D704–D715 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 8 November 2019*
doi: 10.1093/nar/gkz997

The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species

Kent A. Shefchek[1,*], Nomi L. Harris[2], Michael Gargano[3], Nicolas Matentzoglu[4], Deepak Unni[2], Matthew Brush[5], Daniel Keith[1], Tom Conlin[1], Nicole Vasilevsky[5], Xingmin Aaron Zhang[3], James P. Balhoff[6], Larry Babb[7], Susan M. Bello[8], Hannah Blau[3], Yvonne Bradford[9], Seth Carbon[2], Leigh Carmody[3], Lauren E. Chan[10], Valentina Cipriani[11], Alayne Cuzick[12], Maria Della Rocca[13], Nathan Dunn[2], Shahim Essaid[5], Petra Fey[14], Chris Grove[15], Jean-Phillipe Gourdine[5], Ada Hamosh[16], Midori Harris[17], Ingo Helbig[18,19,20,21], Maureen Hoatlin[22], Marcin Joachimiak[2], Simon Jupp[4], Kenneth B. Lett[1], Suzanna E. Lewis[2], Craig McNamara[23], Zoë M. Pendlington[4], Clare Pilgrim[17], Tim Putman[1], Vida Ravanmehr[3], Justin Reese[2], Erin Riggs[24], Sofia Robb[25], Paola Roncaglia[4], James Seager[12], Erik Segerdell[26], Morgan Similuk[27], Andrea L. Storm[13], Courtney Thaxon[28], Anne Thessen[1], Julius O.B. Jacobsen[11], Julie A. McMurry[10], Tudor Groza[23], Sebastian Köhler[29], Damian Smedley[11], Peter N. Robinson[3], Christopher J. Mungall[2], Melissa A. Haendel[1,5], Monica C. Munoz-Torres[1] and David Osumi-Sutherland[4]

[1]Center for Genome Research and Biocomputing, Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR 97331, USA, [2]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94710, USA, [3]The Jackson Laboratory For Genomic Medicine, Farmington, CT 06032, USA, [4]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [5]Oregon Clinical and Translational Research Institute, Oregon Health & Science University, Portland, OR 97239, USA, [6]Renaissance Computing Institute at UNC, Chapel Hill, NC 27517, USA, [7]Broad Institute, Cambridge, MA 02142, USA, [8]The Jackson Laboratory, Bar Harbor, ME 04609, USA, [9]Institute of Neuroscience, University of Oregon, Eugene, OR 97401, USA, [10]College of Public Health and Human Sciences, Oregon State University, Corvallis, OR 97331, USA, [11]William Harvey Research Institute, Barts & The London School of Medicine & Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK, [12]Rothamsted Research, Harpenden AL5 2JQ, UK, [13]Office of Rare Diseases Research (ORDR), National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Bethesda, MD 20892, USA, [14]dictyBase, Center for Genetic Medicine, Northwestern University, Chicago, IL 60611, USA, [15]California Institute of Technology, Pasadena, CA 91125, USA, [16]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD 21205, USA, [17]University of Cambridge, Cambridge CB2 1TN, UK, [18]Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA, [19]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA, [20]Department of Neuropediatrics, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany, [21]Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104, USA., [22]Department of Biochemistry and Molecular Biology, Oregon Health & Science University, Portland, OR 97239, USA, [23]Pryzm Health, 4215 Queensland, Australia, [24]Autism & Developmental Medicine Institute, Geisinger, Danville, PA 17837, USA, [25]Stowers Institute for Medical Research, Kansas City, MO 64110, USA, [26]Xenbase, Cincinnati Children's Hospital, Cincinnati, OH 45229, USA, [27]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA, [28]University of North Carolina Medical School, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516,

[*]To whom correspondence should be addressed. Tel: +1 541 737 5075; Fax: +1 541 737 5077; Email: shefchek@oregonstate.edu

USA and [29]Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

Received September 20, 2019; Revised October 09, 2019; Editorial Decision October 10, 2019; Accepted October 14, 2019

ABSTRACT

In biology and biomedicine, relating phenotypic outcomes with genetic variation and environmental factors remains a challenge: patient phenotypes may not match known diseases, candidate variants may be in genes that haven't been characterized, research organisms may not recapitulate human or veterinary diseases, environmental factors affecting disease outcomes are unknown or undocumented, and many resources must be queried to find potentially significant phenotypic associations. The Monarch Initiative (https://monarchinitiative.org) integrates information on genes, variants, genotypes, phenotypes and diseases in a variety of species, and allows powerful ontology-based search. We develop many widely adopted ontologies that together enable sophisticated computational analysis, mechanistic discovery and diagnostics of Mendelian diseases. Our algorithms and tools are widely used to identify animal models of human disease through phenotypic similarity, for differential diagnostics and to facilitate translational research. Launched in 2015, Monarch has grown with regards to data (new organisms, more sources, better modeling); new API and standards; ontologies (new Mondo unified disease ontology, improvements to ontologies such as HPO and uPheno); user interface (a redesigned website); and community development. Monarch data, algorithms and tools are being used and extended by resources such as GA4GH and NCATS Translator, among others, to aid mechanistic discovery and diagnostics.

INTRODUCTION

The quest to elucidate the genetic basis of disease is hampered by the fragmented landscape of clinical and organismal data. The Monarch Initiative is an open-source resource that has amassed a large collection of genotype–phenotype data: over two million phenotypic associations from dozens of sources covering over 100 species. Monarch provides a bridge between basic and clinical research, developing tools to connect data from multiple sources using ontologies and semantic data integration. Over the past three years, we have introduced and extended numerous integrated ontologies for disease, phenotype, genotype and anatomy, to enable deep semantic integration and cross-species querying. Monarch's data resources, APIs and analysis tools are used both internally and by external groups to bring the power of research organismal data to the clinical domain. Monarch has pioneered the use of research organism data for rare disease diagnosis, and a number of our resources have become global standards.

PRODUCTS OF THE MONARCH INITIATIVE

Ontologies

Monarch develops methods and tools that not only support precision medicine and disease modeling, but also support mechanistic exploration of the relationships between genotype, environment and phenotype across the tree of life. We do this by using ontologies to leverage semantic relationships between biological concepts. Members of the Monarch Initiative and our collaborators have developed key ontological resources such as the Human Phenotype Ontology (HPO); have employed innovative techniques to harmonize phenotype ontologies in the Unified Phenotype Ontology (uPheno); have similarly harmonized disease terminologies in the creation of Mondo; and continue to develop these and other resources that work together to build toward an interoperable semantic landscape. Below, we describe our latest work on ontologies.

The Human Phenotype Ontology (HPO). For proper semantic representation of clinical abnormalities and its associations to diseases we developed the HPO. HPO (https://hpo.jax.org), a flagship of Monarch, is a standardized vocabulary of phenotypic abnormalities associated with over 7000 diseases (1). HPO is the *de facto* standard terminology for clinical 'deep phenotyping' in humans, providing detailed descriptions of clinical abnormalities and computable disease definitions. This ontology enables non-exact matching of sets of phenotypic features (phenotype profiles) against known diseases, other patients and research organisms. The primary labels in the HPO employ medical terminology used by clinicians and researchers. To make the HPO more accessible to patients and non-medical experts, we added layperson synonyms, where appropriate (2). The HPO currently contains 4887 terms with at least one lay synonym, and these are available via the HPO website or in the OWL file with the 'layperson term' subset tag.

Algorithms based on HPO have been implemented in many diagnostic and variant prioritization tools and are used by the UK's 100 000 Genomes Project (3), the US NIH Undiagnosed Diseases Program (4) and Network (5), and thousands of other clinics, labs, tools, and databases worldwide. We have developed strong ties with clinical adopters to continue improving specific areas of the ontology and extend standardized disease descriptions. Initially developed for rare disease phenotyping, HPO also captures many phenotypes for common diseases and can be used as a general resource for patient phenotyping. We continue to explore patient phenotyping from other sources of electronic health records (EHRs) and to promote the adoption of HPO in health care systems (6).

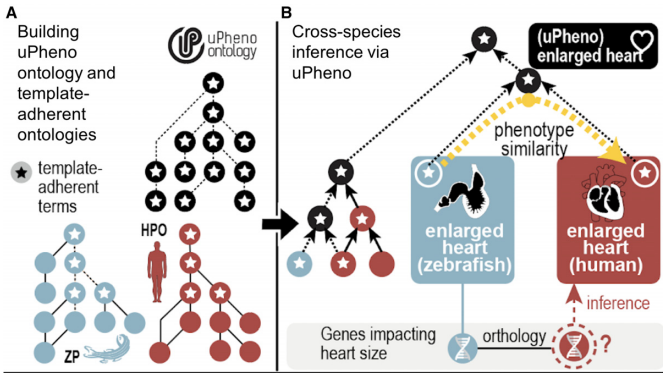

Figure 1. uPheno template-driven ontology development and harmonization. uPheno templates are used to define phenotypes according to agreed upon design patterns. (**A**). Computable definitions specified using uPheno templates are used to automate classification of uPheno and parts of the Zebrafish Phenotype Ontology (ZP (13); dashed lines). (**B**). Computable definitions also drive automated classification of HPO and ZP classes under uPheno classes. For example, enlarged heart in ZP (defined using the zebrafish anatomy heart term) and enlarged heart in HPO are both classified under uPheno enlarged heart (defined using Uberon heart). Algorithms can use this classification under uPheno to predict that human orthologs of zebrafish genes annotated to enlarged heart may cause enlarged heart in humans.

The Unified Phenotype Ontology (uPheno). Different ontologies are used to represent the phenotypes in humans and in each of the major model organism groups. For example, the Mammalian Phenotype (MP) ontology (7) is used by the Mouse Genome Informatics resource (MGI) (8) to annotate mouse phenotypes, HPO by Monarch for human phenotypes and the Drosophila Phenotype Ontology (9) by FlyBase for *Drosophila* Phenotypes. The Zebrafish Information Network (ZFIN) (10) uses a different approach, describing zebrafish phenotypes with terms from reference ontologies such as the Phenotype And Trait Ontology (PATO; http://www.obofoundry.org/ontology/pato) and the Gene Ontology (GO) (11). To enable cross-species querying of phenotypes we have developed an approach for integrating these ontologies based on logical definitions (12), called the Entity-Quality (EQ) approach. In 2013 we implemented this approach in the Unified Phenotype Ontology (uPheno) that integrates organism-specific phenotype ontologies using 'bridging axioms' that connect terms from different ontologies using equivalence or subsumption axioms (13).

The Monarch Initiative uses uPheno (http://obofoundry.org/ontology/upheno) to find candidate genes and potential animal models for human diseases. For this to work well, similar terms both within and between phenotype ontologies need to be described in a logically consistent manner (14). However, while the EQ approach was widely adopted by important model organism communities such ZFIN (10), MGI (8), WormBase (15) and FlyBase (16), the logical definitions themselves were developed in relative isolation, which often caused them to be semantically incompatible. In 2018, we launched a community-wide effort to reconcile and align phenotype ontologies (17) which defined templates based on the Dead Simple Ontology Design Pattern (DOSDP) (18) framework (Figure 1) for the consistent and interoperable representation of phenotypes (for more details on the Reconciliation Effort see section 'Community Engagement'). Based on these patterns, we developed a novel framework (https://github.com/obophenotype/upheno-dev) for the automated construction of uPheno (https://f1000research.com/posters/8-403). We are currently working closely with members from 14 phenotype ontologies and databases to expand the coverage of species-specific phenotypes in uPheno with organisms including *C.elegans*, *Xenopus*, *Dictyostelium discoideum*, *Schizosaccharomyces pombe* and more. This work advances a core mission of the Monarch Initiative: to leverage the wealth of phenotypic knowledge generated by the study of multiple species to understand the genetic nature of human disease.

Mondo: The unified disease ontology. Mondo (http://obofoundry.org/ontology/mondo) is an ontology of diseases and disorders with over 23 000 terms describing a variety of diseases spanning Mendelian, rare, common, complex, infectious and cancer. We created Mondo by integrating available knowledge sources, defining which terms are truly equivalent across different resources. The result thereby enables the integration of associated information, such as treatments, genetics and phenotypes for diagnostics and mechanism discovery. There are many disease terminologies (19), with terms that typically cross-reference each other in ambiguous and conflicting ways. Mondo combines disease information from sources such as OMIM (20), Orphanet (21) and NCIt (National Cancer Institute Thesaurus (22)), in order to leverage the strengths of each resource, including the neoplastic disease classification of NCIt, the rare disease coverage of Orphanet, the Mendelian coverage of OMIM and the common disease coverage of other resources. The Mondo build process uses novel, scalable computational methods to find, untangle and resolve conflicts occurring when disease nomenclatures are merged using cross-references. As a result, Mondo is a logically coherent, merged disease ontology, and it constitutes a scalable solution to the challenge of integrating multiple, partially overlapping and partially conflicting disease terminologies. The Mondo ontology is used in a growing number of bioinformatics resources, including ClinGen (23), the Genetic and Rare Diseases Information Center (GARD, https://rarediseases.info.nih.gov), the European Bioinformatics Institute (EMBL-EBI) as a component of the Experimental Factor Ontology (EFO) (24) and the Kids First Data Resource Portal (https://kidsfirstdrc.org). Initially constructed using semi-automatic methods (25), Mondo is now extensively manually curated. A new release for Mondo is available monthly.

SEPIO. Monarch developed the SEPIO ontology-based data modeling framework for representing evidence and provenance behind scientific assertions (http://obofoundry.org/ontology/sepio). SEPIO stands for Scientific Evidence and Provenance Information Ontology. Initially developed to support harmonized representation of the diverse evidence and provenance information across knowledge sources integrated by Monarch, SEPIO has since been adopted and expanded by external efforts. For ex-

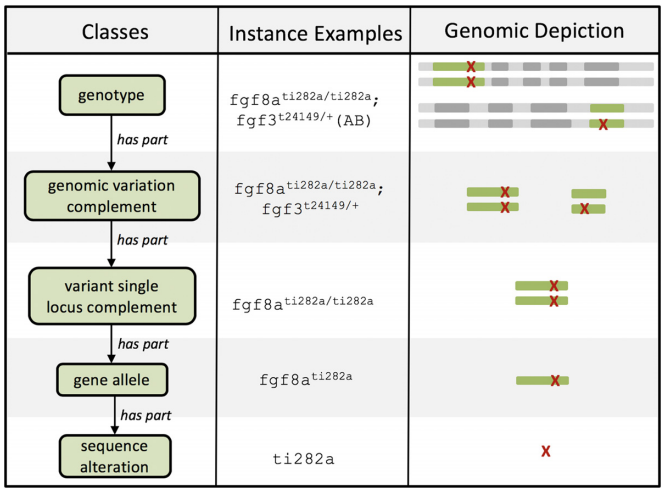

Figure 2. Decomposition of a Zebrafish Genotype. The left panel shows classes in the core genotype partonomy. The center panel shows an example instance of each class from the zebrafish genotype (see also https://zfin.org/ZDB-GENO-161227-1). The right panel shows a graphical depiction of the portion of the genome specified at each level (where the top panel shows a complete genome composed of two sets of homologous chromosomes).

ample, the ClinGen consortium is defining SEPIO-based data models for all five of its curation pipelines, including a Variant Pathogenicity Interpretation data model aligned with the 2015 ACMG Guidelines (https://dataexchange.clinicalgenome.org/interpretation). Several other projects are in the process of defining SEPIO-based data models, including the Variant Interpretation in Cancer Consortium (VICC, https://cancervariants.org), and the GA4GH Genomic Knowledge Standards working group.

GENO. The Genotype Ontology, GENO (http://obofoundry.org/ontology/geno), is an ontology that represents the various components of genotypes, their relationships, and their characteristics. Figure 2 shows classes in the core GENO partonomy (e.g. transitive parts), which decompose a complete genotype into smaller components that reflect the levels at which phenotype annotations are made in different genotype–phenotype resources. For example, ZFIN and MGI annotate full zebrafish and mouse genotypes, respectively, while WormBase annotates gene alleles, and ClinVar (26) annotates individual human sequence alterations (e.g. a single nucleotide variation). The logic encoded in the GENO ontology allows inference of phenotype associations in a way that enables integrated analysis of data across knowledge sources and species. For example, a phenotype annotation made on the zebrafish genotype 'fgf3$^{t24149/+}$(AB)' can be propagated down the partonomy to the 'fgf3^{t24149}' gene allele, allowing direct comparison with data from sources such as WormBase that annotate phenotypes directly on gene alleles.

We have recently evolved GENO to accommodate new use cases from various community partners. For example, we have coordinated with the ClinGen Data Exchange Working Group (http://dataexchange.clinicalgenome.org) to add terms describing copy number variation, allelic phase, allele origin, and allelic state. We are also working with the Alliance of Genome Resources (https://www.alliancegenome.org) to align core GENO terms with genotype-related concepts used in the model organism community. These efforts will facilitate use of GENO to integrate data across a broader set of resources.

The integrated Monarch ontology. Beyond the ontologies developed by the Monarch Initiative, the Monarch platform brings together more than 30 ontologies including GO, the Cell Ontology (CL) (27), Uberon (28) and a number of organism-specific anatomy ontologies, such as FBbt (Drosophila) (https://wiki.flybase.org/mediawiki/images/a/a1/ISB2019_DAO.pdf) and ZFA (zebrafish) (29). To leverage these diverse ontologies in our knowledge graph and in the Monarch app, we have developed the Integrated Monarch Ontology (https://github.com/monarch-initiative/monarch-ontology). One of the key goals of the Monarch Ontology is to ensure interoperability between the integrated source ontologies which often depend on incompatible versions of other ontologies. To that end, we are currently developing a modular strategy based on the Ontology Development Kit (ODK, https://github.com/INCATools/ontology-development-kit). This strategy involves the integration and community wide deployment of so-called 'base' modules—subsets of the ontologies that contain only the native axioms (axioms actually belonging to the ontology) and exclude axioms from their respective ontology dependencies. This ensures that only the latest versions of all ontologies make it into the knowledge graph and stale and possibly incompatible dependencies on old versions are excluded. The Monarch Ontology forms the upper ontological layer of our knowledge and data graph, which will be described in the following sections.

THE MONARCH ARCHITECTURE AND API

Monarch architecture

As previously reported, we developed an ETL (Extract, Transform, Load) pipeline called Dipper (https://github.com/monarch-initiative/dipper), which ingests a wide range of data sources, including genes, mechanisms and context, as well as phenotypic and disease data from disparate sources including research organism and human databases. After ingestion, these data sources are regularized to conform to common association patterns according to the BioLink Model (https://biolink.github.io/biolink-model), are augmented with terms from many ontologies, and are published individually in the Resource Description Framework (RDF) format at https://archive.monarchinitiative.org/latest. These intermediate RDF data files, as well as all referenced ontologies, are unified with SciGraph (https://github.com/SciGraph/SciGraph), an application that wraps a Neo4j graph database and manages the transformation of ontologies and data described using ontologies into a combined graph which is henceforth referred to as the Monarch knowledge graph. This graph database serves as the primary data store for Monarch and its applications. We query and cache results from this graph database, indexing them with an ontology-enhanced Solr instance called GOlr, to provide quick access via our Monarch API, an OpenAPI-compliant data access layer (Figure 3).

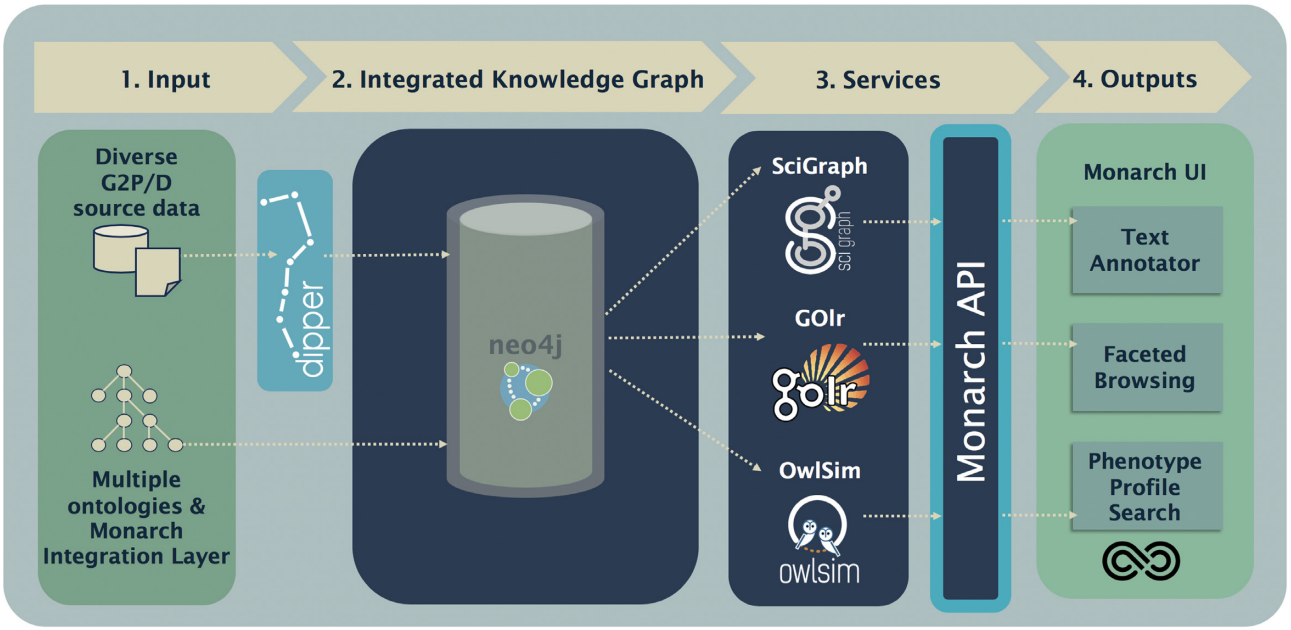

Figure 3. A workflow diagram of the Monarch architecture. Since our last report, we have developed the Monarch API (highlighted) for accessing associations between entities, performing computations on phenotype profiles, executing graph traversal queries, and performing text annotation (https://api.monarchinitiative.org/api).

Data integration

We regularly update our database with the latest gene-to-phenotype data from research organism databases (e.g. MGD (8), ZFIN (10), WormBase (15), FlyBase (16), IMPC (30)), human variants and gene-to-disease data (from OMIM, ClinVar, Orphanet, GWAS Catalog (31)) and other organismal gene-to-phenotype resources (OMIA (32), Animal QTLdb (33)). As well, we ingest other genomic data types, such as GO annotations, gene expression in specific tissues (BgeeDB (34)), protein-to-protein interaction (BioGRID (35)), pathway data (KEGG (36), Reactome (37)), chemical-disease associations (CTD (38)), cell line genotypes-to-disease data (Coriell; https://www.coriell.org), data from the Mouse Phenome Database (39) and from the Mutant Mouse Resource and Research Centers (MMRRC; https://www.mmrrc.org/, Office of the Director grant number OD010921). We recently also added data from the Rat Genome Database (RGD (40)), the Saccharomyces Genome Database (SGD (41)) and data from protein-to-protein interaction networks from STRING (42). The latest release of the Monarch knowledge graph (September 2019, https://archive.monarchinitiative.org/latest) contains over 32.9 million nodes and 160 million edges. In comparison to our previous report, we have 134 244 additional gene-to-phenotype associations. Nearly half (68 640) of the new associations were the result of adding SGD and RGD as new sources of data for Monarch, while 65 604 were added from new data available for mouse, zebrafish, nematode and human combined. Our database now has 26 433 models of disease, a 44% increase since our 2017 report in NAR (43). There are 2 982 400 high-quality protein-protein interactions from STRING from 6 species, and 931 518 from BioGRID. Figure 4 summarizes the sources, their data types, the ontologies used for integration and their delivery within Monarch's knowledge graph.

Gene-to-disease and variant-to-disease data

In order to obtain a fine-grained resolution for genotype–phenotype relationships in humans, Monarch ingests variant-level data and gene-to-disease associations from OMIM, Orphanet, ClinVar and GWAS Catalog. Each source defines specific relationships between genes, variants and diseases depending on different forms of supporting evidence and other factors. Taken together, these sources contain information about 13 different relationships, some directly comparable and some not. Monarch uses these 13 relationships to link genes and variants to diseases, a subset of which is reserved for causal associations between a variation and a disorder, while others are utilized for non-causal associations, such as susceptibility to complex diseases, genome-wide association studies and variants that are likely pathogenic for a condition. The two levels of associations are shown separately in the results tables on our website, and can be queried separately using our API. Our latest data release shows that 3799 protein coding genes are identified as being the main cause for one or more disorders, as opposed to gene variants that contribute to disease susceptibility.

Monarch application programming interface (API)

The Monarch API is a data access layer that sits on top of Monarch's knowledge graph and provides a standard way to access information about entities and association between entities, perform ontology navigation, run semantic similarity queries, perform annotation sufficiency scoring

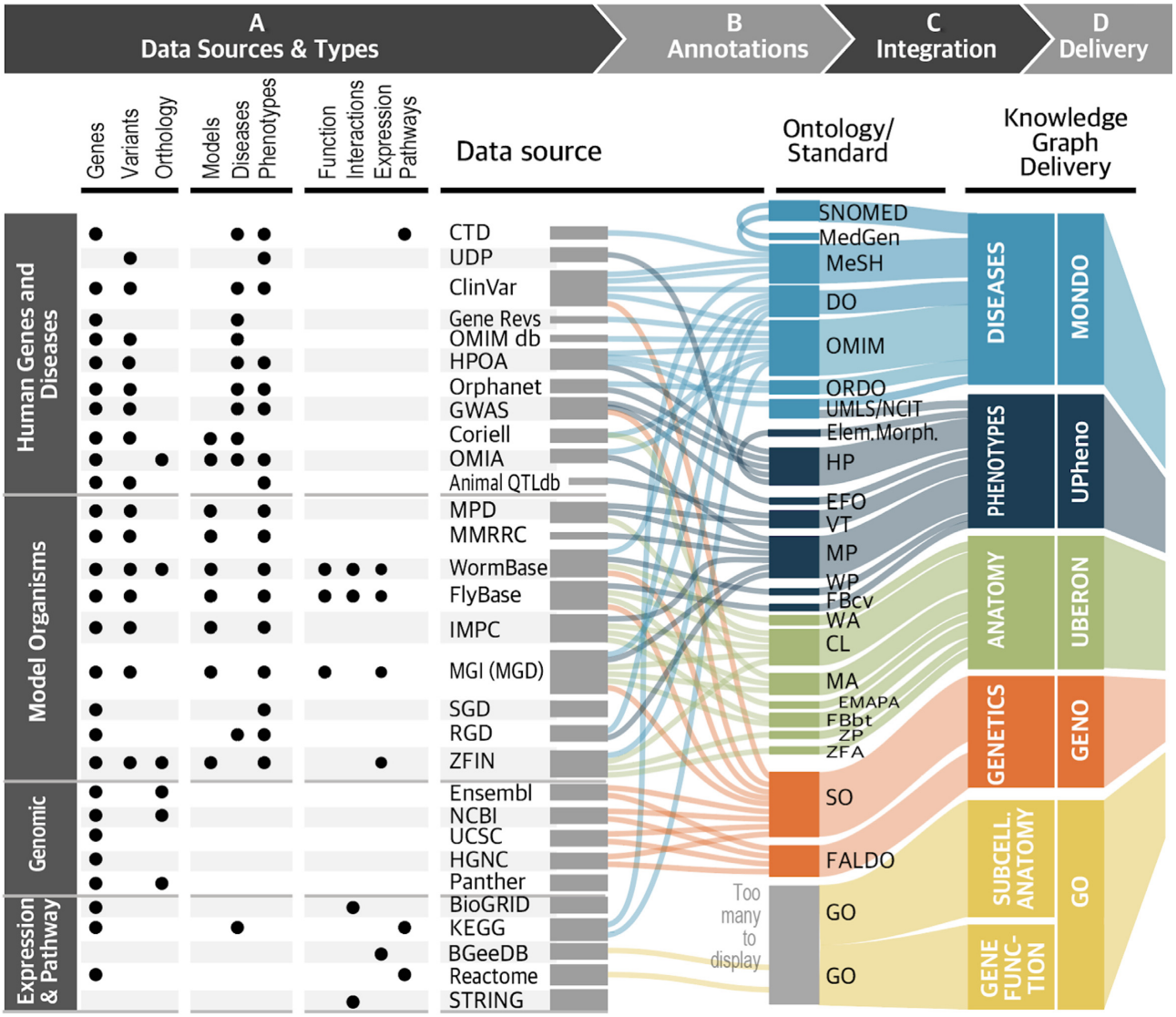

Figure 4. Monarch's data sources. The leftmost set of columns shows the types of data that the integrated data sources serve to Monarch. Note that these sources offer many additional data types that have not yet been integrated into Monarch. Each data source is annotated to specific ontologies and standards, which are, in turn, harmonized using the ontologies indicated in the rightmost panel. Those are used to create an integrated knowledge graph which drives the views and analytics on the Monarch website.

and annotate text with entities via named-entity recognition. Our API has a Swagger-generated documentation interface (https://api.monarchinitiative.org/api/swagger.json) that details all routes available for querying the Monarch database, including required and optional input parameters, JSON schema for the response, and working examples to guide users as they explore each API endpoint. Entity search supports labels, synonyms (e.g. layperson) and definitions. Search has been updated to include genotypes, variants and anatomical entities. All Monarch user interface components are driven by this API (i.e. search, data tables and text annotation). This unified data access layer enables researchers and physicians to explore the Monarch data on our website, and allow us to: (i) run phenotypic comparative analysis making use of semantic similarity software; (ii) annotate text; and (iii) browse data using facets–for example, disease or phenotype categories such as cardiovascular abnormalities (e.g. https://monarchinitiative.org/phenotype/HP:0001626).

In order to simplify programmatic access to Monarch's resources we developed Ontobio, a library written in Python and designed for work with ontologies and ontology associations. It supports a wide range of functionalities such as (i) parsing and using ontologies; (ii) an object model for working with ontologies and their metadata elements; (iii) an API for performing graph operations traversing through an ontology; (iv) ways to access associations from Monarch's knowledge graph; (v) ways to access functional annotations from the Gene Ontology; (vi) tools for performing enrichment analyses with virtually any ontology and associations for that ontology; and (vii) a command-line interface. Ontobio is freely available at https://github.com/biolink/ontobio. As part of this library, we also provide examples of how to use Ontobio for performing the aforementioned function-

alities. Our API makes use of the Ontobio library for communicating with Monarch resources, enabling the Monarch API to remain a lightweight data access layer.

THE NEW MONARCH USER INTERFACE

Monarch will soon be releasing a new, redesigned web interface for Monarch users. The new website presents users with a simplified format for accessing and exploring our data. With the semantic tools available from the Monarch Initiative, researchers, clinicians and the general public can gather, collate and unify disease information across human, research organisms and veterinary species in a single platform. A beta version of the updated user interface (UI) is available at https://beta.monarchinitiative.org. The new UI is a single page application, written in VueJS, that relies on various backend services, primarily our new API (described above), to retrieve and display the data. The Monarch database integrates information from 31 phenotype-related resources (https://beta.monarchinitiative.org/about/data-sources), allowing users to establish connections among biological entities of interest, such as genes, genotypes, gene variants (including SNPs, SNVs, QTLs, CNVs), models (including cell lines, animal strains, species, breeds, as well as targeted mutants), pathways, orthologs, phenotypes and publications. From the new home page, users will be able to explore Monarch using names or identifiers for phenotypes, diseases, genes, publications, variants or models of human disease. All of the information in the Monarch resource is organized using ontologies, rather than free text in isolation. This means that features of the ontology can be used to assist users in search, for example, finding a disease of interest using a synonym, or using the hierarchical organization of a phenotype ontology to group annotations.

The main page for each phenotype, gene, or disease, the 'Overview' tab, offers a summary of all available information in the integrated knowledge graph of the Monarch database, and includes intuitive tools to enable users to navigate through the available data (Figure 5). In addition to finding associated terms from a variety of ontologies (for example, anatomy, function, pathway membership, orthologs, phenotypes) the updated user interface facilitates finding publications in support of all associations displayed for each term. Updated pages documenting the Patient Phenotype Curation Guidelines and the Monarch Phenotype Ontologies Project are also available from the website, as well as details about Exomiser, a Monarch tool used for prioritization of variants and candidate genes from whole exome and whole genome sequencing efforts (described below).

Monarch Annotator

The new site, in addition to offering a range of options to query the Monarch database through searches, includes an updated version of our text annotation tool, which allows users to automatically mark up phenotypes, diseases, anatomical terms, genes and other entities found in text from publications via a UI (Figure 6), or using web services. This functionality was used to mine HPO (1) terms from published case reports in biocuration applications such as PhenoteFX (https://phenotefx.readthedocs.io/en/latest), which helps curators revise or create phenotype annotation records for rare disease and HpoCaseAnnotator, which is used to biocurate pathogenic variants published in scientific literature (https://hpocaseannotator.readthedocs.io).

COMMUNITY ENGAGEMENT

Monarch's overarching goal is to galvanize translational research by developing tools, data services and analysis approaches that help scientists analyze their results in the context of genotype, phenotype and related data from a range of sources. Progress toward this goal is a virtuous loop: by building resources that benefit the community, we also incentivize others to make their tools, ontologies and data more interoperable, thereby making it even easier for a wide range of researchers to leverage Monarch's offerings. Monarch is also collaborating closely with partners such as GA4GH to develop shared standards and resources.

The 'Great phenotype ontology reconciliation effort'

To catalyze phenotype ontology reconciliation, we ran two workshops (Phenotype Ontologies: Traversing All The Organisms, or 'POTATO'). The first installment, which took place at the International Conference on Biological Ontologies (ICBO) in 2018, gathered more than 40 ontology curators, developers and biomedical experts to learn about our tools for pattern-based development and to discuss discrepancies between logical definitions across various phenotype ontologies. As a result of the meeting, representatives of 14 phenotype ontologies and databases covering all major model organisms joined a common Phenotype Ontology Reconciliation Effort (https://github.com/obophenotype/upheno/wiki/Phenotype-Ontologies-Reconciliation-Effort) focused on aligning their respective ontologies by developing and using common sets of design patterns to generate logical definitions (17). The second POTATO workshop (44) was co-located with Biocuration 2019. It focused on developing strategies to deal with the limitations of the Phenotype And Trait Ontology (PATO, http://www.obofoundry.org/ontology/pato), an essential driver of inference in phenotype ontologies. Key accomplishments include the establishment of patterns and workflows for ontological alignment across many organisms; review and implementation of over 100 common design patterns across them, and the elimination of many logical and lexical mistakes (https://douroucouli.wordpress.com/2018/08/06/new-version-of-ontology-development-kit-now-with-docker-support/). Most major phenotype ontologies are now using the Ontology Development Kit (ODK, described in the 'Ontologies' section), which will assist with this standardization and quality control according to OBO Foundry principles (http://www.obofoundry.org/principles/fp-000-summary.html). The template-based generation of phenotype terms now allows *de-novo* generation of born-interoperable, species-specific phenotype ontologies from uPheno templates, reducing the need for manual and error-prone curation of class hierarchies. Examples of ontologies created *de*

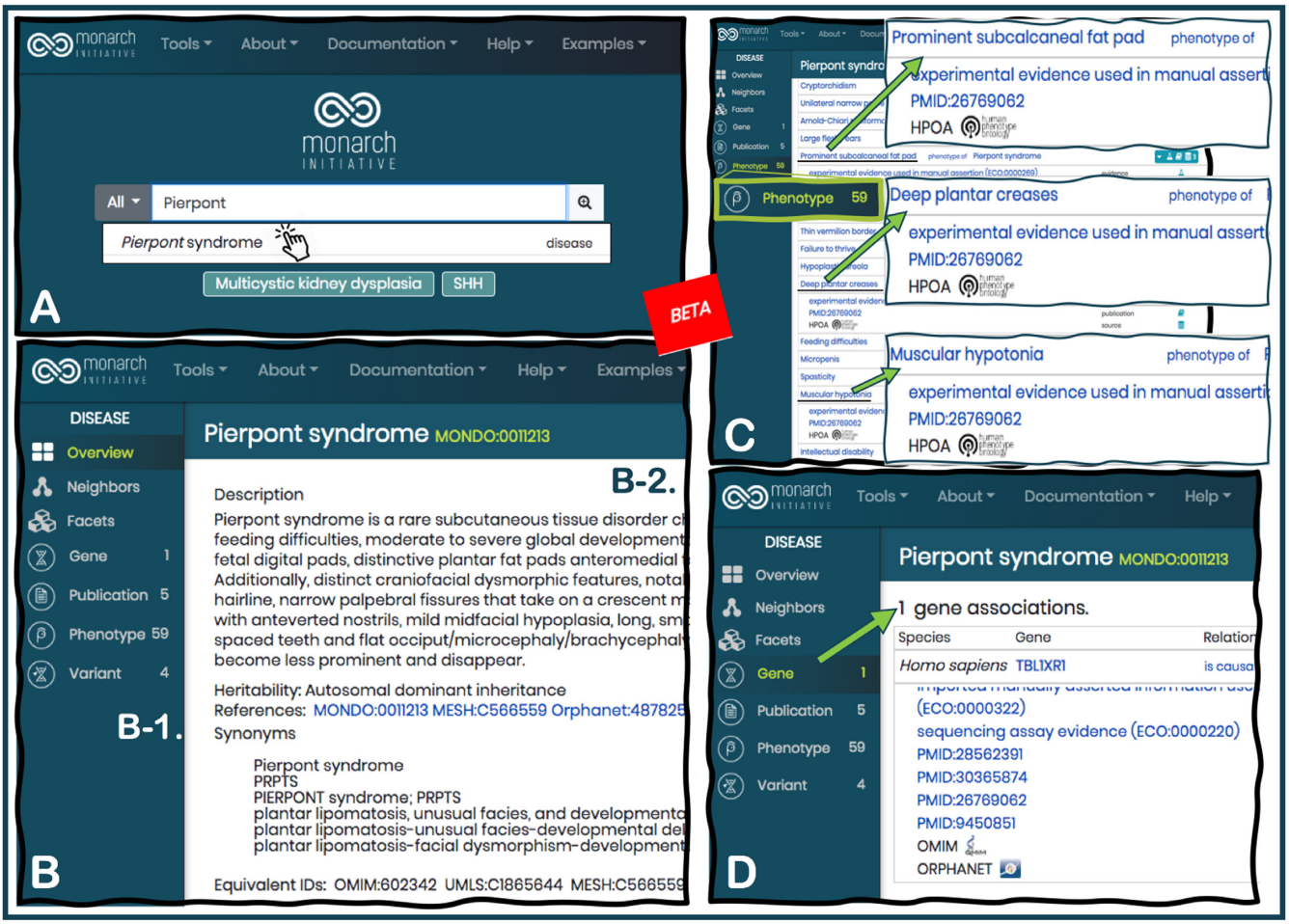

Figure 5. The New Monarch User Interface. A beta version of the new website is available at https://beta.monarchinitiative.org. Entering information on the 'Search' bar, users can navigate directly to terms suggested via autocomplete, or explore more results through the results tables. In this example, a user enters only part of the name of a disease, 'Pierpont syndrome' (**A**). Selecting the term from the auto-complete menu, the user arrives at an overview page, which offers a summary of all available information in the integrated knowledge graph of the Monarch database (**B**). Users can explore all available data using a menu of options shown on a panel on the left (B-1), while the information is updated on the main panel on the right (B-2). In this example, the user learns that Pierpont syndrome, a rare subcutaneous tissue disorder, is characterized by phenotypes that include 'prominent subcalcaneal fat pad' (a term in HPO, with identifier HP:0032276), 'deep plantar creases' (HP:0001869) and 'muscular hypotonia' (HP:0001252), among many others (**C**). Information integrated from the OMIM and Orphanet databases, as well as a number of publications, also support the association of a mutation in one gene, TBL1XR1, as the cause of Pierpont syndrome (**D**).

novo by this approach include the *Xenopus* Phenotype Ontology (45) and the Planarian Phenotype Ontology (http://www.obofoundry.org/ontology/planp.html).

Exomiser

Although whole-exome and genome sequencing have revolutionized rare disease diagnostics, many cases of rare diseases remain unsolved, in part because of the difficulty of prioritizing the hundreds of candidate variants that may remain after removing those identified as common or non-pathogenic. Exomiser (46) is an automated approach developed by the Monarch Initiative to address this problem; it takes patient sequencing data and coded patient phenotypes and analyzes both, informed by a large corpus of gene-to-phenotype associations from humans and model organisms. Exomiser is being used to expedite phenotype-based variant prioritization and improve diagnosis rates in national-level programs such as the rare disease component of the UK 100 000 Genomes Project (https://www.genomicsengland.co.uk), the NIH Undiagnosed Diseases Program (4) and the European Solve-RD project (http://solve-rd.eu), individual hospitals (47) and labs and companies (congenica.com/products-and-services). It has also been used for phenotype profile matching between patients in the Matchbox tool (48) as part of the MatchMaker Exchange (MME) project (https://www.matchmakerexchange.org/) and for automated panel assignment in the Genomics England PanelAssigner tool.

Partnerships within the Global Alliance for Genomics and Health, GA4GH

As an official Driver Project for GA4GH (https://www.ga4gh.org), the Monarch Initiative provides requirements and implementation testbeds, as well as personnel who

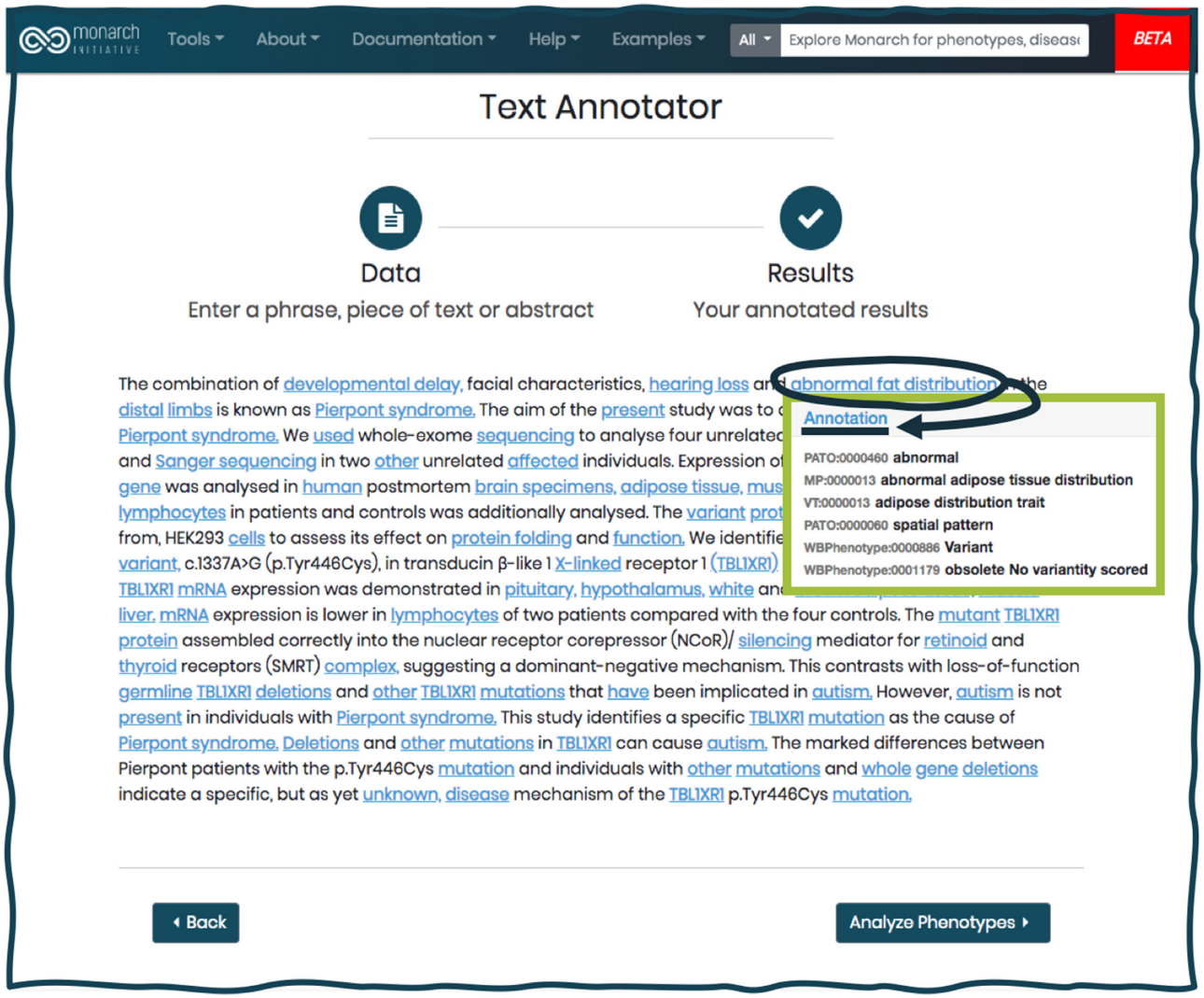

Figure 6. Text annotation widget on the new Monarch website. Users can supply free text and retrieve the resulting marked up text with links to terms in various ontologies. In this example, a user has entered text from a publication entitled 'A specific mutation in TBL1XR1 causes Pierpont syndrome' (51). The Text Annotator tool (in **beta version**) has highlighted terms identified in various ontologies, and hovering over each highlighted term offers details about the marked up annotations, in this case, 'abnormal fat distribution.'

play active leadership roles in both the *Clinical and Phenotypic Data Capture and Exchange* (CP) and the *Genomic Knowledge Standards* (GKS) Work Streams. The GKS work has leveraged several Monarch resources in its emerging standards, including the HPO, Mondo, the genotype ontology GENO and the ontology for evidence and provenance information in science, SEPIO (see 'Ontologies' section above). Because Monarch has been focused on the use of phenotype ontologies and terminologies in clinical and research settings, we have been significant contributors to the new information model for the exchange of clinical and genomic information, Phenopackets (https://github.com/phenopackets/phenopacket-schema). In partnership with HL7 (http://www.hl7.org/) and GA4GH, our workstream is developing a FHIR (http://hl7.org/fhir/) implementation guide for the Phenopackets schema, with the high-level goal of increasing the availability of high-quality standardized phenotypic information for genomic research and genomic medicine across the translational divide.

NCATS Translator

The NCATS Biomedical Data Translator program (https://transltr.io) is building a structured data and machine reasoning ecosystem to address a wide range of questions about human disease posed by researchers, clinicians and patients. The three pillars of the Translator program are (i) data in the form of knowledge graphs, (ii) reusable software modules to perform computational tasks and analysis, and (iii) reasoning systems that can use data relationships and logic to provide possible answers to questions. The Monarch Initiative provides key resources to the Translator project by enabling programmatic queries via the Monarch API and by providing pre-reasoned RDF for ingest into

others' knowledge graphs. The queries retrieve specific curated and structured data on phenotypes, functional annotations and disease variants, as well as corresponding data in model organisms including gene orthologs. The Monarch data serve as the foundation for core Translator computational modules for phenotype comparison and functional similarity, as well as cross-species analytics. These modules have been used in ongoing validations of Translator performance including in a number of data-driven vignettes about chosen exemplar diseases, resulting in the first public outputs from the Translator program.

DISCUSSION

The Monarch Initiative works closely with experts from relevant scientific data providers to ensure that the knowledge graph is useful and the data are correctly represented. This requires significant outreach with several communities of users and data providers, which takes place in the context of face-to-face workshops. Communities engaged in this way include clinicians, curators, toxicologists, exposure scientists, developmental biologists, comparative genomicists, clinical researchers, rare disease researchers and epidemiologists. These workshops provided the opportunity for experts to communicate across domains, develop use cases and provide feedback on data models. In addition to improving the data model and the ontologies, these collaborative workshops have inspired new collaborations and increased general knowledge about bio-ontologies.

The Monarch Initiative has brought together 31 data resources and made a real difference in the lives of patients (49). Monarch has been an important part of projects like GA4GH and the NCATS Translator. Future work will include development of HCLS-Compliant metadata, incorporation of additional datasets, further development of ontologies and extension of the data model to include environmental exposures. While the Monarch Initiative will continue to focus on human disease, data acquisition and model development, it will expand its scope toward a higher diversity of species and domain expert contributions.

DATA AVAILABILITY

The Monarch platform is comprised of multiple components: user interface, data, ontologies, software tools and algorithms. The underlying data are derived from multiple external sources, the use and secondary use of which is governed by the corresponding original license for each source; as we have described in detail, this is not without complex implications (50). It is therefore not possible to provide everything under a unified license. Our integrated data corpus is available for bulk download as RDF formatted files, with subsets of this data as tab-separated value files (https://archive.monarchinitiative.org/latest). Our Neo4j database and Solr index are also publicly available in this archive. Our API provides programmatic access to associations of individual entities. General information about entities, such as definitions, synonyms and cross references, is also available at https://api.monarchinitiative.org/api. A glossary of terms and abbreviations used in this manuscript can be found at https://beta.monarchinitiative.org/glossary.

ACKNOWLEDGEMENTS

We would like to thank the many people who have contributed to the improvements and updates we have reported here, including (but not limited to) Donna Maglott, Anne Pariser and Janine Lewis.

FUNDING

National Institutes of Health (NIH) Office of the Director (OD); The Monarch Initiative [1R24OD011883]; Forums for Integrative Phenomics [1U13CA221044]; Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy [DE-AC02-05CH11231 to S.C., N.L.H., N.D., M.J., S.E.L., C.J.M., J.R., D.U]; EMBL-EBI Core Funds, Open Targets [OTAR005]; European Union's Horizon 2020 Research and Innovation Programme [654248 (CORBEL), 676559 (ELIXIR Excelerate) to S.J., P.R., Z.M.P.]; National Human Genome Research Institute at the US National Institutes of Health [U24 HG002223 to C.G.]; UK Medical Research Council; UK Biotechnology and Biological Sciences Research Council; National Human Genome Research Institute at the US National Institutes of Health [U41HG006627 to A.H.]; Wellcome Trust [104967/Z/14/Z]; National Human Genome Research Institute at the US NIH [U41HG000330 to S.M.B.]; National Human Genome Research Institute (NHGRI) at the US NIH [U41 HG002659 to Y.M.B.]. Funding for open access charge: NIH OD; The Monarch Initiative [1R24OD011883]; Forums for Integrative Phenomics [1U13CA221044].

Conflict of interest statement. None declared.

REFERENCES

1. Köhler,S., Carmody,L., Vasilevsky,N., Jacobsen,J.O.B., Danis,D., Gourdine,J.-P., Gargano,M., Harris,N.L., Matentzoglu,N., McMurry,J.A. *et al.* (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.*, **47**, D1018–D1027.
2. Vasilevsky,N.A., Foster,E.D., Engelstad,M.E., Carmody,L., Might,M., Chambers,C., Dawkins,H.J.S., Lewis,J., Della Rocca,M.G., Snyder,M. *et al.* (2018) Plain-language medical vocabulary for precision diagnosis. *Nat. Genet.*, **50**, 474–476.
3. Turnbull,C., Scott,R.H., Thomas,E., Jones,L., Murugaesu,N., Pretty,F.B., Halai,D., Baple,E., Craig,C., Hamblin,A. *et al.* (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, **361**, k1687.
4. Gall,T., Valkanas,E., Bello,C., Markello,T., Adams,C., Bone,W.P., Brandt,A.J., Brazill,J.M., Carmichael,L., Davids,M. *et al.* (2017) Defining disease, diagnosis, and translational medicine within a homeostatic perturbation paradigm: The national institutes of health undiagnosed diseases program experience. *Front. Med.*, **4**, 62.
5. Ramoni,R.B., Mulvihill,J.J., Adams,D.R., Allard,P., Ashley,E.A., Bernstein,J.A., Gahl,W.A., Hamid,R., Loscalzo,J., McCray,A.T. *et al.* (2017) The undiagnosed diseases network: accelerating discovery about health and disease. *Am. J. Hum. Genet.*, **100**, 185–192.
6. Zhang,X.A., Yates,A., Vasilevsky,N., Gourdine,J.P., Callahan,T.J., Carmody,L.C., Danis,D., Joachimiak,M.P., Ravanmehr,V., Pfaff,E.R. *et al.* (2019) Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *NPJ Digit Med.*, **2**, doi:10.1038/s41746-019-0110-4.
7. Smith,C.L. and Eppig,J.T. (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm. Genome*, **23**, 653–668.

8. Bult,C.J., Blake,J.A., Smith,C.L., Kadin,J.A., Richardson,J.E. and Mouse Genome Database Group (2019) Mouse genome database (MGD) 2019. *Nucleic Acids Res.*, **47**, D801–D806.
9. Osumi-Sutherland,D., Marygold,S.J., Millburn,G.H., McQuilton,P.A., Ponting,L., Stefancsik,R., Falls,K., Brown,N.H. and Gkoutos,G.V. (2013) The Drosophila phenotype ontology. *J. Biomed. Semantics.*, **4**, 30.
10. Van Slyke,C.E., Bradford,Y.M., Howe,D.G., Fashena,D.S., Ramachandran,S., Ruzicka,L. and ZFIN Staff. (2018) Using ZFIN: Data Types, Organization, and Retrieval. *Methods Mol. Biol.*, **1757**, 307–347.
11. The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
12. 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.
13. 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 [version 2; peer review: 3 approved]. *F1000Res.*, **2**, 30.
14. Mungall,C.J., Gkoutos,G.V., Smith,C.L., Haendel,M.A., Lewis,S.E. and Ashburner,M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.*, **11**, R2.
15. Lee,R.Y.N., Howe,K.L., Harris,T.W., Arnaboldi,V., Cain,S., Chan,J., Chen,W.J., Davis,P., Gao,S., Grove,C. *et al.* (2018) WormBase 2017: molting into a new stage. *Nucleic Acids Res.*, **46**, D869–D874.
16. Thurmond,J., Goodman,J.L., Strelets,V.B., Attrill,H., Gramates,L.S., Marygold,S.J., Matthews,B.B., Millburn,G., Antonazzo,G., Trovisco,V. *et al.* (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.
17. Matentzoglu,N., Balhoff,J.P., Bello,S.M., Boerkoel,C.F., Bradford,Y.M., Carmody,L.C., Cooper,L., Grove,C., Harris,N., Köhler,S. *et al.* (2018) Phenotype Ontologies Traversing All The Organisms (POTATO) workshop aims to reconcile logical definitions across species. zenodo doi: https://zenodo.org/record/2382757, 20 December 2018,preprint: not peer reviewed.
18. Osumi-Sutherland,D., Courtot,M., Balhoff,J.P. and Mungall,C. (2017) Dead simple OWL design patterns. *J. Biomed. Semantics*, **8**, 18.
19. Haendel,M.A., McMurry,J.A., Relevo,R., Mungall,C.J., Robinson,P.N. and Chute,C.G. (2018) A census of disease ontologies. *Annu. Rev. Biomed. Data Sci.*, **1**, 305–331.
20. Amberger,J.S., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.*, **47**, D1038–D1043.
21. Pavan,S., Rommel,K., Mateo Marquina,M.E., Höhn,S., Lanneau,V. and Rath,A. (2017) Clinical practice guidelines for rare diseases: The Orphanet Database. *PLoS One*, **12**, e0170365.
22. Sioutos,N., de Coronado,S., Haber,M.W., Hartel,F.W., Shaiu,W.-L. and Wright,L.W. (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.*, **40**, 30–43.
23. 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.
24. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
25. Mungall,C.J., Koehler,S., Robinson,P., Holmes,I. and Haendel,M. (2019) k-BOOM: A Bayesian approach to ontology structure inference, with applications in disease ontology construction. bioRxiv doi: http://dx.doi.org/10.1101/048843, 29 January 2019, preprint: not peer reviewed.
26. Landrum,M.J. and Kattman,B.L. (2018) ClinVar at five years: delivering on the promise. *Hum. Mutat.*, **39**, 1623–1630.
27. 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.
28. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
29. Van Slyke,C.E., Bradford,Y.M., Westerfield,M. and Haendel,M.A. (2014) The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio. *J. Biomed. Semantics.*, **5**, 12.
30. Muñoz-Fuentes,V., Cacheiro,P., Meehan,T.F., Aguilar-Pimentel,J.A., Brown,S.D.M., Flenniken,A.M., Flicek,P., Galli,A., Mashhadi,H.H., Hrabě de Angelis,M. *et al.* (2018) The international mouse phenotyping consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conserv Genet.*, **19**, 995–1005.
31. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
32. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.*, **34**, D599–D601.
33. Hu,Z.-L., Park,C.A. and Reecy,J.M. (2019) Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.*, **47**, D701–D710.
34. Komljenovic,A., Roux,J., Wollbrett,J., Robinson-Rechavi,M. and Bastian,F.B. (2018) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. [version 2; peer review: 2 approved, 1 approved with reservations]. *F1000Res*, **5**, 2748.
35. Oughtred,R., Stark,C., Breitkreutz,B.-J., Rust,J., Boucher,L., Chang,C., Kolas,N., O'Donnell,L., Leung,G., McAdam,R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
36. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
37. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., McKay,S. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
38. Davis,A.P., Grondin,C.J., Johnson,R.J., Sciaky,D., McMorran,R., Wiegers,J., Wiegers,T.C. and Mattingly,C.J. (2019) The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.*, **47**, D948–D954.
39. Bogue,M.A., Grubb,S.C., Walton,D.O., Philip,V.M., Kolishovski,G., Stearns,T., Dunn,M.H., Skelly,D.A., Kadakkuzha,B., TeHennepe,G. *et al.* (2018) Mouse phenome database: an integrative database and analysis suite for curated empirical phenotype data from laboratory mice. *Nucleic Acids Res.*, **46**, D843–D850.
40. Laulederkind,S.J.F., Hayman,G.T., Wang,S.-J., Smith,J.R., Petri,V., Hoffman,M.J., De Pons,J., Tutaj,M.A., Ghiasvand,O., Tutaj,M. *et al.* (2018) A primer for the rat genome database (RGD). *Methods Mol. Biol.*, **1757**, 163–209.
41. Lang,O.W., Nash,R.S., Hellerstedt,S.T., Engel,S.R. and SGD Project. (2018) An introduction to the saccharomyces genome database (SGD). *Methods Mol. Biol.*, **1757**, 21–30.
42. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
43. Mungall,C.J., McMurry,J.A., Köhler,S., Balhoff,J.P., Borromeo,C., Brush,M., Carbon,S., Conlin,T., Dunn,N., Engelstad,M. *et al.* (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
44. Matentzoglu,N., Balhoff,J.P., Bello,S.M., Bradford,Y.M., Carmody,L.C., Cooper,L.D., Grove,A.G., Harris,N.L., Köhler,S., Laporte,M.A. *et al.* (2019) Phenotype Ontologies Traversing All The Organisms (POTATO) workshop. zenodo doi:

http://dx.doi.org/10.5281/zenodo.3352149, 26 July 2019, preprint: not peer reviewed.
45. James-Zorn,C., Ponferrada,V., Fisher,M.E., Burns,K., Fortriede,J., Segerdell,E., Karimi,K., Lotay,V., Wang,D.Z., Chu,S. *et al.* (2018) Navigating xenbase: an integrated xenopus genomics and gene expression database. *Methods Mol. Biol.*, **1757**, 251–305.
46. Smedley,D., Jacobsen,J.O.B., Jäger,M., Köhler,S., Holtgrewe,M., Schubach,M., Siragusa,E., Zemojtel,T., Buske,O.J., Washington,N.L. *et al.* (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc.*, **10**, 2004–2015.
47. Ji,J., Shen,L., Bootwalla,M., Quindipan,C., Tatarinova,T., Maglinte,D.T., Buckley,J., Raca,G., Saitta,S.C., Biegel,J.A. *et al.* (2019) A semiautomated whole-exome sequencing workflow leads to increased diagnostic yield and identification of novel candidate variants. *Cold Spring Harb. Mol. Case Stud.*, **5**, a003756.
48. Arachchi,H., Wojcik,M.H., Weisburd,B., Jacobsen,J.O.B., Valkanas,E., Baxter,S., Byrne,A.B., O'Donnell-Luria,A.H., Haendel,M., Smedley,D. *et al.* (2018) matchbox: an open-source tool for patient matching via the Matchmaker Exchange. *Hum. Mutat.*, **39**, 1827–1834.
49. Zemojtel,T., Köhler,S., Mackenroth,L., Jäger,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.
50. Carbon,S., Champieux,R., McMurry,J.A., Winfree,L., Wyatt,L.R. and Haendel,M.A. (2019) An analysis and metric of reusable data licensing practices for biomedical resources. *PLoS One.*, **14**, e0213090.
51. Heinen,C.A., Jongejan,A., Watson,P.J., Redeker,B., Boelen,A., Boudzovitch-Surovtseva,O., Forzano,F., Hordijk,R., Kelley,R., Olney,A.H. *et al.* (2016) A specific mutation in TBL1XR1 causes Pierpont syndrome. *J. Med. Genet.*, **53**, 330–337.