ANNUAL REVIEWS

Annual Review of Genetics

Scalable Functional Assays for the Interpretation of Human Genetic Variation

Daniel Tabet,[1,2] Victoria Parikh,[3] Prashant Mali,[4] Frederick P. Roth,[1,2] and Melina Claussnitzer[5,6,7]

[1]Donnelly Centre, Department of Molecular Genetics, and Department of Computer Science, University of Toronto, Toronto, Ontario, Canada; email: fritz.roth@utoronto.ca

[2]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada

[3]Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, California, USA

[4]Department of Bioengineering, University of California, San Diego, California, USA

[5]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[6]Center for Genomic Medicine and Endocrine Division, Massachusetts General Hospital, Boston, Massachusetts, USA

[7]Harvard Medical School, Harvard University, Boston, Massachusetts, USA; email: melina@broadinstitute.org

Annu. Rev. Genet. 2022. 56:19.1–19.25

The *Annual Review of Genetics* is online at genet.annualreviews.org

https://doi.org/10.1146/annurev-genet-072920-032107

Keywords

functional assays, genetic variation, variant effect mapping, sequence–function studies, multiplexed assays of variant effect

Abstract

Scalable sequence–function studies have enabled the systematic study and cataloging of hundreds of thousands of coding and noncoding genetic variants in the human genome. This has improved clinical variant interpretation and provided insights into the molecular, biophysical, and cellular effects of genetic variants at an astonishing scale and resolution across the spectrum of allele frequencies. In this review, we explore current applications and prospects for the field, and outline the principles underlying scalable functional assay design, with a focus on the study of single-nucleotide coding and noncoding variants.

INTRODUCTION

Large-scale DNA sequencing has fundamentally transformed our ability to explore the mechanisms of disease biology throughout the life sciences. The accessibility of genomic technologies is altering patient management in the clinic and empowering molecular, biophysical, and cellular studies at increasing scale and resolution. Here, we review the utility and application of sequence–function studies in the large-scale functional assessment of comprehensive sets of human genetic variants, with a focus on single-nucleotide variants (SNVs). We outline the principles underlying scalable functional assay design and present a framework for choosing among alternative model systems and experimental methods. While this review is focused on cell-based assays of SNVs, the same principles may apply across a range of model systems and types of genetic variation.

A primary goal in the field of human disease genetics is to connect genetic loci to disease- and trait-relevant phenotypes. Building on the initial human reference genomes, researchers have made great progress toward cataloging genetic variation (1, 24, 101, 119, 121, 191). Genotyping and sequencing of large population-scale cohorts [e.g., the UK Biobank (25), All of Us (6), and Biobank Japan (141)], together with individual health, demographic, and lifestyle information, have helped link thousands of genetic loci to human traits and diseases and continue to reveal genetic associations with increasing size and diversity of sequenced cohorts (24, 181). Although population-based studies are increasingly well powered to associate common variants with human phenotypes, they are ill suited to making associations for rare and extremely rare variants—which account for most genetic variation and tend to have a larger effect on disease phenotype than common variation (33). Considering that nearly every possible single-nucleotide change currently exists in the human population (122, 195), there is a need for the functional assessment of even the rarest of variants. Here, large-scale sequence–function studies will be essential to inferring the impacts of variants experimentally. Moreover, where common variants are concerned, these studies can help identify the causal variant(s) across the thousands of common trait-associated haplotypes (61). Moving forward, sequence–function studies will be essential to expediting the assessment of genetic variation across the spectrum of allele frequencies, informing our understanding of the mechanisms of variant dysfunction along the way.

Scalable multiplexed assays of variant effect (MAVEs) have thus far collectively probed the functional consequences of hundreds of thousands of genetic variants, encompassing both coding (18, 21, 28, 50, 73, 74, 77, 78, 81, 88, 96, 112, 124, 127, 143, 173, 175, 194, 196) and noncoding elements such as splice sites (15, 20, 72, 97, 102, 197) untranslated messenger RNA (mRNA) regions (163), promoters (108), and enhancers (108, 135, 148). These variant-to-function (V2F) studies can yield variant effect (VE) maps that capture functional impacts of all possible substitutions within a target element, including impacts of variants not yet observed in humans.

A shared vision is now emerging for an atlas of VE maps (11) that encompasses every functional element in the human genome—a key requirement for which is the establishment of a coordinated set of scalable functional assays.

THE PROMISE OF SEQUENCE–FUNCTION STUDIES

Sequence–function mapping can reveal the molecular and mechanistic impacts of coding and noncoding genetic variation, facilitating clinical variant interpretation and by extension therapeutic, preventative, and diagnostic strategies (54, 84). Here, we discuss the potential of sequence–function studies and review progress toward a comprehensive human genome-scale atlas of variant effects.

Molecular and Mechanistic Interpretation of Genetic Variation

Most common genetic variants associated with human traits and diseases rest in noncoding regions of the genome (24). Although these association studies typically use a subset of single-nucleotide polymorphisms (SNPs) as "tags" representing a larger haplotype—which complicates identifying which of typically many variants in the haplotype are causal—disease-associated haplotypes are enriched for sites with increased chromatin accessibility and histone marks associated with enhancer regions, suggesting that causal variants often alter transcriptional regulation (66). Transcriptional regulatory elements can have both proximal and distal effects and often involve multiple upstream regulators and downstream effector genes and noncoding RNAs, which further complicates the identification of functional noncoding variants (34, 83, 169). Beyond the challenge of knowing which variants are functional, a lack of information connecting each element to its target gene(s) obscures the effects of functional variants on downstream cellular and organismal processes.

Systematic annotation of regulatory elements has been undertaken by large consortia [e.g., ENCODE (48), the Roadmap Epigenomics Consortium (157), the GTEx project (82), the Human Cell Atlas (12), FANTOM5 and FANTOM6 (52, 151), and HuBMAP (91)], with a growing consideration for disease-relevant cell types and tissues at differing developmental stages and under varying external influences. Quantitative traits such as chromatin accessibility, transcription factor binding, and gene expression, combined with genotyping or whole-genome sequencing, have provided maps of quantitative trait loci linking gene regulatory traits to genomic regions (19, 30, 82, 105, 177). These efforts have relied on technological, analytical, and conceptual advances and are increasingly shifting to single-cell resolution (188). Furthermore, consortia-based efforts such as the 4D Nucleome Project (41) have started to annotate the spatial organization of the genome in an effort to understand long-distance transcriptional regulation. Sequence–function studies targeting whole genetic elements rather than genetic variants have also been useful, at increasing scales and across varying cell type and cell stage dependencies, for connecting sets of putative regulatory elements to the genes they regulate and to downstream effects (44, 142, 165). However, these have typically lacked the resolution to identify variant-level effects. Fortunately, high-resolution mapping of regulatory elements in which all possible nucleotide substitutions are interrogated is proving useful in detecting the specific nucleotides that are essential for the activities of these elements (108, 135).

In the case of coding variation, both predicting and interpreting the effects of amino acid substitutions on protein fitness are key to our understanding of molecular function and evolution. Sequence–function studies are powerful tools that can reveal underlying biological mechanisms at scale (64). In the field of structural biology, advances in in vitro and in silico methodologies, for example, high-resolution experimental structure determination (16, 183) and structure prediction by DeepMind's AlphaFold (98), have made high-quality protein structures for much of the human proteome widely accessible (190, 201). Combining these with VE maps can further the analysis of structural and functional features, identifying at high resolution regions enriched for intolerance to perturbations. Measurements of variant activity from sequence–function studies have also been used to predict the three-dimensional structure of protein domains (158, 164) and to discriminate among alternative predicted structures (3).

Clinical Variant Classification

The application of large-scale genetic sequencing as a clinical tool has been limited by our inability to infer the functional impact(s) of most genetic variation observed in humans. Indeed, most variants in the ClinVar database (120), which captures clinical variant reports, have been

annotated as variants of uncertain significance (VUSs), meaning that it is unknown whether the variants cause a clinical consequence (171). Many sources of evidence for variant interpretation cannot be readily scaled to match the pace of variant discovery (e.g., family linkage studies) or are less useful for rare variants (e.g., population-level association studies). Functional investigation is often "reactive," in that it is carried out after a genetic variant is identified in the clinic. This model, however, can take months or years to move from variant identification to functional assessment for patient diagnosis. Large-scale sequence–function studies can instead be "proactive" (54, 60), in that exhaustive testing is done for all possible genetic variants, providing functional evidence often in advance of a first clinical presentation.

Recently, the first proactive clinical use of a sequence–function map was reported in pediatric patients presenting with cardiac arrest (60). In one case, clinical testing identified two VUSs in *cis* in the calmodulin 2 (*CALM2*) gene, which encodes a calcium-binding protein, loss of which is associated with cardiac arrhythmias (long-QT syndrome) and catecholaminergic polymorphic ventricular tachycardia (37, 145). Neither variant had previously been observed in the population, making variant interpretation difficult. By consulting a sequence–function map covering nearly all possible missense variants in the calmodulin protein (encoded by *CALM1*, *CALM2*, and *CALM3*), clinicians could conclude that the allele was a contributing cause of the patient's disease phenotype (60, 196).

Clinical variant classification has thus far been most widely applied in the coding portions of the genome for genes linked to medically actionable outcomes (168). Although gene regulatory noncoding variants are frequent causes of both common and rare diseases, clinical variant interpretation for noncoding variants has largely been limited to variants in canonical splice sites and untranslated regions of protein-coding genes (47, 120). Notable exceptions include large-effect noncoding variants in founder populations, such as variants in multiple genes associated with plasma lipid traits (92), and rare large-effect noncoding variants such as those in *MEF2C* causing severe developmental disorders (198) and in *GJB1* causing X-linked Charcot-Marie-Tooth disease (185).

Genetic variants as cataloged in ClinVar and the Human Gene Mutation Database (HGMD) (176) have historically been labeled qualitatively (e.g., pathogenic versus benign) rather than quantitatively (i.e., by magnitude of effect). Noncoding regulatory variation lends itself to the quantitative mapping of effects at the level of transcriptional activity and context-dependent impact on molecular and cellular phenotypes. Examples of such context-dependent impacts include the *FTO* obesity risk locus controlling expression of long-distance target genes *IRX3* and *IRX5* in adipocyte progenitor cells and hypothalamic neurons (34, 170), a vascular disease–associated genetic variant controlling *END1* gene expression in endothelial cells (83), and genetic variants linked to *SORT1* gene expression levels in the context of low-density lipoprotein levels (140).

The clinical annotation of noncoding gene regulatory variants is currently sparse. This is, however, likely to change as whole-genome (as opposed to whole-exome or gene panel) sequencing becomes more commonplace in clinical diagnostics and as systematic sequence–function studies provide proactive evidence to help classify noncoding variants (47).

Progress Toward a Human Variant Effect Atlas

Foundational technologies and sequencing advancements over the last decade have enabled the disease genetics community to study the effects of genetic variation on thousands of nucleotides across the human genome (**Figure 1**). To facilitate coordination and collaboration across more than 100 research labs engaged in systematic sequence–function studies, the MaveRegistry forum enables researchers to share progress on ongoing efforts (38). Examination of MaveRegistry

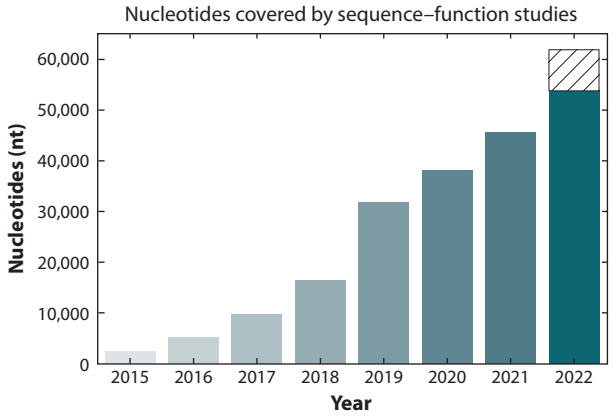

Figure 1

The number of nucleotides in the human genome covered by sequence–function studies—approximately 80% of these are coding nucleotides. Here, we include studies that measure the effects of at least 50% of all possible changes in the relevant coding or noncoding genetic element. For 2022, the solid bar indicates nucleotides covered up to and including June; the dashed lines are an extrapolation for the rest of the year.

suggests that the journey to a complete VE atlas has only just begun; indeed, with maps for less than 100 protein-coding genes published or underway, fewer than 1% of human genes have been covered. There is still more to do for noncoding regions, for which even fewer VE maps have been produced so far.

To store and disseminate sequence–function data, the MaveDB repository was created, with subsequent efforts to comprehensively curate and deposit published data sets (161). MaveDB currently hosts VE maps for only a fraction of human protein-coding genes. Unfortunately, even VE maps described in publications are sometimes unavailable, often because the data could not be accessed. This suggests that deposition of VE map data should be more widely required by journals and funders.

In some sense, this is an overestimate of progress made, considering that sequence–function studies are often carried out on partial sequences and do not capture the entirety of the protein or noncoding element they interrogate. Moreover, VE maps may ultimately need to cover multiple splice isoforms or contain measurements in multiple environments, genetic backgrounds, cell types, or spatiotemporal contexts.

An ecosystem of tools is emerging to support the VE mapping effort. To aid in experimental planning and project selection, MaveQuest aggregates data resources relevant to identifying target protein–coding genes and functional assays (115). Guidelines have been put forward for experimental design, reporting, and evaluation of data (70) and to prioritize genes for sequence–function studies based on their potential for improving variant interpretation in the clinic (114). Software pipelines have been generated for analyzing VE maps (17, 53, 63, 89, 159), and multiple tools exist to visualize data tailored specifically to both coding (90, 200) and noncoding sequences [the Human Genetics Amplifier (HuGeAMP; **https://kp4cd.org**) and the Common Metabolic Diseases Genome Atlas (CMDGA; **https://cmdga.org**)].

The field of sequence–function studies is rapidly growing, with new initiatives still emerging (e.g., the Atlas of Variant Effects Alliance, the International Common Disease Alliance, and the National Human Genome Research Institute's Impact of Genomic Variation on Function Consortium). This is perhaps unsurprising, considering the need for functional, biophysical, and mechanistic data across both variant and network levels. Although the path to a comprehensive

atlas of human variant effects will be long, we are at an inflection point for V2F studies with a plethora of foundational tools in place and promising signs of community coordination emerging.

CONSIDERATIONS FOR SEQUENCE–FUNCTION STUDY DESIGN

In designing a sequence–function study, the criteria for selecting both a target (e.g., a protein, splice site, or regulatory element) and an experimental system will differ widely among individual researchers. For instance, in selecting a target for study, a clinical geneticist might prioritize targets with a substantial VUS burden and for which definitive classification would impact therapy, a molecular geneticist might prioritize a target if it is representative of a class of proteins or regulatory elements for which sequence–structure–function relationships have been less explored, and a systems biologist might prioritize targets that are hubs in a gene regulatory network or signaling process. Though the choice of experimental system is somewhat limited by the target in question, it will be determined largely by researcher-specific interests and constraints, of which there are many. We suggest the use of the three design axes of tractability, fidelity, and granularity as a framework for choosing among alternative model systems and experimental methods.

Tractability reflects the ease of carrying out the assay at scale, which should be evaluated in terms of project duration and availability of necessary infrastructure, as well as resource needs that include supplies, sequencing, and labor costs—in other words, whether the assay can be economically scaled to interrogate many variants. This axis should similarly capture the likely requirements for engineering and validating the assay and should incorporate assay robustness, for example, by capturing estimated failure rates.

Fidelity refers to how well the results of a scalable variant assay correlate with the relevant human phenotype (which may range from the molecular to organismal level). Not every model will capture the full range of a protein's functions, some of which will be more relevant for the disease phenotype. Evaluation of fidelity should extend to how reliably the assay will reflect the human impacts of a variant in different environmental, developmental, and genetic contexts, as assay parameters are varied to approximate these contexts.

Granularity considers the extent to which the assay provides insight into the mechanism(s) of action for a given variant, where a comprehensive understanding of variant effects may require multiple assays that collectively probe variant effects at molecular, cellular, and organismal levels. In the context of gene regulatory variants, this may include the extent to which the assay informs us about steps in the regulatory flow from variant impact to gene product in a given spatiotemporal context.

Choosing an Assay to Inform Variant Classification

The impact of a sequence–function study as evidence for variant classification will be dependent largely on both the target and the assay of choice—we discuss clinical genetic factors that influence assay prioritization. Standards for clinical variant interpretation were established by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (154). These are consensus-based guidelines through which a variant is classified on a discrete spectrum from pathogenic to benign on the basis of available evidence or as a VUS in the event of uncertainty. Clinical and population data, as well as in silico and in vitro studies, are assigned different evidence strengths in favor of a classification. These guidelines consider functional studies to be strong evidence in support of a pathogenic (evidence code PS3) or benign (evidence code BS3) variant impact. Notably, more recent recommendations have described validation approaches that can deem functional evidence as very strong in certain cases (23, 70). Thus, large-scale functional studies have great potential to aid in clinical variant classification at scale.

The ACMG/AMP guidelines for PS3/BS3 evidence codes suggest that functional data be considered only from "well-established" assays, although no criteria for these are provided (154). The Clinical Genome Resource (ClinGen) has since introduced expert panels to refine variant classification guidelines (156), though definitions of the well-established assay provided by these panels have varied widely depending on the gene and disease, suggesting that this process can be highly subjective (99). More recently, objective guidelines for the production and reporting of sequence–function data have been suggested, with statistical procedures to empirically validate evidence strengths (22, 70).

To maximize the clinical impact of sequence–function studies, one can prioritize target selection by considering actionability and "movability," the likelihood that new functional evidence will update the clinical interpretation of a VUS to one that is more informative (114). Estimates of movability vary by gene and variant interpretation methodology and will change as additional evidence sources (e.g., population-level associations and family linkage evidence) become more abundant (54, 114). Functional evidence will also have a greater impact for variants observed more often in the clinic; therefore, a greater priority might be given to a study target for which the average variant is more frequently clinically observed and medically actionable.

A common method for validating an assay's phenotypic readout is to perform small-scale testing using a gold standard set of clinically annotated variants. However, reported variant classifications may be erroneous and thus should not by default be considered definitive (85, 123). Furthermore, it is important to consider that a variant may have an abnormal function and yet not cause disease: Though abnormal function is a necessary condition, it is not sufficient to classify a variant as pathogenic (154). A reduced-function (hypomorphic) variant may still allow normal cellular or organismal function or may contribute to disease manifestation only in particular environmental, developmental, or genetic contexts.

Surrogate Genetics and Nonobvious Disease Models

A nonobvious but crucially important point in assessing the fidelity of potential assays is that the assay phenotype need not *look like* the human disease phenotype. For example, deficiency in the gene *MTHFR* can cause both a molecular endophenotype (elevated blood homocysteine levels) and a range of organismal phenotypes (including intellectual disability and abnormal walking). An assay based on the ability of human *MTHFR* to rescue the growth of *Saccharomyces cerevisiae* cells lacking the orthologous *MET13* gene can be used to assess variant function, and while it may seem surprising that yeast growth rate should predict intellectual disability and other organismal phenotypes, this assay has been shown to faithfully reflect missense variant pathogenicity (125, 167, 194).

Where high-fidelity assays have been found in unicellular model organisms, they have often proven to be highly useful in sequence–function studies given their tractability (18, 143, 179, 194, 196). Human cell lines are evidently an important single-cell model that can enable scalable variant assays, though the choice of cell line will depend largely on the assay in question (78, 112, 124, 172). Where applicable, readily manipulable cell lines (e.g., HeLa or Hek293 cells) are an attractive option due to their high tractability; these can be employed so long as the target element and any known obligate partners function in the cell line of choice and provided the model reliably separates pathogenic variants from benign controls. In certain cases, however, the most highly tractable systems cannot recapitulate complex disease-related functions. Here, one can choose the most readily scalable model capable of successfully capturing the function in question. Where no cognate phenotype in a unicellular organism or single-cell model can be found, multicellular model organisms or human organoid models may be applied, although not yet at a scale compatible with assessing tens of thousands of variants (5, 106, 132).

The search for surrogate functional assays has largely been restricted to a relatively small set of well-described model organisms. To broaden the search outside of the handful of well-characterized models, sets of mutually orthologous genes can be systematically queried to identify cognate phenotypes or phenologs; for instance, human genes associated with angiogenesis tend to have orthologs in yeast that show lovastatin sensitivity (133). Using this approach, diverse uni- and multicellular surrogate genetic models were uncovered for several human genetic diseases, and it is likely that many high-fidelity assays in lesser-used model systems have yet to be discovered.

The Use of In Silico Predictors

Computational advances over the last decade have enabled the development of increasingly sophisticated computational models to predict the impact of both coding and noncoding SNVs. For coding variation, most models are specialized to predict the impact of missense variants on protein function. These often use machine learning to exploit evolutionary conservation, trends in the amino acid substitutions that occur most frequently across protein families, and protein structural features (4, 32, 65, 94, 122, 144, 180, 199). The use of predictive models is an attractive option for efficiently inferring coding variants exome-wide, especially given that a high-quality experimental atlas of sequence–function studies for all human disease-associated proteins and noncoding elements remains largely incomplete. The utility of computational predictors in clinical variant classification, however, remains limited considering the current ACMG/AMP guidelines, which suggest that it be considered supporting evidence at best (154). Large-scale sequence function studies could help refine and validate computational models, improving their accuracy and interpretability and potentially providing more confidence in their clinical utility.

Variant effect prediction has been expanded outside of coding regions to include noncoding gene regulatory SNVs and short insertions and deletions genome-wide (107, 122, 153). Extensive genome-scale measurements of expression, chromatin accessibility, and epigenetic marks have fueled the development of models to predict gene regulatory variant effects from DNA sequences (13, 14, 29, 40, 103, 204). For noncoding gene regulatory SNV effect predictions, sequence-based machine learning models are trained to predict phenotypic outcomes from a sequence on the basis of genome-wide gene regulatory studies, including studies of chromatin accessibility and chromatin immunoprecipitation sequencing (ChIP-seq) that inform about transcription factor binding and histone marks, as well as Cap-Analysis of Gene Expression (CAGE) annotations of gene expression.

While computational models, in combination with statistical fine-mapping and GWAS-QTL (genome-wide association study–quantitative trait locus) colocalization approaches, have begun narrowing associated noncoding regions to driver nucleotides, the resolution and interpretability of such models remain limited. Despite the richness of genome-scale catalogs of expression and epigenetic marks, high-resolution genome-wide gene regulatory data sets assayed across individuals and large-scale ground-truth evidence are still missing. Thus, models are often still insufficient to provide compelling, high-confidence proof for a chain of causation that stretches from variant to regulatory element to gene to downstream phenotype.

Sequence–function studies and computational predictors already complement one another in the sense that, as orthogonal sources of evidence, they are considered independent and can both be applied to interpret a given variant. There is, however, an opportunity to increase the value of complementarity between computational and experimental variant analysis. For instance, some computational predictors are beginning to incorporate VE maps in training (155, 199). Moreover, functional assays could be directed toward blind spots that prove resistant to computational

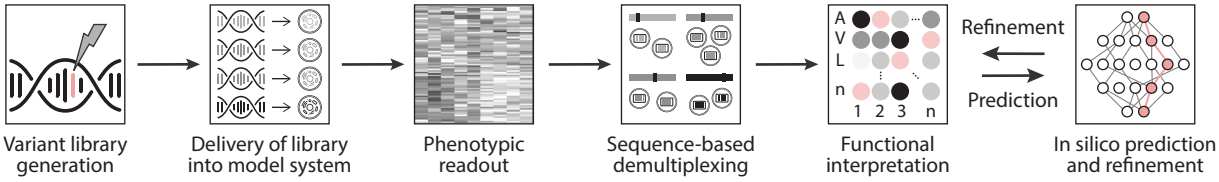

Figure 2

Pipeline overview and methods of multiplexed sequence–function studies. Though each study will vary greatly, they follow a general flow: generation of a variant library, delivery of the library into a model system, phenotypic readout, sequence-based demultiplexing, and functional interpretation, with the option for further in silico prediction and refinement.

prediction (e.g., disordered regions, gain-of-function effects, and positions under strong positive selection).

EXPERIMENTAL METHODS FOR SEQUENCE–FUNCTION STUDIES

Though there is considerable variation in the design of sequence–function studies, they generally share a common flow: generation of variant library or library of guide RNAs for in situ editing, delivery of library to a model system, phenotypic readout, sequence-based demultiplexing, functional interpretation, and optional in silico prediction and refinement (**Figure 2**; **Table 1**). Here, we review each step in the pipeline, bar prediction and refinement, with a particular emphasis on phenotypic readouts. The methods discussed herein are by no means an exhaustive list.

Table 1 Methods discussed in this review are summarized for each stage of a multiplexed sequence–function study

Stages of multiplexed sequence–function studies		Summary of methods discussed in this review
Variant library generation		Error-prone PCR (26, 108), oligonucleotide-targeted editing (59, 89, 95, 109, 116, 196), DNA synthesis (35, 58, 136, 149), genome fragmentation and DNA capture (9, 187, 189, 192, 193)
Library integration into the model system		Targeted integration (128, 129), random integration (79, 124), transient episomal delivery (9, 135), in situ editing (27, 50, 57, 58, 117)
Phenotypic readout	Variant effect on regulation	Connecting regulatory elements to effector genes (44, 68, 69, 165), functional dissection of regulatory elements (9, 189, 192, 193), assessment of variant impacts on regulation (97, 108, 135, 148, 163, 197)
	Variant effect on protein function	Assessment of protein stability and abundance (112, 127), measurement of protein interactions (173, 205), assay of specialized protein function (28, 78, 88, 96, 173, 175)
	Variant effect on cellular phenotype	Fitness-based assays (18, 21, 74, 81, 143, 196), assays of morphology or specialized cellular function (54, 86, 100, 118, 184)
Sequence-based demultiplexing		Direct sequencing of variant alleles (45, 62, 196), sequencing of variant-associated barcodes (7, 131, 196)
Functional interpretation		Computational frameworks for coding variant studies (17, 53, 63, 89, 160) including imputation of unmeasured variants (80, 196), and MPRA-type studies (9, 10, 49, 79, 193)

Abbreviations: MPRA, massively parallel reporter assay; PCR, polymerase chain reaction.

Generating a Variant Library

The first step in a sequence–function study is typically the generation of a variant library carrying nucleotide or amino acid substitutions in the target element (frequently termed saturation mutagenesis). Some libraries are composed of full-length constructs to be introduced later into cells, while other libraries may be collections of reagents enabling in situ mutagenesis.

A straightforward approach to saturation mutagenesis uses error-prone polymerase chain reaction amplification (26). Although this method can be biased toward a subset of nucleotide changes (transitions over transversions), some methods can yield a more balanced mutational spectrum (108). A drawback of this method, however, is that it tends to provide only single-nucleotide substitutions, which, while acceptable for noncoding regions, would fail to encode most of the possible amino acid substitutions in a coding sequence of interest. Although coding variants observed in humans, as well as other species, are primarily single-nucleotide changes, analysis of the full range of amino acid substitutions can provide further insight into the biochemistry underlying a protein's function.

Oligonucleotide-targeted editing is a more flexible, albeit more expensive, option in which libraries of mutagenic oligonucleotides encoding the desired substitutions are synthesized. Variations of this method include POPCode (196), Kunkel (116), PFunkel (59), EMPIRIC (89), PALS (109), and inverse polymerase chain reaction mutagenesis (95). Mutagenic libraries can be synthesized to directly encode the desired substitutions or designed with degenerate sequences, by which all possible substitutions can be reached cost-effectively. For coding libraries, certain degenerate codons (i.e., NNS and NNK) have the advantage of encoding all possible amino acid changes while reducing the frequency of stop codons (relative to NNN mutagenesis), though the use of multiple codons for an amino acid or stop codon substitution at a given position can provide the advantage of internal replication.

Libraries can be also synthesized, allowing for precise control of library composition. Restrictions on fragment length, however, have limited this approach to shorter elements and guide RNA libraries or to methods requiring further assembly in the case of larger constructs (35, 58, 136, 149). Continued advances in DNA synthesis will make this an increasingly tractable option for longer sequences. Alternatively, combining genome fragmentation and DNA capture methods, as often applied in the context of massively parallel reporter assays (MPRAs) and STARR-seq-based assays, can generate high-coverage libraries from genomic DNA (9, 187, 189, 192, 193).

In situ editing for saturation mutagenesis at the endogenous target genomic locus has also been applied using CRISPR-related technologies. In this approach, the synthesized guide RNAs [single-guide RNAs (sgRNAs) or prime editing guide RNAs (pegRNAs)] are cloned into a delivery vector under an RNA polymerase III promoter (e.g., U6) and delivered along with the CRISPR-Cas effector of choice. These approaches are further described in the delivery section below.

There are many applications for libraries with variant combinations. Such libraries may cover single-target elements that carry multiple variants, enabling one to solve protein structures (158, 164), as noted above, or to understand genetic interactions between common and rare variants (194). Alternatively, libraries for two different elements may be usefully combined; for example, two different proteins can be mutagenized simultaneously to identify mutually compensatory changes and other genetic interactions across a protein interaction interface (43).

Delivery of the Variant Library

A variant library can be delivered to a model system of choice by integration into the genome at a safe harbor site, by transient introduction on episomes, or by in situ editing. Considerations for the method of delivery will depend on whether the assay is sensitive to the number of variant and

wild-type alleles and on whether assay fidelity demands the assessment of variants in their native genomic context.

To model phenotypes, it is important to consider the number of alleles at each relevant locus. Where the variant effects under study are phenotypically dominant in the assay, the presence of additional wild-type alleles may be tolerated and indeed may yield more accurate measures of allelic effects (e.g., where one allele has a dominant negative impact on the others) (134). For other assay systems, only one variant allele should be delivered per cell to faithfully associate variant effects with single-cell phenotypes. Where variant effects are recessive, the variant allele should either be the only expressed, or otherwise active, allele of the target element in its cell or be homozygous.

A popular method of variant integration is the Bxb1 landing pad system, designed to ensure that only a single variant-expressing construct is integrated into each cell (128, 129). This system also allows for endogenous alleles of the target to be eliminated where necessary. Alternatively, lentivirus delivery at a low multiplicity of infection can circumvent the need to engineer a landing pad cell line (124). Varying genomic integration sites due to the random integration of lentivirus can, however, lead to variable expression, such that this approach may require averaging over more integrants or be lessened by the introduction of antirepressor elements (79). Additionally, in assays involving state changes, such as cellular differentiation, silencing of lentiviral integrants is an important concern, and here again, a landing pad system integrated at a ubiquitously expressed locus, such as AAVS1, can be beneficial.

In certain applications, it is tolerable, or even desirable, to express many variants of the target of interest within the same cell. This includes assays in which the phenotype of interest is a transcript level, where the variants of interest are in *cis* with transcript-encoding DNA and where the identity of the variant is encoded by the transcript (108, 135, 149). To express many variant constructs in a given cell, one might use extrachromosomal delivery of the library on episomes or genomic integration by lentivirus (110). Episomal expression, however, may fail to capture certain epigenetic effects (e.g., long-range interactions and transcription factor binding) where episomes are not subject to the same mechanisms that control genomically integrated elements (110). Here, endogenous integration can capture more context-dependent effects and is useful when working with cell lines that yield low transfection efficiencies (79).

As noted above, saturation mutagenesis can be performed in situ by directly editing the endogenous locus. To identify regulatory elements, tiled Cas9-mediated cleavage can yield many deletions due to nonhomologous end-joining (NHEJ) repair, thereby creating random disruptions across a target region (27). Where the goal is to understand the impact of specific variants, saturation genome editing can be performed by Cas9-mediated cleavage in the presence of donor templates, using homology-directed repair (HDR) to generate integrated variant libraries at an endogenous locus (57, 58). Given limited HDR efficiency, however, this approach may not be compatible with all cell lines and may yield deletions due to NHEJ (36). Saturation editing can also be performed by fusing a catalytically inactive (dead) Cas9 (dCas9) to a base editor, but this will not generate all possible changes in equal proportions. For example, the cytosine deaminase base editor yields transition mutations (C:G to A:T), but only rarely other substitutions (111, 117). Recently, prime editing, in which a catalytically impaired Cas9 is fused to a reverse transcriptase, was extended to saturation editing in a human cell line by encoding a library of programmed guide RNAs (gRNAs) that simultaneously specify the target site and encode the desired mutation (8, 50). This can be applied in cells with inefficient HDR repair, and while all Cas9-based approaches require the availability of a protospacer-adjacent motif near a target site, this constraint is more relaxed for prime editing.

a

Variant function assessed	Exemplar assay
Transcriptional regulation	**1** – Massively parallel reporter assay
Posttranscriptional regulation	**2** – Minigene splicing assay
Protein stability and abundance	**3** – Antibody labeling, VAMP-seq
Protein–protein interaction	**4** – Yeast two-hybrid system
Specialized protein function	**5** – Kinase or ligand transport assay
Holistic function of genetic element	**6** – Growth rescue in mammalian cell

b

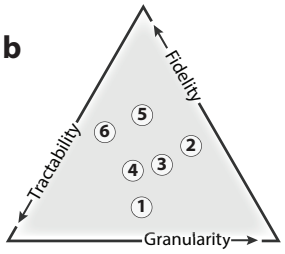

Figure 3

(*a*) Variant effects measurable by sequence–function assays, with common exemplar assays for each (*b*) placed in a ternary plot within the design axes of tractability, fidelity, and granularity to illustrate the inherent trade-off between these features. Abbreviation: VAMP-seq, variant abundance by massively parallel sequencing.

To prevent edits from generating multiple alleles with different variants, in situ editing can be carried out in haploid cell lines, such as the near-haploid Hap1 line (57, 58). Alternatively, individual loci can be haploidized; for example, in situ editing of *NPC1* was carried out at a triploid locus in the HEK293T cell line by introducing deletions to two of the three alleles (50).

Phenotypic Readout

Design and validation of a functional assay is often the most challenging part of a sequence–function study, and an astounding breadth of phenotypic readouts have been used to dissect the molecular, cellular, and clinical effects of both coding and noncoding variants. Here, we classify assay selection for variant interpretation by whether it is assessing the effects of genetic variants on gene regulatory, molecular, and cellular phenotypes. Assays may also range along a spectrum from bespoke (i.e., tailored to an individual protein or noncoding element) to generalizable (i.e., widely applicable to a range of targets).

In **Figure 3**, we place exemplar assays within the design axes of tractability, fidelity, and granularity, illustrating the inherent trade-off between these features. Although quantifying these design parameters is subjective and to a degree case dependent, this conceptual exercise can be helpful in weighing the relative merits of alternative assays.

Variant effects on regulation. The mapping of functional elements within the noncoding genome is an ongoing process, with millions of enhancers and promoters, many with known relevance to diverse human cell types and tissues, already annotated. These annotations, however, provide limited insight into key driver nucleotides and their downstream effects. To detect new elements, connect predicted elements to effector genes and downstream functions, and identify the impacts of noncoding variation within these elements, high-throughput functional assays will be critical.

CRISPR-mediated in situ perturbations have been widely used to connect regulatory elements to their downstream effector genes and functions and, when applied in an inducible system, can do so across various cell stages (75, 76, 87, 147, 150, 182, 202). By combining multiplexed perturbations with scRNA-seq as a readout (e.g., Perturb-seq), transcriptomic profiles can be used to connect regulatory elements to downstream effects and can do so at increasing efficiency and sensitivity when measuring the expression of predetermined targets (e.g., TAP-seq) (2, 39, 44, 69, 126, 165). This was applied to obtain genome-scale transcriptome phenotypes for 200 TP53 and KRAS variants and should prove useful as a generalizable system for discovering more tractable reporter assays to affordably measure the impacts of all variants in target proteins or noncoding elements (186). Alternatively, CRISPR-FlowFISH quantitatively measures the effects of

perturbations in candidate noncoding regulatory elements on the expression of select genes by RNA fluorescence in situ hybridization (68, 71, 152). Recent efforts to tile noncoding regulatory elements by CRISPRi (CRISPR interference) (67) and CREST-seq (42) have showcased the utility of deploying dense tiling as a method to narrow gene regulatory elements down to functional noncoding variants. To further extend readouts to multiple phenotypes simultaneously, methods such as ECCITE-seq can combine perturbations with the measurement of transcriptional state and protein levels (of specific cell surface proteins) in single cells (138).

To quantitatively assess the regulatory activity of noncoding elements, STARR-seq uses self-reporting episomal constructs. In this approach, a sequence library (often of randomly fragmented genomic DNA) is inserted into the 3′ untranslated region (UTR) of a reporter gene, allowing fragments to drive their expression and act as their own transcribed barcodes (9). STARR-seq has limited throughput, however, such that it is not scalable to study saturated libraries from larger genomes. To circumvent this limitation, Cap-STARR-seq makes use of custom microarray capture to select and assay a predetermined set of known or predicted regulatory elements (189). Similarly, ChIP-STARR integrates a ChIP-seq library into the STARR-seq system, allowing interrogation of elements that may be bound by specific DNA-binding proteins (192). To enable genome-wide testing of regulatory regions, HiDRA selectively fragments genomic DNA at regions of open chromatin to densely survey putative transcriptional regulatory elements using the STARR-seq system and thus enable the identification of regulatory elements at up to ∼20-nt resolution (193).

In MPRAs, variant libraries are generally synthesized as oligonucleotides (e.g., on programmable microarrays) and the barcoded sequences are integrated into a reporter construct to enable the efficient dissection of transcriptional regulatory elements (149). With saturation mutagenesis, MPRAs can efficiently reveal the effects of single and combinatorial nucleotide changes on regulatory activity (48, 104, 108, 135) and have been applied in human embryonic stem cells (93, 113) and in vivo, for example, to liver-specific enhancers in mice (148). To map functional elements in untranslated regions, a library of 3′ or 5′ UTR sequences can be inserted downstream of a reporter gene, and the effects of sequence variants on mRNA abundance and stability, as well as on protein production, can be assessed (163, 203). To measure the effects of variants on splicing, minigene assays can be used in combination with libraries spanning multiple intron-exon junctions (15, 20, 72, 97, 102, 197). For MPRAs, and other methods where detecting the phenotype and demultiplexing are both accomplished by sequencing, there is little distinction between phenotypic readout and sequence-based demultiplexing.

Variant effects on protein function. Both coding and noncoding genetic variation can impact protein abundance and alter protein function. Amino acid changes can affect protein function either directly by altering a specific role (e.g., a molecular interaction or enzymatic activity) or more generally by causing an overall loss of stability. Variation in a transcription-regulatory element can impact protein abundance (and function) by altering the expression levels of the element's effector genes. Variation in splice sites, splicing enhancers, or splicing silencer motifs can also change a protein's abundance (e.g., via nonsense-mediated decay) or more radically alter a protein's sequence and structure (e.g., via exon skipping or intron retention).

Phenotypic readouts have been applied to interrogate a wide range of molecular functions, including ubiquitin ligase activity (173), ion channel function (77), protein aggregation (18, 81), ligand interactions (88), DNA mismatch repair (96), and virus–host protein interactions (28, 175). New assays can be specifically tailored to a given function where this is required to achieve high fidelity. Developing and validating these assays can, however, be resource intensive, particularly when considering that they may not be generalizable to other targets. Alternatively, assays that can be more broadly applied are an attractive option with lower recurrent development costs.

An important class of generalizable assays measures variant effects on protein abundance. To assess protein stability directly, VAMP-seq employs a fluorescent reporter construct fused to a protein of interest, such that variants can impact the abundance of the fluorescent reporter via effects on translation rates, folding, stability, and degradation (127). Abundance can also be measured via intra- or extracellular labeling with fluorescent antibodies (112). Where the protein of interest affects the abundance of a secondary protein, this can be harnessed as a generalizable proxy assay (124) and can also be a useful source of generalizable assays for noncoding elements (71). Furthermore, generalizable variant assays may yield multiple phenotypes; for example, ECCITE-seq simultaneously probes the impact of noncoding variants both on the abundance of specific proteins and, more broadly, on transcriptomic profiles (138).

Another class of generalizable assays measures the impact of variation on protein–protein interactions, for example, using the yeast two-hybrid system (56). Here, two proteins of interest are fused to complementary fragments of the transcription factor Gal4. Binding of the two proteins promotes reconstitution of the fragments, such that the reconstituted Gal4 drives the expression of a selectable reporter gene, which is often an auxotrophic or fluorescent marker (46, 174). This tractable system has been used to quantitatively analyze the impacts of variants in the RING domain of *BRCA1* on interaction with the BARD1 RING domain (173). In another example, a system employing a split RNA polymerase biosensor in phages was applied to assess protein interactions (205).

Several studies have also used both generalizable and bespoke assays to gain a more comprehensive understanding of the effects of the mutational landscape on protein function (7, 31, 178); for instance, the tumor suppressor PTEN has been examined with respect to both cellular abundance (by VAMP-seq) and enzymatic activity (127, 137). Where high-fidelity protein-specific assays are not easily scalable, one can leverage a generalizable assay on all variants while performing more resource-intensive bespoke assays on key residues at a smaller scale.

Variant effects on cellular phenotype. Linking genetic variation to disease-relevant cellular phenotypes and programs remains an unmet need in medical genetics. Fitness-based assays, where an element of interest is required for the model's overall fitness, can be harnessed in competitive growth assays, where variants conferring a fitness advantage tend to increase in frequency throughout the population (21, 74). Growth as a selection is useful for several reasons. First, it is easily scalable to high coverage of large variant libraries without relying on expensive machine time (e.g., cell sorting). Second, in the case of assays relying on synthetic growth defects or essential gene complementation, the entire range of a protein's function is at work in the model. Growth-based functional complementation in yeast, in which a human gene rescues the fitness effect conferred by deletion of its ortholog, has been widely used as a phenotypic readout (179, 194, 196). Growth phenotypes in yeast have also been used to define models of unstructured proteins, in which certain conformations of the protein cause dose-dependent toxicity (18, 143). While growth and resistance screens are straightforward and cost-effective, many cellular programs cannot be captured with such assays. Pooled CRISPR interference and activation screens combined with cell sorting–based readouts such as activation of the unfolded protein response (2) or other sortable phenotypes have proven useful to trace both coding and noncoding genetic elements to cellular processes. Cellular readouts such as insulin secretion have further been used to screen type 2 diabetes candidate genes identified in genome-wide association studies in a human pancreatic β-cell line (184).

Distinct image-based cellular phenotypes or morphologies can be similarly effective in capturing a protein's function. Image-based profiling can facilitate the systematic analysis of a diverse

set of spatially resolved cellular and subcellular phenotypes in a scalable format. A recent study, applying high-content fluorescence microscopy to measure the effects of gene perturbations on the yeast endocytic pathway, found that ~30% of the more than 5,000 yeast genes perturbed yielded mutant phenotypes, nearly half of which displayed multiple phenotypes (130). Pooled optical screening can be used in human cells to connect genes, gene regulatory elements, and alleles with rich cellular and subcellular phenotypes and, when combined with in situ sequencing, can resolve dynamic processes in real time (55). Furthermore, optical screening can detect morphological pleiotropy and penetrance, particularly when applied at the single-cell level. Rapid image analysis together with targeted cell labeling by illumination of a photoconvertible fluorescent protein enables cell sorting on the basis of morphology at increasing scale (86, 100), and recent technological advances may soon allow morphology-based cell sorting more directly (146, 166). As platforms improve, pooled optical screening will become an ever more attractive option for assaying variant effects, capturing increasingly complex morphologies at scale (118).

Sequence-Based Demultiplexing

The choice of next-generation DNA sequencing technology used for readout will generally depend on the length of the genetic element in question. Where shorter constructs are concerned, short-read sequencing can be applied (45, 62). Duplex sequencing, in which both strands of each molecule are sequenced, can greatly increase variant calling accuracy. To cover longer stretches of DNA, each of potentially many shorter segments of a larger construct of interest can be sequenced separately (e.g., using the TileSeq approach (196). This approach will, however, fail to detect cases in which a variant within the sequenced segment is connected in *cis* with variants outside the segment, and therefore is less suited to studies looking to understand the impacts of multiple variants per clone.

Alternatively, long-read sequencing can be applied to sequence longer constructs. Though the current cost of long-read sequencing approaches makes them less appropriate for measuring variant frequencies directly, they can be cost-effective when combined with DNA barcodes. Here, the genotype of each clone is connected to a unique barcode sequence by long-read sequencing so that barcodes alone need to be sequenced to read out the results of a multiplexed assay (7, 131, 196).

Functional Interpretation

Following sequencing readout, variant counts must be converted to functional scores. Several computational tools have been designed to handle inputs from sequence–function studies of coding nucleotides, refining outputs (17, 53, 63, 89, 159) and imputing missing scores for unmeasured residues (80, 196), and for MPRA-like studies of noncoding nucleotides (9, 10, 49, 79, 193). Importantly, sequence–function studies result in continuous distributions of effect sizes, and so functionally normal and abnormal variant benchmark sets should be used to calibrate these distributions. For assays on coding variants, functional scores can be rescaled based on the scores of nonsense and synonymous variants. Assays of noncoding elements can be rescaled with known damaging variants or relative to the distribution of random sequences. Important considerations for functional assay scoring and downstream interpretation include the estimation of errors or confidence intervals and procedures for excluding less-well-measured variants (53, 159, 196). Furthermore, score calibration can differ depending on the goals of variant assessment, for instance, estimating the probability of pathogenicity (60) rather than the quantitative biophysical effects on protein function (18, 73).

CONCLUDING REMARKS

Multiplexed sequence–function studies have thus far been applied to hundreds of thousands of genetic variants, yielding substantial clinical, molecular, and biophysical insights. In the process, an impressive collection of experimental methodologies has been developed to enable these experiments at ever-increasing tractability, fidelity, and granularity.

To realize the full potential of these technologies, certain remaining challenges must be addressed. First, while publicly available databases have been developed, the data from many sequence–function studies remain difficult to access. Open data sharing is critically important for the impact of sequence–function studies. Second, sequence–function studies for variant classification have largely focused on Mendelian traits, but understanding how individual variants influence penetrance and expressivity and further interact with other genetic and environmental factors will ultimately be required. Third, for relevance to diverse populations, the genetic contexts of ancestry and sex should be considered in sequence–function studies (162). Finally, high-quality experimental assessments will inevitably conflict with clinical annotations, and these discordances cannot be easily resolved without the need for further interrogation in higher-fidelity models or by comparison to independent patient cohorts.

A complete functional annotation of sequence variants may require phenotyping in specific relevant cellular contexts. For instance, screening of variants in some cell lines might fail to capture functional differences that depend on physiologic or tissue contexts, and these dependencies are often initially unknown. Addressing this aspect, while also dealing with the intrinsic heterogeneity of most multicellular biological systems, remains another overarching challenge.

It is important to note that even small functional effects may be of great interest. Indeed, a small impact on the activity of a given gene may have a profound qualitative effect on a cellular phenotype, and a modest cellular phenotype may correspond to a dramatic organismal phenotype (139). Small effects may also be important where cellular and organismal phenotypes result from the convergence of impacts at many loci, such that even large effects may result from additive polygenic combinations (51). Therefore, the interpretation of a given personal human genome will require integrating the functional impacts of many variants.

It is exciting to consider that an atlas of VE maps covering both coding and noncoding variation could already be generated with currently available technology. Though the pieces to do this are largely in place, linking the full complexity of genetic variation to human phenotypes will necessitate a concerted and interdisciplinary community effort.

DISCLOSURE STATEMENT

F.P.R. advises and holds shares in SeqWell Inc., BioSymetrics Inc., and Constantiam Biosciences Inc., and is a Ranomics Inc. shareholder. P.M. is a scientific cofounder of Shape Therapeutics, Navega Therapeutics, Boundless Biosciences, and Engine Biosciences. The other authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We acknowledge many members of the Atlas of Variant Effects Alliance for their thoughtful input, and we thank Leslie Gaffney for her valuable contributions to the figures in this review. M.C. is supported by the Novo Nordisk Foundation Center for Genomic Mechanisms of Disease at the Broad Institute of MIT and Harvard (National Institute of Diabetes and Digestive and Kidney Diseases DK1261185).

LITERATURE CITED

1. 1000 Genomes Project Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
2. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, et al. 2016. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 167:1867–82.e21
3. Adkar BV, Tripathi A, Sahoo A, Bajaj K, Goswami D, et al. 2012. Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* 20:371–81
4. Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 76:7.20.1–41
5. Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TFC, Stemple DL. 2011. The future of model organisms in human disease research. *Nat. Rev. Genet.* 12:575–82
6. All Us Res. Program Investig., Denny JC, Rutter JL, Goldstein DB, Philippakis A, et al. 2019. The "All of Us" Research Program. *N. Engl. J. Med.* 381:668–76
7. Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, et al. 2021. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* 108:1735–51
8. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, et al. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576:149–57
9. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339:1074–77
10. Ashuach T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. 2019. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* 20(1):183
11. AVE Alliance Found. Members. 2021. The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution. Zenodo. **https://doi.org/10.5281/zenodo.4989960**
12. Aviv R, Teichmann SA, Lander ES, Ido A, Christophe B. 2017. The Human Cell Atlas. *eLife* 6:e27041
13. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, et al. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18:1196–203
14. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* 53:354–66
15. Baeza-Centurion P, Miñana B, Valcárcel J, Lehner B. 2020. Mutations primarily alter the inclusion of alternatively spliced exons. *eLife* 9:e59959
16. Bai X-C, McMullan G, Scheres SHW. 2015. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* 40:49–57
17. Bloom JD. 2015. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinform.* 16:168
18. Bolognesi B, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B. 2019. The mutational landscape of a prion-like domain. *Nat. Commun.* 10:4162
19. Bonder MJ, Smail C, Gloudemans MJ, Frésard L, Jakubosky D, et al. 2021. Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nat. Genet.* 53:313–21
20. Braun S, Enculescu M, Setty ST, Cortés-López M, de Almeida BP, et al. 2018. Decoding a cancer-relevant splicing decision in the *RON* proto-oncogene using high-throughput mutagenesis. *Nat. Commun.* 9:3315
21. Bridgford JL, Lee SM, Lee CMM, Guglielmelli P, Rumi E, et al. 2020. Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood* 135:287–92
22. Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, et al. 2020. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* 12:3
23. Brnich SE, Rivera-Muñoz EA, Berg JS. 2018. Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks. *Hum. Mutat.* 39:1531–41
24. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47:D1005–12

25. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–9
26. Cadwell RC, Joyce GF. 1994. Mutagenic PCR. *Genome Res.* 3(6):S136–40
27. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, et al. 2015. *BCL11A* enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* 527:192–97
28. Chan KK, Dorosky D, Sharma P, Abbasi SA, Dye JM, et al. 2020. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* 369:1261–65
29. Chen KM, Wong AK, Troyanskaya OG, Zhou J. 2022. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* 54:940–49
30. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437:1365–69
31. Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, et al. 2020. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife* 9:e58026
32. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. *PLOS ONE* 7:e46688
33. Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11:415–25
34. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, et al. 2015. *FTO* obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373:895–907
35. Coyote-Maestas W, Nedrud D, Suma A, He Y, Matreyek KA, et al. 2021. Probing ion channel functional architecture and domain recombination compatibility by massively parallel domain insertion profiling. *Nat. Commun.* 12:7114
36. Cox DBT, Platt RJ, Zhang F. 2015. Therapeutic genome editing: prospects and challenges. *Nat. Med.* 21:121–31
37. Crotti L, Johnson CN, Graf E, De Ferrari GM, Cuneo BF, et al. 2013. Calmodulin mutations associated with recurrent cardiac arrest in infants. *Circulation* 127:1009–17
38. Da K, Weile J, Kishore N, Rubin AF, Fields S, et al. 2021. MaveRegistry: a collaboration platform for multiplexed assays of variant effect. *Bioinformatics* 37:3382–83
39. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, et al. 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14:297–301
40. de Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat. Genet.* 54:613–24
41. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, et al. 2017. The 4D Nucleome Project. *Nature* 549:219–26
42. Diao Y, Fang R, Li B, Meng Z, Yu J, et al. 2017. A tiling-deletion-based genetic screen for *cis*-regulatory element identification in mammalian cells. *Nat. Methods* 14:629–35
43. Diss G, Lehner B. 2018. The genetic landscape of a physical interaction. *eLife* 7:e32472
44. Dixit A, Parnas O, Li B, Chen J, Fulco CP, et al. 2016. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167:1853–66.e17
45. Doud M, Bloom J. 2016. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* 8:155
46. Durfee T, Becherer K, Chen PL, Yeh SH, Yang Y, et al. 1993. The retinoblastoma protein associates with the protein phosphatase type 1 catalytic subunit. *Genes Dev.* 7:555–69
47. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, et al. 2021. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. medRxiv 2021.12.28.21267792. **https://doi.org/10.1101/2021.12.28.21267792**
48. ENCODE Proj. Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
49. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, et al. 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* 34:1180–90
50. Erwood S, Bily TMI, Lequyer J, Yan J, Gulati N, et al. 2022. Saturation variant interpretation using CRISPR prime editing. *Nat. Biotechnol.* 40(6):885–895

51. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, et al. 2020. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* 11:3635
52. FANTOM Consort., RIKEN PMI, CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* 507:462–70
53. Faure AJ, Schmiedel JM, Baeza-Centurion P, Lehner B. 2020. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* 21:207
54. Fayer S, Horton C, Dines JN, Rubin AF, Richardson ME, et al. 2021. Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in *BRCA1*, *TP53*, and *PTEN*. *Am. J. Hum. Genet.* 108:2248–58
55. Feldman D, Singh A, Schmid-Burgk JL, Carlson RJ, Mezger A, et al. 2019. Optical pooled screens in human cells. *Cell* 179:787–99.e17
56. Fields S, Song O-K. 1989. A novel genetic system to detect protein–protein interactions. *Nature* 340:245–46
57. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. 2014. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 513:120–23
58. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, et al. 2018. Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* 562:217–22
59. Firnberg E, Ostermeier M. 2012. PFunkel: efficient, expansive, user-defined mutagenesis. *PLOS ONE* 7:e52031
60. Floyd B, Weile J, Kannankeril P, Glazer A, Reuter C, et al. 2022. Proactive variant effect mapping to accelerate genetic diagnosis for pediatric cardiac arrest. Preprints 2022010177. **https://www.preprints.org/manuscript/202201.0177/v1**
61. Flynn E, Lappalainen T. 2022. Functional characterization of genetic variant effects on expression. *Annu. Rev. Biomed. Data Sci.* 5:119–39
62. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, et al. 2010. High-resolution mapping of protein sequence-function relationships. *Nat. Methods* 7:741–46
63. Fowler DM, Araya CL, Gerard W, Fields S. 2011. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* 27:3430–31
64. Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods* 11:801–7
65. Frazer J, Notin P, Dias M, Gomez A, Min JK, et al. 2022. Publisher correction: Disease variant prediction with deep generative models of evolutionary data. *Nature* 601:E7
66. Freund MK, Burch KS, Shi H, Mancuso N, Kichaev G, et al. 2018. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet.* 103:535–52
67. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, et al. 2016. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* 354:769–73
68. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, et al. 2019. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51:1664–69
69. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, et al. 2019. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 176:377–90.E19
70. Gelman H, Dines JN, Berg J, Berger AH, Brnich S, et al. 2019. Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Med.* 11:85
71. Genolet O, Ravid Lustig L, Schulz EG. 2022. Dissecting molecular phenotypes through FACS-based pooled CRISPR screens. *Methods Mol. Biol.* **https://doi.org/10.1007/7651_2021_457**
72. Gergics P, Smith C, Bando H, Jorge AAL, Rockstroh-Lippold D, et al. 2021. High-throughput splicing assays identify missense and silent splice-disruptive *POU1F1* variants underlying pituitary hormone deficiency. *Am. J. Hum. Genet.* 108(8):1526–39
73. Gersing S, Cagiada M, Gebbia M, Gjesing AP, Cote AG, et al. 2022. A comprehensive map of human glucokinase variant activity. bioRxiv 2022.05.04.490571. **https://doi.org/10.1101/2022.05.04.490571**
74. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, et al. 2018. Mutational processes shape the landscape of *TP53* mutations in human cancer. *Nat. Genet.* 50:1381–87

75. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, et al. 2014. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* 159:647–61
76. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, et al. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* 154:442–51
77. Glazer AM, Kroncke BM, Matreyek KA, Yang T, Wada Y, et al. 2020. Deep mutational scan of an *SCN5A* voltage sensor. *Circ. Genom. Precis. Med.* 13(1):e002786
78. Glazer AM, Wada Y, Li B, Muhammad A, Kalash OR, et al. 2020. High-throughput reclassification of *SCN5A* variants. *Am. J. Hum. Genet.* 107:111–23
79. Gordon MG, Inoue F, Martin B, Schubach M, Agarwal V, et al. 2020. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* 15:2387–412
80. Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. 2018. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* 6:116–24.e3
81. Gray VE, Sitko K, Kameni FZN, Williamson M, Stephany JJ, et al. 2019. Elucidating the molecular determinants of Aβ aggregation with deep mutational scanning. *G3* 9:3683–89
82. GTEX Consort. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369:1318–30
83. Gupta RM, Hadaya J, Trehan A, Zekavat SM, Roselli C, et al. 2017. A genetic variant associated with five vascular diseases is a distal regulator of Endothelin-1 gene expression. *Cell* 170:522–33.e15
84. Hanna RE, Hegde M, Fagre CR, DeWeirdt PC, Sangree AK, et al. 2021. Massively parallel assessment of human variants with base editor screens. *Cell* 184:1064–80.e20
85. Harrison SM, Dolinsky JS, Knight Johnson AE, Pesaran T, Azzariti DR, et al. 2017. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genet. Med.* 19:1096–104
86. Hasle N, Cooke A, Srivatsan S, Huang H, Stephany JJ, et al. 2020. High-throughput, microscope-based sorting to dissect cellular heterogeneity. *Mol. Syst. Biol.* 16:e9442
87. Hazelbaker DZ, Beccard A, Angelini G, Mazzucato P, Messana A, et al. 2020. A multiplexed gRNA *piggyBac* transposon system facilitates efficient induction of CRISPRi and CRISPRa in human pluripotent stem cells. *Sci. Rep.* 10:635
88. Heredia JD, Park J, Brubaker RJ, Szymanski SK, Gill KS, Procko E. 2018. Mapping interaction sites on human chemokine receptors by deep mutational scanning. *J. Immunol.* 200:3825–39
89. Hietpas RT, Jensen JD, Bolon DNA. 2011. Experimental illumination of a fitness landscape. *PNAS* 108:7896–901
90. Hilton SK, Huddleston J, Black A, North K, Dingens AS, et al. 2020. *dms-view*: interactive visualization tool for deep mutational scanning data. *J. Open Sour. Softw.* 5(52):2353
91. HuBMAP Consort. 2019. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* 574:187–92
92. Igartua C, Mozaffari SV, Nicolae DL, Ober C. 2017. Rare non-coding variants are associated with plasma lipid traits in a founder population. *Sci. Rep.* 7:16415
93. Inoue F, Kreimer A, Ashuach T, Ahituv N, Yosef N. 2019. Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* 25:713–27.e10
94. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, et al. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99:877–85
95. Jain PC, Varadarajan R. 2014. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal. Biochem.* 449:90–98
96. Jia X, Burugula BB, Chen V, Lemons RM, Jayakody S, et al. 2021. Massively parallel functional testing of *MSH2* missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* 108:163–75
97. Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. 2016. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* 7:11558
98. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–89
99. Kanavy DM, McNulty SM, Jairath MK, Brnich SE, Bizon C, et al. 2019. Comparative analysis of functional assay evidence use by ClinGen Variant Curation Expert Panels. *Genome Med.* 11:77

100. Kanfer G, Sarraf SA, Maman Y, Baldwin H, Dominguez-Martin E, et al. 2021. Image-based pooled whole-genome CRISPRi screening for subcellular phenotypes. *J. Cell Biol.* 220:e202006180
101. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581:434–43
102. Ke S, Anquetil V, Zamalloa JR, Maity A, Yang A, et al. 2018. Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* 28:11–24
103. Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26:990–99
104. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, et al. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* 23:800–11
105. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, et al. 2017. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* 546:370–75
106. Kim J, Koo B-K, Knoblich JA. 2020. Human organoids: model systems for human biology and medicine. *Nat. Rev. Mol. Cell Biol.* 21:571–84
107. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46:310–15
108. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, et al. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* 10:3583
109. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. 2015. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* 12:203–6
110. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, et al. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* 17:1083–91
111. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533:420–24
112. Kozek KA, Glazer AM, Ng C-A, Blackwell D, Egly CL, et al. 2020. High-throughput discovery of trafficking-deficient variants in the cardiac potassium channel K_V11.1. *Heart Rhythm* 17:2180–89
113. Kreimer A, Ashuach T, Inoue F, Khodaverdian A, Deng C, et al. 2022. Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. *Nat. Commun.* 13:1504
114. Kuang D, Truty R, Weile J, Johnson B, Nykamp K, et al. 2021. Prioritizing genes for systematic variant effect mapping. *Bioinformatics* 36:5448–55
115. Kuang D, Weile J, Li R, Ouellette TW, Barber JA, Roth FP. 2020. MaveQuest: a web resource for planning experimental tests of human variant effects. *Bioinformatics* 36:3938–40
116. Kunkel TA. 1985. Rapid and efficient site-specific mutagenesis without phenotypic selection. *PNAS* 82:488–92
117. Kweon J, Jang A-H, Shin HR, See J-E, Lee W, et al. 2020. A CRISPR-based base-editing screen for the functional assessment of BRCA1 variants. *Oncogene* 39:30–35
118. Laber S, Strobel S, Mercader J-M, Dashti H, Ainbinder A, et al. 2021. Discovering cellular programs of intrinsic and extrinsic drivers of metabolic traits using LipocyteProfiler. bioRxiv 2021.07.17.452050. **https://doi.org/10.1101/2021.07.17.452050**
119. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
120. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46:D1062–67
121. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–91
122. Liu X, Li C, Mou C, Dong Y, Tu Y. 2020. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 12:103
123. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469–76

124. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, et al. 2016. Prospective functional classification of all possible missense variants in *PPARG*. *Nat. Genet.* 48:1570–75
125. Marini NJ, Gin J, Ziegle J, Keho KH, Ginzinger D, et al. 2008. The prevalence of folate-remedial MTHFR enzyme variants in humans. *PNAS* 105:8055–60
126. Marshall JL, Doughty BR, Subramanian V, Guckelberger P, Wang Q, et al. 2020. HyPR-seq: single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes. *PNAS* 117:33404–13
127. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, et al. 2018. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50:874–82
128. Matreyek KA, Stephany JJ, Chiasson MA, Hasle N, Fowler DM. 2020. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res.* 48:e1
129. Matreyek KA, Stephany JJ, Fowler DM. 2017. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* 45:e102
130. Mattiazzi Usaj M, Sahin N, Friesen H, Pons C, Usaj M, et al. 2020. Systematic genetics and single-cell imaging reveal widespread morphological pleiotropy and cell-to-cell variability. *Mol. Syst. Biol.* 16:e9243
131. Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, et al. 2016. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife* 5:e15802
132. McDonald D, Wu Y, Dailamy A, Tat J, Parekh U, et al. 2020. Defining the teratoma as a model for multi-lineage human development. *Cell* 183:1402–19.e18
133. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *PNAS* 107:6544–49
134. Meitlis I, Allenspach EJ, Bauman BM, Phan IQ, Dabbah G, et al. 2020. Multiplexed functional assessment of genetic variants in *CARD11*. *Am. J. Hum. Genet.* 107:1029–43
135. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30:271–77
136. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. 2014. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* 42:e112
137. Mighell TL, Evans-Dutson S, O'Roak BJ. 2018. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* 102:943–55
138. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, et al. 2019. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* 16:409–12
139. Mitchell JM, Nemesh J, Ghosh S, Handsaker RE, Mello CJ, et al. 2020. Mapping genetic effects on cellular phenotypes with "cell villages." bioRxiv 2020.06.29.174383. **https://doi.org/10.1101/2020.06.29.174383**
140. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. 2010. From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* 466:714–19
141. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, et al. 2017. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* 27:S2–8
142. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, et al. 2021. Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593:238–43
143. Newberry RW, Leong JT, Chow ED, Kampmann M, DeGrado WF. 2020. Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nat. Chem. Biol.* 16:653–59
144. Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res.* 11:863–74
145. Nyegaard M, Overgaard MT, Søndergaard MT, Vranas M, Behr ER, et al. 2012. Mutations in calmodulin cause ventricular tachycardia and sudden cardiac death. *Am. J. Hum. Genet.* 91:703–12
146. Ota S, Horisaki R, Kawamura Y, Ugawa M, Sato I, et al. 2018. Ghost cytometry. *Science* 360:1246–51
147. Parsi KM, Hennessy E, Kearns N, Maehr R. 2017. Using an inducible CRISPR-dCas9-KRAB effector system to dissect transcriptional regulation in human embryonic stem cells. *Methods Mol. Biol.* 1507:221–33
148. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30:265–70
149. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* 27:1173–75

150. Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, et al. 2013. RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nat. Methods* 10:973–76
151. Ramilowski JA, Yip CW, Agrawal S, Chang J-C, Ciani Y, et al. 2020. Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* 30:1060–72
152. Reilly SK, Gosai SJ, Gutierrez A, Mackay-Smith A, Ulirsch JC, et al. 2021. Direct characterization of *cis*-regulatory elements and functional dissection of complex genetic associations using HCR-FlowFISH. *Nat. Genet.* 53:1166–76
153. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47:D886–94
154. Richards S, Aziz N, Bale S, Bick D, Das S, et al. 2015. 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.* 17:405–24
155. Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* 15:816–22
156. Rivera-Muñoz EA, Milko LV, Harrison SM, Azzariti DR, Kurtz CL, et al. 2018. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum. Mutat.* 39:1614–22
157. Roadmap Epigenomics Consort., Kundaje A, Meuleman W, Ernst J, Bilenky M, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–30
158. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, et al. 2019. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* 51:1170–76
159. Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, et al. 2017. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* 18(1):150
160. Rubin AF, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM. 2016. Enrich2: a statistical framework for analyzing deep mutational scanning data. bioRxiv 075150. **https://doi.org/10.1101/075150**
161. Rubin AF, Min JK, Rollins NJ, Da EY, Esposito D, et al. 2021. MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. bioRxiv 2021.11.29.470445. **https://doi.org/10.1101/2021.11.29.470445**
162. Rusu V, Hoch E, Mercader JM, Tenen DE, Gymrek M, et al. 2017. Type 2 diabetes variants disrupt function of SLC16A11 through two distinct mechanisms. *Cell* 170:199–212.e20
163. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, et al. 2019. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* 37:803–9
164. Schmiedel JM, Lehner B. 2019. Determining protein structures using deep mutagenesis. *Nat. Genet.* 51:1177–86
165. Schraivogel D, Gschwind AR, Milbank JH, Leonce DR, Jakob P, et al. 2020. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* 17:629–35
166. Schraivogel D, Kuhn TM, Rauscher B, Rodríguez-Martínez M, Paulsen M, et al. 2022. High-speed fluorescence image-enabled cell sorting. *Science* 375:315–20
167. Shan X, Wang L, Hoffmaster R, Kruger WD. 1999. Functional characterization of human methylenetetrahydrofolate reductase in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 274:32613–18
168. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, et al. 2016. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.* 99:595–606
169. Smemo S, Tena JJ, Kim K-H, Gamazon ER, Sakabe NJ, et al. 2014. Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* 507:371–75
170. Sobreira DR, Joslin AC, Zhang Q, Williamson I, Hansen GT, et al. 2021. Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of *IRX3* and *IRX5*. *Science* 372:1085–91
171. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, et al. 2017. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* 101:315–25
172. Starita LM, Islam MM, Banerjee T, Adamovich AI, Gullingsrud J, et al. 2018. A multiplex homology-directed DNA repair assay reveals the impact of more than 1,000 BRCA1 missense substitution variants on protein function. *Am. J. Hum. Genet.* 103:498–508

173. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, et al. 2015. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 200:413–22
174. Starling AL, Ortega JM, Gollob KJ, Vicente EJ, Andrade-Nóbrega GM, Rodriguez MB. 2003. Evaluation of alternative reporter genes for the yeast two-hybrid system. *Genet. Mol. Res.* 2:124–35
175. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, et al. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182:1295–310.e20
176. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* 21(6):577–81
177. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. 2007. Population genomics of human gene expression. *Nat. Genet.* 39:1217–24
178. Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, et al. 2020. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *PNAS* 117:5394–401
179. Sun S, Weile J, Verby M, Wu Y, Wang Y, et al. 2020. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Med.* 12:13
180. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, et al. 2018. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50:1161–70
181. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, et al. 2022. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.* **https://doi.org/10.1038/s41591-022-01891-3**
182. Thakore PI, D'Ippolito AM, Song L, Safi A, Shivakumar NK, et al. 2015. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* 12:1143–49
183. Thompson MC, Yeates TO, Rodriguez JA. 2020. Advances in methods for atomic resolution macromolecular structure determination. *F1000Research* 9(Faculty Rev):667
184. Thomsen SK, Ceroni A, van de Bunt M, Burrows C, Barrett A, et al. 2016. Systematic functional characterization of candidate causal genes for type 2 diabetes risk variants. *Diabetes* 65:3805–11
185. Tomaselli PJ, Rossor AM, Horga A, Jaunmuktane Z, Carr A, et al. 2017. Mutations in noncoding regions of *GJB1* are a major cause of X-linked CMT. *Neurology* 88:1445–53
186. Ursu O, Neal JT, Shea E, Thakore PI, Jerby-Arnon L, et al. 2022. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat. Biotechnol.* 40(6):896–905
187. van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, et al. 2019. High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet* 51:1160–69
188. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, et al. 2018. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50:493–97
189. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LTM, et al. 2015. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* 6:6905
190. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50:D439–44
191. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
192. Vockley CM, D'Ippolito AM, McDowell IC, Majoros WH, Safi A, et al. 2016. Direct GR binding sites potentiate clusters of TF binding across the human genome. *Cell* 166:1269–81.e19
193. Wang X, He L, Goggin SM, Saadat A, Wang L, et al. 2018. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* 9:5380
194. Weile J, Kishore N, Sun S, Maaieh R, Verby M, et al. 2021. Shifting landscapes of human MTHFR missense-variant effects. *Am. J. Hum. Genet.* 108:1283–300
195. Weile J, Roth FP. 2018. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Hum. Genet.* 137:665–78
196. Weile J, Sun S, Cote AG, Knapp J, Verby M, et al. 2017. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* 13:957

197. Wong MS, Kinney JB, Krainer AR. 2018. Quantitative activity profile and context dependence of all human 5′ splice sites. *Mol. Cell.* 71:1012–26.e3
198. Wright CF, Quaife NM, Ramos-Hernández L, Danecek P, Ferla MP, et al. 2021. Non-coding region variants upstream of *MEF2C* cause severe developmental disorder through three distinct loss-of-function mechanisms. *Am. J. Hum. Genet.* 108:1083–94
199. Wu Y, Liu H, Li R, Sun S, Weile J, Roth FP. 2021. Improved pathogenicity prediction for rare human missense variants. *Am. J. Hum. Genet.* 108:1891–906
200. Wu Y, Weile J, Cote AG, Sun S, Knapp J, et al. 2019. A web application and service for imputing and visualizing missense variant effect maps. *Bioinformatics* 35:3191–93
201. wwPDB Consort. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 47:D520–28
202. Yeo NC, Chavez A, Lance-Byrne A, Chan Y, Menn D, et al. 2018. An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat. Methods* 15:611–16
203. Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ. 2014. Massively parallel functional annotation of 3′ untranslated regions. *Nat. Biotechnol.* 32:387–91
204. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12:931–34
205. Zinkus-Boltz J, DeValk C, Dickinson BC. 2019. A phage-assisted continuous selection approach for deep mutational scanning of protein–protein interactions. *ACS Chem. Biol.* 14:2757–67