BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 22 2011, pages 3218–3219
doi:10.1093/bioinformatics/btr530

Databases and ontologies
Advance Access publication October 5, 2011

*Dro*PhEA: *Drosophila* phenotype enrichment analysis for insect functional genomics

Meng-Pin Weng and Ben-Yang Liao*
Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County 350, Taiwan, R.O.C.

Associate Editor: Martin Bishop

ABSTRACT

Summary: *Dro*PhEA is a core module of a web application that facilitates research in insect functional genomics through enrichment analysis on mutant phenotypes of fruit fly (*Drosophila melanogaster*). The phenotypes investigated in the analyses can be predefined by FlyBase or customized by users. *Dro*PhEA allows users to specify mutation or ortholog types, displays enriched term results in a hierarchical structure and supports analyses on gene sets of all insect species with a fully sequenced genome.

Availability: http://evol.nhri.org.tw/phenome/DroPhEA/

Contact: liaoby@nhri.org.tw

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 6, 2011; revised on September 9, 2011; accepted on September 16, 2011

1 INTRODUCTION

Enrichment analysis is a promising strategy for investigators to biologically interpret gene lists obtained from high-throughput studies in the post-genomic era. Among the backend annotations employed in different enrichment tools (see review in Huang *et al.*, 2009), mutant phenotype annotation is unique because it represents the consequence of altering the information output of a gene. Therefore, annotations of phenotypes are an ideal means to aid in the understanding of how a set of genes function within the context of the whole organism.

Studies in fruit-fly biology were first documented in the early 1900s (Rubin and Lewis, 2000); over the last century, *Drosophila* has become one of the most widely used model organisms for animal genetics and developmental biology. As an insect species, in addition to basic genetics studies, *Drosophila* has served as a model to assist in agricultural and epidemiological investigations. The richness and comprehensiveness of fly omics data have been invaluable to many fields of biology. Currently, the proportion of mutated and phenotyped fly genes (36.9%; 5606 of 15 191 genes mapped to the genome) is approaching that of the house mouse (*Mus musculus*) [~40% of the genes comprising the mouse genome have been phenotyped; Weng and Liao (2010)]. However, with the exception of the house mouse (Chen *et al.*, 2009; Weng and Liao, 2010), an enrichment tool based on mutant phenotypes of another species has never been developed. Some bioinformatic tools for insect genomics, such as FlyMine (Lyne *et al.*, 2007), do not implement mutant phenotype data. Insects and mammals diverged from each other >900 million years ago (Blair *et al.*, 2005). Due to the substantial anatomical, physiological, developmental and behavioral differences between insects and mammals, it is challenging to utilize mammalian phenotypes to explore insect functional genomics.

We therefore developed *Dro*PhEA (*Drosophila* Phenotype Enrichment Analysis). Similar to *Mam*PhEA, a phenotype enrichment tool for analyses on mammalian genes (Weng and Liao, 2010), *Dro*PhEA provides several useful features. First, *Dro*PhEA allows enrichment analysis not only on phenotypes predefined by FlyBase (http://flybase.org/), but also on customized phenotypes to study complex traits. Second, different types of mutations exhibit distinct impacts on protein function; to remove potential biases caused by the use of data derived from differential mutagenesis approaches (Liao and Zhang, 2008), *Dro*PhEA enables users to perform analyses that exclude phenotypes derived from gain-of-function mutations. Third, *Dro*PhEA generates graphical and downloadable output displaying the enriched or depleted phenotypes according to the hierarchical structure of the phenotypic classification. Finally, due to the paucity of phenotypic data in insect species other than *Drosophila melanogaster*, *Dro*PhEA supports analyses of genes orthologous to *D.melanogaster* for all insect species with a fully sequenced genome (17 to date). Through integration with *Mam*PhEA, *Dro*PhEA is also capable of analyzing mammalian genes.

2 IMPLEMENTATION

Mutant fly phenotypic data were retrieved from FlyBase (version FB2011_03) (http://flybase.org/). Phenotypic entries resulting from mutations affecting multiple loci were excluded. FlyBase applies two hierarchically structured and controlled lexicons to describe phenotypes: (i) ‘Phenotypic class’ (FBcv term) represents the pathology, or the effect of the mutation on the whole organism (e.g. lethal, sterile) and (ii) ‘Anatomy’ (FBbt term) describes the body part manifested by the mutation (e.g. eye, antenna) (Drysdale, 2001). Phenotypic descriptions of a fly mutant allele are often expressed as a compound statement comprising multiple terms. In such cases, we treated each term in the compound statement individually. In fact, we found that each composite term usually contained only one ‘phenotypic class’ FBcv term or ‘anatomy‘ FBbt term, and the remaining terms were used as ‘qualifiers’ (FBcv:0000005; e.g. ‘dominant’) or ‘occurrent’ terms (FBdv:00007008; e.g. ‘pupal stage P5’), among others. Consequently, most of the compound phenotypic descriptions can

*To whom correspondence should be addressed.

simply be expressed as a ‘phenotypic class’ FBcv term, or ‘anatomy’ FBbt term for subsequent analyses. The analysis of gene sets of other insect species was supported by obtaining orthology information of each species to *D.melanogaster* from the InParanoid database (version 7.0, http://inparanoid.sbc.su.se/).

*Dro*PhEA typically compares two gene sets. When one gene set is provided, it is compared with the rest of the genes in the genome. Genes without a ‘phenotypic class’ FBcv term or ‘anatomy’ FBbt term are excluded prior to the analysis. Significantly enriched or depleted phenotypes are detected by Fisher’s exact test. *P*-values can be Bonferroni corrected for multiple tests. *Dro*PhEA allows users to customize phenotypes by combining existing FlyBase controlled lexicons by a keyword search, or by browsing FBcv or FBbt term ontologies. Enrichment analysis applying customized phenotypes gives *Dro*PhEA the capability of exploring complex traits, such as gene essentiality (see Examples 1.1 and 1.2 in Section 3.1 below).

Default *Dro*PhEA generates graphical output displaying a hierarchical structure of enriched/depleted phenotypes. *Dro*PhEA also produces classic output in a simple linear text format showing differentially enriched phenotypes. JavaServer Pages (JSP) and MySQL database on an Apache Tomcat server was used to build the web interface.

3 EXAMPLES

Two examples provided on the *Dro*PhEA website are used to illustrate the use of *Dro*PhEA in hypothesis testing and knowledge discovery.

3.1 Essentiality of evolutionarily conserved (Example 1.1) and lineage-specific genes (Example 1.2)

Genes ubiquitous in genomes across a wide range of the phylogenetic spectrum are likely to perform fundamental biological functions throughout different taxa and taxonomic levels, and are therefore expected to be more critical to individual fitness (survival or reproduction). For the fruit fly, genes vital to fitness (essential genes) can be defined by association with the lethal or sterile mutant phenotype (Liao and Zhang, 2008). Therefore, we created the ‘*essentiality*’ phenotype by combining the controlled lexicons FBcv:0000351 (lethal) and FBcv:0000364 (sterile). Consistent with our expectation, the results indicated 3151 *Drosophila* genes with one human ortholog (retrieved from BioMart version 61) exhibited a strong enrichment in the customized phenotype ‘*essentiality*’ relative to the genomic background ($P = 1.17E$-4). Similarly, we examined 7852 fly lineage-specific genes, defined as genes without an ortholog in the human genome based on the BioMart annotation. Contradictory to the results of the fly-human one-to-one orthologs, the fly lineage-specific genes were significantly depleted with the ‘*essentiality*’ phenotype ($P = 1.91E$-40). This example demonstrates that *Dro*PhEA is a promising bioinformatic tool to explore complex traits.

3.2 Genes associated with mosquito blood-feeding behavior (Example 2)

Blood-feeding behavior is an important characteristic of mosquito species. To elucidate the genetic components in the insect genome that have been acquired or are associated with the adaptation for hematophagy, we downloaded the microarray data of the blood-feeding malaria mosquito (*Anopheles gambiae*) (Marinotti *et al.*, 2005) from VectorBase (release December 2010) (http://www.vectorbase.org/). The 20% (2042/10 207) of mosquito genes shown to exhibit the highest increases in expression signals after a blood meal were analyzed with *Dro*PhEA. Consistent with our current understanding that mosquito hematophagy is required for oocyte development (Dana *et al.*, 2005; Marinotti *et al.*, 2005), the results indicated enrichment of the input gene set with several reproduction-related or cell cycle-related phenotypes; furthermore, results showed the input gene set was depleted in development, nervous system, muscle, sensory organ and behavior phenotypes (Supplementary File 1). We also conducted enrichment analyses using Example 2 gene sets on GO terms and KEGG pathways for comparisons. Significantly enriched KEGG pathways were not detected; however, many enriched/depleted GO terms in Biological Process were reported (Supplementary File 2), and notably consistent with *Dro*PhEA output. Despite the similarity in results, many enriched/depleted terms reported by *Dro*PhEA describe traits at the organismal level (e.g. sterile, viable, circadian rhythm defective, among many others), which are not included in GO or KEGG annotations. Therefore, mutant phenotypes are clearly invaluable sources of complementary data to augment GO/KEGG in bioinformatics analyses in the study of functional genomics.

4 CONCLUSION AND PERSPECTIVE

*Dro*PhEA is an online tool used to explore insect functional genomics through enrichment analysis of *D.melanogaster* phenotypes. Modules for enrichment analyses on phenotypes of model organisms other than mouse and fly will be added with increased availability and improved annotations of eukaryotic phenotypic data in the future. The backend databases of *Dro*PhEA are automatically updated every 2 months. The tutorial of *Dro*PhEA is available online at http://evol.nhri.org.tw/phenome/.

Funding: National Health Research Institutes Intramural Fund and National Science Council Grants (NSC 99-2311-B-400-003-MY2 and NSC 100-2319-B-400-001 to B.-Y.L.).

Conflict of Interest: none declared.

REFERENCES

Blair,J.E. *et al.* (2005) Evolutionary sequence analysis of complete eukaryote genomes. *BMC Bioinformatics*, **6**, 53.

Chen,J. *et al.* (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.

Dana,A.N. *et al.* (2005) Gene expression patterns associated with blood-feeding in the malaria mosquito *Anopheles gambiae*. *BMC Genomics*, **6**, 5.

Drysdale,R. (2001) Phenotypic data in FlyBase. *Brief. Bioinformatics*, **2**, 68–80.

Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Liao,B.Y. and Zhang,J. (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl Acad. Sci. USA*, **105**, 6987–6992.

Lyne,R. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, **8**, R129

Marinotti,O. *et al.* (2005) Microarray analysis of genes showing variable expression following a blood meal in *Anopheles gambiae*. *Insect Mol. Biol.*, **14**, 365–373.

Rubin,G.M. and Lewis,E.B. (2000) A brief history of Drosophila’s contributions to genome research. *Science*, **287**, 2216–2218.

Weng,M.P. and Liao,B.Y. (2010) MamPhEA: a web tool for mammalian phenotype enrichment analysis. *Bioinformatics*, **26**, 2212–2213.