nature communications

Article

https://doi.org/10.1038/s41467-026-73816-4

Genomic landscape of drug binding and pharmacogenetic variation across diverse populations using SNPdrug3D

Received: 23 December 2024

Accepted: 16 May 2026

Published online: 09 June 2026

Check for updates

Ashar J. Malik[1,2,3,4,66], Dimitar Kenanov[5,66], Cheng-Shoong Chong[5,66], Mert Burak Ozturk[6], Qiqi Wang[7], Hong-Yun Chang[8], Maxime Hebrard[9,10,11], Ying Swan Ho[12], Zhenxun Wang[13], Vachiranee Limviphuvadh[5], Jhoann M. T. Miyajima[5], Srinivasaraghavan Kannan[1], Roberto Tirado Magallanes[9], Xueling Sim[14], Jin Fang Chai[14], Sonia Davila[13,15,16], Khung Keong Yeo[13,15,17], Khai Pang Leong[18], Liuh Ling Goh[18], Neerja Karnani[19,20,21], Johan Gunnar Eriksson[19,22,23,24], Ching-Yu Cheng[25,26,27,28], Tien Yin Wong[25,29], Marie Loh[30,31,32], Rinkoo Dalan[33], Chiea Chuen Khor[34], Guo-Liang Chew[35], Weng Khong Lim[15,36,37], Joanne Ngeow[30,38,39], John C. Chambers[10,30,31], E. Shyong Tai[10,14,40], Nicolas Bertin[9], Chris Soon Heng Tan[7], Judice Koh[6,41], Patrick Tan[10,13,42], SG10K_Health Consortium*, Chandra S. Verma[1,43,44] ✉ & Sebastian Maurer-Stroh[5,44,45] ✉

One of the promises of precision medicine is to understand and act on inter-individual genetic differences in drug responses. SNPdrug3D contains the complete genomic landscape of missense single nucleotide variants (SNV) across the human proteome and at a population-wide level that could affect drug binding. Here, we map SNVs in over 80,000 individuals from the Singapore SG10K Health and gnomAD cohorts to identify ~1.17 million variants mapped to residues near ~6000 bound drugs in protein-drug complexes and experimentally verify effects of selected SNVs, including previously uncharacterized variants, on drug binding in relevant proteins ranging from kinases to cytochrome P450s (CYPs). The latter led to a specific predictor for interpreting variants in the CYP family that outperforms existing tools in the prediction of pharmacogenetic effects based on database-annotated (AUROC = 0.9) or assay-based (AUROC = 0.8) test sets. By placing variants and drugs in structural contexts, SNPdrug3D aids drug development by pre-emptively flagging potential resistance sites based on population-specific variability.

The advent of next-generation sequencing technology has generated abundant genetic variation data with potential clinical utility. This includes datasets capturing global variations like gnomAD[1] to a recent genomic database of 10,000 healthy Singaporeans (SG10K Health) that mainly contains variants occurring in the three major ethnic groups of Chinese, Malay and Indians[2,3]. While substantial effort has been devoted to predict[4–8] and understand the clinical relevance of these variants, over 98% of variants have unknown consequences[6]. Specifically, much emphasis has been placed on studying missense SNVs (i.e., missense variants) and their effects in a disease context, but they can also affect drug response by altering drug pharmacokinetic or pharmacodynamic properties. Such pharmacogenetic (PGx) variants

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper.
✉e-mail: chandra@a-star.edu.sg; sebastianms@a-star.edu.sg

occur at least five times in an individual's genome[9], are generally more common[10] than monogenic variants and can affect medication prescribing for a small but growing list of drugs[11]. During drug treatment at the clinic or in clinical trials, they are a cause of adverse drug reactions (ADR) (e.g., 3–5% of all hospital admissions are estimated to be due to ADRs)[12] and therapeutic failure, especially when population genetic diversity is underappreciated[13].

Given their clinical importance and implications for drug development, it is necessary to link genotypes to drug response phenotypes by investigating the molecular mechanisms underlying these phenotypes. One way this may arise is due to variant-induced single amino acid substitutions at or near protein-drug binding sites or through long-distance allosteric effects[14]. For example, some substitutions are known to cause tyrosine kinase inhibitor resistance[15] during anticancer therapy. Similarly, these substitutions can also affect drug metabolism by abrogating the binding of drugs, such as antidepressants to metabolic enzymes, such as CYP2C9 and CYP2C19[16] which are members of the cytochrome P450 (CYP) gene family that is responsible for metabolizing the vast majority of drugs (e.g., 80% of phase I metabolism is mediated by CYP enzymes)[17]. With an estimated 6000–10,000 non-synonymous SNVs[18] per genome, it is expected that some of these missense variants and their associated amino acid substitutions will affect drug binding sites or drug metabolism, but the effects of most remain unknown.

To explore this, there have been past efforts mapping variants to protein structures for both germline and somatic variants[19–26]. However, there has been limited coverage of protein structure space as these works were mostly limited to protein crystal structures and reference proteomes from PDB and UniProt, respectively, and few (e.g., PhyreRisk, mutation3D) were supplemented with protein structure data from prediction models, and none are focused on large-scale drug molecule binding sites. As a result, proteome coverage is limited to UniProt (e.g., 70% coverage by PhyreRisk[20] compared with 72% in SNPdrug3D), and consideration for protein isoforms is lacking (e.g., MISCAST[21] covers 1330 genes as it only includes data from precise reference sequence matches to PDB). With the recent advent of high-quality predictions from AlphaFold, mapping of variants to protein residues is now available for 98.5% of proteins in the UniProt human reference proteome (one sequence per gene)[27]. However, there exists no mapping yet of SNV-drug 3D interaction within these structures. To address these challenges, we introduce SNPdrug3D, an interactive tool that maps and visualizes missense variants in protein structures for 20,442 human genes to assist in the interpretation of variant effects in protein-drug binding sites and apply it to variants in SG10K Health and gnomAD.

Specifically, SNPdrug3D can be used to identify putative functional PGx variants while also providing a predicted molecular mechanism for their effects (e.g., disruption of protein-drug binding at an atomistic level of description of the binding pocket). Here, we demonstrate that its use can improve understanding of functional PGx variants by i) identifying previously uncharacterized amino acid substitutions that affect drug binding to protein targets, which were validated using the cellular thermal shift assay (CETSA)[28] and ii) deriving useful features from protein-drug complexes, such as the proximity of substituted amino acids to ligands for building a machine-learning-based, PGx-inclusive cytochrome P450 (CYP) variant prediction tool (CYPVarPred). Overall, our work provides a comprehensive picture of such variants with an unprecedented annotation depth that not only captures variations in the Singapore population but also global variation via linkage to gnomAD data for a broad coverage of the world population as the ethnic mix in both cohorts are complementary. This provides a complete map of SNV-drug 3D interactions across the human proteome and at a population-wide level.

Results

Data pipeline overview

SNPdrug3D integrates population-level missense variant data from the SG10K Health[2,3] and gnomAD[1] cohorts, 3D protein structure data derived from the Protein Data Bank (PDB)[29], HHPred, Modbase[30] and the AlphaFold Protein Structure Database[31], sequence data from Ensembl, UniProt and NCBI RefSeq and information from DrugBank (academic license)[32].

The workflow for combining different data sources to create SNPdrug3D is shown in Fig. 1. In total, the SNPdrug3D reference proteome comprises of 102,532 unique human proteins (including

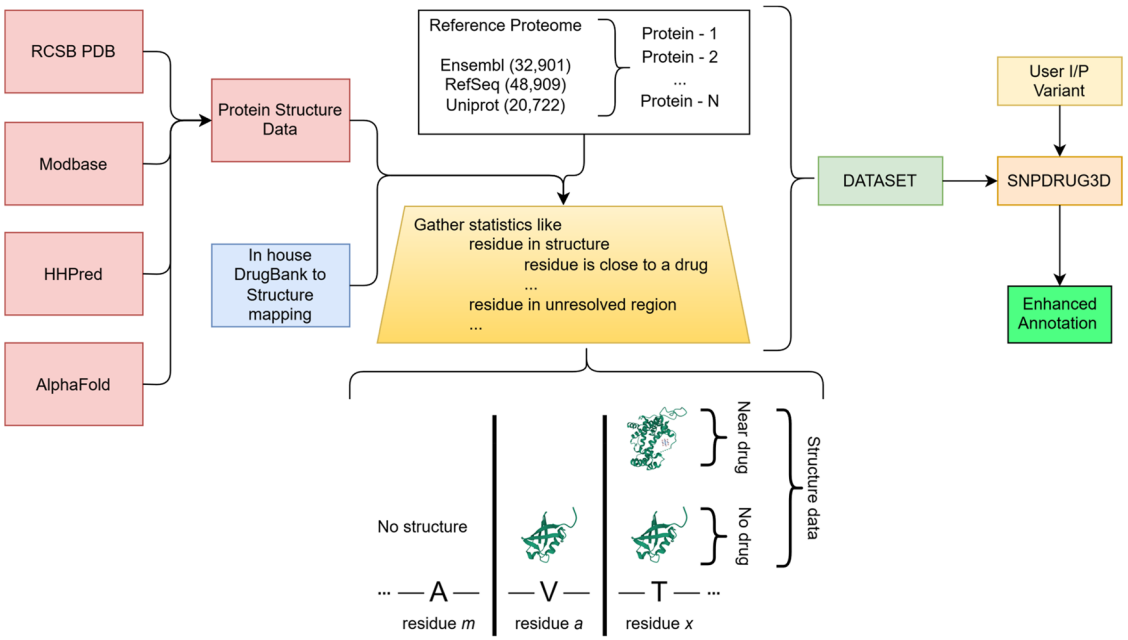

Fig. 1 | SNPdrug3D architecture. SNPdrug3D integrates data from RCSB Protein Data Bank (PDB), Modbase, HHPred, AlphaFold and DrugBank to provide enhanced annotation per missense variant.

https://doi.org/10.1038/s41467-026-73816-4

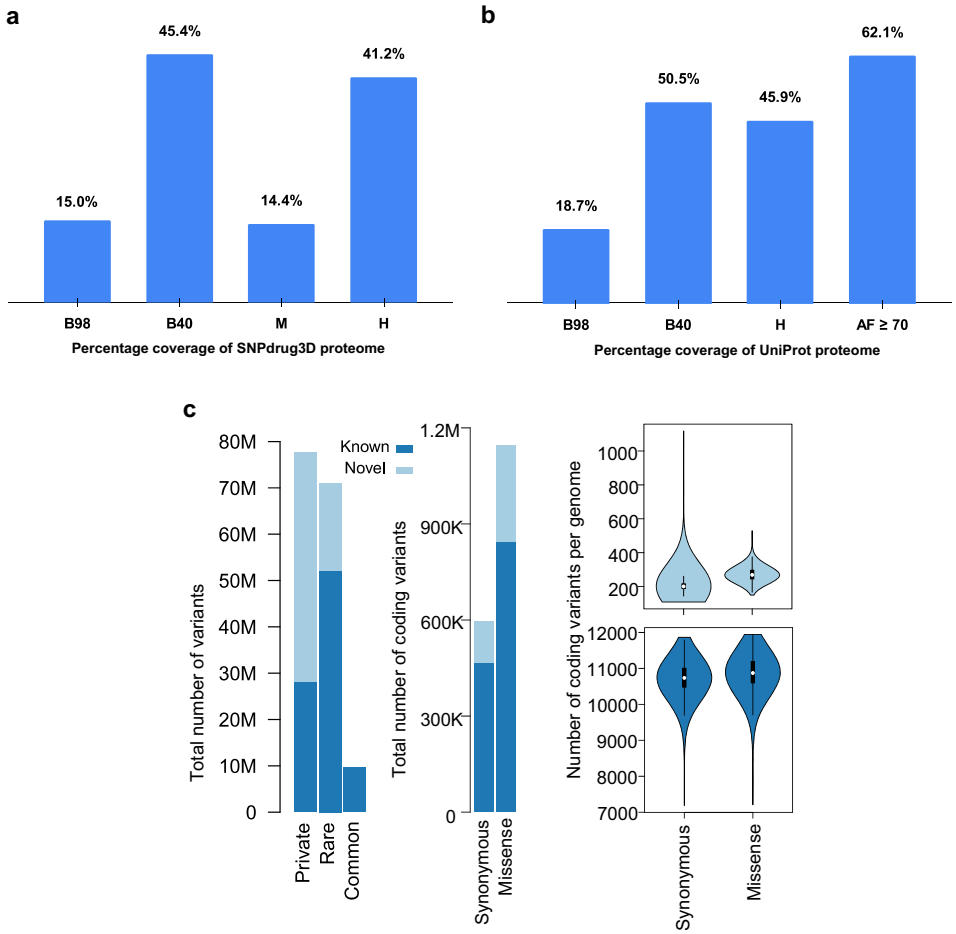

Fig. 2 | Proteome coverage and summary of missense variation data in the SG10K Health cohort. a Coverage of residues in entire SNPdrug3D reference proteome and **b** the Uniprot subset of the reference proteome. Integrating information from all sources increases coverage compared to each method alone. B98, Blast matches from PDB with above 98% similarity; B40, Blast matches from PDB with 40 to 98% similarity; M, Modbase structures; H, HHPred structures; AF, AlphaFold structures (only residues with pLDDT of 70 and above were considered) **c** Total number of known (as of dbSNP v151) and novel (unreported in dbSNP) variants broken down by allele frequency (AF) - Private: found in only one individual, Rare: AF < = 1% and more than one individual, Common: AF > 1%. Total number and individual genome distributions of known and novel protein-coding gene overlapping variants' worse functional consequences: Synonymous, Missense (including stop/start codon loss, inframe insertion/deletion). The distribution of coding variants per genome is represented as a violin plot, the white dot noting the median, the black box illustrate the interquartile rage (IQR), bottom 1st quartile and top 3rd quartile, wicker extend to the adjacent values from the quartiles to +/− 1.5*IQR. Variants' overlap with dbSNP and characterization of variants' protein-coding functional consequence(s) was ascertained using VEP v95 and SG10K Health v5.3.1, which comprised 9770 individuals. Source data are provided as a Source Data file.

isoforms) from 20,442 genes with 65% (~38 million) of amino acid residues in the proteome being mapped to at least one of the 202,229 protein structures from PDB (based on a cut-off of 40% sequence similarity), ModBase, HHPred or AlphaFold (residues with a predicted local distance difference test score (pLDDT) of at least 70). These protein structures may have a ligand or drug as defined by DrugBank and we found that a total of 5962 drug molecules present in this database were associated with a structure out of 28,949 ligands. The remaining 35% (~20 million) of proteome residues mapped to missing residues in crystallized structures and had no homologous residues in non-human proteins (Fig. 1, bottom) or were in low confidence regions in AlphaFold structures (pLDDT <70), possibly associated with intrinsic disorder or are structured only in a complex[27]. Overall, each residue covered will have associated protein structure and sequence data along with information regarding its proximity (i.e., within 8 Å) to one or more drug molecules. Amongst existing efforts, SNPdrug3D has the largest reference proteome size and number of protein structures with the next largest being PhyreRisk, which included 18,874 experimental and 84,818 predicted structures with a coverage of 42,485 UniProt proteins[20]. Furthermore, a key differentiating factor for SNPdrug3D is that it contains binding site information for ~42% (5962) of DrugBank drug molecules (the remainder do not have 3D structures of protein complexes or are not specific in binding) derived from PDB protein structures which is at least 5-fold more than previous efforts, such as mutLBSgeneDB[25] (1324 drugs) and DrugVar[26] (235 drugs).

Mapping variants to structures

Integrating information from various sources allowed SNPdrug3D to reach a full proteome coverage of ~65% (up to 70.5% if low confidence residues in AlphaFold structures are considered), whereas use of protein structure databases individually did not allow for such high coverage (Fig. 2a). To illustrate this, if we considered only structures that had sequences with very close similarity (i.e. above 98%, B98 group) to our reference sequences, a maximum coverage of less than 20% was obtained (Fig. 2a). If only coverage of the UniProt reference proteome (UP000005640_9606) is considered, we show that including AlphaFold predictions alone allowed us to reach a coverage of more than 62% of the residues in the ~20,000 protein sequences found in this proteome (Fig. 2b). If all residues regardless of prediction quality are considered, AlphaFold predictions cover 99% of the UniProt proteome, however, AlphaFold does not include drug binding

https://doi.org/10.1038/s41467-026-73816-4

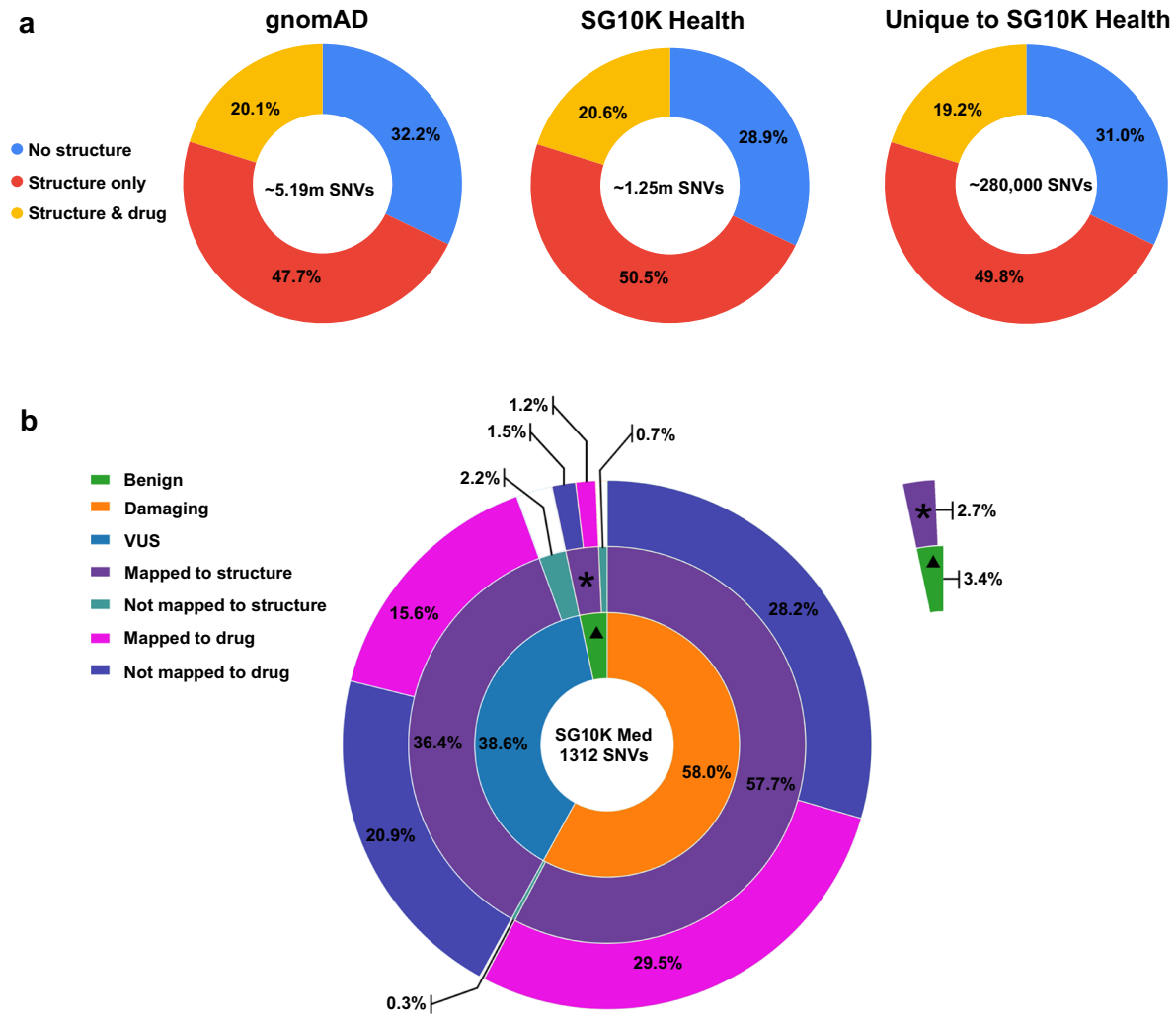

Fig. 3 | Variant annotations. a Mapping of variants from gnomAD, SG10K Health and variants unique to SG10K Health (i.e., not in gnomAD or dbSNP v151) in SNPdrug3D. **b** Subset of curated variants from SG10K Health (SG10K Med) grouped according to their clinically curated predicted variant effects (i.e., damaging, benign and VUS) and then annotated using SNPdrug3D. More than half of the damaging variants and around 40% of VUS variants affected residues near or at drug binding sites. VUS Variant of unknown significance. Source data are provided as a Source Data file.

information and when considering reliable predictions only, the consensus of multiple structure prediction methods used here provides the highest usable coverage of sites in the human proteome. Taken together, we have mapped residues to both experimental and predicted structures while also considering homologous structures (i.e., B40 group—similarity of at least 40 to 98%), which greatly expanded coverage of the SNPdrug3D proteome from 15% (i.e., B98 only) to 65% and of the UniProt proteome from 18.7% to 72%.

Following the above, SNPdrug3D was used to annotate natural missense variants present in the SG10K Health and gnomAD cohorts. As summarized in Fig. 2c, analysis of the latest SG10K Health genetic variation catalog identified 158,331,366 single-nucleotide variants (SNV). Of these, 6.2% (9,770,964) were common (allele frequency, AF of more than 1%), while 49% (77,625,433) and 44.8% (70,934,969) were private (seen in only one individual) or rare (AF < = 1% and in more than one individual), respectively. For each genome, an estimated median of over 11,000 missense SNVs per genome including over 200 novel variants (Fig. 2c, right) were found which is higher than previously estimated[18].

Including gnomAD variation data, there were ~5.8 million unique missense variants in total, with ~5.19 million missense variants from gnomAD (Fig. 3a, left) and ~1.25 million missense variants in SG10K Health (Fig. 3a, middle). Of these 5.8 million variants, around 0.6 million variants were exclusive to the SG10K Health dataset (i.e., not in gnomAD), and further comparison of these variants with dbSNP v151 revealed that ~0.3 million of these were entirely novel (Fig. 3a, right). Around 4.5 million variants were not present in SG10K Health (i.e. were found in gnomAD only) and the rest (~ 676,000) were found in both datasets (Supplementary Fig. 1). While these missense variants may cause amino acid substitutions in the corresponding reference proteins, not all were amenable to be mapped to amino acid residues in protein structures as they may mutate residues in disordered protein regions that were not crystallized or had low confidence predictions if mapped to AlphaFold structures. In total, 68.5% (~ 3.94 of 5.8 million) of variants had protein structure coverage (i.e., mapped to at least one structure with or without a drug molecule bound nearby) with similar proportions of variants structurally annotated in both the gnomAD (~ 67.8%) and SG10K Health (~ 71.1%) datasets.

To understand the contribution of protein-ligand interaction disruptions to disease phenotypes, we sorted a subset of SG10K Health variants (i.e., 1312 missense variants from 508 genes) into three groups (i.e., benign, damaging and VUS) based on annotations from the SG10K Med dataset[33] where variants are already curated to identify highly pathogenic variants. Here, all likely pathogenic or pathogenic variants were consolidated into the damaging group, and similarly, likely benign or benign variants were grouped into the benign class. Using

SNPdrug3D, we observed that ~97% of these 1312 variants were mapped to a protein structure, and of these, ~48% resulted in amino acid substitutions within 8 Å of a drug binding site (Fig. 3b) which is more than two times higher compared to the whole SG10K Health set (21%, Fig. 3a, middle).

By variant pathogenicity class (Fig. 3b), we show that more than half of the damaging variants (~ 51%) affected amino acids in drug-binding sites. As endogenous ligands and drug molecules may share similar binding sites, an amino acid change at these sites may disrupt protein-ligand interactions, leading to a loss of protein function and the manifestation of the disease phenotype. If protein-drug interactions are affected, the drug effect may be diminished, and therapeutic failure may occur in addition to toxicity if drug metabolism is decreased.

Another important role for SNPdrug3D is to help improve the annotation of variants of unknown significance (VUS) with potential effects on drug binding. We observe that ~40% of VUS in this dataset ($n = 204$) affect residues near a drug in a ligand-binding site, suggesting that disruption of protein-ligand interactions may be a possible molecular phenotype of these variants, and besides providing enhanced annotation, makes them prime candidates for further experimental verification of the mechanism underlying the proposed effect.

Using SNPdrug3D to identify missense variants that affect protein-drug binding experimentally

The extensive mapping of missense variants to protein structures allowed us to pinpoint variants that can affect drug binding, possibly leading to varying drug response phenotypes. In the combined global data, we found 1.17 million mapped variants that altered amino acid residues within 8 Å of one or more drugs in at least one protein structure. To delineate and validate the effects of such variants on protein-drug binding, we examined the effects of several variant-induced amino acid substitutions in drug targets and well-known drug metabolizing enzymes (DMEs). The substitutions were selected based on SNPdrug3D annotation of drug proximity to substitutions (within 8 Å) that suggest that they could directly disrupt protein-drug interactions at the drug binding pockets, and they comprise both known and novel examples. Some substitutions away from drug-binding sites (> 8 Å) were also chosen to analyze long-range effects of protein mutations.

For DMEs, we investigated the effects of single amino acid substitutions on two well-characterized enzymes (i.e., CYP2C19 and CYP2D6) where protein-drug binding might lead to alterations in drug pharmacokinetics. Most substitutions were located close to the binding cavity of the drug and the heme group of the holoprotein (Fig. 4a, upper panel) in the enzymes. In CYP2C19, which is responsible for the metabolism of up to 15% of known drugs[34], we tested the effects of two amino acid substitutions (NP_000760.1:p.A297V and p.R433W) on drug binding in an unbiased manner across nine major and minor metabolites of the enzyme based on SNPdrug3D data. Both substitutions are within 5 Å of the drug but p.R433W is known and is part of the no-function CYP2C19*5 haplotype, which has been demonstrated to abrogate metabolism of clopidogrel by decreasing protein stability[35] while p.A297V has an unknown effect. Here, validation of protein-drug binding using a CETSA-luciferase binding assay (see Methods and Supplementary Data 1) demonstrated that the effects of the two CYP2C19 substitutions segregated into three distinct clusters (Fig. 4b). As expected, all R433W-drug pairs clustered together (red diamonds in Fig. 4b) in a distinct cluster away from the WT or control suggesting that the R433W-mutated protein disrupted drug binding along with affecting protein stability. Conversely, all A297V-drug pairs clustered with the WT (green circles) except for A297V-estradiol where a stronger binding is observed compared to WT-estradiol binding. While this would unlikely translate to a clinical impact as estradiol is a minor metabolite of 2C19 and there are other pathways for metabolism, such as by CYP1B1, 2 A, 3 A and sulfation via sulfotransferases (PharmGKB pathway)[36], the result here points to a substrate-dependent effect for A297V-estradiol which will require further validation but has been observed in other substitution-drug pairs in past studies[37].

Next, we analyzed a set of 22 substitutions based on proximity to drug binding site in a CYP2C19-clopidogrel system with 11 substitutions within 8 Å of clopidogrel (i.e., within substitutions) and the other 11 located more than 8 Å from the drug (i.e. far substitutions) (Fig. 4c). Here, we observed four distinct clusters with the blue and green clusters clearly separated and distinct from the other two clusters (purple and red). Points in these latter two clusters are closer to known damaging substitution (i.e., p.R433W, purple cluster). Here, all substitutions in the purple cluster resulted in protein destabilization and reduced binding compared to WT, while substitutions in the red cluster resulted in protein destabilization but improved binding or maintained binding compared to WT. On the contrary, substitutions in the green cluster had similar effects to WT, while the ones in the blue cluster led to protein destabilization with concomitant improvement in binding to clopidogrel compared to WT. Overall, a majority of within substitutions (8 of 11, ~73%) do not cluster with WT and are in destabilizing clusters (purple and red) with no change or less binding compared to WT. Conversely, only 3 of 11 (~ 27%) far substitutions have similar characteristics. The rest are clustered with WT ($n = 3$) or increase binding stability compared to WT ($n = 5$, blue cluster).

For CYP2D6, we experimentally validated the effects of point substitutions on enzyme thermal stability and binding to N-desmethyltamoxifen (NMT, a metabolite of tamoxifen, which is metabolized by CYP2D6 to the active anticancer metabolite endoxifen[38]). Additionally, to diversify testing of drugs for CYP2D6, we also tested the effects of two substitutions on binding of prinomastat—an investigational drug and a potent inhibitor of CYP2D6[39]. For CYP2D6-NMT, we observed that most substitutions ($n = 5$) clustered with WT, including CYP2D6:p.T76M, the only long-range (> 8 Å) substitution we tested (Fig. 4d, red circles) for the enzyme. Two of the three remaining substitutions clustered to the right of the WT (Fig. 4d, green circles) and resulted in mutant proteins with similar binding affinity to NMT compared to WT but were more thermostable (i.e., p.V104A and p.V370I). The outlier is F120I, which, according to assay results, increased enzyme thermostability while also having a better binding affinity to NMT than WT. For prinomastat, as for NMT, a similar effect to WT was observed for p.T76M. However, a substrate-dependent effect was observed for p.F120I, which stabilized the enzyme with minimal changes to prinomastat-CYP2D6 binding affinity as opposed to the increase seen in NMT-CYP2D6 affinity for the same substitution.

To investigate the effects of single amino acid substitutions on drug pharmacodynamics, we also validated their impact on drug binding in drug targets using CETSA. As before, we used SNPdrug3D to identify variant-drug pairs, focusing on missense variants causing amino acid substitutions within 8 Å of bound drugs.

For hematopoietic cell kinase (HCK), we tested the effects of point substitutions on binding of four protein kinase inhibitors to the kinase domain (Fig. 5a), two of which were general pan-kinase inhibitors (i.e., dasatinib and staurosporine), with the remaining being HCK-specific inhibitors (iHCK37 and RK20449) designed for anticancer treatment (i.e., targets HCK dysregulation in acute myeloid leukemia)[40]. For the specific inhibitors, we observed that NP_001165601.1:p.N385K decreased kinase binding of RK20449 but not iHCK27, while being more thermostable than the WT (Fig. 5b). In contrast, p.G268R decreased stability and increased binding to RK20449 but not iHCK27. For the pan-kinase inhibitors, we focused our efforts on the approved drug (dasatinib) and similarly, found that p.N385K impaired kinase binding to dasatinib and also staurosporine, while p.G268R had the opposite effect of increasing the binding affinity of dasatinib to HCK.

https://doi.org/10.1038/s41467-026-73816-4

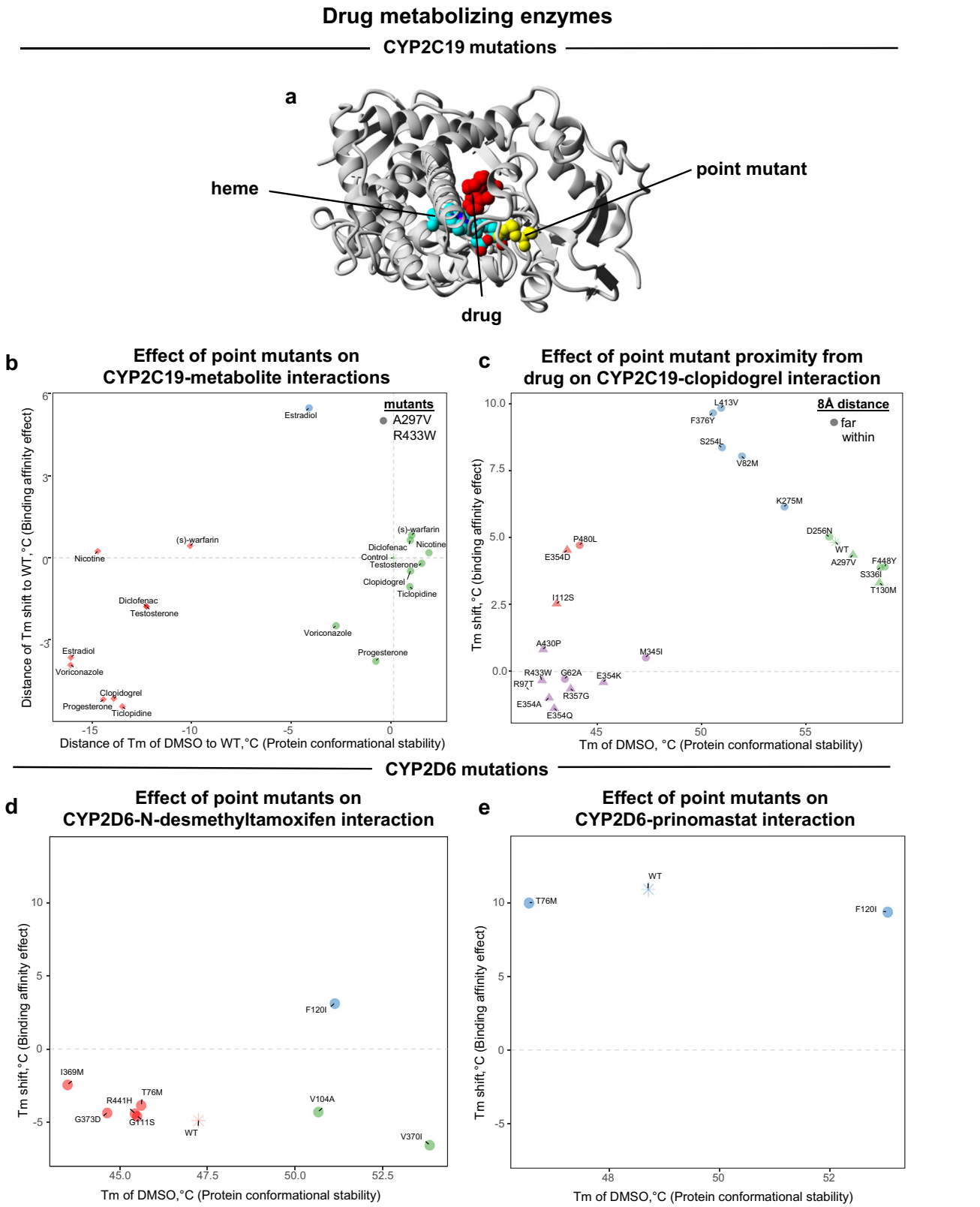

Fig. 4 | Known and previously uncharacterized single amino acid substitutions that affect binding of drugs to drug metabolizing enzymes (CYP2C19 and CYP2D6). a Prototypical structure of a cytochrome P450 enzyme with a heme center that anchors drugs at the catalytic site for metabolism (PDB = 3ME6). **b** Analysis of point substitutions in CYP2C19 on enzyme stability and metabolite binding. **c** Analysis of drug proximity of point substitutions on CYP2C19-clopidogrel binding and enzyme stability. Analysis of point substitutions effects on **d** CYP2D6-NMT or **e** CYP2D6-prinomastat binding and enzyme stability. Points in plots (**b**–**d**) were clustered using k-means clustering with each cluster labeled with a different color. DMSO dimethyl sulfoxide, T_m Half-maximal thermal melting temperature, WT Wild-type. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-73816-4

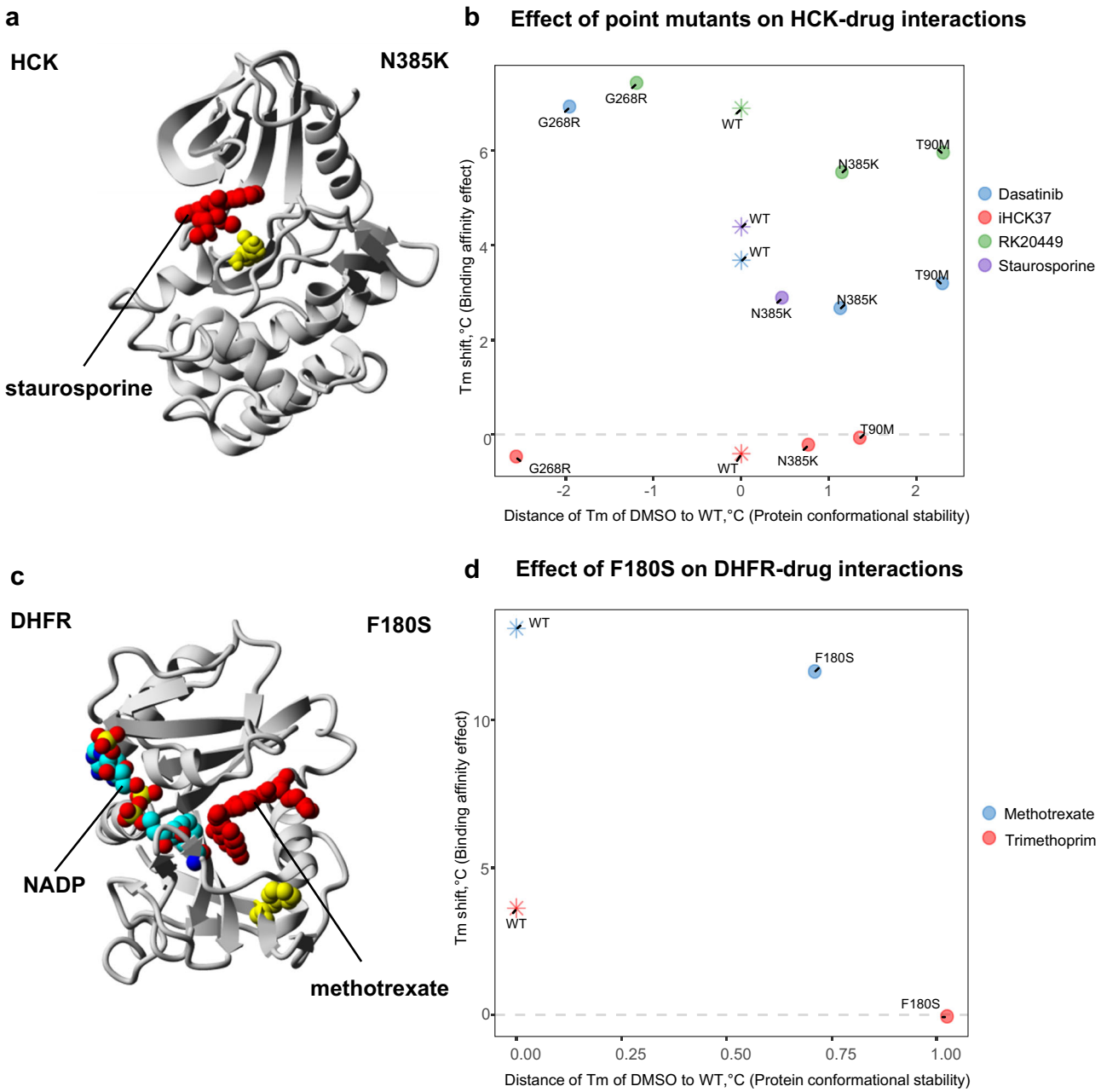

Fig. 5 | Single amino acid substitutions in HCK and DHFR and their effects on enzyme stability and drug binding. a Structure of a HCK homolog with a pan-kinase inhibitor, staurosporine, bound to the kinase domain of the enzyme (PDB = 1QPJ). **b** Analysis of point substitutions effect on HCK-kinase inhibitor binding using HCK-specific and pan-kinase inhibitors. **c** Structure of methotrexate-bound DHFR homolog (PDB = 3CD2). **d** Analysis of F180S effect on DHFR-inhibitor binding with methotrexate and trimethoprim. DMSO dimethyl sulfoxide, Tm Half-maximal thermal melting temperature, WT Wild-type. Source data are provided as a Source Data file.

Overall, the binding of iHCK37 to HCK is insensitive to the substitutions tested, unlike what was observed for RK20449. As such, further development of RK20449 or molecules with a similar scaffold as an HCK-specific inhibitor for anticancer treatment should be subjected to more extensive pharmacodynamics testing for resistance potential, especially if the substitutions are near the kinase binding site.

For dihydrofolate reductase (DHFR), we validated with CETSA the effect of a single substitution with two antifolate drugs, with one being more specific to human DHFR (i.e., methotrexate) while the other, trimethoprim, targets bacterial DHFR (Fig. 5c). Here, we observed a larger thermal shift and hence, greater binding affinity for methotrexate to human DHFR than trimethoprim (Fig. 5d). This aligns with the expectation that trimethoprim is selective for bacterial DHFR but binds weakly to vertebrate DHFR due to active site differences[41] unlike methotrexate. Of interest then are substitutions that can change this dynamic, and here, we found NP000782.1:p.F180S increased protein stability and significantly impairs trimethoprim-DHFR binding but minimally affects methotrexate-DHFR binding compared to WT in human DHFR. Overall, annotation and prediction of population-specific effects on

drug binding is invaluable for drug development efforts which can study and address such effects early in the development pipeline avoiding costly surprises at later stages[42–44]. This especially includes effects in the important CYP family of drug metabolizing enzymes, which display varying variant segregation patterns across populations[45].

Deriving predictive features from SNPdrug3D to build a PGx-inclusive variant effect prediction tool

Beyond mapping and visualization of missense variants in protein-drug complexes, we also hypothesized that information from SNPdrug3D, such as amino acid substitution proximity to drug molecules (or ligands in general) could serve as useful predictive features to build variant effect prediction tools that are able to discriminate PGx variants in addition to disease variants from neutral variants rather than just binary classification of variants into damaging (i.e. disease and PGx) and neutral categories. To the best of our knowledge, tools capable of such ternary classification of missense VUS have not been successfully developed but a previous attempt was made using a random forest classifier which found that PGx and neutral variants had overlapping characteristics[46]. To this end, we focused on variants in the cytochrome P450 (CYP) superfamily of enzymes as variation in CYP genes are known to cause disease (i.e. *CYP21A2* and congenital adrenal hyperplasia[47]) and also alter drug responses (i.e. *CYP2D6* variation and tamoxifen metabolism[48]). The CYP family has a dominant role in drug metabolism with 80% of FDA-approved drugs from 2005 to 2016 being metabolized by 4 of the 57 CYP gene family members[49]. Using supervised machine learning, we built a ternary linear-discriminant analysis-based classifier with a manually curated dataset of 1,301 known CYP missense variants (Supplementary Data 2) of which 712 were disease variants, 397 were neutral variants and 192 were PGx variants. Two categories of predictive features were used: i) predictions and rank scores from dbNSFP (including consensus and input of other methods), ii) distance-based annotations from SNPdrug3D based on whether a mapped variant is near (< 8 Å) or further from at least one drug after being mapped on to protein-drug complexes (see Supplementary Data 3 for the full list of features). Additionally, to leverage on the assay results in the previous section, we also trained a 'boosted' CYPVarPred containing an additional 17 PGx and 8 neutral variants (i.e. PGx if they did not cluster with WT in the assay, neutral otherwise) as listed in Supplementary Table 1.

To assess the classifier ('CYPVarPred'), a held-out test set containing 20% of the total variants in the CYP dataset was used. Comparison of CYPVarPred with a dummy classifier created by k-means clustering of these held-out variants to known variants in each class (i.e. R443W in CYP2C19 for PGx, S486T in CYP2D6 for neutral and P31L CYP21A2 for disease) showed that CYPVarPred vastly outperformed it with a Matthew's correlation coefficient (MCC) of 0.746 and an average class-specific accuracy of 80% across the three predicted variant classes (i.e. disease, neutral and PGx) (Fig. 6a) compared to the dummy classifier (MCC = 0.121, average accuracy = 38%).

To allow direct comparison with classical variant pathogenicity binary predictors, we binarized the outputs of CYPVarPred into neutral/non-neutral (i.e., PGx or disease) predictions and tested all predictors against two separate datasets consisting of either PGx/neutral (n = 18/51) or disease/neutral (n = 113/51) database-annotated variants (Fig. 6b, left). Using defined cut-offs (see Methods) from the various tools to compare with CYPVarPred, we found that CYPVarPred performed the best in classification of variants in both datasets (AUROC = 0.96 and 0.90 on the disease and PGx set, respectively) with no difference in performance between the boosted (Fig. 6b, left and Supplementary Fig. 2a and 2b) and non-boosted CYPVarPred (Supplementary Figs. 3 and 4). A full summary of predictors' performance statistics is available in Supplementary Table 2. Overall, all predictors performed better on the disease/neutral set compared to the PGx/neutral set where CYPVarPred has a clear advantage over the other tools.

To further validate CYPVarPred with data consisting of experimentally assayed CYP variants, we sought an independent dataset of 41 *CYP39A1* variants from a recent study of rare *CYP39A1* variants and their association with exfoliation syndrome[50], along with an additional 14 *CYP39A1* variants uniquely contributed by additional experiments in this study (see Supplementary Table 3) for a total of 55 variants (14 neutral, 41 disease). Across various thresholds and for the specific task of disease vs neutral classification (Fig. 6b, right), CYPVarPred, boosted (Fig. 6b, right and Supplementary Fig. 2c) or unboosted (Supplementary Fig. 3 and 4), performed comparably (AUROC = 0.7) to the state-of-art PrimateAI-3D (PAI3D)[51] (AUROC = 0.72) while both outperform other algorithms. Similarly, we sought to further validate PGx predictions from CYPVarPred and compare its performance to predictors by determining if they could predict substitutions that did not cluster with the wild-type in CETSA assay readouts for the CYP2C19-clopidogrel and CYP2D6-NMT pairs (Fig. 4c, d). There was a total of 25 variants in this test set (Supplementary Table 1), and none of them were in the database-curated CYP variant dataset (Supplementary Data 2). With the assay results as a benchmark, a total of 8 substitutions were clustered with the WT (i.e. neutral) while the rest (n = 17) were not (i.e., PGx). Using AUROC as a metric for comparison, we observed that non-boosted CYPVarPred (AUROC = 0.71) outperformed most predictors except PAI3D (0.75) and SIFT (0.76) (Supplementary Fig. 3 and 4) and when the data was included in the training of CYPVarPred, it has the highest AUROC of 0.80 (Fig. 6b, right and Supplementary Fig. 2d). However, at the default thresholds for the predictors (see Methods), some predictors, such as CADD and CYPVarPred could identify most or all the 17 PGx variants but with low specificity (Supplementary Table 2) suggesting that thresholds for these tools could be adjusted for a better balance between the false and true positive rates. Notably, adding the assay data into training did not cause CYPVarPred performance to drop for any of the test sets and hence, may be useful in training such PGx-inclusive predictors, although validation with a larger assayed PGx/neutral variant dataset will be needed for a more robust assessment of performance.

SNPdrug3D web interface

By combining sequence variant and structure data, SNPdrug3D provides user-friendly access to variant information that is specific to the respective cohorts (e.g., population-specific allele frequencies) in addition to structure and sequence-based information. A user just has to provide a query (e.g. gene name, drug name or SNV coordinates, up to 10 input types) in the webserver's search page (Fig. 7a) and select the relevant variant of interest from the results tables to enter the variant display page (Fig. 7b). The search defaults to variants mapped near to drugs (within 8 Å) but we provide an option to search for mapped variants away from drug binding sites as users could also be interested in protein-specific effects including variant effects on protein-protein interactions and signaling motifs. For each variant, a direct link out to the relevant variant page in gnomAD or the SG10K Health Chorus Variant Browser page is also provided.

At the variant display page (Fig. 7b) and in the sequence feature viewer, the user can inspect up to 14 protein sequence-based features (see Methods) to determine if the variant falls within a protein domain, in a disordered region or a predicted targeting or post-translational modification site. Additionally, nearby amino acid substitutions are also indicated in the viewer along with predictions from SIFT[8] and PolyPhen2[7] to allow the user to find out if a variant of interest is clustered with other potentially damaging variants. In the structure feature viewer, the queried residue (in red) and any nearby drug (in yellow if present and within 8 Å) are visualized along with the 5 or 8 Å binding pockets when this function is toggled. The structure containing a

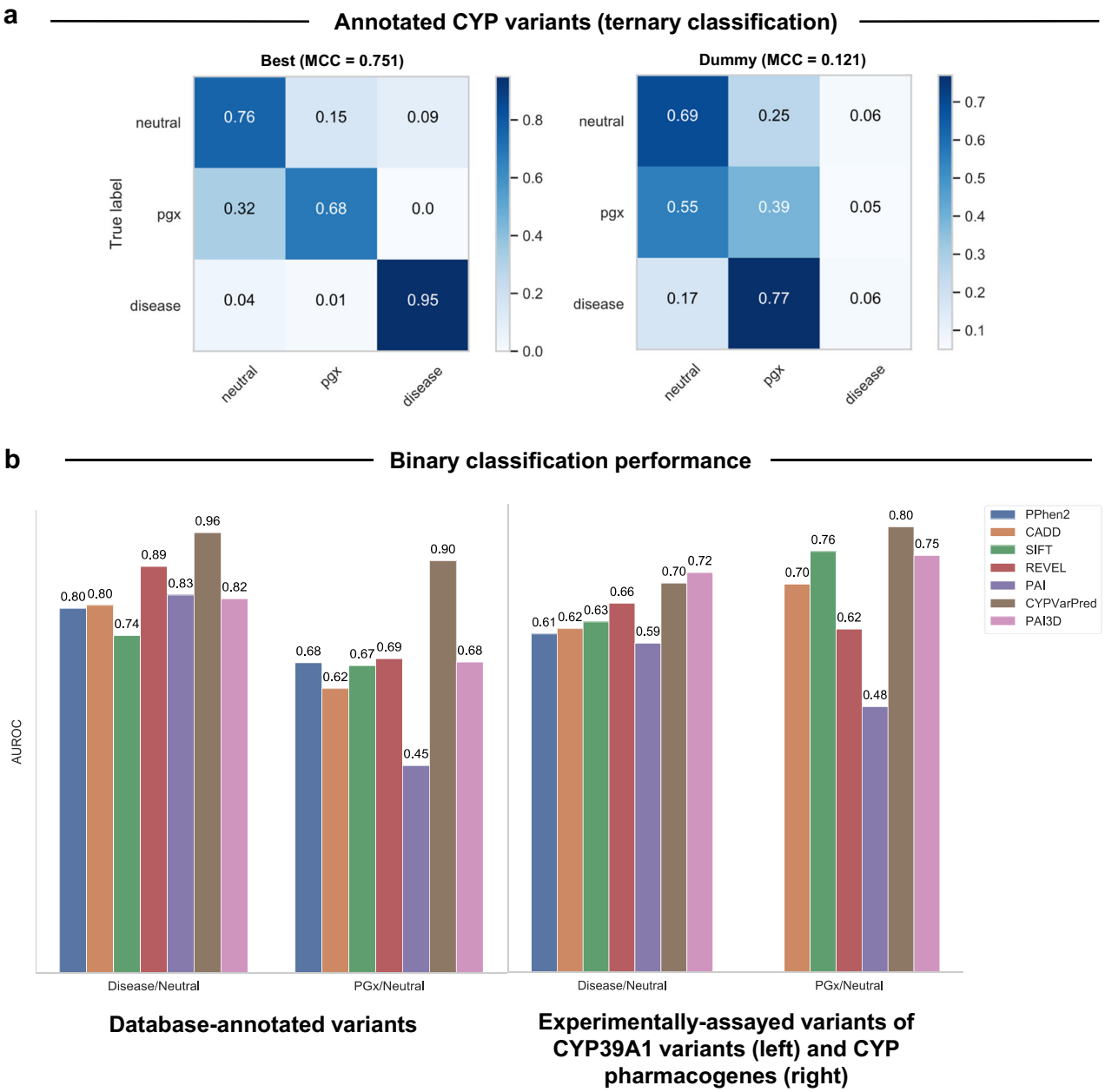

Fig. 6 | Performance and validation of CYPVarPred. a CYPVarPred (upper panel) outperforms (MCC = 0.746) a dummy classifier that leverages k-means clustering with k = 3 to cluster test datapoints (143 disease, 80 neutral and 30 PGx variants) in ternary classification of CYP missense variants. **b** (left) Binary classification performance based on AUROC of CYPVarPred and other oft-cited predictors on database annotated variants with disease/neutral (n = 118/51) or PGx/neutral blind datasets (n = 13/51). (right), Further validation of predictors on a literature-derived test set of 41 disease and 14 neutral *CYP39A1* variants and 17 PGx and eight neutral CYP variants from CETSA assay results. AUROC, area under the receiver operating characteristic curve. Source data are provided as a Source Data file.

nearby drug with sequence closest to the reference sequence is returned although the user can choose to display other structures (experimental or predicted) from a list below the structure viewer.

A benchmarking test conducted by searching ten highly polymorphic genes with at least 2000 hits showed that the average load time for the variant tables was 2 seconds which is 1.7x faster than an older beta version with a load time of 3.4 seconds (see Supplementary Table 4). Nevertheless, for highly polymorphic genes (e.g., *CYP2D6*), we recommend using the filters available to narrow the search to more specific groups of variants for faster search speeds and have implemented a pop-up warning to suggest the use of filters if the search has more than 2000 hits (~90% of searches using a gene name will have less than 2000 hits).

Discussion

SNPdrug3D distinguishes itself from past work of mapping sequence variants to protein-ligand binding sites[19–26,52] because of its unparalleled coverage (~65% of proteome in 3D, 20,442 genes) and size of reference proteome (~100,000 protein sequences including isoforms), as well as variant annotation depth due to inclusion of homology models (ModBase and HHPred), PDB structures and predicted structures from AlphaFold. It must be noted however that ~20% of the residues in AlphaFold structures have pLDDT of less than 70 and were excluded in our analyses. These scores indicate regions of low confidence in the structure prediction, but this is expected in residues falling in loops, linkers and termini that are structurally flexible/disordered and not restricted to rigid conformations. Further,

https://doi.org/10.1038/s41467-026-73816-4

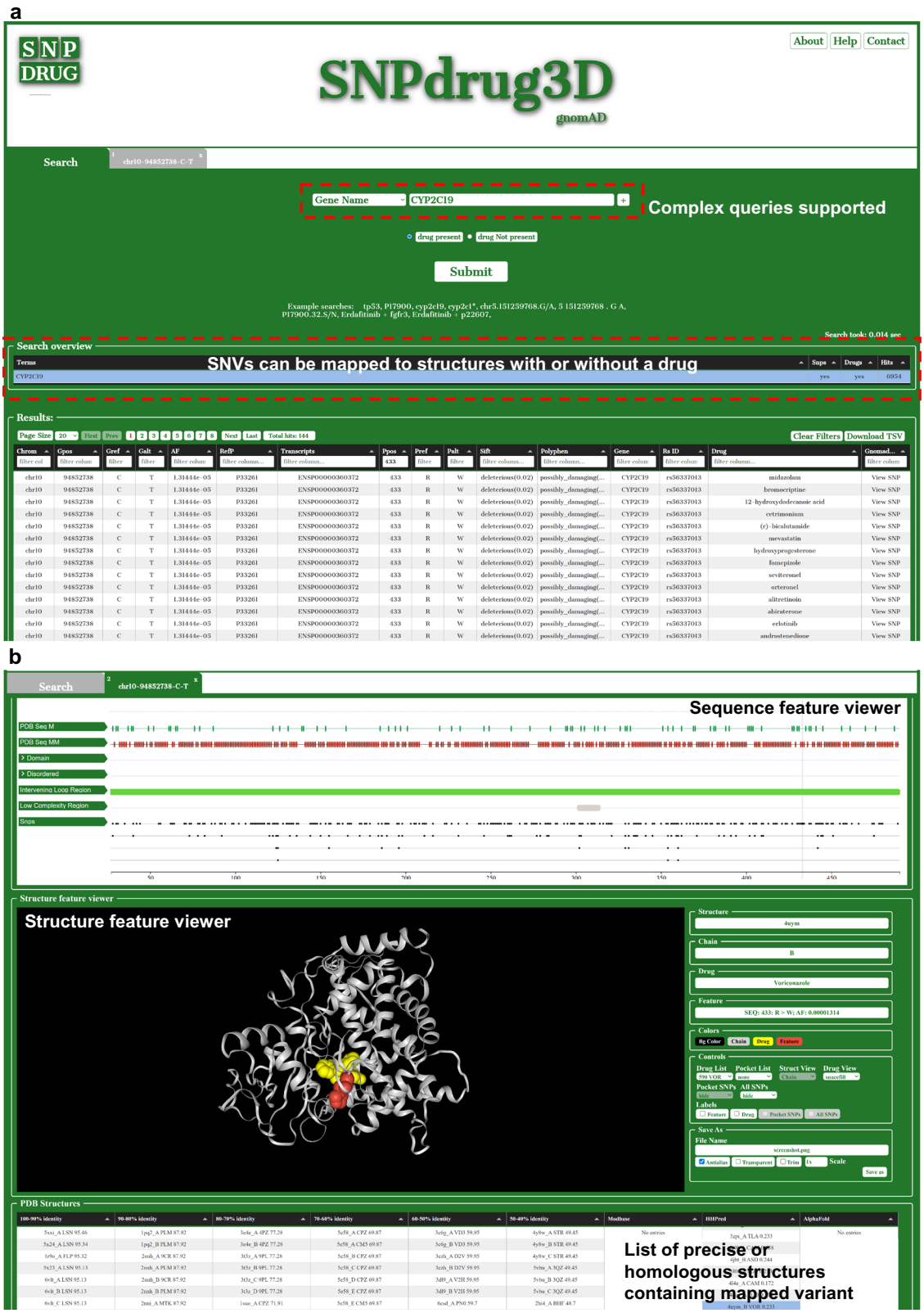

Fig. 7 | SNPdrug3D web interface. a Variant search page that allows users to search (boxes bound by red dotted line) for variants using up to 10 input types (e.g. by Gene Name, Protein ID, DrugBank Name etc). Filters are available in the results table for users to quickly filter out variants of interest. **b** Variant display page with sequence and structure feature viewers containing information about the mapped variant. The structure closest by homology to the queried protein containing a drug with the mapped variant is displayed by default but a user can select other structures containing the mapped variant.

SNPdrug3D also contains information on 6,000 drugs from DrugBank that were identified in PDB structures, and, where appropriate, drug-variant proximity data was calculated using VMD[53]. For further improvements, structural summaries from PDBsum[54] and more complex binding site predictions (e.g., GenProBiS[19]) can be considered. Extensive consideration and analysis of isoform-specific variant effects across protein families were also beyond the scope of the study. However, using molecular dynamics simulations, we did identify two rare variants in p53 (p53:p.I232L and p.I232F) found in SNPdrug3D that are predicted to affect binding only to the DNA-binding domain (DBD) of Y220C-p53 delta133 isoform but not the DBD of untruncated Y220C-p53 (Supplementary Fig. 5).

The data in SNPdrug3D is also of clinical relevance as variation data in the webserver were obtained from the whole-genome sequencing data of more than 80,000 individuals from both gnomAD and SG10K Health cohorts (Fig. 2c). For the Singapore population, the unique variants that do not overlap with dbSNP and gnomAD (~0.3 million variants) (Fig. 3a, right) are of particular interest as functional variants in this group may explain population-specific differences in disease susceptibility and drug response. Specifically, after mapping a small subset of 1312 SG10K Health variants with curated pathogenicity annotations[33] to protein structures, we found most variants annotated to be damaging (~51%) were also mapped to residues near one or more bound drugs in at least one protein structure while around 40% of variants of unknown significance (VUS) were mapped to residues near a bound drug hence providing a rationale for reannotation and testing of these variants. For the VUS group, the use of SNPdrug3D would support more detailed investigations and biochemical characterization of the possible effects of these VUS variants on protein-drug or ligand binding sites whereas for damaging variants, SNPdrug3D results can complement existing pathogenicity predictions (e.g., CADD[55] or REVEL[56] predictions) or annotations by providing a mechanistic explanation behind the predicted deleterious outcome if the damaging variant was found to possibly disrupt a residue important for drug or ligand binding. Complementary assessment with stability predictors may also be useful to investigate if the variant-induced loss of binding is due to protein destabilization in addition to a direct steric disturbance at the binding site. We explored the prediction ability of two well-known tools[57,58] (i.e. FoldX and DUET), and observed that, when correlated with experimental results (Fig. 4c, 22 substitutions in CYP2C19), they were able to predict destabilizing substitutions that clustered away from WT in the experimental data with high accuracy (i.e. 16 out of 19 or 84% and 17 out of 19 or 89% for DUET and FoldX respectively) (Supplementary Table 5).

Importantly, we show that SNPdrug3D can be used to identify variants in DMEs and protein drug targets that can affect protein-drug binding using a simple distance-based threshold of 8 Å which we found to enrich for variants disrupting drug binding in a CYP2C19-clopidogrel system (Fig. 4c). In total, we tested a total of 78 variant-drug pairs (including WT-drug pairs) across 2 DMEs (CYP2C19 and CYP2D6) and 2 drug targets (HCK and DHFR) and 17 unique drugs using CETSA to investigate variant impact on drug pharmacokinetics and pharmacodynamics (Figs. 4 and 5). Most variants tested are rare (<1% allele frequency (AF) in gnomAD or SG10K Health) with a mix of known (e.g. R433W) and unknown variants. Of note are two variants leading to substitutions with substrate-dependent effects on drug binding. The first is in CYP2D6:p.F120I (JF307778.1:c.358 T > A), which has been found to be important for ligand binding in typical CYP2D6 substrates (i.e., positively charged at physiological pH[59]), which is unlike prinomastat, which does not have a positive charge[39]. This might account for the difference in binding affinity observed between NMT and prinomastat (Fig. 4d, e) although further investigation is needed. Further, the variant-induced increase in binding observed only for NMT-CYP2D6 is compelling if this leads to greater catalytic efficiency of NMT to endoxifen as this may have implications for tamoxifen dosing although the variant is very rare (< 0.1% AF in gnomAD and SG10K Health). The other variant of interest - NM_000791.4:c.539 T > C, DHFR:p.F180S - is specific to the SG10K Health cohort, found rarely only in the Malay population (< 0.1% AF) and impairs trimethoprim but not methotrexate binding (Fig. 5d). While this is unlikely to affect methotrexate treatment, it would be compelling to further investigate if having the p.F180S substitution would confer protection against methotrexate toxicity due to reduced folate metabolism with concomitant use of trimethoprim. This is likely more relevant for side effects resulting from antifolate action, such as myelosuppression which is a rare but potentially life-threatening side effect that has been reported even when low-dose methotrexate was given along with trimethoprim[60]. Although we have experimentally validated a diverse set of proteins, variants and drugs, the dataset here represents a very small fraction of the 1.17 million variants we found mapped to drugs in one or more protein structures. While it would be infeasible to test every single one, further work can focus on investigating variant binding effects of established drugs and proteins where genetic variations have been shown to affect clinical outcomes (e.g. FDA's list of Pharmacogenomic Biomarkers).

We also demonstrate that information derived from SNPdrug3D serves as useful predictive features to help build PGx-inclusive missense variant prediction tools, such as CYPVarPred. While most conventional tools (e.g., SIFT[8], PolyPhen2[7]) rely heavily on evolutionary or conservation-based scores to distinguish pathogenic from neutral variants, such features may not be enough to discriminate PGx from neutral variants especially if the disease and PGx variants share dissimilar evolutionary patterns[61]. Here, we show that conventional tools performed reasonably in distinguishing diseased but not PGx from neutral CYP variants (Fig. 6b and Supplementary Fig. 2) with some tools, such as PrimateAI having a high default cut-off (i.e. 0.803)[62] that may be unsuitable for classifying variants in our CYP dataset unless the threshold is adjusted downwards to 0.486 and 0.527 for PrimateAI and PAI3D respectively (see Methods). Given these observations, there have been limited attempts to generally predict and classify PGx variants by redefining cut-offs[61,63] and the inclusion of other features, such as information derived from protein-drug complexes (e.g., from docking models) was also suggested to try to improve the performance of PGx-based classifiers[46]. We show that the latter approach is feasible in a CYP-specific context with good performance in discriminating both disease and PGx variants from neutral variants for both ternary (Fig. 6a) or binary (Fig. 6b) classification tasks. However, we observe that performance across most tools were considerably lower when experimentally assayed variants, instead of database-curated variants, were used for benchmarking the predictors in a disease versus neutral or PGx versus neutral comparison (Fig. 6b and Supplementary Fig. 2). This has been similarly demonstrated in past studies[64,65] where it has been suggested that truth sets consisting of functional assay-based variants should be used to improve evaluation of prediction models to ensure that their predictions are robust and clinically relevant. Beyond CYP variants, we observed that addition of a small dataset of SLCO1B1 variants (i.e. 20 for training, 5 in test set, see Supplementary Table 6) only marginally decreased the performance of the predictor on the database-curated and assay-based test sets (i.e. AUROC dropped by 0.01 and 0.02 respectively compared to the CYP-only model) (see Supplementary Fig. 6). This suggests the possibility of extending predictions beyond CYP variants although a larger dataset of variants in non-CYP pharmacogenes would be needed to reliably assess the multi-protein predictor performance.

It is expected that as more human variation is captured and additional protein structures are obtained, the size and scale of SNPdrug3D will grow along with its utility (e.g. Phase 2 of Singapore's NPM program aims to generate a total of 100,000 Singaporean genomes). We therefore hope that SNPdrug3D will be an integral tool for investigations of missense variants and their effects on ligand binding

to proteins and aid the integration of variant information into a clinical set up for more precise and personalized drug treatments. Specifically, with SNPdrug3D's approach, one can identify candidate variants that influence drug response in a population-specific pattern, such as in age-related macular degeneration where distinct treatment outcomes are notable in Asia[66], thereby also guiding new drug development. Overall, SNPdrug3D will support future precision medicine initiatives through comprehensive interrogation of the molecular phenotype associated with missense VUS and their relation to inter-individual variation in drug responses and diseases.

Methods

SNPdrug3D reference proteome

A reference proteome was built using data from the following sources: Ensemble 96 (https://www.ensembl.org), RefSeq GRCh38.p12 (GCF_000001405.38) and UniProt[67] UP000005640_9606 (both reviewed and unreviewed fasta files, downloaded on 18 September 2019). Variation data from the SG10K Health[2,3] v4.3 cohort was generated by the Genome Institute of Singapore (GIS) and obtained through the National Supercomputing Center of Singapore, in a tab separated file. This cohort consists of healthy individuals of Malay, Chinese and Indian ancestries that were participants in 6 smaller study cohorts in Singapore (e.g. Multi-Ethnic Cohort (MEC) study, Health for Life in Singapore (HELIOS) study and Growing Up in Singapore Towards healthy Outcomes (GUSTO) study). Further information regarding the SG10K Health cohort is available at Singapore's National Precision Medicine portal (https://www.npm.sg/data-access-to-sg10k-health/) and in papers published by the SG10K Consortium[2,3]. Data in the form of variant call format (vcf) files for gnomAD[1] v3.1.2 was obtained from the consortium's website at https://gnomad.broadinstitute.org.

To create a final reference proteome, the protein sequences were cleaned by removing sequences with more than 40% "*" (i.e. translation stop) in their sequences. This is needed for clustering where the cleaned protein sets are merged and then clustered by 100% identity using CD-HIT (v4.6.7)[68] with the following options: -c 1 -d 500 -T 30 -M 2000. The protein sequences which are not included in the reference proteome were mapped to the proteome with BLASTP.

Variation data from both the SG10K Health and gnomAD cohorts were then mapped to the reference proteome and annotated using the Variant Ensembl Predictor[69] (v95 cached). Only variants classified as 'missense_variant' and 'stop_gained' were retained. After mapping, a tab-separated file was generated for further mapping of the variants to protein structures in the subsequent sections.

Obtaining and parsing protein structure data

A complete copy of the DrugBank[32] was acquired on the 1st of Nov, 2021. SNPDrug3D uses ligands identified as drugs by DrugBank and utilizes annotations from it which are displayed alongside the SNP viewer in SNPdrug3D. Structural data was acquired on 1st November 2021 from RCSB PDB[29] and included, the standard structure record (PDB), the extended macro-molecular crystallographic information file (mmCIF), the protein sequence record (SeqRes) and the chemical components library (CCL).

Protein structure data was parsed using VMD (v1.9.3)[53] for identifying protein residues within 5 and 8 Å distance from ligands. Additional data related to ligands (e.g., heavy atom count, chemical formula, three-letter (HET) identifier etc.) were obtained from the CCL. At present, only ligands classified as drugs by DrugBank with a heavy atom count greater than six atoms are included within SNPDrug3D.

Mapping of reference proteome to protein structures

Mapping in the form of an empirical lookup table was generated between the reference proteome residues and those in protein structures. Sequence alignments for this purpose were generated using BLASTP (v2.8.1). To reduce the chances of erroneous mapping between the reference protein and the mapped position in the three-dimensional structure (3D), pairwise alignments were calculated between the reference protein sequence, the PDB fasta record and the sequence of residues within the protein structure. Drug structural proximity data, calculated earlier with VMD, was added to these lookup tables. While using BLASTP, a 40% sequence similarity cut-off was used to include significant results. The blast matches were classified into two categories B98 and B40, where B98 comprises similarity above 98% and B40 comprises matches with similarity between 40% and 98%. For the purposes of detecting remote homology, Modbase (Human genome dataset v.2020)[30] and hhblits (HHpred suite database v.2021)[70] were used. Pre-calculated alignments against proteins in the reference proteome were obtained from Modbase to obtain template structures. Additionally, hhblits was used to identify templates for all proteins in the reference proteome.

Handling and mapping of AlphaFold data

Structural predictions of the human proteome, by AlphaFold[31], as provided by the EBI-AlphaFold database[71] (available at https://alphafold.ebi.ac.uk/) are included in SNPDrug3D. The lookup tables, explained above, include mappings to predictions by AlphaFold to provide enhanced annotations for variants.

The SIFTS[72] record was used to obtain the mapping between Uniprot IDs and PDB structures. Structures with Uniprot IDs for which AlphaFold predictions are available were obtained from the online database mentioned above. Pairwise structural alignment was carried out between all experimentally determined protein structures in RCSB PDB and their AlphaFold counterpart structures which had drugs bound. The rigid structural alignment was carried out using GESAMT[73]. Post-structural alignment, new protein structure files were generated which included the AlphaFold predicted protein in complex with the ligand acquired from the experimentally determined structures. In cases where steric clashes were observed, the complexes were not generated.

Mapping of variants to protein structures

After the creation of the lookup tables in the previous sections where each reference protein sequence is mapped to a structure, we finally mapped each variant found in the gnomAD or SG10K Health cohorts to one or more protein structures. These structures can be from PDB (mostly x-ray crystallographic structures), Modbase (homology models), HHPred (homology models) or AlphaFold (template or non template-based protein structure predictions).

SNPDrug3D webserver

The webserver was built using two programming languages, Perl v5.26 (http://www.perl.org) and JavaScript. For the backend and data analysis or generation, Perl was used while the front end was developed using JavaScript. The building of the search indices was done using Apache Lucy (https://metacpan.org/pod/Lucy) while Mojolicious (https://metacpan.org/pod/Mojolicious) and two in-house Perl modules were used for the webserver's backend. The webserver's front end utilizes the following javascript libraries: Feature-viewer-typescript v.2.2.3[74] (https://github.com/BioComputingUP/feature-viewer-typescript), NGL[75] (https://github.com/nglviewer/ngl), Vue v2.6.10 (https://vuejs.org), Tabulator v4.4.3 (http://tabulator.info), Axios v0.19.0 (https://github.com/axios/axios), Vue-spinner (https://github.com/greyby/vue-spinner), Sweetalert2 (https://sweetalert2.github.io), Color-picker v1.4.2 (https://github.com/taufik-nurrohman/color-picker), Dom-to-image version from 10-06-2017 (https://github.com/tsayen/dom-to-image).

Feature-viewer-typescripts were modified to improve the protein feature representations. Protein sequence features found in the Sequence Feature Viewer of SNPDrug3D were predicted using the following predictors or software: i) GPI anchors (BIGPI v3.2[76]), ii) Coiled

coils (Coil v1.1.1.1[77]), iii) Protein domains (HMMER2[78], hmmpfam v2.3.2[78], ncbi-blast-2.2.26 – rpsblast), iv) Disordered and intervening loop regions (IUPRed2A[79]), v) Myristoylation (NMT v1.4.1[80]), vi) Signal peptides and transmembrane topology (Phobius 1.01[81], SignalP v4.1[82]), vii) Prenylation (FT, GGT1 and GGT2)[83], viii) Peroxisomal targeting signal (PTS1 predictor[84]) and ix) Low complexity region (SEG[85]).

The SNPDrug3D server utilizes the mapping record, along with annotated SG10K Health and gnomAD variant records to allow lookups from users against their query (database deposited at https://doi.org/10.6084/m9.figshare.28080626). By default, the protein structure with the closest sequence similarity to the reference protein affected by the queried variant is displayed. Concurrently, structures from PDB and with drugs in the vicinity of the substituted amino acid are also prioritized for display in the Structure Feature Viewer.

CYPVarPred development

First, known CYP missense variant data was manually curated from UniProt humsavar.txt (2019_07, available at https://www.uniprot.org/docs/humsavar), HGMD[86] Public (free version), ClinVar[87] (11062020) and PharmVar[88] (v4.0.2). A complete list of variants used for training and testing CYPVarPred is available in Supplementary Data 2. Disease variants were 'pathogenic/likely pathogenic' variants in ClinVar, 'Disease' variants in Uniprot or variants that were associated with diseases in HGMD whereas neutral variants were described as 'Polymorphism' in UniProt or 'benign/likely benign' in ClinVar or associated with a normal haplotype but not a deleterious haplotype in PharmVar. Finally, PGx variants were variants that altered drug metabolism or catalytic activity of CYP pharmacogenes in HGMD, 'drug response' variants in ClinVar or are present in at least one haplotype causing altered CYP activity but not in a normal function haplotype in PharmVar. As there may be overlap of labels if a variant is present in more than one database, a consensus approach was undertaken to assign a final label to each variant with preference for non-neutral labels if a tie is observed. All variants were also lifted over to GRCh38 genome coordinates using Picard (v2.22.9) (available at https://github.com/broadinstitute/picard, accessed June 8th 2020). Where a consensus could not be achieved or the amino acid substitutions could not be mapped and lifted over to a specific genomic coordinate due to codon degeneracy, they were excluded from further analyses.

Features for CYPVarPred can be divided into two categories and a full list of features included in the final model is available in Supplementary Data 3:

i) Prediction outputs and rank scores obtained from dbNSFP v4.0[89] plugin via VEP v96.
ii) Binarized SNPDrug3D derived features (e.g., is mutation within 5 A or 8 A of a ligand/drug in the protein of interest?).

With the data and features above, we used supervised machine learning to train a linear discriminant analysis-based classifier using the scikit-learn (v0.20.4) library with Python v2.7. Model was trained using 5-fold cross-validation after splitting the dataset into two smaller subsets – an 80% training set and 20% held-out test set – in a stratified manner. Feature values were scaled using the *RobustScaler* function in sklearn and missing values were imputed using the most frequent value in each feature. Finally, univariate feature selection was implemented during model training by selecting and including only the top 84 features with the highest ANOVA F-values in the final model.

CYPVarPred validation

The predictor was evaluated using the 20% held-out database-cureated test set and further independent datasets of 55 CYP39A1 variants[50] with functional assay data (Supplementary Table 3) and a dataset of CETSA-assayed variants (Supplementary Table 1). All assayed variants were not present in the training set. To evaluate ternary classification performance, CYPVarPred was compared with a dummy classifier that uses k-means clustering with three centroids (i.e. 3 known variants - R433W CYP2C19 (PGx), P31L CYP21A2 (disease) and S486T CYP2D6 (neutral)) to determine if the test set of CYP variants clustered accurately to these centroids.

For binary classification, the outputs from CYPVarPred were binarized into neutral/non-neutral predictions and then compared against other predictors (SIFT[8], PolyPhen2[7], REVEL[56], CADD[55], PrimateAI[62] and PrimateAI-3D[51]) on a neutral/PGx dataset (n = 51/18) and neutral/disease dataset (n = 51/118). For PolyPhen2 we grouped the possibly and probably damaging predictions while the following cut-offs were used to group the variants into two classes for the following predictors: i) SIFT (< = 0.05, deleterious), ii) REVEL (> 0.5, deleterious), iii) CADD (phred score > 20, deleterious), iv) PrimateAI (> 0.486, deleterious) and v) PrimateAI-3D (> 0.527, deleterious). All thresholds used were default thresholds or suggested by the authors of the predictors except for PrimateAI and PrimateAI-3D where multiple thresholds were tested and the thresholds leading to the best binary classification performance on the CYP variant data (Supplementary Data 2) were used.

After obtaining the predictions for all tools, the Matthew's correlation coefficient[90], balanced accuracy, sensitivity and specificity were calculated using the *matthews_corrcoef, balanced_accuracy_score* and *recall_score* functions in sklearn to evaluate the performance of the classifiers for both the ternary and binary classification tasks. Additionally, to evaluate binary classification performance, an ROC curve was plotted after calculation of the TPR and FPR at different thresholds using the following formulae (TN, True Negative; FN, False Negative; TP, True Positive; FP, False Positive):

1) True Positive Rate (TPR)

$$TPR = \frac{TP}{TP + FN}$$

2) False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN}$$

All figures in this section were plotted using the Matplotlib (v0.9.0) and Seaborn (v2.2.3) libraries available in Python.

Prediction of variant-induced effects on protein stability

Predictions for variant induced effects on protein stability were obtained by first repairing the relevant single chain PDB structures (4GQS;CYP2C19, 3QM4;CYP2D6, 2DHF;DHFR and 1AD5;HCK) using the RepairPDB function of the FoldX 5.0 plugin[58] in Yasara Structure[91] (v18.2.7). Following this, the Mutate residue option of the plugin was used to introduce the relevant amino acid substitutions and the average ddG (kcal/mol) calculated from 5 runs at default settings is reported. Here, ddG values of more than 0.5 kcal/mol were considered destabilizing while values below -0.5 kcal/mol were considered stabilizing. We also used another method called DUET[57], and the prediction labels (i.e. destabilizing or stabilizing) and kcal/mol values for each amino acid substitution in each protein were obtained from the DUET webserver at http://biosig.unimelb.edu.au/duet/.

Molecular dynamics simulations to study SNVs effect on drug binding to p53 isoforms

The structure (PDB ID 8A32) of the oncogenic mutant Y220C of the p53 DNA Binding Domain bound with a ligand (nanomolar affinity Kd: 320 nM)[92] was used in this study. This crystal structure contained 4 additional mutations which had to be mutated back to generate Y220C (referred to as Y220C-p53DBD). The structure of the delta133 isoform complexed with the ligand was generated by deleting residues 94 to 132 from Y220C-p53DBD. Both the Y220C-p53DBD – ligand and Y220C-p53DBD ^{delta133} - ligand complexes were subject to MD simulations using

the standard protocols developed in our lab for studies of p53[93]. All atom versions of the Amber 14SB force field (ff14SB)[94] and GAFF[95] forcefield for ligand were used and the simulations were carried out with the *pmemd.CUDA* module of the program Amber18[96].

We first estimated the affinity of the ligand for Y220C-p53DBD^{WT} and for Y220C-p53DBD^{delta133}. This was quantified using the standard MMPBSA approach[97]. For the calculation, 50 frames at equal intervals were selected from the simulation trajectory of Y220C-p53DBD^{WT} – ligand across 100 ns while for Y220C-p53DBD^{delta133} – ligand it was taken from the first 12 ns (during which the ligand remained bound to the protein). This was performed to get an estimate of the effects of the lack of the 94-132 segment in the isoform; the missing region is not in contact with the ligand but is in contact with the region surrounding the ligand.

To estimate the effects of the SNPs (P222L, I232L, I232F), both residues being in the vicinity of the bound ligand, we carried out FEP/MBAR calculations using protocols we have developed earlier[98]. These values are calculated separately for Y220C-p53DBD and for Y220C-p53DBD^{delta133}. Simulation trajectories were visualized using VMD[53] and figures were generated using Pymol[99].

Cell culture and generation of Tet-On reporter cells

HEK293 Flp-In, T-ReX cell line was purchased from Invitrogen (#R78007) and cultured in DMEM containing 10% FBS (Excell), 1% NEAA, 1% Penicillin/Streptomycin (Gibco), 15µg/ml Blasticidin (Invitrogen) and 100µg/ml Zeocin (Invitrogen). The inducible reporter cells were generated by using the Flp-In T-ReX system. Target genes including a GSSG linker, a HiBiT tag at either C- or N-terminus, and attL sites were synthetically synthesized and cloned into the pDONR221 plasmid by Beijing Hitrobio Biotechnology Co., Ltd. Cloned plasmids were used to transfer the target gene into the pFRT-TO-DEST gateway destination vector (Addgene plasmid #106348) via the gateway LR reaction. HEK293 Flp-In, T-ReX cells were transfected with the modified destination vector including a HiBiT-tagged protein of interest and pOG44, expressing Flp recombinase. Cells were exposed to 100 µg/ml Hygromycin (Invitrogen) and 15µg/ml Blasticidin antibiotics for 10 days to eliminate untransformed cells.

Quantitative CETSA-luciferase assay

Transformed HEK293 Flp-In T-ReX reporter cells were seeded with 1µg/ml final concentration of Tetracycline (HARVEYBIO) in DMEM one day before the experiment. The following day, cells were trypsinized and counted. 10^5 cells were prepared in DMEM high glucose without phenol red (HyClone), including protease inhibitor for each different temperature or different dose at a single temperature followed by the addition of compounds and incubation at 37 °C. After the drug treatment, cells were heated for 3 min followed by cooling down for 3 minutes in the thermocycler (VWR) (Doppio). To lyse the cells and measure the bioluminescence signal, the detection reagent that contains LgBiT (Promega), nanoluciferase substrate (Promega), IGEPAL CA-630 (Beyotime), and protease inhibitor (Roche) in PBS was added onto samples. Samples were dispensed to 384-well plate as triplicate followed by a quick spin. The plate was shaken in the microplate shaker for 5 min followed by incubation without shaking for 10 minutes in the dark. The bioluminescence signal was measured in the microplate reader (Tecan Spark).

Data analysis was performed using R (version 4.3.1). All the measured raw luminescence signals were first normalized and scaled to the readings at 37 °C. R package nplr was used to perform 4-parameter logistic regression, determination of half-maximal melting temperature (Tm50), and the estimation of the confidence interval of Tm50. Significance of Tm shift for each drug-gene pairs was tested using overlapping confidence interval approach[100]. The graph plotting was performed using R package ggplot2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Protein sequence data underlying the reference proteome is freely available from Ensemble (http://ftp.ensembl.org/pub/release-96/), NCBI RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/) and UniProt (https://www.uniprot.org/proteomes/?query=human&sort=score) while protein structure data can also be obtained from PDB (https://www.rcsb.org/), ModBase (https://modbase.compbio.ucsf.edu/), HHPred (https://toolkit.tuebingen.mpg.de/tools/hhpred) and the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). Drug-related information is available via Drugbank (https://go.drugbank.com/releases/latest). Access to aggregated variation data from the SG10K Health cohort is available upon free registration and logging into the National Precision Medicine portal (https://chorus.grids-platform.io/) while data from gnomAD is available for download at the consortium's website (https://gnomad.broadinstitute.org/downloads). The entire dataset supporting CYPVarPred development is available in Supplementary Data 2 and the consolidated mapping file for gnomAD variants is available in Figshare (https://doi.org/10.6084/m9.figshare.28080626). Finally, source data are provided as a Source Data file. Source data are provided with this paper.

Code availability

SNPdrug3D for gnomAD variants is freely available at https://snpdrug3d-gnomad.bii.a-star.edu.sg/ but access to SNPdrug3D with SG10K Health data will require free account registration at https://snpdrug3d.bii.a-star.edu.sg/. Additionally, the weights and sample prediction inputs and outputs for CYPVarPred can be found at https://github.com/sd3d-manuscript/sd3d.

References

1. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
2. Wu, D. et al. Large-scale whole-genome sequencing of three diverse asian populations in Singapore. *Cell* **179**, 736–749 (2019).
3. Wong, E. et al. The Singapore national precision medicine strategy. *Nat. Genet.* **55**, 178–186 (2023).
4. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
5. Tian, Y. et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. *Sci. Rep.* **9**, 12752 (2019).
6. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
7. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* Chapter 7, Unit7.20 (2013). https://doi.org/10.1002/0471142905.hg0720s76.
8. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
9. Tabor, H. K. et al. Pathogenic variants for Mendelian and complex traits in exomes of 6517 European and African Americans: implications for the return of incidental results. *Am. J. Hum. Genet.* **95**, 183–193 (2014).
10. Zhou, Y., Ingelman-Sundberg, M. & Lauschke, V. M. Worldwide distribution of cytochrome P450 alleles: a meta-analysis of population-scale sequencing projects. *Clin. Pharm. Ther.* **102**, 688–700 (2017).
11. Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).

12. Fattinger, K. et al. Epidemiology of drug exposure and adverse drug reactions in two swiss departments of internal medicine. *Br. J. Clin. Pharmacol.* **49**, 158–167 (2000).
13. Dugger, S. A., Platt, A. & Goldstein, D. B. Drug development in the era of precision medicine. *Nat. Rev. Drug Discov.* **17**, 183–196 (2018).
14. Tan, Z. W., Tee, W. V., Guarnera, E., Booth, L. & Berezovsky, I. N. AlloMAPS: allosteric mutation analysis and polymorphism of signaling database. *Nucleic Acids Res* **47**, D265–D270 (2019).
15. Chen, Y. -f & Fu, L. -w Mechanisms of acquired resistance to tyrosine kinase inhibitors. *Acta Pharmaceutica Sin. B* **1**, 197–207 (2011).
16. Attia, T. Z. et al. Effect of cytochrome P450 2C19 and 2C9 amino acid residues 72 and 241 on metabolism of tricyclic antidepressant drugs. *Chem. Pharm. Bull.* **62**, 176–181 (2014).
17. Zhou, S. F., Liu, J. P. & Chowbay, B. Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metab. Rev.* **41**, 89–295 (2009).
18. Kumar, S., Dudley, J. T., Filipski, A. & Liu, L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet.* **27**, 377–386 (2011).
19. Konc, J., Skrlj, B., Erzen, N., Kunej, T. & Janezic, D. GenProBiS: web server for mapping of sequence variants to protein binding sites. *Nucleic Acids Res.* **45**, W253–W259 (2017).
20. Ofoegbu, T. C. et al. PhyreRisk: a dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants. *J. Mol. Biol.* **431**, 2460–2466 (2019).
21. Iqbal, S. et al. MISCAST: missense variant to protein structure analysis web suite. *Nucleic Acids Res.* **48**, W132–W139 (2020).
22. Laskowski, R. A., Stephenson, J. D., Sillitoe, I., Orengo, C. A. & Thornton, J. M. VarSite: disease variants and protein structure. *Protein Sci.* **29**, 111–119 (2020).
23. Meyer, M. J. et al. mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum. Mutat.* **37**, 447–456 (2016).
24. Jubb, H. C., Saini, H. K., Verdonk, M. L. & Forbes, S. A. COSMIC-3D provides structural perspectives on cancer genetics for drug discovery. *Nat. Genet.* **50**, 1200–1202 (2018).
25. Kim, P., Zhao, J., Lu, P. & Zhao, Z. mutLBSgeneDB: mutated ligand binding site gene DataBase. *Nucleic Acids Res.* **45**, D256–D263 (2017).
26. Yan, C. et al. Impact of germline and somatic missense variations on drug binding sites. *Pharmacogenomics J.* **17**, 128–136 (2017).
27. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
28. Jafari, R. et al. The cellular thermal shift assay for evaluating drug target interactions in cells. *Nat. Protoc.* **9**, 2100–2122 (2014).
29. Berman, H. M. et al. The protein data bank. *Nucleic acids Res.* **28**, 235–242 (2000).
30. Pieper, U. et al. ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **39**, D465–D474 (2011).
31. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
32. Wishart, D. S. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36**, D901–D906 (2008).
33. Chan, S. H. et al. Analysis of clinically relevant variants from ancestrally diverse Asian genomes. *Nat. Commun.* **13**, 6694 (2022).
34. Guengerich, F. P. *In Cytochrome P450: Structure, Mechanism, and Biochemistry* (ed Paul R. Ortiz de Montellano) 523-785 (Springer International Publishing, 2015).
35. Takahashi, M. et al. Functional characterization of 21 CYP2C19 allelic variants for clopidogrel 2-oxidation. *Pharmacogenomics J.* **15**, 26–32 (2015).
36. Whirl-Carrillo, M. et al. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin. Pharm. Ther.* **110**, 563–572 (2021).
37. Marcath, L. A., Pasternak, A. L. & Hertz, D. L. Challenges to assess substrate-dependent allelic effects in CYP450 enzymes and the potential clinical implications. *Pharmacogenomics J.* **19**, 501–515 (2019).
38. Lim, J. S. et al. Association of CYP2C19*2 and associated haplotypes with lower norendoxifen concentrations in tamoxifen-treated Asian breast cancer patients. *Br. J. Clin. Pharmacol.* **81**, 1142–1152 (2016).
39. Wang, A., Stout, C. D., Zhang, Q. & Johnson, E. F. Contributions of ionic interactions and protein dynamics to cytochrome P450 2D6 (CYP2D6) substrate and inhibitor binding. *J. Biol. Chem.* **290**, 5092–5104 (2015).
40. Saito, Y. et al. A pyrrolo-pyrimidine derivative targets human primary AML stem cells in vivo. *Sci. Transl. Med.* **5**, 181ra152 (2013).
41. Wróbel, A., Arciszewska, K., Maliszewski, D. & Drozdowska, D. Trimethoprim and other nonclassical antifolates an excellent template for searching modifications of dihydrofolate reductase enzyme inhibitors. *J. Antibiotics* **73**, 5–27 (2020).
42. Limviphuvadh, V. et al. Discovering novel SNPs that are correlated with patient outcome in a Singaporean cancer patient cohort treated with gemcitabine-based chemotherapy. *BMC Cancer* **18**, 555 (2018).
43. Heersche, N. et al. Clinical implications of germline variations for treatment outcome and drug resistance for small molecule kinase inhibitors in patients with non-small cell lung cancer. *Drug Resist Updat* **62**, 100832 (2022).
44. Keshava, N. et al. Defining subpopulations of differential drug response to reveal novel target populations. *npj Syst. Biol. Appl.* **5**, 36 (2019).
45. Chong, C. S., Limviphuvadh, V. & Maurer-Stroh, S. Global spectrum of population-specific common missense variation in cytochrome P450 pharmacogenes. *Hum. Mutat.* **42**, 1107–1123 (2021).
46. Li, B. et al. In silico comparative characterization of pharmacogenomic missense variants. *BMC Genomics* **15**, S4 (2014).
47. Simonetti, L. et al. CYP21A2 mutation update: Comprehensive analysis of databases and published genetic variants. *Hum. Mutat.* **39**, 5–22 (2018).
48. Muroi, Y. et al. Functional characterization of wild-type and 49 CYP2D6 allelic variants for N-desmethyltamoxifen 4-hydroxylation activity. *Drug Metab. Pharmacokinet.* **29**, 360–366 (2014).
49. Saravanakumar, A., Sadighi, A., Ryu, R. & Akhlaghi, F. Physicochemical properties, biotransformation, and transport pathways of established and newly approved medications: a systematic review of the top 200 most prescribed drugs vs. the FDA-approved drugs between 2005 and 2016. *Clin. Pharmacokinet.* **58**, 1281–1294 (2019).
50. Genetics of Exfoliation Syndrome, P. et al. Association of rare CYP39A1 variants with exfoliation syndrome involving the anterior chamber of the eye. *JAMA* **325**, 753–764 (2021).
51. Gao, H. et al. The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
52. Stephenson, J. D., Laskowski, R. A., Nightingale, A., Hurles, M. E. & Thornton, J. M. VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics* **35**, 4854–4856 (2019).
53. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
54. Laskowski, R. A., Jablonska, J., Pravda, L., Varekova, R. S. & Thornton, J. M. PDBsum: structural summaries of PDB entries. *Protein Sci.* **27**, 129–134 (2018).
55. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet* **46**, 310–315 (2014).

56. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet* **99**, 877–885 (2016).
57. Pires, D. E., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* **42**, W314–W319 (2014).
58. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).
59. Mackman, R. et al. Active-site topologies of human CYP2D6 and its aspartate-301 → glutamate, asparagine, and glycine mutants. *Arch. Biochem. Biophys.* **331**, 134–140 (1996).
60. Hamid, M. et al. A deadly prescription: combination of methotrexate and trimethoprim-sulfamethoxazole. *J. Commun. Hosp. Intern. Med. Perspect.* **8**, 149–151 (2018).
61. Gerek, N. Z. et al. Evolutionary diagnosis of non-synonymous variants involved in differential drug response. *BMC Med. Genomics* **8**, S6 (2015).
62. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet* **50**, 1161–1170 (2018).
63. Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M. & Lauschke, V. M. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics J.* **19**, 115–126 (2019).
64. Mahmood, K. et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum. Genomics* **11**, 10 (2017).
65. Miosge, L. A. et al. Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci. USA* **112**, E5189–E5198 (2015).
66. Park, U. C. et al. Pharmacogenetic associations with long-term response to anti-vascular endothelial growth factor treatment in neovascular AMD patients. *Mol. Vis.* **20**, 1680–1694 (2014).
67. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
68. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
69. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
70. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
71. Varadi, M. et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
72. Dana, J. M. et al. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).
73. Krissinel, E. Enhanced fold recognition using efficient short fragment clustering. *J. Mol. Biochem* **1**, 76–85 (2012).
74. Paladin, L. et al. The Feature-Viewer: a visualization tool for positional annotations on a sequence. *Bioinformatics* **36**, 3244–3245 (2020).
75. Rose, A. S. et al. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* **34**, 3755-3758 (2018). https://doi.org/10.1093/bioinformatics/bty419.
76. Eisenhaber, B., Bork, P. & Eisenhaber, F. Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.* **292**, 741–758 (1999).
77. Lupas, A., Dyke, M. V. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
78. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–W37 (2011).
79. Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
80. Maurer-Stroh, S., Eisenhaber, B. & Eisenhaber, F. N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J. Mol. Biol.* **317**, 541–557 (2002).
81. Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
82. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971 (2007).
83. Maurer-Stroh, S. & Eisenhaber, F. Refinement and prediction of protein prenylation motifs. *Genome Biol.* **6**, R55 (2005).
84. Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.* **328**, 581–592 (2003).
85. Wootton, J. C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comp. Chem.* **18**, 269–285 (1994).
86. Stenson, P. D. et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet* **133**, 1–9 (2014).
87. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980–D985 (2014).
88. Gaedigk, A., Whirl-Carrillo, M., Pratt, V. M., Miller, N. A. & Klein, T. E. PharmVar and the landscape of pharmacogenetic resources. *Clin. Pharm. Ther.* **107**, 43–46 (2020).
89. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
90. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et. Biophysica Acta (BBA) - Protein Struct.* **405**, 442–451 (1975).
91. Krieger, E. & Vriend, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics* **30**, 2981–2982 (2014).
92. Stephenson Clarke, J. R. et al. Discovery of nanomolar-affinity pharmacological chaperones stabilizing the oncogenic p53 Mutant Y220C. *ACS Pharmacol. Transl. Sci.* **5**, 1169–1180 (2022).
93. Kong, L. R. et al. Targeting codon 158 p53-mutant cancers via the induction of p53 acetylation. *Nat. Commun.* **11**, 2086 (2020).
94. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
95. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
96. D.A. et al. AMBER 2018. *University of California*, San Francisco (University of Columbia, 2018).
97. Kannan, S. et al. Probing the binding mechanism of Mnk inhibitors by docking and molecular dynamics simulations. *Biochemistry* **54**, 32–46 (2015).

98. Kannan, S., Fox, S. J. & Verma, C. S. Exploring gatekeeper mutations in EGFR through computer simulations. *J. Chem. Inf. Model.* **59**, 2850–2858 (2019).
99. DeLano, W. L. Pymol: an open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **40**, 82–92 (2002).
100. Payton, M. E., Greenstone, M. H. & Schenker, N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *J. insect Sci.* **3**, 34 (2003).

Acknowledgements

This project was funded by the A*STAR Industry Alignment Fund (Pre-Positioning) (IAF-PP: H17/01/a0/007) and supported work by all authors. The SG10K Health project is also funded by the Industry Alignment Fund (Pre-Positioning) (IAF-PP: H17/01/a0/007), the project made use of participating study cohorts supported by the following funding sources: (1) HELIOS study by grants from a Strategic Initiative at Lee Kong Chian School of Medicine, the Singapore Ministry of Health (MOH) under its Singapore Translational Research Investigator Award (NMRC/STaR/0028/2017) and the IAF-PP: H18/01/a0/016, (2) GUSTO study by the Singapore National Research Foundation under its Translational and Clinical Research (TCR) Flagship Program and administered by the Singapore MOH's National Medical Research Council (NMRC) Singapore (NMRC/TCR/004-NUS/2008, NMRC/TCR/012-NUHS/2014) with additional funding support available through Agency of Science, Technology and Research (A*STAR) and IAF-PP: H17/01/a0/005, (3) SEED study by NMRC/CIRG/1417/2015, NMRC/CIRG/1488/2018, NMRC/OFLCG/004/2018), (4) MEC by individual research and clinical scientist award schemes from the Singapore National Medical Research Council (NMRC, including MOH-000271-00) and the Singapore Biomedical Research Council (BMRC), the Singapore Ministry of Health (MOH), the National University of Singapore (NUS) and the Singapore National University Health System (NUHS), (5) PRISM cohort study by NMRC/CG/M006/2017_NHCS, NMRC/STaR/0011/2012, NMRC/STaR/0026/2015, Lee Foundation and Tanoto Foundation, (6) TTSH cohort study by NMRC/CG12AUG2017 and CGAug16M012. This research is also supported by the RIE2025 Human Health and Potential (HHP) Industry Alignment Fund Pre-Positioning Program (IAF-PP Grant No. H25J6a0007), administered by the Agency for Science, Technology and Research (A*STAR), Singapore (supported the work of C.S.C).

Author contributions

A.J.M. and D.K. developed annotations tables for SNPdrug3D. D.K. designed the webpage and implemented front-end and back-end of the database. C.S.C. developed CYPVarPred. J.K., M.B.O., W.Q., H.Y.C. and C.S.H.T. performed the CETSA assays and analysis. M.H., R.T.M., and N.B. provided analyses summarizing the characteristics of the SG10K Health cohort. C.C.K., Y.S.H. and Z.W. provided data and analysis for the CYP39A1 variants. S.K. and C.S.V. performed the p53 variant analysis. G.L.C. provided the PrimateAI-3D variant annotations. V.L., J.M.T.M., and P.T. provided feedback for variant analysis and manuscript writing. X.S., J.F.C., S.D., K.K.Y., K.P.L., L.L.G., N.K., J.G.E., C.Y.C., T.Y.W., M.L., R.D., W.K.L., J.N., J.C.C., E.S.T., P.T., R.T.M., M.H., N.B., C.S.V., and S.M.S. are part of the SG10K Health consortium that generated the SG10K Health cohort variation data. A.J.M., D.K., C.S.V., S.M.S., and C.S.C. performed the analysis and wrote the paper. Idea conception and supervision were provided by C.S.V. and S.M.S. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-73816-4.

Correspondence and requests for materials should be addressed to Chandra S. Verma or Sebastian Maurer-Stroh.

Peer review information *Nature Communications* thanks Oliviero Carugo, Sriram Chandrasekaran and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

[1]Atomistic Simulations and Design in Biology, Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [2]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia. [3]Australian Centre for Ecogenomics, The University of Queensland, Brisbane, QLD, Australia. [4]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, VIC, Australia. [5]Protein Sequence Analysis, Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [6]Experimental Drug Development Centre, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [7]Department of Chemistry, Southern University of Science and Technology, Shenzhen, China. [8]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. [9]NPM-Genomic Intelligence & Informatics Engine (NPM-GINIE), Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [10]Precision Health Research Singapore (PRECISE), Singapore, Singapore. [11]Garvan Institute of Medical Research, Sydney, NSW, Australia. [12]Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [13]Duke-NUS Medical School, Singapore, Singapore. [14]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore. [15]SingHealth Duke-NUS Institute of Precision Medicine, Singapore, Singapore. [16]SingHealth Duke-NUS Genomic Medicine Centre, Singapore, Singapore. [17]Department of Cardiology, National Heart Centre Singapore, Singapore, Singapore. [18]Centre for Precision and Genomic Medicine, Tan Tock Seng Hospital, Singapore, Singapore. [19]Institute for Human Development and Potential, Agency for Science, Technology and Research (A*STAR),

Singapore, Singapore. [20]Clinical Data Engagement, Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [21]Department of Biochemistry, National University of Singapore, Singapore, Singapore. [22]Department of Obstetrics & Gynaecology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [23]Folkhalsan Research Center, Helsinki, Finland. [24]Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland. [25]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [26]Ophthalmology & Visual Sciences Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [27]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [28]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [29]Beijing Visual Science and Translational Eye Research Institute (BERI), Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China. [30]Population and Global Health, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [31]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [32]Laboratory of DNA Methylation and Population Epigenetics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [33]Department of Endocrinology, Tan Tock Seng Hospital, Singapore, Singapore. [34]Laboratory of Complex Disease Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [35]Illumina Singapore, Singapore, Singapore. [36]Cancer & Stem Cell Biology Program, Duke-NUS Medical School, Singapore, Singapore. [37]Laboratory of Genome Variation Analytics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [38]Cancer Genetics Service, Division of Medical Oncology, National Cancer Centre, Singapore, Singapore. [39]Oncology Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [40]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [41]Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [42]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [43]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [44]Department of Biological Sciences, National University of Singapore, Singapore, Singapore. [45]Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [66]These authors contributed equally: Ashar J. Malik, Dimitar Kenanov, Cheng-Shoong Chong. ✉e-mail: chandra@a-star.edu.sg; sebastianms@a-star.edu.sg

SG10K_Health Consortium

Tin Aung[25,26], Claire Bellis[46,47], Nicolas Bertin[9], Jin Fang Chai[14], John C. Chambers[10,30,31], Miao Ling Chee[25], Ching-Yu Cheng[25,26,27,28], Paul Chung Pui Cheng[34], Wen Jie Chew[48], Calvin Woon Loong Chin[17,49], Yap Seng Chong[19,22], Stuart Alexander Cook[50,51,52], Paul Eillot[31,53], Johan G. Eriksson[19,22,23,24], Peter D. Gluckman[19,54], Liuh Ling Goh[18], Maxime Hebrard[9,10,11], Pritesh Rajesh Jain[30], Saumya S. Jamuar[15,16,55,56], Justin Jeyakani[9], Rodrigo Toro Jimenez[9], Neerja Karnani[19,20,21], Tat Hung Koh[46], Eng Sing Lee[53,57], Jimmy Lee[53,58], Yung Seng Lee[22,59], Khai Pang Leong[18], Hengtong Li[18,25], Zhihui Li[9], Chia Wei Lim[18], Tock Han Lim[53,60], Weng Khong Lim[15,36,37], Bitong Clarabelle Alexandrine Lin[46], Jianjun Liu[40,61], Marie Loh[30,31,32], Dorrain Low[30], Roberto-Tirado Magallanes[9], Sebastian Maurer-Stroh[5,44,45] ✉, Theresia Mina[30], Shiqi Mok[34], Hong Kiat Ng[30], Joanne Ngeow[30,38,39], Jack Ling Ow[9], Qingsheng Peng[25,26], Shyam Prabhakar[62], Chee Jian Pua[51], Elio Riboli[31,53], Charumathi Sabanayagam[25,26], Nilanjana Sadhu[30], Yee Yen Sia[61], Wey Ching Sim[18], Xueling Sim[14], E. Shyong Tai[10,14,40], Joanna Hui Juan Tan[9], Patrick Tan[10,13,42], Erwin Tantoso[63], Darwin Tay[30], Yik Ying Teo[14], Yih Chung Tham[25,27,28], Li-xian Grace Toh[18], Pi Kuang Tsai[18], Rob M. van Dam[14,64], Chandra Verma[1,43,44], Xiaoyan Wang[30], Tien Yin Wong[25,29], Can Can Xue[25], Chengxi Yang[51], Fabian Yap[13,65] & Khung Keong Yeo[13,15,17]

[46]Extract Platform, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [47]Centre for Genomics and Personalised Health, Genomics Research Centre, School of Biomedical Sciences, Queensland University of Technology, Brisbane, QLD, Australia. [48]Clinical Research & Innovation Office, Tan Tock Seng Hospital, Singapore, Singapore. [49]Cardiovascular Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [50]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore, Singapore. [51]National Heart Research Institute Singapore, National Heart Centre Singapore, Singapore, Singapore. [52]MRC Laboratory of Medical Sciences, London, UK. [53]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [54]Liggins Institute, University of Auckland, Auckland, New Zealand. [55]Genetics Service, Department of Paediatrics, KK Women's and Children's Hospital, Singapore, Singapore. [56]Paediatric Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [57]Clinical Research Unit, National Healthcare Group Polyclinics, Singapore, Singapore. [58]Research Division, Institute of Mental Health, Singapore, Singapore. [59]Khoo Teck Puat-National University Children's Medical ICute, National University Hospital, Singapore, Singapore. [60]National Healthcare Group Eye Institute, Tan Tock Seng Hospital, Singapore, Singapore. [61]Laboratory of Human Genomics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [62]Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [63]Data Management, Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore. [64]Departments of Exercise and Nutrition Sciences and Epidemiology, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA. [65]Department of Paediatrics, KK Women's and Children's Hospital, Singapore, Singapore.