What is this document and how do I use it?

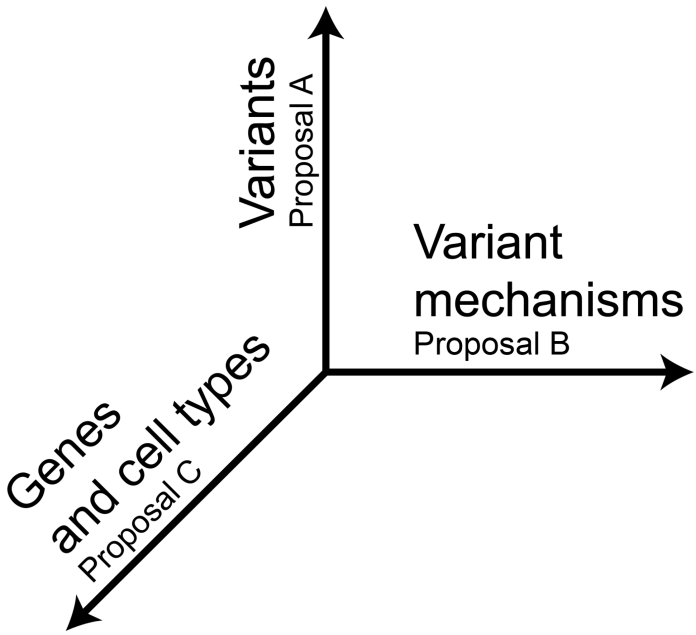
In this document, we articulate 3 straw proposals intended as starting points for our discussion. We also briefly cover the background and current state of the variant effects field. We plan to publish a workshop summary, which will contain elements of the straw proposals and background. **Thus, as you are reading this document please add your thoughts and suggestions as comments!**

AVE 2030: resolving human variants of uncertain significance

Two decades after the International Human Genome Project revealed the first human genome sequence, and with millions of humans having since been sequenced, we remain largely unable to interpret the simplest (single-nucleotide) changes, even in the genome's best-understood (protein-coding) regions harboring the vast majority of known disease-causing variation. These uninterpretable variants end up as Variants of Uncertain Significance (VUS), which cannot be used to diagnose or treat disease.

But, a confluence of experimental and computational technologies have enabled determination of single nucleotide variant effects at scale. Applying these technologies comprehensively would generate an atlas of variant effects, with profound implications for understanding human biology and disease. Having largely overcome the challenges of sequencing human genomes and identifying the functional elements therein, a "third phase" of the genomics revolution would comprehensively evaluate the impact of human variants on functional elements, phenotype and disease risk. Community engagement in this vision is already broad and energetic, e.g., via the Impact of Genomic Variation on Function consortium (IGVF) and Atlas of Variant Effects Alliance (AVE).

The meeting goal is to develop recommendations for a draft atlas that can be realized by 2030, with a focus on resolving variants of uncertain significance. This vision will require prioritizing the types of measurements and predictions that should be made. This draft atlas will be an amalgam of experimentally determined variant effects from a variety of types of assays executed by different labs and production centers as well as variant effect predictions from different algorithms. No single experimental assay can capture all possible mechanisms by which variants can have an effect, nor can any existing assay be scaled to all possible single nucleotide variants by 2030. Likewise, computational variant effect predictions currently cannot shed light on the mechanism by which pathogenic variants act, though they achieve genome scale. Thus, we will discuss the advantages and disadvantages of investing resources in different ways and, thereby, develop a vision for what an atlas should look like in 2030. We will also recommend actions for the community to take.



Below, we describe a series of straw proposals, designed to spark discussion. An important and cross-cutting issue is the tension between generating data to eliminate VUS quickly versus generating data to understand how variants act and empower development of more powerful predictive models. Proposal A is designed to eliminate VUS quickly; Proposal B to explore variant pathogenic mechanisms and yield mechanism-aware predictive models; and Proposal C to explore gene-cell type relationships and yield cell type-aware predictive models. We also give key background information in the four areas where recommendations are needed: technology development, infrastructure and community, clinical translation and data production.

Straw proposal framework

For each straw proposal, we articulate the goal and key assumptions along with principles of variant, assay and gene prioritization. Lastly, we describe the activities to be undertaken early (e.g. in 2025-2027) and those to be undertaken later (e.g. 2027-2030). We crafted the proposals in the context of a budget of ~1MM variant effect measurements. We assume that the average gene in the human genome has ~5,000 possible exonic and intronic SNVs of interest. Each straw proposal highlights a different way in which resources could be used and, consequently, a different vision for an atlas in 2030. The straw proposals are not meant to be endpoints or competing options to choose from, but, instead, starting points for our discussion.

Straw proposal A: Provide functional data and predictions to resolve as many extant and to-be-discovered VUS as possible by 2030

The goal of this proposal is to maximize the direct utility of invested resources in the short term by applying existing assays to a smaller number of genes, focusing on those genes with the highest burden of VUS and most clinical applicability. This proposal will provide at least one piece of high-quality functional data and at least one high-quality prediction for as many current and future VUS as possible. With a budget of ~1MM variant effect measurements, a single measurement could be made for SNVs in ~2,000 genes.

Assumptions

- Coverage of VUS will be maximized by focusing only on protein coding genes
- Sufficient VUS can be covered with existing assays, such that further technology development to capture complex variant effects (e.g. on transcriptomes, cell phenotypes, etc) is lower priority
- Current predictive models, which provide binary predictions of variant effect (e.g. loss of function vs. normal function) are sufficiently good for many genes

Principles of gene prioritization

- Pick genes with the largest number of VUS today
- Pick genes based on clinical utility (e.g. impact on patient care today, etc)
- Pick genes on basis of ability to demonstrate accuracy (e.g. # of control variants)
- Pick genes where currently available and scalable assays will work
- Pick genes where currently available computational predictors function well, to maximize reinterpretation of VUS by combining functional data and predictions

Early emphasis (% allocation of resources)

- Develop innovative approaches to scale existing experimental assays (40%)
- Develop innovative approaches to improve the accuracy of current predictive models (20%)
- Increase gene coverage by developing novel assays that fill important gaps (5%)
- Evaluate utility and deficiencies of already scalable assays (5%)
- Organize community of small scale mapmakers with bespoke assays around VUS elimination (5%)
- Build digital infrastructure for functional data deposition, QC/QA and dissemination into clinical decision support systems, with a focus on eliminating VUS (20%)
- Revise clinical guidelines (ACMG/ClinGen, etc) in anticipation of these functional data arriving (5%).

Late emphasis (% allocation of resources)

- Large-scale generation of data using existing or minimally optimized assays, largely from a small number of functional genome centers (80%)
- Curate and calibrate functional data and predictions for clinical use (5%)
- Improve integration of functional data into clinical workflows (15%)

Straw proposal B: Deeply understand a targeted set of VUS by 2030, focusing on variant pathomechanism

The goal of this proposal is to develop and deploy functional assays and predictive models that can generate a comprehensive understanding of how VUS perturb molecular (e.g. protein activity, stability, localisation, etc) and cellular (e.g. transcriptional, proteomic, and functional state) phenotypes. This proposal will provide multiple pieces of functional data to dissect how VUS exert their effect and will enable the development of models capable of predicting mechanisms of pathogenicity. This proposal would build out a large suite of technologies and apply them to a focused set of genes, developing a comprehensive description of variant pathomechanism. With a budget of ~1MM variant effect measurements, SNVs in ~200 genes could be exhaustively covered with 10 different assays.

Assumptions

- Current technologies can serve some high-need genes to produce useful information, but new methods are needed to provide mechanistic insights and also to assay most genes
- Deep mechanistic characterization of variants in a set of well-chosen genes will enable training of models that can predict the mechanism of a variant's effect
- Generating a holistic characterization of variant mechanism by applying many assays that probe different molecular and cellular functions for fewer genes will be useful clinically and also answer questions about what depth of characterization will be needed for other genes
- Understanding the molecular, cellular and genetic mode of action of variants will enable better diagnosis and also development of variant-specific therapies

Principles of gene and assay prioritization

- Pick genes where knowledge of mechanism for pathogenic variants could be most clinically impactful
- Pick genes with well understood mechanisms of molecular and cellular function that could be extended to the variant level
- Pick genes with incomplete penetrance and variable expressivity
- Pick genes and assays that measure phenotypes in which computational predictors perform poorly
- Explore non-coding variation where there is evidence it causes or modulates disease

Early emphasis (% allocation of resources)

- Develop innovative approaches to characterize variant effects on a range of molecular, cellular and tissue/developmental phenotypes (40%)
- Apply a range of mechanistically informative assays to a basis set of genes (30%)
- Explore computational models that can make mechanistic predictions (20%)
- Revise clinical guidelines to move from "functional vs. non-functional" to incorporate mechanistic insights (10%)

Late emphasis (% allocation of resources)

- Large-scale data generation using newly developed and existing assays, with the goal of exhaustive mechanistic characterization (50%)
- Develop foundation models for mechanistic variant effect predictions (20%)
- Develop a new, "variant pathomechanism" data resource that enables deposition, dissemination, discovery and exploration of mechanistic variant effect information (10%)
- Improve integration of mechanistic functional data into clinical workflows (10%)
- Develop variant-informed or variant-specific therapies (10%)

Straw proposal C: Provide functional data for a small number of variants in many genes and cell types by 2030

The goal of this proposal is to evaluate the effect of enough variants to validate an experimental model in as many genes as possible in as many different specialized cell types as possible. This proposal will reveal gene-specific experimental models and guide the development of cell context-specific predictive models. With a budget of ~1MM variant effect measurements, ~50 SNVs in all ~20,000 genes in the human genome could be covered.

Assumptions

- Current methods for programmed, single nucleotide editing, in combination with current iPS cell-derived specialized cells, are sufficient to enable testing of at least a handful of variants in most genes in the genome
- Genes act in specific cell contexts, and many gene-cell context relationships remain unresolved
- To resolve VUS at scale, a necessary first step is discovery of the relevant cell contexts for each gene-disease dyad

Principles of variant prioritization for each gene

- Include clinical control variants, if they exist
- Include VUS, if they exist
- Include disruptive variants, selected on the basis of computational predictions that leverage conservation and protein structure
- Consider disease associated non-coding variants that have been linked to the gene

Early emphasis (% allocation of resources)

- Develop innovative approaches to scale programmed, single nucleotide editing and iPS assay methods (30%)
- Identify a core set of specialized cell types and develop streamlined and scaled protocols for editing iPS cells and differentiating them to produce these cells (30%)
- Explore computational models that can make cell type-specific variant effect predictions (20%)
- Mine biobanks and other resources to curate/discover gene-disease relationships (20%)

Late emphasis (% allocation of resources)

- Large-scale data generation with the goal of measuring the effect of at least a few variants in nearly every gene in the genome in many cell types (60%)
- Develop foundation models for cell-type specific variant effect prediction (30%)
- Develop a new “variant cell atlas” data resource that enables deposition, dissemination, discovery and exploration of cell contextual variant effect information (10%)

Current state of the technology, data production, clinical translation and infrastructure surrounding an atlas

Technology development

A first generation of mature multiplexed assays of variant effect have been developed and are already widely deployed. These first-generation methods employ comparatively simple model systems and assays. Examples include simple assays in cultured human cell lines (e.g. saturation genome editing, VAMP-seq, MPRA), yeast (e.g. ddPCA) or molecular display. In general, first-generation assays read out phenotypes that are easy to measure (e.g. cell growth or protein abundance). Newly developed or under-development technologies seek to address these limitations. To overcome the first challenge, assays are being developed that enable variant effects to be measured in different genetic, cell type or tissue contexts. To overcome the second challenge, assays are also being developed that can read out richer phenotypes (e.g. cell morphology or transcriptomes). Thus, the atlas effort benefits from a range of experimental technologies that can be deployed at scale, with methods in the pipeline.

Computational variant effect predictors use information about a variant, especially from multiple sequence alignments and protein structure, to predict a variant's effect. Current variant effect predictors are highly accurate when benchmarked against clinical control variants. This high accuracy is the result of the use of more sophisticated modeling approaches, availability of more control pathogenic and benign variants and the rise of metapredictors. Limitations of current predictors are that they generally predict only simple, generalized function (e.g. functional or non-functional) without respect to mechanism or cell context; that evaluating their accuracy can be challenging owing to model circularity; and that many predictors make predictions only for some genes and only for coding variants. To overcome the first challenge, we propose the development of mechanism and cell context aware predictors, which will be greatly empowered by the availability of large volumes of high-quality experimental data. To overcome the second challenge semi-supervised or unsupervised models are being developed. To overcome the third challenge, models that can be extended to all clinically relevant genes are needed. Recently, the rules for using predictor data in clinical variant interpretation workflows have changed dramatically, giving predictions much more weight than was previously possible. Thus, the atlas effort benefits from high-quality, generic predictions of variant effect that can be used effectively in clinical variant interpretation.

Data production

We are in the first few years of scaling multiplexed assays of variant effect for production. Production efforts are underway within the NHGRI IGVF and at the Wellcome Sanger Institute, where the assays most amenable to scaling (e.g. ddPCA, saturation genome editing, VAMP-seq, MPRA) are being scaled. Thus far, these efforts have yielded a significant increase in the availability of multiplexed functional data, perhaps on the order of millions of variant effects measured. However, significant barriers exist to the scaling of even these existing, simpler assays for production. Those barriers are 1) high costs for variant libraries; 2) a large and manual human cell culture footprint; 3) changing incentive structure for collecting data at scale (e.g. prioritizing genes for production, publication and translation happen one gene at a time, currently - what to do with 10s/100s?). To overcome the first challenge we propose technology development around synthetic DNA and toward dedicated library synthesis cores and/or partnerships with industry. To overcome the second challenge, we propose process optimization and technology development aimed at minimizing the number of cells per variant required, as well as development of automation in partnership with industry. The first and second problems can also be addressed by exploiting the synergy between data production and computational variant effect prediction. First, data production can be focused on genes where predictors perform poorly, providing functional data that will reclassify some VUS. Second, data production can be focused on genes where predictors perform well. Here, combining predictor and functional data could lead to the reclassification of most VUS. To overcome the third problem, we propose a transition to user (e.g. biologist and clinician) driven

prioritization of production, with partnerships around genes (and sets of genes) formed before the data is collected.

Clinical implementation

In the last few years, multiplexed functional data has been increasingly incorporated into clinical variant interpretation workflows. The 2015 ACMG/AMP guidance, which has been widely adopted in both the US and internationally, lacked an objective standard for determining the strength of evidence that functional data could generate. This major barrier was addressed to a large degree in 2020 with release by the ClinGen Sequence Variant Interpretation (SVI) of the OddsPath framework as well as some basic guidance around evaluating the validity of functional assays. With this new guidance in place, several groups demonstrated that high-quality multiplexed functional data could provide strong evidence, and that inclusion of such evidence in variant classification workflows could sharply reduce the rate of VUS. Likewise, researchers, industry and clinicians have articulated draft guidance for executing and reporting high-quality MAVEs for clinical use. It is now clear that multiplexed functional data can increase the yield of genetic diagnostic assays, but remaining barriers prevent multiplexed functional data from being used to interpret the average variant.

Those barriers are: 1) most genes lack multiplexed functional data; 2) even for genes where multiplexed functional data exists, a “truth set” of variants sufficient for calibrating the strength of evidence under the OddsPath framework may not be available; 3) current guidance does not address all questions around use of multiplexed functional data, e.g., whether and when to separately calibrate different regions of a protein, how to standardize and incorporate measurement error estimates, what to do when multiple sources of functional data are available, and more; 4) even for genes where multiplexed functional data exists and evidence strength can be appropriately calibrated, clinicians and clinical labs cannot now easily discover and use the data.

To overcome the first problem (lack of functional data), we obviously propose generating more data. Prioritizing targets for MAVE studies should ideally depend on several factors, including a) clinical need (the VUS problem, both current and anticipated, and considering frequency of VUS in non-European populations); b) availability of assays that are (individually or collectively) likely to reflect the pathobiology of variants; c) the cost of carrying out the assay(s), which depends both on the nature of the assays and the length of the protein or genomic element; d) the willingness of appropriate experts in the biological and clinical aspects of the assay to engage; and e) ability to validate the functional assay data once it exists.

To overcome the second problem (lack of a truth set), new strategies are needed. In some cases, clinical control variants could be identified by deeper outreach to clinical labs and researchers testing or studying these genes (particularly the case for benign controls). Alternatively, case/control data generated from large biobanks could be used to demonstrate that functional scores correlate strongly with disease risk, even where no individual variant can be classified confidently.

To overcome the third problem (updating guidance for clinical translation), it will be necessary to work closely with the new committee developing the ACMG/AMP/CAP/ClinGen Sequence Variant Classification standards as well as ongoing updates to those standards (anticipated to be released annually) and guiding an evolved general framework for the use of functional evidence. In addition, it will be necessary to work with ClinGen Expert Panels developing gene-specific guidance for scoring functional evidence.

To overcome the fourth problem (data discovery), infrastructure is needed. In particular extensive curation is necessary to transform multiplexed functional data as it is not typically reported into a format readily usable by clinical laboratories and clinicians. Then, tooling must be built and maintained to share these curated data through ClinVar, the ClinGen Variant Curation Interface, and other places where clinicians and clinical labs discover evidence for variant classification.

Infrastructure, standards and coordination

Recent efforts have established an infrastructure for coordinating the generation, curation, sharing and translation of multiplexed functional data. Here, the NHGRI Impact of Genomic Variation on Function consortium (IGVF) aims to understand how genomic variation affects genome function and produce a catalog of the impact of genomic variants on genome function and phenotypes. The Atlas of Variant Effects Alliance (AVE) is a grass-roots, international effort focused on understanding the effect of variants at nucleotide resolution. AVE has constructed a variety of community resources and infrastructure. Despite the success of these two organizations, significant challenges exist. For example, MaveRegistry is a tool for researchers to announce their intent to generate multiplexed functional data for a particular gene or set of genes. But, use of this tool is voluntary and large-scale production efforts are coordinated informally and occasionally. MaveDB is a database with a large volume of multiplexed functional data and a growing feature set, but it is minimally curated and, being intended as a data warehouse, does not currently support data discovery, exploration or clinical use. A variety of groups have articulated standards for the generation, sharing and clinical use of experimental variant effect data, but these standards are not universally adhered to.

Thus, the major barriers are: 1) lack of active coordination among large production centers and also among smaller/ad hoc efforts and 2) lack of digital infrastructure to house, share, discover and explore multiplex functional data. Overcoming these barriers is important, especially if, as has happened so far, data continues to be produced in an ad hoc, decentralized and uncoordinated manner. To overcome the first barrier, we propose to build upon and extend the MaveRegistry framework. Large production centers can lead the way by agreeing to publish their production schedule. Small, ad hoc data producers would also ideally be incentivized to share their plans. Here, understanding what is preventing sharing (e.g. lack of time/interest/awareness vs. resistance) is the first step. One way to overcome the second barrier is to commit to the continued development of the resources that already exist or are in development, principally the IGVF catalog and MaveDB. Both are under active development and have plans to address many of the needs of the community. But, a systematic study of users (e.g. data generators, clinicians, researchers) would be needed, along with additional resources to develop and sustain these efforts. Another alternative would be to establish new infrastructure. Picking between these two alternatives soon, with broad community engagement, is a necessary first step.

Recommendations

- Generally, sustain strengths, address (important) weaknesses and consider acting on opportunities
- Mixture of proposal A and B
- Tight integration of AI and experiments
- Invest in better standards and descriptions of experiments
- Moving what we can do now into production, and focusing tech dev on the “next” next six years
    - Maybe straw C lives here, since it expands what is assayable (but make sure we are coordinating with IGVF/DepMap, need to make sure somebody does that).