The Deep Genome Project

K. C. Kent Lloyd1, David J. Adams2, Gareth Baynam3,4,5, Arthur L. Beaudet6, Fatima Bosch7, Kym M. Boycott8, Robert E. Braun9, Mark Caulfield10, Ronald Cohn11, Mary E. Dickinson6,12, Michael S. Dobbie13, Ann M. Flenniken14, Paul Flicek15, Sanjeev Galande16, Xiang Gao17, Anne Grobler18, Jason D. Heaney6, Yann Hureau19, Martin Hrabě de Angelis20,21,22, James R. Lupski6, Stanislas Lyonnet23, Ann-Marie Mallon24, Fabio Mammano25, Calum A. MacRae26, Roderick McIntness27, Colin McKerlie11,28, Terrence F. Meehan15, Stephen A. Murray9, Lauryl M. J. Nutter29, Yuichi Obata29, Helen Parkinson16, Michael S. Pepper30, Radislav Sedlacek31, Je Kyung Seong32, Toshihiko Shiroishi29, Damian Smedley33, Glauco Tocchini-Valentini25, David Valle34, Chi-Kuang Leo Wang35, Sara Wells24, Jacqueline White9, Wolfgang Wurst36,37,38, Ying Xu29 and Steve D. M. Brown24

Introduction

In vivo research is critical to the functional dissection of multi-organ systems and whole organism physiology, and the laboratory mouse remains a quintessential animal model for studying mammalian, especially human, pathobiology. Enabled by technological innovations in genome sequencing, mutagenesis and genome editing, phenotype analyses, and bioinformatics, in vivo analysis of gene function and dysfunction in the mouse has delivered new understanding of the mechanisms of disease and accelerated medical advances. However, many significant hurdles have limited the elucidation of mechanisms underlying both rare and complex, multifactorial diseases, leaving significant gaps in our scientific knowledge. Future progress in developing a functionally annotated genome map depends upon studies in model organisms, not least the mouse. Further, recent advances in genetic manipulation and in vivo, in vitro, and in silico phenotyping technologies in the mouse make annotation of the vast majority of functional elements within the mammalian genome feasible. The implementation of a Deep Genome Project—to deliver the functional biological annotation of all human orthologous genomic elements in mice—is an essential and executable strategy to transform our understanding of genetic and genomic variation in human health and disease that will catalyze delivery of the promised benefits of genomic medicine to children and adults around the world.

Rationale

A comprehensive understanding of genetics, at the single locus, gene, and genomic level, and the pathophysiological consequences of gene variation resulting in gene, RNA, or protein dysfunction are crucial to meeting societal expectations of precision medicine and critical to optimizing clinical practice. With over 80% conserved synteny and a high degree of gene orthology, the mouse and human genomes have provided a unique opportunity for comparative functional analysis and the use of genetically altered mice to interrogate the pathobiology of human disease [1]. For example, of the 6000–8000 rare genetic diseases cited by the rare disease community, the genetic basis is known for between 5000 and 6000 [2], many of which were revealed and/or confirmed by studying the causative genetic variants in mice. Nevertheless, when viewed from a genotype perspective, more than 75 to 80% of the computationally annotated ~20,000 genes in the human genome have not had variation in them tied to any specific phenotype [3].

CRISPR/Cas9 has enabled rapid and highly efficient targeted mutagenesis of the mouse genome. Concurrent development of in vivo analytical and imaging technologies has transformed high-throughput pipelines for precise and reproducible phenotyping of mouse mutants [4]. Deep phenotyping of virtually all body systems including cardiovascular, digestive, endocrine, immune, integumentary, lymphatic, muscular, neurological, sensory, reproductive, respiratory, skeletal, and urinary systems is possible. Further, online computational resources such as the MON-ARCH Initiative (www.monarchinitiative.org) that use controlled vocabularies to integrate numeric, text, and image biological information from heterogeneous datasets (e.g., MGI, OMIM, Orphanet) link genotype to phenotype and enable comparisons between mouse and human ontologies.
These and other advances have facilitated the coordination of industrial scale mutant mouse production and phenotyping at costs far less than previously imagined making functional annotation of all human orthologous genomic elements in mice an achievable scientific goal.

Realization of this goal is the foundation of the work of the International Mouse Phenotyping Consortium (IMPC). The IMPC is a coordinated program of 20 research laboratories in 12 countries on 5 continents dedicated to the design, production, and description of the function of human gene orthologs in the mouse genome (www.mousephenotype.org). The magnitude of this global effort reflects the spirit and scale of the Human Genome Sequencing Project. The IMPC uses homologous recombination in embryonic stem (ES) cells and CRISPR/Cas9 technology to create mutants for genes in the mouse genome followed by whole organism phenotyping of female and male cohorts of adult mice and embryos [6]. The focus thus far has been on the production and phenotyping of null protein-coding alleles in the mouse genome [7] recognizing that such resources serve as the fundamental baseline for mammalian gene function upon which the generation and study of allelic series of other mutations—hypomorphic, neomorphic, antimorphic, and hypermorphic—will prosper and will deliver further insights into gene-phenotype relationships.

Gene association with a broad diversity of human diseases, including hearing loss, ocular diseases, metabolic disorders, bone pathologies, developmental abnormalities, and others, differentiated by sex, has been revealed through IMPC-led discovery research and IMPC-fueled studies by the broader scientific community (https://www.mousephenotype.org/data/publications). This work continues on an industrial scale generating novel insights into gene-based disease phenotypes and other scientific domains such as conservation and ecology. This combined mouse production, phenotyping, and informatics approach has recently been applied to ~1/3 of known Mendelian disease genes and detected significant phenotypic similarities between human disease genes and mouse knockouts (i.e., null alleles) of the orthologs for approximately half the genes [8]. At least one clinical phenotype per disease was tested for the majority (95%) of the genes and matches detected across the whole range of body systems [9].

Currently, IMPC has generated null mutations for nearly 9000 genes of which over 6000 have been phenotyped. By July 2021, IMPC will complete comprehensive phenotypic annotation for over 9000 genes, representing about half of the ~18,000 human orthologs in the mouse (Fig. 1). In its 10-year strategic plan for 2021–2030, the IMPC calls for expanding mouse modeling studies to inform precise molecular diagnostics and targeted therapeutics for Mendelian and multifactorial disorders to maximize beneficial impacts on human health (https://www.mousephenotypetest.org/about-impc/).

If revealing the full biological role of every gene is not daunting enough, the pathobiological effects of individual human genetic and genomic variation further escalate the challenge. As exome sequencing (ES), clinical exome sequencing (cES), and whole genome sequencing [5].
considered "much of the human genome remains unexplored and sequences of mutations already exist [11]. As a result, knowledge of biological function and pathological consequences of genes and genetic variations to human development and disease. For instance, the majority of published gene to phenotype studies continue to focus on genes that are well-annotated or for which knowledge of phenotypic heterogeneity and its potentially underlying genetic bases (e.g., locus and allelic heterogeneity, multi-locus variation, modifier loci), as well as our understanding of multiple disease phenotypes converging on a single gene locus, age-dependent penetrance, and variable expressivity, are all limited. This lack of in vivo functional annotation in experimental models contributes to long diagnostic odysseys and is a significant impediment to the development of molecular entities targeted at specific gene products [14].

The significant progress that has been made in deciphering the genetic basis of rare monogenic diseases represents the low-hanging fruit of gene to phenotype relationships. Progress in elucidating complex multi-allelic and multi-locus relationships and the consequences of de novo mutations in complex disorders is also dependent on genome-wide functional descriptions and could be further explored by the development of multi-allelic [15] and multi-locus [16] models. Mouse models have already provided significant insight into complex diseases such as juvenile diabetes [17] and autism spectrum disorders [18]. Non-mouse models have and continue to contribute immensely to this effort, but definitive identification of causal relationships between mutant alleles and human diseases will often require a mammalian model. The molecular, cellular, and physiologic insights gained from mouse studies are critical to directly inform the early recognition of predictive biomarkers before clinical symptoms manifest and to drive the identification and validation of the new therapeutic targets essential for precision medicine.

**Moving forward**
We postulate that four steps undertaken by the collective endeavors of the global community will be needed to drive progress in genomic and precision medicine. These four steps, allied to genome-wide goals for in vitro systems and other model organisms, will deliver a deeper and more comprehensive understanding of individual gene function, make biological resources and data
available to experimentally decipher disease mechanisms, interpret genomic variation, and reduce the diagnostic odysseys of patients with variants of unknown significance. The IMPC’s strategy 2021–2030 (https://www.mousephenotype.org/wp-content/uploads/2019/05/IMPC_Strategy_2021-30.pdf) is also formulated around these four steps.

1) Complete functional annotation of the protein-coding genome. Complete loss-of-function mutations are essential for identifying the phenotypic impact of protein-coding genes and a necessary first step to interpreting clinically relevant human genetic variation causing disease. By 2021, at the conclusion of the IMPC’s current mandate, ~9000 human orthologous genes in mice will remain to be analyzed by the consortium. Stopping at this point, halfway through the genome, would be equivalent to the Human Genome Project halting its sequencing effort after assembling euchromatic sequence of just 11 chromosomes. It will be vital to continue efforts toward the completion of the functional analysis of the remaining unannotated protein-coding genes using mouse models.

2) Establish functional evaluation of the noncoding genome. The entire coding region is only 3–5% of the mammalian genome. The remaining 95% of the genome plays many roles across a variety of biological processes, including DNA replication, transcriptional regulation, and genomic structure. Of particular note, variation in enhancers, silencers, promoters, and insulators can have significant impact on both normal and abnormal gene expression [19] and gene dosage phenomena [20]. Strategies must be implemented for the prioritization and modeling of mutations of conserved noncoding elements in order to fully explore the in vivo function of the darkest part of the genome.

3) Translate functional biological knowledge to clinical knowledge. The emerging field of genomic and precision medicine relies on the ability to interpret the potential pathophysiological consequences of genetic variation in patients. Beyond academic and research considerations, the long-term financial investments in translating human genetic variation to functional phenotypes and disease mechanisms is enormous and growing. Global investments from government agencies and corporate investors are...
predicted to nearly triple from US$79billion to over US$200billion in 10 years (https://bisresearch.com/industry-report/precision-medicine-market.html).

For example, programs from the US All of Us Project, UK Biobank, the UK 100K Genome Project, the Chinese Precision Medicine Initiative, and many others are enrolling volunteers in efforts to gather personal history, clinical information, genome sequences, and environmental metadata to understand the role of genes in health and disease in order to identify new targets for molecular therapies. For these investments to deliver on their promise to improve health outcomes from molecular diagnosis to management and therapeutic intervention, it will be necessary to integrate gene function data generated from the study of mouse and other animal models into clinical databases, such as ClinVar, ClinGen, and others. Continued development of model organism databases, along with improvements in data integration and analysis, is needed to enable mechanistic insight into genetic variation and disease and support future developments in genomic and precision medicine.

4) Enable rapid functional assessment of genomic variation and integrate functional testing into the clinical decision-making process. While statistical inference of human patient data is currently used to discriminate disease causing from benign associated variants, a definitive molecular diagnosis is often not attainable. Even in those cases where this approach is sufficient, the delay in diagnosis is costly—psychologically, socially, and economically. It will be necessary to undertake programs for the rapid creation and analysis of mouse models of human coding variants along with more efficient approaches to phenotyping. These models, along with other mutational variants, will inform diagnostic decisions and targeted treatments. With these programs in place, clinicians and their research colleagues could rely on mouse models as diagnostic and therapeutic testing platforms, examine the pathological significance of a genetic variant in an orthologous mammalian system, interrogate gene/phenotype relationships for different types of alleles (e.g., SNV, etc.), explore potential gene/environment effects, and access comprehensive datasets to help guide clinical decision-making. In turn, mouse genetic experts will need to respond quickly with targeted phenotyping of mouse models in order for them to achieve clinical utility. It will be necessary to optimize funding and bring scientific insights from mouse functional data to inform the application of mouse models as human patient avatars, make the knowledge gained from mouse data available via the electronic medical record, and enhance education and training of clinicians in genetics and genomic medicine.

Conclusion
Despite the incredible scientific advances in genetics at the single gene and genomic level, the collective biomedical community has only begun to scratch the surface of knowledge about the diverse and varied in vivo pathobiological roles of functional elements throughout the human genome. The global mouse genetics community is primed to address the grand challenges that we face to fully comprehend the role of genes and genetics in development, biological homeostasis, systems biology, disease, and medicine, and is ready to launch a new era for the systematic study of the function of the mammalian genome. The time is right to embark on a Deep Genome Project, on the scale of the Human Genome Project, that will fundamentally enhance the knowledgebase across the biomedical sciences. Driven by the enormous potential of mouse genetics and allied to developments in other model organisms and in vitro approaches, this project will be transformative for biology, medicine, and global health.

Authors’ contributions
KCKL and SDBM co-drafted the editorial; KCKL, SDBM, DJA, GB, ALB, FR, KMB, RRB, MC, RC, MED, MSD, AAWF, PF, SG, XG, AG, IDH, YH, MHA, IRL, SL, AMAM, FM, CAM, RM, CM, TFM, SAM, LMJN, YO, HP, MSP, RS, JKS, TS, DS, GTV, DW, CXLW, SW, JW, WW, and YX read, re-drafted, and approved the final version.

Competing interests
PF is a member of the Scientific Advisory Boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. The other authors declare that they have no competing interests.

Author details
1Department of Surgery, School of Medicine, and Mouse Biology Program, University of California, Davis, CA 95618, USA. 2Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. 3Western Australian Register of Developmental Anomalies and Genetic Services of Western Australia, Department of Health, Government of Western Australia, Perth, Australia. 4Division of Paediatrics and Telethon Kids Institute, Faculty of Health and Medical Sciences, University of Western Australia, Perth, Australia. 5Faculty of Science and Engineering, School of Spatial Sciences, Curtin University, Perth, Australia. 6Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. 7Center of Animal Biotechnology and Gene Therapy, Universitat Autònoma Barcelona, Barcelona, Spain. 8Children’s Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON K1H 8L1, Canada. 9The Jackson Laboratory, Bar Harbor, ME 04609, USA. 10Genomics England, William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK. 11The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada. 12Departments of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX 77030, USA. 13Phenomics Australia, The Australian National University, 131 Garan Road, Acton, ACT 2601, Australia. 14The Centre for Phenogenomics, Lunenfeld-Tanenbaum Research Institute, Toronto, ON MST 3H7, Canada. 15European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. 16National Facility for Gene Function in Health and Disease, Department of Biology, Indian Institute of Science, Education and Research (IISER) Pune, Pune, Maharashtra 411008, India. 17SKL of Pharmaceutical Biotechnology and Model Animal Research Center, Collaborative Innovation Center for Genetics and Development, Nanjing Biomedical Research Institute, Nanjing University,
