

Genomics 2 Proteins portal: a resource and discovery tool for linking genetic screening outputs to protein sequences and structures

Received: 3 January 2024

Accepted: 9 August 2024

Published online: 18 September 2024

 Check for updates

Seulki Kwon^{1,11}, Jordan Safer^{1,11}, Duyen T. Nguyen², David Hoksza³, Patrick May⁴, Jeremy A. Arbesfeld⁵, Alan F. Rubin^{6,7}, Arthur J. Campbell¹, Alex Burgin¹ & Sumaiya Iqbal^{1,8,9,10} 

Recent advances in AI-based methods have revolutionized the field of structural biology. Concomitantly, high-throughput sequencing and functional genomics have generated genetic variants at an unprecedented scale. However, efficient tools and resources are needed to link disparate data types—to ‘map’ variants onto protein structures, to better understand how the variation causes disease, and thereby design therapeutics. Here we present the Genomics 2 Proteins portal (<https://g2p.broadinstitute.org/>): a human proteome-wide resource that maps 20,076,998 genetic variants onto 42,413 protein sequences and 77,923 structures, with a comprehensive set of structural and functional features. Additionally, the Genomics 2 Proteins portal allows users to interactively upload protein residue-wise annotations (for example, variants and scores) as well as the protein structure beyond databases to establish the connection between genomics to proteins. The portal serves as an easy-to-use discovery tool for researchers and scientists to hypothesize the structure–function relationship between natural or synthetic variations and their molecular phenotypes.

We live in the era of big biological data where there has been an insurgence of efficient deep learning methods, making millions of high-quality predicted protein structures available to the biomedical community^{1–5}. At the same time, cryo-electron microscopy and other experimental methods are leading to a burst of high-resolution protein structures and assemblies^{6–8}. These advances come when an unprecedented number of genetic variants in the general population and those associated with diseases have been identified and accumulated in multiple databases^{9–13}. Concomitantly, advances in functional

genomics approaches (for example, base editing^{14,15}, prime editing¹⁶ and Perturb-seq¹⁷) have enabled the generation of synthetic mutations and the quantification of their functional impact in different cellular models. Mapping natural or synthetic variants in the context of protein structure provides valuable molecular-level insights and helps hypothesize the structure–function mechanism of the variant.

Challenges remain, however, in connecting genomic data (that is, variants) to protein structural data due to the complexity introduced by diverse RNA transcripts and protein isoforms originating from a

¹Center for the Development of Therapeutics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ²PATTERN, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ³Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

⁴Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. ⁵The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children’s Hospital, Columbus, OH, USA. ⁶Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. ⁷Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia. ⁸Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. ¹⁰Cancer Data Sciences, Dana-Farber/Harvard Cancer Center, Boston, MA, USA. ¹¹These authors contributed equally: Seulki Kwon, Jordan Safer.

 e-mail: sumaiya@broadinstitute.org

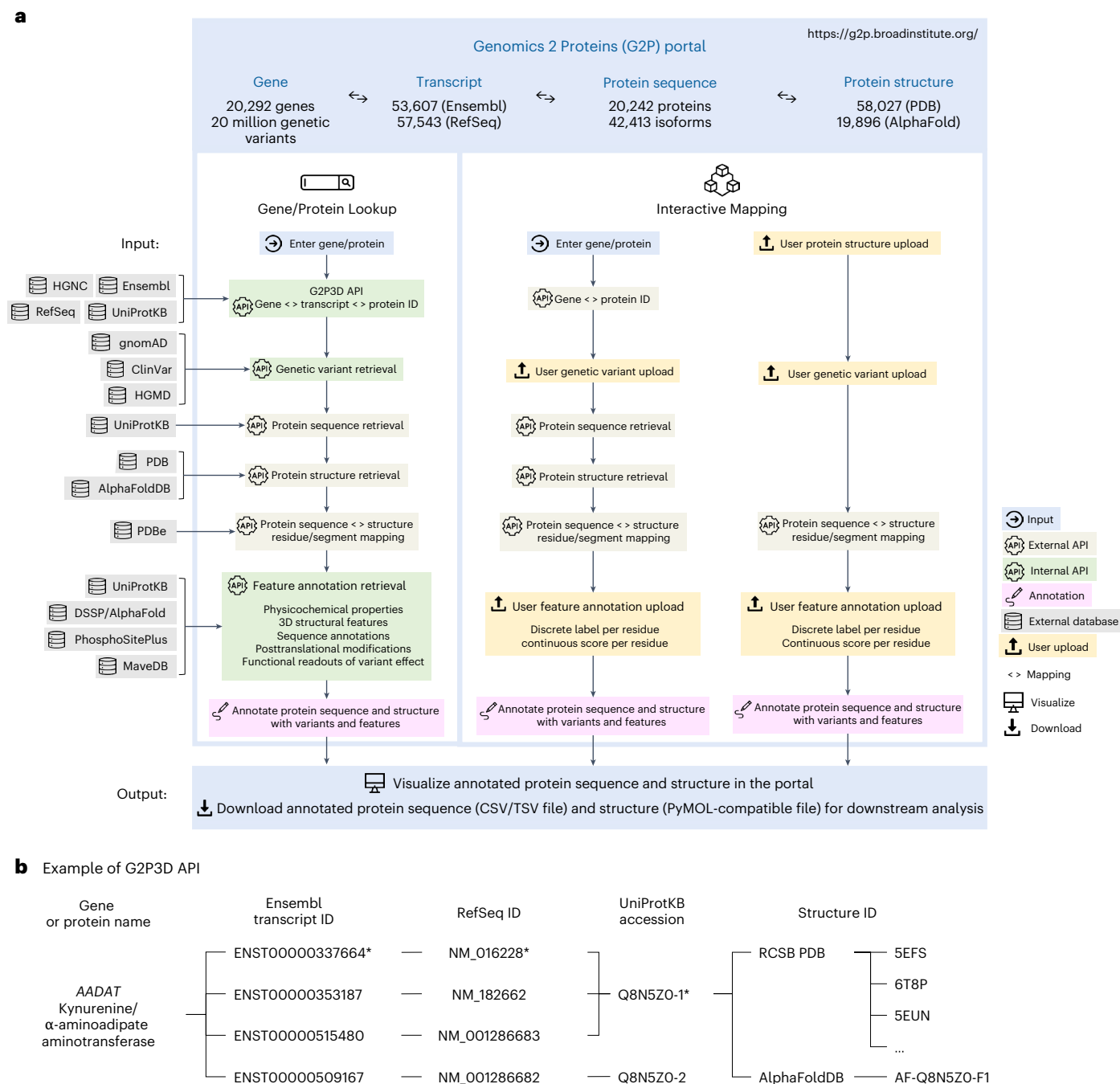


Fig. 1 | The bioinformatic framework of the G2P portal. a, Schematic of data and method integration in the G2P portal and its two main modules: ‘Gene/Protein Lookup’ and ‘Interactive Mapping’. In the Gene/Protein Lookup module, the connections across identifiers of human genes, transcripts, protein sequences and structures were established using an in-house API: G2P3D, for the entire human proteome (see ‘Construction of G2P3D API’ in Methods for details). Variants from databases, such as gnomAD⁹, ClinVar¹⁰ and HGMD¹¹, were subsequently mapped onto protein sequences and structures upon dynamically querying UniProtKB²¹ and structure databases (PDB⁶ and AlphaFoldDB²⁵), respectively. Additionally, protein feature annotations were fetched and calculated from various databases and tools (UniProtKB, DSSP³⁰ and PhosphoSitePlus²⁷). All annotated protein sequences and structures with variants and features are viewable on the portal and downloadable in interoperable formats for further analyses. In the Interactive Mapping module of the portal,

users can upload protein residue-wise annotations of variants and additional features and perform linking genetic data to protein structural data. Users can access this module by starting from a gene and by uploading an in-house protein structure. **b**, An example of G2P3D API output; the API links human genes (HGNC²³) to transcripts (Ensembl²³ and RefSeq²⁴) to protein sequences (UniProtKB) and structures (PDB⁶ and AlphaFoldDB²⁵). In this example, *AADAT* has four Ensembl transcripts and four RefSeq transcripts; three pairs of Ensembl-RefSeq transcripts encode the canonical protein isoform (Q8N5Z0-1*) and the remaining one transcript (ENST00000509167/NM_001286682) corresponds to the noncanonical protein isoform, Q8N5Z0-2. The canonical protein isoform is further dynamically linked to multiple available PDB structures and the AlphaFold structure. In the portal, variants are mapped onto both canonical and noncanonical protein isoforms. Only canonical protein isoform variants are mapped to available protein structures.

single DNA sequence¹⁸, which requires accurate transcript–protein isoform mapping. Another technical hurdle lies in reconciling disparate formats between genomic identifiers of variants such as rsIDs¹⁹ or Human Genome Variation Society (HGVS) notations²⁰ and those of protein data featured on amino acid sequences and their spatial coordinates in structures²¹, often only available for fragments of the full-length protein. Aligning and harmonizing this discrepancy in naming and formatting requires interdisciplinary knowledge or collaborations across researchers from genetics, structural biology and computational biology. Therefore, proteome-wide linking of human genes to protein structures through identifiers and matching variants to their corresponding structural positions thereafter, necessitates efficient computational methods. Additionally, a user-friendly, online method that supports such linking alongside interactive options for uploading data, applying customized filters on variants and exporting interoperable results is needed to enable a broad community toward mapping variants onto proteins, analyzing targets informed by genetics, and thereby designing better therapeutics, which is precisely why we developed the Genomics 2 Proteins (G2P) portal.

The G2P portal is a bioinformatic tool to dynamically query, retrieve and connect genetic variants and transcripts to protein sequence annotations and structures wrapped within an interactive web interface with visualization functions. When writing the paper, ~20 million genetic variants within all human protein-coding genes from public databases are aggregated and annotated within protein sequences and structures with comprehensive protein feature reports. By exploiting experimentally solved and predicted protein structures, the G2P portal covers 99% of all human proteins with corresponding structures.

Results

G2P bioinformatic method and portal

The G2P web portal and the underlying bioinformatic method are built upon a dynamic application programming interface (API) for seamless mapping of identifiers from public databases for genes (HUGO Gene Nomenclature Committee (HGNC))²², transcripts (the Ensembl genome browser²³ and RefSeq²⁴; NCBI Reference Sequence Database), protein sequences (UniProt KnowledgeBase (UniProtKB))²¹ and structures (Protein Data Bank (PDB)⁶ and AlphaFold Database (AlphaFoldDB))²⁵. We call the API Genomics 2 Proteins 3D or G2P3D API ('Construction of G2P3D API' in Methods). As of October 2023, the API links 20,292 human genes that encode 20,242 UniProtKB proteins corresponding to 42,413 isoforms, via 53,607 Ensembl transcripts and 57,543 RefSeq transcripts, to 77,923 protein structures (58,027 PDB structures and 19,896 AlphaFold structures). A schematic overview of the portal and data flow via G2P3D API is illustrated in Fig. 1.

Multi-omics data aggregated in the portal showed that about 47% of all human genes have one unique protein isoform to date, as available in UniProtKB (release 2023_05). The remaining 53% of genes have on average three isoforms by alternative splicing. Overall, 85% of all canonical protein isoforms were translated from the MANE (Matched Annotation from NCBI and EBI) Select transcripts²⁶. Additionally, 99% of human proteins have either PDB or AlphaFold structures. Further statistics of data aggregated via G2P3D API are available in Supplementary Note 1.

The web-based user interface of the G2P portal is built on Google Cloud infrastructure ('G2P Google Cloud infrastructure' in Methods and Extended Data Fig. 1). The portal features two distinct modules (Fig. 1a): (1) Gene/Protein Lookup: a human proteome-wide resource for users to link genetic variants from transcripts to protein sequences and structures; (2) Interactive Mapping: a tool for users to analyze their data, thereby, not limited to publicly available variants or protein structures. The Gene/Protein Lookup module currently aggregates variants from the Genome Aggregation Database (gnomAD)⁹, ClinVar¹⁰ and the Human Gene Mutation Database¹¹ (HGMD) and then maps

them onto amino acid positions upon dynamic retrieval of the protein sequence and structure (Fig. 1b). Moreover, a comprehensive set of protein residue-wise annotations (referred to as 'protein features') are computed and aggregated in the portal, such as UniProt-based features (for example, domain and active site), structural features, posttranslational modifications (PTMs)²⁷ and variants effects from MaveDB²⁸. Finally, variants are mapped onto protein sequences and structures simultaneously with protein features, aiding informed variant-mechanism analysis by assessing the mutational impact on important protein features. Further details of submodules within the two main modules of the portal are available in 'G2P portal sitemap' in Methods and Extended Data Fig. 2.

A suite of visualization tools has been implemented in the portal for intuitive data exploration ('Data visualization tools in the G2P portal' in Methods and Extended Data Fig. 3). The portal undergoes a biannual data update to incorporate the latest information from external databases and maintain the dynamically linked APIs (Supplementary Note 2 and Supplementary Table 1).

Resources in the G2P portal

The Gene/Protein Lookup module of the G2P portal contains variants, protein structures and protein features for all human protein-coding genes (Supplementary Table 2). Genes and proteins are classified by the HGNC²² gene family and protein class from PANTHER knowledgebase²⁹ (Supplementary Note 3.1, Supplementary Fig. 1 and Supplementary Tables 3 and 4).

Variant data. We applied the bioinformatic method to variants in three human genetic variation databases: gnomAD⁹, ClinVar¹⁰ and HGMD¹¹, referred to as gnomAD, ClinVar and HGMD variants, respectively (Fig. 2). As of October 2023, G2P portal aggregates 18,014,632 protein-coding variants annotated in 18,723 human genes from gnomAD v2.1.1, 1,749,628 protein-coding variants in 18,180 genes from ClinVar (10/2023 release) and 312,738 disease-causing mutations in 12,918 genes from HGMD professional 2023.01 ('Variant aggregation' in Methods). In the portal, variants were grouped based on their protein consequence: missense, nonsense, synonymous, frameshift, in-frame indel and others (Fig. 2a–c) and visualized as separate tracks in the protein sequence viewer. Additionally, the portal allows for a customized selection of variants based on database-specific filters: allele frequency (AF) and allele count (AC) for gnomAD, clinical significance for ClinVar, and confidence ascertainment for HGMD variants. This unique functionality facilitates the mapping and interpretation of variants based on their specific protein consequences and clinically relevant criteria.

Most gnomAD variants (~97%) are very rare (Fig. 2d), and the fraction of common variants is the largest for synonymous variants (0.78%). Clinical significance of ClinVar variants (Fig. 2e) showed a correlation with protein consequences: the proportion of benign/likely benign (BLB) and pathogenic/likely pathogenic (PLP) variants are inversely correlated across synonymous (90.1% BLB and 0.1% PLP) and nonsense (0.8% BLB and 83.0% PLP) mutations. Missense mutations hold the highest fraction (87.1%) of variants of uncertain significance (VUS) and conflicting interpretation (CI). Within HGMD data, frameshift and nonsense mutations account for the two largest fractions of high-confidence disease mutations (87% and 81%, respectively), while most synonymous mutations are of low confidence (Fig. 2f). Additional variant statistics are available in Supplementary Note 3.2. The portal maps variants to both canonical and noncanonical protein isoforms (a comparative overview of statistics is shown in Supplementary Fig. 2).

Protein structural data. G2P portal dynamically queries structure databases to map variants annotated on transcripts onto structures of canonical protein isoforms (according to UniProtKB²¹). As of October 2023, variants were mapped onto 58,026 PDB⁶ structures of 7,973 proteins (Supplementary Fig. 3 and Supplementary Table 5) and 19,896

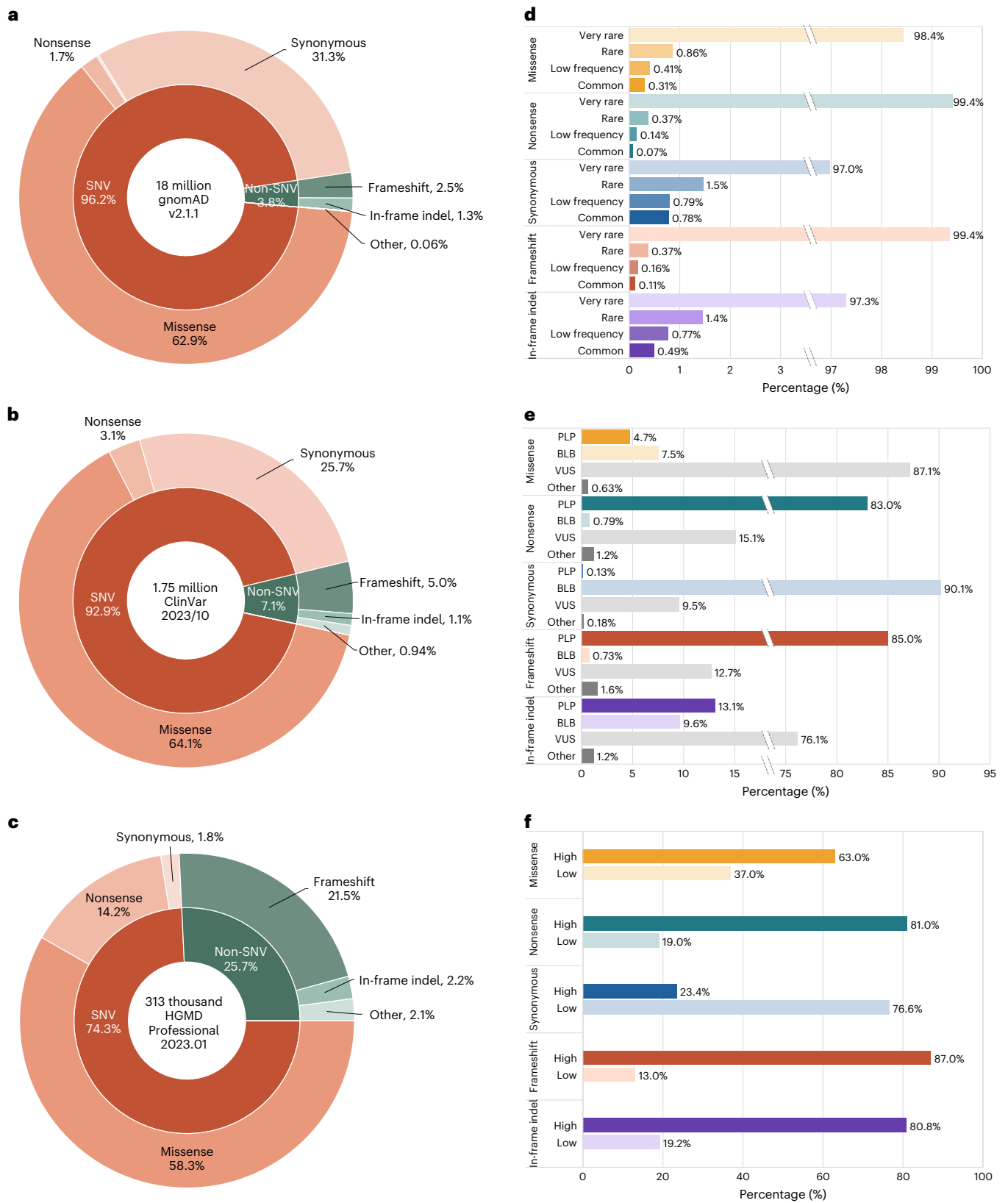


Fig. 2 | Statistics of variants from gnomAD, ClinVar and HGMD databases aggregated in the G2P portal. a–c, Distribution of variant types (single nucleotide variation (SNV) versus non-SNV; insertion, deletion and inversion) and associated protein consequences (missense, synonymous, nonsense, frameshift, in-frame indel and others for all other protein consequences) among 20 million protein-coding variants in gnomAD (a), ClinVar (b) and HGMD (c) databases. Among all databases, a majority of human protein-coding variants are SNV occurring missense mutations. **d,** Distribution of gnomAD variants

categorized by AFs: very rare; $AF < 0.1\%$, rare; $0.1\% \leq AF < 0.5\%$, low frequency; $0.5\% \leq AF < 5\%$, and common; $AF \geq 5\%$. The distributions of each AF group are illustrated across different protein consequences (missense, synonymous, nonsense, frameshift and in-frame indel). **e,** Distribution of the clinical significance of ClinVar variants (PLP, BLB, VUS/CI and others) displayed across different protein consequences. **f,** Distribution of confidence levels (high or low) for HGMD variants across different protein consequences.

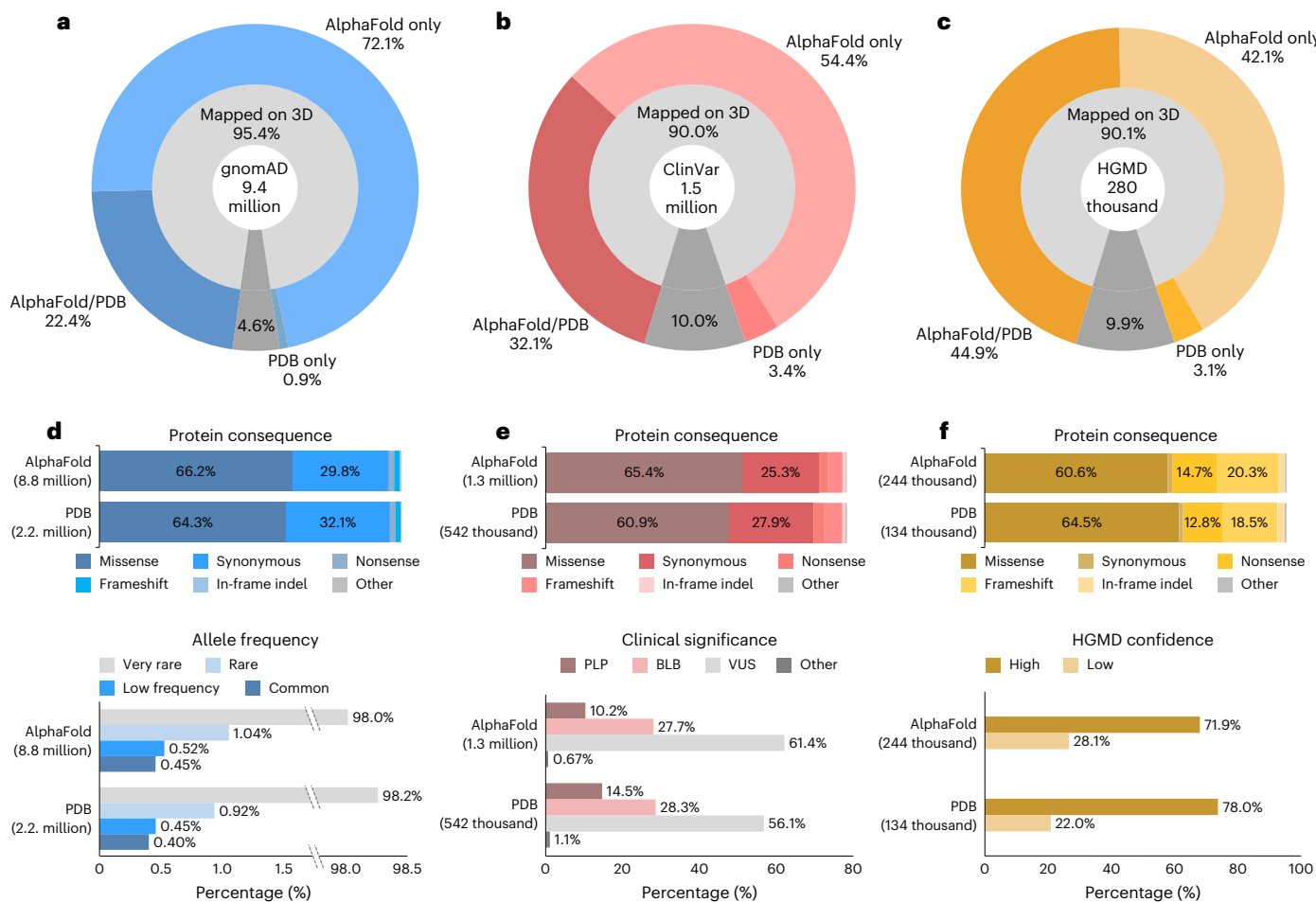


Fig. 3 | Statistics of variants mapped on 3D structures in the G2P portal.

Variants annotated on transcripts corresponding to the canonical protein isoforms were mapped on 3D structures. The total number of canonical protein isoform variants from each database is shown in the middle of the donut chart. **a**, The proportion of 9.4 million gnomAD variants mapped on PDB⁹, AlphaFoldDB²⁵ or both. **b**, The proportion of 1.5 million ClinVar variants mapped on PDB, AlphaFold or both. **c**, The proportion of 280 thousand HGMD variants mapped on PDB, AlphaFold or both. **d**, The distribution of protein consequences (upper)

and AF group (lower) among gnomAD variants mapped on AlphaFold (8.8 million variants) and PDB (2.2 million variants). **e**, The distribution of protein consequences (upper) and clinical significance (lower) among ClinVar variants mapped on AlphaFold (1.3 million variants) and PDB (542 thousand variants). **f**, The distribution of protein consequences (upper) and confidence (lower) among HGMD variants mapped on AlphaFold (244 thousand variants) and PDB (134 thousand variants).

AlphaFold²⁵ structures of 19,972 proteins. AlphaFoldDB covers structures of 12,105 proteins with no experimental structure available yet, with reasonably high predicted local distance difference test (pLDDT)²⁵ (median \approx 75; Supplementary Note 4 and Supplementary Figs. 4 and 5), highlighting the utility of AlphaFold structures for variant mapping and protein three-dimensional (3D) feature analysis of variant positions. As a result, over 90% of gnomAD, ClinVar and HGMD variants were successfully mapped onto protein structures (Fig. 3a–c).

The distribution of protein consequences and database-specific groups of variants (based on AF for gnomAD, clinical significance for ClinVar and confidence for HGMD) mapped on structures illustrated that gnomAD variants do not show differences in AF distributions between variants mapped on PDB and AlphaFold structures (Fig. 3d). A higher fraction of ClinVar PLP variants were mapped on PDB structures (14.5%) compared to AlphaFold structures (10.2%; Fig. 3e). Similarly, a higher fraction of high-confidence disease mutations from HGMD were mapped on PDB structures (78%) than AlphaFold structures (72%; Fig. 3f).

Protein feature data. A comprehensive set of per-residue protein features are integrated into the portal ('Protein features in the G2P portal' in Methods), to help users establish the relationship between

genetic variations and protein structure–function ('Variant and feature mapping onto proteins' in Methods). The features are grouped into (1) physicochemical properties of amino acids, (2) structural features^{30–32} based on PDB and AlphaFold structures, (3) sequence annotations from UniProtKB, (4) PTMs from PhosphoSitePlus²⁷, and (5) readouts from multiplexed assays of variant effect (MAVE) when available in MaveDB²⁸. The portal's proteome-wide feature annotations provide insight into the differential structural and functional characteristics of gnomAD versus ClinVar and HGMD variants. In Fig. 4, we present the abundance of protein features across nine missense variant groups divided by database-specific filters: gnomAD (very rare, rare, low frequency and common), ClinVar (PLP, BLB and VUS) and HGMD (high and low confidence). Notably, gnomAD common and ClinVar BLB missense variants as well as ClinVar PLP and HGMD high-confidence missense variants exhibit similar distributions across all protein features. Similar results for variants leading to different protein consequences are reported in Supplementary Note 5 and Supplementary Figs. 6–10.

Case study—Gene/Protein Lookup module

In the Gene/Protein Lookup module, users can query the G2P portal by a gene or protein name and map variants from gnomAD⁹, ClinVar¹⁰

and HGMD¹¹ from a selected transcript, and protein features onto the protein sequence. Variants and features can thereafter be transferred to a selected PDB⁶ and AlphaFold²⁵ structure. Users can explore the customized annotations of protein sequences and structures with variants and protein features online and download mappings in tabular and PyMOL³³ formats.

Figure 5 presents a case study using *MORC2*, encoding ATPase MORC2 protein, associated with Charcot–Marie–Tooth disease type 2 (CMT2Z)³⁴. Upon selecting a transcript under the ‘variant to protein sequence’ tab (Fig. 5a), users can view variants and protein features mapped in the protein sequence viewer. The portal provides easily selectable filters for variants and features. For example, by filtering for Missense and Pathogenic/Likely pathogenic ClinVar variants, we observe a cluster of PLP missense variants in the N-terminal region (residue 20–470) of MORC2. Protein feature tracks offer further insights into this region: (1) PDBe/SIFT track shows that the N terminus of MORC2 has experimentally solved structures; (2) accessible surface area and AlphaFold confidence (pLDDT) tracks indicate that the N-terminal region exhibits greater structural integrity with lower surface area and higher pLDDT, compared to the C-terminal region; and (3) binding site track highlights predominant presence of catalytic sites (ATP- and ZN²⁺-binding) in the N terminus. Users can visualize these observations on 3D structures using the structure viewer under the ‘variant to protein structure’ tab (Fig. 5b). Here, mapping ClinVar PLP missense variants (yellow, Fig. 5b) alongside the binding site track on the crystal structure of the MORC2 homodimer (PDB 5OF9) reveals the mutations’ proximity to the dimer interface and binding sites (black, Fig. 5b). This case study underscores the utility of G2P portal in unveiling the spatial distribution of *MORC2* pathogenic variants and their protein structure–function relationship.

Case study—Interactive Mapping module

The Interactive Mapping module of the G2P portal allows users to upload genetic variants or sequence annotations, for example, domains, drug-binding pockets and conservation scores, and map them to the target protein’s structure (PDB, AlphaFold or user-provided), extending the capability of linking genomics to proteins beyond existing databases and even human proteome. The portal ensures the security of user-uploaded data (Extended Data Fig. 1); user-provided data are not shared or saved in the portal’s backend. The versatility of the integrated pipeline allows concurrent mapping of multiple data types (genetic variants, discrete feature annotations and continuous scores) from gene to protein sequence and structure and exporting results for downstream analyses. Figure 6 presents a case study using the published base-editing (BE) scanning results of the DNA methyltransferase 3A¹⁵. From ‘start with a gene/protein identifier’, we entered the gene *DNMT3A*, chose a structure (PDB 4U7T) and uploaded annotations such as (1) 34 missense variants (base-edited position) with absolute sgDNA

scores $\geq \pm 2$ standard deviation, (2) sgRNA scores from the BE screen and the pathogenicity prediction scores from AlphaMissense³⁵, and (3) domain annotations (domain) from ref. 15 (Supplementary Table 6). User-uploaded annotations are visible and selectable in the viewer (Fig. 6a), and users can supplement these with additional annotations from integrated resources (described in ‘Resources in the G2P portal’ in Results). By selecting ‘Base-edited position’ and ‘domain’ annotations (Fig. 6a, left), the user can pinpoint the 3D positions of variants within each domain (Fig. 6a, right)—24 and 4 variants are found in MTase and ADD domain, respectively. Figure 6b illustrates the concurrent mapping of user-uploaded and integrated data in the Gene/Protein Lookup module, for example, ClinVar PLP variants and three-class secondary structures; this capability allows users to analyze their variants in the context of known pathogenic variants (Fig. 6b, top) and structural features (Fig. 6b, bottom). Finally, Fig. 6c showcases the mapping of the BE results onto the AlphaFold structure, colored by the AlphaFold confidence scores (pLDDT) and AlphaMissense scores averaged over 20 amino acid substitutions at a reference amino acid position—a capability in the portal that enables users to characterize variants using any state-of-the-art variant effect prediction score. We observed that all 34 base-edited variants mapped to residues with pLDDT > 70, and that 31 variants are pathogenic according to AlphaMissense (score > 0.57). An extension of the *DNMT3A* case study to map user-uploaded annotations on user-provided ESMFold⁴ structure is available in Supplementary Note 6 (Supplementary Fig. 11).

Discussion

Genetic screening is increasingly applied in clinical practice^{36–38}; a persistent bottleneck in translational and clinical genetics, however, is decoding the molecular-level implications of the screening output, that is, genetic variations. Most clinically identified variants remain of uncertain significance³⁹, and determining therapeutically actionable variants is challenging⁴⁰. Bridging genetic variants to structural biology provides a means to connect the potential cause of many diseases to the molecular effect^{41–44}. However, integration of data across genomics, transcriptomics, protein sequence and structure, necessary for connecting variants to protein structure, is not straightforward due to different data types and inherent complexities. Here, we present the G2P portal that overcomes these challenges of multi-omics data integration and establishes the bridge between gene variants and protein structures for the human proteome to help generate hypotheses for translating genetic discovery into molecular consequences and therapeutic mechanisms.

Recognizing the importance, multiple efforts have been made to connect variants to proteins: UniProt feature/variant viewer⁴⁵ displays feature annotations and disease-related variants on protein sequences. RCSB 3D-sequence viewer⁴⁶ shows features from a specific PDB entry and a limited set of variants from UniProt. The Gene2Phenotype⁴⁷,

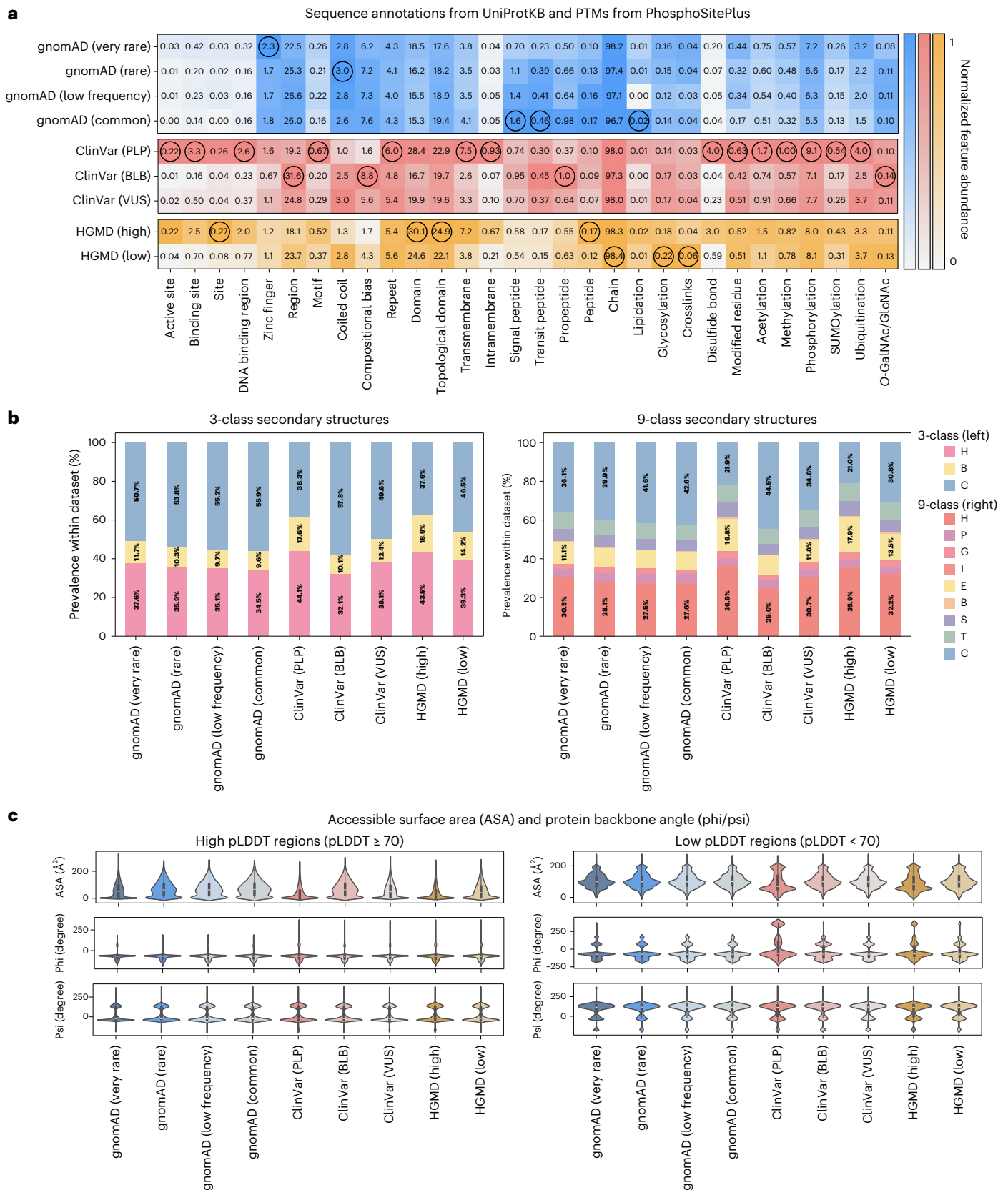
Fig. 4 | Abundance of protein features across nine missense variant datasets.

These variant datasets include gnomAD variants binned by AF: very rare; AF < 0.1%, rare; 0.1% ≤ AF < 0.5%, low frequency; 0.5% ≤ AF < 5%, and common; AF ≥ 5%. ClinVar variants grouped by clinical significance: PLP, BLB, VUS and HGMD disease mutations grouped by confidence levels: high and low. For details about each protein feature, see ‘Protein features in the G2P portal’ in Methods. **a**, The abundance of each sequence annotation from UniProt and PTM site within a given dataset. The calculated abundance of a feature (for example, active site) is denoted as the numerical value at each data point (see Supplementary Fig. 6 for the details of feature abundance calculation). Each point is color coded based on its normalized abundance, wherein the abundance is divided by the maximum value among the nine datasets (denoted as bold and circled) to facilitate comparison of relative abundances across different features. For example, the abundance of the active site is the highest for the ClinVar PLP dataset, represented as 0.23, resulting in the darkest color where normalized abundance equals 1, while the gnomAD common dataset has 0/23 = 0 having the brightest

color. **b**, The proportion of three-class (left) and nine-class (right) secondary structures within variant datasets. Nine secondary structure classes are grouped into three larger classes: helix (H; 3₁₀-helix/G, α -helix/H, π -helix/I and polyproline helix/P), strand (B; β -sheet/E and β -bridge/B) and loop (C; bend/S, turn/T and coil/C). Structured regions (helix and strand) have a higher prevalence of harboring pathogenic variants (~56% of ClinVar PLP variants and HGMD high-confidence disease mutations). **c**, Violin plots showing the distributions of 3D structural features (accessible surface area and backbone phi/psi angles) across different variant datasets. The plots are divided into high (pLDDT ≥ 70, $n = 4,134,666$) and low (pLDDT < 70, $n = 2,544,814$) confidence as predicted by AlphaFold. The violins illustrate the probability density of the data at different values, with the white dot representing the median, the thick black bar in the center representing the interquartile range (IQR), and the thin black line representing the 95% confidence interval. Features of variants summarized in **b** and **c** are computed using AlphaFold structures.

also called G2P, facilitates diagnostic variant filtering given a gene or phenotype, and connects users to DECIPHER⁴⁸ for protein information and mapping of protein-changing variants onto proteins. Additionally, VarMap⁴⁹, in conjunction with ProtVar⁵⁰, offers resources to contextualize missense variations with respect to protein sequence and

structures. The G2P portal stands out from existing tools on multiple fronts, including the scale of integrated resources, the versatility of the user interface and flexible data visualization and export functions tailored toward investigating functional implications of variants within the context of proteins.



Gene/protein lookup module – *MORC2***a** Mapping of variants to the sequence with protein features

MORC2

GeneCards: MORC2²
 HGNC identifier: HGNC:23573³
 HGNC gene symbol: MORC2
 HGNC gene family: Zinc_fingers⁴
 PANTHER protein class: not-available⁵

UniProtKB: Q9Y6X9⁶
 Protein name: ATPase MORC2 (EC 3.6.1.1) (MORC family CW-type zinc finger protein 2) (Zinc finger CW-type coiled-coil domain protein 1)
 Canonical protein isoform: Q9Y6X9-1
 Canonical transcript: ENST00000397641

Protein sequence annotations Variant to protein sequence Variant to protein structure Gene to transcript to protein isoform mapping Additional resources

Select Transcript and Isoform for variant mapping: Q9Y6X9-1: ENST00000397641:NM_001303256*

Select a transcript

Need help with the protein sequence viewer? Show legend

Select Features and Filters

- gnomAD
- ClinVar
- Pathogenicity
- HCMD
- Physicochemical Properties
- Structural Features
- UniProt Sequence Features

Residues 20–470

b Mapping of variants and features to the structure

Protein sequence annotations Variant to protein sequence Variant to protein structure Gene to transcript to protein isoform mapping Additional resources

Select a structure

PDB

Experimentally solved protein structures (source: PDB)

PDB ID	Title	Method	Resolution	Residue count	External Links
S0F9	Crystal structure of human MORC2 (residues 1-603)	X-ray diffraction	1.81	538	PDB of RCSB PDB
S0F9	Crystal structure of human MORC2 (residues 1-603) with central molecular assembly mutation S67L	X-ray diffraction	2.02	541	PDB of RCSB PDB
S0F9	Crystal structure of human MORC2 (residues 1-603) with central molecular assembly mutation T424E	X-ray diffraction	2.57	540	PDB of RCSB PDB

Click to view Click to view Click to view

Click to open structure viewer

AlphaFold Protein Structure Database⁸

Predicted models of protein structures (source: AlphaFold)

AlphaFold ID	Title	Start	End	External Links
AF-Q9Y6X9-F3	ATPase MORC2	1	3032	AlphaFold

Click to view

Fig. 5 | A use case of the Gene/Protein Lookup module for reported variants and protein features of *MORC2*. **a**, The landing page of the Gene/Protein Lookup module shows an overview of the input gene (*MORC2*) information, followed by the protein sequence viewer displaying the aggregated protein features and variants on a selected transcript. **b**, To map variants on a structure, users can

navigate to ‘variant to protein structure’ from the landing page of the Gene/Protein Lookup module, select a structure, and ‘click to view’, which launches the protein structure viewer. The viewer illustrates a structure of PDB S0F9 with concurrent mapping of ClinVar PLP variants track (yellow) and protein feature; Binding site (black).

The G2P portal distinguishes itself from UniProt feature/variant viewer⁴⁵, RCSB 3D-sequence viewer⁴⁶ and DECIPHER⁴⁸ sequence/structure viewers by simultaneously mapping variants from the largest population variant^{9,12} and multiple disease variant databases^{10,11} as separate tracks alongside protein feature tracks. Additionally, users can customize their analysis by filtering variants by protein consequences (for example, missense and synonymous), clinical significance and AF—a feature essential to differentiate local enrichment (or depletion) of pathogenic (or common, putatively benign) variants on the linear protein sequence and 3D structural region associated with a specific protein function. Another differentiating attribute of the G2P portal is its comprehensive protein feature aggregation, encompassing curated annotations from UniProtKB²¹, PDB/SIFTS³¹, AlphaFoldDB²⁵, MaveDB²⁸, PTM data²⁷ and computed features (for example, solvent accessible area). Unlike existing viewers in UniProt⁴⁵, RCSB PDB⁴⁶ and DECIPHER⁴⁸, the G2P portal can map multiple features, including discrete and quantitative scores concurrently with variants, onto structures. This functionality of the portal enables the identification of mutational hotspots in genes/proteins and their characterization by structural/functional features, which reportedly informs variant prioritization^{51,52}. The G2P portal also offers enhanced data export capabilities; users can seamlessly download mapping results in multiple interoperable formats, including tabular CSV/TSV files for protein sequences and PyMOL-compatible files for structures, making the user experience of downstream analysis easy and efficient.

Furthermore, the growing landscape of natural and synthetic variants, as well as predicted protein structures, underscores the urgent, unmet need for a flexible, dynamic and interactive tool for variant mapping on protein structures that go beyond existing databases. The Interactive Mapping module of the portal allows for this—a capability distinct from existing tools. To compare, ProtVar⁵⁰ allows users to upload variants but links them to protein structures in a

variant-by-variant manner. In contrast, the focus of the G2P portal is to offer uploading of variants by a target gene along with protein residue level score or discrete feature annotations and mapping them concurrently over to the target protein’s sequence and structure. Notably, two features of Interactive Mapping make the G2P portal uniquely versatile: the ability to analyze user-uploaded variants and features (1) in the context of public data and (2) on user-uploaded structures. For example, a clinician with an unreported de novo mutation or a molecular biologist with a set of mutations out of a base-editor screen can upload their data by searching the target gene of interest and interactively investigate their data on the target protein’s sequence and structure, alongside already reported variants from population and clinical databases. Similarly, a structural biologist can upload a previously unsolved structure of a drug target or a structure model predicted by AI methods^{1–4} and can map the known set of disease mutations onto the structure for structure-based rationalization of the impact of disease mutations.

In summary, the G2P portal is an open-source discovery tool for human proteome-wide linking of genetic variants to protein sequences and structures. The versatile interface of the portal is built upon an integrated and dynamic bioinformatic method enabling rapid and efficient investigation of genetic variants with user-defined filters from different databases in the context of proteins, which otherwise requires manual labor and is prone to error. We are committed to growing and maintaining the G2P portal, with regular data updates and documentation of release logs. The future expansion plan includes incorporating cross-isoform and cross-species variant mappings, automating sequence similarity-based structure search for variant mapping, predictions of binding pockets⁵³ and free energy change upon mutation^{54,55}. The data and capabilities integrated into the portal will connect researchers across different fields of biology for a holistic understanding of how genetic variants impact protein structure and

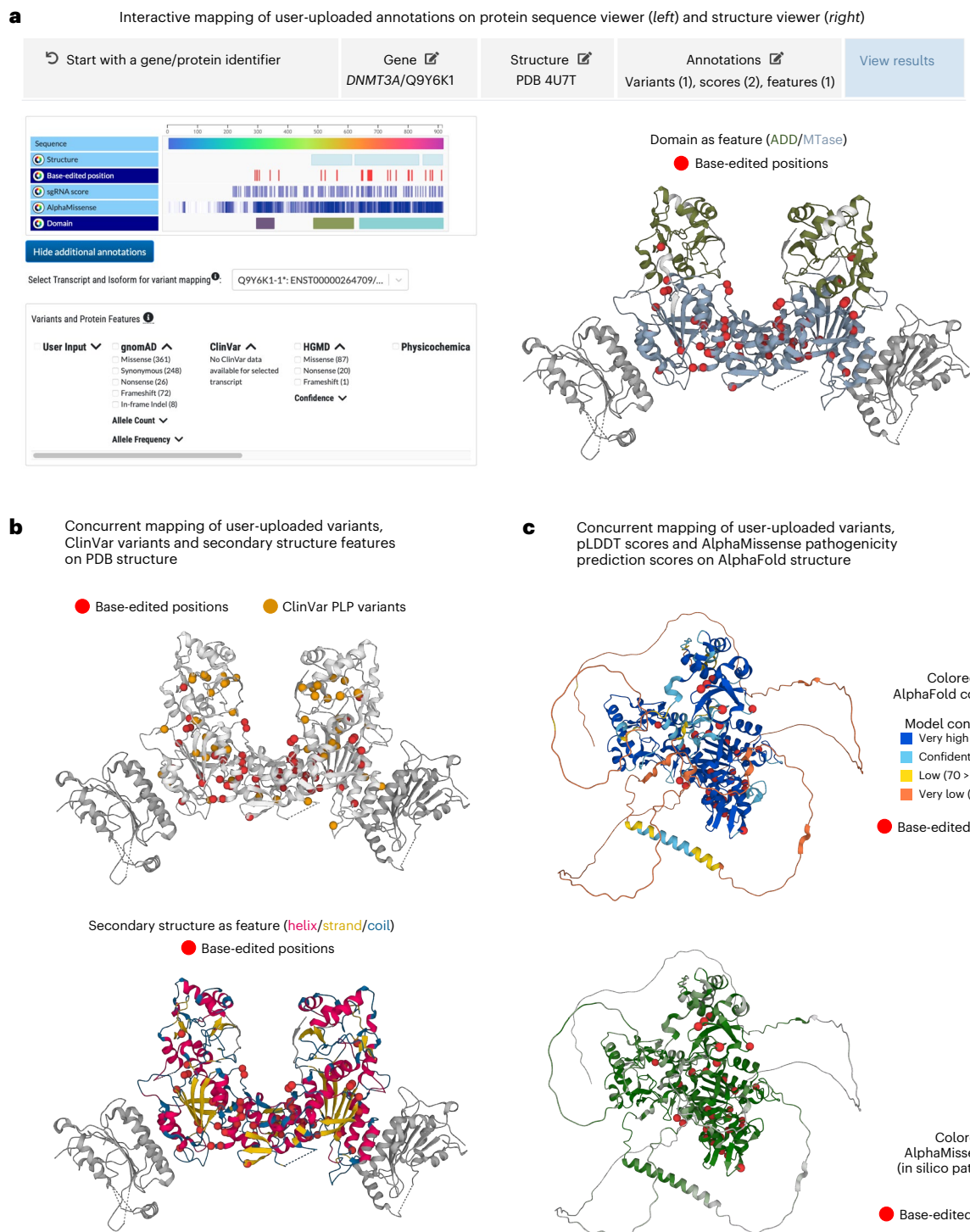


Fig. 6 | A use case of the Interactive Mapping module using *DNMT3A* base-editing screens. **a, The user interface of the Interactive Mapping. From ‘start with a gene/protein identifier’, users are asked to select a gene (*DNMT3A*), structure (PDB 4U7T) and upload annotations (variants, to be shown as spheres; continuous data or scores, to be shown as a heat map; and discrete data or features, to be shown in discrete colors). The selected gene, structure and entered annotations can be edited by going back through the workflow. Finally, in ‘view results’, annotations are visible on sequence (left) and structure (right). The annotation tracks are selectable from the sequence viewer to map specific tracks on the structure. For example, the mapping on the structure viewer (right) is the result of clicking the ‘base-edited position’ and ‘domain’ tracks, where variation data are shown as red spheres and domain annotations are displayed as features in different colors. Colors are editable by the users. **b**, Illustration of the concurrent mapping of user-uploaded variant annotations and data from**

additional G2P-provided resources on the structure (‘Resources in the G2P portal’ in Results). Top, the Base-edited positions (red spheres) and the ClinVar PLP variants (orange spheres) are simultaneously mapped on the structure (MORC2, PDB 7PFP). Bottom, the base-edited positions (red spheres) are displayed in the context of secondary structure annotations (as discrete features) available in the portal. **c**, Illustration of the concurrent mapping of user-uploaded variants, features and scores on the structure. Top, base-edited positions (red spheres) in the context of pLDDT values (four discrete features: very high, confident, low and very low); bottom, user-provided base-edited positions (red spheres) in the context AlphaMissense³⁵ pathogenicity scores (green spectrum) where darker green indicates higher pathogenicity scores. After performing a workflow in the Interactive Mapping, users can download the current mappings as a TSV file (protein residue-wise annotation) and a PyMOL-compatible structure file.

function and thereby facilitate the spectrum of basic biology research, from the translation of genetic discovery into better target selection to drug discovery.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-024-02409-0>.

References

- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
- Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, ead12528 (2024).
- Lin, Z. M. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
- Hekkelman, M. L., Vries, I. D., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).
- Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
- Burley, S. K. et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, gky949 (2018).
- Patwardhan, A. et al. Data management challenges in three-dimensional EM. *Nat. Struct. Mol. Biol.* **19**, 1203–1207 (2012).
- Gudmundsson, S. et al. Variant interpretation using population databases: lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
- Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, gkx1153 (2017).
- Stenson, P. D. et al. The Human Gene Mutation Database (HGMD): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Turner, T. N. et al. denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* **45**, D804–D811 (2017).
- Porto, E. M., Komor, A. C., Slaymaker, I. M. & Yeo, G. W. Base editing: advances and therapeutic opportunities. *Nat. Rev. Drug Discov.* **19**, 839–859 (2020).
- Lue, N. Z. et al. Base editor scanning charts the DNMT3A activity landscape. *Nat. Chem. Biol.* **19**, 176–186 (2023).
- Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
- Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
- Andreadis, A., Gallego, M. E. & Nadal-Ginard, B. Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. *Annu. Rev. Cell Biol.* **3**, 207–242 (1987).
- Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
- den Dunnen, J. T. Describing sequence variants using HGVS nomenclature. in *Genotyping: Methods and Protocols* (eds White S. J. & Cantsilieris S.) 243–251 (Springer New York, 2017).
- Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).
- Seal, R. L. et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2022).
- Hubbard, T. et al. The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
- Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
- Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2021).
- Morales, J. et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
- Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
- Esposito, D. et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).
- Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
- Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
- Dana, J. M. et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).
- Armstrong, D. R. et al. PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* **48**, D335–D343 (2020).
- Schrödinger, L. The PyMOL Molecular Graphics System, version 1.8 (2015).
- Sancho, P. et al. Characterization of molecular mechanisms underlying the axonal Charcot–Marie–Tooth neuropathy caused by mutations. *Hum. Mol. Genet.* **28**, 1629–1644 (2019).
- Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
- Ramos, E. M. et al. Characterizing genetic variants for clinical action. *Am. J. Med. Genet. C Semin. Med. Genet.* **166**, 93–104 (2014).
- Lau, T. K. & Leung, T. N. Genetic screening and diagnosis. *Curr. Opin. Obstet. Gynecol.* **17**, 163–169 (2005).
- Stark, Z. & Scott, R. H. Genomic newborn screening for rare diseases. *Nat. Rev. Genet.* **24**, 755–766 (2023).
- Hoffman-Andrews, L. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *J. Law Biosci.* **4**, 648–657 (2017).
- Carter, T. C. & He, M. M. Challenges of identifying clinically actionable genetic variants for precision medicine. *J. Healthc. Eng.* <https://doi.org/10.1155/2016/3617572> (2016).
- Woodard, J., Iqbal, S. & Mashaghi, A. Circuit topology predicts pathogenicity of missense mutations. *Proteins* **90**, 1634–1644 (2022).
- Iqbal, S. et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl Acad. Sci. USA* **117**, 28201–28211 (2020).
- Iqbal, S. et al. MISCAST: Missense variant to protein Structure Analysis web Suite. *Nucleic Acids Res.* **48**, gkaa361 (2020).
- Costain, G. & Andrade, D. M. Third-generation computational approaches for genetic variant interpretation. *Brain* **146**, 411–412 (2023).

45. Watkins, X., Garcia, L. J., Pundir, S., Martin, M. J. & Consortium, U. ProtVista: visualization of protein sequence annotations. *Bioinformatics* **33**, 2040–2041 (2017).
 46. Bittrich, S. et al. RCSB Protein Data Bank: improved annotation, search and visualization of membrane protein structures archived in the PDB. *Bioinformatics* **38**, 1452–1454 (2022).
 47. Thormann, A. et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun.* <https://doi.org/10.1038/s41467-019-10016-3> (2019).
 48. Bragin, E. et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* **42**, D993–D1000 (2014).
 49. Stephenson, J. D., Laskowski, R. A., Nightingale, A., Hurles, M. E. & Thornton, J. VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics* **35**, 4854–4856 (2019).
 50. Stephenson, J. D. et al. ProtVar: mapping and contextualizing human missense variation. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkae413> (2024).
 51. Hicks, M., Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Functional characterization of 3D protein structures informed by human genetic diversity. *Proc. Natl Acad. Sci. USA* **116**, 8960–8965 (2019).
 52. Iqbal, S. et al. Delineation of functionally essential protein regions for 242 neurodevelopmental genes. *Brain* **146**, 519–533 (2022).
 53. Meller, A. et al. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat. Commun.* **14**, 1177 (2023).
 54. Schymkowitz, J. et al. The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).
 55. Tiberti, M. et al. MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief. Bioinform.* **23**, bbac074 (2022).
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.
- © The Author(s) 2024

Methods

Construction of G2P3D API

We integrated public databases focusing on genes, transcripts and proteins to build an API for seamless mapping of identifiers for genes, transcripts, protein sequences and structures, referred to as the G2P3D API (Fig. 1). The HGNC²² maintains a curated online repository (<https://www.genenames.org/>) of approved genes and their unique symbols and names for human loci. The Ensembl genome browser (<http://useast.ensembl.org/>)²³ offers access to a wide range of genomic annotations. UniProtKB (<https://www.uniprot.org/>)²¹ provides the most current data on protein sequences and functions. These databases each specialize in different aspects of biology and are regularly updated; thus, there would be situations where gene symbols annotated in UniProtKB have been changed or withdrawn in the HGNC, and UniProtKB IDs annotated in the Ensembl browser have been obsolete in the latest release of UniProtKB. To address this issue, G2P3D API has integrated UniProtKB, Ensembl and HGNC to ensure it captures the most comprehensive and up-to-date information on genes, transcripts and proteins.

First, we obtained a list of all human proteins from UniProtKB/Swiss-Prot (indexed by UniProt Accession or UniProtAC) and their corresponding HGNC IDs. Then, we retrieved gene symbols for each protein from HGNC with the provided HGNC ID. Subsequently, all Ensembl and RefSeq transcript identifiers and corresponding UniProtKB protein isoform identifiers were obtained via the Mart View API from Ensembl BioMart⁵⁶ for the human reference genome GRCh38. These data were processed to map each gene symbol (HGNC) to its encoded UniProtAC and then each protein isoform to its corresponding Ensembl and RefSeq transcript when available from Ensembl. Additionally, canonical protein isoform annotations, as defined by UniProtKB, and the canonical Ensembl and MANE Select²⁶ annotation of transcripts were assembled. Next, the PDB⁶ identifiers for the experimentally solved protein structures per protein were obtained using Graph-API (<https://www.ebi.ac.uk/pdbe/graph-api/uniprot/unipdb/:UniProtAC/>) and the identifier for the predicted structure by AlphaFold²⁵ was retrieved using API (<https://alphafold.ebi.ac.uk/api/prediction/:UniProtAC/>). As of October 2023, the G2P3D API, incorporated into the G2P portal (see an example of the API output in Fig. 1b), links 20,292 HGNC genes (Supplementary Table 2) that encode 20,242 UniProtKB/Swiss-Prot human proteins corresponding to 42,413 protein isoforms, via 53,607 Ensembl transcripts and 57,543 RefSeq transcripts, to 77,923 3D protein structures (58,027 experimentally solved and 19,896 computationally predicted).

The G2P3D API is available at <https://g2p.broadinstitute.org/api/gene/:geneName/protein/:UniProtAC/gene-transcript-protein-isoform-structure-map/>. The Swagger user interface for the API and its documentation are available at <https://g2p.broadinstitute.org/api-docs/>.

G2P Google Cloud infrastructure

The schematic of the G2P portal infrastructure is presented in Extended Data Fig. 1. The portal frontend is implemented in React.js, which is served by a Node.js backend running on Google Cloud Platform. The RCSB Saguaro 1D Feature Viewer⁵⁷ and Mol*⁵⁸ are adopted and customized as protein sequence and structure viewers, respectively, to visualize the frontend data on protein sequences and structures. The backend runs on Google App Engine, a serverless and on-demand compute offering that launches a variable number of backend instances proportional to usage.

Google Cloud Storage (GCS) is utilized as the primary data store for variant and protein feature annotations per gene/protein alongside an in-memory datastore used on the backend to track the gene–transcript–protein isoform–protein structure mapping. The static data stored in GCS are collected, processed, formatted and uploaded by the portal admin (Extended Data Fig. 1). To load static data from GCS, the portal requests files directly from the frontend, which reduces

latency by avoiding an additional ‘hop’ where data must first travel to the backend before reaching the frontend. From our testing, the minimum observed time for a backend request is a 60-ms round trip, and by requesting files directly from the frontend, the G2P portal saves a minimum of 60 ms per request. To load data from the in-memory datastore, the portal frontend makes requests to backend APIs, and the backend retrieves and returns the relevant records. The datastore is managed directly by the backend server, not by a separate process. In addition to managed data sources, the portal dynamically requests data from external APIs to provide the most current information possible. The full list of external and internal APIs as well as static and dynamic data maintained in the G2P portal are available in Extended Data Fig. 1.

To this end, the G2P portal web app requests the latest protein sequence and structure records directly from UniProtKB²¹, PDB³², AlphaFoldDB²⁵ and EMBL-EBI APIs⁵⁹. In the ‘Interactive Mapping’ module of the portal, users can provide their data (protein residue-wise annotation of variants, features, scores and protein structures) for joint analysis of user data with G2P-provided resources (‘Resources in the G2P portal’ in Results). The Interactive Mapping module can be securely accessed via Google sign-in, and to further ensure data confidentiality, all user-uploaded data remain within the user’s local browser only; therefore, no user-provided data leaves the user’s device. This ensures that the user has full, secure control over their data while simultaneously providing access to G2P Portal’s variants and protein features for joint analysis. When a user searches a gene or protein via the Gene/Protein Lookup or as part of the Interactive Mapping workflow, static mapping information is fetched directly with the G2P3D API to connect gene to protein to transcript to sequence to structure. Subsequently, detailed gene-specific and protein-specific data are fetched as static data from GCS and dynamic data from external APIs.

G2P portal sitemap

The homepage is the central hub for navigating to two primary modules of the G2P portal: (1) Gene/Protein Lookup and (2) Interactive Mapping, complemented by a top navigation bar featuring tabs for About, Documentation, Statistics, API, Release Logs and Feedback (Extended Data Fig. 2). The disclaimer for using data in the G2P portal is available in the About page. The Statistics page shows the overview of the latest data in the portal. Across the two main modules of the portal, a suite of visualization tools has been implemented for intuitive exploration of the data—protein sequence viewer, variant information and protein feature cards, variant and protein feature tables, protein structure viewer, and mutagenesis output viewer. Details of these viewers are available in ‘Data visualization tools in the G2P portal’ in Methods and Extended Data Fig. 3.

Users can access the Gene/Protein Lookup module by searching for a human gene or protein name. Upon valid input, users are directed to the gene/protein overview page containing the gene family and protein class information for the input gene and a navigation bar with tabs for five submodules, as follows. (1) The ‘protein sequence annotations’ tab hosts a protein sequence viewer that displays a complete list of protein features aggregated within the G2P portal (‘Protein features in the G2P portal’ in Methods). Users can choose a protein isoform identifier from the list of isoforms available for the selected protein, according to UniProtKB²¹. By default, protein features are displayed for the canonical protein isoform. (2) The ‘variant to protein sequence’ tab permits users to select an RNA transcript ID, to map variants from gnomAD⁹, ClinVar¹⁰ and HGMD¹¹ for the selected transcript onto the protein sequence, and displays the mapped variants on the protein sequence viewer on top of protein features (Fig. 5a and Extended Data Fig. 3a). Users can apply filters on variants (different source databases and database-specific filters, for example, AF for gnomAD and pathogenicity for ClinVar) and protein features from an easily (un-)selectable checklist to the left of the sequence viewer. Variant and protein feature data displayed on the protein sequence viewer can also be explored as

a table view and are exportable in CSV and PDF formats. Clicking on a specific variant within the sequence viewer, users can expand the variant and protein feature cards with detailed information on the variant and protein features at the variant position ('Data visualization tools in the G2P portal' in Methods and Extended Data Fig. 3c). (3) Under the 'variant to protein structure' tab, users can find the list of available PDB and AlphaFold protein structures for the selected gene (Fig. 5b). After selecting a structure, users are directed to the 'structure_map' page, where users can map variants and protein features onto structures and view them in the protein structure viewer, coupled with the sequence viewer (Extended Data Fig. 3b). Both protein sequence and structure viewer support dynamic feature and variant selection as described above. Outputs from the structure viewer are exportable in PyMOL-compatible formats. (4) The 'gene to transcript to protein isoform mapping' tab provides a table view of the mapping of identifiers across gene, transcript and protein sequences, downloadable in TSV format. The canonical protein isoform according to the UniProtKB, the canonical transcript in terms of Ensembl and the MANE Select transcript for the input gene are indicated in the table. (5) The 'additional resources' tab offers links to external gene information, such as UCSC⁶⁰, ChEMBL⁶¹, DrugBank⁶², Orphanet⁶³ and OMIM⁶⁴. Moreover, the portal integrates MAVE data from MaveDB for 40 genes²⁸ (Supplementary Table 7). When available, the 'additional resources' tab displays the MAVE data (that is, mutagenesis scores) as heat maps. Additionally, the portal shows the title, description and a short method text describing the MAVE assay. The raw JSON files of scores are available to download alongside a hyperlink to the original source of data.

In the 'Interactive Mapping' modules, users can start their exploration from either a gene/protein identifier or their own protein structures (respective case studies are presented in Fig. 6 and Supplementary Fig. 11). When starting with a gene/protein identifier, users can provide their target gene of interest as input and then choose a structure (PDB or AlphaFold structure). The portal retrieves the protein sequence and the list of available structures dynamically from the UniProt sequence API and PDB/AlphaFold APIs, respectively. Alternatively, users can start with their own protein structures can upload them in PDB format. In both scenarios, the final step prompts a window for annotations, providing a sample format and allowing users to enter their annotations (variants, scores or features). The resulting data are displayed in the 'view results' section (Fig. 6a), featuring both sequence and structure viewers. When starting with a gene/protein identifier, users can also append additional feature annotations, such as protein features and variants, corresponding to the selected transcript or protein isoforms, and map them simultaneously with the user-uploaded data on protein sequences and structures.

Data visualization tools in the G2P portal

Protein sequence viewer. We adopted the RCSB Saguaro 1D Feature Viewer⁴⁴ and customized it for online visualization of variants and protein features mapped onto the protein sequence with dynamic applications of filters on variants and protein features, referred to henceforth as the 'protein sequence viewer' (Extended Data Fig. 3a). The protein sequence viewer in the G2P portal is highly flexible. Variants and features are grouped under collapsible and expandable headers according to variant databases and feature groups and can be easily filtered in and out from the sequence viewer according to AF or pathogenicity criteria (see 'Resources in the G2P portal' in Results and 'Protein features in the G2P portal' in Methods for further details on variants and features integrated in the G2P portal). Users can download the customized mapping data as residue-wise annotations in CSV or PDB format. For example, Extended Data Fig. 3a shows the mapping of *CBS* gnomAD missense variants with the filter 'singleton' and ClinVar missense variants with the filter Pathogenic/Likely pathogenic, in the context of UniProt sequence features alone and other protein features collapsed for clarity.

Protein structure viewer. We integrated the Mol* protein structure viewer³⁸ to visualize variants, protein features and scores on protein structures, simultaneously with protein sequence (Extended Data Fig. 3b). Users can map three types of data from sequence to structure: variants (mutation positions, as spheres), scores (continuous variable, as a heat map) and multiclass features (discrete/categorical variable discretely colored by category). Users can map, review and recolor features as desired, and apply data filters concurrently. For example, a user can filter *CBS* ClinVar missense PLP variants (orange spheres) and gnomAD synonymous singletons (green spheres) and map them concurrently with the domain annotation (light blue) from UniProtKB on the protein structure (Extended Data Fig. 3b). In the Interactive Mapping module, users can map user-uploaded annotations on the structure and can further add variant and feature annotations from available databases, to inspect user-uploaded data in the context of existing data.

The structure viewer is interconnected with the sequence viewer; when a user hovers over residues in sequence, they are highlighted in the structure, and vice versa. The G2P portal is dynamically linked with and loads structures from the PDB⁶ and AlphaFold²⁵. Many AlphaFold structures show high-confidence structured domains surrounded by low-confidence regions, which challenge users to analyze the structure by obscuring structured regions and globular domains. As such, the structure viewer provides additional functionality, allowing users to hide residues on AlphaFold structures based on the AlphaFold confidence of the structure (pLDDT). To export data for subsequent analysis, the structure viewer allows users to download structures and all accompanying features in a prepared PyMOL file, which includes user-uploaded and the G2P portal-provided features as annotations in the PyMOL session.

Variant and protein feature table. Users can view per-residue annotation of variants and protein features per gene (or protein) by clicking 'view as table' on top of the protein sequence viewer (Extended Data Fig. 3a). For gnomAD variants, the table includes the HGVS annotation of variants (HGVS_p, HGVS_c), AC and frequency information, homozygote count, and so on (for example, see <https://g2p.broadinstitute.org/table/LDLR/P01130-1/ENST00000558518/missense/>). For ClinVar variants, the details include genomic and protein consequences, ClinVar variation type, and other clinically relevant information as available in ClinVar (for example, clinical significance, phenotypes and review status; for example, see https://g2p.broadinstitute.org/clinvar/table/LDLR/P01130-1/ENST00000558518/clinvar_single/). Similarly, for HGMD variants, the table lists the variant consequences (genomic and protein), codon change, HGMD confidence and disease annotations (for example, see <https://g2p.broadinstitute.org/hgmdtable/LDLR/P01130-1/ENST00000558518/missense/>). The protein feature table (for example, see <https://g2p.broadinstitute.org/features/LDLR/P01130/P01130-1/>) includes all features described in 'Protein features in the G2P portal' in Methods. Data in these tabular views can be downloaded as machine-readable text files for further usage by users, except for the licensed HGMD professional data. Note that all variant-level information reflects data available in source databases (gnomAD, ClinVar and HGMD) and users are referred to respective databases for the definitions and details of those information.

Variant information and protein feature cards. From the protein sequence viewer, users can click on a variant position to view detailed variant information and protein features for the variant position as summary reports in 'variant information' and 'protein feature' cards, respectively (Extended Data Fig. 3c). These cards include details of a selected variant, which is also available in the 'table view' for the entire gene or protein (as described above in 'Variant and protein feature table'). For example, in the case presented in Extended Data Fig. 3c, users can click on the *CBS* ClinVar missense variant at Gly116 on the

protein sequence viewer, and a card will display below revealing details for variant p.Gly116Arg, such as p.Gly116Arg has been classified as a PLP variant and is associated with homocystinuria. At the same time, the protein feature card shows a summary of five categories of protein features for the residue position Gly116. The summary highlights that the variant p.Gly116Arg substitutes a small, flexible amino acid Gly to a charged amino acid Arg (physicochemical properties), the variant is located at a buried region of the protein structure with an accessible surface area of 7 \AA^2 (structural features), and this missense variant substitutes a known PTM site (PhosphoSitePlus PTMs). Whenever available, each variant and feature information in the cards are linked to their original sources for users to check for any update in the original data source (Extended Data Fig. 3c).

Mutagenesis output viewer. We implemented a mutagenesis output viewer to display the MAVE from MaveDB²⁸, when available (Supplementary Table 7). Users can view MaveDB data under the ‘additional resources’ tab of the Gene/Protein Lookup module (Extended Data Fig. 2). For single missense mutations, a $21 \times N$ heat map is displayed, where N is the range of mutations covered by MAVE perturbations with 21 rows for the 20 different amino acids and the stop codon possible at each position. Each value in the heat map corresponds to the score recorded in the MAVE, or the average of multiple scores if multiple scores were recorded for the same mutation. An example is shown in Extended Data Fig. 3d for *CBS* MAVE readouts collected via DMS-TileSeq at low levels of vitamin B₆. Scores show a clear distinction between residues 90 and 390 (low scores in blue) and residues at the N terminus and C terminus (high scores in red). For double mutant MAVEs, where two different residues were perturbed concurrently, an $N \times N$ heat map is displayed where the row and column each represent one of the two residue positions perturbed in the experiment. As with the single missense mutations, the value in the heat map corresponds to the reported score from the mutation or the average of all scores reported for the residue pair. Different MAVEs utilizing different techniques have different score scales and scores that require interpretation in the context of the methodology used by the corresponding MAVE. To this end, the G2P portal includes a brief description of the experimental technique and scoring methodology of the paper, as provided by MaveDB, and additional links to the score set page in MaveDB and the associated publication such that users can best understand the experimental conditions under which any specific score of interest was collected. To facilitate deeper analysis, the portal includes a downloadable JSON file with all coding and noncoding variants from MAVE data.

Variant aggregation

We downloaded raw VCF files (<https://gnomad.broadinstitute.org/downloads/>) for genome and exome datasets from gnomAD⁹ v2.1.1 and selectively extracted variants that passed all variant filters for quality control (filter = ‘PASS’ flag) and possessed valid HGVS annotation. When the same variant was identified in both genome and exome datasets, we summed the AC and the sample count, subsequently calculating the merged AF value. Variant data from ClinVar¹⁰ (October 2023 release) was downloaded directly from the FTP site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variants_summary_txt.gz). Variants were filtered based on the reference genome GRCh38 and valid HGVS annotation. From HGMD professional release (version 2023.1)¹¹, variants on GRCh38 and with disease-causing state (variantType = ‘DM’ flag, indicating disease mutation) were extracted. Among those variants, we collected variants that have a valid HGVS annotation retrieved from Ensembl Variant Effect Predictor⁶⁵ REST API (https://rest.ensembl.org/vep/human/hgvs:hgvs_notation/). Variants were excluded under the following conditions: (1) reference or altered amino acids are not 20 natural amino acids, or (2) the gene was not included from a list of genes from the G2P3D API, which contains only a reliable set of genes both present in HGNC²² and UniProtKB²¹ databases. The resulting variant

aggregation spans 18,014,632 gnomAD variants, 1,749,628 ClinVar variants and 312,783 HGMD variants mapped on protein sequences.

Variant and feature mapping onto proteins

Genetic variants are annotated on the transcript; for example, variants sourced from gnomAD⁹ are annotated on Ensembl²³ transcripts (ENST-), and those from ClinVar¹⁰ and HGMD¹¹ are annotated on RefSeq²⁴ transcripts (NM-). Each variant aggregated from the databases was linked to its corresponding protein isoform IDs using the in-house G2P3D API (Fig. 1b) and then mapped onto its amino acid position upon fetching the protein sequence using UniProt REST API (<https://rest.uniprot.org/uniprotkb:UniProtAC.json/>). Variants were mapped to both canonical and noncanonical protein sequences but only to structures of canonical protein sequences. Finally, proteins’ functional and structural features were annotated onto variant positions at the protein level (‘Protein Features in G2P portal’). The predicted structures cover the full-length protein sequences; however, the experimental structures often cover only parts of the protein and have gaps. We used the polymer_coverage API (https://www.ebi.ac.uk/pdbe/api/pdb/entry/polymer_coverage/:pdbid) to map experimental structure coverage to the sequence space for each chain. We then mapped protein residue positions from sequence to structure and consequently transferred the variants (that is, protein consequence positions) to protein structures, leveraging Mol*⁵⁸ functionality to properly align variants to positions before and after gaps. We found some limitations with polymer_coverage API and Mol* coverage detection, for example, a gap in a PTEN structure, PDB 5BUG (that is, missing region in crystallographic structure), is incorrectly reported in the API response and incorrectly aligned in the Mol* software.

Protein features in the G2P portal

The G2P portal provides a comprehensive set of protein features on both protein sequences and structures, which include physicochemical properties of amino acids, sequence annotations collected from external databases such as UniProtKB²¹ and PhosphoSitePlus²⁷, 3D structural features collected from PDB⁶ and AlphaFold²⁵ and readouts from the MAVEs when available in MaveDB²⁸.

- (1) The physicochemical properties of reference amino acids: The 20 natural amino acids are grouped into six categories based on physicochemical properties of their side chain R-groups; (i) Aliphatic—alanine (Ala/A), isoleucine (Ile/I), leucine (Leu/L), methionine (Met/M) and valine (Val/V); (ii) Aromatic—phenylalanine (Phe/F), tryptophan (Trp/W) and tyrosine (Tyr/Y); (iii) Polar/neutral—asparagine (Asn/N); glutamine (Gln/Q), serine (Ser/S) and threonine (Thr/T); (iv) Positively charged—arginine (Arg/R), histidine (His/H) and lysine (Lys/K); (v) Negatively charged—aspartic acid (Asp/D) and glutamic acid (Glu/E); (vi) Special—proline (Pro/P; a cyclic side chain and cannot make backbone hydrogen bonds), glycine (Gly/G; does not have a side chain that allows flexibility) and cysteine (Cys/C; a reactive sulfhydryl group -SH in the side chain). In addition to these groupings, the molar mass (g mol^{-1}) and hydrophobicity index (a numerical measure reflecting the hydrophobicity of a side chain—the larger the number is, the more hydrophobic the amino acid) of each amino acid are shown for the protein sequence.
- (2) 3D structural features: The G2P portal provides precomputed annotations on structural features. These features are computed based on AlphaFold-predicted structures, aiming for extensive coverage. Secondary structures of amino acids refer to the local 3D conformations of the polypeptide backbone. DSSP³⁰ (Define Secondary Structure of Protein) is the standard tool for determining secondary structure by classifying each residue into a three-class structure (H, helix; B, β -sheet/strand; C, loop/coil) or a nine-class structure (G, 3_{10} -helix; H, α -helix; I, π -helix;

P, polyproline helix; B, isolated β -bridge; E, parallel β -sheet; S, bend; T, turn; C, loop/coil). We utilized DSSP to annotate both three-class and nine-class secondary structures on AlphaFold structures. When experimental structures are available (for example, from PDBe/SIFTS³¹), we provide PDBe/SIFT secondary structures, which are derived from experimental structures, in a separate track. Additionally, DSSP calculates the accessible surface area (in \AA^2) and the backbone torsional phi/psi angles (in degrees) for each amino acid position within the context of the protein's 3D structures. Furthermore, we include a per-residue confidence score produced in AlphaFold, known as pLDDT. The score ranges from 0 to 100 and categorizes the confidence as 'very high' (pLDDT > 90), 'high' (pLDDT > 70), 'low' (pLDDT > 50) or 'very low' (pLDDT < 50). Residues are color coded accordingly. It is important to note that residues with very low pLDDT scores may indicate that their structures are disordered in isolation.

- (3) Sequence annotation from UniProtKB: We gathered the sequence annotations that describe various regions, domains, or sites of interest for a protein, elucidating its function, binding, sequence motif, domain/site/region, molecular preprocessing and more. The G2P portal offers 31 selected sequence annotations: active site, binding site, chain, coiled coil, compositional bias, cross-link, disulfide bond, DNA binding, domain, glycosylation, initiator methionine, intramembrane, lipidation, modified residue, motif, mutagenesis, non-adjacent residues, non-standard residue, non-terminal residue, peptide, propeptide, region, repeat, sequence conflict, sequence uncertainty, signal, site, topological domain, transit peptide, transmembrane and zinc finger.
- (4) PTM: PTM refers to the covalent and enzyme-mediated modification of proteins to form mature proteins. We collected amino acid positions of seven different PTM types from the PhosphoSitePlus database: (i) acetylation—addition of an acetyl group; (ii) methylation—addition of a methyl group; (iii) *O*-GlcNAc—addition of *N*-acetylglucosamine, also known as *O*-linked *N*-acetylglucosamine; (iv) *O*-GalNAc—addition of *N*-acetylgalactosamine, also known as *O*-linked *N*-acetylgalactosamine; (v) phosphorylation—addition of a phosphoryl group; (vi) SUMOylation—addition of SUMO protein (small ubiquitin-like modifiers); (vii) ubiquitination—attachment of ubiquitin.
- (5) Readouts from MAVE: MaveDB²⁸ is a public repository dedicated to housing datasets from MAVES. These datasets primarily result from deep mutational scanning or massively parallel reporter assay experiments. When a gene/protein is available in MaveDB (Supplementary Table 7), amino acid positions displaying variants whose effect falls within the top and bottom 99th percentile are highlighted in the protein sequence and structure viewer. The rationale behind displaying only the top and bottom 99th percentile was clarity of visualizing the data, but the full data are displayed as heat maps under the additional resource tab of the Gene/Protein Lookup module and are downloadable in JSON format (Extended Data Figs. 2 and 3d).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All resources in this paper are available on the G2P portal website (<https://g2p.broadinstitute.org/>). The G2P3D API is available at <https://g2p.broadinstitute.org/api/gene/:geneName/protein/:uniprotId/gene-transcript-protein-isoform-structure-map/>. The G2P Protein Feature API is available at <https://g2p.broadinstitute.org/api/>

[gene/:geneName/protein/:uniprotId/protein-features/](https://g2p.broadinstitute.org/api/gene/:geneName/protein/:uniprotId/protein-features/). All APIs are easily accessible through the portal Swagger UI at <https://g2p.broadinstitute.org/api-docs/>. The following databases are accessed by the portal: HUGO Genome Nomenclature Committee (<https://www.genenames.org/>), Ensembl browser (<https://www.ensembl.org/>), NCBI Reference Sequence Database (<https://www.ncbi.nlm.nih.gov/refseq/>), UniProtKB (<https://www.uniprot.org/>), RCSB PDB (<https://www.rcsb.org/>), AlphaFold Protein Structure Database (<https://alphafold.ebi.ac.uk/>), gnomAD (<https://gnomad.broadinstitute.org/>), ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), the HGMD (<https://www.hgmd.cf.ac.uk/ac/index.php/>), PhosphoSitePlus (<https://www.phosphosite.org/homeAction.action>), MaveDB (<https://mavedb.org/>) and the PANTHER knowledgebase (<https://www.pantherdb.org/>). Source data are provided with this paper.

Code availability

The G2P portal website is available at <https://g2p.broadinstitute.org/>. An open-source Python library and command line tool for data access via the portal APIs is available at <https://github.com/broadinstitute/g2papi/>. Documentation for API access for the G2P3D API is available at <https://g2p.broadinstitute.org/api-docs/>. The portal relies on third-party software including NodeJS v20.11.0 to run the server. ReactJS 17.0.2, Mol* 3.9.1 and rcsb-saguaro 2.5.12 software packages were used for visualization. All data analysis scripts and the Python client library were written using Python, version 3.10.

References

56. Smedley, D. et al. BioMart—biological queries made easy. *BMC Genomics* **10**, 22 (2009).
57. Segura, J., Rose, Y., Westbrook, J., Burley, S. K. & Duarte, J. M. RCSB Protein Data Bank 1D tools and services. *Bioinformatics* **36**, btaa1012 (2020).
58. Sehna, D. et al. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **49**, W431–W437 (2021).
59. Madeira, F. et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
60. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
61. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
62. Wishart, D. S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
63. Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E. E. & Cornel, M. C. Orphanet: a European database for rare diseases. *Ned. Tijdschr. Geneesk.* **152**, 518–519 (2008).
64. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance In Man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).
65. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* <https://doi.org/10.1186/s13059-016-0974-4> (2016).

Acknowledgements

We thank G. Poncet-Montange and A. Wagner for the scientific discussion. We gratefully acknowledge B. Hajian for the scientific illustration of the G2P portal and the PATTERN team at Broad Institute for the website design feedback. This work was supported by a grant from the Merkin Institute for Transformative Technologies in Healthcare (to S.I.). We also acknowledge the support from NIH/NHGRI grants UM1HG011969 and RM1HG010461 and the Australian Government (to A.F.R.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.I. conceptualized the project, designed the study and led the development of the G2P portal. S.K. performed the data analyses. J.S., D.T.N. and D.H. developed the G2P website. J.S., S.K., P.M., J.A.A. and A.F.R. contributed to the data curation. S.K., J.S. and S.I. wrote the manuscript. D.T.N., D.H., P.M., J.A.A., A.F.R., A.J.C. and A.B. reviewed the final manuscript. S.I., A.J.C. and A.B. contributed to the funding acquisition. S.I. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41592-024-02409-0>.

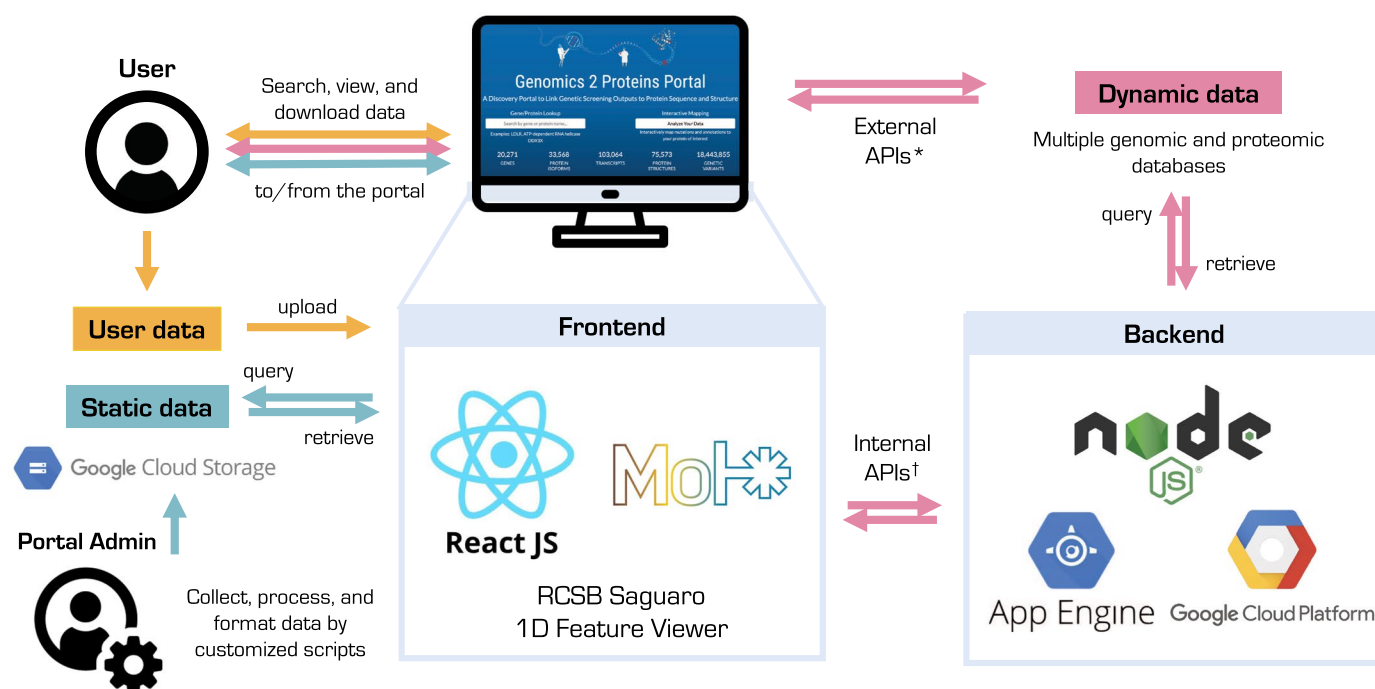
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-024-02409-0>.

Correspondence and requests for materials should be addressed to Sumaiya Iqbal.

Peer review information *Nature Methods* thanks Russ Altman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Genomics 2 Protein Portal infrastructure



† Internal APIs:

List of genes: `/api/genes/options`
 Get gene-wise metadata: `/api/gene/:geneName`
 Get gene-family information: `/api/geneFamily/:familyId`
 G2P3D API: `/api/gene/:geneName/protein/:uniprotId/gene-transcript-protein-isoform-structure-map`
 Protein feature API: `/api/gene/:geneName/protein/:uniprotId/protein-features`

*External APIs:

Get protein sequence per uniprot id: <https://www.ebi.ac.uk/protins/api/protins/:uniprotId>
 Get protein sequence per gene name: <https://rest.uniprot.org/uniprotkb/search?format=fasta&query=:geneName>
 Get PDB structures: <https://www.ebi.ac.uk/pdbe/graph-api/uniprot/unipdb/:uniprotId>
 Get AlphaFold structure: <https://alphafold.ebi.ac.uk/api/prediction/:uniprotId>
 Get sequence annotations from UniProt: <https://rest.uniprot.org/uniprotkb/:uniprotId.json>
 Get structure coverage for protein chains: https://www.ebi.ac.uk/pdbe/api/pdb/entry/polymer_coverage/:pdbId/chain/:chainId

Static data: variants from gnomAD, ClinVar, HGMD, protein features from PhosphoSitePlus, MaveDB data, pre-computed structures features (secondary structures, accessible surface area, dihedral angles, etc.) using DSSP based on AlphaFold structures.

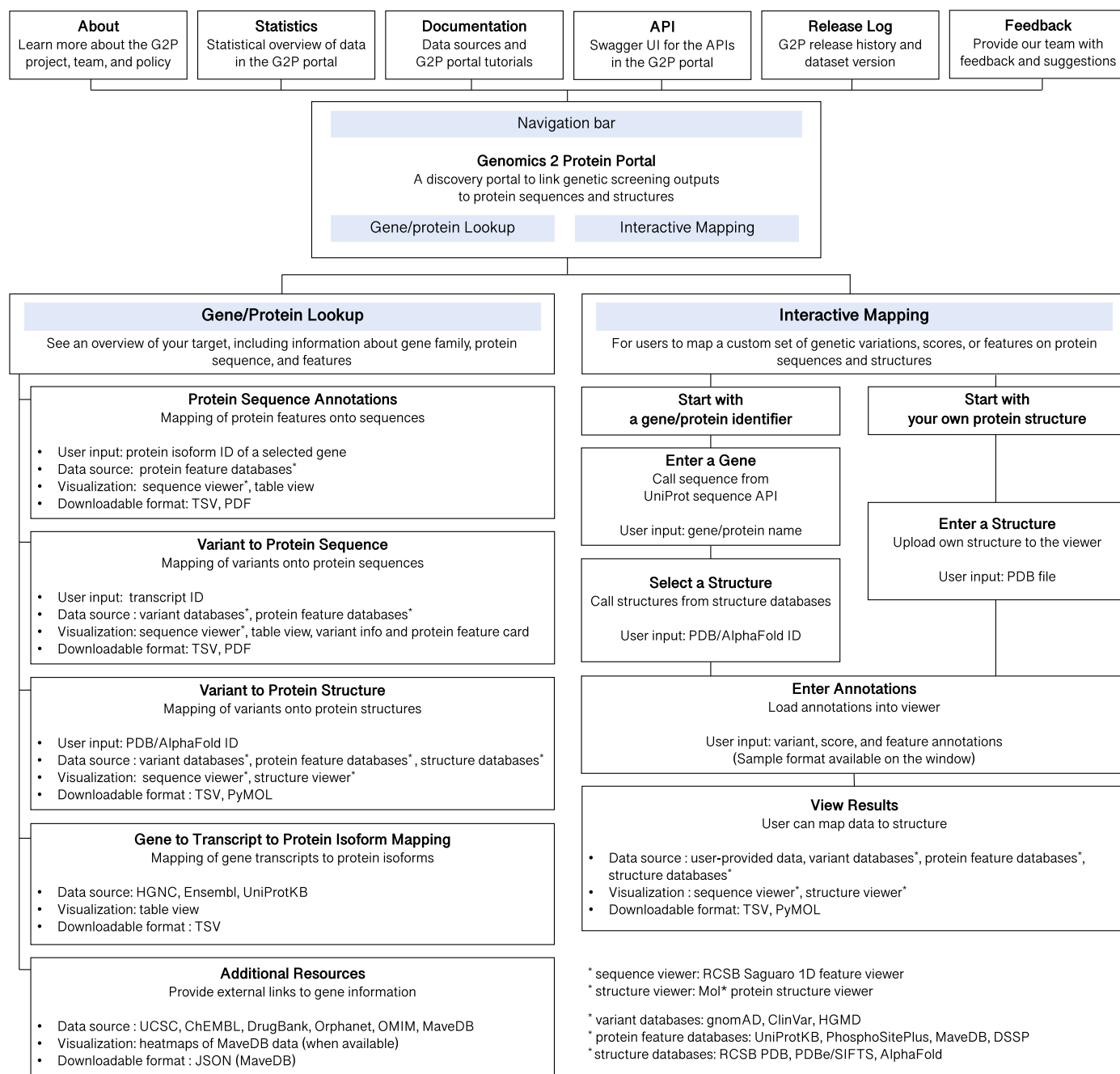
Dynamic data: protein sequences from UniProtKB, protein features from UniProtKB, structural features from SIFTS/PDBe, structures from PDB and AlphaFold

Extended Data Fig. 1 | The Google Cloud infrastructure of the G2P portal.

This figure illustrates the web implementation of the portal. The frontend is implemented in React.js and includes a customized version of RCSB Saguaro 1D Feature Viewer and Mol* as protein sequence and structure viewer, respectively. The backend is implemented in Node.js and uses the Google app engine. Users

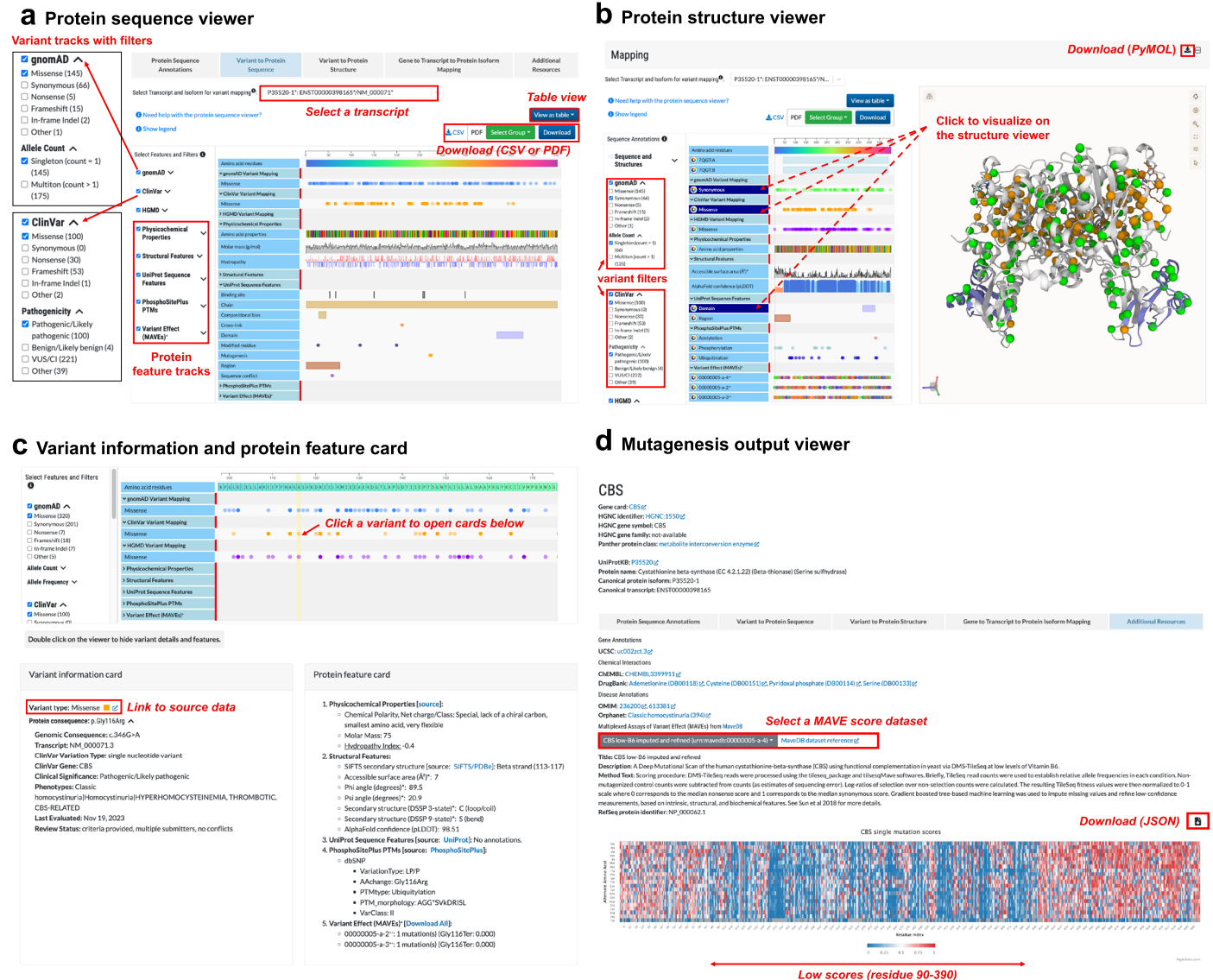
can query, upload, and retrieve data from the portal, and the flow of user-uploaded, static, and dynamic data is shown with arrows in different colors (user-uploaded data in orange, static data in cyan, and dynamic data in pink). All static data are stored in Google cloud storage. All user-uploaded data remain on the users' browser, securing the confidentiality of users' data.

Genomics 2 Protein Portal Sitemap

**Extended Data Fig. 2 | The sitemap of the Genomics 2 Proteins (G2P) portal.**

From the home page, users can access the About, Documentation, Release Logs, APIs in the portal, and Feedback pages, available on the navigation bar at the top of the portal. There are two main modules in the portal: (1) *Gene/Protein Lookup*, accessible via searching by a human gene or protein name; (2) *Interactive Mapping*, accessible via secure Google sign-in upon clicking on the button displayed on the home page. The Gene/Protein lookup module has

five submodules for *protein sequence annotation*, *variant mapping to protein sequence*, *variant mapping to protein structure*, *gene to transcript to protein isoform mapping*, and links to *additional resources*. The *Interactive Mapping* module has two submodules, for allowing users to start with any human gene or a protein structure to map user-uploaded data onto the target protein's sequence and structure. The user input, data sources, visualization methods, and downloadable data formats within each submodule are listed in the figure.



Extended Data Fig. 3 | Data visualization tools in the G2P portal. (a) Protein sequence viewer. This viewer displays protein residue-wise variants and protein features for the selected gene and transcript. Variants can be filtered based on protein consequences and database-specific filters. Data displayed within the viewer can be exported in tabular format (*View as table* button) and downloaded as CSV or PDF formats (*Download* button). The figure shows gnomAD missense (singletons; *blue*) and ClinVar missense (pathogenic/likely-pathogenic; *orange*) for gene *CBS* and transcript NM_000071 along with residue-wise physicochemical properties and UniProt sequence annotations in the protein sequence viewer. **(b) Protein structure viewer.** In the G2P portal, the structure viewer is coupled with the sequence viewer to interactively map variants and protein features on the structure viewer onto the structure. Users can click a track to select variants or features from the sequence viewer to visualize on the structure viewer. Users can download the customized mapping results in a PyMOL-compatible file. The figure displays the concurrent mapping of gnomAD synonymous singleton

variants (*green* spheres), ClinVar missense pathogenic/likely pathogenic variants (*orange* spheres), and the *Domain* annotation from UniProtKB (*light blue*) mapped on the structure (PDB: 7QGT) **(c) Variant information and protein feature cards.** These cards provide a per-variant summary of variant details and protein features for the variant position (see **Methods**: Data visualization tools in the G2P portal, for details). The example in this figure shows the details of *CBS* variant Gly116Arg from ClinVar and the physicochemical, structural, and functional features for the variant position, Gly116. The variant and features are linked to their sources, whenever available. **(d) Mutagenesis output viewer.** This viewer shows the mutagenesis readouts, when available in MaveDB²⁸ for a gene as a heatmap. By hovering over the heatmap, users can view the readouts from the assay and can download the entire score set by clicking the download icon. The figure highlights the residues 90-390 with a differentiating mutagenesis readouts compared to the rest of the protein.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data in this paper are available at the Genomics 2 Protein portal website (<https://g2p.broadinstitute.org/>). The G2P3D API is available at <https://g2p.broadinstitute.org/api/gene/:geneName/protein/:uniprotId/gene-transcript-protein-isoform-structure-map>. The G2P Protein Feature API is available at <https://g2p.broadinstitute.org/api/gene/:geneName/protein/:uniprotId/protein-features>. All APIs are easily accessible through the portal swagger UI at <https://g2p.broadinstitute.org/api-docs/>. An open-source Python library and command line tool for data access via the portal APIs is available at <https://github.com/broadinstitute/g2papi/>.

The following databases are accessed by the portal: HUGO Genome Nomenclature Committee (<https://www.genenames.org/>), Ensembl browser (<https://www.ensembl.org/>), NCBI Reference Sequence Database (<https://www.ncbi.nlm.nih.gov/refseq/>), UniProtKB (<https://www.uniprot.org/>), RCSB Protein Data Bank (<https://www.rcsb.org/>), AlphaFold Protein Structure Database (<https://alphafold.ebi.ac.uk/>), gnomAD (<https://gnomad.broadinstitute.org/>), ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), The Human Gene Mutation Database (<https://www.hgmd.cf.ac.uk/ac/index.php>), PhosphoSitePlus (<https://www.phosphosite.org/homeAction.action>), MaveDB (<https://mavedb.org/>), and the PANTHER knowledgebase (<https://www.pantherdb.org/>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	<input type="text" value="n/a"/>
Reporting on race, ethnicity, or other socially relevant groupings	<input type="text" value="n/a"/>
Population characteristics	<input type="text" value="n/a"/>
Recruitment	<input type="text" value="n/a"/>
Ethics oversight	<input type="text" value="n/a"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<input type="text" value="Sample sizes are the total number of variants available from the gnomAD, ClinVar, and HGMD databases at the time of the analysis. G2P aggregated 18,014,632 protein-coding variants annotated in 18,723 human genes from gnomAD v2.1.1, 1,749,628 protein-coding variants in 18,180 genes from ClinVar (10/2023 release), and 312,738 disease-causing mutations in 12,918 genes from HGMD professional 2023.01."/>
Data exclusions	<input type="text" value="No data was excluded from the analysis."/>
Replication	<input type="text" value="No experiments were conducted as a part of this study, and results reflect a reproducible analysis of variant data from the listed databases."/>
Randomization	<input type="text" value="Randomization was not applicable because this is an analysis of existing databases."/>
Blinding	<input type="text" value="Blinding was not applicable because this is an analysis of existing databases."/>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- n/a | Involved in the study
- Antibodies
 - Eukaryotic cell lines
 - Palaeontology and archaeology
 - Animals and other organisms
 - Clinical data
 - Dual use research of concern
 - Plants

- n/a | Involved in the study
- ChIP-seq
 - Flow cytometry
 - MRI-based neuroimaging

Plants

Seed stocks

n/a

Novel plant genotypes

n/a

Authentication

n/a