Mapping Genetic Variations to Three Dimensional Protein Structures to Enhance Variant Interpretation: a proposed framework

Gustavo Glusman[1,*], Peter W. Rose[2], Andreas Prlić[2,3], Jennifer Dougherty[1], José M Duarte[3], Andrew S. Hoffman[4], Geoffrey J. Barton[5], Emøke Bendixen[6], Timothy Bergquist[7], Christian Bock[7], Elizabeth Brunk[8], Marija Buljan[9], Stephen K. Burley[10,2,3], Binghuang Cai[7], Hannah Carter[8], JianJiong Gao[11], Adam Godzik[12], Michael Heuer[13], Michael Hicks[14], Thomas Hrabe[12], Rachel Karchin[15,16], Julia Koehler Leman[17,18], Lydie Lane[19], David L. Masica[15], Sean D. Mooney[7], John Moult[20,21], Gilbert S. Omenn[22,1], Frances Pearl[23], Vikas Pejaver[7,24], Sheila M. Reynolds[1], Ariel Rokem[24], Torsten Schwede[25], Sicheng Song[7], Hagen Tilgner[26], Yana Valasatava[3], Yang Zhang[22], Eric W. Deutsch[1]

1 Institute for Systems Biology, Seattle, WA, USA
2 San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA
3 RCSB Protein Data Bank, University of California San Diego, La Jolla, CA, USA
4 Human Centered Design & Engineering, University of Washington, Seattle, WA, USA
5 Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK
6 Department of Molecular Biology and Genetics, Aarhus University, 8000 Aarhus, Denmark
7 Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA
8 University of California San Diego, La Jolla, CA, USA
9 Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
10 Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey
11 Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA
12 SBP Medical Discovery Institute, La Jolla, CA, USA
13 AMPLab, University of California, Berkeley, CA, USA
14 Human Longevity, Inc., San Diego, CA, USA
15 Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University
16 Department of Oncology, Johns Hopkins Medicine
17 Flatiron Institute, Center for Computational Biology, Simons Foundation, New York, NY
18 Department of Biology and Center for Genomics and Systems Biology, New York University, New York, NY, USA
19 SIB Swiss Institute of Bioinformatics and University of Geneva, Geneva, Switzerland
20 Institute for Bioscience and Biotechnology Research, University of Maryland, MD, USA
21 Department of Cell Biology and Molecular Genetics, University of Maryland, MD, USA
22 University of Michigan, Ann Arbor, MI, USA
23 School of Life Sciences, University of Sussex, Brighton, UK

24 The University of Washington eScience Institute, Seattle, WA, USA
25 SIB Swiss Institute of Bioinformatics and Biozentrum University of Basel, Basel, Switzerland
26 Weill Cornell Medicine, Brain and Mind Research Institute, New York City, NY, USA

* Correspondence: Gustavo@SystemsBiology.org
Gustavo Glusman
Institute for Systems Biology
401 Terry Ave N
Seattle, WA 98109, USA
Tel: 206 732-1273

Abstract

The translation of personal genomics to precision medicine is predicated on the accurate interpretation of the multitude of genetic variants observed for each individual. However, even when genetic variants are predicted to modify a protein, their functional implications may be unclear. Many diseases are caused by genetic variants affecting important protein features, such as enzyme active sites or interaction interfaces. The scientific community has catalogued millions of genetic variants in genomic databases and thousands of protein structures in the Protein Data Bank. Mapping mutations onto 3D structures enables atomic-level analyses of protein positions that may be important for the stability or formation of interactions; these may explain the effect of mutations and in some cases even open a path for targeted drug development. To accelerate progress in the integration of these data types, we held a two-day Gene Variation to 3D (GVto3D) workshop to report on the latest advances and to discuss unmet needs. The overarching goal of the workshop was to address the question: what can be done together as a community to advance the integration of genetic variants and 3D protein structures that could not be done by a single investigator or lab? Here we describe the workshop outcomes, review the state of the field, and propose the development of a framework with which to promote progress in this arena. The framework will include a set of standard formats, common ontologies, a common application programming interface (API) to enable interoperation of the resources, and a Tool Registry to make it easy to find and apply the tools to specific analysis problems. Interoperability will enable integration of diverse data sources and tools and collaborative development of variant effect prediction methods.

Keywords: genetic variants, protein structure, precision medicine, data integration

Background

Recent progress in DNA sequencing technologies has ushered in an era of rapid and cost-effective genome sequencing, enabling clinical applications [1] and the potential for personalized, systems medicine [2] through the understanding of the individual's genetic risks and by integration with longitudinal phenotype measurements [3]. The detailed knowledge of an individual's genotype poses a significant interpretation challenge: while genetic variants disrupting transcript structure and protein-coding sequences (e.g., nonsense mutations) have long been considered 'low hanging fruit' relative to variants in non-coding sequences, the field

still struggles with interpreting missense mutations, which are more common, and more frequently associated with disease [4]. This has led to an increasing number of variants of uncertain significance (VUS). To address the resulting annotation and reporting challenges [5,6], the American College for Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have released variant interpretation guidelines based on pathogenicity [7]. The interpretation of variants relies on a combination of multiple lines of evidence, including the frequency of the variant in the population (common variants are less likely to be pathogenic), the mode of segregation in pedigrees (e.g., *de novo* mutations not observed in parents are more likely to be pathogenic than those inherited), the mode of presentation in affected individuals (e.g., single dominant variant, single variant in homozygous state, two variants in compound heterozygous state), the predicted effect on RNA and protein sequence and structure, and prior knowledge accumulated in curated databases. Many computational tools have been developed to support these assessments (see Table S1 in Additional File 1). Multiple challenges remain in the quickly evolving field of clinical variant interpretation, including differences in allele frequency among different populations, a growing but still incomplete understanding of how variants affect regulation, sequence and structure of RNA and protein products, and the partial, inconsistently presented and sometimes conflicting knowledge in databases.

To assess the potential pathogenicity of genetic variants, singly or in combinations, it is useful to assess their frequency in control or general populations. Public databases are burgeoning with information about genetic variants in humans and in many model organisms. Resources, such as dbSNP [8], dbVar [9], COSMIC [10], cBioPortal [11], UniProt [12], Kaviar [13], Clinvar [14], HGMD [15], ExAC and gnomAD [16] provide data on hundreds of millions of single nucleotide variants (SNVs) and other types of genetic variations. Each database has a different focus, different sources of data, processing methods, level of coverage, and degree of metadata associated with each variation; some focus only on human variation, while others cover many species. Similarly, each database has differing mechanisms for data access and differing levels of cross-referencing.

The biomedical research community is fortunate to have access to such a wealth of information, but is simultaneously daunted by its sheer size and disparate nature. In addition to public databases hundreds of DNA and RNA sequencing experiments are revealing manifold genetic variants and mutations each year, and an increasing number of these can be linked to protein structure. For example, protein structure analysis of a novel variant in the ubiquitin-protein ligase TRIM11, observed in individuals affected with inflammatory bowel disease, helped determine that the variant is more likely to affect protein-protein interactions rather than protein folding and stability [17]. Functionally important somatic variants in cancer may form statistically significant spatial clusters in three-dimensional protein structure, which are not detectable in one-dimensional sequence, such as kidney cancer-specific variants in the tumor suppressor gene *VHL*, which are proximal to the binding site of VHL for its ubiquitination target HIF1A [18].

Simultaneously, there has been great progress in characterizing the three-dimensional structures of proteins [19,20], both experimentally and computationally. Essentially all publicly

available experimentally derived structures are deposited into the Protein Data Bank (PDB) [21]. When experimentally determined structures are not available for proteins, structural models may be used instead. Protein Model Portal [22] aggregates precomputed models from multiple resources, whereas most methods generate models interactively on request, e.g. I-TASSER [23], ModWeb [24], Phyre2 [25], HHpred [26] or SWISS-MODEL [27]. Currently available homology models with 40 - 50% sequence identity to experimental structures already cover approximately 40% of the residues in the human proteome [28], although this does not always include the full-length protein in the correct quaternary structure, but often only specific domains. Beyond simply having three-dimensional models of proteins, it is crucial to annotate the functional substructures in these models with such information as the locations of binding and active sites, functional domains, regions that are externally accessible versus in the protected interior, interaction interfaces and other structural features that might be related to function [29].

However, the connections between genetic variations and protein structure are not always easy to find. A few computational tools have begun to emerge (cBioPortal [11], COSMIC-3D [30], CRAVAT [31], Jalview [32], MuPIT [33], MutDB [34], STRUM [35], Cancer3D [36]) that enable users to take individual genetic variations, or a list of them, and visualize these in the context of protein structures. For example, CRAVAT [31] allows a user to upload a Variant Call Format (VCF) file [37] (a file format used for representing DNA sequence variations) containing many genetic variants and assess which of those variants map to proteins, and then to explore individual variants in a 3D visualization of each protein when available. STRUM [35] allows users to visualize the structural model of a protein while in addition providing the profiles of the folding free-energy changes induced by the single-nucleotide polymorphisms (SNPs) or mutations. The starting point of STRUM is the wild-type sequence with SNPs or mutations, whereas I-TASSER is used to generate 3D protein models from which the impact of genetic mutations on protein stability can be more accurately calculated compared to the sequence-based approaches. Other tools like Jalview [32] provide a workbench for exploring variants in context with multiple sequence alignments, molecular structures and annotations. COSMIC-3D and cBioPortal [11] map and visualize variants in their databases on 3D protein structures. The VIPUR pipeline [38] goes one step further and allows automatic interpretation of the effect of the mutation on the protein structure. The input to VIPUR is the wild-type sequence and the mutation of interest, and based on availability of a known structure or homology model, the tool maps the mutation onto the structure, and uses Rosetta [39] energy terms (see Glossary) as indicators to report which features are most strongly affected by the mutation. Broad mining of data across thousands of proteins and millions of variants remains challenging due to the computational cost of structure modeling and the limited availability of experimental structures and high-fidelity models.

The confluence of genetic variation information and protein structure knowledge has broad applications to many fields of study, including precision medicine [40]. One can envision a future where an individual’s genetic variants are uploaded to an intelligent system that can flag variants for previously documented functional alterations, and then enable a clinician or genetic counselor to explore potential implications for health and disease of the individual’s variants

based on their predicted effects on the functions of individual proteins. Similarly, the decision of which therapies are indicated may be influenced or directly based on the known function of a drug as it relates to potential variants on the drug's target protein. Such a system remains distant, but the time is right for developing an infrastructure that would enable its development. There are a few ongoing efforts curating functional data and disease association for cancer variants [41–44]. Efforts to computationally model the association of various genomic mutations and human diseases are also underway [45–47].

Although the handful of tools listed above already perform an integration of genetic variation and protein structure at some level, building infrastructure for both large-scale integration as well as broader usage of tools in the lab and in the clinic has yet to be achieved. Large-scale data integration for millions of variants, thousands of genomes, and tens of thousands of structures on platforms such as Apache Spark [48] and Google BigQuery [49,50] will enable complex queries and machine learning approaches to further learn how to predict functional implications of detected variants.

In order to accelerate progress in this field we held a workshop on this topic at the Institute for Systems Biology in Seattle in February 2017. In this article, we summarize the discussions and conclusions of this workshop, and present a comprehensive overview of the field. Finally, we conclude with a proposed architecture for a framework that could allow improved interoperability between the tools in this domain, making it easier for everyone to build on the accomplishments achieved thus far.

The Gene Variation 3D workshop

On February 9th and 10th, 2017, we hosted the Gene Variation to 3D (GVto3D) workshop at the Institute for Systems Biology in Seattle, WA. The goal of the workshop was to explore the state of the field connecting genetic variation and 3D protein structure, and to bring together some of the key researchers working on interpreting genetic variation data. The workshop consisted of a mixture of talks, discussion sessions, and breakout groups. The program is available at the workshop website [51]. Twenty-five speakers provided short (15 minute) summaries of their research; highlights from the talks are available on the meeting website. The oral presentations connected the workshop theme to diverse topics such as RNA sequencing, Big Data technologies, how precision medicine can help with specific diseases, and cancer research.

After all the presentations and discussion sessions concluded, workshop participants separated into two breakout groups to brainstorm about how the research community as a whole could accelerate progress in the field in ways that individual labs could not.
Breakout group 1 discussed existing ontologies, tools and datasets in the field and considered potential architectures for an integrative framework, focusing on how tools and resources could be made more interoperable to enable more widespread use of the tools and integration of inputs and outputs among the tools. Important aspects that emerged in the discussion include:

1. Adoption or development of standardized formats for the various major data types (such as variants, splice isoforms, post-translational modifications, structures, sequence annotations, and phenotypes).

2. Mechanisms to scale up the information exchange to large-scale queries using big data technologies such as DataFrames [52] and BigQuery [49].
3. Use of ontologies to standardize the terminology for the exchange of data and knowledge. These ontologies already mostly exist, and need only be specified as the standard, although some extension may be required.
4. Selection of initial tools that should be part of a pilot phase of the development and initial deployment of the interoperability framework.
5. Development of a tool registry and portal that would serve as a web-accessible resource for finding relevant tools, their inputs and outputs, and also reference data files that can be used to demonstrate and validate the tools and their interoperation.

Breakout group 2 discussed unmet needs, ranging from improvements in structural interpretation of splicing variants to more effective dissemination of knowledge to clinical geneticists, tumor panels and the general public. Salient questions and points that were discussed included:

1. How can we increase the actionability of variants observed in patients? Beyond facilitating access to knowledge on structural impacts of variants, there is need for a metric of confidence in the predicted impact. Gene editing technologies are likely to enhance experimental studies of salient variants.
2. There is also a need to recognize multi-variant interactions within single genes and proteins and mutation effects on protein-protein, protein-nucleic acid, or protein-ligand and drug interactions. Also, annotation of context in which each variant could have an effect is of importance. For instance, information on cell types or cellular conditions in which specific interactions or protein complexes are formed, as well as annotation of epistatic relationships with mutations elsewhere in the genome can help in interpreting a mutation's influence on the cell.
3. How can we improve the interpretation of variants affecting splicing? A proposal was made to create a mechanism for collecting donated RNAseq data to derive a comprehensive set of splice variants and interpreting them in the context of protein structure. It may also be useful to organize data on splice variants by type of alternative splicing (for example, exon swaps, intron retention, and coordinated inclusion of distant alternative exons [53], which are widespread in the human transcriptome and primarily affect protein coding exons (UNDER REVIEW)).
4. How can we standardize annotation pipelines and data integration methods? It was recognized that this has already been partially solved independently by various teams, such as mapping genomic positions onto 3D structures (see "Current State of the Field" section), so there would be benefit from implementing an interoperation framework.
5. Who are the target audiences? Scientists, tumor boards, clinical geneticists, developers of targeted drugs, patients, and lay people with interest in genetic testing.
6. How can we improve documentation and outreach? Suggestions included development of documentation videos and tutorials, and contributing to Wikipedia sections describing the impact of variants on protein structure, building on current experience, like the Protein Standards Initiative [54] of the Human Proteome Organization.

The workshop has already begun to positively impact collaboration and interoperability in the wider research community. For example, an immediate outcome from discussions that occurred during the workshop was that links were added to the Kaviar database of human SNPs [13] and the PeptideAtlas database of proteins detected via mass spectrometry [55,56] pointing researchers to the MuPIT resource [33], so that the variations in the former resources can be depicted using the tools in the latter resource. Engaging members of the research community as we have suggests promising avenues for further work in this direction, including the design of a framework according to principles of user-centered design. Before laying out our vision for the framework, however, we first provide an overview of the current state of the field.

The Current State of the Field

Here we review methods that use 3D structural information from the Protein Data Bank to predict the effect of missense mutations; mapping other types of mutations (e.g., insertions, deletions, splicing effects) remains an open challenge. In Table 1 we present an overview of six classes of prediction methods, summarizing the type of prediction and listing some of their limitations. We have then reviewed the literature and assigned methods to these classes. Table S1 (see Additional File 1) presents an extensive summary of over 30 such methods that have been published in the last decade, and have a current web presence as a web-based user interface, a web service, or a downloadable stand-alone application. In addition, we have captured tools that rely on sequence information only. Prediction tools are trained, tested, and validated on sets of reference proteins and their mutated forms (benchmark datasets). In Table S1 we have included a list benchmark datasets commonly used to train prediction tools.

A first set of methods predicts thermodynamic properties related to mutations: 1) change in protein stability [35,57–71]; and 2) change in binding affinity for protein-protein [65,72–77], protein-nucleic acid [65], and protein-ligand complexes [78]. These methods have been trained on data of wildtype and mutant protein pairs, often using protein stability data from the ProTherm database [79], protein-protein binding affinities from SKEMPI [80], protein-nucleic acid binding affinities from ProNIT [79], and protein-ligand binding affinities from Platinum [81].

A second set of methods [38,57,75,82–87] predicts the phenotypic effect (pathogenicity) of mutations, most often as a binary classification: deleterious or neutral effect. These methods have been trained on data resources that either contain mostly germline mutations, ClinVar [14], HGMD [15], and OMIM [88], or somatic mutations such the Cancer Genome Atlas (TCGA) [89], and COSMIC [10]. Carefully selected benchmark datasets to develop and test prediction methods have been collected: VariBench [90] and VariSNP [91].

Few prediction methods are purely based on 3D structural information, with the exception of FoldX [62], which uses an empirical scoring function to predict the change in protein stability or protein-protein binding. Most methods (Table S1 in Additional file 1) use a combination of structural and sequence features and then formulate a regression problem to predict scalar values (e.g., affinity changes), or a classification problem to predict a mutation as probably deleterious or neutral. Some methods use homology models when experimentally determined structures are not available, to increase structural coverage. The use of structural information

varies from method to method. FoldX uses the 3D atomic coordinates of the protein, whereas most methods extract structural features that characterize changes in the local environment around a mutated residue [38].

Most tools to predict the effect of mutations are available online. However, there is a wide variety of input formats and scope of prediction (i.e., predicting the effect of a single or multiple amino acid mutations). The majority of the 3D protein structure-based tools take PDB residue numbers of the mutated sites as input (Additional File 1, Table S1). There are also tools that exploit structural models predicted by advanced structure modeling algorithms and demonstrate the usefulness of structure predictions compared to those using only sequences, such as FoldX [62] or BindProfX [77]. A smaller number of tools use UniProt/Swiss-Prot residue positions. A minority of tools use chromosome position, dbSNP ID [8], or VCF files as input. A few tools need explicit PDB structures in the wild-type and mutated forms. User interfaces and presentation of results with the available web resources vary significantly; some resources require a user registration, and in some instances results are returned by email.

Several integrated tools have been developed that combine the prediction of the effects of mutations, annotation by functional information, and visual mapping of mutation sites onto 3D protein structures and multiple sequence alignments. Examples include 3DHotspots.org [92], cBioPortal [11], COSMIC-3D [10], CRAVAT [31], Jalview [32], LS-SNP/PDB [93], MOKCA [94], MuPIT [33], RCSB PDB [21], SNP2Structure [95], and Cancer3D [36]. These tools might help elucidate the effect of mutations in the context of both 3D structure and other available annotations. Ensembl's Variant Effect Predictor (VEP) [96] combines several annotation and prediction services, including various considerations of effects on protein products.

A biologist who wants to assess the effect of mutations is confronted with a bewildering set of tools and options. The high variability in the user interfaces and representation and retrieval of results makes a systematic comparison of predictions by multiple tools cumbersome and requires manual input; hence, most tools are not applicable to anything but a small set of selected mutations. A systematic or automated comparison of a list of mutations (e.g., at exome-scale) using multiple tools is generally not possible. For example, it would be useful to run tools that predict multiple effects of mutations simultaneously, such as protein stability and interruption of protein-protein and protein-nucleic acid binding. A further limitation is the input by PDB or UniProt residue position, since SNVs are annotated using genomic coordinates. Mapping between genomic and protein coordinate systems is error prone due, for example, to different genome assembly versions and alternative splicing. Where a mapping from genome to Uniprot is possible, SIFTS [97] and CRAVAT [31] provide consistent residue-level mapping to and from PDB structures and other resources.

Current tools that predict the effect of missense mutations are either based on protein sequence information, 3D structural information, or both. Tools predict either biophysical changes or effect on phenotype. Tools that use 3D structural information and visualization offer additional insights by providing locations of mutations in a 3D context, which is not possible using sequence-based prediction. For example, multiple mutations on a protein can be visualized and potential

hotspots can be identified. In the next section we describe a framework to overcome the large heterogeneity of tools with limits their usefulness, ease of use, and hinders comparative performance assessments.

Proposed framework for making progress as a community

To facilitate innovation in this field, we recommend the development of a framework of common formats and application programming interfaces (APIs) that enable the many resources to interoperate more effectively both at the individual variant level and at large scales. We further recommend the development of a portal that can be used to annotate the current state of tools in the field and guide users on how these tools can interoperate and be used to address different research questions. The outline of the recommended GVto3D Framework takes its lead both from our wider review of the field as well as from the presentations and discussions that occurred amongst those members of the research community in attendance at the workshop; its design incorporates the needs and existing efforts of these researchers.

Figure 1 depicts the recommended components and design of the GVto3D framework. The Tools Registry will act as a central repository of data resources and software tools related to genetic variants, protein sequences, protein structures, variant effect prediction and variant annotation. Metadata about each resource to enable findability of the different software tools will be stored and offered through an interactive web interface and also an application programming interface (API) which in turn enables the development of intelligent software that can automatically discover applicable resources and gather the information about how to communicate with them to obtain the desired results. In addition to name, description, citations, contact information, and uniform resource locators (URLs), each entry will contain information important to the tool's interoperation, such as the inputs and outputs, API support, and reference genome information.

A second component of the portal will be the definition of standard APIs so that information can be sent to and requested from different tools in the same way, thereby reducing software development overhead typically encumbered with different tools use different APIs. It is envisaged that new 3rd party tools will use the API natively while API adapters will be developed in order to bridge with pre-existing 3rd party tools. The API enables seamless interoperability between different variant related tools and also a standard access to a multi-directional mapping among genomic, protein sequence and protein structure coordinates. These mappings will be made available through APIs and as downloadable data files. Mappings will be kept up to date based on the update schedules of the underlying data sources (PDB - weekly, UniProt - monthly), freeing developers from maintaining and updating copies of these data. Once several similar resources support the standard APIs, the site can be further developed into an aggregation portal, where a query at the portal can be automatically farmed out to multiple resources, and the results collated and returned to the user in a single batch. This framework advances the FAIR principles of findability, accessibility, interoperability, and reusability [98] for all tools and resources that participate.

The use of standard file formats and standardized representations of data enable interoperability of prediction tools, e.g., the output from one tool can be passed as input into a second tool, and can thereby simplify the comparison of different methods. The standardized formats are also essential components of a reusable set of integrated tools (software stack), including tools for reading and interpreting data files (file parsers), APIs, and visualization tools. Most of the current tools use a variety of inputs and outputs, placing a large burden on the user to transform data. Standard file formats and uniform APIs will be at the core of future services that will combine and compare different approaches. Various platforms and tools have different schedules and reliability of upgrades; keeping track of versions is important as changes to software may have large effects on the results.

The VCF file format [37], despite its complexity, is the *de facto* standard format for storing variant calls for single nucleotide variants to long insertions and deletions. The Global Alliance for Genomics and Health's Data Working Group File Formats Team defines the VCF specification and its evolution [99]. Variant annotations, e.g., the results of prediction tools, can be captured in the INFO records, which are a set of structured records used to add annotation to VCF files. VCF versions 4.x, including the current version 4.3 [100] define meta-information lines that describe the INFO record data types and enforce standardization [101]. In addition to VCF, a few other formats have been described such as ANN, which defines a different standard for representing variant information in INFO fields; the Variant Effect Predictor [96] supports a simple tab-delimited, as well as JavaScript Object Notation (JSON) output format.

The Human Genome Variation Society aims to foster discovery and characterization of genomic variations including population distribution and phenotypic associations. It has established guidelines and recommendations for the nomenclature of gene variations and serves as an international standard [102].

Progress in this field depends on global collaboration and the sharing and reuse of tools. APIs provide protocols to enable this collaboration. Tools wrapped in standard APIs present a consistent interface to heterogeneous tools, enhancing interoperability, and shield the user from changes to the underlying software. As an example, many prediction tools that use 3D protein structural information define the location of mutations at the protein level using either UniProt or PDB coordinates. Mapping genomic coordinates to 3D protein structure is non-trivial and error prone. Robust APIs that can perform this mapping with up to date 3D information using both types of protein coordinates can augment existing tools that are based on just linear protein sequence coordinates.

Moreover, progress in the prediction of the effect of mutations and use of 3D structural information depends on the availability of well-designed training, test, and validation sets. The tool repository will be a place to share datasets, and protocols and references (metadata) for how these datasets were generated. Validation sets, accompanied by well documented tutorials or vignettes, will include a subset of variants with well understood effects that can be used to test the output of available resources. Eventually these can serve as a set of unit tests for the framework itself.

Conclusions and future perspectives

Although there will continue to be exceptional innovation in this field, the disparate nature of current tools and resources and lack of interoperability contribute to slower progress of the field than might otherwise be possible. The Gene Variation to 3D workshop held in Seattle in February 2017 represents an important step towards spurring collaboration and advancing progress in proteogenomics research. Development of a community-driven interoperability framework for integrating genetic variation resources and protein structure resources promises further expansion of our understanding of the functional implications of genetic variation. While the use of 3D structural features has enabled the atomic level exploration of the effects of mutations, e.g., the identification of mutation hotspots, the accuracy, scope, and scale of predictions are still limited. The proposed framework will enable pooling of data sources and tools and collaborative development.

However, there will be substantial challenges as we move forward with design of the framework. The first challenge is establishing a durable user base for the framework. One possible approach is to engage a few key labs to take the lead as early adopters, and assume that the framework will gain wider community acceptance through the work of these early adopters. We propose a more user-centered design approach, however, which emphasizes an iterative process of engaging multiple communities of practice in conceptualizing, developing, and rolling out the framework; the GVto3D workshop was a first step in this direction. Second, questions of sustainability are also pertinent here insofar as how such a system will be maintained, and who will be responsible for ongoing maintenance. Here, we propose an approach that we envision will become self-sustaining through deployment of open-source technologies in an engaged community. Third, standardization is a key component of any interoperability project, which in this case depends upon work to enhance usage of certain *de facto* standards, and to establish other standards, including the creation of standard APIs. Working closely with the community of potential framework users, as well as with standard-setting bodies such as Global Alliance for Genetics and Health and the Proteomics Standards Initiative of the Human Proteome Organization will be important for helping these standards gain further traction.

Taken together, the user-centered framework we have outlined above – a Tool Registry and a set of standardized formats and common APIs based on deployment of open-source materials – aims at bringing the FAIR principles to bear on current and emerging tools while enabling broader usage across multiple communities of practice. The result promises to be more rapid progress in research that can make use of GVto3D resources and eventual applications to precision medicine, while ensuring that methods and outcomes are findable, accessible, interoperable, and reusable.

List of abbreviations

ACMG: American College of Genetics and Genomics
AMP: Association for Molecular Pathology
API: application programming interface
FAIR: findable, accessible, interoperable, reusable

GVto3D: Gene Variation to 3D
PDB: Protein Data Bank
SNV: single-nucleotide variant
SNP: single-nucleotide polymorphism
URL: universal resource locator
VCF: variant call format
VUS: variant of uncertain significance

Glossary

Benchmark dataset: A curated and well-studied dataset that can be used to evaluate the relative performance of analysis methods and algorithms.
File parser: A computer program module that interprets the structure of input data and breaks the input into well-defined parts that can then be used by other parts of the computer program.
Rosetta energy terms: Rosetta [39] estimates the energetic stability of protein structures as a sum of energy terms, including hydrogen bonding, electrostatic interaction, attractive and repulsive interaction, and solvation terms.
Software stack: A set of software subsystems or components designed to work together as a platform.
Variant Call Format: A standard format of a text file used for storing genome sequence variations relative to a reference genome.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and material
Not applicable.

Competing interests
GG holds stock options in Arivale, Inc. Arivale, Inc. did not fund the study and was not involved in its design, implementation or reporting. MH is an employee at Human Longevity, Inc. The other authors declare that they have no competing interests.

Funding
This work has been funded in part by the National Science Foundation under grant numbers IIS-1636903 and IIS-1636804. JG was supported by a National Cancer Institute Cancer Center Core Grant (P30-CA008748). GG was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number OT3TR002026. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' contributions

AP, EWD, JD, GG and PWR organized the workshop. GG, PWR, AP and EWD wrote the manuscript, with many contributions from other authors. All authors read and approved the final manuscript.

Acknowledgements

We wish to thank Meredith M. Lee and the National Science Foundation's West Big Data Innovation Hub (http://westbigdatahub.org/) for support and encouragement.

Additional file 1

Table S1. Tools that use 3D structural information from the Protein Data Bank to predict the effect of missense mutations.

Figures

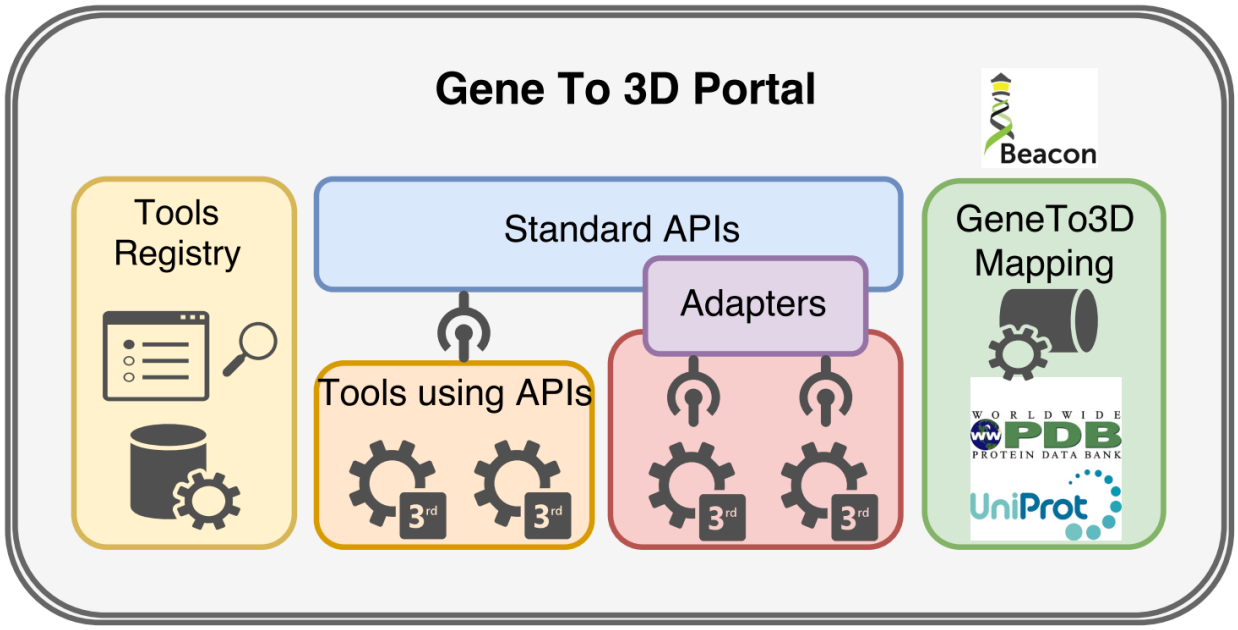

Figure 1. Components of the GVto3D portal. The Tools Registry contains a searchable description and metadata for tools, resources, and reference data sets for third-party variant effect prediction and annotation services. Standardized APIs provide interoperability for data input and output of these 3rd party tools. Custom adapters can provide limited interoperability for tools that cannot adopt the API. A mapping service provides bi-directional mappings from reference genome coordinates to UniProt protein positions and to PDB residue positions. The tools can use the mapping service to accept variant positions in any of the three coordinate systems. A beacon system enables queries about variant positions where 3D structural information and annotation is available.

Tables

Table 1: Classification of methods to predict the effect of missense mutations

Method Type	**Prediction**	**Limitations**
Protein stability	Predicts the difference in unfolding free energy between wild-type and mutant protein	Considers only one possible mechanism that may affect the phenotype
Protein-protein/protein-nucleic acid affinity	Predicts the difference in the binding affinity between binding partners upon mutation	Small training datasets limit the scope of these methods
Protein-ligand affinity	Predicts the difference in ligand binding affinity upon mutation	Small training datasets limit the scope of these methods
Phenotypic effect	Predicts the likelihood that a mutation is deleterious without considering a specific molecular mechanism	Except for Mendelian disease phenotypes, the phenotype may only be observed in a subset of the population (partial penetrance). Databases use different annotation practices and contain contradictory information for some mutations.
Mapping and 3D visualization	Provides a 3D context of the site of mutation and may give atomic level insight into mechanism of action	Visual approach is not suitable for automated whole-exome predictions
Mutation hotspots	Clusters mutations by spatial proximity that are not necessarily close in protein sequence	Clustering may not explain the effect of specific mutations in a hotspot

References

1. Glusman G. Clinical applications of sequencing take center stage. Genome Biol. 2013;14:303.

2. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1:2.

3. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat. Biotechnol. 2017;35:747–56.

4. Katsonis P, Koire A, Wilson SJ, Hsu T-K, Lua RC, Wilkins AD, et al. Single nucleotide variations: biological impact and theoretical interpretation. Protein Sci. Publ. Protein Soc. 2014;23:1650–66.

5. Kassahn KS, Scott HS, Caramins MC. Integrating massively parallel sequencing into diagnostic workflows and managing the annotation and clinical interpretation challenge. Hum. Mutat. 2014;35:413–23.

6. Vears DF, Sénécal K, Borry P. Reporting practices for variants of uncertain significance from next generation sequencing technologies. Eur. J. Med. Genet. 2017;

7. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. Off. J. Am. Coll. Med. Genet. 2015;17:405–24.

8. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.

9. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013;41:D936-941.

10. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.

11. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–4.

12. Pundir S, Magrane M, Martin MJ, O'Donovan C, UniProt Consortium. Searching and Navigating UniProt Databases. Curr. Protoc. Bioinforma. 2015;50:1.27.1-10.

13. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. Bioinforma. Oxf. Engl. 2011;27:3216–7.

14. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862-868.

15. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum. Genet. 2017;136:665–77.

16. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

17. Stittrich AB, Ashworth J, Shi M, Robinson M, Mauldin D, Brunkow ME, et al. Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. Hum. Genome Var. 2016;3:15060.

18. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016;76:3719–31.

19. Zhang Y. Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. 2008;18:342–8.

20. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001;294:93–6.

21. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. 2003;10:980.

22. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, et al. The Protein Model Portal--a comprehensive resource for protein structure and model information. Database J. Biol. Databases Curation. 2013;2013:bat031.

23. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat. Methods. 2015;12:7–8.

24. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, et al. ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 2014;42:D336-346.

25. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 2015;10:845–58.

26. Alva V, Nam S-Z, Söding J, Lupas AN. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res. 2016;44:W410-415.

27. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42:W252-258.

28. Schwede T. Protein modeling: what happened to the “protein structure gap”? Struct. Lond. Engl. 1993. 2013;21:1531–40.

29. Zhang Y. Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. 2009;19:145–55.

30. Modeling Cancer Mutations in 3-D. Cancer Discov. 2017;7:787–8.

31. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinforma. Oxf. Engl. 2013;29:647–8.

32. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinforma. Oxf. Engl. 2009;25:1189–91.

33. Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum. Genet. 2013;132:1235–43.

34. Mooney SD, Altman RB. MutDB: annotating human variation with functionally relevant data. Bioinforma. Oxf. Engl. 2003;19:1858–60.

35. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinforma. Oxf. Engl. 2016;32:2936–46.

36. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. Nucleic Acids Res. 2015;43:D968-973.

37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinforma. Oxf. Engl. 2011;27:2156–8.

38. Baugh EH, Simmons-Edler R, Müller CL, Alford RF, Volfovsky N, Lash AE, et al. Robust classification of protein variation using structural modelling and large-scale data integration. Nucleic Acids Res. 2016;44:2501–13.

39. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–74.

40. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015;526:336–42.

41. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis. Oncol. 2017;1–16.

42. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat. Genet. 2017;49:170–4.

43. Huang L, Fernandes H, Zia H, Tavassoli P, Rennert H, Pisapia D, et al. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J. Am. Med. Inform. Assoc. JAMIA. 2017;24:513–9.

44. Meric-Bernstam F, Johnson A, Holla V, Bailey AM, Brusco L, Chen K, et al. A decision support framework for genomically informed investigational cancer therapy. J. Natl. Cancer Inst. 2015;107.

45. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.

46. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.

47. Gao M, Zhou H, Skolnick J. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. Struct. Lond. Engl. 1993. 2015;23:1362–9.

48. Nothaft FA, Linderman M, Franklin MJ, Joseph AD, Patterson DA, Massie M, et al. Rethinking Data-Intensive Science Using Scalable Analytics Systems. ACM Press; 2015 [cited 2017 Nov 22]. p. 631–46. Available from: http://dl.acm.org/citation.cfm?doid=2723372.2742787

49. BigQuery [Internet]. [cited 2017 Nov 17]. Available from: https://cloud.google.com/bigquery/

50. Reynolds SM, Miller M, Lee P, Leinonen K, Paquette SM, Rodebaugh Z, et al. The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research. Cancer Res. 2017;77:e7–10.

51. Gene Variation to 3D Workshop Program [Internet]. [cited 2017 Aug 23]. Available from: http://genevariation3d.org/schedule/

52. Pandas DataFrame [Internet]. [cited 2017 Nov 17]. Available from: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html

53. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat. Biotechnol. 2015;33:736–42.

54. Deutsch EW, Albar JP, Binz P-A, Eisenacher M, Jones AR, Mayer G, et al. Development of data representation standards by the human proteome organization proteomics standards initiative. J. Am. Med. Inform. Assoc. JAMIA. 2015;22:495–506.

55. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol. 2005;6:R9.

56. Deutsch EW, Sun Z, Campbell D, Kusebauch U, Chu CS, Mendoza L, et al. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. J. Proteome Res. 2015;14:3461–73.

57. Masso M, Vaisman II. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng. Des. Sel. PEDS. 2010;23:683–7.

58. Zhou H, Gao M, Skolnick J. ENTPRISE: An Algorithm for Predicting Human Disease-Associated Amino Acid Substitutions from Sequence Entropy and Predicted Protein Structures. PloS One. 2016;11:e0150965.

59. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. 2006;34:W239-242.

60. Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014;42:W314-319.

61. Witvliet DK, Strokach A, Giraldo-Forero AF, Teyra J, Colak R, Kim PM. ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinforma. Oxf. Engl. 2016;32:1589–91.

62. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382-388.

63. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33:W306-310.

64. Savojardo C, Fariselli P, Martelli PL, Casadio R. INPS-MD: a web server to predict stability of protein variants from sequence and structure. Bioinforma. Oxf. Engl. 2016;32:2542–4.

65. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinforma. Oxf. Engl. 2014;30:335–42.

66. Deutsch C, Krishnamoorthy B. Four-body scoring function for mutagenesis. Bioinforma. Oxf. Engl. 2007;23:3009–15.

67. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006;62:1125–32.

68. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011;79:830–8.

69. Getov I, Petukh M, Alexov E. SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. Int. J. Mol. Sci. 2016;17:512.

70. Worth CL, Preissner R, Blundell TL. SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. 2011;39:W215-222.

71. De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, et al. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res. 2012;40:D935-939.

72. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. 2013;41:W333-339.

73. Brender JR, Zhang Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLoS Comput. Biol. 2015;11:e1004494.

74. Pires DEV, Ascher DB. mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res. 2016;44:W469-473.

75. Li M, Simonetti FL, Goncearenco A, Panchenko AR. MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. Nucleic Acids Res. 2016;44:W494-501.

76. Petukh M, Dai L, Alexov E. SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations. Int. J. Mol. Sci. 2016;17:547.

77. Xiong P, Zhang C, Zheng W, Zhang Y. BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. J. Mol. Biol. 2017;429:426–34.

78. Pires DEV, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci. Rep. 2016;6:29575.

79. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34:D204-206.

80. Moal IH, Fernández-Recio J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinforma. Oxf. Engl. 2012;28:2600–7.

81. Pires DEV, Blundell TL, Ascher DB. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res. 2015;43:D387-391.

82. Luu T-D, Rusu A, Walter V, Linard B, Poidevin L, Ripp R, et al. KD4v: Comprehensible Knowledge Discovery System for Missense Variant. Nucleic Acids Res. 2012;40:W71-75.

83. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat. Methods. 2010;7:248–9.

84. Preeprem T, Gibson G. SDS, a structural disruption score for assessment of missense variant deleteriousness. Front. Genet. 2014;5:82.

85. Gress A, Ramensky V, Büch J, Keller A, Kalinina OV. StructMAn: annotation of single-nucleotide polymorphisms in the structural context. Nucleic Acids Res. 2016;44:W463-468.

86. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J. Mol. Biol. 2014;426:2692–701.

87. Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics. 2011;12 Suppl 4:S3.

88. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514-517.

89. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. Poznan Pol. 2015;19:A68-77.

90. Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. Hum. Mutat. 2013;34:42–9.

91. Schaafsma GCP, Vihinen M. VariSNP, a benchmark database for variations from dbSNP. Hum. Mutat. 2015;36:161–6.

92. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. 2017;9:4.

93. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. Bioinforma. Oxf. Engl. 2009;25:1431–2.

94. Baeissa H, Benstead-Hume G, Richardson CJ, Pearl FMG. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. Oncotarget. 2017;8:21290–304.

95. Wang D, Song L, Singh V, Rao S, An L, Madhavan S. SNP2Structure: A Public and Versatile Resource for Mapping and Three-Dimensional Modeling of Missense SNPs on Human Protein Structures. Comput. Struct. Biotechnol. J. 2015;13:514–9.

96. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

97. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 2013;41:D483-489.

98. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 2016;3:160018.

99. Variant Call Format | 1000 Genomes [Internet]. [cited 2017 Aug 23]. Available from: http://www.internationalgenome.org/wiki/Analysis/variant-call-format

100. Specifications of SAM/BAM and related high-throughput sequencing file formats [Internet]. [cited 2017 Aug 23]. Available from: https://github.com/samtools/hts-specs

101. Variant annotations in VCF format [Internet]. [cited 2017 Aug 23]. Available from: http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf

102. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum. Mutat. 2016;37:564–9.