WEB TOOLS

Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research

[v1; ref status: indexed, http://f1000r.es/p5]

Sebastian Köhler[1,2], Sandra C Doelken[1], Barbara J Ruef[3], Sebastian Bauer[1], Nicole Washington[4], Monte Westerfield[3], George Gkoutos[5], Paul Schofield[6], Damian Smedley[7], Suzanna E Lewis[4], Peter N Robinson[1,2,8], Christopher J Mungall[4]

[1]Institute for Medical and Human Genetics, Chairité-Universitatsmedizin Berlin, Berlin, 13353, Germany
[2]Berlin-Brandenberg Center for Regenerative Therapties (BCRT), Charité-Universitatsmedizin Berlin, Berlin, 13352, Germany
[3]ZFIN, Institute of Neuroscience, University of Oregon, Eugene OR, 97403-5291, USA
[4]Lawrence Berkeley National Laboratory, Berkeley CA, 94720, USA
[5]Department of Computer Science, University of Aberystwyth, Aberystwyth, SY23 2AX, UK
[6]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
[7]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK
[8]Max Planck Institute for Molecular Genetics, Berlin, 14195, Germany

v1 **First Published:** 01 Feb 2013, **2**:30 (doi: 10.12688/f1000research.2-30.v1)
Latest Published: 01 Feb 2013, **2**:30 (doi: 10.12688/f1000research.2-30.v1)

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 zebra fish that contains zebrafi sh 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/.

Article Status Summary

Referee Responses

Referees	1	2	3
v1 published 01 Feb 2013	✓ report	✓ report	✓ report

1 **Larry Hunter**, UC Denver USA

2 **Francisco Couto**, University of Lisbon Portugal

3 **Nicola Mulder**, Institute of Infectious Disease and Molecular Medicine South Africa

Latest Comments

No Comments Yet

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.* (2013) Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research [v1; ref status: indexed, http://f1000r.es/p5] *F1000Research* 2013, **2**:30 (doi: 10.12688/f1000research.2-30.v1)

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

Competing Interests: No competing interests were disclosed.

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[1,2]. 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,[3]), is undertaking systematic phenotyping studies of the knockouts generated by the International Knockout Mouse Consortium (IKMC,[4]). 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_rerio/zmp/) and the data is being made available through the The Zebrafish Model Organism Database (ZFIN,[5]).

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)[7,8]. 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,[9]) 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,[10]). 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[11–15]. 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[16], the Mammalian Phenotype Ontology (MPO,[17]), the Human Phenotype Ontology (HPO,[18,19]), the Worm Phenotype Ontology[20], 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[19,22]. 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 1 a), 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[16,23].

Although several methods, ideas, and applications on cross-species phenotype integration have been presented before[11,12,16,24,25],

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.

Domain	Name (Abbreviation, Reference)	Downloaded File (relative to http://purl.obolibrary.org/obo/)
biochemistry	Chemical Entities of Biological Interest (ChEBI,[29])	chebi.obo
	Gene Ontology (GO,[30])	go.obo
proteins	Protein Ontology (PRO,[31])	pr.obo
cell types	Cell Ontology (CL,[32])	cl.obo
anatomy	Foundational Model of Anatomy (FMA,[33])	fma.obo
	Spatial Ontology (BSPO,-)	bspo.obo
	Mouse adult gross anatomy (MA,[34])	ma.obo
	Zebrafish anatomy and development (ZFA,[35])	zfa.ob
	Multi-species anatomy (UBERON,[36])	uberon.obo
phenotype	Phenotype, Attribute and Trait Ontology (PATO,[22])	pato.obo
	Mouse Pathology (MPATH,[37])	mpath.obo
	Mammalian Phenotype Ontology (MPO,[17])	mp.obo
	Human Phenotype Ontology (HPO,[18])	hp.obo
	Neuro Behavior Ontology (NBO,[38])	nbo.obo

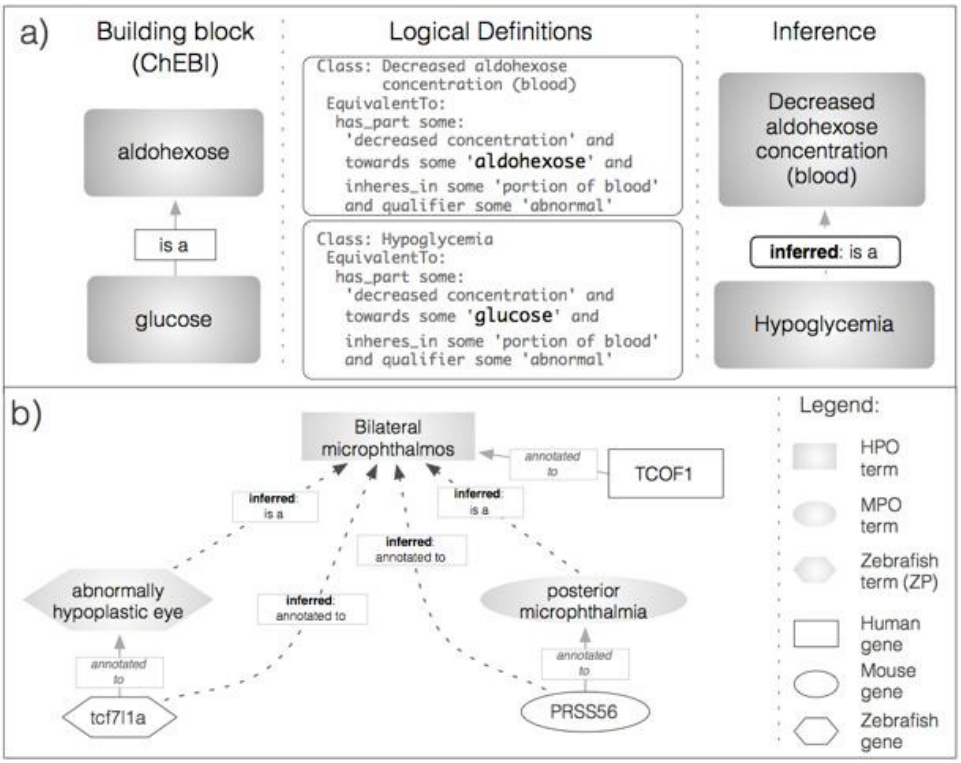

Figure 1 Part **a)** illustrates the main idea how logical definitions and building block ontologies (left) cooperate in order to allow for reasoning procedures to infer new knowledge (right). Note that for the purpose of increased readability, only the term labels are shown and the ontology Uniform 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 *PRSS56* (mouse) are annotated to the phenotype "Bilateral microphthalmos" as well.

accessing such data resources has been complicated by the lack of consistent documentation and distribution of data across heterogenous 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 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 ([25], 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[27] or gene set enrichment analyses[28]. 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 ([39], 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

Table 2 Files required to connect genes and phenotypes as well as to get the orthology relationship between model organism genes and human genes. These files are especially important for Step 4 in Figure 2.

Type	Organism	Obtain from
Orthology to human genes	Mouse	ftp://ftp.informatics.jax.org/pub/reports/HMD_HumanPhenotype.rpt
	Zebrash	http://zfin.org/downloads/ortho.txt
Phenotype annotation	Mouse	ftp://ftp.informatics.jax.org/pub/reports/MGI_PhenoGenoMP.rpt
	Zebrash	http://zfin.org/downloads/pheno.txt
	Human	http://compbio.charite.de/hudson/job/hpo.annotations/lastStableBuild/artifact/misc/phenotype_annotation.tab
Gene-to-disease	Human	<OMIM ftp-site>/**`mim2gene.txt`** and <OMIM ftp-site>/**`genemap`**
	Human	http://www.orphadata.org/data/xml/en_product6.xml

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[40]. 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[16]. The logical semantics are made explicit 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 1 a). 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 in[16] or[23]. 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 GO[16], MPO[12], 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 1,064 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 Tables 1 and 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, anatomical 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-products[36]. 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.**

Table 3 Current statistics on the data contained in the used cross-product files. HPO and MPO files downloaded from http://code.google.com/p/phenotype-ontologies. **Behaviour files downloaded from** http://code.google.com/p/behavior-ontology. **GO-xp file downloaded from** http://obofoundry.org/cgi-bin/detail.cgi?id=biological_process_xp_uber_anatomy.

Ontology	File	Number of classes defined
HPO logical definitions	**hp-equivalence-axioms.obo**	4,666
MPO logical definitions	**mp-equivalence-axioms.obo**	7,278
GO logical definitions using Uberon	**biological_process_xp_uber_anatomy.obo**	1,484
Behavior xp	**behavior_xp.obo**	104

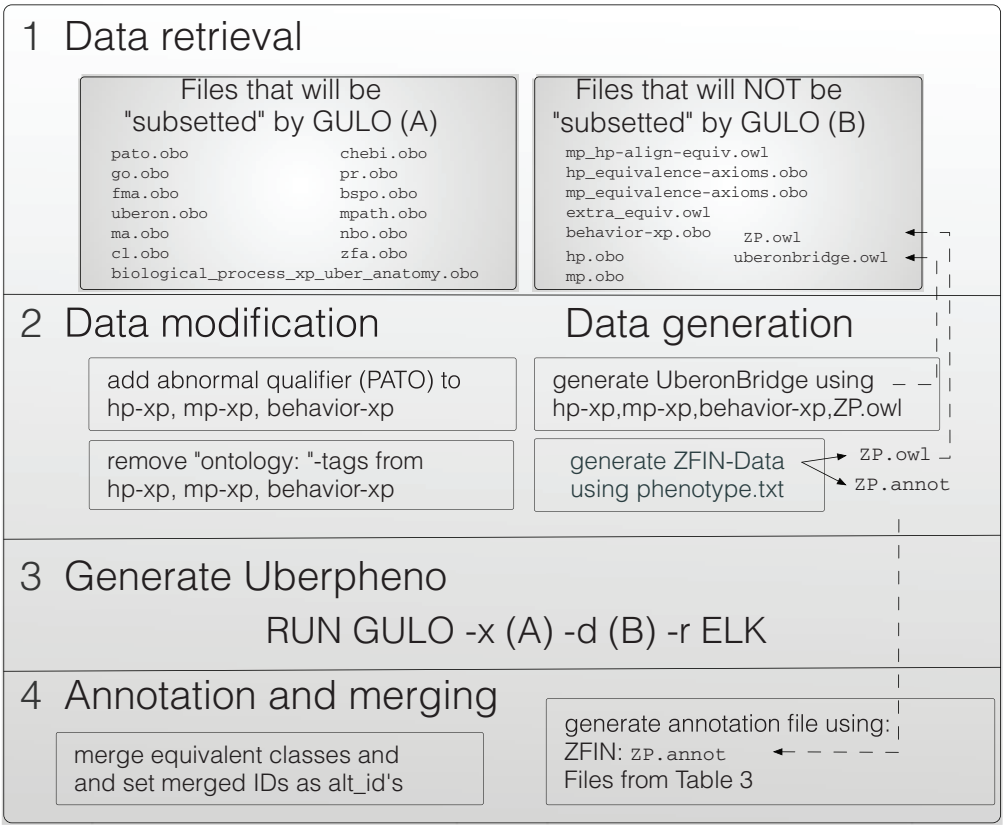

Figure 2 Schematic work- and dataflow illustration for the construction of the *Uberpheno* ontology and the gene annotations.

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 and
   occurs_in some eye
```

Here, the GO process *eye pigmentation (G0:0048069)* 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://compbio.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

```
ZDB-GENE-980605-30;83439;tcf711a;ZFA:0000
107;eye;PATO:0000645;hypoplastic;abnormal
```

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[12], which generates the ontology denoted as **zp.owl.** For every modified gene, a set of post-composed phenotype annotations is stored in **pheno.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

```
Entity=ZFA:0000107 (eye),
Quality=PATO:0000645 (hypoplastic) and
Qualifier=PATO:0000460 (abnormal)
```

generates an OWL class:

```
Class: ZP_0003395
Annotations: label "abnormal(ly) hypoplastic
eye"
EquivalentClassOf:
 has_part some:
   PATO_0000645 and
   inheres_in some ZFA_0000107 and
   qualifier some PATO_0000460
```

Beside generating the ZP-ontology, the annotation relation between the zebrafish genes and ZP-term is written to a file called **zp.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[41] is used to calculate subclass and equivalence relationships between classes. These steps are implemented within the GULO framework[23].

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 **alt_id**-tag for the term. For example, the HPO-term *Gallbladder dysfunction (HP:0005609)* has as **alt_id** the ZP-term *abnormal(ly) decreased functionality gall bladder (ZP:0004170).* The resulting ontology in OBO-format is named **crossSpeciesPheno.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 *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 *Uberpheno* are summarized in Table 1.

The construction pipeline is set up as a job in our continuous integration system accessible at http://compbio.charite.de/hudson, which is already used for data related to the HPO[42]. The job (called **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 http://purl.obolibrary.org/obo/hp/uberpheno/, whereas the file **crossSpeciesPheno.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[43]. 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 **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 *Uberpheno* ontology is shown in Figure 1 b), 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 *posterior microphthalmia* is inferred to be a subclass of the human term *Bilateral microphthalmos* can be used to transfer the information that the mouse gene *PRSS56* is known to cause *Bilateral microphthalmos.* This implies that querying the cross-species ontology for genes related to *Bilateral microphthalmos* will return the human gene *TCOF1,* the mouse gene *PRSS56* and the zebrafish gene *tcf7l1a.*

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 Phenotye Ontology (MPO), or `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 `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.

Statistics	
Uberpheno statistics:	
Phenotype classes:	25,974
HPO-IDs	13,122
MPO-IDs	9,800
ZP-IDs	8,057
Annotation statistics:	
All annotations	235,752
HPO annotations	63,080
-OMIM	49,348
-Orphanet	16,244
MPO annotations	149,164
ZP annotations	23,508

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

```
83439;TCF7L1;abnormal(ly) hypoplastic eye (ZP:0003395);tcf7l1a (ZDB-GENE-980605-30/ZEBRAF)
```

Conclusions

The phenotype resources for mouse, zebrafish, and human are used by several research projects[44–46].

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, *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[27] 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[47]. 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 *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 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 Forschungsge-meinschaft (DFG RO 2005/4-1), the Bundesministerium fur 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.

Acknowledgements

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

References

1. Nadia Rosenthal, Steve Brown: The mouse ascending: perspectives for human-disease models. *Nat Cell Biol.* 2007; **9** (9): 993–9.
2. Graham J Lieschke, Peter D Currie: Animal models of human disease: zebrafish swim into view. *Nat Rev Genet.* 2007; **8** (5): 353–67.
3. Steve D Brown, Mark W Moore: Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Dis Model Mech.* 2012; **5** (3): 289–92.
4. Allan Bradley, Konstantinos Anastassiadis, Abdelkader Ayadi, *et al:* The mammalian gene function resource: the International Knockout Mouse Consortium. *Mamm Genome.* 2012; **23** (9–10): 580–6.
5. Yvonne Bradford, Tom Conlin, Nathan Dunn, *et al:* ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.* 2011; **39** (Database issue): D822–9.
6. Sandra C Doelken, Sebastian Köhler, Christopher J Mungall, *et al:* Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Dis Model Mech.* 2012.
7. Anika Oellrich, Robert Hoehndorf, Georgios V Gkoutos, *et al:* Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases. *PLoS One.* 2012; **7** (6): e38937.
8. Paul N Schofield, Robert Hoehndorf, Georgios V Gkoutos: Mouse genetic and phenotypic resources for human genetics. *Hum Mutat.* 2012; **33** (5): 826–36.
9. Michael Ashburner, Ball CA, Judith A Blake, *et al:* Gene ontology: tool for the unification of biology. *Nat Genet.* 2000; **25** (1): 25–9.
10. Barry Smith, Michael Ashburner, Cornelius Rosse, *et al:* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 2007; **25** (11): 1251–1255.
11. Nicole L Washington, Melissa A Haendel, Christopher J Mungall, *et al:* Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* 2009; **7** (11): e1000247.

12. Christopher J Mungall, Georgios V Gkoutos, Cynthia L Smith, *et al:* Integrating phenotype ontologies across multiple species. *Genome Biol.* 2010; **11** (1): R2.
13. Sebastian Köhler, Marcel H Schulz, Peter Krawitz, *et al:* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet.* 2009; **85** (4): 457–64.
14. Sebastian Köhler, Sandra C Doelken, Ana Rath, *et al:* Ontological phenotype standards for neurogenetics. *Hum Mutat.* 2012; **33** (9): 1333–1339.
15. Sebastian Bauer, Sebastian Köhler, Marcel H Schulz, *et al:* Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics.* 2012; **28** (19): 2502–8.
16. Christopher J Mungall, Michael Bada, Tanya Z Berardini, *et al:* Cross-product extensions of the gene ontology. *J Biomed Inform.* 2011; **44** (1): 80–6.
17. Cynthia L Smith, Carroll-Ann Goldsmith, Janan T Eppig: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* 2005; **6** (1): R7.
18. Peter N Robinson, Sebastian Köhler, Sebastian Bauer, *et al:* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet.* 2008; **83** (5): 610–5.
19. Georgios V Gkoutos, Christopher J Mungall, Sandra Dolken, *et al:* Entity/quality-based logical definitions for the human skeletal phenome using pato. *Conf Proc IEEE Eng Med Biol Soc.* 2009; **1**: 7069–72.
20. Gary Schindelman, Jolene S Fernandes, Carol A Bastiani, *et al:* Worm Phenotype Ontology: integrating phenotype data within and beyond the C. elegans community. *BMC Bioinformatics.* 2011; **12**: 32.
21. Terrence F Meehan, Anna Maria Masci, Amina Abdulla, *et al:* Logical Development of the Cell Ontology. *BMC Bioinformatics.* 2011; **12**: 6.
22. Georgios V Gkoutos, Eain CJ Green, Ann-Marie Mallon, *et al:* Using ontologies to describe mouse phenotypes. *Genome Biol.* 2004; **6** (1): R8.
23. Sebastian Köhler, Sebastian Bauer, Chris J Mungall, *et al:* Improving ontologies by automatic reasoning and evaluation of logical definitions. *BMC Bioinformatics.* 2011; **12**: 418.
24. Robert Hoehndorf, Paul N Schofield, Georgios V Gkoutos: PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.* 2011; **39** (18): e119.
25. Chao-Kung Chen, Christopher J Mungall, Georgios V Gkoutos, *et al:* MouseFinder: Candidate disease genes from mouse phenotype data. *Hum Mutat.* 2012; **33** (5): 858–66.
26. Boris Motik, Peter F Patel-Schneider, Bijan Parsia: OWL 2 Web Ontology Language: structural specification and functional-syle syntax. 2008.
27. Catia Pesquita, Daniel Faria, Andre O Falcão, *et al:* Semantic similarity in biomedical ontologies. *PLoS Comput Biol.* 2009; **5** (7): e1000443.
28. Sebastian Bauer, Steffen Grossmann, Martin Vingron, *et al:* Ontologizer 2.0–a mul-tifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics.* 2008; **24** (14): 1650–1.
29. Paula de Matos, Rafael Alcántara, Adriano Dekker, *et al:* Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.* 2010; **38** (Database issue): D249–54.
30. Gene Ontology Consortium, Harris MA, Clark J, Ireland A, *et al:* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004; **32** (Database issue): D258–61.
31. Darren A Natale, Cecilia N Arighi, Winona C Barker, *et al:* The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.* 2011; **39** (Database issue): D539–45.
32. Jonathan Bard, Seung Y Rhee, Michael Ashburner: An ontology for cell types. *Genome Biol.* 2005; **6** (2): R21.
33. Cornelius Rosse, Jose L Mejino Jr: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform.* 2003; **36** (6): 478–500.
34. Jacqueline H Finger, Constance M Smith, Terry F Hayamizu, *et al:* The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Res.* 2011; **39** (Database issue): D835–41.
35. Judy Sprague, Leyla Bayraktaroglu, Yvonne Bradford, *et al:* The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.* 2008; **36** (Database issue): D768–72.
36. Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, *et al:* Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 2012; **13** (1): R5.
37. Schofield PN, Gruenberger M, Sundberg JP: Pathbase and the MPATH ontology. community resources for mouse histopathology. *Vet Pathol.* 2010; **47** (6): 1016–20.
38. Georgios V Gkoutos, Paul N Schofield, Robert Hoehndorf: The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. *Int Rev Neurobiol.* 2012; **103**: 69–87.
39. Ana Rath, Annie Olry, Ferdinand Dhombres, *et al:* Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat.* 2012; **33** (5): 803–8.
40. John M Hancock, Ann-Marie Mallon, Tim Beck, *et al:* Mouse, man and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm Genome.* 2009; **20** (8): 457–61.
41. Yevgeny Kazakov, Markus Krötzsch, František Simancík: Concurrent classification of EL ontologies. In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Noy, and Eva Blomqvist, editors, Proceedings of the 10th International Semantic Web Conference (ISWC'11). volume 7032 of *LNCS*. Springer, 2011.
42. Christopher J Mungall, Heiko Dietze, Seth J Carbon, *et al:* Continuous Integration of Open Biological Ontology Libraries. *Bio-Ontologies* 2012.
43. John Day-Richter, Midori A Harris, Melissa Haendel, *et al:* OBO-Edit–an ontology editor for biologists. *Bioinformatics.* 2007; **23** (16): 2198–200.
44. Alex Bayés, Louie N van de Lagemaat, Mark O Collins, *et al:* Characterization of the proteome, diseases, evolution of the human postsynaptic density. *Nat Neurosci.* 2011; **14** (1): 19–21.
45. Joanna Amberger, Carol Bocchini, Ada Hamosh: A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum Mutat.* 2011; **32** (5): 564–7.
46. Cynthia L Smith, Janan T Eppig: The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome.* 2012; **23** (9–10): 653–68.
47. Committee on a Framework for Development a New Taxonomy of Disease, National Research Council: Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. The National Academies Press, 2011.

Current Referee Status:

Referee Responses for Version 1

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

Approved: 12 February 2013

Referee Report: 12 February 2013
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.

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

Approved: 11 February 2013

Referee Report: 11 February 2013
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.

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

Approved: 08 February 2013

Referee Report: 08 February 2013
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.