DRAFT

The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution

AVE Alliance Founding Members*

*In the spirit of inclusiveness, this manuscript has a sole attribution: the members of the community who came together to found the Alliance. To fairly give credit, and for the purposes of PubMed, we list the following lead authors and co-authors. We also list AVE Alliance contributing authors, who provided substantial comments and edits.

Lead authors: Douglas M. Fowler[1,2,3] and Matthew Hurles[4]

Co-authors: David J. Adams[4], Anna L. Gloyn[5], William C. Hahn[6,8], Debora S. Marks[7,8], James T. Neal[8], Fritz Roth[9,10], Alan F. Rubin[11,12], Lea M. Starita[1,2,3], Jochen Weile[9,10]

AVE Alliance contributing authors: see Supplementary Table 1

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA
[2]Department of Bioengineering, University of Washington, Seattle, WA, USA
[3]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA
[4]Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK
[5]Division of Endocrinology, Department of Pediatrics, Stanford School of Medicine, Stanford University, CA, US
[6]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[7]Department of Systems Biology, Harvard Medical School
[8]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[9]Donnelly Centre and Departments of Molecular Genetics and Computer Science, University of Toronto, Toronto, ON, Canada
[10]Lunenfeld−Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada
[11]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
[12]Department of Medical Biology, University of Melbourne, Melbourne, VIC, Australia

Table of Contents:

Executive Summary

The Atlas of Variant Effects (AVE) Alliance is a new organization aimed at the systematic, massively-scaled measurement of the biological effect of genetic variation in human, model organism and pathogen genomes to generate an atlas of thousands of variant effect maps for individual genes and associated regulatory elements. The atlas will be transformative, improving the fundamental understanding of biological mechanisms, empowering drug development and advancing clinical diagnosis and disease management.

The Alliance will serve as a key forum to catalyse the assembly of this atlas by expanding the community of scientists who are developing and applying experimental and analytical technologies to generate variant effect maps, by supporting the community deploy resources wisely, and by establishing standards to ensure data are generated, managed and used responsibly and with maximum impact.

This document articulates our vision to generate this atlas of variant effects. We summarise the potential impacts that would be enabled by the atlas. We describe the current state of the field and make a set of recommendations to address key challenges. Finally, we set out the values and organising principles of the AVE Alliance, and invite interested academic and industry researchers, clinicians, funders and others to join.

Mission Statement

The AVE Alliance aims to accelerate systematic and massively-scaled measurement and analysis of the impact of genetic variants on functional elements of human, model organism and pathogen genomes to further the understanding of genes, gene products and their regulation and empower the diagnosis and treatment of human disease.

Vision

The problem. Two decades after publishing the sequence of the first human genome, one million human exomes and genomes will soon have been sequenced. Interpreting the effects of the hundreds of millions of variants thus discovered has become a central challenge for genomics. The genomes of the 7.8 billion people alive today collectively contain essentially all of the ~9 billion possible single nucleotide genetic variants compatible with life[1]. Moreover, within the trillions of cells of an individual, it is likely that every possible single nucleotide genetic variant has arisen through somatic mutation. Thus far, the functional impact of genetic variants has primarily been determined by asking if the variant correlates with having a disease, disorder or other trait. With this approach, scientific and clinical communities have collectively characterised the functional impact of less than 1% of genetic variation in the 1-2% of our DNA that is best understood - the genes that encode proteins. For non-coding variation, the situation is almost certainly worse, given that our knowledge about the locations of non-coding functional elements is more recent and that these elements are arguably more complex given their dependence on the context of cell type, developmental stage and perturbation[2].

Despite our increasingly comprehensive catalog of functional DNA elements in the human genome, including enhancers, splice isoforms, and splice-regulatory elements, the catalog remains incomplete. Moreover, merely knowing the locations of these elements does not reveal which variants within them will alter their function. Most variants have no deleterious effect, but, even when a variant in a well-annotated functional element is known to increase disease risk, the mechanism by which it does so often remains cryptic. Fundamentally, our inability to interpret variants is the major factor limiting the effectiveness of genome sequencing for diagnosing genetic disease, improving medical management and understanding genome function. For example, this limitation is observed in the clinical interpretation of variants implicated in Mendelian disorders when insufficient information about the variant leads to its interpretation as a variant of uncertain significance (VUS). VUS pose a substantial challenge for patients and healthcare providers because an accurate genetic diagnosis is the cornerstone of good clinical care.

A solution. Functional assessment using *in vitro* or cell-based assays can provide strong evidence to interpret the biological impact of variants, and, in principle, can be applied to any variant. However, owing to the resource- and time-intensive nature of such assays, they have generally been undertaken reactively for individual variants, only after, and in most cases long after, the first observation of the variant in an organism. Given the rapid growth in newly discovered variants, we propose to build a systematic, extensive atlas of the impact of variants for most human genes and their attendant regulatory elements. This proactive approach has been enabled by the latest generation of multiplexed assays of variant effect (MAVEs), which profile the biological impact of massive numbers of variants in a single experiment, generating variant effect maps for these genes and regulatory elements[3]. MAVEs can be used to map the effects of many different types of variants including single and multiple nucleotide substitutions, small insertions and deletions and larger events. Using MAVEs, variant functional evidence can be assembled in advance of the discovery of a variant in a patient sample, aiding both rapid biological and diagnostic interpretation.

We anticipate that, when appropriately deployed and coordinated, MAVEs can be harnessed to generate a fundamental and invaluable atlas of variant effects, accessible to all. By describing the effects of nearly every possible single nucleotide variant (SNV), as well as other types of variants, in most genes and regulatory elements in the genome, this atlas will accelerate and empower biological research, drug discovery and medical practice. Building a coherent atlas of variant effects will necessarily be a collective endeavour, drawing together diverse expertise from different communities, including patients, patient advocates, researchers, clinicians, diagnostics companies and drug developers.

Now, thirty years after the formal launch of the international Human Genome Project, and two decades after the launch of the SNP Consortium's effort to generate a public resource of human genetic variation genome-wide, we launch the Atlas of Variant Effects (AVE) alliance, an international collaborative effort that will:

- develop, disseminate and democratize the foundational experimental technologies and computational methods for mapping variant effects,
- facilitate, support and motivate the generation of high quality variant effect maps by a diverse community contributing domain-specific expertise,
- coordinate efforts to apply technologies to generate high quality variant effect maps at scale,
- coordinate efforts to develop computational approaches that leverage variant effect data to model the genotype to phenotype relationship,
- establish standards, infrastructure and partnerships for storing, validating and disseminating the data for maximum impact,
- build partnerships with aligned initiatives (e.g. GA4GH, ClinGen, ICDA), and expand the community of stakeholders including patients, patient advocates, researchers, analysts, clinicians, patients, companies and funders

The ultimate goal to develop a systematic, extensive understanding of the functional impact of variants in human, model organism and pathogen genomes will take many years, however, the field of variant effect mapping is already yielding major impact, and we anticipate that a coordinated and phased approach could be transformative even within the next 3-5 years. The initial efforts of members of the AVE Alliance will focus on quantifying the functional impact of single nucleotide variants at the genomic loci, primarily human protein-coding genes, that offer the greatest and most immediate clinical utility. A first major impact of the atlas of variant effects is expected to be improved genetic diagnosis and precision medicine.

While an initial focus of the Alliance is on clinically relevant human variation, the same technologies and infrastructure are already being deployed to characterise variant effects in human genes unrelated to disease and in other species. Thus, the Alliance will ultimately empower fundamental understanding in many species, including model organisms and human pathogens.

Finally, the Alliance will partner and synergize with other major genomics and genetics initiatives and consortia, especially those that are characterising population variation (e.g. the UK Biobank, https://www.ukbiobank.ac.uk/), describing variation between species (e.g. Earth Biogenome Project, https://www.earthbiogenome.org/), working to understand variation that drives common diseases (e.g. International Common Disease Alliance, http://icda.bio), collating catalogues of clinically relevant variants (e.g. ClinVar[4]) and defining the clinical relevance of genes and variants for use in precision medicine (e.g. ClinGen, https://clincalgenome.org). The alliance is explicitly inclusive, seeking to engage diverse expertise by embracing a broad range of active participants and stakeholders from across the globe.

Where we are now

Historically, the functional impacts of genetic variants have been evaluated using *in vitro* or cell and animal models to inform variant interpretation. However, as traditionally practiced, these functional assays have had numerous limitations. Assays initiated ‘reactively’, after the variant has first been observed often arrive too late to help. Where such experiments have tested only a handful of variants at a time, in different labs at different times, adopting individual protocols and often without sufficient and/or appropriate controls, they cannot be quantitatively validated relative to human phenotypes. Moreover, low throughput functional assays cannot realistically scale to the tens of thousands of possible SNVs in a typical human gene, much less the millions of possible SNVs in the disease-relevant human genome.

Multiplexed assays of variant effect (MAVEs). MAVEs solve these problems by enabling the experimental assessment of thousands of individual variants simultaneously. Although MAVEs can be implemented using a broad range of experimental approaches, each involves mutagenesis of a DNA-encoded protein or regulatory element followed by a multiplexed assay of some aspect of the functionality of that sequence. The first MAVEs were initially applied to small protein domains and short regulatory elements [5,6], termed ‘deep mutational scanning’ and ‘massively parallel reporter assays’ respectively. Some early efforts assayed a single ‘sub-function’ of an element such as ligand interaction[5,7,8] or stability[9,10] in the case of proteins, and ability to serve as a promoter in the case of regulatory elements[6,11]. Other early efforts focused on the ability of an element to carry out its overall function in a cell-based growth assay[12]. Subsequently, MAVEs have been developed for a variety of functions,and have been used to acquire multiple variant effect maps for the same element examining different 'sub-functions' (reviewed in[13–15] and elsewhere).

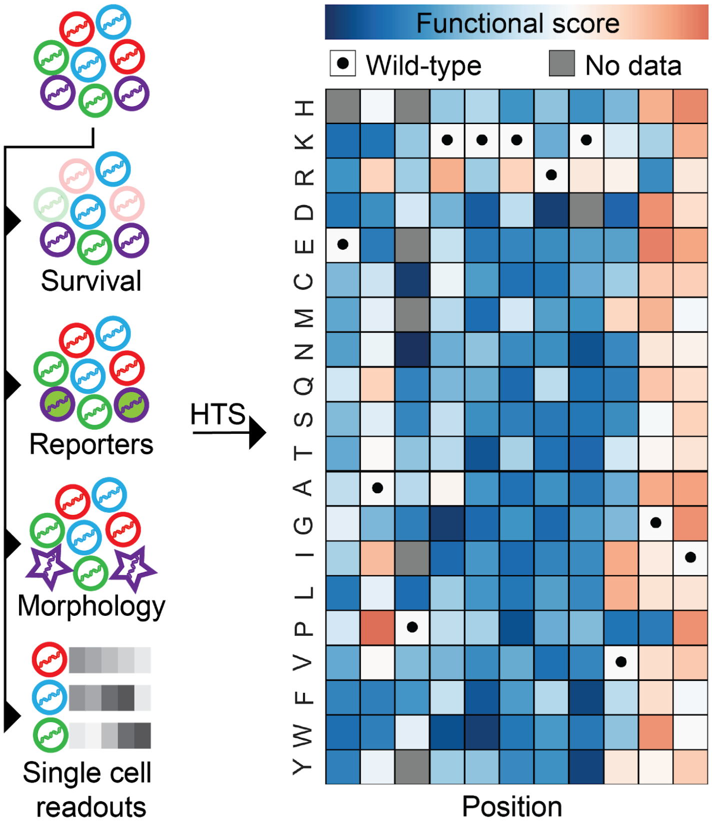

FIGURE 1: Measuring variant effects at scale with multiplexed assays. (left panel) In a multiplexed assay of variant effect (MAVE) a library of variants is expressed in cells such that each cell contains only one variant. Cells expressing the library are then subjected to a functional assay to read out variant effects. Example assays include survival, where variants support cell survival; fluorescent reporter assays, where variants drive a reporter whose signal is used to sort cells according to their fluorescence; cell morphology; or single cell transcriptomic readouts. Cells are then deeply sequenced to track each variant in the assay, and functional scores are computed for every variant. (**right panel)** Functional scores can be arranged into a variant effect map, organized by position in the genome (columns) and variants (rows). Blue indicates low functional scores, white indicates a wild type-like score, and red indicates high functional scores. Gray indicates missing data, and black dots indicate the wild type variant at the indicated position. HTS = high-throughput DNA sequencing

So far, MAVEs have mostly focused on SNVs, as these are the type of variant observed most frequently in human patients, or single amino acid substitutions which allow biochemical insight into the structural and functional role of native amino acids. However, denser mutagenesis yielding combinations of variants can provide insight into intra-gene genetic interactions and thus relationships between the functions of variants at different positions, reveal multiple relevant molecular phenotypes, illuminate the biophysical mechanisms of gene regulation and can even be used to solve protein structures[10,11,16–18].

To date, variant effect maps have been generated for approximately ~80 protein-coding human genes encompassing ~300,000 total variants (**Figure 2**). Many of these maps have been generated in the context of developing the foundational technologies rather than addressing disease-relevant objectives. The existing variant effect maps cover <1% of the clinically relevant human genome. Moreover, no functional element has been mapped in a diverse panel of cell types or across developmental stages.

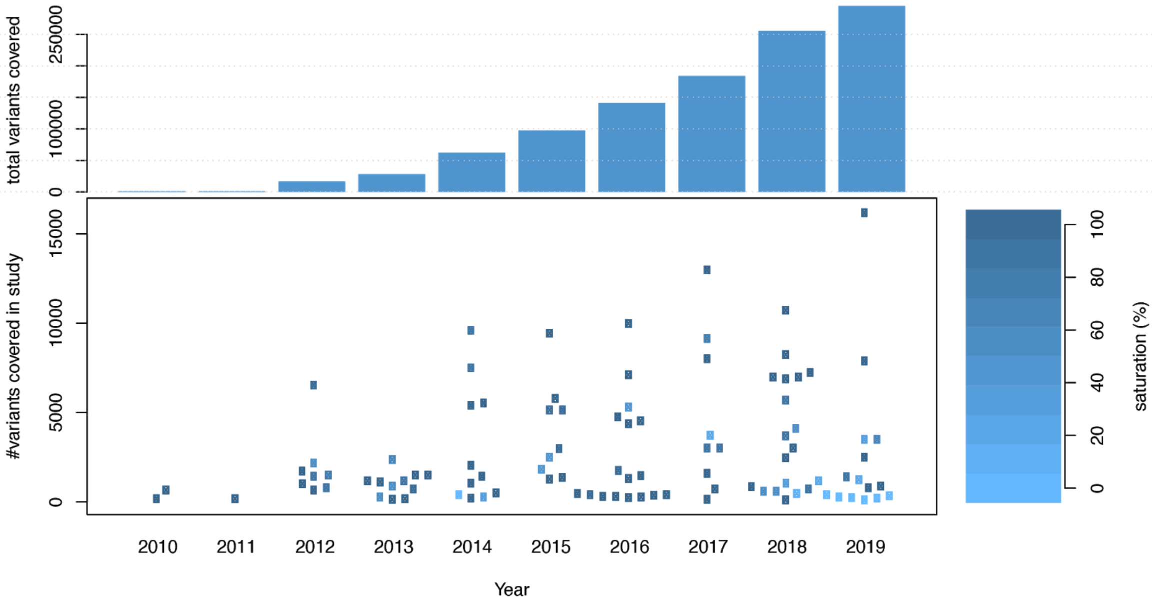

FIGURE 2: Multiplexed variant functional data is rapidly accumulating. Through 2019, ~250,000 single-amino acid change variant effects had been measured using MAVEs of human protein-coding sequences. This includes studies measuring multiple effects for variants studied previously under different conditions. (**top panel**) The cumulative number of variant effects measured using MAVEs by year. (**bottom panel**) The number of variant effects reported in individual studies, where colour indicates the study's saturation of its respective target space.

The value of functional evidence for informing clinical variant interpretation is already well appreciated and has been incorporated within current professional guidelines for genetic diagnosis that are used internationally[19,20]. These guidelines are actively evolving as the optimal weighting of functional evidence is investigated and refined. For example, a recent variant effect map generated in a cell growth assay for *BRCA1* achieved 97% sensitivity and 98% specificity in discriminating germline cancer risk-associated variants[21]. In recently revised recommendations, these *BRCA1* functional data can be used as strong evidence for or against pathogenicity, moving many variants that were previously considered VUS toward classifications of greater clinical utility[20]. MAVE-derived variant functional data has numerous advantages, especially when used to aid clinical variant interpretation. Unlike assaying variants in small batches using different methods in different labs, MAVEs can assay thousands of variants simultaneously, not only improving reproducibility but allowing assessment of variants in the context of the functional effects of all of the variants in that gene, including the effects of known pathogenic and benign variants. The systematic nature of MAVE-derived functional data also enables thorough and statistically rigorous evaluation of the concordance between assay results and clinical interpretations, which is critical for determining how clinically useful the assay is likely to be[20,22], Undoubtedly, frequency-based and clinical reporting-based metrics will continue to be important, but these may have ancestry-dependent biases, whereas MAVEs yield data that are largely independent of ancestry. Thus, the use of multiplexed variant functional data could fundamentally improve the equity of variant interpretation accuracy.

Existing variant effect maps for human genes have been generated by a range of different technologies, from yeast complementation assays to CRISPR-based saturation genome editing in human cells. Each technology has specific advantages and disadvantages. For example, yeast complementation assays are only applicable to a minority of human genes and would not be appropriate for identifying variants that disrupt splicing, whereas CRISPR-based saturation genome editing is currently costly and practical only for growth-based assays. Thus, no single technology can currently be used to generate maps of variant effects for all functional elements, indeed, even within a single gene, multiple assays may be required to assess different pathophysiological mechanisms. Broadly speaking, existing MAVEs can be used to create nearly saturating SNV effect maps for a typical human gene at a reagent cost of $25,000-$75,000, encompassing library preparation, cell culture and sequencing, and it typically takes one researcher 1-3 years to generate a variant effect map. However, current MAVEs require appreciable effort, and the cost of developing new assays can be considerable. Thus, while the existing portfolio of MAVE technologies can be applied to a substantial fraction of genes and other functional elements in the genome, more technology development is required to achieve comprehensive coverage of genomic functional elements, and to increase scalability and reduce costs.

A growing and highly active community of technology developers in academia and industry is extending current MAVE technology, focused on three broad areas:

- improving mutagenesis methods, particularly in the endogenous genomic context;
- developing scalable information-rich assays based on molecular phenotypes and cell morphology/behavior;
- moving into diverse cell types representing broader biological and developmental contexts, and potentially interrogating non-cell autonomous effects;

This highly collaborative community of researchers is actively and openly sharing new methods and tools. For example, the MAVE community has developed a first generation of analytical tools[23–28], established initial data generation and reporting standards[22], and developed a central repository, MaveDB[29]. MaveDB adheres to FAIR principles (Findable, Accessible, Interoperable, and Reusable) and has dramatically improved availability for many new and previously published datasets[30]. The broad range of use cases and users for MAVE datasets requires different downstream tools and interfaces. Well-managed central data repositories serve as key platforms for these downstream tools. For example, new data visualization tools for MAVE datasets such as MaveVis and dms-view use these emerging standards and repository[29,31]. While these tools represent a promising start, further development is needed. For example, more accurate, robust and scalable methods for scoring and assessing confidence in variant effects from raw data, imputation of missing variant effect data and modeling the genotype-phenotype relationship from variant effect data are all urgently needed.

Finally, integration of MAVE evidence with existing clinical workflows is ongoing[20]. A major challenge is combining evidence from multiple MAVEs for variant interpretation. Some assays are specific to only one function of the protein but may miss variants that alter other functions. Other assays may be able to distinguish allelic disorders where two distinct phenotypes with different molecular mechanisms arise from variants in a gene, but not spectrum disorders where a spectrum of overlapping/non-overlapping phenotypes with the same molecular mechanism arise, or vice versa. Thus, structured ways to combine MAVE-derived functional evidence will be increasingly important. Coordination with existing genomic data services (e.g. ClinVar, UniProt, etc.) and standards bodies (e.g. ClinGen, GA4GH, etc.) is needed to ensure MAVE datasets are fully interoperable with current and future workflows in diverse communities.

Impacts

The atlas of variant effects will transform our understanding of genetics by ushering in a new era of nucleotide-resolution knowledge of the genome. The atlas will have major impacts on basic research, translational research and clinical applications of genetic information. These impacts will benefit patients and present new opportunities for industry, as well as advancing our knowledge of fundamental biology. In essence, any effort to determine whether a variant is likely to alter the function of a protein-coding sequence or the activity of a regulatory element will be transformed by having an atlas of variant effects. Some areas of high impact include:

- **Genetic diagnosis.** Accurate, timely diagnosis is the foundation of optimal clinical care. Over 4,000 genes have been robustly associated with single gene disorders (http://www.omim.org), and over 500 genes contain variants that have been causally implicated in cancer (https://cancer.sanger.ac.uk/census). Nucleotide-resolution maps of genes and other functional elements that have already been robustly associated with disease will drive more accurate, more rapid and cheaper genetic diagnostic testing, both for single gene disorders and for cancer. Variant effect maps will be especially informative for disease-associated non-coding elements, where our current inability to predict the effect of a variant is most pronounced. VUS are widely-regarded as a major impediment to the efficacy and wider deployment of genomic medicine. Accurate diagnosis not only informs clinical care, but is critical for conducting informative clinical trials. Thus, improvements in genetic diagnosis is the ‘low-hanging fruit’ of AVE, with major impacts in the short-term.

- **Disease prediction and prevention.** Prevention of adult-onset single gene disorders can be greatly facilitated by accurate determination of disease risk at an earlier age (‘screening’). Adult-onset, single gene disorders include cancer and cardiovascular diseases[32]. Interpretation of genetic variants not seen previously is even more challenging in a screening context than in the diagnostic context, due to the greater risk of false positive results, which can have devastating consequences. Users of AVE will drive more accurate, more rapid screening for diseases that have not yet manifested and thus empower preventative medicine.

- **Personalised medicine.** Optimal therapy and clinical management varies from person-to-person. Genetic variation plays a key role in determining which drugs at which dosages are likely to be safe and efficacious[33]. Interpreting genetic variants that may or may not influence the optimal therapy is challenging, especially for variants that have not previously been observed. AVE will empower pharmacogenomics, enabling more patients to receive drugs that are appropriate for them, reducing severe adverse responses and improving outcomes. Drug resistance, whereby a pathogenic aggressor such as a tumour or an infectious agent, becomes tolerant to a previously efficacious drug, is a major limiting factor in oncology and infectious disease. AVE will encompass genetic variation within the specific targets of drugs and genes involved in drug resistance, as well as within genes encoding the proteins involved in the absorption, distribution, metabolism, and excretion of drugs.

- **Disease association studies.** Associating genetic variation inside a gene or a gene regulatory element with disease underpins all applications of genetics in medicine, including diagnostics, screening and personalised medicine. Currently, very little is known about the contribution of rare genetic variants to common, multifactorial disease risk. Similarly, many (often most) patients with a suspected rare, single gene disorder do not currently receive an informative genetic diagnosis, despite extensive genetic testing. Thus, more disease associations will be found, for all classes of disease, especially for rarer genetic variants that are inherently harder to interpret through purely statistical means. Systematic functional data can empower a next generation of association studies to identify new associations between rare genetic variants and disease risk[34,35]. These new associations will improve genetic prediction of common diseases, which is currently typically limited to genetic risk scores solely comprising common genetic variation, and more generally improve diagnostics, screening and preventative medicine.

- **Disease mechanisms.** Understanding the cellular and molecular mechanisms of a disease is critical for devising appropriate therapeutic strategies, and can inform clinical management. Many genes are associated with more than one monogenic disorder, meaning that different genetic variants in a gene and its regulatory elements can cause distinct clinical conditions. Often these different conditions can result from opposing mechanisms of action, for example, having too much or too little activity of the encoded protein. Sometimes these allelic conditions are distinct (e.g. Hirschsprung disease and multiple endocrine neoplasia caused by mutations in *RET*), but other times less so (e.g. neurodevelopmental disorders caused by mutations in *SCN2A*). Different functional assays, potentially in different cell types, may be required to distinguish different disease mechanisms in the same gene. Moreover, many robust genetic associations with disease are statistical in nature, with the underlying mechanism being unknown. AVE will support efforts to identify and disentangle different mechanisms of variant effect within a given gene.

- **Drug development.** Genetics is increasingly important in drug development. Drugs modulate the expression or activity of specific drug targets, typically proteins. Evidence of genetic association of a drug target with disease is strongly correlated with successful progression through clinical trials[36]. Increasingly, the safety and efficacy of modulating specific drug targets can be predicted by determining the phenotypic effects of genetic variants that increase or decrease the function of the drug target. A key limitation of these ‘dose-response allelic series’ is knowing what effect a variant has on the drug target. It can be hugely informative to identify associations within broadly phenotyped population cohorts of a series of alleles within the same gene of known functional effects. AVE will generate variant effect maps for potential drug targets, which will transform our ability to assess the suitability of different therapeutic strategies, prioritising those with the greatest chance of success. AVE will also empower clinical trials for new drugs for single gene disorders by helping to ensure that patients recruited for those trials have been accurately diagnosed.

- **Sequence/structure/function relationships.** Understanding the relationship between the sequence of a functional element and its function is fundamental to biology[15]. Especially in the case of proteins, the mapping between a given sequence, its structure and its function is complex and remains difficult to predict. Variant effect maps, by virtue of exploring a large swath of sequence space, shed light on these relationships. Already, maps have been used to improve or evaluate computational variant effect prediction, learn about protein structure and understand the composition and mechanisms of regulatory elements[6,11,17,18,37,38]. By revealing the effects of genetic variation in a comprehensive manner, AVE will catalyse the development of new tools for understanding and predicting how sequence drives structure and function, which, in turn will empower all of the impacts described above.

- **Evolutionary genetics.** The contrasting biology of different species is encoded in their genomes. While genetic differences between species can easily be identified, the functional significance of these differences remains largely unexplored. Most genetic differences between species are random changes of no functional significance, and thus discerning the genetic differences that drive functional differences is akin to finding a needle in a haystack. Biological differences between species are not just of academic interest, but underpin a wide variety of different commercial applications. AVE promises to illuminate the functional consequences of genetic changes that define species, and thus improve the understanding of human biology as well as related species.

- **Pathogen biology.** Genetic variation in pathogen genomes influences key characteristics of pathogen biology, including virulence, transmission, immune evasion and drug resistance. The rapid spread of genetic variants that alter pathogen biology can result in major changes in infectious disease burden, and require dramatic public health interventions. Generating maps of variant effects in pathogen genomes will inform the surveillance of pathogen evolution and provide opportunities to respond more rapidly. For example, variant effect maps could empower the design of vaccines that target the most functionally important and least mutable parts of antigens or small molecules with favorable resistance profiles.

Near-term Goals and Recommendations

Developing a systematic and extensive understanding of the functional effect of variants in the human and other key genomes is an ambitious goal, and every herculean undertaking needs a place to start. The Human Genome, 1000 Genomes and ENCODE Projects started by assembling key stakeholders, developing advanced technologies, demonstrating scalability and building the capacity that would lead to success. In addition to developing and refining new technologies, big, collaborative projects require data standards and data dissemination principles that take several iterations to perfect. Despite the rapid uptake of MAVE technologies by diverse labs and the first examples of use of MAVE-derived functional data in the clinic, our community largely lacks consensus on these standards and resources.

Thus, to meet the Mission that we have set ourselves, and to achieve the Impacts set out above, we make the following Recommendations regarding what we, as a community, should aim to achieve over the next 3-5 years. These Recommendations fall within five key areas: technology development, data generation, data analysis, data coordination & sharing, clinical and biological translation. Progress towards these near-term goals will be the key metrics on which the success of the AVE alliance should be judged.

Experimental methods development

Recommendation 1: to develop new MAVE technologies, especially information-rich, disease-relevant cellular assays, to increase coverage of genes in the genome

Recommendation 2: to expand the range of biological contexts in which MAVEs can be performed, including to disease relevant human cell types and multicellular models to enable quantitative assessment of the dependence of variant effects on context

Recommendation 3: to scale up existing MAVE technologies (via automation, miniaturisation, cost reduction) to allow cost-effective data generation of maps at larger scale

Data generation

Recommendation 4: to establish a small number of pilot projects that apply scalable MAVE technologies to a high value set of genes

Recommendation 5: to expand and diversify the AVE community generating and depositing variant effect maps, bringing domain-specific expertise to individual genes and accompanying assays

Recommendation 6: to establish a Registry of active projects generating variant effect maps to facilitate coordination, collaboration and assessment of assay rigor and reproducibility, especially by comparative study of different MAVEs focused on the same genes and regulatory elements and analysed with different available tools, without unintended duplication of effort

Computational methods development

Recommendation 7: to develop and systematically evaluate the second generation of open-source, freely available computational tools for variant effect mapping, including for scoring variant effects, integrating different maps of the same gene, and integrating of maps of variant effects with other informative data (e.g. evolutionary conservation, protein structure) to generate optimal prediction of biological effects and pathogenicity *in vivo*

Data standards and coordination

Recommendation 8: to extend, refine and disseminate the existing experimental design, data generation and reporting standards to ensure the integrity and utility of the aggregate data resource

Recommendation 9: to sustain and develop the existing data coordination infrastructure, ensuring data remain Findable, Accessible, Interoperable, and Reusable (FAIR)

Translation

Recommendation 10: to immediately engage the clinical diagnostic community in all stages of MAVE design, execution and evaluation

Recommendation 11: to develop different ‘channels’ of access to the AVE resource to serve the needs of different user communities, including APIs, data download, embedding in variant annotation workflows, and integration into resources that are already widely used (e.g. ClinVar, Uniprot, DECIPHER)

Recommendation 12: to demonstrate impact across different use cases, including integration in diagnostic interpretation workflows in cancer and single gene disorders

Values

Culture and Principles

The AVE Alliance will act with excellence and integrity, transparency, respect and inclusivity, empowering and engaging its members and stakeholders.

Research excellence and integrity:

- Core principles of the AVE Alliance include ethical design, conduct, reporting and application of research, adhering to current ethical standards, safety practices, relevant legal requirements, local organisational policies, and with the highest level of research integrity
- The Alliance recognises that conflicts of interest may arise, or appear to exist, and expects members to declare these openly
- All data generated must be managed and curated effectively throughout its lifecycle, including archiving, to ensure integrity and privacy as appropriate
- The Alliance views equal opportunity and training as important and expects all members to understand these principles when engaging with or representing the Alliance
- The findings, materials and resources generated by Alliance research will be made freely available to the research community (see below) and the Alliance aims to foster a culture of transparency and honesty

Inclusivity:

The Alliance believes that a diverse and inclusive community is absolutely essential to the achievement of our shared scientific goals. We also believe that it is critical to address disparities in genomics and health research, to ensure that *all* people can contribute to and benefit from the Alliance's work. In part these concerns can be addressed by the systematic and extensive nature of MAVEs, which measure the effect of variants in an unbiased manner. Additionally, the Alliance will engage constructively with underrepresented communities, especially when MAVEs potentially impact them.

Open Access Science:

The AVE Alliance is committed to sharing the data, resources, materials and publications it produces.

Publications, software and data:

- The Alliance encourages the publication of all research findings, including negative findings to allow others to benefit from the work and to avoid unnecessary repetition. Authorship should include all individuals who have made a substantial intellectual contribution as defined by the ICMJE and the contributions of funders should be clearly acknowledged and managed appropriately.
- To maximise public benefit the Alliance strongly encourages open and unrestricted access to publications and encourages the early deposition of manuscripts and sharing of protocols on an open-access preprint server such as bioRxiv.
- Members may develop software during the course of the project and in line with the alliance's ethos of openness and sharing are expected to make their software available to others whenever possible. For example, software should be distributed publicly and with open source, under a free software license, by the time the article employing it is published.
- The AVE Alliance Data Release Policy is intended to ensure that data is as freely and openly available as possible while protecting the rights of data generators to be the first to present or publish large-scale analyses of their results, in keeping with the Bermuda Principles and Fort Lauderdale Agreement.

Organization & Governance

The organization and governance of the Alliance provides a lightweight framework that maximizes impact while retaining the ability to grow and adapt. We envision the Alliance serving a broad array of stakeholders, falling into overlapping categories: developers of experimental and computational methods; data producers and analysts; data consumers; individuals and organizations providing genotypic or trait data; and funders. Below, we outline a provisional structure that will provide strategic leadership, workstreams that will accomplish specific goals and interest groups that will organize around biological or clinical problems (**Figure 3**). We describe the interests of each stakeholder group and explain how to get involved.

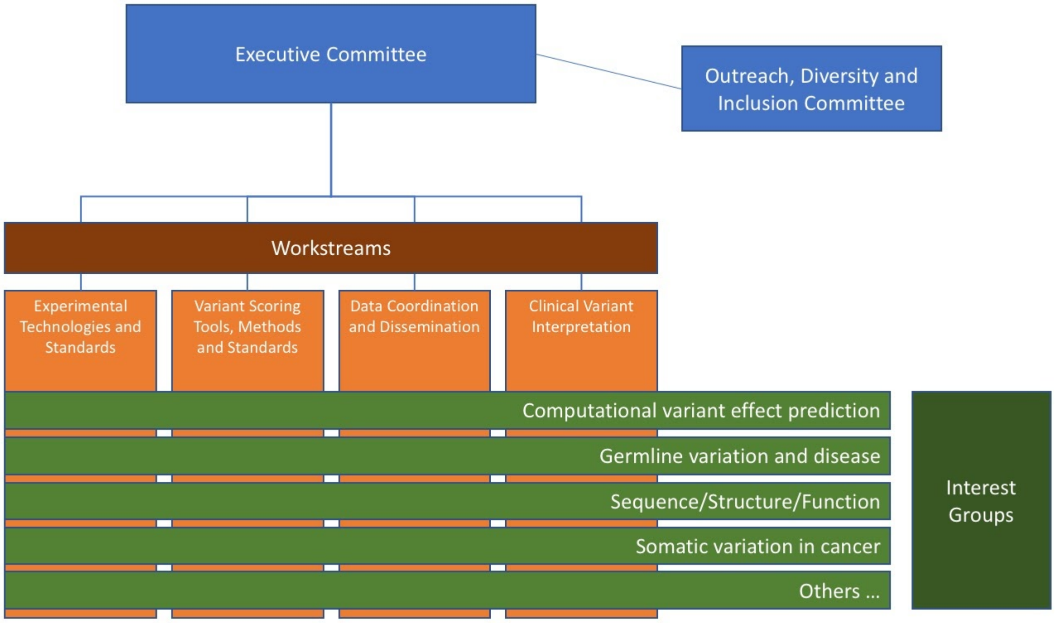
Executive Committee
Outreach, Diversity and Inclusion Committee
Workstreams
Experimental Technologies and Standards
Variant Scoring Tools, Methods and Standards
Data Coordination and Dissemination
Clinical Variant Interpretation
Computational variant effect prediction
Germline variation and disease
Sequence/Structure/Function
Somatic variation in cancer
Others ...
Interest Groups

FIGURE 3: An outline of the organisational structure of the Alliance

Leadership committees

Alliance committees will provide strategic leadership and facilitate partnerships between the Alliance and stakeholder communities. They will consist of Alliance members as well as advisors and stakeholders outside the Alliance. We propose two leadership committees, an Executive Committee and an Outreach, Diversity and Inclusion Committee, with the possibility of adding others bringing together, for example, funders.

Executive Committee

The Executive Committee is responsible for the overall governance of the Alliance. The Executive Committee will handle membership, organize an annual meeting, oversee the efforts of other committees and coordinate the efforts of the working groups. The Executive Committee will also coordinate the efforts of the Alliance with other entities whose work relates to the Alliance such as IGVF, ICDA, ClinGen, Encode, etc. Executive Committee membership will be diverse representation of different stakeholders and working groups as well as balancing regular turnover of members with preservation of institutional memory.

Interim co-chairs: Matt Hurles and Doug Fowler

Interim members: Deborah Marks, JT Neal, Alan Rubin, Lea Starita, Bill Hahn, Fritz Roth, Dave Adams, Anna Gloyn

Outreach, Diversity and Inclusion Committee

The Outreach, Diversity and Inclusion Committee will consist of a cross-section of stakeholders and will serve three primary functions. First, the Outreach, Diversity and Inclusion Committee will work within and outside of the Alliance to ensure that the leadership, membership, work practices and work products of the Alliance reflect our commitment to diversity and inclusion

(Recommendation 5). Second, the Outreach, Diversity and Inclusion Committee will publicize the work of the Alliance, educating members of the scientific, clinical and general communities. The Outreach, Diversity and Inclusion Committee will report their activities to the Executive Committee on a quarterly basis.

Workstreams

Members of the Alliance workstreams will be responsible for realizing the goals of the Alliance by establishing the key infrastructure for turning AVE into reality. Their work will include developing and evaluating methods and tools, setting standards, and disseminating information. Workstreams will develop specific deliverables in collaboration with the Executive Committee, initially taking on responsibility for turning the Recommendations into action. Workstreams activities will require regular discussion, as well as robust asynchronous communication. Each workstream will nominate one member to sit on the Executive Committee and one member to sit on the Outreach, Diversity and Inclusion Committee.

Experimental Technology and Standards workstream

The Experimental Technology and Standards (ETS) workstream will be responsible for facilitating the development, scaling, evaluation, comparison and dissemination of new MAVE methods (Recommendations 1, 2 and 3).

Variant Scoring Tools, Methods and Standards workstream

The Variant Scoring Tools, Methods and Standards (VSTMS) workstream will develop and systematically evaluate computational tools for variant effect mapping and visualization (Recommendation 7). VSTMS will evaluate the impact of experimental design choices like library complexity, number of independent cells and sequencing depth on the accuracy of scoring and error estimation. VSTMS will also be responsible for developing experimental design and reporting standards to enable evaluation of the quality of MAVE datasets in terms of internal controls, replicability and minimal information to be included (Recommendation 8).

Data Coordination and Dissemination workstream

The Data Coordination and Dissemination (DCD) workstream will facilitate the registration of MAVE projects (Recommendation 6). DCD will define and promote infrastructure for MAVE data deposition, coordination and dissemination (Recommendation 9). They will curate and organize MAVE data from the literature into this resource. DCD will also engage with clinical and non-clinical data resources to enable MAVE data sharing (e.g. ClinGen, UniProt, PharmGKB).

Clinical Variant Interpretation workstream

The Clinical Variant Interpretation (CVI) workstream will resolve issues relating to the use of MAVE data to interpret human genetic variants. CVI will establish a small number of pilot projects that apply scalable MAVE technologies to a high value sets of genes (Recommendation 4). CVI will develop approaches for integrating variant effect maps with other sources of information in clinical interpretation in collaboration with clinical standard-setting bodies such as ClinGen (Recommendations 8 and 10). These include best practices such as for curation of gold-standard variants from databases like ClinVar, evaluation of MAVE performance as clinical evidence, and for combination of multiple MAVE data sets, if available.

Interest groups

Alliance interest groups will enable technology developers, data producers/analysts, data consumers and funders to organize around common biological or clinical interests. Interest groups will foster collaboration and coordination, and drive scientific exchange. Interest group members will design and execute projects that contribute data and tools to AVE. Some of these initiatives will act as 'driver projects' that apply the outputs of the workstreams to real-world problems. These driver projects are crucial for influencing the development of the methods, tools and standards being produced by the workstreams. Unlike workstreams, which will be durable and will take on specific tasks (e.g. conducting technology comparisons, developing standards and databases, etc), interest groups will be more flexible and may not have specific deliverables. Alliance members can self-organize interest groups at will, and we intend for this low-stakes option to increase member engagement. We envisage that many Alliance members will wish to be part of both workstreams and specific interest groups.

Some interest groups we anticipate forming are:

- Computational variant effect prediction
- Drug development, target identification and resistance
- Germline variation in common diseases, cancer risk, mendelian disorders and pharmacogenomics
- Pathogens
- Sequence/structure/function relationships
- Somatic variation in cancer
- Variation and evolution

Key Stakeholders

Technology developers

MAVEs are new, and experimental and computational technology for MAVEs are rapidly developing. The Alliance contains a cross-section of members working on MAVE technology and will facilitate performance comparison, technology dissemination, and cross-fertilization.

Data producers and analysts

Existing and newly developed MAVEs are being applied by members of the Alliance. Some members will apply generic MAVEs to tens or hundreds of genes, whereas other members will apply boutique MAVEs on a small scale. The Alliance will facilitate coordination amongst data producers and analysts for the purposes of quality assessment, standardization, data dissemination and avoidance of competition/effort duplication.

Data consumers

MAVE data will be consumed in two distinct, but related, realms: basic science and clinical translation. On the basic science side, biologists, biochemists, genome scientists, and computational/quantitative biologists will use MAVE data to understand proteins, genes, gene regulatory elements and pathways of interest. On the translational side, patients, patient advocacy groups, clinical geneticists, genetic testing companies and biobanks will use MAVE data to interpret the effects of genetic variants. The Alliance will enable data consumers to easily find, assess and utilize MAVE data. Having data consumers at the table before data is collected will maximize the utility of the data.

Funders

The Alliance will consist of independently funded efforts ranging in size from small to large. Similarly, funders may range from foundations interested in one or a few genes to, potentially, government agencies interested in large-scale efforts. The Alliance will enable funders, technology developers, and data producers/analysts to discover each other, define mutual goals and develop proposals.

How to get involved

The Alliance welcomes individuals in any of the stakeholder categories from industry, academia, government or other entities anywhere in the world. Members of AVE Alliance should:

- share the same vision of working towards a comprehensive and freely-available atlas of variant effects,
- agree to adhere to the code of conduct
- be willing to contribute their time.

We encourage you to become a member by visiting http://varianteffect.org.

Acknowledgements

We would like to acknowledge the commentary, discussion and continued involvement of all of the attendees of the 2020 Atlas of Variant Effects Alliance Launch Meeting (**Supplementary Table 2**).

Competing interests

ALG declares that her spouse is an employee of Genentech and holds stock options in Roche. DJA is a consultant for Microbiotica and Astra Zeneca. DSM is a consultant for Insitro, Dyno and Octant. JTN receives research support from Bristol Myers Squibb. FPR holds shares in Ranomics, Inc. and is an investor and advisor for SeqWell, Inc. (each offering services that are potentially useful for variant effect mapping), and has accepted conference travel support from Illumina, Inc. LMS is a consultant for Nostos Genomics. JTN receives research support from Bristol Myers Squibb. W.C.H. is a consultant for Thermo Fisher, Solasta Ventures, MPM Capital, Tyra Biosciences, iTeos, Frontier Medicines, Jubilant, KSQ Therapeutics, RAPPTA Therapeutics and Paraxel. MEH is a consultant for AstraZeneca and co-founder, director, shareholder of Congenica Ltd.

References Cited

1. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-Specific Variants and the Limits of Human Genetics. *Trends Mol. Med.* **22**, 925–934 (2016).
2. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
3. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
4. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–5 (2014).
5. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
6. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
7. Zhang, H. *et al.* Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13456–13461 (2011).
8. Ernst, A. *et al.* Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.* **6**, 1782–1790 (2010).
9. Kim, I., Miller, C. R., Young, D. L. & Fields, S. High-throughput analysis of in vivo protein stability. *Mol. Cell. Proteomics* **12**, 3370–3378 (2013).
10. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16858–16863 (2012).

11. Kinney, J. B., Murugan, A., Callan, C. G., Jr & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9158–9163 (2010).

12. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences* **108**, 7896–7901 (2011).

13. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

14. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Hum. Genet.* **137**, 665–678 (2018).

15. Kinney, J. B. & McCandlish, D. M. Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annu. Rev. Genomics Hum. Genet.* **20**, 99–127 (2019).

16. Otwinowski, J. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).

17. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).

18. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).

19. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

20. Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).

21. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome

editing. *Nature* **562**, 217–222 (2018).

22. Gelman, H. *et al.* Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Med.* **11**, 85 (2019).
23. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).
24. Bloom, J. D. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics* **16**, 168 (2015).
25. Tareen, A., Ireland, W. T., Posfai, A., McCandlish, D. M. & Kinney, J. B. MAVE-NN: Quantitative Modeling of Genotype-Phenotype Maps as Information Bottlenecks. *Cold Spring Harbor Laboratory* 2020.07.14.201475 (2020) doi:10.1101/2020.07.14.201475.
26. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
27. Matuszewski, S., Hildebrandt, M. E., Ghenu, A.-H., Jensen, J. D. & Bank, C. A Statistical Guide to the Design of Deep Mutational Scanning Experiments. *Genetics* **204**, 77–87 (2016).
28. Wu, Y. *et al.* A web application and service for imputing and visualizing missense variant effect maps. *Bioinformatics* **35**, 3191–3193 (2019).
29. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).
30. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
31. Hilton, S. K. *et al.* dms-view: Interactive visualization tool for deep mutational scanning data. *Cold Spring Harbor Laboratory* 2020.05.14.096842 (2020)

doi:10.1101/2020.05.14.096842.

32. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).

33. Weinshilboum, R. M. & Wang, L. Pharmacogenomics: Precision Medicine and Drug Response. *Mayo Clin. Proc.* **92**, 1711–1722 (2017).

34. Rees, M. G. *et al.* Correlation of rare coding variants in the gene encoding human glucokinase regulatory protein with phenotypic, cellular, and kinetic outcomes. *J. Clin. Invest.* **122**, 205–217 (2012).

35. Bonnefond, A. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297–301 (2012).

36. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

37. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst* **6**, 116–124.e3 (2018).

38. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).