VIVID: A Web Application for Variant Interpretation and Visualization in Multi-dimensional Analyses

Swapnil Tichkule [1,2]*, Yoochan Myung [3,4], Myo T. Naung,[1,2] Brendan R.E. Ansell,[1] Andrew J. Guy,[5] Namrata Srivastava,[6] Somya Mehra,[7] Simone M. Cacciò,[8] Ivo Mueller,[1] Alyssa E. Barry,[7,9] Cock van Oosterhout [10], Bernard Pope [11,12,13,14], David B. Ascher [3,4] and Aaron R. Jex[1,15]

[1]Population Health and Immunity, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
[2]Department of Medical Biology, University of Melbourne, Melbourne, Australia
[3]Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Australia
[4]Computational Biology and Clinical Informatics, Baker Heart and Diabetes, Melbourne, Australia
[5]School of Science, RMIT University, Melbourne, Australia
[6]Department of Data Science and AI, Monash University, Melbourne, Australia
[7]Life Sciences Discipline, Burnet Institute, Melbourne, Australia
[8]Department of Infectious Disease, Istituto Superiore di Sanità, Rome, Italy
[9]Institute of Mental and Physical Health and Clinical Translation (IMPACT) and School of Medicine, Deakin University, Geelong, Australia
[10]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK
[11]Melbourne Bioinformatics, University of Melbourne, Melbourne, Australia
[12]Australian BioCommons, Sydney, Australia
[13]Department of Clinical Pathology, University of Melbourne, Melbourne, Australia
[14]Department of Surgery (Royal Melbourne Hospital), University of Melbourne, Melbourne, Australia
[15]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Melbourne, Australia

***Corresponding author:** E-mail: tichkule.s@wehi.edu.au.

Associate editor: Emma Teeling

Abstract

Large-scale comparative genomics- and population genetic studies generate enormous amounts of polymorphism data in the form of DNA variants. Ultimately, the goal of many of these studies is to associate genetic variants to phenotypes or fitness. We introduce VIVID, an interactive, user-friendly web application that integrates a wide range of approaches for encoding genotypic to phenotypic information in any organism or disease, from an individual or population, in three-dimensional (3D) space. It allows mutation mapping and annotation, calculation of interactions and conservation scores, prediction of harmful effects, analysis of diversity and selection, and 3D visualization of genotypic information encoded in Variant Call Format on AlphaFold2 protein models. VIVID enables the rapid assessment of genes of interest in the study of adaptive evolution and the genetic load, and it helps prioritizing targets for experimental validation. We demonstrate the utility of VIVID by exploring the evolutionary genetics of the parasitic protist *Plasmodium falciparum*, revealing geographic variation in the signature of balancing selection in potential targets of functional antibodies.

Key words: **protein structure, variant interpretation, multi-dimensional analysis, data visualization, population genetics, evolution.**

Letter

Introduction

The modern explosion of genomics, population genetics, and genome-wide association studies (GWAS) is producing enormous amounts of polymorphism data linking nucleotide variation to phenotypic outcomes (Luo et al. 2011; Uffelmann et al. 2021) or disease (Duncavage and Tandon 2015; Wu et al. 2016; Giannopoulou et al. 2019). Countless informatic and statistical methods have been developed to compile and quantify patterns in genotypic data that can be distilled into biologically meaningful results (Luo et al. 2011; Uffelmann et al. 2021). However, visualization is one of the most powerful methods for pattern recognition because it is closely aligned to our brain's processing of complex information (Bülthoff and Edelman 1992).

Many tools can display (Variation Viewer; ncbi.nlm.nih.gov/variation/view) and predict the impact of individual single nucleotide polymorphisms (SNPs) on protein

sequence and structure for large-scale population comparative studies and GWAS (Glusman et al. 2017). However, few can display the sum of these SNP locations in protein structures in an automated way. Moreover, current 3D mapping and visualization tools are either manually driven (e.g., Chimera; Pettersen et al. 2004) or restricted to specialist applications. cBioPortal (Cerami et al. 2012), COSMIC-3D (cancer.sanger.ac.uk/cosmic3d/), CRAVAT (Douville et al. 2013), MoKCa database (Richardson et al. 2009), and cancer3D (Porta-Pardo et al. 2015) are limited to cancer-related variants. LS-SNP/PDB (Ryan et al. 2009), MuPIT (Niknafs et al. 2013), PopViz (Zhang et al. 2018), and VarMap (Stephenson et al. 2019) are limited to humans. SNP2Structure (Wang et al. 2015) exclusively accesses genetic variation from dbSNP (Sherry et al. 2001). Finally, to predict the mutational effects of noncancer and nonhuman genetic variants on protein structure, users necessarily depend on third-party software and databases. In turn, this requires multiple webservers and input file formats, which is challenging and time-consuming, hampering comparison and accessibility. Furthermore, although existing tools allow annotation of individual SNPs, they do not scale to support the rendering of complex population-level variant data from any given organism or disease.

We present VIVID, a novel interactive and user-friendly platform that automates mapping of genotypic information and population genetic analysis from Variant Call Format files in 2D and 3D protein-structural space. This platform provides an easy user experience and an integrated analysis environment to yield both individual- and population-level insights, while generalizing to any organism or disease.

New Approach

VIVID is a unique ensemble user interface that enables users to explore and interpret the impact of genotypic variation on the phenotypes of secondary and tertiary protein structures. Using data from a standard VCF, VIVID integrates published algorithms, programs, and databases to map, annotate, analyze, and visualize the effects of mutations on primary protein sequence, 2D protein residue interactions, and a variety of 3D protein-structural renderings at individual and population scales. Figure 1 provides a schematic workflow for VIVID, which consists of three key components: input, data preprocessing, and visualization. Overall, VIVID allows the integration of multiple analyses within one interactive visualization interface, adding new dimensions to variant annotation and functional effect prediction for any organism in any system.

Input Source

VIVID requires four main inputs in the submission page of the webserver. (1) The complete nucleotide coding sequence of a gene in FASTA format. (2) A general feature format (GFF) file to extract genomic coordinates of the queried coding sequence and to map SNPs from (3) a VCF file on (4) a protein structure (PDB file), which can either be retrieved on request through VIVID from the AlphaFold2 predicted protein structure database (Jumper et al. 2021) or provided by users from any source, whether in silico predicted (Roy et al. 2010; Kelley et al. 2015) or accessed via the RCSB Protein Data Bank API (https://www.rcsb.org) (Berman et al. 2000). VIVID also allows users to select codon usage preference by selecting appropriate "genetic code" from the drop-down menu (default: standard code).

Data Preprocessing: web-server Mapping Mutations in 1D–3D

SNPs can be identified and explored within the VIVID interface based on their location in 1D linear display, 2D residue contact maps, and 3D protein-structural renderings. First, the 1D display is generated by mapping SNPs from the VCF file to the primary protein sequence based on the input GFF file and highlighted in the protein sequence viewer. Next, VIVID uses atomic coordinates from the PDB file to calculate the Euclidian distance between each pair of residues and display each mutated amino acid in the context of long-range residue contacts. Finally, VIVID renders each mutated amino acid in the 3D protein structure using NglViewer (Rose et al. 2018).

Data Preprocessing: web-server Annotation

Each mutated amino acid is classified as nonsynonymous or synonymous using an appropriate genetic code selected by the user. Protein sequences and structural renderings can also link to functional and domain-based annotations using data retrieved from UniProt (The UniProt Consortium 2021) and Pfam (Mistry et al. 2021), if available.

Data Preprocessing: Mutational Analyses

VIVID predicts the likely effects of substituted amino acids on protein structure using previously published algorithms, tools, and APIs. Firstly, MODELLER is used to model nonsynonymous mutations (Fiser and Šali 2003). Dynamut2 (Rodrigues et al. 2020) is then used to predict the effects of missense mutations on protein structure stability and flexibility. Arpeggio (Jubb et al. 2017) is further used to calculate interatomic interactions of mutated residues in 3D space. Amino acid conservation score of mutated residues is computed using the Position-Specific Scoring Matrix (PSSM) of PSI-BLAST (Altschul et al. 1997).

BCFtools (http://samtools.github.io/bcftools/howtos/install.html) is used to calculate allele frequencies of each alternate allele from the VCF file. It generates consensus sequences for each sample within VCF file by using a user-supplied coding sequence as a reference, then mapping population genetic observations on protein structures. Finally, a 3D sliding window-based application implemented in BioStructMap (Guy et al. 2018) is used

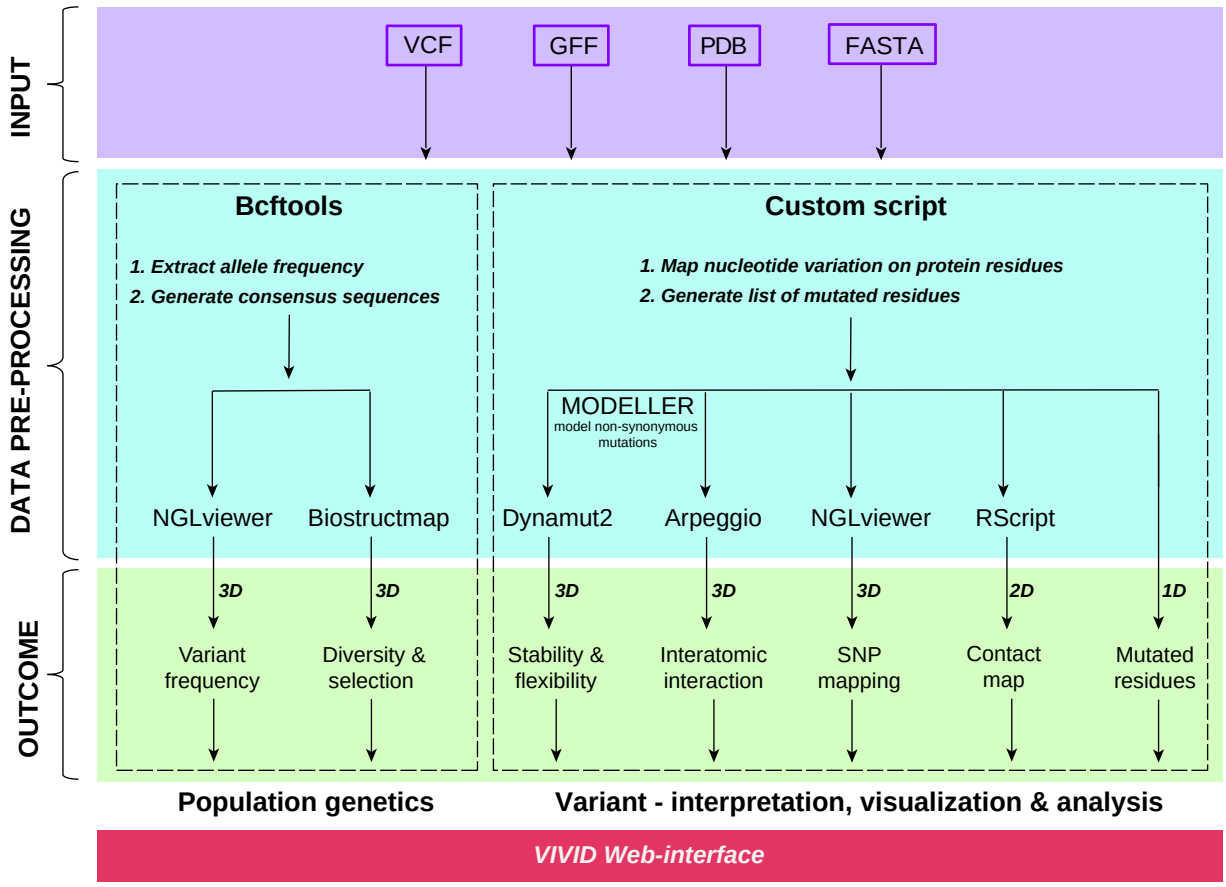


Fig. 1. Schematic workflow of VIVID. The back-end of the VIVID webserver consists of a series of tools for population genetics and mutational analysis. The information of allele frequency in the population and consensus sequence of each sample in the population are extracted from user provided VCF file by using Bcftools v1.8 and are further used by BioStructmap v.0.4.1 to map population genetic indices onto the 3*D* protein structure. The 3*D* structure of nonsynonymous mutations is modeled by MODELLER 10.1 and their effects on protein stability and interatomic interactions are evaluated by Dynamut2 and Arpeggio, respectively. All analysis outcomes (variant frequency, stabilizing/destabilizing effects, interatomic interactions, conservation score, diversity, selection, mutations, etc.) are visualized on 3*D* protein structure using NGLviewer. The front-end of the VIVID webserver was designed with Materialize v1.0.0 and the back-end was based on Python 2.7 via Flask Framework version v1.0.2 on a Linux server running Apache.

to map nucleotide diversity (π) and Tajima's D onto the protein models. This enables users to identify possible links between the effects of natural selection and the structural or biochemical changes in the protein.

Data Preprocessing: web-server Visualization

Nglviewer is used to visualize protein structures in an interactive mode. It also allows the user to map information onto rendered protein structures by selecting options including nonsynonymous and synonymous mutations, PSSM conservation scores, nucleotide diversity, Tajima's D, SNP frequency, etc. In addition, user-selected residues associated with functional and structural domains (retrieved from Uniprot and Pfam databases) in the primary sequence viewer can also be rendered on the protein structure.

Results

The outcome of VIVID in figure 1 is represented in seven panels on the results page of the webserver (fig. 2).

External Information

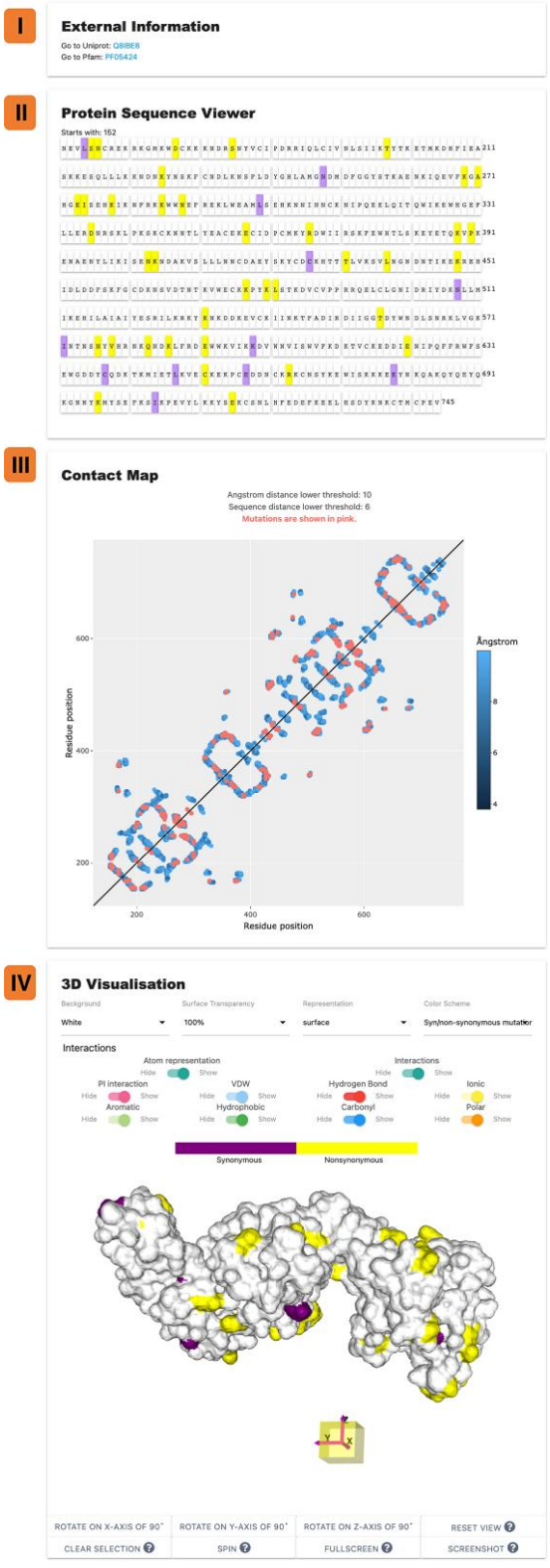
The first panel provides an external link for UniProt and Pfam to allow users to render structural and functional domain information (e.g., conserved structural domains, active sites, etc.) on protein sequence and structure to identify mutational hotspots.

Protein Sequence Viewer

The second panel displays the translated protein sequence of the residues present in the protein structure where synonymous (purple) and nonsynonymous (yellow) mutations are highlighted by default. Users can also click and select protein residues to view them in 3D space in the visualization panel.

Contact Map

The third panel displays an interactive protein contact map where all pairs of residues mapping within a user-defined Euclidean distance threshold (default: 10 Ångstrom) in 3D-space and found more than six



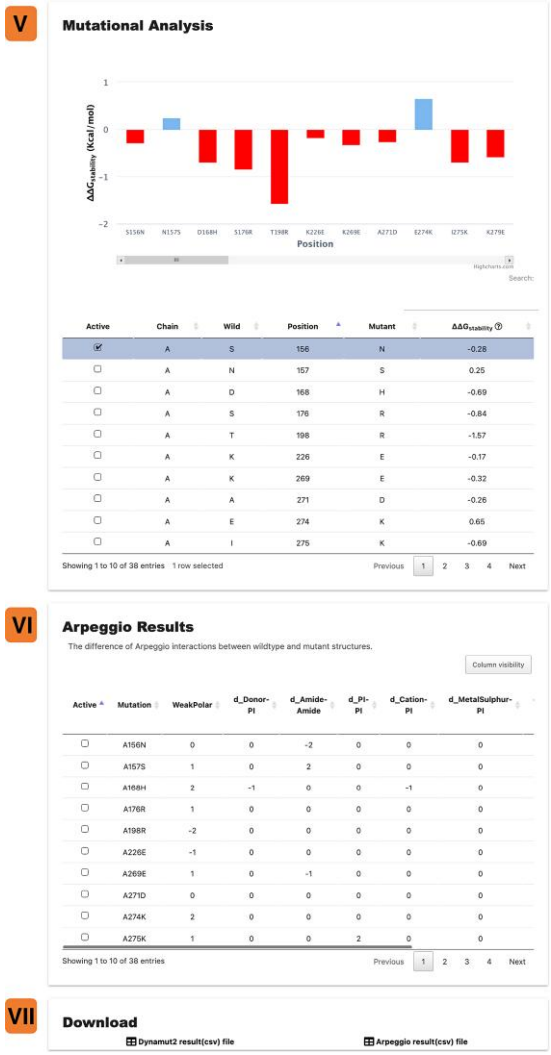


Fig. 2. Seven panels showing the VIVID output on the results page. These panels show the results of integrative analyses of SNPs (data derived from the MalariaGen Pf3*k* version5 dataset) on the EBA175 RII protein.

(default) amino acids apart in the primary sequence are displayed. This highlights long-range contacts that are important for stabilizing protein structure (Vendruscolo et al. 1997; Toth-Petroczy et al. 2016), with pair-wise contacts involving (and potentially disrupted by) mutated residues highlighted in pink.

Fig. 2. Continued

3D Visualization

In the fourth panel, users can visualize the protein structure with different representations and coloring options. Users can map and visualize mutations, annotations, and analyses on the 3D model to detect hotspots of genetic variation by selecting various coloring options. Color coding options include hydrophobicity, non/synonymous mutations, PSSM conservation score, changes of folding free energy ($\Delta\Delta G$), nucleotide diversity, Tajima's D, etc.

Mutational Analysis

The fifth panel displays results from mutational analyses performed using Dynamut2 (Rodrigues et al. 2020). It provides a bar-chart of predicted changes of folding free energy ($\Delta\Delta G$) of substituted amino acids on protein

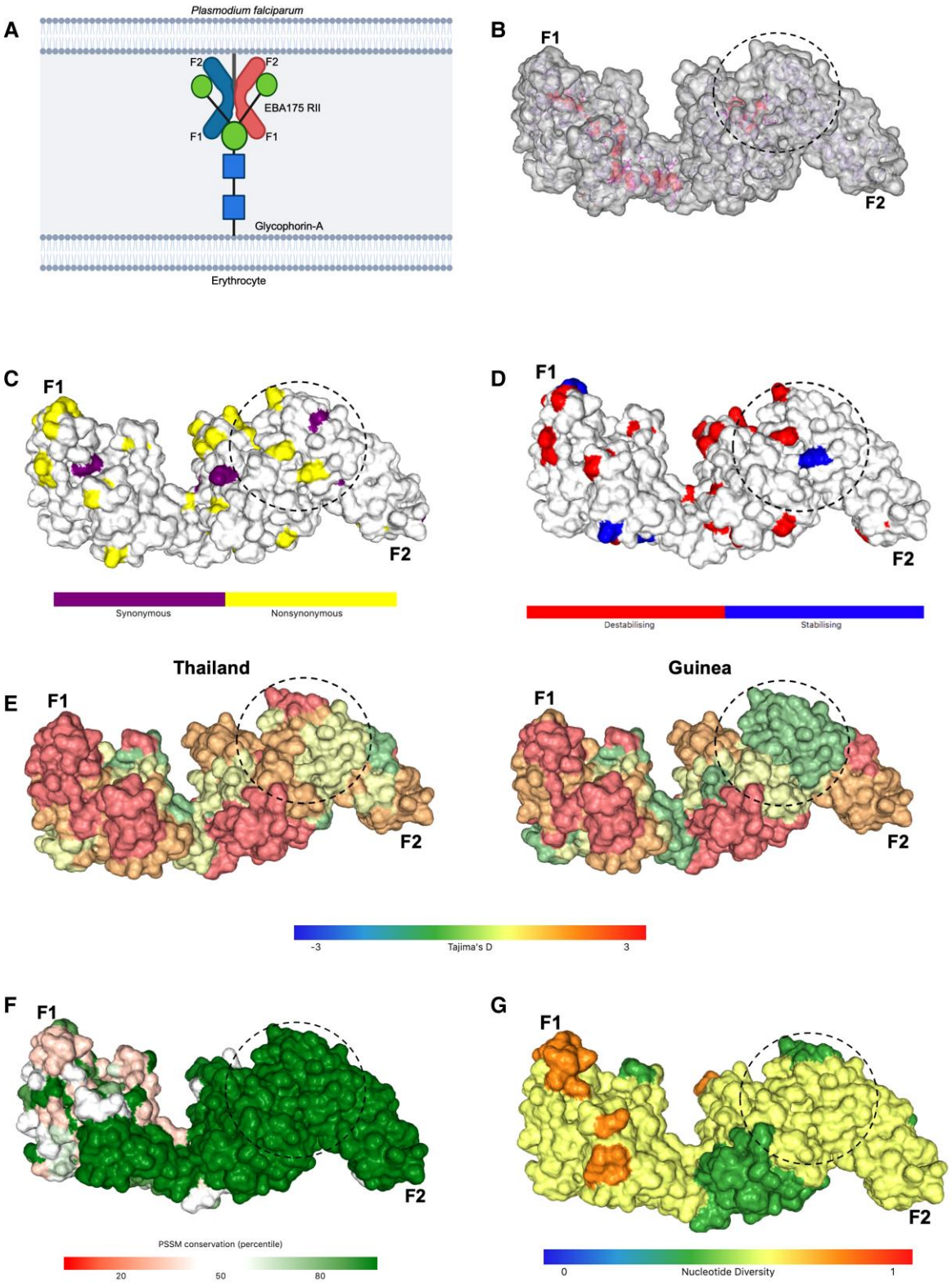


Fig. 3. Analyses of EBA175 RII polymorphisms from two different populations—Thailand (Asia) and Guinea (Africa) using VIVID. A part of the F2 domain (circle) illustrates distinct balancing selection between Thailand and Guinea populations. Unless otherwise stated, only the results from the Thailand population are shown. (*A*) Schematic diagram of the dimerized EBA175 RII ligand binding to glycophorin A human receptor. (*B*) The highlighted residues (circle) indicated the target of known inhibitory antibodies from previous studies (Ambroggio et al. 2013; Chen et al. 2013). (*C*) Map of nonsynonymous and synonymous mutations on EBA175 RII. (*D*) Map of ΔΔG values on EBA175 RII. (*E*) Map of spatial Tajima's D on EBA175 RII. (*F*) Map of PSSM scores on EBA175 RII. (*G*) Map of nucleotide diversity on EBA175 RII.

structure stability and flexibility. The $\Delta\Delta G$ values are also reported in tabular format.

Arpeggio Results

The sixth panel shows interatomic interactions between the substituted residue and nearby residues in protein structures. It reports differences in 20 interatomic interactions between wild-type and mutated residues in tabular format. It is useful to get information about the lost and gained interaction after substitution in a protein structure.

Download

The seventh panel allows users to download Dynamut2 and Arpeggio results.

Case Study

Here, we demonstrate the utility of VIVID and its features by exploring the evolution of vaccine candidates within the parasitic protist *Plasmodium falciparum*, one of the world's primary causes of malaria (Prugnolle et al. 2011). Strain-specific immunity hinders vaccine development, and vaccine escape is likely to occur at sites under balancing selection. Drawing on an ensemble of published tools and algorithms, VIVID provides an intuitive and user-friendly interface to analyze functional divergence in more depth. We used SNP data from the published genomes (MalariaGEN Pf3K v5.1) of naturally occurring *P. falciparum* infections from Guinea ($n = 100$) and Thailand ($n = 148$). We filtered these data for SNPs mapping to region II (AA residues: N152–V745) of the erythrocyte binding antigen protein (EBA175 RII) using the *P. falciparum* 3D7 genome (PlasmoDB release v43). EBA175 mediates binding to human receptor glycophorin A during merozoite invasion of the RBC (Tolia et al. 2005) (fig 3*A*). Region II (RII) of EBA175 is the functional binding domain consisting of two cysteine-rich Duffy binding-like domains called F1 and F2 (Sim et al. 1994). These F1 and F2 domains mediate dimerization of two EBA-175 proteins through disulphide bond formation (Tolia et al. 2005), allowing *P. falciparum* merozoites to bind to the erythrocyte surface (fig 3*A*) during blood-stage invasion (Tham et al. 2012). Antibodies that block dimerization of EBA175 may inhibit glycophorin A binding, which is associated with protection from clinical malaria (Chen et al. 2013; Irani et al. 2015) (see residues highlighted in magenta in fig 3*B*). Most (>90%) of the SNPs mapping to this region in the Guinean (Africa) and Thailand (Asia) populations were nonsynonymous (fig. 3*C*). Spatially derived Tajima's D analysis in VIVID identified a large region of the F1 domain (AA residues E226–L294, I312–K324, and W377–I400) of EBA175 RII as being under balancing selection (i.e., positive Tajima's D values) in both populations (fig. 3*E*). Some of the F1 domain mutations, such as E274K, L482V (also under balancing selection), were predicted to stabilize the parasite protein (i.e., positive $\Delta\Delta G$) (fig. 3*D*). Such polymorphisms can be maintained at intermediate frequencies to help the parasite escape host immune responses while maintaining a functional protein. A second region within the F2 domain (AA residues C551–C591 and Y731–F743) was also identified under balancing selection, but this was specific to the Thailand population (circles in fig. 3*E*). Some of these residues with the F2 domain are targets of functional antibodies (Ambroggio et al. 2013; Chen et al. 2013) (fig 3*B*). In contrast, the F1 domain has high nucleotide diversity, less amino acid (low PSSM) conservation, and is under balancing selection in both the populations (fig 3*E*–*G*). Therefore, the results suggest geographically variable effectiveness of the EBA175 RII vaccine if the formulation is based on a single reference strain.

Conclusion

We developed a web-based application, VIVID, to support improved visualization, analysis, and understanding of how mutations impact protein structure and function in any organism, using a selection of standardized input format files. This method allows users to examine mutations from individuals or populations in multiple dimensions for solved or in silico predicted protein structures, which can either be imported directly from RCSB or AlphaFold2, or uploaded by the user. VIVID can assess SNPs for their impact on protein structure, inferred function, and surface biochemistry (e.g., hydrophobicity) using individual-based data. For population studies, in addition to these features, VIVID can use the SNP data to color any protein model based on localized differences in population genetic metrics. This enables users to visualize protein evolution in 3D and identifies genomic regions under selection. The architecture of this browser-based tool is designed to benefit the broader scientific community, provide an easy user experience, and offer an integrated analysis environment. Our web portal integrates features from multiple programs, algorithms, and databases, providing users with a single platform to explore and interpret variants in multiple dimensions. In addition, the user-friendly and interactive panels allow downloading multiple result outputs that can be used downstream in other programs. In this paper, we have demonstrated the utility of VIVID by using a case study of the evolution of vaccine candidates against malaria. This will be a valuable resource to the research community working in structural and functional proteomics and genomics and will have numerous valuable applications.

Future Developments

We are in the process of expanding the usability of this application by integrating additional features from the available wealth of bioinformatics resources. These include widely used variant repositories for model-organisms, annotation databases in addition to UniProt, and additional protein functional properties to map on 3D protein structures. Moreover, we plan to integrate other developed algorithms and tools from our Biosig server (http://biosig.unimelb.edu.au/biosig). Furthermore, we are in the process

of expanding VIVID features for population and evolutionary genomics analyses and 3D visualization. Finally, we will make available a standalone package for free download. In addition to the current analyses, this will allow users with sufficient computational resources to scale VIVID to perform proteome-wide structural studies.

Acknowledgements

S.T. acknowledges the Australian Society for Parasitology (ASP) for a student conference travel grant and a JD Smyth Postgraduate Travel Award for a Network Researcher Exchange, Training and Travel Award. In addition, S.T. acknowledges the IT team at the Walter and Eliza Hall Research Institute, Australia.

Funding

This work was supported by the Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme to A.R.J and D.B.A. This work was supported by the National Health and Medical Research Council investigator grant (APP1194330) to A.R.J. This work was supported by the National Health and Medical Research Council investigator grant (GNT1174405) to D.B.A. This work was supported by a Victorian Health and Medical Research Fellowship from the Victorian State Government to B.P. This work was supported by an NHRMC Principal Research Fellowship (GNT1155075) to I.M. A.J.G. was supported by a Vice-Chancellor's Postdoctoral Research Fellowship from RMIT University. S.T. was supported by a Walter and Eliza Hall International PhD Scholarship. S.T., Y.M., and M.T.N were further supported by a Melbourne Research Scholarship.

Data Availability

The server is publicly available at web-server http://biosig.unimelb.edu.au/vivid along with the documentation. The SNP dataset used in the case study was obtained from (MalariaGEN Pf3K [version 5.1], https://www.malariagen.net/data/pf3k-5).

References

The UniProt Consortium. 2021. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. **49**:D480–D489.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**:3389–3402.

Ambroggio X, Jiang L, Aebig J, Obiakor H, Lukszo J, Narum DL. 2013. The epitope of monoclonal antibodies blocking erythrocyte invasion by *Plasmodium falciparum* map to the dimerization and receptor glycan binding sites of EBA-175. *PLoS One*. **8**:e56326.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res*. **28**:235–242.

Bülthoff HH, Edelman S. 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc Natl Acad Sci U S A*. **89**:60–64.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, *et al.* 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. **2**:401–404.

Chen E, Paing MM, Salinas N, Sim BK, Tolia NH. 2013. Structural and functional basis for inhibition of erythrocyte invasion by antibodies that target *Plasmodium falciparum* EBA-175. *PLoS Pathog*. **9**:e1003390.

Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R. 2013. CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics*. **29**:647–648.

Duncavage EJ, Tandon B. 2015. The utility of next-generation sequencing in diagnosis and monitoring of acute myeloid leukemia and myelodysplastic syndromes. *Int J Lab Hematol*. **37**(Suppl 1):115–121.

Fiser A, Šali A. 2003. Modeller: generation and refinement of homology-based protein structure models, editors. *Methods in enzymology*. **374**:461–491.

Giannopoulou E, Katsila T, Mitropoulou C, Tsermpini E-E, Patrinos GP. 2019. Integrating next-generation sequencing in the clinical pharmacogenomics workflow. *Front Pharmacol*. **10**:384.

Glusman G, Rose PW, Prlić A, Dougherty J, Duarte JM, Hoffman AS, Barton GJ, Bendixen E, Bergquist T, Bock C, *et al.* 2017. Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. *Genome Med*. **9**:113.

Guy AJ, Irani V, Richards JS, Ramsland PA. 2018. Biostructmap: a Python tool for integration of protein structure and sequence-based features. *Bioinformatics*. **34**:3942–3944.

Irani V, Guy AJ, Andrew D, Beeson JG, Ramsland PA, Richards JS. 2015. Molecular properties of human IgG subclasses and their implications for designing therapeutic monoclonal antibodies against infectious diseases. *Mol Immunol*. **67**:171–182.

Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. 2017. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol*. **429**:365–371.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. **10**:845–858.

Luo L, Boerwinkle E, Xiong M. 2011. Association studies for next-generation sequencing. *Genome Res*. **21**:1099–1108.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, *et al.* 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res*. **49**:D412–D419.

Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, Cooper DN, Karchin R. 2013. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Hum Genet*. **132**:1235–1243.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. **25**:1605–1612.

Porta-Pardo E, Hrabe T, Godzik A. 2015. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res*. **43**:D968–D973.

Prugnolle F, Durand P, Ollomo B, Duval L, Ariey F, Arnathau C, Gonzalez J-P, Leroy E, Renaud F. 2011. A fresh look at the origin of *Plasmodium falciparum*, the most malignant malaria agent. *PLoS Pathog.* **7**:e1001283–e1001283.

Richardson CJ, Gao Q, Mitsopoulous C, Zvelebil M, Pearl LH, Pearl FMG. 2009. MoKCa database—mutations of kinases in cancer. *Nucleic Acids Res.* **37**:D824–D831.

Rodrigues CHM, Pires DEV, Ascher DB. 2020. Dynamut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* **30**:60–69.

Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. 2018. NGL Viewer: web-based molecular graphics for large complexes. *Bioinformatics*. **34**:3755–3758.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* **5**:725–738.

Ryan M, Diekhans M, Lien S, Liu Y, Karchin R. 2009. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics*. **25**:1431–1432.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**:308–311.

Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH. 1994. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science.* **264**:1941–1944.

Stephenson JD, Laskowski RA, Nightingale A, Hurles ME, Thornton JM. 2019. Varmap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics*. **35**: 4854–4856.

Tham WH, Healer J, Cowman AF. 2012. Erythrocyte and reticulocyte binding-like proteins of *Plasmodium falciparum*. *Trends Parasitol*. **28**:23–30.

Tolia NH, Enemark EJ, Sim BK, Joshua-Tor L. 2005. Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. *Cell*. **122**:183–193.

Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS. 2016. Structured states of disordered proteins from genomic sequences. *Cell*. **167**:158–170.e112.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. 2021. Genome-wide association studies. *Nat Rev Methods Primers*. **1**:59.

Vendruscolo M, Kussell E, Domany E. 1997. Recovery of protein structure from contact maps. *Fold Des*. **2**:295–306.

Wang D, Song L, Singh V, Rao S, An L, Madhavan S. 2015. SNP2Structure: a public and versatile resource for mapping and three-dimensional modeling of missense SNPs on human protein structures. *Comput Struct Biotechnol J.* **13**:514–519.

Wu J, Wu M, Chen T, Jiang R. 2016. Whole genome sequencing and its applications in medical genetics. *Quant Biol*. **4**:115–128.

Zhang P, Bigio B, Rapaport F, Zhang S-Y, Casanova J-L, Abel L, Boisson B, Itan Y. 2018. Popviz: a webserver for visualizing minor allele frequencies and damage prediction scores of human genetic variations. *Bioinformatics*. **34**:4307–4309.