Title

The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species



Journal

Nucleic Acids Research, 45(D1)

ISSN

0305-1048

Authors

Mungall, Christopher J

McMurry, Julie A

Köhler, Sebastian

et al.

Publication Date

2017-01-04

DOI

10.1093/nar/gkw1128

Peer reviewed

*D712–D722* *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 29 November 2016*
*doi: 10.1093/nar/gkw1128*

# The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species

**Christopher J. Mungall[1], Julie A. McMurry[2], Sebastian Köhler[3], James P. Balhoff[4], Charles Borromeo[5], Matthew Brush[2], Seth Carbon[1], Tom Conlin[2], Nathan Dunn[1], Mark Engelstad[2], Erin Foster[2], J.P. Gourdine[2], Julius O.B. Jacobsen[6], Dan Keith[2], Bryan Laraway[2], Suzanna E. Lewis[1], Jeremy NguyenXuan[1], Kent Shefchek[2], Nicole Vasilevsky[2], Zhou Yuan[5], Nicole Washington[1], Harry Hochheiser[5], Tudor Groza[7], Damian Smedley[6], Peter N. Robinson[3,8] and Melissa A. Haendel[2,*]**

[1]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA, [2]Department of Medical Informatics and Clinical Epidemiology and OHSU Library, Oregon Health & Science University, Portland, OR, 97239, USA, [3]Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany, [4]RTI International, Research Triangle Park, NC, 27709, USA, [5]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15260, USA, [6]William Harvey Research Institute, Barts & The London School of Medicine & Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK, [7]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia and [8]The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06032mUSA

Received August 22, 2016; Revised October 26, 2016; Editorial Decision October 27, 2016; Accepted November 02, 2016

## ABSTRACT

**The correlation of phenotypic outcomes with genetic variation and environmental factors is a core pursuit in biology and biomedicine. Numerous challenges impede our progress: patient phenotypes may not match known diseases, candidate variants may be in genes that have not been characterized, model organisms may not recapitulate human or veterinary diseases, filling evolutionary gaps is difficult, and many resources must be queried to find potentially significant genotype–phenotype associations. Non-human organisms have proven instrumental in revealing biological mechanisms. Advanced informatics tools can identify phenotypically relevant disease models in research and diagnostic contexts. Large-scale integration of model organism and clinical research data can provide a breadth of knowledge not available from individual sources and can provide contextualization of data back to these sources. The Monarch Initiative (monarchinitiative.org) is a collaborative, open science effort that aims to semantically integrate genotype–phenotype data from many species and sources in order to support precision medicine, disease modeling, and mechanistic exploration. Our integrated knowledge graph, analytic tools, and web services enable diverse users to explore relationships between phenotypes and genotypes across species.**

## INTRODUCTION

A fundamental axiom of biology is that phenotypic manifestations of an organism are due to interaction between genotype and environmental factors over time. In the rapidly advancing era of genomic medicine, a critical challenge is to identify the genetic etiologies of Mendelian disease, cancer, and common and complex diseases, and translate basic science to better treatments. Currently, available human data associates ∼<51% of known human coding genes with phenotype data (based on OMIM (1), ClinVar (2), Orphanet (3), CTD (4) and the GWAS catalog (5)). See Table 1 for a list of database abbreviations. This coverage can be extended to ∼89% if phenotypic information from orthologous genes from five of the most well-studied

*To whom correspondence should be addressed. Tel: +1 503 407 5970; Email: Haendel@ohsu.edu
Present address: Melissa A. Haendel, Department of Medical Informatics and Clinical Epidemiology and OHSU Library, Oregon Health & Science University, Portland, OR, USA.

model organisms is included (Figure 1). Similarly, of the 72% of the 3230 genes in ExAC with 'near-complete depletion of predicted protein-truncating variants have no currently established human disease phenotype' (6), where 88% of these genes without a human phenotype have a phenotype in a non-human organism. However, leveraging these model data for computational use is non-trivial primarily because the relationships between gene and disease (7) and between model system and disease phenotypes (8) are not straightforward.

In recent years, there has been a growth in the number of genotype–phenotype databases available, covering a diversity of domain areas for human, model organisms, and veterinary species. While providing quality inventories of the relevant species and phenotypic data types, most resources are limited to a single species or limit cross-species comparison to direct assertions (e.g. Organism X is a model of Disease Y) or based upon orthology relations (e.g. organism Z is a model of Disease Y due to A and A′ being orthologs). While great strides have been made in text-based search engines, phenotype data remains difficult to search and use computationally due to its complexity and in the use of different phenotype standards and terminologies. Such barriers have made linking and integration with the precision and richness needed for mechanistic discovery across species a significant challenge (9). A newer method to aid identifying models of disease and to discover underlying mechanisms is to utilize ontologies to describe the set of phenotypes that present for a given genotype or disease, what we call a 'phenotypic profile'. A phenotypic profile is the subject of non-exact matching within and across species using ontology integration and semantic similarity algorithms (10,11) in software applications such as Exomiser (12) and Genomiser (13), and this approach has been shown to assist disease diagnosis (14–16). The Monarch Initiative uses an ontology-based strategy to deeply integrate genotype–phenotype data from many species and sources, thereby enabling computational interrogation of disease models and complex relationships between genotype and phenotype to be revealed. The name 'Monarch Initiative' was chosen because it is a community effort to create paths for diverse data to be put to use for disease discovery, not unlike the navigation routes that a monarch butterfly would take.

## Data architecture

The overall data architecture for Monarch is shown in Figure 2. The bulk of the data integration is carried out using our Data Ingest Pipeline (Dipper) tool (https://github.com/monarch-initiative/dipper), which maps a variety of external data sources and databases to RDF (Resource Description Framework) graphs. RDF provides a flexible way of modeling a variety of complex datatypes, and allows entities from different databases to be connected via common instance or class URIs (Uniform Resource Indicators). We use relationship types from the Relation Ontology (RO; https://github.com/oborel/obo-relations) (17) and other vocabularies to connect entities together, along with a number of Open Biological Ontologies (18) (OBOs) to classify these entities. For example, a mouse genotype can be related to a phenotype using the *has_phenotype* relation (RO:0002200), with the genotype classified using a term from the Genotype Ontology (GENO) (19), and the phenotype classified using the Mammalian Phenotype Ontology (MP) (20). We use the Open Biomedical Annotations (OBAN; https://github.com/EBISPOT/OBAN) vocabulary to associate evidence and provenance metadata with each edge, using the Evidence and Conclusions Ontology (ECO) for types of evidence (21).The graphs produced by Dipper are available as a standalone resource in RDF/turtle format at http://data.monarchinitiative.org/ttl.

We also import a number of external and in-house ontologies, for data description and data integration. As these ontologies are all available from the OBO Library in Web Ontology Language (OWL), no additional transformation is necessary. The combined corpus of graphs ingested using Dipper and from ontologies is referred to as the **Monarch Knowledge Graph**. The data integrated within Monarch encompasses a wide range of sources, and includes human clinical knowledge sources as well as genetic and genomic resources covering organismal biology. The list of data sources and ontologies integrated is shown in Figure 3, with a species distribution illustrated in Figure 4. The knowledge graph is loaded into an instance of a SciGraph database (https://github.com/SciGraph/SciGraph/), which embeds and extends a Neo4J database, allowing for complex queries and ontology-aware data processing and Named Entity Recognition. We provide two public endpoints for client software to query these services: https://scigraph-ontology.monarchinitiative.org/scigraph/docs (for ontology access) and https://scigraph-data.monarchinitiative.org/scigraph/docs (for ontology plus data access).

These SciGraph instances provide powerful graph querying capabilities over the complete knowledge graph. Many of the common query patterns are executed in advance and stored in an Apache Solr index, making use of the Gene Ontology 'GOlr' indexing strategy, allowing for fast queries of ontology-indexed associations.

Finally, we also load a subset of the graph into an OwlSim instance, which provides phenotype matching services as well as the ability to perform fuzzy phenotype searches based on a phenotype profile. We also provide phenotype matching services via the Global Alliance for Genomes and Health (GA4GH) Matchmaker Exchange (MME) API MME (22), available at https://mme.monarchinitiative.org.

Many of the data sources we integrate make use of their own terminologies and ontologies. We aggregate these into a unified ontology (https://github.com/monarch-initiative/monarch-ontology/) and make use of bridging ontologies and our curated integrative ontologies to connect these together. In particular:

- The Uber-anatomy ontology (Uberon) bridges species-specific and clinical anatomical and tissue ontologies (23)
- The unified phenotype ontology bridges model organism and human phenotype ontologies and terminologies, using techniques described in (24,25)
- The Monarch Merged Disease Ontology (MonDO) uses a Bayes ontology merging algorithm (26) to integrate multiple human disease resources into a single ontology, and additionally includes animal diseases from OMIA.

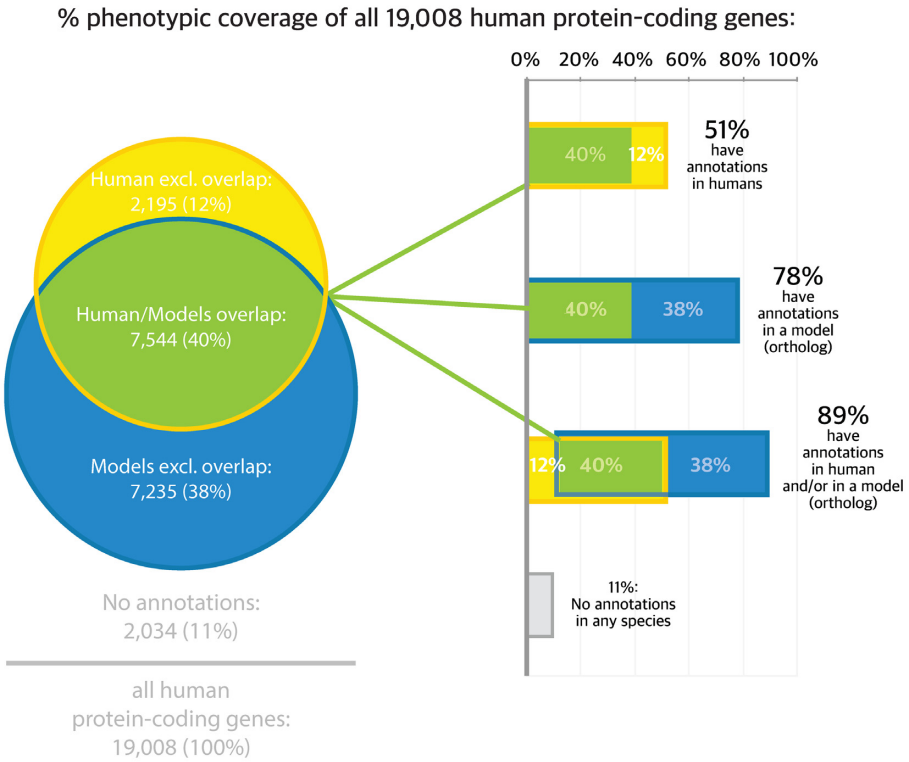


**Figure 1.** The phenotype annotation coverage of human coding genes. Yellow bars show that 51% of those genes have at least one phenotype association reported in humans (HPO annotations of OMIM, ClinVar, Orphanet, CTD and GWAS). The blue bars show that 58% of human coding genes have orthologs with causal phenotypic associations reported in at least one non-human model (MGI, Wormbase, Flybase and ZFIN). The green bars show that 40% of human coding genes have annotations both in human and in non-human orthologs. There are phenotypic associations from humans and/or non-human orthologs that cover 89% of human coding genes.

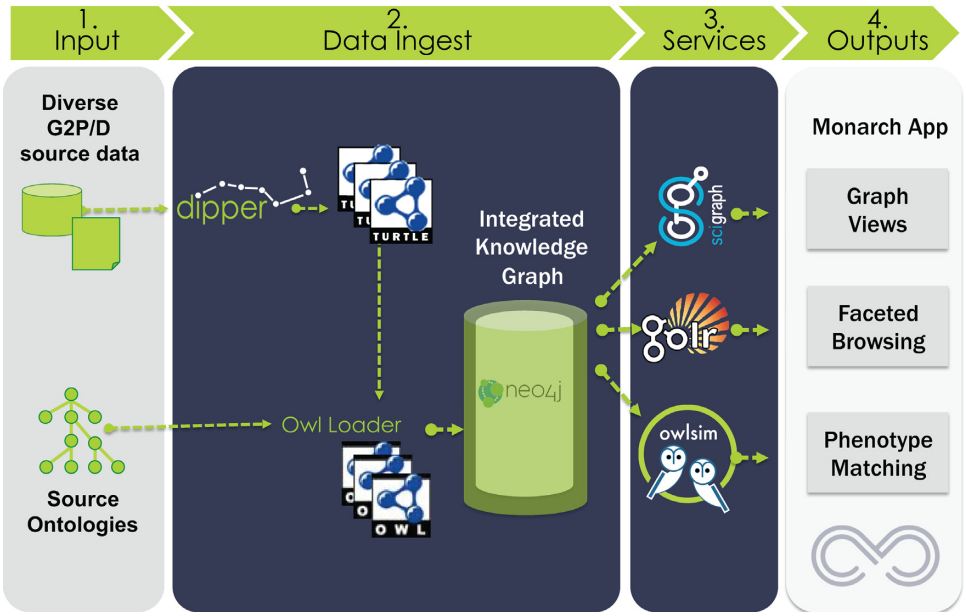


**Figure 2.** Monarch Data Architecture. Structured and unstructured data sources are loaded into SciGraph via Dipper. Ontologies are also loaded into SciGraph, resulting in a combined knowledge and data graph. Data is disseminated via SciGraph Services, an ontology-enhanced Solr instance called GOlr, and to the OwlSim semantic similarity software. Monarch applications and end users access the services for graph querying, application population and phenotype matching.

**Table 1.** Glossary of acronyms

| Acronym | Name | URL | Ref |
|---|---|---|---|
| Bgee | BgeeDb | http://bgee.org/ | (55) |
| BioGrid | Biological General Repository for Interaction Datasets. | https://thebiogrid.org/ | (33) |
| CL | Cell Ontology | http://obofoundry.org/ontology/cl.html | (62) |
| ClinVar | ClinVar | https://www.ncbi.nlm.nih.gov/clinvar/ | (2) |
| CTD | Clinical Toxicology Database | http://ctdbase.org/ | (4) |
| ECO | Evidence and Conclusions Ontology | http://obofoundry.org/ontology/eco.html | (21) |
| ExAC | Exome Aggregation Consortium | http://exac.broadinstitute.org/ | (6) |
| FlyBase | FlyBase | http://flybase.org | (63) |
| GeneNetwork | Gene Network | http://genenetwork.org | (54) |
| GENO | Genotype Ontology | https://github.com/monarch-initiative/GENO-ontology/ | (19) |
| GO | Gene Ontology | http://geneontology.org | (37) |
| GWAS | GWAS Catalog | https://www.ebi.ac.uk/gwas/ | (5) |
| HP | Human Phenotype Ontology | http://human-phenotype-ontology.org/ | (30) |
| KEGG | Kyoto Encyclopedia of Genes and Genomes | http://www.kegg.jp/ | (31) |
| MGI | Mouse Genome Informatics | http://www.informatics.jax.org/ | (36) |
| MonDO | Monarch Merged Disease Ontology | https://github.com/monarch-initiative/monarch-disease-ontology/ | (26) |
| MP | Mammalian Phenotype Ontology | http://obofoundry.org/ontology/mp.html | (20) |
| MPD | Mouse phenome database | http://phenome.jax.org/ | (53) |
| MyGene | MyGene | http://mygene.info | (32) |
| OMIA | Online Mendelian Inheritance in Animals | http://omia.angis.org.au/home/ | (41) |
| OMIM | Online Mendelian Inheritance in Man | http://omim.org | (1) |
| OrphaNet | Portal for rare diseases and orphan drugs | http://www.orpha.net | (3) |
| Panther | PantherDB | http://pantherdb.org | (34) |
| RO | Relation Ontology | http://obofoundry.org/ontology/ro.html | (17) |
| SEPIO | Scientific Evidence and Provenance Information Ontology | https://github.com/monarch-initiative/SEPIO-ontology/ | (59) |
| SO | Sequence Ontology | http://www.sequenceontology.org/ | (27) |
| Uberon | Uber-anatomy ontology | http://uberon.org | (23) |
| Upheno | Unified Phenotype Ontology | https://github.com/obophenotype/upheno/ | (25) |
| WormBase | WormBase | http://wormbase.org | (64) |
| ZFIN | Zebrafish Information Resource | http://zfin.org | (35) |

- The Genotype Ontology (GENO) (19) defines genotypic elements and bridges the Sequence Ontology (SO) (27) and FALDO (28). GENO allows the propagation of phenotypes that are annotated to genotypic elements.

### Entity resolution and unification

One of the many challenges faced when integrating bioinformatics resources is the presence of the same entity in multiple databases, designated by different identifiers (29). This problem is compounded by the different ways the same identifier can be written, using different prefixes or no prefix at all. Taking a Monarch page for a single gene, for example 'fibrinogen gamma chain', *FGG*, (https://monarchinitiative.org/gene/NCBIGene:2266). Monarch has integrated data from a variety of human, model organism, and other biomedical sources such as OMIM (1), Orphanet (3), ClinVar (2), HPO (30), KEGG (31), CTD (4), MyGene (32), BioGrid (33) and via orthology in PantherDB (34) we also incorporate *Fgg* gene data from ZFIN (35) and from MGI (36). No two of these sources represents the identifier for *FGG* in precisely the same way. As part of our data ingest process, we normalize all identifiers using a curated set of database prefixes. These have a defined mapping to an http URI. These curated prefixes have been deposited in the Prefix Commons (https://github.com/prefixcommons), which similarly contains identifier prefixes used within the Gene Ontology (37) and Bio2RDF (38).

In post-processing equivalent identifiers, we perform clique-merging (https://github.com/SciGraph/SciGraph/wiki/Post-processors). We take all edges labeled with either the owl:sameAs or owl:equivalentClasses property and calculate equivalence cliques, based on the symmetric and transitive nature of these properties. We then merge these cliques together, taking a designated 'clique leader' (for instance, NCBI for genes) and mapping all edges in the monarch graph such that they point to a clique leader.

### In-house curation

In addition to ingest of external sources and ontologies, we perform in-house data and ontology curation. For curation of ontology-based genotype–phenotype associations (including disease-phenotypic profiles), we are transitioning to the WebPhenote platform (http://create.monarchinitiative.org), which allows a variety of disease entities to be connected to phenotypic descriptors. We also make use of text mining to create seed disease-phenotype associations using the Bio-Lark toolkit (39), which are then manually curated. Most recently, we have performed a large-scale annotation of PubMed to extract common disease-phenotype associations (40). Most of the in-house curation work involves making smaller resources with free text descriptions of phenotypic information computable, for example, the Online Mendelian Inheritance in Animals (OMIA) resource, with

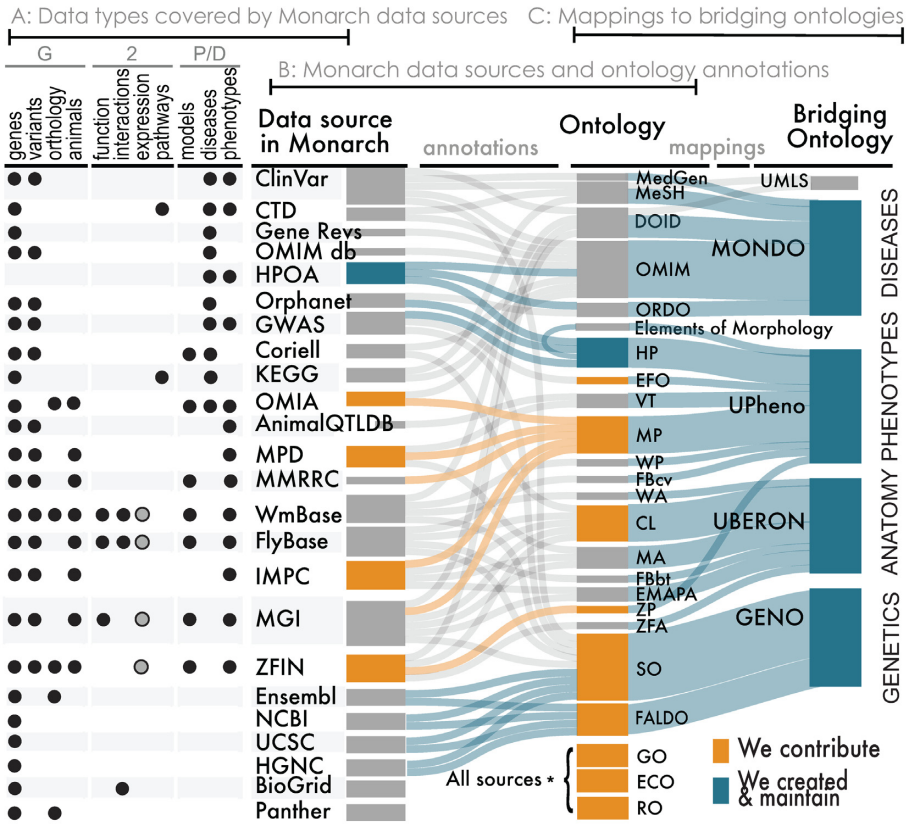


**Figure 3.** Data types, sources, and the ontologies used for their integration into the Monarch knowledge graph. Each data source uses or is mapped to a suite of different ontologies or vocabularies. These are in turn integrated into bridging ontologies for Genetics (GENO), Anatomy (Uberon/CL), Phenotypes (UPheno) and Diseases (MonDO).

whom we have been collaborating to support this curation (41).

## Quality control

External resources and datasets that are incorporated into Monarch are evaluated before incorporation into the Dipper pipeline—we primarily integrate high-quality curated resources. For all ontologies we bring in, we apply automated reasoning to detect inconsistencies between different ontologies. For each release, we perform high-level checks on each integrated resource to ensure no errors in the extraction process occurred, but we do not perform in-depth curation checks of integrated resources. Each release happens once every one to two months.

In order to measure annotation richness, we have also created an annotation sufficiency meter web service (42) available at https://monarchinitiative.org/page/services; this service determines whether a given phenotype profile for any organism is sufficiently broad and deep to be of diagnostic utility. The sufficiency score can be displayed as a five star scale as in PhenoTips (43) and in the Monarch web portal (see below) to aid curation or data entry, and can also be used to suggest additional phenotypic assays to be performed—whether in a patient or in a model organism.

## Monarch web portal

The Monarch portal is designed with a number of different use cases in mind, including:

- A researcher interested in a human gene, its phenotypes, and the phenotypes of orthologs in model organisms and other species
- Patients or researchers interested in a particular disease or phenotype (or groups of these), together with information on all implicated genes
- A clinical scenario in which a patient has an undiagnosed disease showing a spectrum of phenotypes, with no definitive candidate gene demonstrated by sequencing; in this scenario the clinician wishes to search for either known diseases that have a similar presentation, or model organism genes that demonstrate homologous phenotypes when the gene is perturbed
- Researcher looking for diseases that have similar phenotypic feature to a newly identified model organism mutant identified in a screen
- Researchers or clinicians who need to identify potentially informative phenotyping assays for differential diagnosis or to identify candidate genes

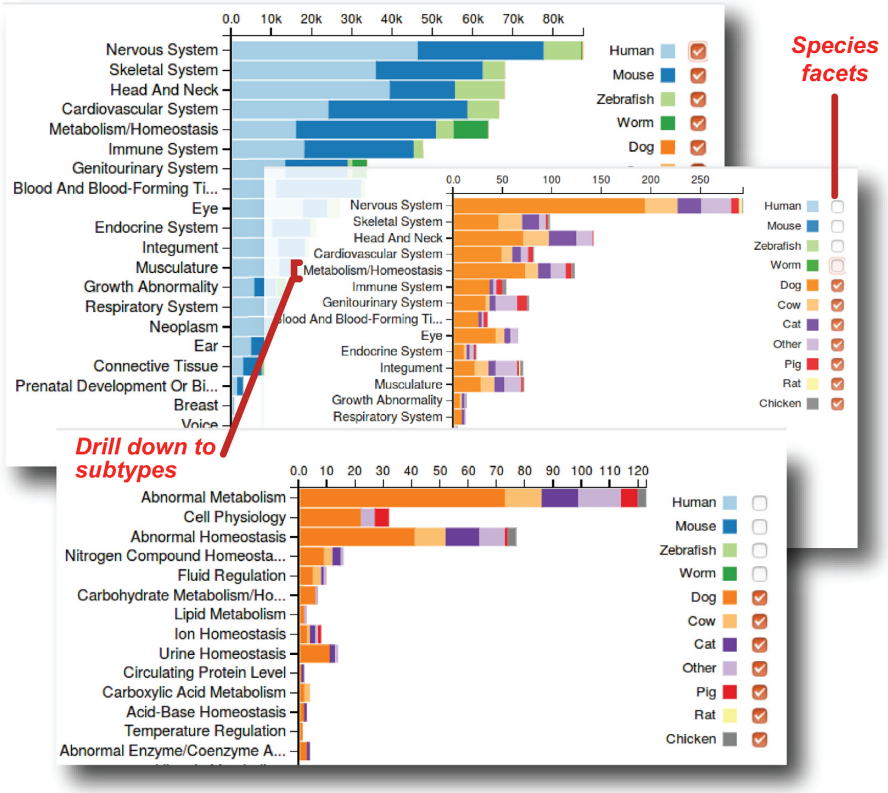


**Figure 4.** Distribution of phenotypic annotations across species in Monarch, broken down by the top levels of the phenotype ontology. The graph can be interactively explored at https://monarchinitiative.org/phenotype/. Note that annotations are currently dominated by human, mouse, zebrafish and *C. elegans* (top panel); the chart is faceted allowing individual species to be switched on and off to see contributions for less data-rich species such as veterinary animals and monkeys (middle panel). Clicking on a given phenotype text allows drilling down to its subtypes (lower panel).

## Features

*Integrated information on entities of interest.* We provide overview pages for entities such as genes, diseases, phenotypes, genotypes, variants and publications. Each page highlights the provenance of the data from the diverse clinical, model organism, and non-model organism sources. These pages can be found either via search (see below) or through an entity resolver. For example, the URL https://monarchinitiative.org/OMIM:266510 will redirect to a page about the disease 'Peroxisome biogenesis disorder type 3B' from the OMIM resource, showing its relationships to other content within the Monarch knowledge graph, such as phenotypes and genes associated with the disease. We make use of MonDO (the Monarch merged disease ontology (26)) to group similar diseases together. Figure 5 shows an example page for Marfan syndrome with related phenotype, gene, model and variant data.

*Basic Search.* The portal provides different means of searching over integrated content. In cases where a user is interested in a specific disease, gene, phenotype etc., these can usually be found via autocomplete. Site-wide synonym-aware text search can also be used to find pages of interest. Because the knowledgebase combines information from multiple species, entities such as genes often have ambiguous symbols. We provide species information to help disambiguate in a search.

*Search by phenotype profile.* One of the most innovative features of Monarch is the ability to query within and across species to look for diseases or organisms that share a set of similar but non-exact set of phenotypes (phenotypic profile). This feature uses a semantic similarity algorithm available from the OWLsim package (http://owlsim.org). Users can launch searches against specific targets: organisms, sets of named gene models, or against all models and diseases available in the Monarch repository. The Monarch Analyze Phenotypes interface (https://monarchinitiative.org/analyze/phenotypes) allows the user to build up a 'cart' of phenotypes, and then perform a comparison against phenotypes related to genes and diseases**.** Results are ranked according to closeness of match, partitioned by species, and are displayed as both a list and in the Phenogrid widget (below).

*Phenogrid.* Given a set of input phenotypes, as associated with a patient or a disease, Monarch phenotypic profile similarity calculations can generate results involving hundreds of diseases and models. The PhenoGrid visualization widget (Figure 6) provides an overview of these similarity results, implemented using the D3 javascript library (44). Phenotypes and models are frequently too numerous to fit

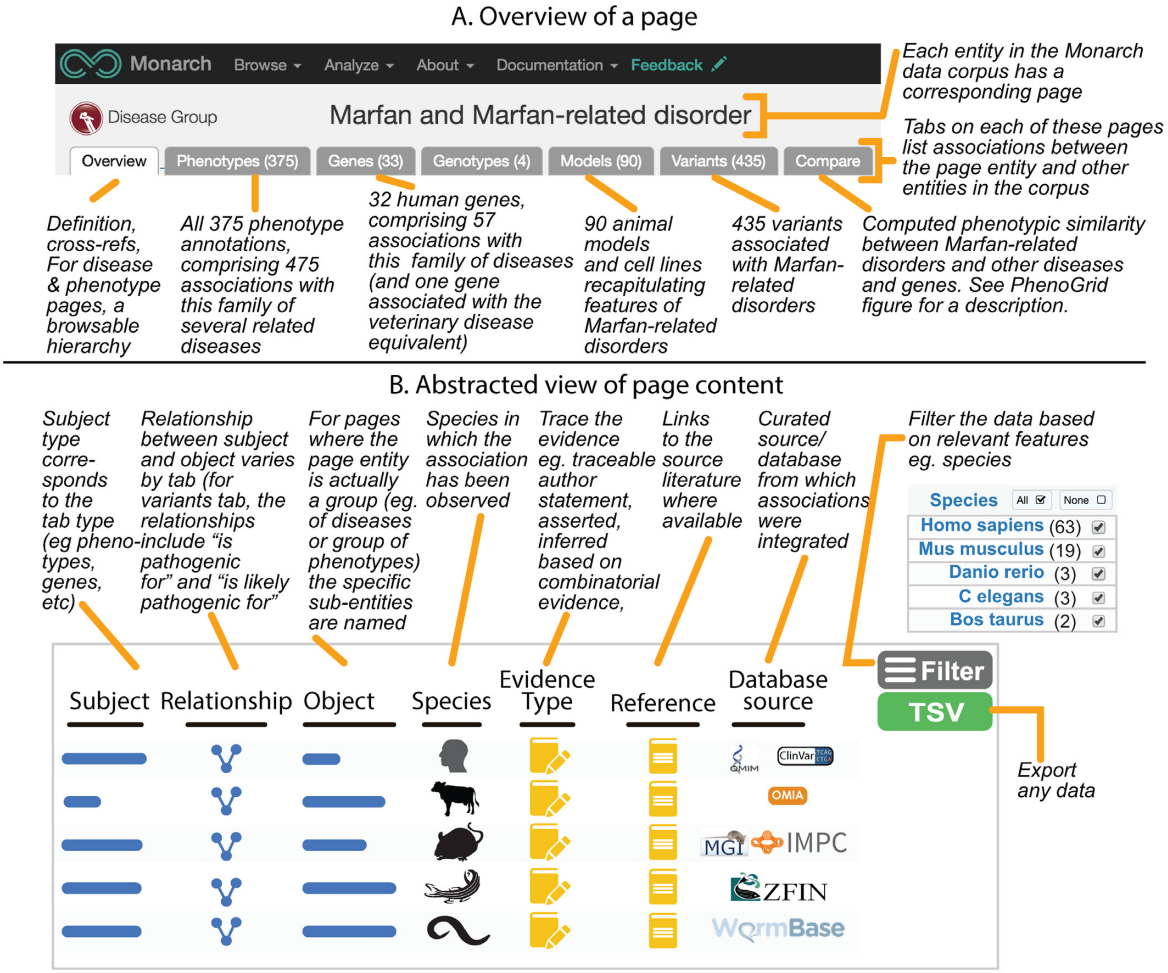


**Figure 5.** Annotated Monarch webpage for Marfan and Marfan Related syndrome. This group of syndromic diseases has a number of different associations spanning multiple entity types—disease phenotypes, implicated human genes, variants and animal models and other model systems. An abstraction of the contents and features of the tabs is shown in the lower panel. Actual contents of the tabs are best viewed in the context of the web app at https://monarchinitiative.org/DOID:14323.

on the initial display; thus scrolling, dragging, and filtering have been implemented. PhenoGrid is available as an open-source widget suitable for integration in third-party web sites, such as for model organism databases as done in the International Mouse Phenotyping Consortium (IMPC) or clinical comparison tools. Download and installation instructions are available on the Monarch Initiative web site.

*Text annotation.* The Monarch annotation service allows a user to enter free text (e.g. a paper abstract or a clinical narrative) and perform an automated annotation on this text, with entities in the text marked up with terms from the Monarch knowledge graph, such as genes, diseases and phenotypes. Once the text is marked up, the user has the option of turning the recognized phenotype terms into a phenotype profile, and performing a profile search, or to link to any of the entity pages identified in the annotation. This tool is also available via services.

*Inferring causative variants.* The Exomiser (12) and more recently, Genomiser tools (45) make use of the Monarch platform and phenotype matching algorithms to rank putative causative variants using a combined variant and phenotype score. These tools have been used to diagnose patients as part of the NIH Undiagnosed Diseases Project (14) and are the first examples of using model organism phenotype data to aid rare disease diagnostics.

## DISCUSSION

The Monarch Initiative provides a system to organize and harmonize the heterogeneous genotype–phenotype data found across clinical and model and non-model organism resources (such as veterinary species), creating a unified overview of this rich landscape of data sources. Some of the challenges we have had to address are that each resource shares data via different mechanisms and uses a different data model. It is particularly important to note that each organism annotates phenotypic data to different aspects of the genotype – one resource might be to a gene, another an allele, another to a set of alleles, a full genotype or a SNP. This not only makes data integration difficult, but it also means that computation over the genotype–phenotype associations must be done with care. Similar is-

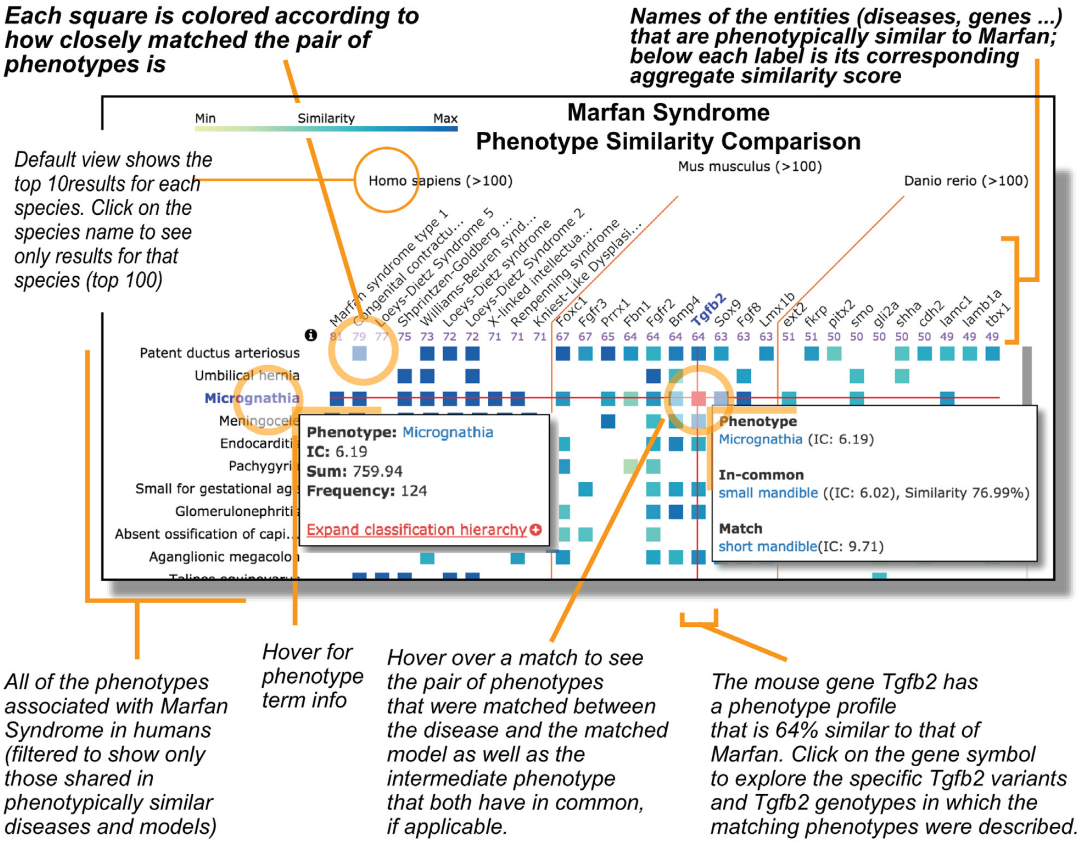


**Figure 6.** Partial screenshot of PhenoGrid showing Marfan syndrome. PhenoGrid shows input phenotypes in rows, models in columns, and cell contents color-coded with greater saturation indicating greater similarity. Disease phenotypes are shown as rows, and phenotypically matching human diseases and model organism genes are shown as columns—the saturation of a cell correlates with strength if phenotypic match. Mouse-over tooltips highlight diseases associated with a selected phenotype (or vice-versa), or details (including similarity scores) of any match between a phenotype and a model. User controls support the selection of alternative sort orders, similarity metrics, and displayed organism(s) (mouse, human, zebrafish or the 10 most similar models for each). Here, we see all diseases or genes that exhibit 'Hypoplasia of the mandible' with the matching mouse gene Tfgb2. Actual PhenoGrid data is best viewed in the context of the web app at https://monarchinitiative.org/Orphanet:284993#compare. Note matches do not need to be exact—here the mouse phenotype of 'small mandible' (Mouse Phenotype Ontology) has a high scoring match to 'micrognathia' (Human Phenotype Ontology) based on the fact that both phenotypes are related to 'small mandible' (Mouse Phenotype Ontology). Advanced PhenoGrid features (not displayed) include the ability to alter the scoring and sorting methods, as well as zoomed-out map-style navigation.

sues at MGI have been described (46). In addition, since most anatomy, phenotype, and disease ontologies describe the biology of one species, it has traditionally been quite difficult to 'map' across species. Some examples are the Human Phenotype Ontology (HPO) (30) and the Mouse Anatomy Ontology (47). Monarch uses four species-neutral ontologies that unify their species-specific counterparts (as shown in Figure 3): GENO for genotypes (19), UPheno for phenotypes (25), UBERON for anatomy (23), and MonDO for diseases (26). Prior efforts to map or integrate species-specific anatomical ontologies (24,48), for example, have been utilized in the construction of these species-neutral ontologies. The end result is a translational platform that allows a unified view of human, model and non-model organism biology.

A comparison between Monarch and existing resources is warranted. InterMine is an open-source data warehouse system used for disseminating data from large, complex biological heterogeneous data sources (49). InterMine provides sophisticated web services to support denormalized query and has been used to improve query and data access to model organism databases (50) and non-model organisms (51). InterMine is a federated approach where individual databases each can adopt and populate their own object-oriented data model, but can also align on certain aspects such as having genomic data models aligned using the SO. However, as yet genotype and phenotype modeling is not aligned, and Intermine does not provide disease matching or phenotypic search. We are currently working with InterMine to achieve harmonization in this area. Other resources, such as KaBOB (52) and Bio2RDF (38) semantically integrate various resources into large triplestores. Bio2RDF typically retains the source vocabulary of the integrated resources, whereas KaBOB is more similar to Monarch in that it maps OBO ontologies (18). Other data integration approaches include the BioThings API, exemplified by the MyVariant system (32) which aggregates variant data from multiple sources. We are currently working with the BioThings API developers to integrate these different approaches within the Dipper framework. Monarch is unique in that it aims to align both genotypic and phenotypic modeling across species and sources.

### Future directions

Future directions include bringing in phenotypic data from specialized sources and databases, incorporating a wider range of datatypes, and to extend and improve analytic methods for making cross-species inferences. Currently the core of Monarch includes primarily qualitative phenotypes described using terms from existing phenotypic vocabularies—we are starting to bring in more quantitative data, from sources such as the MPD (53) and GeneNetwork (54), in addition to expression data annotated to Uberon in BgeeDb (55). We are also extending our phenotypic search methods to incorporate Phenologs, phenotypic groupings inferred on the basis of orthologous genes (56,57). Early comparisons suggest that addition of phenologs to our suite of tools to enable genotype–phenotype inquiry across species will extend our reach in a synergistic manner (58). We therefore plan to implement this type of approach into the Monarch tool suite and website. One of the most important realizations we came across in constructing the Monarch platform was the need to better represent scientific evidence of genotype–phenotype associations. We are currently developing a Scientific Evidence and Provenance Information Ontology (SEPIO) (59) in collaboration with the Evidence and Conclusion Ontology consortium (21) and ClinGen (60) in order to classify associations as complementary, confirmatory, or contradictory. SEPIO will also integrate biological assays from the Ontology of Biomedical Investigations (61). Monarch has also been collaborating with the US National Cancer Institute's Thesaurus (NCIT) team to integrate cancer phenotypes. Finally, Monarch has been working in the context of the Global Alliance for Genomics and Health (GA4GH) to develop a formal phenotype exchange format (www.phenopackets.org) that can aid phenotypic data sharing in numerous contexts such as clinical, model organism research, biodiversity, veterinary, and evolutionary biology.

## ACKNOWLEDGEMENTS

We thank members of the Undiagnosed Disease Program, the International Mouse Phenotyping Consortium, the NCI Semantic Infrastructure team, and NIF/SciCrunch for their contributions.

## FUNDING

National Institutes of Health (NIH) [1R24OD011883]; Wellcome Trust [098051]; NIH Undiagnosed Disease Program [HHSN268201300036C, HHSN268201400093P]; Phenotype RCN [NSF-DEB-0956049]; NCI/Leidos [15x143, BD2K U54HG007990-S2 (Haussler; GA4GH), BD2K PA-15-144-U01 (Kesselman; FaceBase)]; Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy [DE- AC02-05CH11231 to J.N.Y., S.C., S.E.L. and C.J.M.]. Funding for open access charge: NIH [1R24OD011883].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
2. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D88.
3. Rath,A., Olry,A., Dhombres,F., Brandt,M.M., Urbero,B. and Ayme,S. (2012) Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat.*, **33**, 803–808.
4. Davis,A.P., Grondin,C.J., Johnson,R.J., Sciaky,D., King,B.L., McMorran,R., Wiegers,J., Wiegers,T.C. and Mattingly,C.J. (2017) The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.*, doi:10.1093/nar/gkw838.
5. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
6. Lek,M., Karczewski,K.J., Minikel,E V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016 ) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
7. Strohman,R. (2002) Maneuvering in the complex path from genotype to phenotype. *Science*, **296**, 701–703.
8. Houle,D., Govindaraju,D.R. and Omholt,S. (2010) Phenomics: the next challenge. *Nat. Rev. Genet.*, **11**, 855–866.
9. McMurry,J.A., Köhler,S., Washington,N.L., Balhoff,J.P., Borromeo,C., Brush,M., Carbon,S., Conlin,T., Dunn,N., Engelstad,M. *et al.* (2016) Navigating the phenotype frontier: the Monarch Initiative. *Genetics*, **203**, 1491–1495.
10. Smedley,D., Oellrich,A., Köhler,S., Ruef,B., Westerfield,M., Robinson,P., Lewis,S., Mungall,C. and Sanger Mouse Genetics Project (2013) PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database (Oxford)*, bat025.
11. Washington,N.L., Haendel,M.A., Mungall,C.J., Ashburner,M., Westerfield,M. and Lewis,S.E. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.*, **7**, e1000247.
12. Robinson,P.N., Kohler,S., Oellrich,A., Genetics,S.M., Wang,K., Mungall,C.J., Lewis,S.E., Washington,N., Bauer,S., Seelow,D. *et al.* (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, **24**, 340–348.
13. Smedley,D., Schubach,M., Jacobsen,J.O., Köhler,S., Zemojtel,T., Spielmann,M., Jäger,M., Hochheiser,H., Washington,N.L., McMurry,J.A. *et al.* (2016) A Whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *Am. J. Hum. Genet.*, **99**, 595–606.
14. Bone,W.P., Washington,N.L., Buske,O.J., Adams,D.R., Davis,J., Draper,D. *et al.* (2015) Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.*, **18**, 608–617.
15. Zemojtel,T., Kohler,S., Mackenroth,L., Jager,M., Hecht,J., Krawitz,P., Graul-Neumann,L., Doelken,S., Ehmke,N., Spielmann,M. *et al.* (2014) Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.*, **6**, 252ra123.
16. Wright,C.F., Fitzgerald,T.W., Jones,W.D., Clayton,S., McRae,J.F., van Kogelenberg,M., King,D.A., Ambridge,K., Barrett,D.M., Bayzetinova,T. *et al.* (2015) Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, **385**, 1305–1314.
17. Smith,B., Ceusters,W., Klagges,B., Köhler,J., Kumar,A., Lomax,J., Mungall,C., Neuhaus,F., Rector,A.L. and Rosse,C. (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
18. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
19. Brush,M., Mungall,C.J., Washington,N.L. and Haendel,M.A. (2013) What's in a Genotype? An ontological characterization for integration of genetic variation data. In: *International Conference on Biomedical Ontology*, Available from: http://ceur-ws.org/Vol-1060/icbo2013_submission_60.pdf.

20. Smith,C.L. and Eppig,J.T. (2015) Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *J. Biomed. Semantics*, **6**, 11.
21. Chibucos,M.C., Mungall,C.J., Balakrishnan,R., Christie,K.R., Huntley,R.P., White,O., Blake,J.A., Lewis,S.E. and Giglio,M. (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO ). *Database (Oxford)*, **2014**, bau075.
22. Mungall,C.J., Washington,N.L., Nguyen-Xuan,J., Condit,C., Smedley,D. and Köhler,S. (2015)Use of model organism and disease databases to support matchmaking for human disease gene discovery. *Hum. Mutat.*, **36**, 979–984.
23. Haendel,M.A., Balhoff,J.P., Bastian,F.B., Blackburn,D.C., Blake,J.A., Bradford,Y., Comte,A., Dahdul,W.M., Dececchi,T.A., Druzinsky,R.E. *et al.* (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J. Biomed. Semantics*, **5**, 21.
24. Mungall,C., Gkoutos,G., Smith,C., Haendel,M., Lewis,S. and Ashburner,M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.* , **11**, R2.
25. Köhler,S., Doelken,S.C., Ruef,B.J., Bauer,S., Washington,N., Westerfield,M., Gkoutos,G., Schofield,P., Smedley,D., Lewis,S.E. *et al.* (2013) Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research.*, **2**, 30.
26. Mungall,C.J., Koehler,S., Robinson,P., Holmes,I. and Haendel,M. (2016) k-BOOM: a Bayesian approach to ontology structure inference, with applications in disease ontology construction. In: *Phenotype Day*, ISMB., Available from: http://phenoday2016.bio-lark.org/pdf/2.pdf.
27. Mungall,C.J., Batchelor,C. and Eilbeck,K. (2011)Evolution of the sequence ontology terms and relationships. *J. Biomed. Inform.*, **44**, 87–93.
28. Bolleman,J.T., Mungall,C.J., Strozzi,F., Baran,J., Dumontier,M., Bonnal,R.J.P., Buels,R., Hoehndorf,R., Fujisawa,T., Katayama,T. *et al.* (2016) FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. *J. Biomed. Semantics*., **7**, 39.
29. McMurry,J., Muilu,J., Dumontier,M., Hermjakob,H., Conte,N., Gormanns,P., Gonzalez-Beltran,A., Gormanns,P., Hastings,J., Haendel,M.A. *et al.* (2015) 10 Simple rules for design, provision, and reuse of identifiers for web-based life science data. *Zenodo*, doi:10.5281/zenodo.18003.
30. Köhler,S., Doelken,S.C., Mungall,C.J., Bauer,S., Firth,H.V., Bailleul-Forestier,I., Black,G.C., Brown,D.L., Brudno,M., Campbell,J. *et al.* (2014) The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.
31. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
32. Xin,J., Mark,A., Afrasiabi,C., Tsueng,G., Juchler,M., Gopal,N., Stupp,G.S., Putman,T.E., Ainscough,B.J., Griffith,O.L. *et al.* (2016) High-performance web services for querying gene and variant annotation. *Genome Biol.*, **17**, 91.
33. Chatr-Aryamontri,A., Breitkreutz,B-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
34. Mi,H., Poudel,S., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*, **44**, D336–D342.
35. Ruzicka,L., Bradford,Y.M., Frazer,K., Howe,D.G., Paddock,H., Ramachandran,S., Singer,A., Toro,S., Van Slyke,C.E., Eagle,A.E. *et al.* (2015) ZFIN, The zebrafish model organism database: Updates and new directions. *Genesis*, **53**, 498–509.
36. Eppig,J.T., Richardson,J.E., Kadin,J.A., Ringwald,M., Blake,J.A. and Bult,C.J. (2015) Mouse Genome Informatics (MGI): reflecting on 25 years. *Mamm. Genome.*, **26**, 272–284.
37. The Gene Ontology Consortium. (2014) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
38. Dumontier,M., Callahan,A., Cruz-Toledo,J., Ansell,P., Emonet,V. and Belleau,F. (2014) Bio2RDF release 3: a larger connected network of linked data for the life sciences. *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*. **1272**, pp. 401–404.
39. Groza,T., Köhler,S., Doelken,S., Collier,N., Oellrich,A., Smedley,D., Couto,F.M., Baynam,G., Zankl,A., Robinson,P.N. *et al.* (2015) Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database (Oxford)*, **2015**, bav005.
40. Groza,T., Köhler,S., Moldenhauer,D., Vasilevsky,N., Baynam,G., Zemojtel,T., Schriml,L.M., Kibbe,W.A., Schofield,P.N., Beck,T. *et al.* (2015) The human phenotype ontology: semantic unification of common and rare disease. *Am. J. Hum. Genet.*, **97**, 111–124.
41. Faculty of Veterinary Science U of S. Online Mendelian Inheritance in Animals. http://omia.angis.org.au.
42. Washington,N., Haendel,M. and Köhler,S. (2013) How good is your phenotyping? Methods for quality assessment. In: *Phenoday2014Bio-LarkOrg*. Available from: http://phenoday2014.bio-lark.org/pdf/6.pdf.
43. Girdea,M., Dumitriu,S., Fiume,M., Bowdin,S., Boycott,K.M., Chénier,S., Chitayat,D., Faghfoury,H., Meyn,M.S., Ray,P.N. *et al.* (2013) PhenoTips: patient phenotyping software for clinical and research use. *Hum. Mutat.*, **34**, 1057–1065.
44. Bostock,M., Ogievetsky,V. and Heer,J. (2011) $D^3$: Data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, **17**, 2301–2309.
45. Smedley,D., Schubach,M., Jacobsen,J.O.B., Köhler,S., Zemojtel,T., Spielmann,M., Jäger,M., Hochheiser,H., Washington,N.L., McMurry,J.A. *et al.* (2016) A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.*, **99**, 595–606.
46. Bello,S.M., Smith,C.L. and Eppig,J.T. (2015) Allele, phenotype and disease data at Mouse Genome Informatics: improving access and analysis. *Mamm. Genome*, **26**, 285–294.
47. Hayamizu,T.F., Baldock,R.A. and Ringwald,M. (2015) Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. *Mamm. Genome*, **26**, 422–430.
48. Hayamizu,T.F., de Coronado,S., Fragoso,G., Sioutos,N., Kadin,J.A. and Ringwald,M. (2012) The mouse-human anatomy ontology mapping project. *Database (Oxford)*, **2012**, bar066.
49. Kalderimis,A., Lyne,R., Butano,D., Contrino,S., Lyne,M., Heimbach,J., Hu,F., Smith,R., Štěpán,R., Sullivan,J. *et al.* (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, **42**, W468–W472.
50. Lyne,R., Sullivan,J., Butano,D., Contrino,S., Heimbach,J., Hu,F., Kalderimis,A., Lyne,M., Smith,R.N., Štěpán,R. *et al.* (2015) Cross-organism analysis using InterMine. *Genesis*, **53**, 547–560.
51. Elsik,C.G., Tayal,A., Diesh,C.M., Unni,D.R., Emery,M.L., Nguyen,H.N., Hagen,D.E. *et al.* (2016) Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Res.*, **44**, D793–D800.
52. Livingston,K.M., Bada,M., Baumgartner,W.A. Jr and Hunter,L.E. (2015) KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*, **16**, 126.
53. Grubb,S.C., Bult,C.J. and Bogue,M.A. (2014) Mouse phenome database. *Nucleic Acids Res.*, **42**, D825–D834.
54. Mulligan,M.K., Mozhui,K., Prins,P. and Williams,R.W. (2016) GeneNetwork – a toolbox for systems genetics. *Syst. Genet. Methods Mol. Biol.*, **9**.
55. Bastian,F., Parmentier,G., Roux,J., Moretti,S., Laudet,V. and Robinson-Rechavi,M. (2008) Bgee: Integrating and comparing heterogeneous transcriptome data among species. In: *Data Integration in the Life Sciences*., p., 124–131.
56. McGary,K.L., Park,T.J., Woods,J.O., Cha,H.J., Wallingford,J.B. and Marcotte,E.M. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6544–6549.
57. Woods,J.O., Singh-Blom,U.M., Laurent,J.M., McGary,K.L. and Marcotte,E.M. (2013) Prediction of gene-phenotype associations in humans, mice, and plants using phenologs. *BMC Bioinformatics.*, **14**, 203.
58. Laraway,B. (2015) *Comparative Analysis of Semantic Similarity and Gene Orthology Tools for Identification of Gene Candidates for Human Diseases*. Oregon Health & Science University. Available from: http://digitalcommons.ohsu.edu/etd/3741.

59. Brush,M., Shefchek,K. and Haendel,M.A. (2016) SEPIO: a semantic model for the integration and analysis of scientific evidence. In: *International Conference on Biomedical Ontology and BioCreative (ICBO BioCreative 2016)*. Corvallis, Oregon. Available from: http://icbo.cgrb.oregonstate.edu/.
60. Rehm,H.L., Berg,J.S., Brooks,L.D., Bustamante,C.D., Evans,J.P., Landrum,M.J., Ledbetter,D.H., Maglott,D.R., Martin,C.L., Nussbaum,R.L. *et al.* (2015) ClinGen–the clinical genome resource. *N. Engl. J. Med.*, **372**, 2235–2242.
61. Bandrowski,A., Brinkman,R., Brochhausen,M., Brush,M.H., Bug,B., Chibucos,M.C., Clancy,K., Courtot,M., Derom,D., Dumontier,M. *et al.* (2016) The Ontology for Biomedical Investigations. *PLoS One*, **11**, e0154556.
62. Diehl,A.D., Meehan,T.F., Bradford,Y.M., Brush,M.H., Dahdul,W.M., Dougall,D.S., He,Y., Osumi-Sutherland,D., Ruttenberg,A., Sarntivijai,S. *et al.* (2016) The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics.*, **7**, 44.
63. Attrill,H., Falls,K., Goodman,J.L., Millburn,G.H., Antonazzo,G., Rey,A.J., Marygold,S.J. and the FlyBase consortium (2016) FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic Acids Res.*, **44**, D786–D792.
64. Howe,K.L., Bolt,B.J., Cain,S., Chan,J., Chen,W.J., Davis,P., Done,J., Down,T., Gao,S., Grove,C. *et al.* (2016) WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res.*, **44**, D774–D780.