Published in final edited form as:
Nature. 2024 September ; 633(8028): 47–57. doi:10.1038/s41586-024-07510-0.

Deciphering the impact of genomic variation on function

IGVF Consortium

Abstract

Our genomes influence nearly every aspect of human biology—from molecular and cellular functions to phenotypes in health and disease. Studying the differences in DNA sequence between individuals ("genomic variation") could reveal novel mechanisms of human biology, uncover the basis of genetic predispositions to diseases, and guide the development of new diagnostics and therapeutics[1–3]. Yet, understanding how genomic variation alters genome function to influence phenotype has proven challenging. To unlock these insights, we need a systematic and comprehensive catalog of genome function and the molecular and cellular effects of genomic variants. Toward this goal, the Impact of Genomic Variation on Function (IGVF) Consortium will combine approaches in single-cell mapping, genomic perturbations, and predictive modeling to investigate the relationships among genomic variation, genome function, and phenotypes. IGVF will create maps across hundreds of cell types and states describing how coding variants alter protein activity, how noncoding variants change the regulation of gene expression, and how such effects connect through gene regulatory and protein interaction networks. These experimental data, computational predictions, and accompanying standards and pipelines will be integrated into an open resource that will catalyze community efforts to explore how our genomes influence biology and disease across populations.

Introduction

Since the initial sequencing of the human genome, genetic studies have been immensely productive[1–3]. Exome and genome sequencing studies have identified hundreds of millions of genomic variants, including single-nucleotide variants (SNVs), small insertions and deletions (indels), and larger structural variants (Fig. 1)[4,5]. Comparisons within families, case-control cohorts, and population-scale biobanks have discovered hundreds of thousands of associations between such variants and phenotypes in both health and disease[6–12].

Correspondence to: markerpaper@igvf.org.

Author Contributions
Jesse M. Engreitz, Heather A. Lawson, and Harinder Singh co-led the Writing Group. Jesse M. Engreitz, Heather A. Lawson, Harinder Singh, Lea Starita, Gary C. Hon, Hannah Carter, Nidhi Sahni, Timothy E. Reddy, Xihong Lin, Yun Li, Nikhil Munshi, Maria Chahrour, Benjamin Hitz, and Ali Mortazavi wrote initial text based on input from PIs, the Writing Group, and Working Group and Focus Group Co-Chairs. Jesse M. Engreitz, Alan Boyle, and Jayoung Ryu developed figures. All authors contributed to developing the vision and goals of the IGVF Consortium, outlining the project, and editing the manuscript. The role of the NHGRI Program Management in the preparation of this paper was limited to coordination and scientific management of the IGVF Consortium.

Competing Interests
R. D. Steiner has been a consultant for Leadiant Biosciences, Mirum Pharmaceuticals, PTC Therapeutics, and Travere. He has received honoraria from Medscape and is an employee and shareholder of PreventionGenetics, part of Exact Sciences. B. P. Kleinsteiver is a co-inventor on patents and patent applications that describe genome engineering technologies, and is on the scientific advisory board of Acrigen Biosciences, Life Edit Therapeutics, and Prime Medicine.
The other authors declare no competing interests.

The next challenge is to understand how genomic variation affects molecular and cellular processes to influence organismal phenotype (Fig. 1). At a molecular level, genomic variation can impact the temporal-spatial and quantitative expression of genes or the activity and localization of proteins. Altered gene expression or protein activity can, in turn, impact other genes and proteins, for example via gene regulatory and protein-protein interaction networks. Changes in such molecular networks impact the properties of cells and tissues, and in doing so can influence organismal phenotypes. Here we use the term "genome function" to refer to these processes encoded by the genome, and note that this does not necessarily imply "function" in terms of evolutionary selection.[13,14]

Previous and ongoing efforts have produced breakthroughs in mapping various aspects of genome function, including locating and annotating millions of noncoding regulatory elements in the human genome[15,16]; mapping associations between genomic variants and effects on gene or protein expression across dozens of human tissues[17–19]; profiling hundreds of cell types and states through single-cell measurements of gene expression[20,21]; applying saturation mutagenesis to analyze coding variants in selected disease genes[22–24]; and characterizing how genes and proteins interact genetically or physically in molecular networks[25–27]. These efforts, as well as disease-specific consortia and other studies, have also demonstrated how mapping the impact of genomic variation on genome function can reveal molecular mechanisms in human biology and disease, guide genetic diagnosis and clinical management, and facilitate the development of novel therapeutics (Fig. 1, Box 1, reviewed in[1,28,29]).

Yet, connecting genomic variants to functions and phenotypes continues to prove challenging and slow. The molecular mechanisms underlying most genetic associations for common diseases remain to be established[2,29], genetic diagnosis for rare diseases continues to be hindered by the preponderance of variants of uncertain significance (VUS)[7,30], and the impact of genomic variation across diverse groups and populations remains poorly studied[31,32]. New approaches are needed to accelerate research throughout the community and thereby unlock the vast unrealized potential for understanding human biology and improving human health[33,34].

Advances in experimental and computational genomics now promise to overcome some of the key challenges:

i. Regulatory elements and genes can have cell-type or context-dependent activities, which have been challenging to analyze comprehensively. Emerging single-cell technologies now enable the generation of comprehensive maps of chromatin state and gene expression in nearly any cell type in the body[20,21], and computational analysis of these datasets can help to locate candidate regulatory elements, identify transcription factor (TF) binding regions and footprints, and delineate gene regulatory networks[35–38].

ii. Previously it has been difficult to uncover the causal relationships between genomic variation and genome function, including due to challenges of linkage disequilibrium between common variants. New approaches in statistical fine-mapping now enable improved interpretation of genome-wide association studies

(GWAS) and quantitative trait loci (QTL) studies[39–42], and high throughput technologies for designed genomic perturbations, such as with CRISPR screens[43–50] and massively parallel reporter assays (MPRAs)[51–57], provide a powerful means to systematically characterize the effects of variants, elements, and genes.

iii. The scale of the problem is immense. With billions of possible single-nucleotide genomic variants, 20,000 genes, and thousands of cell types, we cannot expect to experimentally map the effects of all possible variants on all aspects of genome function in all possible contexts. To address this, recent studies have highlighted the possibility of training computational models that can generalize to make predictions about genome function for untested variants, cell types, and/or contexts[58–64].

iv. While previous efforts have largely focused on particular types of genome variation or individual diseases, integrative analysis of coding and noncoding variation into molecular networks, and comparisons across diverse cellular contexts and diseases, could greatly accelerate progress[28,65–67].

v. Finally, recent successes by CASP[68], ENCODE[15], and others[17,21,69] have highlighted how uniting a diverse community of investigators under a common framework can catalyze advances throughout the global scientific community by developing uniform standards and analysis pipelines, creating uniformly processed, AI-readable datasets amenable to predictive modeling, and enabling the comparison and synthesis of alternative strategies.

With these challenges and opportunities in mind, the National Human Genome Research Institute (NHGRI) launched the IGVF Consortium in 2021, with the goal of developing a systematic understanding of the effects of genomic variation on genome function and how these effects shape phenotypes. The Consortium consists of >120 laboratories collaborating on five key activities to address the above challenges: (i) Mapping Centers, to analyze regulatory element and gene activity at single-cell resolution across hundreds of cell types; (ii) Functional Characterization Centers, to systematically characterize the molecular and cellular effects of introducing variants or perturbing elements and genes; (iii) Predictive Modeling Projects, to develop and apply computational approaches to comprehensively model the impact of genomic variation on genome function and guide experimental design; (iv) Regulatory Network Projects, to advance network-level understanding of the influence of genetic variation and genome function on cellular and organismal phenotypes; and (v) a Data and Administrative Coordinating Center, to lead development of resources and infrastructure to share IGVF data, standards, and pipelines with the scientific community. IGVF membership and activities are expanding further via Affiliate Membership, a process by which any researcher or research project can apply to join IGVF to drive its vision and execution. Through these activities, the IGVF Consortium aims to generate a catalog that can be broadly deployed for exploring genome function and the impact of genetic variation on human biology and diseases in diverse populations. Below we describe the goals, strategies, and anticipated deliverables of IGVF (Box 2).

Map-perturb-predict framework

To create a comprehensive catalog of the effects of genomic variation, IGVF has developed a strategy that integrates three complementary components (Fig. 2). One component will be to quantify the activity of regulatory elements and the expression of genes via single-cell mapping. Another will conduct systematic perturbations of variants, regulatory elements, and genes. A third will seek to generalize results to new, unstudied genomic variants and cellular contexts via predictive modeling. Integration of these three components in a map-perturb-predict framework will create substantial synergy across the consortium.

Single-cell mapping

Identifying noncoding regulatory elements and genes and mapping their activities across cell types and states is foundational for understanding where, when and how genomic variation might impact genome function. Yet, many previous efforts have lacked this level of resolution. We will collect single-cell data across hundreds of cell types and states (see below for biological systems and contexts). We will apply primarily single-nucleus (sn)ATAC-seq and snRNA-seq, including in multiomic formats, to enable integration of IGVF data with other emerging datasets (Fig. 2). Individual projects will explore additional single-cell approaches including TF binding, histone modifications, chromatin interactions, protein levels and activity, and clonal tracing. Key assays (including 10x Multiome, SHARE-seq[70], and Parse Evercode[71]) will be directly compared and calibrated on the same samples, and the performance of computational analyses and predictive models will be assessed as a function of sequencing depth. These data will provide a foundation for interpreting the effects of functional characterization experiments and building cell-type-specific maps of variant effects.

Genomic perturbations

Perturbation experiments will be crucial for understanding the causal relationships among variants, regulatory elements, genes, and phenotypes, but until recently have been challenging to apply at sufficient scale. New enabling technologies include high-throughput screens using CRISPR genetic or epigenetic perturbations or over-expression strategies[22,23,43–50]; reporter assays for enhancer or promoter activities[51–57]; and fine-mapping of different types of quantitative trait loci (QTLs)[17,42,72] including single-cell eQTLs[73–75]. IGVF plans to conduct >2 million experimental perturbations, including to directly study the effects of naturally occurring or designed DNA variants, and to perturb regulatory elements and genes to build maps of genome function (Fig. 2). We will characterize the effects of these perturbations using diverse assays including measurements of chromatin accessibility[76], gene expression[77–79], protein expression and activity[25,80–83], and molecular and cellular phenotypes[84]. These data will enable directly characterizing variants of interest, such as those associated with disease, and provide data to train or evaluate predictive models of variant effects.

Predictive modeling

Genome function is complex, and we cannot expect to experimentally map the effects of all possible variants on all possible activities in all possible cellular contexts. Predictive models will be needed to make predictions that generalize across contexts — for example, to link genetic variants to effects on TF binding and chromatin accessibility[57,61–64]; connect regulatory elements to their target genes[64,85,86]; or identify causal genes and cell types enriched for heritability for complex diseases or traits[87–92]. We will leverage advances in machine learning and large-scale perturbation datasets to tackle key prediction problems — including mapping aspects of genome function, interpreting the impact of genomic variation, and guiding the design of future experimental assays such that the data produced will be maximally informative for subsequent predictive modeling. To systematically evaluate and calibrate such models, we will build benchmarking pipelines that compare predictions to perturbation data, including both from IGVF functional characterization experiments and external sources such as QTL, GWAS, and genome sequencing studies[87,88,93,94]. In areas where data collection is already advanced, we will engage the external community by designing prediction challenges with held-out assessment datasets produced by IGVF.

Application areas

Together, these three activities will form an iterative map-perturb-predict framework that IGVF will apply to explore a wide array of cell types, cellular phenotypes, and diseases (Fig. 2). Projects will apply distinct but overlapping sets of experimental assays and computational models, enabling a broad exploration of possible strategies and integration of insights across biological systems.

IGVF projects have flexibility to study diverse biological models, prioritized based on relevance to human disease, expertise of consortium members, tractability, and other considerations. Current models include human embryonic and induced pluripotent stem cells (iPSC) differentiated into lineages spanning all germ layers in 2D and 3D (*e.g.*, gastruloids, cardiomyocytes, and neurons); primary cell types relevant to disease areas of interest (*e.g.*, smooth muscle cells for coronary artery disease); and human and mouse tissues *in vivo* to inform how cell-cell interactions and environment alter genome function (*e.g.*, liver and lung in the presence of bacterial lipopolysaccharide) (Fig. 2). Selected models include dynamic biological processes that will provide insights into how regulatory networks change over time, such as B cell activation and differentiation or fibroblast-to-iPSC reprogramming.

While the primary objective of IGVF is to characterize variation and function of the human genome, IGVF studies will also study and create resources for mouse models, such as for comparing the effects of variants, elements, and genes across individuals with different genetic backgrounds, and for *in vivo* genomic perturbation experiments to understand how variants or genes affect cellular phenotypes in a tissue environment. IGVF will leverage the genetic diversity found in the collaborative cross (CC)[95], which includes more than 15 million SNVs between the 8 founder strains. These strains include the reference C57BL/6J strain, mouse disease models such as NOD, and recombinant inbred CC strains. Current efforts include collecting single-cell mapping data across 8 tissues in male and female adults to identify cell-type specific cellular programs and QTLs and compare to matching human samples.

The map-perturb-predict framework will enable integration across biological systems and models. For example, to enable integrative analysis across all projects studying gene regulation, we will generate and harmonize multiomic snRNA+ATAC data as a reference map in each cellular model. To compare genomic perturbation datasets across projects, we will deploy consistent data processing pipelines, quantify reproducibility, and assess power. To integrate information across experimental assays and cellular models, we will train predictive models that learn from diverse data types and can generalize to new, unstudied cell types.

Throughout, a unifying analysis framework will be to consider and evaluate which cellular models and assays provide the best ability to distinguish or enrich for genomic variants associated with disease. For example, studies of coding variation in known Mendelian disease genes will validate the relevance of their cellular assays by comparison to known pathogenic and benign variants. Studies of noncoding variants associated with a common, complex disease might select a cellular model whose regulatory elements are globally enriched for containing risk variants. Such comparisons to human genomic variation will provide an external benchmark applicable to evaluating many methods and design decisions throughout IGVF (see below).

A map of genome function and variant effects

IGVF will deliver a preliminary variant effect map that integrates three key aspects of genome function: gene expression, protein function, and molecular networks (Fig. 3). This draft map would allow querying, for any possible SNV in the genome: Is this variant measured or predicted to (i) impact transcription factor binding, regulatory element activity, and target gene expression in particular cell contexts, for noncoding variants; (ii) impact protein function, for coding variants; and (iii) connect to other genes/proteins via gene regulatory networks and/or protein-interaction networks, for both coding and noncoding variants?

For each of these aspects of genome function, computational models have shown promise but much work is needed to improve their accuracy. Toward this goal, this preliminary map will integrate annotations of the different aspects of genome function for the first time, and to establish benchmarking pipelines to quantify the accuracy of all predictions against perturbation data and external human genetics datasets. We will encode this map of genome function, along with benchmarks and external data, in a multi-relational knowledge graph[96–99] as part of the IGVF Catalog (see below). This Catalog will provide a foundation for an iterative and ongoing effort extending beyond IGVF to improve the accuracy of this map over time (Fig. 3).

Effects on gene regulation

In the 99% of our genome that does not encode for proteins, noncoding variants can impact genome function by altering gene expression, splicing, chromatin state, or other aspects of gene regulation. Despite advances by ENCODE, GTEx, and other projects, we still lack models that can make accurate causal inferences about how genomic variation affects gene regulation[94,100,101]. We will seek to build genome-wide annotations of key components of this *cis*-regulatory code: Which SNVs affect transcription factor binding sites, regulatory element activity, and gene expression in *cis*, in which cell types or states, with what magnitude and direction of effect?

To do so, IGVF plans to (i) generate multiomic snRNA+ATAC-seq data at a depth needed to identify candidate cis-regulatory elements, detect likely transcription factor binding sites[35,102], and predict enhancer-gene relationships[36,37,64,86,88,93]; (ii) test >1 million noncoding variants in enhancer activity reporter assays[51,52,56,57,103]; (iii) test thousands of noncoding variants for effects on expression through fine-mapping of eQTLs or direct CRISPR-based genome editing[17,42–45,47]; (iv) measure >100,000 putative regulatory interactions between candidate regulatory elements and nearby genes, for example using dCas9-based epigenome editing[85,104–107]; (v) and perturb transcription factors to read out effects on gene expression using Perturb-seq[77–79]. These experiments will be conducted in multiple cellular models, so that the data can be used to develop predictive models that generalize across many cell types. These cellular models will include several previously studied in ENCODE[15] and GTEx[17], to enrich and benefit from rich existing datasets.

Effects on protein function

For protein-coding sequences, our ability to interpret the functions of genomic variation is based on our knowledge of the genetic code for protein synthesis — which has enabled identifying open reading frames encoding novel proteins and understanding nonsense or frameshift variants. However, most coding variants, including missense variants and inframe indels, remain difficult to interpret, and we still lack a comprehensive understanding of how changes in protein sequence might affect different aspects of protein structure, expression, dynamics, and activity.

We will improve the annotation of protein-coding missense variants by applying high-throughput technologies[25,80–83] to experimentally characterize the impacts of >200,000 missense variants on protein and cellular properties, including protein stability, subcellular localization, cell viability, cell morphology, and protein-protein interactions. These experiments will directly characterize thousands of variants in clinically relevant genes, such as those associated with Mendelian diseases, and provide data to refine or develop new models to predict the likely impact of coding variants in other genes across the genome.

Molecular networks and cellular phenotypes

Upon linking a variant to effects on gene expression or protein activity in *cis*, we will seek to annotate the sets of other genes and proteins linked to the variant in *trans* through molecular networks in a given cell type or state. Specifically, we will focus on defining three types of molecular networks: (i) gene expression programs, described by sets of genes whose expression levels are correlated across single cells; (ii) gene regulatory networks that describe which transcription factors regulate which target genes via specific noncoding regulatory sequences; and (iii) sets of interacting proteins or protein complexes. We will also examine dynamic changes to these molecular networks across cell fate or state transitions, and, to a more limited extent, explore links to downstream cellular phenotypes.

To build these maps, we will collect longitudinal multiomic data across dynamic cellular processes including differentiation and reprogramming[70,108–110]; study how genes and proteins interact in molecular networks, including by mapping protein-protein interactions[25] and conducting large-scale Perturb-seq[77–79]; and assess how CRISPR-based perturbations or natural genetic variation across individuals affects cellular phenotypes including differentiation, gene expression programs, and cellular states. Such time-resolved datasets will be used to build dynamical regulatory models that incorporate feedback and feed forward loops and account for cell fate or state transitions.

We anticipate that many aspects of this map will be cell-type specific, with annotations for each of the hundreds of cell types, states, and contexts studied by IGVF. For example, predictive models that use snRNA-seq and ATAC-seq as inputs could be developed using data from cellular models, in which predictions can be directly evaluated with matching perturbation data, and applied to make cell-type specific predictions in cell types from primary tissues[36,37,86,111].

Exploring the impact of variation on disease

The map-perturb-predict framework and IGVF variant effect map will provide new resources for the community to study the impact of genomic variation on human diseases and phenotypes, but this goal presents additional challenges. For many diseases, an individual's risk is likely to be determined by a combination of thousands of independently acting variants[112,113] — including for diseases presumed to follow Mendelian inheritance patterns, where penetrance and expressivity may include a polygenic component[114]. A single variant may have pleiotropic effects on multiple genes and pathways, only one or several of which may be important for disease[1,17,88,89]. Disease susceptibility can involve many different cell types, possibly at specific timepoints, with effects accumulating over decades or in specific environmental contexts[115]. The impact of genomic variation on genome function and phenotype can also differ across age, sex, populations, and ancestry, for example due to differences in allele frequencies[116] or possible genetic or environment interactions[117–121].

Toward addressing some of these challenges, we will focus on assessing how IGVF maps and methods can be best applied to: (i) inform clinical variant interpretation, particularly for rare diseases; (ii) learn about molecular and cellular mechanisms underlying risk for common and rare diseases; and (iii) ensure that lessons about the impact of genomic variation on genome function are applicable across diverse populations. Notably, each of these questions represents a major research area involving many strategies beyond those pursued in IGVF[7–9,28,122–124], and these exploratory efforts will seek to integrate with other efforts in the field.

Informing genetic diagnosis

IGVF will apply variant effect maps of coding variation to inform the clinical interpretation of VUS in genes with known and suspected links to Mendelian genetic diseases. Data from multiplexed assays of variant effect can be translated into powerful evidence for clinical variant interpretation, for example moving 50% of VUS in *BRCA1*, 70% in *TP53*, 74% in *MSH2*, and 90% in *DDX3X* into more definitive pathogenic or benign classifications[81,125,126]. These studies have improved genetic test results for cancer risk and ended diagnostic odysseys for families with neurodevelopmental disease.

To expand this approach, IGVF labs will experimentally measure the effects of hundreds of thousands of variants in known disease genes, with a particular focus on those where identification of loss-of-function variants is clinically actionable[127,128]. We will assess the extent to which experimental data or computational predictions correctly identify variants previously classified as either pathogenic or benign, and calibrate these analyses for clinical applications[129,130]. Clinicians routinely use experimental and predictive data to interpret the effects of coding variants, but do not yet do so for noncoding variants. Thus we will explore whether IGVF data and predictions could also improve the clinical interpretation of noncoding variants. IGVF will deliver variant effect maps and calibrated predictions that will ultimately substantially reduce the VUS burden in etiological diagnosis of rare disease[124]. Integration of maps for both coding and noncoding variants could also aid in the development of the next-generation polygenic risk score methodologies for better risk characterization in complex phenotypes[117].

Molecular mechanisms of disease risk

Improved variant effect maps could be transformative for identifying new biological mechanisms that influence genetic risk for disease. In particular, we will seek to understand how best to combine the map-perturb-predict framework and variant effect maps with human genetic data to nominate variants, genes, cell types, and cellular programs that influence disease risk.

We will study a variety of diseases and traits guided by the expertise of consortium members, including highly powered quantitative traits with simpler biological architectures, such as lipid and hematological traits, as well as complex diseases involving many cell types such as systemic lupus erythematosus, coronary artery disease, and Alzheimer’s disease. Comparison of strategies between these systems will be informative. As one example, IGVF investigators are studying variants associated with lipid traits, where GWAS and whole-exome sequencing studies have already identified hundreds of associated noncoding and coding variants, and where certain key genetic pathways involved in lipid handling are already known[11,131–133]. By conducting CRISPR screens to identify variants and regulatory elements that affect lipid uptake in cellular models enriched for trait heritability, testing variant effects on enhancer activity in massively parallel reporter assays, and applying state-of-the-art predictive models, we will evaluate which combinations of experiments and/or predictive models provide the best ability to predict disease-associated variation and known causal genes. To complement these high-throughput maps, certain projects will conduct detailed studies of mechanisms of particular GWAS loci or known disease genes, including in animal models. These combined efforts will reveal mechanisms of genetic risk for selected diseases, inform the molecular genetic architecture of complex traits, and help to develop strategies to identify causal variants, genes, and pathways for any complex disease.

Impact of variation across populations

IGVF aims to ensure that insights about the impact of genomic variation are applicable to and inclusive of people of diverse groups. To do so, we will promote diversity in functional genomics studies, experimentally study and computationally annotate variants observed in diverse populations, study diseases disproportionately affecting disadvantaged or under-represented populations, and explore the extent to which particular variants might exert the same or different effects due to interactions with genetic background or environment[134–136].

We will employ both experimental and computational strategies. In the current design phase, we have incorporated variants from diverse populations, including from the 1000 Genomes Project[137], Millions Veterans Program[138], and cross-ancestry GWAS meta-analyses[131,139–142]. Biological models include iPSCs derived from individuals from different populations (including European, East Asian, and African), and genetically diverse mouse lines from the Collaborative Cross[143]. Saturation mutagenesis will be employed to measure variant effects in clinically relevant protein-coding sequences to enable interpretation of variants observed in any individual[144]. We will deploy computational models to make context-specific predictions for SNVs across the genome, including methods to predict individual-specific effects of noncoding variants on chromatin state and gene expression[61,63,64]. These data and analyses will provide insights into variant effects across groups and provide a valuable resource for investigating the effects of variants discovered in diverse populations.

Data release and resources

A major goal of IGVF is to catalyze future research to understand the relationships between genome function, genomic variation, and phenotype. To do so, we will build the IGVF Data Resource to enable biomedical researchers across diverse disciplines to access and apply IGVF datasets, predictions, and methods (https://igvf.org).

For researchers who want to explore data and predictions, we will create the IGVF Catalog. The IGVF Catalog will consist of one or more web portals that enable searching for information about specific variants, genomic loci, or genes, and will draw from processed data, analysis products, and computational predictions generated by IGVF as well as external data sources (Fig. 3). To support users who want programmatic access to perform integrative analyses or to develop web applications, we will also provide an application programming interface (API) to the underlying knowledge graph.

For researchers who want to access raw or processed data, we will develop the IGVF Data Portal. The Data Portal will provide web-browser and programmatic access to uniformly processed datasets, analysis products, and rich metadata, enabling users to develop new computational methods, analyze IGVF data in new ways, or compare their data to IGVF standards. The IGVF Data Portal will follow principles of making data Findable, Accessible, Interoperable, and Reusable (FAIR)[145]. Data will be stored in cloud file buckets to facilitate computing on the data in place. Some IGVF data may not have consent for public sharing; such data will be deposited in NHGRI's Analysis, Visualization and Informatics Lab (AnVIL) platform to provide access control in adherence to NIH Policy[146].

For researchers who want to apply IGVF methods and strategies to additional systems, the Data Portal will also share documentation on IGVF standards, protocols, and best practices for experimental design, data analysis, and predictive modeling. These resources will include computational methods, data formats, and consensus data processing pipelines for key assays and analysis products, such as for single-nucleus RNA-seq and ATAC-seq, CRISPR experiments, MPRAs, eQTL studies, and others. Data analysis tools will include approaches to assess replicates, quantify experimental noise, and assess power. All data processing code will be released with open-source licenses to enable others to analyze similar data in an identical fashion, and we will strive to make sure that it can be run on compute resources accessible to researchers throughout the global research community.

For all researchers, we will provide training and support on how to access these IGVF resources. For up-to-date information on where to find instructional streaming videos, online notebooks and tutorials, and schedules for seminars and webinars, visit www.igvf.org. Altogether, we expect that these resources will enable a wide range of scientific activities, expanding far beyond the specific studies undertaken by the IGVF Consortium.

Finally, IGVF is committed to rapid release of data and results. Data and predictions will be released upon quality control and no later than manuscript submission, and manuscripts will be posted on preprint servers prior to manuscript submission.

Collaborations and community

Understanding genomic variation and genome function is a grand challenge that demands global and interdisciplinary collaboration. IGVF welcomes collaboration with, and input from, the broader scientific community. Researchers interested in joining IGVF can apply for Affiliate Membership. Affiliate members can participate fully in working groups and other IGVF collaborations, and thereby drive the vision, goals, and execution of consortium activities. For more information, visit https://igvf.org/affiliate-membership/.

IGVF is actively coordinating with other consortia, including ClinGen[8], the Genomics Research to Elucidate the Genetics of Rare diseases (GREGoR) consortium, and the Atlas of Variant Effects (AVE) Alliance[144]. These collaborations will facilitate the open exchange and interoperability of genomic data and resources, for example to use common variant naming schema, genome and transcriptome builds, and analysis pipelines.

Similarly, IGVF activities will benefit from close interactions with efforts to characterize human genomic variation and assemblies, such as the Human Pangenome Reference Consortium (HPRC)[147]; with efforts to catalog disease-associated variation across ancestries, including All of Us[148] and TOPMed[10]; with efforts to build atlases using single-cell tools, such as the Human Cell Atlas[21] and HuBMAP[20]; and with efforts to compare and evaluate strategies for interpreting genetic variation associated with disease, such as the International Common Disease Alliance[28].

Outlook

With the rapid expansion of human genetics studies linking variation to disease, the interpretation of the impact of genomic variation on function is currently a rate-limiting step for delivering on the promise of precision medicine. The IGVF Consortium will pursue a unique, coordinated strategy for accelerating progress (Box 2).

Success for IGVF will involve creating resources and generating scientific advances not possible through individual efforts. Key outcomes include (i) insights into genome biology and advances in genetic diagnosis enabled by the map-perturb-predict framework and variant effect maps; (ii) an interoperable ecosystem of data, predictions, and models that will be used by IGVF and the broader scientific community to derive insights into genome function, genomic variation, and phenotype; (iii) massive, uniformly processed datasets spanning single-cell and functional characterization assays that directly assay large swaths of the genome and serve as an enduring, foundational resource for developing predictive models; (iv) a Catalog that provides web and programmatic access to look up integrative predictions and experimental data regarding variants, genomic elements, and genes across many cell types and contexts; and (v) new methods and strategies for studying genome variation and function, derived through systematic comparisons of methods. Altogether, these activities will set in motion community efforts to expand on this framework by collecting additional datasets, training improved models, generating more accurate maps, and expanding the approach to additional cell types and aspects of genome function.

While ambitious, IGVF activities do have limitations in scope. IGVF aims for systematic analysis of certain aspects of genome function, but others—including effects on nuclear organization; RNA splicing, localization, and translation; protein signaling and metabolism; and cellular phenotypes, cell-cell interactions, and tissue organization—are of great interest but will require efforts beyond the current membership of the Consortium. IGVF projects span a great breadth of cellular models and disease areas, but are not necessarily designed for comprehensive analysis of any single disease. IGVF will use cellular models to develop predictive models that are applicable to understanding variants in many systems, but systematic analysis to map epistatic interactions among variants, environment, time, and other variables will require deeper studies and alternative approaches. IGVF welcomes interactions with or membership of projects that aim to explore or systematically address these areas of interest.

Many challenges lie ahead. Genomic technologies, both experimental and computational, are developing rapidly, and balancing the implementation of the newest scalable tools with continuing standards to ensure data interoperability will require attention. While data generation technologies have increased throughput exponentially over the last 15 years, the amount of data needed to build accurate models of genome function is unknown, and fully realizing the goal of mapping the impact of genomic variation on function will require additional advances in both experimental and computational methods. For all of these challenges, the framework developed by the IGVF Consortium to develop and benchmark methods, refine best practices and standards, and share data and methods will drive scientific discoveries in human health and disease for years to come.

Acknowledgements

This work was supported by the NIH NHGRI IGVF Program (UM1HG011966, UM1HG011969, UM1HG011972, UM1HG011989, UM1HG011996, UM1HG012003, UM1HG012010, UM1HG012053, UM1HG011986, UM1HG012076, UM1HG012077, U01HG011952, U01HG011967, U01HG012009, U01HG012022, U01HG012039, U01HG012064, U01HG012069, U01HG012041, U01HG012047, U01HG012051, U01HG012059, U01HG012079, U01HG012103, U24HG012012, U24HG012070), NIH NCI (R01CA197774), and the Novo Nordisk Foundation (NNF21SA0072102). Figures were illustrated by SciStories LLC and Vivian Yeung. We thank members of the IGVF External Consultants Panel (Guillaume Bourque, Prashant Mali, Judy Cho, Barbara Engelhardt, and Olga Troyanskaya) for critical feedback on the manuscript.

Author List

***Writing group* (ordered by contribution):** Jesse M. Engreitz, Heather A. Lawson, Harinder Singh, Lea M. Starita, Gary C. Hon, Hannah Carter, Nidhi Sahni, Timothy E. Reddy, Xihong Lin, Yun Li, Nikhil V. Munshi, Maria H. Chahrour, Alan P. Boyle, Benjamin C. Hitz, Ali Mortazavi, Mark Craven, Karen L. Mohlke, Luca Pinello, Ting Wang

***Steering Committee Co-Chairs* (alphabetical by last name):** Anshul Kundaje, Karen L. Mohlke, Feng Yue

***Code of Conduct Committee* (alphabetical by last name):** Sarah Cody, Nina P. Farrell, Michael I. Love, Lara A. Muffley, Michael J. Pazin, Fairlie Reese, Eric Van Buren

Working Group and Focus Group Co-Chairs (alphabetical by last name):

Catalog: Kushal Dey, Benjamin C. Hitz, Michael I. Love, Lea M. Starita, Feng Yue

Characterization: Gary C. Hon, Martin Kircher, Timothy E. Reddy

Computational Analysis, Modeling, and Prediction: Xihong Lin, Jian Ma, Predrag Radivojac

Project Design: Brunilda Balliu, Jesse M. Engreitz, Nidhi Sahni

Mapping: Nina P. Farrell, Brian A. Williams

Networks: Hannah Carter, Danwei Huangfu

Standards and Pipelines: Anshul Kundaje, Luca Pinello

Cardiometabolic: Nikhil V. Munshi, Chong Y. Park, Thomas Quertermous

Cellular Programs and Networks: Hannah Carter, Jishnu Das

Coding Variants: Michael A. Calderwood, Douglas M. Fowler, Predrag Radivojac, Lea M. Starita, Marc Vidal

CRISPR: Lucas Ferreira, Luca Pinello

Defining and Systematizing Function: Mark Craven, Sean D. Mooney, Vikas Pejaver

Enumerating Variants: Benjamin C. Hitz, Jingjing Zhao

Evolution: Steven Gazal, Evan Koch, Steven K. Reilly, Shamil Sunyaev

Imaging: Anne E. Carpenter

Immune: Jason D. Buenrostro, Christina S. Leslie, Rachel E. Savage

Impact on Diverse Populations: Stefanija Giric, Yun Li

iPSC: Chongyuan Luo, Kathrin Plath

MPRA: Alejandro Barrera, Michael I. Love, Max Schubach

Noncoding Variants: Jesse M. Engreitz, Andreas R. Gschwind, Jill E. Moore, Nidhi Sahni

Neuro: Nadav Ahituv, Maria H. Chahrour

Phenotypic Impact and Function: Kushal Dey, Xihong Lin, S. Stephen Yi

QTL/Statgen: Brunilda Balliu, Ingileif Hallgrimsdottir, Kyle Gaulton, Saori Sakaue

Single Cell: Sina Booeshaghi, Anshul Kundaje, Eugenio Mattei, Ali Mortazavi, Surag Nair, Lior Pachter, Austin Wang

Characterization Awards (contact PI, MPIs (alphabetical by last name), other members (alphabetical by last name)):

UM1HG011966: Jay Shendure[1,5,171,172,173], Nadav Ahituv[2], Martin Kircher[3,4], Vikram Agarwal[1,174], Andrew Blair[2], Theofilos Chalkiadakis[4], Florence M. Chardon[1], Pyaree M Dash[4], Chengyu Deng[2], Nobuhiko Hamazaki[1], Pia Keukeleire[3], Connor Kubo[1], Jean-Benoît Lalanne[1], Thorben Maass[3], Beth Martin[1], Troy A. McDiarmid[1], Mai Nobuhara[2], Nicholas F Page[2], Sam Regalado[1], Max Schubach[4], Jasmine Sims[2], Aki Ushiki[2], Jingjing Zhao[2]

UM1HG011969: Lea M. Starita[1,5], Douglas M. Fowler[1,5], Sabrina M. Best[1], Gabe Boyle[1], Nathan Camp[6], Silvia Casadei[1], Estelle Y. Da[7], Moez Dawood[5,8], Samantha C. Dawson[6], Shawn Fayer[1], Audrey Hamm[1], Richard G. James[6], Gail P. Jarvik[1], Abbye E. McEwen[1,5,9], Nick Moore[7], Lara A. Muffley[1], Sriram Pendyala[1], Nicholas A. Popp[1], Mason Post[1], Alan F. Rubin[7], Jay Shendure[1,5,171,172,173], Nahum T. Smith[1], Jeremy Stone[5], Malvika Tejura[1], Ziyu R. Wang[1], Melinda K. Wheelock[1], Ivan Woo[1], Brendan D. Zapp[1]

UM1HG011972: Jesse M. Engreitz[10,11,12,61], Thomas Quertermous[13], Dulguun Amgalan[10,11], Aradhana Aradhana[10], Sophia M. Arana[10], Michael C. Bassik[10], Julia R. Bauman[10], Asmita Bhattacharya[10], Xiangmeng Shawn Cai[10,11,22], Ziwei Chen[14], Stephanie Conley[10,11], Salil Deshpande[15], Benjamin R. Doughty[10], Peter P. Du[10], James A. Galante[10], Casey Gifford[10,11,16,17], William J. Greenleaf[10,161], Andreas R. Gschwind[10], Katherine Guo[10,11], Revant Gupta[10], Sarasa Isobe[18], Evelyn Jagoda[12,13], Nimit Jain[10, 215], Hank Jones[10,11], Helen Y. Kang[10,11], Samuel H. Kim[162], YeEun Kim[163], Sandy Klemm[10], Anshul Kundaje[10,14], Ramen Kundu[13], Soumya Kundu[14], Mauro Lago-Docampo[18], Yannick C.

Lee-Yow[10,11], Roni Levin-Konigsberg[10], Daniel Y. Li[13], Dominik Lindenhofer[19], X. Rosa Ma[10,11], Georgi K. Marinov[10], Gabriella E. Martyn[10,11], Chloe V. McCreery[10], Eyal Metzl-Raz[10], Joao P. Monteiro[13], Michael T. Montgomery[10,11], Kristy S. Mualim[11,20,164], Chad Munger[10,11], Glen Munson[12], Tri C. Nguyen[10,11], Trieu Nguyen[13], Brian T. Palmisano[13], Anusri Pampari[14], Chong Y. Park[13], Marlene Rabinovitch[18], Markus Ramste[13], Judhajeet Ray[12], Kevin R. Roy[10,21], Oriane M. Rubio[11], Julia M. Schaepe[22], Gavin Schnitzler[13], Jacob Schreiber[10], Disha Sharma[13], Maya U. Sheth[10,11,22], Huitong Shi[13], Vasundhara Singh[12], Riya Sinha[23], Lars M. Steinmetz[10,19,21], Jason Tan[10,11,14], Anthony Tan[10,11], Josh Tycko[10], Raeline C. Valbuena[10], Valeh Valiollah Pour Amiri[10], Mariëlle J.F.M. van Kooten[10], Alun Vaughan-Jackson[10], Anthony Venida[10], Chad S. Weldy[13], Matthew D. Worssam[13], Fan Xia[10,11], David Yao[10], Tony Zeng[10,11], Quanyi Zhao[13], Ronghao Zhou[10,11]

UM1HG011989: Marc Vidal[24,25], Michael A. Calderwood[24,25,26], Anne E. Carpenter[27], Zitong Sam Chen[27], Beth A. Cimini[27], Georges Coppin[24,25,26,28], Atina G. Coté[29,30,31], Marzieh Haghighi[27], Tong Hao[24,25,26], David E. Hill[24,25,26], Jessica Lacoste[29,30], Florent Laval[24,25,26,28,184], Chloe Reno[29,30], Frederick P. Roth[29,30,31,32], Shantanu Singh[27], Kerstin Spirohn-Fitzgerald[24,25], Mikko Taipale[29,30], Tanisha Teelucksingh[29], Maxime Tixhon[24,25,26,185], Anupama Yadav[24,25,26], Zhipeng Yang[24,25,26]

UM1HG011996: Gary C. Hon[33,34,35], W. Lee Kraus[33,34], Nikhil V. Munshi[36,37], Daniel A. Armendariz[33], Maria H. Chahrour[38,211,212,213,214], Ashley E. Dederich[39], Ashlesha Gogate[38], Lauretta El Hayek[38], Sean C. Goetsch[36], Kiran Kaur[38], Hyung Bum Kim[33], Melissa K. McCoy[39], Mpathi Z. Nzima[33], Carlos A. Pinzón-Arteaga[40], Bruce A. Posner[39], Daniel A. Schmitz[37], Sushama Sivakumar[36,37], Anjana Sundarrajan[33], Lei Wang[33], Yihan Wang[33], Jun Wu[37], Lin Xu[40,41], Jian Xu[42], Leqian Yu[37], Yanfeng Zhang[40], Huan Zhai[33], Qinbo Zhou[40]

UM1HG012003: Hyejung Won[43,44], Michael I. Love[43,204], Karen L. Mohlke[43], Jessica L. Bell[43,44], K. Alaine Broadaway[43], Katherine N. Degner[43,44], Amy S. Etheridge[43], Stefanija Giric[43], Beverly H. Koller[43], Yun Li[43,204], Won Mah[43,44], Wancen Mu[204], Kimberly D. Ritola[44,205], Jonathan D. Rosen[43], Sarah A. Schoenrock[43,44], Rachel A. Sharp[43,44]

UM1HG012010: Luca Pinello[45,61], Daniel Bauer[47,48], Guillaume Lettre[49,50], Richard Sherwood[51], Basheer Becerra[47,48], Logan J. Blaine[45,52], Eric Che[45,47,217], Lucas Ferreira[52,53], Matthew J. Francoeur[51], Ellie N. Gibbs[51], Nahye Kim[45,54,217], Emily M. King[45,54,217,218], Benjamin P. Kleinstiver[45,54,217], Estelle Lecluze[49], Zhijian Li[45,46], Zain M. Patel[45,46], Quang Vinh Phan[51], Jayoung Ryu[45,52], Marlena L Starr[53], Ting Wu[48,53]

UM1HG012053: Charles A. Gersbach[55,56], Gregory E. Crawford[56,57], Timothy E. Reddy[58], Andrew S. Allen[58], William H. Majoros[58], Nahid Iglesias[55,56], Alejandro Barrera[56,58], Ruhi Rai[56], Revathy Venukuttan[56], Boxun Li[55,56], Taylor Anglen[56,59], Lexi R. Bounds[55,56], Marisa C. Hamilton[56], Siyan Liu[56], Sean R. McCutcheon[55,56], Christian D. McRoberts Amador[56,60], Samuel J. Reisman[56,59], Maria A. ter Weele[55,56], Josephine C. Bodle[55,56], Helen L. Streff[55,56], Keith Siklenka[58], Kari Strouse[58]

Mapping Awards (contact PI, MPIs (alphabetical by last name), other members (alphabetical by last name)):

UM1HG011986: Jason D. Buenrostro[61,62], Bradley E. Bernstein[61,63], Juliana Babu[61,62], Guillermo Barreto Corona[61], Kevin Dong[61], Fabiana M. Duarte[61,62], Neva C. Durand[61], Charles B. Epstein[61], Kaili Fan[61,62,98], Nina P. Farrell[61], Elizabeth Gaskell[61], Amelia W. Hall[61], Alexandra M. Ham[61], Mei K. Knudson[61], Eugenio Mattei[61], Rachel E. Savage[61,62], Noam Shoresh[61], Siddarth Wekhande[61], Cassandra M. White[61], Wang Xi[61,62]

UM1HG012076: Ansuman T. Satpathy[64,65,66], M. Ryan Corces[73,74,187], Serena H. Chang[73,74,187], Iris M. Chin[73,74,187], James M. Gardner[75,76], Zachary A. Gardell[73,74,187], Jacob C. Gutierrez[64,66], Alia W. Johnson[73,74,187], Lucas Kampman[73,74,187], Maya Kasowski[64,72], Caleb A. Lareau[64,65,66], Vincent Liu[64,66], Leif S. Ludwig[67,68], Christopher S. McGinnis[64,65,66], Shreya Menon[73,74,187], Anita Qualls[75,76], Katalin Sandor[64,65,66], Adam W. Turner[73,74,187], Chun J. Ye[69,70,71], Yajie Yin[64,66], Wenxi Zhang[64]

UM1HG012077: Ali Mortazavi[188,189], Barbara J. Wold[190,191], Sina Booeshaghi[190], Maria Carilli[192], Dayeon Cheong[188], Ghassan Filibam[188], Kim Green[188,193], Ingileif Hallgrimsdottir[190], Shimako Kawauchi[189], Charlene Kim[190], Heidi Liang[189], Rebekah Loving[190], Laura Luebbert[190], Grant MacGregor[188], Angel G Merchan[190], Lior Pachter[190,194], Elisabeth Rebboah[188], Fairlie Reese[188,189], Narges Rezaie[188,189], Jasmine Sakr[189,195], Delaney K. Sullivan[190], Nikki Swarna[192], Diane Trout[190], Sean Upchurch[190], Ryan Weber[188], Brian A. Williams[190]

Predictive Modeling Awards (contact PI, MPIs (alphabetical by last name), other members (alphabetical by last name)):

U01HG011952: Alan P. Boyle[196,197], Christopher P. Castro[196], Elysia Chou[196], Fan Feng[196], Andre Guerra[198], Yuanhao Huang[196], Linghua Jiang[196], Jie Liu[196], Ryan E. Mills[196,197], Weizhou Qian[196], Tingting Qin[196], Maureen A. Sartor[196,198], Rintsen N. Sherpa[196], Jinhao Wang[196], Yiqun Wang[196], Joshua D. Welch[196], Zhenhao Zhang[196], Nanxiang Zhao[196]

U01HG011967: Andrew S. Allen[58], William H. Majoros[58], Sayan Mukherjee[77,78,79], C. David Page[58], Shannon Clarke[58], Richard W. Doty[58], Yuncheng Duan[80], Raluca Gordan[58,79], Kuei-Yueh Ko[58], Shengyu Li[58], Boyao Li[58], Timothy E. Reddy[58], Alexander Thomson[58]

U01HG012009: Soumya Raychaudhuri[51,82], Alkes Price[83,84,82], Shamil Sunyaev[51,52], Thahmina A. Ali[81], Kushal K. Dey[81,83], Arun Durvasula[83,85], Manolis Kellis[46,220], Evan Koch[52], Saori Sakaue[51,82]

U01HG012022: Predrag Radivojac[86], Lilia M. Iakoucheva[87], Tulika Kakati[87], Sean D. Mooney[88], Yile Chen[88], Mariam Benazouz[88], Vikas Pejaver[89,90], Shantanu Jain[86], Daniel Zeiberg[86], M. Clara De Paolis Kaluza[86], Michelle Velyunskiy[86]

U01HG012039: Mark Craven[91], Audrey Gasch[92], Kunling Huang[93], Yiyang Jin[91], Qiongshi Lu[91], Jiacheng Miao[91], Michael Ohtake[94], Eduardo Scopel[92], Robert D. Steiner[95,96,97], Yuriy Sverchkov[91]

U01HG012064: Zhiping Weng[98], Manuel Garber[98], Xihong Lin[84,100], Yu Fu[98], Natalie Haas[98], Xihao Li[43,84,204], Nishigandha Phalke[98], Shuo C. Shan[98], Nicole Shedd[98], Eric Van Buren[84], Tianxiong Yu[98], Yi Zhang[101], Hufeng Zhou[84]

U01HG012064: Anshul Kundaje[10,14], Alexis Battle[102,103,104,105], Ziwei Chen[14], Salil Deshpande[15], Jesse M. Engreitz[10,11,12,61], Livnat Jerby[10], Eran Kotler[10], Soumya Kundu[10,14], Andrew R. Marderstein[64], Georgi K. Marinov[10], Stephen B. Montgomery[10,64,106], Surag Nair[14], AkshatKumar Nigam[10,14], Evin M. Padhi[64], Anusri Pampari[14], Aman Patel[14], Jonathan Pritchard[10], Ivy Raine[10], Vivekanandan Ramalingam[10], Kameron B. Rodrigues[64], Jacob M. Schreiber[10], Arpita Singhal[14], Riya Sinha[15], Valeh Valiollah Pour Amiri[10], Austin T. Wang[14]

Network Projects (contact PI, MPIs (alphabetical by last name), other members (alphabetical by last name)):

U01HG012041: Harinder Singh[107], Jishnu Das[107], Nidhi Sahni[108,110], Marisa Abundis[111], Deepa Bisht[108], Trirupa Chakraborty[107], Jingyu Fan[107], David R. Hall[107], Zarifeh H. Rarani[107], Abhinav K. Jain[108], Babita Kaundal[108], Swapnil Keshari[107], Daniel McGrail[113,114], Nicholas A. Pease[107], Vivian F. Yi[107], S. Stephen Yi[115,116]

U01HG012047: Hao Wu[117], Sreeram Kannan[118], Hongjun Song[119], Jingli Cai[120], Ziyue Gao[117], Ronni Kurzion[119], Julia I. Leu[117], Fan Li[117], Dongming Liang[117], Guo-li Ming[119], Kiran Musunuru[120], Qi Qiu[117], Junwei Shi[121], Yijing Su[119], Sarah Tishkoff[117], Ning Xie[117], Qian Yang[119], Wenli Yang[120], Hongjie Zhang[117], Zhijian Zhang[119]

U01HG012051: Danwei Huangfu[122,123], Michael A. Beer[124], Sharon Adeniyi[122,123], Hyein Cho[122,123], Ronald Cutler[125], Rachel A. Glenn[122,123,126], David Godovich[122,123], Anna-Katerina Hadjantonakis[122,123], Nan Hu[122,123], Svetlana Jovanic[122,123], Renhe Luo[122,123], Jin Woo Oh[124], Milad Razavi-Mohseni[124], Dustin Shigaki[124], Simone Sidoli[125], Thomas Vierbuchen[122,123], Xianming Wang[122,123], Breanna Williams[122,123], Jielin Yan[122,123], Dapeng Yang[122,123], Yunxiao Yang[124]

U01HG012059: Maike Sander[127], Hannah Carter[128], Kyle J. Gaulton[127], Bing Ren[129,130], Weronika Bartosik[129], Hannah S. Indralingam[129], Adam Klie[131], Hannah Mummey[131], Mei-Lin Okino[132], Gaowei Wang[127], Nathan R. Zemke[129],Kai Zhang[129], Han Zhu[127]

U01HG012079: Chongyuan Luo[133], Kathrin Plath[134], Noah Zaitlen[135], Brunilda Balliu[136,137,138], Jason Ernst[134,137], Justin Langerman[134], Terence Li[133], Yu Sun[134]

U01HG012103: Christina S. Leslie[199], Alexander Y. Rudensky[200,201], Preethi K. Periyakoil[199], Vianne R. Gao[199], Melanie H. Smith[202], Norman M. Thomas[199], Laura T. Donlin[202,203], Amit Lakhanpal[202], Kaden M. Southard[199], Rico C. Ardy[199]

Data and Administrative Coordinating Center Awards (contact PI, MPIs (alphabetical by last name), other members (alphabetical by last name)):

U24HG012103: J. Michael Cherry[10], Mark B. Gerstein[166,167,168,169,170], Kalina Andreeva[10], Pedro R. Assis[10], Beatrice Borsari[166,167], Eric Douglass[10], Shengcheng

Dong[10], Idan Gabdank[10], Keenan Graham[10], Benjamin C. Hitz[10], Otto Jolanki[10], Jennifer Jou[10], Meenakshi S. Kagda[10], Jin-Wook Lee[10], Mingjie Li[10], Khine Lin[10], Stuart R. Miyasato[10], Joel Rozowsky[166,167], Corinn Small[10], Emma Spragins[10], Forrest Y. Tanaka[10], Ian M. Whaling[10], Ingrid A. Youngworth[10], Cricket A. Sloan[10]

U24HG012103: Ting Wang[139,140], Feng Yue[175,176], Eddie Belter[140], Xintong Chen[175], Rex L. Chisholm[178], Sarah Cody[140], Patricia Dickson[180], Changxu Fan[139], Lucinda Fulton[140], Heather A. Lawson[139], Daofeng Li[139], Tina Lindsay[140], Yu Luan[175], Yuan Luo[179], Huijue Lyu[175], Xiaowen Ma[139], Jian Ma[165], Juan Macias-Velasco[139], Karen H. Miga[186], Kara Quaid[139], Nathan Stitziel[181], Barbara E. Stranger[177], Chad Tomlinson[140], Juan Wang[175], Wenjin Zhang[139], Bo Zhang[182], Guoyan Zhao[139,183,216], Xiaoyu Zhuo[139]

IGVF Affiliate Member Projects (Contact PIs, other members (alphabetical by last name)):

Kristen Brennand[219]

Alberto Ciccia[210], Samuel B. Hayward[210], Jen-Wei Huang[210], Giuseppe Leuzzi[210], Angelo Taglialatela[210], Tanay Thakar[210], Alina Vaitsiankova[210]

Kushal K. Dey[46,141], Thahmina A. Ali[141]

Steven Gazal[142,143,144], Artem Kim[142]

H. Leighton Grimes[209], Nathan Salomonis[209]

Rajat Gupta[13], Shi Fang[13], Vivian Lee-Kim[13]

Matthias Heinig[145,146,147], Corinna Losert[145,146]

Thouis R. Jones[12], Elisa Donnard[12], Maddie Murphy[12], Elizabeth Roberts[12], Susie Song[12]

Jill E. Moore[98]

Sara Mostafavi[221,112], Alexander Sasse[221], Anna Spiro[221]

Len A. Pennacchio[148,151], Momoe Kato[148], Michael Kosicki[148], Brandon Mannion[148], Neil Slaven[148]

Axel Visel[148,151]

Katherine S. Pollard[152,153,154], Shiron Drusinsky[152,153], Sean Whalen[152]

John Ray[1,172,208], Ingrid A. Harten[172], Ching-Huang Ho[172]

Steven K. Reilly[109]

Neville E. Sanjana[149,150], Christina Caragine[149,150], John A. Morris[149,150]

Davide Seruggia[155,156], Ana Patricia Kutschat[155,156], Sandra Wittibschlager[155,156]

Han Xu[108], Rongjie Fu[108], Wei He[108], Liang Zhang[108]

S. Stephen Yi[157,158], Daniel Osorio[157,158]

NHGRI Program Management (alphabetical by last name): Zo Bly[159], Stephanie Calluori [160,206], Daniel A. Gilchrist[160], Carolyn M. Hutter[160], Stephanie A. Morris[160], Michael J. Pazin[160], Ella K. Samer[160,207]

Affiliations:

1. Department of Genome Sciences, University of Washington, Seattle, WA, USA

2. Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA

3. Institute of Human Genetics, University Medical Center Schleswig-Holstein, University of Lübeck, 23562 Lübeck, Germany

4. Exploratory Diagnostic Sciences, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany

5. Brotman Baty Institute for Precision Medicine, Seattle, WA., USA

6. Center of immunotherapy and Immunity, Seattle Children’s Research Institute, Seattle, WA, USA

7. Bioinformatics Division, WEHI, Parkville, VIC, Australia

8. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

9. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA

10. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

11. Basic Science and Engineering Initiative, Stanford Children’s Health, Betty Irene Moore Children’s Heart Center, Stanford, CA, USA

12. The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA, USA

13. Division of Cardiovascular Medicine, School of Medicine, Stanford University

14. Department of Computer Science, Stanford University School of Medicine, Stanford, CA, USA

15. Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA

16. Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA

17. Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA

18. Division of Pediatric Cardiology and Cardiovascular Institute, Stanford University School of Medicine, Stanford University

19. European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany

20. Department of Biology, Stanford University, Stanford, CA, USA

21. Stanford Genome Technology Center, Palo Alto, CA., USA

22. Department of Bioengineering, Stanford University School of Engineering, Stanford, CA, USA

23. Department of Biomedical Informatics, Stanford University School of Medicine, Stanford, CA, USA

24. Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA

25. Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA

26. Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA

27. Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, Massachusetts

28. Laboratory of Viral Interactomes, GIGA Institute, University of Liège, Liège, Belgium

29. Donnelly Centre for Cellular and Biomolecular Research (CCBR), University of Toronto, Toronto, Ontario, Canada

30. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

31. Lunenfeld-Tanenbaum Research Institute (LTRI), Sinai Health System, Toronto, Ontario, Canada

32. Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

33. Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA

34. Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA

35. Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, US

36. Department of Internal Medicine, Division of Cardiology, University of Texas Southwestern Medical Center, Dallas, TX, USA

37. Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA

38. Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA

39. Department of Biochemistry, University of Texas Southwestern Medical Center, TX, USA

40. Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, U.S.A

41. Department of Pediatrics, Division of Hematology/Oncology, University of Texas Southwestern Medical Center, Dallas, TX, U.S.A

42. Children's Medical Center Research Institute, University of Texas Southwestern Medical Center, TX, USA

43. Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

44. Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

45. Department of Pathology, Harvard Medical School, Boston, MA, USA

46. Broad Institute of MIT and Harvard, Boston, MA, USA

47. Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA

48. Department of Pediatrics, Harvard Medical School, Boston, MA, USA

49. Montreal Heart Institute, Montreal, Quebec, H1T 1C8, Canada

50. Département de Médecine, Université de Montréal, Montréal, Quebec, H3T 1J4, Canada

51. Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

52. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

53. Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA

54. Center for Genomic Medicine and Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

55. Department of Biomedical Engineering, Duke University, Durham, NC, USA

56. Center for Advanced Genomic Technologies, Duke University, Durham, NC

57. Department of Pediatrics, Duke University, Durham, NC, USA

58. Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC

59. Department of Cell Biology, Duke University Medical Center, Durham, NC, USA

60. Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC, USA

61. Gene Regulation Observatory, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

62. Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA

63. Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA

64. Department of Pathology, Stanford University, Stanford, CA, USA

65. Parker Institute for Cancer Immunotherapy, San Francisco, CA, United States

66. Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, 94158, USA.

67. Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

68. Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany

69. Institute for Human Genetics, Department of Medicine, Division of Rheumatology, University of California, San Francisco, CA, United States

70. Parker Institute for Cancer Immunotherapy, San Francisco, CA, United States

71. Chan Zuckerberg Biohub, San Francisco, CA, United States

72. Sean N Parker Center for Allergy and Asthma Research, Stanford University, Stanford, CA, USA

73. Gladstone Institute of Neurological Disease, San Francisco, CA, USA

74. Department of Neurology, University of California San Francisco, San Francisco, CA, USA

75. Department of Surgery, University of California San Francisco, San Francisco, CA, USA

76. Diabetes Center, University of California San Francisco, San Francisco, CA, USA

77. Department of Statistical Science, Duke University, Durham, NC, USA

78. Department of Mathematics, Duke University, Durham, NC, USA

79. Department of Computer Science, Duke University. Durham, NC, USA

80. Department of Biology, Duke University, Durham NC, USA

81. Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York NY, USA

82. Department of Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

83. Department of Epidemiology, Harvard T.H.Chan School of Public Health, Boston, MA, USA

84. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

85. Department of Genetics, Harvard Medical School, Boston, MA, USA

86. Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA

87. Department of Psychiatry, University of California San Diego, La Jolla, CA 92093, USA

88. Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, USA

89. Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

90. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

91. Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

92. Department of Genetics, University of Wisconsin, Madison, WI, USA

93. Department of Statistics, University of Wisconsin, Madison, WI, USA

94. Department of Computer Sciences, University of Wisconsin, Madison, WI, USA

95. Department of Pediatrics, University of Wisconsin, Madison, WI, USA

96. PreventionGenetics Inc., Part of Exact Sciences, Marshfield, WI, USA

97. Marshfield Clinic Health System, Marshfield, WI, USA

98. Program in Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA, USA

99. Department of Data Science, Dana-Farber Cancer Institute, Boston, MA

100. Department of Statistics, Harvard University, Cambridge, MA

101. Department of Data Science, Dana-Farber Cancer Institute, Boston, MA

102. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

103. Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA

104. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

105. Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA

106. Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

107. Departments of Immunology and Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA

108. Department of Epigenetics and Molecular Carcinogenesis, University of Texas MD Anderson Cancer Center, Houston, TX, USA

109. Department of Genetics, Yale School of Medicine, New Haven, CT, USA

110. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

111. Departments of Immunology and Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA

112. Canadian Institute for Advanced Research, Toronto, ON, Canada

113. Center for Immunotherapy and Precision Immuno-Oncology, Cleveland Clinic, Cleveland, OH, USA

114. Center for Immunotherapy and Precision Immuno-Oncology, Cleveland Clinic, Cleveland, OH, USA

115. Livestrong Cancer Institutes, Department of Oncology, and Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA

116. Interdisciplinary Life Sciences Graduate Programs (ILSGP), and Oden Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX, USA

117. Department of Genetics, University of Pennsylvania, Philadelphia, PA., USA

118. Department of Electrical & Computer Engineering, University of Washington, Seattle, WA., USA

119. Department of Neuroscience, University of Pennsylvania, Philadelphia, PA., USA

120. Department of Medicine, University of Pennsylvania, Philadelphia, PA., USA

121. Department of Cancer Biology, University of Pennsylvania, Philadelphia, PA., USA

122. Developmental Biology Program, Sloan Kettering Institute, New York, NY, USA

123. Center for Stem Cell Biology, Sloan Kettering Institute for Cancer Research, New York, NY 10065, USA

124. Department of Biomedical Engineering and McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University; Baltimore, MD 21218, USA

125. Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA

126. Weill Cornell Graduate School of Medical Sciences, Weill Cornell Medicine, 1300 York Avenue, New York, NY 10065, USA

127. Department of Pediatrics, University of California, San Diego, USA

128. Department of Medicine, University of California, San Diego, USA

129. Department of Cellular and Molecular Medicine, University of California, San Diego, CA, USA

130. Center for Epigenomics, University of California, San Diego

131. Bioinformatics and Systems Biology Program, University of California, San Diego, CA USA

132. Biomedical Sciences Program, University of California, San Diego, CA USA

133. Department of Human Genetics, University of California Los Angeles, Los Angeles, CA USA

134. Department of Biological Chemistry, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

135. Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA

136. Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, CA, USA

137. Department of Computational Medicine, University of California Los Angeles, Los Angeles, CA, USA

138. Department of Biostatistics, University of California Los Angeles, Los Angeles, CA, USA

139. Department of Genetics, Washington University, St. Louis, MO., USA

140. McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO, USA

141. Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA

142. Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, CA, USA

143. Department of Quantitative and Computational Biology, University of Southern California, CA, USA

144. Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, CA, USA

145. Institute of Computational Biology, Helmholtz Zentrum Munich, Neuherberg, Germany

146. Department of Computer Science, School of Computation, Information and Technology, Technical University Munich, Munich, Germany

147. Munich Heart Alliance, DZHK (German Center for Cardiovascular Research), Munich, Germany

148. Lawrence Berkeley National Laboratory, Berkeley, CA 94720

149. New York Genome Center, New York, NY USA

150. Department of Biology, New York University, New York, NY USA

151. DOE Joint Genome Institute, Berkeley, CA USA

152. Gladstone Institutes, San Francisco, CA USA

153. University of California, San Francisco, CA USA

154. Chan Zuckerberg Biohub - San Francisco, San Francisco, CA USA

155. St. Anna Children’s Cancer Research Institute (CCRI), Vienna, Austria

156. CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

157. Livestrong Cancer Institutes, Department of Oncology, and Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA

158. Interdisciplinary Life Sciences Graduate Programs (ILSGP), and Oden Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX, USA

159. Division of Genomic Medicine, National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, MD, USA

160. Division of Genome Sciences, National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, MD, USA

161. Department of Applied Physics, Stanford University, Stanford, CA, USA 94305

162. Cancer Biology Program, Stanford University School of Medicine, Stanford, CA, USA 94305

163. Immunology Graduate Program and Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA 94305

164. Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA

165. Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

166. Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT, USA

167. Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA

168. Department of Computer Science, Yale University, New Haven, CT, USA

169. Department of Statistics & Data Science, Yale University, New Haven, CT, USA

170. Department of Biomedical Informatics & Data Science, Yale University, New Haven, CT, USA

171. Howard Hughes Medical Institute, Seattle, WA, USA

172. Systems Immunology, Benaroya Research Institute, Seattle, WA, USA

173. Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA

174. mRNA Center of Excellence, Sanofi Pasteur Inc., Waltham, MA, USA

175. Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine Northwestern University, Chicago, IL, USA

176. Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL, USA

177. Center for Genetic Medicine, Department of Pharmacology, Northwestern University, Chicago, IL, USA

178. Center for Genetic Medicine and Department of Cell and Developmental Biology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

179. Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University

180. Department of Pediatrics, Washington University, St. Louis, MO., USA

181. Department of Medicine, Washington University, St. Louis, MO., USA

182. Department of Developmental Biology, Washington University, St. Louis, MO., USA

183. Department of Pathology and Immunology, Washington University, St. Louis, MO., USA

184. TERRA Teaching and Research Centre, University of Liège, Gembloux, Belgium

185. Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium

186. UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA

187. Gladstone Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA

188. Department of Developmental and Cell Biology, UC Irvine, Irvine, CA., USA

189. Center for Complex Biological Systems, UC Irvine, Irvine, CA., USA

190. Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA USA

191. Richard N. Merkin Institute for Translational Research, California Institute of Technology, Pasadena, CA., USA

192. Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA., USA

193. Department of Neurobiology and Behavior, UC Irvine, Irvine, CA., USA

194. Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA., USA

195. Department of Pharmaceutical Sciences, UC Irvine, Irvine, CA., USA

196. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

197. Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

198. Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

199. Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA

200. Howard Hughes Medical Institute and Immunology Program at Sloan Kettering Institute, New York, NY, USA

201. Ludwig Center for Cancer Immunotherapy, Memorial Sloan Kettering Cancer Center, New York, NY, USA

202. Division of Rheumatology, Department of Medicine, Hospital for Special Surgery, New York, NY, USA

203. Weill Cornell Medical College and Graduate School, New York, NY, USA

204. Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

205. Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

206. Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, New York, NY, USA

207. Masters of Physician Assistant Studies Program, Colorado Mesa University, Grand Junction CO, USA

208. Department of Immunology, University of Washington, Seattle, WA, USA

209. Cincinnati Children’s Hospital, Cincinnati OH, USA

210. Department of Genetics and Development, Institute for Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA

211. Department of Neuroscience, University of Texas Southwestern Medical Center, Dallas, TX, USA

212. Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX, USA

213. Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA

214. Peter O'Donnell Jr. Brain Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA

215. Altos Labs Inc., 1300 Island Drive, Redwood City, CA, USA

216. Department of Neurology, Washington University, St. Louis, MO., USA

217. Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

218. PhD Program in Biological and Biomedical Sciences, Harvard University, Boston, MA, USA

219. Departments of Psychiatry and Genetics, Division of Molecular Psychiatry, Department of Genetics, Wu Tsai Institute, Yale University School of Medicine, New Haven, CT, USA

220. MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

221. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

References

1. Claussnitzer M et al. A brief history of human disease genetics. Nature 577, 179–189 (2020). [PubMed: 31915397] This review describes progress in identifying genomic variants associated with common and rare diseases, and approaches needed to combine these data with maps of genome function to advance diagnostic and therapeutic strategies.
2. Loos RJF 15 years of genome-wide association studies and no signs of slowing down. Nat. Commun. 11, 5900 (2020). [PubMed: 33214558]
3. Green ED et al. Strategic vision for improving human health at The Forefront of Genomics. Nature 586, 683–692 (2020). [PubMed: 33116284]
4. Karczewski KJ et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). [PubMed: 32461654]
5. Collins RL et al. A structural variation reference for medical and population genetics. Nature 581, 444–451 (2020). [PubMed: 32461652]
6. Sollis E et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 51, D977–D985 (2023). [PubMed: 36350656]
7. Landrum MJ et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, D980–5 (2014). [PubMed: 24234437]
8. Rehm HL et al. ClinGen--the clinical genome resource. N. Engl. J. Med. 372, 2235–2242 (2015). [PubMed: 26014595]
9. Zhou W et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom 2, 100192 (2022). [PubMed: 36777996]
10. Taliun D et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021). [PubMed: 33568819]
11. Backman JD et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature 599, 628–634 (2021). [PubMed: 34662886]

12. Karczewski KJ et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genomics 2, 100168 (2022). [PubMed: 36778668]

13. Doolittle WF, Brunet TDP, Linquist S & Gregory TR Distinguishing between 'function' and 'effect' in genome biology. Genome Biol. Evol. 6, 1234–1237 (2014). [PubMed: 24814287]

14. Kellis M et al. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. U. S. A. 111, 6131–6138 (2014). [PubMed: 24753594]

15. ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710 (2020). [PubMed: 32728249] An exemplary team science effort which has led to development of methods, data resources, and standards enabling fundamental advances in understanding gene regulation and genome function.

16. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]

17. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020). [PubMed: 32913098] This latest flagship manuscript from the Genotype-Tissue Expression Consortium maps how genomic variation regulates gene expression across human tissues, providing a resource for interpreting the molecular effects of variants associated with common diseases.

18. Võsa U et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. 53, 1300–1310 (2021). [PubMed: 34475573]

19. Li YI et al. RNA splicing is a primary link between genetic variation and disease. Science 352, 600–604 (2016). [PubMed: 27126046]

20. Consortium HuBMAP. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature 574, 187–192 (2019). [PubMed: 31597973]

21. Regev A et al. The Human Cell Atlas. eLife 6, (2017).

22. Matreyek KA et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. 50, 874–882 (2018). [PubMed: 29785012]

23. Findlay GM et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature 562, 217–222 (2018). [PubMed: 30209399]

24. Esposito D et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biology vol. 20 Preprint at 10.1186/s13059-019-1845-6 (2019).

25. Luck K et al. A reference map of the human binary protein interactome. Nature 580, 402–408 (2020). [PubMed: 32296183]

26. Szklarczyk D et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 49, D605–D612 (2021). [PubMed: 33237311]

27. Pacini C et al. Integrated cross-study datasets of genetic dependencies in cancer. Nat. Commun. 12, 1661 (2021). [PubMed: 33712601]

28. International Common Disease Alliance. ICDA Recommendations and White Paper. https://icda.bio.

29. Abdellaoui A, Yengo L, Verweij KJH & Visscher PM 15 years of GWAS discovery: Realizing the promise. Am. J. Hum. Genet. 110, 179–194 (2023). [PubMed: 36634672]

30. Rehm HL & Fowler DM Keeping up with the genomes: scaling genomic variant interpretation. Genome Med. 12, 5 (2019). [PubMed: 31892366]

31. Bentley AR, Callier S & Rotimi CN Diversity and inclusion in genomic research: why the uneven progress? J. Community Genet. 8, 255 (2017). [PubMed: 28770442]

32. Sirugo G, Williams SM & Tishkoff SA The Missing Diversity in Human Genetic Studies. Cell 177, (2019).

33. Lappalainen T & MacArthur DG From variant to function in human disease genetics. Science 373, 1464–1468 (2021). [PubMed: 34554789]

34. Findlay GM Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. Hum. Mol. Genet. 30, R187–R197 (2021). [PubMed: 34338757]

35. Hu Y et al. Single-cell multi-scale footprinting reveals the modular organization of DNA regulatory elements. bioRxiv (2023) doi:10.1101/2023.03.28.533945.

36. Granja JM et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. 53, 403–411 (2021). [PubMed: 33633365]

37. Kartha VK et al. Functional inference of gene regulation using single-cell multi-omics. Cell Genom 2, (2022).

38. Kotliar D et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. (2019) doi:10.7554/eLife.43803.

39. Benner C et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501 (2016). [PubMed: 26773131]

40. Wang G, Sarkar A, Carbonetto P & Stephens M A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. J. R. Stat. Soc. Series B Stat. Methodol. 82, 1273–1300 (2020). [PubMed: 37220626]

41. Weissbrod O et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat. Genet. 52, (2020).

42. Wang QS et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nat. Commun. 12, 3394 (2021). [PubMed: 34099641]

43. Cuella-Martin R et al. Functional interrogation of DNA damage response variants with base editing screens. Cell 184, 1081–1097.e19 (2021). [PubMed: 33606978]

44. Morris JA et al. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. Science 380, eadh7699 (2023). [PubMed: 37141313]

45. Martin-Rufino JD et al. Massively parallel base editing to map variant effects in human hematopoiesis. Cell 186, 2456–2474.e24 (2023). [PubMed: 37137305]

46. Anzalone AV, Koblan LW & Liu DR Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nat. Biotechnol. 38, 824–844 (2020). [PubMed: 32572269]

47. Hanna RE et al. Massively parallel assessment of human variants with base editor screens. Cell 184, 1064–1080.e20 (2021). [PubMed: 33606977]

48. Klann TS et al. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nature Biotechnology vol. 35 561–568 Preprint at 10.1038/nbt.3853 (2017).

49. Fulco CP et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science 354, 769–773 (2016). [PubMed: 27708057]

50. Canver MC et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, (2015). This study applied CRISPR-Cas9 screens to dissect a GWAS-nominated enhancer of BCL11A, a negative regulator of fetal haemoglobin expression during erythropoiesis, and motivated the development of enhancer-targeting CRISPR therapeutics for sickle cell disease.

51. Arnold CD et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074–1077 (2013). [PubMed: 23328393]

52. Melnikov A et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. 30, 271–277 (2012). [PubMed: 22371084]

53. Bergman DT et al. Compatibility rules of human enhancer and promoter sequences. Nature 607, 176–184 (2022). [PubMed: 35594906]

54. Vockley CM et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res. 25, 1206–1214 (2015). [PubMed: 26084464]

55. Klein JC et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods 17, 1083–1091 (2020). [PubMed: 33046894]

56. Patwardhan RP et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. 27, 1173–1175 (2009). [PubMed: 19915551]

57. Agarwal V et al. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. bioRxiv (2023) doi:10.1101/2023.03.05.531189.

58. Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). [PubMed: 34265844]

59. Alipanahi B, Delong A, Weirauch MT & Frey BJ Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 33, 831–838 (2015). [PubMed: 26213851]

60. Zhou J & Troyanskaya OG Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods 12, 931–934 (2015). [PubMed: 26301843] The study develops a deep learning framework (DeepSEA) trained on chromatin profiling data to predict effects of single-nucleotide genomic variants on transcription factor binding and chromatin state.

61. Avsec Ž et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. 53, 354–366 (2021). [PubMed: 33603233] This study introduces the BPNet model, a convolutional neural network to predict basepair-resolution epigenomic data from DNA sequence, and applies this framework to learn rules of the regulatory syntax underlying transcription factor binding.

62. Beer MA Predicting enhancer activity and variant impact using gkm-SVM. Hum. Mutat. 38, 1251–1258 (2017). [PubMed: 28120510]

63. Chen KM, Wong AK, Troyanskaya OG & Zhou J A sequence-based global map of regulatory activity for deciphering human genetics. Nat. Genet. 54, 940–949 (2022). [PubMed: 35817977]

64. Avsec Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods 18, 1196–1203 (2021). [PubMed: 34608324]

65. Boyle EA, Li YI & Pritchard JK An expanded view of complex traits: from polygenic to omnigenic. Cell 169, 1177 (2017). [PubMed: 28622505]

66. Rentzsch P, Witten D, Cooper GM, Shendure J & Kircher M CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894 (2018).

67. Han J-DJ Understanding biological functions through molecular networks. Cell Res. 18, 224–237 (2008). [PubMed: 18227860]

68. Kryshtafovych A, Schwede T, Topf M, Fidelis K & Moult J Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins 89, 1607–1617 (2021). [PubMed: 34533838] Work by CASP over almost 20 years illustrates how community efforts to develop gold-standard data, benchmarks, and critical assessments can facilitate development of predictive models of protein structure and function, with CASP XIV marking a major advance through the introduction of AlphaFold2.

69. The Critical Assessment of Genome Interpretation Consortium. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. Genome Biol. 25, 53 (2024). [PubMed: 38389099] This paper reports a collaborative effort to independently assess computational models for interpreting the effects of variants on molecular phenotypes and disease risk, and demonstrates their utility in clinical and research applications.

70. Ma S et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell 183, 1103–1116.e20 (2020). [PubMed: 33098772] This study introduces SHARE-seq and demonstrates how single-cell multiomic data enables mapping dynamics of regulatory element activity across differentiation states by correlating distal enhancers with target genes.

71. Tran V et al. High sensitivity single cell RNA sequencing with split pool barcoding. bioRxiv 2022.08.27.505512 (2022) doi:10.1101/2022.08.27.505512.

72. Xu Y et al. An atlas of genetic scores to predict multi-omic traits. Nature 616, 123–131 (2023). [PubMed: 36991119]

73. Nathan A et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. Nature 606, 120–128 (2022). [PubMed: 35545678]

74. Perez RK et al. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. Science 376, eabf1970 (2022). [PubMed: 35389781]

75. Yazar S et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science 376, eabf3041 (2022). [PubMed: 35389779]

76. Gate RE et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. 50, 1140–1150 (2018). [PubMed: 29988122]

77. Adamson B et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell 167, 1867–1882.e21 (2016). [PubMed: 27984733]

78. Dixit A et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell 167, 1853–1866.e17 (2016). [PubMed: 27984732]

79. Replogle JM et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell (2022) doi:10.1016/j.cell.2022.05.013.

80. Sahni N et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell 161, 647–660 (2015). [PubMed: 25910212] Systematic ORF screens showed that a majority of coding variants in Mendelian disorders affect protein interaction networks, providing a resource to benchmark predictors of variant effects.

81. Fayer S et al. Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. Am. J. Hum. Genet. 108, 2248–2258 (2021). [PubMed: 34793697] This study illustrates how experimentally derived variant effect maps can have high clinical utility in interpreting variants for Mendelian diseases.

82. Starita LM et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics 200, 413–422 (2015). [PubMed: 25823446]

83. Sun S et al. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 26, 670–680 (2016). [PubMed: 26975778]

84. Bray M-A et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat. Protoc. 11, 1757–1774 (2016). [PubMed: 27560178]

85. Fulco CP et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat. Genet. 51, 1664–1669 (2019). [PubMed: 31784727]

86. Sakaue S et al. Tissue-specific enhancer-gene maps from multimodal single-cell data identify causal disease alleles. medRxiv (2022) doi:10.1101/2022.10.27.22281574.

87. Weeks EM et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Nat. Genet. 55, 1267–1276 (2023). [PubMed: 37443254]

88. Nasser J et al. Genome-wide enhancer maps link risk variants to disease genes. Nature 593, 238–243 (2021). [PubMed: 33828297]

89. Schnitzler GR et al. Convergence of coronary artery disease genes onto endothelial cell programs. Nature (2024). 10.1038/s41586-024-07022-x

90. Finucane HK et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. 50, 621–629 (2018). [PubMed: 29632380]

91. Forgetta V et al. An effector index to predict target genes at GWAS loci. Hum. Genet. 141, 1431–1447 (2022). [PubMed: 35147782]

92. Ghoussaini M et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 49, D1311–D1320 (2021). [PubMed: 33045747]

93. Gschwind AR et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. bioRxiv 2023.11.09.563812 (2023) doi:10.1101/2023.11.09.563812.

94. Karollus A, Mauermeier T & Gagneur J Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biol. 24, 56 (2023). [PubMed: 36973806]

95. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 36, 1133–1137 (2004). [PubMed: 15514660]

96. Hogan A et al. Knowledge Graphs. arXiv [cs.AI] (2020).

97. Feng F et al. GenomicKB: a knowledge graph for the human genome. Nucleic Acids Res. 51, D950–D956 (2023). [PubMed: 36318240]

98. Chandak P, Huang K & Zitnik M Building a knowledge graph to enable precision medicine. Sci Data 10, 67 (2023). [PubMed: 36732524]

99. Lobentanzer S et al. Democratizing knowledge representation with BioCypher. Nat. Biotechnol. 41, 1056–1059 (2023). [PubMed: 37337100]

100. Ambrosini G et al. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. Genome Biol. 21, 114 (2020). [PubMed: 32393327]

101. de Boer CG & Taipale J Hold out the genome: a roadmap to solving the cis-regulatory code. Nature 625, 41–50 (2024). [PubMed: 38093018]

102. Yuan H & Kelley DR scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat. Methods 19, 1088–1096 (2022). [PubMed: 35941239]

103. Inoue F et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 27, 38–52 (2017). [PubMed: 27831498]

104. Xie S, Duan J, Li B, Zhou P & Hon GC Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell 66, 285–299.e5 (2017). [PubMed: 28416141]

105. Gasperini M et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 1516 (2019). [PubMed: 30849375]

106. Reilly SK et al. Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR–FlowFISH. Nat. Genet. 53, 1166–1176 (2021). [PubMed: 34326544]

107. Schraivogel D et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat. Methods 17, 629–635 (2020). [PubMed: 32483332]

108. McGinnis CS et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods 16, 619–626 (2019). [PubMed: 31209384]

109. Daniel B et al. Divergent clonal differentiation trajectories of T cell exhaustion. Nat. Immunol. 23, 1614–1627 (2022). [PubMed: 36289450]

110. Rebboah E et al. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. Genome Biol. 22, 286 (2021). [PubMed: 34620214]

111. Pratapa A, Jalihal AP, Law JN, Bharadwaj A & Murali TM Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods 17, 147–154 (2020). [PubMed: 31907445]

112. Zhang Y, Qi G, Park J-H & Chatterjee N Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat. Genet. 50, 1318–1326 (2018). [PubMed: 30104760]

113. O'Connor LJ The distribution of common-variant effect sizes. Nat. Genet. 53, 1243–1249 (2021). [PubMed: 34326547]

114. Lewis CM & Vassos E Polygenic risk scores: from research tools to clinical instruments. Genome Med. 12, 44 (2020). [PubMed: 32423490]

115. Hekselman I & Yeger-Lotem E Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat. Rev. Genet. 21, 137–150 (2020). [PubMed: 31913361]

116. Uffelmann E et al. Genome-wide association studies. Nature Reviews Methods Primers 1, 1–21 (2021).

117. Weissbrod O et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat. Genet. 54, 450–458 (2022). [PubMed: 35393596]

118. Heid IM et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat. Genet. 42, 949–960 (2010). [PubMed: 20935629]

119. Goossens GH, Jocken JWE & Blaak EE Sexual dimorphism in cardiometabolic health: the role of adipose tissue, muscle and liver. Nat. Rev. Endocrinol. 17, 47–66 (2021). [PubMed: 33173188]

120. Rajabli F et al. Ancestral origin of ApoE ε4 Alzheimer disease risk in Puerto Rican and African American populations. PLoS Genet. 14, e1007791 (2018). [PubMed: 30517106]

121. Blue EE, Horimoto ARVR, Mukherjee S, Wijsman EM & Thornton TA Local ancestry at APOE modifies Alzheimer's disease risk in Caribbean Hispanics. Alzheimers Dement. 15, 1524–1532 (2019). [PubMed: 31606368]

122. Baxter SM et al. Centers for Mendelian Genomics: A decade of facilitating gene discovery. Genet. Med. 24, 784–797 (2022). [PubMed: 35148959]

123. Costanzo MC et al. The Type 2 Diabetes Knowledge Portal: An open access genetic resource dedicated to type 2 diabetes and related traits. Cell Metab. 35, 695–710.e6 (2023). [PubMed: 36963395]

124. Richards S et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. 17, 405–424 (2015). [PubMed: 25741868]

125. Scott A et al. Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. Genome Biol. 23, 266 (2022). [PubMed: 36550560]

126. Radford EJ et al. Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. medRxiv 2022.06.10.22276179 (2022) doi:10.1101/2022.06.10.22276179.

127. Wojcik MH et al. Beyond the exome: what's next in diagnostic testing for Mendelian conditions. ArXiv (2023) doi:10.1002/ajmg.a.63053.

128. Miller DT et al. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. 24, 1407–1414 (2022). [PubMed: 35802134]

129. Pejaver V et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am. J. Hum. Genet. 109, 2163–2177 (2022). [PubMed: 36413997]

130. Brnich SE et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. 12, 1–12 (2019). [PubMed: 31892350]

131. Graham SE et al. The power of genetic diversity in genome-wide association studies of lipids. Nature 600, 675–679 (2021). [PubMed: 34887591] This study demonstrates how expanding genomic studies to include people of non-European ancestries will improve identification of functional variants and the portability of polygenic risk scores to diverse groups.

132. Musunuru K & Kathiresan S Genetics of Common, Complex Coronary Artery Disease. Cell 177, 132–145 (2019). [PubMed: 30901535]

133. Hamilton MC et al. Systematic elucidation of genetic mechanisms underlying cholesterol uptake. Cell Genom 3, 100304 (2023). [PubMed: 37228746]

134. Martin AR et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. 51, 584–591 (2019). [PubMed: 30926966]

135. Shi H et al. Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat. Commun. 12, 1098 (2021). [PubMed: 33597505]

136. Hou K et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat. Genet. 55, 549–558 (2023). [PubMed: 36941441]

137. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]

138. Gaziano JM et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. 70, 214–223 (2016). [PubMed: 26441289]

139. Kanoni S et al. Implicating genes, pleiotropy, and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis. Genome Biol. 23, 268 (2022). [PubMed: 36575460]

140. Sakaue S et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. 53, 1415–1424 (2021). [PubMed: 34594039]

141. Aragam KG et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nat. Genet. 54, 1803–1815 (2022). [PubMed: 36474045]

142. Tcheandjieu C et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. Nat. Med. 28, 1679–1692 (2022). [PubMed: 35915156]

143. Threadgill DW, Miller DR, Churchill GA & de Villena FP-M The collaborative cross: a recombinant inbred mouse population for the systems genetic era. ILAR J. 52, 24–31 (2011). [PubMed: 21411855]

144. Fowler DM et al. An Atlas of Variant Effects to understand the genome at nucleotide resolution. Genome Biol. 24, 147 (2023). [PubMed: 37394429]

145. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). [PubMed: 26978244]

146. Schatz MC et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genom 2, (2022).

147. Wang T et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature 604, 437–446 (2022). [PubMed: 35444317]

148. All of Us Research Program Investigators et al. The 'All of Us' Research Program. N. Engl. J. Med. 381, 668–676 (2019). [PubMed: 31412182]

149. Gilbert LA et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell 159, (2014).

150. Sherry ST et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001). [PubMed: 11125122]

151. Frankish A et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2019). [PubMed: 30357393]

152. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531 (2023). [PubMed: 36408920]

153. Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000). [PubMed: 10802651]

154. Del Toro N et al. The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res. 50, D648–D653 (2022). [PubMed: 34761267]

155. Vasilevsky NA et al. Mondo: Unifying diseases for the world, by the world. medRxiv (2022) doi:10.1101/2022.04.13.22273750.

156. Köhler S et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 49, D1207–D1217 (2021). [PubMed: 33264411]

157. Amberger JS, Bocchini CA, Scott AF & Hamosh A OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 47, D1038–D1043 (2019). [PubMed: 30445645]

158. Musunuru K et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, (2010).

159. Sort1, Encoded by the Cardiovascular Risk Locus 1p13.3, Is a Regulator of Hepatic Lipoprotein Export. Cell Metab. 12, 213–223 (2010). [PubMed: 20816088]

160. Claussnitzer M et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. 373, 895–907 (2015). [PubMed: 26287746]

161. Smemo S et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507, 371 (2014). [PubMed: 24646999]

162. Graham DB & Xavier RJ Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature 578, 527–539 (2020). [PubMed: 32103191]

163. Kim S, Eun HS & Jo E-K Roles of Autophagy-Related Genes in the Pathogenesis of Inflammatory Bowel Disease. Cells 8, (2019).

164. Singh NK, Singh NN, Androphy EJ & Singh RN Splicing of a Critical Exon of Human Survival Motor Neuron Is Regulated by a Unique Silencer Element Located in the Last Intron. Mol. Cell. Biol. 26, 1333 (2006). [PubMed: 16449646]

165. Hua Y et al. Antisense correction of SMN2 splicing in the CNS rescues necrosis in a type III SMA mouse model. Genes Dev. 24, 1634 (2010). [PubMed: 20624852]

166. Frangoul H et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. N. Engl. J. Med. 384, 252–260 (2021). [PubMed: 33283989]

167. Sankaran VG et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science 322, 1839–1842 (2008). [PubMed: 19056937]

Box 1:

Mapping the impact of genomic variation on genome function can reveal biological mechanisms and advance precision health

Selected examples (see also refs [1,28,29]):

Learning basic and disease biology:

- eQTL and gene knockdown studies of a CAD GWAS locus identified sortilin (*SORT1*) as a regulator of LDL cholesterol levels and elucidated its molecular function in LDL uptake[158,159]
- Epigenomic maps and variant-to-function studies revealed a role for transcription factors *IRX3/5* in regulating adipocyte browning to influence obesity[160,161]
- Characterization of risk variants for inflammatory bowel disease has identified multiple genes involved in autophagy, including *ATG16L1* and LRRK2, revealing new roles in myeloid and intestinal epithelial cells[162,163]

Guiding genetic diagnosis:

- Saturation genome editing of *BRCA1* led to improved diagnosis of inherited risk for breast and ovarian cancer[23]
- Functional variant annotations improve the applicability of polygenic risk scores across populations[117]

Guiding therapeutic development:

- Designed mutagenesis of *SMN2* identified an intronic splice enhancing sequence that guided development of antisense oligonucleotides to treat spinal muscular atrophy[164,165]
- Dissection of a GWAS locus led to identification of *BCL11A* as a repressor of fetal hemoglobin and development of CRISPR editors for sickle cell disease[50,166,167]

Box 2:

IGVF goals and approaches

- Characterize the impact of genomic variants, regulatory elements, and genes on molecular and cellular phenotypes — by analyzing naturally occurring or designed genomic perturbations across dozens of cellular models.
- Identify where and when regulatory elements and genes are active with resolution for individual cell types and states — by applying single-cell mapping technologies across hundreds of biological samples including cellular models, tissues, and environmental contexts.
- Predict the consequences of genomic variation on genome function and phenotype for previously unstudied variants and/or cellular contexts — by developing predictive computational models that can generalize across contexts and establishing benchmarking pipelines to evaluate and calibrate their accuracy.
- Study diverse cellular and disease systems, types of genomic variation, and aspects of genome function — by developing and applying a "map-perturb-predict" framework in which single-cell mapping, genomic perturbations, and predictive modeling are synergistically combined.
- Create an initial map that annotates the predicted effects of every possible single-nucleotide variant in the human genome on key aspects of genome function — by integrating models for how coding variants might alter protein function, how noncoding variants might affect gene expression, and how noncoding and coding variants might connect within molecular networks.
- Advance our understanding of the impact of genomic variation on disease — by exploring how best to apply IGVF resources to inform genetic diagnosis and to identify biological mechanisms of disease risk.
- Ensure that these advances are applicable to and inclusive of people of diverse sexes, ancestries, and populations — by studying individuals with different genetic backgrounds, assaying and predicting effects of variants observed in diverse populations, and studying diseases disproportionately affecting disadvantaged or under-represented populations.
- Catalyze research by others toward the long-term goal of understanding the impact of genomic variation — by partnering with the broader research community and developing resources and infrastructure to share IGVF data, methods, standards, and pipelines.

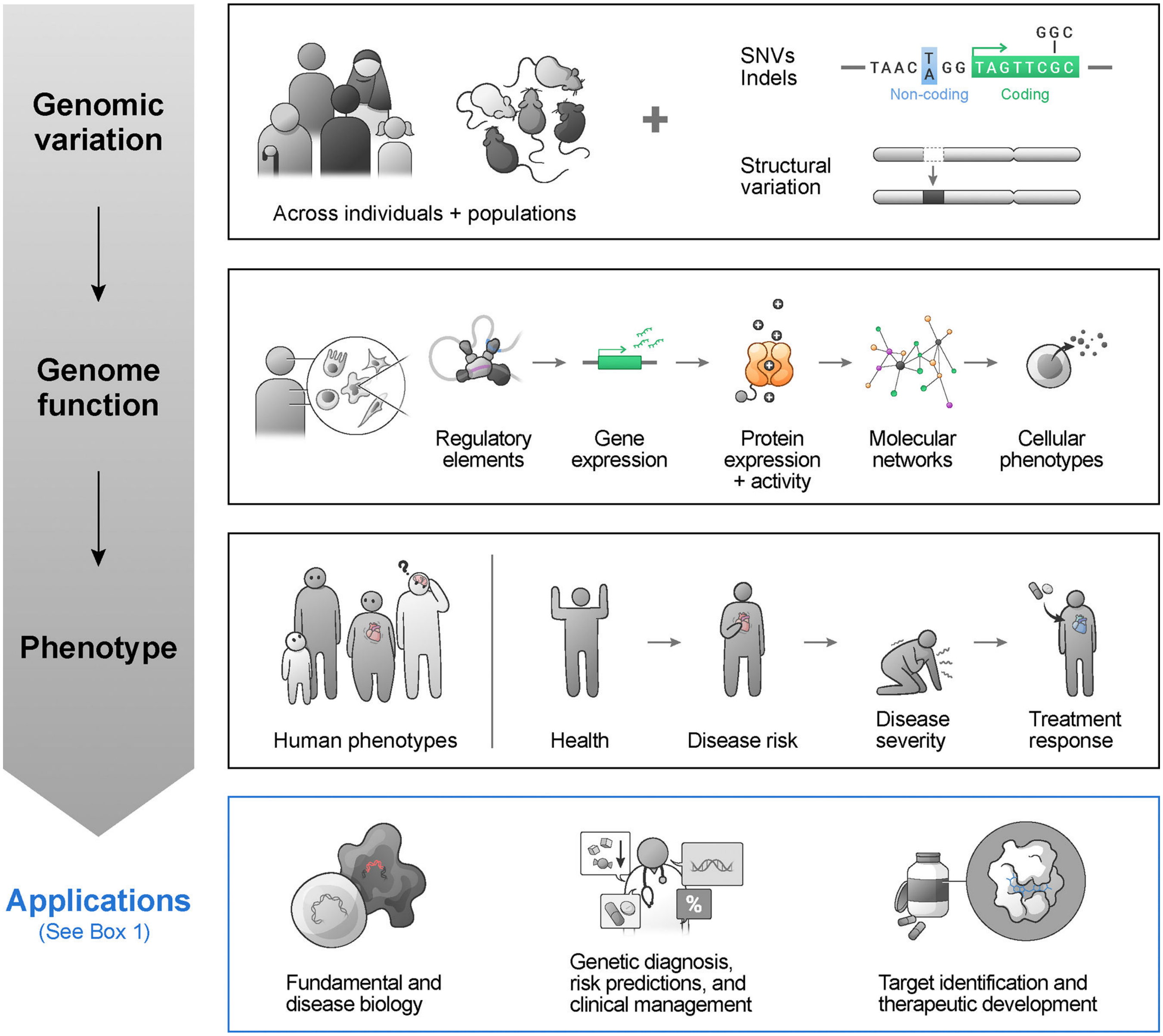

Figure 1. Genomic variation influences genome function and phenotype.

Genomic variation includes SNVs, indels, and structural variants, which can alter protein-coding sequences or noncoding sequences. Genome function encompasses the cell-type specific activities and interactions among regulatory elements, genes and proteins within molecular networks that underlie cellular phenotypes. Organismal phenotypes include quantitative and binary traits in health and disease.

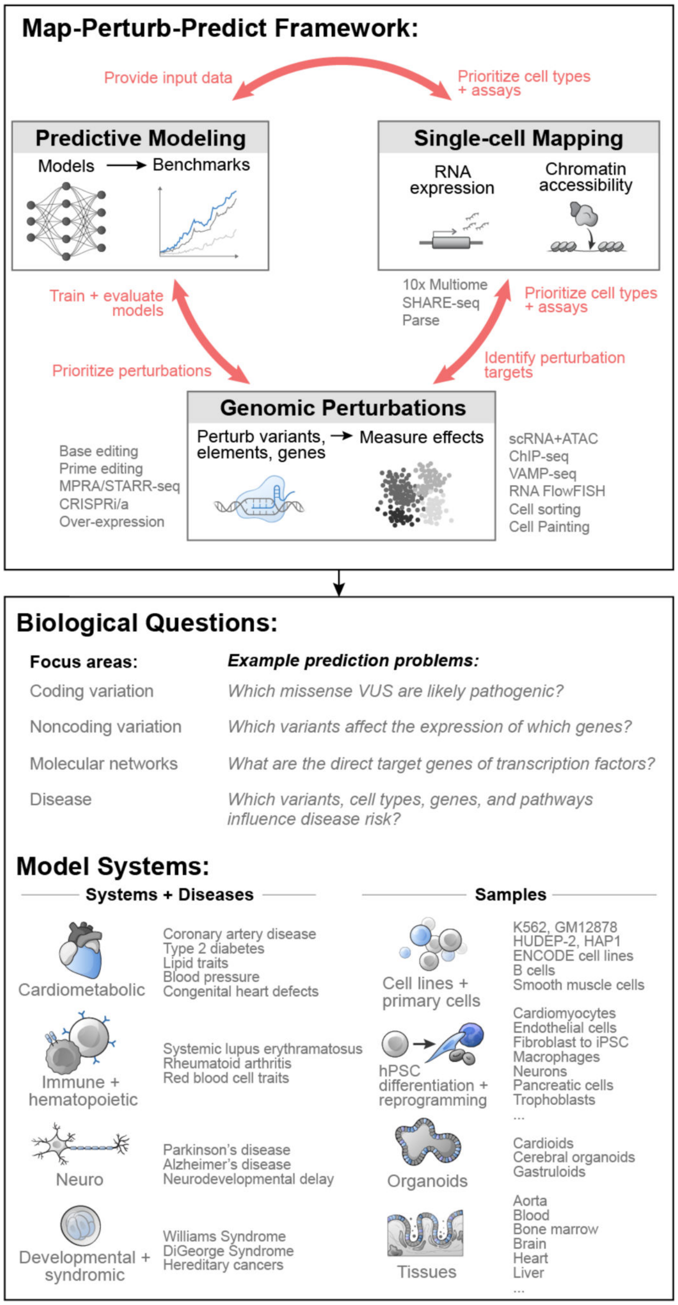

Figure 2. A map-perturb-predict framework to connect genome variation to genome function and phenotype.

(top) IGVF projects will apply single-cell mapping, genomic perturbations, and predictive modeling, which will interact in a synergistic and iterative fashion (red text); Gray: Examples of experimental approaches, including 10x Multiome, simultaneous high-throughput ATAC and RNA expression with sequencing (SHARE-seq)[70], and Parse Evercode (split-pool combinatorial indexing single-cell RNA-seq)[71], massively parallel reporter assays (MPRA)[52,56,57], Self-Transcribing Assay of RNA Reporters (STARR-seq)[51], CRISPR interference and activation (CRISPRi/a)[149], Variant Abundance by Massively Parallel sequencing (VAMP-seq)[22], RNA FlowFISH[85], and Cell Painting[84]. (bottom) IGVF projects will address a wide variety of biological questions and utilize diverse biological systems, models, and samples. hPSC: Human pluripotent stem cells, including embryonic stem cells and iPSCs.

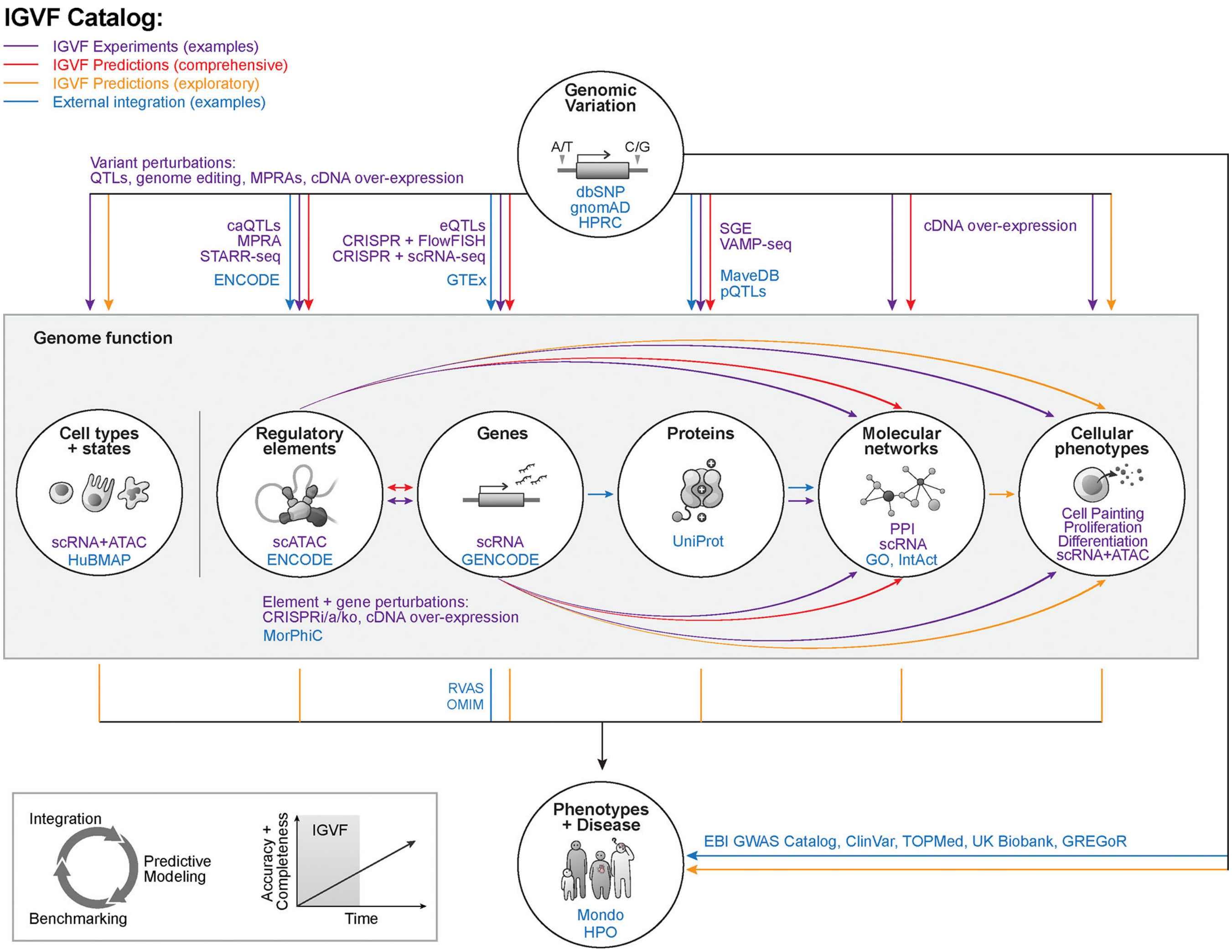

Figure 3. The IGVF Catalog of genome function and the impact of genomic variation.

IGVF will create a catalog linking genomic variation (top) to genome function (middle box) to phenotype (bottom). Purple: Examples of experimental methods applied by IGVF. Red: Relationships where IGVF plans to develop and apply computational models to comprehensively annotate all possible single-nucleotide variants across many cell types. Orange: Relationships where IGVF plans to develop and apply computational methods in a more targeted fashion, for example in the context of certain cellular phenotypes or diseases. Blue: Examples of external resources or ontologies that could interact with and/or be incorporated into this catalog. We note that the listed set of edges represent current plans and are not exhaustive with respect to topics relevant to interpreting genomic variation. Abbreviations and citations: dbSNP[150], gnomAD[4], ENCODE[15], GTEx[17], chromatin accessibility (ca)QTLs, saturation genome editing (SGE)[23], Variant Abundance by Massively Parallel sequencing (VAMP-seq)[22], MaveDB[24], HuBMAP[20], GENCODE[151], UniProt[152], Gene Ontology (GO)[153], protein-protein interactions (PPI), IntAct Molecular Interaction Database[154], NHGRI Molecular Phenotypes of Null Alleles in Cells (MorPhiC)

Consortium, Mondo Disease Ontology[155], Human Phenotype Ontology (HPO)[156], rare variant association studies (RVAS), Online Mendelian Inheritance in Man (OMIM)[157].