Abstract

Many drugs are known to be ineffective for some patients, carrying certain non-synonymous single nucleotide variants (nsSNV). But understanding the biophysical rationales of nsSNVs' implications towards drug efficacy remains difficult. Recent advancements in both population-level next-generation sequencing (NGS) and high-resolution protein-drug co-crystal determination provide a way to address this challenge. In this study, we developed a supervised learning method referred as GenoDock to predict nsSNVs which may lead to protein-drug binding disruptions. Specifically, we collected the protein-drug complexes with high resolution structures available and mapped their somatic and germline nsSNVs onto the structures. According to whether the nsSNVs can impair the binding, they were further grouped into two classes and used as target labels. We integrated genomics, structural and physiochemical features from nsSNVs, protein structures and drug ligands and trained GenoDock to do the prediction. Cross-validation result shows that the GenoDock can effectively predict disruptive nsSNVs (with AUC=0.97). Drug resistance effect towards gefitinib by T790M mutation in EGFR gene was applied to validate the prediction of GenoDock as a case study. We make our GenoDock method publicly available as a web interface at http://genodock.molmovdb.org/.

Introduction

In recent years, the immense growth of both genetic variation [1] and protein structure datasets [2] which benefit from great advancement in related techniques has enabled us to study in depth the impact of genomic variants onto protein structures and functions [3]. People have taken great efforts to get the insights of how genetic variants cause various diseases at a population level in order to potentially enhance drug effectiveness in the era of personalized medicine [4-6]. Variant annotation tools such as SIFT, Polyphen-2, CADD, and GERP are some examples of such achievements, which mainly focus on sequence conservation within and across species to assign general impact of a non-synonymous single nucleotide variant (nsSNV) [7-10]. In general, studies for this purpose are usually limited due to the lack of variation data as well as the corresponding high resolution protein or protein-drug complex structures [11]. Recently, with more variant and structural data are available, many efforts have been made to relate genomic variants with protein crystal structures to better bridge the increasing gap between genomic variation and protein structure, and to better understand how certain protein function alterations origin from genomic variants [12-16].

Linking protein 3D structures and genomics, i.e. genetic diversity across large population, using computational models has been proved to be a powerful and innovative approach for precise medicine [16]. This association will help accumulate evidence as guidance towards clinical practice. Here, we choose protein-drug interactions as our primary focus. We aim to investigate the mechanism of how a nsSNV potentially perturbs the interaction between the associated protein and drug ligands [12, 17]. Studies have shown that many drugs are effective towards only a limited fraction of individuals due to different responses from patients to specific drugs [18-20]. One of the reasons of loss of efficacy for drugs arises from genetic variants that each patient carries [20, 21]. Thus, a patient's genetic-centric prescription may be a reasonable approach to address the problem of drug ineffectiveness since recent advances of sequencing

techniques make it more practical and affordable for high-throughput personal genomic analysis. Once personal carried genetic variants are identified, the focus can then be shifted to how single point alternation of protein residues caused by nsSNVs would influence drug efficacy. Thus, a well-constructed database that directly links genetic variants to reliable human drug-protein co-crystal structures is in great need. Also, there is a call for a systematic pipeline to accurately predict if a nsSNV of interest would destabilize protein-drug binding activity.

To embody this idea, we develop a pipeline, GenoDock, to bridge nsSNVs on a large population scale and protein-drug co-crystal structures in the study. Our primary focus is to investigate how and how likely a given variant would affect protein-ligand binding affinity. We first construct our database by mapping germline and somatic variants onto their associated protein residues and drug molecules present in that protein structure. We then examined the binding affinity change (ΔBA) between the native and mutated protein structures associated with each nsSNV in our database through molecular docking-based method. We grouped the variants based on whether they would lead to a positive shift in binding affinity ($\Delta BA > 0$) or not ($\Delta BA \leq 0$). The former class of SNVs is our main focus in this study due to their high potential to cause drug-resistance activity. Among different types of nsSNVs in our database, we find that the portion of nsSNVs that would cause a positive ΔBA increase from common (6%), rare (7%), passenger (9%) to driver (15%) groups. Next, we describe a novel supervised learning model based on random forest algorithm to predict the probability of a given nsSNV to destabilize protein-drug binding by integrating genomic, structural and physiochemical features from nsSNV annotations, protein structures and drug ligands. Finally, we present GenoDock program suite together with a web interface (http://genodock.molmovdb.org/), which can be used to rapidly and efficiently prioritize nsSNV candidates that disrupt protein-drug binding.

Results

GenoDock *database and toolkit*

Figure 1a shows our strategy to construct the database that is publicly available from our GenoDock website (http://genodock.molmovdb.org/). The database contains 10,283 non-synonymous SNVs (nsSNV) from 228 proteins in *Homo sapiens*, and 113 FDA-approved drug ligands, which have co-crystal structures with at least one protein. We screen all the human proteins with high resolution (<3.0Å) X-ray-solved protein PDB structures (https://www.rcsb.org/) [22] and keep these with at least one FDA approved drug ligand in the co-crystal structures. After removing the structural redundancy based on the result of sequence alignment, we map the germline nsSNVs from Exome Aggregation Consortium (ExAC) [23] and the somatic nsSNVs from The Cancer Genome Atlas (TCGA) dataset [24-26] to these 228 protein structures according to BioMart-derived human gene and transcript ID [27]. In total, we collected 8,565 nsSNVs in 166 PDB structures for ExAC germline variants, and 1,718 nsSNVs in 135 PDB structures for TCGA somatic mutations. The nsSNVs, protein structures, and drug ligands form SNV-Ligand-PDB 3-tuple entries in our database. For each SNV-Ligand-PDB entry, as visualized in Figure 1b, we use Modeller program suite [28] to generate the mutated structure using homology modelling. We then use Auto Dock Vina [29] to calculate the binding affinity score for wild type protein and the corresponding ligand (ΔG_{WT}) and that after the residue is mutated (ΔG_{MUT}) in order to get the score change (ΔBA) in kcal/mol ($\Delta BA = \Delta G_{MUT} - \Delta G_{WT}$). The ΔBA value set serve as the reference set for GenoDock program suite.

The binding affinity change between the native and the point mutation structure with their drug ligand is the target label that GenoDock aims to predict based on a random forest classifier. As shown in Figure 1c, we category ΔBA values for each SNV-Ligand-PDB entry into two classes: if ΔBA is positive, we tag it as "Class 1"; if ΔBA is non-positive, we tag it as "Class 2". A positive shift in binding affinity indicates that it requires less energy to break the binding between the protein and the ligand, and thus the point mutation plays a disruptive role that could potentially cause drug resistance if the protein serves as a drug target. We integrate selected

genomic, structural and physiochemical features of SNVs, PDBs, and ligands to train the classifier: SNV annotation features include allele frequency, SIFT [10], PolyPhen-2 [30], and GERP [8] score; ligand features include molecular weight, hydrogen-bond donor and acceptor count, rotatable bond count and polar surface area of the ligand; protein structure features include binding site, side chain polarity and volume change, and distance of the mutated residue from ligand (see 'Methods' for details of random forest model construction and feature selection; Figure 1, Figure 4 and Supplementary Figure 1 & 2).

Amino acid mutation landscape in **GenoDock** *dataset*

After the construction of GenoDock dataset, we then analyze the mutation landscape of TCGA somatic and ExAC germline variants in our dataset which provides us with the opportunity to analyze known amino acid changes and mutation trends that are under high selective constraints or potentially lead to human disease. As depicted in Figure 2a, the two most abundant mutations recorded in our GenoDock database are arginine to cysteine and arginine to histidine. This is within our expectation. First, arginine is the most frequently occurred amino acid among the somatic mutations and germline variants that can be mapped on to a PDB structure in our protein pool (14% in wildtype distribution, see Figure 2a); second, arginine to cysteine mutation is also found to be the most common mutation that cause human disease in disease-associated variant datasets such as Human Gene Mutation Database (HGMD), the Online Database of Mendelian Inheritance in Man (OMIM), and ClinVar [31-34]; third, arginine to histidine is also identified as a mutation signature that is very enriched in cancers. These observations are of similar landscape described in Szpiech et al. [35].

Analyzing the mutation landscape of our database is very useful for our following study of how a point mutation affects drug efficacy, which is further tailored to how side-chains interact with ligand differently before and after the replacement. This drives us to focus on certain physicochemical properties of side-chains such as volume change and polarity change between

the pool of wildtype residues with mutated ones (see 'Methods' for details). We observe that ~ 1/3 of somatic nsSNVs lead to point mutations from a charged amino acid residue to a polar one; whereas among the germline variants, the most frequently occurred mutations are between two hydrophobic amino acids (Supplementary Figure 3). Previous literature also shows that the cancer mutation signature, arginine to histidine mutation, can confer protein pH sensitivity to the mutant and thus alters protein function leading to diseases [35-37]. Due to the cancer-associated nature of the somatic mutations in our database, further bio-physical and biochemical studies on how a nsSNV might alter protein functions provides valuable insights towards cancer drug responses from patients.

Distributions of ΔBA in different groups of nsSNVs

With those ExAC germline nsSNVs in our dataset, our interest is to see whether there is a significant difference between the rare and the common nsSNV groups in terms of drug-binding destabilization. Rare and ultra-rare nsSNVs are in general interpreted as more likely to be deleterious than those common ones. The allele frequency values in population level studies also indicate varying degrees of constraint during natural selection. Similarly, we divide the TCGA somatic nsSNVs into highly deleterious driver nsSNVs and neutral passenger nsSNVs to investigate different impacts of the two groups on drug binding. Recognizing driver SNVs from a larger body of passenger nsSNVs remains a big challenge in cancer genomics [38] (see 'Methods' for details regarding common, rare, passenger and driver SNV tagging).

In Figure 2b, we visualize the distributions of binding affinity change for each group, especially for "Class 1" nsSNVs that positively shift ΔBA, which contribute to 6.0% and 8.9% of all nsSNVs in our ExAC and TCGA data source (Supplementary Figure 4). Though we do not observe a significant difference in ΔBA distributions between common and rare nsSNVs, when we bring together the top common and rare germline nsSNVs with positive ΔBA (the "outlier" region in the boxplot), top rare nsSNVs have a significantly higher ΔBA than those common

ones. It implies that rare nsSNVs pool contains more extremely deleterious samples in terms of disrupting drug-protein binding than those from common nsSNV pool (e.g. the top 50 group has *p-value = 1.4e-4 from two-sample Wilcoxon t-test*; Supplementary Figure 5). This observation is intuitively consistent with our expectation as rare variants tend to have greater impacts on protein stability as a result of higher selective constraints.

Based on remarkable efforts made in characterization of cancer genomes [24, 25, 39], people have validated the important roles of driver nsSNVs in driving cancer progression [40, 41]. These facts motivate us to probe the impacts of nsSNVs from driver genes on perturbing interactions between associated protein residues and drug ligands. Indeed, our analysis shows a significant difference between passenger and driver nsSNVs. Those cancer-associated driver nsSNVs tend to destabilize protein-drug binding to a bigger extent compared with neutral passenger ones (*p-value = 3.60e-4 from two-sample Wilcoxon test*). In Figure 2b, we also plot the percentage of nsSNVs that lead to a non-positive ΔBA ("Class 2") together with the percentage of nsSNVs that do not change the binding affinity upon point mutation ($\Delta BA = 0$). We find that the portion of nsSNVs that would cause a non-positive ΔBA decrease from common (94%), rare (93%), passenger (91%) to driver (85%) groups. This indicates that in the driver nsSNV group there is a heavier portion of variants that impair drug binding compared with the other groups. Next, we conduct further analysis to see more difference in Class 1 and Class 2 variants in terms of genomic, structural and physiochemical properties. Specific properties with different responses from the two classes of variants will serve as features in our later learning method to separate binding-disruptive nsSNVs from the rest.

Differential effects of features on drug-resistant and non-drug-resistant nsSNVs

The GenoDock project aims to provide a pipeline that could efficiently distinguish variants that destabilize protein drug binding activities ("Class 1") from the rest ("Class 2"). Genomic,

structural and physicochemical properties (features) of variants, proteins and ligands are playing important roles in discerning the two classes of variants. Thus we extract and define a list of features that discriminate the "Class 1" nsSNVs from those in "Class 2" and serve as training reference in our classifier (see 'Methods' for details on feature selection and construction). For each 'SNV-Ligand-PDB' in GenoDock database, we construct three groups of features (Figure 3; Supplementary Figure 6): SNV annotation features (Figure 3a); protein structure features (Figure 3b), and drug ligand features (Figure 3c) to see if these features are sensitive to differentiate the two classes of nsSNVs.

In Figure 3a, we use SIFT and Polyphen-2 scores to show whether the nsSNVs associated with protein residues are intra-species conserved across a population [7, 10, 30, 42]. A lower SIFT score indicates a greater chance that a nsSNV being "deleterious" due to high inter-species residue conservation and high selective constrains [43]. Similarly, a higher Polyphen-2 score denotes a greater likelihood that a nsSNV being "possibly damaging" [30]. We also employ GERP score to measure whether the point mutation is on inter-species conserved [44-46], indicated by a higher GERP score [8]. We observe nsSNVs in "Class 1" have a significantly lower mean SIFT score (mean = 0.101 and mean = 0.149, respectively) and a significantly higher Polyphen-2 score (mean = 0.665 and mean = 0.516, respectively) than those from "Class 2" (*p-value for SIFT is 1.21e-6 and p-value for Polyphen-2 is 2.20e-18; both from two-sample Wilcoxon test*), indicating that nsSNVs with a lower SIFT or a higher Polyphen-2 score are more likely to cause a positive shift on ΔBA. The median GERP scores for the two classes also differ significantly (*p-value = 0.0101 from two-sample Wilcoxon test*). nsSNVs that cause positive ΔBA are likely to be mapped onto more conserved regions on protein structure (mean = 3.32) than the other group (mean = 2.99).

In Figure 3b, we show the box plot distributions of the two classes of nsSNVs regarding protein structure features. Distance between mutated amino acid residue and drug molecule is

perhaps the most direct feature to tell whether a point mutation would be likely to affect ligand binding. We observe that more nsSNVs that impair binding activity are in the binding pocket (mean = 6.29Å) than the other class (mean = 19.8Å, *p-value = 1.27e-143 from two-sample Wilcoxon test*). If the distance is bigger than our threshold (8Å), the mutation is less likely to affect the protein and drug ligand binding due to the weaker van der Waals interaction. Another important physical property affecting drug binding is side-chain volume change between wildtype and mutated residue. Upon our definition of volume change index, we observe that nsSNVs which disrupt ligand binding are more likely to result from a decreased side chain volume (mean = -0.177, see "Methods" for definition of volume change index), whereas on average the nsSNVs that lead to a non-positive ΔBA have a bulkier side chain volume (mean = 0.0343; *p-value = 1.68e-20 from two-sample Wilcoxon test*). Side chain polarity change is another feature in context of ligand-protein interaction. For example, side chain polarity decreasing from a charged residue to a hydrophobic one may break the hydrogen bond network or salt bridge between the wild type residue to drug ligand (*see "Discussion" for detailed case analysis*) [47-50]. Here we observe the two groups of nsSNVs have a significant difference in this feature as well (*p-value = 0.0217 from two-sample Wilcoxon test*).

Figure 3c depicts the difference from the drug ligand in the co-crystal protein structure that nsSNVs are mapped to. In order to study nsSNVs' impacts towards protein-ligand binding, ligand properties are also an important part. We extract five features among various of physicochemical properties for each drug molecule in our database (Figure3a; Supplementary Figure 6). We observe that those nsSNVs with a positive ΔBA reside in a protein structure with a heavier drug ligand (mean = 361g/mol) than the other group (mean = 341g/mol), and this difference is significant (*p-value = 2.14e-3 from two-sample Wilcoxon test*). Also, we notice that the polar surface area of the drug ligands with a nsSNV that lead to positive ΔBA tend to be smaller (mean = 94.6Å^2), compared with the other group (mean = 105Å^2; *p-value = 5.13e-5 from two-*

sample Wilcoxon test). One reason may arise from the sensitivity of a heavier ligand and of a ligand with smaller polar surface area is higher in response to the side chain volume or polarity change upon point mutation.

These genomic, structural and physiochemical properties that act differently for "Class 1" SNVs and "Class 2" SNVs shown in Figure 3 provide good training feature candidates for our learning method to prioritize SNV candidates that lead to a positive protein ligand binding affinity change. We find SNV annotation scores including Polyphen-2, SIFT and GERP; ligand molecule properties such as polar surface area, and protein structural alteration including side-chain volume change are all promising input features to our GenoDock classification model present below.

Performance evaluation of GenoDock toolkit in classifying binding affinity change

In this study, we present GenoDock classifier to predict binding affinity score change upon point mutations based on docking calculations as gold-standard set for ΔBA, aiming to help with potential nsSNVs that cause ligand-binding disruption and drug resistance. We implemented a machine learning approach to achieve this purpose with additional steps integrated into our pipeline for evaluating our predictions. To make sure our evaluation towards GenoDock classifier is unbiased, we design a method which involves a cross-validation step to pick up the best performed model among a set of chosen learning methods; a grid-search-based model selection step to optimize the parameters for learning model construction, and an evaluation step using an independent test set isolated from the learning set (Supplementary Figure 7; see "Methods" for details). As we provide four independent models depending on information availability (SNV annotations only; SNV annotations + Structure; SNV annotations + Ligand; SNV annotations + Structure + Ligand), we apply the procedure above onto each model to make our pipeline a

uniform one. Our tryout for different learning methods shows that random forest classifier is the best one (Supplementary Figure 8; see "Methods" for model selection). During our preparation of training data, we tune the number of samples of "Class 1" (nsSNVs cause positive ΔBA) and "Class 2" (nsSNVs cause non-positive ΔBA) to be 1:1 in our training set to avoid potential bias from imbalanced sample volume of two classes, while keeping the original sample ratio of two classes unchanged in the test set. For the models in which only one of PDB structure or ligand molecule is present, we evaluate the classification performance with "Bind Site" feature included and excluded during the training process, separately. As depicted in Figure 4a, we test the classifier with default setting that all nsSNVs are mapped onto binding site residues for an upper limit of probability that the nsSNVs of interest to be disruptive towards ligand binding (in GenoDock web interface, users can also choose "Bind Site" to be "False" when nsSNV of interest is out of binding pocket). The area under the receiver-operator characteristic curve (AUC of ROC) for predictions of four models are 0.73 (SNV annotations only), 0.92 (SNV annotations + Structure), 0.96 (SNV annotations + Ligand), and 0.97 (SNV annotations + Structure + Ligand), respectively. If whether target nsSNVs are in binding pocket or not remain unknown, the performance of GenoDock is shown in Figure 4b, with "Bind Site" feature excluded during training and test process for "SNV + PDB" and "SNV + ligand" model. AUC values here for these two models become 0.76 and 0.79, respectively. After all, as we feed the GenoDock classifier with more and more features, the performance of predictions keeps improving: when input integrates all of the three feature groups, our method is able to identify most of the nsSNVs that lead to a positive shift towards binding affinity with an AUC of 0.97. Using the same learning pipeline, we compare the performance of GenoDock with some other nsSNV impact annotation tools from our model including SIFT, Polyphen-2 and GERP, together with another tool that is not used in our model, the Combined Annotation Dependent Depletion (CADD) [9]. GenoDock gives the highest AUC value among these tools since it is specifically developed for addressing the impact of nsSNVs on ligand-binding affinity change instead of a general

annotation towards potential benign or deleterious influences onto protein function (Supplementary Figure 9).

We then apply Gini distance to identify relevant importance of different features during the decision-making process, as shown in Figure 4c. The Gini distance helps visualize the relative importance of each feature in different models [51]. We observe that the relative importance of the repeating features such as the nsSNV annotations and binding site remain stable across models, revealing the robustness of our method. The relative importance across genomic and structural features under a uniform learning pipeline provides us a reasonable way to draw insights on how a nsSNV would make impacts towards ligand binding. Thus, we further apply C5 [52] decision tree algorithm to construct a knowledge model using significant features identified by Gini distance analysis in order to explain the classification result of GenoDock. The knowledge model also helps prioritize features that we should consider when determining whether a given nsSNV could cause ligand binding destabilization (see 'Methods' for decision tree construction; Supplementary Figure 10 and Figure 5a-5b). We observe that whether the protein residue associated with target nsSNV on binding site or not initiates the top branch of the tree, followed up by distance between mutated residue and ligand, side chain volume change, and GERP score etc., leading to the classification result of a given nsSNV. Whether an nsSNV is associated with a binding pocket residue is the most important factor to determine its effect on binding affinity.

Based on our performance evaluation results, we have shown that by integrating features from SNV annotations, protein structures and drug ligand properties, GenoDock can clearly identify nsSNVs that lead to a positive ΔBA shift from the rest candidates with high accuracy. From our C5 knowledge model, relative importance of different features provides some reasonable insights of how some of these variants contribute to protein-ligand binding disruptions. We then run GenoDock classifier on certain SNV-PDB-Ligand entries to validate the reliability of prediction based on established clinical experiment results. We also identify new

nsSNV candidates that will potentially impair protein-drug binding.

GenoDock helps identify known and unknown nsSNVs that disrupt protein-ligand binding

Figure 5a depicts the decision-making process that GenoDock reaches to the prediction that T790M mutation (rs55181378) from human epidermal growth factor receptor (EGFR; PDB ID: 2ity) is very likely to impair the binding between one of its tyrosine kinase inhibitors (TKIs), gefitinib, and the kinase domain (possibility of $\Delta BA > 0$ is 64%). Through molecular and clinical studies, people have shown that the resistance towards gefitinib arise from the substitution of a bulkier methionine residue for threonine at position 790 in the kinase domain [53-57]. Further studies on the EGFR-gefitinib co-crystal structure show that the larger methionine residue lead to steric hindrance of the aromatic moieties of gefitinib molecule, preventing the accessibility of gefitinib to the binding pocket of EGFR kinase domain [53, 54, 57, 58]. We show our knowledge model to visualize the decision making logic behind GenoDock prediction result in Figure 5a. From the top level, the mutated residue is mapped in the binding pocket of the kinase domain, and the side chain volume is increased by 1/3 from threonine to methionine, which may potentially block the interaction of the ligand to the binding pocket. Furthermore, the functional annotations of the nsSNV associated with T790M mutation indicate that this variant is highly likely to be deleterious and the mutated residue resides in a highly conserved region, which strengthens the confidence that this variant would impair the protein-ligand binding. Together with the next fact that the side chain polarity decreases from the polar threonine to the hydrophobic methionine, GenoDock classifies this nsSNV to be very likely to cause a positive shift towards binding affinity.

In figure 5b, we show the process of our knowledge model showing how GenoDock helps identify novel mutations that could potentially lead to drug resistance. Farnesyl diphosphate synthase (FPPS) is an important target for the bisphosphonate class of drugs such as zoledronate

(ZOL). ZOL targets FPPS as an immunomodulator which alters macrophages from a tumor-promoting to a tumor-killing phenotype [59-64]. ZOL is a highly hydrophilic binder to FPPS via electrostatic and hydrogen bond interactions [65]. We visualized the interaction between ZOL and FPPS (PDB ID: 4p0w) in Figure 5b, in which ZOL ligand is binding to ARG112A via a "salt bridge" between the positive charged guanidium with the negative charged sulfate group of ZOL. However, with the mutation R112H (rs155317993), this binding network no longer exists. GenoDock classifies this nsSNV as $\Delta BA > 0$ with a probability of 99.8%, followed by a similar decision-making pipeline discussed in the previous case. This prediction indicates a novel mutation that could very likely impair the inhibitor effectiveness. We validate this prediction using AutoDock Vina, which gives a positive binding affinity shift, 0.31kcal/mol. More biological functional assays can be performed in the future in addition to the computational validation.

GenoDock web interface

To make our pipeline accessible to the public, we provide a web interface, the GenoDock web server (http://genodock.molmovdb.org/). We tailored GenoDock into four individual models based on the accessibility of input features to broaden the application landscape of our tool, with different level of prediction accuracy. The users can import their sample data using our GenoDock graphic user interface with different feature set combination: SNV annotation info only; SNA annotation and PDB info; SNV annotation and ligand info; SNV annotation, PDB and ligand info. The predicted result will be feedback in form of a HTML webpage. The calculation page can be reached at http://genodock.molmovdb.org/calculation/0. Users can also download our open source python code for four GenoDock models through GitHub (https://github.com/gersteinlab/GenoDock_local) to run large scale inputs on local computers or on HPC clusters.

Discussion

In this study, we developed a high-throughput computational pipeline to bridge nsSNVs with their annotations from different sequencing datasets onto high resolution protein structures for downstream analysis; a highly sensitive classification model to prioritize nsSNV candidates that could potentially cause protein drug binding disruption based on integration of genomic annotations and structural properties, and a user-friendly GUI, the GenoDock server, that rapidly provides predictions of binding affinity change for nsSNVs of interest.

For the construction of GenoDock database, we employ nsSNVs from ExAC Consortium and TCGA project as our germline variant and somatic variants feed, respectively. From a pool of ~2.5M ExAC germline variants and ~1M Pan-Cancer somatic mutations, we successfully map ~10K nsSNVs onto ~0.3K human proteins binding with FDA-approved drug ligands which are solved as high resolution co-crystal structures. We identified 735 nsSNVs that lead to a positive shift towards binding affinity, present in 123 protein structures, covering 85 drug ligands (see "Additional File: table 1"). For prioritization of nsSNVs that would cause binding disruption, we demonstrate GenoDock is an efficient classifier with 0.97 AUC when all features are available. With investigations of relative feature importance, we provide reasonable insights of how genomic, structural and physiochemical features affect a given nsSNVs impacts towards the interaction dynamics of the associated protein residue with its surroundings, particularly, a drug ligand that binds to it.

While our approach can identify novel nsSNV candidates that potentially impair protein-drug binding with a rapid yet accurate manner, the method is still limited by two aspects. First, the lack of high-resolution co-crystal structures of proteins and their associated drug ligands. The unbalanced availability of structure data and variant data leads to only 1% of the SNVs mapped

from our SNV data source onto the protein-drug co-crystal structures. Fortunately, as protein characterization techniques such as NMR, electron microscopy and cryo-electron microscopy (cryo-EM) [66] advance, we can foresee that more and more highly reliable protein-drug structural data will be available. In addition, remarkable progress in putative 3D protein-drug interaction models based on homology modelling techniques may also potentially expand the structure pool [67, 68]. Together with tremendous progress in revealing the mutational landscape of human genomes via large-scale sequencing projects such as The UK 10,000 Project and the International Cancer Genomics Consortium, we will periodically include new SNV-Ligand-PDB entries into our classification pipeline for better prediction accuracy, and for nominations of additional novel nsSNV candidates that cause protein-drug binding disruptions. Second, our binding affinity change data is based on docking calculations at current stage, which limits the upper boundary of our prediction accuracy. So far, it is not practical to obtain or conduct experimental binding affinity change results between wild-type and mutant ligand-protein complex by ligand binding assays (LBA) [69] for each SNV-Ligand-PDB entry in our database. Therefore, we construct our gold-standard ΔBA reference set based on AutoDock Vina, which is well established and wildly used in pharmaceutical research projects. We further validate the consistence of the ΔBA results for each SNV-Ligand-PDB entry via AutoDock to get confidence towards the quality of our gold-standard set (Supplementary Figure 11). If we have LBA data later on, we plan to update the ΔBA reference values with experimental results under the same pipeline to further enhance the reliability of GenoDock predictions. Third, we fix the protein backbone while conducting docking calculations to avoid concerns and problems raised from protein flexibilities, which makes it hard to probe influence towards binding activities by protein motions or conformational changes.

We demonstrate that GenoDock is a useful tool to predict nsSNV candidates that could potentially disrupt protein-ligand binding activities, which could be further employed as a metric

to gain valuable mechanistic insights into drug resistance activities and to design personalized disease therapies for individual patients accordingly. Though our random forest classifier is very successful at predicting binding affinity change associated with an nsSNV and at estimating the relative importance of our features, the classifier is an ensemble of 10000 decision trees, which is vague to learn the decision making process explicitly. To further explain the predictions by GenoDock, we apply C5 [52] decision tree algorithms to generate our knowledge model, which is a consensus tree depicting an universal decision-making process. Compared with C4.5 [70] and ID3 [71] algorithms, C5 is faster and tends to generate simple and clean trees. C5 also reduces overfitting and tends to use high relevant features affecting the ligand protein binding process with reduced-error pruning [52]. Though this decision tree cannot guarantee to identify global optimum, the decision rules extracted from our C5 tree truthfully reflect the biophysical knowledge and mechanistic insights of binding affinity changes predicted by GenoDock.

GenoDock framework integrates genomic, structural, and physicochemical features for predictions of nsSNV impacts towards drug response. Particularly, to cater the fast growing variant and structural data, GenoDock is an efficient and reliable toolkit to prioritize and filter variants into a subset of highly promising candidates for downstream analysis, for example, drug resistance studies of a target system. We believe that GenoDock will continuously help to better understand and to predict the impacts of variants as more datasets, advanced molecular docking software, and LBS experimental data being used into our method.

Methods

GenoDock Database preparation

Germline variants were collected from Exome Aggregation Consortium(ExAC) release 1[23] (download source: ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/). Somatic variants came from The Cancer Genome Atlas (TCGA) network (http://cancergenome.nih.gov; download source: http://portal/gdc.cancer.gov/repository). "Simple Nucleotide Variation", "Masked Somatic Mutation" and "MuTect2 Variant Aggregation and Masking" were served as filters for "Data Category", "Data Type", and "Workflow Type", respectively. The list of FDA approved drug ligands was directly obtained from DrugBank [72]. Human protein PDB structures with a resolution higher than 3.0 Å were downloaded from the Protein Data Bank (https://www.rcsb.org/) [22]. A careful curation to filter out PDB that contains FDA approved drug molecules was conducted. The mapping of the variants from both the ExAC and TCGA datasets to the curated co-crystal PDB structures was done using a modified version of a previously published method [17] (See Supplementary Method for detailed steps of mapping SNVs onto PDB structures).

Mutant structure and binding affinity change calculation

For each "SNV-Ligand-PDB" entry recorded in our database, we generated a mutant structure associated with that nsSNV based on homology modelling via Modeller (ver. 9.18) [28] using the mutant sequence and the native protein structure. During the modelling process, adjustments were made to the target residue under stereo-chemical and homology-derived restraints, followed by a minimization step of the restraints to deliver the final mutant structure. In this project, 10,283 mutant PDB structures were generated in total.

For each native-mutated protein structure pair, we used AutoDock Vina [29] to evaluate the drug ligand binding affinity change: $\Delta BA = \Delta G(MUT) - \Delta G(WT)$, in kcal/mol, where $\Delta G(MUT)$ and $\Delta G(WT)$ are binding affinity of the mutated and native protein-drug complex evaluated by AutoDock Vina, respectively. During the calculation, we fixed the protein structure to avoid concerns from protein flexibility. "Local optimization" was applied for ligand binding

model, and "Vina score" was set as the scoring function. Due to the lack of experimental LBS data, we validated the calculations of Vina by applying the same procedure with AutoDock Tools (ver. 6.2.6) [73] to check the consistency of the two methods. If for a given structure pair, ΔBA values calculated by two scoring methods were of the same sign (both positive, indicating both tools assigned a drug binding disruptive role to the SNV; or both non-positive), then we regard the result as consistent. The two methods achieved a consistency of 84%. Also, the two sets of results from Vina and AutoDock Tools reached a Pearson product-moment correlation (PMCC) of 0.89 (Supplementary Figure 11), indicating a strong consistency.

Features extraction and construction for machine learning method

SNV annotation features including SIFT score, Polyphen-2 score for somatic and germline nsSNVs in our study are directly extracted from the "INFO" column of VCF files from ExAC consortium and TCGA project. GERP scores were retrieved directly from Sidow lab (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html) [8].

Ligand features including molecular weight, H-bond donor and acceptor count, rotatable bond count, and polar surface area for each drug molecule in our database are extracted from PubChem database [74].

We construct structural and physicochemical features as following. Amino acid side chain volume change index is defined as $\Delta V_{index} = log_2(\frac{V_{MUT}}{V_{WT}})$, where V_{MUT} and V_{WT} stand for van der Waals volume [75] of mutant and wildtype protein residue, respectively. For each amino acid, we assign a polarity index. Positive charged including ARG and LYS has an index of 1; polar residues including GLN, ASN, HIS, SER, THR, and TYR has an index of 0.5; hydrophobic residues including ALA, ILE, LEU, MET, PHE, VAL, PRO, and GLY has an index of 0; negative

charged residues including GLU and ASP has an index of -1. Amino acid side chain polarity change index is defined as $\Delta\text{polarity} = \text{polarity(mutant)} - \text{polarity(wildtype)}$. The distance between a protein residue to a ligand is defined as the shortest distance of a heavy atom of that residue to a heavy atom of the associated ligand. If a residue has a distance less than 8Å from the target ligand in the co-crystal structure, we consider that this residue is in the binding pocket, which forms the "binding site" feature. For the other two models where only one of drug ligand or protein structure is available, we assign the mutated residue to be in the binding pocket by default (SNV annotation + PDB), or we assume that the nsSNV will be mapped onto binding site once the co-crystal complex structure is available (SNV annotation + Ligand). In this way, we are able to predict the maximal probability of the target nsSNV to be ligand-binding disruptive. Users are also free to choose "binding side" to "OFF" if they want the prediction for the protein residues of associated variants are not in binding sites.

Training, testing, and evaluating the performance of machine learning method

GenoDock dataset is separated into training set (70%) and test set (30%) in a random manner. We also prepare a validation set for specific case studies so that samples of interest are separated from the training and testing pipeline. To avoid potential bias raised from imbalanced composition of two classes of samples in our dataset (735 entries from "Class 1"; 9458 entries from "Class 2"), we count the number of samples from "Class 1" ($\Delta BA > 0$) and randomly select equal number of samples from "Class 2" ($\Delta BA \leq 0$) to make up the balanced training set. Scikit-learn package [76] is used for learning model development. We test classification methods including Lasso Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). We train each learning model through a 10-fold grid-search cross-validation process. For each training, the rest 30% data is tested for performance

evaluation. Based on the AUC values, RF has the highest AUC among all methods (Supplementary Figure 8). Compared with RF, LR and SVM heavily rely on assumptions that target dataset have linearly separable patterns by line or hyperplane; GBDT requires strict parameter tuning in order to perform well. In addition, we pick up RF because we can easier gain mechanistic insights based on feature significance along the decision-making process to better explain how nsSNVs affect ligand-protein binding activity. Feature selection is performed by evaluation of AUC for each feature respectively. If the selection power of a feature is near or worse than random selection, we remove it from our feature pool (e.g. allele frequency). With the same procedure, we trained and optimized a random forest model for each of the four feature combinations (SNV annotations only; SNV annotations + Structure; SNV annotations + Ligand; SNV annotations + Structure + Ligand) for GenoDock. All source code and scripts are free to download at https://github.com/gersteinlab/GenoDock_local.

Construction of knowledge model to explain GenoDock prediction result

C5 decision tree is generated using "C50" package in R to explain GenoDock predictions. We selected highly ranked features (side chain polarity and volume change, GERP, distance between mutated residue and drug ligand, polar surface area of ligand, and bind site) based on Gini distance metric to construct the tree.

Protein-ligand complex visualization

All figures regarding protein-ligand complex are generated by the PyMOL molecular graphics system, Version 2.0 Schrödinger, LLC. [77]

References

[1] Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111:E455-64.

[2] Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 2015;43:D345-56.

[3] Sethi A, Clarke D, Chen J, Kumar S, Galeev TR, Regan L, et al. Reads meet rotamers: structural biology in the age of deep sequencing. Curr Opin Struct Biol. 2015;35:125-34.

[4] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372:793-5.

[5] Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol. 2001;19:491-6.

[6] Laing RE, Hess P, Shen Y, Wang J, Hu SX. The role and impact of SNPs in pharmacogenomics and personalized medicine. Curr Drug Metab. 2011;12:460-86.

[7] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.

[8] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.

[9] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310-5.

[10] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073-81.

[11] Glusman G, Rose PW, Prlic A, Dougherty J, Duarte JM, Hoffman AS, et al. Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med. 2017;9:113.
[12] Clarke D, Sethi A, Li S, Kumar S, Chang RWF, Chen J, et al. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. Structure. 2016;24:826-37.
[13] Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112:E5486-95.
[14] Meyer MJ, Lapcevic R, Romero AE, Yoon M, Das J, Beltran JF, et al. mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat. 2016;37:447-56.
[15] Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet. 2016;48:827-37.
[16] Meyer MJ, Beltran JF, Liang S, Fragoza R, Rumack A, Liang J, et al. Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods. 2018.
[17] Kumar S, Clarke D, Gerstein M. Localized structural frustration for evaluating the impact of sequence variants. Nucleic Acids Res. 2016;44:10062-73.
[18] Meyer UA, Zanger UM, Schwab M. Omics and drug response. Annu Rev Pharmacol Toxicol. 2013;53:475-502.
[19] Spear BB, Heath-Chiozzi M, Huff J. Clinical application of pharmacogenetics. Trends Mol Med. 2001;7:201-4.
[20] Wilkinson GR. Drug metabolism and variability among patients in drug response. N Engl J Med. 2005;352:2211-21.
[21] Madian AG, Wheeler HE, Jones RB, Dolan ME. Relating human genetic variation to variation in drug responses. Trends Genet. 2012;28:487-95.
[22] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235-42.
[23] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis

of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285-91.
[24] Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061-8.
[25] Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519-25.
[26] Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113-20.
[27] Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford). 2011;2011:bar049.
[28] Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Protein Sci. 2016;86:291-2937.
[29] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455-61.
[30] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248-9.
[31] Peterson TA, Doughty E, Kann MG. Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. Journal of Molecular Biology. 2013;425:4047-63.
[32] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514-7.
[33] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980-5.
[34] Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum

Genet. 2014;133:1-9.
[35] Szpiech ZA, Strauli NB, White KA, Ruiz DG, Jacobson MP, Barber DL, et al. Prominent features of the amino acid mutation landscape in cancer. PLoS One. 2017;12:e0183273.
[36] Reichold M, Zdebik AA, Lieberer E, Rapedius M, Schmidt K, Bandulik S, et al. KCNJ10 gene mutations causing EAST syndrome (epilepsy, ataxia, sensorineural deafness, and tubulopathy) disrupt channel function. Proc Natl Acad Sci U S A. 2010;107:14490-5.
[37] Zhang Z, Miteva MA, Wang L, Alexov E. Analyzing effects of naturally occurring missense mutations. Comput Math Methods Med. 2012;2012:805827.
[38] Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. J Mol Biol. 2013;425:3919-36.
[39] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39:D945-50.
[40] Hong MK, Macintyre G, Wedge DC, Van Loo P, Patel K, Lunke S, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. Nat Commun. 2015;6:6605.
[41] Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. 2014;6:5.
[42] Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88:440-9.
[43] Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7:61-80.
[44] Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56-65.
[45] Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative

annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342:1235587.
[46] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337:64-9.
[47] Boccuto L, Aoki K, Flanagan-Steet H, Chen CF, Fan X, Bartel F, et al. A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. Hum Mol Genet. 2014;23:418-33.
[48] Doss CG, Nagasundaram N. Investigating the structural impacts of I64T and P311S mutations in APE1-DNA complex: a molecular dynamics approach. PLoS One. 2012;7:e31677.
[49] Kumar A, Rajendran V, Sethumadhavan R, Purohit R. Molecular dynamic simulation reveals damaging impact of RAC1 F28L mutation in the switch I region. PLoS One. 2013;8:e77453.
[50] Zhang Z, Norris J, Kalscheuer V, Wood T, Wang L, Schwartz C, et al. A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome. Hum Mol Genet. 2013;22:3789-97.
[51] Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;10:213.
[52] Pandya RPaJ. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. International Journal of Computer Applications. 2015;117:18-21.
[53] Balak MN, Gong Y, Riely GJ, Somwar R, Li AR, Zakowski MF, et al. Novel D761Y and common secondary T790M mutations in epidermal growth factor receptor-mutant lung adenocarcinomas with acquired resistance to kinase inhibitors. Clin Cancer Res. 2006;12:6494-501.
[54] Kobayashi S, Boggon TJ, Dayaram T, Janne PA, Kocher O, Meyerson M, et al.

EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med. 2005;352:786-92.
[55] Kosaka T, Yatabe Y, Endoh H, Yoshida K, Hida T, Tsuboi M, et al. Analysis of epidermal growth factor receptor gene mutation in patients with non-small cell lung cancer and acquired resistance to gefitinib. Clin Cancer Res. 2006;12:5764-9.
[56] Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, Zakowski MF, et al. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med. 2005;2:e73.
[57] Janne PA. Challenges of detecting EGFR T790M in gefitinib/erlotinib-resistant tumours. Lung Cancer. 2008;60 Suppl 2:S3-9.
[58] Daub H, Specht K, Ullrich A. Strategies to overcome resistance to targeted protein kinase inhibitors. Nat Rev Drug Discov. 2004;3:1001-10.
[59] Coscia M, Quaglino E, Iezzi M, Curcio C, Pantaleoni F, Riganti C, et al. Zoledronic acid repolarizes tumour-associated macrophages and inhibits mammary carcinogenesis by targeting the mevalonate pathway. J Cell Mol Med. 2010;14:2803-15.
[60] Kunzmann V, Bauer E, Wilhelm M. Gamma/delta T-cell stimulation by pamidronate. N Engl J Med. 1999;340:737-8.
[61] Martin MB, Grimley JS, Lewis JC, Heath HT, 3rd, Bailey BN, Kendrick H, et al. Bisphosphonates inhibit the growth of Trypanosoma brucei, Trypanosoma cruzi, Leishmania donovani, Toxoplasma gondii, and Plasmodium falciparum: a potential route to chemotherapy. J Med Chem. 2001;44:909-16.
[62] Russell RG. Bisphosphonates: the first 40 years. Bone. 2011;49:2-19.
[63] Shipman CM, Croucher PI, Russell RG, Helfrich MH, Rogers MJ. The bisphosphonate incadronate (YM175) causes apoptosis of human myeloma cells in vitro by inhibiting the mevalonate pathway. Cancer Res. 1998;58:5294-7.
[64] Wood J, Bonjean K, Ruetz S, Bellahcene A, Devy L, Foidart JM, et al. Novel antiangiogenic effects of the bisphosphonate compound zoledronic acid. J Pharmacol Exp Ther. 2002;302:1055-61.
[65] Liu YL, Lindert S, Zhu W, Wang K, McCammon JA, Oldfield E. Taxodione and

arenarone inhibit farnesyl diphosphate synthase by binding to the isopentenyl diphosphate site. Proc Natl Acad Sci U S A. 2014;111:E2530-9.
[66] Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci. 2015;40:49-57.
[67] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One. 2011;6:e28766.
[68] Zhan Y, Guo S. Three-dimensional (3D) structure prediction and function analysis of the chitin-binding domain 3 protein HD73_3189 from Bacillus thuringiensis HD73. Biomed Mater Eng. 2015;26 Suppl 1:S2019-24.
[69] Benore M. Response to review of Fundamental Laboratory Approaches for Biochemistry and Biotechnology. Biochem Mol Biol Educ. 2010;38:64.
[70] Quinlan JR. C4.5: programs for machine learning: Morgan Kaufmann Publishers Inc.; 1993.
[71] J.R.Quinlan. Induction of Decision Trees. MACH LEARN. 1986;1:81--106.
[72] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074-D82.
[73] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30:2785-91.
[74] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016;44:D1202-13.
[75] Darby NJ, Creighton TE. Dissecting the disulphide-coupled folding pathway of bovine pancreatic trypsin inhibitor. Forming the first disulphide bonds in analogues of the reduced protein. J Mol Biol. 1993;232:873-96.
[76] Pedregosa FaV, G. and Gramfort, A. and Michel, V., and Thirion BaG, O. and Blondel, M. and Prettenhofer, P., and Weiss RaD, V. and Vanderplas, J. and Passos, A. and, Cournapeau DaB, M. and Perrot, M. and Duchesnay, E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825--30.

[77] The PyMOL Molecular Graphics System. Schrodinger, LLC.

Main Figure Captions and main Figures

Figure 1. Framework of the GenoDock Project – from dataset preparation to model construction.

(a) A flowchart for collecting, cleaning and processing raw data to construct GenoDock database from the protein structure data source (RCSB PDB), SNV data source (ExAC and TCGA), and

drug ligand data source (PubChem Compound).

(b) Illustration of protein-ligand binding affinity change upon point mutation. In this case, the MET on chain A resides in the catalytic domain of human phosphodiesterase 4B (PDB ID: 1xos; Ligand ID: VIA, sidelnafil) and is mutated to CYS by an nsSNV (rs66368865). The uncharged CYS demonstrates weaker binding to the ligand, indicated by a positive shift of binding affinity change (0.07, by AutoDock Vina).

(c) Construction of the random forest model to predict the direction of protein-ligand binding affinity change ($\Delta BA > 0 \; or \; \Delta BA \leq 0$). Several SNV annotation features (i.e. SIFT, GERP, Polyphen-2), ligand features (i.e. molecular weight, hydrogen bond donor/acceptor count), and structural features (i.e. binding site, side chain volume and polarity change) are combined to predict the direction of protein-ligand binding affinity change.

Figure 2. Heat map for amino acid mutation landscape and boxplot of ligand binding affinity changes for different types of SNVs in GenoDock

(a) Heat map for amino acid mutation landscape in GenoDock database. X-axis and y-axis refers to types of mutated amino acids and wild type amino acids, respectively. Different counts for each mutation pair is colored from white to cyan. Percentage distribution in wildtype and mutated amino acid pools are shown on top the heat map in green and purple, respectively. In the heat map, the two most abundant mutation pairs are arginine to cysteine and arginine to histidine, which are referred as "mutation signatures" in previous literatures.

(b) An overall comparison of common, rare, passenger and driver SNVs in terms of binding affinity change from GenoDock data source. nsSNVs that cause $\Delta BA > 0$ are plotted in order to compare the extent of destabilization towards ligand binding activities by each nsSNV group. The mean values for those SNVs leading to ligand-binding disruption for common, rare, passenger, and driver SNVs from ExAC and TCGA dataset are 0.117kcal/mol, 0.129 kcal/mol, 0.159 kcal/mol, and 0.236 kcal/mol, respectively. The difference in common and rare SNVs from ExAC dataset is not significant; the difference of passenger and driver SNVs from TCGA is significantly different, with a p-value of 3.60e-4 from two-sample Wilcoxon test, where driver nsSNVs have a bigger extent in disrupting ligand binding compared with other groups. The green-dot line and

pink-dot line in the figure show the percentage of SNVs from each group that lead to non-positive shift of binding affinity($\Delta BA \geq 0$; 94%, 93%, 91%, 85%, respectively), and those that do not change the binding affinity ($\Delta BA = 0$; 88%, 87%, 87%, 77%, respectively). It is clear that cancer driver nsSNVs have a greater probability to result in a positive binding affinity change compared with the other three groups .

Fig. 3. Boxplot distribution between "Class 1" nsSNVs (positive binding affinity shift) and "Class 2" nsSNVs (non-positive binding affinity shift) regarding different features groups:

(a) PolyPhen-2, SIFT and GERP score as SNV annotation features. We observe that Polyphen-2, SIFT, and GERP scores for the two groups of SNVs are all significantly different with p-values smaller than 0.05 from two-sample Wilcoxon tests. nsSNVs that disrupt ligand protein binding have a higher mean Polyphen-2 score (mean Polyphen-2 value: 0.665 and 0.516 for Class 1 and Class 2, respectively) and a lower SIFT score (mean SIFT value: 0.101 and 0.149 for Class 1 and Class 2, respectively), both indicating a more deleterious role of disruptive nsSNVs on protein

function. In terms of GERP score, nsSNVs lead to positive binding affinity change are more likely to be associated with protein residues from more conserved regions, indicating by a higher mean GERP score (mean GERP value: 3.32 and 2.99 for “Class 1” and “Class 2”, respectively).

(b) Side-chain volume and polarity change as protein structure features; distance between ligand and mutated residue when co-crystal structure is present. Amino acid side chain volume and polarity change before and after mutation will directly affect interaction of protein residue with ligand. We observe that the mean value of both side chain volume and polarity are statistically significant. On average, nsSNVs that destabilize ligand binding have decreased side chain volumes compared with the other class of ns SNVs (mean volume change index: -0.177 and 0.0343 for “Class 1” and “Class 2”, respectively). For side chain polarity change, there is also a significant difference between the two classes of nsSNVs (mean polarity change index: 0.0224 and 0.0856 for “Class 1” and “Class 2”, respectively). When protein-drug co-crystal structures present, we directly calculate the distance of the mutated protein residue from the drug ligand. Within our expectation, the nsSNVs which will positively shift binding affinity are more likely to be mapped on to residues within binding pocket (mean distance from ligand: 6.29Å and 19.8Å for

"Class 1" and "Class 2", respectively).

(c) Polar surface area and molecular weight as ligand features. Within the context of protein drug ligand interaction, physiochemical features of drug molecules play vital roles to interpret nsSNV implications. We observe that nsSNVs that disrupt binding affinity, the drug ligands tend to have a significant smaller average polar surface area that those corresponded with nsSNVs in the other class (mean ligand polar surface area: 94.62Å^2 and 105.5Å^2 for "Class 1" and "Class 2", respectively). We also observe that the average molecular weight of drug ligands interacting with disruptive nsSNVs is significantly higher than those corresponded with the other class (mean molecular weight of ligand: 361.0g/mol and 341.2g/mol for "Class 1" and "Class 2", respectively).

Figure 4. Performance and implementation of GenoDock classifier for binding affinity change prediction.

(a) ROC plots for four models with different input feature groups (with "Binding Site" feature

included during training process in “SNV annotation + PDB” and “SNV annotation + ligand” model). Our classifier achieved AUC of 0.73 (SNV annotations only), 0.92 (SNV annotations + Structure), 0.96 (SNV annotations + Ligand), and 0.97 (SNV annotations + Structure + Ligand), respectively. For “SNV annotation + PDB” and “SNV annotation + ligand” models, we train the model including binding site information, and we test the data assuming those nsSNVs will be mapped onto protein residues within binding pocket in order to estimate the upper limit of likelihood to disrupt ligand binding activity.

(b) ROC plots for four GenoDock models with different input feature groups (with “bind site” feature excluded during training process in “SNV annotation + PDB” and “SNV annotation + ligand” model). Our classifier achieved AUC of 0.73 (SNV annotations only), 0.76 (SNV annotations + Structure), 0.79 (SNV annotations + Ligand), and 0.97 (SNV annotations + Structure + Ligand), respectively. For “SNV annotation + PDB” and “SNV annotation + ligand” models, we train and test the model without “binding site” feature to predict the influence of nsSNVs onto binding affinity change in case we cannot tell whether the associated protein residue is on binding site or not. In GenoDock web interface, users can switch “binding site” to be known

or unknown for predictions of interest.

(c) Gini distance for relative feature significance in four models. We employ Gini distance as a measurement for feature importance in 4 models of GenoDock. We find GERP score, amino acid side chain volume change, polar surface area of drug ligand, distance between mutated amino acid residue and drug ligand are the most important features in SNV annotation features, PDB features, ligand features, and co-crystal structure features, respectively. While more features feeding into our classifier, significance of each feature are stable across different models. Particularly, binding site is an important feature if there is at least one structural component (protein PDB, drug ligand or co-crystal structure) present during the classification process of GenoDock. If the protein residue associated with nsSNV of interest is not on binding pocket, the probability of this nsSNV to disrupt the drug-protein binding is much smaller than those nsSNVs that are associated with binding pocket residues.

Fig.5. Case study: GenoDock identifies known and unknown drug-resistance mutations.

(a) Identification of T790M mutation on EGFR with gefitinib-resistant effect. The threonine on chain A in human EGFR protein (PDB ID: 2ity) is mutated to methionine by a somatic nsSNV (rs55181378). T790M is a well-studied mutation in clinical research. Patients with somatic activating mutations in the EGFR gene would develop resistance to tyrosine kinase inhibitors (TKIs) such as gefitinib (Ligand ID: IRE). With the T790M mutation, drug resistance arises from the steric hindrance of gefitinib binding due to the increased side chain volume of methionine, leading to a positive shift to binding affinity. GenoDock correctly predicts this shift step by step along its decision-making process.

(b) Identification of an unknown mutation potentially leading to drug resistance: resistance effect towards zoledronate acid by R112H mutation on human ASH1L. The arginine on chain A in ASH1L protein (PDB ID: 4p0w) is mutated to histidine by a somatic SNV (rs155317993). Due to the breaking of the salt bridge between the ARG side chain and the drug ligand zoledronic acid (Ligand ID: ZOL), the resulting uncharged HIS binds to the ligand much weaker, indicated by a positive shift of binding affinity change, which is correctly predicted by GenoDock.