Comprehensive assessment of Indian variations in the druggable kinome landscape highlights distinct insights at the sequence, structure and pharmacogenomic stratum.

Gayatri Panda[1‡], Neha Mishra[1‡], Disha Sharma[2,3], Rahul C. Bhoyar[3], Abhinav Jain[2,3], Mohamed Imran[2,3], Vigneshwar Senthilvel[2,3], Mohit Kumar Divakar[2,3], Anushree Mishra[3], Parth Garg[1], Priyanka Banerjee[4], Sridhar Sivasubbu[2,3], Vinod Scaria[2,3], Arjun Ray[1*]

1 Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla, India.

2 Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India.

3 CSIR-Institute of Genomics and Integrative Biology, Mathura Road, Delhi-110020, India.

4 Institute for Physiology, Charité-University Medicine Berlin, 10115 Berlin, Germany.

‡These authors contributed equally to this work.

* arjun@iiitd.ac.in

Keywords

Indian genetic variations, IndiGenome Consortium, Pharmacogenmics, single nucleotide variants, docking, adverse drug reactions

Abstract

The population diversity in India contains a treasure of clinically relevant rare mutations which may have evolved differently in different subpopulations. While there are many subgroups present in the nation, the publicly available database like the 1000 Genome data (1KG) contains limited samples for indian ethnicity. Such databases are critical for the pharmaceutical and drug development industry where the diversity plays a crucial role in identifying genetic disposition towards adverse drug reactions. A qualitative and comparative sequence and structural study utilizing variant information present in the recently published, largest curated Indian genome database (Indigen) and the 1000 Genome data was performed for variants belonging to the kinase coding genes, the second most targeted group of drug targets. The sequence level analysis identified similarities and differences among different populations based on the SNVs and amino acid exchange frequencies whereas comparative structural analysis of IndiGen variants was performed with pathogenic variants reported in UniProtKB Humsavar data. The influence of these variations on structural features of the protein, such as structural stability, solvent accessibility, hydrophobicity, and the hydrogen-bond network were investigated. In-silico screening of the known drugs to these Indian variation-containing proteins reveal critical differences imparted in the strength of binding due to the variations present in the Indian population. In conclusion, this study constitutes a comprehensive investigation into the understanding of common variations present in the second largest population in the world, and investigating its implications in the sequence, structural and pharmacogenomic landscape.

Introduction

The presence of single nucleotide polymorphisms imparts a genetic basis of human complex diseases and human phenotypic variations [A.J. Marian, 2013]. As per various reports, SNPs are found to be responsible for defining the risk of an individual's susceptibility to various drug responses and illnesses [Alwi, 2005]. The distribution of allele frequency of SNPs provides relevant information about the evolution, migration, and genetic structure of a population [Sanghera et al., 2008]. Most of the genetic variant-related data come from databases like the

1000 Genome database, GnomAD, Exac Database, containing ethnicity-wise variant information which is largely Eurocentric. It is so because majority of the studies that are performed to associate genetic variants with diseases, like the Genome-Wide Association Studies (GWAS) have been conducted mainly on the European population (78%) followed by Asian(10%), African(2%), Hispanic(1%), and other ethnicities (<1%) [Sirugo et al., 2019] neglecting the Indian population. It creates an information bias leading to a population-specific disease assessment analysis leaving the African and Indian populations under-studied and under-consulted. These population-specific SNPs deviate in variation patterns from other over-represented populations causing health and diagnosis disparities[Chan et al., 2015] [Wei et al., 2012].

Adverse drug reactions (ADRs) are a major contributor to morbidity and mortality. The presence of a genomic variation in genes coding for drug transport and metabolism have been associated with inter-individual differences in drug response and ADR risks. Several SNP-related studies have shown that variants can modulate the efficacy of a drug leading to adverse drug reactions (ADRs) [Impicciatore et al., 2001] [Sanghera et al., 2008]. Drug Gene Interaction Database (DGIdb) organizes the drug-gene interactions from various papers, databases and web resources[Freshour et al., 2021]. dbSNP, a curated database alone contains 38 million SNPs which makes timely maintenance, integration, and correction a cumbersome process [Sherry et al., 2001]. SNPs are a vital and decisive factor for finalizing a therapeutic approach and selection of drug and their dosages [Alwi, 2005]. European population being primary conduct of drug trials prior to approval and marketing of drugs could be one of the factors on the occurrence of ADRs[Clinical and Guidelines, 2006]. Hence, this prioritizes the need for population-specific pharmacogenomic analysis and integration of gene, drug, pathway, and potential drug-target related information.

Genetic studies of populations from the Indian subcontinent are important due to India's large share of the global population, complex demographic background, and unique social structure. Indo-genomic variation is fascinating due to the diverse ancestral components, social categorization of people, endogamy practised in different cultures, and dynamic and ancient admixture events that the Indian population has experienced over a long period of time.[Bamshad et al., 2001]. Reports suggest that the population expansion in India (post-agriculture) has led to the emergence of a huge amount of genomic diversity exceeding the genetic diversity of the whole of Europe[Sengupta et al., 2016]

The practice of endogamy in various communities disturbs the frequency of a disease in different sub-groups of the Indian population [Nakatsuka et al., 2017], indicating that genetic divergence can also affect the efficacy of the drug. Globally, India is the largest generic drug provider [Bhosle et al., 2016](16). Regardless of the Indian genetic diversity, the current healthcare system in India follows the same drug therapy as in Europe and America. The use of genetic information, experiments, and other types of molecular screening helps a practitioner to choose an appropriate therapy for the first time, avoiding the time-consuming and expensive trial-and-error medication cycle. Extensive research on the population diversities and related SNPs causing the different inter-individual drug responses is the need of the hour for efficient treatment design. IndiGen programme was initiated with an aim to collect sequencing data of thousands of individuals from diverse ethnic groups in India and develop public health technologies applications using this population genome data[Jain et al., 2021].

In our present work, we conducted the first exhaustive and comparative study of common Indian-specific variants (using IndiGen data) with other populations to identify the population-specific variations causing a difference in drug responses and ADRs. This pharmacogenomic study was executed by keeping a focus on druggable genes of kinase's family, the second most targeted group of drug targets after the G-protein coupled receptors [Bhullar et al., 2018]. The human genome encodes 538 protein kinases[Berndt et al., 2017]. Many of these kinases are associated with deadly diseases like cancer [Paul and Mukhopadhyay, 2012]. Most of the kinase targeting drugs have been tested and approved based on the trials done on European populations and it is possible that the same drugs might exhibit a deviation in efficacy and response on Indian population. The presence of a SNP in functionally important genes have higher chances of deleterious impact by either affecting drug-gene interaction or by causing structural changes at the protein level leading to disruption of the drug-binding sites [Lee, 2010].

As a result, interpreting the number of mutations and their effect on the structure, stability, and function of the protein is crucial. Any destabilising non-synonymous SNP (nsSNP) will cause the drug's metabolic process to be disrupted. This study was carried out at both sequence and structure level to examine the effect of missense mutations in Drug-Gene interaction as well as the structural changes caused by these mutations at the protein level.The sequence-level analysis was implemented to perceive the similarities and differences among different populations based on the single nucleotide variants (SNVs) and amino acid exchange frequencies. The effect of these variants on structural properties of the protein, like structural stability, solvent-accessibility, hydrophobicity, and the hydrogen-bond network were measured by utilizing different structural analysis tools. Any modification in protein-ligand binding due to the presence of SNVs was analyzed by molecular docking method. A comparative structural analysis was conducted using UniProtKB Humsavar data.This work will help us understand the variability caused by these variants and thus could guide us in deciphering the effect of SNP in the efficacy of the drug-protein/gene interaction.

Results

Indian variations in the kinome landscape

To first get an overview of the Indian variations present in the druggable kinome landscape, an exhaustive annotation of variation containing 545 kinase coding genes found in the IndiGen data and the families along with the number of drugs associated with them were mapped (Figure 1). It was observed that despite having more drug-gene interactions, very few genes from the atypical protein kinases family contained missense mutations. The SNVs in a conserved protein region can influence the protein structure and its stability and can affect the protein-protein or protein-drug binding affinity. A gene with more variation and multiple marketed drugs has a greater tendency of causing ADRs. It was found that the tyrosine kinase family, which has a maximum (1978) number of FDA-approved drugs consists of the maximum (5013) number of variations. The class of kinases other than TK (Tyrosine Kinase) like the CMGC (cyclin-dependent kinase (CDK), mitogen-activated protein kinase (MAPK), glycogen synthase kinase

(GSK3), CDC-like kinase (CLK), TLK (Serine/threonine-protein kinase tousled-like 1) and AGC (PKA, PKC, PKG) contain a large number of variations i.e., 10518, 1193, and 2943 respectively but the number of drugs with known Drug-Gene interactions were limited to 213, 185, and 339, which was comparatively less than the Tyrosine Kinase family. The CK1(casein kinase 1) class among all others contains the lowest (275) number of variations and lowest (18) drug-gene interactions. Kinase families associated with 545 kinase coding genes with number of drugs and SNPs observed in each class are shown in Supplementary Table S10.

Analysis of the sequence-level differences of Indian variations in context with other populations

The genetic variation pattern in the Indian population was elucidated by generating an amino acid exchange matrix for all SNPs reported for 545 druggable kinase genes in IndiGen data. Figure 2A represents an amino-acid exchange matrix for the Indian population where the X and Y axis correspond to amino acids at the reference and alternative alleles in IndiGen data. Results from the analysis revealed that nearly 68% of Arginine(R) converts to Tryptophan (W) i.e., a hydrophobic amino acid converting to a basic polar amino acid. Similarly, 58% of Cystine (C) observed at reference SNP sites gets converted into Tyrosine (Y) i.e, a polar uncharged amino acid converting to polar aromatic amino acid. Other amino acid conversions with moderate frequency (40-50%) were Leucine(L) to Phenylalanine(F) both non-polar amino-acids, Lysine(K) to Glutamic acid(E) which involved basic to acidic conversion, and Asparagine(N) to Aspartic acid(D), an amidic to acidic conversion. It was worth noticing that regardless of having a maximum number of codons (6) coding for Serine(S) and Leucine(L), the amino acid exchange for these two residues were comparatively lower than Tyrosine (Y) and Tryptophan (W) which have only one associated codon.

In order to comprehend the inter-conversion distribution of the chemical groups present in mutating amino acids and develop a coherent relation of these amino-acid exchanges with physicochemical property, a chemical shift analysis was performed. The mutating amino acids were classified on the basis of their R-groups into 12 chemical classes (Aliphatic, Hydroxyl, Cyclic, Aromatic, Basic, Acidic, Sulpho, Amides, Non-polar, Uncharged polar, Hydrophobic

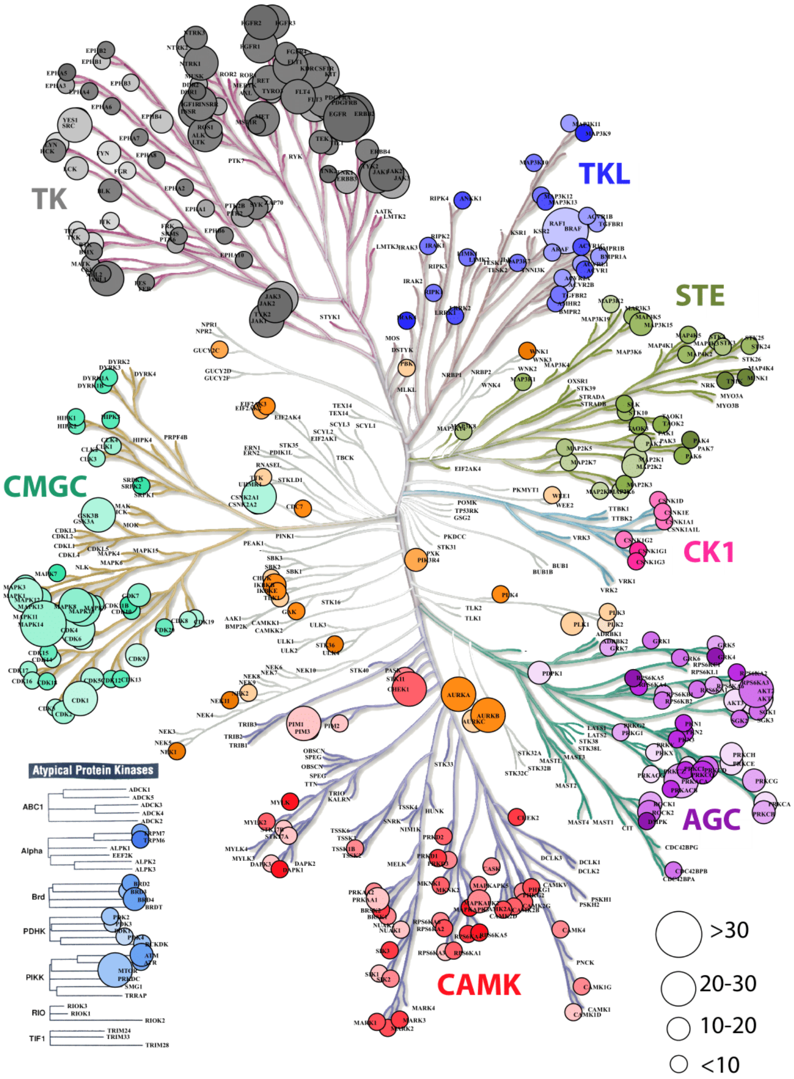

Fig 1. Dendrogram representation kinase coding genes in IndiGen data using KinMapbeta. The circle size represents the number of drug molecules available for a gene with known drug-gene interaction. The class of kinase is highlighted with a unique colour and the colour gradient in each data circle represents the number of variations present in IndiGen data for that gene.

, and Hydrophillic). In Figure 2B, X axis represents 12 chemical classes while on the Y-axis the distribution of the delta amino acid count of reference and altered amino acids for each chemical class (shown by 12 colors)has been shown. It can be observed that most of the residues from the hydrophobic class (Gly, Ala, Val, Pro, Leu, Ile, Met, Trp, Cys, and Phe) have mutated to either nonpolar (Gly, Ala, Val, Pro, Leu, Ile, Met, Trp, Phe), other hydrophobic (Gly, Ala, Val, Pro, Leu, Ile, Met, Trp, Phe, Cys) or aliphatic (Gly, Ala, Val, Leu, Ile) amino acid classes.

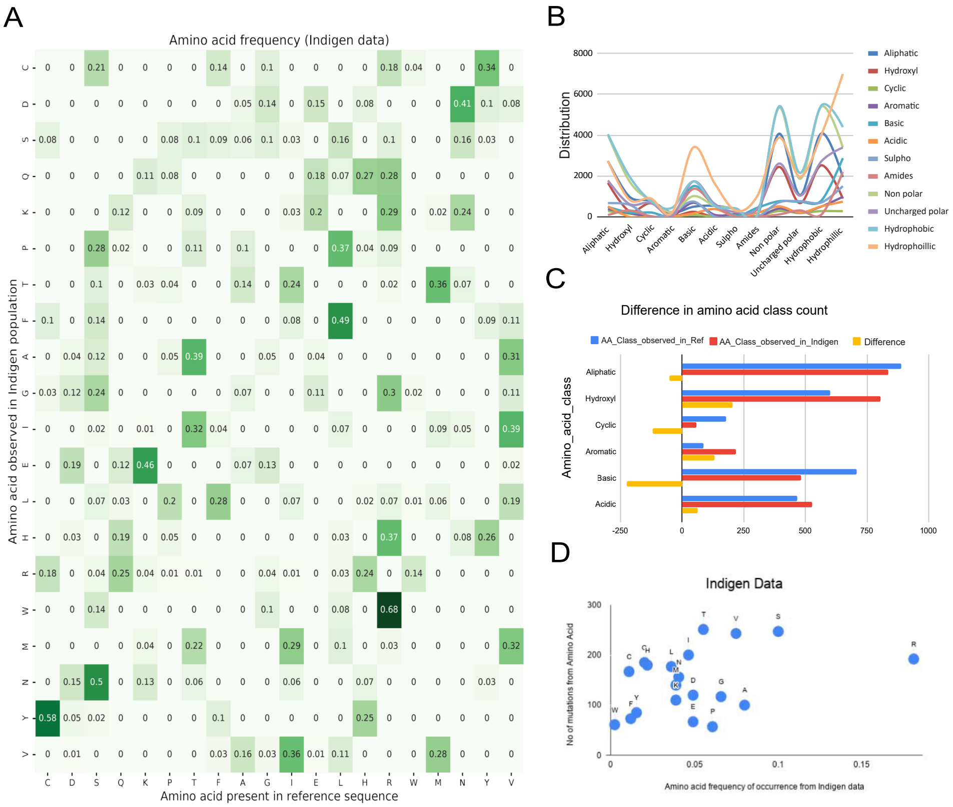

Fig 2. Sequence Analysis using SNPs reported for 545 druggable kinase Coding genes in IndiGen Data: A. Amino-acid exchange matrix for reference and altered amino acids of SNPs in Indigen data. B. Chemical shift observed among the reference and altered amino acids at SNP sites reported in Indigen data. C. Chemical changes observed among the reference amino acid in RefSeq(hg38) and altered amino acids at SNP sites reported in Indigen data. D. Scatter plot of mutability scores for each amino acid type in Indigen data

Inter-class or intra-class amino acid exchanges were also explored by looking at the classes associated with the peaks of each of the distributions. Intra-class conversions were observed for amino-acids belonging to hydrophilic, hydrophobic, and non-polar classes (peaks for the same class) supporting conservative replacement[French and Robson, 1983]. Additionally, several mutating amino acids have shown inter-class conversions such as aliphatic and hydroxyl amino acids converted to hydrophobic or non-polar amino acids as well as amino-acids in basic and acidic classes have converted to amino-acid from hydrophillic class. It was observed that many amino-acids have shown tendency for conversion to an amino-acid belonging to non-polar or hydrophobic amino acids(6/12 classes). The distribution for hydrophilic class was slightly different from others with a very prominent peak at basic class, indicating that these amino acids are more likely to exchange with the basic amino acids like Lys, Arg and His, apart from intra-class conversions.

In support of this, one more analysis was performed in which the reference amino-acids were taken as per RefSeq hg38 sequence whereas altered amino-acid at the same SNP site was taken from Indigen data. These amino-acids were classified into six different 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)) to avoid any repetition of amino acids. The difference in amino-acid counts at the SNP site for each class was then plotted. In Figure 2C, Y-axis represents six chemical classes of amino acids with respect to the amino acid counts in RefSeq(hg38) and IndiGen data. This chemical shift analysis confirms that there is a net loss in basic, cyclic and aliphatic amino acid class whereas a net gain is observed in the hydroxyl, aromatic and acidic amino acid classes. It is important to note here that while the hydroxyl, Aromatic and Acidic amino acid class contains 2,3, and 2 amino acids respectively, it still contributes to the net gain; while the aliphatic class, with maximum number of amino acid, showed a net loss in amino acid count. This clarifies that the net gain or loss in any amino acid class is independent of its size.

In order to understand the relationship between the mutational frequency of a specific amino acid with its frequency of occurrence in the IndiGen data, a mutability score for each amino acid type was calculated. In Figure 2D, mutability scores for amino acids observed in IndiGen data are shown. The plot shows that Arginine (R) is the most observed amino acid with >0.15 frequency of occurrence whereas Tryptophan(W) is the least observed residue at the reference SNP site in IndiGen data. Amino acids like Valine, Serine, and Threonine have shown a greater propensity to get mutated as compared to other amino acids. These observations are also in agreement with the inferences made from the amino acid exchange matrix(Figure 2A). In Figure 2A, Arginine(R) can be seen as the most mutable amino acid with the greatest amino-acid exchange frequency(maximum frequency - 0.68) and Tryptophan as the least mutable amino-acid (maximum frequency-0.14).

After establishing an in-depth description of the Indian population, a comparative sequence analysis was performed for the variants in IndiGen data with other populations, such as European (EUR), American (AMR), African (AFR), South Asian (SAS), and East Asian (EAS) populations from the 1000 genome data. In Figure 3A, we observe that the mutation from Cystine(C) to Tyrosine(Y), and Arginine(R) to Tryptophan(W) was quite prevalent in all the populations except in American(AMR). A similar pattern of amino acid exchange and mutability is observed among different population although the frequencies varied.

Reports have suggested about the relationship between allele frequency and ethnicity of SNPs[Mattei et al., 2009, Mori et al., 2005]. Allele frequency(AF) plot (Figure 3B) was generated by calculating the minor allele frequency of variants in each ethnic group so as to explore how these variants differed among different populations(Indian and 1000 genome populations). The analysis revealed that allele frequency curve followed by SNPs in IndiGen and South-Asian were quite similar and comparatively different from others with very high AF for some variants belonging to GRK4 gene,i.e, Y292A and V486A. This indicates there is a considerable difference in allele frequency between Eurocentric and the understudied (AFR, Indian) populations. A similar AF plot (Figure 3C) was generated by comparing allele frequencies for SNPs in IndiGen data with their allele frequencies in different publicly available databases.

In order to identify all the common and rare population-specific SNPs among variants of different population, analysis was carried out using SNPs reported for twelve genes present in our structure data (without any allele frequency filter). In Figure 3D, a Venn diagram showing unique and common SNPs for twelve genes among different populations (Indian, SAS, EUR, AFR, EAS and AMR) is shown. It was observed that the IndiGen variants have very less overlap with the variants of other population(majorly European and South Asian) in 1000 genome data indicating specificity of IndiGenic variants. These non-overlapping variants draw a distinction between Indian and 1000 genome population. The South-Asian population contains samples for Gujrati Indian from Houston (GIH), Punjabi from Lahor, Pakistan (PJL), Bengali from Bangladesh (BEB), Sri Lankan Tamil from the UK (STU) and Indian Telugu from the UK (ITU). Despite containing variants from Indian ethinicity, South-Asian SNPs have shown less overlap with IndiGenic variants supporting the hypothesis that specific subgroups have conserved mutation that has spread through that population and evolved differently through time [Christensen et al., 2003]. This observation stresses on the fact that behavioural and environmental changes(epigenetics) might lead to genetic differences among populations.

Upon having an in-depth understanding of the effects of variations on the sequence, we next explored the effect on the protein's structure. Firstly, protein domain analysis was done to find out the number of SNPs falling within the domains and the number of SNPs that are falling before and after the domains(Figure 3E). In order to understand the impact of SNVs at protein structure, the protein sequences were divided into three parts – domain regions, post-domain region and pre-domain region, indicating the position of a variant based on its presence before, within or after protein domain. It was observed that for Indigen data 952 variants were falling within the domain while 226 variants for were present in post domain region whereas only twelve variants were observed in the pre-domain region. Similarly for variants in 1000 genome data for European, American, African, East Asian and South Asian populations were categorised into pre-domain, post domain and within domain variants. Surprisingly all the populations from 1000 genome and Indigen data revealed a larger bias for a SNV to fall in within the protein domain or post-domain region as compare to pre-dromain region.

Structure level comparison of IndiGen and Disease-causing variants

To further understanding the SNV's effect on the protein structures, IndiGen structure dataset was constructed by taking into account only variants of druggable kinases lying within the crystal length, thus giving only twelve kinase genes and corresponding 22 variants. Disease causing variants corresponding to these 12 genes were extracted from Humsavar data (217 variants) and compared. The structural characteristics like distribution of solvent -accessibility, secondary structure, conservation score and change in hydrophobicity of variants/variant residues in IndiGen structure data and Humsavar data were compiled and compared. For solvent accessibility comparison (in Figure 4A), a cutoff of 5% solvent exposure was applied onto the Naccess results for variants in both datasets to distinguish between buried and exposed residues. The results revealed most mutations are observed in the exposed residues in both the datasets. This is in line with the conventional study shown by a group that states more than

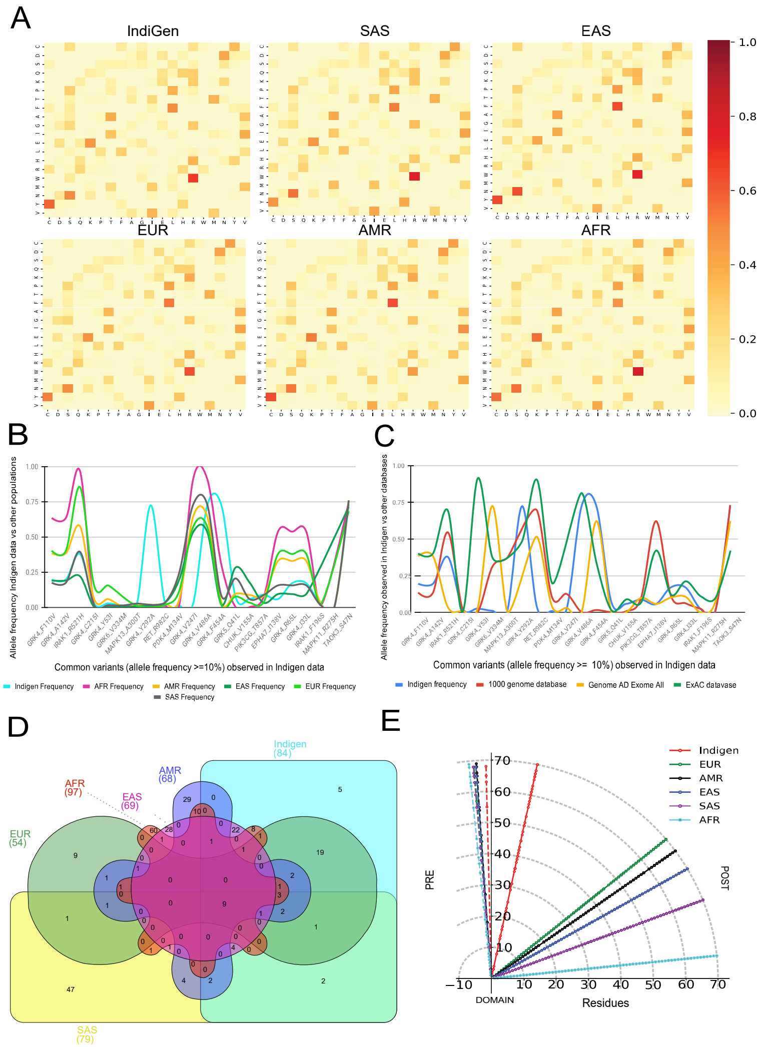

Fig 3. A. Comparing the trend of amino acid exchange among different populations from 1000 genome project and with Indian population. The heatmap was generated on the basis of the allele frequency of variants in IndiGen and other populations of 1000 genome data. The colour intensity of each cell is proportional to the frequency of amino acid exchange from one specific amino acid to another among all the databases. B. Comparing IndiGen specific variations (22 variants) with allele frequency ≥10% different populations with 1000 genome data. On the X-axis common IndiGen variants qualifying the filters used for structure data are shown with gene and variant names(22 variants) whereas on the Y-axis, allele frequency for these variants in IndiGen and other populations is plotted. C. IndiGen specific SNPs(22 variants) with AF≥ 10% observed in different databases like 1000 genome project, Genome AD exome data and Exac database; with IndiGen variations on X-axis and their allele frequencies in different databases on Y-axis. D. Venn-diagram of common Indian variants (allele frequency (AF) ≥10%) among different populations. E. Variations lying pre-, within and post domain was mapped where the angle of the lines are a function of the number of variations, with the y-axis ”Domain” location as zero, and where larger variations in a population shall bear larger weight. Variations were plotted on the basis of their distance from post and pre-domain location.

60% of solvent exposed SNPs have a disease association[Gong and Blundell, 2010]. In IndiGen data, 81.8% of variant residues (22 residues) were found to be exposed which was roughly equal to solvent exposure of residues in Humsavar data with 81.1% exposed residues (74 residues). No appreciable difference was observed in solvent accessibility for variants in both datasets. The secondary structure preference of variants in both the datasets revealed that variant residues in IndiGen data have a slight preference to occur on alpha-helix part of the protein while the variants in Humsavar data share equal secondary structure preference for their occurrence either in alpha-helix or in loop/random coil of a protein (Figure 4B).

Residue conservation scores for variants in IndiGen structure data (22 residues) and in Humsavar data (74 residues) were calculated using Consurf [Ashkenazy et al., 2016]. A density plot showing the distribution of conservation score for variants in both the datasets is shown in Figure 4C. The Humsavar density curve follows nearly normal distribution while the IndiGen curve follows a bimodal distribution with two peaks. Moreover, the median line divides the area under the curve into two equal halves. The median line for Humsavar data (0.007) was present closer to 0 than IndiGen data's median (0.358). Hence, in order to elucidate the percentage of residues with more or less conservation, a threshold value of -1/+1 relative conservation score was considered. It was observed that the percentage of highly conserved residues (with Consurf conservation score greater than -1) was more in Humsavar distribution (steeper) than in IndiGen. Likewise, the percentage of highly variable residues (with conservation score >1) adhering to the area under the curve on the right of +1 was more for IndiGen data than for Humsavar data, indicating that Humsavar data has a higher percentage of residues that are involved in variations, being more conserved.

The distribution of change in hydrophobicity from reference to altered residue for variants in Humsavar and IndiGen structure data is shown in Figure44D. The medians for both the distributions were found next to each other and very close to 0, suggesting that the percentage of variations with increase or decrease in hydrophobicity is almost equal in both the datasets. In order to find out the percentage of residues with some significant change in hydrophobicity, a threshold value of -2 was considered for increase in hydrophobicity whereas +2 threshold was taken for decrease in hydrophobicity. It was observed that the percentage of varying residues with significant increase in hydrophobicity was observed for IndiGen structure data whereas the percentage of residues with significant decrease in hydrophobicity was found for Humsavar data.

Effect of SNVs on structural properties of the protein

Structural stability of generated variants

Prior to investigation of the structural properties of nsSNPs in IndiGen Structural Data, the thermodynamic stability of minimized native and mutant structures was evaluated using FoldX. The influence of genetic variation on protein's stability and flexibility was predicted using Dynamut by calculating Δ ΔG (change in folding energy) value for all the 22 variants. Dynamut implements normal mode analysis for predicting the effect of SNP on native protein structure. The results from Dynamut revealed that 11/22 variants had Δ ΔG negative suggesting destabilization after mutation. The FoldX and Dynamut energy values were visualized in the alluvial plot and shown in Figure 4E. The plot shows 12 genes, their native protein structures (PDB IDs: 4YHJ, 5TQY, 3NYO, 6GQ7, 4TNB, 6BFN, 3GC9, 6BDN, 6I83, 4EYJ, 3NRU and 3D2R) and 22 mutants linked with their corresponding energy values. The gene names and the PDB codes for native protein structures were shown in the first two columns followed by ΔG (in kcal/mol) for all natives given by FoldX and Δ ΔG (in kcal/mol) given by Dynamut for all the variants. The PDB names in the plot were arranged on the basis of the decreasing number of mutations reported for them. As per Dynamut predictions, a mutant F454A of PDB code 4YHJ has shown Δ ΔG of -2.767 kcal/mol (Destabilizing) and change in Vibrational Entropy Energy between Wild-Type and mutant (Δ ΔS-Vib) as 1.178 kcal.mol-1. K-1 showing an increase of the molecular flexibility after mutation.

Secondary Structure Annotation and Relative Solvent Accessibility of mutated residues

The secondary structure of a protein includes largely α-helix and β-pleated sheet structures, which is involved in local interactions between stretches of a polypeptide chain. The ability of a protein to interact with other molecules depends on amino acid residues located on the surface with high solvent accessibility. Any alterations in these residues may affect the protein's

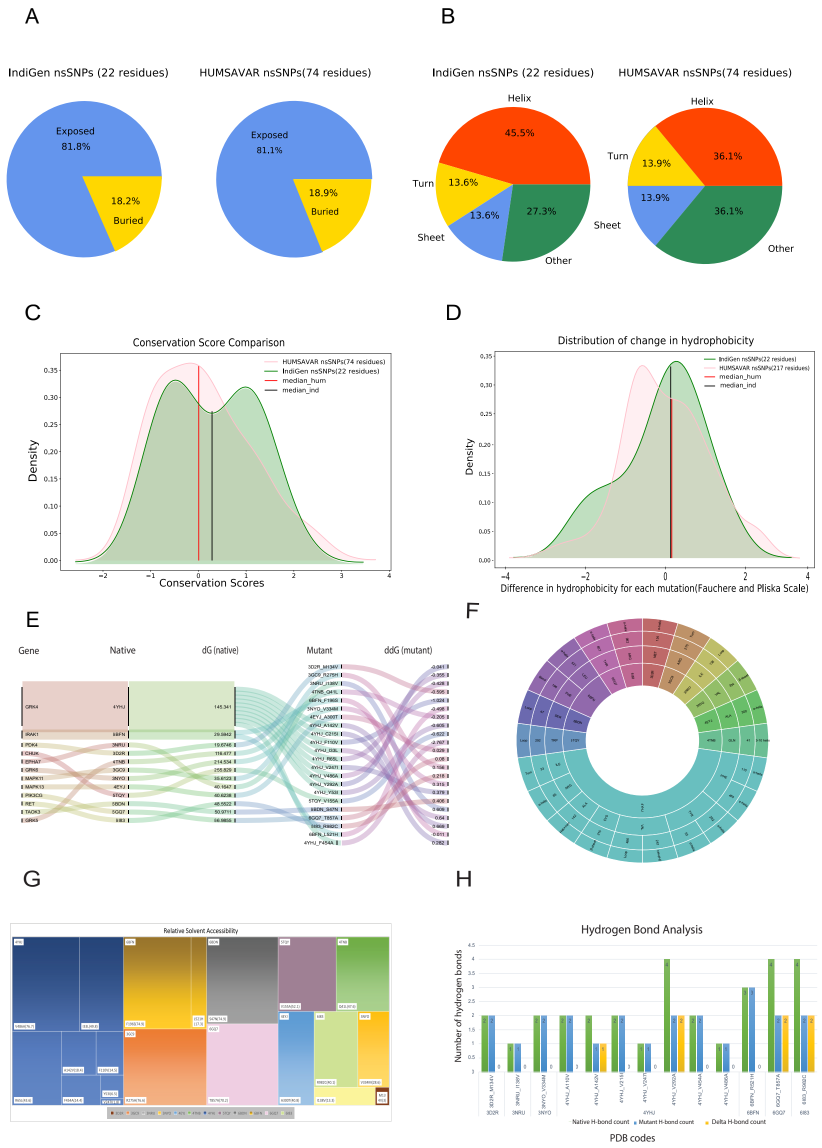

Fig 4. Comparison of structural characteristics of variants in IndiGen and Humsavar data: A. Solvent accessibility for the variants in both datasets. B. Secondary structure in which each of the variants occurs in both datasets. C. Conservation score and ΔHydrophobicity distribution of variants in Humsavar and IndiGen data. D. The area under the curve present on the left of -2 (ΔHydrophobicity) belongs to the percentage of residues for which a significant increase in hydrophobicity after mutation was observed while the exact opposite was observed for percentage of residue present on the right of +2 on x-axis. E. Alluvial plot representing FoldX Energy plot for 12 native PDBs (δ G Native column) and change in folding energy for 22 variants (δ δ G) by Dynamut (in kcal/mol). F. Sunburn Plot representing secondary structure assignment done by DSSP for mutant residues. G. Treemap showing relative solvent accessibility calculated by Naccess for mutated residues. H. HBPLUS results showing the number of hydrogen bonds made by mutated residue before mutation (green -bar), after mutation (blue-bar), and ΔH-bonds (yellow bars)

functioning thereby increasing the importance behind the study of structural properties of mutated residues. Solvent accessibility (using Naccess) and the secondary structure properties (using DSSP) of mutated residues were studied. The Figure 4F is a sunburn plot showing results for secondary structure assignment by DSSP. The plot consists of four concentric circles with innermost circle comprising 12 PDB IDs, second-inner circle comprising 3-letter code of reference amino acid present at mutant site, third-inner circle shows the mutant position and outermost circle contains the secondary annotation for that residue given by DSSP. The color coding was done on the basis of native PDBs. Majority of the variants were found to be present in alpha-helix region as compared to other regions of the protein.

In the Figure 4G, the results obtained from Naccess for relative solvent accessibility of mutated residue was represented by a Treemap. The area of rectangles represents the relative solvent accessibility scale associated with mutated residue. All 12 PDB IDs are shown with 12 different colors forming a hierarchy. The color-coding was done on the basis of associated PDB IDs. The relative solvent accessibility of two mutated residues belonging to PDB code 4YHJ (Y53I and C215I) was zero hence not shown in the figure. The area of rectangle for R275H and V486A mutants of 3GC9 and 4YHJ pdbs were largest with rel. solvent accessibility more than 75 suggesting that these two reference amino acids, arginine of 3GC9 at 275th position and valine at 486th position were relatively more accessible than others. The results from this plot disclosed that there were 5 residues with more than 60 relative solvent accessibility (Arginine, Valine, Phenylalanine and Serine) belonging to 3GC9, 4YHJ, 6BDN and 6BFN PDB IDs.

Effect of SNP in hydrophobicity and hydrogen bonding

A single amino acid change may result in alteration of hydrophobicity or disruption of the hydrogen-bond network thus modifying the structure and function of the protein as well[Kumar and Biswas, 2019]. The change in hydrophobicity observed in mutants in IndiGen structure data were arranged according to Fauchere and Pliska scale [FAUCHÈRE et al., 1988] (Supplemental_Fig_S1-A). In the IndiGen structure data, 12 out of the 22 variants exhibited decrease in hydrophobicity whereas an increase in net hydrophobicity was observed in the rest. The number of hydrogen bonds made by the altered residue before and after the mutation were calculated using the HBPLUS program (Figure 4H). Variants 4YHJ_A142V showed a loss of 1 hydrogen bond, while 4YHJ_V292A, 6GQ7_T857A and 6I83_R982C resulted in loss of two hydrogen bonds.

Effect of SNP on Ligand Binding

Given the pharmacological importance of kinase proteins, molecular docking was performed to comprehend the effect of SNP in the drug-gene interaction. All FDA approved drugs available in DGIdb for genes present in IndiGen structure data were docked against the native and mutant protein structures. In 25 out of 62 protein-drug pairs, changes in binding affinity (0.7 to -9.1 kcal/mol) was observed in native and mutant forms, whereas for remaining pairs, no change in binding affinity was observed. The Figure 5A represents the change in binding affinity observed for the 25 protein-drug pairs. In 20 protein-drug pairs a decrease in binding energy was observed while 5 pairs have shown an increase in binding-energy; indicative that the presence of an SNP destabilizes the complex. One protein-drug pair, T857A mutant of gene PIK3CG (PDB ID: 6GQ7), which when bound to drug Zinc sulfate (DrugBank id - DB09322) revealed a stark decrease in binding energy(-9.1 kcal/mol) when comparing the native- (-13.0 kcal/mol) versus mutant- (-3.9 kcal/mol) drug pair. These 25 protein-drug pairs with difference in binding affinity were further considered for binding site and ligand similarity.

It was observed that the binding pocket of the ligands in native and mutant forms for their respective receptors was the same, stipulating that presence of SNP didn't change the binding site of drugs with their target protein. A snapshot of the first pose of ligand docked in the protein was taken in PyMol for all native protein-drug complexex. The mutated residue in every complex is shown in red-color with sticks representation which was away from the binding pocket of the ligands in all cases(except in the case of 6GQ7-T857A). Ligand binding pockets (post docking) shown in mesh representation with different colors in Supplemental_Fig _S2-(A-G).

In an attempt to find out the reason behind the huge decrease in binding affinity in case of mutant T857A(PDB ID: 6GQ7)-zinc-sulfate(DrugBank id– DB09322) complex the binding site residues of this drug in native and mutant complex were compared and visualized in PyMol [Schrödinger and DeLano] and LigPlot+ [RA and MB, 2011], shown in 5D. It was observed that the location of the binding pocket-residues in mutant and native forms was unchanged and the main binding pocket was away from the mutated residue. However, a decrease in one hydrogen bond was observed in ligand interaction diagram of native and mutant complexes..

Binding Site Similarity Analysis

Fpocket was used to detect the binding pockets present in a protein structure [Le Guilloux et al., 2009]. For every protein in IndiGen structure data, the best binding pose of its ligand was considered as main pocket which aligned to detected pockets by Fpocket. Only the pocket which perfectly aligned were considered for the analysis (12 pockets). Pocket similarity score (distance between a particular pocket pair) for each protein-pocket pair was calculated using DeeplyTough tool shown in (Figure 5B). Similarity is proportional to the score, less negative means more similar.After applying a zscore cutoff (-/+0.70), all the pockets pairs were classified as similar, dissimilar or intermediate, resulting in 12 pairs of similar and dissimilar pockets. The difference in distribution of PS scores of similar and dissimilar pocket pairs is visible from the box plot in Figure 5C. Statistical test (Mann–Whitney U test) revealed statistically significant (p¡0.05) difference between the similar and dissimilar pocket pairs.

Ligand Similarity/diversity Analysis

The 25-protein drug pairs with delta binding energy observed after docking were considered for this analysis. In total there were five different PDB structures (6GQ7, 5TQY, 3GC9,4TNB, 6I83) with five respective mutations and 24 drugs as shown in (Supplemental_Table_S7). All drug-like chemicals from our ligand dataset were considered for chemical similarity analysis. Two drugs- DB09332 (Zinc Sulphate) and DB00040 (Glucagon) were excluded in this analysis as zinc-sulfate contains counter-ion and glucagon is a peptide hormone. From this analysis, it was observed that all the associated drugs exhibit a great molecular diversity (Figure 6). The maximum pairwise similarity for Morgan2 fingerprints and MACCS fingerprints has a Tanimoto score of 0.40 and 0.70, respectively. On the other hand, the pairwise dissimilarity (1-similarity) for Morgan2 fingerprints and MACCS fingerprints has a Tanimoto score of 0.98 and 0.90, respectively. The computational prediction platform ProTox-II, which includes

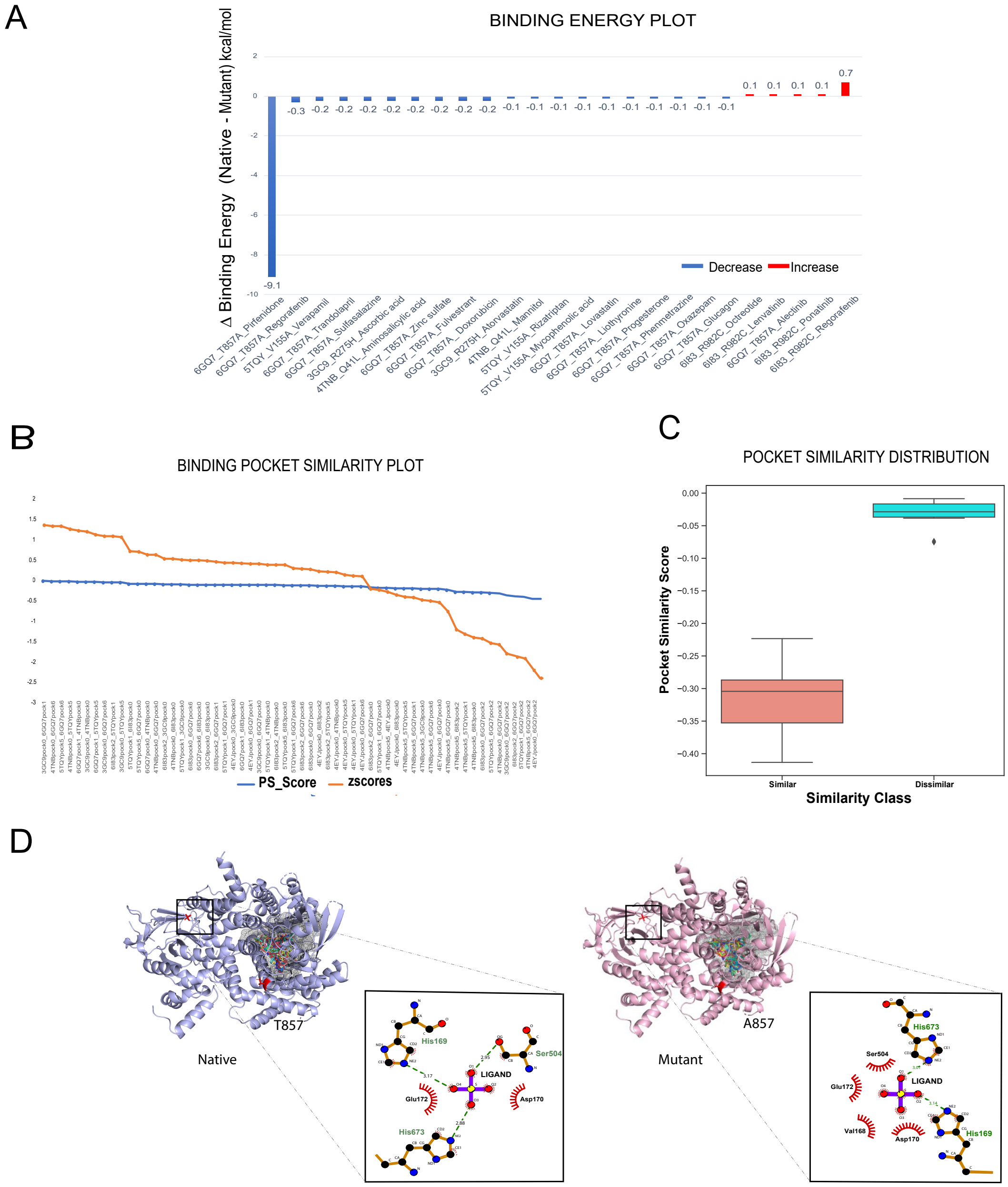

Fig 5. A. Bar plot showing docking results for 25 protein-drug pairs on x-axis and change in binding affinity observed on y-axis. Blue bars represent a decrease in binding affinity and red bars represent increase in binding affinity after mutation. B. Pocket Similarity Curve for proteins in IndiGen structure data. C. Box-plot showing distribution of PS Scores of similar and dissimilar pocket pairs. Using Mann–Whitney U test, p-value(0.000018) was calculated and used to interpret the result of the test. D. Ligand interaction diagram of native 6GQ7(PIK3CG gene) and its mutant T857A bound to Zinc Sulfate (DB09322) and main binding pocket (grey pocket) where majority of ligands docked.

cheminformatics-based machine learning models for predicting 46 toxicity endpoints, was used to predict toxicity profiles of compounds/drugs. For the prediction of various toxicity endpoints, such as acute toxicity (LD50 values), hepatotoxicity, cytotoxicity, carcinogenicity, mutagenicity, immunotoxicity, adverse outcomes pathways (Tox21), and toxicity targets, ProTox-II integrates many statistical methodologies such as molecular similarity, pharmacophores, and fragment propensities, as well as machine learning models (off-targets). In vitro assays (e.g. Tox21 assays, Ames bacterial mutation assays, hepG2 cytotoxicity assays, Immunotoxicity assays) and in vivo cases were used to create the predictive models (e.g. carcinogenicity, hepatotoxicity). These models have been validated on separate external datasets and have shown to be effective and well-cited.[Banerjee et al., 2018].

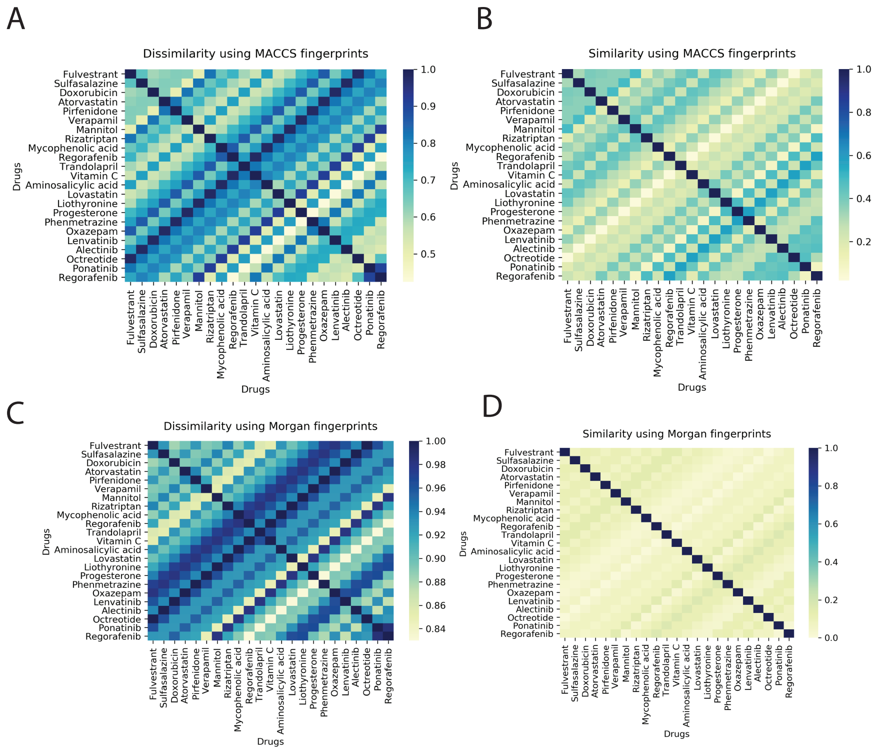

Fig 6. Heat maps representing drug pairwise similarity and dissimilarity. A. Ligand dissimilarity using MACCS fingerprints. B. Ligand similarity using MACCS fingerprints C. Ligand dissimilarity using Morgan fingerprints. D. Ligand similarity using Morgan fingerprints. Similarity and dissimilarity (1-similarity) score is represented using Tanimoto coefficient (taking a value between 0 and 1, with 1 corresponding to maximum similarity)

As per the predictions made by ProToxII (Supplemental_Table_S10), it is observed that the mycophenolic acid (DB01024) which is an immunosuppresant drug , interacting with PDB structure 6GQ7, mutations T857A, is hepatotoxic, immunotoxic and cytotoxic. It also inhibits SR-MMP(mitochondrial membrane potential) with a confidence score of 0.79. Another interesting observation is the drug Regorafenib (DB08896) which is also predicted to be hepatotoxic, and is active in two different stress response pathways SR-MMP, and SR-p53. Regorafenib is associated with adverse events like hypertension, stomatitis, abnormal liver function.[Krishnamoorthy et al., 2015]. However, the exact mechanism of developing hypertension is not very well-defined. Abnormalities in liver function is also reported in case of Regorafenib[De Wit et al., 2014]. The drug progesterone (DB00396) is predicted to be active in six adverse outcome pathways(AOPs). Like progesterone, many other drugs can result in such molecular inhibition/ activation of NR-AR by progesterone, and can result in reduced AR signalling /impaired follicle recruitment as cellular or tissue level response and may be impaired fertility in organism[Pivonello et al., 2020]

Phenotypic drug-drug similarity

In order to look for phenotypically similar drugs in IndiGen data a list of protein IDs and drug molecules associated with them was considered (Supplemental_Table_S7). This information could be useful to get insights about similar drugs present in IndiGen structure data. A correlogram was plotted with drug names on x/y axis. The positive and negative correlation was shown by blue and red color circles. The color intensity and circle size depends on correlation coefficient. (Supplemental_Fig_S1-B). A strong correlation (more blues dots in Figure S1-B) can be observed from this plot indicating promiscuous nature of drugs (binding to multiple targets) or target proteins. For instance drugs Fulvestrant and Rizatriptan are chemically dissimilar (similarity score 0.20 in Figure 6). However, in terms of phenotypic drug-drug similarity - they are highly similar as they bind to the same protein target highlighting the differential binding ability of kinases to a set of fairly specific inhibitors.

Protein-Protein Interaction Network and Biological Processes involved

Since an evident decrease in binding energy was observed in case of T857A mutant of gene PIK3CG with drug Zinc sulfate (DrugBank id - DB09322), this difference in binding affinity can affect the structure and functioning of this protein and others associated with it, thereby understanding its significance and functions linked with is important. PIK3CG gene phosphorylates phosphatidylinositol 4,5-bisphosphate and generates phosphatidylinositol 3,4,5-trisphosphate (PIP3) which is responsible for the recruitment of PH domain-containing proteins to membrane, therefore activating signaling cascades involved in cell growth, survival, proliferation, motility and morphology[M. Christopher, 2016b]. PI3Ks play a pivotal role in human cancers leading to the discovery of small inhibitors of these lipid kinases.[Wang et al., 2015]. The physical and functional association of the protein PDB-6GQ7 were studied by giving the gene name (PIK3CG) as input to the STRING database.[Szklarczyk et al., 2019] The gene PIK3CG was found to have 10 predicted functional partners, i.e., HRAS, KRAS, NRAS, PIK3R6, PIK3R2 PIK3R5, PIK3R1, PIK3R3, AKT1 and PDPK1 shown in Supplemental_Fig_S3A. The information related to the biological processes in which these genes are involved was obtained from Gonet webserver [Pomaznoy et al., 2018], as shown in Supplemental_Fig_S3B.

Discussion

Adverse drug reactions are often associated with genes that are more prone to variations and targeted by multiple drugs. Firstly, to have a global understanding of the distribution of the common variations present in India, the kinome tree for all the druggable kinase genes was constructed (Figure 1). This revealed that tyrosine kinase class consisted large number of variations and was found to be associated with numerous drugs. Receptors tyrosine kinases (RTKs) are involved in broad range of functions such as proliferation, differentiation and apoptosis of cells and have been extensively used as drug target in cancer studies. Many of the tyrosine kinase inhibitors are antibody-based drugs used in treatment of tumors, malignancies and inflammatory diseases[Bennasroune et al., 2004]. The sequence based analysis(Figure 3B) of IndiGen variants disclosed that Indian population is genetically very different from the other populations. Conservative mutations can affect the protein's stability which can modulate its functioning and catalytic pattern followed by it in different organisms[Rodriguez-Larrea et al., 2010]. Studies have shown there is a strong correlation between frequency of occurrence of amino acids in the human genome and number of associated codons[Alwi, 2005]. On the contrary, observation made in amino-acid exchange matrix and chemical shift analysis(Figure 2) suggested that mutation from one amino-acid type to other was independent of number of codons coding for any amino-acid. The changes in chemical classes for majority of amino-acids were found to be conserved indicating more intra-class mutations than inter-class mutations. The mutability plot (Figure 2D) revealed that Arginine (R) is more mutable than other amino acids and the probable reason behind this could be the presence of CpG dinucleotide in the codons coding for Arginine which is relatively vulnerable to mutations[M. Christopher, 2016a].

Ancestry has a very important role to play in evolution of a SNP in different ethnic groups of a population. This also indicates that there is a relationship between allele frequency and ethnicity of the population. Even a fractional exchange of amino acids can have a completely different impact on different populations. Amino acid frequency comparison study stipulated that the variant frequency pattern followed a similar trend in all the populations except Indigen(Figure 3B). Some variants were found to be common in Indian population and rare in other populations(population-specific variants) indicating that it will be affecting Indian population with higher frequency than others(Figure 3D). On comparing allele frequency of Indian mutations with the ones present in publicly available databases it was inferred that many conserved mutations in IndiGen data are still understudied as none of the existing databases contains these mutations (referring to IndiGen data=samples from 1000 individuals of strict Indian ethnicity)(Figure 3C). Protein domain regions are stable conserved parts of a protein sequence and its 3D structure. Therefore, variants present inside the protein domains are more likely to affect the protein structure, stability and function. The comparative study of variants on the basis of their position with respect to domain location suggested that many Indian variants were present either within the domain or in the post-domain region.(Figure 3E)

One of the most useful predictors of the phenotypic effects of missense mutations is protein structural information and stability. Missense mutations can disrupt protein structure and function in one of two ways: they can destabilise the entire protein fold or they can change functional residues, such as active sites or protein-protein interactions, and pathogenic mutations are enriched in both the buried cores of proteins and in protein interfaces[Gerasimavicius et al., 2020]. Reports have claimed that buried amino acids are often observed to be associated with diseases and commonly observed in functional sites. [M. Christopher, 2016a]. On the contrary in relative structural analysis of IndiGen and Humsavar dataset it was found that residues with relatively higher solvent accessible surface were more prone to mutations.(Figure 4A) [M. Christopher, 2016a]

Mutations that occur in properly structured part of a protein are more likely to be pathogenic than mutations that do not, due to their strong destabilizing effect on protein structure. According to stability analysis performed by Dynamut, 11 variants were found to destabilize protein's structure and from 11 destabilizing variants, 7 were found to be present in the helix region of the protein. IndiGen variants occur more in the alpha-helix region while Humsavar variants share equal secondary structure preference for their occurrence either in alpha-helix or in loop/random coil of a protein.(Figure 4B) Several studies have suggested that secondary structure elements like sheets and helices vary a lot in their ability to tolerate mutations. This differential tolerance of mutations could be due to difference in number of non-covalent residue interactions within these secondary structure units.[Abrusán and Marsh, 2016].The conservation score distribution implied a higher percentage of residues with greater conservation in that Humsavar data than in IndiGen data. Since Humsavar variants are reported to be associated with a disease it his highly likely that their presence in highly conserved region could be a reason behind their disease occurrence. Hydrophobic interactions and hydrogen bonds are the two most prevalent interactions present in protein structure. Hydrophobes as the name suggests tend to isolate themselves from water molecules due to which many hydrophobic amino acids are often found to be buried inside the protein structure. Contrasting results were observed in hydrophobicity distribution with significant increase in hydrophobicity for IndiGen structure data whereas the decrease in hydrophobicity was found for Humsavar data. (Figure 4C)

Occurence of SNPs at the ligand binding sites (LBSs) can influence protein's structure, stability and binding affinity with small molecules. Interesting findings claimed that ligand binding residues have a significantly higher mutation rate than other parts of the protein [Kim et al., 2017]. In order to validate whether a single amino acid substitution can change the binding affinity of a ligand with its target protein or not, molecular docking of ligands(FDA approved drugs) with native and mutant structure was performed. The docking results suggested that since the mutated residue was away from the binding pocket not much difference in binding affinity was observed in native and mutant forms except in T857A mutant in which a polar amino acid has converted to a non-polar amino acid leading to loss of two hydrogen bonds (4H), thereby decreasing the binding affinity of ligand(Zinc-sulphate) with protein. Binding site similarity analysis on the basis of PS score and Z score cut-off revealed that many drugs in our dataset share a similar binding site(Figure 6B). These drugs are more similar based on substructure features (local similarity) using MACCS fingerprints.(Figure 6). Moreover, the molecular diversity of 12 drugs binding to 6GQ7 (PIK3GA) suggest the promiscuous nature of the kinase and enabling insights which are relevant for understanding polypharmacology and negative side-effects. Further analysis of these and other inhibitors that bind to PIK3GA, clustered by phenotype information, can give us deeper insights into targeted kinase inhibitor design. The PPI and Gene Ontology analysis revealed that PIK3CG gene is functionally associated with ten other genes and most of them are involved in signal transduction, response to stress, anatomical structure development, immune system process, cellular protein modification process and biosynthetic process.(Supplemental_Fig_S3). PIK3CG gene is altered (Mutation, Amplification, Loss) in 2.68% of all cancers. It is found to be associated with lung, colon, and endometrial adenocarcinoma, cutaneous melanoma, prostate cancer, and breast invasive ductal carcinoma.

While in this study, we have explored common variants present in the Indian population, sampling lower allele frequencies shall be also useful, in the future, to understand the underlying fundamentals of rare diseases. Additionally, experimental validation of the findings in this study shall provide further credence to the results.This study on IndiGen variant data may assist in redesigning the healthcare system from “One Size Fits for All” to “Population or Individual Specific Drug System” and a big step towards the effective treatment of patients due utilisation of drugs with less side-effects.

Materials and Methods

Variant Data collection

The combined variant data of Indian population was curated from over 1029 whole genome sequences collected as part of the IndiGen programme to represent diverse Indo-ethnicities. The variant data comprised of single nucleotide variants and indels which were annotated using Annovar[Wang et al., 2010]. Only SNVs were considered for our study.

Assembling druggable genes

The Drug Gene Interaction Database (DGIdb) version 3 is a database that contains information on all currently approved drugs as well as other future targets of interest.[Freshour et al., 2021]. Genes were annotated in this database with respect to known drug-gene interactions and potential druggability. It normalizes its content from 30 open-source databases like DrugBank [Wishart et al., 2008], therapeutic target database (TTD)[Chen et al., 2002], PharmGKB [Boom et al., 2013], The Druggable genome and other web resources like Oncology Knowledge Base (OncoKB) [With et al., 2017], cancer genome interpreter (CGI) [Tamborero et al., 2018], etc. A list of 545 druggable kinases and associated FDA approved drugs was retrieved from the DGIdb using browse category search while limiting the categories to specific resources i.e ‘GuideToPharmacologyGenes’(Supplemental_Table_S1). The Guide to Pharmacology is a curated repository of ligand-activity-target relationships, with the most of its information derived from high-quality pharmacological and medicinal literature. This druggable kinase gene list was further enriched by adding features like Ensembl ID, PDB ID, RefSeq Match Transcript, gene start - gene end, Uniprot ID, sequence length and structure length etc. using BioMart resource [Smedley et al., 2009] and is automated using python.

Data Preparation

Sequence Data Preparation

Dataset used for sequence analysis contained 545 druggable kinase genes and its associated variants. Protein sequences for these genes were downloaded from NCBI Genbank and mutant sequences were prepared by adding the variants to the native sequence as per the Annovar data.

Structure Data Preparation

Structure Data was prepared by collecting all druggable kinase genes for which a crystallised protein structure (maximum crystal length) was available in UniProt [Bateman, 2019]. The variants from IndiGen data with an allele frequency $\geq$ 10%, falling within the crystal length were accounted for in this analysis. After applying these filters, 12 genes and their corresponding 22 variants were left, and were referred to as IndiGen Structure data (Supplemental_Table_S4). In an attempt to conduct a comparative structural analysis, Humsavar (Human polymorphisms and disease mutations) data was taken. It lists all missense variants annotated in human UniProtKB/Swiss-Prot entries (Release: 2020_04 of 12-Aug-2020). In this data the variants were classified as disease causing (31132- 64.1%), Polymorphisms (39464-23%) and Unclassified (8381- 12.9%). The variants associated to the genes present in IndiGen Structure data were extracted from Humsavar complete list of variants. This dataset was referred to as Humsavar dataset which consisted of total 217 variants, and used for benchmarking structural analysis (Supplemental_Table_S5).

Data Processing and Visualization

Drug, Gene and Variant Tree

The primary goal of this analysis was to have a quantitative and qualitative insight about frequency of occurrence of variation in family of kinases and availability of drugs against it. This will aid in gathering information related to the family of kinases with more variations and drugs reported. An online tool, KinMap [Eid et al., 2017], was used for an interactive exploration of kinase coding genes present in IndiGen data. The genes associated with 545 druggable kinases, number of variations and drugs reported against each gene in DGIdb was given as an input to this tool.

Amino-acid Conversions and Mutabilities

The tendency of conversion of an amino acid type to another type and identification of any pattern in this conversion can guide in understanding the change in physicochemical property of a protein sequence. This analysis was conducted using a python script and the reported variants for kinases were taken into account. The script generated a 20X20 matrix which gave a normalized count of each amino acid with respect to other amino acids i.e percent conversion of each amino acid. Normalized count = (Amino acid count in samples)/(Amino acid count from refseq)*100. This amino-acid exchange matrix was correlated with chemical properties of mutating amino acids by analysing the chemical shifts associated with variants among different populations and databases. The overall amino acid count for each class of amino-acids was summed up for reference and altered residues and the difference in the counts was called as chemical shift. The mutability of an amino-acid is defined as the ratio of total number of mutations for a specific amino acid in the data and the frequency of occurrence for that amino acid in the reference human genome.This mutational frequency was calculated for all the variants in IndiGen(AF >10%).

Multiple Sequence Alignment and Protein Domain Analysis

To understand the effect of SNPs on protein's function it was checked whether the observed variation (SNPs) is conserved and falls under a protein domain or not. Clustal Omega [Sievers and Higgins, 2014] was implemented to perform the multi-sequence alignment (MSA). The protein sequence files in FASTA format were generated using a python script. For protein domain analysis, Pfam Scan (Embl-ebi n.d.) web server maintained by EMBL-EBI was used. A single file of all protein sequences in FASTA format was provided to it as input (default parameters). It gave an output file consisting of domain name, its start and end position corresponding to every input sequence (hmm_name, hmm_start, hmm_end) and other information. Mutations which were observed within domain region (hmm_start - hmm_end) annotated as 0 for others the distance of mutation from domain region was also calculated.

Variant Protein Structure Generation

Computational protein structure prediction helps in generating a three-dimensional structure of proteins. The prediction here is based on in-silico techniques and relies on principles from known protein structures mostly obtained by X-Ray crystallography, NMR Spectroscopy and physical energy function. Before proceeding to the structure analysis few filters were added to the base data. These filters were, 1. Availability of protein crystal structure, 2. Availability of drug molecules against the protein, 3. Crystal structure and sequence coverage $\geq 70\%$, 4. Allele frequency of the nsSNP observed in the IndiGen population $\geq 10\%$, 5. SNP coverage to the crystal structure. In view of the fact that the native crystal structure was already available in Protein Data Bank, we only require to mutate a single amino acid position by taking the reference and altered amino acids present in IndiGen structure data for a particular gene/protein. This single reference amino acid of the protein was mutated using rotkit function of PyMol that allows access to its mutagenesis feature. The crystal structure of the protein based on the requirements mentioned above were downloaded from RCSB PDB and mutated using the rotkit function. This process was automated by python code. It was followed with energy minimization and refinement of these mutant structures (22 variants) using Chimera [Pettersen et al., 2004]. The parameters used for minimization of energy include 1000 steepest descent steps with step size of 0.02 Ang and force-field AMBER ff14SB. For the assessment of structural stability of the native and mutant protein structures, FoldX [Schymkowitz et al., 2005] was implemented. FoldX calculates energy differences that come close to experimental values. The impact of mutations on protein conformation, flexibility and stability was predicted by Dynamut[Rodrigues et al., 2018]. The structural differences in native and mutant forms were analyzed using several tools like DSSP (28) for secondary structure annotation of mutated residue, HBPLUS(29) to study gain or loss of hydrogen bonds after the mutation and Naccess (27) to compare the solvent accessible surface area of the mutated residue.

Molecular Docking

Receptor-ligand docking was performed in order to study the drug-gene interaction and analyze the effect of SNP in binding affinity of drug with its target protein before and after the occurrence of mutation. A set of kinase genes with FDA approved drugs available in DGIdb were taken into account. Only 7/12 genes (CHUK, EPHA7, GRK5, MAPK11, MAPK13, PI4K2B, PIK3CG) from IndiGen structure data were found to exhibit drug-gene interactions given drugs were FDA approved. The protein structure files (in Protein Data Bank as a PDB format) for these 7 genes and their 7 modelled variants were considered as receptors. Since our dataset comprised 62 ligands that were to be docked with 14 receptors, a virtual screening was performed using AutoDock vina[Trott and Olson, 2009]. The drugs/ligands were downloaded from DrugBank and PubChem [Kim et al., 2019] in PDB format. The preparation of receptors (removal of water, missing hydrogens,etc.) and ligand was followed with their conversion to PDBQT format. In the absence of any prior information about the target binding site, blind docking was performed for all the protein-ligand pairs. The docking was performed to the center of the binding cavity using Cartesian coordinates that differed for every protein calculated using PyRx[Dallakyan, Sargis; Olson, 2015]. The docking grid with a dimension of 60 Å x 60 Å x 60 Å was used in each docking calculation with an exhaustiveness option of 100 (average accuracy). The maximum number of binding modes to generate was kept 500 with an energy range of 20kcal/mol. 50 iterations of these parameters for every target protein was followed.

Binding site comparison

Binding site similarity comparison was computed based on the fact that the binding sites on proteins are more conserved than the rest of the protein structure. Detecting ligand-binding sites similarities in globally unrelated proteins can help in the repurposing of new drugs, predicting side-effects, severe toxicity, and drug-target interactions. There exists a basic principle that similar pockets or cavities in a protein structure recognize similar type of ligands, so as to validate this principle, several protein-ligand binding site comparison methods are available which are utilized in many drug discovery scenarios, one such tool is DeeplyTough [Simonovsky and Meyers, 2020]. Since the proteins used in this work belong to the kinase family, it is highly likely that they share similar binding pockets. Fpocket [Le Guilloux et al., 2009] was used to locate all the binding pockets present in the protein. For every target protein, only those pockets which aligned to the best binding pose of docked ligand were given as input to DeeplyTough for assessment of similar binding sites. This tool gave Pocket similarity score as an output for each input protein-pocket pair. Since the difference in PS score among the input pairs was very small, Z-score (orange line) was calculated for every PS-score in order to claim similar pocket pairs with some statistical significance. In order to classify the pockets pairs as similar, dissimilar or intermediate, a Z-score cut-off was considered (-/+0.70).

Ligand similarity/diversity analysis

Molecular similarity of the ligands (drugs) can be assessed using their structural features (e.g., shared substructures, ring systems, functional groups, topologies, etc.) of the compounds and their representations in the N-dimensional chemical space. These descriptors are often defined by mathematical functions of molecular structures. In this analysis, MACCS (Molecular ACCess System) keys with 166 keys and circular -Morgan fingerprints with radius 2 were used[Fernández-De Gortari et al., 2017]. These fingerprint-based similarity computations were implemented using the popular chemoinformatics package RDkit [Bento et al., 2020] in python. Tanimoto similarity coefficient was used to compute a quantitative score in order to measure the degree of ligand similarity and dissimilarity (1-similarity)- using weighted values of molecular descriptors.

Phenotypic drug-drug similarity

The tendency of a drug to bind to multiple targets is called drug polypharmacology, it is well known property of drugs. Reports have suggested about the association of drug polyphamacology with the target protein family and binding site similarity of their primary targets [Jalencas and Mestres, 2013]. If two drug molecules target same gene product then they are expected to have similar activities and mechanism of action[Prinz et al., 2016]. Thus, repurposed form of similar drugs can act as alternative to the ones with adverse drug reactions. On the basis of drug-gene interaction data obtained from DGIdb several drugs were observed to have same target protein.

Author contribution

G.P., N.M., A.R. conceptualized the study, performed analysis and wrote the manuscript. D.S., R.C.B., A.J., M.I., V.S., M.K.D., A.M., S.S., and V.S. generated the IndiGen data and assisted in the inputs in the manuscripts. P.G. performed the domain analysis. G.P., N.M., and P.B. performed the pharmacogenomic analysis and P.B., S.S. and V.S. gave critical insights during the manuscript writing.

References

G. Abrusán and J. A. Marsh. Alpha Helices Are More Robust to Mutations than Beta Strands. *PLoS Computational Biology*, 12(12):1–16, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1005242.

M. A.J. Marian. Nihms346353.Pdf. 159(2):64–79, 2013. doi: 10.1016/j.trsl.2011.08.001. Molecular.

Z. B. Alwi. The Use of SNPs in Pharmacogenomics Studies. *The Malaysian journal of medical sciences : MJMS*, 12(2):4–12, 2005. ISSN 1394-195X. URL `http://www.ncbi.nlm.nih.gov/pubmed/22605952{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3349395`.

H. Ashkenazy, S. Abadi, E. Martz, O. Chay, I. Mayrose, T. Pupko, and N. Ben-Tal. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1):W344–W350, 2016. ISSN 13624962. doi: 10.1093/nar/gkw408.

M. Bamshad, T. Kivisild, W. S. Watkins, M. E. Dixon, C. E. Ricker, B. B. Rao, J. M. Naidu, B. V. Prasad, P. G. Reddy, A. Rasanayagam, S. S. Papiha, R. Villems, A. J. Redd, M. F. Hammer, S. V. Nguyen, M. L. Carroll, M. A. Batzer, and L. B. Jorde. Genetic evidence on the origins of Indian caste populations. *Genome Research*, 11(6):994–1004, 2001. ISSN 10889051. doi: 10.1101/gr.GR-1733RR.

P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner. ProTox-II: A webserver for the prediction of toxicity of chemicals. *Nucleic Acids Research*, 46(W1):W257–W263, 2018. ISSN 13624962. doi: 10.1093/nar/gky318.

A. Bateman. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1): D506–D515, 2019. ISSN 13624962. doi: 10.1093/nar/gky1049.

A. Bennasroune, A. Gardin, D. Aunis, G. Crémel, and P. Hubert. Tyrosine kinase receptors as attractive targets of cancer therapy. *Critical Reviews in Oncology/Hematology*, 50(1):23–38, 2004. ISSN 10408428. doi: 10.1016/j.critrevonc.2003.08.004.

A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij, and A. R. Leach. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12(1):1–16, 2020. ISSN 17582946. doi: 10.1186/s13321-020-00456-1. URL `https://doi.org/10.1186/s13321-020-00456-1`.

N. Berndt, R. M. Karim, and E. Schönbrunn. Advances of small molecule targeting of kinases. *Current Opinion in Chemical Biology*, 39:126–132, 2017. ISSN 18790402. doi: 10.1016/j.cbpa.2017.06.015.

D. Bhosle, A. Sayyed, A. Bhagat, H. Shaikh, A. Sheikh, V. Bhopale, and Z. Quazi. Comparison of Generic and Branded Drugs on Cost Effective and Cost Benefit Analysis. *Annals of International medical and Dental Research*, 3(1):1–6, 2016. ISSN 23952814. doi: 10.21276/aimdr.2017.3.1.pc1.

K. S. Bhullar, N. O. Lagarón, E. M. McGowan, I. Parmar, A. Jha, B. P. Hubbard, and H. P. Rupasinghe. Kinase-targeted cancer therapies: Progress, challenges and future directions. *Molecular Cancer*, 17(1):1–20, 2018. ISSN 14764598. doi: 10.1186/s12943-018-0804-2.

D. V. D. Boom, M. Wjst, and R. E. Everts. PharmGKB: The Pharmacogenomics Knowledge Base Caroline. *Methods in Molecular Biology*, 1015:71–85, 2013. doi: 10.1007/978-1-62703-435-7. URL `http://link.springer.com/10.1007/978-1-62703-435-7`.

S. L. Chan, S. Jin, M. Loh, and L. R. Brunham. Progress in understanding the genomic basis for adverse drug reactions: A comprehensive review and focus on the role of ethnicity. *Pharmacogenomics*, 16(10):1161–1178, 2015. ISSN 17448042. doi: 10.2217/PGS.15.54.

X. Chen, Z. L. Ji, and Y. Z. Chen. TTD: Therapeutic Target Database. *Nucleic Acids Research*, 30(1):412–415, 2002. ISSN 03051048. doi: 10.1093/nar/30.1.412.

T. M. Christensen, Z. Vejlupkova, Y. K. Sharma, K. M. Arthur, J. W. Spatafora, C. A. Albright, R. B. Meeley, J. P. Duvick, R. S. Quatrano, and J. E. Fowler. Conserved subgroups and developmental regulation in the monocot rop gene family. *Plant Physiology*, 133(4):1791–1808, 2003. ISSN 0032-0889. doi: 10.1104/pp.103.029900. URL `http://www.plantphysiol.org/content/133/4/1791`.

G. Clinical and P. Guidelines. Drug development research in resource-limited countries. *Group*, (December 2005):1–101, 2006.

A. Dallakyan, Sargis; Olson. Participation in global governance: Coordinating ”the voices of those most affected by food insecurity”. *Global Food Security Governance*, 1263:1–11, 2015. ISSN 0717-6163. doi: 10.1007/978-1-4939-2269-7.

M. De Wit, C. B. Boers-Doets, A. Saettini, K. Vermeersch, C. R. De Juan, J. Ouwerkerk, S. S. Raynard, A. Bazin, and C. Cremolini. Prevention and management of adverse events related to regorafenib. *Supportive Care in Cancer*, 22(3):837–846, 2014. ISSN 14337339. doi: 10.1007/s00520-013-2085-z.

S. Eid, S. Turk, A. Volkamer, F. Rippmann, and S. Fulle. Kinmap: A web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics*, 18(1):1–6, 2017. ISSN 14712105. doi: 10.1186/s12859-016-1433-7. URL `http://dx.doi.org/10.1186/s12859-016-1433-7`.

J.-L. FAUCHÈRE, M. Charton, L. B. Kier, A. Verloop, and V. Pliska. Amino acid side chain parameters for correlation studies in biology and pharmacology. *International Journal of Peptide and Protein Research*, 32(4):269–278, 1988. ISSN 13993011. doi: 10.1111/j.1399-3011.1988.tb01261.x.

E. Fernández-De Gortari, C. R. García-Jacas, K. Martinez-Mayorga, and J. L. Medina-Franco. Database fingerprint (DFP): an approach to represent molecular databases. *Journal of Cheminformatics*, 9(1):1–9, 2017. ISSN 17582946. doi: 10.1186/s13321-017-0195-1.

S. French and B. Robson. What is a conservative substitution? *Journal of Molecular Evolution*, 19(2):171–175, 1983. ISSN 00222844. doi: 10.1007/BF02300754.

S. L. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. J. Song, M. Griffith, O. Griffith, and A. H. Wagner. Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Research*, 49(D1):D1144–D1151, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1084.

L. Gerasimavicius, X. Liu, and J. A. Marsh. Identification of pathogenic missense mutations using protein stability predictors. *Scientific Reports*, 10(1):1–10, 2020. ISSN 20452322. doi: 10.1038/s41598-020-72404-w. URL `https://doi.org/10.1038/s41598-020-72404-w`.

S. Gong and T. L. Blundell. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PLoS ONE*, 5(2), 2010. ISSN 19326203. doi: 10.1371/journal.pone.0009186.

P. Impicciatore, I. Choonara, A. Clarkson, D. Provasi, C. Pandolfini, and M. Bonati. Incidence of adverse drug reactions in paediatric in/out-patients: A systematic review and meta-analysis of prospective studies. *British Journal of Clinical Pharmacology*, 52(1):77–83, 2001. ISSN 03065251. doi: 10.1046/j.0306-5251.2001.01407.x.

A. Jain, R. C. Bhoyar, K. Pandhare, A. Mishra, D. Sharma, M. Imran, V. Senthivel, M. K. Divakar, M. Rophina, B. Jolly, A. Batra, S. Sharma, S. Siwach, A. G. Jadhao, N. V. Palande, G. N. Jha, N. Ashrafi, P. K. Mishra, A. K. Vidhya, S. Jain, D. Dash, N. S. Kumar, A. Vanlallawma, R. J. Sarma, L. Chhakchhuak, S. Kalyanaraman, R. Mahadevan, S. Kandasamy, B. M. Pabitha, R. E. Rajagopal, R. J. Ezhil, D. P. P. Nirmala, A. Bajaj, V. Gupta, S. Mathew, S. Goswami, M. Mangla, S. Prakash, K. Joshi, Meyakumla, S. Sreedevi, D. Gajjar, R. Soraisham, R. Yadav, Y. S. Devi, A. Gupta, M. Mukerji, S. Ramalingam, B. K. Binukumar, V. Scaria, and S. Sivasubbu. IndiGenomes: A comprehensive resource of genetic variants from over 1000 Indian genomes. *Nucleic Acids Research*, 49(D1):D1225–D1232, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa923.

X. Jalencas and J. Mestres. On the origins of drug polypharmacology. *MedChemComm*, 4(1): 80–87, 2013. ISSN 20402511. doi: 10.1039/c2md20242e.

P. Kim, J. Zhao, P. Lu, and Z. Zhao. MutLBSgeneDB: Mutated ligand binding site gene DataBase. *Nucleic Acids Research*, 45(D1):D256–D263, 2017. ISSN 13624962. doi: 10.1093/nar/gkw905.

S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, 2019. ISSN 13624962. doi: 10.1093/nar/gky1033.

S. K. Krishnamoorthy, V. Relias, S. Sebastian, V. Jayaraman, and M. W. Saif. Management of regorafenib-related toxicities: A review. *Therapeutic Advances in Gastroenterology*, 8(5): 285–297, 2015. ISSN 17562848. doi: 10.1177/1756283X15580743.

A. Kumar and P. Biswas. Effect of site-directed point mutations on protein misfolding: A simulation study. *Proteins: Structure, Function and Bioinformatics*, 87(9):760–773, 2019. ISSN 10970134. doi: 10.1002/prot.25702.

V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:1–11, 2009. ISSN 14712105. doi: 10.1186/1471-2105-10-168.

N. H. Lee. Pharmacogenetics of drug metabolizing enzymes and transporters: effects on pharmacokinetics and pharmacodynamics of anticancer agents. *Anti-cancer agents in medicinal chemistry*, 10(8):583–92, 2010. ISSN 1875-5992. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3770187{&}tool=pmcentrez{&}rendertype=abstract`.

A. M. L. S. M. Christopher. Genetic and epigenetic heterogeneity in acute myeloid leukemias. *Physiology & behavior*, 176(1):100–106, 2016a. doi: 10.1016/j.str.2015.03.028.Insights.

A. M. L. S. M. Christopher. HHS Public Access. *Physiology & behavior*, 176(1):100–106, 2016b. doi: 10.1159/000373949.Phosphatidylinositol-3.

J. Mattei, L. D. Parnell, C. Q. Lai, B. Garcia-Bailo, X. Adiconis, J. Shen, D. Arnett, S. Demissie, K. L. Tucker, and J. M. Ordovas. Disparities in allele frequencies and population differentiation for 101 disease-associated single nucleotide polymorphisms between Puerto Ricans and non-Hispanic whites. *BMC Genetics*, 10:1–12, 2009. ISSN 14712156. doi: 10.1186/1471-2156-10-45.

M. Mori, R. Yamada, K. Kobayashi, R. Kawaida, and K. Yamamoto. Ethnic differences in allele frequency of autoimmune-disease-associated SNPs. *Journal of Human Genetics*, 50(5): 264–266, 2005. ISSN 14345161. doi: 10.1007/s10038-005-0246-8.

N. Nakatsuka, P. Moorjani, N. Rai, B. Sarkar, A. Tandon, N. Patterson, G. S. Bhavani, K. M. Girisha, M. S. Mustak, S. Srinivasan, A. Kaushik, S. A. Vahab, S. M. Jagadeesh, K. Satyamoorthy, L. Singh, D. Reich, and K. Thangaraj. The promise of discovering population-specific disease-associated genes in south asia. *Nature Genetics*, 49(9):1403–1407, sep 2017. ISSN 1061-4036. doi: 10.1038/ng.3917. URL `http://www.nature.com/doifinder/10.1038/ng.3917`.

M. K. Paul and A. K. Mukhopadhyay. Tyrosine kinase – Role and significance in Cancer. *International Journal of Medical Sciences*, 1(283):101–115, 2012. ISSN 1449-1907. doi: 10.7150/ijms.1.101.

E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004. ISSN 01928651. doi: 10.1002/jcc.20084.

C. Pivonello, G. Muscogiuri, A. Nardone, F. Garifalos, D. P. Provvisiero, N. Verde, C. De Angelis, A. Conforti, M. Piscopo, R. S. Auriemma, A. Colao, and R. Pivonello. Bisphenol A: An emerging threat to female fertility. *Reproductive Biology and Endocrinology*, 18(1), 2020. ISSN 14777827. doi: 10.1186/s12958-019-0558-8.

M. Pomaznoy, B. Ha, and B. Peters. GOnet: A tool for interactive Gene Ontology analysis. *BMC Bioinformatics*, 19(1):1–8, 2018. ISSN 14712105. doi: 10.1186/s12859-018-2533-3.

J. Prinz, I. Vogt, G. Adornetto, and M. Campillos. A Novel Drug-Mouse Phenotypic Similarity Method Detects Molecular Determinants of Drug Effects. *PLoS Computational Biology*, 12 (9):1–29, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1005111.

L. RA and S. MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51:2778–2786, 2011.

C. H. Rodrigues, D. E. Pires, and D. B. Ascher. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research*, 46(W1):W350–W355, 2018. ISSN 13624962. doi: 10.1093/nar/gky300.

D. Rodriguez-Larrea, R. Perez-Jimenez, I. Sanchez-Romero, A. Delgado-Delgado, J. M. Fernandez, and J. M. Sanchez-Ruiz. Role of conservative mutations in protein multi-property adaptation. *Biochemical Journal*, 429(2):243–249, 2010. ISSN 02646021. doi: 10.1042/BJ20100386.

D. K. Sanghera, L. Ortega, S. Han, J. Singh, S. K. Ralhan, G. S. Wander, N. K. Mehra, J. J. Mulvihill, R. E. Ferrell, S. K. Nath, and M. I. Kamboh. 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 Medical Genetics*, 9(Ci):1–9, 2008. ISSN 14712350. doi: 10.1186/1471-2350-9-59.

L. Schrödinger and W. DeLano. Pymol. URL `http://www.pymol.org/pymol`.

J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The FoldX web server: An online force field. *Nucleic Acids Research*, 33(SUPPL. 2):382–388, 2005. ISSN 03051048. doi: 10.1093/nar/gki387.

D. Sengupta, A. Choudhury, A. Basu, and M. Ramsay. Population stratification and underrepresentation of Indian subcontinent genetic diversity in the 1000 genomes project dataset. *Genome Biology and Evolution*, 8(11):3460–3470, 2016. ISSN 17596653. doi: 10.1093/gbe/evw244.

S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001. ISSN 03051048. doi: 10.1093/nar/29.1.308.

F. Sievers and D. G. Higgins. Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, 1079:105–116, 2014. ISSN 10643745. doi: 10.1007/978-1-62703-646-7_6.

M. Simonovsky and J. Meyers. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *Journal of chemical information and modeling*, 60(4):2356–2366, 2020. ISSN 1549960X. doi: 10.1021/acs.jcim.9b00554.

G. Sirugo, S. M. Williams, and S. A. Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, mar 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.02.048. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867419302314`.

D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk. BioMart - Biological queries made easy. *BMC Genomics*, 10:1–12, 2009. ISSN 14712164. doi: 10.1186/1471-2164-10-22.

D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Von Mering. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019. ISSN 13624962. doi: 10.1093/nar/gky1131.

D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, C. de Torres, R. Dienstmann, A. Gonzalez-Perez, and N. Lopez-Bigas. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1):25, mar 2018. doi: 10.1186/s13073-018-0531-8. URL `http://dx.doi.org/10.1186/s13073-018-0531-8`.

O. Trott and A. J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):NA–NA, 2009. ISSN 01928651. doi: 10.1002/jcc.21334.

K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164, sep 2010. doi: 10.1093/nar/gkq603. URL `http://dx.doi.org/10.1093/nar/gkq603`.

X. Wang, J. Ding, and L. H. Meng. PI3K isoform-selective inhibitors: Next-generation targeted cancer therapies. *Acta Pharmacologica Sinica*, 36(10):1170–1176, 2015. ISSN 17457254. doi: 10.1038/aps.2015.71.

C. Y. Wei, M. T. Michael lee, and Y. T. Chen. Pharmacogenomics of adverse drug reactions: Implementing personalized medicine. *Human Molecular Genetics*, 21(R1):58–65, 2012. ISSN 09646906. doi: 10.1093/hmg/dds341.

D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(SUPPL. 1):901–906, 2008. ISSN 03051048. doi: 10.1093/nar/gkm958.

P. With, M. Oncokb, U. S. Food, N. Comprehensive, C. Network, R. To, C. Genomics, C. Oncokb, I. The, U. S. Food, and N. Com. OncoKB : A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, 2017.