BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 17 2010, pages 2212–2213
doi:10.1093/bioinformatics/btq359

Databases and ontologies

Advance Access publication July 6, 2010

MamPhEA: a web tool for mammalian phenotype enrichment analysis

Meng-Pin Weng and Ben-Yang Liao*

Division of Biostatistics & Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County 350, Taiwan, R.O.C.

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: MamPhEA is a web application dedicated to understanding functional properties of mammalian gene sets based on mouse-mutant phenotypes. It allows users to conduct enrichment analysis on predefined or user-defined phenotypes, gives users the option to specify phenotypes derived from null mutations, produces easily comprehensible results and supports analyses on genes of all mammalian species with a fully sequenced genome.

Availability: http://evol.nhri.org.tw/MamPhEA/

Contact: liaoby@nhri.org.tw

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

Received on April 27, 2010; revised on June 9, 2010; accepted on June 29, 2010

1 INTRODUCTION

To biologically interpret gene lists generated from high-throughput studies, enrichment analyses on annotations of gene features, such as terms of Gene Ontology (GO; http://www.geneontology.org), have become routine in genomic research. Numerous enrichment tools have been developed in the past few years (see review by (Huang *et al*., 2009). The implementation of novel statistics has provided new tools to deal with gene sets of special property, e.g. Gene Set Enrichment Analysis (GSEA) for genes with an ordered structure (Subramanian *et al*., 2005). However, the power of enrichment analyses still relies largely on the biological nature of backend annotations used for finding enrichment.

Most mouse (*Mus musculus*) mutants generated in laboratories have been designed to aid the study of human genetics, physiology or diseases progression. According to the current version of Mouse Genome Informatics (MGI, http://www.informatics.jax.org/; Blake *et al*., 2009), nearly 9000 genes in the mouse genome have been mutated and phenotyped in the way that closely resembles human clinical examinations. The mutant phenotypes characterize the consequence of disturbing or disrupting the information output of a gene and thus are ideal to aid understanding of how genes function at the systems level. These unique attributes make the enrichment analyses incorporating mouse-mutant phenotypes outperform those using GO or KEGG pathways (http://www.genome.jp/kegg/), the two most widely used backend databases in enrichment analyses (Huang *et al*., 2009), in human disease candidate gene prioritization (Chen *et al*., 2007) and in work to understand genetic bases of neurological disorders (Webber *et al*., 2009). These cases clearly suggest a pressing need for a versatile enrichment tool from the aspect of mutant phenotypes in studying mammalian biology.

We have therefore developed MamPhEA (Mammalian Phenotype Enrichment Analysis). In comparison to ToppGene (Chen *et al.*, 2009), the only other enrichment tool employing mouse phenotypes, MamPhEA has several unique features. First, MamPhEA allows users to perform enrichment analysis not only on phenotypes predefined by MGI, but also on user-defined phenotypes to study complex traits. Second, different types of mutations impact protein functions distinctly; to remove potential biases caused by the use of data derived from differential mutagenesis approaches (Liao and Zhang, 2008), MamPhEA allows users to perform analysis of phenotypes exclusively derived from loss-of-function mutations. Third, in order to give users easily understood results, MamPhEA generates graphical and downloadable output displaying the enriched or depleted phenotypes according to the hierarchical structure of the phenotypic classification. And, finally, MamPhEA supports analyses of genes of all mammalian species with a fully sequenced genome (35 to date).

2 IMPLEMENTATION

The core phenotype data, including phenotype descriptions of each mouse mutant strain (represented as MP IDs) and Mammalian Phenotype Ontology (MPO), were retrieved from MGI. The MP IDs are hierarchically structured; thus, if a mouse gene is annotated for an MP ID, it is assigned to all the parental MP IDs. Data of the mouse strains with multiple mutated genes were excluded. Only phenotypes of mouse strains with homozygous mutation are considered. Loss-of-function phenotypes, here, are defined as phenotypes resulted from mutations generated by targeted gene deletions, random gene disruptions or gene trapping (Liao and Zhang, 2007). Finally, 8824 genes with MP IDs and 5982 genes with loss-of-function MP IDs were identified. To support analysis on other mammalian genomes, we obtained the orthology information of mammalian genes from Ensembl (http://www.ensembl.org/). MamPhEA accepts all gene identifiers cross-referenced by Ensembl (e.g. RefSeq ID, Ensembl ID and Affymatrix ID).

MamPhEA typically compares two gene sets. When one gene set is given, it is compared to the rest of the genes in the genome. Users need to specify the scale of phenotypes to be examined. Significantly enriched or depleted phenotypes are detected by Fisher's exact test. *P*-values are Bonferroni corrected for multiple tests. One novel feature of MamPhEA is allowing users to customize phenotypes

*To whom correspondence should be addressed.

of their own interest by combining existing MP IDs. Potentially relevant MP IDs can be searched by keyword or by browsing MPO. The option of customizing phenotypes broadens the applicability of MamPhEA in studying complex traits or diseases, such as gene essentiality (see Example in Section 3.2).

MamPhEA generates graphical output displaying enriched or depleted phenotypes that are hierarchically structured according to MPO. MamPhEA also produces classic output in a simple linear text format showing differentially enriched MP IDs. The Web interface of MamPhEA is built by JavaServer Pages (JSP) and MySQL database on an Apache Tomcat server.

3 EXAMPLES

We provide two examples to illustrate the use of MamPhEA in hypothesis testing and knowledge discovery.

3.1 Tissue expression and gene function (Example 1 on the MamPhEA home page)

Genes have to be expressed to function in the cell; thus, it is expected that deleting genes highly expressed in a certain tissue from the genome will lead to dysfunction of tissues where the deleted genes are expressed in. We compared human genes highly expressed in liver (gene set 1) with those highly expressed in testis (gene set 2) (Supplementary Dataset 1). Highly expressed genes are defined as genes with top 20% expression level genome widely in a tissue, based on the human RNAseq data (Wang *et al*., 2008). As expected, liver-expressed genes (gene set 1) are enriched with phenotypes of abnormal homeostasis, abnormal metabolism and abnormal morphology or physiology of liver/biliary system (Supplementary Figs 1 and 2). Therefore, contrary to the assertion of the neutral model of transcriptome evolution (Khaitovich *et al*., 2004; Yanai *et al*., 2004), we confirm that expression of a gene is an important component of gene function and should be targeted by natural selection during evolution (Liao and Zhang, 2006).

3.2 Expression level and gene essentiality (Example 2 on the MamPhEA home page)

A gene is said to be essential if its deletion would render the fitness of the organism zero. Accordingly, the phenotype 'essentiality' of mammals can be defined as premature death or infertility of the mutant and can be customized through merging existing MGI-predefined MP IDs (MP:0002080 prenatal lethality, MP:0002081 perinatal lethality, MP:0002082 post-natal lethality, MP:0002160 abnormal reproductive system morphology and MP:0001919 abnormal reproductive system physiology) (Supplementary Fig. 3) on the basis of loss-of-function mutations (Liao and Zhang, 2007). It has been found that highly expressed genes tend to be more essential than lowly expressed genes in several prokaryotes and eukaryotes (Hannay *et al*., 2008). Using MamPhEA, we examined whether this is also true for the mouse. Mouse gene list of high expression levels (top 20%, gene set 1) and that of low expression levels (bottom 20%, gene set 2) were selected according to RNAseq data from the mouse brain (Mortazavi *et al*., 2008) (Supplementary Dataset 2), and were fed to MamPhEA to find differential enrichment of the customized phenotype 'essentiality'. Consistent with previous findings in other organisms (Hannay *et al*., 2008), highly expressed genes were found to be more essential (609/1197 = 51% essential genes) than lowly expressed genes (273/718 = 38% essential genes) for the mouse ($P = 4.95E-8$) (Supplementary Fig. 4).

4 CONCLUSION AND PERSPECTIVE

MamPhEA is designed to perform enrichment analysis on mammalian mutant phenotypes. The international genetics community has initiated the Knockout Mouse Project to individually delete and phenotype every gene in the mouse genome (Austin *et al*., 2004). With the rapid growth of mouse phenotype data in the near future, the power of MamPhEA is expected to continuously increase. We update the databases used to build MamPhEA regularly. The online tutorial of MamPhEA is available at http://evol.nhri.org.tw/MamPhEA/tutorial.html.

ACKNOWLEDGEMENTS

We thank Chung-Yen Lin and Chieh-Hua Lin for valuable comments on the user interface.

Funding: Intramural funding from the National Health Research Institutes (to B.-Y. L.).

Conflict of Interest: none declared.

REFERENCES

Austin,C.P. *et al.* (2004) The knockout mouse project. *Nat. Genet.*, **36**, 921–924.

Blake,J.A. *et al.* (2009) The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res.*, **37**, D712–D719.

Chen,J. *et al.* (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.

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

Hannay,K. *et al.* (2008) Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation. *BMC Genomics*, **9**, 609.

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.

Khaitovich,P. *et al.* (2004) A neutral model of transcriptome evolution. *PLoS Biol.*, **2**, 682–689.

Liao,B.Y. and Zhang,J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.*, **23**, 530–540.

Liao,B.Y. and Zhang,J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.*, **23**, 378–381.

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.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Webber,C. *et al.* (2009) Forging links between human mental retardation-associated CNVs and mouse gene knockout models. *PLoS Genet.*, **5**, e1000531.

Yanai,I. *et al.* (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS*, **8**, 15–24.