Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium

Terrence F Meehan1,27, Nathalie Conte1,27, David B West2,27, Julius O Jacobsen3, Jeremy Mason1,4, Jonathan Warren1, Chao-Kung Chen4, Ilincu Tudose1, Mike Relac1, Peter Matthews1, Natasha Karp4, Luis Santos5, Tanja Fiegel5, Natalie Ring5, Henrik Westerberg5, Simon Greenaway5, Duncan Sneddon5, Hugh Morgan5, Gemma F Codner5, Michelle E Stewart5, James Brown5, Neil Horner5, Luis Santos5

The International Mouse Phenotyping Consortium

Although next-generation sequencing has revolutionized the ability to associate variants with human diseases, diagnostic rates and development of new therapies are still limited by a lack of knowledge of the functions and pathobiological mechanisms of most genes. To address this challenge, the International Mouse Phenotyping Consortium is creating a genome- and phenome-wide catalog of gene function by characterizing new knockout-mouse strains across diverse biological systems through a broad set of standardized phenotyping tests. All mice will be readily available to the biomedical community. Analyzing the first 3,328 genes identified models for 360 diseases, including the first models, to our knowledge, for type C Bernard–Soulier, Bardet–Biedl-5 and Gordon Holmes syndromes. 90% of our phenotype annotations were novel, providing functional evidence for 1,092 genes and identified models for 360 diseases, including the first models, to our knowledge, for type C Bernard–Soulier, Bardet–Biedl-5 and Gordon Holmes syndromes. Finally, we describe our role in variant functional validation with The 100,000 Genomes Project and others.

With its extensive toolkit for genome modification and its capacity for recapitulating human disease, the laboratory mouse is arguably the preferred model organism for studying and validating the effects of genetic variants in mendelian disease, as well as identifying previously unsuspected disease-associated genes. Null mouse mutations have been generated and described in the literature for approximately one-half of the genes in the genome. However, hypothesis-driven phenotyping of these mutants has led to discoveries in areas that largely reflect the expertise and specific research questions of individual investigators. As a result, the extent of functional annotation, the potential to fully uncover pleiotropy and the opportunity to exploit mutant mouse models for disease-agnostic interrogation is limited. Furthermore, the...
lack of reproducibility in knockout experiments is a well-documented challenge in drug-target development and behavioral and other translational studies. This lack of reproducibility is commonly due to using poorly defined statistical methods, performing studies in only one sex and practicing bias in animal selection. The development of a comprehensive reference phenotype database by using fully validated, standardized and automated phenotyping procedures across all body systems in mutants of both sexes provides a robust dataset to corroborate disease-causing factors in humans.

The International Mouse Phenotyping Consortium (IMPC) is creating just such a catalog of mammalian gene function that systematically associates mouse genotype-to-phenotype data and enables researchers to formulate hypotheses for biomedical and translational research as well as purpose-driven preclinical studies. The IMPC adult phenotyping pipeline analyzes cohorts of male and female knockouts on an isogenic C57BL/6N background from embryonic-stem-cell resources produced by the International Knockout Mouse Consortium comprising targeted null mutations with reporter-gene elements. Homozygotes are characterized, except in those strains (approximately 30%) in which gene inactivation necessitates the use of heterozygotes to study mice that are embryonic/perinatal lethal or subviable. The pipeline measures a total of 509 phenotyping parameters that encompass diverse biological and disease areas including neurological, behavioral, metabolic, cardiovascular, pulmonary, reproductive, respiratory, sensory, musculoskeletal and immunological parameters. Standardized and harmonized protocols developed by the IMPC are used to decrease phenotypic variance across the centers and build upon experience from the pilot EUMODIC project, in which a limited 7% discordant-phenotype rate has been observed for a large set of 22 common reference mutant lines. Rigorous data quality control is applied to the captured data from the ten phenotyping centers, and an automated statistical analysis pipeline (PhenStat; Online Methods) identifies mutants with statistically significant phenotypic abnormalities.

In the current IMPC data release 5.0 (2 August 2016), 3,328 genes have been fully or partially phenotyped, thus generating over 20 million data points and 28,406 phenotype annotations. Complementing the physiological, behavioral and structural phenotype datasets, the IMPC also provides annotated expression of LacZ data across multiple organ and tissue systems for 1,413 genes and extensive histopathological analysis of adult tissues for 333 genes. The IMPC portal (URLs) provides a single point of access to phenotype data, embryonic-stem-cell and Cas9-RNA-guided nuclease resources and mutant mouse strains. Sophisticated query interfaces for both gene and phenotype data are provided, as well as tools to visualize phenotypes encompassing quantitative, categorical and image data. Periodic data releases provide the most recent genotype-phenotype associations.

In our current analysis, we (i) identified new mouse models for human mendelian disorders with a known genetic basis, (ii) uncovered candidate disease-associated genes for human mendelian disorders for which only a genomic location had been associated and (iii) identified new mouse disease models involving genes with little or no previous functional annotation. A summary of the results is presented in Figure 1 and is described in more detail in the following sections.

Results

Comparison of IMPC findings with previous knowledge

We first investigated the concordance of our phenotype annotations with previously reported data for mouse lines involving the same genes. We found that 621 genes assessed by us have previous mouse-model annotations, on the basis of a literature review of knockout lines by the Mouse Genome Informatics (MGI) group. An assessment of the corresponding 2,547 MGI gene-phenotype associations previously assessed by an IMPC procedure showed that 958 (38%) were detected, and 62% (385 out of 621) of the genes had at least one phenotype reproduced (Supplementary Table 1). This result was in line with previous reports describing the reproducibility of biomedical models. A lack of reproducibility may be due to several factors including different genetic backgrounds and variations in experimental methods and statistical methods. For example, evidence for the previously reported increase in circulating glucose in Gad2 mice was found in our data (URLs) but was not considered statistically significant by our robust methods. Despite our best efforts to be as broad based as possible in the context of a high-throughput project, there were an additional 10,068 MGI phenotypes for these genes that we were unable to assess or that would have required a different type of allele to be introduced to observe the effect. However, as shown below, our pipeline covers all major disease areas. Additionally, use of our resources by the research community to publish over 1,300 new publications is generating numerous new annotations from MGI’s curation of the literature and upcoming changes to our pipeline, such as phenotyping a subset of genes in mice at later ages (12–18 months) and including new behavioral tests, will increase our coverage. Furthermore, we generated extensive new knowledge: 90% (8,984 of 9,942) of the gene-phenotype annotations described by IMPC have not previously been described in the literature.

Models of human mendelian disease

The high volume and complexity of data produced by the IMPC present challenges for identifying relevant models of human diseases. To facilitate discovery, we developed a translational pipeline to
automatically detect phenotypic similarities between the IMPC strains and over 7,000 rare diseases described in the Online Mendelian Inheritance in Man (OMIM)\textsuperscript{17} and Orphanet databases\textsuperscript{18}. The pipeline utilizes the human phenotype ontology (HPO)\textsuperscript{19} annotations for rare diseases, maintained by the Monarch Initiative\textsuperscript{20}, our Mammalian Phenotype Ontology (MP)\textsuperscript{21} annotations of phenotypic abnormalities and the PhenoDigm algorithm, which has also been developed by the Monarch Initiative\textsuperscript{22}. The methodology is based on previous work demonstrating superior identification of disease models, as compared with defining mouse strains solely by ontology or using other methods of calculating phenotype similarity\textsuperscript{22}, and the results provide a quantitative measure of how well an adult mouse model recapitulates the clinical features of a disease.

From the ~15% of mouse protein-coding genes phenotyped by IMPC to date, 889 known rare disease–gene associations represented within OMIM and Orphanet have an orthologous IMPC mouse strain and display at least one phenotype (Supplementary Table 2). By comparing human and mouse phenotypes, our automated pipeline identified 185 adult disease–gene associations for which the IMPC mutant mouse strain modeled the human disease; most associations (134) involved genes for which a mutant mouse strain either had not been generated or had not been reported as a model of that disease, on the basis of the MGI curation (Table 1 and Supplementary Tables 2 and 3). Each of the 889 associations had a mean of 14.7 ± 27.8 (mean ± s.d.) candidate genes associated with the disease, on the basis of the algorithm, with a median rank of 3 for the true associated gene in the 185 sets of recalled associations.

| Category                      | Frequency     |
|-------------------------------|---------------|
| Automated IMPC disease model  | 134/889 (15.1%)|
| Automated IMPC disease models | 185/889 (20.8%)|
| Additional manual lethality   | 75/889 (19.7%)|
| Total IMPC disease models     | 360/889 (40.5%)|

Table 1 Frequency of IMPC models corresponding to mendelian disease–gene associations in OMIM or Orphanet

The range of human mendelian diseases with matching mouse phenotypes was broad and included multiple biological systems (Table 2). Three examples of mouse models that, to our knowledge, have not previously been reported (Fig. 2) are for Bernard–Soulier syndrome, type C (MIM231200), Bardet–Biedl syndrome-5 (MIM615983) and Gordon Holmes syndrome (MIM212840). Bernard–Soulier syndromes are bleeding disorders that result from mutations in genes

![Figure 2](https://example.com/figure2.png)

**Figure 2** Mouse models for mendelian disease. (a,b) Gp9: Bernard–Soulier syndromes result from mutations in the glycoprotein lb platelet membrane receptor complex. Gp9\textsuperscript{m1.LKOMP/MWP12} homozygotes have increased platelet volume (a) and decreased platelet numbers (b). WT, wild type. In a, female control, n = 479 mice, female homozygous, n = 8; male control, n = 428; male homozygous, n = 8; linear mixed-effects model without weight, P = 0. In b, female control, n = 439; female homozygous, n = 8; male control, n = 428; male homozygous, n = 8; linear mixed-effects model without weight, P = 2.31 × 10^-6. Asterisks, significant difference between mutant and same-sex controls; mixed-effects-model P < 0.0000. (c,d) Bbs5: Bbs5 is associated with Bardet–Biedl syndrome (BBS). Bbs5\textsuperscript{m1.EUCOMM/MW51} homozygotes have increased body-fat percentage (c) and impaired glucose tolerance (d). In c, female control, n = 1,276; female homozygous, n = 8; male control, n = 1,296; male homozygous, n = 8; linear mixed-effects model without weight, P = 1.99 × 10^-11. In d, blood glucose levels were determined after a 16-h fast followed by intraperitoneal glucose injection. Female control, n = 491; female homozygous, n = 8; male control, n = 509; male homozygous, n = 8; linear mixed-effects model without weight, P = 2.85 × 10^-12. X-ray visualization of Bbs5 homozygotes and controls. In a–d, box limits, first and third quartiles; line, median; whiskers, minimum and maximum values. (e,f) Rnf216: Gordon Holmes syndrome is associated with Rnf216 and is characterized by hypogonadism and cerebellar ataxia. (e) X-ray image of a 14-week-old Bbs5 homozygous-null female mouse shows the large body size of this strain compared with an age-matched, wild-type control. (f) Rnf216\textsuperscript{m1.EUCOMM/MW51} homozygous-null male mice are infertile. Histopathology images (20× magnification) show seminiferous-tubule degeneration and atrophy with Leydig-cell hyperplasia and epididymal aspermatia in null mice, as compared with the unaffected seminiferous tubules and epididymis in control mice. Scale bars, 100 µm.
encoding protein products of the glycoprotein Ib complex, which serves as the platelet-membrane receptor for von Willebrand factor. Glycoprotein Ib is composed of four subunits encoded by four separate genes, GP1BA, GP1BB, GP9 and GP5, and mutations in all these genes are associated with an autosomal recessive (AR) disorder characterized by prolonged bleeding times, enlarged platelets, an inability to clot and incomplete penetrance of thrombocytopenia23. 

Table 2  Examples of IMPC disease models across diverse biological systems

| Biological system | Disease-associated gene | Human mendelian disease | Relevant human phenotype | Overlapping mouse phenotype |
|-------------------|-------------------------|-------------------------|--------------------------|-----------------------------|
| Bone              | SCARF2                  | Van Den Ende–Gupta syndrome | Long metacarpals          | Increased length of long bones |
| Cardiovascular    | LMNA                    | Cardiomyopathy dilated 1a | Dilated cardiomyopathy    | Increased heart weight       |
| Craniofacial      | MSK1                    | Orofacial cleft 5        | Cleft palate              | Cleft palate                 |
| Embryo            | PSIPH                   | Phosphoserine phosphatase deficiency | Intrauterine growth retardation | Abnormal embryo size         |
| Growth/body size  | GHRHR                   | Isolated growth hormone deficiency, type 1b | Short stature             | Decreased body length        |
| Hearing           | SLC52A2                 | Brown–Vialetto–Van Laere syndrome 2 | Sensorineural hearing impairment | Increased or absent threshold for auditory brainstem response |
| Hematopoietic     | GP9                     | Bernard–Soulier syndrome | Thrombocytopenia          | Thrombocytopenia             |
| Metabolism        | KCNJ11                  | Non-insulin-dependent diabetes mellitus | Type II diabetes mellitus | Impaired glucose tolerance   |
| Muscle            | COL6A2                  | Bethlem myopathy         | Distal muscle weakness    | Dropped grip strength        |
| Neurological      | GOSR2                   | Epilepsy, progressive myoclonic, 6 | Difficulty walking        | Abnormal gait                |
| Reproductive System | RNF216                | Gordon Holmes syndrome   | Infertility               | Male infertility             |
| Retina            | BBS5                    | Bardet–Biedl syndrome 5  | Rod-cone dystrophy        | Abnormal retina morphology   |

Human mendelian disease is caused by a variety of mutations resulting in complete loss or partial loss or gain of function under various modes of inheritance. The IMPC phenotypes only the null allele in a homozygous state or, in the case of embryonic/perinatal lethality, in the heterozygous state. Thus, IMPC mouse strains are suitable for identification of putative disease-associated genes, as opposed to variant identification, by identifying which genes expressing variants of unknown importance are pathogenic. The 889 human diseases associated with genes orthologous to those of the IMPC mouse model strains were inherited with approximately equal frequency by autosomal dominant (AD) or AR genetics (379 AD compared with 423 AR and 87 unknown/X-linked diseases). The frequency of inheritance by AD and AR genetics was also equivalent for the 185 adult disease–gene associations for which the IMPC mutant mouse line modeled the human disease (82 AD compared with 94 AR and 7 unknown/X-linked diseases). These results indicated that the mouse models effectively model human disease independently of the mode of human inheritance. Human AR disease is likely to be a consequence of a mutation resulting in complete or partial loss of function, for which haploinsufficiency is not adequate to produce symptoms. As would be expected, AR human disease was more frequently modeled by homozygous–null mouse mutants: 65% (61/94) of the AR models were viable and phenotyped as homozygous mice, whereas 35% (33/94) were subviable/lethal as homozygotes, and therefore heterozygous mice were phenotyped. AD inheritance can be attributed to either haploinsufficiency or gain-of-function mutations, and we found that 46% of the dominant human mutations were modeled by heterozygous mouse mutants, in agreement with a haploinsufficiency mechanism for almost half of the diseases. Interestingly, 227 of the 423 (54%) tested AR associations were homozygous lethal/subviable in mice, thus leading us to consider whether early mortality might occur in people with these diseases or might occur without extensive medical intervention. Lethality matches are not detectable by our automated algorithm, because human lethality is rarely recorded in the disease HPO annotations, and for 74 of the 889 associations (8%), homozygous lethality is the only mouse phenotype that we have detected to date. To address this potential lack of detection, we manually investigated whether the
associations involving mouse homozygous lethal/subviable strains were associated in OMIM/Orphanet with human embryonic or early death (before 2 years of age) or with severe early-onset disorders in people not likely to survive through puberty without substantial medical support (for example, cleft palate is a lethal phenotype in mice but is easily treatable in humans). This procedure uncovered a further 97 new mouse–human disease associations (Supplementary Table 2, column J annotated with Y-L) for which human lethality was recorded and another 78 for which the disease would probably have been lethal without medical intervention (Supplementary Table 2, column J annotated with Y-PL). Most of these lethality matches were inherited with AR genetics (73%, 122 of the 166 diseases with reported inheritance from OMIM/Orphanet) and modeled by homozygous mouse mutants, in agreement with the conclusion that homozygous loss-of-function mutations in essential genes in humans produce either early death or severe congenital medical conditions requiring advanced medical support for survival. Examples of mouse and human embryonic/early-onset lethality include ventriculomegaly with cystic kidney disease, which results in in utero or neonatal human fatality (MIM219730; gene, CRB2) and Stuve–Wiedermann Syndrome (MIM601559; gene, LIFR). Diseases that would probably have been lethal without medical support and that had a corresponding lethal/subviable mouse strain included COACH disease (MIM216360; gene, RGRP1PL1), Meier Gorlin syndrome 1 (MIM224690; gene, ORC1) and human phospherine phosphatase deficiency (MIM614023; gene, PSPH). For human phospherine phosphatase deficiency, microcomputed tomography indicated that a homozygous lethal mutant mouse in Phosphatase-deficient (EUCCOMM)Whs has structural abnormalities at embryonic day (E) 15.5 that closely resemble the developmental and structural defects in human phospherine phosphatase deficiency (Fig. 3).

When we included these manually curated lethality matches, 40.5% (360) of the disease models had phenotypic overlap with the 889 disease-associated genes (Table 1), and the majority (78%; 279 of 360) represented the first report, to our knowledge, of a candidate mouse model for these diseases. The discovery rate of disease models in our analysis was comparable to that in previous reports on smaller high-throughput mouse phenotyping studies that have found modeling of 46% of 59 and 33% of 42 associations through manual investigation of data.26,27

For cases in which we did not detect a model despite testing for at least one equivalent phenotype (54%; 484), the explanations included differences between human and mouse biology, differences in genetic background, a null allele not being appropriate to model the disease or differing methodologies used for annotation. For example, rarely observed phenotypes for a disease are often recorded in the HPO annotations and would probably fall below the statistical significance threshold if they were similarly nonpenetrant in mice. Finally, there is a slight possibility that some previous descriptions of alleles may have influenced disease modeling in cases in which a hypomorph rather than a null mutant is possible (e.g., the 90 tm1a mutant) or in which a retained neomycin cassette may have altered expression of genes in cis (e.g., 90 tm1a and 10 KOMP1 mutants).

Functional knowledge and candidate genes associated with mendelian diseases
The second major clinical use for the IMPC data is providing new data on the phenotypes and functions of genes. Thus, IMPC has prioritized genes with no known disease associations or minimal Gene Ontology (GO) annotation. On the basis of MGI’s literature curation of mutant strains involving any allele type except conditional mutations, 1,830 of the 3,328 genes phenotyped in this IMPC release have not previously had a mouse mutant produced. No GO molecular-function or biological-process annotations were available for 189 genes, whereas another 903 genes had annotations inferred from computational analysis.28 (Fig. 4a). The phenotypes of these mutant strains provide insights into the functions of a large class of genes (sometimes described as the ‘ignorome’)29, for which there is little or no existing functional information (Supplementary Table 5).

Examples of candidate genes for human mendelian disease with previously little functional information include family with sequence similarity 53 member B (Fam53b), for which no phenotypic variants had been reported in humans or mice. The gene is differentially expressed in adult definitive erythrocytes compared with primitive erythrocytes, with more than a sixfold log2 change, as shown in the Expression Atlas (URLs)30,31. Homozygous Fam53b(+/−)(EUCCOMM)Menga knockout mice showed increased mean corpuscular hemoglobin and decreased erythrocyte cell numbers (Fig. 4b,c), thus suggesting that the gene is involved in hematopoiesis and is a candidate for macrocytic hyperchromic anemias. PhenoDigm identified this gene as having a phenotype mimicking that of Diamond–Blackfan anemia (MIM105650), a group of 15 unique anemias generally attributable to defects in ribosome synthesis but for which known mutations account for only approximately 54% of all affected people32. A single functional study has suggested that Fam53b is required for Wnt signaling, which is key in determining cell fate, cell proliferation, stem cell maintenance and formation of the anterior–posterior axis.33 The Fam53b knockouts thus suggest a new candidate pathway that may account for the 46% of people with Diamond–Blackfan anemia with unknown genetic causes.
As well as providing fundamental insights into the functions of genes with few or no previous functional annotations, the phenotype analyses are also identifying numerous new candidate disease models that may provide a foundation for relating gene function to disease phenotype. This new data and biological resource may be used to detect novel genotype-to-phenotype associations in diseases for which simply considering existing human data would lead to causative variants being overlooked among the overwhelmingly abundant associated variants of unknown importance, as occurs in many exome-sequencing studies. Over half of diagnosed rare diseases still have no known associated genes, and diagnostic rates in most high-throughput mendelian disease sequencing projects are 20–30%, largely because of a lack of functional information for most genes. The IMPC data can be used to remedy this deficiency and to start achieving better diagnostic rates. As a demonstration of the potential application of IMPC data to the discovery of novel disease-associated genes, we identified candidate genes potentially associated with mendelian diseases through unknown molecular mechanisms, for which broad genetic localization was available in OMIM from previous studies. Our disease-matching algorithm identified 135 associations in which our predicted disease-associated gene was found to be within these loci (Supplementary Table 6).

For example, adult mice heterozygous for the Kldc2tm1b(EUCOMM)Hmgu allele have a complex syndrome of abnormalities including altered electrocardiogram findings. The phenotypes match the clinical signs described for the AD disease arrhythmogenic right ventricular dysplasia 3 (ARVD3; MIM60286) associated with cardiac arrhythmias caused by fibrofatty replacement of the right-ventricle myocardium. Kldc2 is syntenic with the ARVD3 locus, thus suggesting that it may be a candidate gene. Usmg5-null mice recapitulate the clinical symptoms of muscle weakness and abnormal gait seen in people with the dominant intermediate A form of Charcot–Marie–Tooth disease. The human ortholog USMG5 is located within a 9.8-Mb critical region identified in people with this disease. Whereas the implicated human loci are sometimes megabases in length and encompass hundreds of genes, these examples illustrate how IMPC phenotype data allow for the scoring of candidate genes with disease-causing variants and have important implications for current genetic projects investigating rare diseases through next-generation-sequencing technologies.

**DISCUSSION**

By analyzing phenotypic similarities between IMPC’s mouse strains and human diseases, we provided new disease models and novel functional knowledge for a substantial and growing proportion of protein-coding genes. These models are readily available and can be exploited to study disease mechanisms, to develop new gene therapies and pharmacological treatments, and to improve understanding of gene function. The novelty of the IMPC lies in both the scale of the vision to produce a comprehensive catalog of mammalian gene function across all genes and the non-hypothesis-driven, standardized approach to phenotyping. This resource facilitates novel discoveries about the functions of genes and their roles in diseases, as highlighted above.

The potential of IMPC phenotypic comparisons for prioritizing candidates in human mendelian syndromes is accessible to clinical researchers performing next-generation-sequencing-based diagnostics through inclusion within the Exomiser software package, which combines an assessment of variant pathogenicity with gene candidacy on the basis of the similarity of human disease phenotypes to known phenotypic knowledge from humans, mice and fish. Exomiser is being applied within the NIH Undiagnosed Disease Program and the 100,000 Genomes Project, which embeds genomics into a national healthcare system.
the IMPC will also be of value in recently launched precision-medicine efforts, whose goals are to improve treatment through the customization of healthcare according to the genomic information of individual patients and environmental factors. Harmonization of phenotypic traits captured in diverse formats across multiple centers will be critical to the stratification of disease populations for improved treatment as well as using model-organism data to better identify disease-causing gene variants.

Whereas advances in technologies based on CRISPR and induced pluripotent stem cells have vastly expanded researchers’ toolkits, the work of the IMPC highlights the continuing importance of mouse models in understanding disease mechanisms. Mice are vertebrate mammals with physiological characteristics that recapitulate those of all major human biological systems, thus allowing the study of processes that cannot be investigated in vitro, including the effects of behavioral, inflammatory, endocrine and sex-specific processes on disease. Whereas CRISPR-based methodologies now allow for genome engineering in nearly every species, mice have other characteristics that have made them a widely used model organism for over a century. Inbred mouse strains such as the C57BL/6N strain used by the IMPC have standardized, uniform genetic backgrounds that decrease phenotypic variability, and most strains have a 2-year lifespan that allows for comprehensive studies in a timely manner.

Reproducibility of results in translational studies is a major issue, and we found that the overlap of phenotypes between IMPC mouse strains and previously published mutant strains was in line with results from other studies investigating reproducibility. This finding highlights the importance of high-quality phenotype annotation of human clinical records and mouse phenotypes and demonstrates the importance of open sharing of data. Toward this end, the IMPC adheres to the ARRIVE guidelines for reproducibility of animal-model experiments, including making all data available and allowing for transparent statistical analysis via free distribution of our PhenStat software12.

In conclusion, the IMPC has established an ever-expanding knowledge base of mammalian gene function, a large resource of novel disease models and the capacity for functional validation of variants identified in disease sequencing projects. This information should be of great value to the human disease research community.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by NIH grants U54 HG006370 (T.F.M., P.F., A.-M.M., H.P., D.S. and S.D.M.B.), U42 OD011185 (S.A.M.), U54 HG006332 (R.E.B. and K.L.S.), U54 HG006348-81 and OD011174 (A.L.B.), 1R24OD011883 (C.I.M., M.H., N.W. and D.S.), HG006364-0381, U54H G906364 (K.C.K.L. and C.I.M.) and U42 OD011175 (C.M. and K.C.K.L.). Additional support was provided by the Wellcome Trust, Medical Research Council Strategic Award 53658 (S.W. and S.D.M.B.); the government of Canada through Genome Canada and Ontario Genomics (OGI-051) (C.M. and S.D.M.B.); the National Centre for Scientific Research (CNRS); the French National Institute of Health and Medical Research (INSERM); the University of Strasbourg (UDS); the Centre Européen de Recherche en Biologie et en Médecine; the Agence Nationale de la Recherche under the framework program Investissements d’Avenir labeled ANR-10-IDEX-0002-02, ANR-10-INBS-07 PHENOMIN (Y.H.); the German Federal Ministry of Education and Research through Infrafrontier grant 01KX1012 (S.A.M., V.G.-D. and M.H.d.A.); and the ‘EUCOMM: Tools for Functional Annotation of the Mouse Genome’ (EUROMITOOLS) project, grant agreement FP7-HEALTH-F4-2010-261492 (W.M.).

AUTHOR CONTRIBUTIONS

T.F.M., D.B.W., N.C. and D. Smeddon contributed to data analysis, writing the paper and the design and execution of the work. N.H., M.H., N.W., C.J.M., P.M., J.O.J., C.-K.C., J.T., H.M., M.R., N.K., J.W., H.W., J.M. and D. Smeddon contributed to development of the software, statistical analysis, database and APIs. L.S., T.F., N.R. and S.G. performed quality control of the phenotype data J.B., J.K.W., S.Y.C., G.E.C., M.E.S., C.L.R., J.G., V.G.-D., T.S., G.P. and L.R. led the experimental work and data production. J.M., J.S., A.B., M.E.D., M.H.d.A., M.M., Y.H., G.P.T.-V., K.C.K.L., X.G., C.M., M.J.I., S.A.M., K.L.S., R.E.B., S.W., A.-M.M., P.F., H.P., J.W., A.L.B., W.C.S., D.J.A., S.D.M.B., W.W., S.N., A.M.F., L.M.J.N., Y.O. and J.K.S. were senior principal investigators of the key programs that contributed to the paper and were critical in the design, management and execution of the study; and the writing and reviewing of the manuscript. The additional IMPC consortium members all contributed to data acquisition and data handling.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

URLs

IMPC portal, http://www.mousephenotype.org/; glucose results for Gad2, http://mousephenotype.org/data/charts?accession=MGI:95634&allele_accession=MGI:5548398&parameter_stable_id=1MPC_IPG_010_001&metadatad_group=297b1cf545ae86ea013b14aca71ef1&zygosity=homozygote&phenotyping_center=HMGU/; IMPC FTP site, ftp://ftp.ebi.ac.uk/pub/databases/impc/latest/; IMPC publications, http://www.mousephenotype.org/data/allerelefs/; International Mouse Phenotyping Resource of Standardised Screens (IMPreSS), http://www.mousephenotype.org/impress/; IMPC data access, http://www.mousephenotype.org/data/documentation/index/; IMPC ARRIVE guidelines, http://www.mousephenotype.org/about-impc/arrive-guidelines/; IMPC GO annotations, https://www.mousephenotype.org/data/gene2go/; ExpressionAtlas result for Fam53b, http://www.ebi.ac.uk/gxa/genes/ENSMUSG00000030956?b=%7B%7bus%7B%20organism%20part%20%7B%20true%7D%7D&d=%7B%7D - differential; GEMM, https://www.har.mrc.uk/gemm-call-guidance-applicants/; PhenStat, http://www.bioc杜绝.org/packages/release/bioc/vignettes/PhenStat/inst/doc/PhenStatUsersGuide.pdf; MGI, http://www.informatics.jax.org/; MGI downloads, ftp://ftp.informatics.jax.org/pub/reports/; Monarch Initiative, https://monarchinitiative.org/; OWLtools, https://github.com/owllcollab/owltools/.

REFERENCES

1. Belo, S.M., Smith, C.L. & Eppig, J.T. Allele, phenotype and disease data at Mouse Genome Informatics: improving access and analysis. Mamm. Genom. 26, 285–294 (2015).
2. Begley, C.G. & Ellis, L.M. Drug development: raise standards for preclinical cancer research. Nature 483, 531–533 (2012).
3. Fonio, E., Golani, I. & Benjamin, Y. Measuring behavior of animal models: faults and remedies, Nat. Methods 9, 1167–1170 (2012).
4. Kilkenney, C., Browne, W.J., Cuthill, I.C., Emerson, M. & Altman, D.G. Improving bioscience research reporting; the ARRIVE guidelines for reporting animal research. PLoS Biol. 8, e1000412 (2010).
5. Brown, S.D.M. & Moore, M.W. The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. Mamm. Genom. 23, 632–640 (2012).
6. Hrabé de Angelis, M. et al. Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. Nat. Genet. 47, 969–978 (2015).
7. Skarnes, W.C. et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature 474, 337–342 (2011).
8. Bradley, A. et al. The mammalian gene function resource: the International Knockout Mouse Consortium. Mamm. Genom. 23, 580–586 (2012).
9. Rosen, B., Schick, J. & Wurst, W. Beyond knockouts: the International Knockout Mouse Consortium delivers modular and evolving tools for investigating mammalian genes. Mamm. Genom. 26, 456–466 (2015).
10. Dickinson, M.E. et al. High-throughput discovery of novel developmental phenotypes. Nature 537, 508–514 (2016).
11. Adams, D. et al. Bloombury report on mouse embryo phenotyping. recommendations from the IMPC workshop on embryonic lethal screening. Dis. Model. Mech. 6, 571–579 (2013).
12. Kurbatova, N., Mason, J.C., Morgan, H., Meehan, T.F. & Karp, N.A. PhenStat: a tool kit for standardized analysis of high throughput phenotypic data. PLoS One 10, e0131274 (2015).

13. West, D.B. et al. A lacZ reporter gene expression atlas for 313 adult KOMP mutant mouse lines. Genome Res. 25, 598–607 (2015).

14. Adissu, H.A. et al. Histopathology reveals correlative and unique phenotypes in a high-throughput mouse phenotyping screen. Dis. Model. Mech. 7, 515–524 (2014).

15. Koscielny, G. et al. 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 (2014).

16. Freedman, L.P., Cockburn, I.M. & Simcoe, T.S. The economics of reproducibility in preclinical research. PLoS Biol. 13, e1002165 (2015).

17. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 43, D789–D798 (2015).

18. Rath, A. et al. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum. Mutat. 33, 803–808 (2012).

19. Köhler, S. et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 45, D1, D865–D876 (2017).

20. Mungall, C.J. et al. Use of model organism and disease databases to support matchmaking for human disease gene discovery. Hum. Mutat. 36, 979–984 (2015).

21. Smith, C.L. & Eppig, J.T. 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 (2015).

22. Smedley, D. et al. PhenODig: analyzing curated annotations to associate animal models with human diseases. Database (Oxford) 2013, bat025 (2013).

23. Savoia, A. et al. Spectrum of the mutations in Bernard-Soulier syndrome. Hum. Mutat. 35, 1033–1045 (2014).

24. Khan, S.A. et al. Genetics of human Bardet-Biedl syndrome, an updates. Clin. Genet. 90, 3–15 (2016).

25. Margolin, D.H. et al. Axiala, dementia, and hypogonadotropism caused by disordered ubiquitination. N. Engl. J. Med. 368, 1992–2003 (2013).

26. Santos, P. et al. RNF216 mutations as a novel cause of autosomal recessive Huntington-like disorder. Neurology 84, 1760–1766 (2015).

27. White, J.K. et al. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell 154, 452–464 (2013).

28. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43, D1049–D1056 (2015).

29. Pandey, A.K., Lu, L., Wang, X., Homayouni, R. & Williams, R.W. Functionally enigmatic genes: a case study of the brain ignoreme. PLoS One 9, e88889 (2014).

30. Petryszak, R. et al. Expression Atlas update: an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 44, D746–D752 (2016).

31. Kingsley, P.D. et al. Ontogeny of erythroid gene expression. Blood 121, e5–e13 (2013).

32. Borja, I. et al. The ribosomal basis of Diamond-Blackfan anemia: mutation and database update. Hum. Mutat. 31, 1269–1279 (2010).

33. Kizil, C. et al. Simplet/Fam53b is required for Wnt signal transduction by regulating β-catenin nuclear localization. Development 141, 3529–3539 (2014).

34. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat. Protoc. 10, 2004–2015 (2015).

35. Bone, W.P. et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet. Med. 18, 608–617 (2016).

36. Harkness, J.H., Shi, X., Janowsky, A. & Phillips, T.J. Trace amines: a case study of the brain ignorome. Nucleic Acids Res. 44, 608–617 (2015).

37. Cade, B.E. et al. Genetic associations with obstructive sleep apnea traits in Hispanic/Latino Americans. Am. J. Respir. Crit. Care Med. 194, 886–897 (2016).

38. Knowles, J.W. et al. Identification and validation of N-acetyltransferase 2 as an insulin sensitivity gene. J. Clin. Invest. 126, 403 (2016).

39. Lang, B. et al. Recurrent deletions of ULK4 in schizophrenia: a gene crucial for neuritogenesis and neuronal motility. J. Cell Sci. 127, 630–640 (2014).

40. McIntyre, R.E. et al. A genome-wide association study for regulators of micronucleus formation in mice. G3 (Bethesda) 6, 2343–2354 (2016).

41. Levy, R., Mott, R.F., Iraqi, F.A. & Gabet, Y. Collaborative cross mice in a genetic association study reveal new candidate genes for bone microarchitecture. BMC Genomics 16, 1013 (2015).
Data quality control (QC). IMPC Data Coordination Centre (DCC) through an internal QC web inter-cycle occurred when data were uploaded from the phenotyping center to the (e.g., insufficient sample or incorrect instrument calibration) and are detailed that account for known variations in experimental workflow and design of package developed for IMPC. PhenStat is a suite of statistical analysis tools Statistical analysis. and experimental metadata. With IMPC data release version 5.0, (2 August measured on the same day with an equal number of control mice, the compari-
care for the breeding scheme used at that center (exact details are available in the IMPC data portal) and supplied the analysis results to the IMPC DCC. All available wild-type and mutant mice were used in the analysis, with center-specific blindings strategies used during group allocation, no specific inclu-
sion or exclusion criteria, and no randomization approach beyond relying on mendelian inheritance to randomize, as detailed in our ARRIVE guideline document (URLs). All analysis presented in this publication was based on the binary assignment of a significant deviation (or lack thereof) from wild type and the associated phenotype term. Detailed output of our statistical analysis for each test is presented in our portal pages (URLs), including all raw data, summaries, visualizations, variances and calculated P values for genotype–phenotype associations.

Matching mouse phenotypes to OMIM and Orphanet disease descriptions. Automated PhenoDigm. We used the Human Phenotype Ontology (HPO) annotations available from the Monarch Initiative (accessed 2 September 2016) describing the clinical phenotypic features of over 7,000 diseases reported in OMIM and Orphanet. These HPO terms were semantically compared with the phenotypic features (MP annotations) of IMPC mouse strains by using the PhenoDigm algorithm, developed by us and fellow members of the Monarch Initiative, as reflected in authorship, to generate an overall score indicating the similarity of the phenotype of a given mouse strain to that of a particular disease. PhenoDigm calculates the individual score for each HPO–MP phenotype match, on the basis of the proximity of the two terms in the overall cross-species ontology (Jaccard index; simJ) and the observed frequency of the phenotype in common from the overall disease and mouse annotations (Information Content; IC); i.e., exact clinical and mouse phenotype matches involving rarely observed phenotypes scored highest. The geometric mean of the IC and simJ was used to generate the HPO–MP pairwise score. The overall PhenoDigm percentage score was a comparison of the best and mean scores for all the pairwise HPO–MP comparisons relative to the maximum possible scores for a mouse model perfectly matching the disease. The disease models described in this paper were selected by applying a threshold of at least one HPO–MP match with a score greater than 1.35, thus maximizing precision and recall relative to those of other similarity thresholds of 1.0, 1.25, 1.5 or 1.75 (Supplementary Fig. 1).

Known human genes and regions associated with diseases were extracted from OMIM and Orphanet, and matching mouse orthologs were identified from HomoloGene. Comparisons to previous mouse mutants from the MGI resource were achieved by downloading and processing a file named MGI_GenePheno.rpt, which contains literature curation of mouse lines associated with all allele types except those involving conditional mutations, and ALL_OMIM.rpt, which curates any literature assertions of a particular mouse line being a mouse model of a particular OMIM disease (URLs; downloaded 2 September 2016).

Lethality matching. Screening for lethal or potentially lethal genes from data within the OMIM database could not be automated. For the set of mouse genes that were homozygous preweaning suvivable or lethal and that also had OMIM records, we manually inspected the OMIM records to identify those...
with reported in utero or early deaths (before two years of age) and coded these in Supplementary Table 1 as yes–lethal (Y-L), thus indicating that for some human cases with mutations for these genes, the phenotype of human lethality matched the phenotype of mouse subviability. We also screened for OMIM records with severe congenital defects and/or rapid progression of early-onset severe disease in human patients requiring substantial medical support for survival. Mice with similar phenotypes would not be likely to survive through weaning in the absence of medical support and therefore were scored as yes–probable lethal (Y-PL), thus indicating a probable match of the human phenotype to the mouse subviable phenotype.

Matching candidate-gene phenotypes to human traits from OMIM linkage and cytogenetic findings. Diseases with no known molecular mechanism but a narrowed-down cytogenetic region containing the likely causative gene were extracted from OMIM (accessed 2 September 2016). Ensembl was used to identify the human genes within these regions and their mouse orthologs retrieved from HomoloGene. The overlaps between these genes and candidates from the PhenoDigm analysis of the same disease were then flagged within our database and are highlighted in both our portal and Supplementary Tables 1–6.

Identifying novel gene–phenotype relationships from the IMPC database. An online tool in the IMPC portal (URLs) imports GO annotations daily from the Quick GO resource45 and categorizes them on the basis of the evidence codes assigned by GO curators. Annotations were analyzed on 24 March 2017. We started with 2,668 genes that had IMPC nonlethal phenotypes. Categories incorporated the following evidence codes:

- Experimental: inferred from experiment (EXP), inferred from direct assay (IDA), inferred from physical interaction (IPI), inferred from mutant phenotype (IMP), inferred from genetic interaction (IGI) and inferred from expression pattern (IEP).
- Curated computational: inferred from sequence or structural similarity (ISS), inferred from sequence orthology (ISO), inferred from sequence alignment (ISA), inferred from genomic context (IGC), inferred from biological aspect of ancestor (IBA), inferred from biological aspect of descendant (IBD), inferred from key residues (IKR), inferred from rapid divergence (IRD) and inferred from reviewed computational analysis (RCA).

Automated electronic: inferred from electronic annotation (IEA), traceable author statement (TAS), nontraceable author statement (NAS) and inferred by curator (IC).

No biological data available: no biological data available (ND) and not listed as a gene in GO (no evidence code).

Ethical approval. Mouse production, breeding and phenotyping at each center was done in compliance with each center’s ethical animal care and use guidelines in addition to the applicable licensing and accrediting bodies, in accordance with the national legislation under which each center operates. Details of each center’s ethical review organization, processes and licenses are provided in Supplementary Table 7. All efforts were made to minimize suffering by considerate housing and husbandry. All phenotyping procedures were examined for potential refinements disseminated throughout the IMPC. Animal welfare was assessed routinely for all mice.

Code availability. The automated phenotype comparisons were performed with the open-source OWLTools package provided by the Monarch Initiative.

Data availability. All data presented here are openly available from the IMPC portal via our FTP site (ftp://ftp.ebi.ac.uk/pub/databases/impc/latest/). We also provide regular data exports to the MGI group, which provides public access to all available mouse data, and the Monarch Initiative, which integrates genotype–phenotype data from humans and numerous other species.