A Comprehensive Resource for Human Disease Models: The International Mouse Phenotyping Consortium

Terrence F. Meehan1, Nathalie Conte1, David B. West2, Julius O. Jacobsen3, Jeremy Mason1, Jonathan Warren1, Chao-Kung Chen1, Ilincu Tudose1, Mike Relac1, Peter Matthews1, Natasha Karplus1, Luis Santos1, Tanja Fiegel2, Natalie Ring2, Henrik Westerberg2, Simon Greenaway3, Duncan Sneddon3, Hugh Morgan3, Gemma FCCodner3, Michelle E. Stewart4, James Brown5, Neil Horner5, The International Mouse Phenotyping Consortium5, Melissa Haendel1, Nicole Washington2, Christopher J. Munagat1, Corey L. Reynolds1, Juan Gallegos1, Valerie Gallus-Durner1, Tania Song1,12,13,14, Guillaume Pavlovic1,12,13,14, Lynette R. Bower5, Mark Moore16, Iva Morse17, Xiang Gao18, Glauro P. Tocchini-Valentini19, Yuichi Obata20, John Seavitt9, Arthur L. Beaudet21, Mary E. Dickinson5, Yann Herault1,12,13,14, Wolfgang Wurst10, Martin Hrabe de Angelis10, K.C. Kent Lloyd16, Ann M Flenniken21, Lauryl M. Nutter21, Susan Newbigging21, Colin McKenzie21, Monica J. Justice25, Stephen A. Murray22, Karen L. Svenson22, Robert E. Braun22, Jacqueline K. White2, Allan Bradley2, Paul Flice1, Sara Wells1, William C. Skarnes4, David J. Adams4, Helen Parkinson1, Ann-Marie Mallon1, Steve D.M. Brown2, Damian Smedley23

1European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
2Children’s Hospital Oakland Research Institute, Oakland, California 94609, USA
3William Harvey Research Institute, Queen Mary University of London, London, E1 4NS, UK
4The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
5Medical Research Council Harwell (Mammalian Genetics Unit and Mary Lyon Centre), Harwell, Oxfordshire OX11 0RD, UK
6The International Mouse Phenotyping Consortium
7Department of Medical Informatics and Clinical Epidemiology and OHSU Library, Oregon Health & Science University, Portland, OR, 97239, USA
8Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
9Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA
10Helmholtz Zentrum Münchener, German Research Center for Environmental Health, Institute of Experimental Genetics, Neuherberg 85764, Germany
11CELPHEDIA, PHENOMIN, Institut Clinique de la Souris (ICS), 1 rue Laurent Fries, F-67404 Illkirch-Graffenstaden, France
12Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Université de Strasbourg, Illkirch, France
13Centre National de la Recherche Scientifique, UMR7104, Illkirch, France
14Institut National de la Santé et de la Recherche Médicale, U964, Illkirch, France
15Mouse Biology Program, University of California, Davis, California 95618, USA
16IMPC, San Anselmo, California 94960, USA
17Charles River Laboratories, Wilmington, Massachusetts 01887, USA
18SKL of Pharmaceutical Biotechnology and Model Animal Research Center, Collaborative Innovation Center for Genetics and Development, Nanjing Biomedical Research Institute, Nanjing University, Nanjing 210061, China
19Monterotondo Mouse Clinic, Italian National Research Council (CNR), Institute of Cell Biology and Neurobiology, Monterotondo Scalo I-00015, Italy
20RIKEN BioResource Center, Tsukuba, Ibaraki 305-0074, Japan
The Centre for Phenogenomics, Toronto, Ontario M5T 3H7, Canada

Mouse Imaging Centre, The Hospital for Sick Children, Toronto, Ontario M5T 3H7, Canada

The Jackson Laboratory, Bar Harbor, Maine 04609, USA

Contributed equally

Correspondence to:
Dr Damian Smedley
William Harvey Research Institute
Queen Mary University of London
London, E1 4NS
Email: d.smedley@qmul.ac.uk

Additional consortium authors from The International Mouse Phenotyping Consortium

Although next generation sequencing has revolutionised the ability to associate variants with human diseases, diagnostic rates and development of new therapies are still limited by our lack of knowledge of function and pathobiological mechanism for most genes. To address this challenge, the International Mouse Phenotyping Consortium (IMPC) is creating a genome- and phenome-wide catalogue of gene function by characterizing new knockout mouse strains across diverse biological systems through a broad set of standardised phenotyping tests. Here we describe the plethora of new disease models made readily available to the biomedical community as well as basic and translational knowledge that has arisen from analysis of the first 3328 genes.

With its extensive toolkit for genome modification and capacity for recapitulating human disease, the laboratory mouse is arguably the preferred model organism for studying and validating the effect of genetic variants in Mendelian disease, as well as identifying previously unsuspected disease genes. Null mouse mutations have hitherto been generated and described in the literature for approximately one-half of the genes in the genome. However, hypothesis-driven phenotyping of these mutants reveals discoveries in areas that largely reflect the expertise and specific research questions of individual investigators. As a result, the extent of functional annotation, potential to fully uncover pleiotropy, and opportunity to exploit mutant mouse models for disease-agnostic interrogation is limited. Furthermore, the lack of replicability in knockout experiments is a well-documented challenge in drug target development, behavioural and other translational studies and 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 employing fully validated, standardised and automated phenotyping procedures across all body systems that are applied to mutants of both sexes provides a robust dataset to corroborate disease-causing factors in humans.

The IMPC is creating just such a catalogue of mammalian gene function that systematically associates mouse genotype-phenotype data and enables researchers to formulate hypotheses for biomedical and translational research, and purpose-driven preclinical studies. The IMPC adult phenotyping pipeline analyses cohorts of male and female knockouts on an isogenic C57BL/6N background from Embryonic Stem (ES) cell resources produced by the International Knockout Mouse Consortium comprising targeted null mutations with reporter gene elements. Homozygotes are characterised, except in those strains (approximately 30%) where gene inactivation necessitates the
use of heterozygotes to study mice that are embryonic/perinatal lethality or subviable\textsuperscript{10,11}. The pipeline measures a total of 509 phenotyping parameters that encompass diverse biological and disease areas including neurological, behavioural, metabolic, cardiovascular, pulmonary, reproductive, respiratory, sensory, musculoskeletal, and immunological.\textsuperscript{24} Standardised and harmonised protocols developed by the IMPC are used to reduce phenotype variance across the centers and builds off experience from the pilot EU-MODIC project where a limited 7\% discordant phenotype rate was observed for a large set of 22 common reference mutant lines.\textsuperscript{6}

Rigorous data quality control is applied to the captured data from the 10 phenotyping centers and an automated statistical analysis pipeline (PhenStat\textsuperscript{12}, see Methods) identifies mutants with statistically significant phenotype abnormalities. In the current IMPC data release 5.0 (August 2016), 3,328 genes have been fully or partially phenotyped, generating over 20 million data points and 28,406 phenotype annotations. Complementing the physiological, behavioural and structural phenotype datasets, the IMPC also provides annotated expression of LacZ data across multiple organ and tissue systems for 1413 genes\textsuperscript{13} and extensive histopathology analysis of adult tissues for 333 genes\textsuperscript{14}. The IMPC portal (http://www.mousephenotype.org) is the single point of access to phenotype data, ES cell and Cas9-RNA-guided nuclease resources and mutant mouse strains. Sophisticated query interfaces of both gene and phenotype data are provided, as well as tools to visualise phenotypes encompassing quantitative, categorical, and image data\textsuperscript{15}. Periodic data releases provide the latest genotype-phenotype associations.

In our current analysis, we identify 1) new mouse models for human Mendelian disorders with a known genetic basis, 2) uncover candidate disease genes for human Mendelian disorders where to date only a genomic location is associated, and 3) identify new mouse disease models involving genes with previously little or no functional annotation. A summary of the results is presented in \textbf{Figure 1} and described in more detail in the following sections.

\textbf{Comparison of IMPC findings with previous knowledge}

Prior to assessing the utility of our resource for human disease researchers, we first investigated how concordant our phenotype annotations are with previously reported data for mouse lines involving the same genes. 621 genes have phenotype annotations available from the IMPC as well as pre-existing mouse models curated from literature review of knockout lines by the Mouse Genome Informatics (MGI) group\textsuperscript{1}. An assessment of the 2547 MGI gene-phenotype associations that have already been
assessed by an IMPC procedure revealed 958 (38%) were detected, with 62% (385 out of 621) of the genes having at least one phenotype reproduced (Supplement Table 1).

This is in line with previous reports describing reproducibility of biomedical models. Where phenotypes are not reproduced this could be due to several factors including the null allele being tested in different genetic backgrounds in the previous report and variations in experimental technique and statistical methods e.g. evidence for the previously reported observation of increased circulating glucose for Gad2 mice is seen in our data (http://mousephenotype.org/data/charts?accession=MGI:95634&allele_accession=MGI:5548938&parameter_stable_id=IMPC_IPG_010_001&metadata_group=297b1cf545aee8ee0113b14aca71ef1&zygosity=homozygote&phenotyping_center=HMGU) but not considered statistically valid by our robust methods. As well as phenotypes that are not reproduced, there will be previously reported gene functions and phenotypes that we have not assessed despite our best efforts to be as broad-based as feasible in the context of a high-throughput project, or that require a different type of allele to be introduced to observe the effect. For example, the genes analysed for reproducibility have an additional 10068 gene-phenotype associations curated by MGI that are not addressed by our current pipeline. However, we show below that our pipeline is capable of detecting animal models across all the major disease areas and the lines that we make readily available to the research community are already generating numerous additional annotations through MGI’s literature curation of over 1300 papers published to date using our resources. Future modifications to our pipeline such as phenotyping a subset of genes at later ages (12-18 months) and new behavioural tests will increase our coverage and we already generate extensive new knowledge, as discussed further below, with 90% (8984 of 9942) gene-phenotype annotations described by IMPC having not been described in the literature before.

Models of Human Mendelian Disease

The high volume and complexity of data produced by the IMPC presents challenges for finding relevant human disease models. To facilitate discovery, we developed a translational pipeline to automatically detect phenotype similarities between the IMPC strains and over 7000 rare diseases described in the Online Mendelian Inheritance in Man (OMIM) and Orphanet databases. The pipeline utilises the human phenotype ontology (HPO) annotations for rare diseases maintained by the Monarch Initiative and our Mammalian Phenotype Ontology (MP) annotations of phenotype abnormalities and the PhenoDigm algorithm, also developed by the Monarch Initiative.

5
The results provide a quantitative measure of how well an adult mouse model recapitulates clinical features of a disease and is based on previous work that demonstrated superior identification of disease models compared to defining mouse strains solely by orthology or by other methods of calculating phenotype similarity. From the ~13% of mouse protein-coding genes phenotyped thus far by IMPC, 889 known rare disease-gene associations represented within OMIM and Orphanet have an orthologous IMPC mouse strain and display at least one phenotype (Supplement Table 2). By comparing human and mouse phenotypes, our automated pipeline identified 185 adult disease-gene associations where the IMPC mutant mouse strain modelled the human disease, with the majority (134) involving genes that have not had a mouse generated before or reported as a model of that disease from the curation efforts of MGI (Table 1, Supplement Tables 2 & 3). Each of the 889 associations had a mean of 14.7 ± 27.8 (SD) candidate genes for the disease from the algorithm, with a median rank of 1 for the true associated gene in the 185 recalled associations.

The range of human Mendelian diseases with matching mouse phenotypes was broad and included multiple biological systems (Table 2). Three examples of new mouse models first reported here (Figure 2) are for Bernard-Soulier syndrome, Type C (OMIM:231200), Bardet-Biedl syndrome-5 (OMIM:615983) and Gordon Holmes syndrome (OMIM:212840). Bernard-Soulier syndromes are bleeding disorders that result from mutations in genes encoding protein products of the glycoprotein Ib (GP Ib) complex that serves as the platelet membrane receptor for von Willebrand factor. GP Ib is composed of 4 subunits encoded by 4 separate genes: GP1BA, GP1BB, GP9, and GP5 with mutations in all these genes being associated with an autosomal recessive disorder characterised by prolonged bleeding times, enlarged platelets, an inability to clot, and incomplete penetrance of thrombocytopenia. GP5m1.1(KOMP)Vlcg null homozygotes have a decreased number of platelets with a larger cell volume (Figure 2 A,B), recapitulating key features of the disease and adding evidence that the point mutations associated with disease in humans lead to a functionally null complex.

Bardet-Biedl syndromes (BBS) are heterogeneous autosomal recessive ciliopathies characterised by retinitis pigmentosa, obesity, kidney dysfunction, polydactyly, behavioral dysfunction, and hypogonadism. The disorder is associated with no fewer than 19 genes whose products form the BBSome, a protein complex involved in signaling receptor trafficking within cilia, which may also have functions not involving cilia. Twenty mutations that include splice site, missense/nonsense, insertion, indels and deletion mutations within the BBS5 gene account for 4% of all BBS cases.
null mice exhibit abnormal retina morphology resembling the retinal
dystrophy observed in Bardet-Biedl syndrome patients. Other phenotypes were also
observed in null mice recapitulating many hallmarks of BBS including obesity (Figure
2C,E) as well as other features such as impaired glucose homeostasis (Figure 2D).

Gordon Holmes syndrome is another autosomal recessive disorder characterised by
hypogonadism as well as cerebellar ataxia that has been associated with RNF216
Male infertility was observed in Rnf216 homozygous null mice with
histopathology identifying seminiferous tubule degeneration and atrophy characterised
by diffuse absence of most or all germ cells and presence of occasional multinucleated
spermatids with pyknotic nuclei within tubules that are lined by vacuolated Sertoli cells.
Seminiferous changes were accompanied by diffuse interstitial cell (Leydig cell)
hyperplasia. The epididymis was devoid of spermatozoa (epididymal aspermia) (Figure
2F).

From the 704/889 known associations where we did not detect an IMPC model,
48 have not yet been tested in the mouse for a phenotype that could recapitulate any of
the clinical phenotypes, leaving 656 (74%) associations where our automatic algorithm
did not detect a potential disease phenotype from the IMPC pipeline. In order to
evaluate the sensitivity of the automated human-mouse phenotype matching, we
manually evaluated 100 randomly selected examples of these associations where no
match was detected (Supplement Table 4), leading to 12 additional discoveries where
the phenotype matches fell below the similarity threshold used in our algorithm. For
example, decreased startle reflex indicates a match for the clinical phenotype of
deafness but is low scoring by PhenoDigm. PhenoDigm is optimised to maximise
precision and recall (see Methods and Supplementary Figure 1) and reducing our
threshold to detect such matches would introduce many false positives. In addition, 1
further model was found by inclusion of additional histopathology annotation data from
our database. As manual assessment is not feasible in the long term given the ever-
increasing number of strains and new data for existing strains, we will instead focus on
other sources of phenotype data such as histopathological annotations in future
versions of the PhenoDigm algorithm.

Human Mendelian disease is caused by a variety of complete loss, partial loss or
gain-of-function mutations under various modes of inheritance. The IMPC only
phenotypes the null allele in a homozygous state or, if embryonic/perinatal lethal, in the
heterozygous state. Thus, IMPC mouse strains are suitable for putative disease gene,
as opposed to variant, identification e.g. for identifying which genes expressing variants
of unknown significance are pathogenic. The 889 human diseases associated with
genes orthologous to the IMPC mouse strains were inherited with roughly equal frequency by autosomal dominant (AD) or recessive (AR) genetics (n = 379 AD vs 423 AR, and 87 unknown/X-linked). The frequency of inheritance by AD and AR genetics was also equivalent for the 185 adult disease-gene associations where the IMPC mutant mouse line modelled the human disease (n = 82 AD vs 94 AR, and 7 unknown/X-linked). This indicates that the mouse models were effectively modelling human disease independent of the mode of human inheritance. Human AR disease is likely a consequence of a complete or partial loss-of-function mutation where haploinsufficiency is not adequate to produce symptoms. As would be expected, AR human disease was more frequently modelled by homozygous null mouse mutants: 65% (61/94) of the AR models were viable and phenotyped as homozygous mice, while 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 modelled by heterozygous mutants in the mouse, consistent with a haploinsufficient mechanism for almost half of the diseases.

Interestingly, 227 of the 423 (54%) tested AR associations were homozygous lethal/subviable in the mouse leading us to consider whether early mortality occurred in these patients or would have occurred without extensive medical intervention. Lethality matches are not detectable by our automated algorithm, as 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 we have detected so far. To address this, we manually investigated whether the associations involving mouse homozygous lethal/subviable strains were associated in OMIM/Orphanet with human embryonic or early death (< 2 years) or with severe, early onset disorders in patients not likely to survive through puberty without significant medical support e.g. cleft palate is a lethal phenotype in mice but easily treatable in humans. This uncovered a further 97 new mouse/human disease associations (Supplement Table 2, column J annotated with Y-L) where human lethality was recorded and another 78, where the disease would probably have been lethal without medical intervention (Supplement Table 2, column J annotated with Y-PL). The majority of these lethality matches were inherited with AR genetics (73%, 122 of the 166 diseases with reported inheritance from OMIM/Orphanet) and modelled by homozygous mouse mutants, consistent with the conclusion that homozygous loss of function mutations in essential genes in humans produces either early death or severe congenital medical conditions requiring advanced medical support for survival. Examples of mouse and human embryo/early-onset lethality include:
ventriculomegaly with cystic kidney disease that results in in utero or neonatal human
fatality (OMIM:219730, gene: CRB2) and Stuve-Wiedermann Syndrome (OMIM:601559,
gene: LIFR). Diseases that would probably have been lethal without medical support
with a corresponding lethal/subviable mouse strain include: Coach Disease
(OMIM:216360, gene: RPGIP1), Meier Gorlin Syndrome 1 (OMIM:224690, gene:
ORC1), and Human phosphoserine phosphatase deficiency (OMIM:614023, gene: PSPH).
In the latter, a homozygous lethal mouse mutant in Psph
\(^{tm1a(EUCOMM)/Wtsi}\) has structural
abnormalities at Embryo day (E) 15.5 detected by micro-computed tomography (micro-
CT) that closely resemble the developmental and structural defects in the human
phosphoserine phosphatase deficiency patients (Figure 3).

When we include these manually curated lethality matches, 40.5\% (360) disease
models have phenotype overlap with the 889 disease genes (Table 1) with the majority
(78%: 279 of 360) being the first report of a candidate mouse model for these diseases.
The discovery rate of disease models in our analysis is comparable to previous reports
on smaller high-throughput mouse phenotyping studies that found modelling of 46% of
59 and 33% of 42 associations using manual investigation of data\(^{5,27}\). In the 54.4% (484) cases where we did not yet detect a model despite testing for at least one
equivalent phenotype (Supplement Table 2), explanations could range from real
differences in human and mouse biology, the genetic background and a null allele not
being appropriate to model the disease or the differing methodologies used for
annotation e.g. rarely observed phenotypes for a disease are often recorded in the HPO
annotations and would likely fall below the statistical threshold if similarly, non-penetrant
in mice. Finally, 90 of the 889 assessed known disease-gene associations involve the
tm1a allele where there is a small possibility that a hypomorph rather than null was
classified, affecting the ability to recapitulate the disease, as well as a slight chance
of altered expression of genes in cis to the retained neomycin selection cassette leading
to false positive disease models (see Methods). Another 10 involved the KOMP1 allele,
also with the possible neomycin effect. These models were phenotyped based on
prioritisation by centers as it takes a year or more of breeding to remove the floxed
critical exon and neomycin selection cassette.

New Functional Knowledge and Mendelian Disease-Gene Candidates

The second major clinical use case for the IMPC’s data is providing new data on
the phenotypes and functions of genes. IMPC has prioritised genes with no known
disease associations or minimal GO annotation to address this. Based on MGI’s
literature curation of mutant strains involving any allele type except conditional
mutations, 1830 of the 3,328 genes phenotyped in this IMPC release have never had a mouse produced before. No Gene Ontology (GO) molecular function or biological process annotations are available for 189 genes, while another 903 genes had inferred annotations from computational analysis (Figure 4A). The phenotypes of these mutant strains provide substantive insights into the function of a large class of genes (sometimes described as the ignorome) for which there is little or no existing functional information (Supplement Table S5).

Examples of candidate genes for human Mendelian disease with previously little functional information include Family with sequence similarity 53 member B (Fam53b), which had no reported phenotypic variants in human or mouse. The gene is differentially expressed in adult definitive erythrocytes compared to primitive erythrocytes with a >6-fold $\log_2$ change (http://www.ebi.ac.uk/gxa/genes/ENSMUSG00000030956?bs=%7B"mus+musculus"%3Atrue%7D&ds=%7B%7D - differential). Homozygous Fam53b$^{tm1b(EUCOMM)Hmgu}$ knockout mice showed increased mean corpuscular hemoglobin and decreased erythrocyte cell number (Figure 4B,C), suggesting the gene is involved in hematopoiesis and is a candidate for macrocytic hyperchromic anemias. PhenoDigm identified this gene as a phencopy for Diamond-Blackfan Anemia (DBA; OMIM:105650), a group of fifteen unique anemias generally attributable to defects in ribosome synthesis but for which known mutations only account for approximately 54% of all DBA patients. A single functional study suggested that Fam53b is required for Wnt signaling, which is a key step in determining cell fate, cell proliferation, stem cell maintenance and anterior-posterior axis formation. The Fam53b knockouts thus implicate a new candidate pathway to be considered for the 46% of DBA patients where genetic causes are not known.

As well as providing fundamental insights into the function of genes with little 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 resources may be used to detect novel genotype to phenotype associations in disease where simply considering existing human data would lead to causative variants being overlooked among the overwhelmingly abundant associated variants of unknown significance, as happens in many exome sequencing studies: over half of diagnosed rare diseases still have no known causative gene and diagnostic rates in most high-throughput Mendelian disease sequencing projects are 20-30%, largely due to a lack of functional information for most genes. To remedy this and to start achieving better diagnostic rates we can
utilise the data that the IMPC provides. As a demonstration of the potential of IMPC data for novel disease gene discovery, we identified candidate genes for Mendelian diseases with an unknown molecular mechanism but where a broad genetic localization was available in OMIM from previous studies. Our disease matching algorithm identified associations where our predicted disease gene falls within these loci (Supplement Table S6).

For example, adult mice heterozygous for the Klhdc2

Right Ventricular Dysplasia 3 (ARVD3; OMIM:602086) that presents with cardiac arrhythmias caused by fibro-fatty replacement of right ventricle myocardium. Klhdc2 is syntenic with the ARVD3 locus, suggesting it as a candidate gene. Usmg5 null mice recapitulated the clinical symptoms of muscle weakness and abnormal gait seen in patients with the dominant intermediate A form of Charcot-Marie-Tooth disease. The human orthologue (USMG5) is located within a 9.8Mb critical region identified in patients with this disease. While the implicated human loci are sometimes Mbps in length and encompass hundreds of genes, these examples illustrate how IMPC phenotype data allows for the scoring of candidate genes with disease causing variants and has important implications for current rare disease genetic projects that are using next generation sequencing technologies (see Discussion).

Discussion

By analysing phenotype similarities between the IMPC’s mouse strains and clinical features of human disease, we have provided new human disease models and identified novel functional gene information for a significant and growing proportion of protein coding genes. These models are made readily available and can be exploited to further understand disease mechanisms, develop new gene therapy and pharmacological treatments, and further our understanding of gene function. The novelty of the IMPC is both the scale of the vision to produce the first comprehensive catalogue of mammalian gene function across all genes as well as the non-hypothesis driven, standardised approach to the phenotyping. The latter facilitates novel discoveries about the function of genes and their role in disease as already demonstrated above for the potential involvement of Klhdc2, Usmg5 and Fam53b in Arrhythmogenic Right Ventricular Dysplasia 3, Charcot-Marie-Tooth disease and Diamond-Blackfan Anemia respectively.
The IMPC mouse strains illustrate the value of matching mouse mutant phenotypes to human Mendelian syndromes without candidate genes in order to prioritise candidates for follow-up with additional molecular analyses. Our phenotype comparisons are made amenable to clinical researchers performing next-generation sequencing based diagnostics through inclusion within the Exomiser software package that combines an assessment of variant pathogenicity with gene candidacy based on similarity of the patients’ phenotypes to known phenotypic knowledge from human, mouse and fish. Exomiser is being applied within the NIH Undiagnosed Disease Program (UDP) and Network (UDN) as well as the NHS England 100,000 Genomes Project that aims to embed genomics into a national healthcare system for the first time.

The IMPC goes beyond modelling human Mendelian syndromes by leveraging its existing global infrastructure to address complex biological questions. IMPC mouse strains are widely used with over 1300 publications that cite use of IMPC and IKMC resources in research across every major biological system (http://www.mousephenotype.org/data/allelerel) including studies that use IMPC mice and data to help validate SNPs associated with complex traits in both humans and mice. Such studies will be supported by ongoing work using the IMPC phenotyping pipeline to characterise the eight founder inbred mouse strains for the collaborative cross (CC) resource used in the study of complex traits and in targeted non-coding mutant mice strains (11 miRNA, 4 lincRNA in the current data release) to study regulatory activities. Other collaborative multi-centre efforts are using IMPC mouse strains to study gene function in hearing, vision, metabolism, and pervasive sexual dimorphism. Starting this year, a significant fraction (~15%) of mutant strains will be re-phenotyped at 12-18 months of age to identify genes involved with late-onset disease.

In addition, the major change in our strategy has been the adoption of new CRISPR based methods to increase the production rate of IMPC strains as well as providing additional opportunities to characterise mutant mouse strains containing the same single nucleotide variants or indels as human patients. CRISPR technologies allow the co-generation of these precision models alongside an IMPC null allele, providing an economy of scale, as well as tailored production of these specific variants by IMPC centers such as MRC Harwell via the recent establishment of the MRC Genome Editing Mice for Medicine service (GEMM; https://www.har.mrc.ac.uk/gemm-call-guidance-applicants). Production and characterisation of mice modelling several rare disease patient variants identified through whole genome sequencing by the 100,000 Genomes Project has been initiated. Here the goal is to either functionally validate variants of unknown significance (VUS) to produce a diagnosis and/or facilitate...
mechanistic and therapeutic studies in collaboration with the clinicians and researchers associated with this translational, healthcare project. Development of new therapies for rare disease still lags far behind the discovery of new disease genes by next-generation sequencing technologies, and generation of these precision medicine models is key to addressing this issue. Within the next year, candidate non-coding, regulatory variants in undiagnosed 100,000 Genomes Project cases will be selected and become amenable to functional validation in collaboration with the IMPC centers. The IMPC plans similar relationships with other large-scale sequencing projects such as the UDN as well as the US Precision Medicine Initiative. We believe many of the lessons we have learned establishing the IMPC will also be of value to the recently launched precision medicine efforts whose goals are to improve treatment through the customisation of healthcare based on a patient's genomic information and environmental factors. Harmonisation of phenotype 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 causal disease gene variants.

While advances in CRISPR and induced-Pluripotent Stem cells (iPSc) technologies have now vastly expanded the researchers' toolkit, the work of the IMPC highlights the continuing importance of mouse models to understanding disease mechanisms. Mice are vertebrate mammals with physiological characteristics that recapitulate all major human biological systems, allowing study of processes not possible with in vitro studies including the impacts of behavioral, inflammatory, endocrine, and gender-specific processes on disease. While 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 reduce phenotypic variability, with most strains having a 2-year lifespan that allows for comprehensive studies in a timely manner.

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

13
In conclusion, the IMPC has established an ever-expanding knowledgebase of mammalian gene function, a large resource of novel disease models and the capacity for functional validation of variants identified in disease sequencing projects that will be of great value to the human disease community.

METHODS

Mouse Production:

Targeted ES cell clones obtained from the International Knockout Mouse Consortium (IKMC) resource,7,42 were injected into mouse blastocysts for chimera generation. The resulting chimeras were mated to C57BL/6N mice, and the progeny were screened to confirm germline transmission. Following the recovery of germline-transmitting progeny, the majority of strains (82%) were crossed with a coisogenic C57BL/6N transgenic strain bearing a germ-line-expressing Cre recombinase to excise the floxed neomycin selection cassette (neo) and critical exon for EUCOMM alleles and generate a true deletion. For the rest, the requirement early on for establishment and testing of the pipeline without additional breeding meant lines were characterised that contained either tm1a alleles (16%; rely on a stop codon that could potentially be spliced around and retain the neo cassette that can alter transcriptional activity of other genes in cis) or the KOMP1 allele (2%; retain neo cassette). The resulting C57BL/6N heterozygotes were intercrossed to determine viability and generate homozygous mutants. All strains are made accessible from http://www.mousephenotype.org/.

Mouse Phenotyping and Experimental Design:

Based upon previous analysis on appropriate sample sizes to detect significant effects by out statistical framework (see below), a minimum of seven male and seven female mice were characterised from 9 weeks of age until 16 weeks of age using a broad-based phenotyping pipeline that assessed every major biological system. IMPC centers employed a common control strategy where cohorts of age-matched, wild-type C57BL/6N mice are phenotyped in a continuous manner alongside mutant C57BL6/N strains. These cohorts are used in quality control (e.g. baseline drift) and in statistical analysis of the data. A centralised database of consensus IMPC standard operating procedures (SOPs), IMPReSS (http://www.mousephenotype.org/impress), ensured that all phenotyping data and metadata are collected in a reproducible and standardised format. Cohorts of at least seven homozygous mice of each sex per line were generated. If no homozygotes were obtained from 28 or more offspring of heterozygote intercrosses during production, the strain was scored non-viable. Similarly, if less than
13% of the pups resulting from intercrossing were homozygous, the strain was scored subviable. For non-viable and subviable strains, heterozygous mice were committed to the phenotyping pipelines. The individual mouse was considered the experimental unit within the studies.

Data quality control (QC):

Defined criteria were established for QC failures (e.g. insufficient sample, incorrect instrument calibration) and detailed within IMPReSS to provide valid reasons for discarding data. A second QC cycle occurred when data was uploaded from the phenotyping center to the IMPC Data Coordination Centre (DCC) using an internal QC web interface. Data was only QC failed from the dataset if clear technical reasons were identified for a measurement being an outlier and this was tracked within the database.

Wild-type–knockout comparisons:

Wild-type vs null comparisons, i.e. a dataset, were restricted to data collected at one centre and were assembled by selecting data from knockout and wild-type mice that had data collected from the same versioned protocols and with the same metadata parameters (e.g. instrument). As wild-type mice are measured every week, a null strain is generally compared to data from hundreds of wild-type control mice. In the case when all members of a null mouse strain are measured on the same day with an equal number of control mice, the comparison is restricted to this smaller set of data to eliminate batch effects.

A dataset consists of the collection of data values (mutant and control) for a single measured variable (parameter) with the same allele, zygosity, center, and experimental metadata. Using IMPC data release version 5.0, (2nd August 2016), IMPC has analysed 352,729 continuous datasets and 944,270 categorical datasets produced from 10 phenotyping centres. These raw data are available at the IMPC web portal with a page detailing the various methods by which data can be extracted (http://www.mousephenotype.org/data/documentation/index).

Statistical analysis:

Statistical analysis was performed using the PhenStat R package developed for the IMPC. PhenStat is a statistical analysis tool suite that accounts for known variation in experimental workflow and design of phenotyping pipelines. Briefly, categorical data analysis was completed using a Fisher’s Exact test and continuous data analysis was completed with the PhenStat Mixed Model (MM) framework which uses linear mixed models that treat batch as a random effect. Details of the implementation including
decision tree model, descriptions and the lower FDR rates associated with multi-batch data are available in the PhenStat package user’s guide (http://goo.gl/tfbA5k), and described in the literature. For this analysis, results from one batch, low batch (mutants measured in batches between 2-4 times, and multi-batch (5 or greater) experiments were used. For viability and fertility data, the center conducting the experiment used a statistical method appropriate for the breeding scheme utilised at that center (exact details are available on 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 blinding strategies during group allocation, no specific inclusion/exclusion criteria, and no randomisation approach beyond relying on Mendelian inheritance to randomise as detailed in our ARRIVE guidance compliance document (http://www.mousephenotype.org/about-impc/arrive-guidelines). All analysis presented in this paper is based on the binary assignment of a significant deviation (or not) from wild-type and the associated phenotype term. Detailed output of our statistical analysis for each test are presented on our portal pages (mousephenotype.org) including all raw data, the distribution and variance and calculated p-value for the genotype being associated with the phenotype.

Matching Mouse Phenotypes to OMIM and Orphanet Disease Descriptions:

Automated PhenoDigm:

We utilised the Human Phenotype Ontology (HPO) annotations available from the Monarch Initiative (Accessed 2nd September 2016) describing the clinical phenotype features of over 7,000 diseases reported in OMIM (http://omim.org; accessed 7th November, 2015) and Orphanet (http://www.orpha.net; accessed 2nd September 2016). These HPO terms were semantically compared with the phenotype features (MP annotations) of IMPC mouse strains using the PhenoDigm algorithm, developed by us and fellow members of the Monarch Initiative as reflected in authorship, to generate an overall score for how phenotypically similar a mouse strain is to a particular disease. PhenoDigm calculates the individual score for each HPO-MP phenotypic match based on 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 score highest. The geometric mean of the IC and simJ is used to generate the HPO-MP pairwise score. The overall PhenoDigm percentage score is a comparison of the best and mean scores for all the pairwise HPO-MP comparison relative to the maximum possible scores for a...
perfectly matching mouse model to that 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 which maximised precision and recall compared to other similarity thresholds of 1.0, 1.25, 1.5, or 1.75 (Supplementary Figure 1).

Known human genes and regions associated with diseases were extracted from OMIM and Orphanet and matching mouse orthologues were identified from HomoloGene (http://www.ncbi.nlm.nih.gov/homologene). Comparisons to previous mouse mutants from the MGI resource were achieved by download and processing a file named MGI_GenePheno.rpt containing 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 (both available from ftp://ftpinformatics.jax.org/pub/reports/ and downloaded 2nd 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 pre-weaning subviable or lethal, and also had OMIM records, we manually inspected the OMIM records to identify those with reported in utero or early deaths (prior to two-years of age) and coded these in Supplementary Table 1 as Yes-Lethal (Y-L) indicating that for some human cases with mutations for these genes, the phenotype of human lethality matched the phenotype of mouse sub-viability. We also screened for OMIM records with severe congenital defects and/or rapid progression of early onset severe disease in human patients requiring significant 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) indicating a probable match of the human phenotype with 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 (downloaded 2nd September 2016). Ensembl was used to identify the human genes contained within these regions and their mouse orthologues retrieved from HomoloGene. The overlap between these genes and candidates from the PhenoDigm analysis of the same

17
disease were then flagged within our database and are highlighted on both our portal as well as the supplementary tables presented here.

### Identifying Novel Gene-Phenotype Relationships from the IMPC Database:

An online tool on the IMPC portal ([https://www.mousephenotype.org/data/gene2go](https://www.mousephenotype.org/data/gene2go)) imports GO annotations daily from the Quick GO resource ([https://www.ebi.ac.uk/QuickGO](https://www.ebi.ac.uk/QuickGO)) and categorises them based on the evidence codes assigned by GO curators. Annotations were analysed on 24 March 2017. We started with 2668 genes that had IMPC non-lethal phenotypes. Categories incorporate 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), 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 Sequence Model (ISM), 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), Inferred from Reviewed Computational Analysis (RCA)
- **Automated electronic:** Inferred from Electronic Annotation (IEA), Other: Traceable Author Statement (TAS), Non-traceable Author Statement (NAS), Inferred by Curator (IC)
- **No biological data available:** No biological Data available (ND), 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 centers’ ethical animal care and use guidelines in addition to their applicable licensing and accrediting bodies, reflecting the national legislation under which they operate. Details of each centers’ 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 that were disseminated throughout the IMPC. Animal welfare was assessed routinely for all mice.

### Data availability:

Deleted:...

Deleted:...

Deleted:...

Deleted:...

Deleted:...

Formatted: Font: 12 pt

Formatted: Underline
All data presented here is openly available from the IMPC portal (mousephenotype.org) via our FTP site: ftp://ftp.ebi.ac.uk/pub/databases/impc/latest.

We also provide regular data exports to the MGI group (http://www.informatics.jax.org/) who provide public access to all available mouse data and the Monarch Initiative (https://monarchinitiative.org) who integrate genotype-phenotype data from human and numerous other species.

Code availability

The automated phenotype comparisons were performed using the open-source OWLtools package (https://github.com/owlcollab/owltools) provided by the Monarch Initiative (https://monarchinitiative.org).

ACKNOWLEDGEMENTS

This work was supported by NIH grants U54 HG006370 (T.F.M., P.F., A.-M.M., H.E.P., D.S., S.D.M.B.), U42 OD011185 (S.A.M.), U54 HG006332 (R.E.B., K.S.), U54 HG006348-S1 and OD011174 (A.L.B.), 1R24OD011883 (C.J.M., M.H, N.W., D.S.), HG006364-03S1 and U42 OD011175 (C.M, K.C.K.L.), and additional support provided by the The Wellcome Trust, Medical Research Council Strategic Award 53658 (S.W., S.D.M.B.), Government of Canada through Genome Canada and Ontario Genomics (OGI-051),(C.M., S.D.M.B.), Wellcome Trust Strategic Award, 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 frame programme “Investissements d’Avenir” labelled ANR-10-IDEX-0002-02, ANR-10-INBS-07 PHENOMIN to (Y.H.), The German Federal Ministry of Education and Research by Infrafrontier grant 01KX1012 (S.M., V.G.D., H.F., M.H.d.A.) ‘EUCOMM: Tools for Functional Annotation of the Mouse Genome’ (EUCOMMTOLS) project - grant agreement no [FP7-HEALTH-F4-2010-261492] (W.G.W)

AUTHOR CONTRIBUTIONS

K.L.S, R.E.B, S.W, A-M.M, P.F, H.E.P, J.W, A.B, W.C.S, D.J.A, S.D.M.B, W.W, S.N, A.M.F, L.M.J.N are senior PIs of the key programs that contributed to the paper, and were critical for the design, management, and execution of the study and writing and reviewing of manuscript. The additional IMPC consortium members all contributed to data acquisition and data handling.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests
Bibliography


21. Smith, C. L. & Eppig, J. T. Expanding the mammalian phenotype ontology to

22. Check spelling and grammar.

23. Don’t adjust space between Latin and Asian text. Don’t adjust space between Asian text and numbers.
support automated exchange of high throughput mouse phenotyping data

generated by large-scale mouse knockout screens. J. Biomed. Semantics 6, 11

Smedley, D. et al. PhenoDigm: analyzing curated annotations to associate animal


Margolin, D. H. et al. Ataxia, dementia, and hypogonadotropism caused by

Santens, P. et al. RNF216 mutations as a novel cause of autosomal recessive

White, J. K. et al. Genome-wide generation and systematic phenotyping of


(2014).

Petryszak, R. et al. Expression Atlas update—an integrated database of gene and
protein expression in humans, animals and plants. Nucleic Acids Res. 44, 2015
(gkv1045 (2015).


Boria, I. et al. The ribosomal basis of Diamond-Blackfan Anemia: mutation and

Kizil, C. et al. Simplet/Fam53b is required for Wnt signal transduction by

Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with

Bone, W. P. et al. Computational evaluation of exome sequence data using

Harkness, J. H. et al. Trace Amine-Associated Receptor 1 Regulation of

Cade, B. E. et al. Obstructive Sleep Apnea Traits in Hispanic/Latino Americans.

Knowles, J. W. et al. Identification and validation of N-acetyltransferase 2 as an

Lang, B. et al. Recurrent deletions of ULK4 in schizophrenia: a gene crucial for

McIntyre, R. E. et al. A Genome-Wide Association Study for Regulators of

Levy, R. et al. Collaborative cross mice in a genetic association study reveal new

Ringwald, M. et al. The IKMC web portal: A central point of entry to data and
resources from the International Knockout Mouse Consortium. Nucleic Acids Res.
39, 2011).

Karp, N. A., Melvin, D., Sanger Mouse Genetics Project & Mott, R. F. Robust and

Seyers, E. W. et al. Database resources of the National Center for Biotechnology

TABLES

Table 1. Frequency of IMPC Models that correspond to Mendelian Disease-Gene Associations in OMIM or Orphanet.

650 known rare disease-gene associations covered by OMIM and Orphanet have a phenotyped IMPC strain involving the orthologous mouse gene. The PhenoDigm automated pipeline and manual curation approaches identified matching phenotypes between mouse strains and human disease. A correspondence between the human disease and mouse model was defined when at least one of the human clinical phenotypes was recapitulated by the IMPC line. Novel models were defined when MGI contains no curated mouse line or literature asserted disease model for the gene. The manual lethality matching category corresponds to IMPC mutant strains for which homozygosity produced embryo or neonatal lethality/subviability and matched reports of human lethality/subviability in the OMIM/Orphanet summaries (see methods).

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>Deleted:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated IMPC Disease Model (novel only)</td>
<td>134/889 (15.1%)</td>
<td>88/650 (13.5%)</td>
</tr>
<tr>
<td>Automated IMPC Disease Models (all)</td>
<td>185/889 (20.8%)</td>
<td>129/650 (19.8%)</td>
</tr>
<tr>
<td>Additional Manual Lethality IMPC Disease Models (all)</td>
<td>175/889 (19.7%)</td>
<td>133/650 (20.4%)</td>
</tr>
<tr>
<td>Total IMPC Disease Models (all)</td>
<td>360/889 (40.5%)</td>
<td>262/650 (40.2%)</td>
</tr>
</tbody>
</table>
### Table 2. Examples of IMPC Disease Models Across Diverse Biological Systems

<table>
<thead>
<tr>
<th>Biological system</th>
<th>Disease Gene</th>
<th>Human Mendelian disease</th>
<th>Relevant Human Phenotype</th>
<th>Overlapping Mouse phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bone</td>
<td>SCARF2</td>
<td>Van Den Ende-Gupta Syndrome</td>
<td>Long metacarpals</td>
<td>increased length of long bones</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>LMNA</td>
<td>Cardiomyopathy Dilated 1a</td>
<td>Dilated cardiomyopathy</td>
<td>increased heart weight</td>
</tr>
<tr>
<td>Craniofacial</td>
<td>MSX1</td>
<td>Orofacial Cleft 5</td>
<td>Cleft palate</td>
<td>Cleft palate</td>
</tr>
<tr>
<td>Embryo</td>
<td>PSPH</td>
<td>Phosphoserine Phosphatase Deficiency</td>
<td>Intrauterine growth retardation</td>
<td>abnormal embryo size</td>
</tr>
<tr>
<td>Growth/Body size</td>
<td>GHRHR</td>
<td>Isolated Growth Hormone Deficiency, Type Ib</td>
<td>Short stature</td>
<td>decreased body length</td>
</tr>
<tr>
<td>Hearing</td>
<td>SLC52A2</td>
<td>Brown-Vialetto-Van Laere Syndrome 2</td>
<td>Sensorineural hearing impairment</td>
<td>increased or absent threshold for auditory brainstem response</td>
</tr>
<tr>
<td>Hematopoietic</td>
<td>GP9</td>
<td>Bernard-Soulier Syndrome</td>
<td>Thrombocytopenia</td>
<td>Thrombocytopenia</td>
</tr>
<tr>
<td>Metabolism</td>
<td>KCNJ11</td>
<td>Diabetes Mellitus, Noninsulin-Dependent</td>
<td>Type II diabetes mellitus</td>
<td>impaired glucose tolerance</td>
</tr>
<tr>
<td>Muscle</td>
<td>COL6A2</td>
<td>Bethlem Myopathy</td>
<td>Distal muscle weakness</td>
<td>Decreased grip strength</td>
</tr>
<tr>
<td>Neurological</td>
<td>GOSR2</td>
<td>Epilepsy, Progressive Myoclonic, 6</td>
<td>Difficulty walking</td>
<td>abnormal gait</td>
</tr>
<tr>
<td>Reproductive System</td>
<td>RNF216</td>
<td>Gordon Holmes Syndrome</td>
<td>Infertility</td>
<td>male infertility</td>
</tr>
<tr>
<td>Retina</td>
<td>BBS5</td>
<td>Bardet-Biedl Syndrome 5</td>
<td>Rod-cone dystrophy</td>
<td>abnormal retina morphology</td>
</tr>
</tbody>
</table>
Figure 1. IMPC Mutant Models of Human Disease and Gene Function

Human disease models were identified by measuring the degree of phenotype similarity between IMPC null mutant mouse strains and their orthologous human disease genetic loci. **Models of Mendelian Disease**- of 889 potential disease models, 360 mutant strains had both phenotype overlap and an orthologous null allele to diseases with known mutations as described in OMIM and Orphanet; **Novel Mendelian Disease Candidates**- 135 strains had phenotype overlap and null alleles syntenic to linkage or cytogenetic regions associated with human diseases with unknown molecular mechanisms; **New Functional Knowledge**- of 2564 genes with a non-lethal IMPC phenotype, IMPC data provide the first functional experimental evidence for 1092 of these genes based on Gene Ontology Annotation.
**Figure 2. New Mouse Models for Mendelian Human Disease**

**Gp9**- Bernard-Soulier syndromes are bleeding disorders that result from mutations in the glycoprotein Ib (GP Ib) platelet membrane receptor complex. *Gp9<sup>-/-</sup>* Bernard-Soulier syndrome are bleeding disorders that result from mutations in the glycoprotein Ib (GP Ib) platelet membrane receptor complex. *Gp9<sup>tm1.1(KOMP)Vlcg</sup>* homozygotes have abnormal platelet development as demonstrated by increased platelet volume (A) and decreased platelet numbers (B). *Bbs5* - *BBS5* is associated with Bardet-Biedl syndrome (BBS), a ciliopathy with multisystem involvement with severe and early-onset of symptoms. *Bbs5<sup>tm1b(EUCOMM)Wtsi</sup>* homozygous mice display profoundly increased body fat percentage (C), impaired glucose tolerance (D) and size (E). *Rnf216* - Gordon Holmes syndrome is associated with *RNF216* and is characterised by hypogonadism and cerebellar ataxia. *Rnf216<sup>tm1b(EUCOMM)Wtsi</sup>* homozygous null male mice are infertile. Histopathology identified seminiferous tubule degeneration and atrophy with Leydig cell hyperplasia (Fa, 20x magnification) and epididymal aspermia (Fb, 20x magnification) in null mice compared to unaffected seminiferous tubules (Fc, 20x magnification) and epididymis (Fd, 20x magnification) in control mice.
Psph. Phosphoserine phosphatase deficiency (OMIM: 614023) is an autosomal recessive disorder characterised by prenatal and postnatal growth retardation, psychomotor retardation and facial dysmorphologies with the severity of the symptoms requiring medical support for survival. Complete preweaning lethality was observed in Psph<sup>tm1.1(KOMP)V</sup>lcg homozygous null mice (A) (# of pups, asterisks indicate no surviving homozygotes). LacZ reporter expression regulated by the Psph promoter in asymptomatic heterozygous E12.5 embryos shows extensive gene expression (bar 1mm) (B). Gross images of E15.5 homozygous mutant embryos confirmed growth retardation, haemorrhage, and facial dysmorphologies (bar 5mm) (C). Imaging of E15.5 embryos by microCT showed significant growth retardation, as well as facial dysmorphologies consistent with the human Mendelian disorder (D).
Figure 4. Novel Mouse Models of Disease – Over 40% of IMPC strains are for genes that lack experimental evidence for function according to the Gene Ontology Consortium (A in grey). Fam53b. Fam53b<sup>tm1b(EUCOMM)Hmgu</sup> homozygous mutant mice had significantly decreased red blood cell counts (B), and enlarged erythrocytes (C), consistent with Diamond-Blackfan Anemia (DBA, OMIM: 105560). Dnajc5b. Dnajc5b<sup>tm1b(EUCOMM)Hmgu</sup> homozygous mutants displayed significantly shortened QT interval as measured by electrocardiogram (D), supporting a role for DNAJC5b variants associated with human variability to statin effects on cardiovascular incident frequency.

Data are presented as means ± standard deviation, asterisks indicate a significant difference between mutant and same sex controls using the mixed model with a p < 0.00001. Sample sizes for Fam53b mutants were 8 for both male and female groups, and >500 for same sex controls and for Dnajc5b mutants were 8 and 9 respectively compared to 8 for same sex controls measured on the same day.
Table S1. Reproducibility of 2547 MGI curated gene-phenotype associations that have also been assessed by the IMPC

Table S2. Comparison of human Mendelian disease caused by known gene mutations with targeted null mice

Table S3. Summary of phenotypes for human Mendelian disease mapping to mouse mutations with adult mutant phenotypes

Table S4. Manual curation of human disease and mouse phenotypes for 100 genes

Table S5. Mutant mouse gene IDs with phenotypes having no or minimal Gene Ontology Annotations

Table S6. Candidate genes for genetically mapped human Mendelian disease

Table S7. Contributing institute animal welfare approvals
Supplementary Figure 1. Precision and recall of the automated PhenoDigm algorithm.

IMPC mouse strains were ranked by the PhenoDigm algorithm for diseases involved in 650 known gene associations from OMIM and Orphanet then the precision and recall of the lines involving the orthologue of the known disease gene were plotted. Performance was measured where models were excluded below a threshold of either 1.0, 1.25, 1.35, 1.5 or 1.75 for the geometric mean of the information content and Jaccard index of the best phenotype match.
Novel complex disease-gene candidates

Over the past decade many genetic variants have been associated with complex, non-Mendelian, human disease by Genome Wide Association Studies (GWAS)\(^{36,37}\), with data archived at the NHGRI-EBI GWAS Catalog (http://www.ebi.ac.uk/gwas/). Findings from GWAS link a complex disease phenotype with DNA sequence variation near or within genes, but often a paucity of functional information precludes disease gene identification. The IMPC phenotyping pipeline may provide functional information to support GWAS associations and develop hypotheses for follow-up studies. In order to evaluate the potential of the IMPC mouse phenotype data for validation of human GWAS findings, we mapped the 2432 mouse genes with abnormal phenotypes to human orthologues in Ensembl release 84\(^{38}\), and identified a subset of 326 orthologous genes with significant GWAS associations. For these 326 genes, we manually compared the human and mouse phenotypes and identified 59 possible Human GWAS association/mouse phenotype matches, or 18% of the evaluated GWAS findings (Supplement Table S6).

Several interesting examples illustrate how the mutant mouse phenotype may inform GWAS associations. DNAJC5B encodes a DnaJ/Hsp40 family member protein with no known function, though other protein family members are involved in the co-chaperoning of heat-shock proteins to specific sub-cellular compartments\(^{39}\). DNAJC5B has been associated with differential coronary heart disease event reduction by statin therapy in three separate GWAS heart health studies\(^{40}\). Dnajc5b\(^{tm1b(EUCOMM)Hmgu}\) homozygous null mice have a significantly shortened QT period which represents the time taken for ventricular depolarisation and repolarization, suggesting a mechanism by which allelic variants may contribute to cardiovascular health (Figure 4D). Human QT syndromes are abnormal heart rhythm conditions associated with functional disruption of the potassium channel human ether-a-go-go related (hERG) gene. Other members of the DNAJ/HSP40 family, DJA1 and DJA2, have been shown to be key modulators of hERG degradation regulation, consistent with a potential role for DNAJC5b in cardiac event reduction\(^{41}\).

ATPase, Ca++ transporting, cardiac muscle, slow twitch 2 (ATP2A2) codes for a sarcoplasmic/endoplasmic reticulum Ca\(^{++}\)-ATPase and is expressed in muscle
and brain. A GWAS association between schizophrenia and two SNPs located within the \(ATP2A2\) gene has been reported\(^{42}\), and the gene is associated with Darrier-White disease (OMIM:124200), a dominantly inherited skin disorder that also has a high incidence of psychoses and affective disorders\(^{43}\). \(Atp2a2^{tm1b(EUCOM)Hmgu}\) heterozygous mice had altered prepulse inhibition (female decreased \(p < 1.6 \times 10^{-4}\); male increased \(p < 2.3 \times 10^{-6}\)), a behavioural phenotype that is a hallmark of human schizophrenic disease\(^{44}\), supporting the GWAS linkage findings for schizophrenia.

Another informative example is \(NOS1AP\) (nitric oxide synthase 1 [neuronal] adaptor protein), linked to ECG abnormalities and QT length in two separate GWAS studies\(^{45,46}\). \(NOS1AP\) is essential for regulating intracellular Ca\(^{++}\) in cardiac muscle\(^{47}\). Consistent with the GWAS ECG associations, the homozygous \(Nos1ap^{tm1b(KOMP)Wtsi}\) null mice had increased heart weight (\(p<8.5 \times 10^{-9}\)); and increased left ventricular wall thickness (\(p< 6 \times 10^{-6}\)) which are both indicators of heart disease.

Matching Candidate Gene Phenotypes to Human Traits from Genome Wide Association Studies (GWAS):

The 2042 genes that had IMPC non-lethal phenotypes (IMPC data release 4.2) were mapped to human orthologues using Ensembl (Release 83)\(^{38}\). The GWAS oracle database was queried to retrieve phenotype information at the given human gene locus (GWAS version from February 2016) and phenotype comparisons were manually reviewed.