Published in final edited form as:

Mamm Genome. 2012 October ; 23(9-10): 669–679. doi:10.1007/s00335-012-9404-4.

Computational tools for comparative phenomics; the role and promise of ontologies

Georgios V. Gkoutos^{1,3,*}, Paul N. Schofield², and Robert Hoehndorf¹

¹Department of Genetics, University of Cambridge, Downing Street, Cambridge, Cambridge CB2 3EH, UK

²Department of Physiology, Development and Neuroscience, University of Cambridge, Downing street, Cambridge CB2 3EG, UK

³Department of Computer Science, University of Aberystwyth, Old College, King Street, SY23 2AX, UK

Abstract

A major aim of the biological sciences is to gain an understanding of human physiology and disease. One important step towards such a goal is the discovery of the function of genes that will lead to better understanding of the physiology and pathophysiology of organisms ultimately providing better understanding, diagnosis, and therapy. Our increasing ability to phenotypically characterise genetic variants of model organisms coupled with systematic and hypothesis-driven mutagenesis is resulting in a wealth of information that could potentially provide insight to the functions of all genes in an organism. The challenge we are now facing is to develop computational methods that can integrate and analyse such data. The introduction of formal ontologies that make their semantics explicit and accessible to automated reasoning promises the tantalizing possibility of standardizing biomedical knowledge allowing for novel, powerful queries that bridge multiple domains, disciplines, species and levels of granularity. We review recent computational approaches that facilitate the integration of experimental data from model organisms with clinical observations in humans. These methods foster novel cross species analysis approaches, thereby enabling comparative phenomics and leading to the potential of translating basic discoveries from the model systems into diagnostic and therapeutic advances at the clinical level.

1 Introduction

The discovery of gene function is one of the key aims of biomedical science in the 21st century. It is hoped that with this will provide an understanding not only of the normal biology of organisms, but also of their pathophysiology. The original promise of genome sequencing was that it would provide insights into gene function which would allow us to identify genes involved in human diseases and their predispositions [59]. The full genome sequence of humans and a wide variety of model organisms such as the mouse, fly, and worm are available today, but despite the development innovative structural and homology based approaches [61, 60], we have been largely unsuccessful in using sequence alone to reliably predict gene function in the context of the whole organism. Currently in the mouse (http://www.informatics.jax.org/mgihome/homepages/stats/all_stats.shtml/ accessed July 2011 and C. Smith. pers.comm) we have only 13,000 genes (about half the protein-coding genes) with any experimentally-based functional annotations in the Gene Ontology (GO)

*Corresponding Author : g.gkoutos@gen.cam.ac.uk.

molecular function tree [5]. Of the 8,600 GO molecular function terms, most have very few associated gene products. Furthermore, only 7,000 genes have biological process annotations.

Candidate genotype–disease associations are being established at a rapid rate as a result of the proliferation of Genome-wide associated studies (GWAS), and together with the accelerating phenotype characterisation of naturally-occurring and engineered genetic variants of model organisms, we are accumulating a wealth of complementary information that is beginning to provide us with those promised insights into gene function.

Gathering the phenotype data for humans and model organisms is arguably a greater challenge than the genome sequencing projects as the range of phenotype measurements and the complexity of the data present major problems. Not only are the datasets extremely large for some data types, such as images, but they require the development of novel semantic approaches to enable computation and data integration. We face two major problems; the fragmentation of phenotype data across many sources [33, 28], and the inability to accurately integrate this data computationally due to semantic inconsistency [81]. The latter is a problem even within the same organism when data are coded using different formalisms or free text, but trying to integrate and co-analyse datasets between organisms presents us with even more challenging problems resulting from different terminologies, different assays, and in many cases differing conceptualisations of phenotype over and above the obvious mismatches of species -specific anatomies, behaviour, and physiology. However, the rewards for enabling data gathering and integration between species are potentially enormous, and much effort has been expended in recent years to address these barriers [12, 15, 17, 18, 34, 57, 66, 29, 75, 93, 91].

The conservation of gene function and expression across species is remarkable [98]. The ability of orthologs to substitute for each other in different species is one example of this [3, 92]. While the phenotype manifestations of mutations in homologous genes might be expected to give rather diverse phenotypes in different organisms, it has been shown that in many cases, particularly between vertebrates, phenotypes are remarkably conserved, implying that the underlying physiological pathways in which these genes function are themselves highly conserved [70]. Correlating these pathways and interacting networks with disease is a first step on the road to *systems pathobiology*. This conservation is self-evident from the successes that phenotype comparisons using gene orthology have had in recent years [34, 24, 63]. However, orthology is only useful when good phenotype information is available for one or more species, and it does not permit the discovery of novel relationships between phenotypes and the pathways involved in normal physiology or disease, for example where a phenotype is associated with mutations in a series of unrelated genes in a pathway. Text mining has also been used to collect phenotype information [34, 91, 63] but whilst useful within species, establishing relations between phenotypes of different species is deeply problematical where there is semantic ambiguity, and in many cases a difference in the concepts underlying lexically identical terms and vice versa.

Ideally, we would have complete, semantically standardised datasets for the whole genome, underlying the complete genotype/phenotype map for humans and their model organisms. The phenotype landscape that we have for vertebrates, including mice, fish, and humans, is currently mostly gathered from the literature by skilled curators but lacking such semantic standardisation. Human data is in many ways more complete than that for model organisms, but is much more scattered and difficult to access, being distributed across databases and repositories ranging in organisation and coverage from the locus specific databases (LSDBs) to Online Mendelian Inheritance in Man (OMIM) and Orphanet [89, 4, 94]. Integration and

use of all of this gathered data provides a richness of genotype/phenotype relationships that data from one organism cannot provide [71].

While a complete view of the phenome of any metazoan organism is a long way off, recent developments in mouse biology and genetics have set us well on the way to having this “encyclopedia for the mouse genome” within ten years [1, 2]. The International Knockout Mouse Consortium (IKMC) (<http://www.knockoutmouse.org/>) set out to generate a knockout embryonic stem (ES) cell for every protein coding gene in the mouse genome.[80] The mice being generated from this resource are now being used by the International Mouse Phenotyping Consortium (IMPC) (<http://www.mousephenotype.org/>) [14] to generate systematic phenotype information for each of the viable lines. This primary line phenotyping includes a wide range of tests, for example behaviour, dysmorphology, blood chemistry, immunology, with the aim of identifying abnormal phenotypes in organ systems or specific physiological processes that may be picked up for more detailed secondary investigation. The dataset produced by this large-scale international programme will, in principle, give us the best coverage of the phenome for any higher organism, and will set the paradigm for standardising, recording, and archiving phenotype data. High quality phenotype data is now available for many species including *Danio rerio* (Zfin database [13]), *Drosophila* (Flybase [21]), *Caenorhabditis elegans* (Wormbase [38]) and yeast (SGD [23]).

If data from the IMPC and hypothesis-driven experiments in model organisms is to be integrated with the data from humans, representation of these datasets must be semantically consistent across the various resources that store them; without such standardisation, the full value of this data cannot be realised [77]. The integration strategy which has proved to be the most successful to date has been the use of biomedical ontologies to describe data and experiments [81], and the ontology-driven approach to integration and analysis of phenotype information is the focus in this review.

2 The promise of ontologies

The use of ontologies as an approach to semantic standardisation was proposed more than a decade ago [35, 5] and since then has become the dominant methodology used to semantically categorise phenodeviance [26]. The biomedical research community has invested considerable amount of effort and resources in the development and establishment of ontologies that are becoming increasingly successful as information management and integration tools in many disparate scientific fields allowing interoperability and semantic information processing between diverse biomedical resources and domains [83, 30].

2.1 Formal ontology

In computer science, an ontology is a specification of a conceptualization of a domain of knowledge [36, 37]. Ontologies commonly distinguish between *classes* (also called *concepts*, or *categories* or *universals*) and *individuals* within a domain of knowledge. A class is an entity that can have *instances*, while individuals are entities that cannot be instantiated [40]. Examples of classes include *Tower* or *Triathlon*, while examples of individuals include the Eiffel tower or the 2009 Ironman Triathlon in Hawaii. The Eiffel tower can be an instance of the class *Tower*, and the 2009 Ironman Triathlon an instance of *Triathlon*. The meaning of classes is specified by stating what must be true of their instances.

In addition to classes and individuals, ontologies often include *relations*. Relations hold between entities, they are the “the glue that holds things together, the primary constituents of the facts that go to make up reality” [9].

In *formal* ontologies, the specification of classes and relations follows the axiomatic-deductive method. Given a set of terms that are used within a domain and whose meaning we wish to specify, we begin by providing *explicit definitions* for some terms, potentially introducing new terms. An explicit definition of a term t is a statement that can replace every occurrence of t in any sentence.

Eventually, a set of *primitive terms* remains that are not further defined. Following the axiomatic method [41], using only the primitive terms, we can construct complex sentences. Based on the intended meaning of the primitive terms, we consider some of these sentences true and some of them false in our domain. We select some of the true sentences as *axioms* which provide the core of our ontology. Ideally, the axioms are chosen so that all true sentences in the domain we intend to represent follow by means of logical deduction from the axioms. More commonly, however, only *some aspects* of the intended meaning are formally represented while other aspects are omitted either due to limitations in language expressivity or due to their irrelevance to the problem for which an ontology is developed.

2.2 Reasoning

Based on the axioms and definitions, we can use deduction to infer statements that logically follow from the axioms. The process of automatically deducing sentences from axioms is called *automated reasoning*. Automated reasoning allows users of an ontology to carry out key activities: verifying the ontology's consistency, inferring hidden knowledge and thereby performing powerful queries. An ontology is formally inconsistent if there is a statement ϕ such that ϕ and its negation $\neg\phi$ can be inferred from the ontology's axioms. If an ontology is formally inconsistent, *every* statement can be inferred from the ontology [10].

Automated reasoning can further determine whether classes in an ontology are unsatisfiable: a class C is unsatisfiable, if it is impossible for the class to have any instances. Unsatisfiable classes in an ontology are commonly the result of a contradictory class definition. Automated reasoning in the Web Ontology Language (OWL) can be employed to automatically compute the generalization hierarchy underlying an ontology as well as for verification of data consistency and complex queries [95, 74, 44, 43]. Highly efficient automated reasoners are available to process OWL ontologies [79, 90, 65, 56]. To decrease the computational complexity of automated reasoning over OWL ontologies further, several OWL profiles were developed that define subsets of the OWL language and guarantee tractable (i.e., polynomial-time) automated reasoning [64]. In particular the OWL EL profile was found to provide the expressivity required for most biomedical ontologies [7, 78, 45], and highly optimized OWL EL reasoners are available or under development to support reasoning over very large ontologies [7, 55, 56].

A high expressivity is required to accurately specify complex axioms that constrain the domain under investigation, and languages with higher expressivity than OWL are often required in the biomedical domain to achieve this goal [46, 42]. On the other hand, automated reasoning over large ontologies and associated datasets benefits from languages with a low complexity of inferences in which complex axioms cannot be formulated.

2.3 Interoperability between biomedical data sources

The combination of formal ontologies and automated reasoning can enable interoperability between biomedical databases, web services, and software tools [30, 44]. Ontologies provide controlled vocabularies that can be shared across different data repositories and therefore facilitate the integration of these databases. Ontologies further provide a graph structure based on their taxonomy and axioms, and this structure can be used to enable data retrieval [8], clustering [96] and integrated data analysis [87].

The recent addition of further axioms to some widely used biomedical ontologies has further enabled the use of automated reasoning for powerful queries and complex retrieval operations [68, 44, 43]. For example, in an ontology of disease, a class *Arthritis* could be defined as an *Inflammation* that **occurs in** a *Joint*. Such an axiom is a definition in which *Arthritis* is related to a process (*Inflammation*) and an anatomical location (*Joint*); the type of relation is specified as the **occurs in** relation. An automated reasoner could use this information together with the background information contained in an anatomy ontology to retrieve *Arthritis* as a disease that affects (**occurs in**) the skeletal system (of which a *Joint* is a part).

3 Key challenges for interoperability of phenotype resources

3.1 Formally representing the knowledge

In order for ontologies to realize their potential, they must provide rich, explicit, and consistent descriptions for their terms so that automated systems are able to process and understand their meaning, thereby enabling their use to infer new information. For this purpose, such descriptions are currently being created for numerous ontologies within the biomedical domain. These descriptions are being increasingly expressed, in formal languages, such as the Web Ontology Language (OWL) [32]. However, in order to make use of these descriptions it is imperative that their semantics are explicit and accessible to automated reasoning. More precisely, it is imperative that their definitions need to include precise descriptions of the relationships that are employed as well as ensure the consistency of the knowledge represented.

3.1.1 OBO and OWL—While some ontologies are currently being developed directly in OWL, the majority of ontologies relevant for descriptions of phenotypes are available in the OBO Flatfile Format [53]. The OBO Flatfile Format is a semi-formal language of which a fragment has been embedded in OWL while other parts remain largely informal [53, 67]. Establishing an accurate representation of biomedical ontologies in a formal language such as OWL would decrease ambiguity in biomedical ontologies and improve their interoperability.

While several projects have proposed OWL as a representation language for biomedical ontologies and biomedical knowledge, two aspects are important to consider. On one hand, the representation language needs to be rich enough to express *relevant* distinctions in biomedical applications, and in many cases, OWL does not provide sufficient expressivity to represent even basic biomedical or chemical facts [46, 42]. On the other hand, biomedical ontologies are often large so that *efficient* processing of the knowledge becomes important.

The first trade-off that must be considered in selecting a language are expressivity and decidability. For example, the logic underlying OWL 2 is designed to allow *decidable* subsumption: it is possible to design an algorithm which will always terminate and determine whether one class in an ontology is a subclass of another or not. The semantics of the OBO Flatfile Format 1.4 (draft, August 19 2011), on the other hand, extends OWL 2 and does not guarantee decidability. Similarly, the RNA Ontology [42] uses *second order logic* to express some of its axioms, and second order logic is undecidable. For example, while it is possible to *express* that molecules are maximally connected structures in second order logic, it is not, in principle, possible to design an algorithm that can explore all the consequences of this statement. Instead, theorem provers for second-order logics often rely on incomplete algorithms that can infer some, but not all, consequences.

The second trade-off that must be considered is expressivity vs. complexity of automated reasoning. A logic may be decidable, yet algorithms may require exponential (or doubly-

exponential in the case of OWL 2) time to determine the answer to a problem. Weaker logics, such as EL++ [7], may provide strongly reduced expressivity while supporting efficient, polynomial-time answers [56]. For example, there has been significant debate about the use of negation to formalize phenotypes such as *Absent appendix* [16, 47], and while the use of negation generally leads to more flexibility and expressive inferences, it also requires significant amount of time to perform these inferences and does not scale to large datasets [49]. As a result, less expressive forms of formalizing absence are in use [66] and can efficiently be applied to large volumes of data.

Finally, expressivity and usability is another trade-off. Languages with higher expressivity often require specialized background knowledge, while the use of less expressive languages may be easier to learn and less likely to lead to errors. For example, the OBO Flatfile Format 1.2 specifies a basic graph-based representation of ontologies which provides ontology creators with visual feedback for structuring information [31], while formal languages such as OWL usually require the manipulation of axioms and the use of automated reasoning for provision of feedback. User-interfaces, graphical or natural language representations of formal languages, may support the correct use of complex language features. For example, Common Logic [54] allows for a syntactic representation of axioms based on conceptual graphs [85], and similar methods can be used to improve the usability of complex languages.

3.1.2 Relations—Relations in biomedical ontologies are used both to interrelate individual entities and classes. For example, a **part-of** relation may hold between an individual ear and a head, but also between the class *Ear* and the class *Head*. The latter relations are patterns that stand for complex axioms, and some of these axiom patterns are listed in the OBO Relation Ontology [82]. For example, according to the OBO Relation Ontology, the parthood relation between *Ear* and *Head* is translated as an axiom that states that all instances of *Ear* must be a part of (between individuals) some instance of *Head*. Such translation patterns can now be specified in the OBO Flatfile Format [48], thereby allowing for flexible definitions of relations in a variety of ontologies.

3.1.3 Consistency—Once phenotype ontologies are integrated by aligning their content and standardizing their relations, contradictions may arise that make it impossible to use these ontologies for automated reasoning [47]. Automated reasoners can be employed to verify the consistency of integrated ontologies, and some studies have shown that phenotype ontologies as well as anatomy ontologies may give rise to contradictory class definitions as a result of combining the axioms in two or more ontologies [68, 69, 44]. Consequently, methods are needed that allow not only the detection of such errors but also the explanation and subsequently the repair of them.

3.2 Making ontologies interoperable on a large scale

The explicit semantics of ontologies are rarely taken advantage of by software systems due to the issues of tractability arising from the high complexity of reasoning over formal ontologies. As a consequence, current ontology-based resources such as the various model organism databases, search engines, ontology repositories, ontology browsers and interfaces, make little or no use of the semantic power of the ontologies at all, which consequently diminishes their utility towards facilitating data integration and interoperability. Unless an ontology's semantics can be employed by ontology-based applications and methods, the original goal of ontologies to facilitate data integration and interoperability cannot be achieved, thereby diminishing the value of the ontology development and maintenance efforts of the past decade.

The solution seems to arise from modularization methods [45, 72, 58] as well as recent progress in implementing highly optimized automated reasoners [56]. In particular, reasoners that are capable of processing the EL subset of OWL enable tractable automated reasoning and ensure that ontologies can now achieve their goal of data integration and interoperability, not only in a static sense that is applied in database annotations, but in the more important dynamic sense that is determined by how these ontologies are used [78, 88, 6, 7, 45].

3.3 Bridging domains and levels of granularity to cross species

By eliminating inconsistencies and thereby enabling the formalisation of the relation and class definitions in biomedical ontologies, it becomes possible to employ the resulting ontologies for integrated data analysis across multiple domains and levels of granularity, across different types of knowledge and between different species. To achieve these goals, it is necessary to formally express the connections between domains, granularity scales, species, and types of information in ontologies while at the same time increasing the coverage of information within each domain. For example, although ontologies can provide connections between processes on a cellular scale and how these cellular processes contribute to organ-scale physiological processes, such information is not widely included in current ontologies due to the complexity of the underlying phenomena. However, inclusion of such information in ontologies is important if they are to facilitate cross-domain and cross-species analyses. The generation of such connections between biomedical ontologies has the potential to make biomedical information retrieval a knowledge-driven discipline based on formalized ontologies that make their semantics explicit and accessible to automated reasoning, thus resulting in the capability to answer novel, powerful queries that bridge multiple domains, disciplines, species and levels of granularity.

4 Computational analysis of phenotypes

4.1 Enabling Comparative Phenomics

Comparative phenomics exploits the evolutionary conserved physiologic and pathophysiologic mechanisms in order to enable our ability to study human disease and its treatment. Our understanding of the function of a gene can be enhanced by our increasing ability to compare mutant and ‘wild type’ phenotypes associated with it both within a single organism as well as between species. This is the main reason that the biomedical community to invest in a numerous Model Organism Databases (MODs) that organise and store, amongst other genetic information, phenotype information associated with specific mutations and genetics variants for a particular species.

However, the description of phenotypes presents a major conceptual and practical problem due to the complexity of the underlying domain as well as the need to compare and integrate phenotype knowledge across various domains and species to facilitate comparative phenomics [27]. To address this issue, a formal species and domain-independent method for the description of phenotypes that requires combinations of orthogonal ontologies with the ability to correlate factors depending on experimental values has been developed, termed PATO - the Phenotype And Trait Ontology [29].

4.2 PhenomeBlast

PATO has the ability to unify phenotype species specific statements and can be employed to provide formal definitions for the species-specific phenotype ontologies that the various MODs utilise for the description of their data. It is now increasingly common for the various databases to either directly annotate their phenotype data based on this method or employ it in order to define the classes of the species and domain specific phenotype ontologies they

use for annotation in order to accurately express the meaning of their phenotype term, and to perform inferences over them [84, 22, 73, 86, 97, 76, 62].

PATO-based phenotype definitions can be formally represented in OWL which can be formalised and combined with ontology modularization [45] that can be utilised for efficient automated reasoning. Such formalisation and modularization can be implemented by *PhenomeBlast* software that aligns phenotypes across species and enables the generation of a single, unified, and logically consistent representation of phenotype data for multiple species. *PhenomeBLAST* provides the foundation for the development of a comparative phenomics framework that scales across model organism databases and high-throughput phenotype experiments.

4.3 PhenomeNet

PhenomeNet [52] utilises the *PhenomeBLAST* approach and forms a cross-species network of phenotype similarity between genotypes and diseases. Based on the semantically and logically consistent cross-species ontology created through *PhenomeBlast*, *PhenomeNet* incorporates phenotype annotations for mouse, zebrafish, fly, yeast, and worm available from 5 different MODs. *PhenomeNet* also includes human phenotypes associated with inherited diseases that are found in the Online Mendelian Inheritance in Man (OMIM) database [4]. The resulting ontology contains more than 500,000 classes and more than a million and a half axioms allowing for the generation of a phenotype network that contains more than 111,000 complex phenotype nodes, each of which represents a complex phenotype observed in an animal organism or the phenotype associated with a human disease.

PhenomeNet is amenable to efficient automated reasoning through ontology modularization [45] and design patterns for the description of phenotype information associated with human diseases or with experimental data derived from model organisms. The resulting integration of phenotype ontologies allows for a direct comparison of phenotypes across species and domains of knowledge which can then be ranked, based on a measure of semantic similarity, in respect to diseases as well as other phenotype characterisations.

As a result *PhenomeNet* can be utilized, with a high success rate [52], for predicting genes that participate in the same pathway, orthologous genes as well as gene-disease associations. To measure the performance of *PhenomeNet* in these tasks, we analyze the receiver operating characteristic (ROC) curve [25]. A ROC curve can be used to visualize the performance of a classifier and plots the classifier's true positive rate as a function of the false positive rate. The area under the ROC curve (AUC) is a quantifiable measure of the classifier's performance and is equivalent to the probability that the classifier ranks a randomly chosen positive example higher than a randomly chosen negative one [25]. We use the mouse models that are associated with an OMIM disease in the MGI database [11] as positive examples of genotype-disease associations, and the remaining as negative examples. The resulting ROC curve is displayed in Figure 1 and achieves an AUC of 0.868.

PhenomeNet significantly outperforms any other phenotype-based approaches for predicting gene-disease associations and moreover its performance matches gene prioritization methods that are based on prior information about molecular causes of the disease. The underlying difference between the *PhenomeNet* approach and other disease gene prioritization methods is that while the latter rely on additional sources of information other than the phenotype involved (e.g. functional, pathway and literature annotations), *PhenomeNET*'s predictions are based on information about phenotypes alone. As such, it can be applied to identify candidate genes for diseases with an unknown molecular basis.

PhenomeNET can further identify relevant animal models of a particular strain and within a certain environment and thereby improve the speed and reduce the cost required to test and validate novel candidate gene-disease associations. Therefore, it will allow the targeting of particular strains and genotypes, and utilize the results from large-scale phenotyping projects such as the IMPC [19].

4.3.1 PhenomeBrowser—*PhenomeBrowser* [51] provides a web interface that allows users to access both the crossspecies integration of phenotypes and a variety of similarity-based comparison of sets of phenotypes across species. It includes all the data (experimental, genes, alleles, disease) and provides links to the databases from which the phenotypes originate. It allows users to perform queries based on diseases, genes, or genotypes using either their names or corresponding identifiers employed by model organism databases and explore them based on their phenotype annotations and their predicted associations to diseases.

4.4 PhenomeDrug

One of the major aims of pharmacological research lies with its potential ability to *repurpose* or *reposition* existing drugs for new indications. Identifying new targets for existing drugs and new indications for known mechanisms of actions is both a great challenge and an opportunity that the community faces today. Our understanding of the physiology and pathobiology that governs diseases and their phenotype manifestations will greatly facilitate our ability to find new targets for existing drugs. The pharmacogenomic community has long been aware of the value of in silico approaches for the analysis of the increasing amount of information available both in public and private databases, as well as their usefulness for proposing new potential drug indications and novel drug discovery. One of the areas though that still remains to be fully exploited is the wealth of phenotype information that is increasingly becoming available from a variety of studies on animal models of human disease.

PhenomeDrug [50] is a method for predicting novel associations between drugs and diseases based on the *PhenomeNET* method for comparing phenotypes across species. Currently, it utilises data from the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB)[39], a central repository containing a wealth of relationships between genetic, genomic, drug-response related phenotype data, and clinical information. It then combines the predictions of disease-gene associations from *PhenomeNET* with the drug-gene associations available from the PharmGKB to suggest new diseases in which a drug may be active.

Since the *PhenomeNET* method performs a comparison of phenotypes directly, it can prioritize genes for orphan diseases of which the molecular basis is unknown and hence *PhenomeDrug*'s predictions can suggest potentially new drugs for rare and orphan diseases. The use of PhenomeNET further provides direct links to animal models that can be used to investigate the drug and disease mechanisms as well as the drug's role in the disease.

5 Outlook: methodological advances and viable areas of investigation

Our analysis of the success rate of phenotype-based approaches for revealing human gene-disease association based on experimental data from animal organisms reveals several areas where further research and development is urgently needed. In particular, novel methods need to focus on improving our computational ability to exploit the power of phenotype information for a better understanding of the function of a gene and its role in the disease. A key limiting factor in our ability to recover known and validated models is the completeness and quality of phenotype descriptions of human disease. A large number of

the current phenotype descriptions that are associated with OMIM diseases are either incomplete or completely lack explicit phenotype descriptions. In some cases particular areas of the phenotype are exhaustively described and annotated, while others, no less important to the clinical manifestation of the disease, are sparsely, though accurately, annotated. The result is a skewing in the significance of matches between human and model organism phenotypes.

Furthermore, OMIM diseases lack feature frequency data which, if present, would allow us to perform our analysis by applying different filters based on the frequency of the manifested observations in patients. Since our ability to map human disease phenotype manifestations to animal traits depends on the existence of such descriptions, it is vital that this issue is addressed where possible. Orphanet (<http://www.orpha.net>) does capture frequency data where it is reliably available and integration of this into disease descriptions is highly desirable.

Tightly linked to this issue is the completeness of the animal data (both experimental and literature) characterisations. Sources of skewing in the annotation of diseases in model organisms comes from the hypothesis-based investigation of mutants, where investigators may only have examined one aspect of the phenotype, either because of their research interests or their inability to carry out more comprehensive phenotyping. The systematic broad-based phenotyping underway in the IMPC produces the nearest to an ideal phenotype dataset from this point of view, with standardized knockout strategy, background strains and assays, and should eliminate a lot of the variation in the quality of the existing datasets. The areas not currently covered in the proposed primary pipeline, such as embryological or age-related phenotypes, are likely to be dealt with in targeted studies by some of the phenotyping centres involved.

Our ability to integrate experimental data from model organisms with clinical observations in humans depends on the power of the expressivity of our representation. As such, ontology-based integration approaches reveal that there are several biomedical knowledge domains that are underrepresented or not covered. One such area is the domain of behaviour. Behaviour and neurological disease is one the most challenging domains since it needs to account both for direct observations as well as inferences that are based on them. Furthermore, it needs to account for our interpretations and correlations of animal behaviour to human associations. However, it represents a vital area of research since animal models of behaviour and neurological disease have been shown to be highly successful for unveiling the genetic basis of many diseases such as Parkinson's, Huntington and Alzheimer disease.

Unifying and integrating phenotype data between animal models and humans in the domain of physiology presents another challenge due to the high complexity of the underlying domain. To address this issue, we have to be able to account for attributes of biological processes and relate them to the attributes of their parts and participants [20]. Furthermore, there is a need to establish a link between structural components of biological systems and the processes and interactions in which they participate. A consistent and comprehensive representation of the physiology domain would allow us to reveal the biological underpinnings of the phenotype manifestations we observe as well as account for the fact the attributes of processes are often measured indirectly and inferred from other attributes.

Research in these areas may then lead to the potential of exploiting phenotype data for predicting novel associations between drugs and diseases, as the pioneering PhenomeDrug study has shown [50] and are likely to be useful in drug repurposing and finding uses for orphan drugs. Currently, such approaches are impaired by the distribution of pharmacological knowledge across multiple heterogeneous databases, and integrating the

pharmacological and pharmacogenomics resources, like PharmGKB, DrugBank, and CTD, will be a precondition for further large-scale integrative analyses.

Finally, the computational representation, integration, and analysis of phenotype information has the potential for being applied not only to describing the large amount of phenotype information that will be generated from large scale phenotype projects such as IMPC but it could be employed to suggest and identify possible gene disease associations based on the minimal information that is produced from the phenotype pipelines (designed for breadth in order to recover any potentially significant phenotype) such efforts employ. Potentially, such a task will revolutionize the results of such projects, since it can guide, based on the minimal phenotype information generated from these screens, the prioritisation of secondary phenotype pipelines that are designed to reveal the phenotype depth of the mutation. Furthermore, it can be employed to optimize the phenotype assays applied, their order, their applicability etc. or even take into account other dimensions such as their cost effectiveness.

The approach described here works both within a single organism and between organisms and, being completely independent of any considerations of genetic orthology, presents an entirely phenotype view of the relations between diseases, pathways, and gene function. As human phenotype data becomes deeper, more standardized and interoperable as part of the drive towards precision medicine, and more individuals are subjected to whole genome or whole exon sequencing, the power of a phenotype-based approach to determining pathogenicity of mutations and understanding the underlying pathobiology will be fully realized.

References

1. Abbott A. Mouse megascience. *Nature*. 2010; 465:526. [PubMed: 20520665]
2. Abbott, Alison. Mouse project to find each gene's role. *Nature*. May.2010 465(7297)
3. Al-Hasani, Keith; Vadolas, Jim; Knaupp, Anja; Wardan, Hady; Voullaire, Lucille; Williamson, Robert; Ioannou, Panayiotis. A 191-kb genomic fragment containing the human α -globin locus can rescue α -thalassemic mice. *Mammalian Genome*. 2005; 16:847–853.10.1007/s00335-005-0089-9 [PubMed: 16284800]
4. Amberger, Joanna; Bocchini, Carol; Hamosh, Ada. A new face and new challenges for online mendelian inheritance in man (OMIM). *Hum Mutat*. 2011; 32:564–567. [PubMed: 21472891]
5. Ashburner, Michael; Ball, Catherine A.; Blake, Judith A.; Botstein, David; Butler, Heather; Cherry, Michael J.; Davis, Allan P.; Dolinski, Kara; Dwight, Selina S.; Eppig, Janan T.; Harris, Midori A.; Hill, David P.; Tarver, Laurie I.; Kasarskis, Andrew; Lewis, Suzanna; Matese, John C.; Richardson, Joel E.; Ringwald, Martin; Rubin, Gerald M.; Sherlock, Gavin. Gene ontology: tool for the unification of biology. *Nature Genetics*. May; 2000 25(1):25–29. [PubMed: 10802651]
6. Baader, F.; Lutz, C.; Suntisrivaraporn, B. Is tractable reasoning in extensions of the description logic *EL* useful in practice?. *Proceedings of the Methods for Modalities Workshop (M4M-05)*; Berlin, Germany. 2005.
7. Baader, F.; Lutz, C.; Suntisrivaraporn, B. CEL – a polynomial-time reasoner for life science ontologies. In: Furbach, U.; Shankar, N., editors. *Proceedings of the 3rd International Joint Conference on Automated Reasoning (IJCAR'06)*, volume 4130 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag; 2006. p. 287-291.
8. Bada, Michael; Stevens, Robert; Goble, Carole; Gil, Yolanda; Ashburner, Michael; Blake, Judith A.; Cherry, Michael J.; Harris, Midori; Lewis, Suzanna. A short study on the success of the gene ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*. Feb; 2004 1(2): 235–240.
9. Barwise, J. *The Situation in Logic*. CSLI; Stanford, CA: 1989.
10. Barwise, Jon; Etchemendy, John. *Language, Proof and Logic*. Center for the Study of Language and Inf; Apr. 2002

11. Blake, Judith A.; Bult, Carol J.; Kadin, James A.; Richardson, Joel E.; Eppig, Janan T. the Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Research*. 2011; 39(suppl 1):D842–D848. [PubMed: 21051359]
12. Bodenreider O, Hayamizu TF, Ringwald M, De Coronado S, Zhang S. Of mice and men: aligning mouse and human anatomies. *AMIA Annu Symp Proc*. 2005:61–65. [PubMed: 16779002]
13. Bradford, Yvonne; Conlin, Tom; Dunn, Nathan; Fashena, David; Frazer, Ken; Howe, Douglas G.; Knight, Jonathan; Mani, Prita; Martin, Ryan; Moxon, Sierra A.; Paddock, Holly; Pich, Christian; Ramachandran, Sridhar; Ruef, Barbara J.; Ruzicka, Leyla; Schaper, Holle Bauer; Schaper, Kevin; Shao, Xiang; Singer, Amy; Sprague, Judy; Sprunger, Brock; Van Slyke, Ceri; Westerfield, Monte. ZFIN: enhancements and updates to the zebrafish model organism database. *Nucleic acids research*. Jan.2011 39(Database issue)
14. Brown, Steve; Moore, Mark. An encyclopaedia of gene function; the international mouse phenotyping consortium. 2012. in press
15. Burgun, Anita; Mougins, Fleur; Bodenreider, Olivier. Two approaches to integrating phenotype and clinical information. *AMIA ... Annual Symposium proceedings / AMIA Symposium*. AMIA Symposium; 2009; 2009. p. 75-79.
16. Ceusters, Werner; Elkin, Peter; Smith, Barry. Referent tracking: The problem of negative findings. *Stud Health Technol Inform*. 2006
17. Chen, Jing; Bardes, Eric E.; Aronow, Bruce J.; Jegga, Anil G. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*. Jul.2009 37(Web Server issue):gkp427+.
18. Chen, Jing; Xu, Huan; Aronow, Bruce; Jegga, Anil. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*. 2007; 8(1):392. [PubMed: 17939863]
19. Collins FS, Finnell RH, Rossant J, Wurst W. A new partner for the international knockout mouse consortium. *Cell*. 2007; 129(2):235. [PubMed: 17448981]
20. Cook, Daniel L.; Bookstein, Fred L.; Gennari, John H. Physical properties of biological entities: An introduction to the ontology of physics for biology. *PLoS ONE*. 2011; 6(12):e28708, 12. [PubMed: 22216106]
21. Rachel Drysdale and FlyBase Consortium. FlyBase : a database for the drosophila research community. *Methods in molecular biology* (Clifton, NJ). 2008; 420:45–59.
22. Rachel Drysdale and FlyBase Consortium. FlyBase : a database for the drosophila research community. *Methods in molecular biology* (Clifton, NJ). 2008; 420:45–59.
23. Engel, Stacia R.; Balakrishnan, Rama; Binkley, Gail; Christie, Karen R.; Costanzo, Maria C.; Dwight, Selina S.; Fisk, Dianna G.; Hirschman, Jodi E.; Hitz, Benjamin C.; Hong, Eurie L.; Krieger, Cynthia J.; Livstone, Michael S.; Miyasato, Stuart R.; Nash, Robert; Oughtred, Rose; Park, Julie; Skrzypek, Marek S.; Weng, Shuai; Wong, Edith D.; Dolinski, Kara; Botstein, David; Michael Cherry, J. *Saccharomyces genome database provides mutant phenotype data*. *Nucleic acids research*. Jan.2010 38(Database issue)
24. Espinosa, Octavio; Hancock, John M. A gene-phenotype network for the laboratory mouse and its implications for systematic phenotyping. *PLoS ONE*. 2011; 6(5):e19693, 05. [PubMed: 21625554]
25. Fawcett, Tom. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006; 27(8):861–874. *ROC Analysis in Pattern Recognition*.
26. Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D. Building mouse phenotype ontologies. *Pac Symp Biocomput*. 2004:178–189. [PubMed: 14992502]
27. Gkoutos, GV.; Green, ECJ.; Mallon, AM.; Hancock, JM.; Davidson, D. In: Altman, Russ B.; Dunker, Keith A.; Hunter, Lawrence; Jung, Tiffany A.; Klein, Teri E., editors. *Building mouse phenotype ontologies; Proceedings of the 9th Pacific Symposium on Biocomputing (PSB 2004); Hawaii, USA*. Jan 6–10; London: World Scientific; 2004. p. 178-89.
28. Gkoutos, Georgios V. Towards a phenotypic semantic web. *Current Bioinformatics*. May; 2006 1(2):235–246.
29. Gkoutos, Georgios V.; Green, Eain C.; Mallon, Ann-Marie M.; Hancock, John M.; Davidson, Duncan. Using ontologies to describe mouse phenotypes. *Genome biology*. 2005; 6(1)

30. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*. 2008; 41(5):687–693. 10. [PubMed: 18358788]
31. Golbreich, Christine; Horrocks, Ian. The obo to owl mapping, go to owl 1.1!. In: Golbreich, Christine; Kalyanpur, Aditya; Parsia, Bijan, editors. *Proceedings of OWL: Experiences and Directions 2007 (OWLED-2007)*. CEUR-WS.org; 2007.
32. Grau B, Horrocks I, Motik B, Parsia B, Patelschneider P, Sattler U. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*. Nov; 2008 6(4): 309–322.
33. Groth, Philip, Weiss, Bertram. Phenotype data: A neglected resource in biomedical research? *Current Bioinformatics*. Aug; 2006 1(3):347–358.
34. Groth P, Pavlova N, Kaley I, Tonov S, Georgiev G, Pohlenz HD, Weiss B. PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res*. Jan.2007 35(Database issue)
35. Gruber, TR. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: Guarino, N.; Poli, R., editors. *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Deventer, The Netherlands: Kluwer Academic Publishers; 1993.
36. Gruber, Thomas R. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*. 1995; 43(5–6)
37. Guarino, Nicola. Formal ontology and information systems. In: Guarino, Nicola, editor. *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*. IOS Press; 1998. p. 3-15.
38. Harris, Todd W.; Antoshechkin, Igor; Bieri, Tamberlyn; Blasiar, Darin; Chan, Juancarlos; Chen, Wen J.; De La Cruz, Norie; Davis, Paul; Duesbury, Margaret; Fang, Ruihua; Fernandes, Jolene; Han, Michael; Kishore, Ranjana; Lee, Raymond; Müller, Hans-Michael; Nakamura, Cecilia; Ozersky, Philip; Petcherski, Andrei; Rangarajan, Arun; Rogers, Anthony; Schindelman, Gary; Schwarz, Erich M.; Tuli, Mary A.; Van Auker, Kimberly; Wang, Daniel; Wang, Xiaodong; Williams, Gary; Yook, Karen; Durbin, Richard; Stein, Lincoln D.; Spieth, John; Sternberg, Paul W. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*. Jan; 2010 38(suppl 1):D463–D467. [PubMed: 19910365]
39. Hernandez-Boussard, Tina; Whirl-Carrillo, Michelle; Hebert, Joan M.; Gong, Li; Owen, Ryan; Gong, Mei; Gor, Winston; Liu, Feng; Truong, Chuong; Whaley, Ryan; Woon, Mark; Zhou, Tina; Altman, Russ B.; Klein, Teri E. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic acids research*. Jan.2008 36(Database issue)
40. Herre, Heinrich; Heller, Barbara; Burek, Patryk; Hoehndorf, Robert; Loebe, Frank; Michalek, Hannes. *Onto-Med Report*. Vol. 8. IMISE, University of Leipzig; Leipzig, Germany: 2006. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0].
41. Hilbert, David. *Axiomatisches Denken*. *Mathematische Annalen*. 1918; 78:405–415.
42. Hoehndorf, Robert; Batchelor, Colin; Bittner, Thomas; Dumontier, Michel; Eilbeck, Karen; Knight, Rob; Mungall, Chris J.; Richardson, Jane S.; Stombaugh, Jesse; Westhof, Eric; Zhirbel, Craig L.; Leontis, Neocles B. The RNA ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology*. Apr; 2011 6(1):53–89.
43. Hoehndorf, Robert; Dumontier, Michel; Gennari, John H.; Wimalaratne, Sarala; Bono, Bernard de; Cook, Daniel L.; Gkoutos, Georgios V. Integrating systems biology models and biomedical ontologies. *BMC Systems Biology*. Aug.2011 5(1):124+. [PubMed: 21835028]
44. Hoehndorf, Robert; Dumontier, Michel; Oellrich, Anika; Rebholz-Schuhmann, Dietrich; Schofield, Paul N.; Gkoutos, Georgios V. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLOS ONE*. Jul.2011 6(7):e22006. [PubMed: 21789201]
45. Hoehndorf, Robert; Dumontier, Michel; Oellrich, Anika; Wimalaratne, Sarala; Rebholz-Schuhmann, Dietrich; Schofield, Paul; Gkoutos, Georgios V. A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*. Apr; 2011 27(7):1001–1008. [PubMed: 21343142]
46. Hoehndorf, Robert; Kelso, Janet; Herre, Heinrich. The ontology of biological sequences. *BMC bioinformatics*. Nov.2009 10(1):377+. [PubMed: 19919720]

47. Hoehndorf, Robert; Loebe, Frank; Kelso, Janet; Herre, Heinrich. Representing default knowledge in biomedical ontologies: Application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*. 2007; 8(1):377. [PubMed: 17925014]
48. Hoehndorf, Robert; Oellrich, Anika; Dumontier, Michel; Kelso, Janet; Rebholz-Schuhmann, Dietrich; Herre, Heinrich. Relations as patterns: Bridging the gap between OBO and OWL. *BMC Bioinformatics*. 2010; 11(1):441+. [PubMed: 20807438]
49. Hoehndorf, Robert; Oellrich, Anika; Rebholz-Schuhmann, Dietrich. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*. 2010; 26(24):3112–3118. 10. [PubMed: 20971987]
50. Hoehndorf, Robert; Oellrich, Anika; Rebholz-Schuhmann, Dietrich; Schofield, Paul N.; Gkoutos, Georgios V. Linking pharmgkb to phenotype studies and animal models of disease for drug repurposing. *Pacific Symposium on Biocomputing (PSB)*. 2012:388–399.
51. Hoehndorf, Robert; Schofield, Paul N.; Gkoutos, Georgios V. Phenomebrowser. 2011. <http://phenomebrowser.net>
52. Hoehndorf, Robert; Schofield, Paul N.; Gkoutos, Georgios V. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*. 2011; 39(18):e119. [PubMed: 21737429]
53. Horrocks, Ian. Technical report. University of Manchester; Mar. 2007 OBO flat file format syntax and semantics and mapping to OWL Web Ontology Language. <http://www.cs.man.ac.uk/~horrocks/obo/>
54. ISO. Technical report. ISO International Standard; 2007. Information technology – common logic (cl): a framework for a family of logic-based languages.
55. Kazakov, Yevgeny. Consequence-driven reasoning for Horn SHIQ ontologies. *Proceedings of the 21st International Conference on Artificial Intelligence (IJCAI 2009)*; July 11–17 2009; p. 2040-2045.
56. Kazakov, Yevgeny; Krötzsch, Markus; Simancík, František. Unchain my *EL* reasoner. *Proceedings of the 23rd International Workshop on Description Logics (DL'10), CEUR Workshop Proceedings*; CEUR-WS.org. 2011.
57. Kitsios GD, Tangri N, Castaldi PJ, Ioannidis JP. Laboratory mouse models for the human genome-wide associations. *PLoS One*. 2010; 5(11):e13782. [PubMed: 21072174]
58. Kutz, Oliver; Mossakowski, Till. A modular consistency proof for dolce. *AAAI*. 2011
59. Lander, Eric S. Initial impact of the sequencing of the human genome. *Nature*. Feb; 2011 470(7333):187–197. [PubMed: 21307931]
60. Lee, David; Redfern, Oliver; Orengo, Christine. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*. Dec; 2007 8(12):995–1005.
61. Loewenstein, Yaniv; Raimondo, Domenico; Redfern, Oliver; Watson, James; Frishman, Dmitriy; Linial, Michal; Orengo, Christine; Thornton, Janet; Tramontano, Anna. Protein function annotation by homology-based inference. *Genome Biology*. 2009; 10(2):207+. [PubMed: 19226439]
62. Masuya, Hiroshi; Makita, Yuko; Kobayashi, Norio; Nishikata, Koro; Yoshida, Yuko; Mochizuki, Yoshiki; Doi, Koji; Takatsuki, Terue; Waki, Kazunori; Tanaka, Nobuhiko; Ishii, Manabu; Matsushima, Akihiro; Takahashi, Satoshi; Hijikata, Atsushi; Kozaki, Kouji; Furuichi, Teiichi; Kawaji, Hideya; Wakana, Shigeharu; Nakamura, Yukio; Yoshiki, Atsushi; Murata, Takehide; Fukami-Kobayashi, Kaoru; Mohan, Sujatha; Ohara, Osamu; Hayashizaki, Yoshihide; Mizoguchi, Riichiro; Obata, Yuichi; Toyoda, Tetsuro. The RIKEN integrated database of mammals. *Nucleic Acids Research*. Jan; 2011 39(suppl 1):D861–D870. [PubMed: 21076152]
63. McGary, Kriston L.; Park, Tae J.; Woods, John O.; Cha, Hye J.; Wallingford, John B.; Marcotte, Edward M. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*. Apr; 2010 107(14):6544–6549.
64. Motik, Boris; Grau, Bernardo Cuenca; Horrocks, Ian; Wu, Zhe; Fokoue, Achille; Lutz, Carsten. Recommendation. *World Wide Web Consortium (W3C)*; 2009. Owl 2 web ontology language: Profiles.
65. Motik, Boris; Shearer, Rob; Horrocks, Ian. Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research*. 2009; 36:165–228.

66. Mungall, Christopher; Gkoutos, Georgios; Smith, Cynthia; Haendel, Melissa; Lewis, Suzanna; Ashburner, Michael. Integrating phenotype ontologies across multiple species. *Genome Biology*. 2010; 11(1):R2+. [PubMed: 20064205]
67. Mungall, Christopher J. Technical report. Lawrence Berkeley National Laboratory; 2011. OBO flat file format 1.4 syntax and semantics [draft]. <http://berkeleybop.org/~cjm/obo2owl/obo-syntax.html>
68. Mungall, Christopher J.; Bada, Michael; Berardini, Tanya Z.; Deegan, Jennifer; Ireland, Amelia; Harris, Midori A.; Hill, David P.; Lomax, Jane. Cross-product extensions of the gene ontology. *Journal of biomedical informatics*. 2011; 44(1):80–86. [PubMed: 20152934]
69. Mungall, Christopher J.; Batchelor, Colin; Eilbeck, Karen. Evolution of the sequence ontology terms and relationships. *Journal of Biomedical Informatics*. Mar.2010
70. Oti M, Brunner HG. The modular nature of genetic diseases. *Clinical Genetics*. 2007; 71:1–11. [PubMed: 17204041]
71. Oti, Martin; Huynen, Martijn A.; Brunner, Han G. The biological coherence of human phenome databases. *Am J Hum Genet*. Dec; 2009 85(6):801–808. [PubMed: 20004759]
72. Rector, Alan L. Modularisation of domain ontologies implemented in description logics and related formalisms including owl. K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture; New York, NY, USA. ACM Press; 2003. p. 121-128.
73. Robinson PN, Koehler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*. 2008; 83(5):610–615. [PubMed: 18950739]
74. Rutenber, Alan; Clark, Tim; Bug, William; Samwald, Matthias; Bodenreider, Olivier; Chen, Helen; Doherty, Donald; Forsberg, Kerstin; Gao, Yong; Kashyap, Vipul; Kinoshita, June; Luciano, Joanne; Scott Marshall, M.; Ogbuji, Chimezie; Rees, Jonathan; Stephens, Susie; Wong, Gwendolyn; Wu, Elizabeth; Zaccagnini, Davide; Hongsermeier, Tonya; Neumann, Eric; Herman, Ivan; Cheung, Kei H. Advancing translational research with the semantic web. *BMC Bioinformatics*. 2007; 8(Suppl 3):S2+. [PubMed: 17493285]
75. Sardana, Divya; Vasa, Suresh; Vepachedu, Nishanth; Chen, Jing; Gudivada, Ranga Chandra; Aronow, Bruce J.; Jegga, Anil G. PhenoHM: human-mouse comparative phenome-genome server. *Nucleic Acids Research*. 2010
76. Schindelman, Gary; Fernandes, Jolene; Bastiani, Carol; Yook, Karen; Sternberg, Paul. Worm phenotype ontology: integrating phenotype data within and beyond the c. elegans community. *BMC Bioinformatics*. 2011; 12(1):32. [PubMed: 21261995]
77. Schofield, Paul N.; Hoehndorf, Robert; Gkoutos, Georgios V. Mouse genetic and phenotypic resources for human genetics. 2012 in press.
78. Schulz, Stefan; Suntisrivaraporn, Boontawee; Baader, Franz; Boeker, Martin. SNOMED reaching its adolescence: Ontologists' and logicians' health check. *International Journal of Medical Informatics*. 2009; 78(Supplement 1):S86–S94. [PubMed: 18789754]
79. Sirin, Evren; Parsia, Bijan. In: Haarslev, Volker; Möller, Ralf, editors. Pellet: An OWL DL reasoner; Proceedings of the 2004 International Workshop on Description Logics, DL2004, Whistler, British Columbia, Canada, Jun 6–8, volume 104 of CEUR Workshop Proceedings; Aachen, Germany. CEUR-WS.org; 2004.
80. Skarnes, William C.; Rosen, Barry; West, Anthony P.; Koutsourakis, Manousos; Bushell, Wendy; Iyer, Vivek; Mujica, Alejandro O.; Thomas, Mark; Harrow, Jennifer; Cox, Tony; Jackson, David; Severin, Jessica; Biggs, Patrick; Fu, Jun; Nefedov, Michael; de Jong, Pieter J.; Francis Stewart, A.; Bradley, Allan. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*. Jun; 2011 474(7351):337–342. [PubMed: 21677750]
81. Smedley, Damian; Swertz, Morris A.; Wolstencroft, Katy; Proctor, Glenn; Zouberakis, Michael; Bard, Jonathan; Hancock, John M.; Schofield, Paul. Solutions for data integration in functional genomics: a critical assessment and case study. *Briefings in bioinformatics*. Nov; 2008 9(6):532–544. [PubMed: 19112082]
82. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. *Genome Biol*. 2005; 6(5):R46. [PubMed: 15892874]

83. Smith, Barry; Ashburner, Michael; Rosse, Cornelius; Bard, Jonathan; Bug, William; Ceusters, Werner; Goldberg, Louis J.; Eilbeck, Karen; Ireland, Amelia; Mungall, Christopher J.; Leontis, Neocles; Serra, Philippe R.; Ruttenberg, Alan; Sansone, Susanna A.; Scheuermann, Richard H.; Shah, Nigam; Whetzel, Patricia L.; Lewis, Suzanna. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech.* 2007; 25(11):1251–1255.
84. Smith, Cynthia L.; Goldsmith, Carroll-Ann W.; Eppig, Janan T. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology.* 2004; 6(1):R7. [PubMed: 15642099]
85. Sowa, John F. Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole; Pacific Grove: 2000.
86. Sprague, Judy; Bayraktaroglu, Leyla; Bradford, Yvonne; Conlin, Tom; Dunn, Nathan; Fashena, David; Frazer, Ken; Haendel, Melissa; Howe, Douglas G.; Knight, Jonathan; Mani, Prita; Moxon, Sierra A.; Pich, Christian; Ramachandran, Sridhar; Schaper, Kevin; Segerdell, Erik; Shao, Xiang; Singer, Amy; Song, Peiran; Sprunger, Brock; Van Slyke, Ceri E.; Westerfield, Monte. The zebrafish information network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucl Acids Res.* Nov.2007 :gkm956+.
87. Subramanian, Aravind; Tamayo, Pablo; Mootha, Vamsi K.; Mukherjee, Sayan; Ebert, Benjamin L.; Gillette, Michael A.; Paulovich, Amanda; Pomeroy, Scott L.; Golub, Todd R.; Lander, Eric S.; Mesirov, Jill P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America.* 2005; 102(43):15545–15550. [PubMed: 16199517]
88. Suntisrivaraporn, Boontawe. Empirical evaluation of reasoning in lightweight DLs on life science ontologies. *Proceedings of the 2nd Mahasarakham International Workshop on AI (MIWAI'08);* 2008.
89. Thorisson, Gudmundur A.; Muilu, Juha; Brookes, Anthony J. Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nature reviews. Genetics.* Jan; 2009 10(1):9–18.
90. Tsarkov, D.; Horrocks, I. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 4130. LNAI; 2006. FaCT++ description logic reasoner: System description; p. 292-297.
91. van Driel, Marc A.; Bruggeman, Jorn; Vriend, Gert; Brunner, Han G.; Leunissen, Jack AM. A text-mining analysis of the human phenome. *European Journal of Human Genetics.* Feb; 2006 14(5): 535–542. [PubMed: 16493445]
92. Wallace, Helen A.; Marques-Kranc, Fatima; Richardson, Melville; Luna-Crespo, Francisco; Sharpe, Jackie A.; Hughes, Jim; Wood, William G.; Higgs, Douglas R.; Smith, Andrew J. Manipulating the mouse genome to engineer precise functional syntenic replacements[no-break space]with human sequence. *Cell.* Jan; 2007 128(1):197–209. [PubMed: 17218265]
93. Washington, Nicole L.; Haendel, Melissa A.; Mungall, Christopher J.; Ashburner, Michael; Westerfield, Monte; Lewis, Suzanna E. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* 2009; 7(11):e1000247, 11. [PubMed: 19956802]
94. Weinreich, Steffanie S.; Mangon, R.; Sikkens, JJ.; Teeuw, MC.; Cornel, ME. Orphanet: a european database for rare diseases. *Ned Tijdschr Geneesk.* Mar; 2008 9(152):518–9. [PubMed: 18389888]
95. Wolstencroft K, Lord P, Taberner L, Brass A, Stevens R. Protein classification using ontology classification. *Bioinformatics.* Jul; 2006 22(14):e530–e538. [PubMed: 16873517]
96. Xu, Tao; Du, LinFang; Zhou, Yan. Evaluation of GO-based functional similarity measures using *s. cerevisiae* protein interaction and expression profile data. *BMC Bioinformatics.* 2008; 9(1):472. [PubMed: 18986551]
97. Yamazaki, Yukiko; Jaiswal, Pankaj. Biological ontologies in rice databases. an introduction to the activities in gramene and oryzabase. *Plant Cell Physiol.* 2005; 46(1)
98. Zheng-Bradley, Xiangqun; Rung, Johan; Parkinson, Helen; Brazma, Alvis. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology.* 2010; 11(12):R124. [PubMed: 21182765]

Figure 1.

A plot of the true positive rate vs. the false positive rate for the task of identifying associations between mouse models and diseases. The set of true positive instances is taken from the MGI database while negative and unknown associations between a mouse model and a disease constitute negative instances. The area under the ROC curve (AUC) of *PhenomeNet* for this task is 0.868.