Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research
[v2; ref status: indexed, http://f1000r.es/2td]

Sebastian Köhler¹,², Sandra C Doelken¹, Barbara J Ruef³, Sebastian Bauer¹, Nicole Washington⁴, Monte Westerfield³, George Gkoutos⁵, Paul Schofield⁶, Damian Smedley⁷, Suzanna E Lewis⁴, Peter N Robinson¹,²,⁸, Christopher J Mungall⁴

¹Institute for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Berlin, 13353, Germany
²Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Berlin, 13353, Germany
³ZFIN, Institute of Neuroscience, University of Oregon, Eugene OR, 97403-5291, USA
⁴Lawrence Berkeley National Laboratory, Berkeley CA, 94720, USA
⁵Department of Computer Science, University of Aberystwyth, Aberystwyth, SY23 2AX, UK
⁶Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
⁷Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK
⁸Max Planck Institute for Molecular Genetics, Berlin, 14195, Germany

Abstract
Phenotype analyses, e.g. investigating metabolic processes, tissue formation, or organism behavior, are an important element of most biological and medical research activities. Biomedical researchers are making increased use of ontological standards and methods to capture the results of such analyses, with one focus being the comparison and analysis of phenotype information between species.

We have generated a cross-species phenotype ontology for human, mouse and zebrafish that contains classes from the Human Phenotype Ontology, Mammalian Phenotype Ontology, and generated classes for zebrafish phenotypes. We also provide up-to-date annotation data connecting human genes to phenotype classes from the generated ontology. We have included the data generation pipeline into our continuous integration system ensuring stable and up-to-date releases.

This article describes the data generation process and is intended to help interested researchers access both the phenotype annotation data and the associated cross-species phenotype ontology. The resource described here can be used in sophisticated semantic similarity and gene set enrichment analyses for phenotype data across species. The stable releases of this resource can be obtained from http://purl.obolibrary.org/obo/hp/uberpheno/.
Corresponding authors: Sebastian Köhler (sebastian.koehler@charite.de), Christopher J Mungall (CJMungall@lbl.gov)

How to cite this article: Köhler S, Doelken SC, Ruef BJ et al. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research [v2; ref status: indexed, http://f1000r.es/2td] F1000Resarch 2014, 2:30 (doi: 10.12688/f1000research.2-30.v2)

Copyright: © 2014 Köhler S et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by grants of the Deutsche Forschungsgemeinschaft (DFG RO 2005/4-1), the Bundesministerium für Bildung und Forschung (BMBF project number 0313911), the MGD grant from the National Institutes of Health, HG000330, the ZFIN grant from the National Institutes of Health, U41-HG002659, and the grants from the National Institutes of Health, R01-HG004838 and R24-OD011883.

Competing interests: No competing interests were declared.

First published: 01 Feb 2013, 2:30 (doi: 10.12688/f1000research.2-30.v1)
First indexed: 11 Feb 2013, 2:30 (doi: 10.12688/f1000research.2-30.v1)
Introduction

Research on model organisms is crucial for discovering the function of genes and DNA elements and for understanding the phenotypic effects of mutations on these genes, which is leading to a better understanding of the pathobiology of human disease\(^1\). The amount of phenotypic information derived from targeted mutations and hypothesis-driven studies is increasing rapidly, and is now being further augmented by high-throughput international efforts to systematically analyse the effects of genomic variation on model organism phenotypes. For example, the International Mouse Phenotyping Consortium (IMPC\(^1\)), is undertaking systematic phenotyping studies of the knockouts generated by the International Knockout Mouse Consortium (IKMC\(^1\)). This means that there will soon be structured phenotype data for loss-of-function mutants for every protein-coding gene in the mouse. Similar approaches are being taken in zebrafish (Danio rerio) by the Zebrafish Mutation Project (ZMP, http://www.sanger.ac.uk/Projects/D_reario/zmp/) and the data is being made available through the Zebrafish Model Organism Database (ZFIN\(^1\)).

Model organism phenotype/genotype datasets are extremely valuable as they can provide clues to human gene functions and involvement in disease processes where no data is available for the human ortholog. At the time of writing, 2,358 human genes are associated with Mendelian phenotypes, but more importantly there are 5,492 human genes with no such phenotype associations, where an orthologous mouse or zebrafish gene does have phenotype data (Data obtained by analysing the file HSgenes_crossSpeciesPhenoAnnotation.txt from http://purl.obolibrary.org/obo/hp/uberpheno/). We have previously demonstrated the power of this approach in determining likely pathogenicity of genes within the intervals of recurrent copy number variation (CNV) diseases\(^6\) and it can be applied much more widely in, for example, prioritizing candidate genes identified through human genome wide association studies (GWAS)\(^7\). Historically, a major problem has been the lack of common semantics across databases, with each project using some combination of free-text descriptions or in-house vocabularies. Thus, phenotype information is not easily integrated across different species. This inhibits comparisons based on phenotype alone, and where orthology is useful phenotypic comparisons cannot be used to their full potential. This is made even more complicated by different conceptualizations of phenotypes in different species and the impact of species-specific anatomies. As the ability of investigators to mobilise this growing collection of model organism data has become more important, it is crucial to develop appropriate ontologies and computational strategies to describe phenotypes such that phenotype descriptions can be objectively related to each other, both within and between species. This becomes even more important as the divergence between the number of human genes with phenotype information and the amount of systematically phenotyped model organism genes is expected to increase in the near future due to high throughput-screens\(^1\).

The application of controlled vocabularies and ontologies has accelerated over recent years; the Gene Ontology (GO) being probably the most successful example in the field of biomedical ontologies. Many other ontologies exist, each of which has been developed for a specific domain in biomedicine. Now a major goal is to increase semantic and syntactic interoperability between those ontologies (e.g. the Open Biomedical Ontologies (OBO) Foundry\(^9\)). One approach is to develop ontologies by defining complex ("pre-composed") classes in terms of other more elementary (atomic) classes (building blocks) that are species-agnostic. If several ontologies make use of shared building block ontologies, interoperability can be facilitated across a larger domain. For example ontologies that contain classes concerned with DNA-replication in different organisms or cells should refer to a shared class representing DNA-replication-process, enabling computers to detect that the same class is referenced.

We have previously shown how phenotype information can be linked and used in cross-species phenotype analyses\(^10\). A crucial part of this strategy is the use of logical definitions to render ontology terms in a way that is computable. Recently, logical definitions of terms representing classes of phenotypic deviations have been developed by several groups. Developers of OBO Foundry ontologies, such as the GO\(^9\), the Mammalian Phenotype Ontology (MPO\(^10\)), the Human Phenotype Ontology (HPO\(^8\)), the Worm Phenotype Ontology\(^9\), and also the Cell Ontology\(^21\), are now creating logical definitions of their ontology-classes using terms from other building block ontologies. In this effort the Phenotype, Attribute and Trait Ontology (PATO), an ontology of phenotypic qualities, is a key tool\(^9,22\). Examples for building block ontologies that are used for the representation of classes of phenotypic abnormalities are given in the upper part of Table 1.

Objectives

Given that logical definitions exist for most classes of an ontology, automatic reasoners can be applied. These implement algorithms for computing the logical consequences that can be inferred from a set of asserted axioms. An example can be seen in Figure 1a), where logical definitions are used to automatically infer that Hypoglycemia is a subclass of Decreased aldohexose concentration (blood) based on the asserted subclass relationship between 'glucose' and 'aldohexose' in ChEBI. This means that reasoners are able to use computable, logical definitions to infer the positions of classes in a subsumption hierarchy. Thus, those definitions can be helpful tools for the development and maintenance of ontologies\(^23\).

Although several methods, ideas, and applications on cross-species phenotype integration have been presented before\(^9,10,14,16,25\), accessing such data resources has been complicated by the lack of consistent documentation and distribution of data across heterogeneous resources. For example, some ontologies are provided in the Web Ontology Language (OWL\(^26\)) and others in the Open Biomedical Ontologies (OBO) format. Although the OBO-format
Table 1. Typical building block ontologies: here the focus lies on ontologies that can be used to represent complex classes of phenotype abnormalities in zebrafish, mouse, and human.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Name (Abbreviation, Reference)</th>
<th>Downloaded file (relative to <a href="http://purl.obolibrary.org/obo/">http://purl.obolibrary.org/obo/</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>biochemistry</td>
<td>Chemical Entities of Biological Interest (ChEBI)</td>
<td>chebi.obo</td>
</tr>
<tr>
<td></td>
<td>Gene Ontology (GO)</td>
<td>go.obo</td>
</tr>
<tr>
<td>proteins</td>
<td>Protein Ontology (PRO)</td>
<td>pr.obo</td>
</tr>
<tr>
<td>cell types</td>
<td>Cell Ontology (CL)</td>
<td>cl.obo</td>
</tr>
<tr>
<td>anatomy</td>
<td>Foundational Model of Anatomy (FMA)</td>
<td>fma.obo</td>
</tr>
<tr>
<td></td>
<td>Spatial Ontology (BSPO)</td>
<td>bsopo.obo</td>
</tr>
<tr>
<td></td>
<td>Mouse adult gross anatomy (MA)</td>
<td>ma.obo</td>
</tr>
<tr>
<td></td>
<td>Zebrafish anatomy and development (ZFA)</td>
<td>zfa.ob</td>
</tr>
<tr>
<td></td>
<td>Multi-species anatomy (UBERON)</td>
<td>uberon.obo</td>
</tr>
<tr>
<td>phenotype</td>
<td>Phenotype, Attribute and Trait Ontology (PATO)</td>
<td>pato.obo</td>
</tr>
<tr>
<td></td>
<td>Mouse Pathology (MPATH)</td>
<td>mpath.obo</td>
</tr>
<tr>
<td></td>
<td>Mammalian Phenotype Ontology (MPO)</td>
<td>mp.obo</td>
</tr>
<tr>
<td></td>
<td>Human Phenotype Ontology (HPO)</td>
<td>hp.obo</td>
</tr>
<tr>
<td></td>
<td>Neuro Behavior Ontology (NBO)</td>
<td>nbo.obo</td>
</tr>
</tbody>
</table>

Figure 1. Part a) illustrates the main idea how logical definitions and building block ontologies (left) cooperate in order to allow for reasoning procedures to infer new knowledge (right). Note that for the purpose of increased readability, only the term labels are shown and the ontology Uniform Resource Identifier (URIs) are skipped. Part b) illustrates an excerpt of the Uberpheno ontology to show how information on phenotype abnormalities in different organisms can be combined. It also illustrates how the annotations of genes can be transferred across different species by means of orthology relationships of genes. For example, after reasoning one could easily request all genes that are known to be related to the phenotype description “Bilateral microphthalmos” from the HPO. In Uberpheno “abnormally hypoplastic eye” from zebrafish (ZP) and “posterior microphthalmia” from MPO, are inferred to be subclasses of “Bilateral microphthalmos”. These inferences can be used to infer that the genes *tcf7l1a* (zebrafish) and *PRSS6* (mouse) are annotated to the phenotype “Bilateral microphthalmos” as well.
focuses especially on human readability and ease of parsing, OWL is often needed to enable complex reasoning tasks. Unfortunately, the power and complexity of OWL may discourage some researchers.

For example, the OWLSim package (http://owlsim.org) provides the ability to execute a number of standard semantic similarity techniques. Although access to the results of OWLSim in phenotype analyses is available (23, http://www.mousemodels.org), there is at the moment no single set of gene annotations linked to a single integrated ontology.

The Uberpheno-ontology is similar to the “phene.owl” ontology distributed as part of the phenomeblast-project (http://code.google.com/p/phenomeblast/) and generated as part of a phenotype data analysis executed within PhenomeNET(24). These two ontologies differ in a number of characteristics. The first characteristic is the underlying OWL model, and the set of external ontologies that are brought in to enrich the ontology - it is not yet clear how far the OWL model or some of these external ontologies affect the resulting structure of the ontology. Also it is likely that both Uberpheno and “phene.owl” will converge on the same model and a standard set of imported ontologies. The second characteristic is the breadth of species covered, with “phene.owl” including fly, worm and yeast; in contrast, Uberpheno focuses on human, mouse and zebrafish, yielding a smaller more focused ontology. Further investigations are required to determine the extent to which the adding of more distant organisms help or hinder analyses. Another difference is that Uberpheno is intended for a wide range of biomedical researchers, some of who may be unfamiliar with OWL or OWL reasoning.

Our objective here is to provide an OBO-format ontology (Uberpheno), which we update at regular intervals and which can easily be used for downstream analysis, e.g. by applying semantic similarity measures(25) or gene set enrichment analyses(26). Of similar importance are the data that link into such an ontology by means of the annotation relation. To the best of our knowledge, no single integrated cross-species ontology together with annotation of all genes in human and model organisms (here mouse and zebrafish) has been made easily available for researchers and kept up-to-date on a regular basis.

Materials and methods

Model organism data

Cross-species ontology-based approaches offer a promising new methodology to reliably detect phenotypic similarities between human disease manifestations and model organism phenotypes(6,11,24,25). They can pave the way to gain clinically relevant insights from the almost 5,500 genes for which, currently, only mouse and zebrafish phenotype information is available. Both the Mouse Genome Informatics (MGI) and the ZFIN data resources provide manually curated assignments of their model organism genes to human genes. They are available from the corresponding website (see Table 2).

The annotation of genes to phenotypes are also accessible online. Zebrafish genes are annotated by Entity-Quality (EQ) statements. Mouse genes are annotated with terms from the MPO and are downloadable from the MGI website. To associate human genes with terms from the HPO, the annotation of human diseases is required. By using further files from OMIM (http://omim.org) and Orphanet(27), (http://www.orphadata.org/) diseases can be mapped to the disease-causing genes. These two steps allow the transfer of phenotype information to the underlying genes. All required files and their corresponding links are summarized in Table 2.

Phenotype descriptions

The approach taken to logically define phenotype descriptions is termed the Entity-Quality approach (EQ), in which phenotype descriptions can be partitioned into (minimally) two parts. The first part represents the affected entity, i.e. the thing for which an observation is made. This can be entities of various domains, e.g., a chemical or an anatomical structure. The second part represents the quality of the entity and is described in a qualitative or quantitative way(22). In the typical setting, a phenotype is described using a class expression consisting of a PATO quality class differentiated by a bearer entity class using the inheres_in relation from the OBO Relation Ontology(28). To give an example for logical definitions, consider the HPO term Hypoglycemia and its EQ definition, specified in OWL as shown in Figure 1 (center).

The word Hypoglycemia refers to an abnormally decreased concentration of glucose in the blood. The logical definition uses relations and follows the pattern described in previous work on the definition of phenotypes(24). The logical semantics are made explicit

<table>
<thead>
<tr>
<th>Type</th>
<th>Organism</th>
<th>Obtain from</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthology to human genes</td>
<td>Mouse</td>
<td>ftp://ftp.informatics.jax.org/pub/reports/HMD_HumanPhenotype.rpt</td>
</tr>
<tr>
<td>Phenotype annotation</td>
<td>Mouse</td>
<td><a href="http://zfin.org/downloads/ortho.txt">http://zfin.org/downloads/ortho.txt</a></td>
</tr>
<tr>
<td></td>
<td>Zebrafish</td>
<td><a href="http://zfin.org/downloads/pheno.txt">http://zfin.org/downloads/pheno.txt</a></td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td><a href="http://compbio.charite.de/hudson/job/hpo.annotations/lastStableBuild/artifact/misc/phenotype_annotation.tab">http://compbio.charite.de/hudson/job/hpo.annotations/lastStableBuild/artifact/misc/phenotype_annotation.tab</a></td>
</tr>
<tr>
<td>Gene-to-disease</td>
<td>Human</td>
<td>&lt;OMIM ftp-site&gt;\mim2gene.txt and &lt;OMIM ftp-site&gt;/genemap</td>
</tr>
</tbody>
</table>
when translating the definitions to OWL. Currently, the translation to OWL is performed using a “has part some”-semantics implemented in the OBO-format library (http://code.google.com/p/oboformat). The translation is shown in Manchester syntax in Figure 1a). In the example, the class Hypoglycemia is defined as the equivalent of the intersection of all classes of things that are “A concentration which is lower relative to the normal” (decreased concentration), “deviate from the normal or average” (abnormal), with respect to (towards) glucose, and inhering in “blood” (using the term portion of blood from the FMA). More details can be found in16 or17. Automated reasoning logically infers then that the asserted knowledge in ChEBI induces Hypoglycemia to be a subclass of Decreased aldohexose concentration (blood). The files used to define phenotype classes are summarized in Table 3.

Uberpheno construction

The general work- and data-flow of the cross-species ontology generation is illustrated in Figure 2. In steps one to three, the aforementioned EQ definitions are used to generate a single cross-species phenotype ontology (Uberpheno) for human, mouse, and zebrafish phenotypes. Step four generates files that make it very convenient to use the generated data for several research purposes, because genes are linked to the terms of the generated cross-species phenotype ontology, which is very lightweight and available in the convenient OBO-format.

Step 1. Logical definitions are being developed for GO16, MPO13, and HPO.19. Almost all logical definitions refer to classes from other ontologies. A set of logical definitions is again an ontology itself. These bridging ontologies (also called cross-product files) are available on the main OBO Foundry website, as well as from the individual repositories for each of the projects. An example for a logical definition is presented in the previous section and in Figure 1. A major fraction of HPO and MPO terms are currently defined by means of EQ statements and a summary of the logical definition files that are used can be found in Table 3. These files provide axioms that connect phenotype classes to multiple classes in most of the ontologies listed in Table 1.

The HPO and MPO logical definitions were augmented with pairwise equivalence axioms generated by lexical matching. These mappings are represented in a file mp_hp-align-equiv.owl (see the phenotype ontologies archive on Google code at http://code.google.com/p/phenotype-ontologies). A total of 1,864 such lexically derived equivalence axioms were derived in this way and used to supplement the semantic analysis.

In step one, all of the required files are pulled from the web (see Table 1 and Table 3). Note, that there are ontologies that are required in their entirety (denoted (B) in Figure 2). In contrast, several building block ontologies (denoted (A) in Figure 2) are only referred in parts by the logical definitions.

When defining phenotypes using the EQ model, the affected entity can either be a biological function or process from GO, or an anatomical entity. Some of the ontologies used to create the definitions are largely species-independent (GO, ChEBI). However, anatomically entities are mostly defined by referring anatomy ontologies that are specific for one species. In order to enable reasoning across these vertebrate anatomies, the metazoan, species-independent Uberon ontology is used in constructing anatomically-based cross-products16. In order to construct Uberpheno, an equivalence axiom was generated between every class in Uberon that contains a cross-reference to a species anatomy ontology class. Note that very general terms from Uberon such as tissue are excluded, which can be identified by their membership to the subset upper_level in Uberon. The generated file is called uberonbridge.owl.

One of the files (see Table 3) defines GO process terms by the anatomy term to which the process is related. For example,

Class: eye pigmentation
EquivalentTo: pigmentation andoccurs_in some eye

Here, the GO process eye pigmentation (GO:0048069) is logically defined as being equivalent to everything that is a pigmentation (GO:0043473) and also “occurs_in” an eye (UBERON:0000970). In order to use these definitions, the different relationships used therein, such as occurs_in, are made interpretable for the reasoner. For our purposes, an additional ontology called http://comphio.charite.de/svn/hpo/trunk/misc/go_xp_misc/extra_equiv.owl was created in which these relationships are made a subPropertyOf of inheres_in.

Step 2. In step two a data preprocessing is required, because for zebrafish no pre-composed ontology of phenotype abnormalities exists (e.g. no phenotype term such as abnormally hypoplastic eye exists). Instead, the ZFIN project makes use of so-called “post-composed” annotations, using a combination of classes in the EQ model. The ZFIN file pheno.txt (Table 2) contains lines such as


<table>
<thead>
<tr>
<th>Ontology</th>
<th>File</th>
<th>Number of classes defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPO logical definitions</td>
<td>hp-equivalence-axioms.obo</td>
<td>4,666</td>
</tr>
<tr>
<td>MPO logical definitions</td>
<td>mp-equivalence-axioms.obo</td>
<td>7,278</td>
</tr>
<tr>
<td>GO logical definitions using Uberon</td>
<td>biological_process_xp_uber_anatomy.obo</td>
<td>1,484</td>
</tr>
<tr>
<td>Behavior xp</td>
<td>behavior_xp.obo</td>
<td>104</td>
</tr>
</tbody>
</table>
For legibility the tab-separators are replaced in this example by the semicolon. In order to use these annotations for reasoning, a translation table was implemented, as described before\[^1\], which generates the ontology denoted as \(zp\text{.}owl\). For every modified gene, a set of post-composed phenotype annotations is stored in \(pheno\text{.}txt\). For every unique annotation for zebrafish genes, a class in the ZP identifier space is created. Again, the aforementioned "has part some"-translation to OWL is applied. For example, a zebrafish gene annotation with

\[
\text{Entity=ZFA:0000107 (eye),} \\
\text{Quality=PATO:0000645 (hypoplastic) and} \\
\text{Qualifier=PATO:0000460 (abnormal)}
\]

generates an OWL class:

\[
\text{Class: ZP_0003395} \\
\text{Annotations: label "abnormal(ly) hypoplastic eye"} \\
\text{EquivalentClassOf:}
\]

Beside generating the ZP-ontology, the annotation relation between the zebrafish genes and ZP-term is written to a file called \(zp\text{.}annot\), which is also available for download.

Since some logical definitions of phenotypes are lacking the qualifier abnormal we ensure consistency, by adding this qualifier to all of the definitions. We also remove the inconsistently used ontology-tags from the xp-files.

Steps 3 and 4. At first, a single, merged OWL ontology is created from all the ontologies and bridging axioms. The ELK reasoner\[^4\] is used to calculate subclass and equivalence relationships between classes. These steps are implemented within the GULO framework\[^2\].

To increase the usability of the ontology, the Ontologizer API\[^28\] was used to merge all clusters of equivalent classes together into a single class. The HPO identifier is taken as the primary identifier.
if present and the identifiers of other phenotype classes are stored under \textit{alt}_id\text{-}tag for the term. For example, the HPO-term Gallbladder dysfunction (HP:0005609) has as \textit{alt}\_id\_the ZP-term abnormal(ly) decreased functionality gall bladder (ZP:0004170). The resulting ontology in OBO-format is named \textsf{crossSpecies-Pheno.obo} and contains only phenotype classes from the HPO, MPO, and ZP.

Finally a cross-species annotation file is generated, in which all human genes are associated with terms from the \textit{Uberpheno}. The annotations are either stemming from human or model organisms, whereby the model organism annotations are stemming from the ortholog gene.

**Results and discussion**

All of the above described methods are integrated into a single pipeline. This pipeline automatically downloads required files, preprocesses the data and applies a reasoning procedure to the obtained set of ontology classes. The ontologies used to construct \textit{Uberpheno} are summarized in Table 1.

The construction pipeline is set up as a job in our continuous integration system accessible at \url{http://compbio.charite.de/hudson}, which is already used for data related to the HPO\cite{1}. The job (called \textsf{hpo.ontology.uberpheno}) is configured to run once a week, ensuring that the most recent version of all ontologies and annotation files are used. Only stable releases of the generated files are made available to the users and errors are immediately forwarded to us via email. The generated build artifacts are available at \url{http://purl.obolibrary.org/obo/hp/uberpheno/}, whereas the file \textsf{crossSpecies-Pheno.obo} contains the cross-species phenotype ontology in OBO-format. The resulting ontology has a light footprint (3.5 MB) and can easily be explored by using tools such as example OBO-Edit\cite{2}. Note that only phenotype classes are present in the ontology and classes from the referenced building block ontologies are filtered out. Each build also generates the file \textsf{HSgenes_crossSpeciesPheno-Annotation.txt}, which contains the annotation of all human genes to terms of HPO, MPO, and ZP. A summary of the data contained in the two files is given in Table 4.

An excerpt of the \textit{Uberpheno} ontology is shown in Figure 1b), demonstrating how the phenotype descriptions from different ontologies are combined and automatically organised into a single integrated hierarchy. For instance, the fact that the mouse term \textit{posterior microphthalmia} is inferred to be a subclass of the human term \textit{Bilateral microphthalmos} can be used to transfer the information that the mouse gene PRSS56 is known to cause \textit{Bilateral microphthalmos}. This implies that querying the cross-species ontology for genes related to \textit{Bilateral microphthalmos} will return the human gene \textit{TCOF1}, the mouse gene \textit{PRSS56} and the zebrafish gene \textit{tcf7l1a}.

In total, the annotation file contains approx. 235,000 annotations of human genes with phenotype classes (see Table 4). For example the human gene \textit{TCF7L1} is associated with the zebrafish phenotype abnormal(ly) hypoplastic eye because the ortholog zebrafish gene \textit{tcf7l1a}, ZDB-GENE-980605-30 is annotated with this phenotype. Thus, the generated file \textsf{HSgenes_crossSpeciesPhenoAnnotation.txt} contains the line:

\begin{verbatim}
83439;TCF7L1;abnormal(ly) hypoplastic eye (ZP:0003395);tcf7l1a(ZDB-GENE-980605-30/ZEBRAF)
\end{verbatim}

**Table 4. Statistics of the build artifacts generated (build #63).** ‘Phenotype classes’ denotes the number of classes that are either from the Human Phenotype Ontology (HPO), Mammalian Phenotype Ontology (MPO), or \textit{zp.owl} (ZP). Note that the sum of HPO-, MPO-, and ZP-IDs is higher than the total number total ‘Phenotype classes’ because some MPO- and ZP-IDs are listed as \textit{alt}_id of an HPO-class and are not listed as separate ‘Phenotype class’. Also, the number of human annotations is less than the sum of annotations supported by OMIM or Orphanet entries, because some annotations have evidence from both databases.

<table>
<thead>
<tr>
<th>Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uberpheno statistics:</strong></td>
<td></td>
</tr>
<tr>
<td>Phenotype classes:</td>
<td>25,974</td>
</tr>
<tr>
<td>HPO-IDs</td>
<td>13,122</td>
</tr>
<tr>
<td>MPO-IDs</td>
<td>9,800</td>
</tr>
<tr>
<td>ZP-IDs</td>
<td>8,057</td>
</tr>
<tr>
<td><strong>Annotation statistics:</strong></td>
<td></td>
</tr>
<tr>
<td>All annotations</td>
<td>235,752</td>
</tr>
<tr>
<td>HPO annotations</td>
<td>63,080</td>
</tr>
<tr>
<td>-OMIM</td>
<td>49,348</td>
</tr>
<tr>
<td>-Orphanet</td>
<td>16,244</td>
</tr>
<tr>
<td>MPO annotations</td>
<td>149,164</td>
</tr>
<tr>
<td>ZP annotations</td>
<td>23,508</td>
</tr>
</tbody>
</table>

**Conclusions**

The phenotype resources for mouse, zebrafish, and human are used by several research projects\cite{34-36}.

The problem of comparing phenotypes between species can be overcome by using formal logical definitions that make use of species agnostic ontologies together with a multi-species anatomy ontology, Uberon. The approach to implementing the paradigm that we report in this paper constructs a single, integrated, cross-species phenotype ontology, \textit{Uberpheno}, based on the logical definitions of human and the main model species, mouse and zebrafish. The resulting construct is continuously updated and automatically constructed as the constituent ontologies are updated and augmented, making it a dynamic and current resource available to the community.

Increasingly model organism data are being used for gene set enrichment, pathogenicity prediction and semantic similarity analyses\cite{37} and the high throughput phenotyping projects newly underway promise rich genome-wide phenotypic coverage within a decade. This will complement the new initiatives to systematically gather high precision, formally coded, phenotype data from clinical studies\cite{38}. The promise that all this data holds can only be realized if the informatics tools are available to handle and analyse this rich resource and we believe that \textit{Uberpheno} is an accessible and widely applicable resource with which this may be achieved.
Author contributions
SK, CJM, SEL, PS and PNR conceived the study. SK, SB, CJM and DS set up the code to create the ontology and the annotation files. BJR, SCD, DS, NW, GVG, PS and MW helped with the data preparation and processing. SK, CJM, PS, BJR, PNR and NW wrote the manuscript. All authors read and approved the manuscript.

Competing interests
No competing interests were declared.

Grant information
This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by grants of the Deutsche Forschungsgemeinschaft (DFG RO 2005/4-1), the Bundesministerium für Bildung und Forschung (BMBF project number 0313911), the MGD grant from the National Institutes of Health, HG000330, the ZFIN grant from the National Institutes of Health, U41-HG002659, and the grants from the National Institutes of Health, R01- HG04838 and R24-OD011883.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
We would like to thank Anika Oellrich for extensive proofreading of the draft version of the manuscript.

References


Open Peer Review

Current Referee Status: ☑️ ☑️ ☑️

Version 2

Referee Report 27 March 2014
doi:10.5256/f1000research.3649.r4261

Larry Hunter
Center for Computational Pharmacology & Computational Bioscience Program, UC Denver, Aurora, CO, USA

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 23 January 2014
doi:10.5256/f1000research.3649.r3279

Nicola Mulder
Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Cape Town, South Africa

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 22 January 2014
doi:10.5256/f1000research.3649.r3271

Francisco Couto
Large-Scale Informatics Systems Laboratory (LASIGE) Lisbon, University of Lisbon, Campo Grande, Lisbon, Portugal

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 12 February 2013
doi:10.5256/f1000research.905.r764

Nicola Mulder
Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Cape Town, South Africa

This paper describes a cross-species phenotype ontology, which promises to be extremely useful. It was not clear to me how much of the data preparation was purely computational versus some biological input. One has to be wary of trying to make too many terms apply to multiple species, but given the collective experience of these authors I am sure this has been taken into consideration.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 11 February 2013
doi:10.5256/f1000research.905.r761

Francisco Couto
Large-Scale Informatics Systems Laboratory (LASIGE) Lisbon, University of Lisbon, Campo Grande, Lisbon, Portugal

The availability of a cross-species phenotype ontology built as a mashup of different other ontologies is a great contribution to the community and results from the implementation of a well-designed integration process. The process combines knowledge from different sources taking in account their provenance, which guarantees the continuous update of the proposed ontology. The successful exploitation of logical definitions is also a very good contribution to promoting the usage of more formal definitions which I believe will, for example, help to enhance the reliability of semantic similarity and enrichment analyses.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 08 February 2013
doi:10.5256/f1000research.905.r757
Larry Hunter  
Center for Computational Pharmacology & Computational Bioscience Program, UC Denver, Aurora, CO, USA

This is really nice work, clearly demonstrating the value of logical definitions of ontology terms and of inference over those definitions. The automatic updating of the inference, ensuring that UberPheno reflects current annotations (and definitions, which change less frequently) should be adopted as a ‘best practice’ by other providers of derived information.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

**Competing Interests:** No competing interests were disclosed.