Homophila: human disease gene cognates in Drosophila

Samson Chien1,2, Lawrence T. Reiter2, Ethan Bier2 and Michael Gribskov1,2,*

1San Diego Supercomputer Center and 2Department of Biology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0349, USA

Received August 16, 2001; Revised and Accepted October 10, 2001

ABSTRACT

Although many human genes have been associated with genetic diseases, knowing which mutations result in disease phenotypes often does not explain the etiology of a specific disease. Drosophila melanogaster provides a powerful system in which to use genetic and molecular approaches to investigate human genetic diseases. Homophila is an inter-genomic resource linking the human and fly genomes in order to stimulate functional genomic investigations in Drosophila that address questions about genetic disease in humans. Homophila provides a comprehensive linkage between the disease genes compiled in Online Mendelian Inheritance in Man (OMIM) and the complete Drosophila genomic sequence. Homophila is a relational database that allows searching based on human disease descriptions, OMIM number, human or fly gene names, and sequence similarity, and can be accessed at http://homophila.sdsc.edu.

INTRODUCTION

The continuing progress in the sequencing of the human genome will accelerate the identification of many genes involved in human diseases. Although a map location, nucleotide sequence and even the identity of the protein involved in a specific disease may be known, it is often difficult to decipher the etiology of the disease without employing an experimental organism. One approach to deciphering the role of these genes in specific diseases is to investigate the function of cognate genes in model organisms. A number of groups (1–4) have examined various sets of genes for cognates in model organisms. A number of settings. The result of this procedure was a list of 5283 protein sequence records using the LocusLink database was required. Each Drosophila disease entry was looked up in the LocusLink mim2loc table, which relates OMIM entries to NCBI locus records. Each locus record was then used to locate the correct protein sequence entries associated with 911 OMIM disease loci and 666 matching Drosophila genes (Table 1).

The continuing progress in the sequencing of the human genome will accelerate the identification of many genes involved in human diseases. Although a map location, nucleotide sequence and even the identity of the protein involved in a specific disease may be known, it is often difficult to decipher the etiology of the disease without employing an experimental organism. One approach to deciphering the role of these genes in specific diseases is to investigate the function of cognate genes in model organisms. A number of groups (1–4) have examined various sets of genes for cognates in model organisms. A number of settings. The result of this procedure was a list of 5283 protein sequence records using the LocusLink database was required. Each Drosophila disease entry was looked up in the LocusLink mim2loc table, which relates OMIM entries to NCBI locus records. Each locus record was then used to locate the correct protein sequence entries associated with 911 OMIM disease loci and 666 matching Drosophila genes (Table 1).

Homophila integrates information from four main sources: human disease gene information from OMIM, information relating OMIM entries to specific sequences from LocusLink (7), Drosophila nucleotide and protein sequence data (8), and annotation of Drosophila genes from FlyBase (9).

Construction of Homophila began with a list of OMIM disease entries (ones that either appear in the Morbid Map or contain an allelic variant notation). Because of the narrative nature of the OMIM database, which often discusses entirely unrelated proteins that may have been excluded as the causes of the disease, it is not possible to simply look up the sequences related to each disease in OMIM. A more involved procedure relying on the NCBI LocusLink database was required. Each OMIM disease entry was looked up in the LocusLink mim2loc table, which relates OMIM entries to NCBI locus records. Each locus record was then used to locate the correct protein sequence records using the LocusLink loc2UG, loc2acc and loc2ref tables, which specify entries in the NCBI UniGene, protein and RefSeq databases, respectively.

Each of the protein sequence entries was compared to the complete Drosophila genome sequence using the BLASTP program (10). BLAST comparisons were performed using BLAST v.2.09 with the standard BLOSUM 62 and expect = 1 × 10^-10 settings. The result of this procedure was a list of 5283 protein sequence entries associated with 911 OMIM disease loci and 666 matching Drosophila genes (Table 1).

*High hit indicates a BLAST E-value of 1 × 10^-30 or lower.

cross-genomic analysis bridges the gap between the human disease and the Drosophila genome databases (6). Furthermore, this information is available online in a searchable format supported by a relational database management system (RDBMS).

DATABASE CONTENT

Homophila integrates information from four main sources: human disease gene information from OMIM, information relating OMIM entries to specific sequences from LocusLink (7), Drosophila nucleotide and protein sequence data (8), and annotation of Drosophila genes from FlyBase (9).

Table 1. Statistical summary of the information contained in the Homophila database

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of OMIM entries analyzed</td>
<td>1858</td>
</tr>
<tr>
<td>Number of OMIM entries with protein sequence</td>
<td>1224</td>
</tr>
<tr>
<td>Number of human protein sequences BLASTed</td>
<td>5283</td>
</tr>
<tr>
<td>Number of high hit OMIM entries</td>
<td>911</td>
</tr>
<tr>
<td>Number of high hit Drosophila genes</td>
<td>666</td>
</tr>
<tr>
<td>Number of Drosophila genes with P elements</td>
<td>133</td>
</tr>
</tbody>
</table>

A relational database has been implemented to allow queries on these results and is available online (http://homophila.sdsc.edu) using the MySQL RDBMS (11). PERL scripts using the DBI package are used to convert queries entered on the Homophila web pages to SQL queries to the actual RDBMS.

A complete list of P-element locations in the Drosophila genomic sequence was kindly provided by FlyBase (9). This information is added to the database searches in order to identify cognate genes for which null mutants in genes already exist (e.g. the P-element falls within the protein coding sequence of a gene) or for which it would be straightforward to generate null deletion mutations (e.g. by imprecise P-element excision).

ACCESS

Homophila is available for both browsing and searching online at http://homophila.sdsc.edu. The database content is also available in a relational version or as flat files upon request. Many OMIM disease entries have multiple protein sequences linked to the disease through LocusLink. The BLAST search results for each of the probe sequences are merged and used to create a list of best matching sequences (Fig. 1).

The precompiled list of best matches obviously gives an incomplete view of the correspondence between the gene probes for a specific disease and their Drosophila cognates. More complete information is available by directly searching the database. Searches based on OMIM entry number, human and Drosophila gene names and symbols, human disease description and text keywords are available. All entries matching the search query are displayed in a summary output (Fig. 2).

The information stored in Homophila is changing rapidly as new disease loci are sequenced. Homophila is updated approximately every 2 months using a semi-automated process to import source data and perform the requisite analyses.

FUTURE DEVELOPMENT PLANS

1. Complete automation of data update and analysis
2. Extension of analyses to other genomes: Dictyostelium, Caenorhabditis elegans, Saccharomyces cerevisiae and Mus musculus.
3. Inclusion/linkage of more complete information about human diseases and Drosophila genes so that searches based on known human disease phenotypes and Drosophila mutant phenotypes can be used to identify potentially novel functional groupings of human and fly genes.

ACKNOWLEDGEMENT

This work is supported in part by the National Institutes of Health through a National Center for Research Resources program grant (P 41 RR08605-06) to the National Biomedical Computation Resource at the San Diego Supercomputer Center.
REFERENCES


