INFORMATICS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Use of Model Organism and Disease Databases to Support Matchmaking for Human Disease Gene Discovery

Christopher J. Mungall,[1]* Nicole L. Washington,[1] Jeremy Nguyen-Xuan,[1] Christopher Condit,[2] Damian Smedley,[3] Sebastian Köhler,[4] Tudor Groza,[5] Kent Shefchek,[6] Harry Hochheiser,[7] Peter N. Robinson,[4] Suzanna E. Lewis,[1] and Melissa A. Haendel[6]

[1]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California; [2]San Diego Supercomputing Center, UC San Diego, La Jolla, California; [3]Wellcome Trust Sanger Institute, Mouse Informatics group, Hinxton, UK; [4]Charité - Universitätsmedizin Berlin, Institute for Medical and Human Genetics, Berlin, Germany; [5]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, Australia; [6]Department of Biomedical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon; [7]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania

For the Matchmaker Exchange Special Issue

Received 4 May 2015; accepted revised manuscript 22 July 2015.
Published online 13 August 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22857

ABSTRACT: The Matchmaker Exchange application programming interface (API) allows searching a patient's genotypic or phenotypic profiles across clinical sites, for the purposes of cohort discovery and variant disease causal validation. This API can be used not only to search for matching patients, but also to match against public disease and model organism data. This public disease data enable matching known diseases and variant–phenotype associations using phenotype semantic similarity algorithms developed by the Monarch Initiative. The model data can provide additional evidence to aid diagnosis, suggest relevant models for disease mechanism and treatment exploration, and identify collaborators across the translational divide. The Monarch Initiative provides an implementation of this API for searching multiple integrated sources of data that contextualize the knowledge about any given patient or patient family into the greater biomedical knowledge landscape. While this corpus of data can aid diagnosis, it is also the beginning of research to improve understanding of rare human diseases.

Hum Mutat 36:979–984, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: rare disease; informatics; ontology; phenotype; model systems; Matchmaker Exchange

Introduction

The last two decades have witnessed enormous progress in our understanding of the genome due to large-scale projects that have interrogated the sequence and variation of the human genome as well as functional genomics projects. The need to understand the relationships between genomic variation and human disease has brought about a number of large-scale projects such as UK 100,000 Genomes (http://www.genomicsengland.co.uk/the-100000-genomes-project/), NIH Undiagnosed Diseases Program/Network [Tifft and Adams, 2014], The Cancer Genome Atlas [Weinstein et al., 2013], and others. However, despite such efforts, we know the phenotypic consequences of only approximately 38% of the human coding genome, and associated genes have not been identified for approximately half of known heritable diseases [Boycott et al., 2013]. Further, since each of us harbors some 100 genuine loss-of-function variants with around 20 genes completely inactivated [MacArthur et al., 2012], prioritization based solely on variant frequency and pathogenicity cannot reliably identify the causative mutation in all cases—the ability to compute on phenotype as well as sequence is necessary.

Phenotyping in Humans and Model Organisms

There are an ever-increasing number of large-scale projects to catalog phenotypic abnormalities in model organisms, for example, the International Mouse Phenotyping Consortium (IMPC) [Koscielny et al., 2014] and comparable efforts in zebrafish [Kettleborough et al., 2013]. This is particularly relevant for rare and undiagnosed diseases, since phenotype data are available from rat, mouse, zebrafish, worm, and fruitfly via orthology for approximately 80% of the human coding portion of the genome. However, similar to human informatics resources, much of the genomic data in such projects are standardized for exchange and computation but the phenotypic data remain fairly unstructured, are very diverse, and are much less amenable to computation.

The development of resources for the computational analysis of disease has been substantially slower for a number of reasons, mainly including the complexity of computational representations of disease manifestations (phenotype), of disease causation (etiology), and of the development of disease manifestations and complications with time (disease course and natural history). Deep phenotyping (the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described), represents an important prerequisite for the success of the precision medicine endeavor. To maximize the utility of the results, deep phenotyping requires three main components: (1) controlled vocabularies or ontologies to precisely, accurately, and comprehensively describe phenotypic abnormalities in humans and model organisms (see Fig. 1); (2) use of these

*Correspondence to: Christopher J. Mungall, Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA. E-mail: cjmungall@lbl.gov

Contract grant sponsors: NIH (R24OD011883); Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy (contract no. DE-AC02-05CH11231).

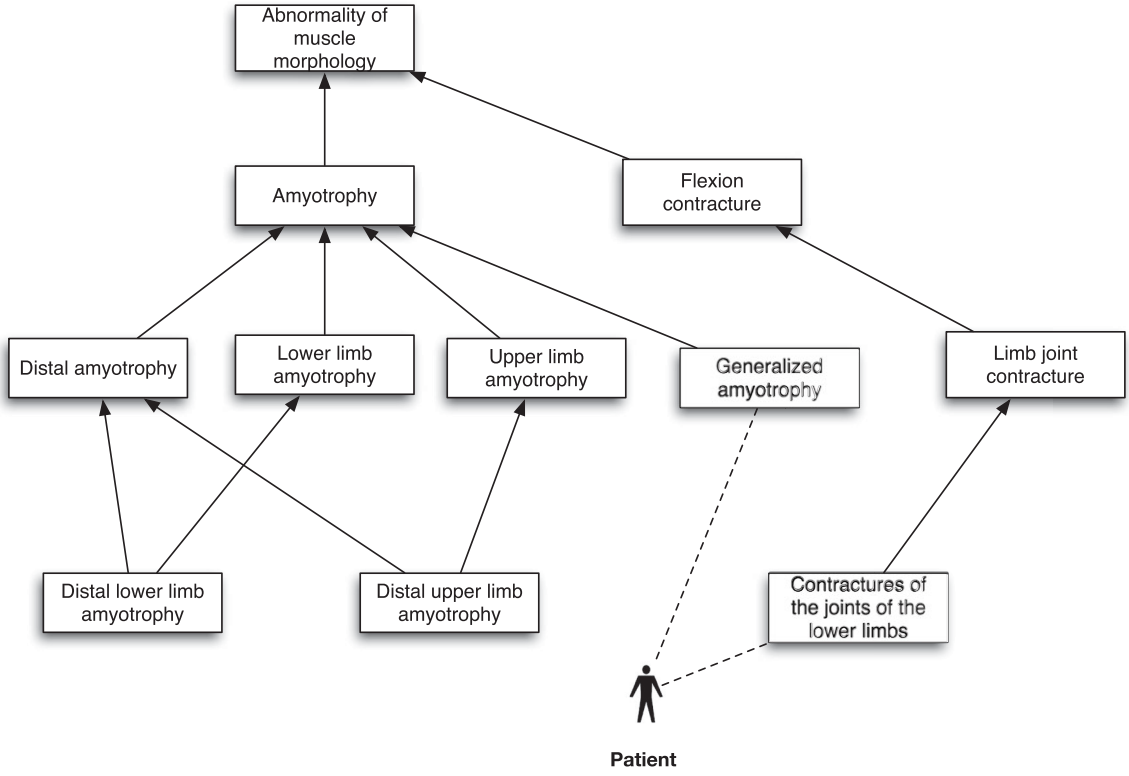

Figure 1. Example of a patient annotation to a subset of the HPO. A hypothetical patient is annotated with two phenotypes, "generalized amyotrophy" and "contractures of the joints of the lower limbs" (annotations are indicated using dashed lines). Phenotypes can be described at different levels of granularity or specificity (more general terms are shown near the top of the figure). Any individual patient can be assigned any number of HPO terms.

controlled vocabularies to describe disease manifestations and thereby provide computational models of human diseases or of medically relevant animal models of disease; and (3) algorithms and tools that represent a foundation for the computational analysis of disease.

The Monarch Initiative is an international consortium that aims to utilize improved deep phenotyping for the purposes of disease diagnostics and mechanism discovery. We provide vocabularies for the description of patients that are integrated with model organism vocabularies, and a system for computing over these descriptions. These vocabularies are intended to meet the demonstrated need for deep phenotyping vocabularies that focus on scientific investigation of the patient rather than billing or quality of care estimates (e.g., the International Classification of Disease ICD). In support of this goal, the human phenotype ontology (HPO) [Robinson et al., 2008; Köhler et al., 2014] not only provides more granular patient phenotypes, but also because of its underlying logic, can be utilized for phenotypic comparison of model systems. We have similarly helped develop the Mammalian Phenotype Ontology [Smith and Eppig, 2015], the zebrafish ontology [Slyke et al., 2014], and many others that have been semantically integrated [Mungall et al., 2010] with HPO to assist in the use of deep phenotyping data from model organisms in combination with human data. Toward this end, we have also constructed a large data corpus (www.monarchinitiative.org) of genotype–phenotype associations from human clinical sources and model organism sources, which have been semantically integrated using the aforementioned phenotype ontologies, together with a suite of genotype, anatomy, and sequence ontologies. The complete set of sources (currently 18) integrated is visible on http://monarchinitiative.org/sources. This corpus is the basis for the semantic similarity algorithms that have been implemented in tools such as Exomiser [Robinson et al., 2013] and PhenIX [Zemojtel et al., 2014] for the purposes of clinical variant prioritization. In these tools, clinical exome sequencing produces a list of candidates that can be further prioritized by utilizing what is known about the phenotypic effects of orthologs genes and interacting proteins in other species. These tools make use of the graph nature of the phenotype ontologies in order to score a match of an overall set of phenotypes [Robinson and Webber, 2014].

Matchmaker Exchange

The Matchmaker Exchange (MME) API (application programming interface) allows for the discovery of patients with shared genetic or phenotypic profiles across different sites [Buske et al., 2015; Phillipakis et al., 2015]. Although it is possible for a Matchmaker to implement only gene matching [e.g., Sobreira et al., 2015], we focus here on phenotype matching. The phenotypic profile of a patient (i.e., the set of individual phenotypes exhibited throughout the course of their disease or disorder) and/or their genotype at one site is matched against profiles at the partner site. The closest matching patients are returned, together with information about which portion of the profile matches. Sites can be paired in a variety of configurations, with a key exchange mechanism and tiered levels of availability of genotype and phenotype data to support privacy.

While the API has been designed and constructed with the specific purpose of performing patient-to-patient matchmaking, in practice, it can be extended to other use cases, such as patient-to-disease and

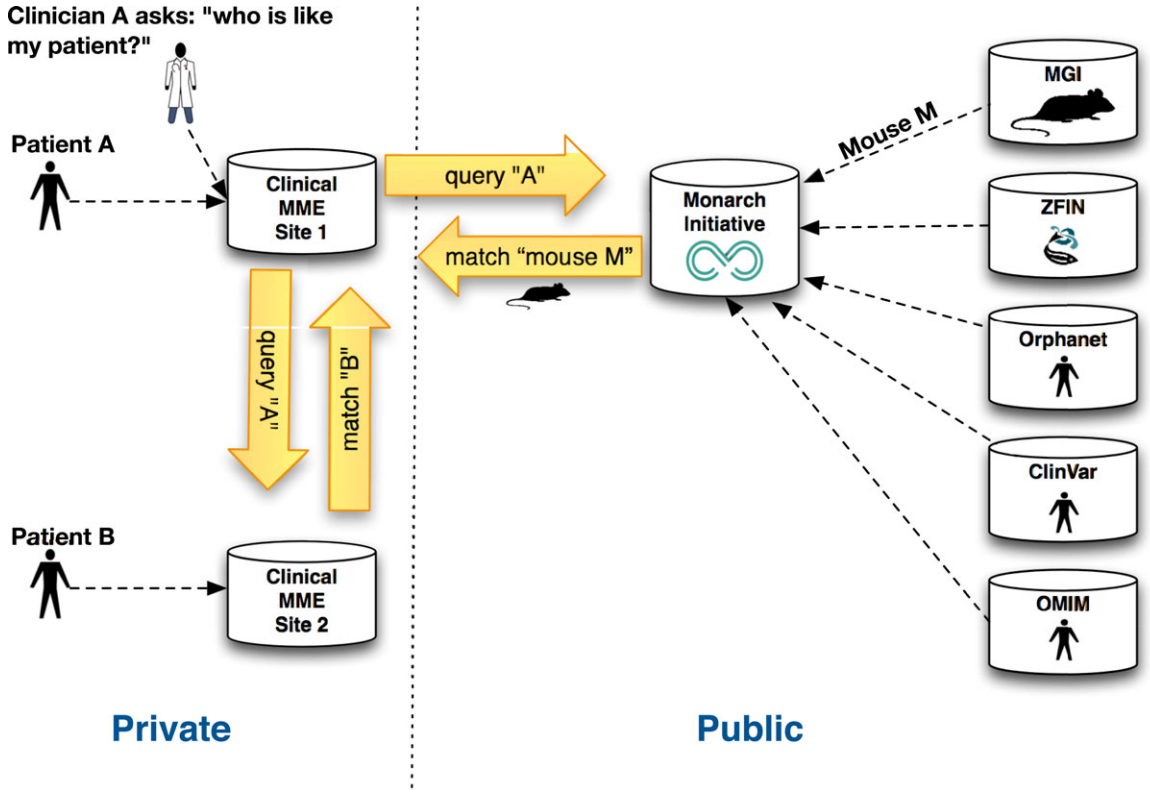

Figure 2. Monarch in the MME landscape. The MME API is used to facilitate discovery of patient B in a remote patient MME database, for a particular patient A that exhibits matching phenotypic features. The same clinical site can connect to Monarch to discover a range of models (e.g., mouse and zebrafish) and other aggregated diseases and variants associated with similar phenotype profiles. Note the diagram only shows a subset of the many phenotypic knowledge sources feeding into the Monarch platform.

patient-to-model matchmaking. This is because, conceptually, the elements being matched all have phenotype profiles, independently of the encapsulating entity—that is, patient data, disease description, or model system. Figure 2 shows the placement and interaction of the Monarch approach within a hypothetical constellation of Matchmakers. Here, we review the Monarch matchmaking service and the implementation of the MME API.

Results

The Monarch Matchmaking System

The Monarch disease model comparison system is based on a large curated and aggregated knowledge base of phenotypic effects of variants covering both humans and model organisms such as mouse, zebrafish, and drosophila. These come from sources, among others, including MGI [Blake et al., 2014], IMPC [Koscielny et al., 2014], ZFIN [Howe et al., 2013], and for human, OMIM [Amberger et al., 2015], ClinVar [Landrum et al., 2014], and Orphanet [Rath et al., 2012]. For a number of sources, we also perform extensive manual curation of gene–disease–phenotype associations [e.g., for human, see Köhler et al., 2014]. We have constructed a data warehousing pipeline that regularly pulls data from these external sources and consolidates them into the integrated Monarch knowledgebase.

In contrast to other Matchmaker efforts, the underlying Monarch repository has to make use of a wider set of vocabularies than HPO or International Consortium for Human Phenotype Terminologies (ICHPT). Many of the terms in these vocabularies are inappropriate for or not complete enough for description of phenotypes in other species. Our approach therefore relies on an ontology-matching strategy. Previously, we have shown how phenotype ontologies from human, mouse, and more distant species such as zebrafish and *C. elegans* can be integrated together [Mungall et al., 2010; Köhler et al., 2013] through the use of multispecies organ system ontologies [Mungall et al., 2012; Haendel et al., 2014]. These exploit functional analogies and evolutionary homologies across species—for example, a human bone marrow phenotype would be matched against a zebrafish "head kidney" (a structure found in teleost fish such as zebrafish that is distinct from the kidney and shares function with the mammalian bone marrow) phenotype based on homology and shared function between these tissues. Conversely, a "head kidney" should not match to a human kidney despite the lexical match. This ontology also allows connection across levels of scale (e.g., between Purkinje cells and the cerebellum) or based on shared developmental origins.

The Monarch interspecies phenotype-matching algorithm has been previously described [Washington et al., 2009; Smedley et al., 2013] and is available as a Java standalone tool or Web services called OwlSim (http://owlsim.org). Figure 3 shows an illustrative example of how the algorithm works. Note that we never expect a model system to completely recapitulate the features of a human disease, and even when individual phenotypes are shared, the different modalities offered by a clinical setting and a laboratory setting means that these phenotypes are frequently observed at different levels of granularity. Our algorithms take account of this, and score according to closeness of an individual's phenotypic features weighted over the whole profile.

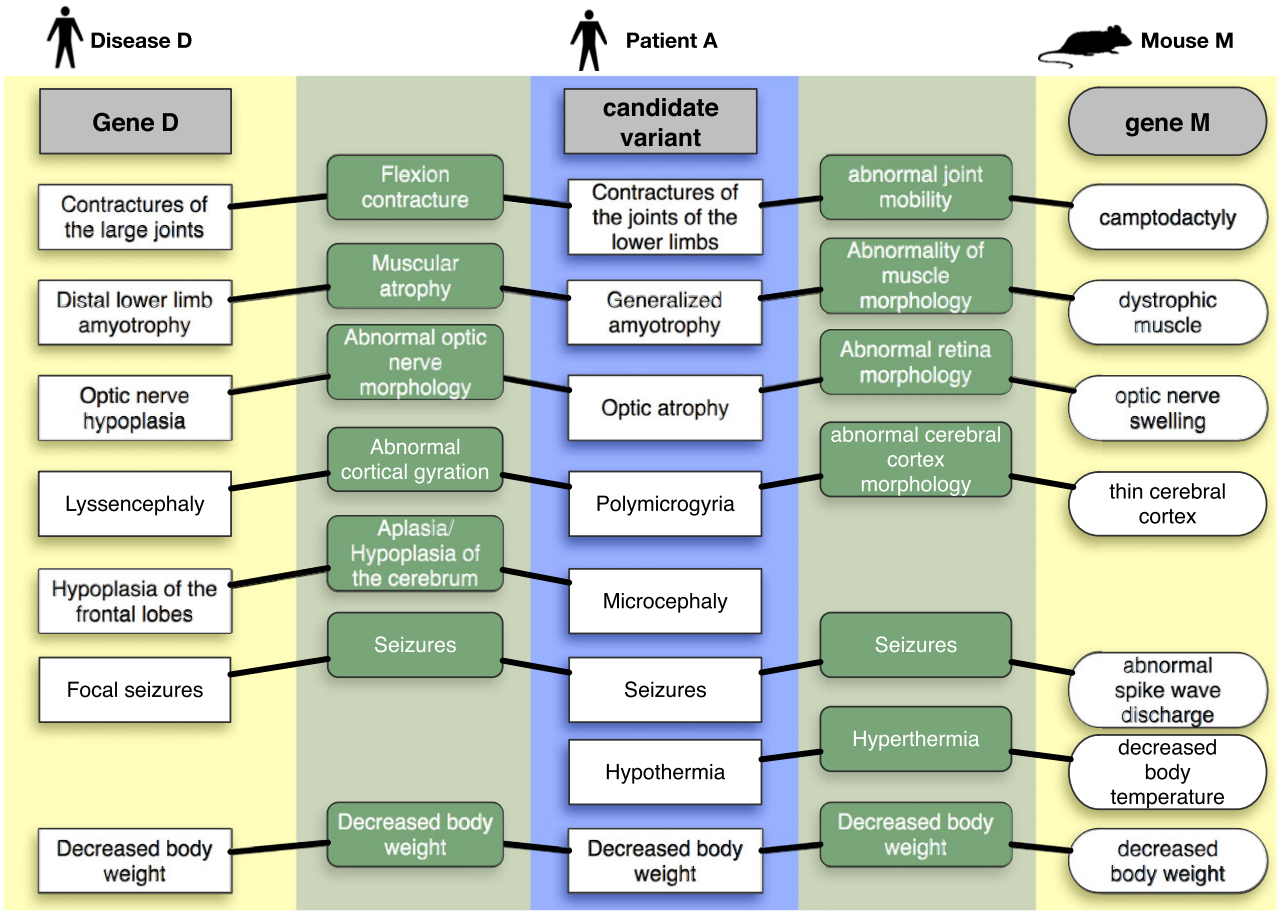

Figure 3. Patient matching against known diseases and model organisms. This figure illustrates matching from a patient to both a human disease and a mouse gene, using synthetic data. The center blue box represents the phenotypic profile of the undiagnosed disease patient encoded using HPO. The left yellow portion of the figure shows the closest matching known disease to this patient, where the disease information comes from public sources annotated with HPO by Monarch. On the right is a matching animal model, described using terms from the mammalian phenotype ontology (MPO). The green portions of the diagram show the commonality between the matched phenotype terms both within species (left) and across species (right). Note that there are missing phenotypes in disease D or model M, another reason why comparison against the largest possible corpus is warranted. Patient A is based on a known undiagnosed disease patient that was solved based on phenotypic similarity to mouse and interactome data.

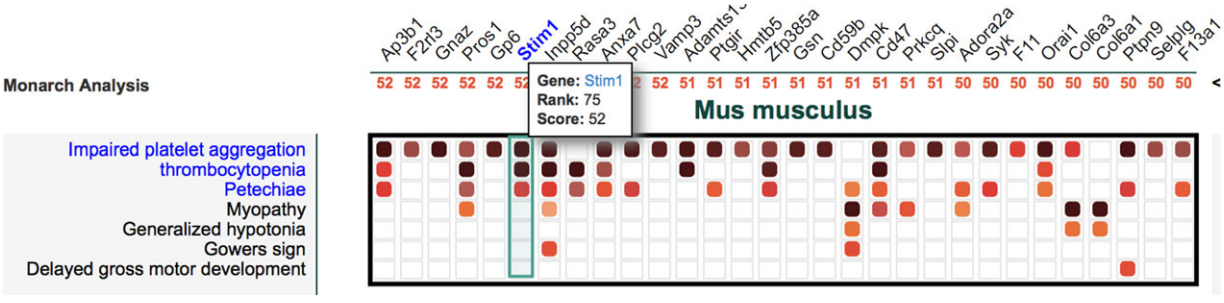

Figure 4. Visualizing patient similarities to known diseases and model organisms. An undiagnosed disease program patient's phenotypes are on the left and match against the best genetic models in mice. The darker the square, the more in common the phenotypes are between the patient and the matched profile. Mouse models are shown here for comparison purposes. Note that a mouse mutant in the ortholog of STIM1 has three matching phenotypes with the patient's profile and when combined with exome analysis assisted the diagnosis of this patient. MME implementations of Phenogrid would enable comparison of input patient profiles against other patients accessible through MME protocols as well as known diseases and models.

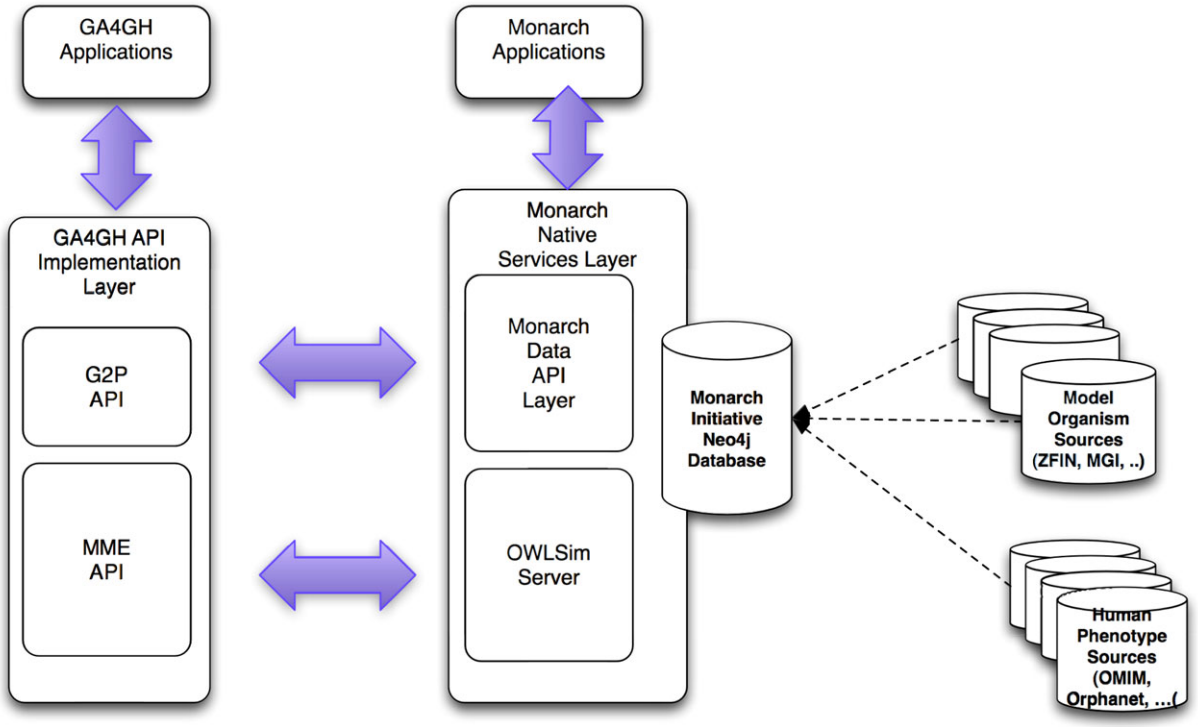

Figure 5. Monarch architecture in the context of some other global alliance APIs, such as the G2P API. The GA4GH APIs are implemented as a layer on top of our own REST services, which are backed by a SciGraph/Neo4J graph database.

Even for moderately complex patients, matchmaking results can be significantly more complex than the example given in Figure 3. A given query with *n* input phenotypes might identify *m* candidate models, each of which having up to *n* phenotypes similar to the input profile. A moderately sized query potentially matching dozens of candidate models might lead to hundreds of phenotype similarities requiring interpretation. In our experience in developing Monarch tools, textual displays of these results have proven insufficient for critical tasks including comparisons of models, examination of individual phenotypes shared across multiple models, and identification of gaps—particularly in terms of phenotypes that are not recapitulated in otherwise similar models. To address these difficulties, we have developed the PhenoGrid visualization tool (Fig. 4), which provides a compact tabular display of the pairwise similarity of phenotypic features between patients, diseases, and/or models. Currently deployed on the Monarch Initiative Website, PhenoGrid is available as a reusable Web widget (http://www.github.com/monarch-initiative/phenogrid) that can be easily adapted and integrated into MME installations for the purposes of comparing patients, known diseases, and model organisms. Figure 4 illustrates a patient profile from the Undiagnosed Disease Program that prioritized a mutation in STIM1 [Markello et al., 2015] based upon phenotypic similarity to a mouse mutant when combined with exome data using the Exomiser tool [Robinson et al., 2013].

Methods

Implementation of the MME API Within Monarch

The implementation of the phenotype matching algorithm used by Monarch is called OwlSim. It is open source and can be installed in a variety of settings. OwlSim has its own APIs, both at the Java and REST levels, but this predates and is different from the MME API. In order to implement the MME API we created a bridge layer on top of our API to expose the services. This bridge is written in Scala using the Play framework (https://www.playframework.com/). A request to this MME implementation will be propagated to the OwlSim API, and the data coming back is transformed to match the MME API specifications before being sent to the user. The API can be queried on https://mme.monarchinitiative.org.

Monarch and the GA4GH Ecosystem

The MME API has been adopted as part of the Global Alliance for Genomes and Health (GA4GH), an international coalition that aims to facilitate data sharing to advance human health. The GA4GH provides an integrated system of APIs (https://github.com/ga4gh/) that can be implemented and consumed by a variety of data and tool providers. The MME API is a part of this constellation of APIs, but has a distinct specification (https://github.com/MatchMakerExchange/mme-apis).

The Monarch consortium are participants in the GA4GH and are helping shape the APIs that address genotype to phenotype data exchange, both within the MME and in the genotype-to-phenotype (G2P) team. Figure 5 shows how Monarch fits into this ecosystem. We provide our own REST APIs that are highly tuned to some of the unique services we provide; all our data are modeled using rich semantic graphs, making use of the SciGraph framework (https://github.com/SciGraph/SciGraph). We implement the GA4GH APIs as an additional layer on top of this, providing a unified means of access to a broader range of tools and applications.

Conclusions

Model systems can help diagnose disease and uncover novel disease–gene associations. The Monarch system provides a means of matching phenotypes between a human and a model organism. Through the standardized schemas of the MME, the Monarch Initiative contributes valuable knowledge for researchers, clinicians, and their IT personnel to integrate into MME-based interfaces, connecting model organism phenotypic resources and expertise with the clinic.

Acknowledgment

Disclosure statements: The authors declare no conflict of interest.

References

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res 43 (Database issue):D789–D798.

Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res 42 (Database issue):D810–D817.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14:681–691.

Buske O, Schiettecatte F, Hutton B, Dumitriu S, Misyura A, Huang L, Hartley T, Girdea M, Sobreira N, Mungall C, Brudno M. 2015. The matchmaker exchange API: Automating patient matching through the exchange of structured phenotypic and genotypic profiles. Hum Mutat 36:922–927.

Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, Hayamizu TF, Ibrahim N, et al. 2014. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J Biomed Semantics 5:21.

Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, et al. 2013. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res 41(Database issue):D854–D860.

Kettleborough RNW, Busch-Nentwich EM, Harvey SA, Dooley CM, de Bruijn E, van Eeden F, Sealy I, White RJ, Herd C, Nijman IJ, Fenyes F, Mehroke S, et al. 2013. A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature 496:494–497.

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, et al. 2014. The Human Phenotype Ontology Project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 42(Database issue):D966–D974.

Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, Robinson PE, Mungall CJ. 2013. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. Version 2. F1000Res 2:30.

Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, Blake A, Chen CK, Easty R, Di Fenza A, Fiegel T, Griffiths M, et al. 2014. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res 42(Database issue):D802–D809.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42 (Database issue):D980–D985.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828.

Markello T, Chen D, Kwan JY, Horkayne-Szakaly I, Morrison A, Simakova O, Maric I, Lozier J, Cullinane AR, Kilo T, Meister L, Pakzad K, et al. 2015. York platelet syndrome is a CRAC channelopathy due to gain-of-function mutations in STIM1. Mol Genet Metab 114:474–482.

Mungall C, Gkoutos G, Smith C, Haendel M, Lewis S, Ashburner M. 2010. Integrating phenotype ontologies across multiple species. Genome Biol 11:R2.

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. Genome Biol 13:R5.

Phillipakis A, Azzariti D, Beltra, S, et al. 2015. The matchmaker exchange: A platform for rare disease gene discovery Hum Mutat 36:915–921.

Rath A, Olry A, Dhombres F, Miličić Brandt M, Urbero B, Ayme S. 2012. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. Human Mutat 33:803–808.

Robinson P, Köhler S, Oellrich A, Wang K, Mungall C, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, Gilissen C, Haendel MA, Smedley D. 2013. Improved exome prioritization of disease genes through cross species phenotype comparison. Genome Res 24:340–348.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610–615.

Robinson PN, Webber C. 2014. Phenotype ontologies and cross-species analysis for translational research. PLoS Genet 10:e1004268.

Smedley D, Oellrich A, Köhler S, Ruef B, Westerfield M, Robinson P, Lewis S, Mungall C. 2013. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database (Oxford) 2013:bat025.

Smith CL, Eppig JT. 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.

Sobreira N, Schiettecatte F, Valle D, Hamosh A. 2015. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat 36:928–930.

Tifft CJ, Adams DR. 2014. The National Institutes of Health undiagnosed diseases program. Curr Opin Pediatr 26:626–533.

Slyke V, Ceri E, Bradford YM, Westerfield M, Haendel MA. 2014. The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio Rerio. J Biomed Semantics 5:12.

Washington NL, Haendel Ma, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. 2009. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology 7:e1000247.

Weinstein JN, Collisson EA, Mills GB, KRM Shaw, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The cancer genome atlas pan-cancer analysis project. Nat Genet 45:1113–1120.

Zemojtel T, Kohler S, Mackenroth L, Jager M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehmke N, Spielmann N, Oien C, Schweiger MR, et al. 2014. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med 6:252ra123.