modPhEA: model organism Phenotype Enrichment Analysis of eukaryotic gene sets

Meng-Pin Weng and Ben-Yang Liao*

Institute of Population Health Sciences, National Health Research Institutes, Miaoli County 350, Taiwan, R.O.C.

*to whom correspondence should be addressed

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY RESULTS

The results of six groups of Example Datasets are described and discussed below. Example Dataset 1 includes human genes that exhibit expression changes in response to Zika virus (ZIKV) infection. The analyses of Examples 1.1 and 1.2 are based on human phenotypes and mouse phenotypes, respectively. Example Datasets 2.1 and 2.2 include genes of budding yeast and fission yeast (Schizosaccharomyces pombe), respectively, that are involved in the lineage-specific duplication events occurred after the divergence of bud-ting and fission yeasts. Example Dataset 3 includes zebrafish orthologs that are expressed at lower levels in the eyes of the cave-dwelling Sinocyclocheilus species compared to the surface-dwelling Sinocyclocheilus species. The analyses of Examples 3.1 and 3.2 are based on knockdown-derived phenotypes and mutagenesis-derived phenotypes, respectively. Example Dataset 4 includes C. elegans genes that are expressed in gonads. Example Dataset 5 includes genes that are highly expressed after a blood-meal by a malaria vector mosquito and Example Dataset 6 includes genes that are highly expressed in vascular smooth muscle cells of patients with giant cell arteritis (GCA). To allow users to compare the results obtained from modPhEA with the enrichment analysis on Gene Ontology (GO) terms or KEGG pathways, we have additionally conducted enrichment analyses on GO terms and KEGG pathways. Please click this hyperlink to see the enriched GO terms or KEGG pathways that were found for each Example Dataset.

Example Dataset 1: Human genes exhibiting expression changes in response to Zika virus (ZIKV) infection

A previous ZIKV outbreak in Brazil was associated with a marked increase in the number of infants born with microcephaly (Mlakar, et al., 2016). Although it is known that ZIKV can be transmitted from mother to child (Mysorekar and Diamond, 2016), it
was unclear whether ZIKV directly causes microcephaly (Mlakar, et al., 2016; Pacheco, et al., 2016; Rasmussen, et al., 2016). This gene set was defined based on differentially expressed genes that were identified by Tang, et al. (2016) between Zika virus-infected (ZIKV) and mock-infected human embryonic cortical neural progenitor cells (hNPCs). Following an analysis of this gene set by modPhEA (Example 1.1), several human phenotypes were enriched, including those related to cranial development (e.g., “Abnormality of skull size” [HP:0000240, FDR-corrected \( P < 10^{-7} \)], “Abnormality of the calvaria” [HP:0002683, FDR-corrected \( P = 0.018 \)], etc.) and brain development (e.g., “Abnormality of brain morphology” [HP:0012443, FDR-corrected \( P < 10^{-8} \]), “Microcephaly” [HP:0000252, FDR-corrected \( P = 0.007 \]), etc.). Consistent with these findings based on human phenotypes, an analysis of mouse phenotypes (Example 1.2) also showed enriched phenotypes of “abnormal cranium size” (MP:0010031, FDR-corrected \( P = 0.019 \), “abnormal brain development” (MP:0000913, FDR-corrected \( P < 10^{-4} \)), etc. These results suggest a molecular basis and a direct contribution of ZIKV infection to microcephaly.

It should be noted that phenotypes related to abnormal limb development were also identified in both analyses based on human phenotypes (e.g., “Abnormal appendicular skeleton morphology” [HP:0011844, FDR-corrected \( P = 0.045 \); Example 1.1]. Intriguingly, deformed limbs have recently been reported in newborns infected by ZIKV (van der Linden, et al., 2016). Thus, the enriched phenotypes reported by modPhEA could be biomedically important, and can guide the design of experiments aiming to understand human diseases.

**Example Datasets 2:** Budding yeast genes (Example Dataset 2.1) and fission yeast (*Schizosaccharomyces pombe*) genes (Example Dataset 2.2) arising from lineage-specific duplication events after the divergence of budding and fission yeasts.

Gene duplication has been proposed to be an important mechanism underlying organismal adaptation (Ohno, 1970; Zhang, 2003), including that of yeasts (Qian and Zhang, 2014). Based on orthology information annotated by Ensembl Fungi v29 (http://fungi.ensembl.org), a list of paralogs in the budding yeast genome and a list of paralogs in the fission yeast genome that arose after the latest common ancestor of these two yeast species were obtained. Enriched analyses of these two lists of genes (paralogs), defined as Example Datasets 2.1 and 2.2, respectively, were conducted based on the budding yeast phenotypes. Budding yeast can thrive under strictly anaerobic conditions, while fission yeast cannot (Heslot, et al., 1970). Correspondingly, the enriched phenotypic term, “anaerobic metabolism (APO:0000210, FDR-corrected \( P = 0.003 \))”, was identified for duplicated genes in the budding yeast lineage (Example
Datasets 2.1). In addition, the enriched term, “resistance to chemicals” (APO:0000087, FDR-corrected $P < 10^{-2}$), that was identified in this gene set may be related to the more advanced metabolic capacity of budding yeast in fermentation, during which yeast cells are exposed to various stresses. The analysis of genes duplicated in the fission yeast lineage (Example Dataset 2.2) identified “interaction with host/environment” (APO:0000287, FDR-corrected $P = 0.021$), “septum formation” (APO:0000221, FDR-corrected $P = 0.024$) and “position of spindle pole body” (APO:0000214, FDR-corrected $P = 0.031$). The reported phenotype of “interaction with host/environment” (APO:0000287) is consistent with the lineage-specific adaptation of fission yeast that includes the survival of fission yeast as spores in the gut of insect vectors (Coluccio, et al., 2008). The other two terms, “septum formation” (APO:0000221) and “position of spindle pole body” (APO:0000214), may be related with the “binary fission” form of asexual reproduction that characterize S. pombe.

**Example Dataset 3:** zebrafish orthologs showing reduced expression in eyes of cave-dwelling Sinocyclocheilus species

Normal eyes characterize the surface-dwelling Sinocyclocheilus (Cypriniformes: Cyprinidae) teleost fish species (S. angustiporus), while the cave-dwelling Sinocyclocheilus species (S. anophthalmus) has small eyes that are buried deeply within adipose tissue that is covered with skin. To understand the molecular basis of these differences, whole-eye transcriptomes for both S. angustiporus and S. anophthalmus were profiled in Meng, et al. (2013). Genes with an RNA-seq signal that was decreased by > 50% in the eyes of S. anophthalmus compared with the eyes of S. anophthalmus were identified and defined as this gene set. When an enrichment analysis of zebrafish knockdown-derived phenotypes was performed with modPhEA (Example 3.1), “eye photoreceptor cell” (ZFA:0009154, FDR-corrected $P < 10^{-3}$) and “retinal cone cell” (ZFA:0009262, FDR-corrected $P < 10^{-3}$) were identified. When an enrichment analysis of zebrafish mutagenesis-derived phenotypes was performed (Example 3.2), “retinal outer plexiform layer” (ZFA:0001330, FDR-corrected $P = 0.027$) was identified. The enriched terms found in Examples 3.1 and 3.2 were not overlapped, but were all related with vision and are consistent with the reduced retinal cell density and photoreceptor cell height that histologically characterize S. anophthalmus eyes according to Meng, et al. (2013). The results of Example 3.1 and 3.2 indicate that analyses focusing on phenotypic data derived from different approaches could produce results that are complement with each other.

In addition to the abovementioned terms, this gene set of reduced mRNA
expression is enriched in knockdown-derived phenotypes (see the result of Example 3.1) manifested in "epiphysis" (ZFA:0000019, FDR-corrected \( P = 0.006 \)), a circadian clock pace maker that contains photoreceptor cells. Interestingly, a recent study comparing genomes of three Sinocyclocheilus species found that Skp1-Cul1-Fbxl3 (SCF) protein complex, the most relevant in clock mechanism in mammals, has been degenerated in S. anophthalmus (Yang, et al., 2016). In the future, it will be interesting to examine if S. anophthalmus lacks circadian rhythms.

**Example Dataset 4: C. elegans genes expressed in gonads**

The transcriptomes and proteomes of fruit fly and nematode were simultaneously compared in a previous study, and the results suggest that the evolution of the transcriptome is largely neutral (Schrimpf, et al., 2009). That is, protein expression appears to be largely controlled at the level of protein translation, and mRNA expression signals do not necessarily predict gene functions. When an enrichment analysis was performed for genes expressed in an isolated gonad of C. elegans as determined by RNA-seq (Ortiz, et al., 2014), many reproductive phenotypes were identified, including: gonad morphology (e.g., “gonad morphology variant” [WBPhenotype:0001355, FDR-corrected \( P < 10^{-28} \)]) and fertility (e.g., “fertility reduced” [WBPhenotype:0001384], FDR-corrected \( P < 10^{-90} \)). Therefore, in contrast with the neutral model of transcriptome evolution that has been proposed (Khaitovich, et al., 2004; Schrimpf, et al., 2009), our results from modPhEA indicate that mRNA expression levels strongly influence functions of the tissue in which the genes are expressed.

**Example Dataset 5: Genes that are highly expressed after a blood-meal by malaria vector mosquito.**

Blood-feeding behavior is an important characteristic of mosquitos. To elucidate the genetic components associated with the hematophagy of these animals, microarray data for the malaria mosquito (Anopheles gambiae) (Marinotti, et al., 2005) were downloaded. The top 20% of the mosquito genes that exhibited the greatest increases in gene expression after a blood meal were defined. The remaining 80% of the A. gambiae genes were used as background data for performing an enrichment analysis of fruit fly phenotypes. Consistent with our current understanding that mosquito hematophagy is required for oocyte development, the results indicated that enrichment of this gene set included several reproduction-related phenotypes (e.g., “female sterile” [FBcv:0000366, FDR-corrected \( P < 10^{-8} \)]) and cell cycle-related.
phenotypes (e.g., “cell cycle defective” [FBcv:0000671, FDR-corrected \( P < 10^{-8} \)])). This example dataset and its results have been presented and discussed in our previous study (Weng and Liao, 2011).

**Example Dataset 6:** Highly expressed human genes in vascular smooth muscle cells of patients with giant cell arteritis (GCA).

Human diseases can be characterized by phenotypic abnormalities described with human phenotype ontology (HPO) terms. For example, GCA has been described by Groza, et al. (2015) with the human phenotypic terms, vasculitis (HP:0002633), granulomatosis (HP:0002955), amaurosis fugax (HP:0100576), facial palsy (HP: 0010628), renal amyloidosis (HP: 0001917), dysphagia (HP: 0002015), trismus (HP: 0000211), and encephalopathy (HP: 0001298). Accordingly, we customized a phenotype, “giant cell arteritis (GCA),” by combining the above seven HPO terms. The gene set provided in this example contains genes with a processed expression signal > 8 in at least one of the vascular smooth muscle cell samples analyzed from GCA patients (downloaded from GSE63425 of NCBI GEO). The analysis conducted by modPhEA showed that this gene set is enriched with the customized term, “giant cell arteritis (GCA)” \( (P < 10^{-4}) \), demonstrating the capability of modPhEA in investigating complex traits/diseases.

**SUPPLEMENTARY REFERENCES**


