordinates. Go to HymenopteraMine. It is recommended that you login to MyMine so you can save your work.
 - a. Click the Regions Tab in the HymenopteraMine navigation bar.
 - b. Select “A. mellifera” from the “1. Select Organism” pull-down menu.
 - c. Click the square next to “2. Select Feature Types” to uncheck all options and then click the box next to “gene” as the chosen feature option.
 - d. If you created a spreadsheet with 3 columns, highlight all columns simultaneously and copy them into the text box. The columns entered will automatically be tab-delimited. If you choose to use the locations provided in Table 5 of this chapter, be aware that they are formatted differently than what was described for the Excel/OOCalc spreadsheet. Each cell in Table 5 contains an entire SNP location, with chromosome id, start and end coordinate. Table 5 has 3 columns to save space within the chapter, but the columns must not be selected simultaneously. Highlight each column and

paste it, one-at-a time, into the Regions search text box, so that you are always extending the list when you add the next column.

- e. Type “50kb” into the text box of “4. Extend your regions at both sides.”
- f. Click “Search” to run the Regions Search tool (Fig. 6A).
- g. After the Region Search is successfully run you are presented with an output page listing each of the regions and the numbers of genes identified within the regions (Fig. 6B). To save a list of all the genes, use the “Create List by feature type” button above the output after selecting “Gene” in the pull-down menu and click “Go”. This action creates a new list and produces a List Analysis page.
- h. Click on the histogram in the “Gene Source” column header to see the column summary (Fig. 6C). You will notice that the list contains two gene sources, RefSeq and the *A. mellifera* official gene set (amel_OGSv3.2). Within the column summary pop-up box, filter the list by checking “amel_OGSv3.2”, clicking the blue filter box, and selecting “Restrict table to matching rows”. Now save the list of amel_OGSv3.2 genes using “Save as List” above the table, clicking “Gene (333 Genes)” and in the pop-up menu, naming the list “OGSv3.2 genes within 50kb of SNPs for scouting”, and clicking “Create List”. We will use the OGS gene list for the remainder of this example, but if you wish to also save the RefSeq genes, you could now click “Undo” above the table to remove the filter, then filter using the Gene Source column summary to restrict table to rows matching “Amel_RefSeq”.

- i. While in the process of saving the lists, you may have noticed enrichment widgets appearing below the table. At this point it is not advisable to rely on these enrichments, because they were performed on the original list containing identifiers from two gene sets. The enrichment widgets do not alter the original list even after you filter it.
- j. To see a valid enrichment, click the Lists tab in the navigation bar. If you are presented with the Lists Upload page rather than the View page, click “View” in the brown bar below the HymenopteraMine navigation bar. Once on the list view page, click your list name, “OGSv3.2 genes within 50kb of SNPs for scouting.” The Gene Ontology enrichment shows some significantly enriched GO terms; however this analysis is still not correct, because the default background population set is all genes in a species annotated with the appropriate data type. HymenopteraMine has two gene sets for *A. mellifera* annotated with GO terms, so two gene sets were used as the background population. To correct this, click “Change” below “Background Population.” The resulting pull-down menu shows all of your lists as well as some premade lists. For this analysis, select “A. mellifera all amel_OGSv3.2 Genes (15314)” (Fig. 7).

3. To continue with the goal of identifying differentially expressed (DE) genes [35] within 50kb of the SNPs associated with scouting behavior, the next step is to create a DE gene list. Go to the supporting online material webpage (http://science.sciencemag.org/content/suppl/2012/03/07/335.6073.1225.DC1?_ga

[=1.205106516.512133959.1337265718](#)) and download Table S3B

(http://science.sciencemag.org/highwire/filestream/593583/field_highwire_adjunct_files/0/1213962_SuppTable_S3B.xlsx). You will use the gene ids listed in column E.

- a. In HymenopteraMine, click the Lists tab in the navigation bar. If you are presented with a view of your lists rather than the Upload page, click “Upload” in the brown bar just below the navigation bar.
- b. By reading the publication [35], we know that the gene ids used in this study were from the old *A. mellifera* gene set (amel_OGSv1.0). Therefore, we need to create a list of alias identifiers and later we will convert them to amel_OGSv3.2 gene ids. Choose “Alias Name” from the “Select Type” pull-down menu and “A. mellifera” from the “for Organism” menu (Fig. 8A).
- c. Paste all the ids from Column E of the publication supplemental table into the text box. There is no need to be concerned with the 260 non-amel_OGSv1.0 identifiers that are also found in this column.
- d. Click “Create List”. HymenopteraMine performs a look-up and removes any identifiers that are not found. Notice that 775 Alias Names of the 959 amel_OGSv1.0 identifiers entered are found in HymenopteraMine. The missing 260 identifiers are from other species, such as *Drosophila melanogaster*, or OGSv1 genes that did not map to a gene in the updated *A. mellifera* gene set (Fig. 8B).

- e. Enter a name for the list such as “Alias OGSv1 DE Genes Scout vs Recruit” and click the green “Save a list of 775 AliasNames” button. You are taken to a page showing the list you created. Notice that there are no enrichment widgets as this is not a Gene list (Fig. 8C).
4. The next step is to convert your alias id list to a gene id list using a template query.
 - a. Click the Home tab in the navigation bar, and then click the ALIAS AND DBXREF tab in the template categories bar, halfway down the page.
 - b. Click the template “Alias ID → Gene ID”, or if you cannot see that template listed, click “More Queries” and then click the template name. In the template query pop-up menu, you change the default alias id to your list of ids by checking “constrain to be”, making sure “IN” is selected, and then selecting your list name “Alias OGSv1 DE Genes Scout vs Recruit”. Make sure that *A. mellifera* is selected under “Organism > Short Name”. Toggle on the optional constraint to specify a gene source, otherwise both RefSeq and amel_OGSv3.2 gene ids will be included in the output. Make sure that “amel_OGSv3.2” is selected (Fig. 9A).
 - c. Click “Show Results”. The output is a table with 855 rows showing amel_OGSv1 ids and amel_OGSv3.2 ids. Note that there is not always a one-to-one correspondence in ids due to changes in gene models in the updated gene set. For example, you will notice that the amel_OGSv1 id GB11731 is listed in both the first and second rows, with cross references

to amel_OGSv3.2 genes GB40009 and GB40010, due to the original gene being split in the new gene set (Fig. 9B).

- d. Now you must save a list of amel_OGSv3.2 gene ids by clicking “Save as List” above the table, clicking “Gene (851 Genes)” and naming the list such as “OGSv3.2 DE Genes Scout vs Recruit”. Notice that you will be saving 851 genes, even though the table has 885 rows. This is due to some of the amel_OGSv3.2 gene ids being listed more than once, due to multiple genes of the old gene set being merged in the new gene set. Finally, click “Create List” (Fig. 9C).

5. The next step is to determine if any of the 851 genes in the list “OGSv3.2 DE Genes Scout vs Recruit” are found in the list “OGSv3.2 genes within 50kb of SNPs for scouting.” Click the Lists tab in the navigation bar. If necessary, switch from the Upload page to the View page by clicking “View” in the brown bar below the navigation bar. Check the boxes next to the lists you made, “OGSv3.2 DE Genes Scout vs Recruit” and “OGSv3.2 genes within 50kb of SNPs for scouting.” Click “Intersect” in the white “Actions” bar above your list of lists. Name the list for the intersection, such as “DE vs genes within 50kb SNP intersection” and click “Save” (Fig. 10). Now you have a list of 22 DE genes that are located within 50kb of the SNPs associated with scouting behavior .

Once you have your final list of genes, you would like to get more information about them. You have many options. We will show two approaches: using column

manager to add gene descriptions and pathway information and using the list in a template query to retrieve GO terms.

6. The first step is to use the column manager on the table provided on the list analysis page.
 - a. If necessary, return to the table by clicking on the name of the list “DE vs genes within 50kb SNP intersection” shown on the Lists “View” page.
 - b. Click “Manage Columns” above the table. First we will remove columns that are not of interest (Fig. 11A). Click the red circle next to all columns except “Gene >> DB identifier”. We are not interested in “Gene >> Symbol” or “Gene >> Name”, because OGS gene sources are not assigned symbols or names.
 - c. Now we would like to retrieve additional information, including gene symbols, descriptions, and pathways, associated with RefSeq genes (Fig. 11B). Click the green “+ Add a Column” button on the upper right. Scroll down the data model tree to find “Db Cross References” and click the “+” button. Along the way, you may wish to close the “Organism”, “Chromosome Location” and “Chromosome” parts of the tree, which were open because their attributes were included in the original table. Within “Db Cross References”, open the “Cross Reference” subtree. Under “Cross Reference”, select “Symbol”, “Description” and “DB identifier”. Once selected, each will be highlighted in a blue bar. Scroll further down within the “Db Cross References” subtree and open the sub-subtree for

“Pathways”. Be sure that the “Pathways” you open is descended from “Db Cross References”, rather than the one you could find further down the page directly descended from “Gene”. Within the correct “Pathways”, select “Identifier” and “Name” so that each is highlighted in a blue bar. Now all of the columns you will be adding are highlighted in blue.

- d. Click the green “Add 5 new columns” box. You are returned to the column list, which includes your new columns listed in the order of output. If you wish to change the output order, use the arrows next to the column name to move the position (Fig. 11C).
- e. Click “Apply Changes”. You will notice that you have only 9 rows of output, even though you started with 22 gene ids. This is due to the lack of pathway information for 13 genes. You would still like to see symbols and gene descriptions for all genes, so you can modify the default requirement that a relationship must be present in order for output to be provided.
- f. Click the “Manage Relationships” button. Toggle the relationship for “Gene >> Db Cross References >> Cross Reference Pathways” to be optional, and click “Apply Changes” (Fig. 12A). The table now has a row for each gene. When you make a relationship optional, the table format is modified so that rows contain embedded subtables, each indicating the number of results in that subtable. Clicking on a subtable opens it for viewing. Your output shows that most genes have zero or one pathway annotation. GB52164 (cyclin-dependent kinase 7) and GB52468 (cytochrome b-c1 complex subunit 8) are each annotated with two

pathways. Click “2 pathways” in a cell to view the pathway names (Fig. 12B). If you choose to export the table using the “Export” button, and download all columns, the pathway information will no longer be shown as embedded rows. For this table, the exported file has 24 rows, since two genes each have two pathways, and 14 of the rows have no information in the pathway columns.

7. The final task is to use a template query to retrieve gene ontology terms for our intersection gene list. Click “Home” in the navigation bar and click “Function” in the template category bar on the home page.
 - a. Click the template name “Gene → GO Terms” (Fig. 13A). In the template pop-up menu, under “Gene > DB identifier” check “constrain to be”, select “IN”, and select your intersection list, “DE vs genes within 50kb SNP intersection”. There is no need to select the organism, since these identifiers are unique to *A. mellifera*, but if you are unsure, you could toggle on the “Organism → Short Name” constraint and select “A. mellifera”.
 - b. Click “Show Results”. The output provides GO identifiers and terms. Clicking on the histogram symbol to show the column summary for Gene DB Identifier shows that 19 of your 22 genes are annotated with GO terms (Fig. 13B). Mousing over either a GO term identifier or name within the table allows you to see a description of the term. The Namespace column indicates whether each term is part of the biological process, molecular

function or cellular component ontology, and you can filter for any one of these using the column header tools. The Qualifier column shows whether any term is annotated with a term such as “NOT” or “Contributes to”, which would indicate that the given function is not intended to describe the actual function of the gene. The grey “NO VALUE” result represents the usual case in which the gene is annotated to have the function listed. It is always a good idea to view the column summary in the Qualifier column to see if it contains any information needed to properly interpret a GO annotation, and perhaps filter out rows with “NOT” or keep only rows with “NO VALUE”.

3. Notes

3.1 HymenopteraMine Release for this Chapter

The most current release of HymenopteraMine can be accessed from the main HGD navigation bar, or with the following URL:

<http://hymenopteragenome.org/hymenopteramine/>. This chapter is based on HymenopteraMine release 1.2, which will be maintained here once it is no longer the current release: <http://hymenopteragenome.org/hymenopteramine-release-1.2/>. You can perform this example anonymously, but it is advisable to login to your MyMine account so that you can save your work.

3.2 Computing Database Cross References

We compute database cross references for gene identifiers based on overlapping coding exons on the same strand. When alternate assemblies are used for generating the primary gene sets, we first use the UCSC LiftOver tool [37] to map OGS genes to the newest RefSeq genome assembly. We do not attempt to compute database cross references for individual transcripts.

3.3 Computing Aliases

To assign HGD identifiers to consortium gene sets that do not have gene ids, we identify transcripts originating from the same gene locus based on overlapping coding sequences. Each group of overlapping transcripts is assigned a gene identifier in the form of species code plus a set of digits (e.g. SINV10017). Transcripts are assigned identifiers that include the gene id plus a suffix in the form of “-RA”, where the last letter varies to distinguish transcript isoforms. There is a 1-to-1 relationship between consortium transcript identifiers and HGD transcript identifiers.

In addition to assigning HGD identifiers, we compute aliases for old and new gene sets of the same source (e.g. aliases between *A. mellifera* OGSv1 and OGSv3.2). In order to do so, we first map the old gene set to a newer assembly (if necessary) using the UCSC LiftOver tool, and then compute overlapping coding sequences on the same strand to identify coding transcripts from the same gene locus. We then assign as Aliases any pair of genes that have any coding sequence in common. Most gene ids have 1-to-1 alias relationships, but some genes have no alias or multiple aliases due to disagreement in gene models.

3.5 Computing Gene Expression Levels and Variant Effects

We download fastq files for *A. mellifera* Illumina runs with reads of at least 100bp from the SRA, trim for adaptors using Fastq-MCF (<https://code.google.com/p/ea-utils/wiki/FastqMcf>), trim for quality using DynamicTrim [38] and align reads to the *A. mellifera* genome assembly Amel_4.5 using TopHat2 [39]. We determine FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and normalized read counts for each expression dataset for transcripts in the amel_OGSv3.2 and RefSeq gene sets using cuffquant and cuffnorm, which are part the Cufflinks package [40]. We also use HTSeq [41] to determine raw read counts per transcript, and use the raw counts to compute RPKM (Reads Per Kilobase of transcript per Million mapped reads). We compute *A. mellifera* SNP effects using SnpEff [42].

3.6 Orthologues

HymenopteraMine includes orthologues from OrthoDB [29], which identifies orthologous groups of genes that are descended from a single ancestral gene. We use the OrthoDB data set computed based on a common insect ancestor to allow the inclusion of *Drosophila melanogaster* (a Dipteran). This means that any orthologous group in HymenopteraMine can include duplicated genes that emerged after divergence from the common insect ancestor. All pairwise relationships within an orthologous group are called orthologues even if some might be classified as paralogues in an analysis of a smaller taxonomic group (e.g. the relationship between *A. mellifera* and *A. florea* genes

since divergence from the *Apis* ancestor). Users may investigate gene lineages in OrthoDB to clarify relationships.

3.7 *Drosophila melanogaster* data

In order to leverage orthology with well-annotated fly genes, we use the same FlyBase release that was used in the OrthoDB release, and it may not be the most recent FlyBase release.

3.8 Sequence Identifiers used in BLAST Databases

Sequence identifiers used in BLAST databases of genome assemblies are the same identifiers used in JBrowse. RefSeq or GenBank chromosome and scaffold identifiers are used for all species except *A. mellifera*, for which the original consortium identifiers are used. The identifiers used for RNA or protein BLAST databases are either the RefSeq identifier or the HGD official gene set identifier. For species with new assigned HGD identifiers, the original consortium ids are also provided in the BLAST output.

Acknowledgements

The authors would like to thank Colin M. Diesh for his contributions to the development of HymenopteraMine, HGD BLAST and HGD JBrowse/Apollo. This work was supported by USDA National Institute of Food and Agriculture Hatch Project 1009273 and Agriculture and Food Research Initiative Competitive grant no. 2010-65205-20407.

4. References

1. Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE (2016) Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Res* 44 (D1):D793-800. doi:10.1093/nar/gkv1208
2. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B, Elhaik E, Evans JD, Foster LJ, Graur D, Guigo R, teams Hp, Hoff KJ, Holder ME, Hudson ME, Hunt GJ, Jiang H, Joshi V, Khetani RS, Kosarev P, Kovar CL, Ma J, Maleszka R, Moritz RF, Munoz-Torres MC, Murphy TD, Muzny DM, Newsham IF, Reese JT, Robertson HM, Robinson GE, Rueppell O, Solovyev V, Stanke M, Stolle E, Tsuruda JM, Vaerenbergh MV, Waterhouse RM, Weaver DB, Whitfield CW, Wu Y, Zdobnov EM, Zhang L, Zhu D, Gibbs RA, Honey Bee Genome Sequencing C (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* 15:86. doi:10.1186/1471-2164-15-86
3. Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443 (7114):931-949
4. Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, Grimmelikhuijzen CJ, Hasselmann M, Lozier JD, Robertson HM, Smagghe G, Stolle E, Van Vaerenbergh M, Waterhouse RM, Bornberg-Bauer E, Klasberg S, Bennett AK, Camara F, Guigo R, Hoff K, Mariotti M, Munoz-Torres M, Murphy T, Santesmasses D, Amdam GV, Beckers M, Beye M, Biewer M, Bitondi MM, Blaxter ML, Bourke AF, Brown MJ, Buechel SD, Cameron R, Cappelle K, Carolan JC, Christiaens O, Ciborowski KL, Clarke DF, Colgan TJ, Collins DH, Cridge AG, Dalmay T, Dreier S, du Plessis L, Duncan E, Erler S, Evans J, Falcon T, Flores K, Freitas FC, Fuchikawa T, Gempe T, Hartfelder K, Hauser F, Helbing S, Humann FC,

Irvine F, Jermini LS, Johnson CE, Johnson RM, Jones AK, Kadowaki T, Kidner JH, Koch V, Kohler A, Kraus FB, Lattorff HM, Leask M, Lockett GA, Mallon EB, Antonio DS, Marxer M, Meeus I, Moritz RF, Nair A, Napflin K, Nissen I, Niu J, Nunes FM, Oakeshott JG, Osborne A, Otte M, Pinheiro DG, Rossie N, Rueppell O, Santos CG, Schmid-Hempel R, Schmitt BD, Schulte C, Simoes ZL, Soares MP, Swevers L, Winnebeck EC, Wolschin F, Yu N, Zdobnov EM, Aqrabi PK, Blankenburg KP, Coyle M, Francisco L, Hernandez AG, Holder M, Hudson ME, Jackson L, Jayaseelan J, Joshi V, Kovar C, Lee SL, Mata R, Mathew T, Newsham IF, Ngo R, Okwuonu G, Pham C, Pu LL, Saada N, Santibanez J, Simmons D, Thornton R, Venkat A, Walden KK, Wu YQ, Debyser G, Devreese B, Asher C, Blommaert J, Chipman AD, Chittka L, Fouks B, Liu J, O'Neill MP, Sumner S, Puiu D, Qu J, Salzberg SL, Scherer SE, Muzny DM, Richards S, Robinson GE, Gibbs RA, Schmid-Hempel P, Worley KC (2015) The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol* 16:76. doi:10.1186/s13059-015-0623-3

5. Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, Hernandez A, Yandell M, Ence D, Holt C, Yocum GD, Kemp WP, Bosch J, Waterhouse RM, Zdobnov EM, Stolle E, Kraus FB, Helbing S, Moritz RF, Glastad KM, Hunt BG, Goodisman MA, Hauser F, Grimmelikhuijzen CJ, Pinheiro DG, Nunes FM, Soares MP, Tanaka ED, Simoes ZL, Hartfelder K, Evans JD, Barribeau SM, Johnson RM, Massey JH, Southey BR, Hasselmann M, Hamacher D, Biewer M, Kent CF, Zayed A, Blatti C, 3rd, Sinha S, Johnston JS, Hanrahan SJ, Kocher SD, Wang J, Robinson GE, Zhang G (2015) Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. *Science* 348 (6239):1139-1143. doi:10.1126/science.aaa4788

6. Kocher SD, Li C, Yang W, Tan H, Yi SV, Yang X, Hoekstra HE, Zhang G, Pierce NE, Yu DW (2013) The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome Biol* 14 (12):R142. doi:10.1186/gb-2013-14-12-r142
7. Nygaard S, Zhang G, Schiott M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M, Pan H, Hauser F, Krogh A, Grimmelikhuijzen CJ, Wang J, Boomsma JJ (2011) The genome of the leaf-cutting ant *Acromyrmex echinator* suggests key adaptations to advanced social life and fungus farming. *Genome Res* 21 (8):1339-1348. doi:10.1101/gr.121392.111
8. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Fave MJ, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkampf M, Hu H, Johnson BR, Kim J, Marsh SE, Moeller JA, Munoz-Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S, Tsutsui ND, Viljakainen L, Wissler L, Yandell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genetics* 7 (2):e1002007. doi:10.1371/journal.pgen.1002007
9. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, Zhang P, Huang Z, Berger SL, Reinberg D, Wang J, Liebig J (2010) Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* 329 (5995):1068-1071. doi:10.1126/science.1192428
10. Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, Weichselgartner T, Kemena C, Stokl J, Schultner E, Wurm Y, Smith CD, Yandell M, Heinze J, Gadau J, Oettler J (2014)

Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat Commun* 5:5495. doi:10.1038/ncomms6495

11. Oxley PR, Ji L, Fetter-Pruneda I, McKenzie SK, Li C, Hu H, Zhang G, Kronauer DJ (2014)

The genome of the clonal raider ant *Cerapachys biroi*. *Current biology* : CB 24 (4):451-458.

doi:10.1016/j.cub.2014.01.018

12. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E,

Elsik CG, Fave MJ, Fernandes V, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE,

Helmkampf M, Holley JA, Hu H, Viniegra AS, Johnson BR, Johnson RM, Khila A, Kim JW, Laird

J, Mathis KA, Moeller JA, Munoz-Torres MC, Murphy MC, Nakamura R, Nigam S, Overson RP,

Placek JE, Rajakumar R, Reese JT, Robertson HM, Smith CR, Suarez AV, Suen G, Suhr EL, Tao

S, Torres CW, van Wilgenburg E, Viljakainen L, Walden KK, Wild AL, Yandell M, Yorke JA,

Tsutsui ND (2011) Draft genome of the globally widespread and invasive Argentine ant

(*Linepithema humile*). *Proc Natl Acad Sci U S A* 108 (14):5673-5678.

doi:10.1073/pnas.1008617108

13. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H,

Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave MJ, Fernandes V,

Gibson JD, Graur D, Gronenberg W, Grubbs KJ, Hagen DE, Viniegra AS, Johnson BR, Johnson

RM, Khila A, Kim JW, Mathis KA, Munoz-Torres MC, Murphy MC, Mustard JA, Nakamura R,

Niehuis O, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Suen G, Tao S, Torres

CW, Tsutsui ND, Viljakainen L, Wolschin F, Gadau J (2011) Draft genome of the red

harvester ant *Pogonomyrmex barbatus*. *Proc Natl Acad Sci U S A* 108 (14):5667-5672.

doi:10.1073/pnas.1007901108

14. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra MB, Oettler J, Comtesse F, Shih CJ, Wu WJ, Yang CC, Thomas J, Beaudoin E, Pradervand S, Flegel V, Cook ED, Fabbretti R, Stockinger H, Long L, Farmerie WG, Oakey J, Boomsma JJ, Pamilo P, Yi SV, Heinze J, Goodisman MA, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L (2011) The genome of the fire ant *Solenopsis invicta*. *Proc Natl Acad Sci U S A* 108 (14):5679-5684.

doi:10.1073/pnas.1009690108

15. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Nasonia Genome Working G, Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJ, Kitts P, Lynch JA, Murphy T, Oliveira DC, Smith CD, van de Zande L, Worley KC, Zdobnov EM, Aerts M, Albert S, Anaya VH, Anzola JM, Barchuk AR, Behura SK, Bera AN, Berenbaum MR, Bertossa RC, Bitondi MM, Bordenstein SR, Bork P, Bornberg-Bauer E, Brunain M, Cazzamali G, Chaboub L, Chacko J, Chavez D, Childers CP, Choi JH, Clark ME, Claudianos C, Clinton RA, Cree AG, Cristino AS, Dang PM, Darby AC, de Graaf DC, Devreese B, Dinh HH, Edwards R, Elango N, Elhaik E, Ermolaeva O, Evans JD, Foret S, Fowler GR, Gerlach D, Gibson JD, Gilbert DG, Graur D, Grunder S, Hagen DE, Han Y, Hauser F, Hultmark D, Hunter HCt, Hurst GD, Jhangian SN, Jiang H, Johnson RM, Jones AK, Junier T, Kadowaki T, Kamping A, Kapustin Y, Kechavarzi B, Kim J, Kim J, Kiryutin B, Koevoets T, Kovar CL, Kriventseva EV, Kucharski R, Lee H, Lee SL, Lees K, Lewis LR, Loehlin DW, Logsdon JM, Jr., Lopez JA, Lozado RJ, Maglott D, Maleszka R, Mayampurath A, Mazur DJ, McClure MA, Moore AD, Morgan MB, Muller J, Munoz-Torres MC, Muzny DM, Nazareth LV, Neupert S, Nguyen NB, Nunes FM, Oakeshott JG, Okwuonu GO, Pannebakker BA, Pejaver VR, Peng Z, Pratt SC, Predel R, Pu LL, Ranson H, Raychoudhury R,

Rechtsteiner A, Reese JT, Reid JG, Riddle M, Robertson HM, Romero-Severson J, Rosenberg M, Sackton TB, Sattelle DB, Schluns H, Schmitt T, Schneider M, Schuler A, Schurko AM, Shuker DM, Simoes ZL, Sinha S, Smith Z, Solovyev V, Souvorov A, Springauf A, Stafflinger E, Stage DE, Stanke M, Tanaka Y, Telschow A, Trent C, Vattathil S, Verhulst EC, Viljakainen L, Wanner KW, Waterhouse RM, Whitfield JB, Wilkes TE, Williamson M, Willis JH, Wolschin F, Wyder S, Yamada T, Yi SV, Zecher CN, Zhang L, Gibbs RA (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327 (5963):343-348. doi:10.1126/science.1178028

16. Rago A, Gilbert DG, Choi JH, Sackton TB, Wang X, Kelkar YD, Werren JH, Colbourne JK (2016) OGS2: genome re-annotation of the jewel wasp *Nasonia vitripennis*. *BMC Genomics* 17:678. doi:10.1186/s12864-016-2886-9

17. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491. doi:10.1186/1471-2105-12-491

18. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* 14 (8):R93. doi:10.1186/gb-2013-14-8-r93

19. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, Holmes IH (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 17:66. doi:10.1186/s13059-016-0924-1

20. Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W, Wurm Y (2015) Sequenceserver: a modern graphical user interface for custom BLAST databases. *BioRxIV*. doi:https://doi.org/10.1101/033142

21. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* 28 (23):3163-3165. doi:10.1093/bioinformatics/bts577
22. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, Tyers M (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 45 (D1):D369-D379. doi:10.1093/nar/gkw1102
23. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29 (1):308-311
24. Gramates LS, Marygold SJ, Santos GD, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, Falls K, Goodman JL, Hu Y, Ponting L, Schroeder AJ, Strelets VB, Thurmond J, Zhou P, the FlyBase C (2017) FlyBase at 25: looking to the future. *Nucleic Acids Res* 45 (D1):D663-D671. doi:10.1093/nar/gkw1016
25. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, K€orninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res* 44 (D1):D481-487. doi:10.1093/nar/gkv1351
26. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res* 43 (Database issue):D1049-1056. doi:10.1093/nar/gku1179
27. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat

S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* 45 (D1):D190-D199.

doi:10.1093/nar/gkw1107

28. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* 40 (Database issue):D109-114. doi:10.1093/nar/gkr988

29. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV (2017) OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.

Nucleic Acids Res 45 (D1):D744-D749. doi:10.1093/nar/gkw1119

30. NCBI Resource Coordinators (2017) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 45 (D1):D12-D17.

doi:10.1093/nar/gkw1071

31. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference

sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44 (D1):D733-745. doi:10.1093/nar/gkv1189

32. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Consortium (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 40 (Database issue):D54-56. doi:10.1093/nar/gkr854

33. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45 (D1):D158-D169. doi:10.1093/nar/gkw1099

34. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C (2015) The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res* 43 (Database issue):D1057-1063. doi:10.1093/nar/gku1113

35. Liang ZS, Nguyen T, Mattila HR, Rodriguez-Zas SL, Seeley TD, Robinson GE (2012) Molecular determinants of scouting behavior in honey bees. *Science* 335 (6073):1225-1228. doi:10.1126/science.1213962

36. Southey BR, Zhu P, Carr-Markell MK, Liang ZS, Zayed A, Li R, Robinson GE, Rodriguez-Zas SL (2016) Characterization of Genomic Variants Associated with Scout and Recruit Behavioral Castes in Honey Bees Using Whole-Genome Sequencing. *PLoS One* 11 (1):e0146430. doi:10.1371/journal.pone.0146430

37. Hickey G, Paten B, Earl D, Zerbino D, Haussler D (2013) HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* 29 (10):1341-1342. doi:10.1093/bioinformatics/btt128

38. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485. doi:10.1186/1471-2105-11-485

39. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14 (4):R36. doi:10.1186/gb-2013-14-4-r36
40. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 41 (5):511-515. doi:10.1038/nbt.1621
14. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31 (2):166-169. doi:10.1093/bioinformatics/btu638
42. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6 (2):80-92. doi:10.4161/fly.19695

Figure Legends

Figure 1. The Hymenoptera Genome Database (HGD) includes three divisions – BeeBase, NasoniaBase and the Ant Genomes Portal. The navigation bar on the HGD home page provides access to each of the divisions, as well as to resources available to all three divisions (e.g. HymenopteraMine and BLAST). The navigation bar for each division provides tabs to division specific resources (e.g. JBrowse and data download pages) as well as links the HGD shared resources and the HGD homepage.

Figure 2. The HymenopteraMine navigation bar is available on all HymenopteraMine pages. The HymenopteraMine home pages provides the Quick Search and Quick List tools, and access to categorized template queries.

Figure 3. The Quick Search tool performs a full text search and retrieves all data objects containing the searched term. A faceted selector to the left of the result list allows you to filter results by data class or organism.

Figure 4. The hierarchical data model tree, accessible by clicking “Browse Data Model” on the QueryBuilder page, shows the data classes in HymenopteraMine and numbers of data objects for each class. Clicking an “i” symbol provides a definition for a data class.

Figure 5. The Model Browser and Query Overview are used in query construction. A) Query construction is initiated by clicking CONSTRAIN. A pop-up menu allows entry of a constraint, such as an identifier. After clicking “Add to Query” the constraint is shown in the Query

Overview. B) During query construction, to easily navigate to the correct subtree to add additional attributes related to a particular data class, click on the word in brown font shown next to the data class in the Query Overview. Here, clicking “Gene” next to “Cross Reference” allows you to navigate to the correct area of the Model Browser to add the symbol of the cross reference gene. C) The first subclass under “Homologues Homologue” is “Gene”, with a red up arrow, indicating that it is a reference to the parent “Gene” class, rather than the homologous gene. Look further down to see “Homologue Gene”. Clicking “Show” next to “DB identifier” under “Homologue Gene” adds an output column for the gene id of the homologue. D) The “DB Identifier” selected to show here is for the “Cross Reference Gene” within the “Db Cross References x Ref” collection that is a subclass of “Homologue Gene” rather than the “Db Cross References x Ref” collection that is a subclass of “Gene”. E) This final Query Overview shows two query constraints and eight output columns. F) After query construction is complete, you can rearrange columns in the “Columns to Display” section. At the bottom of the page are options to save, export and run the query. The “Start building a template query” button shows up if you are logged into MyMine.

Figure 6. A) The Regions search interface takes chromosome coordinate input lists in several formats. B) The output of a Regions search consists of pages showing results for each region. The “Create List by feature type” function allows you to create a list of identifiers for further use in HymenopteraMine. C) After creating a list, the list is presented in a List Analysis page. Since this is a list of all *A. mellifera* genes within the searched regions, it includes genes from two gene sets. The table can be filtered for one gene set using the column summary tool in the column header. D) A new list can be created from the filtered table.

Figure 7. Enrichment Widgets appear when you view any list of gene identifiers. However, the default background population gene list is usually not the appropriate dataset for the analysis. The background population can be changed and replaced with one of the premade gene lists for each species, or you can use the Lists tool to create a custom background population gene list for your study.

Figure 8. A) The Lists Upload page accepts any identifier, but the identifier will not be validated in the HymenopteraMine lookup step unless the correct data class is selected. Here “Alias Name” is selected, because the *A. mellifera* identifiers in the entered list are from the old (amel_OGSv1) gene set. B) After performing a lookup, some identifiers are eliminated from the list because they are not found in HymenopteraMine. C) Clicking “Save a list of 775 AliasNames” in the previous figure leads to this list of alias identifiers. Enrichment widgets are not provided since this is not a list of primary gene set identifiers.

Figure 9. A) The list of alias ids can be used in the “Alias ID → Gene ID” template query to convert the identifiers to primary gene set ids. The optional constraint for Gene > Source is turned on, and amel_OGSv3.2 is selected so that only ids from that gene set will be returned. B) The output of the template query includes amel_OGSv1 ids and amel_OGSv3,2 ids. The first two rows demonstrate that there is not always a one-to-one correspondence in ids due to changes in gene models in the updated gene set, as amel_OGSv1 id GB11731 is listed in both rows, with cross references to amel_OGSv3.2 genes GB40009 and GB40010, due to the original gene being split in the new gene set. C) Saving the final list of amel_OGSv3.2 gene ids by clicking “Save as

List” above the table, shows that 851 genes will be saved, even though the table has 885 rows. This is due to multiple genes of the old gene set being merged in the new gene set.

Figure 10. A List intersection is performed on the Lists View page by selecting each list, clicking “Intersection”, and providing a name for the new list.

Figure 11. A) The “Manage Columns” button above the HymenopteraMine table leads to an interface displaying the columns in order. A column can be deleted by clicking the red circle, and the column position can be modified using the up or down arrows. B) Clicking “+ Add a Column” in the previous menu leads to a hierarchical display of the data model, similar to the Model Browser in the QueryBuilder, but without the means for adding constraints. A column is selected by clicking the name so it is highlighted in a blue bar. C) Once clicking “Apply Changes” after selecting columns, the column list is updated to include the new columns. Clicking “Apply Changes” once more leads to the updated table.

Figure 12. A) The Manage Relationships interface allows you to modify the default property that all relationships must be present in order for a row to be present in the output. In our example, we started with 22 genes but after adding new columns (Fig. 11), the number of genes in the table is reduced to 9. This is due to lack of pathway information for 13 genes. We can modify the relationship to remove the requirement for the existence of a pathway. B) After modifying the pathway relationship, all 22 genes are included in the output table. Making the relationship optional also changes the format of the output, so that the column(s) included in the optional

relationship are now shown as embedded subtables that can be clicked on to view. Exporting the table changes the format back to inline with missing values for some rows.

Figure 13. A) Our gene list can be used in the Gene → GO Term template query. It is not necessary to constrain the organism since the identifiers are unique to *A. mellifera*. B) The output of the template query shows that 19 of the genes are annotated with GO terms. The faded words “NO VALUE” in the Qualifier column are an important output in this table. “NO VALUE” is the result we are looking for. We may wish to filter the output for Qualifiers other than “NO VALUE” (such as “NOT”) so that we interpret the GO Terms correctly.

Tables

Table 1. Species in the Hymenoptera Genome Database

HGD Division	Species	Common Name or Group	Genome/Gene Set Reference(s)
BeeBase	<i>Apis mellifera</i>	European honey bee	[2, 3]
	<i>Apis dorsata</i>	giant honey bee	
	<i>Apis florea</i>	dwarf honey bee	
	<i>Bombus impatiens</i>	common Eastern bumble bee	[4]
	<i>Bombus terrestris</i>	buff-tailed bumble bee	[4]
	<i>Eufriesea mexicana</i>	orchid bee	[5]
	<i>Dufourea novaeangliae</i>	sweat bee	[5]
	<i>Habropoda laboriosa</i>	Southeastern blueberry bee	[5]
	<i>Lasioglossum albipes</i>	sweat bee	[6]
	<i>Megachile rotundata</i>	alfalfa leafcutting bee	[5]
	<i>Melipona quadrifasciata</i>	stingless bee	[5]
Ant Genomes Portal	<i>Acromyrmex echinator</i>	Panamanian leaf cutter ant	[7]
	<i>Atta Cephalotes</i>	leaf cutter ant	[8]
	<i>Camponotus floridanus</i>	Florida carpenter ant	[9]
	<i>Cardiocondyla obscurior</i>		[10]
	<i>Cerapachys biroi</i>	clonal raider ant	[11]
	<i>Harpegnathos saltator</i>	jumping ant	[9]
	<i>Linepithema humile</i>	Argentine ant	[12]
	<i>Pogonomyrmex barbatus</i>	red harvester ant	[13]
	<i>Solenopsis invicta</i>	red fire ant	[14]
	<i>Wasmannia auropunctata</i>	little fire ant	
NasoniaBase	<i>Nasonia vitripennis</i>	parasitoid jewel wasp	[15,16]

Table 2. HymenopteraMine External Data Sources

Data Source	Reference
BioGrid	[22]
dbSNP	[23]
FlyBase	[24]
FlyReactome	[25]
Gene Ontology	[26]
InterPro	[27]
KEGG	[28]
OrthoDB	[29]
PubMed	[30]
Reactome	[25]
RefSeq	[31]
SRA	[32]
UniProt	[33]
UniProt-GOA	[34]

Table 3. Gene set and alias data source names in HymenopteraMine. “HGD” and “C” (for Consortium) indicate alternative identifiers for official gene sets. OrthoDB aliases may change in the future. “G” or “T” indicates whether alias ids are for genes (G) or transcripts (T).

Species	Primary Gene Set Data Source Name(s)	Alias Source Name(s) (G or T)
A. dorsata	Ador_RefSeq	Ador_OrthoDB (G)
A. echinator	aech_OGSv3.8_HGD, Aech_RefSeq	aech_OGSv3.8_C (T)
A. florea	Aflo_RefSeq	Aflo_OrthoDB (G)
A. mellifera	amel_OGSv3.2, Amel_RefSeq	amel_OGSv1 (G)
B. impatiens	bimp_OGSv1.0, Bimp_RefSeq	Bimp_OrthoDB (G)
B. terrestris	Bter_RefSeq	Bter_OrthoDB (G)
C. biroi	armyant.OGS.V1.8.6_HGD, Cbir_RefSeq	armyant.OGS.V1.8.6_C (T)
C. floridanus	cflo_OGSv3.3_HGD, Cflo_RefSeq	cflo_OGSv3.3_C (T)
C. obscurior	cobs_OGSv1.4	
D. novaeangliae	Dufourea_novaeangliae_v1.1_HGD, Dnov_RefSeq	Dufourea_novaeangliae_v1.1_C, Dnov_OrthoDB (T)
E. mexicana	Eufriesea_mexicana_v1.1_HGD	Eufriesea_mexicana_v1.1_C, Emex_OrthoDB (T)
H. laboriosa	Habropoda_laboriosa_v1.2_HGD	Habropoda_laboriosa_v1.2_C, Hlab_OrthoDB (T)
H. saltator	hsal_OGSv3.3_HGD, Hsal_RefSeq	hsal_OGSv3.3_C (T)
L. albipes	lalb_OGSv5.42_HGD	lalb_OGSv5.42_C (T)
L. humile	lhum_OGSv1.2, Lhum_RefSeq	
M. rotundata	Megachile_rotundata_v1.1_HGD, Mrot_RefSeq	Megachile_rotundata_v1.1_C, Mrot_OrthoDB (T)
M. quadrifasciata	Melipona_quadrifasciata_v1.1_HGD	Melipona_quadrifasciata_v1.1_C, Mqua_OrthoDB (T)
N. vitripennis	nvit_OGSv1.2, Nvit_EviGene, Nvit_RefSeq	
P. barbatus	pbar_OGSv1.2, Pbar_RefSeq	
S. invicta	sinv_OGSv2.2.3_HGD, Sinv_RefSeq	sinv_OGSv2.2.3_C (T)
W. auropunctata	Waur_RefSeq	Waur_OrthoDB (G)

E=Evidential Gene Set (*N. vitripennis*), F=FlyBase, O=OGS, R=RefSeq, U=UniProt.

[illegible]

Table 5. SNP locations from [9] for use in the example described in Subheading 2.4.11.

Group1:2097573..2097573	Group6:16625187..16625187	Group12:2253248..2253248
Group1:2097586..2097586	Group7:9560908..9560908	Group12:2253284..2253284
Group1:2179980..2179980	Group8:12005663..12005663	Group12:2253301..2253301
Group1:7106109..7106109	Group9:4087742..4087742	Group12:2253308..2253308
Group1:9605960..9605960	Group9:7003402..7003402	Group12:2253502..2253502
Group1:9612560..9612560	Group9:11035994..11035994	Group12:2253524..2253524
Group1:9976934..9976934	Group10:5450230..5450230	Group12:2253528..2253528
Group2:3079339..3079339	Group10:8609127..8609127	Group12:2253589..2253589
Group2:7060217..7060217	Group10:8624631..8624631	Group12:2253812..2253812
Group2:7064150..7064150	Group12:1118838..1118838	Group12:2253824..2253824
Group2:8478663..8478663	Group12:1137567..1137567	Group12:2278672..2278672
Group2:9476451..9476451	Group12:1219747..1219747	Group12:2282983..2282983
Group2:14317970..14317970	Group12:1294097..1294097	Group12:2302941..2302941
Group4:5736174..5736174	Group12:1325665..1325665	Group12:2383470..2383470
Group4:5812780..5812780	Group12:1421540..1421540	Group12:2425407..2425407
Group4:5816424..5816424	Group12:1473174..1473174	Group12:2484094..2484094
Group4:5823271..5823271	Group12:1473195..1473195	Group12:2618272..2618272
Group4:6418141..6418141	Group12:1640082..1640082	Group12:2872625..2872625
Group4:6472951..6472951	Group12:1744256..1744256	Group12:2872631..2872631
Group5:471817..471817	Group12:1770057..1770057	Group12:2996797..2996797
Group5:471818..471818	Group12:1805764..1805764	Group12:3535567..3535567
Group5:10104581..10104581	Group12:1814177..1814177	Group12:3535855..3535855
Group5:10232114..10232114	Group12:1814180..1814180	Group12:3540115..3540115
Group5:11006033..11006033	Group12:1820238..1820238	Group12:3607075..3607075
Group5:11012508..11012508	Group12:1842196..1842196	Group12:3614603..3614603
Group5:11012515..11012515	Group12:1842255..1842255	Group12:3638525..3638525
Group5:11012516..11012516	Group12:1842386..1842386	Group12:3682798..3682798
Group5:11012538..11012538	Group12:1842411..1842411	Group12:3845071..3845071
Group5:11048772..11048772	Group12:1842525..1842525	Group12:4531251..4531251
Group5:11062231..11062231	Group12:1859484..1859484	Group12:4636018..4636018
Group5:11063028..11063028	Group12:1859516..1859516	Group12:5139164..5139164
Group5:11095696..11095696	Group12:1882625..1882625	Group12:5500706..5500706
Group5:11104858..11104858	Group12:1916728..1916728	Group12:9498320..9498320
Group5:11134225..11134225	Group12:2160939..2160939	Group13:9587246..9587246
Group5:11286749..11286749	Group12:2195252..2195252	Group15:6478761..6478761
Group6:1062671..1062671	Group12:2196250..2196250	Group16:1670311..1670311
Group6:4544720..4544720	Group12:2206117..2206117	Group16:1670313..1670313
Group6:12559697..12559697	Group12:2215036..2215036	Group16:6118876..6118876

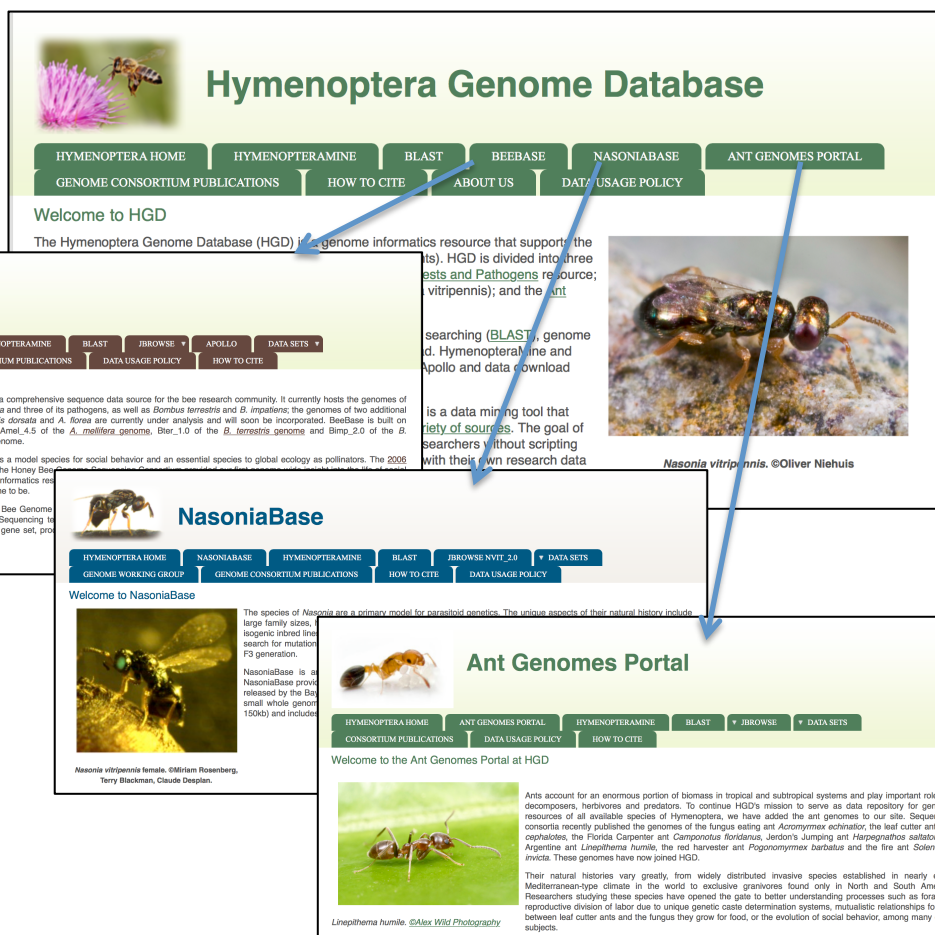




Figure 1


HymenopteraMine v1.2 An integrated data warehouse for the [Hymenoptera Genome Database](#)
[tutorial](#) | [about](#) | [cite](#)

[Home](#)
[MyMine](#)
[Templates](#)
[Lists](#)
[QueryBuilder](#)
[Regions](#)
[Data Sources](#)
[Data Model](#)
[Help](#)
[API](#)
[HGD BLAST](#)

[Contact Us](#) | beeminer@beebase.org | [Log out](#)


Search: [GO](#)



Quick Search

Search HymenopteraMine. Enter **names, identifiers or keywords** for genes, proteins, ontology terms, authors, etc. (e.g. GB45565, GB40018, NM_001011573.1, NMDA receptor 1, ACEP17531, SINV10001).

SEARCH



Quick List

Enter a list of identifiers.

e.g. GB41586, Sec61Beta, TRAM, Mocs1, mal, GB41212, GB53085, ses8, Hsp90

advanced

ANALYZE

About v1.2 and Templates

HymenopteraMine v1.2 integrates genomic data for bees, ants and parasitoid jewel wasp. Expression and variation data are provided for *A. mellifera*. The tabs below show query template categories. "Alias and DBxref" templates are for id conversion between same-species gene sets. "Entire Gene Set" templates are for querying an organism's entire gene set.

TUTORIAL

GENES
 GENE EXPRESSION
 PROTEINS
 HOMOLOGY
 FUNCTION
 VARIATION
 ALIAS AND DBXREF
 ENTIRE GENE SET

HymenopteraMine includes GO annotations for most species, KEGG pathways for *A. mellifera* and *D. melanogaster*, Reactome and Fly Reactome pathways for *D. melanogaster*, and interactions (BioGRID and IntAct) for *D. melanogaster*. You can use orthologous relationships to leverage *A. mellifera* and *D. melanogaster* information for other species.

Query for function:

- Gene ➔ GO Terms
- Gene ➔ Pathways
- GO Term ➔ Gene
- Gene ➔ *A. mellifera* Homologues ➔ *A. mellifera* Pathways, *A. mellifera* Transcripts, *A. mellifera* Expression

Figure 2

Quick Search

Search Hymenoptera

Enter names, identifiers or keywords for genes, proteins, ontology authors, etc. (e.g. GB45565, C NM_001011573.1, NMDA receptor ACEP17531, SINV10001).

Vitellogenin

SEARCH

Search results 1 to 100 out of 275 for *Vitellogenin*

<< First < Previous | Next > Last >>

2.5%

Categories

Hits by Category

- UniProt Feature: 123
- Publication: 57
- mRNA: 33
- Protein: 29
- Gene: 26
- Protein Domain: 6
- GO Term: 1

Hits by Organism

- A. florea: 9
- C. biroi: 7
- A. echinator: 6
- A. mellifera: 6
- H. saltator: 6
- L. humile: 6
- W. auropunctata: 6
- C. floridanus: 5
- D. melanogaster: 5
- D. novaeangliae: 5
- N. vitripennis: 5
- P. barbatus: 4
- S. invicta: 4
- H. laboriosa: 3
- M. quadrifasciata: 3
- A. dorsata: 2
- B. impatiens: 2
- B. terrestris: 2
- M. rotundata: 2

Type	Details	Score
Gene	406088 - - Vg <div>Source: Amel_RefSeq Status: Protein Coding Length: 6177 FASTA Chromosome: Group4: 4020743-4026919 Location: A. mellifera</div>	*****
Gene	100122102 - - - <div>Source: Nvit_RefSeq Status: Protein Coding Length: 6049 FASTA Chromosome: NC_015868.2: 27386020-27392068 Location: N. vitripennis</div>	*****
GO Term	GO:0008196 vitellogenin receptor activity Description: Receiving vitellogenin, and delivering vitellogenin into the cell via endocytosis.	*****
Protein	VIT_APIME Q868N5 <div>Organism . Name: Apis mellifera Length: 1770 FASTA</div>	*****
Protein	404088ADL8 ADL8 - - - <div>Source: Cbir_RefSeq Status: Protein Coding Length: 20373 FASTA</div>	*****
Protein		
Protein		

Search results 1 to 26 out of 26 for *Vitellogenin*

Category restricted to Gene

0.054%

Categories

Category: Gene

show all

Hits by Organism

- L. humile: 3
- W. auropunctata: 3
- A. echinator: 2
- A. florea: 2
- A. mellifera: 2
- C. biroi: 2
- N. vitripennis: 2
- P. barbatus: 2
- A. dorsata: 1
- B. impatiens: 1
- B. terrestris: 1
- C. floridanus: 1
- D. novaeangliae: 1
- H. saltator: 1
- M. rotundata: 1
- S. invicta: 1

Type	Details	Score
<input type="checkbox"/> Gene	406088 - - Vg <div>Source: Amel_RefSeq Status: Protein Coding Length: 6177 FASTA Chromosome: Group4: 4020743-4026919 Location: A. mellifera</div>	*****
<input type="checkbox"/> Gene	100122102 - - - <div>Source: Nvit_RefSeq Status: Protein Coding Length: 6049 FASTA Chromosome: NC_015868.2: 27386020-27392068 Location: N. vitripennis</div>	*****
<input type="checkbox"/> Gene	105278563 - - - <div>Source: Cbir_RefSeq Status: Protein Coding Length: 20373 FASTA</div>	*****

Figure 3

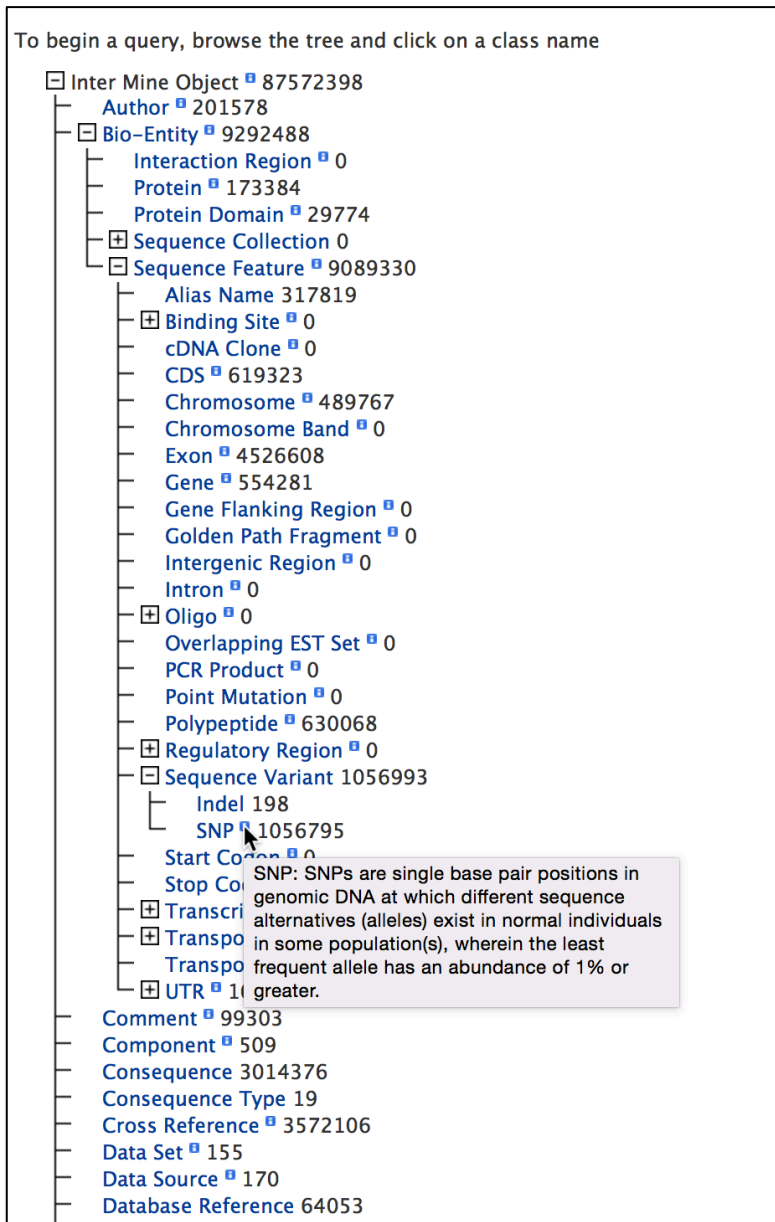


Figure 4



Figure 5

Overlap features search from a new list of Genomic Regions

Search for features that overlap a list of genome coordinates you enter or upload, e.g. scaffold0008056297.58300

Genome coordinates help

- Select Organism:
- Select Feature Types:

☐ Alias Name ☐ Indel ☐ SNP
☐ CDS ☐ mRNA ☐ tRNA
☐ Exon ☐ miRNA ☐ Three Prime UTR
☐ Five Prime UTR ☐ Polypeptide ☐ Transcript
☒ Gene ☐ Primary Transcript
- Type/Paste in genomic regions in ☐ base coordinate ☐ interbase coordinate
 (example for input format chr:1..1000)
 (example for input format chr:1..1000)
 (example for tab delimited input format)

```

Group12 4531251 4531251
Group12 4636018 4636018
Group12 5130164 5130164
Group12 5500706 5500706
Group12 9498020 9498020
Group13 9587246 9587246
Group13 6478761 6478761
Group16 1670311 1670311
Group16 1670313 1670313
Group16 6118876 6118876
      
```

 or Upload genomic regions from a .txt file...
- Extend your regions at both sides:

A

Selected organism: *A. mellifera*
 Selected feature types: Gene
 Extend Regions: 50 kbp

Hide

Export for all regions: or Create List by feature type: Page size:

GENOME REGION	FEATURE	FEATURE TYPE	LOCATION
Group1:2047573..2147573 Original input: Group1:2097573..2097573 <input type="button" value="TAB"/> <input type="button" value="CSV"/> <input type="button" value="GFF3"/> <input type="button" value="FASTA"/> <input type="button" value="BED"/>	LOC100576700 100576700	Gene	Group1:1992915..2144812
	GB47717	Gene	Group1:2017774..2075237
	TRNAS-CGA 107966080	Gene	Group1:2094643..2094724
	GB47702	Gene	Group1:2118956..2119189
Group1:2047586..2147586 Original input: Group1:2097586..2097586 <input type="button" value="TAB"/> <input type="button" value="CSV"/> <input type="button" value="GFF3"/> <input type="button" value="FASTA"/> <input type="button" value="BED"/>	LOC100576700 100576700	Gene	Group1:1992915..2144812
	GB47717	Gene	Group1:2017774..2075237
	TRNAS-CGA 107966080	Gene	Group1:2094643..2094724
	GB47702	Gene	Group1:2118956..2119189
Group1:2129980..2229980 Original input: Group1:2179980..2179980 <input type="button" value="TAB"/> <input type="button" value="CSV"/> <input type="button" value="GFF3"/> <input type="button" value="FASTA"/> <input type="button" value="BED"/>	LOC100576700 100576700	Gene	Group1:1992915..2144812
	GB47703	Gene	Group1:2170987..2171256
	LOC100576602 100576602	Gene	Group1:2174156..2181262
	GB47715	Gene	Group1:2174287..2174669
	GB47714	Gene	Group1:2177313..2177683
	GB47713	Gene	Group1:2177986..2178846

List Analysis for all_regions_Gene_list_1 (650 Genes)

Showing rows 1 to 25 of 650

Rows per page:

Gene DB Identifier	Gene Secondary Identifier	Gene Symbol	Gene Name	Gene Source	Gene Chromosome	Gene Chromosome Location . Start	Gene Chromosome Location . End	Gene Organism
100576143	NO VALUE	LOC100576143	NO VALUE	Amel_Ret	NO VALUE	6589597	16802942	A. mellifera
100576212	NO VALUE	LOC100576212	NO VALUE	Amel_Ret	NO VALUE	655933		
100576246	NO VALUE	LOC100576246	NO VALUE	Amel_Ret	NO VALUE	659923		
100576274	NO VALUE	LOC100576274	NO VALUE	Amel_Ret	NO VALUE	1053124		
100576282	NO VALUE	LOC100576282	NO VALUE	Amel_Ret	NO VALUE	673690		
100576302	NO VALUE	LOC100576302	NO VALUE	Amel_Ret	NO VALUE	827533		
100576428	NO VALUE	LOC100576428	NO VALUE	Amel_Ret	NO VALUE	113780		
100576458	NO VALUE	LOC100576458	NO VALUE	Amel_Ret	NO VALUE	117596		
100576468	NO VALUE	LOC100576468	NO VALUE	Amel_Ret	NO VALUE	1184510		

2 Gene Sources

333 Items Selected

Filter values

Gene Source

Count

amel_OGSv3.2 333

Amel_RetSeq 317

Filter

Restrict table to matching rows

Exclude matching rows from table

Group12 117596

List Analysis for all_regions_Gene_list_1 (650 Genes)

Showing rows 1 to 25 of 333

Gene > Organism (1 Organism)											
Pick items from the table											
<div><div>Create List</div><div>Add to List</div></div>											
Gene DB Identifier	Gene Secondary Identifier	Gene Symbol	Gene Name	Gene Source	Gene Chromosome Location . Start	Gene Chromosome Location . End	Gene Organism				
GB40183	NO VALUE	NO VALUE	GB40183	amel_OGSv3.2	NO VALUE	NO VALUE	3158	Group12	1677746	1680627	A. mellifera
GB40184	NO VALUE	NO VALUE	GB40184	amel_OGSv3.2	NO VALUE	NO VALUE	3003	Group12	1674005	1677162	A. mellifera
GB40185	NO VALUE	NO VALUE	GB40185	amel_OGSv3.2	NO VALUE	NO VALUE	3222	Group12	1669687	1672689	A. mellifera
GB40186	NO VALUE	NO VALUE	GB40186	amel_OGSv3.2	NO VALUE	NO VALUE	3714	Group12	1664759	1667980	A. mellifera
GB40187	NO VALUE	NO VALUE	GB40187	amel_OGSv3.2	NO VALUE	NO VALUE	581	Group12	1660152	1663865	A. mellifera
GB40188	NO VALUE	NO VALUE	GB40188	amel_OGSv3.2	NO VALUE	NO VALUE	581	Group12	1658872	1659452	A. mellifera

Go to # on this page

Figure 6

Upload | View

Search: e.g. Nasonia vitripennis, Api GO

Lists

View your own and public lists, search by keyword and compare or combine the contents of lists. Click on a list to view graphs and summaries in an analysis page, select lists using checkboxes to perform set operations. Click 'Upload' above to import a new list.

Filter: Filter: -- filter by a tag -- Reset

Actions: Union | Intersect | Subtract | Asymmetric Difference | Copy Delete Options: ☒ Show descriptions

QGSv3.2 Genes Within 50kb SNPs for Scouting 333 Genes

MY

all_regions_Gene_list_1 650 Genes

MY

A. dorsata all RefSeq Genes 11507 Genes

Can be used as background population in enrichment.

Gene Ontology Enrichment

GO terms enriched for items in this list.

Number of Genes in this list not analysed in this widget: 227

Test Correction Max p-value Ontology

Holm-Bonferroni 0.05 biological_process

Background population

Default Change

View Download

GO Term

☐ detection of chemical stimulus involved in sensory perception of smell [GO:0050907]

9.163670

☐ sensory perception of smell [GO:0007608]

1.002234

☐ detection of chemical stimulus involved in sensory perception [GO:0050907]

1.095532

☐ detection of stimulus involved in sensory perception [GO:0050906]

1.095532

☐ detection of stimulus [GO:0051606]

1.306822

☐ response to chemical [GO:0042221]

2.193839

☐ response to chemical [GO:0042221]

3.132757

Change background population

☐ Save your preference for next time

Filter...

A. dorsata all RefSeq Genes (11507)

A. echinator all RefSeq Genes (12253)

A. mellifera all RefSeq Genes (14061)

A. mellifera all amel_OGSv3.2 Genes (15314)

A. florea all RefSeq Genes (11600)

B. terrestris all RefSeq Genes (11083)

D. novaeangliae all RefSeq Genes (10043)

Can be used as background list for enrichment.

Gene Ontology Enrichment

GO terms enriched for items in this list.

Number of Genes in this list not analysed in this widget: 227

Test Correction Max p-value Ontology

Holm-Bonferroni 0.05 biological_process

Background population

A. mellifera all amel_OGSv3.2 Genes Change

View Download

GO Term	p-Value	Matches
<input type="checkbox"/> detection of chemical stimulus involved in sensory perception of smell [GO:0050911]	1.114879e-6	18
<input type="checkbox"/> sensory perception of smell [GO:0007608]	1.255940e-6	18
<input type="checkbox"/> detection of chemical stimulus [GO:0009593]	1.413388e-6	18
<input type="checkbox"/> detection of chemical stimulus involved in sensory perception [GO:0050907]	1.413388e-6	18
<input type="checkbox"/> detection of stimulus involved in sensory perception [GO:0050906]	1.588375e-6	18
<input type="checkbox"/> detection of stimulus [GO:0051606]	2.245124e-6	18
<input type="checkbox"/> response to chemical [GO:0042221]	2.443369e-6	22

Figure 7

1 Upload list of identifiers
2 Verify identifier matches
List analysis

Create a new list

Select the type of list to create and either enter in a list of identifiers or upload identifiers from a file. A search will be performed for all the identifiers in your list.

- Separate identifiers by a **comma, space, tab or new line**.
- Qualify any identifiers that contain whitespace with double quotes like so: "even skipped".

Select Type:

Alias Name

for Organism:

A. mellifera

Type/Paste in identifiers

(click to see an example)

GB12793
GB15018
BB17002
GB19995
GB11118
GB11462
GB18662
GB14621
NW_0012
GB13587

or Upload identifiers from a .txt file...

Choose File
no file selected

☐ Match on case

Reset

Create List

Choose a name for the list

Alias OGSv1 DE Genes Scout vs Recruit (e.g. Smith 2013)

Add additional matches
You entered: 959 identifiers
We found: 775 AliasNames

Save a list of 775 AliasNames

Why are the numbers different? See below.

Summary

Download summary

Synonyms

Page 1 of 155 1 2 3 4 5 ... 155

5 rows per page

Identifier you provided	Match	source	class
GB19014	GB19014	amel_OGSv1.0	AliasName
GB13034	GB13034	amel_OGSv1.0	AliasName
GB15288	GB15288	amel_OGSv1.0	AliasName
GB19343	GB19343	amel_OGSv1.0	AliasName
GB16776	GB16776	amel_OGSv1.0	AliasName

No matches found

XR_0149, GB20017, GB10290, GB30290, DB76102, GB17401, DB75251, GB15980, GB18895, GB13044, GB10570, DB74208, GB14813,

List Analysis for Alias OGSv1 DE Genes Scout vs Recruit (775 Alias Names)

Manage Columns
Export

Manage Filters
Generate Python code

Manage Relationships
Save as List

Rows per page: 25

< < > >

page 1

Showing rows 1 to 25 of 775

Alias Name	Alias ID	Alias Name	Alias source
GB10005		amel_OGSv1.0	
GB10017		amel_OGSv1.0	
GB10025		amel_OGSv1.0	
GB10039		amel_OGSv1.0	
GB10048		amel_OGSv1.0	
GB10054		amel_OGSv1.0	
GB10094		amel_OGSv1.0	
GB10095		amel_OGSv1.0	
GB10097		amel_OGSv1.0	
GB10118		amel_OGSv1.0	
GB10124		amel_OGSv1.0	

Figure 8

Alias ID → Gene ID

Given an Alias ID, retrieve Gene ID, optionally constrained by gene source.

Alias Name > Alias ID

☒ constrain to be IN

Organism > Short Name

Gene > Source

☐ optional ON | OFF ☐ constrain to be IN

A

Trail: Query

Alias ID → Gene ID

Given an Alias ID, retrieve Gene ID, optionally constrained by gene source.

Showing rows 1 to 25 of 855

Rows per page: 25

Gene Organism	Aliases Alias ID	Aliases Alias source	Gene DB identifier	Gene Source
A. mellifera	GB11731	amel_OGSv1.0	GB40009	amel_OGSv3.2
A. mellifera	GB11731	amel_OGSv1.0	GB40010	amel_OGSv3.2
A. mellifera	GB16776	amel_OGSv1.0	GB40012	amel_OGSv3.2
A. mellifera	GB15020			
A. mellifera	GB12206			
A. mellifera	GB12261			
A. mellifera	GB19979			
A. mellifera	GB17788			
A. mellifera	GB10231			
A. mellifera	GB15428			

B

Home MyMine Temp

Trail: Query

Alias ID → Gene ID

Given an Alias ID, retrieve Gene ID, optionally constrained by gene source.

Showing rows 1 to 25 of 855

Rows per page: 25

Gene Organism	Aliases Alias ID	Aliases Alias source	Gene DB identifier	Gene Source
A. mellifera	GB11731	amel_OGSv1.0	GB40009	amel_OGSv3.2
A. mellifera	GB11731	amel_OGSv1.0	GB40010	amel_OGSv3.2
A. mellifera	GB16776	amel_OGSv1.0	GB40012	amel_OGSv3.2
A. mellifera	GB15020			

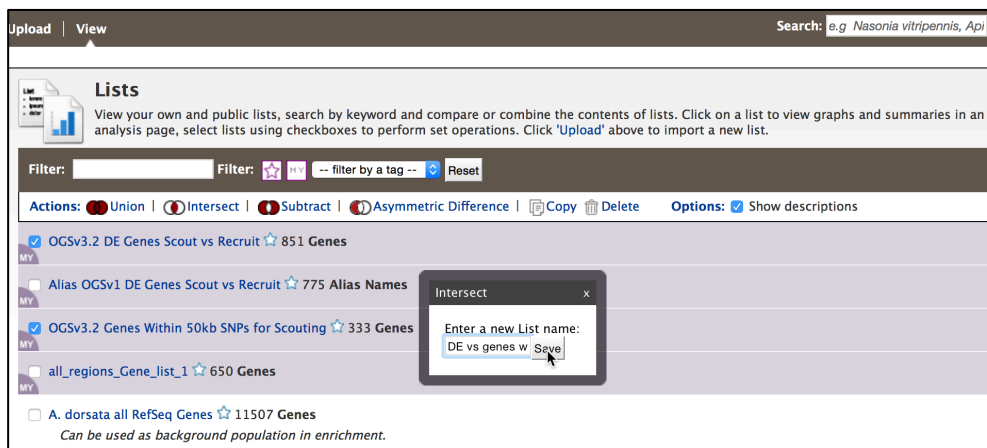
Create a new List of 851 Genes

List Name

List Description

C

Figure 9



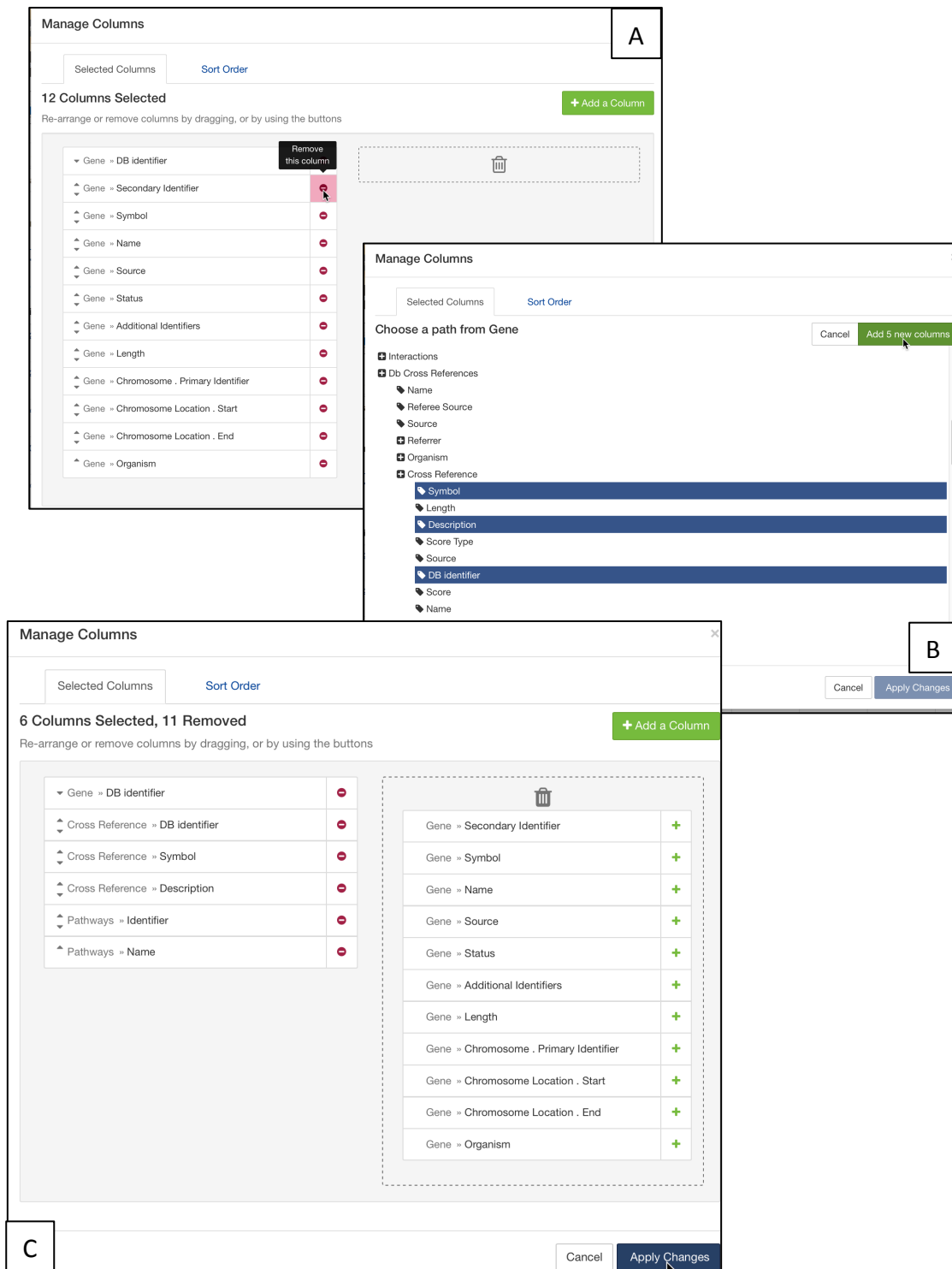


Figure 11

Gene

GO Terms

Given a gene id, retrieve GO terms.

Gene > DB identifier

=

☒ constrain to be IN

Organism > Short Name

optional

ON | OFF

Show Results

[web service URL](#) [Perl](#) [Python](#) [Ruby](#) [Java](#) [help](#) [export XML](#)

showing rows 1 to 25 of 42

Rows per page: 25

page 1

Gene DB identifier	Ontology Term Namespace	GO Annotation Qualifier
GB40221	membrane	cellular_component
GB40222	kinase activity	molecular_function
GB40222		molecular_function
GB42436		biological_process
GB42436	activity	molecular_function
GB43005	y, acting on CH-OH group of donors	molecular_function
GB43005	ptide binding	molecular_function
GB44469	membrane	cellular_component
GB44520	of ribosome	molecular_function
GB44520		cellular_component
GB44520		biological_process
GB44837		molecular_function
GB44837		molecular_function
GB44838	GO:0003676	nucleic acid binding
GB44839	GO:0003676	nucleic acid binding
GB44839	GO:0046872	metal ion binding
GB50414	GO:0006886	intracellular protein transport
GB50414	GO:0008565	protein transporter activity

Figure 13