



Understanding large scale sequencing datasets through changes to protein folding

David Shorthouse , Harris Lister, Gemma S. Freeman and Benjamin A. Hall 

Corresponding author: Benjamin A. Hall. Tel.: +44 20 76798418; E-mail: b.hall@ucl.ac.uk

Abstract

The expansion of high-quality, low-cost sequencing has created an enormous opportunity to understand how genetic variants alter cellular behaviour in disease. The high diversity of mutations observed has however drawn a spotlight onto the need for predictive modelling of mutational effects on phenotype from variants of uncertain significance. This is particularly important in the clinic due to the potential value in guiding clinical diagnosis and patient treatment. Recent computational modelling has highlighted the importance of mutation induced protein misfolding as a common mechanism for loss of protein or domain function, aided by developments in methods that make large computational screens tractable. Here we review recent applications of this approach to different genes, and how they have enabled and supported subsequent studies. We further discuss developments in the approach and the role for the approach in light of increasingly high throughput experimental approaches.

Keywords: genomics; mutation; folding; DDG; structure

INTRODUCTION

DNA sequencing has revolutionized biomedical research. The ability to routinely sequence samples with increasing sensitivity and continually reducing costs has enabled major new insights into diseases and ageing. In turn, this has led to a growth in the routine use of sequencing in the clinic; for example, in the UK, the NHS now routinely sequences at risk prenatal foetuses [1] for mutations in a panel of genes known to cause disorders, and this type of testing is now taking place in several countries [2]. This collection of data is likely to become increasingly widespread as costs continue to decline, and with that become more available in lower-middle income countries. In principle this deluge of data should improve our understanding of disease, as we become able to link variants to specific clinical phenotypes. However, a fundamental issue with genomic variants is in their interpretability. It is estimated that roughly two-thirds of disease causing variants are nucleotide substitutions, causing either a truncation of the protein (e.g. through introducing a termination codon, or a splice site) or a substitution of amino acids in the protein sequence (missense mutations) [3]. Missense mutations can cause either loss or gain of resultant protein activity, or not alter gene function at all. Whilst some missense mutations occur reliably in hotspots, the wide diversity of variants that present in the clinic lead to problems of interpreting variants of uncertain significance (VUS). This is a problem for relatively common diseases, such as the RASopathy Noonan's syndrome (affecting 1 in 2000 individuals), where almost 60% of prenatally detected mutations in associated

genes were found to be VUS [4], but presents an even more serious barrier for diagnosis of rare diseases. Given that the frequency of rare diseases as a cohort is high, affecting 3.5–5.9% of individuals globally [5], interpretation and confirmation of the role of a missense mutation in disease is paramount. This is of particular importance for genes that are not tractable to assay (e.g. many membrane proteins), where the commonly used assay does not reflect all aspects of the disease (e.g. assays for fumarate hydratase function, [6]), or where collection of materials is not possible (e.g. RASopathies, which can present prenatally [7]).

Bioinformatic and genomic approaches offer one route to understanding such variants. FATHMM [8], GenePy [9] and PolyPhen-2 [10] are examples of statistical tools to predict pathogenicity of variants, using machine learning algorithms trained on recorded observations. Whilst these approaches are powerful, they also share common limitations. Whilst they can illustrate statistical correlations between variant sites and types and disease, they do not offer biophysical mechanisms for variant action. This is necessary for understanding the type of perturbation caused by the variant— an 'edgetic' modification would influence a specific function or protein interaction, whilst 'global' modifications would alter overall levels of gene activity on all downstream elements [11]. The type of mutation is associated with specific clinical phenotypes or cellular behaviours, and understanding this is therefore necessary for predicting clinical outcomes. Understanding specific mechanisms can also provide routes to drug development. For example, mutations to the gene

David Shorthouse is a lecturer at UCL using data science to study cancer patient heterogeneity.

Harris Lister and Gemma Freeman are undergraduate students currently studying Biomedical Engineering and Infection and Immunity at UCL, respectively. Ms Freeman is a Laidlaw Scholar.

Benjamin A. Hall is a principal research fellow at UCL, using computational models to study how mutations alter cells in healthy ageing and carcinogenesis.

Received: October 8, 2023. Revised: February 26, 2024. Accepted: March 1, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Methods for computing the energy change of folding on mutation

Method	Computational Cost for Saturation Screen	Comments
Forcefield-based methods (e.g. Foldx)	Low – 10-100 s of CPU hours per structure	High efficiency, easy to parallelize, based on a set of energy parameters (forcefield), no incorporation of dynamics
Ensemble-based methods (e.g. Rosetta FlexDDG)	Medium – 1000-10 000 s of CPU hours per structure	Medium efficiency, possible to parallelize, collects ensembles of structures from small-scale dynamics calculations to generate energies
Dynamics-based methods (e.g. Alchemical methods)	High – >10 000 s of CPU hours per structure	Poor efficiency for high-throughput, requires extensive CPU time for each mutation, can be parallelized, mutates a wildtype protein over a small molecular dynamics simulation

and one predicted destabilizing, were validated *in vitro*, demonstrating that the predictions were correct. The authors further identified a mutation at the same site of a known deleterious variant that was expected to be stable, and confirmed experimentally that this alternative missense mutation was indeed tolerated. This work was followed up by Figueiredo, who adapted similar experimental and theoretical approaches to study one missense variant [27]. However, arguably the most substantial impact of this notable work is the impact on assessment and curation guidelines for mutations of *CDH1* ([28], reviewed in [29]). These explicitly used the evidence taken from this work to alter the rules for clinical classification of *CDH1* variants based on predictions from protein folding calculations. This highlights the potential value of computational analysis such as this in assessing VUS when taken alongside other data.

Retinitis pigmentosa (RP) is one of the most common retinopathies, causing progressive loss of peripheral and night vision, which can, in turn, lead to loss of central vision [30]. This is frequently caused by mutations to rhodopsin. Rakoczy *et al.* studied 103 known RP causing mutations in rhodopsin with a mixture of structural bioinformatics tools, including misfolding calculations [31]. A total of 62 of the variants were found to cause protein misfolding, whilst others could be understood in terms of known protein biology or membrane insertion. Notably they observed a clear correlation between calculated $\Delta\Delta G$ values and both vision loss onset and average age of night blindness, suggesting that the severity of the disorder could be understood through the relative destabilization of the protein fold. This work led to several other studies, including experimental exploration of small groups of mutations [32], and high throughput assays that were directly compared with data, challenging some of the conclusions drawn [33]. Perhaps some of the most impactful work that arises from this study however is the insight that treating the misfolding directly might reverse some or all of the symptoms. The use of small chaperone proteins as a therapy was computationally explored in [34], and the wider set of treatments for RP (including chaperone therapy) reviewed in [35].

Fumarate hydratase is an oncoenzyme, whose loss is associated with hereditary leiomyomatosis and renal cell cancer (HLRCC) [36]. This is a late-onset disease, for whom improved mutant classification could support clinical diagnosis of VUS, but in principle routine sequencing could identify individuals at higher risk. Shorthouse *et al.* [21] used misfolding alongside other measures of biophysical properties in the computational equivalent of a multiplexed assay for variant effects (MAVE) to determine the protein structural features that drove disease

based on publicly available data from the fumarate hydratase database [37]. This classifier was able to identify three structural features that correctly predicted the impact of mutations-misfolding (accounting for 2/3 of deleterious mutations), modifications proximal to the active site and substitutions at hinges in the protein. Notably, the importance of dynamics had not been observed previously and its analysis was enabled by use of elastic network modelling, verified by molecular dynamics. Furthermore, this approach ruled out possible mechanisms speculated in the literature of the involvement of an allosteric 'B' site in this disease [38]. The model was verified against publicly available metabolomics data, and classification and data were made available with publication. This has enabled the reuse of the data in clinical case reports, where novel mutations were assessed and classified using the approach [39]. A more complicated problem for understanding VUS is the known issues in clinical assays used to assess enzyme activity in the clinic, which identify fumarate hydratase deficiency in individuals who do not develop renal cancer [6]. The same computational approach applied to small sets of variants reveals that variants that are identified in the assay but do not lead to cancer have lower misfolding energies than disease causing variants and are distant from the enzyme binding site [40]. This finding is consistent with the variants being sufficiently disruptive to trigger a response in the assay but insufficient to cause more serious disease.

Studies of the COVID spike protein have made extensive use of folding calculations to understand the patterns of evolving pathogenicity. Learning how the spike changes is important both for understanding how COVID adapts to human hosts, but also for its role in immune evasion. One pre-alpha study by Laha *et al.* reported alignments of whole genome sequencing, identifying frequently mutated regions and co-occurrence of mutation pairs between different proteins [41]. Using structures available at the time, models of mutations were constructed using a combination of SWISS-MODEL [26] to model missing loops, and FoldX to introduce point mutations. The impact of mutations were assessed using both FoldX overall energy estimates, alongside empirical approaches that estimated the energy of substitution through summing the impact of specific contacts made and broken and portioning energies. This study made several important contributions that aided the interpretation of later datasets. The identification of the recurrent mutation D614G as a stabilizing substitution supported later work on its impact on viral infectivity [42], and the underlying mechanism of selection was further extended to consider the impact on binding affinity [43]. This observation was used to support several later sequencing studies

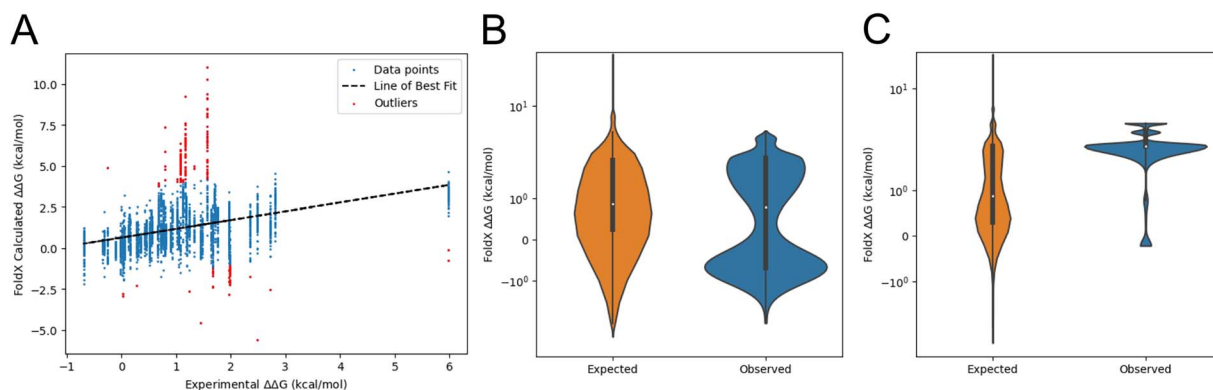


Figure 2: Exploring challenges for prediction in future. **(A)** Estimates of correlation between experimental and calculated values of $\Delta\Delta G$ of mutation can be low for individual proteins, but this may not substantially limit their ability to classify. A systematic analysis of $\Delta\Delta G$ analyses applied to structures of PIN1 shows a low Spearman correlation of 0.46. However, a linear regression highlights that 80% (4/5) outliers (red) result from four substitutions of ~ 800 . **(B)** Gain of function mutations may not show the same enrichment for misfolding as loss of function mutations. PTPN11 $\Delta\Delta G$ estimates here show a distinct distribution from the null hypothesis when applied to the whole gene, with a shift towards stabilization. **(C)** PTPN11 $\Delta\Delta G$ estimates when applied to the autoinhibitory SH2 domain show that misfolding is favoured.

of loss of function versus gain of function mutations observed in Gerasimavivius *et al.* [63].

FUTURE DEVELOPMENTS

As a tool for understanding mutagenesis, misfolding calculations are now well established. As they are used increasingly in exhaustive screens, it is an open question how the technology will develop and continue to be applied in future. One key feature that reflects its maturity is the range of tooling that is becoming available, particularly evident in the analyses of COVID spike protein, as scripts developed by different groups become refined. MutateX [67] is one example of a devoted set of scripts to automate the calculation of saturation screens, validated against several well understood systems. Related to folding calculations, the same tools can further be adapted to calculations of binding affinity, as achieved in RosettaDDG, a Rosetta based pipeline for calculating folding energies and affinity [17]. As high-quality models of protein structure have become more widely available due to alphafold predictions [68, 69], we can further expect more studies that perform calculations on these models, as has been done with neurexin [70]. An additional likely future development is the incorporation of these calculations into existing machine learning methods for variant prediction. As folding energy has been demonstrated to be descriptive for predicting mutation effects, it follows that this energy would be a useful additional feature for these tools. For many frequently mutated genes, there would also be a value in pre-calculating folding energies and making them available through public databases.

Substantial challenges and questions still remain however for the method. One noteworthy feature is that whilst there have been successes in developing folding based classifiers, correlations reported in the literature between theoretical predictions and experimental observations vary widely, for example, 0.2, 0.5 and 0.8 depending on structure [24, 71, 72]. This would suggest that success depends at least in part on the gene of interest and whether its structure is amenable to computational saturation screens (e.g. globular proteins are likely to perform better). We note that correlation, however, may not be an appropriate tool in isolation to measure success, as our comparison of theoretical and experimental estimates from different structures of PIN1 reveals that 80% (4/5) of outlier estimates arise from only 4 of ~ 800

substitutions (Figure 2A). Further systematic analysis of multiple structures, such as [58, 73], would support our understanding of the limitations of FoldX and other predictive tools.

A further issue is how applicable this approach will be for the analysis of pathogenic gain of function mutations. It is intuitive that misfolding can cause loss of gene activity, but there is some evidence that alterations of folding of specific domains may enable gain of function. This was explored in the selection of mutations to *FBXW7* in aged skin [65], which may be expected to have gain of function mutations under positive selection due to its relationship with *NOTCH1* in other tissues. In this situation, apparent selection of stabilizing mutations across the gene was influenced by the strong selection of non-destabilizing mutations to the substrate binding site, and was not apparent once they were excluded. More generally, we might expect that for gain of function mutations, the role of specific domains or regions in gene function becomes more important (as suggested in [63]). For example, mutations to *PTPN11* can lead to different RASopathies through gain of function. Whilst the impact of mutation on misfolding across the whole protein sequence shows a selection for stabilization of the protein (Figure 2B), mutation to the autoinhibitory SH2 domains shows strong selection of destabilizing mutations (Figure 2C). Dedicated studies of well characterized gain of function mutations in single genes will better illuminate the utility of folding calculations here.

A more complicated question is what the role of these folding calculations is in light of more advanced experimental techniques for measuring folding. Tsuboyama *et al.* recently published a landmark paper presenting cDNA display proteolysis, a technique for rapidly measuring folding stability [74] applied to the study of single and double mutants. This work further presents a uniquely powerful resource for understanding protein stability for a large set of proteins measured under consistent conditions. In the context of this and other folding based deep mutation scans, the question arises—do we need to attempt to predict what we can measure? Just as the widespread availability of high-quality protein structure models from AlphaFold [68] does not negate the value of novel experimentally derived structures, experimental tools to measure protein folding effects of mutations do not negate the utility of computational predictions of folding energy. Theoretical approaches offer unique insights into molecular mechanisms alongside experimental data, and

both schools have a long tradition of being accelerated by the availability of new technologies, through working synergistically together. Two further factors also challenge the primacy of single experimental approaches. Experimental assays have inherent limitations, which have been noted to reflect lab specific experimental conditions [25], and even highly similar but different tissue environments influence patterns of mutagenesis *in vivo* [65]. As such it cannot be assumed that single assays are 'correct' for all situations. Second, imperfect measures (whether experimentally or computationally determined) have been shown to be highly effective, as illustrated by examples cited here that use homology models successfully in the absence of experimental structures [25]. One clear target for combined computational and experimental targets to address is the origin of the apparent gap between statistically derived pathogenicity scores, particularly as recent methods show increasing accuracy [75]. In the longer term, we can expect folding forcefields and calculation methods to improve, and folding and dynamics calculations will continue to offer a uniquely powerful and relevant window into experimentally inaccessible problems.

Key Points

- The estimation of folding energies through general forcefields has become increasingly accurate whilst computing costs have continued to come down.
- This enables the application of computational saturation screens that explore the mutational landscape of different genes.
- For individual genes, this has enabled the creation of predictive screens that support analysis of newly observed variants.
- It further enables powerful aggregate analysis of collections of mutations across multiple genes, revealing fundamental shared features and interactions across networks.

ACKNOWLEDGEMENTS

BAH acknowledges support from the Royal Society (grant no. UF130039). GSF is supported by a Laidlaw scholarship.

AUTHOR CONTRIBUTIONS

David Shorthouse (Writing – review & editing [equal]), Harris Lister (Investigation [supporting], Writing – review & editing [supporting]), Gemma Freeman (Investigation [supporting], Writing – review & editing [supporting]), and Benjamin Hall (Writing – original draft [lead], Writing – review & editing [equal])

REFERENCES

1. Woods J. R21: Rapid prenatal exome sequencing, <https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/r21-rapid-prenatal-exome-sequencing/>.
2. Ravitsky V, Roy M-C, Haidar H, et al. The emergence and global spread of noninvasive prenatal testing. *Annu Rev Genomics Hum Genet* 2021;**22**:309–38.
3. Eichler EE. Genetic variation, comparative genomics, and the diagnosis of disease. *N Engl J Med* 2019;**381**:64–74.
4. Leach NT, Wilson Mathews DR, Rosenblum LS, et al. Comparative assessment of gene-specific variant distribution in prenatal and postnatal cohorts tested for Noonan syndrome and related conditions. *Genet Med* 2019;**21**:417–25.
5. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020;**28**:165–73.
6. Alam NA, Rowan AJ, Wortham NC, et al. Genetic and functional analyses of FH mutations in multiple cutaneous and uterine leiomyomatosis, hereditary leiomyomatosis and renal cancer, and fumarate hydratase deficiency. *Hum Mol Genet* 2003;**12**:1241–52.
7. Dempsey E, Homfray T, Simpson JM, et al. Fetal hydrops – a review and a clinical approach to identifying the cause. *Expert Opin Orphan Drugs* 2020;**8**:51–66.
8. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;**34**:57–65.
9. Mossotto E, Ashton JJ, O'Gorman L, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics* 2019;**20**:254.
10. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
11. Kiel C, Serrano L. Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Mol Syst Biol* 2014;**10**:727.
12. Weidemann F, Jovanovic A, Herrmann K, Vardarli I. Chaperone therapy in Fabry disease. *Int J Mol Sci* 2022;**23**:1887.
13. Fitipaldi H, Franks PW. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022. *Hum Mol Genet* 2023;**32**:520–32.
14. Newport TD, Sansom MSP, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Res* 2018;**47**:D390–7.
15. Hall BA, Halim KBA, Buyan A, et al. Sidekick for membrane simulations: automated ensemble molecular dynamics simulations of transmembrane helices. *J Chem Theory Comput* 2014;**10**:2165–75.
16. Wassenaar TA, Pluhackova K, Moussatova A, et al. High-throughput simulations of dimer and trimer assembly of membrane proteins. The DAFT approach. *J Chem Theory Comput* 2015;**11**:2278–91.
17. Sora V, Laspiur AO, Degn K, et al. RosettaDDGPrediction for high-throughput mutational scans: from stability to binding. *Protein Sci* 2023;**32**:e4527.
18. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* 1991;**222**:67–88.
19. Kotler E, Shani O, Goldfeld G, et al. A systematic p53 mutation library links differential functional impact to cancer mutation pattern and evolutionary conservation. *Mol Cell* 2018;**71**:178–190.e8.
20. Song H, Wu J, Tang Y, et al. Diverse rescue potencies of p53 mutations to ATO are predetermined by intrinsic mutational properties. *Sci Transl Med* 2023;**15**:eabn9155.
21. Shorthouse D, Hall MWJ, Hall BA. Computational saturation screen reveals the landscape of mutations in human fumarate hydratase. *J Chem Inf Comput Sci* 2021Articles ASAP;**61**:1970–80.
22. Wang Z, Moul J. SNPs, protein structure, and disease. *Hum Mutat* 2001;**17**:263–70.
23. 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.

24. Schymkowitz J, Borg J, Stricher F, et al. The FoldX web server: an online force field. *Nucleic Acids Res* 2005;**33**:W382–8.
25. Simões-Correia J, Figueiredo J, Lopes R, et al. E-cadherin destabilization accounts for the pathogenicity of missense mutations in hereditary diffuse gastric cancer. *PLoS One* 2012;**7**:e33783.
26. Bienert S, Waterhouse A, de Beer TA, et al. The SWISS-MODEL repository—new features and functionality. *Nucleic Acids Res* 2017;**45**:D313–d319.
27. Figueiredo J, Mercadillo F, Melo S, et al. Germline CDH1 G212E missense variant: combining clinical, in vitro and in vivo strategies to unravel disease burden. *Cancers (Basel)* 2021;**13**:4359.
28. Lee K, Krempely K, Roberts ME, et al. Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Hum Mutat* 2018;**39**:1553–68.
29. van der Post RS, Vogelaar IP, Carneiro F, et al. Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline CDH1 mutation carriers. *J Med Genet* 2015;**52**:361–74.
30. Massof RW, Finkelstein D. Two forms of autosomal dominant primary retinitis pigmentosa. *Doc Ophthalmol* 1981;**51**:289–346.
31. Rakoczy EP, Kiel C, McKeone R, et al. Analysis of disease-linked rhodopsin mutations based on structure, function, and protein stability calculations. *J Mol Biol* 2011;**405**:584–606.
32. Herrera-Hernández MG, Razzaghi N, Fernandez-Gonzalez P, et al. New insights into the molecular mechanism of rhodopsin retinitis pigmentosa from the biochemical and functional characterization of G90V, Y102H and I307N mutations. *Cell Mol Life Sci* 2022;**79**:58.
33. Wan A, Place E, Pierce EA, Comander J. Characterizing variants of unknown significance in rhodopsin: a functional genomics approach. *Hum Mutat* 2019;**40**:1127–44.
34. Behnen P, Felling A, Comitato A, et al. A small chaperone improves folding and routing of rhodopsin mutants linked to inherited blindness. *iScience* 2018;**4**:1–19.
35. Athanasiou D, Aguila M, Bellingham J, et al. The molecular and cellular basis of rhodopsin retinitis pigmentosa reveals potential strategies for therapy. *Prog Retin Eye Res* 2018;**62**:1–23.
36. Skala SL, Dhanasekaran SM, Mehra R. Hereditary Leiomyomatosis and renal cell carcinoma syndrome (HLRCC): a contemporary review and practical discussion of the differential diagnosis for HLRCC-associated renal cell carcinoma. *Arch Pathol Lab Med* 2018;**142**:1202–15.
37. Bayley J-P, Launonen V, Tomlinson IPM. The FH mutation database: an online database of fumarate hydratase mutations involved in the MCUL (HLRCC) tumor syndrome and congenital fumarase deficiency. *BMC Med Genet* 2008;**9**:20.
38. Picaud S, Kavanagh KL, Yue WW, et al. Structural basis of fumarate hydratase deficiency. *J Inher Metab Dis* 2011;**34**:671–6.
39. Franke K, Vagher J, Boyle J, et al. Rare variant in the fumarate hydratase gene found in patients with clinical features of hereditary leiomyomatosis and renal cell cancer (HLRCC): a case series. *Clin Case Rep* 2022;**10**:e05513.
40. Andreou A, Shorthouse D, Cole Y et al. Refining Renal Cell Carcinoma (RCC) risk predictions for germline FH variants. In: *UK Cancer Genetics Group (UKCGG) Spring Meeting 2023, Leeds*.2023.
41. Laha S, Chakraborty J, Das S, et al. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infect Genet Evol* 2020;**85**:104445.
42. Zhang L, Jackson CB, Mou H, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* 2020;**11**:6013.
43. Bello M, Hasan MK, Hussain N. Energetic and structural basis for the differences in infectivity between the wild-type and mutant spike proteins of SARS-CoV-2 in the Mexican population. *J Mol Graph Model* 2021;**107**:107970.
44. He J, Mei Q, Peng Y, et al. Are the original SARS-CoV-2 novel mutants from in vitro culture able to escape the immune response? *J Med Virol* 2023;**95**:e28931.
45. Gu SH, Song DH, Yun H, et al. Molecular characterization of SARS-CoV-2 from the saliva of patients in the Republic of Korea in 2020. *Health Sci Rep* 2022;**5**:e856.
46. Umair M, Ikram A, Rehman Z, et al. Genomic diversity of SARS-CoV-2 in Pakistan during the fourth wave of pandemic. *J Med Virol* 2022;**94**:4869–77.
47. Teng S, Sobitan A, Rhoades R, et al. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. *Brief Bioinform* 2020;**22**:1239–53.
48. Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. *Bioinformatics* 2008;**24**:2397–8.
49. Deng X, Garcia-Knight MA, Khalid MM, et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell* 2021;**184**:3426–3437.e8.
50. Bæk KT, Mehra R, Kepp KP. Stability and expression of SARS-CoV-2 spike-protein mutations. *Mol Cell Biochem* 2023;**478**:1269–80.
51. De Marco C, Veneziano C, Massacci A, et al. Dynamics of viral infection and evolution of SARS-CoV-2 variants in the Calabria area of southern Italy. *Front Microbiol* 2022;**13**:934993.
52. Akcesme B, Erkal B, Donmez ZY. Structural and functional characterization of SARS-CoV-2 nucleocapsid protein mutations identified in Turkey by using in silico approaches. *Acta Virol* 2023;**67**:59–68.
53. Kim H, Chung SH, Kim HS, et al. Investigation of SARS-CoV-2 lineages and mutations circulating in a university-affiliated hospital in South Korea analyzed using Oxford Nanopore MiniION sequencing. *Osong Public Health Res Perspect* 2022;**13**:360–9.
54. Rhoades R, Sobitan A, Mahase V, et al. In-silico investigation of systematic missense mutations of middle east respiratory coronavirus spike protein. *Front Mol Biosci* 2022;**9**:933553.
55. Sharma P, Kumar M, Tripathi MK, et al. Genomic and structural mechanistic insight to reveal the differential infectivity of omicron and other variants of concern. *Comput Biol Med* 2022;**150**:106129.
56. Pereson MJ, Flichman DM, Martínez AP, et al. Evolutionary analysis of SARS-CoV-2 spike protein for its different clades. *J Med Virol* 2021;**93**:3000–6.
57. Shorthouse D, Hall BA. SARS-CoV-2 variants are selecting for spike protein mutations that increase protein stability. *J Chem Inf Model* 2021;**61**:4152–5.
58. Mehra R, Kepp KP. Structural heterogeneity and precision of implications drawn from cryo-electron microscopy structures: SARS-CoV-2 spike-protein mutations as a test case. *Eur Biophys J* 2022;**51**:555–68.
59. Jalal D, Samir O, Elzayat MG, et al. Genomic characterization of SARS-CoV-2 in Egypt: insights into spike protein thermodynamic stability. *Front Microbiol* 2023;**14**:1190133.
60. Tahsin A, Ahmed R, Bhattacharjee P, et al. Most frequently harboured missense variants of hACE2 across different populations exhibit varying patterns of binding interaction with spike glycoproteins of emerging SARS-CoV-2 of different lineages. *Comput Biol Med* 2022;**148**:105903.
61. Arruda HRS, Lima TM, Alvim RGF, et al. Conformational stability of SARS-CoV-2 glycoprotein spike variants. *iScience* 2023;**26**:105696.

62. Cheng TMK, Goehring L, Jeffery L, et al. A structural systems biology approach for quantifying the systemic consequences of missense mutations in proteins. *PLoS Comput Biol* 2012;**8**: e1002738.
63. Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat Commun* 2022;**13**: 3895.
64. Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. *Sci Rep* 2020;**10**:15387.
65. Hall MWJ, Shorthouse D, Alcraft R, et al. Mutations observed in somatic evolution reveal underlying gene mechanisms. *Commun Biol* 2023;**6**:753.
66. Fowler JC, King C, Bryant C, et al. Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discov* 2021;**11**:340–61.
67. Tiberti M, Terkelsen T, Degn K, et al. MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform* 2022; **23**:bbac074.
68. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
69. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2021;**50**:D439–44.
70. Rhoades R, Henry B, Prichett D, et al. Computational saturation mutagenesis to investigate the effects of Neurexin-1 mutations on AlphaFold structure. *Genes* 2022;**13**:789.
71. Kumar V, Rahman S, Choudhry H, et al. Computing disease-linked SOD1 mutations: deciphering protein stability and patient-phenotype relations. *Sci Rep* 2017;**7**:4678.
72. Song X, Wang Y, Shu Z, et al. Engineering a more thermostable blue light photo receptor *Bacillus subtilis* YtvA LOV domain by a computer aided rational design method. *PLoS Comput Biol* 2013;**9**:e1003129.
73. Sapozhnikov Y, Patel JS, Ytreberg FM, Miller CR. Statistical modeling to quantify the uncertainty of FoldX-predicted protein folding and binding stability. *BMC Bioinformatics* 2023;**24**:426.
74. Tsuboyama K, Dauparas J, Chen J, et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* 2023;**620**:434–44.
75. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492.