The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-122220-112147

Annual Review of Biomedical Data Science

Integration of Protein Structure and Population-Scale DNA Sequence Data for Disease Gene Discovery and Variant Interpretation

Bian Li,[1] Bowen Jin,[2] John A. Capra,[3] and William S. Bush[4]

[1]Department of Biological Sciences and Center for Structural Biology, Vanderbilt University, Nashville, Tennessee, USA

[2]Graduate Program in Systems Biology and Bioinformatics, Department of Nutrition, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

[3]Bakar Computational Health Sciences Institute and Department of Epidemiology and Biostatistics, University of California, San Francisco, California, USA; email: tony@capralab.org

[4]Cleveland Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA; email: wsb36@case.edu

Keywords

protein 3D structure, population genetics, variant interpretation, disease gene discovery, data integration

Abstract

The experimental and computational techniques for capturing information about protein structures and genetic variation within the human genome have advanced dramatically in the past 20 years, generating extensive new data resources. In this review, we discuss these advances, along with new approaches for determining the impact a genetic variant has on protein function. We focus on the potential of new methods that integrate human genetic variation into protein structures to discover relationships to disease, including the discovery of mutational hotspots in cancer-related proteins, the localization of protein-altering variants within protein regions for common complex diseases, and the assessment of variants of unknown significance for Mendelian traits. We expect that approaches that integrate

these data sources will play increasingly important roles in disease gene discovery and variant interpretation.

INTRODUCTION

In November 1949, Linus Pauling and coworkers reported a difference in electrophoretic mobility between normal and sickle cell hemoglobins (1). Eight years later, in 1957, Vernon Ingram pinpointed the cause of the electrophoretic difference through chromatography: the replacement of a glutamate residue with a valine residue at position six in the beta chain of hemoglobin (2) (**Figure 1*a***). Pauling and Ingram's discoveries were groundbreaking in at least two aspects. First, Pauling's discovery showed that the mechanism underlying a genetic disease could be traced to an alteration in a biophysical property of a protein, and second, Ingram's result demonstrated that the alteration was due to a change in the amino acid of the corresponding peptide chain (3). When

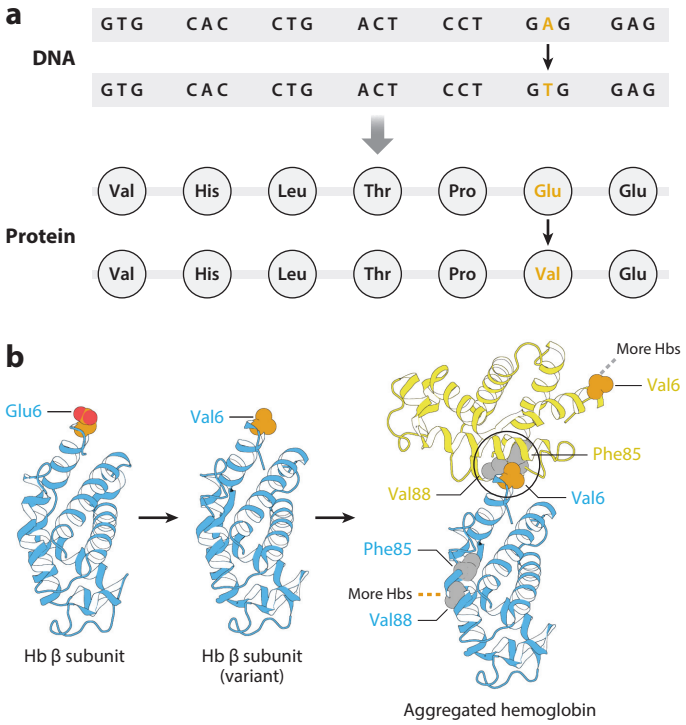

Figure 1

Discovery of amino acid substitutions in the *HBB* gene that induce the sickle cell trait through the abnormal polymerization of hemoglobins (Hbs). (*a*) Variation visualized at the sequence level: A single-nucleotide substitution (A to T) in the *HBB* gene at codon 6 results in a replacement of a glutamate (Glu) residue with a valine (Val) residue. (*b*) 3D structure of Hb gives mechanistic insights into the pathogenic nature of the amino acid substitution: The exposed Val residue in the variant form [Protein Data Bank (PDB) ID: 2HBS; PDB ID for the unsubstituted form: 1A3N; one subunit is shown in both structures] becomes a hinge point for Hb polymerization under conditions of hypoxia. The Val residue in the variant form makes hydrophobic contacts with Ala70, Phe85, and Val88 from an adjacent Hb molecule. For aggregated Hbs, one subunit from each of the two neighboring Hbs is shown (PDB ID: 2HBS). In all structures, the heme prosthetic group is hidden for clarity.

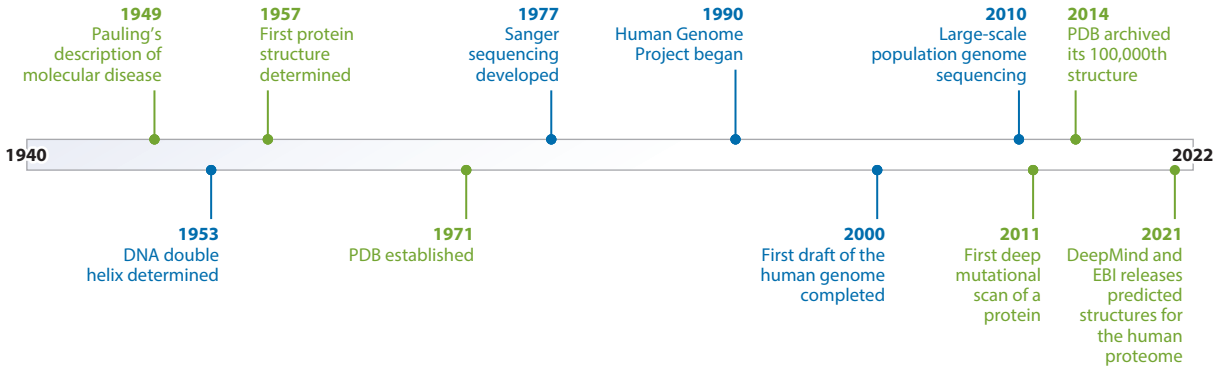

Figure 2

Timeline of key events in the coevolution of the fields of genomics (*blue*) and protein structure and function (*green*) from 1940 to 2022. Abbreviations: EBI, European Bioinformatics Institute; PDB, Protein Data Bank.

viewed in the context of the tertiary structure of hemoglobin obtained via X-ray crystallography by Max Perutz (4), the cause of sickle cell anemia is elegantly explained at the molecular level: Hydrophobic contacts form between the valine residues of one hemoglobin molecule and the alanine, phenylalanine, and leucine residues of another adjacent hemoglobin molecule, causing the fibers to stretch and deform the red blood cell (5) (**Figure 1*b***). This convergence of genetics, biochemistry, and structural biology is a foundational example of molecular medicine.

Many other inherited disorders have been explored in a similar fashion by first identifying genomic regions and genetic variants cosegregating with the disorder in families and then providing a molecular characterization of the influential variants. Huntington's disease was first investigated through genetic linkage analysis, which identified a region on chromosome 4 (6) that nearly ten years later was localized to a trinucleotide expansion (CAG) in the *HTT* gene (7). Molecular genetics work later revealed the protein encoded by *HTT* to be an activating transcription factor for the *BDNF* gene whose activity is lost with pathogenic levels of the CAG expansion (8). Identification of the *CFTR* gene for cystic fibrosis followed a similar path (reviewed in 9). The discovery of these influential proteins for disease spawned extensive efforts to understand their function through a variety of techniques that ultimately culminated in the determination of high-resolution 3D structures (10, 11). Key events in the coevolution of the science of protein structure/function and genetics/genomics are outlined in **Figure 2**.

While this general process of integrating genetic and structural understanding has proven fruitful for several genetic disorders, it is a slow process, with genetic analyses and discovery often siloed away from molecular and biochemical characterization of proteins. Explaining the molecular consequences of genetic variants and how they influence disease risk remains a major goal in the field of human genetics. In this review, we discuss how recent advances in both genomic sequencing technology and protein structure determination are poised to disrupt this traditional scientific workflow by considering genetic variation in the context of the 3D structure of the protein to advance genomic discovery and our understanding of molecular function simultaneously.

Protein Data Bank (PDB): a database of information about the 3D shapes of proteins and other macromolecules, stored in a standardized format

PROTEIN STRUCTURE AND GENETIC VARIATION DATABASES ARE MASSIVE, BUT LARGELY INDEPENDENT

Recent Advances in Protein Structure Determination

The Protein Data Bank (PDB) just celebrated its 50th anniversary and now grows at a rate of approximately 10% per year with over 10,000 new protein structures released annually since 2015

(12, 13). While the bulk of these are derived from traditional structural biology techniques like X-ray crystallography and nuclear magnetic resonance, a growing percentage of these structures are derived from cryo-electron microscopy (cryo-EM) due to major technological advances in that field (14). While cryo-EM structures are generally not as high resolution as structures from X-ray crystallography, advances in cryo-EM methods can resolve the structure of membrane proteins and intrinsically disordered proteins, both of which are extremely difficult to obtain with classic approaches (15, 16). For example, while the HTT and CFTR proteins have been studied extensively for decades, only partial structure models were available. The full structures for CFTR (17) and HTT (18) were only recently resolved from cryo-EM in 2018.

The PDB, with its large number of experimentally determined protein structures, also provides templates for computationally deriving structures of other proteins (19). When a protein of interest is sequence-similar to a protein with an existing high-resolution structure, homology modeling can be applied to produce a structure of the similar protein (20). Large efforts like SWISS-MODEL (21) and ModBase (22) have systematically applied this approach for large numbers of human proteins to generate computational predictions.

In contrast to homology modeling, de novo computational prediction attempts to estimate a protein structure from its amino acid sequence alone and has long been an extremely challenging problem in biology (23, 24). While these approaches have had very limited success, recent deep learning approaches have shown dramatic improvements in structure prediction. Structure prediction algorithms are evaluated through the biennial Critical Assessment of protein Structure Prediction (CASP) experiment (25), which compares method performance on soon-to-be released protein structures held back by the PDB. Deep learning approaches have made significant advances in the CASP experiment (26, 27), and in 2020 the AlphaFold2 approach produced accurate predictions of unsolved protein structures that are essentially indistinguishable from experimental structures for many proteins (28, 29). Deep learning approaches from the Rosetta community have also recently achieved substantial accuracy at predicting structures (30), protein interfaces (31), and protein design (32). While computational predictions still have some limitations, these approaches dramatically extend our catalog of structures to cover nearly all human proteins.

Critical Assessment of protein Structure Prediction (CASP): a biennial experiment within the structure prediction community involving a competition to predict an unpublished set of experimentally derived protein structures

Genome Aggregation Database (gnomAD): a project to aggregate and harmonize exome and genome sequencing data from large-scale sequencing projects to make summary data available for the wider scientific community

dbSNP: a public domain archive of genetic variation managed by the US National Center for Biotechnology Information

Recent Advances in Population-Scale Determination of Genetic Variation

Since Pauling and Ingram's work, techniques for sequencing DNA have undergone several revolutions. Building on the advance provided by Sanger sequencing, the Human Genome Project was completed in the early 2000s (33). The resulting reference genome enabled the development of second-generation short-read sequencing technologies with dramatically lower cost (34). The rapid and continuing decrease in the cost of exome and genome sequencing has made the identification of genetic variants present in an individual commonplace, resulting in many high-resolution catalogs of genetic variation that have empowered large-scale genetic association studies (35–37).

Recent large human genetic variation datasets generated by high-throughput DNA sequencing have identified millions of genetic variants, the vast majority of which are exceptionally rare. For example, in the Alzheimer's Disease Sequencing Project whole-exome sequencing of ~10,000 individuals, 97% of identified variants were found to have a minor allele frequency of less than 1%, and 23% of variants were observed in only a single sample (singleton variants) (38). The Genome Aggregation Database (gnomAD) project aggregated tens of thousands of exomes and genomes from multiple disease phenotype studies and found that 99% of variants have a frequency of less than 1% (39). The most recent release (v3.1) called over 200 million genetic variants (40). The National Center for Biotechnology Information's database of genetic variants, dbSNP (41), has long served as a clearinghouse for genetic variant information, and its latest build (build 155)

reports over 1 billion distinct reference single-nucleotide polymorphism clusters thanks to the recent incorporation of sequencing studies. While the bulk of these variants are located in noncoding regions of the human genome, nearly 15 million variants lie within exonic regions, and the 1000 Genomes Project estimates that each individual carries 10,000 to 12,000 protein-altering variants (42). These rare variants' potential effects on diseases of interest remain largely unknown because we lack the statistical power to detect associations between single low-frequency variants and phenotypes. However, unlike the glutamate-to-valine substitution in hemoglobin, most of these identified genetic variants likely have very small effects, only modestly impacting disease risk over the life course (43).

While nearly all of sequence data contributing to these database expansions are based on short-read sequencing technologies, new long-read sequencing methods are now being employed to fill in the gaps from telomere to telomere of human chromosomes (44). The Telomere-2-Telomere (T2T) Consortium recently published updated assemblies of chromosome X (45) and chromosome 8 (46) and is building the most comprehensive map of the genome to date (47). Sequencing studies built from these newer maps will expand our catalogs of genetic variation even further to include previously inaccessible regions of the genome and complex structural variants.

Telomere-to-Telomere (T2T) Consortium: an open, community-based effort to generate the most complete assembly of the human genome to date using long-read sequencing technologies

Genome-wide association study (GWAS): a population-based analysis of large-scale genetic data used to identify commonly occurring genetic variants that associate with disease

CURRENT METHODS FOR DISEASE GENE DISCOVERY AND VARIANT INTERPRETATION DO NOT FULLY LEVERAGE PROTEIN STRUCTURE

Genetic Methods for Disease Gene Discovery

Genetic linkage analysis was used to discover the *HTT* and *CFTR* genes, and it has been used extensively to study hundreds of inherited Mendelian diseases (48). Through the estimation of shared chromosomal segments within families, linkage analysis identifies the coinheritance of genomic regions with disease. Additional fine-mapping studies and molecular characterization of genomic regions are required to further identify causal variants within genes of the region. Linkage analysis has also had some success in identifying genes for familial forms of genetically complex diseases, such as the discovery of *BRCA1/2* genes for breast cancer (49, 50) and the *APOE* gene for Alzheimer's disease (AD) (51). Despite these successes, linkage analysis is dramatically underpowered for gene discovery within families with more sporadic occurrence of other complex diseases. The development of microarray technologies in the late 1990s and early 2000s enabled large-scale genotyping, and since the mid-2000s the genome-wide association study (GWAS) has been the primary genetic study design (35). Most GWAS are based on large-population-based samples and attempt to associate specific base pair changes that are relatively frequent in a population with disease risk. This approach has had dramatic success at identifying genetic associations for many traits where linkage analysis failed. The vast majority of these variants, however, only modestly alter disease risk through changes to gene regulation (52). Due to the need for exceptionally large sample sizes and the confounding effect of genetic ancestry, GWAS have focused largely on individuals of European descent, which reduces the generalization of genetic effects and the identification of disease variants distinct to other human populations (53). For studies of rare variants, numerous efforts (including dedicated funding support) are ongoing to enhance the diversity of genomic sequencing studies going forward (54).

Because of their low frequency, rare variants are especially difficult to associate with traits due to very limited statistical power. Unit-based approaches have been developed to alleviate this lack of power by testing the collective effect of a group of rare variants. These approaches share the common strategy of grouping variants by a (presumably) functional unit (e.g., a gene or transcript) and then estimating an effect based on a function of those variants. The most basic

Sequence kernel association test (SKAT): a statistical approach that combines the effects of multiple rare variants within an analysis unit to improve power for rare variant association studies

of these is a collapsing test, which assigns individuals a binary value based on the presence or absence of a variant within the unit (55). If some individuals carry multiple variants within the same unit, collapsing tests discard this potentially useful information. To address this limitation, a burden test creates a discrete variable based on the number of variants each individual carries within the unit and then estimates a per-variant effect on the trait under study. Key assumptions of the burden test are that all variants within the unit affect risk in the same direction (typically they are assumed to be deleterious) and that all variants have approximately equivalent effects on the phenotype. The sequence kernel association test (SKAT) eliminates these assumptions using a statistical framework that estimates the variance in the phenotype explained by rare variants in the unit through kernel regression (56). This test estimates a measure of genetic sharing among all pairs of individuals within a dataset and estimates the relationship between this genetic sharing and phenotype similarity among all pairs of individuals. Using SKAT, variants can have different and opposing directions of effect within the same unit and contribute to the test statistic. The typical implementation of SKAT, however, does not estimate effects for singleton variants (as they are not shared among two or more individuals to allow for estimation of their impact on trait variance). Several extensions to the SKAT framework have been developed to further improve statistical power by accounting for singleton effects (SKAT-O) (57), better estimating the null distribution (58), and computing exact *p*-values using likelihood ratio tests (59), among many others (60–63).

Despite these statistical developments, power for rare variant association tests remains somewhat limited, especially when neutral or benign variants are included in the test (58). As with most statistical tests, a straightforward way to improve power is to increase the sample size; however, for sequencing studies, adding more samples will also result in the identification of additional rare variants and singletons. Adding additional variants to unit-based tests may actually decrease power if they are nonfunctional or do not otherwise influence the trait (58). Therefore, adding more samples may increase power for some variants, but it also expands the problem by identifying additional variants that may reduce the overall power of unit tests.

Multiple approaches have been developed for prioritizing variants as benign, or likely to alter biological functions and potentially pathogenic, and it is common practice to include variant annotations as either a criterion for including variants in a unit test or as a metric as part of the test statistic itself. Most studies have used coarse annotations for this filtering (i.e., missense only, putative loss of function, etc.) (64, 65) or have combined scoring approaches from multiple annotation methods [e.g., VAAST (Variant Annotation, Analysis, and Search Tool), STAAR (set test for association using annotation information)] (66, 67). While some of these approaches include computational predictions of variant effects that may leverage protein information as part of their predictions [such as PolyPhen (68)], these represent a very limited use of information about the functional context of the variants—protein structure—in tests of rare variant association.

High-Throughput Experiments for Missense Variant Interpretation

According to the American College of Medical Genetics and the Association for Molecular Pathology guidelines, functional data are the strongest evidence for classifying a variant as benign or pathogenic (69). Traditionally, the functional effect of variants has often been studied via site-directed mutagenesis followed by functional characterization in a model system. While this one-at-a-time approach usually generates accurate experimental data supporting variant classification and insights into the mechanism underlying the functional effect, it probes only a tiny fraction of the possible genetic variation in a gene. This severely limits our understanding of how the global landscape of genetic variation influences function and results in phenotypic consequences (70).

Recent developments in multiplex assays of variant effect (MAVEs) (71, 72) allow for the examination of massive numbers of variants in a single experiment. For example, cellular growth-based deep mutational scanning (73, 74) has enabled the systematic characterization of all possible, or a large fraction of, missense variants of several disease genes (71). More generally, MAVEs use cell-based assays where a protein is expressed from a plasmid or virus that can be mutated to form screening libraries. These libraries are then coupled with a selection process to identify the variants enriched and depleted in functional cells. They have the distinct advantage of testing all variants simultaneously under the same experimental conditions so that measurements for different variants are directly comparable to each other. Due to their high-throughput nature, MAVEs are well suited for efficient functional characterization of large-scale newly discovered variants (71), or for reclassifying variants of unknown significance (VUS) that may emerge from clinical sequencing (75), and may help build a foundation for functional interpretation of other variants that might be identified in the future (76). While MAVEs have been implemented for certain protein functions like antibiotic resistance (77, 78) or protein-binding affinity (79), given the diversity of protein functions, the design and validation of a well-performing functional assay for protein-altering variants are a major challenge for broad implementation of these assays (80).

Multiplexed assays of variant effect (MAVEs): large-scale experimental approaches that couple functional assays with high-throughput sequencing to assess the functional consequences of thousands of genetic variants simultaneously

Variant of unknown significance (VUS): a genetic variant identified within a disease-associated gene that may or may not have implications for disease risk

The coarseness of the functional assays implemented in most MAVEs does not provide detailed information on the mechanism by which each variant exerts an effect. A recently developed assay called variant abundance by massively parallel sequencing (VAMP-seq) measures the effect of variants on the steady-state cellular abundance of a protein, which is ultimately dictated by the stability of variant proteins (81). VAMP-seq is mechanistically appealing because it can separate variants with modest effects on stability from those that are substantially destabilizing (70). Compared to other approaches, VAMP-seq is also more cost effective because it can be generalized to many proteins and provides more information about the thermodynamic effects of individual variants on the protein (81). At the molecular level, a variant may induce a wide range of changes beyond compromising stability and structure, including changes to conformational dynamics and reaction kinetics (82), disrupted macromolecular interaction (82, 83), and ablation of posttranslational modification sites (83), among others (84). Extensions of the VAMP-seq approach will hopefully allow for high-throughput characterization of these other aspects of protein function as well.

While these approaches represent exciting new developments in high-throughput assays of variant effect, they are currently cost prohibitive for conducting on a broad scale. Thus, a comprehensive, experimentally constructed resource for prioritizing variant function for most proteins is unlikely to be available in the foreseeable future.

Computational Variant Effect Prediction Methods

Computational variant effect prediction is a valuable approach that can provide important evidence to support variant classification (69), in particular, evidence based on biological first principles (85). Numerous other reviews have covered computational variant effect prediction in detail from various perspectives (86–90). Despite this immense body of literature, several substantial limitations in these approaches are evident.

First, for many methods the precise nature of what is being predicted is unclear (i.e., pathogenicity, loss/gain of function, damaging effect, or deleteriousness). This issue has been generally underappreciated (91), although recent work points out that computational tools that claim to predict variant effect often fail to define "effect" (90). Key examples in the literature are methods trained to discriminate disease variants from benign variants across orthologous proteins; these approaches may interpret their effect as functional even though no explicit functional evaluation has been made (91).

Second, many methods also operate as a black box, providing a prediction with no underlying model, rationale, or justification for making it. With this limitation, reporting a potential molecular effect for a functional variant will be extremely difficult (92). There are notable exceptions to this trend (93, 94), and methods for deriving meaning from complex deep learning models have been developed (95) and may be applied to better understand the biological features used to derive predictions.

Third, variant effect prediction methods are still not very accurate (especially for use in a clinical setting) and often generate predictions that agree poorly with one another (96, 97). Variant classification methods are ultimately trained using en masse datasets of known (or presumed) pathogenic variants across thousands of individuals (68, 98, 99). While the refinement of these resources is an ongoing effort in the genetics community, they are likely subject to severe ascertainment and reporting biases, inconsistencies in data quality and assessment, and reporting errors (100). Classification or scoring approaches also assume a uniform threshold of deleteriousness across the whole genome, which is a poor assumption given decades of research on variable penetrance (101). Presumably these issues can largely be resolved by training models on a gene- (102), gene-family- (92), or disease-specific (103, 104) basis using homogeneously derived variant annotation data, such as a KCNQ1-specific model (102), or the epilepsy gene-specific model of Traynelis et al. (103).

Predictions of variant impact may prove useful for rare variant association tests for identifying new disease genes, but the impact of this area extends beyond genomic discovery. Once disease genes are known, these data have the potential to transform the practice of medicine by enabling the delivery of precision care based on patients' genome sequences (84–86, 105–107). However, a major roadblock to achieving this goal is our inability to reliably interpret the molecular, phenotypic, and therapeutic consequences of the millions of genetic variants that are being revealed by sequencing patients' genomes (71, 105, 107, 108). The Critical Assessment of Genome Interpretation (CAGI) is an analog to CASP that provides the scientific community with a framework for assessing new methods for variant impact (109), but none of the tools developed for the interpretation of genetic variants over the past two decades is sufficiently accurate or informative to be used to interpret novel variants in a general clinical setting (96, 110, 111).

THE POTENTIAL FOR INTEGRATING GENETIC VARIANTS AND 3D STRUCTURE TO IMPROVE GENOMIC VARIANT ANALYSIS

Integrating the large catalogs of human genetic variants with 3D protein structures represents a promising new approach to variant prioritization for both rare and common diseases. In the following subsections we first describe several preliminary analyses demonstrating that the patterns of germline and somatic variants in protein 3D space provide valuable functional context. We then highlight initial efforts to integrate structural context into algorithms for (*a*) common disease gene discovery and (*b*) variant interpretation in rare disease. An overview of relevant methods is provided in **Table 1**.

Analysis of Variants in 3D Identifies Disease Hotspots

When a new protein structure is described, a typical qualitative analysis of the protein involves identification of protein domains and their molecular function, but many studies also examine the location of disease-associated mutations in their new 3D context, such as the mapping of disease-associated mutations in the structure of mitochondrial complex II (90). Beyond these qualitative discussions of genetic variation in the context of proteins, the earliest quantitative analyses of variants within proteins focused mainly on the identification of mutational hotspots via the analysis of

Table 1 Overview of relevant methods in genetic variant analysis

Category	Reference(s)	Method	Features	Key limitations
Set-based variant association tests	55	Collapsing/ burden test	Estimates average effect of multiple rare variants on the trait	Rare variants are assumed to have the same direction and magnitude of effect on the trait
	56	SKAT	Extends collapsing and burden tests to allow variants to have different and opposing directions of effect on the trait	Does not estimate effects for singleton variants
	57, 58	SKAT-O	Improves statistical power of SKAT by dynamically combining SKAT and burden tests to account for singleton effects	Singleton variants are assumed to have a consistent direction of effect on the trait
Variant prioritization	66, 67	VAAST/STAAR	Dynamically prioritizes rare variants by their biological function or pathogenicity	Dynamic weighting of various annotations makes it difficult to interpret the precise biological impact of the variants influencing the trait
	68	PolyPhen2	Uses protein secondary-structure information to aid variant consequence prediction	Limited incorporation of protein structure information in the overall variant scoring approach
	124	Pathogenic proximity	Uses 3D protein structure and a set of known pathogenic variants to prioritize variants based on spatial location within the protein	Requires more than three known pathogenic variant examples, which may not exist for all proteins
	125	COSMIS	Uses evolutionary constraints within protein structures to predict variant impact	Does not account for severity of amino acid substitutions
3D structure–based variant distribution analyses	115	HotMAPS	Estimates somatic mutation density metric for each residue of a 3D protein structure	Assumes a uniform null distribution of variants within the protein, which may not hold for all protein structures
	113	CLUMPS	Identifies mutation clusters in 3D protein structures using cluster analysis methods	Sensitive to user-specified distance thresholds, which can alter cluster results
	114	mutation3D	Generates hierarchical clustering of mutations within 3D protein structures	Sensitive to user-specified distance thresholds, which can alter cluster results
	116	HotSpot	Identifies somatic mutation clusters within 3D protein structures using graph theory metrics	Sensitive to user-specified distance thresholds, which can alter cluster results
	119	Ripley's K	Detects both clustering and dispersion of variants using a spatial statistic over multiple distance thresholds	Clustering or dispersal patterns may not be obvious due to dynamic distance thresholds used in the test

(*Continued*)

Table 1 (*Continued*)

Category	Reference(s)	Method	Features	Key limitations
3D structure–based association tests	121	POINT	Aids single-variant tests by incorporating the influence of neighboring variants relative to the target variant	Computationally intensive and sensitive to allele frequencies of analyzed variants
	122	PSCAN	Hierarchically groups rare variants into windows based on their Euclidean distances, and tests each window against the phenotype to identify signal regions of the protein	Relies on underlying SKAT and collapsing test statistics
	123	POKEMON	Uses a kernel test to directly compare the spatial distribution of rare missense variants among cases and controls	Unable to pinpoint specific risk variants within the protein that are driving the effect

Abbreviations: SKAT, sequence kernel association test; SKAT-O, optimized SKAT; STAAR, set test for association using annotation and information; VAAST, Variant Annotation, Analysis, and Search Tool.

somatic cancer variants (112–118). Recent work has expanded these analyses to germline variants (119). The approaches for performing these analyses are closely related to spatial cluster analysis, with a general approach that first maps missense variants into the 3D space of a protein structure, and then uses Euclidean distances among all variants to compute a clustering metric. The statistical significance of the metric is evaluated by generating an empirical null distribution via resampling a random set of variants across the protein (120).

Several variations of these analyses on somatic variants in cancer have been published. Tokheim et al. (115) developed the HotMAPS algorithm and applied it to somatic mutations from The Cancer Genome Atlas. HotMAPS estimates a local mutation density metric for each residue of a protein by summing the number of missense mutations at a given residue and within 10 Å. From this analysis, they identified 216 tumor-type-specific hotspot regions in 54 genes and found differences in the mutational density of oncogenes (more focal) versus tumor-suppressor genes (more heterogeneous). In contrast to the density-based metric of HotMAPS, Kamburov et al. (113) developed the CLUMPS approach based on the pairwise Euclidean distances among all variants and their frequencies. Applied to the PanCancer compendium, CLUMPS identified 50 proteins with mutations clustering at protein–protein interaction interfaces. Stehr et al. (112) also considered Euclidean distances among somatic mutations from eight cancer types, but they first divided the protein into structurally defined domains for clustering analysis. Their spatial analysis results were used to identify 24 oncogenes and tumor-suppressor genes. Meyer et al. (114) developed a method called mutation3D, which employs a variant of a complete-linkage (or sometimes called furthest-neighbor) hierarchical clustering using a user-specified maximum cluster diameter. This approach is the most liberal in its definition of somatic variant clusters, but it identified many genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) that are significantly enriched for known cancer-driver genes. Niu et al. (116) also developed a clustering method called HotSpot on The Cancer Genome Atlas tumor sequencing data that iteratively defines variant clusters based on Euclidean distance. Their approach identified 38 clusters, 35 of which occurred within known cancer genes, and recapitulated many known patterns of somatic mutations within cancer proteins.

We developed a framework for testing hypotheses about the 3D spatial distribution of variants that was motivated by Ripley's K, a spatial statistic used in ecology and epidemiology (119). Ripley's

K is computed by identifying mutated residue pairs with a Euclidean distance less than a specified threshold *t*. Random samples of residues (of equivalent size to the variant set) are selected from the protein to generate an empirical distribution for K at a given threshold, and multiple threshold values are evaluated to generate a distribution of observed K values relative to permuted values. This enables the computation of *p*-values for observed deviations from a random spatial distribution. We applied the Ripley's K approach to COSMIC recurrent somatic mutations, known Mendelian disease variants from ClinVar, and likely benign variants from large sequencing projects. We identified 25 proteins with significant somatic variant clustering. Applying Ripley's K to germline missense variants for Mendelian diseases revealed that disease variants tend to cluster in 3D space (119).

While each of these approaches used somewhat distinct methods for identifying variant clustering, all methods identified significant clustering of missense somatic mutations within *BRAF*, *FBXW7*, *EGFR*, and *PIK3CA*. *BRAF*, *EGFR*, and *PIK3CA* are well-established proto-oncogenes that are commonly mutated across multiple cancer types, and *FBXW7* is a critical tumor suppressor that is also frequently mutated in multiple cancers. The consistency of these findings across multiple somatic mutation datasets and with multiple methods demonstrates the robustness of the somatic clustering within these proteins.

Integration of 3D Structure and Population-Scale Genetic Variation Improves Gene Discovery

The consistent identification of spatial clustering of cancer-driving variants and rare missense germline disease variants suggests that the position of missense variants within the 3D structure of a protein could be used to improve disease gene discovery and variant interpretation. Several methods have recently been developed using protein structure to increase the statistical power of unit-based rare variant tests, such as POINT (121), PSCAN (122), and POKEMON (123). **Figure 3** illustrates these three approaches applied to a simulated effect within the *CSF1R* protein and the information contributed by the spatial distribution of variants.

POINT is a single-variant association test that improves power by borrowing information from spatially neighboring rare variants (121). For a target variant *m*, the Euclidean distance to all other variants is calculated and weighted by a parameter *c*, which selects and scales the influence of neighboring variants relative to the target variant (*c* is adaptively chosen based on the dataset). The matrix derived from the Euclidean distances is then used within a SKAT or burden framework to determine the association of a variant with the phenotype of interest (in the context of the local region of the protein structure). POINT was applied to data from the ACCORD (Action to Control Cardiovascular Risk in Diabetes) clinical trial and demonstrated associations of *PCSK9* variants with low-density lipoprotein cholesterol levels and *ANGPTL4* and *CETP* variants with high-density lipoprotein (HDL) cholesterol levels, respectively.

In contrast, PSCAN is a unit test evaluating the effects of all rare variants to generate a single test statistic. PSCAN scans the entire set of variants, hierarchically collapses the variants into windows based on their Euclidean distances, and tests each window against the phenotype using either a SKAT or a burden test framework (122). PSCAN returns collections of signal regions, which are spatially colocated variants that contribute significantly to the trait of interest. PSCAN was applied to National Heart, Lung, and Blood Institute Exome Sequencing Project data to examine associations to HDL and triglyceride traits, and to the Alzheimer's Disease Sequencing Project data for associations to AD. PSCAN identified several protein regions for AD risk, notably including the amyloid binding regions of SORL1.

While the POINT and PSCAN approaches explicitly integrate protein structure information into association tests, both tests are constructed from frequency-based tests and are subject to

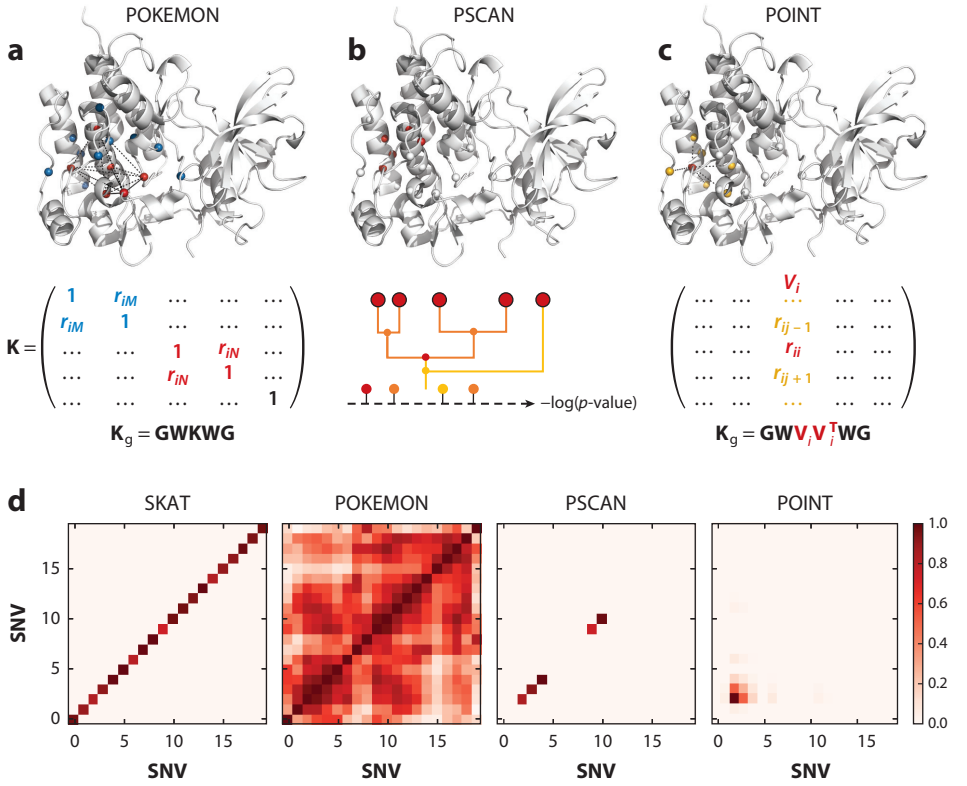

Figure 3

Comparison of POKEMON, PSCAN, and POINT tests, which integrate 3D protein structure for disease gene discovery. Here, a comparison is made among simulated variants from a case/control study on the structure of the protein CSF1R (PDB: 2OGV). We simulated a clustering pattern by distributing influential variants within the core of the protein structure and scaling the variant effects (odds ratios) proportionally to their distance from the core. We then randomly sampled 20 variants from the protein. The minor allele frequencies for all the variants were randomly sampled from a log-transformed uniform distribution within an interval (−4, −2.3), which restricts to variants with minor allele frequencies within the range (0.0001, 0.005) or singletons. The number of case and control subjects was set as 5,000. (*a*) The cluster of variants in CSF1R carried mainly by case subjects identified by POKEMON is indicated by red spheres connected with black dashed lines. The blue spheres represent neutral variants that do not associate with the simulated phenotype. In the matrix **K**, r_{iM} and r_{iN} are blocks within the kernel, representing the genetic similarity contributed from neutral and risk variants, respectively. The matrix **K** will be directly used for calculating the genetic similarly kernel $\mathbf{K}_g$, shown in the equation below, with genotype matrix **G** and a Blossum62 weight matrix **W**. (*b*) The signal region of CSF1R identified by PSCAN is indicated by red spheres. The diagram below shows the hierarchical clustering from PSCAN. The variants are clustered hierarchically according to their 3D position on the protein. Every node in the cluster tree is tested against the phenotype and the resulting *p*-values are combined using the Cauchy method to generate a single test statistic for the signal region. (*c*) The variant identified as the risk variant by POINT is highlighted in red. Yellow spheres indicate the neighboring variants contributing to the test of risk variant. The matrix below contains the signal variant–focused vector, with r_{ii} representing the scale weights from the risk variant, while r_{ij+1} and r_{ij-1} are scale weights for the contribution from neighboring variants to the risk variant. The scale weights are used to construct a diagonal matrix, which is then used for calculating the genetic similarly kernel $\mathbf{K}_g$. (*d*) The pairwise SNV-based **K** matrix for SKAT, POKEMON, PSCAN (SKAT framework), and POINT (SKAT framework). For SKAT and POKEMON, the weight matrix is used in the association test. For PSCAN and POINT, the weight matrix is for the selected signal variant or the signal region. These diagrams demonstrate how neighboring variants contribute to the test statistic. Abbreviations: PDB, Protein Data Bank ID; SKAT, sequence kernel association test; SNV, single-nucleotide variant.

the limitations of SKAT and burden test approaches. In contrast, the POKEMON approach is a kernel-based test solely based on the spatial distribution of rare missense variants within case and control subjects, with less emphasis on the variant frequency (123). POKEMON is based on a kernel matrix of the genetic similarity between two individuals based on the variants they carry. Pairs of variants are weighted by their distance within a specific protein structure, and spatially close variants share high weights, while spatially distant variants share low weights. POKEMON also accounts for the magnitude of amino acid substitutions via the BLOSUM62 matrix. When used in a case/control analysis, a significant result from POKEMON indicates that case subjects share more spatially clustered or dispersed rare variants than the controls (or vice versa). While this test was explicitly designed for extremely low-frequency variants (singletons and doubletons), the kernel can be combined with the traditional beta-scaled variant frequency kernel employed in SKAT tests. Similar to PSCAN, the POKEMON structure kernel identified a spatial pattern of rare variants within *SORL1*, as well as *TREM2*, a known AD gene association driven by rare variants, among others.

Missense Variant Interpretation Using Protein Structure Information and Population-Based Genetic Variation

The differences in the spatial distribution of Mendelian disease variants and benign missense variants from population-scale studies suggest that the integration of protein structural information has great potential for the identification of variants with presumably large effect sizes (e.g., in rare Mendelian diseases) (119). Illustrating this approach, multiple new VUS for familial interstitial pneumonia were mapped into a homology model of the RTEL1 protein (124). Comparing spatial proximity to known pathogenic variants relative to putatively benign variants from the 1000 Genomes Project successfully classified five of six disease-segregating VUS as pathogenic. This basic approach built only on spatial proximity achieved an accuracy competitive with two other commonly used variant classification tools (PolyPhen2 and evolutionary sequence conservation defined by ConSurf).

Expanding on this idea, we developed a general method for mapping the tolerance of spatial neighborhoods in protein structures to genetic variation (125). Using a sequence-context-aware mutational model and genetic variants observed in more than 100,000 individuals from gnomAD, we computed an amino acid site–specific score termed COSMIS. This score compares the observed number of missense variants in the connected 3D neighborhood of the site of interest to the expected number of variants. Sites with significantly fewer variants than expected in their connected 3D neighborhood highlight regions of proteins under substantial functional constraint in recent human evolution that are thus more likely to lead to dysfunction when disrupted. Using this approach and leveraging protein 3D structure resources, like the AlphaFold protein structure database (126), we characterized the 3D spatial constraint on amino acid sites in the human proteome at near proteome scale. We find that the COSMIS score can distinguish pathogenic from benign variants in many settings and identify highly constrained proteins and amino acid sites that have not previously been associated with disease phenotypes.

FUTURE ROADMAP

As sequencing studies continue to advance our understanding of genetic variation within the genomes of diverse individuals, and experimental and computational approaches expand the catalog of protein 3D structures, we have an opportunity to gain a new and deeper understanding of protein-level variation and its impact on protein function and disease. Characterizing missense variant effects in the context of protein 3D structures helps identify variants that may cause the

disruption of native intra- or intermolecular interactions or residue packing in the protein core (127–130). For example, one fundamental, structural level effect of a missense mutation is that it may decrease the thermodynamic stability or cause misfolding of the protein, leading to insufficient cellular abundance of the protein to perform its function. Destabilization or misfolding is often caused by disruption of native residue packing in the protein core and is a common origin of inherited diseases (70, 127, 128). Therefore, computational predictions of changes in protein stability are likely useful for variant interpretation (70, 128). Recent studies also illustrate that thermodynamic stability may be predictable from the biophysical properties of missense substitutions via deep learning approaches (131), and that it could be incorporated into pathogenicity predictions.

Quantifying the locations of tolerated missense variants within protein structures provides a map of recent functional constraint that can inform functional assessments of proteins (125, 132–134). The current catalog of human population genetic variation is expected to expand further. At the same time, advanced machine learning techniques are being applied to solve structural biology problems at a rapid pace (31, 135, 136). These technologies are expected to tremendously expand our accessibility of the space of protein 3D structures in the near future (137). Thus, merging genetic variation and protein structure data should be a focus for the community. We anticipate

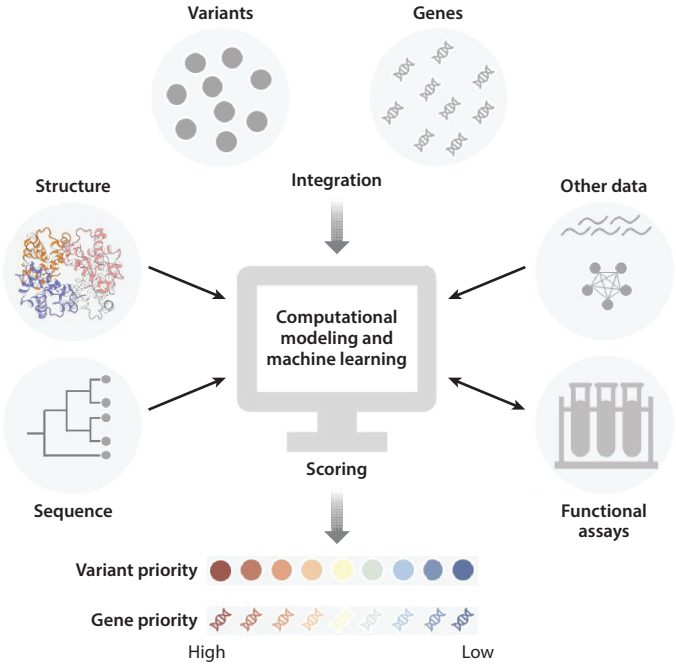

Figure 4

Integration of multiple sources of data for the prioritization of disease variants and genes. Untangling the relationship between genotype and phenotype is a complex task and a grand challenge in human genetics. We propose that integrating multiple sources of data into a heuristic computational framework that uses machine learning has the potential to generate desirable solutions to this challenge. In this review, we focus our discussion on sequence variation and 3D protein structure data, as well as saturated mutagenesis data generated by high-throughput functional assays. However, data sources need not be limited to these; for example, gene expression and protein–protein interaction networks could also be integrated. Functional assays play a special role because they can often be synergistically coupled with computation and modeling to generate insights into variant/gene effects that could not be obtained by either approach alone.

that new approaches to integrating spatial quantifications of constraint with other complementary features about protein structure and function will further improve computational variant interpretation methods.

Statistical association tests that incorporate protein information are also in their infancy. Information from a variety of variant-induced molecular changes could be added to these tests. Advances in our understanding of molecular genetics will inform these tests as well. As most rare variants are heterozygous (affecting only one of two chromosome copies), advances in modeling allele-specific expression (138), haploinsufficiency effects (139), and protein translation rates (140) may explain additional variation in disease risk due to missense variants. In the coming years, these new types of data, along with sophisticated modeling approaches (**Figure 4**), will empower a multiscale view of variant function, providing informative links from the genome, through protein function, to molecular and disease phenotypes.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to thank the following funding sources: American Heart Association Postdoctoral Fellowship 20POST35220002 (to B.L.) and National Institutes of Health awards R01GM126249 (to W.S.B) and R01LM013434 (to J.A.C).

LITERATURE CITED

1. Pauling L, Itano HA, Singer SJ, Wells IC. 1949. Sickle cell anemia, a molecular disease. *Science* 110(2865):543–48
2. Ingram VM. 1957. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* 180(4581):326–28
3. Strasser BJ. 1999. "Sickle cell anemia, a molecular disease." *Science* 286(5444):1488–90
4. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT. 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature* 185(4711):416–22
5. Vekilov PG. 2007. Sickle-cell haemoglobin polymerization: Is it the primary pathogenic event of sickle-cell anaemia? *Br. J. Haematol.* 139(2):173–84
6. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306(5940):234–38
7. MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, et al. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell.* 72(6):971–83
8. Zuccato C, Ciammola A, Rigamonti D, Leavitt BR, Goffredo D, et al. 2001. Loss of huntingtin-mediated BDNF gene transcription in Huntington's disease. *Science* 293(5529):493–98
9. Tsui LC, Dorfman R. 2013. The cystic fibrosis gene: a molecular genetic perspective. *Cold Spring Harb. Perspect. Med.* 3(2):a009472
10. Zhang Z, Liu F, Chen J. 2018. Molecular structure of the ATP-bound, phosphorylated human CFTR. *PNAS* 115(50):12757–62
11. Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J. 2017. Molecular structure of the human CFTR ion channel. *Cell* 169(1):85–95.e8
12. wwPDB (Worldw. Protein Data Bank) Found. 2022. *Deposition statistics*. Web Resour. wwPDB Found., Piscataway, NJ, accessed Jan. 1. **https://www.wwpdb.org/stats/deposition**

13. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 47(D1):D520–28
14. Nogales E. 2016. The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* 13:24–27
15. Bonomi M, Vendruscolo M. 2019. Determination of protein structural ensembles using cryo-electron microscopy. *Curr. Opin. Struct. Biol.* 56:37–45
16. Thonghin N, Kargas V, Clews J, Ford RC. 2018. Cryo-electron microscopy of membrane proteins. *Methods* 147:176–86
17. Zhang Z, Liu F, Chen J. 2018. Molecular structure of the ATP-bound, phosphorylated human CFTR. *PNAS* 115(50):12757–62
18. Guo Q, Huang B, Cheng J, Seefelder M, Engler T, et al. 2018. The cryo-electron microscopy structure of huntingtin. *Nature* 555(7694):117–20
19. Muhammed MT, Aki-Yalcin E. 2019. Homology modeling in drug discovery: overview, current applications, and future perspectives. *Chem. Biol. Drug Des.* 93(1):12–20
20. Haddad Y, Adam V, Heger Z. 2020. Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLOS Comput. Biol.* 16(4):e1007449
21. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46(W1):W296–303
22. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, et al. 2014. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* 42:D336–46
23. Dill KA, Ozkan SB, Shell MS, Weikl TR. 2008. The protein folding problem. *Annu. Rev. Biophys.* 37:289–316
24. Li B, Fooksa M, Heinze S, Meiler J. 2018. Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Crit. Rev. Biochem. Mol. Biol.* 53(1):1–28
25. Moult J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins* 23(3):ii–v
26. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, et al. 2019. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinforma.* 87(12):1141–48
27. Xu J, Wang S. 2019. Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins Struct. Funct. Bioinforma.* 87(12):1069–81
28. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–89
29. Lupas AN, Pereira J, Alva V, Merino F, Coles M, Hartmann MD. 2021. The breakthrough in protein structure prediction. *Biochem. J.* 478(10):1885–90
30. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557):871–76
31. Humphreys I, Pei J, Baek M, Krishnakumar A, Anishchenko I, et al. 2021. Computed structures of core eukaryotic protein complexes. *Science* 374(6573):eabm4805
32. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, et al. 2021. De novo protein design by deep network hallucination. *Nature* 600(7889):547–52
33. Abdellah Z, Ahmadi A, Ahmed S, Aimable M, Ainscough R, et al. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011):931–45
34. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, et al. 2017. DNA sequencing at 40: past, present and future. *Nature* 550(7676):345–53
35. Bush WS, Moore JH. 2012. Chapter 11: genome-wide association studies. *PLOS Comput. Biol.* 8(12):e1002822
36. Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Investig.* 118(5):1590–605
37. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, et al. 2017. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101(1):5–22
38. Butkiewicz M, Blue EE, Leung YY, Jian X, Marcora E, et al. 2018. Functional annotation of genomic variants in studies of late-onset Alzheimer's disease. *Bioinformatics* 34(16):2724–31

39. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285–91
40. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809):434–43
41. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 29(1):308–11
42. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, et al. 2015. A global reference for human genetic variation. *Nature* 526(7571):68–74
43. Visscher PM, Yengo L, Cox NJ, Wray NR. 2021. Discovery and implications of polygenicity of common diseases. *Science* 373(6562):1468–73
44. Wrighton K. 2021. Filling in the gaps telomere to telomere. In *Nature Milestones: Genomic Sequencing*, p. S21. London: Springer Nature. **https://www.nature.com/articles/d42859-020-00117-1**
45. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585(7823):79–84
46. Logsdon GA, Vollger MR, Hsieh PH, Mao Y, Liskovykh MA, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* 593(7857):101–7
47. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, et al. 2021. The complete sequence of a human genome. bioRxiv 2021.05.26.445798. **https://doi.org/10.1101/2021.05.26.445798**
48. Amberger JS, Bocchini CA, Scott AF, Hamosh A. 2019. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. 47(D1):D1038–43
49. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, et al. 1994. Localization of a breast cancer susceptibility gene, *BRCA2*, to chromosome 13q12-13. *Science* 265(5181):2088–90
50. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250(4988):1684–89
51. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, et al. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261(5123):921–23
52. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 47(D1):D1005–12
53. Sirugo G, Williams SM, Tishkoff SA. 2019. The missing diversity in human genetic studies. *Cell* 177(4):26–31
54. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, et al. 2018. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* 19(3):175–85
55. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. 2019. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* 20(12):747–59
56. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89(1):82–93
57. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, et al. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91(2):224–37
58. Wang K. 2016. Boosting the power of the sequence kernel association test by properly estimating its null distribution. *Am. J. Hum. Genet.* 99(1):104–14
59. Schweiger R, Weissbrod O, Rahmani E, Müller-Nurasyid M, Kunze S, et al. 2017. RL-SKAT: an exact and efficient score test for heritability and set tests. *Genetics* 207(4):1275–83
60. Lee S, Teslovich TM, Boehnke M, Lin X. 2013. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* 93(1):42–53
61. Sun J, Zheng Y, Hsu L. 2013. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol.* 37(4):334–44
62. Lin DY, Tang ZZ. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* 89(3):354–67
63. Chen H, Meigs JB, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* 37(2):196–204

64. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, et al. 2021. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 597(7877):527–32
65. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, et al. 2018. Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* 25:1859–75
66. Kennedy B, Kronenberg Z, Hu H, Moore B, Flygare S, et al. 2014. Using VAAST to identify disease-associated variants in next-generation sequencing data. *Curr. Protoc. Hum. Genet.* 81:6.14.1–6.14.25
67. Li X, Li Z, Zhou H, Gaynor SM, Liu Y, et al. 2020. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* 52(9):969–83
68. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* 7(4):248–49
69. Richards S, Aziz N, Bale S, Bick D, Das S, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17(5):405–24
70. Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. 2019. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.* 44(7):575–88
71. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, et al. 2017. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* 101(3):315–25
72. Kinney JB, McCandlish DM. 2019. Massively parallel assays and quantitative sequence-function relationships. *Annu. Rev. Genom. Hum. Genet.* 20:99–127
73. Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods.* 11(8):801–7
74. Fowler DM, Stephany JJ, Fields S. 2014. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* 9(9):2267–84
75. Glazer AM, Wada Y, Li B, Muhammad A, Kalash OR, et al. 2020. High-throughput reclassification of *SCN5A* variants. *Am. J. Hum. Genet.* 107(1):111–23
76. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, et al. 2019. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* 20:223
77. Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, et al. 2012. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J. Mol. Biol.* 424(3–4):150–67
78. Stiffler MA, Hekstra DR, Ranganathan R. 2015. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* 160(5):882–92
79. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, et al. 2015. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 200(2):413–22
80. Gasperini M, Starita L, Shendure J. 2016. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* 11(10):1782–87
81. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, et al. 2018. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50(6):874–82
82. Nussinov R, Tsai CJ, Jang H. 2019. Protein ensembles link genotype to phenotype. *PLOS Comput. Biol.* 15(6):e1006648
83. Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, et al. 2012. Systematic functional prioritization of protein posttranslational modifications. *Cell* 150(2):413–25
84. Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell* 152(6):1237–51
85. Yourshaw M, Taylor SP, Rao AR, Martín MG, Nelson SF. 2015. Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Brief. Bioinform.* 16(2):255–64
86. Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* 12(9):628–40
87. Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.* 30(5):703–14
88. Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genom. Hum. Genet.* 7:61–80

89. Niroula A, Vihinen M. 2016. Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.* 37(6):579–97
90. Zhu C, Miller M, Zeng Z, Wang Y, Mahlich Y, et al. 2020. Computational approaches for unraveling the effects of variation in the human genome and microbiome. *Annu. Rev. Biomed. Data Sci.* 3:411–32
91. Rost B, Radivojac P, Bromberg Y. 2016. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* 590(15):2327–41
92. Heyne HO, Baez-Nieto D, Iqbal S, Palmer DS, Brunklaus A, et al. 2020. Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci. Transl. Med.* 12(556):aav2848
93. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, et al. 2020. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* 11:5918
94. Wang C, Balch WE. 2018. Bridging genomics to phenomics at atomic resolution through variation spatial profiling. *Cell Rep.* 24(8):2013–28.e6
95. Tsang M, Cheng D, Liu Y. 2018. Detecting statistical interactions from neural network weights. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. **https://openreview.net/forum?id=ByOfBggRZ**
96. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, et al. 2015. Comparison of predicted and actual consequences of missense mutations. *PNAS* 112(37):E5189–98
97. Itan Y, Casanova JL. 2015. Can the impact of human genetic variations be predicted? *PNAS* 112(37):11426–27
98. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46(3):310–15
99. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, et al. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99(4):877–85
100. Care MA, Needham CJ, Bulpitt AJ, Westhead DR. 2007. Deleterious SNP prediction: Be mindful of your training data! *Bioinformatics* 23(6):664–72
101. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* 132(10):1077–130
102. Li B, Mendenhall JL, Kroncke BM, Taylor KC, Huang H, et al. 2017. Predicting the functional impact of KCNQ1 variants of unknown significance. *Circ. Cardiovasc. Genet.* 10(5):e001754
103. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, et al. 2017. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* 27(10):1715–29
104. Evans P, Wu C, Lindy A, McKnight DA, Lebo M, et al. 2019. Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets. *Genome Res.* 29(7):1144–51
105. Manolio TA, Rowley R, Williams MS, Roden D, Ginsburg GS, et al. 2019. Opportunities, resources, and techniques for implementing genomics in clinical care. *Lancet* 394(10197):511–20
106. Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, et al. 2019. Genomic medicine for undiagnosed diseases. *Lancet* 394(10197):533–40
107. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, et al. 2019. Pharmacogenomics. *Lancet* 394(10197):521–32
108. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46(D1):D1062–67
109. Andreoletti G, Pal LR, Moult J, Brenner SE. 2019. Reports from the fifth edition of CAGI: the critical assessment of genome interpretation. *Hum. Mutat.* 40(9):1197–201
110. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* 508(7497):469–76
111. Eilbeck K, Quinlan A, Yandell M. 2017. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* 18(10):599–612
112. Stehr H, Jang S-HJ, Duarte JM, Wierling C, Lehrach H, et al. 2011. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol. Cancer* 10:54

113. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, et al. 2015. Comprehensive assessment of cancer missense mutation clustering in protein structures. *PNAS* 112(40):E5486–95
114. Meyer MJ, Lapcevic R, Romero AE, Yoon M, Das J, et al. 2016. mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum. Mutat.* 37(5):447–56
115. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, et al. 2016. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.* 76(13):3719–31
116. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, et al. 2016. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* 48(8):827–37
117. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, et al. 2017. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* 9:4
118. Kumar S, Clarke D, Gerstein MB. 2019. Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. *PNAS* 116(38):18962–70
119. Sivley RM, Dou X, Meiler J, Bush WS, Capra JA. 2018. Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am. J. Hum. Genet.* 102(3):415–26
120. Martinez-Ledesma E, Flores D, Trevino V. 2020. Computational methods for detecting cancer hotspots. *Comput. Struct. Biotechnol. J.* 18:3567–76
121. West RM, Lu W, Rotroff DM, Kuenemann MA, Chang SM, et al. 2019. Identifying individual risk rare variants using protein structure guided local tests (POINT). *PLOS Comput. Biol.* 15(2):e1006722
122. Tang ZZ, Sliwoski GR, Chen G, Jin B, Bush WS, et al. 2020. PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biol.* 21:217
123. Jin B, Capra JA, Benchek P, Wheeler N, Naj AC, et al. 2022. An association test of the spatial distribution of rare missense variants within protein structures identifies Alzheimer's disease-related patterns. *Genome Res.* 32(4):778–90
124. Sivley RM, Sheehan JH, Kropski JA, Cogan J, Blackwell TS, et al. 2018. Three-dimensional spatial analysis of missense variants in *RTEL1* identifies pathogenic variants in patients with Familial Interstitial Pneumonia. *BMC Bioinform.* 19:18
125. Li B, Roden DM, Capra JA. 2021. The 3D spatial constraint on 6.1 million amino acid sites in the human proteome. bioRxiv 2021.09.15.460390. **https://doi.org/10.1101/2021.09.15.460390**
126. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50(D1):D439–44
127. Wang Z, Moult J. 2001. SNPs, protein structure, and disease. *Hum. Mutat.* 17(4):263–70
128. Yue P, Li Z, Moult J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353(2):459–73
129. Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* 10(10):709–20
130. Gao M, Zhou H, Skolnick J. 2015. Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure* 23(7):1362–69
131. Li B, Yang YT, Capra JA, Gerstein MB. 2020. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLOS Comput. Biol.* 16(11):e1008291
132. Perszyk RE, Kristensen AS, Lyuboslavsky P, Traynelis SF. 2021. Three-dimensional missense tolerance ratio analysis. *Genome Res.* 31(8):1447–61
133. Silk M, Pires DEV, Rodrigues CHM, D'Souza EN, Olshansky M, et al. 2021. MTR3D: identifying regions within protein tertiary structures under purifying selection. *Nucleic Acids Res.* 49(W1):W438–45
134. Hicks M, Bartha I, di Iulio J, Venter JC, Telenti A. 2019. Functional characterization of 3D protein structures informed by human genetic diversity. *PNAS* 116(18):8960–65
135. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, et al. 2021. Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034. **https://doi.org/10.1101/2021.10.04.463034**

136. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold—making protein folding accessible to all. bioRxiv 2021.08.15.456425. **https://doi.org/10.1101/2021.08.15.456425**
137. Callaway E. 2021. DeepMind's AI predicts structures for a vast trove of proteins. *Nature* 595(7869):635
138. Castel SE, Aguet F, Mohammadi P, Aguet F, Anand S, et al. 2020. A vast resource of allelic expression data spanning human tissues. *Genome Biol*. 21:234
139. Cummings BB, Karczewski KJ, Kosmicki JA, Seaby EG, Watts NA, et al. 2020. Transcript expression-aware annotation improves rare variant interpretation. *Nature* 581(7809):452–58
140. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, et al. 2015. Impact of regulatory variation from RNA to protein. *Science* 347(6222):664–67