Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2025, Volume 5, Page No: 104-125

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Population-Specific Variants in the Human Kinome: Insights from the IndiGen Cohort Reveal Distinct Sequence, Structural, and Pharmacogenomic Implications for Drug Response in India

Andre Pacheco[1], José Cardoso[1], Miguel Faria[1]*, Rui Tavares[1]

[1]Department of Natural Products Biotechnology, Faculty of Pharmacy, University of Algarve, Faro, Portugal.

***E-mail** ✉ miguel.faria.npb@outlook.com

Received: 24 January 2025; **Revised:** 06 May 2025; **Accepted:** 12 May 2025

ABSTRACT

India accounts for over 17% of the global population and exhibits extensive genetic heterogeneity, including many clinically important rare variants distributed among numerous sub-groups. These variants remain underrepresented in widely used international reference datasets such as the 1000 Genomes Project (1KG), which contains comparatively few samples from Indian backgrounds. Such resources are crucial for drug discovery and pharmaceutical research, where population diversity is essential for identifying genetic factors that affect predisposition to adverse drug reactions. In this work, we conducted a qualitative, comparative evaluation of sequence- and structure-level differences between kinase-gene variants documented in the newly released IndiGen database—the most comprehensive curated Indian genome dataset—and those cataloged in 1KG. Kinase genes represent one of the most frequently exploited drug-target classes. The sequence-based comparisons highlighted shared and divergent patterns among populations using nsSNVs and amino acid substitution trends, while structural analyses contrasted IndiGen variants with pathogenic alterations curated in the UniProtKB Humsavar dataset. We assessed how these substitutions influence protein features such as folding stability, hydrophobic interactions, solvent exposure, and hydrogen-bonding networks. Computational docking against known drugs targeting proteins harboring Indian-specific variants revealed several notable shifts in binding affinity attributable to these population-specific alterations. Overall, this study provides an in-depth overview of common variants in the world's second-largest population and explores their implications for sequence properties, structural behavior, and pharmacogenomic relevance. These initial findings, which highlight Indian-specific ADR-linked variants, may support both pre-clinical and post-marketing surveillance workflows aimed at reducing adverse drug events in India.

Keywords: Indian genetic variations, IndiGenome Consortium, Pharmacogenomics, Single nucleotide variants, Docking, Adverse drug reactions

How to Cite This Article: Pacheco A, Cardoso J, Faria M, Tavares R. Population-Specific Variants in the Human Kinome: Insights from the IndiGen Cohort Reveal Distinct Sequence, Structural, and Pharmacogenomic Implications for Drug Response in India. Spec J Pharmacogn Phytochem Biotechnol. 2025;5:104-25. https://doi.org/10.51847/UtxcClcWDU

Introduction

Single nucleotide changes form a key genetic basis for complex traits and human phenotypic diversity [1]. When single-nucleotide variants (SNVs) occur in more than 1% of a population, they are categorized as SNPs. Numerous studies show that SNPs contribute to individual susceptibility to various diseases and drug responses [2]. The allele-frequency distribution of SNVs provides insight into migration patterns, evolutionary history, and population structure [3, 4].

Genetic diversity significantly influences how individuals respond to medications, and populations with different geographic ancestries often display distinct variant frequencies [5-7]. As a result, drug risk profiles vary between groups. Variants that alter the function of common drug targets can influence pharmacokinetics (PK) or pharmacodynamics (PD), potentially resulting in adverse drug reactions.

Most variant data originate from widely used repositories—including the 1000 Genomes dataset [8] and gnomAD [9]—both of which are predominantly Eurocentric. This bias stems from the fact that the majority of Genome-

Wide Association Studies (GWAS) have focused on the European population (78%), followed by Asian (10%), African (2%), Hispanic (1%), and <1% from other groups [10], leaving Indian individuals insufficiently represented. Such imbalances lead to misinformed population-specific disease assessments and leave African and Indian groups under-investigated. These unique SNV patterns may differ substantially from those of overrepresented populations, contributing to disparities in diagnosis and clinical interpretation [11, 12].

Globally, adverse drug reactions (ADRs) remain a major cause of illness and mortality [13]. Genetic variation in pathways involved in drug transport and metabolism has long been associated with differences in drug tolerance and the likelihood of ADRs. Numerous SNV-focused studies have demonstrated that variants can directly influence drug efficacy and toxicity [3, 14].

The Drug–Gene Interaction Database (DGIdb) compiles drug–gene interactions from multiple datasets and publications [15]. The dbSNP repository [16] alone includes 38 million SNP entries, making its maintenance and curation a substantial undertaking. SNVs often guide decisions regarding therapeutic selection and appropriate dosage [2]. Because clinical drug trials are disproportionately conducted in European populations before approval and commercialization [17], ADR patterns observed globally may reflect this bias. Therefore, population-focused pharmacogenomic studies—which integrate variant, gene, pathway, and drug-target information—are essential for addressing these gaps.

Indian genetic diversity is shaped by a complex mix of deep ancestral lineages, long-standing social stratification, extensive endogamy, and multiple ancient admixture events that have taken place over thousands of years [18-20]. These population-level distinctions can translate into variation in drug response, with certain groups experiencing unexpected adverse outcomes [21].

India also represents the world's leading supplier of generic pharmaceuticals [22].

Despite this, clinical care in India relies largely on treatment regimens derived from European and North American populations, with little consideration for local genomic variation. Integrating genetic screening, molecular profiling, and related technologies into routine practice could allow clinicians to select suitable therapies at the outset, avoiding lengthy and costly trial-and-error drug adjustments. Given the magnitude of India's genetic heterogeneity, robust investigation of population-level variation and relevant SNVs that drive drug-response differences is critically needed.

The IndiGen initiative was launched to sequence thousands of individuals from diverse Indian groups and to develop public health applications based on these genomic datasets [23].

In this study, we conducted a broad comparative assessment of common Indian-enriched variants (from IndiGen) against those found in other global populations to highlight population-specific changes that may influence drug sensitivity and adverse drug reactions (ADRs). Our pharmacogenomic investigation focused on kinase-related drug targets, which form the second largest class of therapeutic targets after G-protein–coupled receptors [24]. The human genome contains 538 protein kinases [24], many of which are implicated in disorders such as cancer [25]. Since most kinase-directed therapies were tested predominantly in European cohorts, their performance could differ in individuals from India. A single SNV that disrupts a critical functional region may alter drug–gene interactions or induce structural changes in the protein, potentially modifying drug-binding behavior [26]. Therefore, assessing the number and effects of such substitutions on protein integrity, stability, and activity is essential. Any destabilizing non-synonymous SNV (nsSNV**)** can interfere with normal drug metabolism. To capture these effects, our study integrated both sequence-based and structure-centric analyses of missense variants, evaluated their influence on drug-related interactions, and examined mutation-induced structural perturbations. Sequence-level comparisons were used to examine SNV patterns and amino-acid exchange frequencies across populations, while structural assessments evaluated the impact on stability, solvent exposure, hydrophobic features, and hydrogen-bond networks using multiple computational tools.

Alterations in protein–ligand interactions caused by SNVs were assessed using molecular docking, and we also compared structural changes associated with IndiGen variants against pathogenic variants cataloged in UniProtKB Humsavar.

Together, this framework enhances our understanding of how Indian-specific variation may influence drug efficacy.

Materials and Methods

Variant data collection

Genomic variants and their allele frequencies for the Indian population were obtained from 1,029 whole-genome sequences of unrelated individuals sampled across diverse Indo-ethnic groups as part of the IndiGen program [23]. The dataset included single-nucleotide variants and indels, all annotated relative to the GRCh38 reference genome using Annovar [27]. Only non-synonymous SNVs (nsSNVs) were selected for downstream analyses [23]. For comparison, the publicly accessible 1000g2015aug_all VCF from the 1000 Genomes Project [28] was used.

Assembling druggable genes

The Drug-Gene Interaction Database (DGIdb) v3 aggregates data on existing therapeutic agents and emerging drug targets [15]. Genes are categorized according to known drug interactions and predicted druggability, with content sourced from 30 public databases, including DrugBank [29], the Therapeutic Target Database (TTD) [30], PharmGKB [31], OncoKB [32], and the Cancer Genome Interpreter (CGI) [33]. Using the database's "GuideToPharmacologyGenes" filter, we extracted a list of 545 kinase genes and their corresponding FDA-approved drugs. These kinase entries were annotated further with details such as Ensembl IDs, PDB structures, RefSeq transcripts, genomic coordinates, UniProt accessions, sequence lengths, and structural coverage, using BioMart tools [34].

Data preparation

Sequence data preparation

For the sequence-level analyses, we worked with 545 kinase genes considered druggable along with their mapped variants. Protein sequences for these genes were obtained from NCBI GenBank, and modified (mutant) versions were produced by introducing substitutions specified in the Annovar annotations.

Structure data preparation

Structural analysis required further refinement of the sequence dataset using the following criteria:

1. A corresponding experimental protein structure had to be available.
2. A drug compound targeting the protein needed to exist.
3. The SNV location had to be represented within the available crystal structure.
4. Sequence-to-structure alignment had to meet a minimum 70% coverage.
5. The IndiGen allele frequency for each nsSNV had to be ≥10%.

After applying these filters, 12 kinase genes with 22 variants remained; these formed the IndiGen Structure dataset.

Three genes—EPHA7, RET, and TAOK3—showed structure coverages below 70% for PDB IDs 3NRU, 6I83, and 6BFN, respectively. They were kept, however, because the IndiGen-reported nsSNV residues appeared within the crystallized regions, satisfying the SNV-mapping requirement.

For comparative purposes, we also used Humsavar, a UniProtKB/Swiss-Prot resource (Release 2020_04; 12-Aug-2020) cataloging human missense variants. In Humsavar, variants are classified as:

- Disease-associated: 31,132 (64.1%)
- Polymorphisms: 39,464 (23%)
- Unclassified: 8,381 (12.9%)

From this full set, we extracted variants linked to the 12 IndiGen Structure genes, yielding 217 variants, which served as a benchmarking set.

Data processing and visualization

Drug, gene, and variant tree

This step aimed to quantify and visualize how often variants appear across kinase families and to relate this to available drug molecules. We used KinMap [35] to explore kinase genes listed in the IndiGen dataset interactively. As input, the tool received information on 327 druggable kinases, along with the annotated variants and DGIdb-reported drugs linked to each gene.

Amino acid conversions and mutabilities

To evaluate how often one amino acid type substitutes for another and to look for trends affecting biochemical properties, we used a custom Python script (repository link: https://github.com/raylab-projects/Pharmacogenomics).
The analysis included all kinase variants [36-38].
The script produced a 20×20 normalized amino-acid exchange matrix representing the percentage conversion from each residue type to every other: Normalized count = (Residue count in sample variants) / (Residue count in RefSeq) × 100.
We compared these exchange patterns with chemical characteristics of the residues, assessing chemical shifts by subtracting reference and alternate residue totals for each amino-acid class.
A residue's mutability was defined as the number of observed mutations for that residue, divided by its frequency in the human reference proteome. Mutability was calculated using IndiGen variants with AF > 10%.

Statistical analysis of amino acid conversions

To test whether specific amino-acid substitutions differed significantly between IndiGen and 1000 Genomes populations (EAS, AMR, AFR, SAS, EUR), we applied a one-proportion z-test. Multiple-testing adjustments were done using the Benjamini–Hochberg FDR procedure. Corrected p-values < 0.05 were regarded as significant.

Multiple sequence alignment and protein domain analysis

To assess potential functional impacts of SNVs, we examined whether each variant fell within conserved residues or annotated protein domains. Multiple-sequence alignments were created using Clustal Omega [39], and mutant FASTA sequences were generated using Python scripts.
Protein-domain locations were identified via the PfamScan server [40], using a combined FASTA file of all proteins (default settings). The output included domain identifiers, start and end coordinates (hmm_name, hmm_start, hmm_end), and associated metadata. Variants located inside domain boundaries (hmm_start–hmm_end) were labeled as 0, and for variants outside domains, the distance from the nearest domain edge was calculated.

Variant protein structure generation

Computational modeling enables the reconstruction of three-dimensional protein structures using in-silico strategies built upon knowledge derived from experimentally solved structures, including X-ray crystallography, NMR spectroscopy, and energy-based modeling. Following application of the structural-filtering criteria, the native PDB structures corresponding to the 12 genes in the IndiGen Structure dataset were retrieved from the RCSB Protein Data Bank. These crystallized forms served as templates for generating mutant models by substituting one reference residue with its corresponding alternate amino acid.
Single-site mutations were introduced using the rotkit mutagenesis tool within PyMol [41]. The entire procedure—downloading structures, modifying residues, and organizing outputs—was automated through Python scripts. The resulting 22 mutant models were then subjected to energy minimization in Chimera [42], applying:

- 1000 steepest-descent cycles,
- 0.02 Å step size,
- AMBERff14SB force field.

Predicted changes in flexibility, structural stability, and overall conformation were obtained using DynaMut [43]. Structural comparisons between wild-type and mutant proteins were further evaluated using:

- DSSP [44] to annotate secondary-structure changes,
- HBPLUS [45] to detect hydrogen-bond gain/loss,
- Naccess [46] to measure solvent-accessible surface area differences.

Molecular docking

Docking simulations were used to determine how each SNV alters the affinity of FDA-approved drugs toward their kinase targets. Only 10 of the 12 IndiGen Structure genes—CHUK, EPHA7, GRK5, MAPK11, MAPK13, PI4K2B, PIK3CG, GRK4, TAOK3, and IRKA1—had documented drug interactions in DGIdb, adding up to 69 approved compounds.

Protein structures for these ten genes and their 20 modeled variants (total receptors = 30) were used for docking. All 69 ligands were obtained in PDB format from DrugBank and PubChem.

Receptor preparation involved removing heteroatoms and water, adding polar hydrogens, and ligand preparation required assigning AutoDock4 atom types, Gasteiger charges, and aromatic-carbon identification. Receptors and ligands were saved in PDBQT format.

With no predefined binding-site information available, blind docking was executed using AutoDock Vina [47]. Binding-site coordinates (per protein) were determined with PyRx [48]. Each docking run used:

- grid box: 60 Å × 60 Å × 60 Å,
- exhaustiveness = 100,
- 500 maximum poses,
- energy range = 20 kcal/mol,
- 50 iterations per protein.

Ligand similarity/diversity and toxicity analysis

Previous work has shown that shared chemical features are linked to similar ADR profiles [49, 50]. To identify such patterns within our ligand set, we analyzed chemical similarity and toxicity for all compounds.

Ligand similarity was assessed through molecular descriptors capturing structural motifs such as functional groups, ring systems, and substructure patterns, represented in multi-dimensional chemical space. Two fingerprinting methods were applied:

- MACCS 166-key fingerprints
- Morgan circular fingerprints with radius 2 [51].

Similarity calculations were performed using RDKit [52], and the Tanimoto coefficient was used to quantify similarity and diversity (dissimilarity = 1 − similarity). Drug toxicity predictions were carried out using ProTox-II [53].

Phenotypic drug–drug similarity

Many drugs interact with multiple targets—a phenomenon known as polypharmacology. Prior studies have demonstrated that such multitarget behavior often reflects structural conservation within protein families or similarities between primary binding pockets [54].

If two compounds modulate the same gene product, they typically exhibit comparable mechanisms of action and pharmacological effects [55]. Consequently, structurally related or target-sharing drugs may serve as replacement options when adverse reactions occur. Analysis of DGIdb-derived drug–gene interactions revealed several compounds converging on the same protein targets.

Results and Discussion

Indian variations across the kinome

To obtain an initial perspective on how Indian-specific variants are distributed within the drug-relevant kinome, the kinase genes in the IndiGen dataset that overlap with the 327 drug–gene interactions cataloged in DGIdb were mapped onto a phylogeny using KinMap [35] **(Figure 1)**. Within this landscape, the atypical kinase group showed 148 drug–gene associations and 1,224 amino-acid–altering substitutions. Interestingly, although this class displayed numerous drug connections, only a limited subset of these atypical kinases carried missense alterations. Changes occurring within conserved segments of a protein may modify folding, dynamic behavior, or ligand-binding affinity, ultimately altering drug responsiveness. Genes harboring a larger pool of variants and multiple approved therapeutic agents are more likely to contribute to adverse drug outcomes [56]. Among all kinase categories, the Tyrosine Kinase group exhibited the highest count of FDA-approved agents (1,978) and simultaneously the most substitutions (5,013). Other kinase groups, such as CMGC (CDKs, MAPKs, GSK3s, CLKs, TLKs) and AGC (PKA/PKC/PKG), displayed substantial variation—10,518, 1,193, and 2,943 variants respectively—but possessed fewer drugs with validated drug–gene associations (213, 185, and 339 respectively) compared to the Tyrosine Kinase class. The CK1 family showed the smallest number of substitutions (275) and the lowest number of drug–gene links (18).

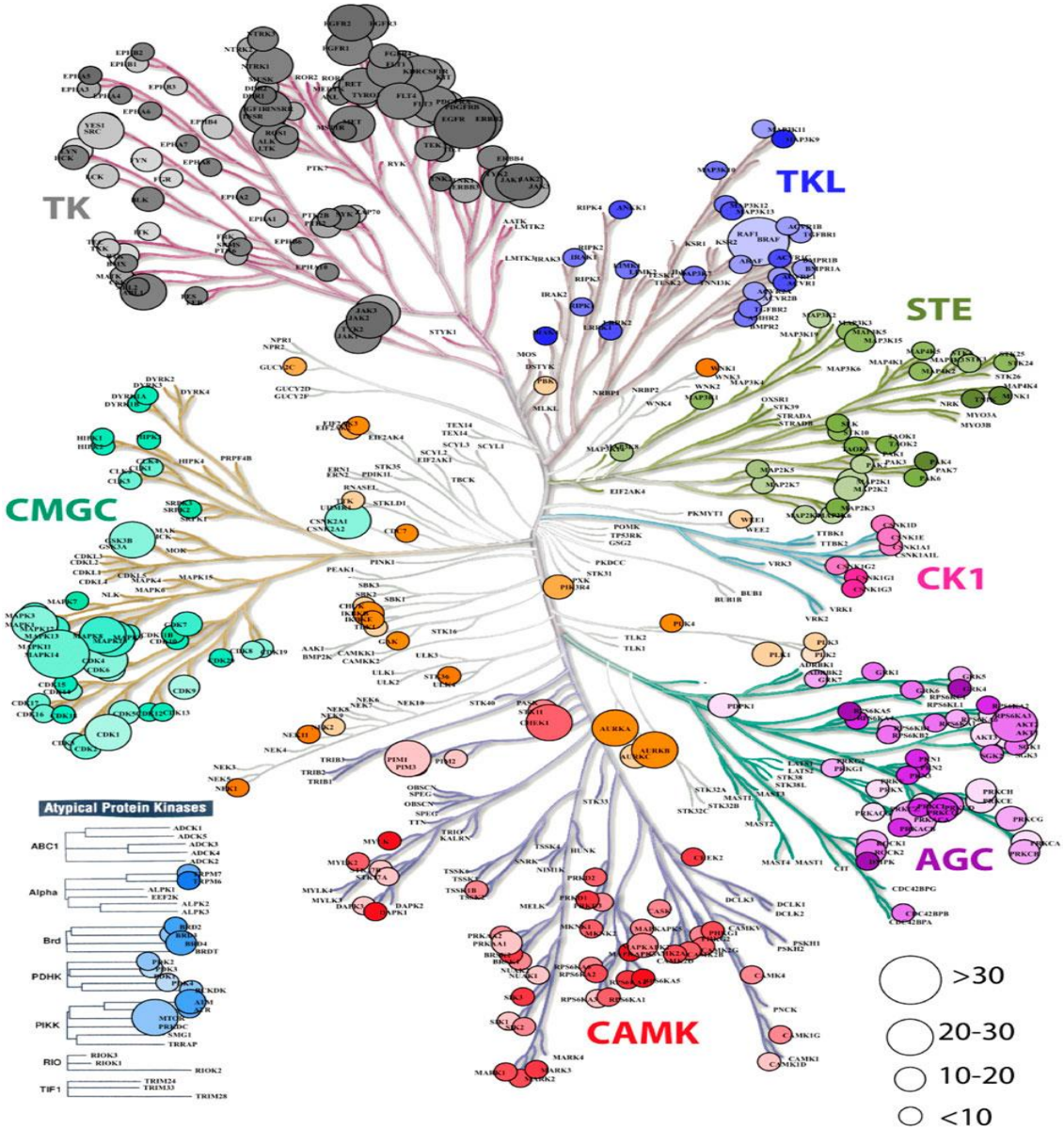


Figure 1. Dendrogram of IndiGen kinase-coding genes created with KinMapbeta. Circle diameter corresponds to the number of therapeutics known to interact with each gene, while color gradients reflect the quantity of IndiGen variants associated with that gene. Each kinase family is visually distinguished by a unique color.

Comparative evaluation of amino-acid substitutions in druggable kinase genes among Indians and 1000G populations

To characterize mutation tendencies in the Indian cohort, an amino-acid substitution matrix was generated for all SNVs across the 545 druggable kinase genes included in the IndiGen dataset. For every nsSNV, the frequency with which a given reference residue (per RefSeq) was replaced by an alternative residue was calculated to identify residues undergoing recurrent alteration. These normalized exchange proportions are displayed in **Figure 2a**.

The analysis showed that approximately 68% of Arginine (R) residues were replaced by Tryptophan (W), indicating a transition from a positively charged polar residue to a hydrophobic aromatic one. Likewise, 58% of Cysteine (C) sites changed to Tyrosine (Y), representing a shift from a polar uncharged to an aromatic polar residue. Other frequently observed conversions in the 40%–50% range included Leucine (L) → Phenylalanine (F) (a non-polar to non-polar shift), Lysine (K) → Glutamic acid (E) (basic to acidic), and Asparagine (N) → Aspartic acid (D) (amide to acidic). Notably, despite Serine (S) and Leucine (L) having six codons each, their substitution rates were lower compared to single-codon residues like Tyrosine (Y) and Tryptophan (W).

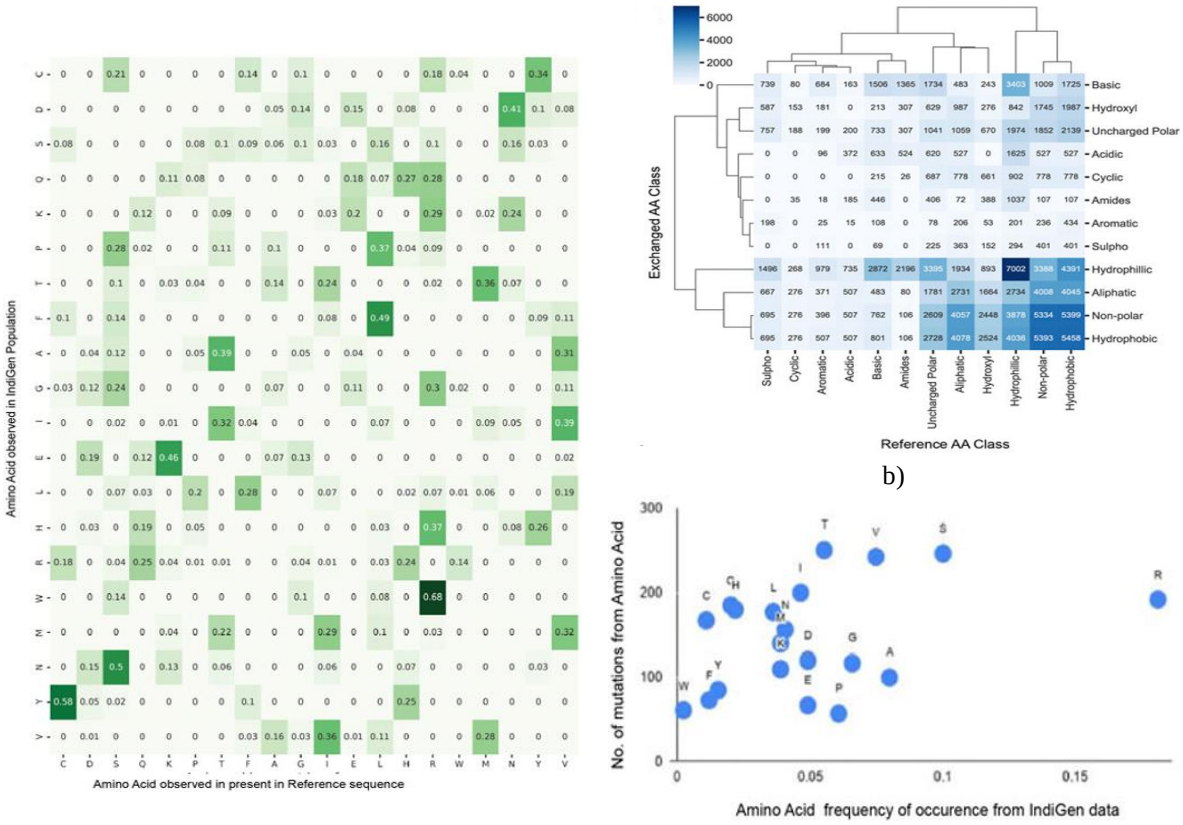


Figure 2. Amino-acid–level analyses for the 545 druggable kinase genes in IndiGen.
(a) Exchange-frequency matrix of reference vs. alternate residues.
(b) Heat-cluster map displaying chemical-group shifts for all amino-acid conversions.
(c) Scatter plot indicating mutability values for each residue type.

To further interpret how chemical properties shift following amino-acid substitutions, a chemical-group transition analysis was performed. Mutated residues were grouped into 12 chemical categories (Aliphatic, Hydroxyl, Cyclic, Aromatic, Basic, Acidic, Sulfur-containing, Amides, Non-polar, Uncharged polar, Hydrophobic, Hydrophilic). **Figure 2b** shows a cluster map detailing the number of conversions between these classes.

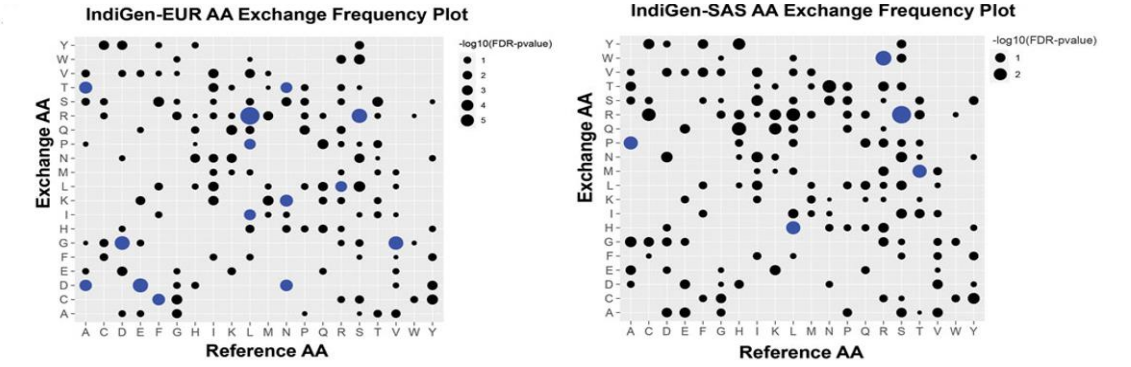
Conservative intra-class changes were prevalent for Hydrophilic residues (Ser, Thr, Tyr, Asn, Gln, Asp, Glu, Lys, Arg, His), Hydrophobic residues (Gly, Ala, Val, Pro, Leu, Ile, Met, Trp, Cys, Phe), and Non-polar residues (Gly, Ala, Val, Pro, Leu, Ile, Met, Trp, Phe), in agreement with earlier observations [57]. However, numerous groups exhibited more inter-class than intra-class conversions. For example, residues from Non-polar, Hydroxyl, Aliphatic, and Uncharged-polar categories frequently shifted into the Hydrophobic class. In parallel, residues belonging to Amide, Basic, Acidic, Aromatic, Cyclic, and Sulfur-containing categories tended to convert into Hydrophilic residues.

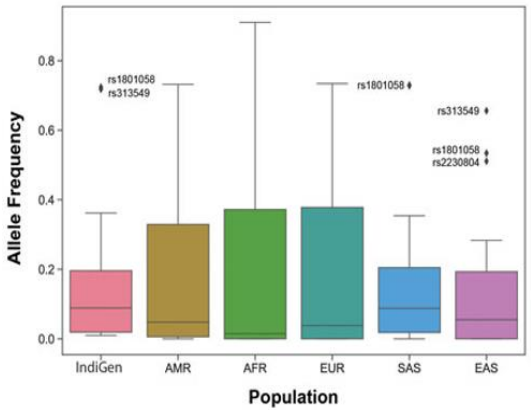
In line with the earlier findings, an additional evaluation was carried out in which the reference amino-acid identities were taken from the RefSeq hg38 annotation, while the substituted residues at the same SNV positions were drawn from the IndiGen dataset. To prevent overlap among residue categories, all amino acids were grouped into six chemical classes: Aliphatic (Gly, Ala, Val, Leu, Ile), Hydroxyl (Ser, Thr), Cyclic (Pro), Aromatic (Phe, Tyr, Trp), Basic (Lys, Arg, His), and Acidic (Asp, Glu). For each class, the change in residue counts at SNV sites was computed and displayed. A horizontal bar chart illustrates how the six categories differ when comparing amino-acid frequencies in RefSeq (hg38) versus IndiGen. This chemical shift assessment revealed an overall reduction in aliphatic, cyclic, and basic categories, while the hydroxyl, aromatic, and acidic groups showed a net increase. Notably, even though the hydroxyl, aromatic, and acidic groups contain only 2, 3, and 2 residues respectively, all three contribute to net gains; in contrast, the aliphatic category—despite having the largest number of amino acids—shows an aggregate loss. This pattern indicates that the direction of gain or loss within a chemical class does not depend on the class size.

To explore how often specific residues undergo substitution relative to their abundance in the IndiGen cohort, mutability scores were calculated for each amino-acid type. These scores are displayed in **Figure 2c**. Arginine (R) exhibited the highest observed frequency (>0.15) at mutated sites, whereas Tryptophan (W) appeared least frequently as the reference residue in IndiGen. Valine, Serine, and Threonine also showed a greater tendency to mutate than many other residues. This aligns with the trends noted earlier in the amino-acid exchange matrix **(Figure 2a)**. In that matrix, Arginine (R) had the strongest substitution signal (maximum exchange frequency ≈ 0.68), while Tryptophan (W) showed the weakest (maximum frequency ≈ 0.14).

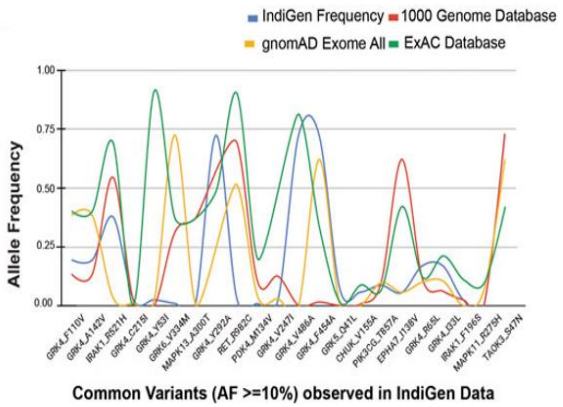
Following this detailed characterization of the Indian dataset, a comparative amino-acid–level analysis was conducted against populations from the 1000 Genomes Project, including European (EUR), American (AMR), African (AFR), South Asian (SAS), and East Asian (EAS) groups. For each population, the number of exchanges from the RefSeq residue to any alternative amino acid was computed in the same manner as for IndiGen **(Figure 2a)**. Differences between the IndiGen exchange profile and those of the 1000G populations were assessed using a proportion z-test. A total of 144 non-null reference–alternate residue pairs were examined. Each exchange was assigned a p-value and corresponding z-score. After adjusting the p-values, the negative log of the corrected values was visualized in **Figure 3a**.

Figure 3a shows a bubble diagram where the original residue is on the X-axis and the substituted residue is on the Y-axis. Bubble size increases as the FDR-corrected p-value decreases. Significant exchanges identified between IndiGen–EUR and IndiGen–SAS are highlighted in blue. The counts of statistically meaningful differences (FDR $p < 0.05$) between IndiGen and AMR, AFR, EUR, EAS, and SAS were 17, 16, 14, 12, and 5 respectively. These results indicate that IndiGen variants resemble the South Asian patterns more closely than those of the other populations. This is consistent with the composition of the SAS group, which includes Gujrati Indians in Houston (GIH), Punjabi from Lahore (PJL), Bengali from Bangladesh (BEB), Sri Lankan Tamil in the UK (STU), and Indian Telugu in the UK (ITU), all of which share regional ancestry with IndiGen.





d)



e)

Figure 3. (a) Comparison of amino-acid exchange behavior between IndiGen and major 1000G populations. Bubble sizes reflect −log10(FDR-corrected p-values) for each reference–alternate pair; significant cases ($p < 0.05$) are colored blue. (b) UpSet visualization of significant exchanges shared across IndiGen relative to the five 1000G populations. (c) Grouped bar chart showing variant counts occurring before (green), inside (pink), or after (violet) functional domains for IndiGen and 1000G groups. (d) Boxplot comparing allele-frequency distributions for the 22 IndiGen variants (AF ≥ 10%) included in the structural dataset across global 1000G populations.
(e) Occurrence of IndiGen-specific SNVs (22 variants with AF ≥ 10%) across other resources, including 1000 Genomes, gnomAD exomes, and ExAC, with allele frequencies plotted per variant.

The population-wise distribution of amino-acid substitutions was examined by constructing an UpSet visualization **(Figure 3b)**, relying on all statistically relevant exchanges identified in **Figure 3a**. This UpSet diagram contains four components: a bar panel quantifying overlap sizes, a matrix-like panel beneath it identifying the intersecting population combinations, labels above each bar indicating shared substitutions, and a small bar graph on the left summarizing total substitution counts per dataset. Bars positioned above a single filled node indicate substitutions confined to one population group. Five substitutions occurred solely in the IndiGen–AMR and IndiGen–EAS comparisons; four appeared only in IndiGen–SAS and IndiGen–EUR; and two substitutions were exclusive to IndiGen–AFR. A replacement from Alanine to Threonine (A→T) occurred in every population set except IndiGen–SAS. Sets of three-population overlaps included Ser→Ala and Arg→Ser for IndiGen vs. AFR/EAS/AMR, Leu→Pro for IndiGen vs. EAS/AFR/AMR, and Leu→Arg for IndiGen vs. AFR/EUR/AMR, all representing conversions between non-polar or basic categories. Larger intersections were observed for IndiGen vs. AMR/EUR and IndiGen vs. AFR/AMR, each involving three shared exchanges: Phe→Cys; Val→Gly; Asn→Lys; and Gln→Glu; Gln→Arg; Gln→Leu, reflecting shifts from uncharged polar residues toward acidic, basic, or non-polar classes.

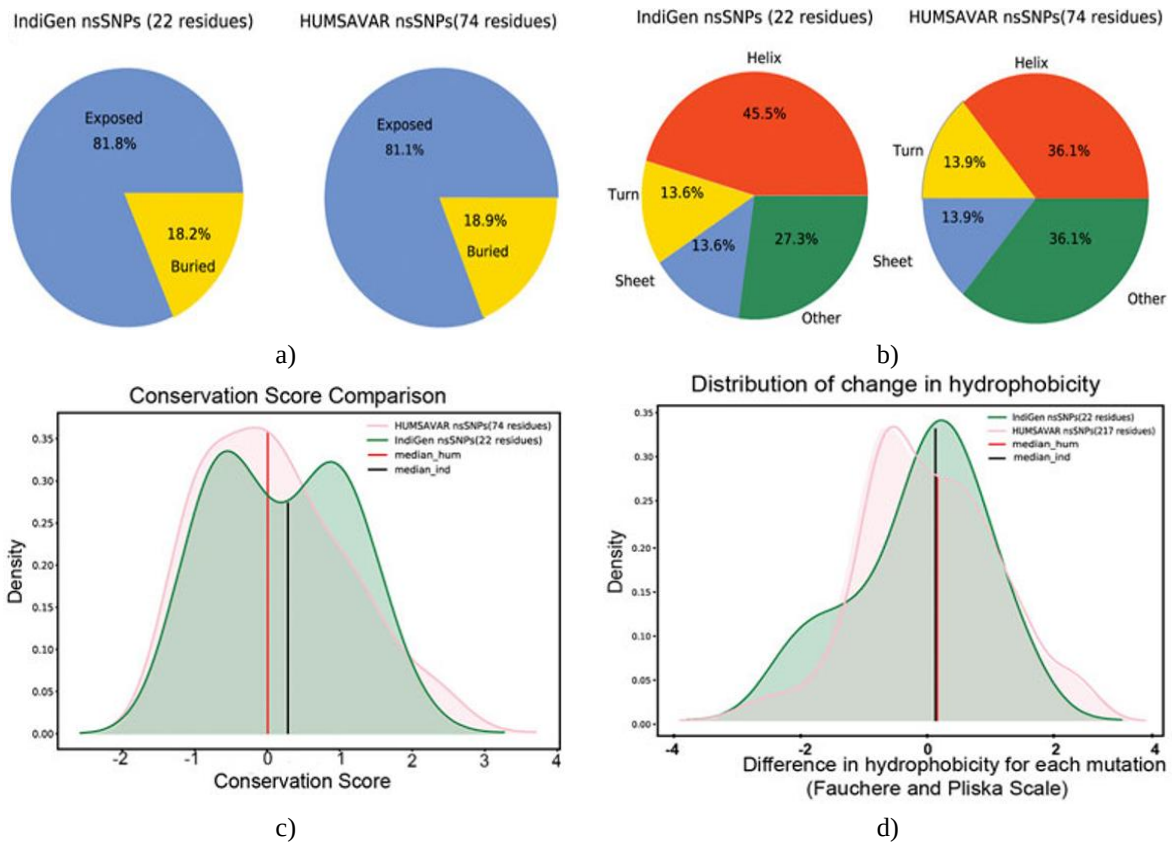
After evaluating how sequence-level substitutions differ among populations, we next evaluated structural consequences. Protein domains constitute evolutionarily conserved regions with stable folds; therefore, variants falling inside these domains are more likely to influence folding, stability, or function. To quantify how many single-nucleotide variants (SNVs) are positioned within, before, or beyond conserved domains, domain-based classification was performed across IndiGen and the 1000 Genomes datasets (African, American, European, East Asian, and South Asian). In **Figure 3c**, the grouped bar figure displays six populations along the X-axis, while the Y-axis shows counts of SNVs lying inside domains (pink), upstream of domains (green), or downstream of domains (violet). All datasets showed comparatively fewer variants in pre-domain regions, indicating that both IndiGen and the 1000G populations tend to have more SNVs within or after a domain than before it. For EAS, AFR, AMR, and EUR, the tallies for within-domain and post-domain variants were nearly the same. In IndiGen, most substitutions (952) occurred inside domains, 226 were located in post-domain segments, and only twelve variants appeared in the pre-domain zone. Given that most SNVs occur within domain boundaries, amino-acid changes in these regions are likely to influence protein architecture, stability, and functional output.

Previous studies indicate that SNV allele frequencies can vary across ethnic backgrounds [58, 59]. To evaluate this, allele-frequency distributions for IndiGen variants (AF ≥10%; 22 structural variants) were contrasted with those from the 1000G cohorts, as shown in the boxplot in **Figure 3d**. The comparison demonstrated minimal

variation among the examined groups. The allele-frequency patterns for IndiGen, South Asian, and East Asian populations were nearly identical, with similar median values and overlapping outliers—specifically rs1801058 and rs313549 from the GRK4 gene (Y292A and V486A). An additional boxplot **(Figure 3e)** used the same 22 variants to compare IndiGen allele frequencies against those from the gnomAD exomes, ExAC, and the 1000 Genomes datasets.

Structural comparison of indigen variants and disease-linked mutations

To more deeply investigate how IndiGen SNVs might alter protein structures, we curated an IndiGen structural subset containing only variants from drug-targeted kinases that fall within crystallographically resolved regions. This yielded twelve kinase genes and 22 corresponding variants. For comparison, 217 disease-associated variants from these same genes were extracted from Humsavar. Structural parameters—including solvent accessibility, secondary-structure classification, conservation scores, and hydrophobicity shifts—were evaluated for both datasets. For solvent exposure **(Figure 4a)**, residues were categorized as buried or exposed using a 5% Naccess threshold [46]. Both datasets showed that exposed residues accumulate more substitutions. Gong and Blundell (2010) similarly reported that over 60% of solvent-facing SNVs are disease-linked. In IndiGen, 81.8% of the 22 residues were exposed, nearly identical to Humsavar, where 81.1% of the 74 mutated residues were solvent-accessible [60]. Thus, exposure patterns did not differ substantially between the two groups. For secondary-structure localization **(Figure 4b)**, IndiGen variants showed a slight bias toward alpha-helical regions, whereas Humsavar variants exhibited an even split between helices and loop/random-coil regions.



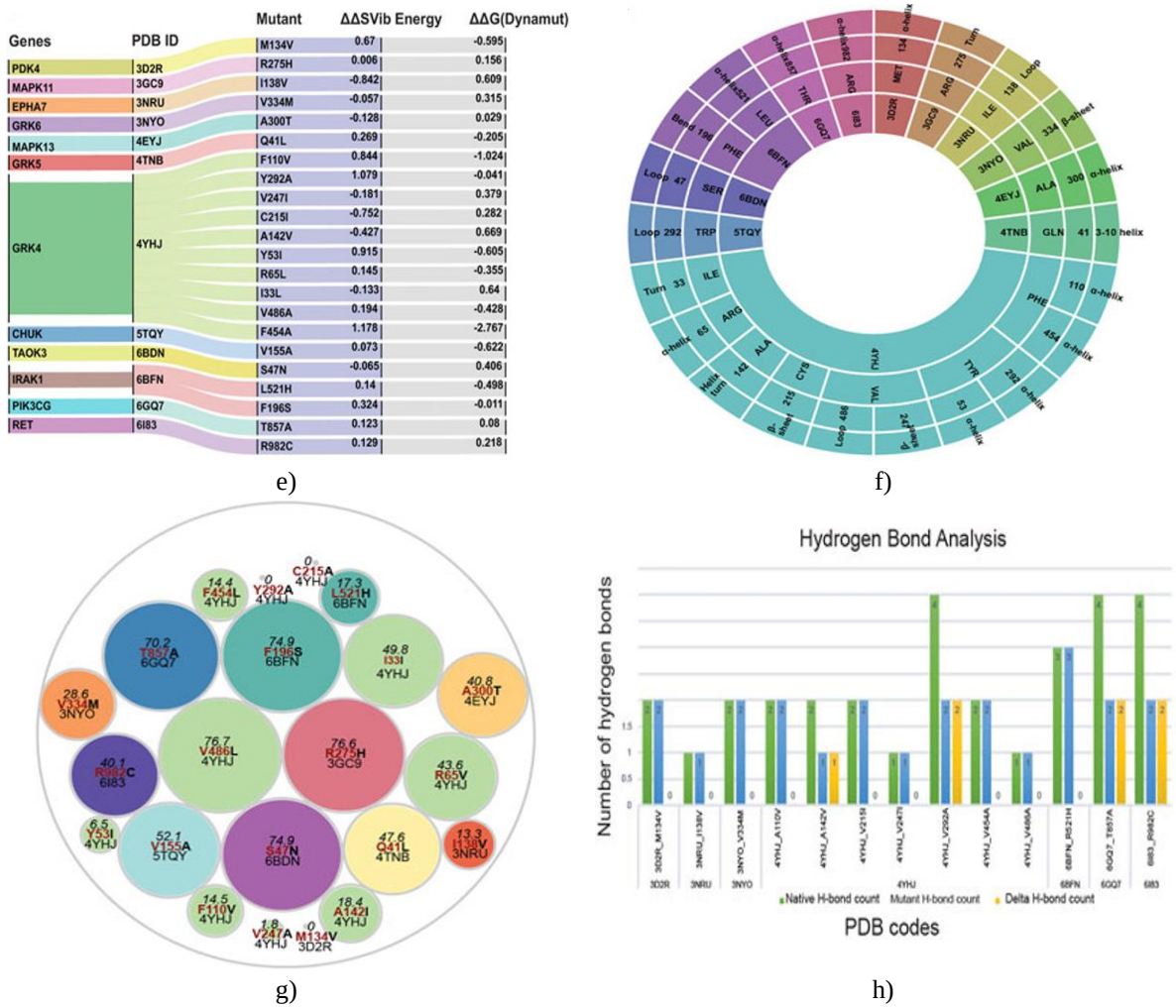


Figure 4. Comparative overview of structural features of variants in the IndiGen and Humsavar datasets: (a) Solvent exposure levels of variants across both collections. (b) Secondary structural elements where each variant is located in the two datasets. (c) Distributions of conservation scores and ΔHydrophobicity values for Humsavar and IndiGen variants. (d) The portion of the curve left of −2 on the ΔHydrophobicity axis corresponds to residues showing a notable rise in hydrophobicity after mutation, whereas residues to the right of +2 indicate those with a marked reduction. (e) Alluvial diagram illustrating changes in folding free energy (δδG, kcal/mol) and Dynamut-derived vibrational entropy for 22 variants. (f) Sunburst diagram depicting mutant residue secondary structure assignments computed using DSSP. (g) Circle-packing visualization of relative solvent accessibility (by Naccess) for mutated residues across 22 variants from 12 proteins. (h) HBPLUS output showing hydrogen bond counts for each residue before mutation (green), after mutation (blue), and ΔH-bond changes (yellow).

To evaluate evolutionary conservation among altered residues, conservation scores for IndiGen structural variants (22 residues) and Humsavar variants (74 residues) were determined via Consurf [61]. **Figure 4c** displays their score distributions. The Humsavar pattern approximated a normal distribution, whereas the IndiGen set exhibited a bimodal profile. The median line bisected the curve into equivalent areas; the Humsavar median (0.007) was closer to zero than the IndiGen median (0.358). To estimate proportions of residues with higher or lower conservation, a relative conservation threshold of −1/+1 was applied. A greater percentage of strongly conserved residues (Consurf score > −1) appeared in the Humsavar dataset, producing a steeper curve. Conversely, a larger fraction of highly variable sites (score > 1) fell within the IndiGen distribution, suggesting that Humsavar variants are more often located at conserved positions.

The distribution of hydrophobicity changes between reference and mutated residues for the two datasets is presented in **Figure 4d**. The medians for both were nearly identical and near 0, implying that increases and decreases in hydrophobicity occurred at comparable frequencies. To quantify residues with substantial

hydrophobicity alterations, −2 was used as the cutoff for increases and +2 for decreases. IndiGen variants showed a higher percentage of residues with significant hydrophobicity gain, while residues with substantial hydrophobicity loss were more frequent in the Humsavar group.

Effects of nsSNVs on protein structural attributes

Stability assessment of modeled variants

Before analyzing structural features of IndiGen nsSNVs, the influence of mutations on protein stability and flexibility was examined using Dynamut [43]. The tool employs normal mode analysis and a machine-learning approach to estimate ΔΔG (change in folding free energy, kcal/mol) and ΔΔS (vibrational entropy difference, kcal/mol/K) between wild-type and mutant structures. Dynamut results indicated that 11 of 22 variants exhibited negative ΔΔG values, implying destabilization, while 14 of 22 variants showed positive ΔΔS values, suggesting enhanced flexibility following mutation **(Figure 4e)**. The plot presents 12 genes with corresponding PDB structures (4YHJ, 5TQY, 3NYO, 6GQ7, 4TNB, 6BFN, 3GC9, 6BDN, 6I83, 4EYJ, 3NRU, 3D2R) and their 22 associated mutants with respective energy metrics. Dynamut predicted that GRK4 variants F454A and F110V (PDB ID: 4YHJ) display ΔΔG values of −2.767 kcal/mol and −1.024 kcal/mol, indicating destabilization, and ΔΔS-Vib values of 1.178 kcal/mol/K and 0.844 kcal/mol/K, reflecting increased flexibility. The observed loss of aromaticity—conversion of phenylalanine to alanine or valine—may account for reduced stability, as aromatic rings offer substantial structural reinforcement.

Secondary structure annotation and relative solvent accessibility of mutated residues

Protein secondary structure mainly consists of α-helices and β-sheets, which arise from localized interactions along segments of the polypeptide chain. A protein's capacity to associate with other molecules is strongly influenced by surface-exposed amino acids that exhibit high solvent accessibility. Alterations in these residues can influence protein behavior, emphasizing the relevance of analyzing structural features at mutation sites. Solvent accessibility was calculated using Naccess [46], while DSSP [62] was used to evaluate secondary structure properties for each mutated residue. **Figure 4f** presents a sunburst diagram summarizing DSSP-assigned secondary structural classes. This visualization contains four rings: the innermost displays 12 PDB identifiers, the second includes the three-letter code of the native residue at each mutation site, the third ring marks the mutation position, and the outermost ring provides the DSSP-derived structural annotation. Colors correspond to each residue's original PDB structure. Most variants were located within α-helical segments rather than other structural elements.

Figure 4g shows a circle-packing plot illustrating the relative solvent accessibility for mutated residues from 22 variants spanning 12 proteins. Circle size reflects the accessibility score, and circles belonging to nsSNPs from the same gene share the same color. The native residue and its position are highlighted in red to indicate their associated accessibility. Two mutations in structure 4YHJ (Y53I and C215I) had accessibility values of zero. In contrast, residues Arg275 and Val486 in mutants R275H (PDB ID: 3GC9) and V486A (PDB ID: 4YHJ) displayed values above 75, signifying substantially higher exposure. Overall, five residues—arginine, valine, phenylalanine, and serine—from PDB entries 3GC9, 4YHJ, 6BDN, and 6BFN exhibited relative solvent accessibility exceeding 60.

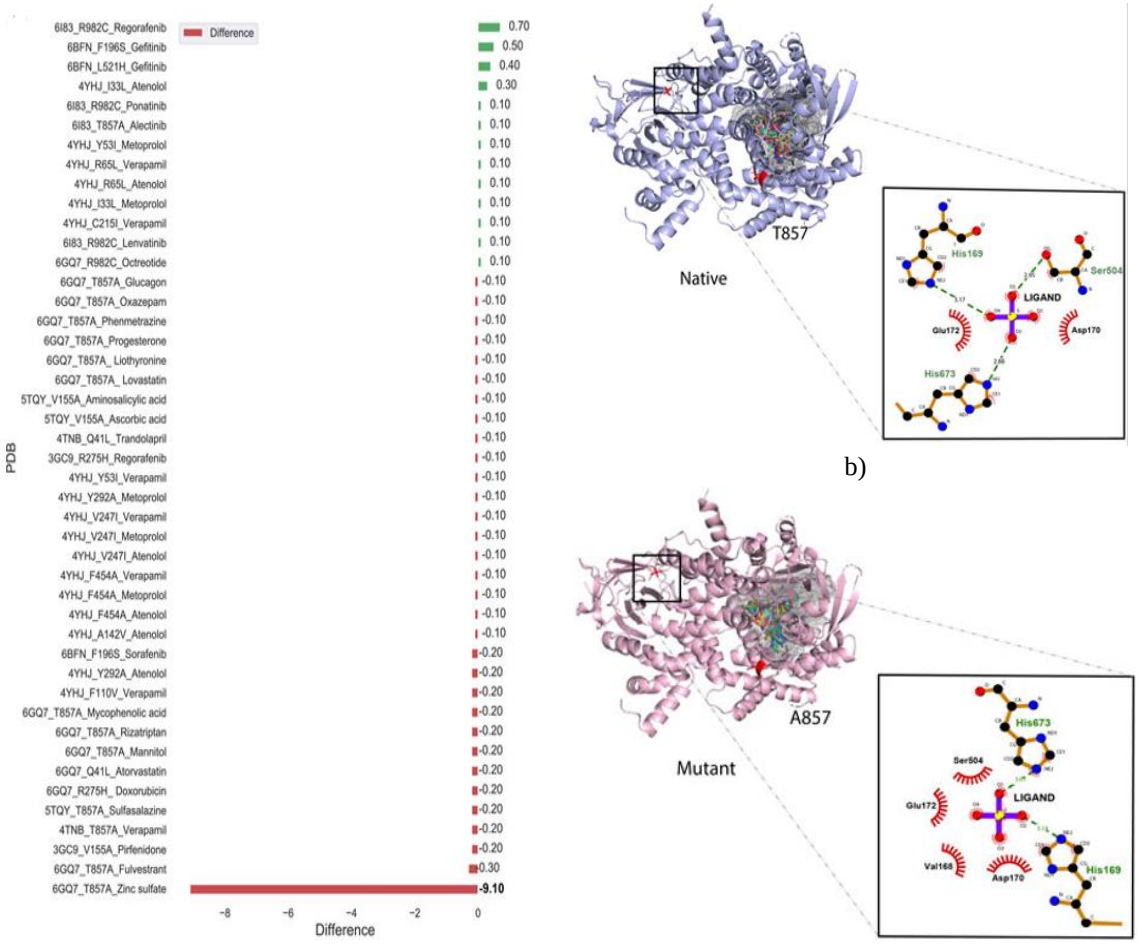
Effect of nsSNV on hydrophobicity and hydrogen bonding

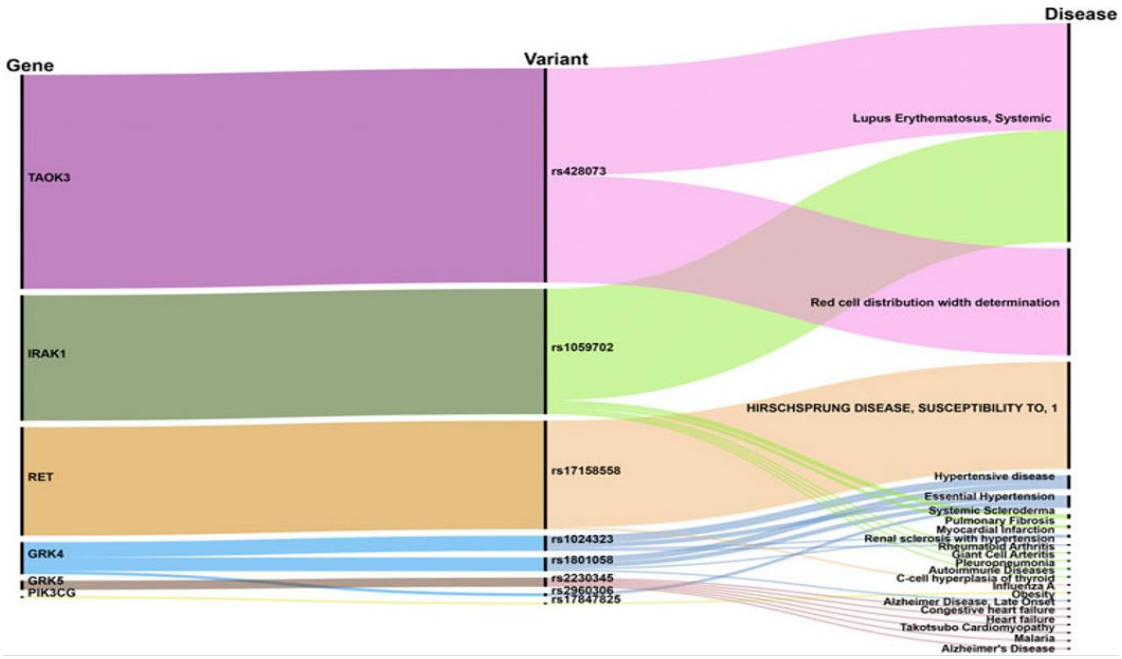
A single amino acid substitution can modify hydrophobic properties or interrupt hydrogen-bond networks, potentially altering both the structure and function of the protein [63]. Hydrophobicity changes in the IndiGen structural variants were assessed based on the Fauchère and Pliska scale [64]. Within the 22 variants analyzed, 12 showed reduced hydrophobicity, whereas the remaining variants displayed an overall increase. Hydrogen bonds formed by the substituted residue before and after mutation were determined using HBPLUS **(Figure 4h)**. Variant 4YHJ_A142V resulted in the loss of one hydrogen bond, while variants 4YHJ_V292A, 6GQ7_T857A, and 6I83_R982C each exhibited the loss of two hydrogen bonds.

Effect of nsSNV on ligand binding

Because kinases play a central role in pharmacology, molecular docking analyses were carried out to evaluate how an SNV might alter drug–target interactions. Every FDA-approved compound listed in DGIdb for genes

represented in the IndiGen structural dataset was docked to both the reference and variant protein conformations. AutoDock Vina was used to estimate the binding energy (Gibbs free energy, ΔG in kcal/mol) for each receptor–ligand pair, and the values obtained for native and altered models were compared. Among 69 evaluated protein–drug combinations, 45 showed shifts in binding energy ranging from 0.7 to −9.1 kcal/mol, while the remainder showed no detectable difference. The distribution of these changes is summarized in **Figure 5a**. Of the 45 affected pairs, 32 exhibited reduced affinity and 13 demonstrated a gain in binding strength, suggesting that in many instances an nsSNV compromises the stability of the complex. One notable case involved the T857A substitution in PIK3CG (PDB: 6GQ7), where docking with zinc sulfate (DrugBank DB09322) showed a pronounced reduction in binding energy (−9.1 kcal/mol), with the native protein scoring −13.0 kcal/mol versus −3.9 kcal/mol for the mutant. These 45 altered pairs were subjected to additional assessments of similarity in ligand and binding-site characteristics.





d)

Figure 5. (a) Bar chart depicting docking outcomes for the 45 protein–drug combinations (x-axis) and the magnitude of affinity change (y-axis). Red bars denote loss of affinity, and green bars indicate increased affinity following mutation. (b) Interaction schematic of native 6GQ7 (PIK3CG) with zinc sulfate (DB09322). (c) Interaction schematic of the T857A mutant of PIK3CG (6GQ7) bound to zinc sulfate (DB09322), including the principal binding pocket (grey) used by most ligands. (d) Alluvial diagram illustrating the connections among genes, variants, and diseases, with line thickness between variant and disease proportional to the VDA score from DisGeNET.

Inspection of docked poses revealed that the ligand-binding cavities remained unchanged between native and mutated receptors, indicating that the SNVs did not alter the actual docking site. The altered residue is highlighted in red stick format in each structure and, except for the 6GQ7-T857A case, was consistently located away from the binding region.

To understand why the T857A variant (6GQ7) produced such a large decline in affinity when interacting with zinc sulfate (DB09322), the interacting residues in the wild-type and mutant complexes were compared using PyMol [41] and LigPlot+ [65]; these comparisons are displayed in **Figures 5b** and 5C. The spatial arrangement of the residues forming the pocket remained unchanged in both conformations, and the main pocket lay distant from the substituted site. Nonetheless, the interaction diagrams showed a reduction of one hydrogen bond in the mutant complex relative to the native form.

Gene, variant and disease association

Databases such as OMIM [66], DisGeNET [67], PharmGKB [68], and CTD [69] can be utilized to extract disease links or adverse reactions associated with gene targets or specific variants. DisGeNET aggregates records from CTDTM [69], UniProt [70], Orphanet [71], MGD [72], and RGD [73] for gene–disease mapping, and from ClinVar [74], the NHGRI-EBI GWAS Catalog [75], and GWASdb [76] for variant–disease associations. The platform computes a Gene-Disease Association (GDA) or Variant-Disease Association (VDA) score, ranging from 0 to 1 on a 1–10 scale, based on supporting evidence and data sources.

The rsIDs from the structure dataset were submitted to the DisGeNET interface, and a summary of outcomes is provided in **Figure 5d**. The alluvial diagram links genes, variants, and diseases, where the width of each variant–disease connection corresponds to its VDA value. Variants rs428073, rs1059702, and rs17158558 were identified as associated with Systemic Lupus Erythematosus, hematological-trait susceptibility, and HIRSCHSPRUNG disease, each reaching a VDA score of 0.7, reflecting support from at least one curated entry.

Drug-response implications of the variants in our structural dataset were probed using PharmGKB [68]. The gene set interacted with 69 drugs according to DGIdb, and 22 variants were linked to these interactions. Phenotype

information derived from PharmGKB indicated that rs1024323 and rs1801058 (F110V, A142V, Y292A, V486A, and F454A) influenced clinical outcomes—specifically Hypertension, Nephrosclerosis, and renal complications—by modifying the interaction between the GRK4 gene and the β-blocker Metoprolol.

Ligand similarity/diversity and toxicity analysis

For this section, 45 protein–drug combinations that showed changes in binding energy after docking were examined. Overall, seven PDB entries (6GQ7, 5TQY, 3GC9, 4TNB, 6I83) containing sixteen mutations and 28 pharmaceuticals were included. Every compound categorized as drug-like in the ligand library was used to assess chemical similarity. The results indicated that the drug set exhibits substantial structural heterogeneity **(Figure 6)**. The highest pairwise similarity, using Morgan2 and MACCS fingerprints, produced Tanimoto values of 0.40 and 0.70, respectively. Correspondingly, pairwise dissimilarity (1 − similarity) reached 0.98 (Morgan2) and 0.90 (MACCS).

To estimate toxicity, the ProTox-II platform [53] was applied. This computational system integrates cheminformatics and machine-learning models covering 46 toxicity categories. Predictions include acute toxicity (LD50), liver injury, cytotoxic effects, carcinogenic risk, mutagenicity, immune-related toxicity, Tox21 pathways, and off-target toxicity endpoints, using information from molecular similarity, pharmacophore cues, fragment behavior, and ML classifiers. Both in vitro datasets (e.g., Tox21, Ames mutagenicity, HepG2 cytotoxicity, immunotoxicity) and in vivo evidence (e.g., carcinogenicity, hepatotoxicity) underpin these models.

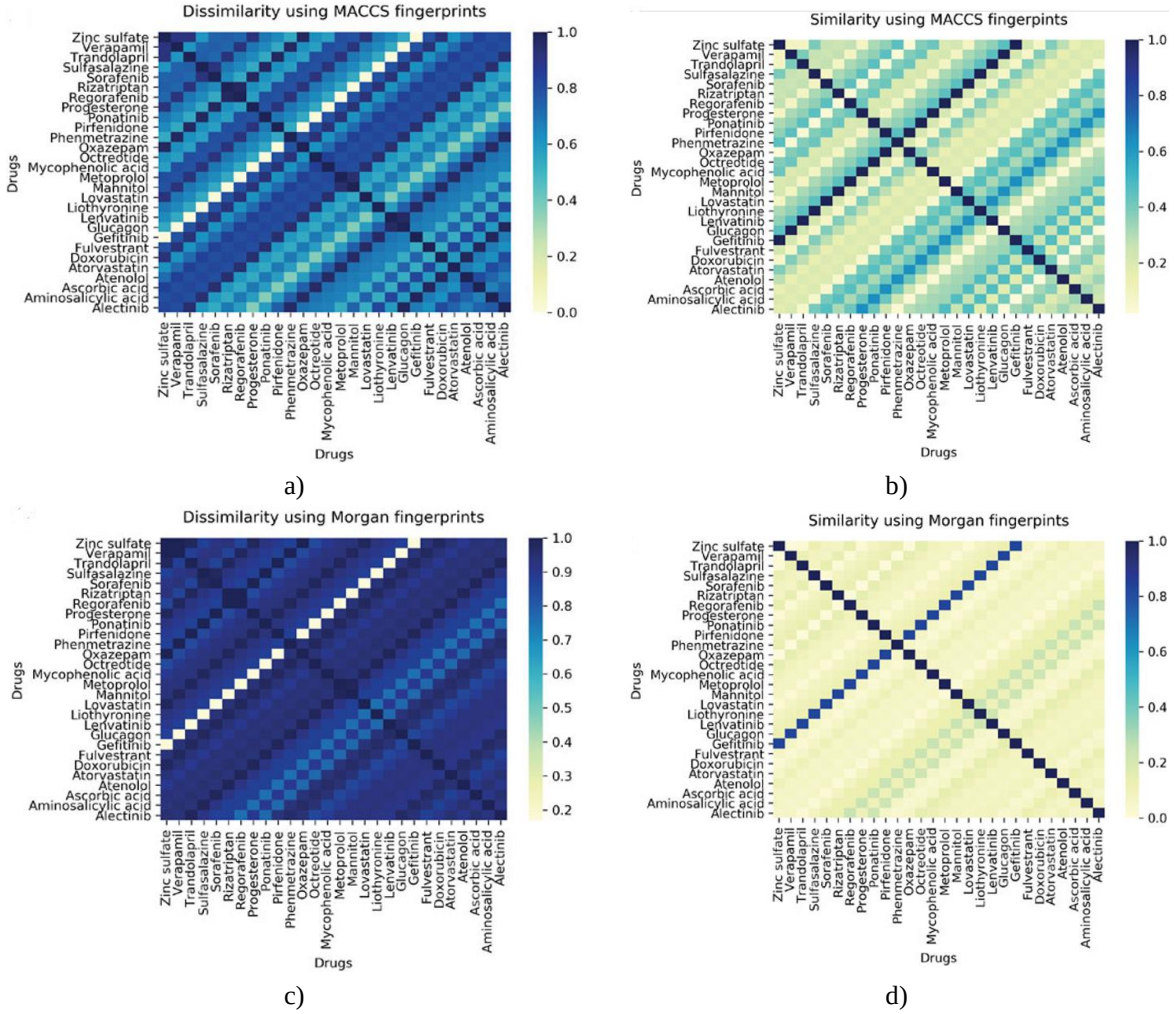


Figure 6. (a) MACCS-based dissimilarity heatmap. (b) MACCS-based similarity heatmap. (c) Morgan fingerprint dissimilarity heatmap. (d) Morgan fingerprint similarity heatmap. Tanimoto scores (0–1 scale) denote similarity or dissimilarity (1 − similarity).

According to ProTox-II predictions, mycophenolic acid (DB01024)—an immunosuppressive compound—was predicted to show hepatotoxic, immunotoxic, and cytotoxic properties and to interact with PIK3CG (PDB: 6GQ7, T857A variant). It also inhibited SR-MMP (mitochondrial membrane potential) with a confidence of 0.79. Regorafenib (DB08896) was another compound predicted to be hepatotoxic and active in two stress-mediated pathways (SR-MMP and SR-p53). Clinical reports associate Regorafenib with hypertension, stomatitis, and liver dysfunction [77, 78], although the mechanism for hypertension remains unclear. Progesterone (DB00396) was predicted to participate in six AOPs. Similar to progesterone, additional drugs may modulate NR-AR, leading to reduced AR signaling, compromised follicular activation, and possible fertility effects at the tissue or organism level [79].

Phenotypic drug-drug similarity

A single compound can interact with several protein targets, just as a single target can be recognized by multiple drugs. Such cross-interactions may contribute to therapeutic effects as well as clinically relevant side effects [80]. Exploring drug–target interaction patterns is essential for identifying new therapeutic uses of existing medications (repurposing) and for understanding adverse outcomes.

To identify drugs with similar phenotypic behavior in the IndiGen dataset, protein identifiers and associated ligands were compiled. This enabled the recognition of related drug behaviors within the structural dataset. A correlogram was constructed with drug names displayed on both axes; blue and red circles denote positive and negative correlations, respectively, with intensity proportional to the correlation value. Numerous strong correlations (dense blue markers) suggest that many compounds or targets display promiscuous binding.

For example, Fulvestrant and Rizatriptan are structurally unrelated **(Figure 6)**, yet they show high phenotypic similarity because they engage the same protein, demonstrating how kinases can accommodate diverse inhibitors. Another notable case involves Metoprolol (β1 blocker) and Atenolol (beta-blocker), which share a ligand similarity of 0.20 (Tanimoto coefficient) but have a phenotypic correlation of 1. Both drugs display overlapping adverse reactions—hypertension and renal/urinary disorders [1, 81], also supported by the EU ADR reporting data. Both medications are linked to GRK4, and multiple studies have documented that genetic alterations in GRK4 influence drug-induced hypertension [82] and renal impairment [83, 84]. Insights into variants affecting salt sensitivity and blood-pressure regulation could inform new therapeutic strategies and improve drug outcomes in the Indian population. Additionally, Metoprolol is metabolized poorly by CYP2D6, leading to elevated circulating levels [85]. Prior studies highlight how CYP2D6 genotypes modify cardiovascular treatment responses to beta-blockers [86].

Adverse drug reactions frequently arise from genes that harbor numerous variants and serve as targets for multiple therapeutic agents [87]. To examine how IndiGen variants are distributed across the human kinome, a kinome dendrogram was generated for all kinase genes with known drug interactions **(Figure 1)**. This analysis showed that the tyrosine kinase group contained the highest number of substitutions (5013) and was linked to a large repertoire of drugs (1978). Receptor tyrosine kinases (RTKs) regulate essential cellular events—cell growth, differentiation, and programmed cell death—and have long been exploited as pharmacological targets in oncology. Many TKIs are monoclonal antibody–based therapeutics used to treat cancers, immune-related disorders, and chronic inflammation [88]. Chemical-shift evaluation showed intra-class conversions between hydrophilic and non-polar amino acid categories. Even conservative shifts of this type can influence folding stability, enzymatic behavior, and overall structural integrity across species [89]. Although amino acid use in the genome typically mirrors codon availability [2], the amino-acid exchange profiles and chemical-shift patterns observed here **(Figure 2)** indicated that mutation trends did not correlate with codon redundancy. Instead, residues displayed a higher likelihood of transitioning into hydroxyl, aromatic, or acidic classes. The mutability distribution **(Figure 2c)** highlighted Arginine (R) as the most mutation-prone, likely due to CpG dinucleotides within its codons—sites known to mutate at elevated rates [90].

Population background is a major determinant of how single-nucleotide variants arise and propagate across groups, indicating a clear relationship between allele frequencies and ancestry. Even subtle amino-acid substitutions can manifest very differently across populations. When comparing IndiGen amino-acid exchange frequencies with those from the 1000 Genomes cohorts **(Figure 3b)**, AMR samples differed markedly from IndiGen patterns, whereas SAS individuals displayed strong similarity to IndiGen data. The Alanine-to-Threonine (A→T) substitution was statistically enriched and appeared in all datasets except IndiGen–SAS, underscoring shared genetic features between Indian and South Asian lineages.

A subset of variants exhibited high allele frequencies in Indian individuals but were rare in other groups, indicating population-specific mutations that may exert greater functional impact within India **(Figure 3d)**. Cross-referencing IndiGen allele frequencies with public variant databases revealed that many recurrent IndiGen substitutions remain uncharacterized, as these databases lack entries for them (IndiGen representing 1000 rigorously curated Indian genomes) **(Figure 3e)**. Because protein domains are structurally conserved and functionally essential, variants occurring within domain boundaries—or immediately after them—are more likely to influence folding, dynamics, and activity. A comparative positional analysis showed that numerous Indian variants were located in domain or post-domain regions **(Figure 3c)**.

Protein structural stability is among the most reliable metrics for estimating the phenotypic consequences of missense alterations. Such mutations can impair protein behavior either by destabilizing the fold or by altering residues critical for catalysis or binding interfaces. Disease-associated substitutions are often enriched at buried sites and at interaction surfaces [91], and previous studies have shown that buried amino acids frequently coincide with pathogenicity [92]. Contrary to these trends, structural comparisons between IndiGen and Humsavar variants demonstrated that residues with higher solvent exposure were more frequently mutated in IndiGen **(Figure 4a)**.

Mutations occurring in well-structured elements tend to have stronger destabilizing effects and are thus more likely to be pathogenic. Dynamut-based stability profiling identified 11 substitutions predicted to reduce structural stability, 7 of which were located within alpha-helical segments. IndiGen variants favored the helical regions, while Humsavar variants appeared equally distributed between helices and loop/coil portions **(Figure 4b)**. Several investigations have noted that secondary-structure elements differ markedly in their mutation tolerance, likely due to distinct patterns of non-covalent interactions stabilizing sheets versus helices [93]. Conservation analysis showed that Humsavar variants were concentrated in highly conserved residues, which aligns with their disease association. Hydrophobic packing and hydrogen bonding dominate protein stabilization; hydrophobic residues, in particular, tend to cluster inside the core. Interestingly, hydrophobicity trends diverged sharply: IndiGen variants showed a clear rise in hydrophobicity, while Humsavar variants exhibited reduced hydrophobic character **(Figure 4c)**.

Alterations in ligand-binding regions (LBSs) through nsSNVs can reshape a protein's conformation, its thermodynamic stability, and the strength of its interactions with small molecules. Earlier reports indicated that amino acids participating in ligand recognition accumulate mutations at a markedly higher frequency than residues elsewhere in the protein [94]. To examine whether a single residue change can modify a ligand's affinity for its receptor, we performed docking studies of FDA-approved compounds using both the native and variant protein models. Overall, the docking outcomes showed minimal shifts in binding energies when the substituted residue was located outside the active pocket; however, the T857A variant—where a polar residue is replaced with a hydrophobic one—resulted in the loss of two hydrogen bonds (4H) and consequently reduced the interaction strength between the ligand (zinc-sulfate) and the protein.

Additionally, the diverse binding characteristics of 12 different drugs interacting with 6GQ7 (PIK3CG) highlight the broad specificity of this kinase and provide valuable clues for studies involving polypharmacology and potential adverse outcomes. Expanding the evaluation of these and other PIK3CG-targeting agents, especially when grouped by phenotypic effects, may shed light on features useful for designing more selective kinase inhibitors. Furthermore, structurally unrelated medications may exhibit comparable ADR profiles if they produce similar phenotypic signatures, as seen with Metoprolol and Atenolol. Both are known to engage with GRK4 gene variants. Clarifying how these variants influence blood-pressure regulation and salt sensitivity could lead to new therapeutic pathways and more refined drug responses in Indian patients.

This work considered multiple dimensions that contribute to ADR occurrence—from chemical diversity and target engagement to genetic variability and its association with specific adverse reactions. The complexity of genome regulation and heterogeneous clinical outcomes means ADRs may stem from the intended target or from far more intricate biological interactions. In this context, our strategy can assist in predicting novel ADRs, clarifying mechanisms behind known reactions, and informing future experimental validation.

Although we focused on common variants within the Indian population, examining lower-frequency alleles in subsequent work will be essential for insights into rare disorders. Experimental corroboration will further strengthen the interpretations proposed here. Insights derived from IndiGen variant information could support a transition from a uniform drug-use paradigm to more population-tailored or even individualized therapy, ultimately improving treatment outcomes by minimizing side effects. Establishing a dedicated ADR-reporting

database would also be crucial for assessing risks associated with poly-therapy and mitigating drug-drug interaction issues in the Indian context.

The analyses conducted here—aimed at identifying ADRs enriched in the Indian population—will gain greater value when compared directly with established national ADR datasets. At present, this exploratory work may help shape approaches for both pre-clinical safety assessments and post-authorization surveillance of pharmaceuticals used in India. Agencies such as the EMA and FDA already maintain robust ADR-reporting platforms, generating real-world evidence (RWE) that provides insight into the circumstances under which ADRs occur. A major limitation in India is the lack of widespread awareness and the absence of a publicly accessible, comprehensive ADR-monitoring system. Integrating RWE with polypharmacology- and polytoxicity-based analyses would significantly advance the development of safer therapeutic practices that prioritize patient well-being.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Marian AJ. Molecular genetic studies of complex phenotypes. Transl Res. 2012;159(2):64-79.
2. Alwi ZB. The Use of SNPs in Pharmacogenomics Studies. Malays J Med Sci. 2005;12(2):4-12.
3. Sanghera DK, Ortega L, Han S, Singh J, Ralhan SK, Wander GS, et al. Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. BMC Med Genet. 2008;9:59.
4. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18(1):77.
5. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A. 2016;113(4):E440-9.
6. Schärfe CPI, Tremmel R, Schwab M, Kohlbacher O, Marks DS. Genetic variation in human drug-related genes. Genome Med. 2017;9(1):117.
7. Lauschke VM, Zhou Y, Ingelman-Sundberg M. Novel genetic and epigenetic factors of importance for inter-individual differences in drug disposition, response and toxicity. Pharmacol Ther. 2019;197:122-52.
8. 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
9. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
10. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177(1):26-31. doi:10.1016/j.cell.2019.02.048. Erratum in: Cell. 2019;177(4):1080.
11. Wei CY, Lee MT, Chen YT. Pharmacogenomics of adverse drug reactions: implementing personalized medicine. Hum Mol Genet. 2012;21(R1):R58-65.
12. Chan SL, Jin S, Loh M, Brunham LR. Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. Pharmacogenomics. 2015;16(10):1161-78.
13. Khalil H, Huang C. Adverse drug reactions in primary care: a scoping review. BMC Health Serv Res. 2020;20(1):5.
14. Impicciatore P, Choonara I, Clarkson A, Provasi D, Pandolfini C, Bonati M. Incidence of adverse drug reactions in paediatric in/out-patients: a systematic review and meta-analysis of prospective studies. Br J Clin Pharmacol. 2001;52(1):77-83.
15. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res. 2021;49(D1):D1144-D1151.

16. 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(1):308-11.
17. Alteri E, Awadzi K, Close P, Couper M, Dodoo A, Fomundam H, et al. Good clinical practice guidelines: drug development research in resource-limited countries. Geneva: CIOMS; 2005.
18. Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, et al. Genetic evidence on the origins of Indian caste populations. Genome Res. 2001;11(6):994-1004.
19. Sengupta D, Choudhury A, Basu A, Ramsay M. Population Stratification and Underrepresentation of Indian Subcontinent Genetic Diversity in the 1000 Genomes Project Dataset. Genome Biol Evol. 2016;8(11):3460-70.
20. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of discovering population-specific disease-associated genes in South Asia. Nat Genet. 2017;49(9):1403-7.
21. Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: the genetics of variable drug responses. Circulation. 2011;123(15):1661-70.
22. Bhosle D, Sayyed A, Bhagat A, Shaikh H, Sheikh A, Bhopale V, et al. Comparison of generic and branded drugs on cost effective and cost benefit analysis. Asian J Med Diagn Res. 2016;3(1):1-6.
23. Jain A, Bhoyar RC, Pandhare K, Mishra A, Sharma D, Imran M, et al. IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. Nucleic Acids Res. 2021;49(D1):D1225-D1232.
24. Bhullar KS, Lagarón NO, McGowan EM, Parmar I, Jha A, Hubbard BP, et al. Kinase-targeted cancer therapies: progress, challenges and future directions. Mol Cancer. 2018;17(1):48.
25. Paul MK, Mukhopadhyay AK. Tyrosine kinase - Role and significance in Cancer. Int J Med Sci. 2004;1(2):101-15.
26. Lee NH. Pharmacogenetics of drug metabolizing enzymes and transporters: effects on pharmacokinetics and pharmacodynamics of anticancer agents. Anticancer Agents Med Chem. 2010;10(8):583-92.
27. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
28. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. 2017;45(D1):D854-D859.
29. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36(Database issue):D901-6.
30. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. Nucleic Acids Res. 2002;30(1):412-5.
31. van den Boom D, Wjst M, Everts RE. MALDI-TOF mass spectrometry. Methods Mol Biol. 2013;1015:71-85.
32. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017:PO.17.00011.
33. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10(1):25.
34. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart--biological queries made easy. BMC Genomics. 2009;10:22.
35. Eid S, Turk S, Volkamer A, Rippmann F, Fulle S. KinMap: a web-based tool for interactive navigation through human kinome data. BMC Bioinformatics. 2017;18(1):16.
36. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007;9(03):90-5.
37. McKinney W. Data structures for statistical computing in Python. scipy. 2010;445(1):51-6.
38. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357-62.
39. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol. 2014;1079:105-16.
40. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47(W1):W636-W641.
41. Schrödinger L, DeLano W. Pymol [Internet]. 2020 [cited 2025 Dec 8]. Available from: http://www.pymol.org/pymol. [Dataset]
42. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

43. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. 2018;46(W1):W350-W355.
44. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.
45. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol. 1994;238(5):777-93.
46. Hubbard SJ, Thornton JM. Naccess [Computer program]. London: Department of Biochemistry Molecular Biology, University College London; 1993.
47. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.
48. Dallakyan S, Olson A. Participation in global governance: Coordinating 'the voices of those most affected by food insecurity'. Glob Food Secur Gov. 2015;1263:1-11.
49. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, et al. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem. 2007;2(6):861-73.
50. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics. 2010;26(12):i246-54.
51. Fernández-de Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform. 2017;9:9.
52. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12(1):51.
53. Banerjee P, Eckert AO, Schrey AK, Preissner R. ProTox-II: a webserver for the prediction of toxicity of chemicals. Nucleic Acids Res. 2018;46(W1):W257-W263.
54. Jalencas X, Mestres J. On the origins of drug polypharmacology. Med Chem Comm. 2013;4(1):80-7.
55. Prinz J, Vogt I, Adornetto G, Campillos M. A Novel Drug-Mouse Phenotypic Similarity Method Detects Molecular Determinants of Drug Effects. PLoS Comput Biol. 2016;12(9):e1005111.
56. Bachtiar M, Lee CG. Genetics of population differences in drug response. Curr Genet Med Rep. 2013;1(3):162-70.
57. French S, Robson B. What is a conservative substitution?. J Mol Evol. 1983;19(2):171-5.
58. Mori M, Yamada R, Kobayashi K, Kawaida R, Yamamoto K. Ethnic differences in allele frequency of autoimmune-disease-associated SNPs. J Hum Genet. 2005;50(5):264-6.
59. Mattei J, Parnell LD, Lai CQ, Garcia-Bailo B, Adiconis X, Shen J, et al. Disparities in allele frequencies and population differentiation for 101 disease-associated single nucleotide polymorphisms between Puerto Ricans and non-Hispanic whites. BMC Genet. 2009;10:45.
60. Gong S, Blundell TL. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. PLoS One. 2010;5(2):e9186.
61. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44(W1):W344-50.
62. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39(Database issue):D411-9.
63. Kumar A, Biswas P. Effect of site-directed point mutations on protein misfolding: A simulation study. Proteins. 2019;87(9):760-73.
64. Fauchère JL, Charton M, Kier LB, Verloop A, Pliska V. Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res. 1988;32(4):269-78.
65. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model. 2011;51(10):2778-86.
66. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat. 2011;32(5):564-7.
67. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833-D839.

68. Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. Methods Mol Biol. 2013;1015:311-20.
69. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 2019;47(D1):D948-D954.
70. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480-D489.
71. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum Mutat. 2012;33(5):803-8.
72. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE; Mouse Genome Database Group. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 2015;43(Database issue):D726-36.
73. Shimoyama M, De Pons J, Hayman GT, Laulederkind SJ, Liu W, Nigam R, et al. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res. 2015;43(Database issue):D743-50.
74. 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(D1):D862-8.
75. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(Database issue):D1001-6.
76. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431-2.
77. Krishnamoorthy SK, Relias V, Sebastian S, Jayaraman V, Saif MW. Management of regorafenib-related toxicities: a review. Therap Adv Gastroenterol. 2015;8(5):285-97.
78. De Wit M, Boers-Doets CB, Saettini A, Vermeersch K, de Juan CR, Ouwerkerk J, et al. Prevention and management of adverse events related to regorafenib. Support Care Cancer. 2014;22(3):837-46.
79. Pivonello C, Muscogiuri G, Nardone A, Garifalos F, Provvisiero DP, Verde N, et al. Bisphenol A: an emerging threat to female fertility. Reprod Biol Endocrinol. 2020;18(1):22.
80. Kanji R, Sharma A, Bagler G. Phenotypic side effects prediction by optimizing correlation with chemical and target profiles of drugs. Mol Biosyst. 2015;11(11):2900-6.
81. Charles C, Ferris AH. Chronic Kidney Disease. Prim Care. 2020;47(4):585-95.
82. Frey MK, Dao F, Olvera N, Konner JA, Dickler MN, Levine DA. Genetic predisposition to bevacizumab-induced hypertension. Gynecol Oncol. 2017;147(3):621-5.
83. Armando I, Villar VA, Jose PA. Genomics and Pharmacogenomics of Salt-sensitive Hypertension. Curr Hypertens Rev. 2015;11(1):49-56.
84. Sanada H, Yoneda M, Yatabe J, Williams SM, Bartlett J, White MJ, et al. Common variants of the G protein-coupled receptor type 4 are associated with human essential hypertension and predict the blood pressure response to angiotensin receptor blockade. Pharmacogenomics J. 2016;16(1):3-9.
85. Banerjee P, Dunkel M, Kemmler E, Preissner R. SuperCYPsPred-a web server for the prediction of cytochrome activity. Nucleic Acids Res. 2020;48(W1):W580-W585.
86. Rau T, Heide R, Bergmann K, Wuttke H, Werner U, Feifel N, et al. Effect of the CYP2D6 genotype on metoprolol metabolism persists during long-term treatment. Pharmacogenetics. 2002;12(6):465-72.
87. Wilke RA, Lin DW, Roden DM, Watkins PB, Flockhart D, Zineh I, et al. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. Nat Rev Drug Discov. 2007;6(11):904-16.
88. Bennasroune A, Gardin A, Aunis D, Crémel G, Hubert P. Tyrosine kinase receptors as attractive targets of cancer therapy. Crit Rev Oncol Hematol. 2004;50(1):23-38.
89. Rodriguez-Larrea D, Perez-Jimenez R, Sanchez-Romero I, Delgado-Delgado A, Fernandez JM, Sanchez-Ruiz JM. Role of conservative mutations in protein multi-property adaptation. Biochem J. 2010;429(2):243-9.
90. de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Comput Biol. 2013;9(12):e1003382.
91. Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. Sci Rep. 2020;10(1):15387.

92. Iqbal S, Pérez-Palma E, Jespersen JB, May P, Hoksza D, Heyne HO, et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. Proc Natl Acad Sci U S A. 2020;117(45):28201-28211.
93. Abrusán G, Marsh JA. Alpha Helices Are More Robust to Mutations than Beta Strands. PLoS Comput Biol. 2016;12(12):e1005242.
94. Kim P, Zhao J, Lu P, Zhao Z. mutLBSgeneDB: mutated ligand binding site gene DataBase. Nucleic Acids Res. 2017;45(D1):D256-D263.