Nancy Garnhart
University of New Hampshire
2010

Homologue Inspector
(HomeIn)

Table of Contents

Introduction

We have created a software application, Homology Inspector (HomIn), that enables users to look at subsets of gene homolog families created by large-scale sequence comparison techniques. This program is written in Java 1.6 using the database tool db4o for Java version 7.12 (Versant) to store and query a set of gene homolog families. It is designed to process the results of the pipeline described by Cooper, et al (Cooper, Vohr et al. 2010).

Two different types of homologous relationships are used to classify the genes in newly sequenced genomes such that functional and evolutionary information can be derived: orthologs and paralogs. Gene orthologs are homologous genes derived by vertical descent from a single ancestral gene in the last common ancestor of the compared species. Orthologs typically have the same or similar function in different species so this relationship can be used to assign putative function to previously unstudied gene sequences.

Gene paralogs are homologous genes which have evolved by duplication of an ancestral gene. Paralogs are likely to have evolved toward functional diversification so they are not generally used to infer same or similar function. However, they are a valuable tool that can be used to look for patterns in the composition of the families which might indicate either small or large-scale gene duplication. Exploring patterns of gene duplication is central to understanding genome evolution.

Using sequence comparison techniques, such as BLAST (Altschul, Gish et al. 1990) , families of homologues (both orthologs and paralogs) between any number of organisms can be derived. When this family contains exactly one member from each species, it can be considered to contain only orthologs and same or similar function is often inferred.

For families that contain paralogs, patterns in the copy number from each genome can be used to explore possible genome duplications or deletions. Further, when functional annotations are available, patterns of functional bias related to the duplications can also be detected and used as a jumping off point for further study.

The number of families of homologues generated by comparing two or more genomes can be large (Sanzol 2010), (Proost, Van Bel et al. 2009), (Procter, Thompson et al. 2010). Because of this, visual inspection of the data for patterns of duplication is not feasible. HomIn provides a user-friendly, advanced query system to access specified subsets of homologue families. The subsets can be based on either genome composition, Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, Goto et al. 2010), (Kanehisa, Goto et al. 2006), (Kanehisa and Goto 2000) Orthology Database(KO) categories, Clusters of Orthologous Groups of proteins (COGs) (Tatusov, Koonin et al. 1997), (Tatusov, Fedorova et al. 2003), Genome Ontology (GO) categories (Ashburner, Ball et al. 2000) or any other annotation.

Related Work

Expressed sequence tags (ESTs) provide scientists with a valuable source of data for the large-scale characterization of duplicated genes in genomes that are not fully sequenced (Sanzol 2010). EST collections represent a core (underutilized?) resource for the understanding of genome functionality and an important jumping off point for the exploration of genome duplications and functional diversification. Despite this, many studies using ESTs to characterize genomes are based on custom scripts and databases (Sanzol 2010) so there are very few, if any, publicly available tools to explore EST collections.

There are several databases and programs aimed at generating and/or presenting paralogy and orthology relationships. Not all, however, are generally useful: Some are for specific genomes and address the needs of a particular research community, including PLAZA (Proost, Van Bel et al. 2009) for plants, GreenPhylDB (Conte, Gaillard et al. 2008) for *O. sativa* and *A. thaliana*, FlyPhy (Wu, Xu et al. 2009) for *Drosophila* and the bursal EST database project for Chicken (Abdrakhmanov, Lodygin et al. 2000). Others are designed only for fully sequenced genomes. These programs include PANTHER (Mi, Dong et al. 2010) and Synteny Database (Catchen, Conery et al. 2009).

There are also several programs available that identify protein families in order to target individual genes and identify function for gene expression studies and pharmaceutical target identification. Examples include Panning for Genes, (Retief, Lynch et al. 1999) and PANTHER version 7, (Mi, Dong et al. 2010) and GO-Diff (Chen, Ye et al. 2010), a Linux-based program that uses GO terms to search for functional differentiation between EST-based transcriptomes by comparing unique genes in the organism of interest to GO and Unigene families.

As far as we know this is the only effort to produce a tool that allows a user to explore a set of protein families generated by any method and on any platform.

Extended Definition of Problem and Approach

Introduction

We designed this program to allow users to easily explore sets of gene homologue families in an interactive graphical environment.   There are two types of data that can be imported as text formatted files: sets of gene homologues and annotations for these genes.

The first step in setting up the interactive environment is to create a database by importing a list of families of gene homologues. These families are all that is necessary to create a database that users can easily explore in the interactive graphical environment.

The second optional step in setting up the interactive environment is to import annotations for any or all of the genes in the families. These annotations can be used as a basis for further exploration of the families.   KEGG annotations can be directly imported into the program in order to add KEGG Ontology (KO), COG and GO functional annotations, however, any type of annotation can be used.

Once the database has been created and populated, users can begin to explore by making queries based on genome composition or annotation. Query results summaries are displayed as text and in chart format.   The user can also export a tab-delimited text file of any result set as well as link directly to on-line information for KEGG data associated with any gene.

Database building

This stand-alone application, which should run on any machine with the latest Java Virtual Machine (JVM), is written in Java 1.6 and uses db4o for Java, version 7.12. To build the database of gene families both a tab-delimited text file containing family ids followed by a list of member genes and a tab delimited text file listing any unique genes from the included genomes are imported. Including a the list of unique genes is important for building more realistic functional category distribution statistics. KEGG annotation data which includes KO, COG and GO database annotations are automatically built into the database so that when KEGG annotation files are imported, the necessary background information is available.

Accessing DB in an interactive environment

Once the database is built and KEGG annotations are imported, two basic types of queries are available: queries based on genome composition and queries based on annotation.

Genome composition queries allows users to query for families with particular numbers of members from each genome or with a particular gene. For example, a user could ask for families with exactly one gene from one genome and at least two genes from another.

Functional Category queries allow the user to search for families with any set of functional categories. The functional categories used as criteria can be KO first level classes, COG functional categories, GO SLIM categories or any other category the user has imported.

Query results are displayed in a tree format that summarizes the number of families and the number of genes with each annotation. Any KO, COG or GO annotation summary displays can be double-clicked to go directly to their primary resource webpage.

KO and COG statistics can also be displayed in chart format and GO SLIM statistics can be displayed in table format. This allows the biologist to see at a glance which functional categories might be over or under represented in a given set of families. The total number of genes per genome annotated with a given category is used as the expected representation for comparison.

All queries the biologist makes are automatically stored in the database so that the user can view them anytime the database is opened.

Genome Composition queries can also be run using an "any" designation which enables the user to effectively run more than one query at once. If the user wants to search for families with exactly one member from any genome and more than one from the others, the "any" tool allows the user to perform all the potential combinations as one search.

KO annotations

KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by BLAST comparisons against the manually curated KEGG GENES database. The result contains KO assignments and automatically generated KEGG pathways. The KO database consists of manually defined ortholog groups that correspond to KEGG pathway nodes and KEGG BRITE functional hierarchy nodes. Once genes are assigned the KO identifiers, or the K numbers, by KAAS, the organism-specific pathway maps and BRITE hierarchies can be automatically generated.

GO annotations

The GO describes how gene products behave in a cellular context. The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. It is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains.

GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms and are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene product function is required. Our database utilizes the generic GO slim provided by GO which, like the GO itself, is not species specific. When a KO annotation is imported, any associated GO annotations that are not already part of the generic GO SLIM are converted to their parent SLIM(s) terms before being stored in the database.

COG annotations

COG is a Phylogenetic classification of proteins encoded in complete genomes. COGs were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.   COGS are intended to be used to facilitate both functional annotation of genomes and large-scale evolutionary studies.

Appendix A: User Guide

Getting Started

Operating Systems

Tested on Mac, Linux and Windows 7. If you have trouble opening the executable, try going to www.java.com and downloading the latest version of Java.

Before Running

You should have a copy of the executable file as well as a folder named "data". The executable file may appear as "HomIn" or "HomIn.jar", depending on your Operating System. The data folder includes several files with data necessary for building your database along with the necessary programs to rebuild/update these files. The three included programs are:

1. OBO2db4o: an executable to convert the GO OBO format database into a db4o database and create a new KO to GO mapping file with the GO ids converted to Generic GO SLIMS.
2. getKOs.pl: a Perl script for rebuilding the KO data files.
3. getCOGs.pl: a Perl script to rebuild the COG data files.

The executable and the data file should stay in the same directory (folder).

Annotation Data Preparation

The data folder contains all the data needed to run the application and build a new database. You may, however, wish to update these data files with the latest release of the appropriate annotation database.

Updating KO information files

If you wish to build your database with the most recent KEGG Orthology release:

1. Download the latest release at ftp://ftp.genome.jp/pub/kegg/genes/
2. Run parseKOs.pl
    a. Before running the script, you may need to change the permissions for the file (chmod 744 getKOs).
    b. From a command line interface:
        i. navigate to the "data" folder
        ii. run the script with the command ./getKOs.pl <kodbfilename>

Five data files will be written by this script: kos.category, kos.cog, kos.definitions, kos.go and kos.names. These files are used whenever a new HomIn database is created.

Updating COG information files

Updating the COG information should not be necessary. However, if you wish to do so, follow the instructions in “Updating KO information files” step 2a to change permissions and run getCOGs.pl. Two data files will be generated: cogdefs and cogfuncats.

Creating a new, updated db4o GO database

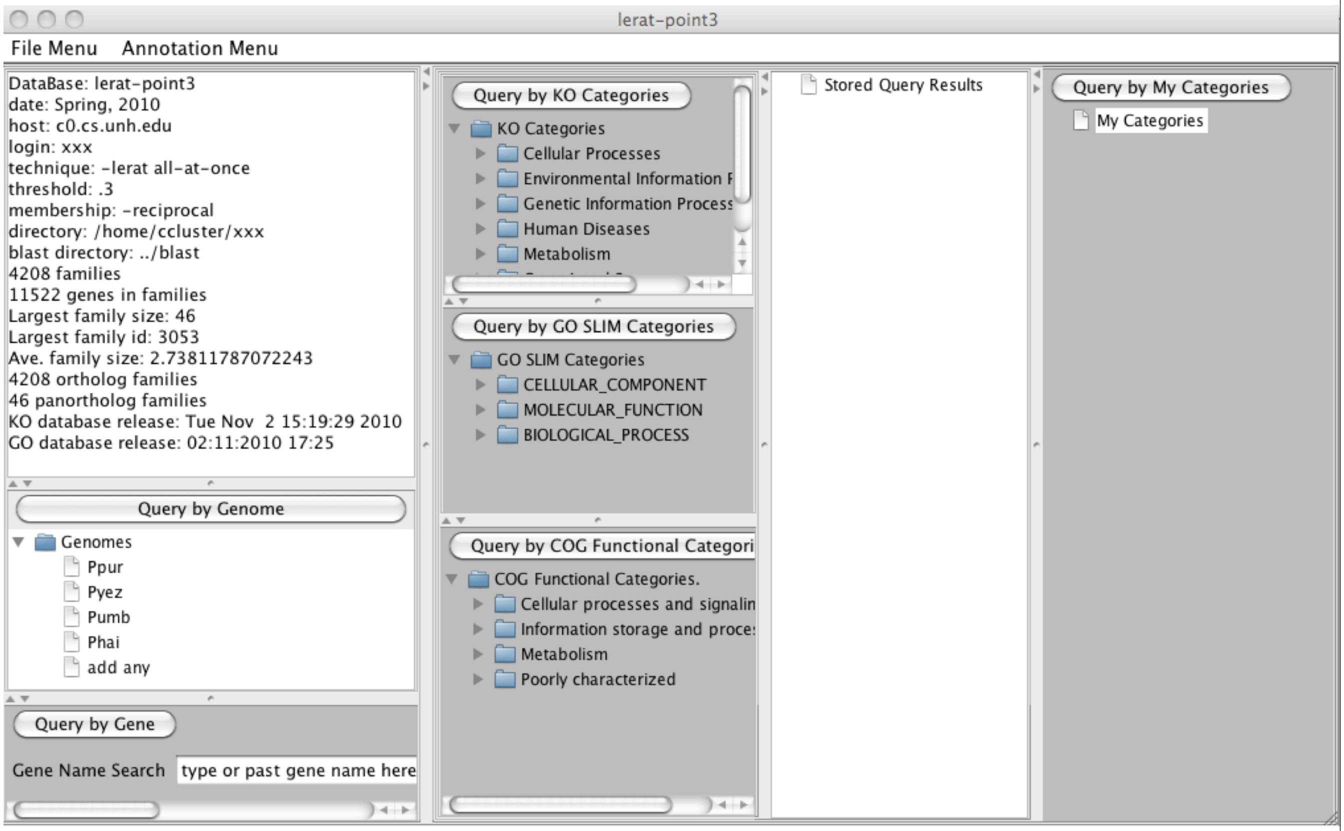
To download the latest version of the GO database and convert it to db4o, follow these steps:

1. Download the latest full ontology GO database in OBO format at http://www.geneontology.org/GO.downloads.ontology.shtml
2. Double click to open the OBO2db4o application.
3. Select New->Create db4o GO db
4. In the File Chooser menu, select the GO file you just downloaded
5. Select New-> Convert KO GOs to SLIMs
6. In the File Chooser menu, select db4oGO (the database you just created)
7. In the next File Chooser menu, select kos.go (included in the data file).

Loading The Sample Data

Create a new database

1. Double click on the executable file.
2. Select Lerat->New Lerat Data Base.
3. In the Choose directory dialog box, select the *folder* "lerat-point3" and click "Choose Directory".
4. If you are asked if you want to overwrite an existing database, select "Yes".
5. This may take up to a few minutes.
6. The Choose KO results file dialog box will appear when the database has been built. Navigate to and select the "KAAS results" file. Click "Choose".
7. In the "KO genomes" dialog box, select "All" from the drop-down menu and click "OK".
8. This may take several minutes. A new database exploration window will appear when your database is ready. It should look something like Figure 1 and contain the following panels, starting from the top, left corner and going clock-wise:
    a. MetaData
    b. Query by KO Categories
    c. Stored Query Results
    d. Query by My Categories
    e. Query by COG Functional Categories
    f. Query by Gene
    g. Query by Genome
    h. In the middle: Query by GO SLIM panel.



DataBase: lerat-point3
date: Spring, 2010
host: c0.cs.unh.edu
login: xxx
technique: -lerat all-at-once
threshold: .3
membership: -reciprocal
directory: /home/ccluster/xxx
blast directory: ../blast
4208 families
11522 genes in families
Largest family size: 46
Largest family id: 3053
Ave. family size: 2.73811787072243
4208 ortholog families
46 panortholog families
KO database release: Tue Nov 2 15:19:29 2010
GO database release: 02:11:2010 17:25

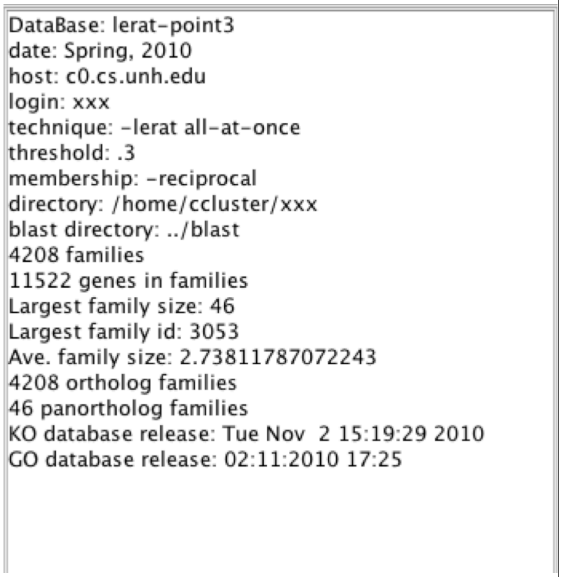
Gene Name Search: type or past gene name here

Figure 1: Exploration Window

Exploring the sample database

MetaData

The metadata panel should contain essentially the same text as Figure 1. The dates of the KO and GO database release may be more recent than those shown in the figure.

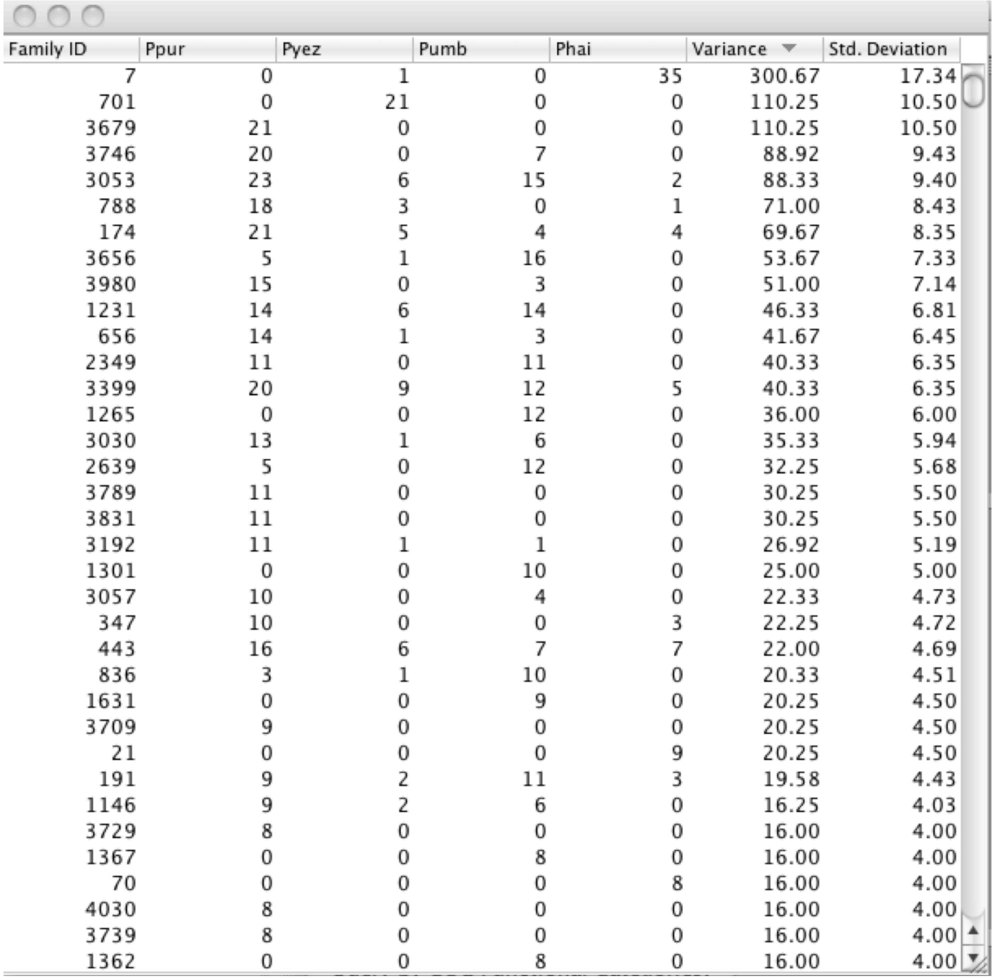


```
DataBase: lerat-point3
date: Spring, 2010
host: c0.cs.unh.edu
login: xxx
technique: -lerat all-at-once
threshold: .3
membership: -reciprocal
directory: /home/ccluster/xxx
blast directory: ../blast
4208 families
11522 genes in families
Largest family size: 46
Largest family id: 3053
Ave. family size: 2.73811787072243
4208 ortholog families
46 panortholog families
KO database release: Tue Nov  2 15:19:29 2010
GO database release: 02:11:2010 17:25
```

Figure 2: MetaData Frame

Family Composition Statistics

Select *File Menu->View Count Stats* to show a table summarizing the composition of all families in the imported data set (Figure 3). This table includes the Variance and Std Deviation based on the expectation of each family having the same number of genes from each genome. In Figure 3 data has been sorted by the Standard Deviation by clicking two times in the Label for the Standard Deviation column.



Family ID	Ppur	Pyez	Pumb	Phai	Variance	Std. Deviation
7	0	1	0	35	300.67	17.34
701	0	21	0	0	110.25	10.50
3679	21	0	0	0	110.25	10.50
3746	20	0	7	0	88.92	9.43
3053	23	6	15	2	88.33	9.40
788	18	3	0	1	71.00	8.43
174	21	5	4	4	69.67	8.35
3656	5	1	16	0	53.67	7.33
3980	15	0	3	0	51.00	7.14
1231	14	6	14	0	46.33	6.81
656	14	1	3	0	41.67	6.45
2349	11	0	11	0	40.33	6.35
3399	20	9	12	5	40.33	6.35
1265	0	0	12	0	36.00	6.00
3030	13	1	6	0	35.33	5.94
2639	5	0	12	0	32.25	5.68
3789	11	0	0	0	30.25	5.50
3831	11	0	0	0	30.25	5.50
3192	11	1	1	0	26.92	5.19
1301	0	0	10	0	25.00	5.00
3057	10	0	4	0	22.33	4.73
347	10	0	0	3	22.25	4.72
443	16	6	7	7	22.00	4.69
836	3	1	10	0	20.33	4.51
1631	0	0	9	0	20.25	4.50
3709	9	0	0	0	20.25	4.50
21	0	0	0	9	20.25	4.50
191	9	2	11	3	19.58	4.43
1146	9	2	6	0	16.25	4.03
3729	8	0	0	0	16.00	4.00
1367	0	0	8	0	16.00	4.00
70	0	0	0	8	16.00	4.00
4030	8	0	0	0	16.00	4.00
3739	8	0	0	0	16.00	4.00
1362	0	0	8	0	16.00	4.00

Figure 3: Stats summary table

Query By Genome

To query the families by genome composition select genomes in the Query by Genome Frame Figure 4. Use the following steps to try it:

1. Select all of the genomes in this database: Ppur, Pyez, Pumb and Phai
    a. Select all four by holding down the command (Mac)/Ctrl(Windows) key as you select each one.
    b. Ignore "add any" for now.
2. Click the *Query by Genome* button.
3. A new dialogue window will appear (FIGURE 5). This window allows you to choose the minimum and maximum number of genes from each genome per family that you would like to search for. For now, leave them all on the value 1 and click the search button. A dialogue box will appear that allows you to customize the name of your query. Change the name of this query to "panorthologs" and click *OK*.

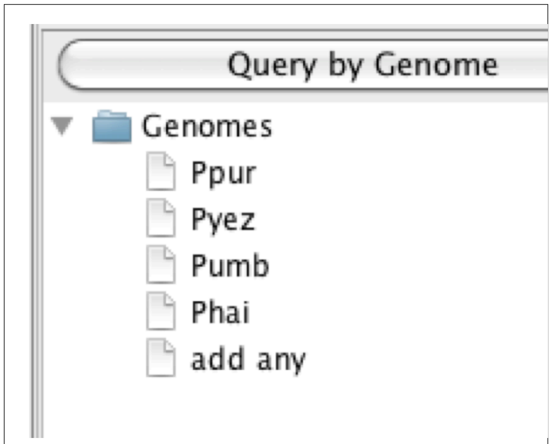
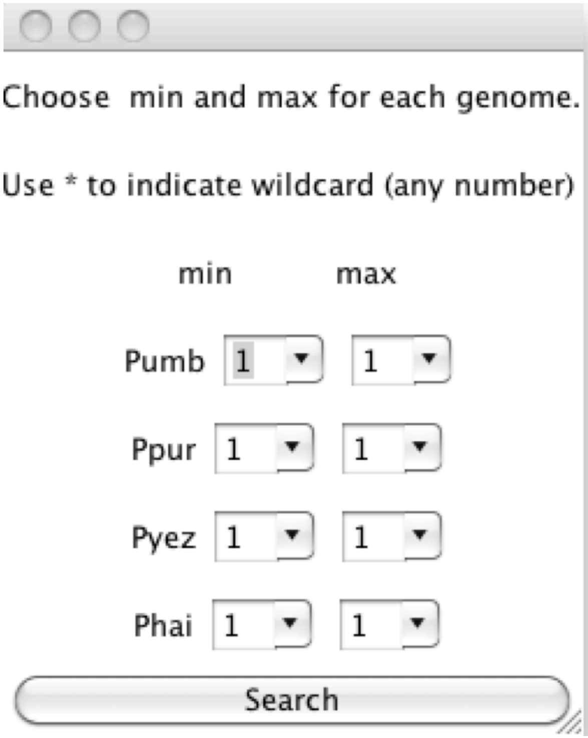


Figure 4: Query by Genome panel



Choose min and max for each genome.

Use * to indicate wildcard (any number)

	min	max
Pumb	1	1
Ppur	1	1
Pyez	1	1
Phai	1	1

FIGURE 5: Composition Dialogue

Exploring the Query Results

Your Query: You may need to expand the *Stored Query Results* folder in order to see your query result folder which should be labeled "*panorthologs*".   You should see that there are 46 families in this database that are panorthologs (Figure 6.). Expanding the *panorthologs* query folder reveals annotation statistics for these 46 families.

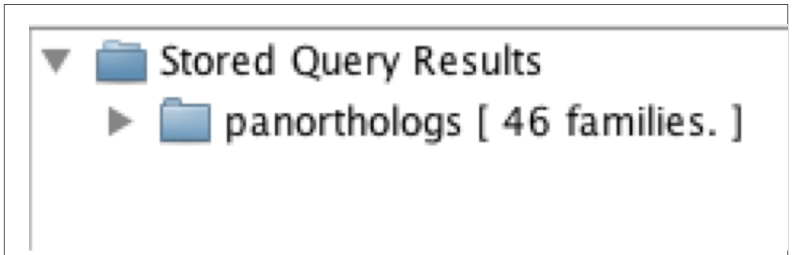
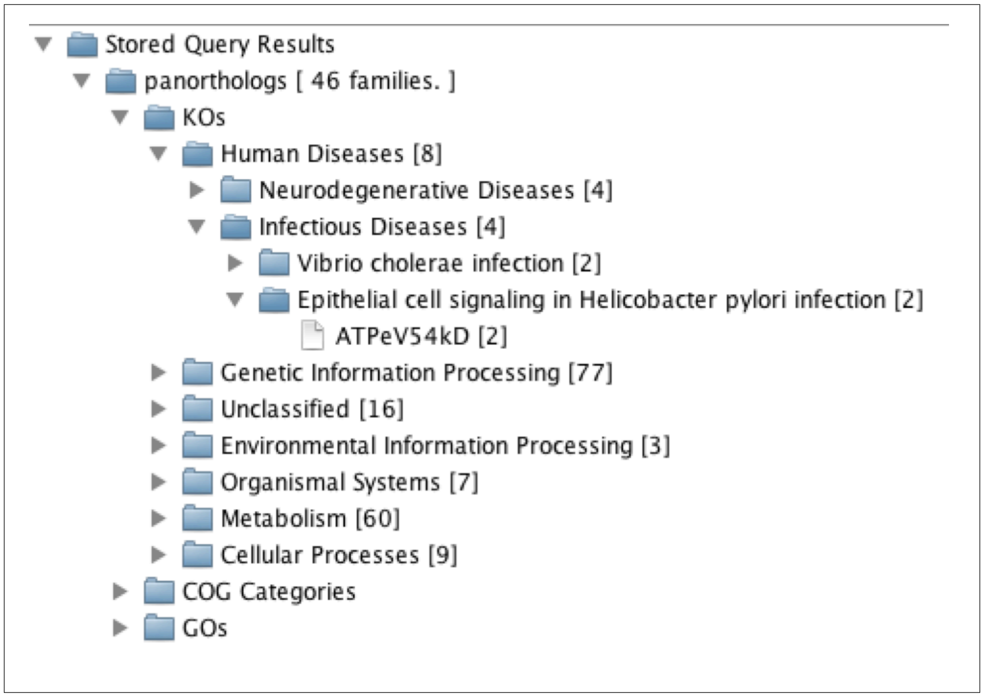


Figure 6: Query Results

KO results: The partially expanded *KOs* folder should look like Figure 8. Double click on the *KOs* folder to bring up a chart showing the distribution of KO First Level Categories in the *panorthologs* (Figure 7). Double clicking on any of the sub-folders of the *KOs* folder should take you directly to an applicable web page for that category.



- ▼ Stored Query Results
    - ▼ panorthologs [46 families.]
        - ▼ KOs
            - ▼ Human Diseases [8]
                - ▶ Neurodegenerative Diseases [4]
                - ▼ Infectious Diseases [4]
                    - ▶ Vibrio cholerae infection [2]
                    - ▼ Epithelial cell signaling in Helicobacter pylori infection [2]
                        - ATPeV54kD [2]
            - ▶ Genetic Information Processing [77]
            - ▶ Unclassified [16]
            - ▶ Environmental Information Processing [3]
            - ▶ Organismal Systems [7]
            - ▶ Metabolism [60]
            - ▶ Cellular Processes [9]
        - ▶ COG Categories
        - ▶ GOs

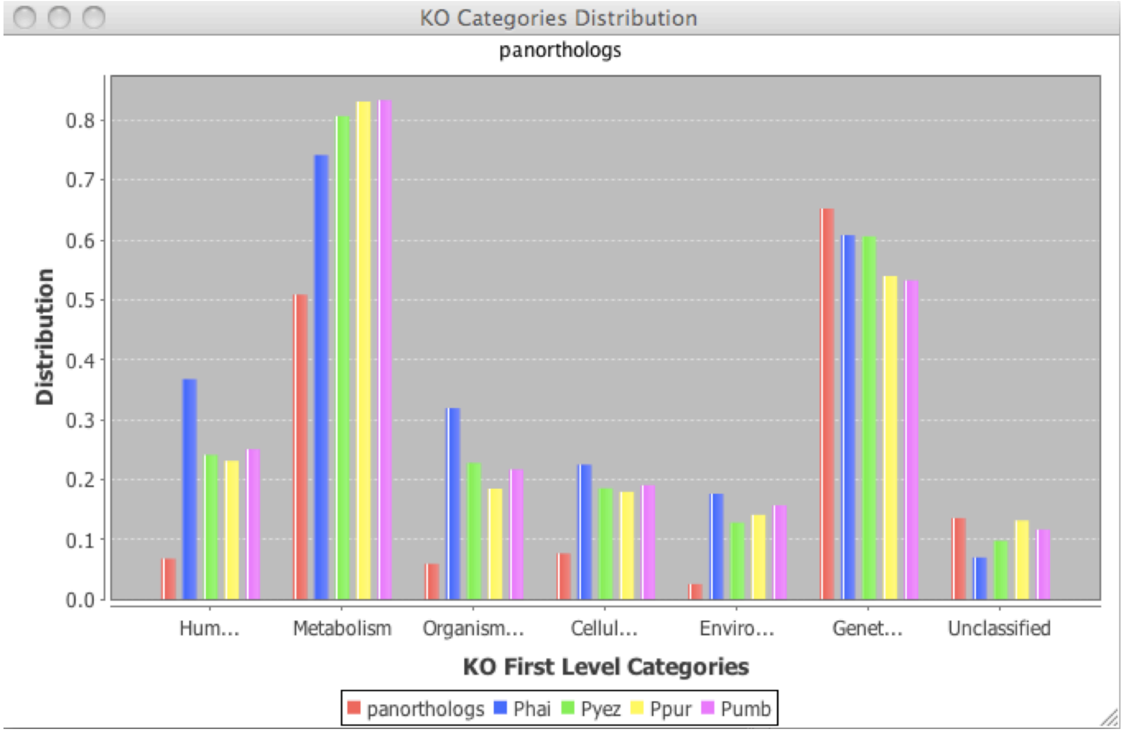
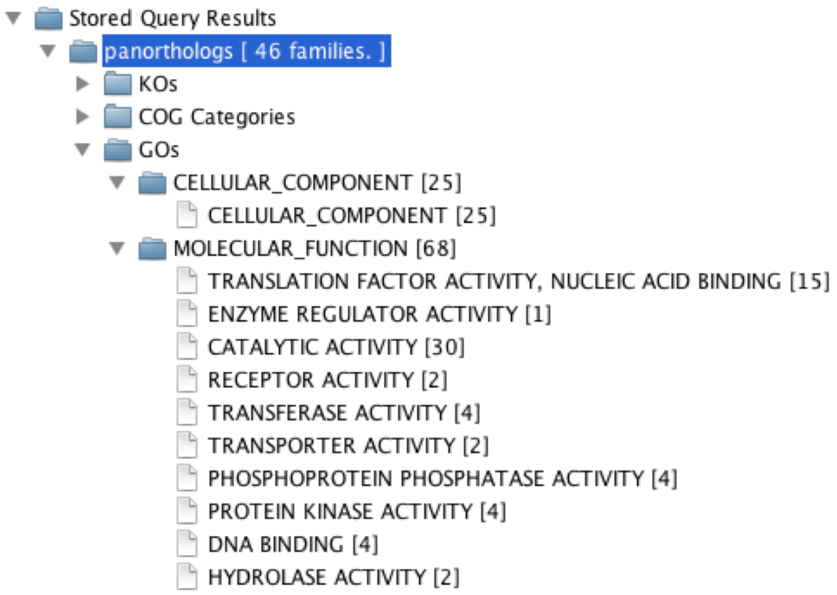


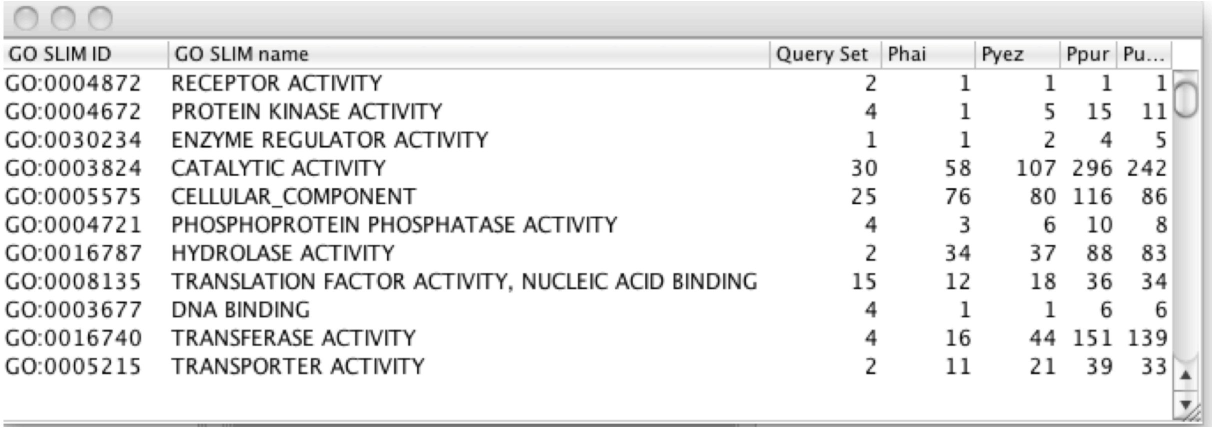
Figure 7: KO distribution chart

GO results Expanding the *GOs* folder should look like Figure 10 and double-clicking on the *GOs* folder will create a new window with a table like Figure 9 (with some resizing of the window and columns). Again, try double clicking on any of the sub-folders to go to the applicable web page for that category.



- Stored Query Results
  - panorthologs [46 families.]
    - KOs
    - COG Categories
    - GOs
      - CELLULAR_COMPONENT [25]
        - CELLULAR_COMPONENT [25]
      - MOLECULAR_FUNCTION [68]
        - TRANSLATION FACTOR ACTIVITY, NUCLEIC ACID BINDING [15]
        - ENZYME REGULATOR ACTIVITY [1]
        - CATALYTIC ACTIVITY [30]
        - RECEPTOR ACTIVITY [2]
        - TRANSFERASE ACTIVITY [4]
        - TRANSPORTER ACTIVITY [2]
        - PHOSPHOPROTEIN PHOSPHATASE ACTIVITY [4]
        - PROTEIN KINASE ACTIVITY [4]
        - DNA BINDING [4]
        - HYDROLASE ACTIVITY [2]

Figure 10: GO folder expanded



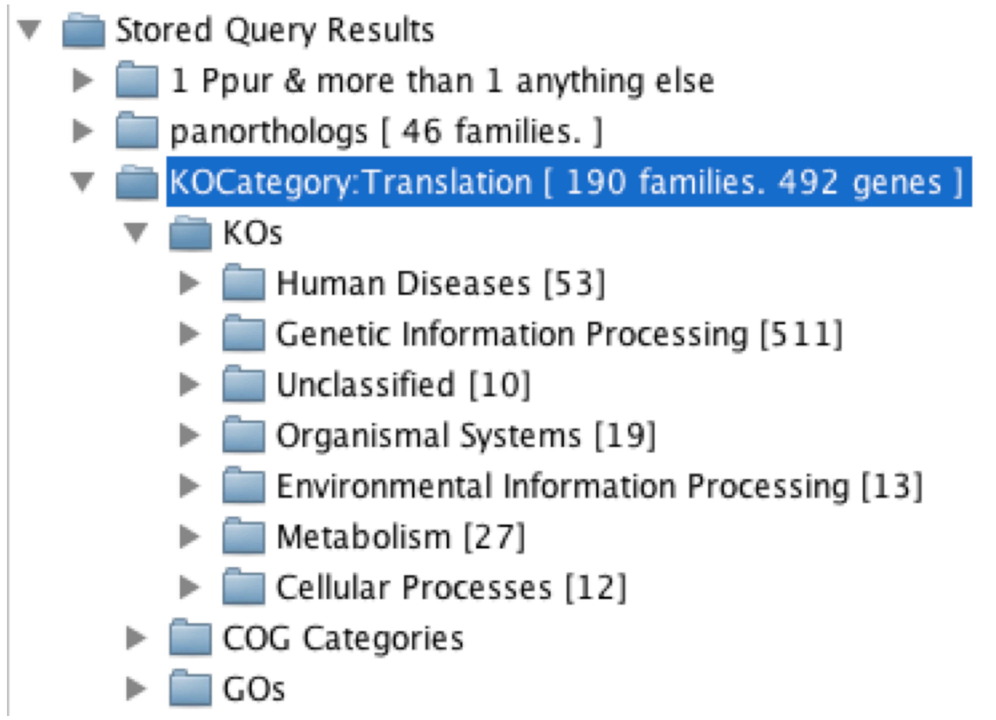
GO SLIM ID	GO SLIM name	Query Set	Phai	Pyez	Ppur	Pu...
GO:0004872	RECEPTOR ACTIVITY	2	1	1	1	1
GO:0004672	PROTEIN KINASE ACTIVITY	4	1	5	15	11
GO:0030234	ENZYME REGULATOR ACTIVITY	1	1	2	4	5
GO:0003824	CATALYTIC ACTIVITY	30	58	107	296	242
GO:0005575	CELLULAR_COMPONENT	25	76	80	116	86
GO:0004721	PHOSPHOPROTEIN PHOSPHATASE ACTIVITY	4	3	6	10	8
GO:0016787	HYDROLASE ACTIVITY	2	34	37	88	83
GO:0008135	TRANSLATION FACTOR ACTIVITY, NUCLEIC ACID BINDING	15	12	18	36	34
GO:0003677	DNA BINDING	4	1	1	6	6
GO:0016740	TRANSFERASE ACTIVITY	4	16	44	151	139
GO:0005215	TRANSPORTER ACTIVITY	2	11	21	39	33

Figure 9: GO Slim Table

Query by Annotation

To Query the sample database by an annotation:

1. In the KO Categories panel, expand the *KO Categories* folder (if it isn't already)
2. Expand the Genetic Information and Processing folder
3. Select Translation
4. Click the *Query by KO Categories* button
5. A dialogue box will appear which allows you to name the query. Name it and click "OK"
6. The Stored Query Results Panel should now look something like Figure 11, depending on what you name your query and how many of the folders you expand.



Stored Query Results
1 Ppur & more than 1 anything else
panorthologs [46 families.]
KOCategory:Translation [190 families. 492 genes]
KOs
Human Diseases [53]
Genetic Information Processing [511]
Unclassified [10]
Organismal Systems [19]
Environmental Information Processing [13]
Metabolism [27]
Cellular Processes [12]
COG Categories
GOs

Figure 11: Query by Annotation

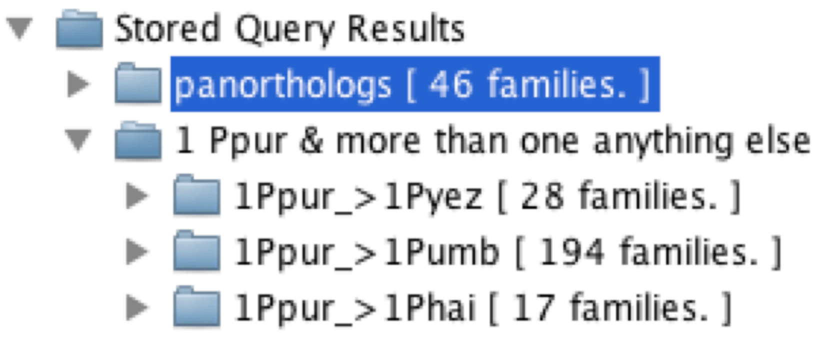
Using the "Any" Genome Category

The advantage to using the "any" genome option is that you can run more than one query at once. To test the "any" designation:

1. Select *Ppur* and *add any* in the Query by Genome panel.
2. Click the *Query by Genome* button.
3. A new dialogue window will appear.
    a. Leave both the minumum and maximum count for Ppur as 1.
    b. For AnyA, change the minimum count to 2 and the maximum count to the wildcard (*). You can use the drop down menu or type the 2 and the * in.
4. Click the *search* button.
5. While the queries are being run, you will get a series of four dialogue boxes.
    a. The first will give you the option to re-name the master query.
    b. The following three let you re-name the three sub-queries.

The results of this query should look something like Figure 12 depending on how you chose to name the queries. In this example, the master query is named "1 Ppur & more than one anything else" . The names of each sub-query were left on the default settings.

The first sub-query, "1Ppur_>1Pyez", contains all the families in the result set that contain exactly one gene from Ppur and more than one gene from Pyez. The other two queries are families that contain one gene from Ppur and more than one from the other two genomes.



▼ Stored Query Results
  ▶ panorthologs [46 families.]
  ▼ 1 Ppur & more than one anything else
    ▶ 1Ppur_>1Pyez [28 families.]
    ▶ 1Ppur_>1Pumb [194 families.]
    ▶ 1Ppur_>1Phai [17 families.]

Figure 12: Results of a Query using "any"

Using Your Own Annotations

Any annotations can be imported from a tab-delimited text file into your database. An example file, colors, is included in the test files folder. This file contains the name of a gene in the first column and an annotation in the second. To import these annotations into the database:

1. Select *Annotation Menu->Add My Category Data*.
2. In the open dialogue, navigate to the *colors* file, select it and click the *Open* button.
3. A new dialogue will appear asking you to name this category. Give it a name and click *OK.*

The Query by My Categories panel should look like Figure 13 when the folders are both fully expanded. These categories are now available for querying the database.

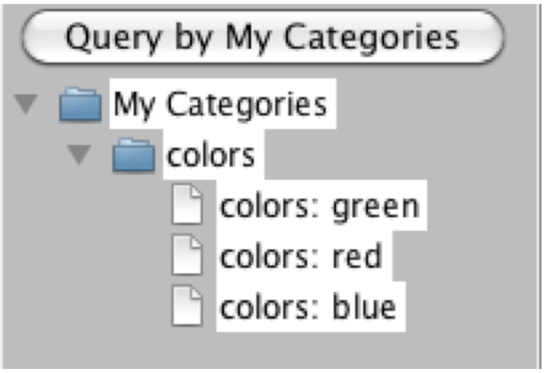


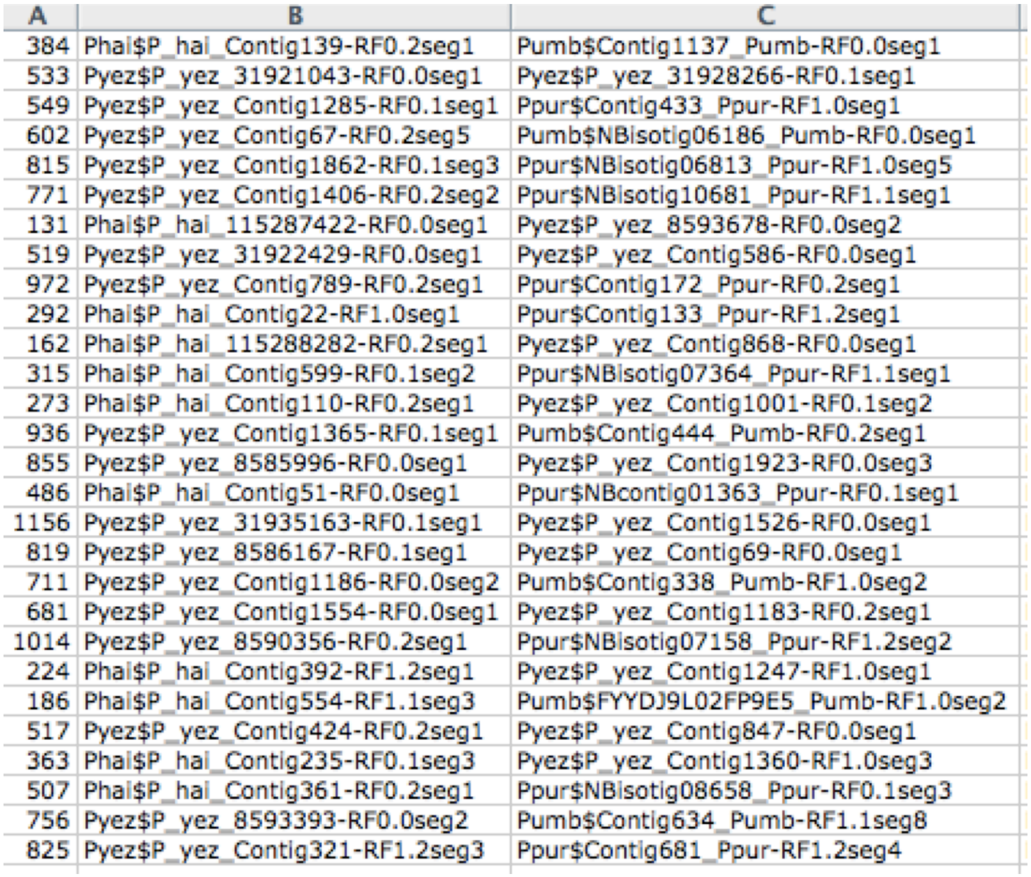
Figure 13: Your Own Annotations

Exporting Queries

All of the families in a saved query can be exported to a text file by following these steps:

1. Select the query 1Ppur_>1Pyez .
2. Go to File Menu->Export Selected Query Families.

Figure 14 shows what the first three columns of this file look like if it's open in a spreadsheet program. The first column contains the IDs of the 28 families that have exactly one gene from Ppur and more than one gene from Pyez. The following columns contain the list of genes in each family.



A	B	C
384	Phai$P_hai_Contig139-RF0.2seg1	Pumb$Contig1137_Pumb-RF0.0seg1
533	Pyez$P_yez_31921043-RF0.0seg1	Pyez$P_yez_31928266-RF0.1seg1
549	Pyez$P_yez_Contig1285-RF0.1seg1	Ppur$Contig433_Ppur-RF1.0seg1
602	Pyez$P_yez_Contig67-RF0.2seg5	Pumb$NBisotig06186_Pumb-RF0.0seg1
815	Pyez$P_yez_Contig1862-RF0.1seg3	Ppur$NBisotig06813_Ppur-RF1.0seg5
771	Pyez$P_yez_Contig1406-RF0.2seg2	Ppur$NBisotig10681_Ppur-RF1.1seg1
131	Phai$P_hai_115287422-RF0.0seg1	Pyez$P_yez_8593678-RF0.0seg2
519	Pyez$P_yez_31922429-RF0.0seg1	Pyez$P_yez_Contig586-RF0.0seg1
972	Pyez$P_yez_Contig789-RF0.2seg1	Ppur$Contig172_Ppur-RF0.2seg1
292	Phai$P_hai_Contig22-RF1.0seg1	Ppur$Contig133_Ppur-RF1.2seg1
162	Phai$P_hai_115288282-RF0.2seg1	Pyez$P_yez_Contig868-RF0.0seg1
315	Phai$P_hai_Contig599-RF0.1seg2	Ppur$NBisotig07364_Ppur-RF1.1seg1
273	Phai$P_hai_Contig110-RF0.2seg1	Pyez$P_yez_Contig1001-RF0.1seg2
936	Pyez$P_yez_Contig1365-RF0.1seg1	Pumb$Contig444_Pumb-RF0.2seg1
855	Pyez$P_yez_8585996-RF0.0seg1	Pyez$P_yez_Contig1923-RF0.0seg3
486	Phai$P_hai_Contig51-RF0.0seg1	Ppur$NBcontig01363_Ppur-RF0.1seg1
1156	Pyez$P_yez_31935163-RF0.1seg1	Pyez$P_yez_Contig1526-RF0.0seg1
819	Pyez$P_yez_8586167-RF0.1seg1	Pyez$P_yez_Contig69-RF0.0seg1
711	Pyez$P_yez_Contig1186-RF0.0seg2	Pumb$Contig338_Pumb-RF1.0seg2
681	Pyez$P_yez_Contig1554-RF0.0seg1	Pyez$P_yez_Contig1183-RF0.2seg1
1014	Pyez$P_yez_8590356-RF0.2seg1	Ppur$NBisotig07158_Ppur-RF1.2seg2
224	Phai$P_hai_Contig392-RF1.2seg1	Pyez$P_yez_Contig1247-RF1.0seg1
186	Phai$P_hai_Contig554-RF1.1seg3	Pumb$FYYDJ9L02FP9E5_Pumb-RF1.0seg2
517	Pyez$P_yez_Contig424-RF0.2seg1	Pyez$P_yez_Contig847-RF0.0seg1
363	Phai$P_hai_Contig235-RF0.1seg3	Pyez$P_yez_Contig1360-RF1.0seg3
507	Phai$P_hai_Contig361-RF0.2seg1	Ppur$NBisotig08658_Ppur-RF0.1seg3
756	Pyez$P_yez_8593393-RF0.0seg2	Pumb$Contig634_Pumb-RF1.1seg8
825	Pyez$P_yez_Contig321-RF1.2seg3	Ppur$Contig681_Ppur-RF1.2seg4

Figure 14: Exported Query Results

Using Your Own Data

Data Preparation

Results of the pipeline described by Cooper, et al can be used as is. To build the database, leave these original results files in the directory created by the pipeline. The directory should have the same name as the prefix of the result files. The names of the result files should also be left as is: The file containing the families must be named *prefix.family*. The file containing the unique gene must be names *prefix.unique*. Any statistics, in a file names *prefix.stats*.

If you are using files generated with another method, the family file is tab-delimited text with each family listed on one line. The first column is a unique ID for the family optionally followed immediately by a colon (no space or tab). The following columns are the genes in the family. The name for each gene should be prefixed by the name of the genome it is from and '$'. The gene names in the .unique file should follow the same format. Please see the included *.stats* file for an example of how to format that file.

Preparing Annotation Data

KEGG annotations data, which includes KO, COG and GO database annotations, can be read directly from the text formatted result file returned by KEGG Automatic Annotation Server (KAAS). Any GO annotation associated with a KO will be automatically converted to its parent generic GO SLIM term(s).

Any other annotations the user wishes to include in the analysis can be imported from a tab delimited text file. This file should be given a name that describes this annotation for example "color". The first column in each entry in this file contains gene names and the second contains a string assigning the annotation such as "blue".

Getting Started

Once your database has been created, you can immediately search for families based on *Genome Composition* using *Query by Genome*. Other search criteria options are available once annotation data have been added to the database.

Reopening the Database You've Created

All Query Results are stored permanently in the database unless deleted by the User (using File Menu -> Delete Selected Queries). To reopen your database:

1. Select Lerat-> Open Lerat Data Base
2. in the open dialogue, navigate to the folder containing your original result files. Your database will have been saved in this folder by default.

Abdrakhmanov, I., D. Lodygin, et al. (2000). "A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function." Genome research 10(12): 2062-2069.

> Chicken B cells create their immunoglobulin repertoire within the Bursa of Fabricius by gene conversion. The high homologous recombination activity is shared by the bursal B-cell-derived DT40 cell line, which integrates transfected DNA constructs at high rates into its endogenous loci. Targeted integration in DT40 is used frequently to analyze the function of genes by gene disruption. In this paper, we describe a large database of >7000 expressed sequence tags (ESTs) from bursal lymphocytes that should be a valuable resource for the identification of gene disruption targets in DT40. ESTs of interest can be recognized easily by online or keyword searches. Because the database reflects the gene expression profile of bursal lymphocytes, it provides valuable hints as to which genes might be involved in B-cell-specific processes related to immunoglobulin repertoire formation, signal transduction, transcription, and apoptosis. This large collection of chicken ESTs will also be useful for gene expression studies and comparative gene mapping within the chicken genome project. Details of the bursal EST sequencing project and access to database search forms can be found on the DT40 web site (http://genetics.hpi.uni-hamburg.de/dt40.html).

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology 215(3): 403-410.

> A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature genetics 25(1): 25-29.

Catchen, J. M., J. S. Conery, et al. (2009). "Automated identification of conserved synteny after whole-genome duplication." Genome research 19(8): 1497-1505.

An important objective for inferring the evolutionary history of gene families is the determination of orthologies and paralogies. Lineage-specific paralog loss following whole-genome duplication events can cause anciently related homologs to appear in some assays as orthologs. Conserved synteny-the tendency of neighboring genes to retain their relative positions and orders on chromosomes over evolutionary time-can help resolve such errors. Several previous studies examined genome-wide syntenic conservation to infer the contents of ancestral chromosomes and provided insights into the architecture of ancestral genomes, but did not provide methods or tools applicable to the study of the evolution of individual gene families. We developed an automated system to identify conserved syntenic regions in a primary genome using as outgroup a genome that diverged from the investigated lineage before a whole-genome duplication event. The product of this automated analysis, the Synteny Database, allows a user to examine fully or partially assembled genomes. The Synteny Database is optimized for the investigation of individual gene families in multiple lineages and can detect chromosomal inversions and translocations as well as ohnologs (paralogs derived by whole-genome duplication) gone missing. To demonstrate the utility of the system, we present a case study of gene family evolution, investigating the ARNTL gene family in the genomes of Ciona intestinalis, amphioxus, zebrafish, and human.

Chen, Z., H. Ye, et al. (2010). "A gene family-based method for interspecies comparisons of sequencing-based transcriptomes and its use in environmental adaptation analysis." Journal of genetics and genomics = Yi chuan xue bao 37(3): 205-218.

> We describe a new method for sequencing-based cross-species transcriptome comparisons and define a new metric for evaluating gene expression across species using protein-coding families as units of comparison. Using this measure transcriptomes from different species were evaluated by mapping them to gene families and integrating the mapping results with expression data. Statistical tests were applied to the transcriptome evaluation results to identify differentially expressed families. A Perl program named Pro-Diff was compiled to implement this method. To evaluate the method and provide an example of its use, two liver EST transcriptomes from two closely related fish that live in different temperature zones were compared. One EST library was from a recent sequencing project of Dissosticus mawsoni, a fish that lives in cold Antarctic sea waters, while the other was newly sequenced data (available at: http://www.fishgenome.org/polarbank/) from Notothenia angustata, a species that lives in temperate near-shore water of southern New Zealand. Results from the comparison were consistent with results inferred from phenotype differences and also with our previously published Gene Ontology-based method. The Pro-Diff program and operation manual can be downloaded from: http://www.fishgenome.org/download/Prodiff.rar.

Conte, M. G., S. Gaillard, et al. (2008). "GreenPhylDB: a database for plant comparative genomics." Nucleic acids research 36(Database issue): D991-998.

> GreenPhylDB (http://greenphyl.cirad.fr) is a comprehensive platform designed to facilitate comparative functional genomics in Oryza sativa and Arabidopsis thaliana genomes. The main functions of GreenPhylDB are to assign O. sativa and A. thaliana sequences to gene families using a semi-automatic clustering procedure and to create 'orthologous' groups using a phylogenomic approach. To date, GreenPhylDB comprises the most complete list of plant gene families, which have been manually curated (6421 families). GreenPhylDB also contains all of the phylogenomic relationships computed for 4375 families. A total of 492 TAIR, 1903 InterPro and 981 KEGG families and subfamilies were manually curated using the clusters created with the TribeMCL software. GreenPhylDB integrates information from several other databases including UniProt, KEGG, InterPro, TAIR and TIGR. Several entry points can be used to display phylogenomic relationships for A. thaliana or O. sativa sequences, using TAIR, TIGR gene ID, family name, InterPro, gene alias, UniProt or protein/nucleic sequence. Finally, a powerful phylogenomics tool, GreenPhyl Ortholog Search Tool (GOST), was incorporated into GreenPhylDB to predict orthologous relationships between O. sativa/A. thaliana protein(s) and sequences from other plant species.

Cooper, V. S., S. H. Vohr, et al. (2010). "Why genes evolve faster on secondary chromosomes in bacteria." PLoS computational biology 6(4): e1000732.

In bacterial genomes composed of more than one chromosome, one replicon is typically larger, harbors more essential genes than the others, and is considered primary. The greater variability of secondary chromosomes among related taxa has led to the theory that they serve as an accessory genome for specific niches or conditions. By this rationale, purifying selection should be weaker on genes on secondary chromosomes because of their reduced necessity or usage. To test this hypothesis we selected bacterial genomes composed of multiple chromosomes from two genera, Burkholderia and Vibrio, and quantified the evolutionary rates (dN and dS) of all orthologs within each genus. Both evolutionary rate parameters were faster among orthologs found on secondary chromosomes than those on the primary chromosome. Further, in every bacterial genome with multiple chromosomes that we studied, genes on secondary chromosomes exhibited significantly weaker codon usage bias than those on primary chromosomes. Faster evolution and reduced codon bias could in turn result from global effects of chromosome position, as genes on secondary chromosomes experience reduced dosage and expression due to their delayed replication, or selection on specific gene attributes. These alternatives were evaluated using orthologs common to genomes with multiple chromosomes and genomes with single chromosomes. Analysis of these ortholog sets suggested that inherently fast-evolving genes tend to be sorted to secondary chromosomes when they arise; however, prolonged evolution on a secondary chromosome further accelerated substitution rates. In summary, secondary chromosomes in bacteria are evolutionary test beds where genes are weakly preserved and evolve more rapidly, likely because they are used less frequently.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research 28(1): 27-30.

> KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The genomic information is stored in the GENES database, which is a collection of gene catalogs for all the completely sequenced genomes and some partial genomes with up-to-date annotation of gene functions. The higher order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. The PATHWAY database is supplemented by a set of ortholog group tables for the information about conserved subpathways (pathway motifs), which are often encoded by positionally coupled genes on the chromosome and which are especially useful in predicting gene functions. A third database in KEGG is LIGAND for the information about chemical compounds, enzyme molecules and enzymatic reactions. KEGG provides Java graphics tools for browsing genome maps, comparing two genome maps and manipulating expression maps, as well as computational tools for sequence comparison, graph comparison and path computation. The KEGG databases are daily updated and made freely available (http://www. genome.ad.jp/kegg/).

Kanehisa, M., S. Goto, et al. (2010). "KEGG for representation and analysis of molecular networks involving diseases and drugs." Nucleic acids research 38(Database issue): D355-360.

> Most human diseases are complex multi-factorial diseases resulting from the combination of various genetic and environmental factors. In the KEGG database resource (http://www.genome.jp/kegg/), diseases are viewed as perturbed states of the molecular system, and drugs as perturbants to the molecular system. Disease information is computerized in two forms: pathway maps and gene/molecule lists. The KEGG PATHWAY database contains pathway maps for the molecular systems in both normal and perturbed states. In the KEGG DISEASE database, each disease is represented by a list of known disease genes, any known environmental factors at the molecular level, diagnostic markers and therapeutic drugs, which may reflect the underlying molecular system. The KEGG DRUG database contains chemical structures and/or chemical components of all drugs in Japan, including crude drugs and TCM (Traditional Chinese Medicine) formulas, and drugs in the USA and Europe. This database also captures knowledge about two types of molecular networks: the interaction network with target molecules, metabolizing enzymes, other drugs, etc. and the chemical structure transformation network in the history of drug development. The new disease/drug information resource named KEGG MEDICUS can be used as a reference knowledge base for computational analysis of molecular networks, especially, by integrating large-scale experimental datasets.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic acids research 34(Database issue): D354-357.

> The increasing amount of genomic and molecular information is the basis for understanding higher-order biological systems, such as the cell and the organism, and their interactions with the environment, as well as for medical, industrial and other practical applications. The KEGG resource (http://www.genome.jp/kegg/) provides a reference knowledge base for linking genomes to biological systems, categorized as building blocks in the genomic space (KEGG GENES) and the chemical space (KEGG LIGAND), and wiring diagrams of interaction networks and reaction networks (KEGG PATHWAY). A fourth component, KEGG BRITE, has been formally added to the KEGG suite of databases. This reflects our attempt to computerize functional interpretations as part of the pathway reconstruction process based on the hierarchically structured knowledge about the genomic, chemical and network spaces. In accordance with the new chemical genomics initiatives, the scope of KEGG LIGAND has been significantly expanded to cover both endogenous and exogenous molecules. Specifically, RPAIR contains curated chemical structure transformation patterns extracted from known enzymatic reactions, which would enable analysis of genome-environment interactions, such as the prediction of new reactions and new enzyme genes that would degrade new environmental compounds. Additionally, drug information is now stored separately and linked to new KEGG DRUG structure maps.

Mi, H., Q. Dong, et al. (2010). "PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium." Nucleic acids research 38(Database issue): D204-210.

Protein Analysis THrough Evolutionary Relationships (PANTHER) is a comprehensive software system for inferring the functions of genes based on their evolutionary relationships. Phylogenetic trees of gene families form the basis for PANTHER and these trees are annotated with ontology terms describing the evolution of gene function from ancestral to modern day genes. One of the main applications of PANTHER is in accurate prediction of the functions of uncharacterized genes, based on their evolutionary relationships to genes with functions known from experiment. The PANTHER website, freely available at http://www.pantherdb.org, also includes software tools for analyzing genomic data relative to known and inferred gene functions. Since 2007, there have been several new developments to PANTHER: (i) improved phylogenetic trees, explicitly representing speciation and gene duplication events, (ii) identification of gene orthologs, including least diverged orthologs (best one-to-one pairs), (iii) coverage of more genomes (48 genomes, up to 87% of genes in each genome; see http://www.pantherdb.org/panther/summaryStats.jsp), (iv) improved support for alternative database identifiers for genes, proteins and microarray probes and (v) adoption of the SBGN standard for display of biological pathways. In addition, PANTHER trees are being annotated with gene function as part of the Gene Ontology Reference Genome project, resulting in an increasing number of curated functional annotations.

Procter, J. B., J. Thompson, et al. (2010). "Visualization of multiple alignments, phylogenies and gene family evolution." Nature methods 7(3 Suppl): S16-25.

> Software for visualizing sequence alignments and trees are essential tools for life scientists. In this review, we describe the major features and capabilities of a selection of stand-alone and web-based applications useful when investigating the function and evolution of a gene family. These range from simple viewers, to systems that provide sophisticated editing and analysis functions. We conclude with a discussion of the challenges that these tools now face due to the flood of next generation sequence data and the increasingly complex network of bioinformatics information sources.

Proost, S., M. Van Bel, et al. (2009). "PLAZA: a comparative genomics resource to study gene and genome evolution in plants." The Plant Cell 21(12): 3718-3731.

The number of sequenced genomes of representatives within the green lineage is rapidly increasing. Consequently, comparative sequence analysis has significantly altered our view on the complexity of genome organization, gene function, and regulatory pathways. To explore all this genome information, a centralized infrastructure is required where all data generated by different sequencing initiatives is integrated and combined with advanced methods for data mining. Here, we describe PLAZA, an online platform for plant comparative genomics (http://bioinformatics.psb.ugent.be/plaza/). This resource integrates structural and functional annotation of published plant genomes together with a large set of interactive tools to study gene function and gene and genome evolution. Precomputed data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, intraspecies whole-genome dot plots, and genomic colinearity between species. Through the integration of high confidence Gene Ontology annotations and tree-based orthology between related species, thousands of genes lacking any functional description are functionally annotated. Advanced query systems, as well as multiple interactive visualization tools, are available through a user-friendly and intuitive Web interface. In addition, detailed documentation and tutorials introduce the different tools, while the workbench provides an efficient means to analyze user-defined gene sets through PLAZA's interface. In conclusion, PLAZA provides a comprehensible and up-to-date research environment to aid researchers in the exploration of genome information within the green plant lineage.

Retief, J. D., K. R. Lynch, et al. (1999). "Panning for genes--A visual strategy for identifying novel gene orthologs and paralogs." Genome research 9(4): 373-382.

We have developed a rapid visual method for identifying novel members of gene families. Starting with an evolutionary tree, 20-50 protein query sequences for a gene family are selected from different branches of the tree. These query sequences are used to search the GenBank and expressed sequence tag (EST) DNA databases and their nightly updates using the tfastx3 or tfasty3 programs. The results of all 20-50 searches are collated and resorted to highlight EST or genomic sequences that share significant similarity with the query sequences. The statistical significance of each DNA/protein alignment is plotted, highlighting the portion of the query sequence that is present in the database sequence and the percent identity in the aligned region. The collated results for database sequences are linked using the WWW to the underlying scores and alignments; these links can also be used to perform additional searches to characterize the novel sequence further. With traditional "deep" scoring matrices (BLOSUM50) one can search for previously unrecognized families of large protein superfamilies. Alternatively, by using query sequences and EST libraries from the same species (e. g., human or mouse) together with "shallow" scoring matrices and filters that remove high-identity sequences, one can highlight new paralogs of previously described subfamilies. Using query sequences from the glutathione transferase superfamily, we identified two novel mammalian glutathione transferase families that were recognized previously only in plants. Using query sequences from known mammalian glutathione transferase subfamilies, we identified new candidate paralogs from the mouse class-mu, class-pi, and class-theta families.

Sanzol, J. (2010). "Dating and functional characterization of duplicated genes in the apple (Malus domestica Borkh.) by analyzing EST data." BMC plant biology 10(1): 87.

ABSTRACT: BACKGROUND: Gene duplication is central to genome evolution. In plants, genes can be duplicated through small-scale events and large-scale duplications often involving polyploidy. The apple belongs to the subtribe Pyrinae (Rosaceae), a diverse lineage that originated via allopolyploidization. Both small-scale duplications and polyploidy may have been important mechanisms shaping the genome of this species. RESULTS: This study evaluates the gene duplication and polyploidy history of the apple by characterizing duplicated genes in this species using EST data. Overall, 68% of the apple genes were clustered into families with a mean copy-number of 4.6. Analysis of the age distribution of gene duplications supported a continuous mode of small-scale duplications, plus two episodes of large-scale duplicates of vastly different ages. The youngest was consistent with the polyploid origin of the Pyrinae 37-48 MYBP, whereas the older may be related to gamma-triplication; an ancient hexapolyploidization previously characterized in the four sequenced eurosid genomes and basal to the eurosid-asterid divergence. Duplicated genes were studied for functional diversification with an emphasis on young paralogs; those originated during or after the formation of the Pyrinae lineage. Unequal assignment of single-copy genes and gene families to Gene Ontology categories suggested functional bias in the pattern of gene retention of paralogs. Young paralogs related to signal transduction, metabolism, and energy pathways have been preferentially retained. Non-random retention of duplicated genes seems to have mediated the expansion of gene families, some of which may have substantially increased their members after the origin of the Pyrinae. The joint analysis of over-duplicated functional categories and phylogenies, allowed evaluation of the role of both

polyploidy and small-scale duplications during this process. Finally, gene expression analysis indicated that 82% of duplicated genes, including 80% of young paralogs, showed uncorrelated expression profiles, suggesting extensive subfunctionalization and a role of gene duplication in the acquisition of novel patterns of gene expression. CONCLUSIONS: This study reports a genome-wide analysis of the mode of gene duplication in the apple, and provides evidence for its role in genome functional diversification by characterising three major processes: selective retention of paralogs, amplification of gene families, and changes in gene expression.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC bioinformatics 4(Journal Article): 41.

> BACKGROUND: The availability of multiple, essentially complete genome sequences of prokaryotes and eukaryotes spurred both the demand and the opportunity for the construction of an evolutionary classification of genes from these genomes. Such a classification system based on orthologous relationships between genes appears to be a natural framework for comparative genomics and should facilitate both functional annotation of genomes and large-scale evolutionary studies. RESULTS: We describe here a major update of the previously developed system for delineation of Clusters of Orthologous Groups of proteins (COGs) from the sequenced genomes of prokaryotes and unicellular eukaryotes and the construction of clusters of predicted orthologs for 7 eukaryotic genomes, which we named KOGs after eukaryotic orthologous groups. The COG collection currently consists of 138,458 proteins, which form 4873 COGs and comprise 75% of the 185,505 (predicted) proteins encoded in 66 genomes of unicellular organisms. The eukaryotic orthologous groups (KOGs) include proteins from 7 eukaryotic genomes: three animals (the nematode Caenorhabditis elegans, the fruit fly Drosophila melanogaster and Homo sapiens), one plant, Arabidopsis thaliana, two fungi (Saccharomyces cerevisiae and Schizosaccharomyces pombe), and the intracellular microsporidian parasite Encephalitozoon cuniculi. The current KOG set consists of 4852 clusters of orthologs, which include 59,838 proteins, or approximately 54% of the analyzed eukaryotic 110,655 gene products. Compared to the coverage of the prokaryotic genomes with COGs, a considerably smaller fraction of eukaryotic genes could be included into the KOGs; addition of new eukaryotic genomes is expected to result in substantial increase in the coverage of eukaryotic genomes with KOGs. Examination of the

phyletic patterns of KOGs reveals a conserved core represented in all analyzed species and consisting of approximately 20% of the KOG set. This conserved portion of the KOG set is much greater than the ubiquitous portion of the COG set (approximately 1% of the COGs). In part, this difference is probably due to the small number of included eukaryotic genomes, but it could also reflect the relative compactness of eukaryotes as a clade and the greater evolutionary stability of eukaryotic genomes. CONCLUSION: The updated collection of orthologous protein sets for prokaryotes and eukaryotes is expected to be a useful platform for functional annotation of newly sequenced genomes, including those of complex eukaryotes, and genome-wide evolutionary studies.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science (New York, N.Y.) 278(5338): 631-637.

> In order to extract the maximum amount of information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. Comparison of proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allowed the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This relation automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

Versant db4o open source object database, Versant.

Wu, J., X. Xu, et al. (2009). "FlyPhy: a phylogenomic analysis platform for Drosophila genes and gene families." BMC bioinformatics 10(Journal Article): 123.

BACKGROUND: The availability of 12 fully sequenced Drosophila species genomes provides an excellent opportunity to explore the evolutionary mechanism, structure and function of gene families in Drosophila. Currently, several important resources, such as FlyBase, FlyMine and DroSpeGe, have been devoted to integrating genetic, genomic, and functional data of Drosophila into a well-organized form. However, all of these resources are gene-centric and lack the information of the gene families in Drosophila. DESCRIPTION: FlyPhy is a comprehensive phylogenomic analysis platform devoted to analyzing the genes and gene families in Drosophila. Genes were classified into families using a graph-based Markov Clustering algorithm and extensively annotated by a number of bioinformatic tools, such as basic sequence features, functional category, gene ontology terms, domain organization and sequence homolog to other databases. FlyPhy provides a simple and user-friendly web interface to allow users to browse and retrieve the information at multiple levels. An outstanding feature of the FlyPhy is that all the retrieved results can be added to a workset for further data manipulation. For the data stored in the workset, multiple sequence alignment, phylogenetic tree construction and visualization can be easily performed to investigate the sequence variation of each given family and to explore its evolutionary mechanism. CONCLUSION: With the above functionalities, FlyPhy will be a useful resource and convenient platform for the Drosophila research community. The FlyPhy is available at http://bioinformatics.zj.cn/fly/ .