


Disease insights through cross-species phenotype comparisons

Melissa A. Haendel¹ · Nicole Vasilevsky¹ · Matthew Brush¹ · Harry S. Hochheiser² · Julius Jacobsen³ · Anika Oellrich³ · Christopher J. Mungall⁴ · Nicole Washington⁴ · Sebastian Köhler⁵ · Suzanna E. Lewis⁴ · Peter N. Robinson⁵ · Damian Smedley³ 

Received: 27 March 2015 / Accepted: 20 May 2015 / Published online: 20 June 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract New sequencing technologies have ushered in a new era for diagnosis and discovery of new causative mutations for rare diseases. However, the sheer numbers of candidate variants that require interpretation in an exome or genomic analysis are still a challenging prospect. A powerful approach is the comparison of the patient's set of phenotypes (phenotypic profile) to known phenotypic profiles caused by mutations in orthologous genes associated with these variants. The most abundant source of relevant data for this task is available through the efforts of the Mouse Genome Informatics group and the International Mouse Phenotyping Consortium. In this review, we highlight the challenges in comparing human clinical phenotypes with mouse phenotypes and some of the solutions that have been developed by members of the Monarch Initiative. These tools allow the identification of mouse models for known disease-gene associations that may otherwise have been overlooked as well as candidate genes

may be prioritized for novel associations. The culmination of these efforts is the Exomiser software package that allows clinical researchers to analyse patient exomes in the context of variant frequency and predicted pathogenicity as well the phenotypic similarity of the patient to any given candidate orthologous gene.

Introduction

Despite the many recent successes in identifying causative mutations for human heritable diseases through the use of new sequencing technologies, an associated gene has not been identified for approximately half of the ~7000 diseases (Boycott et al. 2013) with current progress at 150–200 new disease-gene identifications per year (<http://www.irdirc.org>). Discovery of these genotype-to-phenotype relationships is the critical first step towards understanding the mechanism of these heritable diseases and developing potential new treatments.

Although new technologies such as whole exome sequencing (WES) are cost effective and fast, they typically generate thousands of potential candidate variations that need to be interpreted in light of what is known or can be predicted about the variant and the affected gene. One of the most powerful lines of evidence comes from whether the patient's clinical signs and symptoms show similarity to phenotype data previously associated with mutations in the gene.

A wealth of data for this task is available in the Mouse Genome Database (MGD) (Eppig et al. 2015) through the curation efforts of the Mouse Genome Informatics (MGI) group and from the high throughput phenotyping of the International Mouse Phenotyping Consortium (IMPC) (Koscielny et al. 2014). The paper by Meehan et al. in this

✉ Damian Smedley
ds5@sanger.ac.uk

¹ University Library and Department of Medical Informatics and Epidemiology, Oregon Health & Science University, Portland, OR, USA

² Department of Biomedical Informatics and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15206, USA

³ Skarnes Faculty Group, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

⁴ Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

⁵ Computational Biology Group, Institute for Medical Genetics and Human Genetics, Universitätsklinikum Charité, Augustenburger Platz 1, 13353 Berlin, Germany

issue describes how IMPC aims to complete the functional catalogue of all protein-coding genes by 2020, strengthening the existing status of the mouse as the premier model organism for investigating human disease.

The MGI and IMPC website resources are available to clinical researchers to assess individual human disease variant candidates. However, until recently this data have been under-utilized and not used in an automated, systematic approach due to the challenges in comparing human and mouse phenotypes and the lack of tools allowing clinicians and researchers to perform these comparisons (Gkoutos et al. 2012). In this review, we discuss the challenges in comparing phenotypes across species and integration with exome analysis, some of the solutions that have been developed in the context of the Monarch Initiative (www.monarchinitiative.org), and emerging tools for rare disease exome analysis that exploit these comparisons.

Clinical and model organism phenotype data

Data on the ~7000 known genetic and other rare human diseases are stored in the Online Inheritance in Man (OMIM) (Amberger et al. 2015). OMIM contains substantial amounts of descriptive data on the objective signs and subjective symptoms for each disease. However, as this data are represented as free text, it is less amenable to computational analysis, e.g. related diseases cannot easily be discovered using these descriptions. The Human Phenotype Ontology (HP) was developed to describe such phenotypes in a standardized manner that allows such analyses (Köhler et al. 2014a) and there are now over 11,000 terms in HP. The results of an ongoing curation effort by the Monarch Initiative, and members of the rare disease community such as Orphanet (Ayme 2003), are made publicly available from <http://www.human-phenotype-ontology.org> and currently contain annotations for 9019 DECIPHER, OMIM, and Orphanet disorders.

The largest source of mouse phenotype data is the MGD, containing curated annotation of mouse mutants described in literature and also by the import of large-scale projects such as IMPC. Phenotypes are described using the well-established Mammalian Phenotype Ontology (MP) developed precisely for this curation effort. MP currently contains 10,000 terms (Smith and Eppig 2012). MGD contains 278,701 phenotype annotations for over 53,000 different mouse strains involving disruptions in 10,753 genes. The IMPC database contains data for 1470 strains, each with a presumptive null mutation in a unique gene, and 5725 phenotype annotations. The IMPC pipeline involves a sequential set of tests collecting data on parameters covering all major adult organs and most major

disease areas (Koscielny et al. 2014). Given the focussed nature of most published studies, phenotypes that are not assigned to a MGD strain cannot be assumed to be absent. In contrast, for the standardized IMPC pipeline, every assayed phenotype can be assumed to be negative if not reported. However, the pipeline only covers a defined but limited range of phenotypes.

At present some 3400 human genes have HP annotations assigned to them based on their association with disease(s). Mouse mutants involves only a single gene disruption and MP annotation(s) exist for 9974 genes, with only 2341 overlapping with the set of human disease genes. Therefore there is an abundance of genes with genotype–phenotype information available only in the mouse and potentially translatable to human disease studies.

The Monarch Initiative (www.monarchinitiative.org) is an international consortium that aims to integrate data from a large number of diverse resources for human and model organisms (including from IMPC, MGD, OMIM, Orphanet, etc.) describing diseases, phenotypes, environmental factors, drugs, literature, research resources, etc. for the purposes of disease mechanism discovery and diagnosis. The foundation of the Monarch Initiative is the semantic integration of genotype–phenotype data into a single knowledge base that provisions for the application of graph-based computational analyses through the OWLSim software package, including phenotypic profile matching (Washington et al. 2009). Flexible tools for data access and retrieval through APIs and Web widgets suitable for inclusion in third-party sites support the customization and use of this data for diverse purposes.

Cross-species phenotype mapping

The biggest barrier to computational use of the mouse genotype–phenotype associations for human disease research is the use of different phenotype ontologies by the two communities. For example a computer, or even a non-specialist researcher, would not know that the HP term *craniosynostosis* (HP:0001363) is equivalent to the MP term *premature suture closure* (MP:0000081). Mungall et al. 2010 described a process called “logical decomposition” that could be used to define the species-specific phenotype terms using generic, species-agnostic ontologies to computationally define the terms in the species-specific ontologies. Each term is broken down to a combination of a quality (Q), representing what is abnormal about the entity, and an entity (E), representing the anatomical structure or biological process (Köhler et al. 2013; Washington et al. 2009). The entity terms come from well-established ontologies such as the Gene Ontology (GO 2015), the Chemical Entities of Biological Interest [CHEBI; (Hastings

et al. 2013)] ontology, or the UBERON multi-species anatomy ontology (Mungall et al. 2012; Haendel et al. 2014). The Phenotype and Trait Ontology (PATO) is used for the qualities. In the above example, both the HP and MP terms are represented by the *premature closure* (PATO:0002166) of the *suture* (UBERON:0000969) and therefore can be detected as equivalent by an algorithm. In this manner, the logic underlying HP and MP is being co-developed by members of the Monarch Initiative and MGI.

This approach has been applied to human disease, mouse, and zebrafish datasets. Known disease genes were detected with high specificity and sensitivity by semantic phenotype comparisons (Köhler et al. 2013; Washington et al. 2009). The algorithm performs pairwise comparisons between each disease and animal phenotype. Related but non-exact matches can be detected by taking advantage of the hierarchical structure of the ontologies; e.g. a clinical phenotype of *speech articulation problems* and a mouse mutant exhibiting abnormal *larynx morphology* would share a common phenotype of *abnormality of the larynx*. Each match is scored using measures of semantic similarity (Pesquita et al. 2009) such as the Jaccard index or the Information Content of the common phenotype match. The similarity between the disease and animal model is then given by an aggregated score between all the matches, such as the average score across all possible matches or the score of the best pairwise match.

Tools for exploring mouse models of human disease

A number of resources have taken advantage of the cross-species phenotype matching approach to develop websites to generate a ranked list of mouse models for a chosen human disease (Chen et al. 2012; Hoehndorf et al. 2011; Smedley et al. 2013). Here we will describe the features available in some of the various tools developed by members of the Monarch Initiative before describing the Monarch Initiative website itself that integrates data from many other sources and allows users to visualize the phenotypic similarities.

PhenoDigm

PhenoDigm allows users to query for copy number variant (CNV) syndromes from DECIPHER (Bragin et al. 2014) as well as rare diseases from OMIM and Orphanet. Ranked results from mouse and zebrafish phenotype comparisons are displayed along with the information on whether the mutation in the gene is known to be associated with the disease or is located in a critical region for diseases not yet

associated with any gene. Clicking on a gene presents the results from individual animal models associated with that gene so the affect of different alleles, zygosity, and genetic background can be compared to select the optimal model. Many of these mouse models can then be ordered from public repositories for hypothesis-driven mechanistic or therapeutic target validation or purpose-driven therapeutic target effect experiments, e.g. the European Mouse Mutant Archive (Wilkinson et al. 2010). The individual matched phenotypes for each model can also be explored. Figure 1 shows an example where a disease (Craniosynostosis, type 1 OMIM:123100) associated with mutations of *TWIST1* is queried to discover that suitable *Twist1* mouse models of this disease exist and are available in public repositories. These tools can also be used to suggest candidate disease genes for diseases with no known molecular association.

PhenogramViz

The cross-species phenotype comparison approach can also be used to assess the contribution of multiple genes within CNV regions to the disease phenotype (Doelken et al. 2013). Cases can be seen where the whole CNV syndrome can be explained by the disruption of only one of the affected genes, as well as others where different aspects of the syndrome are linked to different genes. PhenogramViz is a Cytoscape plug-in that allows clinicians to explore their own CNV patients by entering the deleted or duplicated region along with patient phenotypes (Köhler et al. 2014b).

International Mouse Phenotyping Consortium

Elsewhere in this issue, Meehan et al. describe the IMPC standardized phenotyping pipeline and portal. The data being generated by the IMPC's controlled and robust statistical analysis framework are likely to be significantly more reproducible than literature-reported findings. Here, the IMPC has also mapped their quantitative assays to the MP, which enables semantic comparison using the PhenoDigm methodology to present high-quality, potential disease models in the IMPC pages. Rather than simple searches for results on individual diseases, faceted, combinatorial searches are allowed using factors such as disease category e.g. *cardiac*, and whether they are associated with known gene associations or with predicted associations from cross-species phenotype comparisons. Figure 2 shows an example where a novel candidate (*ARHGEF11*) is identified for Cone-Rod dystrophy 8 (OMIM:605549) based on phenotype matches to the IMPC model and the location of the gene in a previously identified critical region.

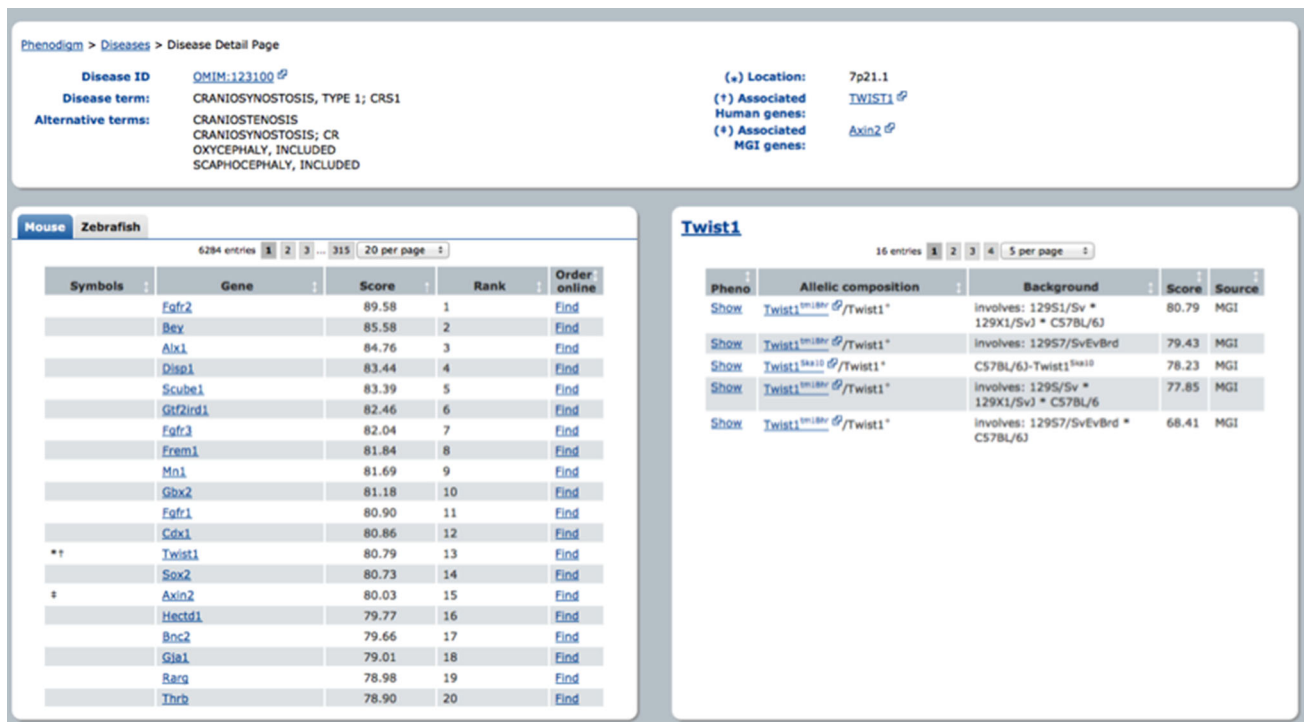


Fig. 1 Cross-species phenotype comparisons using PhenoDigm identify an animal model for Craniosynostosis, type 1. Craniosynostosis, type 1 (OMIM:123100) is already known to be associated with mutations in *TWIST1* (top panel). The bottom left panel reveals that mouse mutants of *Twist1* represent a good phenotypic match to the

clinical signs of this disease. The bottom right-hand panel shows the scores and evidence for different mouse mutants involving *Twist1*, allowing researchers to follow the *Order online* link to obtain the most relevant mouse strain for further mechanistic studies or therapeutic development

The Monarch PhenoGrid

The integrated genotype–phenotype data held within Monarch can be utilized to drive the identification of models for disease research and disease diagnostics (as described above and for Exomiser below). Such integrated data can also be utilized for visualization of the relationships between the different data types. For example, PhenoGrid (Fig. 3), available on the Monarch website, highlights the phenotypic similarity of patient or disease profiles against the most similar mouse models. For software developers, PhenoGrid is available as an open-source widget suitable for integration in third-party websites (www.github.com/monarch-initiative/phenogrid), and is customizable with respect to organism, genotypes versus genes, and user-specified comparisons.

Clinical application to rare disease diagnostics

Many incidences of rare disease remain undiagnosed after exome or genome sequencing due to the sheer number of candidate variants. Even after removing low quality and common variants and those deemed unlikely to be

pathogenic, 10–100's of variants remain. It is already known that each of us harbour ~100 genuine loss of function variants with ~20 genes completely inactivated (MacArthur et al. 2012), so prioritization based solely on variant frequency and pathogenicity is unlikely to identify the causative variant. The additional strategies of studying multiple-affected individuals, linkage data, identity-by-descent inference, de novo heterozygous mutations from trio analysis, or prior knowledge of affected pathways to narrow down to the causative variant are often not possible or successful.

In the last few years, a number of tools have been developed that utilize phenotype data associated with the patient as well as the results of sequencing (Javed et al. 2014; Robinson et al. 2014; Sifrim et al. 2013; Zemojtel et al. 2014). One of these tools, Exomiser, uses an algorithm termed PHenotypic Interpretation of Variants in Exomes (PHIVE) to combine data on the rarity of the variant and its predicted pathogenicity along with the similarity of the patient-to-mouse models for each candidate gene in the exome. A high scoring variant will be: (i) rarely or never observed in the 1000 Genomes Project and Exome Variant Server datasets, (ii) predicted to be highly pathogenic by PolyPhen, SIFT, and/or

Disease: Cone-Rod Dystrophy 8

Name	Cone-Rod Dystrophy 8
Synonyms	-
Locus	1q12-q24
Associated Human Genes	-
Mouse Orthologs	-
Source	OMIM:605549

Mouse Models associated by gene orthology ?

No mouse models associated with OMIM:605549 by orthology to a human gene.

Potential Mouse Models predicted by phenotypic similarity ?

Show entries
Search:

Mouse Gene Symbol	In Disease Locus	MGI Phenotype Similarity Score	IMPC Phenotype Similarity Score
Arhgef11	Yes		55.53

OMIM:605549 Disease Phenotype Terms

- Abnormality of the optic nerve
- Macular degeneration
- Photophobia
- Night blindness
- Abnormal retinal pigmentation
- Peripheral visual field loss
- Retinal arteriolar constriction
- Increased lacrimation

Associated Mouse Models (PhenoDigm predicted)

55.53: [Arhgef11^{tm1b\(KOMP\)Wtsi}](#)/[Arhgef11^{tm1b\(KOMP\)Wtsi}](#) C57BL/6NTac (Source: IMPC)

- abnormal retina morphology
- abnormal retinal pigmentation
- improved glucose tolerance
- impaired glucose tolerance

Fig. 2 Identification of a novel candidate for Cone-Rod dystrophy 8 using cross-species phenotype comparisons at the IMPC portal. A high scoring phenotype match for OMIM:605549 is obtained for an IMPC mouse strain involving disruption of the mouse *Arhgef11* gene

where abnormalities of the retina are reported in both the disease and the model. In addition, the tool highlights the human orthologue that lies within the previously reported locus at 1q12-24

MutationTaster, and (iii) be located in a gene with a mouse model that exhibits very similar phenotypes to the patient.

For the phenotype comparisons, PHIVE uses the same OWLSim methodology used in the tools above and mouse phenotype data from MGI and IMPC. Benchmarking was performed on 100,000 simulated disease exomes containing known disease variants from HGMD added to unaffected exomes from the 1000 Genomes Project. The variant-based scores (frequency and pathogenicity) were found to combine synergistically with the phenotype scores to optimize the identification of the known causative variant as the top hit. The correct gene was recalled as the top hit in up to 83 % of samples and performance was improved by up to 54 fold by including phenotype information.

Although 88 % of the disease genes assessed had mouse strains with mutation in the orthologous gene, there were obviously some tested exomes where mouse phenotype data were missing and therefore performance will be expected to improve as the IMPC nears its goal of complete coverage of the genome. In the mean time, coverage has been increased by including human and zebrafish phenotypes as well as a guilt-by-association approach using protein-protein associations for those genes that have no data in any of the species. This modified algorithm (hiP-HIVE) was able to detect the known disease-gene associations as the top hit in 97 % of the benchmarking exomes. In a strategy where the known human disease-gene phenotypes were masked, representing discovery of a novel

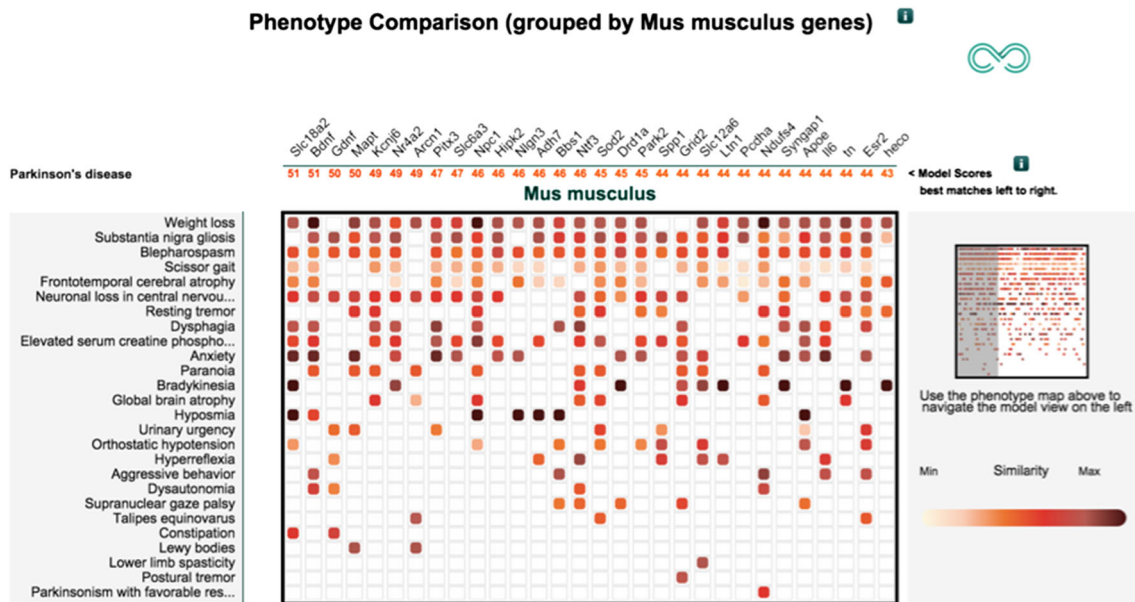


Fig. 3 Monarch PhenoGrid showing a phenotypic comparison of Parkinson's disease with the most phenotypically similar mouse models. Matching phenotypes are displayed in *rows*, matching models in *columns* (indicated here by the gene that is mutated), and *cell* contents *colour* coded with greater saturation indicating greater

similarity. 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. This example can be seen in the Compare tab at <http://monarchinitiative.org/disease/DOID:14330> (Color figure online)

association, the correct variant was detected as the top hit in 87 % of the benchmarking exomes. This version of Exomiser is being used by a number of groups as part of their analysis pipeline, such as the NIH Undiagnosed Disease Program (Gahl et al. 2012). The downloadable, command-line version of Exomiser requires no additional installation steps and is easily integrated into any bioinformatic pipeline.

Conclusions

In this review we have highlighted the latest achievements in the computational analysis of mutations in mouse genes, mouse phenotypes, and mouse genotype–phenotype associations for novel insights into human disease. That any of this has been possible is testament to the remarkable ability of mouse models to recapitulate disease phenotypes, and the advances made in using ontologies to annotate and query disease and model organism data.

Improvements to the ontologies and algorithms are needed in particular disease areas (Oellrich et al. 2014; Robinson and Webber 2014). Beyond these technical challenges, a cultural shift is still needed to encourage collection of higher-quality phenotype data. For efficient and accurate diagnosis of rare disease patients, detailed and comprehensive clinical phenotypes need to be collected to be used alongside the new sequencing technologies in analysis (see <http://monarch->

[initiative.blogspot.com/2015/01/how-to-annotate-patients-phenotypic.html](http://monarchinitiative.blogspot.com/2015/01/how-to-annotate-patients-phenotypic.html) for further detail). Use of tools such as PhenoTips (Girdea et al. 2013) can greatly facilitate informative patient phenotyping. On the mouse side, although IMPC will collect and annotate phenotype data on all protein-coding genes, the additional published phenotypes on these and other strains of mice will be vital for the successful interpretation of human genotype and phenotype data.

The role MGI plays in collecting these extra annotations will still be critical but the development of journal data submission rules for phenotypes would also be a welcome improvement. For example, if authors were required to describe all negative phenotypes (phenotypes measured but found to show no significant difference from wild type) then this highly relevant data could be incorporated into the phenotype matching algorithms. The Monarch Initiative is developing an online phenotyping tool to facilitate easy capture of phenotype data for any model organism and validate the genotypes with the correct nomenclature authorities. This will be critical to ensure publication of sufficient information to adequately link the phenotypic consequences of mutation to the specific genotype (Vasilevsky et al. 2013). The tool will also indicate whether or not the phenotypic profiles of the models are sufficient for comparison against all other known models of disease.

Assuming these challenges continue to be addressed, and with the completion of the IMPC's dataset on functional consequences of mutation in all genes and the further

development of these computational approaches, the next few years promise to be an exciting era for furthering our understanding of human disease by comparison analysis with mouse models.

Acknowledgments This work was supported by core infrastructure funding from the Wellcome Trust and National Institutes of Health (NIH) Grant [1 U54 HG006370-01] and NIH Office of the Director Grant #5R24OD011883. We are grateful to Cynthia Smith of MGI for her help in developing logical definitions for MP.

Conflict of interest The author(s) have no conflict of interest to declare.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 43:D789–D798
- Ayme S (2003) Orphanet, an information site on rare diseases. *Soins: la revue de reference infirmiere* 672:46–47
- 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
- Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ (2014) DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res* 42:D993–D1000
- Chen CK, Mungall CJ, Gkoutos GV, Doelken SC, Köhler S, Ruef BJ, Smith C, Westerfield M, Robinson PN, Lewis SE, Schofield PN, Smedley D (2012) MouseFinder: candidate disease genes from mouse phenotype data. *Hum Mutat* 33:858–866
- Doelken SC, Köhler S, Mungall CJ, Gkoutos GV, Ruef BJ, Smith C, Smedley D, Bauer S, Klopocki E, Schofield PN, Westerfield M, Robinson PN, Lewis SE (2013) Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Dis Model Mech* 6:358–372
- Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res* 43:D726–D736
- Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, Carlson-Donohoe H, Gropman A, Pierson TM, Golas G, Wolfe L, Groden C, Godfrey R, Nehrebecky M, Wahl C, Landis DM, Yang S, Madeo A, Mullikin JC, Boerkoel CF, Tift CJ, Adams D (2012) The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med* 14:51–59
- Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chenier S, Chitayat D, Faghfoury H, Meyn MS, Ray PN, So J, Stavropoulos DJ, Brudno M (2013) PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat* 34:1057–1065
- Gkoutos GV, Schofield PN, Hoehndorf R (2012) Computational tools for comparative phenomics: the role and promise of ontologies. *Mamm Genome* 23:669–679
- GO Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res* 43:D1049–D1056
- Haendel M, Balhoff J, Bastian F, Blackburn D, Blake J, Bradford Y, Comte A, Dahdul W, Dececchi T, Druzinsky R, Hayamizu T, Ibrahim N, Lewis S, Mabee P, Niknejad A, Robinson-Rechavi M, Sereno P, Mungall C (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J Biomed Semantics* 5:21
- Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 41:D456–D463
- Hoehndorf R, Schofield PN, Gkoutos GV (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res* 39:e119
- Javed A, Agrawal S, Ng PC (2014) Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods* 11:935–937
- Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, Robinson PN, Mungall CJ (2013) Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Res* 2:30
- 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, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jahn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park SM, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AO, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BB, Washington NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN (2014a) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 42:D966–D974
- Köhler S, Schoeneberg U, Czeschik JC, Doelken SC, Hehir-Kwa JY, Ibn-Salem J, Mungall CJ, Smedley D, Haendel MA, Robinson PN (2014b) Clinical interpretation of CNVs with cross-species phenotype data. *J Med Genet* 51:766–772
- 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, Horne A, Karp NA, Kurbatova N, Mason JC, Matthews P, Oakley DJ, Qazi A, Regnart J, Retha A, Santos LA, Szeddon DJ, Warren J, Westerberg H, Wilson RJ, Melvin DG, Smedley D, Brown SD, Flicek P, Skames WC, Mallon AM, Parkinson H (2014) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res* 42:D802–D809
- MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335:823–828
- Mungall CJ, Gkoutos GV, Smith CL, Haendel MA, Lewis SE, 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

- Oellrich A, Koehler S, Washington N, Mungall C, Lewis S, Haendel M, Robinson PN, Smedley D (2014) The influence of disease categories on gene candidate predictions from model organism phenotypes. *J Biomed Semantics* 5:S4
- Pesquita C, Faria D, Falcao AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 5:e1000443
- Robinson PN, Webber C (2014) Phenotype ontologies and cross-species analysis for translational research. *PLoS Genet* 10:e1004268
- Robinson PN, Köhler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, Gilissen C, Haendel M, Smedley D (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 24:340–348
- Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y (2013) eXtasy: variant prioritization by genomic data fusion. *Nat Methods* 10:1083–1084
- Smedley D, Oellrich A, Köhler S, Ruef B, Sanger Mouse Genetics P, Westerfield M, Robinson P, Lewis S, Mungall C (2013) PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database* 2013:bat025
- Smith CL, Eppig JT (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome* 23:653–668
- Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, Haendel MA (2013) On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* 1:e148
- 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 Biol* 7:e1000247
- Wilkinson P, Sengerova J, Matteoni R, Chen CK, Soulat G, Ureta-Vidal A, Fessele S, Hagn M, Massimi M, Pickford K, Butler RH, Marschall S, Mallon AM, Pickard A, Raspa M, Scavizzi F, Fray M, Larrigaldie V, Leyritz J, Birney E, Tocchini-Valentini GP, Brown S, Herault Y, Montoliu L, de Angelis MH, Smedley D (2010) EMMA—mouse mutant resources for the international scientific community. *Nucleic Acids Res* 38:D570–D576
- Zemojtel T, Köhler S, Mackenroth L, Jager M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehmke N, Spielmann M, Oien NC, Schweiger MR, Kruger U, Frommer G, Fischer B, Kornak U, Flottmann R, Ardeshirdavani A, Moreau Y, Lewis SE, Haendel M, Smedley D, Horn D, Mundlos S, Robinson PN (2014) Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* 6:252ra123