Functional characterization of 3D protein structures informed by human genetic diversity

Michael Hicks[a,1], Istvan Bartha[b,1], Julia di Iulio[c], J. Craig Venter[d], and Amalio Telenti[c,2]

[a]Human Longevity, Inc., San Diego, CA 92121; [b]École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [c]The Scripps Research Institute, La Jolla, CA 92037; and [d]J. Craig Venter Institute, La Jolla, CA 92037

Edited by Jean-Laurent Casanova, The Rockefeller University, New York, NY, and approved March 25, 2019 (received for review December 12, 2018)

Sequence variation data of the human proteome can be used to analyze 3D protein structures to derive functional insights. We used genetic variant data from nearly 140,000 individuals to analyze 3D positional conservation in 4,715 proteins and 3,951 homology models using 860,292 missense and 465,886 synonymous variants. Sixty percent of protein structures harbor at least one intolerant 3D site as defined by significant depletion of observed over expected missense variation. Structural intolerance data correlated with deep mutational scanning functional readouts for PPARG, MAPK1/ERK2, UBE2I, SUMO1, PTEN, CALM1, CALM2, and TPK1 and with shallow mutagenesis data for 1,026 proteins. The 3D structural intolerance analysis revealed different features for ligand binding pockets and orthosteric and allosteric sites. Large-scale data on human genetic variation support a definition of functional 3D sites proteome-wide.

protein structure | genome constraint | exome | deep mutational scanning

Recent large-scale sequencing projects of the human genome and exome detail the extent of genetic diversity in the human population (1–3). To date, there are over 4.5 million amino acid-changing (missense) variants reported in the human exome. Much attention has been directed to the association of variants with disease (3, 4). However, these data also represent an unprecedented opportunity to characterize protein structure–function relationships in vivo. In particular, the pattern of distribution of genetic variants describes the functional limits to structural and functional modifications for a given protein. Inference of critical 3D sites could also be informative for drug development and mechanisms of action, including selectivity, lack of response, and toxicity.

Finding important sites within these structures has been done through a variety of methods. Genetics-based scoring metrics can measure the deleteriousness of genetic variants in a protein, a property that strongly correlates with both molecular functionality and pathogenicity (5, 6). Scores may also consider interspecies conservation (7) to discover "constrained elements" indicative of putative functional elements. Previous approaches have emphasized gene-level features (e.g., essentiality, burden of variation) and linear analyses of variation in a gene rather than the distribution of variants in 3D space. However, additional methods have been created in the field of cancer to assess the clustering of somatic variants in protein structures. Ryslik et al. (8–11) described Identification of Protein Amino acid Clustering (iPAC), Spatial Protein Amino acid Clustering (SpacePAC), Graph Protein Amino acid Clustering (GraphPAC), and Quaternary Protein Amino acid Clustering (QuartPAC). Fujimoto et al. (12), Tokheim et al. (13), and Meyer et al. (14) analyzed 3D position and clustering of mutations using exome sequence data from The Cancer Genome Atlas (TCGA) from up to 7,215 samples and 23 types of cancer and over 975,000 somatic mutations. A comparison of algorithms for the detection of cancer drivers at subgene resolution was just published (15). It should be noted that scoring methods in oncology emphasize mutational clustering, as critically relevant in cancer biology, and not intolerance to variation in the human proteome at large.

Recent sequencing efforts of human genomes and exomes identify several hundreds of thousands of missense variants, which can be used to derive human-specific intolerant sites when aggregated in 3D space (3, 6, 16). Most studies that analyze the relationship between point mutations and experimentally observed 3D protein structures published to date have been limited to individual proteins. Bhattacharya et al. (17) manually analyzed one single nucleotide variant in each of 374 human protein structures to assess the effects of genetic variation on structure, function, stability, and binding properties of the proteins. Arodź and Płonka (18) analyzed a limited set of pairs of proteins of the same length differing by a single amino acid. Recently, Sivley et al. (19) presented a comprehensive analysis of the spatial distribution of missense variants in the human proteome. They identified 215 proteins with significant spatial constraints on the distribution of disease-causing missense variants in protein structures. Glusman et al. (20) reported on a workshop titled "Gene Variation to 3D (GVto3D)." The overarching goal of the workshop was to provide the framework to advance the integration of genetic variants and 3D protein structures.

Tolerance to Amino Acid Changes in the 3D Space of the Human Proteome

A thorough analysis of the proteome requires a large study population to observe enough genetic variation to allow the detection

Significance

Increasing numbers of human genome population sequences provide new detail on the genetic variability of the human proteome. It is possible to identify proteins that are depleted in genetic variation, and this approach can now be extended to the identification of 3D features and structures that are uniquely intolerant to variation. We speculated that 3D features that are intolerant to variation correspond to privileged functional domains of the protein. We approached this question with sequence data nearly 140,000 individuals with modeling of >8,500 protein structures. In keeping with the hypothesis, structural predictions correlated with experimental functional readouts. We believe that information derived from human variation complements other metrics at the structural level and can serve to inform drug development.

Author contributions: A.T. designed research; I.B. designed the ProtC browser; M.H., I.B., and J.d.I. performed research; J.C.V. contributed key scientific input and resources; M.H., I.B., and A.T. analyzed data; and M.H., I.B., J.C.V., and A.T. wrote the paper.

Conflict of interest statement: M.H. is an employee of Human Longevity, Inc. J.C.V. owns stock in Human Longevity, Inc.

This article is a PNAS Direct Submission.

Data deposition: Final scores and intermediate results from the genome to proteome mapping, including the UniProt–Protein Data Bank pairwise alignments, are available at https://doi.org/10.5281/zenodo.1311198. The source code is available at doi.org/10.5281/zenodo.2628193. An interactive browser is available at protc.labtelenti.org.

[1]M.H. and I.B. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: atelenti@scripps.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820813116/-/DCSupplemental.

Published online April 15, 2019.

of intolerance and tolerance to mutation of spatial neighborhoods. To advance this field, we initiated a study that uses human genetic variation from 138,632 human exomes and genomes and 31,116 X-ray protein structures (corresponding to 4,715 proteins) to model tolerance to amino acid changes in the 3D space. To understand variation in the structural proteome, we first identified structures that fulfilled our inclusion criteria: X-ray crystal structures with a defined resolution and a minimum chain length greater than 10 amino acids. In addition, we mapped 139,535 Uniprot features [a combination of "structure-based" features, composed of helices, strands, and turns, and "all" features, which includes a list of features from the UniProt Knowledgebase (UniprotKB) defined in *Matarials and Methods*] to the structures and extracted a 3D context for each feature defined as the union of the 5-Å-radius spheres around every atom of a feature, hereafter referred to as a 3D site. We identified 860,292 missense variants for these proteins from the analysis of 138,632 individuals' exomes. From these contextualized data, we constructed a model that describes functional constraints in 3D protein structures (*Materials and Methods* section and Fig. 1*A*). The strength of intolerance to missense variation was summarized by the mean of a posterior distribution that accounts for both observed missense variation and expected missense variation at the level of 3D sites (*Materials and Methods* section), termed the three-dimensional tolerance score (3DTS). While we used a 5-Å-radius space to generalize the analysis proteome-wide, the same approach can be applied to scoring whole domains as well or to tailor to the protein of interest. Below, we show the impact of varying the radius space on functional prediction of selected proteins.

We describe the distribution of 3DTS values in Fig. 1*B*. In total, 3,097 (66%) proteins had at least one intolerant 3D site defined at the 20th percentile proteome-wide (3DTS = 0.14). The most intolerant 3D sites corresponded to DNA binding sites, zinc fingers, and intramembrane domains, while the most tolerant 3D sites included nonstandard residues (i.e., selenocysteines), glycosylation sites, and transit peptides. Structural features (helix, turn, strand) showed median 3DTS values close to the proteome-wide median (Fig. 1*C*), which holds true for interspecies conservation (genomic evolutionary rate profiling, GERP++) as well (*SI Appendix*, Fig. S1). The rank correlation of the medians of the different feature types between 3DTS and GERP++ is 0.45.

The precise interpretation of 3DTS values requires the assessment of functional consequences of amino acid changes in intolerant versus tolerant 3D sites. However, a challenge of functional testing proteome-wide is the requirement of cellular assays that are disease and gene relevant, robust, and scalable—a serious limitation that explains that to date, the experimental characterization of all possible missense variants in a mammalian gene [deep mutational scanning (21, 22)] has been limited to a handful of proteins: PPARG (23); MAPK1/ERK2 (24); p53 (25); PTEN and TPMT (26); UBE2I, SUMO1, TPK1, CALM1, CALM2, and CALM3 (27); and two single-protein domains of BRCA1 (the RING domain) and YAP65 (the WW domain) (21, 28). We therefore sought to validate 3DTS against the available functional data for the complete human proteins for which there is comprehensive deep mutation scanning (nine proteins covering ∼2,300 amino acid positions and ∼40,000 mutants). In addition, we evaluated 1,026 proteins with shallow mutagenesis (approximately 2,100 individual experimental mutational data from Uniprot) to show that 3DTS identifies functional mutations as intolerant preferentially.

Functional Readout of 3D Tolerance Scores

To introduce the approach, we first assessed the structure–function relationship for peroxisome proliferator-activated receptor gamma (PPARG). PPARG is a drug target for thiazolidinediones and newer partial PPARG modulators used in the treatment of diabetes (22). PPARG exemplifies the challenge of classifying newly identified variants even in a well-studied protein implicated in disease. In the original work (23), functional interpretation of PPARG variants required the construction of a cDNA library consisting of all possible amino acid substitutions in the protein. The library was introduced into human macrophages edited to lack the endogenous PPARG and stimulated with PPARG agonists to trigger the expression of CD36, a canonical target of PPARG. Sorted CD36+ and CD36− cell populations were sequenced to determine the distribution of each PPARG variant in relation to CD36 activity. We showed good correlation (r^2 = 0.41, P = 2.6E-5) between the 3D sites defined by 3DTS on the structure [Protein Data Bank (PDB) ID code 3DZY] and the functional scores described in Majithia et al. (23). Specifically, both the in vitro and in silico scores identified the DNA-binding and ligand-binding sites as intolerant to missense variation, while the hinge domain reflected increased tolerance to missense variation (Fig. 2*A*). Additionally, Majithia et al. (23) indicated that their transgene library may not have detected all possible functional effects of coding variation, suggesting that the concordance between in vitro and in silico readouts should be interpreted as conservative.

While we use PPARG as an example of the implementation of 3DTS, we also analyzed the other proteins with existing deep mutational scanning data. Fig. 2*B* shows the distributions of Pearson r^2 values for all structures (ranging from 0 to 0.72 for CALM1, 0 to 0.54 for CALM2, 0.02 to 0.33 for ERK2, 0.17 to 0.41 for PPARG, 0.21 to 0.39 for PTEN, 0 to 0.83 for SUMO1, 0.13 to 0.22 for TPK1, 0.09 to 0.17 for TPMT, and 0 to 0.62 for UBE2I) that cover at least 70% of the canonical isoform under four different 3DTS conditions: two different sets of 3D features and two different models of rate variation. Precision–recall curves and average precision for the comparison of deep mutational screen data of 3DTS and the various in silico methods is shown in *SI Appendix*, Fig. S2. EVmutation has the highest average precision (0.75). Importantly, different structures for the same protein differ in the correlation value; the median r^2 and the distributions tend to be large both within and between conditions and genes. These variations could occur for a variety of reasons such as alternative protein interaction partners, different structural coverages of the protein, varied crystallization conditions, etc. We speculate that 3DTS might serve to identify functionally relevant conformations for a given protein; that is, for a protein with multiple available structures, the best correlations may represent the most parsimonious and functionally plausible structures. Data regarding the optimal structures are available in Dataset S1.

We compared the functional prediction of 3DTS with 23 published scores: CADD (5), SIFT (29), PROVEAN (30), FATHMM (31), MutationAssessor (32), fathmm-MKL (33), FitCons (34), DANN (35), MetaSVM/MetaLR (36), GenoCanyon (37), Eigen-PC (38), M-CAP (39), REVEL (40), PhyloP (41), PhastCons (42), GERP++ (7), SiPhy (43), Polyphen-2 (44), and EVmutation (45). Importantly, we bring these scores to the 3D environment, as the purpose of this analysis is the definition of functional regions and not the prediction of deleteriousness at single-amino acid level resolution. These various scores trained under a range of assumptions, most commonly interspecies conservation, coevolution, and pathogenicity. Overall, 3DTS performs comparably to these other methods in the 3D space (Fig. 2*C*). In the future, use of ensemble methods (modeling on multiple scores) is expected to perform better than single scores (for a comparison of all structures and methods, see *SI Appendix*, Fig. S3 and Dataset S2). The diversity and complementarity of the various methods suggest that users should analyze proteins under various assumptions and models. Here, 3DTS adds a dimension that has not been included in previous predictors. The availability of multiple proteins with deep mutational screening data also supported a more formal assessment of the effect of varying the size of the 3D sites and confirming the general validity of the use of the 5-Å radius (*SI Appendix*, Fig. S4).

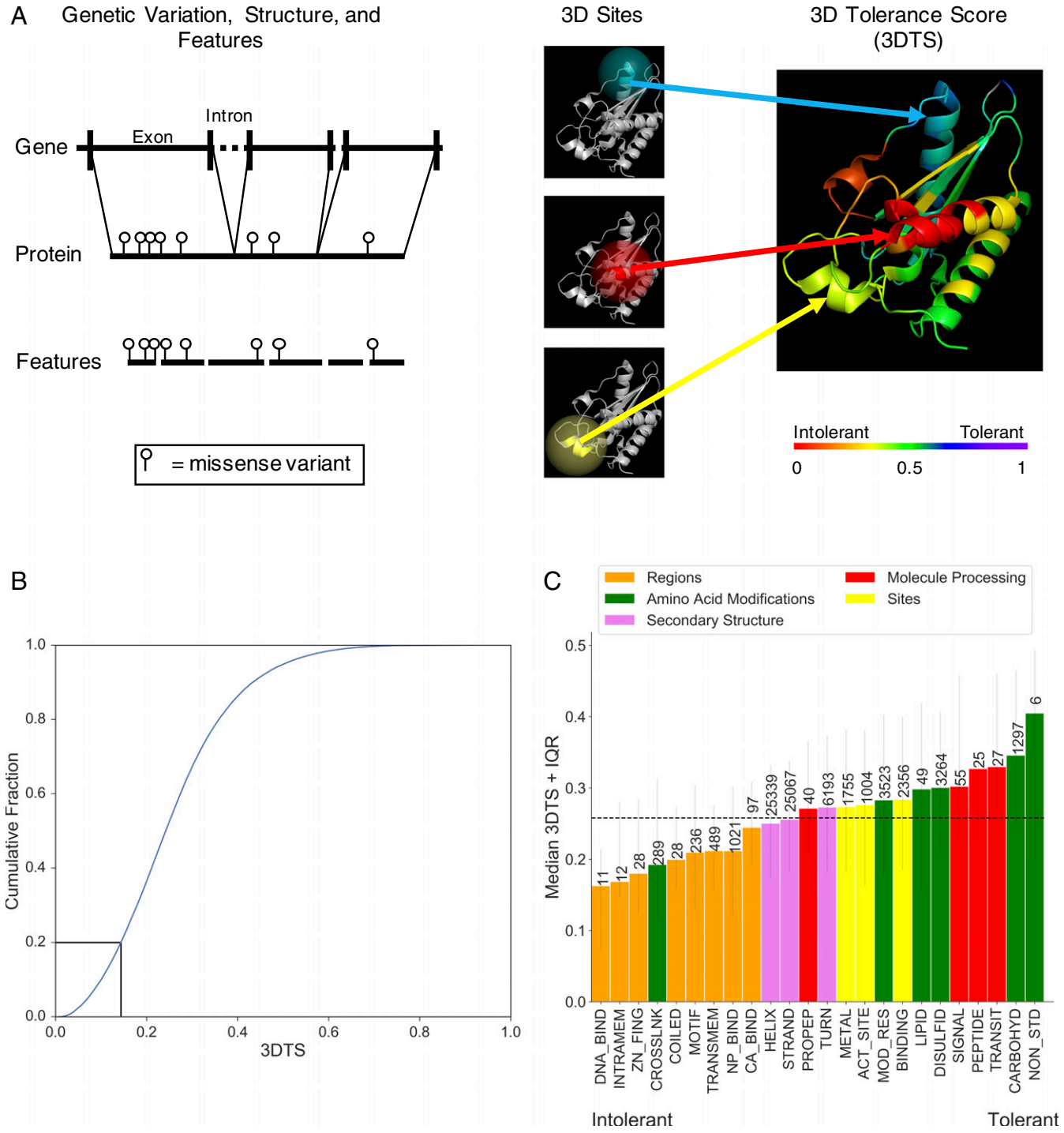

Fig. 1. Three-dimensional tolerance to variation in the proteome. (*A*) Missense variation data from genome and exome sequencing projects are mapped to 3D protein structures. Features extracted from Uniprot are also mapped to the 3D structures. Using these features as reference points, a 3D context is constructed, and the corresponding genetic data are extracted. A 3DTS is generated from this information. The 3DTS values are projected back onto the 3D structure. (*B*) The distribution of tolerance values across the structural proteome for 139,535 3D sites for structures representing 4,715 proteins. The 3DTS value at the 20th percentile (3DTS < 0.14) is used to define intolerant sites. (*C*) Median 3DTS for a subset of feature types with the interquartile ranges (IQR). The number of each feature type with a 3DTS value is shown above each column. The overall median across the structural proteome is represented by a horizontal dashed line. Feature types are colored by subsections defined by Uniprot (https://www.uniprot.org/help/sequence_annotation).

We then extended the evaluation to a large corpus of functional readouts for 1,026 proteins for which shallow mutational information was available. The median 3DTS score for 4,428 3D functional sites (those that carry an experimentally tested "loss of function" variant) is lower than the proteome background (Kolmogorov–Smirnov two-sided test P value = 3.7E-42), which may yet include undescribed functional sites. Importantly, at any level of global gene essentiality, functional sites are systematically more constrained than the rest of the protein (Fig. 2*D*). In summary, the in silico 3DTS values may provide functional prediction

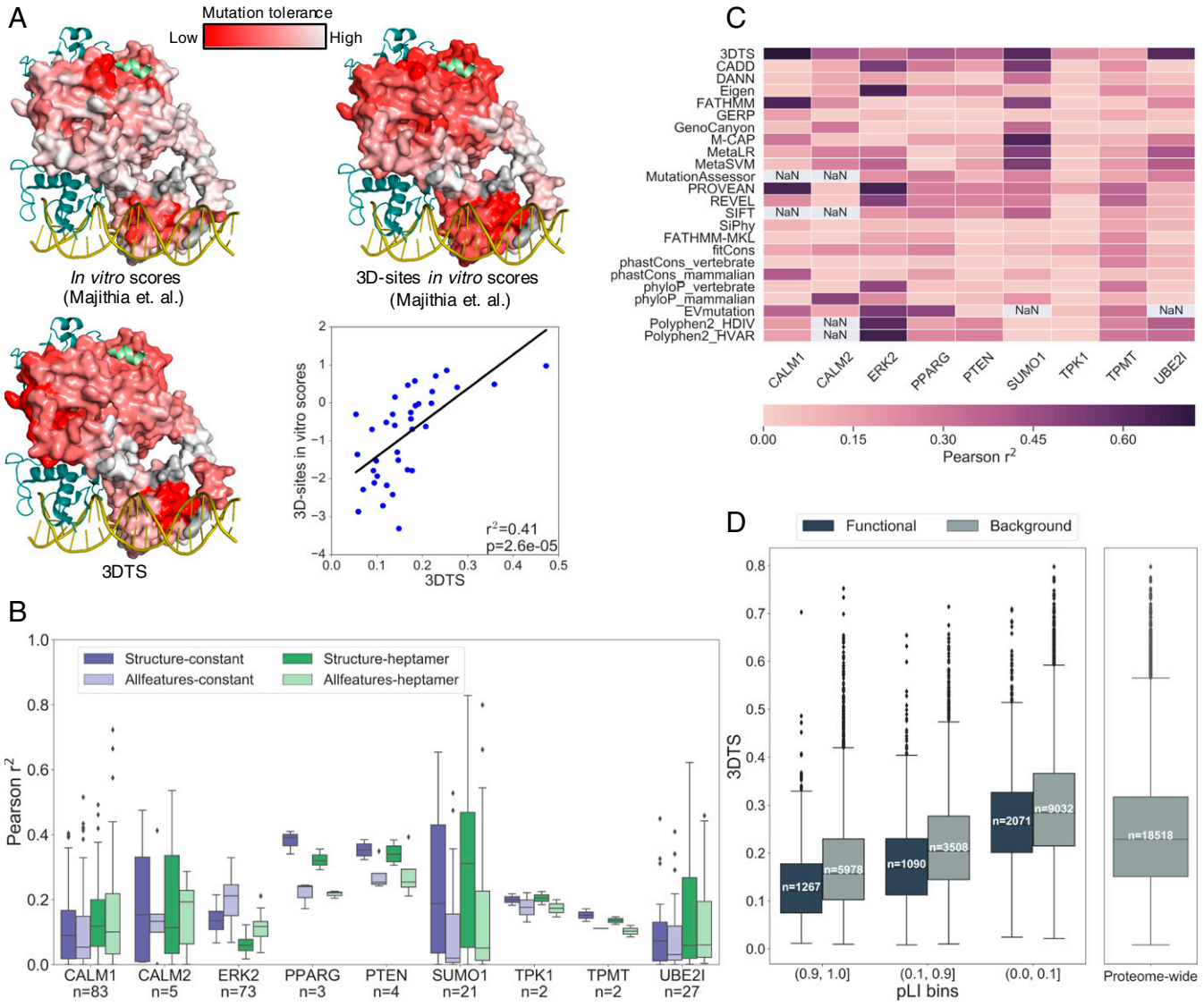

Fig. 2. Validation of 3DTS. (*A*) Comparison of deep mutational screen data and in silico 3DTS data for the DNA-binding and ligand-binding domains of PPARG. (*Top*) Projection of the functional scores described in Majithia et al. (23) for each amino acid and the scores averaged across the 3DTS-defined sites for the crystal structure 3dzy (32). The color scheme is chosen to match the one described in Majithia et al. (*Bottom*) A projection of 3DTS onto PPARG is seen on the *Left*, and the 3D site level correlation between 3DTS and the 3D site averaged in vitro functional scores is shown in the plot on the *Right*. (*B*) Comparison of deep mutational screen data and 3DTS under different modeling assumptions for all available PDB structures covering 70% of the canonical protein length for nine genes. "Structure" refers to 3D sites defined by secondary structure elements, and "Allfeatures" uses 3D sites defined by all Uniprot features as detailed in the *Materials and Methods*. "Constant" and "heptamer" refer to the mutation rates as discussed in the *Materials and Methods*. (*C*) Comparison of the optimal 3DTS model to 23 other scoring methods at the 3D site level for nine genes. Pearson r^2 values for comparisons of deep mutational screen data and in silico data at the 3D site level for the nine genes are provided. "NaN" refers to methods with unavailable scores. (*D*) Shallow mutagenesis data proteome-wide. Here, 3DTS identifies functional sites (loss of function) as more constrained (lower 3DTS values) at all levels of global gene essentiality compared with the rest of the protein. pLI > 0.9 (essential gene) functional to background Kolmogorov–Smirnov two-sided test *P* value = 9.3E-31; 0.1 > pLI > 0.9 functional to background Kolmogorov–Smirnov two-sided test *P* value = 2.3E-20; pLI < 0.1 functional to Kolmogorov–Smirnov two-sided test *P* value = 1.1E-18.

without engaging in extensive and time-consuming in vitro assays and dedicated functional readouts; this is critical given the paucity of human proteins that have been subjected to deep mutational scanning and functional testing.

Three-Dimensional Tolerance to Amino Acid Change of Drug Target Sites

One application of the present work could involve prioritization of drug target sites. Protein structure-based methods are now routinely used at all stages of drug development, from target identification to optimization (46). Central to all structure-based discovery approaches is the knowledge of the 3D structure of the target protein or complex because the structure and dynamics of the target determine which ligands it binds (46). The characterization of human-specific intolerant sites and tolerance to genetic variation can be used to parse structural information to define active sites and also to define functionally important topographically distinct sites that can support allosteric interactions for small molecules to modulate protein function (47). We analyzed the 3D intolerance characteristics for 97 proteins that included known drug targets with a bound ligand and proteins with known allosteric sites (Dataset S3). The corresponding proteins carried a median number of one unique nonoverlapping intolerant 3D site (range 0–7). Overall, 17 proteins lacked an intolerant site, while 26 had more than one unique intolerant site. In

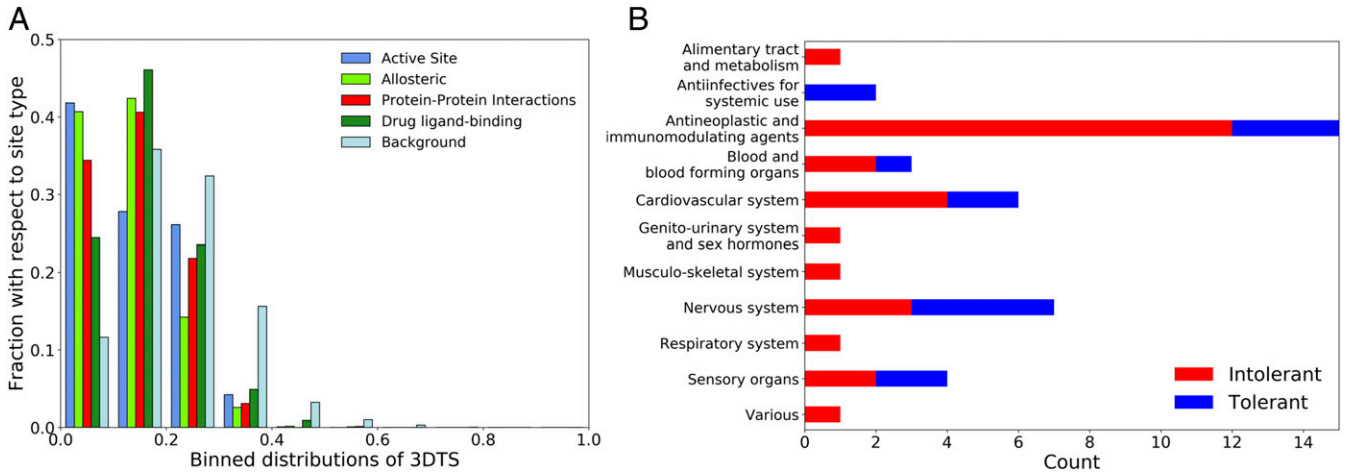

Fig. 3. Characteristics of druggable sites. (*A*) Binned 3DTS scores describing active sites, allosteric sites, protein–protein interaction sites, drug ligand-binding sites, and background. The sum of each site type is 1. Active-site background Kolmogorov–Smirnov two-sided test *P* value = 4.9E-110. Allosteric background Kolmogorov–Smirnov two-sided test *P* value = 1.1E-84. Protein–protein interactions background Kolmogorov–Smirnov two-sided test *P* value = 1.8E-89. Drug ligand-binding background Kolmogorov–Smirnov two-sided test *P* value = 3.0E-75. (*B*) Counts of tolerant and intolerant drug ligand-binding sites grouped by therapeutic area. Here, tolerant is defined as 3DTS > 0.24 (50th percentile of 3DTS), while intolerant is defined as described in the text (3DTS < 0.14; 20th percentile of 3DTS); drug binding sites between these 3DTS values are not included. See Dataset S3 for full details about this dataset.

the most intolerant bin, active sites were most constrained, followed by allosteric, protein–protein interaction, and ligand-binding pockets (Fig. 3*A* and Dataset S3). The higher scores of allosteric sites (more tolerant) relative to their orthosteric counterparts are consistent with the existing knowledge indicating that these sites tend to be under lower evolutionary conservation pressure (47). We also observed an unequal distribution of tolerant and intolerant binding sites across therapeutic classes (Fig. 3*B* and Dataset S3). For example, antineoplastic and immunomodulating agents preferentially target intolerant sites. The identification of multiple intolerant 3D sites and domains in many drug targets could be exploited for rational drug design and for analysis of drug screening results.

Recently, we and others evaluated genome constraint based on depletion of human variation data in linearly defined regions in coding (48, 49) and in noncoding regions (16). The current study extends this approach to regions defined by tertiary structure. The increasing detail of the limits of protein diversity that can be gathered through large-scale sequencing of the human population and 3D proteins structures offers additional data on orthosteric, allosteric, and additional functional sites that could be harnessed for drug development.

Materials and Methods

Detailed information is provided in *SI Appendix*.

Genomic and Variant Data. We included a set of 123,136 exomes and 15,496 whole human genomes from gnomAD (https://gnomad.broadinstitute.org/). Feature annotations were taken from Uniprot text files that were cross-referenced from Gencode. We used pairwise global sequence alignment to align the Uniprot amino acid sequence to the Gencode transcript. X-ray structure data from the Protein Data Bank were used if they were linked within the Uniprot text files. The PyMol molecular visualization system was used to identify any residue within 5 Å of a defined Uniprot feature (also referred to as a 3D site).

Creation of a 3D Tolerance Score. We group variants based on their spatial proximity in 3D protein space and based on Uniprot feature annotation. We term these groups 3D sites. We calculate the expectation on the probability that the 3D site is intolerant to missense mutation using a model, which accounts for the differences among loci in the rates of neutral missense variation due to the genetic code, differential sample availability, and regional mutation rates.

Functional Data and Pathogenicity Scores. Deep mutational scanning data are available for PPARG (23); MAPK1/ERK2 (24); p53 (25), PTEN and TPMT (26); UBE2I, SUMO1, TPK1, CALM1, CALM2, and CALM3 (27); and two single-protein domains of BRCA1 (the RING domain) and YAP65 (the WW domain) (21, 28). For most scores, comparative method data were sourced from dbNSFPv3.5a (36, 50) except for EVmutation (45) data. Scores resulting in missense variants were averaged across a nucleotide (where applicable), then an amino acid position, and, last, a 3D site.

Drug Ligand Data Set and Analyses. A set of structures defined as therapeutic targets of FDA-approved drugs was used. Therapeutic targets were taken from the supplementary information of Santos et al. (51). Ligand-binding sites were defined as those residues within 5 Å of any of the bound therapeutic molecule residues. Drug liganded molecules were assigned to their ATC codes using the supplementary information of Santos et al. (51). We used the Allosteric Database (release no. 3.06) (52). A nonredundant list of protein active sites was included for those structures found in the Drug Ligand Data Set and the Allosteric Data Set. Additionally, protein–protein interfaces were included if those structures were found in the Drug Ligand Data Set, Active Site Data Set, and the Allosteric Data Set.

Statistics. Statistics were calculated using the NumPy (www.numpy.org) and SciPy (https://www.scipy.org) libraries in Python and in-house statistical software in Scala.

Public Resources. We provide final scores and intermediate results from the genome to proteome mapping, including the UniProt–PDB pairwise alignments, at https://doi.org/10.5281/zenodo.1311198 (53). We provide the source code at doi.org/10.5281/zenodo.2628193 (54). There is an interactive browser at protc.labtelenti.org.

ACKNOWLEDGMENTS. We thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data for this resource. A full list of contributing groups can be found at https://gnomad.broadinstitute.org/about.

1. Auton A, 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
2. Telenti A, et al. (2016) Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci USA* 113:11901–11906.
3. Lek M, Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291.
4. Biesecker LG, Green RC (2014) Diagnostic clinical genome and exome sequencing. *N Engl J Med* 371:1170.
5. Kircher M, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46:310–315.
6. Cassa CA, et al. (2017) Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet* 49:806–810.

7. Davydov EV, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6:e1001025.
8. Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H (2013) Utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics* 14:190.
9. Ryslik GA, et al. (2014) A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC Bioinformatics* 15:231.
10. Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H (2014) A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics* 15:86.
11. Ryslik GA, Cheng Y, Modis Y, Zhao H (2016) Leveraging protein quaternary structure to identify oncogenic driver mutations. *BMC Bioinformatics* 17:137.
12. Fujimoto A, et al. (2016) Systematic analysis of mutation distribution in three dimensional protein structures identifies cancer driver genes. *Sci Rep* 6:26483.
13. Tokheim C, et al. (2016) Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* 76:3719–3731.
14. Meyer MJ, et al. (2016) mutation3D: Cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum Mutat* 37:447–456.
15. Porta-Pardo E, et al. (2017) Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat Methods* 14:782–788.
16. di Iulio J, et al. (2018) The human noncoding genome defined by genetic diversity. *Nat Genet* 50:333–337.
17. Bhattacharya R, Rose PW, Burley SK, Prlić A (2017) Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One* 12:e0171355.
18. Arodź T, Płonka PM (2012) Effects of point mutations on protein structure are non-exponentially distributed. *Proteins* 80:1780–1790.
19. Sivley RM, Dou X, Meiler J, Bush WS, Capra JA (2018) Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am J Hum Genet* 102:415–426.
20. Glusman G, et al. (2017) Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: A proposed framework. *Genome Med* 9:113.
21. Fowler DM, et al. (2010) High-resolution mapping of protein sequence-function relationships. *Nat Methods* 7:741–746.
22. Fowler DM, Fields S (2014) Deep mutational scanning: A new style of protein science. *Nat Methods* 11:801–807.
23. Majithia AR, UK Monogenic Diabetes Consortium, Myocardial Infarction Genetics Consortium, UK Congenital Lipodystrophy Consortium (2016) Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* 48:1570–1575.
24. Brenan L, et al. (2016) Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell Rep* 17:1171–1183.
25. Kato S, et al. (2003) Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci USA* 100:8424–8429.
26. Matreyek KA, et al. (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* 50:874–882.
27. Weile J, et al. (2017) A framework for exhaustively mapping functional missense variants. *Mol Syst Biol* 13:957.
28. Starita LM, et al. (2015) Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 200:413–422.
29. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–1081.
30. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7:e46688.
31. Shihab HA, et al. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 34:57–65.
32. Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8:R232.
33. Shihab HA, et al. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31:1536–1543.
34. Gulko B, Hubisz MJ, Gronau I, Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 47:276–283.
35. Quang D, Chen Y, Xie X (2015) DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31:761–763.
36. Dong C, et al. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 24:2125–2137.
37. Lu Q, et al. (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 5:10576.
38. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48:214–220.
39. Jagadeesh KA, et al. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 48:1581–1586.
40. Ioannidis NM, et al. (2016) REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 99:877–885.
41. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20:110–121.
42. Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.
43. Garber M, et al. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25:i54–i62.
44. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249.
45. Hopf TA, et al. (2017) Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 35:128–135.
46. Webb B, et al. (2014) Comparative modeling of drug target proteins. *Elsevier Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*, ed Reedijk J (Elsevier, Waltham, MA).
47. Wenthur CJ, Gentry PR, Mathews TP, Lindsley CW (2014) Drugs for allosteric sites on receptors. *Annu Rev Pharmacol Toxicol* 54:165–184.
48. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR (2019) A map of constrained coding regions in the human genome. *Nat Genet* 51:88–95.
49. Hayeck TJ, et al. (2019) Improved pathogenic variant localization via a hierarchical model of sub-regional intolerance. *Am J Hum Genet* 104:299–309.
50. Liu X, Wu C, Li C, Boerwinkle E (2016) dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat* 37:235–241.
51. Santos R, et al. (2017) A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 16:19–34.
52. Shen Q, et al. (2016) ASD v3.0: Unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Res* 44:D527–D535.
53. Bartha I, Hicks M, Telenti A (2018) Functional characterization of 3D-protein structures informed by human genetic diversity - data. Zenodo. Available at https://doi.org/10.5281/zenodo.1311198. Deposited July 12, 2018.
54. Bartha I, Hicks M, Telenti A (2018) Functional characterization of 3D-protein structures informed by human genetic diversity - source code. Zenodo. Available at doi.org/10.5281/zenodo.2628193. Deposited April 4, 2019.