Comparative Analysis of Semantic Similarity and Gene Orthology Tools for Identification of Gene Candidates for Human Diseases

Master's Thesis

Bryan Laraway
Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University
September 2015

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Bryan Laraway

has been approved.

Kemal Sonmez, PhD

Melissa Haendel, PhD

Chris Mungall, PhD

Shannon McWeeney, PhD

Abstract

In the study of rare and undiagnosed diseases it is of critical importance to identify potential gene candidates for those disease in order to establish an appropriate treatment protocol. However, there may be insufficient resources available to investigate rare and undiagnosed diseases, so it may be necessary to identify potential gene candidates from patient clinical observations to narrow the focus of the investigation. Two approaches for identifying gene candidates are investigated and compared: the OWLSim package for comparing the semantic similarity of phenotypes and the Phenolog approach for identifying orthologous phenotypes by searching for orthologous gene enrichment between phenotypes.

In comparing the two approaches by measuring ranked-recall of known disease causative genes, the Phenolog approach performed better than the OWLSim package. However, each approach was able to exclusively identify known disease causative genes, indicating that a combined approach would provide a better retrieval of correct gene candidates than either method alone.

Background

A fundamental goal in biology is to acquire a greater understanding of physiology, morphology, and disease. Through extensive observation and experimentation in model organisms, there is a wealth of knowledge available on the molecular and cellular processes in a wide range of species, from single-celled organisms to multicellular plants, insects, and animals. In addition, there is an increasing amount of information on how genetic mutation results in alterations of those processes, and how those alterations ultimately result in changes that manifest as a mutant phenotype or disease. Despite the availability of this immense body of knowledge, the underlying mechanisms of many human diseases are poorly understood or unknown, particularly in the cases of rare or undiagnosed diseases.

In the study of rare and undiagnosed diseases, it is of critical importance to identify the underlying genetic mechanisms that produce the phenotypes associated with a disease. However, with rare and undiagnosed diseases where funding may be limited for investigating a disease with a small

prevalence, additional tools are necessary to identify gene candidates that may be responsible for the phenotypes presented by the patient. Even as the costs of whole genome sequencing decrease and the analytical throughput increases, it is still beneficial to identify gene candidates to help narrow the focus of investigations for the underlying mechanism of an undiagnosed disease.

As discussed by McGary el. al.[1] (citing data from McKusick's Online Mendelian Inheritance in Man[2], the Mouse Genome Database[3], WormBase[4], Saccharomyces Genome Database[5], and the Saccharomyces cerevisiae Morphological Database[6]), the number of gene to phenotype associations in model organisms has far outpaced the number of gene to phenotype associations in humans. Therefore it is expedient to utilize the existing gene-phenotype association data in model organisms in order to accelerate the number of gene-phenotype association discoveries in humans. However, there are significant barriers to comparing phenotypes of model organisms to clinical observations of human diseases, including differences in anatomy between organisms, dissimilar manifestations of phenotypes between organisms, and equating genes across species. In this study, two approaches for identifying gene candidates for human diseases were investigated that attempt to overcome these interspecies comparison barriers: identifying semantic similarity of phenotypic profiles using OWLSim and identification of orthologous phenotypes, or Phenologs, through testing for enrichment of orthologous genes between phenotypes.

In order to properly describe approaches for identifying gene candidates for human diseases in model organisms, we must first properly define the relationship between diseases, phenotypes, genotypes, genes, and gene variants. A gene variant represents a mutation of a gene that is "abnormal" from the "normal" or "wild-type" version of the gene, resulting in many potential types of variation in the gene product including changes in expression, reduced function or loss of function, or accumulation of the gene product through reduced degradation. A mutation of a gene can be associated with an observable phenotype, with the sum total of the phenotypes associated with the gene mutation representing the phenotypic profile of that gene mutation.

A genotype is the sum total of gene mutations for an organism, which includes the background strain of the organism, if known, as compared to the reference genome of the organism. The variant genotype can result in one or more variant phenotypes, with the sum total of those phenotypes representing the phenotypic profile of the genotype (**Figure 1A**). For genotypes with more than one gene variant, there may be additional phenotype members of the phenotypic profile than found in the single variant genotypes due to direct or indirect interactions between the gene variants and/or their products.

A disease can also be described as a collection of phenotypes, which for human disease may be comprised of clinical observations, with the sum total of the collection of phenotypes representing the phenotypic profile of the disease (**Figure 1B**). If the genotype of the diseased individual is known, a direct connection can be inferred between the genotype and the disease, with the caveat that in humans this relation is not as clear as the relation between genotypes and diseases in inbred strains of laboratory animals. However, in the absence of genotype data, the phenotypic profiles of diseases and genotypes can potentially be matched to infer an association between the genotype and the disease (**Figure 1C**).

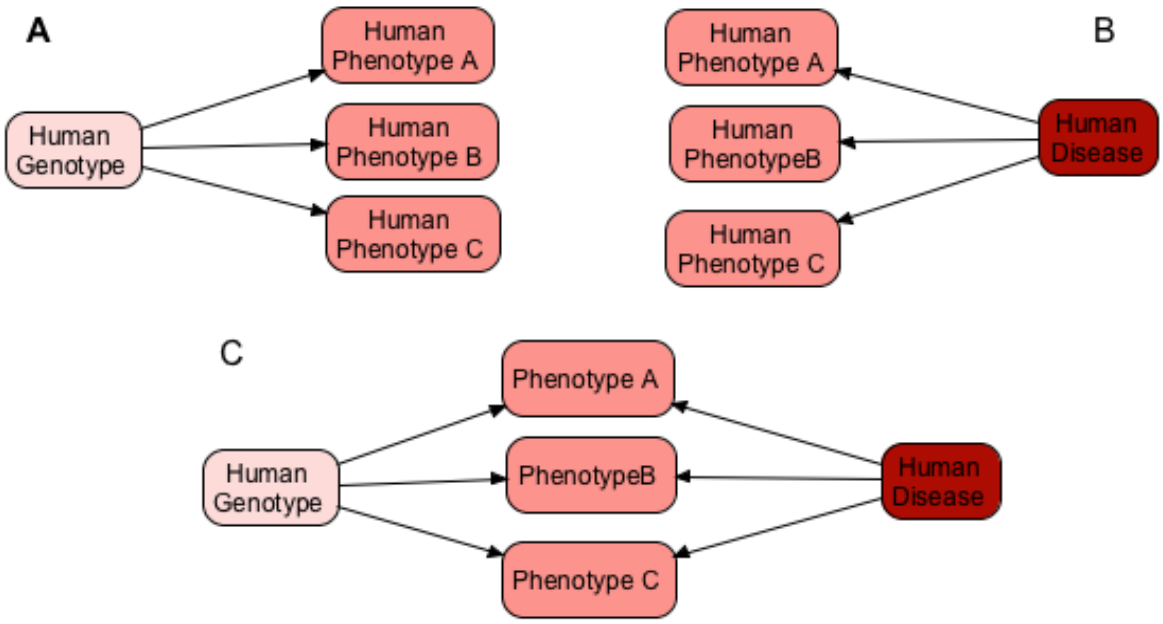

Figure 1: Phenotypic profiles of a genotype (**A**) and a disease (**B**). By matching the phenotypic profiles between the genotype and disease, a genotype-disease relation can be inferred (**C**).

The manifestation of all phenotypes in the phenotypic profile associated with a gene variant or disease is not required for an afflicted individual/organism, but instead the phenotypic profile represents the sum total of potential phenotypes that are associated with the gene variant or disease. The expression of the phenotype may be dependent upon developmental stage of the organism, environmental factors, or experimental conditions, and may include variability between individuals. By making comparisons between the full set of potential phenotypes included in a phenotypic profile, we can attempt to identify relationships between diseases and gene variants across species while allowing for variability in the presentation of the phenotypic profile.

In order to identify gene candidates for human diseases from model organisms, we must identify similar phenotypic profiles across species by comparing the phenotypic profiles of human diseases and model organism gene variants. One existing tool for identifying similarities between phenotypic profiles is the OWLSim package, which provides semantic similarity metrics for entities (diseases, genes) and their attributes (phenotypes). A second tool that can be extended from single phenotype comparisons to comparisons of phenotypic profiles is the identification of Phenologs, or orthologous phenotypes, by identifying enrichment of orthologous genes between phenotypes. Both tools are described below.

OWLSim

The OWLSim package[7,8] provides tools for calculating the semantic similarity of entities based on their attributes through the utilization of ontologies, computational reasoners, and semantic similarity metrics. For the purpose of this study, the entities are gene variants or diseases, while the attributes are the phenotype members of the phenotypic profiles of the gene variants or diseases. Ontologies provide a controlled vocabulary for a domain of knowledge, describing the types, properties, and relationships between attributes while providing a hierarchical organization of attributes. Computational reasoners are used to reason across ontologies, moving across hierarchical relationships to identify inferred equivalencies at varying levels of detail within the ontology. Semantic similarity metrics are used to provide similarity scores for the described entities and their attributes based upon the rarity of the matching inferred attributes.

For example, when comparing Pax6 mutations between humans and mice, the "opaque cornea" phenotype in humans and the "small eyed" phenotype in mice would not normally match when performing computational string matching (**Figure 2**). However, using the OWLSim package the terms can be inferred to more general descriptions that match by computational string matching in a process known as subsumption. Scoring of the matching terms is dependent upon the frequency that the matching term appears in the set of annotations and the term's depth in the ontology, with less frequent terms receiving higher scores. The scores are known as the Information Content (IC) score[8,9], which is defined as the negative log of the probability that the description would be used to define a gene, genotype, or allele, with the probability of the description being equal to the total number of features annotated with that description divided by the total number of features. The highest scoring subsumed description for a pair of phenotypes is known as the least common subsumer.

OWLSim can be used to identify candidate genes for human diseases in which the genetic basis is unknown by determining the semantic similarity of the phenotypic profiles of the disease and model organism gene variants[8]. OWLSim provides four semantic similarity metrics (maxIC, simIC, ICCS, and simJ) that provide a measure of the similarity of two entities (diseases/genes) and their attributes (phenotypic profiles). By comparing the phenotypic profile of a human disease in which the causative gene is unknown and the phenotypic profiles of variant genes in model organisms, candidate genes can be identified in model organisms whose orthologs may be associated with the disease in humans, either directly or through involvement in the same or similar biological pathway. In point of fact, the work performed by Washington et al. in determining the phenotypic similarity of genes between species often found that the most phenotypically similar gene variants between species were not always the sequence orthologs[8].

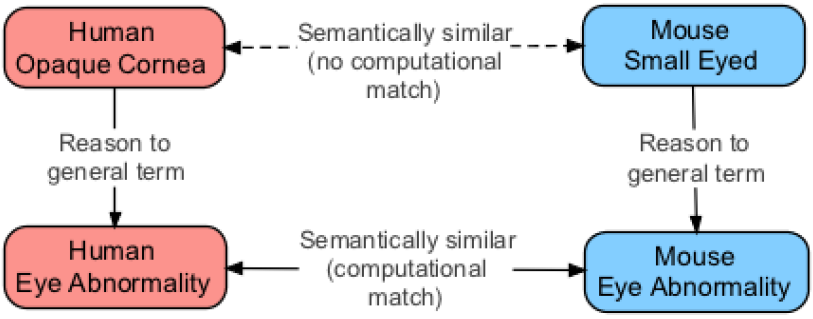

Figure 2: Using ontologies and computational reasoners, a non-matching pair of attributes can be converted to matching attributes through hierarchical relationships represented in the ontology.

While the OWLSim package can be used to match diseases and genes that have semantically similar phenotypic profiles, it will naturally have difficulty in identifying similar phenotypes in organisms that arise from similar genes but present in phenotypes that are not semantically similar. This may occur due to the two compared organisms being highly disparate and not sharing anatomy (such as comparing humans to yeast), it may be due to insufficient annotation of the phenotypes or insufficient integration or mapping of ontologies across species, or may occur due to an abnormality of how the phenotype is presented in different organisms despite the phenotypes arising from the same mutated gene. In order to identify matching phenotypic profiles despite the presence of these "non-obvious" phenotype pairs, an alternative method to semantic similarity must be employed.

Phenologs

Mutations in a pair of genes that are orthologous between two disparate species may present in those species as phenotypes that are not obviously similar due to differences in anatomy, but at a molecular level the biochemical pathways involved may be analogous between the two organisms. Identification of these non-obvious phenotypic relations requires a different approach then matching phenotypes based on semantic similarity. For example, attempting to identify similar phenotypes between humans and yeast based on phenotypic expression is difficult due to the lack of anatomical features in yeast, a single-celled organism. Comparisons can be made at the cellular and molecular levels, but not with the emergent properties of the phenotype in the tissues and anatomical features of a multicellular

organism. For example, an abnormal angiogenesis phenotype in mice would not be obviously associated with reduced growth rates of yeast when exposed to the hypercholesterolemia drug lovastatin. However, the two phenotypes are associated with genes that are known to be orthologous[1].

The second approach for identifying gene candidates for human diseases in this study is through the identification of Phenologs[10,11]. In the initial work by McGary et al., the authors extended the concept of orthologous genes to orthologous phenotypes by testing for the enrichment of orthologous genes between cross-species phenotype pairs (**Figure 3**). If a pair of phenotypes are determined to be a Phenolog, any unmatched orthologs from one organism would be considered gene candidates for the other organism and vice versa. Continuing that work, Woods et al. devised a matrix-based approach for incorporating multiple Phenologs to identify potential gene-phenotype associations and provide weighted rankings for those gene candidates.

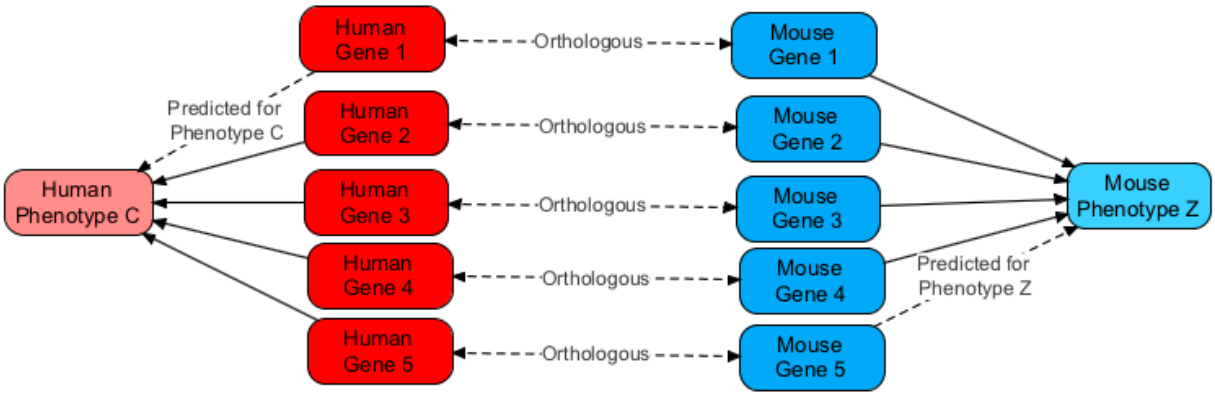

Figure 3: Identification of cross-species phenotype associations through the matching of orthologous genes. Pairs of phenotypes with matching orthologous genes are assessed by hypergeometric probability to determine if a signicant association between the phenotypes exists. Orthologs that are associated with a phenotype in one organism but not the orthologous phenotype are considered gene candidates for the orthologous phenotype.

An important distinction between the current study and the prior Phenolog implementations for identifying gene candidates is that instead of directly comparing model organism phenotypes and human diseases, this study follows the approach used by OWLSim in treating a disease as a collection of phenotypes, represented as a phenotypic profile. This distinction requires that gene candidate predictions for multiple model organism phenotypes be combined, corresponding to the phenotypic profile of the

human disease. This also allows for the identification of gene candidates for undiagnosed human diseases, as gene candidate predictions can be assembled from Phenologs that correspond to the patient's clinical observations. Here I test two approaches for combining the gene candidate prediction scores: a maximum scoring method and an additive scoring method.

The goal of this study in examining the ability of OWLSim and Phenologs to identify gene candidates for human diseases is to compare the performance of these tools and identify whether or not a combination of the two approaches would be beneficial for identifying gene candidates for human diseases. Both approaches are tested for the recall of known disease causative genes as described in the OMIM morbid map, and the results of these tests are reported.

Methods

Data preparation

Pre-scrubbed human (Human Phenotype Ontology: Annotations - HPO), mouse (Mouse Genome Informatics - MGI), and zebrafish (The Zebrafish Model Organism Database - ZFIN) data were obtained from the Neuroscience Information Framework (NIF, http://neuinfo.org). Human data was retrieved from the NIF production views of HPO on 5/19/15, built on data retrieved from HPO on 11/22/14. Mouse data was retrieved from the NIF production views of MGI on 5/19/15, built on data retrieved from MGI on 10/18/14. Zebrafish data was retrieved from the NIF production views of ZFIN on 5/19/15, built on data retrieved from ZFIN on 11/22/14. PANTHER data for identifying orthologs was also obtained from a NIF production view of PANTHER on 4/13/15, built on data retrieved from PANTHER on 11/25/14. For the purposes of this study, multi-gene genotypes were removed, as the contribution of each gene to a given phenotype is uncertain without additional investigation.

OWLSim

The OWLSim package produces several semantic similarity metrics when evaluating two phenotypic profiles: MaxIC, ICCS, SimIC, and SimJ. MaxIC is the highest IC score out of all matching EQ statements between two phenotypes. ICCS is a metric that takes all of the EQ statements for two

features and attempts to find the highest matching EQ statement for each EQ statement of the two features. The set of highest scoring EQ statements are known as the common subsumers, and the average IC score of the set is the ICCS. SimIC takes two phenotypic profiles and uses the ratio of the sum of the IC scores of the common EQ statements between the phenotypes over the sum of the IC scores of the total EQ statements between phenotypes. In addition to these three metrics, OWLSim also utilizes simJ, or the "Jaccard index," which does not use IC values. SimJ is the ratio of the total number of descriptions in common between two phenotypic profiles over the total number of descriptions in both profiles. The scores provided by these four metrics represent the similarity of two phenotypic profiles of human diseases and model organism genes as determined by OWLSim, and were used for comparison with the gene candidate predictions identified using Phenologs.

To identify potential gene candidates for human diseases, the phenotypic profiles of human diseases and mouse/zebrafish genes were compared using OWLSim. In order to compare phenotypic profiles between human diseases and mouse/zebrafish genes using OWLSim, a local instance of the OWLSim server, including the necessary ontologies, was downloaded and implemented. The OWLSim server and ontologies were downloaded from the OWLTools GitHub repository (https://github.com/owlcollab/owltools) and the phenotype-ontologies repository (https://code.google.com/p/phenotype-ontologies/) on 6/14/15. Phenotypic profile comparison queries were pre-assembled for all disease-gene comparisons, then used to query the OWLSim server. The results from successful queries were assembled for each disease-gene comparison, and then sorted by each OWLSim semantic similarity metric (maxIC, ICCS, SimIC, and SimJ) for comparison to Phenologs.

Phenologs

Using the matrix-based gene-candidate prediction approach as described in the Phenolog extension paper by Woods et al.[12], weighted gene-predictions were assembled for human diseases from mouse and zebrafish gene-phenotype associations. For each species included in the analysis, a gene-phenotype association list was assembled, and then converted to ortholog-phenotype lists using PANTHER. Using PANTHER, the gene IDs were converted to the corresponding ortholog IDs using

orthologs with either the least diverged ortholog (LDO) or ortholog (O) designation. Groups of genes from the same orthogroup with identical PANTHER IDs, were collapsed into a single instance of the ortholog's PANTHER ID to prevent overrepresentation of genes from the same orthogroup in subsequent ortholog enrichment calculations. Using these ortholog-phenotype lists, a matrix was created with rows corresponding to orthologs and columns corresponding to phenotypes. Each *i*,*j* position in the matrix was set to 1 if an association between the i-th gene/ortholog and the *j*-th phenotype was present.

Based on the work of Woods et al.[12], the following settings were used for the implementation of gene predictions for the Phenolog algorithm in this study. To determine the orthogroup-phenotype association predictions, the additive method was used to calculate X_{ij}, the probability that orthogroup i is involved in phenotype j. In the additive method, X_{ij} is calculated by taking the sum of all nearest neighbor phenotypes k, weighted by the similarity between phenotypes j and k. For the weighting function, the hypergeometric cumulative distribution function was used, while the Pearson sample correlation was used for the similarity function. The maximum number of nearest neighbor phenotypes k used in the phenotype matrix Φ was ten.

To properly compare the Phenolog gene candidate predictions for single phenotypes to the OWLSim gene candidates for diseases (which are the gene candidates for the phenotypic profile of the disease, a combination of phenotypes), Phenolog scores for diseases were assembled from individual phenotypes corresponding to the phenotype members of the phenotypic profile of the disease. Two approaches were assessed for combining Phenolog gene candidate scores: a maximum score method where the highest prediction score for a gene candidate from the phenotypic profile member phenotypes was used; and an additive score method, in which the scores for a gene candidate were added from all phenotype members of the phenotypic profile. The gene candidates were then sorted by rank for each human disease and compared to the ranked gene candidate predictions obtained from the OWLSim package.

Recall of known OMIM disease-gene associations

To compare the performance of the OWLSim and Phenolog approaches in identifying known gene candidates for human diseases, each approach was tested for the ability to retrieve disease-gene associations present in the OMIM morbid map (http://omim.org/downloads) at the time of retrieval, which contains known disease causative genes. Out of the 6922 human disease-gene association records present in the OMIM morbid map, 4192 disease-gene associations had an available mouse ortholog for the disease-associated gene, while 3800 human disease-gene associations had an available zebrafish ortholog for the disease-associated gene. These disease-gene associations were further reduced when selecting for the mouse and zebrafish orthologs that are present in the gene-phenotype association data obtained from MGI and ZFIN, leaving 3834 records with matching mouse orthologs and 1694 records with matching zebrafish orthologs present in the MGI and ZFIN datasets, respectively.

Ranked retrieval of the mouse and zebrafish orthologs of the OMIM morbid map genes was calculated and plotted for all morbid map disease-gene associations as well as for disease-specific comparisons between OWLSim and Phenolog approaches. All recall-rank and scatterplot graphs were created using Stata 12.1.

All additional code created to perform the project, including Stata data analysis scripts, can be found on GitHub (https://github.com/bryanlaraway/phenothrowdown). Majority of the code is written using Python, version 3.4.3.

Results

Recall of known human disease-gene associations

To compare the performance of the OWLSim and Phenolog approaches in identifying known gene candidates for human diseases, each approach was tested for the ability to retrieve disease-gene associations present in the OMIM morbid map, which contains known disease causative genes. **Figure 4** shows a comparison of the OWLSim metrics and Phenolog approaches in retrieving mouse and zebrafish orthologs of human genes associated with OMIM diseases. The Phenolog additive method performed

best, with 47.2% of the mouse/zebrafish orthologs being returned for the OMIM disease-gene associations within the first 50 gene candidates and 61.7% in the top 500 gene candidates (**Table 1**). The Phenolog max method performed second best, with 45.5% of the mouse/zebrafish orthologs being returned for the OMIM disease-gene associations within the first 50 gene candidates and 60.4% in the top 500 gene candidates. Amongst the OWLSim semantic similarity metrics, maxIC returned the largest percentage of correct orthologs with 25.2% within the top 50 gene candidates and 51.5% within the top 500 gene candidates, with ICCS, simIC, and simJ returning successively fewer genes within the top 50 and 500 gene candidates (**Table 1**).

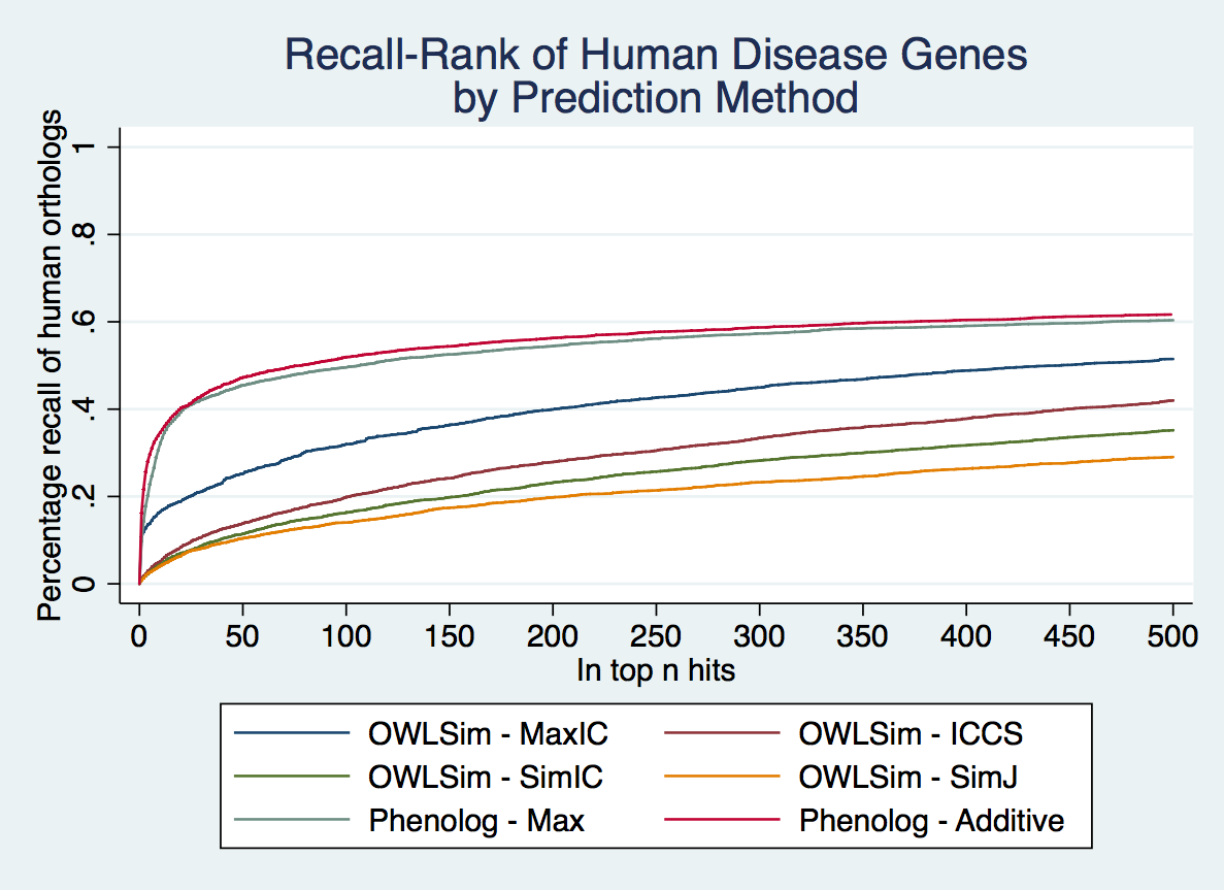

Figure 4: Recall of correct ortholog using OWLSim and Phenolog gene candidate predictions for human diseases from mouse and zebrafish data.

Table 1: Recall of OMIM morbid map disease-gene associations.

Metric	Recall in top 50 hits	Recall in top 500 hits	Total Recall
OWLSim maxIC	25.2%	51.5%	87.1%
OWLSim ICCS	13.8%	42.0%	87.1%
OWLSim simIC	11.4%	35.2%	87.1%
OWLSim simJ	10.4%	29.0%	87.1%
Phenolog max	45.5%	60.4%	66.8%
Phenolog additive	47.2%	61.7%	66.8%

Total observations = 4235

When examining the contribution of mouse (**Figure 5**) and zebrafish (**Figure 6**) data individually to the identification of correct orthologs for known disease-gene associations for known OMIM human disease-gene associations, the mouse data provided a greater percentage of human orthologs within the top 50/500 gene candidates than the zebrafish data. The top scoring OWLSim metric, MaxIC, only returned 11.5% of the correct orthologs within the top 500 gene candidates using zebrafish data, but returned 52.4% of the correct orthologs in the top 500 gene candidates when using mouse data (**Table 2**). In comparison, the additive and maximum Phenolog methods returned 32.1% and 31.8% of the correct orthologs within the top 500 gene candidates when using zebrafish data, respectively, and 61.8% and 60.3% when using mouse data. The greater performance of the Phenolog methods is to be expected given that humans and mice are closer than humans and zebrafish in regards to phylogenetic distance, as there are larger barriers for determining semantically similar phenotypes due to greater differences in anatomy and physiology between humans and zebrafish than humans and mice.

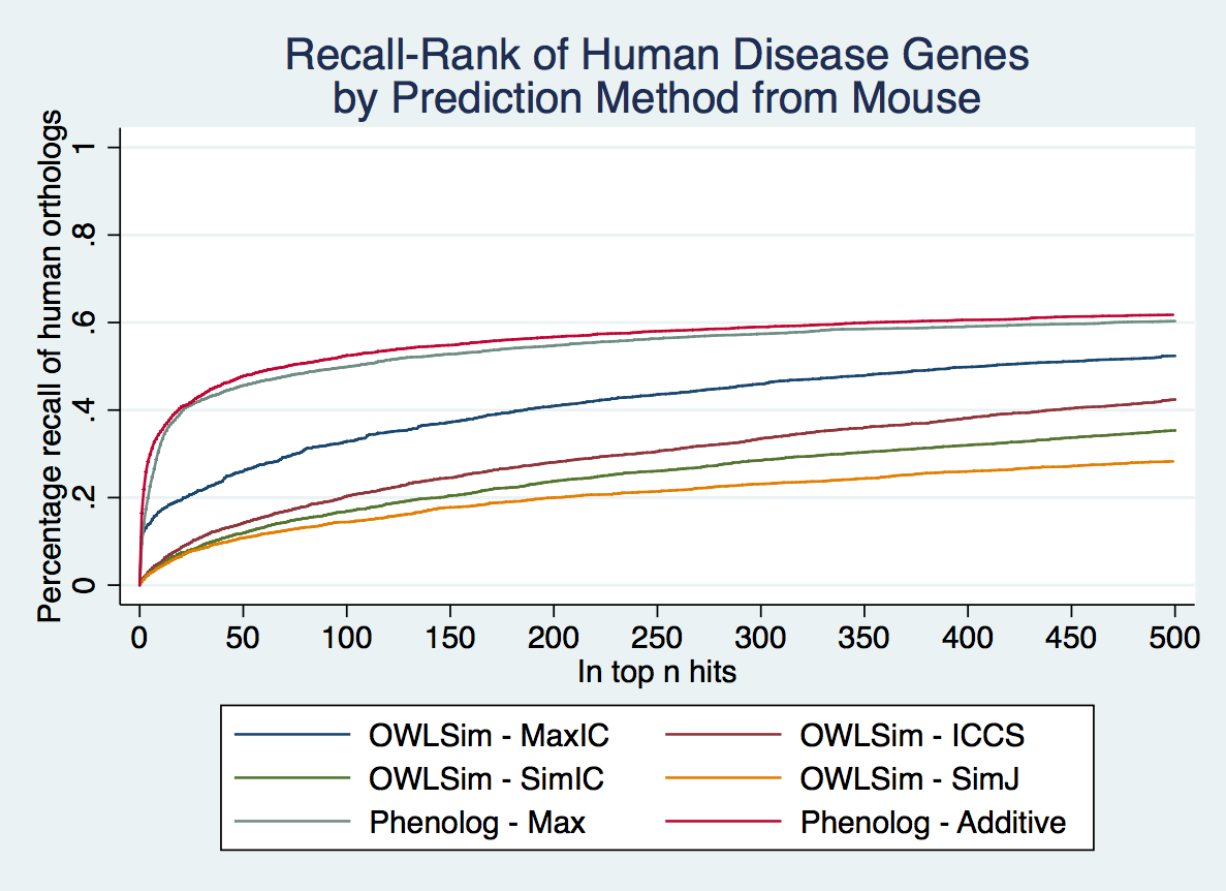

Figure 5: Recall of correct ortholog using OWLSim and Phenolog gene candidate predictions for human diseases from mouse data.

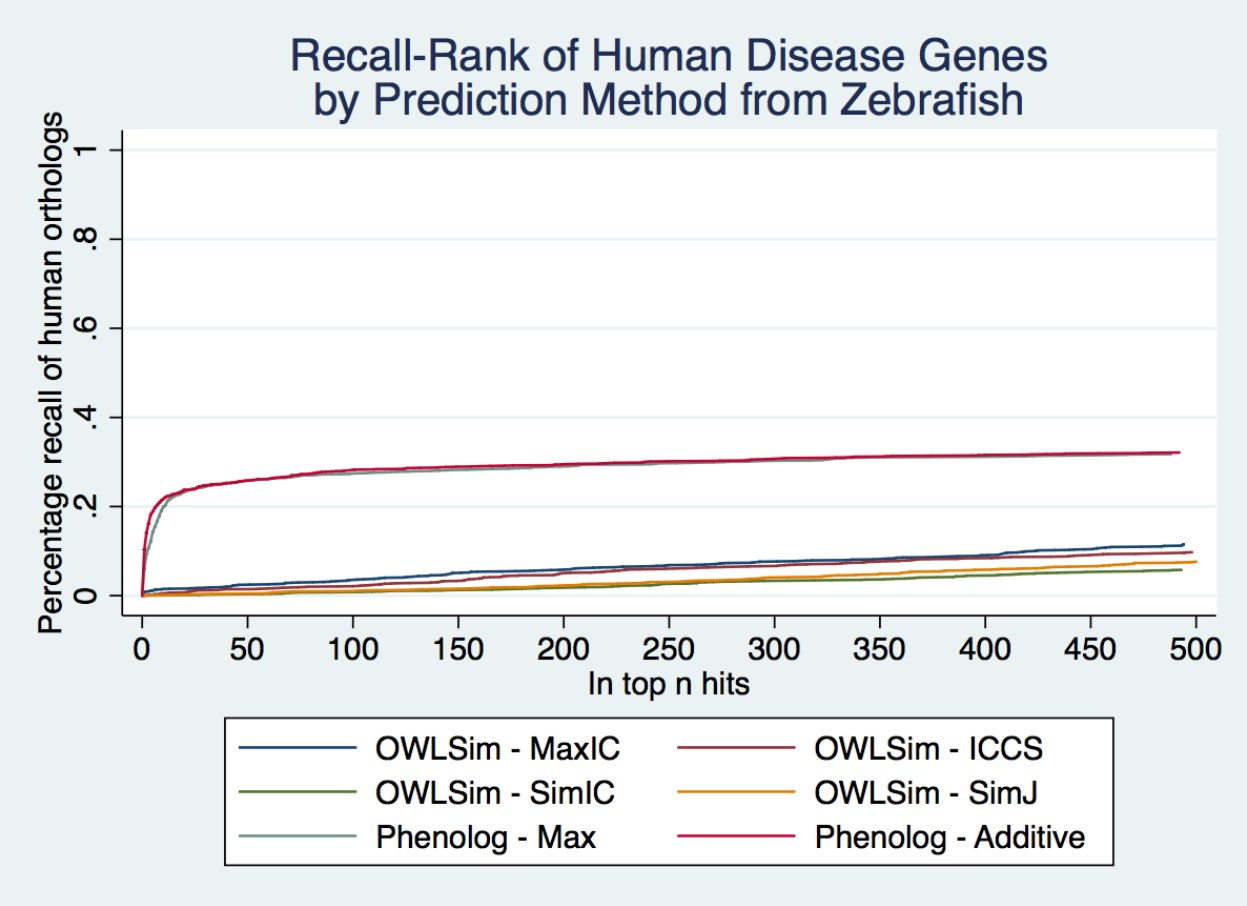

Figure 6: Recall of correct ortholog using OWLSim and Phenolog gene candidate predictions for human diseases from zebrafish data.

Table 2: Recall of OMIM morbid map disease-gene associations using only mouse or zebrafish datasets.

Metric	Recall in top 500 mouse hits	Recall in top 500 zebrafish hits
OWLSim maxIC	52.4%	11.5%
OWLSim ICCS	42.4%	9.7%
OWLSim simIC	35.3%	5.8%
OWLSim simJ	28.3%	7.6%
Phenolog max	60.3%	31.8%
Phenolog additive	61.8%	32.1%

Total observations = 4235

Overlap of known human disease-gene association retrieval

Due to the greater phylogenetic distance between humans and zebrafish than humans and mice and the corresponding greater differences in anatomy and physiology, it is expected that OWLSim would have increased difficulty in identifying correct gene candidates using zebrafish data than the Phenolog approaches. In support of this prediction, it is necessary to examine the overlap in known disease causative gene retrieval for each method when varying the datasets. When using the full dataset (comparing OWLSim maxIC and Phenolog additive metrics), 1705 known orthologs are identified by both OWLSim and Phenolog methods, with the Phenolog additive metric exclusively identifying an additional 772 correct orthologs and the OWLSim maxIC metric exclusively identifying an additional 363 correct orthologs (**Figure 7A**). When using only mouse data, the results are similar, with 1646 orthologs identified by both methods, 723 orthologs identified by the Phenolog Additive metric, and 363 orthologs identified by the OWLSim MaxIC metric (**Figure 7B**). However, when using only zebrafish data, 101 orthologs identified by both methods, 444 orthologs identified by the Phenolog Additive metric, and 95 orthologs identified by the OWLSim MaxIC metric (**Figure 7C**), indicating that the Phenolog method has greater performance in retrieving known disease causative genes than OWLSim between species with greater phylogenetic distances. Additionally, due to each gene candidate prediction method uniquely identifying disease causative genes across all datasets, it is clear that a combined gene candidate prediction approach will perform better than either approach in isolation.

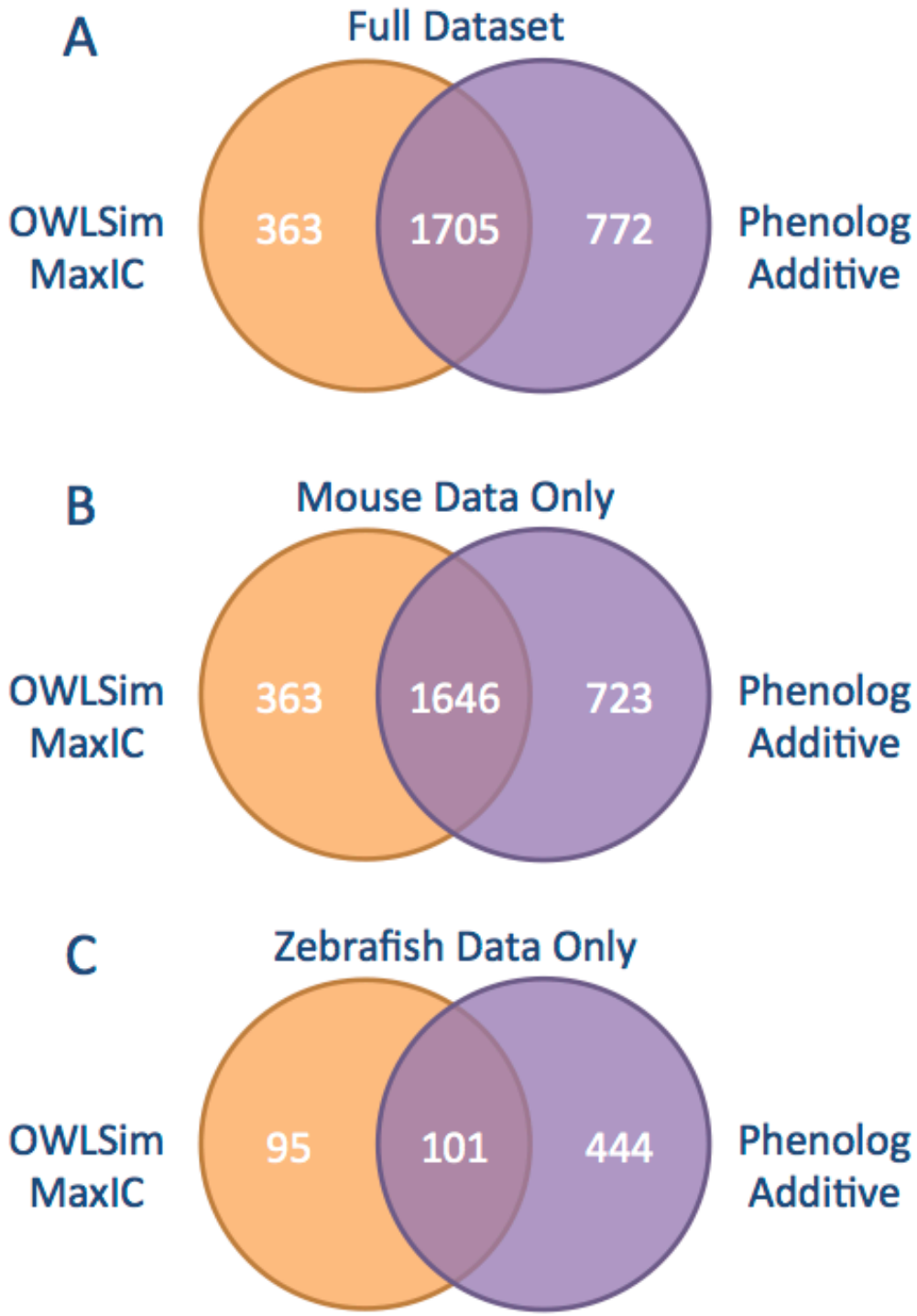

Figure 7: Recall of correct ortholog using OWLSim and Phenolog gene candidate predictions for human diseases from indicated dataset.

Discussion

Recall of known disease causative genes

In evaluating the OWLSim and Phenolog approaches to identifying gene candidates for human diseases, it is necessary to take a closer look at the retrieval of known disease causative genes for a selection of human diseases. The retrieval of known disease causative genes for breast cancer, osteoporosis, Pfeiffer Syndrome, and SHEP2 are described below. As the OWLSim maxIC metric and Phenolog additive method performed best, the examination will focus on these metrics unless otherwise noted.

Breast Cancer

Breast cancer is associated with one phenotype in the HPO dataset, breast carcinoma, and the disease is directly associated with nine genes: KRAS, TSG101, TP53, BRIP1, SLC22A18, PPM1D, RB1CC1, PIK3CA, and AKT1. The breast carcinoma phenotype is associated with 22 genes, which includes the nine genes associated directly with the disease. Within the OMIM morbid map, breast cancer is indicated to be associated with 23 genes, with 21 genes having phenotyped mouse or zebrafish orthologs in the MGI and ZFIN datasets. The retrieval of the mouse and zebrafish orthologs of these genes by OWLSim (maxIC) and Phenologs (additive method) is shown in **Figure 8**, along with the scoring of additional gene candidates. Some known disease causative genes are only retrieved by one method or the other, as indicated by the positioning of known disease causative genes scores along either the maxIC or Phenolog additive axis. OWLSim was able to retrieve nine disease causative orthologs while five orthologs were retrieved by Phenologs, with an overlap of two orthologs retrieved by both methods. The gene candidate with the best combined scoring of both methods was the zebrafish ortholog for TP53 (maxIC score: 6.45, Phenolog additive score: 10.98). However, while the mouse ortholog for TP53 had the same Phenolog score (Phenolog additive score: 10.98), it was not retrieved by any OWLSim metric. This is particularly interesting given that mice have greater anatomical similarities to humans than zebrafish, particularly in regards to the presence of breast tissue, which mice possess but zebrafish do not. Given these anatomical differences, it would be expected that genes associated with a

disease of the breast in humans would be more likely to be retrieved by OWLSim from organisms with the same anatomical features, but in this circumstance that is not the case.

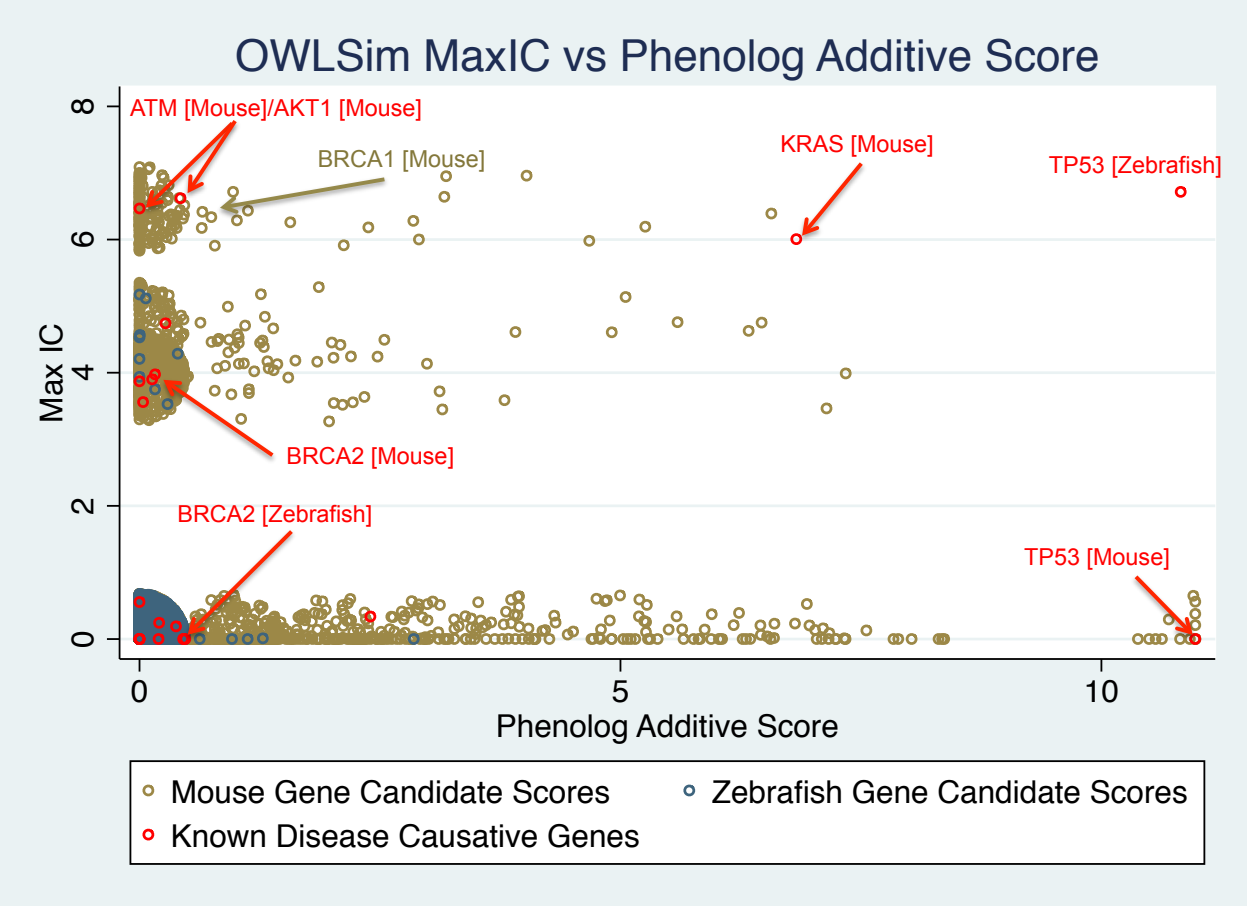

Figure 8: Scoring of known breast cancer causative genes and additional gene candidates from OWLSim (maxIC) and Phenolog (additive) methods. Random jitter applied to allow for visualization of overlapping data points.

In addition to the genes present in the OMIM morbid map, five mouse orthologs tied for the top OWLSim maxIC score (6.55): large tumor suppressor (Lats1); neuregulin 3 (Nrg3), which is involved in mammary gland morphogenesis[13]; relaxin/insulin-like family peptide receptor 1 (Rxfp1), a gene shown to be involved in cancer growth and metastasis[14]; relaxin 1 (Rln1), a gene involved in inhibition of tumor growth[15]; and hairless (Hr), which in addition to relations to loss of hair[16] is also associated with tumor development[17] and preventing apoptosis[18]. The top Phenolog gene candidate was transformation related

protein 63 (Trp63), a gene known to be associated with mammary cancer stem cell renewal[19], which tied with the score for the mouse and zebrafish orthologs of TP53 (Phenolog additive score: 10.98).

Using GeneMANIA[20] to analyze the top 21 Phenolog gene candidates not included in the OMIM morbid map, there are several genes present that are involved with stem cell regulation and proliferation, the Wnt signaling pathway, and morphogenesis of glands, branching structures, and epithelial layers, all of which can be related to breast cancer.

The ten nearest-neighbor phenotypes that informed the Phenolog scores were the zebrafish phenotypes 'abnormal(ly) neoplastic whole organism' and 'abnormal(ly) disrupted protein phosphorylation', and the mouse phenotypes 'increased esophageal papilloma incidence', 'increased pilomatricoma incidence', 'prognathia', 'increased teratocarcinoma incidence', 'abnormal umbilical cord blood vessel morphology', 'abnormal urethral gland morphology', 'increased melanoma incidence', and 'abnormal head size'.

Colorectal Cancer

Colorectal cancer is associated with five phenotype in the HPO dataset: neoplasm of the stomach; uterine leiomyosarcoma, renal cell carcinoma, hereditary nonpolyposis colorectal carcinoma, and transitional cell carcinoma of the bladder. These five phenotypes are associated with 84 genes, which includes nineteen genes associated directly with the disease. Within the OMIM morbid map, colorectal cancer is indicated to be associated with 23 genes, with 22 genes having phenotyped mouse or zebrafish orthologs in the MGI and ZFIN datasets. The retrieval of the mouse and zebrafish orthologs of these genes by OWLSim (maxIC) and Phenologs (additive method) is shown in **Figure 9**, along with the scoring of additional gene candidates. Some known disease causative genes are only retrieved by one method or the other, as indicated by the positioning of known disease causative genes scores along either the maxIC or Phenolog additive axis.

OWLSim was able to retrieve sixteen disease causative orthologs for colorectal cancer, while fifteen orthologs were retrieved by Phenologs, with an overlap of twelve orthologs retrieved by both methods. The gene candidate with the best combined scoring of both methods was the zebrafish ortholog

for TP53 (maxIC score: 7.33, Phenolog additive score: 30.83). However, while the mouse ortholog for TP53 had the same Phenolog score (Phenolog additive score: 30.83), it was not retrieved by any OWLSim metric, similarly to the results for breast cancer. This is particularly interesting as humans, mice, and zebrafish do have anatomical similarities in regards to the gastrointestinal tract, which should allow for comparisons by OWLSim between the three species where the lack of breast tissue in zebrafish may have prevented those same comparisons. This indicates that the lack of retrieval of the mouse ortholog of TP53 for breast cancer has less to do with the comparison of breast anatomy between species when comparing phenotypic profiles, and instead points to a deficiency somewhere else, such as in the annotation of TP53 in mice, annotation of breast cancer related phenotypes in mice, or insufficient annotation of breast cancer in the mouse phenotype ontology.

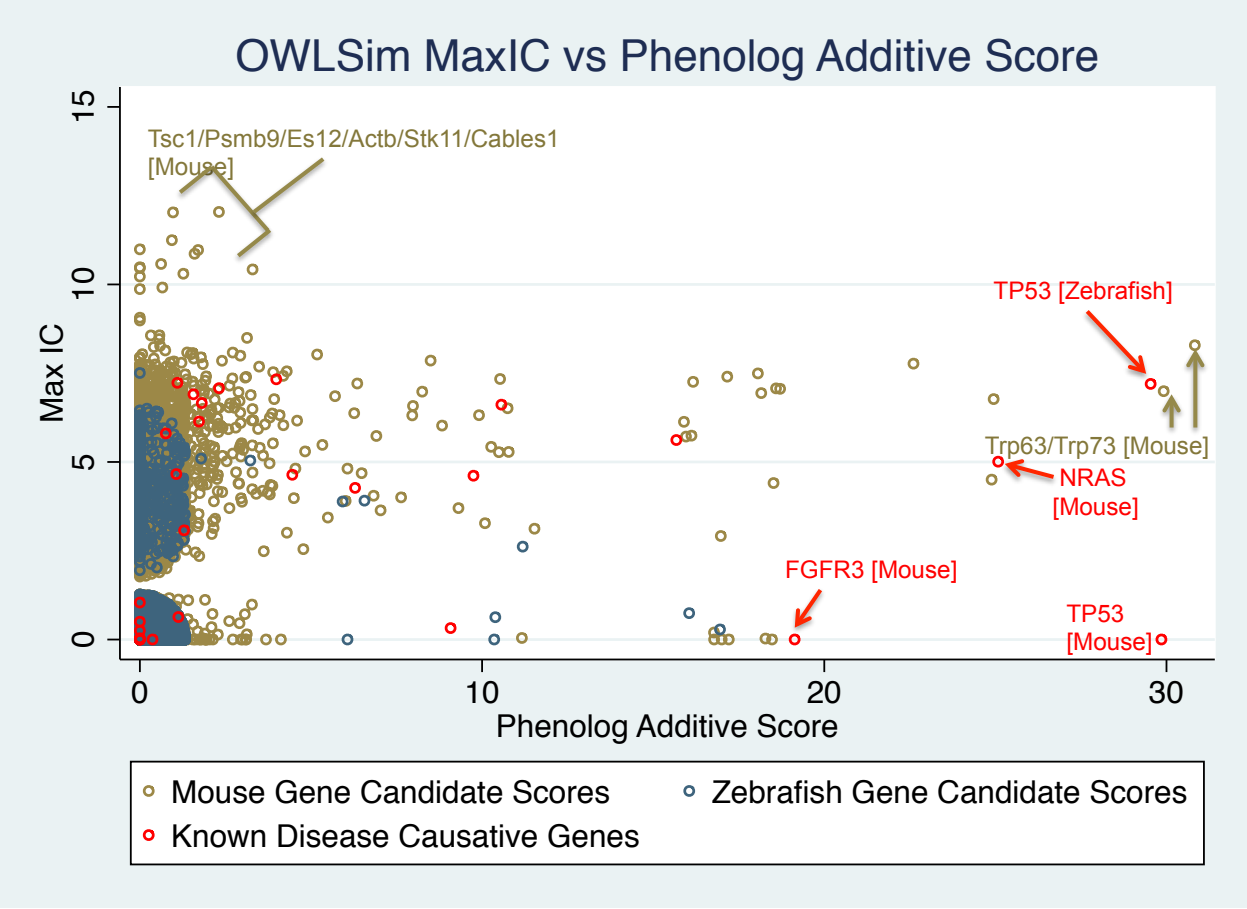

Figure 9: Scoring of known colorectal cancer causative genes and additional gene candidates from OWLSim (maxIC) and Phenolog (additive) methods. Random jitter applied to allow for visualization of overlapping data points.

In addition to the genes present in the OMIM morbid map, four mouse orthologs tied for the top OWLSim maxIC score (11.67): beta-actin (Actb); polo-like kinase 3 (Plk3), which acts as a tumor suppressor and is inferred to bind to TP53[21]; tuberous sclerosis 1 (Tsc1), a tumor suppressor; and proteasome subunit, beta type 9 (Psmb9), which also has a role as a tumor suppressor. As with breast cancer, the top Phenolog gene candidates included transformation related protein 63 (Trp63), but also included transformation related protein 73 (Trp73), which is a member of the p53 family of transcription factors along with Trp63, both of which tied with the score for the mouse and zebrafish orthologs of TP53 (Phenolog additive score: 30.83).

Osteoporosis

Osteoporosis is directly associated with three genes in HPO (COL1A1, VDR, COL1A2) and a single phenotype, osteoporosis. The osteoporosis phenotype is associated with 78 genes in the HPO database. Of the eight genes associated with the disease in the OMIM morbid map, there are six phenotyped genes with mouse and zebrafish orthologs available. Of these six human genes, OWLSim was able to retrieve five orthologs for five genes within the top 500 hits: mouse orthologs for COL1A2 (collagen, type I, alpha 2), LRP5 (low density lipoprotein receptor-related protein 5), VDR (vitamin-D receptor), and CALCR, and the zebrafish ortholog for COL1A1 (collagen, type I, alpha 1). Phenolog gene candidate predictions included the mouse orthologs for LRP5, VDR, and COL1A2. The retrieval of the mouse and zebrafish orthologs of these genes by OWLSim (maxIC) and Phenologs (additive method) is shown in **Figure 10**, along with the scoring of additional gene candidates.

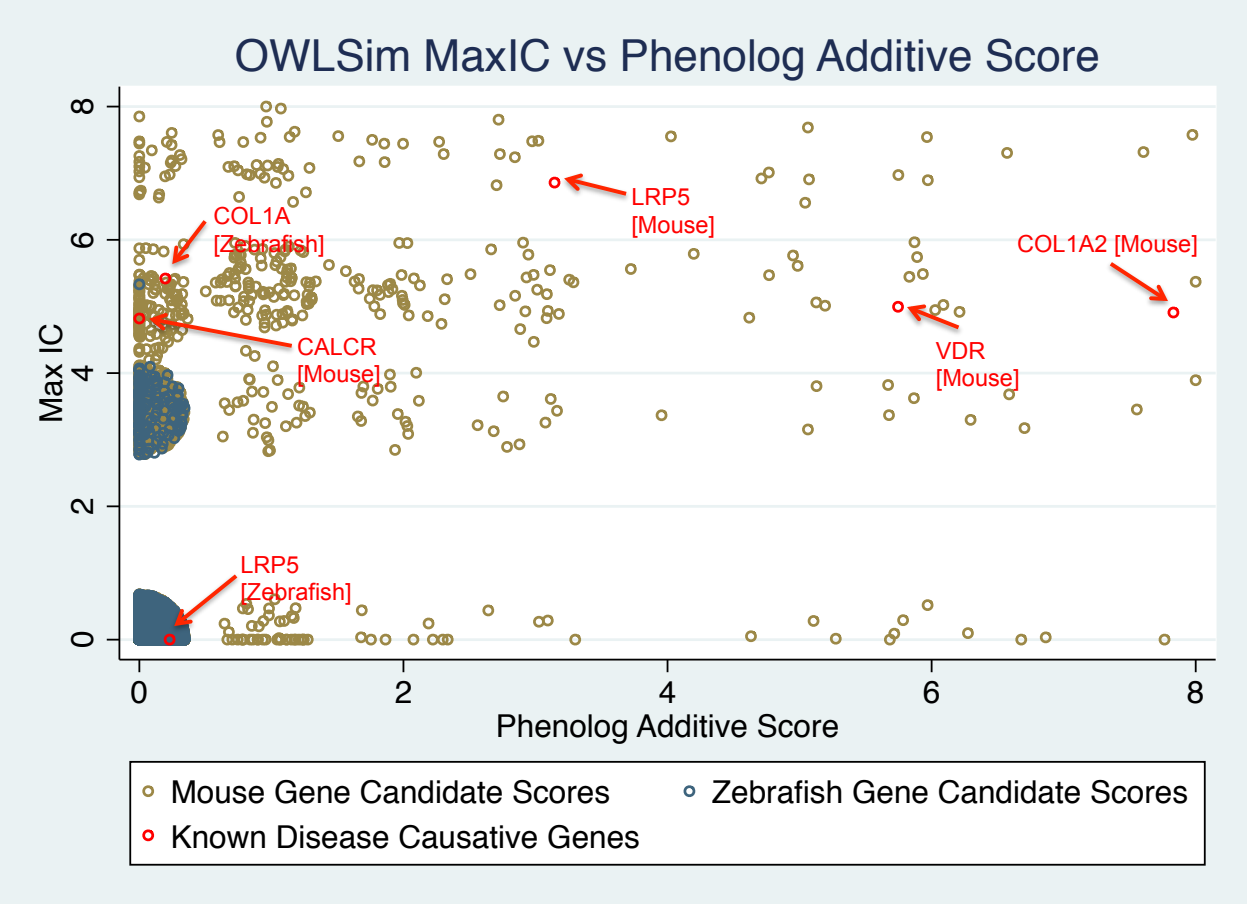

Figure 10: Scoring of known osteoporosis causative genes and additional gene candidates from OWLSim (maxIC) and Phenolog (additive) methods. Random jitter applied to allow for visualization of overlapping data points.

Additional Phenolog gene candidates include members of the collagen family (Col1a1, Col2a1, Col9a1, Col10a1, Col11a1, and Col27a1, all tying for the top prediction), thyroid hormone receptors (Thra and Thrb), and fibroblast growth factor receptors (FGFR2 and FGFR3). Using GeneMANIA[20] to analyze the top 18 Phenolog gene candidates not included in the OMIM morbid map, there is enrichment of mouse orthologs related to collagen, connective tissues, extracellular matrix, and skeletal system morphogenesis. In comparison, the top gene candidates identified by OWLSim are enriched in orthologs related to ossification, bone and tissue remodeling, and bone resorption.

The ten nearest neighbor phenotypes used to assemble the Phenolog gene candidate predictions for osteoporosis were 'abnormal femur morphology', 'abnormal compact bone morphology', 'increased width of hypertrophic chondrocyte zone', 'increased long bone epiphyseal plate size', 'osteoarthritis', 'decreased

bone mineral density', 'abnormal long bone epiphysis morphology', 'abnormal epiphyseal plate morphology', 'decreased interleukin-9 secretion', and 'abnormal joint capsule morphology'.

Pfeiffer Syndrome

Pfeiffer Syndrome is a syndrome characterized by craniosynostosis and abnormalities of the hands and feet[22,23]. It is directly associated with two genes in the HPO database, FGFR1 and FGFR2, and 28 phenotypes. Through the additional gene-phenotype annotations available in HPO for these 28 phenotypes and the corresponding gene-phenotype annotations, Pfeiffer Syndrome is associated with an additional 1153 genes. As Pfeiffer Syndrome has both a large phenotypic profile for OWLSim to compare with the phenotypic profiles of mouse and zebrafish genes, as well as a large set of gene-phenotype associations for use by Phenologs, it makes a good test case for comparing the two approaches.

The retrieval of known disease causative genes and additional gene candidates by OWLSim (maxIC) and Phenologs (additive method) is shown in **Figure 11.** Note that in the figure, the one low scoring disease causative gene is an additional zebrafish ortholog for FGFR1. All least diverged orthologs were successfully retrieved by Phenologs. OWLSim was able to retrieve three out of four potential mouse/zebrafish orthologs for FGFR1 and FGFR2, missing the mouse ortholog for FGFR2. The rankings for the zebrafish orthologs were rather low, with the ortholog for FGFR1 ranking 1341st and the ortholog for FGFR2 ranking 6192nd. However, the retrieval of the mouse ortholog for FGFR1 ranked 139th.

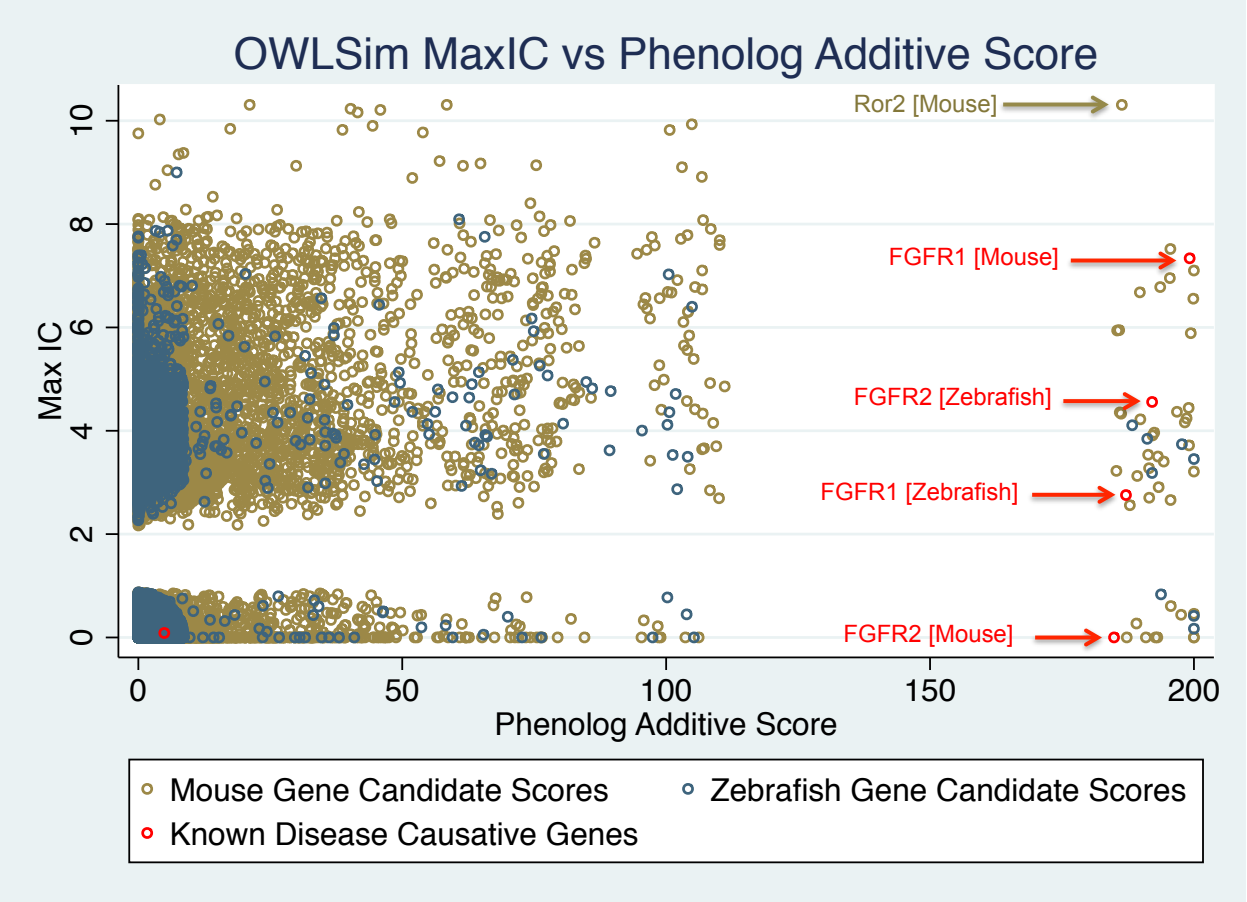

Figure 11: Scoring of known Pfieffer Syndrome causative genes and additional gene candidates from OWLSim (maxIC) and Phenolog (additive) methods. Random jitter applied to allow for visualization of overlapping data points.

Through the combination of the 10 nearest neighbor phenotypes for each of Pfeiffer Syndrome's 28 phenotypes, Pfeiffer has a total of 188 unique nearest neighbor phenotypes, consisting of 132 mouse phenotypes and 56 zebrafish phenotypes. From this large set of nearest neighbor phenotypes, there is naturally a large set of gene candidates for Pfeiffer syndrome. Using the Phenolog max score method, the zebrafish and mouse orthologs for FGFR1 and FGFR2 were tied for the 52nd highest ranked orthologs with a max score of 9.997 (maximum score is 10.0). However, using the Phenolog additive score method, the zebrafish and mouse orthologs for FGFR1 and FGFR2 were tied for the highest ranked orthologs with an additive score of 193.46, with the next highest score being 103.33. In total, 53 orthologs had the highest Phenolog additive score, consisting of 10 zebrafish orthologs and 43 mouse orthologs. Using GeneMANIA, the zebrafish orthologs are enriched in development, regeneration, and wound healing

functions. The mouse orthologs are enriched in transmembrane receptor kinase activity, cell-signaling pathways, protein phosphorylation, growth and development pathways, and cell motility.

SHEP2

Skin/Hair/Eye Pigmentation 2 (SHEP2) is a condition resulting in red hair, fair skin, and freckles, as well as increased susceptibility to UV-induced skin damage due to greater expression of pheomelanin/reduced expression of eumelanin[24]. Within the HPO dataset, SHEP2 is associated with one phenotype, red hair, which is associated with six genes: OCA2, PDE4D, POMC, PRKAR1A, TYRP1, and ZNF469. However, within HPO, SHEP2 is not directly related with any genes. Although in the OMIM morbid map there is one gene associated with SHEP2, the melanocortin 1 receptor (MC1R), this gene association is not present in the HPO dataset for either the gene-disease association with SHEP2 or the gene-phenotype association with red hair. This provides an opportunity to test the retrieval of a known gene association when the gene is not currently associated with the disease or associated phenotype in the dataset, serving as an example for finding a correct novel gene candidate for an undiagnosed disease.

The retrieval of orthologs of the known disease causative gene (MC1R) and additional gene candidates by OWLSim (maxIC) and Phenologs (additive method) is shown in **Figure 12**. While the zebrafish ortholog for MC1R was not retrieved by either Phenolog method and the OWLSim maxIC and ICCS metrics, it was retrieved by the OWLSim simIC and simJ metrics. However, both of these OWLSim metrics ranked the zebrafish ortholog poorly, with the simIC metric ranking the ortholog at position 907th and simJ ranking the ortholog at position 819th. In contrast, for the mouse ortholog of MC1R, the Phenolog additive and max methods both returned a rank of 4th. In addition, the maxIC OWLSim metric provided a ranking of 1st, with rankings of 251st, 555th, and 2051st for ICCS, simIC, and simJ, respectively.

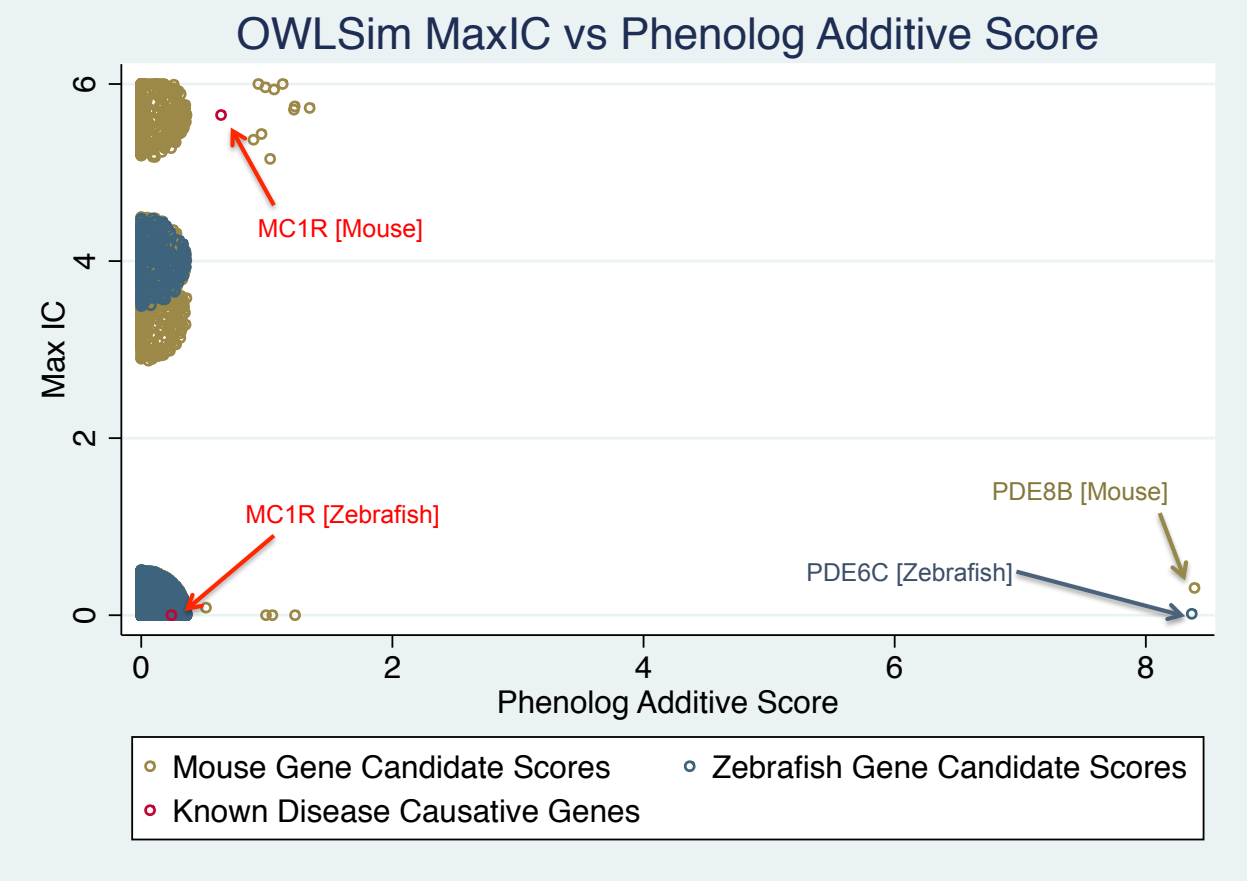

Figure 12: Scoring of known SHEP2 causative genes and additional gene candidates from OWLSim (maxIC) and Phenolog (additive) methods. Random jitter applied to allow for visualization of overlapping data points.

Conclusions

In the comparison of the OWLSim and Phenolog approaches to identifying gene candidates for human diseases from model organism data, the Phenolog additive approach performed best overall in recalling known disease causative genes. While the Phenolog additive and maximum methods performed similarly, the additive method had a small performance advantage over the maximum method. In addition, the additive method appears to return known disease causative genes at higher ranks than the maximum method. Both OWLSim and Phenolog approaches are able to exclusively identify known disease causative genes, indicating that each method can provide uniquely valuable gene candidate predictions for undiagnosed diseases. However, the Phenolog method was able to exclusively identify a greater portion of known disease causative genes between humans and zebrafish, indicating that

Phenologs will perform better than OWLSim when comparing species that are further apart in regards to phylogenetic distance. Although the Phenolog approach performed better than OWLSim, a combined approach will naturally yield a greater return of positive disease gene candidates than either method alone.

References

1. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A*. 2010;107:6544-6549. doi:10.1073/pnas.0910200107.

2. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue):D514-D517. doi:10.1093/nar/gki033.

3. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. The mouse genome database (MGD): New features facilitating a model system. *Nucleic Acids Res*. 2007;35. doi:10.1093/nar/gkl940.

4. Harris TW, Antoshechkin I, Bieri T, et al. Wormbase: A comprehensive resource for nematode research. *Nucleic Acids Res*. 2009;38. doi:10.1093/nar/gkp952.

5. Dwight SS, Harris MA, Dolinski K, et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*. 2002;30:69-72. doi:10.1093/nar/30.1.69.

6. Saito TL, Ohtani M, Sawai H, et al. SCMD: Saccharomyces cerevisiae Morphological Database. *Nucleic Acids Res*. 2004;32:D319-D322. doi:10.1093/nar/gkh113.

7. Chen CK, Mungall CJ, Gkoutos G V., et al. Mousefinder: Candidate disease genes from mouse phenotype data. *Hum Mutat*. 2012;33:858-866. doi:10.1002/humu.22051.

8. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*. 2009;7. doi:10.1371/journal.pbio.1000247.

9. Schlicker A, Albrecht M. FunSimMat: A comprehensive functional similarity database. *Nucleic Acids Res*. 2008;36(SUPPL. 1). doi:10.1093/nar/gkm806.

10. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A*. 2010;107:6544-6549. doi:10.1073/pnas.0910200107.

11. Woods JO, Singh-Blom UM, Laurent JM, McGary KL, Marcotte EM. Prediction of gene-phenotype associations in humans, mice, and plants using phenologs. *BMC Bioinformatics*. 2013;14:203. doi:10.1186/1471-2105-14-203.

12. Woods JO, Singh-Blom UM, Laurent JM, McGary KL, Marcotte EM. Prediction of gene-phenotype associations in humans, mice, and plants using phenologs. *BMC Bioinformatics*. 2013;14:203. doi:10.1186/1471-2105-14-203.

13. Howard B, Panchal H, McCarthy A, Ashworth A. Identification of the scaramanga gene implicates Neuregulin3 in mammary gland specification. *Genes Dev*. 2005;19(17):2078-2090. doi:10.1101/gad.338505.

14. Feng S, Agoulnik IU, Truong A, et al. Suppression of relaxin receptor RXFP1 decreases prostate cancer growth and metastasis. *Endocr Relat Cancer*. 2010;17(4):1021-1033. doi:10.1677/ERC-10-0073.

15. Radestock Y, Hoang-Vu C, Hombach-Klonisch S. Relaxin reduces xenograft tumour growth of human MDA-MB-231 breast cancer cells. *Breast Cancer Res*. 2008;10(4):R71. doi:10.1186/bcr2136.

16. Thompson CC. Hairless is a nuclear receptor corepressor essential for skin function. *Nucl Recept Signal*. 2009;7:e010. doi:10.1621/nrs.07010.

17. Kim H, Casta A, Tang X, et al. Loss of hairless confers susceptibility to UVB-induced tumorigenesis via disruption of NF-kappaB signaling. *PLoS One*. 2012;7(6):e39691. doi:10.1371/journal.pone.0039691.

18. O'Driscoll C, Bressler JP. Hairless expression attenuates apoptosis in a mouse model and the COS cell line; involvement of p53. *PLoS One*. 2010;5(9):e12911. doi:10.1371/journal.pone.0012911.

19. Memmi EM, Sanarico AG, Giacobbe A, et al. p63 Sustains self-renewal of mammary cancer stem cells through regulation of Sonic Hedgehog signaling. *Proc Natl Acad Sci U S A*. 2015;112(11):3499-3504. doi:10.1073/pnas.1500762112.

20. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008;9 Suppl 1:S4. doi:10.1186/gb-2008-9-s1-s4.

21. Yang Y, Bai J, Shen R, et al. Polo-like kinase 3 functions as a tumor suppressor and is a negative regulator of hypoxia-inducible factor-1 alpha under hypoxic conditions. *Cancer Res*. 2008;68(11):4077-4085. doi:10.1158/0008-5472.CAN-07-6182.

22. Cohen MM. Pfeiffer syndrome update, clinical subtypes, and guidelines for differential diagnosis. *Am J Med Genet*. 1993;45(3):300-307. doi:10.1002/ajmg.1320450305.

23. PFEIFFER RA. [DOMINANT HEREDITARY ACROCEPHALOSYNDACTYLIA]. *Z Kinderheilkd*. 1964;90:301-320. http://www.ncbi.nlm.nih.gov/pubmed/14316612. Accessed September 16, 2015.

24. Valverde P, Healy E, Jackson I, Rees JL, Thody AJ. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet*. 1995;11(3):328-330. doi:10.1038/ng1195-328.