Phenotype and gene ontology enrichment as guides for disease modeling in *C. elegans*

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

January 10, 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 these data, and generating hypotheses from them 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. The best-known ontology for biological research is the Gene Ontology (GO), which provides a controlled language to describe molecular and cellular functions of genes []. In *C. elegans*, curated tissue and phenotype ontologies exist which provide controlled languages with which to describe *C. elegans* anatomy and phenotypes respectively []. However, enrichment tools only exist for gene and tissue ontologies in the community today (see for example []). 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.

To study a clinically relevant disease in a non-human, 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* by identifying disease-associated human genes in an

unbiased manner through genome-wide association studies (GWAS) and identified candidate homolog genes in *C. elegans*. The homologs can be used to identify *C. elegans* disease phenologs, which can in turn be used as the basis for screens to identify genes that are associated with that phenolog. Approaches similar to this have been successfully used in the past to make non-obvious links between phenotypes in different species [].

The concept of a phenolog can also be useful when applied within a species. In *C. elegans*, not all phenotypes are equally easy to study. Although genome-wide measurements can help elucidate the genetic network underlying a phenotype, devising screens to test which genes are functionally important can be difficult. A common strategy to study phenotypes that are difficult to screen is to select an easier-to-screen phenolog, and to test positive hits for the true phenotype of interest afterwards []. Currently, selection of screening phenotypes is performed based on researcher experience. By formalizing phenotype enrichment analysis as a tool with which to analyze gene sets, researchers should be able to formally establish phenologs, which has consequences for screen design.

An additional problem with genome-wide queries of *C. elegans* states (be they developmental, neuronal, or other) is that they do not always have a straightforward interpretation in terms of phenotypes. In these situations, researchers must rely on intuition to select a phenotype to screen for. As a result, many hits may go unexplored that would prove fruitful. The question of how to design a screen that is maximally informative is an important question that has so far not been addressed within this community.

To facilitate understanding of large datasets, and to make discovery of phenologs easier, we have completed an enrichment tool suite in WormBase that allows users to rapidly perform phenotype, tissue and gene ontology enrichment analyses (PEA, TEA and GEA respectively) 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 systemic lupus erythematosus, obesity and obesity-related traits, and rheumatoid arthritis 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

Human disease phenolog identification

We used the GWAS EBI-NHGRI catalog [] to extract information on all genome-wide association studies deposited there. We only selected traits that had > 300 associated genes. We identified 24 traits that met our criteria. Next, we used DIOPT [] to identify candidate homologs for the genes associated with these traits. Briefly, DIOPT combines a large number of methods for identifying homologs and returns homolog candidates associated with a compound score. Depending on the score, homologs can be considered 'high', 'moderate' or 'low' rank, reflecting confidence in the homology. Many-to-one and one-to-many homology relationships are allowed in DIOPT, reflecting a mixture of uncertainty and family expansion/reduction. For our study, we only accepted homolog candidates with 'high' or 'moderate' scores and we did not insist on a one-to-one relationship between genes.

After we identified worm homologs for each trait, we reassessed how many traits still had > 100 gene candidates, and dropped all traits that had less than this for our analysis. We identified 18 traits that met this criteria. The gene lists for each of these 18 traits were then analyzed for gene, tissue and phenotype enrichment. Tissue enrichment was performed using the WormBase Tissue Enrichment Analysis tool (TEA) [].

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 was 251, annotated by 9,169 genes for the version XXXXX); and we generated a dictionary that included terms with at least 50 annotating genes or more and had a similarity threshold of 0.95 for GEA (the total number of terms in this dictionary is 271, annotated by 14,636 genes for the version XXXX). 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 'myofibril' and 'striated muscle dense body'. 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

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 GWSA 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

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. In total, 341/614 genes had at least one phenotype annotation; 548/614 had at least one gene ontology term annotation; and 427/614 had at least one tissue term annotation.

Top results for obesity-related traits included 'acetylcholinesterase inhibitor response variant' (38 genes, $q < 10^{-6}$), 'neurite morphology variant' (21 genes, $q < 10^{-2}$), and 'thin' (31 genes, $q < 10^{-2}$). Terms involving locomotion were significantly enriched, as were terms involving body shape and food consumption ($q < 10^{-1}$). Concomitant with these phenologs was a tissue enrichment in neuron-related terms. GO enrichment suggested that these genes are participating in 'iron ion binding' (40 genes, $q < 10^{-20}$) and 'tetrapyrrole binding' (37 genes, $q < 10^{-13}$).

Tissue and phenotype enrichment therefore suggest that obesity-related traits may be studied in *C. elegans* through neuron physiology and function, specifically with respect to acetylcholinesterase inhibitors. Moreover, GO enrichment implicates iron and tetrapyrrole binding as metabolic components of the obesity-related phenologs in *C. elegans*.

Systemic lupus erythematosus

Systemic lupus is an autoimmune disease that is believed to be polygenic in nature []. It mainly affects women and is characterized by painful and swollen joints, hair loss, and fatigue []. Since worms do not have a cellular immune system, we were interested in what phenologs corresponded to this disorder in *C. elegans*. To establish phenolog candidates, we obtained 283 genes associated with the disease via GWAS studies, and found 135 homolog candidates in *C. elegans*.

Lupus-associated homologs were reasonably well annotated. Slightly more than half of the genes had at least one phenotype annotation (76/135) and almost all genes were annotated to at least one tissue or gene ontology term (104/135 and 115/135 genes respectively). We found that Lupus-associated homologs were enriched in 'aneuploidy' (7 genes, $q < 10^{-1}$) and 'meiotic chromosome segregation' (8 genes, $q < 10^{-1}$). 'Cell fate transformation'

(6 genes, $q < 10^{-1}$), and 'excess intestinal cells' (5 genes, $q < 10^{-1}$) were also overrepresented, as was 'male tail morphology' (6 genes, $q < 10^{-1}$). Finally, the phenotype 'nonsense mRNA accumulation' was also enriched (5 genes, $q < 10^{-1}$). Meanwhile, TEA suggested that the 'excretory duct cell' (5 genes, $q < 10^{-2}$) and the 'posterior gonad arm' are overrepresented in this dataset. We also found that the Pn.p cells P3.p through P8.p were enriched in this dataset (5 genes, $q < 10^{-1}$). GO enrichment pointed at 'modification-dependent macromolecule catabolic process' (23 genes, $q < 10^{-15}$) as a molecular function that characterizes this dataset. However, this GO term was enriched only due to a single gene family, the *skr* gene family. Almost the entire *skr* family was considered a candidate homolog to the SKP1 human gene, making the GO enrichment suspect.

Enrichment of the terms for 'aneuploidy', 'meiotic chromosome segregation', and 'excess intestinal cells' were largely driven by the same gene group, which includes *cki-1*, and several *skr* genes. On the other hand, 'cell fate transformation' and 'male tail morphology' reflected the involvement of developmental genes *let-23*, and *lin-12* among others. The term 'nonsense mRNA accumulation' was the result of *pept-3*, *smg-7*, *tsr-1*, *dhcr-7* and *F08B4.7*. Therefore, we conclude that systemic lupus erythematosus is potentially represented by a combination of three phenotypes in *C. elegans*: A cell proliferation phenotype (either increased or decreased), probably marked by increased aneuploidy; a developmental phenotype involving cell fate transformation and leading to dysmorphias; and a molecular phenotype involving impairment of the nonsense-mediated decay pathway. The results from the tissue enrichment analysis highlighted three tissues that are particularly sensitive to *lin* mutations (the gonad, the excretory duct cell and the vulval precursor cells), and the gonad arms undergo large quantities of nuclear proliferation.

Rheumatoid arthritis

Rheumatoid arthritis is an auto-immune disease that is characterized by swollen and painful joints that progressively deteriorate []. Unlike lupus, rheumatoid arthritis is not life-threatening [], and comorbidity between rheumatoid arthritis and lupus is low [], suggesting that they may have at least partially distinct genetic causes. We found 309 genes associated with rheumatoid arthritis, for which we found 124 worm homolog candidates.

The only phenotype that was enriched for these homologs was 'short' (10 genes, $q < 10^{-4}$), even though 64 homologs were associated with at least one phenotype term. No tissue was enriched in this dataset. Because 82 genes are annotated to have expression in at least one tissue, the lack of enrichment does not reflect ignorance about the sites of expression of these genes. GEA showed that enriched molecular functions for these genes include 'collagen trimer' (22 genes, $q < 10^{-15}$). However, this term was enriched as the result of degeneracy in the homolog candidates for the SFTPD gene. Other terms included 'glycosylation' (10 genes, $q < 10^{-4}$) and 'Golgi apparatus' (11 genes, $q < 10^{-3}$), but this enrichment was also the result of degenerate homolog candidates for the human gene B3GNT7 which encodes a beta-1,3-N-acetylgalactosaminyltransferase.

The 'short' phenotype was the result of the *cat-4*, *dpy-7*, *rnt-1*, *sem-4*, *unc-116*, *ocrl-1* and some genes in the *fat* family. Although these genes are bound by a common phenotype, any genetic relationships between these genes are not immediately clear. Some genes, like *sem-4* and *rnt-1* are likely transcription factors with roles in development (including hypodermal development). Others are molecular motors (*unc-116*) that are broadly expressed throughout the body of *C. elegans*. Yet others have known roles in neuron and muscle function, such as *ocrl-1* and *cat-4*. The 'short' phenotype is a subset of the 'body length variant' phenotype. Body length in *C. elegans* can be controlled via cell size, shape and number []; alternatively, cuticle development can alter body shape []; finally, muscles can alter the effective body length artificially [].

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 exper-

iment 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. PEA reveals that the ciliated neuron transcriptome is enriched for genes that are typically associated with 'meiotic chromosome segregation' (46 genes, $q < 10^{-5}$), 'aneuploidy' (42 genes, $q < 10^{-5}$) and 'spindle defective early embryos' (45 genes, $q < 10^{-2}$).

In addition, TEA points at the *C. elegans* gonad primordium, the somatic gonad and early embryonic cells as the sites where genes associated with ciliary neurons are enriched. The 'male distal tip cell' is a tissue that is overrepresented in this dataset, but 'distal tip cell' is not enriched, which suggests that structures that are present only in the male tissue are overrepresented in this dataset.

Although one interpretation of the results would be that microtubule genes are driving the enrichment of these terms, another possibility is that there are cell-cycle genes that are driving the enrichment of these phenotypes and tissues. Indeed, GO enrichment shows terms such as 'DNA replication' (29 genes, $q < 10^{-5}$), and 'purine NTP-dependent helicase activity' (15 genes, $q < 10^{-1}$). Visual inspection of list in question reveals that cell-cycle and DNA replication/repair genes are abundant in this transcriptome and include genes such as *atm-1*, *dna-2*, or *hpr-17*. This analysis reveals that the ciliary neuron transcriptome is enriched in genes associated with microtubules, but the cell-cycle machinery is also carefully regulated.

Deconstructing phenotype-tissue relationships

Tissue enrichment on the Egl gene set reveals cellular components of the phenotype

How does a phenotype emerge? We realized that with the tools that we have developed, it is possible to understand what tissues contribute to a phenotype in a probabilistic framework. In other words, we can extract all genes associated with a particular phenotype, then search for tissue terms that are enriched to understand how a phenotype arises from interactions between anatomical regions. As a test of this, we selected the egg-laying defective (Egl) phenotype. In *C. elegans*, egg-laying is a complex behavior that involves a large number of tissues []. The somatic gonad acts as a repository for the eggs, the uterine seam cells help protect the uterus, and a variety of muscles help contract the uterus and open the vulva to lay an egg []. The vulva must be well-formed to allow passage of an egg, and the hermaphrodite-specific neuron (HSN) is involved in the egg-laying control []. The complexity of the interactions that happen to allow egg-laying make understanding the Egl phenotype in terms of tissues a challenging task.

We extracted all of the *C. elegans* genes that have been associated with an Egl phenotype and we used TEA to understand what tissues are enriched. The HSN was enriched more than five-fold above background ($q < 10^{-7}$) as were vulD, vulC, vulE and vulF ($q < 10^{-6}$). The vulA, vulB2 and vulB1 were enriched at slightly lower levels ($q < 10^{-5}$), whereas the uterine muscles and uterine seam cells were enriched more than twice above background levels ($q < 10^{-2}$). Therefore, the Egl phenotype would seem to emerge primarily from defects in the HSN, secondarily from defects in the vulva, and only sometimes from defects in the uterine seam cells or muscles. It is notable that all vulval cells were not equally enriched. Although all the 'vul' cells are annotated to a similar degree (between 50–70 genes for each cell type), the vulD and vulC cells had the largest en-

richment effect size and the lowest q-values, suggesting that these cells are more likely to be associated with an Egl phenotype than the others. This may reflect the fact that vulD and vulE are the site of attachment for four vulval muscles, vm1. Perhaps this attachment is particularly fragile, and perturbations to these cells prevent adequate function of these muscles. In support of these observations, P7.pa had the largest fold-enrichment of any tissue. In *C. elegans*, P7.pa gives rise to vulD and vulC. However, vulF is also attached to a set of four additional vulval muscles, vm2. Why is vulF less associated with an Egl phenotype?

Quantifying the anatomy-phenotype mapping via Bayesian probabilities

Another way to understand the phenotype-anatomy mapping is by considering how informative a given anatomy term is on a particular phenotype, or vice-versa. To this end, we calculated two conditional probabilities that helped us answer this question. The first conditional probability,

$$P(\text{a gene has Egl annotation}|\text{it is expressed in } X) \tag{1}$$

answers the question: For a gene with an expression pattern that includes the tissue term X (i.e., the gene is expressed at least in X), what is the probability that this gene has an Egl phenotype (i.e., the phenotype annotations for this gene include Egl)? For simplicity, we can re-write this equation more succintly by removing a few words. The calculation of this probability is straightforward and follows from the definition of conditional probability:

$$P(\text{Egl}|X) = \frac{N_{\text{genes annotated Egl and X}}}{N_{\text{genes annotated with X}}}. \tag{2}$$

Equation 2 measures how likely a gene is to be annotated with an Egl phenotype given that its expression pattern includes the term X. A related quantity (which is neither the inverse nor the complement) is the conditional probability that a gene which is annotated with at least the Egl phenotype is expressed in tissue X. That is to say,

$$P(X|\text{Egl}) = \frac{N_{\text{genes annotated Egl and X}}}{N_{\text{genes annotated with Egl}}}. \tag{3}$$

Equation 3 tells us how probable it is that any given gene that is annotated with an Egl phenotype includes X as a tissue term. Taken together, equations 2 and 3 help us understand how predictive anatomic expression is of phenotypes, and how predictive phenotypes are of anatomic expression.

We calculated the conditional probability that a gene has an Egl phenotype given that it's expression pattern includes a tissue term X and we searched for the tissue terms that maximized this probability. The list of terms that maximized this probability reflected the results from running TEA on the subset of genes that have an Egl phenotype. We also calculated the conditional probability that a gene has expression in a tissue term X given that it is annotated with an Egl phenotype and we searched for terms that maximized this probability. The terms that maximized this probability were 'nervous system', 'pharynx' (a body part with a lot of neurons), 'sex organ' and 'tail' (a body part with neurons and hypodermis). In general, the terms that had a high $P(\text{Egl}|X)$ did not have a high $P(X|\text{Egl})$. Additionally, the terms that had a high $P(X|\text{Egl})$ are broad terms that include a lot of cells, whereas the terms that had a high $P(\text{Egl}|X)$ were considerably more specific. We conclude that the Egl phenotype arises from a small set of tissues. The Egl phenotype can be best predicted by genes with expression patterns that include at least one of a small number of cells (mainly vul cells, HSN). On the other hand, answering whether the expression pattern of a gene includes a particular anatomic region or tissue given that the gene has an Egl phenotype is hard to do for small tissues or single cells. However, guesses about what functional system or broad anatomic region may be affected by an Egl gene can be answered with confidence ($\sim 70\%$ of the time, the nervous system is the system affected by an Egl mutant).

Table 1. Conditional probabilities for various tissues. The first column shows the conditional probability that a gene has an Egl phenotype given that it has expression in tissue X (given by the row). The second column shows the conditional probability that a gene has expression in the anatomy term X given that it has an Egl phenotype. The first 9 terms are the terms for which $P(\text{Egl}|X)$ is maximized. The last three terms are the terms which have the highest $P(X|\text{Egl})$. For clarity, the Pn.p cells are not shown even though $P(\text{Egl}|\text{Pn.p}) \sim 0.24$.

Tissue	$P(\text{Egl}\|X)$	$P(X\|\text{Egl})$
P7.pa	0.30	0.04
HSN	0.27	0.11
vulC	0.27	0.06
vulD	0.26	0.07
vulE	0.25	0.06
vulF	0.24	0.06
vulA	0.24	0.05
vulB2	0.23	0.05
vulB1	0.22	0.05
nervous system	0.02	0.72
pharynx	0.00	0.46
sex organ	0.07	0.41
tail	0.05	0.33

Conclusions

We have highlighted three possible uses for phenotype and GO enrichment analyses. Taken together, the WormBase Enrichment Suite can help guide researchers as they search for a phenologue that is representative of a human disease that has no immediately obvious counterpart in the worm. Such phenologues may benefit from the fact that they represent an unbiased approach to disease modeling as long as the human genes are selected from unbiased screens. GWAS data suggests that certain aspects of lupus biology may be best understood from a developmental perspective in *C. elegans*. On the other hand, our results suggest that rheumatoid arthritis might not be effectively modeled in the worm, although it is possible that our inability to find phenologues more concretely is the result of the GWAS screens: If GWAS screens identified factors that lead to worse prognosis of rheumatoid arthritis, but not factors that cause onset of the disease, this could explain the lack of phenotype enrichment in *C. elegans*.

The addition of GO and Phenotype Ontology enrichment testing to WormBase marks an important step towards a unified set of analyses that can help researchers to understand genomic datasets. By offering tissue, phenotype and gene ontology enrichment on a single site, researchers will benefit both in the speed at which the analysis is completed and from the fact that the analyses are all carried out with the same methodology and the same underlying algorithms.