Phenotype Enrichment Discovers Phenologs for Disease Modeling in *C. elegans*

David Angeles-Albores[1,†] Paul W. Sternberg[1,*]

January 4, 2017

Introduction

The last decade has seen an explosion of techniques capable of genome-wide measurements. Some examples of genome-wide tools include RNA-seq [] to measure gene expression, CHIP-seq [] to measure binding, or genome-wide methylation profiling to understand epigenetic changes. These tools are capable of generating large quantities of data. Understanding this data, and generating hypotheses from it remains challenging. A common approach used to understand these datasets is to reduce the dimensionality of the data via enrichment analyses of ontologies, which helps researchers understand what terms are overrepresented beyond random levels. By analyzing overrepresented terms in aggregate, researchers can better understand what biological processes were most affected in a given experiment, and form hypotheses about what is happening []. This approach is limited by what ontologies can be tested for enrichment. In *C. elegans*, only the gene ontology (GO) and tissue ontology are available for testing [] even though curated gene, tissue and phenotype ontologies exist. Another limitation is that tissue enrichment testing is not offered on the same websites as GO enrichment testing, which requires users to test their data on different websites that may or may not use different methodologies to detect enrichment.

Another way to use enrichment tools is for evolutionary comparison purposes. In molecular biology it is often useful to know when a gene is homologous between two species—that is to say, common by descent—because knowledge of homology often brings with it knowledge of function []. Indeed, many important gene regulatory networks (GRNs) are conserved between organisms as highly diverged as nematodes and humans (for example, see []). While genes and GRNs may be conserved between species, their outputs often differ however. For example, although homeobox genes are known to be important for limb formation in many animals [], these genes do not form limbs in nematodes. The concept of a phenolog has been put forward to explain relationships between phenotypes that have the same underlying genetic regulatory network []. Formally, two phenotypes are phenologs of each other if the homologs of the genes that cause a phenotype in an organism cause a second phenotype in another.

In order to study a clinically relevant disease, an appropriate model has to be established. A straightforward method towards establishing a disease model in *C. elegans* is to link a disease to a causal gene, then to identify the homologous gene in *C. elegans* and then to study the function of the genetic homolog to extrapolate back to humans. However, this method relies on the existence of known disease genes and requires that the homolog have a phenotype that can be reliably identified and studied. A fundamentally different way to establish a disease model in *C. elegans* would be to identify the phenologs of the disease to be studied in *C. elegans*, and to use that phenolog as the basis for screens to identify genes that are associated with that phenolog. This approach has been successfully used in the past to make non-obvious links between phenotypes in different species [].

In order to facilitate understanding of large datasets, and to make discovery of phenologs easier, we have developed a complete enrichment tool suite in WormBase that allows users to rapidly perform all analyses on curated *C. elegans* ontologies using the same methodology for each one. They are located at We applied our tools towards the unbiased discovery of phenologues of multigenic, complex diseases including schizophrenia, type-2 diabetes, crohn's disease and prostate cancer by using genes associated with these diseases via genome-wide association studies. We also illustrate the utility of the complete enrichment suite for finding new relationships in complex data by analyzing a ciliary neuron transcriptome [].

Methods

Results

Developing the WormBase Enrichment Suite

We developed the dictionaries for PEA and GEA using the same procedure as was used for TEA []. We generated a dictionary that included terms with at least 50 annotating genes or more and had a similarity threshold of 0.95 for PEA (the total number of terms in the dictionary is XXX); and we generated a dictionary that included terms with at least 100 annotating genes or more and had a similarity threshold of 0.95 for GEA (the total number of terms in this dictionary is XXX). Next, we benchmarked the dictionaries on the same gene sets as TEA and obtained enrichment of all the expected categories. For example, on a gene set enriched for embryonic muscle genes [], the top two enriched phenotype terms by q-value were 'muscle system morphology variant' and 'body wall muscle thick filament variant'; the top two enriched GO terms were 'developmental process' and 'contractile fiber'. For all the benchmarking results, see supplementary information. Having generated and validated our dictionaries, we proceeded to identify phenologs for several common human diseases.

Applying the WormBase Enrichment Suite

In order to discover phenologs, we first needed to identify genes that contribute to a disease in an unbiased manner. One way to discover gene associations in an unbiased manner is to perform genome-wide association studies (GWAS) in human populations. Therefore, we used the GWAS NHGRI-EBI Catalog [] to identify genes associated with human diseases. We found the best nematode candidate homologs for these genes using DIOPT [] and applied our enrichment suite to each of these gene regulatory networks.

Obesity-related traits

Human genes in Obesity-related traits: 957
Worm genes in Obesity-related traits: 614

Obesity-related traits is a category within the GWAS NHGRI-EBI catalog that pools studies that have measured obesity and other traits associated with obesity, such as heart rate, physical activity, hormone levels, body composition and cholesterol levels. Since this category includes many parameters, we expected there would be many phenologs. GWAS studies have identified 957 genes associated with these traits. Using DIOPT, we found 614 homologs for these genes, of which XXX had a phenotype.

Top results for obesity-related traits included 'acetylcholinesterase inhibitor response variant', 'behavior variant', 'neurite morphology variant' and 'thin'. Terms involving locomotion were significantly enriched, as were terms involving body shape and food consumption ($q < 0.1$). Concomitant with these phenologs was an enrichment in tissues including the *C. elegans* 'tail', 'sex organ', and an eclectic collection of neuronal tissues and cells.

Schizophrenia

Human genes in schizophrenia: 899
Worm genes in schizophrenia: 433

Schizophrenia results included 'acetylcholinesterase inhibitor response variant', multiple terms involving vulval formation and induction, cell division defects (‘sister chromatid segregation defective early embryo’, ‘cytokinesis variant’) and terms involving lethality or arrest at multiple stages. Tissue enrichment showed that the anal depressor and sphincter muscles were significantly overrepresented, as were neuronal tissues. Somatic gonad, including the spermatheca and the distal tip cell were also significantly overrepresented in this dataset.

Type-2 Diabetes

Human genes in diabetes: 396
Worm genes in diabetes: 224

The phenolog results for type-2 diabetes included ‘somatic gonad morphology variant’, and ‘Q neuroblast lineage migration variant’.

Crohn’s Disease

Human genes in Crohn’s Disease: 390
Worm genes in Crohn’s Disease: 213

Crohn’s disease in the worm presents traits including ‘brood size variants’, ‘P granule defective’, ‘cell fate transformation’, ‘male mating defective’ as well as various defects involving the *C. elegans* gonad. Reflecting this, TEA showed that Crohn’s associated genes in the worm were enriched in the reproductive tract and somatic gonad, as well as the dorsal, ventral and lateral nerve cords and the nerve ring.

Prostate Cancer

Human genes in prostate cancer: 273
Worm genes in prostate cancer: 273

Genes associated with prostate cancer in nematodes were associated with cell fate transformations, adult lethal phenotypes and P granule localization defects, and enriched tissues included the nerve cords, the sex organs and the tail of the animal.

Ontology Enrichment as an aid for Screen Design

An additional use for a tool like PEA would be as a tool to help guide and design screens to identify genes from an RNA-seq or other genome-wide experiment for further study. This would be particularly useful in cases when researchers may not know what phenotype to expect, in which case PEA can guide selection of a phenotype. Another use case is a scenario where the phenotype under study is not easy to screen for. By finding phenologs to the phenotype of interest, the researcher can design an easier screen for genes that affect the phenolog in question, then re-test genes for the original phenotype of interest.

Enrichment in the Ciliary Neuronal Transcriptome

As an example of how ontology enrichment can improve our understanding of transcriptomes, we selected a ciliary neuron dataset [] and ran the complete WormBase Enrichment Suite on it. Ciliary neurons are present in the *C. elegans* male tail, but they are also present in the male cephalic sensillum and hermaphrodites also have ciliated neurons. The results are illuminating. PEA reveals that the ciliated neuron transcriptome is enriched for genes that are typically associated with chromosome segregation, aneuploidy and spindle defects. This makes sense—cilia employ microtubules for structural integrity, and microtubules are required for chromosome segregation. Whereas PEA suggests that this dataset is enriched in microtubule defects, TEA points at the *C. elegans* gonad primordium, somatic gonad and germline precursors as the sites where genes associated with ciliary neurons are enriched. *C. elegans* has a stereotyped cell lineage [], and no cellular division happens after a certain point in post-embryonic development, with the exception of the germline.

Conclusions