An integrative approach to ortholog prediction for disease-focused and other functional studies

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

| Citation         | Hu, Yanhui et al. “An Integrative Approach to Ortholog Prediction for Disease-focused and Other Functional Studies,” BMC Bioinformatics 12.1 (2011): 357. Web. 9 Mar. 2012. |
|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| As Published     | http://dx.doi.org/10.1186/1471-2105-12-357                                                                                                                                                       |
| Publisher        | Springer (Biomed Central Ltd.)                                                                                                                                                                  |
| Version          | Final published version                                                                                                                                                                       |
| Accessed         | Mon Dec 10 16:04:22 EST 2018                                                                                                                                                                   |
| Citable Link     | http://hdl.handle.net/1721.1/69633                                                                                                                                                            |
| Terms of Use     | Creative Commons Attribution                                                                                                                                                                  |
| Detailed Terms   | http://creativecommons.org/licenses/by/2.0                                                                                                                                                      |
An integrative approach to ortholog prediction for disease-focused and other functional studies

Yanhui Hu, Ian Flockhart, Arunachalam Vinayagam, Clemens Bergwitz, Bonnie Berger, Norbert Perrimon and Stephanie E Mohr

Abstract

Background: Mapping of orthologous genes among species serves an important role in functional genomics by allowing researchers to develop hypotheses about gene function in one species based on what is known about the functions of orthologs in other species. Several tools for predicting orthologous gene relationships are available. However, these tools can give different results and identification of predicted orthologs is not always straightforward.

Results: We report a simple but effective tool, the Drosophila RNAi Screening Center Integrative Ortholog Prediction Tool (DIOPT; http://www.flyrnai.org/diopt), for rapid identification of orthologs. DIOPT integrates existing approaches, facilitating rapid identification of orthologs among human, mouse, zebrafish, C. elegans, Drosophila, and S. cerevisiae. As compared to individual tools, DIOPT shows increased sensitivity with only a modest decrease in specificity. Moreover, the flexibility built into the DIOPT graphical user interface allows researchers with different goals to appropriately ‘cast a wide net’ or limit results to highest confidence predictions. DIOPT also displays protein and domain alignments, including percent amino acid identity, for predicted ortholog pairs. This helps users identify the most appropriate matches among multiple possible orthologs. To facilitate using model organisms for functional analysis of human disease-associated genes, we used DIOPT to predict high-confidence orthologs of disease genes in Online Mendelian Inheritance in Man (OMIM) and genes in genome-wide association study (GWAS) data sets. The results are accessible through the DIOPT diseases and traits query tool (DIOPT-DIST; http://www.flyrnai.org/diopt-dist).

Conclusions: DIOPT and DIOPT-DIST are useful resources for researchers working with model organisms, especially those who are interested in exploiting model organisms such as Drosophila to study the functions of human disease genes.

Background

It is well established that the evolutionary conservation of proteins correlates with conservation at the level of biological and/or biochemical functions. Thus, identification of putative orthologs among species can be a helpful first step in functional genomics studies, as functional information obtained in one species can be used to formulate testable hypotheses about function in another. Indeed, for many approaches, identification of orthologs is the first step to using model systems such as S. cerevisiae, C. elegans, and Drosophila melanogaster (Drosophila) to study human genes, such as the set of disease genes in the Online Mendelian Inheritance in Man (OMIM) database [1-4] or the results of a genome-wide association study (GWAS) [5,6]. Similarly, orthologs of genes functionally characterized in model organisms may identify as yet unknown candidate genes relevant to human health. Thus, in addition to its value in evolutionary studies, mapping of orthologs among species can serve important purposes for functional analyses.

To date, a large number of approaches and corresponding databases have been developed to predict putative orthologous genes among various species. Most of these are based on either phylogeny-based algorithms (e.g. TreeFam, Phylome, Ensembl Compara) or on sequence
similarity (*e.g.* InParanoid, orthoMCL and OMA) [7-12]. More recently, additional tools have been reported based on functional similarity among ortholog groups identified using protein-protein interaction (PPI) networks (*e.g.* NetworkBLAST, IsoBase) [13-15]. Notably, PPI-based tools have strong potential to improve in performance over time, as PPI data sets are improving in coverage and quality at a rapid pace. Because of their differing approaches, user interfaces, output formats and so on, these otherwise powerful tools and algorithms can be difficult to use by a biologist without expertise in bioinformatics, particularly when handling large-scale data sets such as the results of a full-genome RNAi screen or large-scale proteomics analysis.

Furthermore, determining the sensitivity and specificity of a given ortholog prediction approach is not easy. Only a few studies have compared the relative quality of results obtained with different tools. Evaluation efforts are hindered by the very reason that motivates a researcher to develop an ortholog prediction tool—*i.e.* we simply do not have enough functional data. Nevertheless, some studies have looked at the comparative performance of some tools using available approaches. Altenhoff *et al.*, for example, assessed several tools, including OMA, orthoMCL, Inparanoid and Compara, using gene ontology (GO), enzyme category (EC), gene expression profile and gene neighborhood conservation. The results of their comparison and corresponding estimate of the relative quality of the tools suggest that simple bidirectional best hit (BBH)-based ortholog predictions are more likely to yield functionally consistent predicted orthologs. Their results also highlight that different tools have different advantages [16].

As each of the existing tools has different strengths, combining results generated by more than one tool might be the best way forward. At least two such integrative methods have been reported. HCOP is a human-centered ortholog prediction tool that in its original published form, facilitated comparison of ortholog assertions made by PhIGs, HomoloGene, Ensembl Compara, Inparanoid and MGI [17,18]. Currently, the HCOP website also includes predictions from Evola, HGNC, OMA, OPTIC, TreeFam, UCSC and ZFIN and has removed PhIGs http://www.genenames.org/cgi-bin/hcop.pl. Another tool, COMPARE, is not limited to human-centered relationships and combines results from Inparanoid, orthoMCL, Ensembl Compara and TreeFam [19]. These combined approaches are useful in that they allow for identification and comparison of ortholog results from various sources, allowing the end-user to make informed choices. However, neither tool incorporates the full range of approaches available (*i.e.* sequence-based, tree-based and network-based approaches) and facilitates comparison among any pair of common model organisms. For example, HCOP only provides predictions relevant to human genes, and COMPARE provides predictions among human, mouse, worm, fly, chicken, zebrafish and sea squirt genes but not yeast genes. Thus, room for significant refinement of integrative approaches remains.

After successful ortholog prediction, a next challenge can be to combine functional information obtained in multiple organisms to uncover conserved functions that provide insights into human diseases. Insight into disease using ortholog prediction tools can be twofold. First, identification of orthologs can make model organisms available to perform molecular genetic studies of conserved disease genes that are not possible to perform in other organisms. Second, conserved genes first identified in a model organism can become candidate disease genes and/or provide new insights relevant to disease. We also note that genes that are not conserved among specific groups of species or have distinct differences may be important for evolutionary studies, and may be good targets for the development of biocides, such as to control disease vector, parasite, or crop pest populations [20,21].

The increasing number of human genes that have been directly or indirectly associated with diseases provides a good starting point for identification of conserved orthologs in model organisms for disease-focused studies. To date, a large number of disease-associated genes have been reported and the information compiled into various resources. Perhaps top among these is the NCBI Online Mendelian Inheritance in Man (OMIM) database [1-4]. Since the completion of the human genome and following various technological advances, an increasing number of genome-wide association studies (GWAS) provide evidence for possible disease association of many previously uncharacterized genes [5,6]. In addition, there have been several automated efforts to evaluate the links between human diseases and genes, such as based on gene annotations or the published literature. The Disease Ontology project, for example, assigns disease information based on GeneRIF annotation [22], and resources such as BITOLA, MedGene and HuGE Navigator [23-25] provide systematic summaries of disease-gene relationships culled from the literature in NCBI PubMed.

Facilitating the identification of orthologs between a model organism and humans is of particular relevance to genes associated with human diseases. *Drosophila* is a model organism of particular interest for which a wide variety of molecular genetic tools are readily available. Moreover, *Drosophila* models have been developed for a number of human diseases. For example, *Drosophila* is an established model for a wide variety of nervous
system-related diseases (reviewed in JEIBMANN and PAULUS 2009) [26]; various muscular dystrophies [27-29]; responses to infection by human pathogens (e.g. [30-34]); multi-symptom inherited disorders [35,36]; heart disease [25,37,38]; and cancer [9,39-42]. *Drosophila* is an emerging model for the study of asthma [43], lipotoxicity [44], and metabolism [45-47], and is being used to study responses to therapeutic treatments and natural products [48-50]. Overall, conservation between humans and *Drosophila* at the gene, pathway, cognate organ and behavioral levels, together with available tools and assays, suggest that *Drosophila* can serve as an excellent system for rapid functional characterization of GWAS candidates and other disease-associated genes, such as in placing them in specific conserved pathways or implicating them in specific metabolic, physiological or developmental roles.  

Despite the utility and potential of *Drosophila* in studying disease, however, the number and scope of resources linking human disease genes to known or putative *Drosophila* orthologs is still limited. OrthoDisease and Homophila are two tools aimed at translation of OMIM disease gene annotations to genes of other organisms. OrthoDisease maps orthologous gene relationships among human genes and genes in 100 genomes, including *Drosophila*, using the Inparanoid tool [51]. Homophila was designed specifically for *Drosophila* and uses BLASTP to compare human protein sequences to the *Drosophila* proteome [52,53]. Both approaches have value and are used extensively. Nevertheless, each has important caveats. For example, OrthoDisease may miss relevant orthologs due to the limited coverage provided by use of just one approach to ortholog mapping, and Homophila can generate inappropriate gene mapping via uncurated one-way BLASTP results. Moreover, neither tool currently accommodates the increasing amount of GWAS data available to the community. Thus, we have developed a new online tool that integrates the results of several ortholog-mapping tools based on different algorithms, and used the approach to create an online-searchable analysis of relationships among genes in OMIM or GWAS data sets and genes in the mouse, zebrafish, *C. elegans*, *Drosophila*, and *S. cerevisiae* genomes.

### Results and Discussion

**An integrative tool to identify putative orthologs among several model organisms**

More than three dozen large-scale ortholog prediction tools or resources have been developed in the past decade, and more than a dozen of these are currently available online in relatively accessible formats. Based on published reviews and comparative studies evaluating different tools, we selected nine tools that provide a reasonable balance of specificity and sensitivity [7-12,15,16,54-59]. We excluded tools that were developed prior to the availability of full-genome sequence and/or that have not been maintained and updated. However, we note that the data available from tools we included are based on different genome releases (see the Documentation Page, http://www.flyrnai.org/DRSC-OPT.html). Three of the nine tools we selected are representative of the phylogenetic tree approach and five are representative of tools that cluster genes based on genome-wide sequence comparison. In addition, the recent availability of large datasets and networks of protein-protein interactions (PPIs) has opened the doors to proteomics-based approaches to identification of functional orthologs, *i.e.* predictions based on the similarity of the PPI networks surrounding the potentially orthologous proteins. A leading resource for ortholog prediction based on this approach is IsoBase [15,60,61], which is aimed at identification of functionally related proteins based on integration of sequence similarity and protein-protein network-based information. As our principle goal in developing the tool is a focus on functional orthologs, we included IsoBase in addition to the other eight tools in our integrative approach.

Integration of the results of these nine tools required a number of pre-processing steps and decisions, as the outputs from the different tools differ both in terms of their use of gene and protein identifiers, and in terms of the format and complexity of output files. The major identifiers used by the nine tools are Ensembl protein identifiers, NCBI gene or protein identifiers, Uniprot identifiers, and species-specific identifiers such as from FlyBase and WormBase (Additional file 1). Another difference among the tools that we had to address was that some tools make predictions based on comparison with all protein isoforms of a gene, whereas others first normalized the proteins at the gene level by selecting the longest isoform for inclusion in the analysis. To facilitate merging and comparing the outputs from the nine tools we included, we first programmed an algorithm to standardize different identifiers to Entrez gene IDs using mapping files, such as from NCBI or species-specific databases. Our use of gene IDs eliminated redundancy due to inclusion by some tools of multiple isoforms. Recognizing that many users of our tool would prefer to be inclusive rather than to inadvertently exclude genes that might be relevant to a given disease-focused or other study, we chose to include not only ortholog predictions but also inparalog or co-ortholog predictions. Although paralogous genes may have diverged in function more than orthologs, those that have arisen from recent gene duplications may be members of a function-oriented ortholog group [57,62]. When building the tool, we first obtained the data from each site; then standardized various identifiers and converted the different output
formats; next integrated all the predictions into our FlyR-NAi database [63]; and finally launched an online application (see Materials and Methods).

The new tool is named DIOPT, for DRSC Integrative Ortholog Prediction Tool, and it is freely accessible online at http://www.flyrnai.org/diopt (Figure 1). Using Drosophila or human data as inputs, we find that DIOPT identified 28605 fly and human orthologous relationships, of which 2537 are one-to-one relationships. As expected, DIOPT identified more human-fly many-to-one relationships than one-to-many relationships (i.e. 5504 versus 1672), consistent with the lower redundancy of the fly genome. DIOPT achieves an approximately 70% to 428% increase in the number of Drosophila-human gene relationships identified as compared with use of any individual tool. This corresponds to 20% to 116% more coverage of the Drosophila genome and 30% to 172% more coverage of the human genome (Table 1). DIOPT allows users to use all nine tools or any subset in a search. Although a small number of differences in predictions might be due to the use of different genome release versions by different tools, most differences will be due to the specific algorithms and approaches used in the various tools. DIOPT calculates a simple score indicating the number of tools that support a given orthologous gene-pair relationship, such that the maximum score for most species pairs is nine (see the Document Page, http://www.flyrnai.org/DRSC-OPT.html), assuming that the user selects to accept outputs from all tools (i.e. the default setting). Because all nine sources cover the entire genome of each species included in the search, we reasoned that this simple count of how many tools agree with each other is an appropriate first indicator of confidence (see below). We next focused on methods for evaluation and improved display of the results, towards the goals of generating higher-confidence outputs, making the tool user-friendly for biologists, and providing additional gene- or protein-associated information and functionality.

Assessment of functional consistency of ortholog pairs predicted using DIOPT

As the field lacks a commonly accepted “gold-standard” test set for assessment of quality, we opted to test a few approaches that together help give an overall picture of the quality of DIOPT results. Our first assessment was based on molecular function annotation of gene ontology (GO MF). GO MF is widely used to assess the functional consistency of two genes. As a first test of DIOPT performance, we examined the similarity of GO MF annotations for all Drosophila-human ortholog pairs predicted using the default DIOPT settings (i.e. incorporating results from all nine tools). We recognized that a potential caveat to using GO MF annotations to analyze performance is the danger for ‘circular arguments’, as some GO MF annotations are themselves based on information about orthologs. Indeed, 40% of GO functional annotations for Drosophila genes and 55% for human genes were assigned based on sequence similarity without manual inspection. We were able to address this by taking into consideration the evidence codes for GO MF assignments. Namely, we opted to perform the analysis with two subsets; i.e., the subset of GO MF annotations supported by experimental data or curator/author statements, and applying an even more stringent rule, only those GO MF annotations that are supported by experimental data. We used GOSemSim [64], an R package http://www.r-project.org/ for semantic similarity computation, to calculate the functional similarity score for each orthologous pair by considering the set of selected GO terms assigned to each gene. The results of our analysis demonstrate that ortholog pairs with lower scores (i.e. predicted by fewer tools) have nearly comparable functional consistency as compared with ortholog pairs with high scores. However, higher score pairs do show a slightly higher level of functional similarity (Figure 2a). Thus, DIOPT scores provide one means of prioritizing or limiting gene list outputs.

The lack of a standardized benchmark remains a significant challenge in comparing and evaluating ortholog prediction tools. The “Quest for Orthologs” initiative is currently attempting to provide benchmarking for orthology predictions but that resource is not yet available [65]. To obtain some measure of DIOPT performance relative to individual tools, we used two different approaches based on available resources. First, we used a manually assembled set of high-confidence Drosophila-human orthologous pairs made up of 159 orthologous pairs of trans-membrane proteins generated by manual phylogenetic analysis [54], 89 pairs of kinases generated manually based on kinase domain similarity as well as phylogeny [66] and 71 pairs of oxidative phosphorylation genes assembled manually by sequence alignment and analysis of exon/intron structures [67,68]. All these data sets, with the exception of the OXPHOS set, have been previously used for evaluation of ortholog prediction tools. The idea here was to identify a set of proteins where orthology relationships can be established with high confidence by approaches other than those discussed here. To this end, the dataset’s composition is biased towards types of proteins for which such independent verification is feasible. However, such a bias can distort the evaluation results: the coverage of the selected proteins (especially, trans-membrane proteins) in PPI data is low and consequently, the performance of a PPI-based method like IsoBase is likely to be underestimated. Nevertheless, we believe that
Figure 1 DIOPT facilitates identification of predicted orthologs based on various approaches, with flexible inputs and outputs. The DIOPT graphical user interface facilitates input of small or large gene or protein lists and is compatible with one or more of a variety of identifiers. Three filters are available: two of them exclude relationships that are predicted by < 2 or < 3 tools unless the only match score is equal to or lower than the threshold. The third one and most stringent filter limits outputs to the best-matching ortholog(s) per gene, as judged by the number of tools supporting the prediction. DIOPT provides a score based on the number of tools supporting the prediction and a weighted score based on the average of GO semantic similarity of orthologous pairs predicted by each tool using high quality GO functional annotation. In addition, DIOPT also provides the option to view the original score of each tool and a protein-protein alignment based on an updated proteome annotation (RefSeq release 44). Percent protein identity is calculated for both the overall alignment and domain regions.
testing with this set can provide one useful way to ask if
as predicted, the integrative approach provides improved
results as compared with use of individual tools.

DIOPT is able to identify 301 of the total of 319 rela-
tionships in the high-confidence manually assembled test
set. Thus, DIOPT achieved 94% sensitivity, comparing
favorably to the 39% to 81% sensitivity observed with
individual tools. We also used this test set to look at spe-
cificity and observed the classical trade-off between spe-
cificity and sensitivity. (See Figure 3 for definitions of
these terms.) Without filters, the specificity of DIOPT is
38%, as compared to 48% to 71% observed with indivi-
dual tools (Figure 3a). This decrease may be acceptable
for some applications of the tool; however, for other
applications, users may prefer to limit the results. As
mentioned above, we reasoned that the simple DIOPT
score (i.e. the number of tools supporting the prediction)
might provide a way to achieve higher specificity, and
this indeed proves true. If we filter out the relationships
with score equal to one, i.e. relationships supported by
only one of the 9 tools, the sensitivity of DIOPT remains
at about 90% but specificity increases from 38% to 53%.
To take best advantage of this or of a user’s own assess-
ment, we provide users with several options for filtering
results. Users can include any subset of the nine tools in
a search; apply a filter that removes predictions sup-
ported by only one tool (i.e. scores < 2 are removed); or
apply a filter that returns only the predictions with the
best score (i.e. if a query gene matches to more than one
ortholog, only the ortholog(s) with the highest score are
displayed).

We performed a different type of test for specificity
using a negative test set; i.e. a set of 1,130 human genes
defined as human lineage-specific by Eukaryotic Clusters
of Orthologous Groups (KOG) [37]. We found that
DIOPT shows lower specificity (82%) as compared to
individual tools using this test set (93%-99%) (Figure 3b).
This may be partially due to the fact that the ‘lineage-
specific’ set in KOG was assembled over a decade ago,
when the available approaches may not have been sensi-
tive enough to detect meaningful gene relationships that
exist in this group. This idea is supported by the observa-
tion that 34% of the genes from the KOG set that were
identified as not human specific by DIOPT are predicted
by multiple tools to have orthologs in the fly.

Online Features of DIOPT
More than a dozen ortholog prediction tools are currently
available online. However, most tools do not allow for
searching in batch mode (i.e. with a long list of genes or
other identifiers), making them impractical to use for
many types of analysis, including analysis of the typically
long lists of genes or proteins resulting from the increasing
number of high-throughput RNAi, proteomics and other
studies. Although many of the tools do allow a researcher
to install the program locally, potentially facilitating batch
searches, this is a significant hurdle for a researcher with-
out expertise in bioinformatics, and even after installation,
large datasets may still be difficult to process and subse-
quently manage as outputs. To make searching with short
or long lists of identifiers possible online, we developed
DIOPT to have the flexibility of accepting one or a few
typed entries, a long list of entries, or an uploaded list of
gene or protein identifiers (Figure 1). Moreover, we built
gene and protein mapping into DIOPT and accommodate
a large number of NCBI, species-specific and other differ-
ent identifiers. Thus, after choosing the input and output
species of interest, researchers can use most common
identifiers, or even mix different types of identifiers, in
their input lists. For expert users, we also provide the
option of specifying the type of identifier, such that other
matches may not be included. We also make it possible to
search with RNAi reagent IDs (i.e. amplicon IDs) from the
Drosophila RNAi Screening Center (DRSC) or fly stock

| Tool       | D.mel-H.sap relationships | Unique relationships* | D.mel genes covered | H.sap genes covered |
|------------|---------------------------|-----------------------|---------------------|---------------------|
| Compara    | 16880                     | 3994                  | 8416                | 9826                |
| Homologene | 5423                      | 82                    | 5015                | 4762                |
| Inparanoid | 10204                     | 695                   | 5855                | 7441                |
| IsoBase    | 9051                      | 1639                  | 8065                | 7251                |
| OMA        | 5673                      | 296                   | 4499                | 5222                |
| orthoMCL   | 12853                     | 2352                  | 6435                | 7890                |
| Phyloyme   | 8690                      | 1496                  | 4560                | 6251                |
| RoundUp    | 7480                      | 398                   | 5383                | 6706                |
| TreeFam    | 12894                     | 1347                  | 7156                | 9106                |
| DIOPT      | 28605                     | NA                    | 9724                | 12971               |

D.mel, Drosophila melanogaster; H.sap, Homo sapiens
* Unique Relationships, relationships not predicted by any of the other eight tools.

Table 1 The number of Drosophila-human ortholog pairs predicted using individual published ortholog tools or the integrative tool DIOPT
evaluate orthologs we have also calculated weighted scores. To do this, the nine tools were weighted based on average GO semantic similarity of orthologous pairs annotated with high-quality GO MF annotation (GO MF2, see above). The average GO semantic similarity scores are very similar for each of the 9 tools, ranging from 0.612 to 0.691 (Figure 2b). To assign weights, we normalized the scores by setting the median to one and defined the weights accordingly (0.903 to 1.019). Weighted scores serve as an additional way for users to make informed decisions when multiple orthologs with the same simple count score are retrieved.

We have added the option to view the original prediction/evaluation scores from each of the tools, along with the protein alignment. To help address the caveat that the original predictions were made by various tools using different genome releases, the protein-protein alignments we provide are based on a current proteome annotation for any given output pair, including highlighting of regions that correspond to domain annotations (RefSeq release 44; Figure 1). The sequence identities of the entire aligned sequence, as well as identity within the annotated domain regions, which typically have higher identity, are summarized and displayed. For example, for the fly gene *par-1* (FlyBase ID FBgn0260934) and its human ortholog MARK4 (NCBI Gene ID 57787), DIOPT shows that the overall protein identity is 47% whereas the identity of the kinase catalytic domain is 88% (Figure 1). This feature allows researchers to quickly zoom in on functionally critical motifs, amino acids, etc. based on their specific expertise and interests. Finally, the database and application of DIOPT were designed in a way that it is easy to be updated and expanded, such as when new and improved ortholog prediction tools become available, or when existing tools are updated.

**Using DIOPT results to identify genes in model organisms related to human diseases**

High-throughput genomics technologies generate a vast amount of data, making it possible for researchers to search for genetic causes of human diseases. Concomitant with the rise in disease genomics is an increase of publicly accessible resources that associate loci, mutations and genes with putative or known roles in disease. Perhaps top among these is Online Mendelian Inheritance in Man (OMIM), which focuses on inherited diseases. The information contained in OMIM has been extracted from publications and undergone manual curation [1-4]. GWAS data sets, such as those cataloged by the National Genome Research Institute (NHGRI; see http://www.genome.gov/gwastudies/), provide associations among loci or genes and diseases or traits based on linkage analysis using single nucleotide polymorphisms as markers. Due to technological advances in genotyping,
Figure 3 Comparing sensitivity and specificity among the nine individual ortholog prediction tools and the integrative tool DIOPT

Comparing sensitivity and specificity using a manually assembled reference set (3a), or using the KOG human-specific gene set (3b). Sensitivity is defined as the percent of manually assembled pairs that can be identified by each tool versus all the manually assembled orthologous pairs (same for Figures 3a and 3b). In 3a, specificity is defined as the percent of manually assembled pairs that can be identified by each tool versus all the orthologous pairs identified by each tool if queried with either the Drosophila or the human genes from the test set. In 3b, specificity is defined as the percent of putative human-specific genes that do not have fly ortholog versus all the human specific genes from KOG list.
the amount of GWAS data now available has increased dramatically [5,6]. With GWAS data, identification of a specific disease-causing mutation in a gene may be missing, with the results pointing to one or more candidate genes that may be relevant. These candidates require further functional characterization, and the wealth of phenotypic information available in Drosophila and other model organisms may be of help to narrow down the list of genes in a disease-associated locus.

Drosophila has been used extensively to study human diseases but as discussed previously, resources linking fly or other model organism genes to human diseases are still limited. The tools OrthoDisease and Homophila were developed to help identify disease genes in model organisms. These tools incorporate OMIM but not GWAS information and their use of OMIM data sets differ from one another. OMIM annotation includes both gene/locus terms and disease phenotype terms. OrthoDisease includes most phenotype terms, whereas Homophila is focused on gene/locus terms, with neither tool fully incorporating all available terms at OMIM. Moreover, OrthoDisease uses Inparanoid predictions to map human genes to genes in model organisms, such that some ortholog pairs identified using DIOPT-DIST are missed when searching OrthoDisease (e.g. mapping of human TP53 to Drosophila p53). By contrast, Homophila uses one-way BLASTP, which is so inclusive that it tends to include erroneous predictions. For example, ALDH2 is identified in OMIM as ‘causing acute alcohol sensitivity’ (OMIM IDs 610251 and 100650).

For ALDH2, OrthoDisease predicts one corresponding gene in Drosophila (Aldh) and Homophila predicts 11 corresponding genes in the fly. In contrast, DIOPT predicts two fly genes: Aldh, which is also predicted using OrthoDisease and Homophila, and the additional gene CG31075, which shares 59% identity and 74% similarity with human ALDH2 and is predicted to be an ALDH2 ortholog by four tools.

Thus, we reasoned that developing a tool based on DIOPT might improve upon results from existing tools, as well as provide an opportunity to incorporate newly available GWAS data. The tool we developed was named DIOPT-DIST as it combines DIOPT ortholog mapping with disease and trait information http://www.flyrnai.org/diopt-dist. With DIOPT-DIST, we linked C. elegans, Drosophila, zebrafish, mouse or yeast genes to human diseases by extracting disease-gene information at OMIM http://www.ncbi.nlm.nih.gov/omim and GWAS data cataloged by NHGRI http://www.genome.gov/gwastudies and then using DIOPT to map orthologs (Figure 4 & 5; see also Materials & Methods and the Documentation Page, http://www.flyrnai.org/DRSC-OPT.html). DIOPT-DIST includes both disease phenotype terms and gene/locus terms from OMIM. At the DIOPT-DIST website, users can query with one or more genes from a model organism to get a list of human orthologs along with their disease or trait annotations. Conversely, users can query with a disease term, disease category (see below) or OMIM IDs to get a list of human genes as well as their corresponding orthologs in a specified model organism. The DIOPT score in the results table links to a page displaying a protein alignment and details about the DIOPT results.

To develop the DIOPT-DIST tool, we first had to normalize the disease and trait information using a standardized classification of disease terms. The 5700 disease or trait descriptions from the OMIM and GWAS data sets are in free text and show significant variation in word choice, phrasing and sentence structure in describing diseases. Indeed, in some cases the terminologies vary even within one resource, including spelling variations or typing errors. These variations hinder systematic searches for disease phenotypes. By contrast, Medical Subject Heading (MeSH) terms are a previously established controlled vocabulary describing disease phenotypes (C01-C27) and mental disorders (F03) with a hierarchical structure allowing more general disease terms to be distinguished from sub-classes. MeSH accommodates most disease synonyms and is used to index medical literature in PubMed [71]. Thus, we chose to classify OMIM and GWAS disease terms based on MeSH general disease categories in order to facilitate disease search and data mining.

To do this, we took advantage of OMIM clinical synopsis heading/sub-heading text [72], and were able to map about 3000 OMIM disease terms to the corresponding MeSH terms. Using MeSH synonym annotation, we were able to classify most of the remaining OMIM and GWAS terms. Some cannot be mapped using MeSH clinical synopsis heading/sub-headings or synonyms because they are traits rather than diseases (e.g. eye or hair color) or they are disease risk factors, diagnoses or treatments, and neither of those categories is covered by MeSH disease annotations. Therefore, we created two additional categories, “Disease risk factor, diagnosis or treatment” and “Trait”, and mapped to these categories as appropriate (Table 2). Because both OMIM and GWAS are updated periodically, and particularly as GWAS data is accumulating at a rapid pace, we plan to regularly update DIOPT-DIST, including conversion to our standard MeSH-based terms. The date of the last update will be indicated on the DIOPT-DIST search page.

We found a total of about 6000 human genes from OMIM and GWAS data associated genetically with a wide range of diseases and traits. Using DIOPT without any filter, we can identify 4282 corresponding orthologs in Drosophila. The numbers drop to 3620 and 3382
\emph{Drosophila} genes, respectively, when we apply filters that exclude predictions supported by scores of < 2 or < 3. Of these high-confidence orthologs, 33% were from both OMIM and GWAS, 37% were from OMIM only, and 30% were from GWAS only (Figure 4). Our results suggest that molecular functions relevant to a large number of human disorders and GWAS associations could potentially be studied using \emph{Drosophila}. Table 2 summarizes the distribution of these \emph{Drosophila} genes over 25 disease categories in MeSH, plus our two added categories. The categories with the most genes are “Congenital, Hereditary, and Neonatal Diseases and Abnormalities”, “Nervous System Diseases”, “Disease risk factor, diagnosis or treatment” and “Musculoskeletal Diseases”. As expected, DIOPT-DIST also performs well with examples such as those mentioned above, \textit{i.e.} human TP53 mapping to \emph{Drosophila} p53 (\textit{e.g.} p53 appears among results of a disease full text search for “cancer” with \emph{Drosophila} as the output species) and the mapping of ALDH2 to \emph{Drosophila} Aldh and CG31075 (\textit{e.g.} resulting from search with OMIM ID 610251).

The annotated results of our analysis should be a useful resource for prioritizing genes from \emph{Drosophila} and four additional model organisms (\textit{S. cerevisiae}, \textit{C. elegans}, zebrafish, or mouse). Users can query with a disease or disease category to generate a corresponding list of predicted orthologs of the human genes in one of the five model organisms included in the tool (Figure 5). For example, a search using “diabetes” currently identifies 217 predicted \emph{Drosophila} orthologs corresponding to human diabetes-related genes (see Additional file 2). This list includes 30 genes that have previously been studies in metabolic diseases in fly, including \textit{Akt1}, InR and \textit{chico} (reviewed in \cite{45,73}). This list also provides additional candidates such as \textit{sima} and \textit{Spargel}. The availability of a wealth of molecular genetic tools for \emph{Drosophila} makes it possible to study candidate disease-associated genes identified using DIOPT-DIST.
Comparing disease associations in the literature for Drosophila and human orthologs

In addition to curated sources like OMIM, information about gene and disease associations also exist in the form of millions of published literature citations in free text format. For example, Bodenreider and Burgun recently showed that the MeSH index of biomedical articles provides a framework for comparing disease phenotypes of mouse and human orthologous genes [71]. Moreover, they showed that the categories “Neoplasms” and “Nervous System Diseases” have the most disease annotations in common between mouse and human orthologs. For our analysis, we extracted disease phenotypes based on MeSH indexing of biomedical articles in PubMed literature citations for the 28605 orthologous gene pairs between Drosophila and humans (Table 1). We identified 3557 fly genes with at least one disease phenotype associated in the literature. Of these, 2739 (77%) have at least one putative human ortholog as revealed by DIOPT. In addition, 2283 Drosophila genes share at least one disease category with their human orthologs based on independent studies (Additional file 3). These Drosophila genes cover 23 of the 26 disease categories (Table 3). “Cancer”, “Digestive system diseases” and “Nervous system diseases” have the highest percentages of orthologous genes sharing the same disease phenotype(s), whereas parasitic diseases, eye diseases and infectious diseases have the lowest percentages of orthologous genes with the same disease phenotype(s).

We further looked at the publications for the 4282 fly disease genes identified by mapping human OMIM and GWAS data. 1534 of them have been published as related to disease using the fly model and 888 (58%) are associated with the same disease category in the literature as that assigned by OMIM/GWAS mapping (Table 4). This supports the idea that orthologous genes are likely to have similar disease phenotypes across species and also validates disease category assignments made when we mapped OMIM/GWAS annotation. This analysis also demonstrates that there are still many disease candidate genes to be studied in model organisms. OMIM and GWAS annotation are ongoing efforts, enriching for genes mutated in the germ-line and associated with inherited human diseases, traits, predisposition, etc. Conversely, literature mining uses existing information to identify additional disease-gene associations, and provided more than 2000 additional candidate disease-related genes in Drosophila, a useful supplement to the set of genes annotated at OMIM and GWAS.

Conclusions

Existing approaches to identification of predicted orthologs provide an excellent starting point for the development of integrative tools. DIOPT provides a simple but
Table 2 Disease genes from OMIM and GWAS (sorted by # of human genes)

| Disease category | # of disease terms | # of human genes | # of Drosophila genes | # of Drosophila genes filter1 | # of Drosophila genes filter2 |
|------------------|--------------------|------------------|-----------------------|-----------------------------|-----------------------------|
| Congenital, Hereditary, and Neonatal Diseases and Abnormalities | 3518 | 2803 | 2614 | 2076 | 1890 |
| Nervous System Diseases | 1921 | 1816 | 1901 | 1437 | 1339 |
| Disease risk factor, diagnosis or treatment | 273 | 1459 | 1768 | 1398 | 1292 |
| Musculoskeletal Diseases | 1523 | 1332 | 1433 | 1110 | 1035 |
| Eye Diseases | 1410 | 1214 | 1356 | 1038 | 953 |
| Skin and Connective Tissue Diseases | 1100 | 1117 | 1349 | 1039 | 949 |
| Digestive System Diseases | 901 | 1063 | 1397 | 1058 | 935 |
| Otorhinolaryngologic Diseases | 933 | 879 | 996 | 750 | 704 |
| Male Urogenital Diseases | 782 | 863 | 1069 | 838 | 765 |
| Cardiovascular Diseases | 843 | 856 | 1140 | 869 | 756 |
| Female Urogenital Diseases and Pregnancy Complications | 771 | 830 | 1041 | 819 | 741 |
| Nutritional and Metabolic Diseases | 400 | 701 | 896 | 698 | 608 |
| Stomatognathic Diseases | 686 | 637 | 761 | 559 | 528 |
| Mental Disorders | 344 | 626 | 906 | 610 | 543 |
| Neoplasms | 500 | 616 | 673 | 493 | 463 |
| Pathological Conditions, Signs and Symptoms | 539 | 608 | 489 | 337 | 319 |
| Hemoric and Lymphatic Diseases | 628 | 584 | 775 | 614 | 545 |
| Endocrine System Diseases | 442 | 511 | 627 | 451 | 427 |
| Immune System Diseases | 323 | 480 | 523 | 371 | 337 |
| Respiratory Tract Diseases | 447 | 479 | 639 | 493 | 445 |
| Trait | 208 | 434 | 536 | 374 | 336 |
| Virus Diseases | 23 | 65 | 48 | 41 | 39 |
| Bacterial Infections and Mycoses | 19 | 32 | 36 | 24 | 24 |
| Parasitic Diseases | 7 | 16 | 28 | 19 | 18 |
| Substance-Related Disorders | 7 | 14 | 17 | 13 | 13 |

Methods

Data source

Genome-wide orthologous prediction results of the 9 tools selected were retrieved from their corresponding websites (see the Document Page, http://www.flyrnai.org/DRSC-OPT.html). The test set of manually assembled orthologous pairs was retrieved from three resources and fly-human relationships were selected. The trans-membrane proteins were retrieved from Multi-Paranoid http://multiparanoid.sbc.se/stats.html; protein kinase annotations were retrieved from OrthoInspector http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024942/bin/1471-2105-12-11-S1.XLS; and the oxidative phosphorylation nuclear genes were extracted from MitoComp2 http://www.mitocomp.uniba.it/. The set of putative human-specific genes were retrieved from KOG http://www.ncbi.nlm.nih.gov/COG/. Human disease genes were extracted from OMIM morbidmap and mim2gene at NCBI. A clinical synopsis for each disease was extracted from the omim.txt file using an in-house developed JAVA program. GWAS information was extracted from the NGRI GWAS publications list http://www.genome.gov/gwastudies/, which can be downloaded via FTP http://www.genome.gov/admin/gwascatalog.txt[5]. The MeSH entry term lookup file was generated using an in-house developed JAVA program. The MeSH indexes were extracted from a local copy of PubMed with an in-house developed JAVA program.
Categorizing the disease terms

CS headings of OMIM terms were manually matched to MeSH disease category terms. For GWAS disease/trait terms and OMIM phenotype terms without CS headings, terms were first simplified and then mapped to MeSH using the MeSH entry term lookup file (e.g. "breast cancer", "breast tumor" are the entry terms for the MeSH term "breast neoplasms"). Terms corresponding to traits (e.g. "hair color") or disease risk factors, diagnoses and treatments (e.g. terms as divergent as "HDL cholesterol," "hip bone size," "telomere length," and "tanning") could not be mapped in this way because they are not covered by MeSH terms (i.e. they are not diseases). Therefore, we created two additional MeSH term-like categories, Y01 for “Disease risk factor, diagnosis or treatment” and Y02 for “Trait,” and mapped terms to these categories as appropriate (Table 2). Following mapping to existing MeSH categories and the two new categories, about 100 terms remained unmapped. These turned out to be rare disorders and were mapped to MeSH terms manually.

Implementation of the DIOPT user interface and disease gene query interface

The DIOPT user interface is implemented as a collection of CGI scripts written in Perl. They are hosted on a shared server provided by the Research IT Group (RITG) at Harvard Medical School. The database is hosted on a MySQL server also provided by RITG.

Protein alignments

Protein sequences were retrieved from the NCBI RefSeq database release 44. The longest isoform was selected if
there were multiple isoforms per gene. Protein alignments were computed in advance using the EMBOSS water program and stored in the database for display. The water program was run using the BLOSSUM62 matrix and all defaults. The alignment display features the aligned portion of the two protein sequences with a match line indicating the degree to which each base pair contributes to the alignment score, with “*” indicating an exact match between the bases; “.” indicating a base pair that scores positively on the BLOSSUM62 matrix; “.” indicating an alignment between any other two bases, and a space indicating an area where a gap was inserted into one sequence or the other. CDD domain annotations were extracted from RefSeq protein records and are highlighted in green on the aligned sequences. The locations of the domains on the alignment sequences, and the identity scores of the domain regions and the entire alignment, are computed at the time of display.

| Disease Category                                      | Fly disease genes mapped from OMIM/GWAS | Fly disease genes from OMIM/GWAS that are studied in disease in literature | Fly disease genes from OMIM/GWAS with disease categories confirmed by literature | % of fly disease genes with disease categories confirmed by literature |
|-------------------------------------------------------|------------------------------------------|---------------------------------------------------------------------------|--------------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| Congenital, Hereditary, and Neonatal Diseases and Abnormalities | 2614                                     | 707                                                                       | 539                                                                                 | 76%                                                                   |
| Otorhinolaryngologic Diseases                         | 996                                      | 12                                                                        | 8                                                                                    | 67%                                                                   |
| Nervous System Diseases                               | 1901                                     | 669                                                                       | 431                                                                                 | 64%                                                                   |
| Musculoskeletal Diseases                              | 1433                                     | 88                                                                        | 56                                                                                  | 64%                                                                   |
| Cardiovascular Diseases                               | 1140                                     | 95                                                                        | 58                                                                                  | 61%                                                                   |
| Stomatognathic Diseases                               | 761                                      | 14                                                                        | 8                                                                                    | 57%                                                                   |
| Skin and Connective Tissue Diseases                   | 1349                                     | 72                                                                        | 36                                                                                  | 50%                                                                   |
| Digestive System Diseases                             | 1397                                     | 66                                                                        | 31                                                                                  | 47%                                                                   |
| Hemic and Lymphatic Diseases                          | 775                                      | 64                                                                        | 27                                                                                  | 42%                                                                   |
| Neoplasms                                             | 673                                      | 382                                                                       | 145                                                                                 | 38%                                                                   |
| Male Urogenital Diseases                              | 1069                                     | 202                                                                       | 75                                                                                  | 37%                                                                   |
| Eye Diseases                                           | 1356                                     | 334                                                                       | 120                                                                                 | 36%                                                                   |
| Respiratory Tract Diseases                            | 659                                      | 6                                                                         | 2                                                                                    | 33%                                                                   |
| Immune System Diseases                                | 523                                      | 56                                                                        | 18                                                                                  | 32%                                                                   |
| Female Urogenital Diseases                            | 1041                                     | 236                                                                       | 72                                                                                  | 31%                                                                   |
| Diseases and Pregnancy Complications                  |                                           |                                                                           |                                                                                     |                                                                       |
| Mental Disorders                                      | 906                                      | 345                                                                       | 105                                                                                 | 30%                                                                   |
| Endocrine System Diseases                             | 627                                      | 94                                                                        | 27                                                                                  | 29%                                                                   |
| Nutritional and Metabolic Diseases                    | 896                                      | 255                                                                       | 71                                                                                  | 28%                                                                   |
| Parasitic Diseases                                    | 28                                       | 122                                                                       | 2                                                                                   | 2%                                                                   |
| Pathological Conditions, Signs and Symptoms           | 489                                      | 1074                                                                      | 143                                                                                 | 13%                                                                   |
| Bacterial Infections and Mycoses                      | 36                                       | 133                                                                       | 14                                                                                  | 11%                                                                   |
| Disorders of Environmental Origin                     | 1768                                     | 358                                                                       | 0                                                                                   | 0%                                                                   |
| Virus Diseases                                        | 48                                       | 24                                                                        | 0                                                                                    | 0%                                                                   |
| Substance-Related Disorders                           | 17                                       | 0                                                                         | 0                                                                                   | 0%                                                                   |
| All Categories                                         | 4282                                     | 1534                                                                      | 888                                                                                 | 58%                                                                   |

Additional material

Additional file 1: summary of different identifiers used by different ortholog prediction tools this file shows that the gene and/or protein identifiers used by different tools vary a lot.
Acknowledgements and Funding

The authors would like to thank members of the Perrimon lab and Lizabeth Perkins for helpful suggestions and discussions. We also thank Daniel Park and Rohit Singh for critical comments on the manuscript. This work was supported by NIH R01 GM067671 to N.P. In addition, N.P. is an Investigator of the Howard Hughes Medical Institute, S.E.M. is supported in part by the Dana Farber/Harvard Cancer Center and C.B. is supported by NIDDK of the Howard Hughes Medical Institute, S.E.M. is supported in part by the Dana Farber/Harvard Cancer Center and C.B. is supported by NIDDK of the Howard Hughes Medical Institute, S.E.M. is supported in part by the Dana Farber/Harvard Cancer Center.

Authors’ contributions

YH designed and tested the application, implemented the back-end of the application, and drafted the manuscript. IF implemented the user interface and contributed to database design. AV contributed to tool evaluation. CB contributed to disease classification and manuscript editing. 8B participated in the design of the application and manuscript editing. NP provided critical input on key features and edited the manuscript. SEM provided oversight, contributed to disease classification and manuscript editing. BB participated in web application, and drafted the manuscript. IF implemented the user interface and contributed to disease classification and manuscript editing. YH designed and tested the application, implemented the back-end of the application, and drafted the manuscript. IF implemented the user interface and contributed to database design. AV contributed to tool evaluation. CB contributed to disease classification and manuscript editing. 8B participated in the design of the application and manuscript editing. NP provided critical input on key features and edited the manuscript. SEM provided oversight, contributed to disease classification and manuscript editing. BB participated in web application, and drafted the manuscript. IF implemented the user interface and contributed to disease classification and manuscript editing.

Received: 11 May 2011 Accepted: 31 August 2011

Published: 31 August 2011

References

1. McKusick VA: On the naming of clinical disorders, with particular reference to eponyms. Medicine (Baltimore) 1998, 77(1):1-2.
2. Amberger J, Bocchini CA, Scott AF, Hamosh A: McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 2009, , 37 Database: D793-796.
3. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005, , 33 Database: DS14-517.
4. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009, 106(23):9362-9367.
5. Johnson AD, O’Donnell CJ: An open access database of genome-wide association results. BMC Med Genet 2009, 10.6.
6. Hi L, Coghlan A, Ruan J, Coin LJ, Herich JK, O’Connor L, Liu T, Zhang Z, Bolund L, et al: TreeFam: a curated database of phylogenetic trees of animal genome families. Nucleic Acids Res 2006, , 34 Database: DS52-580.
7. Huerta-Cepas J, Bueno A, Dopazo J, Gabaldon T: PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res 2008, , 36 Database: D491-496.
8. Veltia AJ, Severin J, Ureta-Vidal A, Heng L, Dubin R, Birney E: EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 2009, 19(2):375-385.
9. O’Brien KP, Remm M, Sonnhammer EL: Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 2005, , 33 Database: D476-480.
10. Li L, Stoeckert CJ Jr, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003, 13(9):2178-2189.
11. Schneider A, Desimoz C, Gonnert GH: OMA Browser-exploring orthologous relationships across 352 complete genomes. Bioinformatics 2007, 23(16):2180-2182.
12. Kalaev M, Smoot M, Ideker T, Sharan R: NetworkBLAST: comparative analysis of protein networks. Bioinformatics 2008, 24(4):594-596.
13. Bandopadhyay S, Sharan R, Ideker T: Systematic identification of functional orthologs based on protein network comparison. Genome Res 2006, 16(3):428-435.
14. Park D, Singh R, Baym M, Liao CS, Berger B: IsoBase: a database of functionally related proteins across PPI networks. Nucleic Acids Res 2009, 37 Database: D295-300.
15. Alterhoff AM, Desimoz C: Phylogenetic and functional assessment of orthologs: inference projects and methods. PLoS Comput Biol 2009, 5(1):e1000262.
16. Wright MW, Eyre TA, Lush MJ, Povey S, Bruford EA: HCP: the HGNC comparison of orthology predictions search tool. Mamm Genome 2005, 16(1):827-828.
17. Eyre TA, Wright MW, Lush MJ, Bruford EA: HCP: a searchable database of mammalian orthology predictions. Brief Bioinform 2007, 8(1):5-5.
18. Salgado D, Gimmenez G, Coulier F, Marcelli C: COMPARE, a multi-organism system for cross-species data comparison and transfer of information. Bioinformatics 2008, 24(3):447-449.
19. Vinet J, Smith CA, Backeljaaf CA, Bae J, Brown L, Basso D: The chitinase PCHT1 from the human malaria parasite Plasmodium faliparum lacks prosexoglycan and chitin-binding domains and displays unique substrate preferences. Proc Natl Acad Sci USA 1999, 96(24):14061-14066.
20. Beatty JM, Smagge G, Ogura T, Nakagawa Y, Spindler-Barth M, Henrich VC: Properties of ecdysonoid receptors from diverse insect species in a heterologous cell culture system: a basis for screening novel insecticidal candidates. Insect Biochem Mol Biol 2009, 39(11):987-995.
21. Osborne JD, Halsall J, Holko M, Lin SM, Kibbe WA, Zhi LJ, Danila MF, Geng C, Chisholm RL: Annotating the human genome with Disease Ontology. BMC Genomics 2009, 10(Suppl 1):S6.
22. Hu T, Hines LM, Wang H, Zuo D, Rivera M, Richardson A, LaBaer J: Analysis of genomic and proteomic data using advanced literature mining. J Proteome Res 2003, 2(6):405-412.
23. Hirotsu D, Peterlin B, Mitchell JA, Humphrey SM: Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005, 74(2-4):289-298.
24. Yu W, Cline M, Dolan SM, Yesupriya A, Wulf A, Liu T, Khoury MJ, Gwinn M: GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC Bioinformatics 2008, 9:205.
25. Witt AE, Hines LM, Collins NL, Hu Y, Gunawardane RN, Moreira D, Raphael J, Jepson D, Koundinya M, Rolfs A, et al: Functional proteomics approach to investigate the biological activities of cDNAs implicated in breast cancer. J Proteome Res 2008, 7(5):599-610.
26. Chartier A, Benoit B, Simonelig M: PathoMLO: A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. Cell Host Microbe 2008, 4(6):305-307.
31. Apidianakis Y, Rahme LG. Drosophila melanogaster as a model host for studying Pedomonas aeruginosa infection. Nat Protoc 2009, 4(9):1285-1294.

32. Buchon N, Bioderick NJ, Poidevin M, Pradenvand S, Lemaître B. Drosophila intestinal response to bacterial infection: activation of host defense and stem cell proliferation. Cell Host Microbe 2009, 5(2):200-213.

33. Vietsou I, Dowling AJ, Evans IR, Waterfield N, ffrench-Constant RH, Wood W. Drosophila embryos as model systems for monitoring bacterial infection in real time. PLoS Pathog 2009, 5(7):e1000518.

34. Castanguy-Vianier J, Vial L, Tremblay J, Deziel E. Drosophila melanogaster as a model host for the Burkholderia cepacia complex. PLoS One 2010, 5(7):e11607.

35. Grant J, Saldanha JW, Gould AP. A Drosophila model for primary coenzyme Q deficiency and dietary rescue in the developing nervous system. Dis Model Mech 2010, 3(11-12):799-806.

36. Melicharek DJ, Ramerez LC, Singh S, Thompson R, Marenda DR. Kismet/CHD7 regulates axon morphology, memory and locomotion in a Drosophila model of CHARGE syndrome. Hum Mol Genet 2010, 19(21):4253-4266.

37. Tatoulis RV, Fedorova ND, Jackson JD, Jacobs AR, Kinyuin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003, 4:41.

38. Nishimura M, Ocorr K, Bodmer R, Cartry J. Thyroid Extending healthy life span. BMC Biol 2010, 8:480.

39. Roeder T, Isermann K, Kabesch M. Promising? A systematic analysis of e1000374. Genetcaselm:1285-1294.

40. Yedvobnick B, Moberg K. A Drosophila model for primary coenzyme Q deficiency and dietary rescue in the developing nervous system. Dis Model Mech 2010, 3(11-12):799-806.

41. Jang AC, Starz-Gaiano M, Montell DJ.可以从 Drosophila 中分离出右向和同源基因，它们在多种基因组中的共同存在和功能。Nucleic Acids Res 2009, 37(3):149-151.

42. Reiter LT, Potocki L, Chien S, Gribskov M, Bier E. A systematic analysis of human disease orthologs and inparalogs shared by multiple proteomes. Bioinformatics 2006, 22(14):e9-15.

43. Roark M, Legleiter J, Marsh JL, Thompson LM, Lindquist S, Misfolding and reduces toxicity in Huntington disease models. Hum Mol Genet 2006, 15(18):2743-2751.

44. Pfleger CM, Wang J, Friedman L, Vittorino R, Conley LM, Ho L, Fivecoat HC, Fasini RT. Grape-seed polyphenolic extract improves the eye phenotype in a Drosophila model of tapathy. Int J Alzheimers Dis 2010, 2010.

45. Raymond FL, Tarpey P. The genetics of mental retardation. Hum Mol Genet 2006, 15(Spec No 2):R110-116.

46. O’Brien KP, Westerlund I, Sonnhammer EL. OrthoDisease: a database of human disease orthologs. Hum Mutat 2004, 24(2):112-119.

47. Reiter LT, Potocki L, Chien S, Gribskov M, Bier E. A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. Genome Res 2001, 11(6):1114-1125.

48. Chien S, Reiter LT, Bier E, Gribskov M. Homophila: human disease gene cognates in Drosophila. Nucleic Acids Res 2002, 30(1):149-151.

49. Alexeyenko A, Tamás I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 2006, 22(14):e9-15.

50. Hulsen T, Huynen MA, de Vlieg J, Groenew PM. Benchmarking ortholog identification methods using functional genomics data. Genome Biol 2006, 7(4):R31.