How good is your phenotyping? Methods for quality assessment

Nicole L. Washington[1], Melissa A. Haendel[2], Sebastian Köhler[3], Suzanna E. Lewis[1], Peter Robinson[3], Damian Smedley[4], Christopher J. Mungall[1]

1 Lawrence Berkeley National Laboratory, Berkeley, CA; 2 Oregon Health & Sciences University, Portland, OR; 3 Institut für Medizinische Genetik und Humangenetik, Charité - Universitätsmedizin Berlin, Berlin, Germany; 4 Wellcome Trust Sanger Institute, Hinxton, UK

1 INTRODUCTION

Semantic phenotyping has been shown to be an effective means to aid variant prioritization and characterization by comparison to both known Mendelian diseases and across species with animal models (Robinson et al 2013). This process, whereby symptoms and characteristic phenotypic findings are curated with species-specific ontology terms, has generated a baseline set of disease-phenotype descriptions for more than 7,000 Mendelian diseases (Kohler et al 2014a) as well as many thousands of descriptions of additional animal models. By leveraging the knowledge encoded in the ontology graph and methods drawn from information theory, similarities can be computed between any two sets of phenotype descriptions (Washington et al 2009). This very powerful technique has the potential to be used for disease diagnosis, particularly for novel and rare diseases when the underlying genetic cause is unknown.

The robustness of semantic similarity methods is heavily dependent on the quality of both the knowledgebase as well as the phenotype profile being studied. Therefore, capturing the highest-quality phenotypic profiles is necessary. Until now, these phenotypic profiles have been typically captured by specialized curators, but as we want to move this technique into the diagnostic setting it will need to move into a physician's hands. This process of acquiring structured phenotype annotations for individual patients may seem daunting and unnecessarily complex for physicians with high demands on their time. Annotation tools such as Phenotips (Girdea et al 2013) greatly facilitate recording rigorous phenotype annotations in the clinic, but don't themselves provide guidance about what constitutes annotations sufficient for comparative phenotype analysis. Since clinicians are not used to providing structured phenotype data, it is necessary to provide a measurement of how a given patient phenotype profile compares against the corpus of available genotype-phenotype annotations, including that of known diseases, animal models, and other patients in the system. A metric to gauge overall complexity and diagnostic capability of a phenotype profile generated in this way would greatly enhance the ability to use structured phenotyping in the clinical setting for comparative analysis. Conversely, such a metric can also be utilized in the context of any systematic model organism phenotyping efforts.

Here, we present a method to assess the sufficiency of a phenotype profile, by investigating the necessary and sufficient information characteristics required to identify disease similarity based on phenotypes alone. This scoring method is being provided as a REST service through the Monarch Initiative API.

2 METHODS.

2.1 Data and Ontologies

Data and ontologies for analysis were downloaded on 2014-03-23. Human disease-phenotype annotations were obtained from http://human-phenotype-ontology.org, and treated as our "gold standard" set, which contained annotations for approx. 7,500 diseases. Mouse genotype-phenotype annotations were obtained from MGI (www.informatics.jax.org). Zebrafish genotype-phenotype annotations were obtained from ZFIN (www.zfin.org). All annotation data, pre-formatted for use in OWLSim, is available for download[1]. This data is also regularly updated in the Monarch Initiative website and services.

We used the Human Phenotype Ontology (HP) (http://purl.obolibrary.org/obo/hp.obo) in pairwise comparisons of diseases in this study, and the integrated phenotype ontology for multi-species analysis (Kohler S. et al 2014b), which includes the HP, the mouse phenotype ontology (MP), and

[1] http://code.google.com/p/phenotype-ontologies/

a zebrafish phenotype ontology (ZP) derived from the post-composed Entity-Quality annotations used by ZFIN (derived from the Zebrafish Anatomy and PATO quality ontologies).

2.2 Derived Disease Profiles

We generated new disease profiles derived from the set of disease-phenotype profiles described above. Briefly, one or more synthetic disease profiles D' was created for each disease D by removing, replacing, or altering phenotypes in the profile. These were generated in several ways: removing entire phenotypic categories (Method 2.3), replacing some or all annotations with less-specific superclass(es), or choosing random subsets. For any given derived disease, a set of controls were generated in parallel in order to assess any significant difference in similarity score between the derived disease and the original parent disease. For category-depletion derived diseases, we used only those diseases were there was >1 annotated category.

2.3 Categorical classifiers

We used the 1st degree subclasses in the upper level of the HP (typically divisions based on anatomical systems) to assess the role of broad phenotypic categories in the specificity of a profile. The 20 classes are listed in Table 1.

Table 1 Classes used to assess the role of broad phenotypic categories. The HPO identifier and abbreviated label is shown.

Category	ID
abdomen	HP:0001438
blood	HP:0001871
breast	HP:0000769
cardiovascular	HP:0001626
connective tissue	HP:0003549
ear	HP:0000598
endocrine	HP:0000818
eye	HP:0000478
genitourinary	HP:0000119
growth	HP:0001507
head/neck	HP:0000152
immune	HP:0010987
integument	HP:0001574
metabolism	HP:0001939
musculature	HP:0003011
neoplasm	HP:0002664
nervous system	HP:0000707
prenatal	HP:0001197
respiratory	HP:0002086
skeletal	HP:0000924

2.4 Similarity methods

All similarity comparisons were performed using OWLSim (owlsim.org), which enables a set of ontological entities to be compared against one or more other sets. Briefly, the HP and disease-phenotype associations were loaded and Information Content (IC) scores generated for each class based on the frequency of annotations (directly or inferred). Similarity scores were computed using OWLSim, as described in Smedley et al (2013), between the derived diseases (both cases and controls) and all diseases in the set.

Receiver Operator Characteristic (ROC) analysis was performed using the R ROCR package (rocr.bioinf.mpi-sb.mpg.de/) to assess the precision/recall of derived disease profiles when compared against their parent diseases.

2.5 Profile scores

Scores for any phenotype profile can be obtained via REST services, described in our documentation at monarchinitiative.org . We utilized IC measurements to generate scored annotation profiles for all diseases in our corpus of annotations. We generate three scores for an annotation profile as follows.

A *simple score* is calculated to assess the richness (measured by sumIC) and depth/strength (measured by maxIC and meanIC) of a profile as compared to all other annotated profiles (diseases or genes), without regard to the underlying shape of the ontology. The simple score is calculated using all phenotypes in the profile (where D is an alias for it's set of phenotypes $P_{1..n}$. Here, α, β, and γ coefficients were chosen to independently weigh the effects of sumIC, maxIC, and meanIC, respectively (where $\alpha+\beta+\gamma=1$). This results in a score in the range of (0..1). Our initial implementation weighs each factor equally.

$$simple\_score(D) = \alpha \frac{sumIC(D)}{mean(sumIC(D_{1..n}))} + \beta \frac{\max IC(D)}{mean(\max IC(D_{1..n}))} + \gamma \frac{meanIC(D)}{mean(meanIC(D_{1..n}))}$$

We can account for the shape of the ontology by assessing scores based on high-level categories in the ontology. A *categorical score* can be calculated using a similar formula to the simple score, but by taking the subset of phenotypes that are subclasses of a single phenotype category, and scaled using the mean obtained only from diseases with annotations to that category. The overall *categorical score* for a profile is averaged for all *c* categories (in our initial tests, there are *c=20* categories as described above). We do not yet

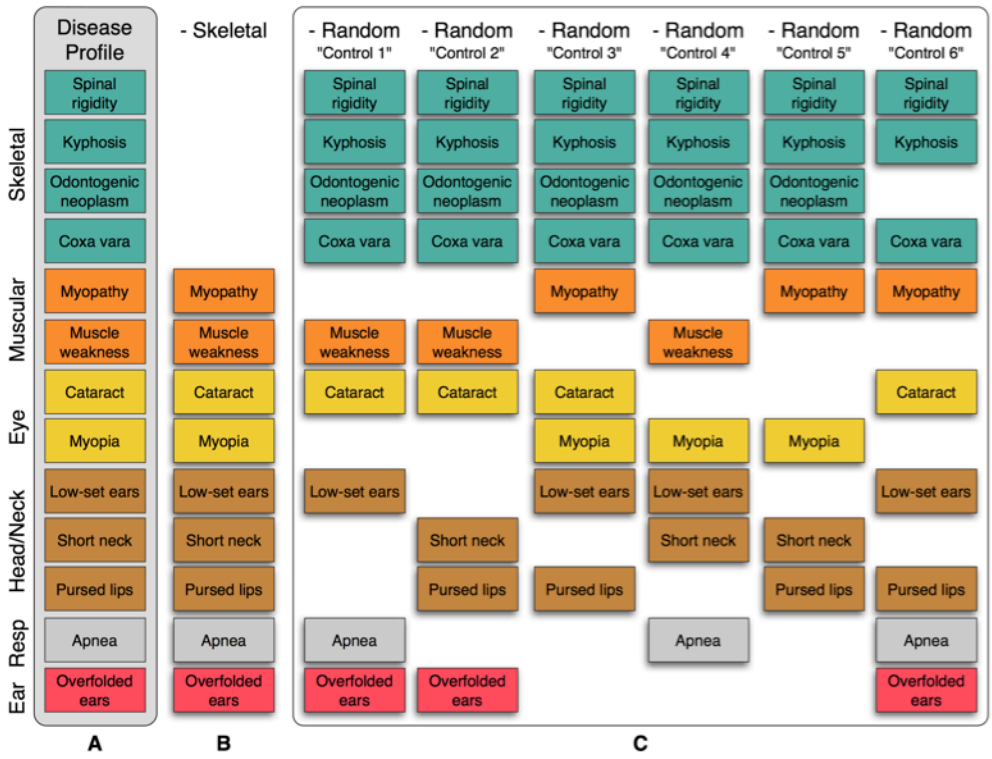

Figure 1. Illustration of original and derived disease-phenotype profiles for Schwartz-Jampel Syndrome, Type I. (A) Original phenotype profile with color-coded phenotype categories. (B) Derived phenotype profile with all skeletal phenotypes ($n=4$) removed. (C) A set of control profiles created by random removal of n annotations. (Only a subset of phenotypes is indicated for illustrative purposes.)

correct for phenotype classes that are subclasses of multiple categories (asserted or inferred).

$$categorical\_score(D) = \frac{\sum_{1}^{c} simple\_score\_per\_category(D)}{number\_of\_categories}$$

We calculate a *scaled score* per profile by incorporating the categorical score in a weighted formula, with the initial δ=0.25:

$$scaled\_score(D) = (1-\delta)(simple\_score(D)) + \delta(categorical\_score(D))$$

The Monarch Initiative REST services currently use the α,β,γ,δ coefficients presented here.

3 RESULTS & DISCUSSION

To explore the creation of metrics to evaluate the sufficiency of a phenotype profile, we first integrated and analyzed semantically curated phenotypic characteristics and their properties of more than 7,500 genetic diseases from OMIM, Decipher, and Orphanet, together with a catalog of approximately 47,000 mouse and 14,000 zebrafish genotypes with curated phenotypes from MGI and ZFIN, respectively.

In order to approximate sub-optimal and/or more-general patient profiles that might be obtained in the clinic, we created a synthetic series of disease-phenotype profiles derived and permuted from the known disease profiles. These derived profiles were compared to all known diseases (including the original "parent" disease) using OWLSim in order to obtain a similarity score and rank. Furthermore, for any derived profile we create a set of control profiles to test for significance of similarity score changes. This method is illustrated in Figure 1 for Schwartz-Jampel Syndrome (OMIM:255800). In order to test the influence of skeletal phenotypes in the phenotype profile (Figure 1A), we created a derived disease by removing all skeletal phenotypes (Figure 1B) from the phenotype set, together with a set of controls where an equivalent number of random non-skeletal phenotypes were removed (Figure 1C). These derived phenotype profiles were then compared to the entire corpus of diseases, including the parent disease. In this example, removing all skeletal phenotypes resulted in a similarity score of 86% when compared to the original disease, as opposed to the controls, which were significantly more similar (with 91+/- 0.78% similarity). This suggests that, for this disease, skeletal phenotypes are significantly more influential compared to random. This result makes intuitive sense because skeletal phenotypes comprise 40% of the phenotypic profile for Schwartz-Jampel Syndrome, and removing them would appear to present a very different disease. However, when compared to all other known diseases, the skeletal-depleted derived disease profile is still more

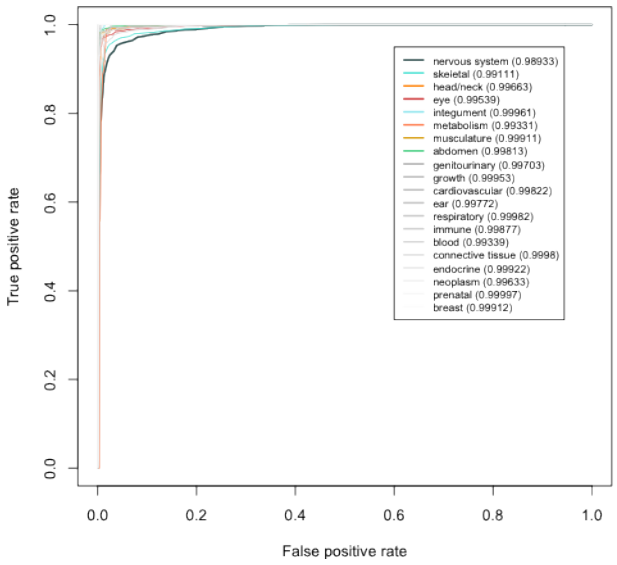

Figure 2. ROC curve indicates robustness of the OWLSim similarity algorithm when entire phenotypic categories were removed. ROCR was performed using similarity scores comparing all derived diseases with all other original disease phenotype profiles. A comparison was classified true if a derived disease was compared to it's parent disease, otherwise it was false. These were grouped into bins and plotted according to the category of phenotypes that was depleted in the derived diseases. AUC was calculated for each category, and ranged from a minimum of 0.9893 for nervous system-depleted to a maximum of 0.99997 for prenatal-depleted profiles.

similar than any other disease in the annotation corpus.

If we take the derived disease profiles created for all multi-categorical diseases (n=5948) and compare them to known diseases, 92% of these derived diseases are still most-phenotypically-similar to their parent-disease. There was little difference in Area Under the Curve (AUC) scores and shape of the ROC curve when assessing the derived disease comparisons to all known diseases for each category (Figure 2). This result suggests that the semantic similarity algorithm and approach are very robust; faced with many missing phenotypes, even entire categories, a suboptimal disease profile is still sufficient to compare and obtain the correct disease.

As described in the Methods, we have implemented a computation of a sufficiency score, available dynamically as a REST service from http://monarchinitiative.org/page/services, which can be utilized by third-party applications. The *scaled score*, which is a measurement of the uniqueness, depth, and complexity of a phenotype profile, is prominently displayed (transformed to 0-5 stars) on any disease, gene, or genotype page in the Monarch website so users can immediately understand how the phenotype profile of a given entity compares against the rest of the corpus. As applied to animal models, it can aid researchers when assessing the quality of a phenotype match; for example a highly-similar cross-species match might be less meaningful if it only has a 2-star sufficiency score (which probably indicates it is poorly annotated). For clinicians, a 5-star graphical display has been added to the Phenotips (www.phenotips.org) interface in order to provide feedback to clinicians recording patient phenotype profiles in the clinic.

We plan to continue our analysis using these same methods to create additional synthetic phenotype profiles for comparison as mentioned in the methods, by varying several factors: overall information content (sumIC) and number of annotations can be tested by simply removing one or more annotations; maximum information content (maxIC) can be tested by removing one or more of the most-significant annotations; specificity of annotations can be tested by "lifting" annotations to more-generic superclasses. Finally, we can take into account the co-occurrence frequency for any pair or set of phenotypes. The additional derived datasets will also help us examine potential limitations of our method that might be due to incompleteness of our baseline set. We will use the results of these analyses to derive optimal weighting coefficients for the different factors in order to refine our initial implementation of the sufficiency score.

4 REFERENCES

Girdea M et al. (2013) PhenoTips: patient phenotyping software for clinical and research use. *Hum Mut.*, **34**, 1057-65

Kohler S. et al (2014a) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucl. Acids Res.* 42 (D1): D966-D974 doi:10.1093/nar/gkt1026

Kohler S. et al (2014b) Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000 Res.* **2**, 30

Robinson P et al (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, **24**, 340-348.

Sing T et al (2005) ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-3941

Smedley D et al (2013) PhenoDigm: Analyzing curated annotations to associate animal models with human diseases *Database* **2013**, bat025

Washington, NL et al (2009) Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation *PLoS Biol.* 10.1371/journal.pbio.1000247