Integrating genetic and structural features: building a hybrid model to characterize variants for protein-drug interactions

Bo Wang[1,6], Chengfei Yan[2,3,6], Shaoke Lou[2,3], Prashant Emani[2,3], Bian Li[2,3], Min Xu[2,3], Xiangmeng Kong[2,3], William Meyerson[2,5], Yucheng T. Yang[2,3], Donghoon Lee[2], Mark Gerstein[2,3,4,7*]

[1]Department of Chemistry, Yale University, New Haven, CT 06520, USA

[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

[4]Department of Computer Science, Yale University, New Haven, CT 06520, USA

[5]Yale School of Medicine, Yale University, New Haven, CT 06520, USA

[6]Co-first author

[7]Lead Contact

[*]Correspondence: pi@gersteinlab.org

Summary

Many drugs are known to be ineffective for some patients, carrying certain non-synonymous single nucleotide variants (SNV). Prioritization of SNVs that disrupt drug efficacy remains difficult due to lack of experimentally measured data of ligand binding assays (LBA). Here, with recent developments in both population-level next-generation sequencing (NGS) and high-resolution protein-drug co-crystal structure determination, we bootstrap physical calculations of binding affinity change as pseudo gold standard to construct a supervised learning method referred to as GenoDock to prioritize from gigantic SNVs candidates for those disrupting ones. Specifically, we collected the protein-drug complexes with high resolution structures and mapped associated somatic and germline SNVs onto the protein residues. We classify SNVs as disruptive and non-disruptive according to whether they can impair the binding from molecular docking calculations. We integrated genomics, structural and physicochemical features from SNVs, protein structures and drug ligands and trained GenoDock to do the prediction (with AUC=0.97).

s

2.

[illegible]

Introduction

In recent years, the immense growth of both genetic variation (Zuk et al., 2014) and protein structure datasets (Rose et al., 2015) which benefit from great advancement in related techniques has enabled us to study in depth the impact of genomic variants on protein structures and functions (Sethi et al., 2015). Great efforts have been taken to get the insights into how genetic variants associate with various diseases at a population level in order to potentially enhance drug effectiveness in the era of personalized medicine (Collins and Varmus, 2015; Ginsburg and McCarthy, 2001; Laing et al., 2011). Variant annotation tools such as SIFT (Adzhubei et al., 2013), Polyphen-2 (Adzhubei et al., 2013), and CADD (Kircher et al., 2014) 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 (SNV). In general, studies for this purpose are usually limited to the available experimentally measured SNV implication characterizations on native and mutant protein samples, compared with fast-growing amount of variants which can to be mapped onto protein structures (Glusman et al., 2017). Conceptionally, we can map SNV data with associated proteins to quantitatively investigate how related physical properties are altered upon point mutation. In practice, the experimentally measured data are highly limited for certain mutations. Up to date, compared with more than 1 million exonic SNVs that have been identified by various consortium projects such as TCGA and ExAC, the available experimental measurements on characterizing variant implications such as protein-ligand binding affinity change are even more scarce (Pires et al., 2015). When we enlarge the scope to structural bioinformatics, the available protein structure PDB files are in a larger scale: there are about 41,000 protein structures available from *Homo sapiens* in RSCB Protein Data Bank database (Berman et al., 2000); more than 175,000 exonic variants can be mapped with at least one protein PDB file with at least a 2.8Å resolution from RCSB PDB database(Kumar et al., 2016). Advance of computational methods for physical property calculations in past decades provides a practical way to bridge the gap.

Computational simulation of proteins has been validated as a crucial method (De Vivo et al., 2016) to study protein dynamics and conformations and to calculate associated physical properties such as free energy change, especially when the capacity of conducting experimental measurements is limited. Great advancements have been made from pioneering molecular dynamics (MD) work by Levitt et al. (Levitt and Warshel, 1975) and McCammon et al. (McCammon et al., 1977) decades ago, to more recent structure modeling and docking tools such as UCSF DOCK (Kuntz et al., 1982), Rosetta (Rohl et al., 2004), AutoDock (Morris et al., 2009), and MODELLER (Webb and Sali, 2016). With the growth of genetic variation data in a population level, linking protein 3D structures and genomics, i.e. genetic diversity across large population, using computational models has been proven to be a powerful and innovative approach for precision medicine (Meyer et al., 2018). Here, we choose protein-drug interactions as our primary focus. We aim to investigate how likely a SNV perturbs the interaction between the associated protein and drug ligands. Studies have shown that many drugs are effective towards only a limited fraction of individuals due to different responses from patients to specific drugs (Meyer et al., 2013; Spear et al., 2001; Wilkinson, 2005). One of the reasons of loss of efficacy for drugs is drug-resistant genetic variants carried by patients (Madian et al., 2012; Wilkinson, 2005) [=>Ref1.1]. 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 alteration of protein residues caused by SNVs would influence drug efficacy. Thus, a well-constructed database that directly links genetic variants to reliable human drug-protein co-crystal structures, as well as a robust methods to accurately predict if a SNV of interest would disrupt the binding of a drug to its protein target would help to investigate how individual carried variants would potentially affects drug efficacy.

To embody this idea, we developed a supervised learning method, GenoDock, to bridge SNVs on a large population scale and protein-drug co-crystal structures in the study. Our primary goal is to investigate how a given variant affects protein-ligand binding affinity. We first construct our database by mapping germline and somatic variants onto their associated protein residues with drug molecules present in the protein structure. We then examined the binding affinity change (ΔBA) between the native and mutated protein structures associated with each SNV in our database through molecular docking. We grouped the variants based on whether they would lead to a positive shift in binding affinity ($\Delta BA > 0$) or not ($\Delta BA \leq 0$). A positive shift in binding affinity indicates that the corresponded SNV is a disruptive one. The disruptive SNVs are our main focus in this study due to their high potential to associate with drug-resistance. Due to the available experimentally measured ligand binding affinity change data is highly limited (Benore, 2010), it is not practical to train a supervised learning model based on experiment data. We fill this gap by constructing a calculated binding affinity change set as our "pseudo gold standard" using docking program suites. This enables us to train a novel supervised learning model using random forest algorithm to predict the probability of a given SNV to destabilize protein-drug binding by integrating genomic, structural and physicochemical features from SNV 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 SNV candidates that disrupt protein-drug binding.

Results

GenoDock dataset and toolkit

Figure 1a highlights our strategy to construct the dataset that is publicly available from our GenoDock website (http://genodock.molmovdb.org/). The database contains 10,283 non-synonymous SNVs (SNV) from 228 proteins in *Homo sapiens*, and 113 FDA-approved drug ligands, which have co-crystal structures with at least one of the 228 proteins. We screened from over 30K human proteins with high resolution (better than 3.0Å) X-ray-solved protein PDB structures (https://www.rcsb.org/) and kept those 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 mapped the germline SNVs from Exome Aggregation Consortium (ExAC) (Lek et al., 2016) and the somatic SNVs from The Cancer Genome Atlas (TCGA) dataset (Cancer Genome Atlas Research, 2008, 2012; Cancer Genome Atlas Research et al., 2013) to these 228 protein structures according to BioMart-derived human gene and transcript ID (Kasprzyk, 2011). In total, we collected 8,565 SNVs in 166 PDB structures for ExAC germline variants, and 1,718 SNVs in 135 PDB structures for TCGA somatic mutations. The SNVs, protein structures, and drug ligands form SNV-Structure-Ligand 3-tuple entries in our database. For each SNV-Structure-Ligand entry, as visualized in Figure 1b, we used Modeller program suite (Webb and Sali, 2016) to generate a putative structural model of the point mutation through homology modelling. We then used AutoDock Vina (Trott and Olson, 2010) to calculate the binding affinity score for the 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, or "pseudo gold standard" for GenoDock program suite.

The change in binding affinity of the drug ligand after the protein target is mutated is the target label that GenoDock aims to predict based on a random forest classifier. 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. As shown in Figure 1c, we categorize ΔBA values for each SNV-Structure-Ligand entry into two classes: if ΔBA is positive, we tag it as "disruptive"; if ΔBA is non-positive, we tag it as "non-disruptive". We integrated selected genomic, structural and physicochemical features of SNVs, protein structures, and ligands to train the classifier: SNV annotation features include allele frequency, SIFT (Kumar et al., 2009), PolyPhen-2 (Adzhubei et al., 2010), and GERP (Davydov et al., 2010) scores; ligand features include molecular weight, hydrogen-bond donor and acceptor count, rotatable bond count and polar surface area; protein structure features include binding site, side chain hydropathy 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 S1 & S2).

Ideally, we need experimentally measured binding affinity assay (LBA) data to characterize the impact of SNV on protein-drug binding, but the LBA data is far from enough compared with the number of variants mapped on to protein residues. For example, Platinum database (Pires et al., 2015) is a recent effort to collect experimentally measured LBA data for over 1,000 mutations, which could potentially serve as the real gold standard of binding affinity change. However, only around 100 mutations of Platinum dataset are associated with human proteins. By constructing the pseudo gold standard set for each of the SNV-Structure-Ligand entry, we expand the number of entries of the gold standard set to ~10k. The GenoDock model thus enables us to prioritize SNVs that may potentially disrupt protein-drug binding on large scale of drug ligand, protein structure and exonic SNV datasets. For instance, there are more than 10 million exonic variants sequenced from consortium projects such as ExAC and TCGA; more than 175,000 exonic variants can be mapped onto at least one protein structure with at least a 2.8Å resolution from RCSB PDB database (Kumar et al., 2016); DrugBank database (Wishart et al., 2018) contains around 2,700 approved small molecule drugs. All these datasets could potentially be screened with GenoDock program suite to prioritize the disruptive SNVs.

by 15x

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. 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. Within the GenoDock database, we find that the two most abundant mutations are arginine to cysteine and arginine to histidine (Supplementary Figure S3). 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 wild-type distribution, see Figure S3); 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) (Stenson et al., 2014), the Online Database of Mendelian Inheritance in Man (OMIM) ((Hamosh et al., 2005), and ClinVar (Hamosh et al., 2005; Landrum et al., 2014; Peterson et al., 2013; Stenson et al., 2014); third, many cancer mutation signatures are enriched in the arginine to histidine mutation (Peterson et al., 2013). Previous literature shows that mutation from arginine to histidine can confer protein pH sensitivity to the mutant and thus alters protein function leading to diseases (Reichold et al., 2010; Szpiech et al., 2017; Zhang et al., 2012). Overall, we observe that ~ 1/3 of somatic SNVs 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 S4).

Distributions of ΔBA for common, rare, passenger, and driver SNVs

With these ExAC germline SNVs in our dataset, our interest is to see whether there is a significant difference between the rare and the common SNV groups in terms of destabilization of the protein-drug complex. Rare and ultra-rare SNVs are in general interpreted as of higher impact 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 SNVs into highly deleterious driver SNVs and neutral passenger SNVs to investigate different impacts of the two groups on drug binding (Stefl et al., 2013) (see 'Methods' for details regarding common, rare, passenger and driver SNV tagging).

In Figure 2, we visualize the distributions of binding affinity change for each group, especially for disruptive SNVs that positively shift ΔBA, which contribute to 6.0% and 8.9% of all SNVs in our ExAC and TCGA data source (Supplementary Figure S5). Though we do not observe a significant difference in ΔBA distributions between common and rare SNVs, when we bring together the top common and rare germline SNVs with positive ΔBA (the "outlier" region in the boxplot), top rare SNVs have a significantly higher ΔBA than those common ones. It implies that rare SNVs pool contains more extremely deleterious samples in terms of disrupting drug-protein binding than those from common SNV pool (e.g. the top 50 group has *p = 3.5e-7*; Supplementary Figure S6). 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 efforts made in characterization of cancer genomes (Cancer Genome Atlas Research, 2008, 2012; Forbes et al., 2011), people have validated the important roles of driver SNVs in driving cancer progression (Hong et al., 2015; Raphael et al., 2014). These facts motivate us to probe the impacts of SNVs from driver genes on perturbing interactions between associated protein residues and drug ligands. Indeed, our analysis shows a significant difference between passenger and driver SNVs. Those cancer-associated driver SNVs tend to destabilize protein-drug binding to a bigger extent compared with neutral passenger ones (*p = 3.60e-4*). In Figure 2, we also plot the percentage of SNVs that lead to a non-positive ΔBA together with the percentage of SNVs that do not change the binding affinity upon point mutation ($\Delta BA = 0$). We find that the portion of SNVs 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 SNV 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 disruptive and non-disruptive variants in terms of genomic, structural and physicochemical 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 SNVs from the rest.

Feature engineering and exploration to classify disruptive and non-disruptive SNVs

This work aims to provide a pipeline that could efficiently distinguish variants that destabilize protein drug binding activities (disruptive) from the rest (non-disruptive). 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 disruptive SNVs from those in non-disruptive and serve as training reference in our classifier (see 'Methods' for details on feature selection and construction). For each SNV-Structure-Ligand entry in GenoDock database, we construct three groups of features: 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 SNVs.

In Figure 3a, we observe that disruptive SNVs 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 non-disruptive ones (*p-value for SIFT is 1.21e-6 and p-value for Polyphen-2 is 2.20e-18*), indicating that those more deleterious SNVs (indicated by a lower SIFT or a higher Polyphen-2 score (Adzhubei et al., 2013; Adzhubei et al., 2010; Gonzalez-Perez and Lopez-Bigas, 2011; Kumar et al., 2009; Tennessen et al., 2012)) are more likely to cause a positive shift on ΔBA. The median GERP scores for the two classes also differ significantly (*p = 0.0101*). SNVs that cause positive ΔBA are likely to be mapped onto more conserved regions (indicated by a higher GERP score) (Genomes Project et al., 2012; Khurana et al., 2013; Tennessen et al., 2012) 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 SNVs 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 SNVs that impair binding activity are in the binding pocket (mean = 6.29Å) than the other class (mean = 19.8Å, *p = 1.27e-143*). 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 wild-type and mutated residue. Upon our definition of volume change index, we observe that SNVs 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 SNVs that lead to a non-positive ΔBA have a bulkier side chain volume (mean = 0.0343; *p = 1.68e-20*). Side chain hydropathy change is another feature in context of ligand-protein interaction. For example, side chain hydropathy score (Kyte and Doolittle, 1982) increasing from a hydrophilic 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*) (Boccuto et al., 2014; Doss and Nagasundaram, 2012; Kumar et al., 2013; Zhang et al., 2013). We observe this trend from the SNVs in our database, the SNVs with a positive ΔBA have a higher hydropathy score (mean = 0.63) than the other class (mean = 0.35), indicating that the disruptive SNVs tend to have a less hydrophilic character (*p = 0.0217*).

Figure 3c depicts the difference from the drug ligand in the co-crystal protein structure that SNVs are mapped to. In order to study SNVs' 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 S7). We observe that those SNVs 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 = *2.14e-3*). Also, we notice that the polar surface area of the drug ligands with a SNV that lead to positive ΔBA tend to be smaller (mean = 94.6Å^2), compared with the other group (mean = 105Å^2; p = *5.13e-5*). 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 hydropathy change upon point mutation.

After a feature exploration and engineering process based on differential effects of each feature has on disruptive and non-disruptive SNVs, we select good training feature candidates shown in Figure 3 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.

Construction and 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 using docking calculations as the gold-standard for ΔBA, aiming to help identifying potential SNVs 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 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 S8; see "Methods" for details). When applying GenoDock for practical use, it is possible that some of the feature groups are not available. For example, an user may only has an SNV and a drug ligand of interest to investigate whether the SNV would be disruptive towards ligand binding, and there is yet no protein structure accessible. Thus we provide four independent models depending on information availability (SNV only; SNV + Structure; SNV + Ligand; SNV + Structure + Ligand), we apply the procedure above onto each model to make our pipeline a uniform one. Model selection for different learning methods shows that random forest classifier is the best one (Supplementary Figure S9; see "Methods" for model selection).

During our preparation of training data, we tune the number of samples of disruptive SNVs and non-disruptive SNVs 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 "Binding Site" feature included and excluded during the training process, separately. As depicted in Figure 4a, we test the classifier trained with SNVs' "Binding Site" feature ("Binding Site" is "known") to get the probability of SNVs to disrupt binding. The area under the receiver-operator characteristic curve (AUC of ROC) for predictions of four models are 0.73 (SNV only), 0.91 (SNV + Structure), 0.96 (SNV + Ligand), and 0.97 (SNV + Structure + Ligand), respectively. If whether target SNVs are in binding pocket or not remains unknown, we then train our classifier with "Binding Site" feature excluded ("Binding Site" is "unknown") during training and test process for "SNV + Structure" and "SNV + Ligand" model. In Figure 4b, AUC values for these two models become 0.74 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 SNVs that lead to a positive shift towards binding affinity with an AUC of 0.97. Using the same learning pipeline, we back test the performance of

GenoDock with the performance of SIFT, Polyphen-2, GERP, and the Combined Annotation Dependent Depletion (CADD) (Kircher et al., 2014), independently. GenoDock gives the highest AUC value among these tools since it is specifically developed for addressing the impact of SNVs on ligand-binding affinity change instead of a general annotation towards potential benign or deleterious influences onto protein function (Supplementary Figure S10).

We then apply Gini importance to identify relevant importance of different features during the decision-making process. (Menze et al., 2009). We observe that the relative importance of the features such as the SNV annotations and binding site remain stable across our different 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 an SNV would make impacts towards ligand binding (Supplementary Figure S11).

Performance evaluation of GenoDock using experimentally measured data

To further evaluate the performance of GenoDock, we apply the program suite on an independent test set parsed from Platinum database, serving as the gold standard set (see 'Methods' for details on dataset preparation). For the 87 data entries parsed from Platinum, the AUC of ROC reaches 0.62, which shows reasonable and acceptable accuracy of GenoDock benched with experimentally measured results. We then evaluate the precision of GenoDock predictions on Platinum dataset by tuning the cutoff between "disruptive" and "non-disruptive" based on predicted probability of $\Delta BA > 0$. For example, when cutoff is set to be 0.7, those SNVs with a probability of $\Delta BA > 0$ greater than 0.7 will be assigned to be a "disruptive" one; otherwise we assign the SNV to be a "non-disruptive" one. We count the number of true positive and false positive entries benched with the gold standard set and calculated the precision. With cutoff to be 0.5, the precision reaches at 0.84 (Supplementary Figure S12).

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 SNVs that lead to a positive ΔBA shift from the rest candidates with high accuracy. The performance from the independent test set based on experimentally measured LBA results further validates the prediction reliability by GenoDock. In this study, we also identify SNV candidates that may potentially impair protein-drug binding.

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

We present an example of the implicit decision-making process of GenoDock in Figure 5, based on the overall importance score rankings of different features for the "SNV + Structure + Ligand" model. As shown in Figure 5a, GenoDock successfully reaches to the prediction that somatic T790M mutation (rs55181378) on 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 EGFR kinase domain (probability 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 (Balak et al., 2006; Janne, 2008; Kobayashi et al., 2005; Kosaka et al., 2006; Pao et al., 2005). 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 (Balak et al., 2006; Daub et al., 2004; Janne, 2008; Kobayashi et al., 2005). This biophysical rationale is traced in the classification process of GenoDock. From the decision flow in Figure 5a, 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 SNV associated with T790M mutation indicate that this variant is of high impact, which strengthens the confidence that this variant would impair the protein-ligand binding. Together with the next fact that the side chain hydropathy changes from the hydrophilic threonine to the hydrophobic methionine, GenoDock classifies this SNV to be very likely to cause a positive shift towards binding affinity.

In Figure 5b, we present an example representing the method by which GenoDock helps identify new variant candidates that could potentially lead to drug resistance, using "SNV + Structure + Ligand" model. 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 (Coscia et al., 2010; Kunzmann et al., 1999; Martin et al., 2001; Russell, 2011; Shipman et al., 1998; Wood et al., 2002). ZOL is a highly hydrophilic binder to FPPS via electrostatic and hydrogen bond interactions (Liu et al., 2014). 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 SNV as a disruptive one with a probability of 99.8%, followed by a similar decision-making pipeline discussed in the previous case. The disruptive role of R112H in ZOL binding to FPPS has not yet characterized through experiment assays, however, GenoDock provide evidence that this variant is highly possible to impair the inhibitor effectiveness and is worth further investigation. We validate the predictions of both examples using AutoDock, which gives a positive binding affinity shift. More biological functional assays can be performed in the future in addition to the computational validation. In addition of the two specific case above, GenoDock is also used to process large scale of SNV candidates for disruptive variant screening.

An application of GenoDock on large scale disruptive variant screening for drug ligands

GenoDock program suite can be applied for many purposes. Previous studies revealed that variants have direct impact upon protein structures, which could have significant consequences upon drug binding (Collins and Varmus, 2015; Ginsburg and McCarthy, 2001; Laing et al., 2011). However, no computational tool yet provides a large-scale analysis for the implications of variations on the drug efficacy. Here, we apply the program to evaluate how likely a drug ligand can be disrupted with large amount of somatic SNVs carried by individuals (Figure 6). The 290,515 somatic exonic variants are retrieved from original TCGA dataset, with SNV annotation features (SIFT, PolyPhen-2, and GERP) aligned for each SNV. We then select the two drug ligand from previous case study, gefitinib (IRE) and zoledronate (ZOL), together with eight other randomly picked-up drug ligands from GenoDock database: risedronate (RIS), sildenafil (VIA), acetazolamide (AZM), imatinib (STI), progesterone (STR), testosterone (TES), androstenedione (ASD), and dorzolamide (ETS). We run GenoDock ("SNV + Ligand" model) for each of the ten drug ligand with every SNV in the pool to calculate the probability of being disruptive of a variant for a certain drug. We assume that each variant is associated with amino acid residues locating within the binding pocket for that drug in order to evaluate the maximum probability of this variant to be disruptive. For each drug ligand, we plot the density distribution curve. Each curve represents ~0.3 million probability of being disruptive for each variant. Based on the cutoff we optimized based on Platinum dataset test result, SNV with a probability higher than 0.5 is highly likely to impair the binding for corresponded drug ligand. Thus, we can get a rough sense of how easily a drug ligand is affected when interacting with protein residues associated with various somatic variants. The higher portion of SNVs with probability higher than 0.5 for a drug ligand, the drug is more likely to be affected in its efficacy. For the ten drug ligands, imatinib has the highest number of SNVs that could potentially disrupt its binding with protein residues, whereas only 7% of the somatic SNVs could potentially impair binding activity for acetazolamide. This approach provides a reasonable method to evaluate drug ligand candidates with similar functionality. The drug ligand with less SNVs that could disrupt its binding may be a relatively better choice compared with other candidates with higher portion of disruptive SNVs within a given variant pool.

The GenoDock web interface

To make our method accessible, 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 feature only; SNA feature and structure feature; SNV feature and ligand feature, and all three groups of features. The predicted result will be feedback in form of a HTML webpage. The calculation page can be reached through http://genodock.molmovdb.org/calculation/0. Users can also download our open source python code to run large scale inputs on local computers or on HPC clusters.

Discussion

In this study, we constructed a dataset to bridge SNVs with their annotations from different sequencing datasets onto high resolution protein structures for downstream analysis; a highly sensitive classification model to prioritize SNV candidates that could potentially cause protein drug binding disruption based on integration of genomic annotations and structural properties, and a user-friendly web interface, the GenoDock server, that rapidly provides predictions of binding affinity change for SNVs of interest. The GenoDock method is a “hybrid model” that leverage physical calculations as “pseudo gold standard” to train statistical learning model when available experimentally measured “real gold standard” is highly limited.

For the construction of GenoDock database, we employed SNVs from the ExAC Consortium and the TCGA project as the source of germline variant and somatic variants feed, respectively.

From a pool of ~2.5 million ExAC germline variants and ~1 million Pan-Cancer somatic mutations, we successfully mapped ~10,000 SNVs onto ~300 human proteins for each of which high-resolution co-crystal structure with FDA-approved drug ligand is available. We identified 735 SNVs with predicted positive shift in binding affinity from 123 proteins of that ~300 protein pool, covering 85 drug ligands (see "Additional File: table 1"). For the prioritization of SNVs that would cause binding disruption, we demonstrated that GenoDock is an efficient classifier with an AUC of 0.97 when all features are available. The independent test set on Platinum experimentally measured binding affinity change results further shows acceptable and reasonable prediction sensitivity and precision with an AUC of 0.62 and with an precision of 0.84 (Supplementary Figure S13).

The major challenge of this study is to construct the gold standard set for binding affinity change native and wildtype protein-ligand co-crystal structures due to the lack of corresponded experimental measured ligand binding assay results. For example, Platinum contains about 1k mutations, and less than 10% of the mutations (87) are associated with SNVs mapped on human protein. This 87 experiment results of binding affinity change serves as our "real" gold standard set in the study. However, this dataset is far from enough to construct our supervised learning model. To fill this gap, we construct our "pseudo" gold standard set of binding affinity change for each of the ~10k SNVs in our GenoDock dataset via docking calculations. The prediction result of GenoDock based on the pseudo gold standard is acceptable when benched with the independent test set with 87 real gold standard entries. By conducting physical calculations to construct a relatively large enough gold standard to train our statistical leaning model, we are then enabled to process large scale of variants and structure datasets.

While our approach can identify SNV candidates that potentially impair protein-drug binding in a rapid yet accurate manner, the method is still limited in two aspects. First, the lack of high-resolution co-crystal structures of protein-drug complexes which limits the size of our "pseudo" gold standard set. As structural data is sparse, only 1% of exome SNVs were mapped onto protein-drug co-crystal structures. Fortunately, with the development of protein structure determination techniques such as NMR, electron microscopy and cryo-electron microscopy (cryo-EM) (Bai et al., 2015) , we can foresee that the number of highly reliable protein-drug structural data will increase rapidly. In addition, remarkable progress in putative 3D protein-drug interaction models based on homology modelling techniques may also potentially expand the structure pool (Marks et al., 2011; Zhan and Guo, 2015). 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 update the GenoDock dataset with new SNV-Structure-Ligand entries for better prediction results.

CONTRA

Second, our binding affinity change data is calculated based on docking calculations at current stage, which limits the upper boundary of our prediction accuracy. Calculation or prediction of binding affinity change between protein and ligand molecule is a challenging task. Developments of docking methods in recent years give us higher confidence of using the calculated results (Ballester et al., 2014; Smith et al., 2016; Yan et al., 2016). Thus we construct our "pseudo" gold standard set based on binding affinity calculations from AutoDock Vina, which is a state-of-the-art and a well-established program suite wildly used in pharmaceutical research projects (Castro-Alvarez et al., 2017; Wang et al., 2016). We further validated the consistency of the ΔBA results for each SNV-Structure-Ligand entry via AutoDock . If we have enough experimentally measured LBA data for mutations recoded in GenoDock database later on, we plan to update the ΔBA 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 “hybrid model” that joint physical calculations and statistical learning for predicting SNV candidates that could potentially disrupt protein-ligand binding, 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. We believe that GenoDock will continuously help to predict the impacts of SNVs on protein-drug interaction when datasets of larger size and better quality, advanced molecular docking software, and more LBA experimental data being used in our method.

Methods

GenoDock Database preparation

Germline exonic variants were collected from Exome Aggregation Consortium(ExAC) release 1(Lek et al., 2016) (download source: ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/). Somatic exonic 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 (Wishart et al., 2018). Human protein PDB structures with a resolution better than 3.0 Å were downloaded from the Protein Data Bank (https://www.rcsb.org/) (Berman et al., 2000). A careful curation to filter in 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 (Kumar et al., 2016). For tagging common and rare variants from ExAC dataset, a cutoff of 1 was used to differentiate rare SNVs from common ones: if a variant occurred only once in the ExAC dataset, we tagged this SNV as rare; if a variant occurrence count is bigger than 1, we tagged it as a common SNV [=>Ref2.2]. For tagging driver and passenger SNVs from TCGA dataset, SNVs were tagged as enriched-in-driver variants if they were variants in cancer driver genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC), version 83. If a variant was not in a cancer driver gene, we tagged it as a passgener one. Not all SNVs in driver genes are driver variants, but they are more likely to be driver variants, which is sufficient for our purpose in this study.

Mutant structure and binding affinity change calculation

For each entry recorded in our database, we generated a mutant structure associated with that SNV through homology modelling using Modeller (ver. 9.18) (Webb and Sali, 2016), using the corresponded native co-crystal structure as template. 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 (Trott and Olson, 2010) to evaluate the change in drug binding affinity to setup the pseudo gold standard set: $\Delta BA = \Delta G(MUT) - \Delta G(WT)$, in kcal/mol, where $\Delta G(MUT)$ and $\Delta G(WT)$ are binding affinities of the drug with the mutated and wild-type protein target evaluated using 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 experimentally measured LBA data for every entry in GenoDock dataset, we validated the calculations of Vina by applying the same procedure with AutoDock Tools (ver. 6.2.6) (Morris et al., 2009) 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 12), indicating a strong consistency.

Features extraction and construction for machine learning method

SNV features:

SIFT and PolyPhen-2: SIFT score and Polyphen-2 score for somatic and germline exonic SNVs in our study were directly extracted from the “INFO” column of VCF files from ExAC consortium and TCGA project.

GERP: GERP scores were retrieved directly from Sidow lab (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html) (Davydov et al., 2010).

Ligand features:

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 were extracted from PubChem database (Kim et al., 2016).

Structure features:

Amino acid side chain volume change index: defined as $\Delta V_{index} = log_2(\frac{V_{MUT}}{V_{WT}})$, where V_{MUT} and V_{WT} stand for van der Waals volume (Darby and Creighton, 1993) of mutant and wild-type protein residue, respectively.

Amino acid side chain hydropathy change: for each amino acid, we employed amino acid hydropathy scale by Kyte and Doolittle (Kyte and Doolittle, 1982) as the hydropathy metric [=>Ref2.3]. Amino acid side chain hydropathy change index is defined as $\Delta\text{hydropathy} = \text{hydropathy}(\text{mutant}) - \text{hydropathy}(\text{WT})$.

Distance between mutation and drug ligand: the distance between a protein residue to a ligand was defined as the shortest distance of a heavy atom of that residue to a heavy atom of the associated ligand.

Binding site ("on"/"off"): This is a binary feature describing whether the mutation is on or off the binding site. If a residue has a distance less than 8Å from the target ligand in the co-crystal structure, we consider that this residue is on the binding pocket . Though this feature is grouped into the structure feature set, it could still be used when only one of drug ligand or protein structure is available. We construct the "SNV + Ligand" model and "SNV + Structure" model under two scenarios: "Binding Site" is known and "Binding Site" is unknown. The former model was trained with "Binding Site" feature included, and users need to tell GenoDock whether the SNV of interest is associated with residues on or off the binding pocket. The later model was trained with "Binding Site" feature excluded. In practice , with "Binding Site" being "on", we are able to predict the maximal probability of the target SNV to be ligand-binding disruptive. On the other hand, users are also free to set "Binding Site" being "off" if they want the prediction for the protein residues of associated variants are not in binding sites. When the user does not care about binding status when applying "SNV + Ligand" model or "SNV + Structure" models, they can use remove this feature and make "Bind Site" to be unknown. We engineered GenoDock source scripts for both application scenarios.

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

GenoDock dataset was separated into training set (70%) and test set (30%) in a random manner. To avoid potential bias raised from imbalanced composition of the two classes of samples in our dataset (735 entries for disruptive SNVs; 9,458 entries for non-disruptive SNVs), we counted the number of disruptive SNV samples ($\Delta BA > 0$) and randomly select equal number of non-disruptive SNV samples from ($\Delta BA \leq 0$) to make up the balanced training set. Scikit-learn package (Pedregosa et al., 2011) is used for learning model development. We tested classification methods including Lasso Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). We trained each learning model through a 10-fold grid-search cross-validation process. For each training, the rest 30% data was tested for performance evaluation. Based on the AUC values, RF has the highest AUC among all methods (Supplementary Figure S9). Feature selection was performed by evaluation of AUC for each feature respectively. If the selection power of a feature was near or worse than random selection, we removed 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 only; SNV + Structure; SNV + Ligand; SNV + Structure + Ligand) for GenoDock.

Curation of independent test set based on experimental measurements

We also prepared an independent bench set comprising experimentally measured binding affinity change upon for mutations from Platinum database (Pires et al., 2015). Briefly, the full Platinum database content was downloaded as a flat comma-separated file from http://biosig.unimelb.edu.au/platinum/. Amino acid mutations other than single-point mutations and those found in species other than human beings were excluded. In addition, mutations that are not resulted from single-nucleotide variation were also removed because GenoDock uses the GERP score as one of the predictive features and the GERP score is position-specific. Further, mutations that cannot be mapped onto their associated UniProtKB canonical amino acid sequences were discarded. In the end, 87 unique data points were obtained (two data points with the same mutation but different ligands were considered to be different) and used as the independent test set. Each data point in this set was labeled as "disruptive" if its associated fold change in binding affinity upon mutation, ($\frac{BA_{WT}-BA_{MUT}}{BA_{WT}}$), is negative or "non-disruptive" otherwise. Note that the curation of this test set was conducted in a manner that is completely blinded from the training of GenoDock. Ligand features and structure features are assigned for each mutation entry for the database to run GenoDock. We then apply "SNV + Structure + Ligand" model (auROC = 0.97) on the dataset to evaluate the reliability of GenoDock predictions. [=>Ref1.3; 2.5, 2.6]

Protein-ligand complex visualization

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

SUPPLEMENTAL INFORMATION

Supplemental Information includes twelve figures and one additional information table.

Author Contributions

B.W., C.Y., and M.G. conceived and designed the study. B.W. carried out the study, developed scripts, constructed the web-interface, produced the figures, and wrote the paper. C.Y. and M.X. calculated docking experiments. S.L., P.E., X.K., W.M., and D.L. prepared and processed datasets. All authors edited the manuscript. M.G. and C.Y. oversaw the project.

Acknowledgements

This work is supported by XXX. We thank Sushant Kumar for sharing the original source code for mapping variants onto protein structures, as well as Jing Zhang and Declan Clarke for helpful discussions and feedback.

References

The PyMOL Molecular Graphics System. Schrodinger, LLC.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*, Unit7 20.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Bai, X.C., McMullan, G., and Scheres, S.H. (2015). How cryo-EM is revolutionizing structural biology. Trends Biochem Sci *40*, 49-57.

Balak, M.N., Gong, Y., Riely, G.J., Somwar, R., Li, A.R., Zakowski, M.F., Chiang, A., Yang, G., Ouerfelli, O., Kris, M.G.*, et al.* (2006). Novel D761Y and common secondary T790M mutations in epidermal growth factor receptor-mutant lung adenocarcinomas with acquired resistance to kinase inhibitors. Clin Cancer Res *12*, 6494-6501.

Ballester, P.J., Schreyer, A., and Blundell, T.L. (2014). Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? J Chem Inf Model *54*, 944-955.

Benore, M. (2010). Response to review of Fundamental Laboratory Approaches for Biochemistry and Biotechnology. Biochem Mol Biol Educ *38*, 64.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Boccuto, L., Aoki, K., Flanagan-Steet, H., Chen, C.F., Fan, X., Bartel, F., Petukh, M., Pittman, A., Saul, R., Chaubey, A.*, et al.* (2014). 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 *23*, 418-433.

Cancer Genome Atlas Research, N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068.

Cancer Genome Atlas Research, N. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120.

Castro-Alvarez, A., Costa, A.M., and Vilarrasa, J. (2017). The Performance of Several Docking Programs at Reproducing Protein-Macrolide-Like Crystal Structures. Molecules *22*.

Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. N Engl J Med *372*, 793-795.

Coscia, M., Quaglino, E., Iezzi, M., Curcio, C., Pantaleoni, F., Riganti, C., Holen, I., Monkkonen, H., Boccadoro, M., Forni, G.*, et al.* (2010). Zoledronic acid repolarizes tumour-associated macrophages and inhibits mammary carcinogenesis by targeting the mevalonate pathway. J Cell Mol Med *14*, 2803-2815.

Darby, N.J., and Creighton, T.E. (1993). 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 *232*, 873-896.

Daub, H., Specht, K., and Ullrich, A. (2004). Strategies to overcome resistance to targeted protein kinase inhibitors. Nat Rev Drug Discov *3*, 1001-1010.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol *6*, e1001025.

De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of Molecular Dynamics and Related Methods in Drug Discovery. J Med Chem *59*, 4035-4061.

Doss, C.G., and Nagasundaram, N. (2012). Investigating the structural impacts of

I64T and P311S mutations in APE1-DNA complex: a molecular dynamics approach. PLoS One *7*, e31677.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res *39*, D945-950.

Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Ginsburg, G.S., and McCarthy, J.J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol *19*, 491-496.

Glusman, G., Rose, P.W., Prlic, A., Dougherty, J., Duarte, J.M., Hoffman, A.S., Barton, G.J., Bendixen, E., Bergquist, T., Bock, C.*, et al.* (2017). Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med *9*, 113.

Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet *88*, 440-449.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res *33*, D514-517.

Hong, M.K., Macintyre, G., Wedge, D.C., Van Loo, P., Patel, K., Lunke, S., Alexandrov, L.B., Sloggett, C., Cmero, M., Marass, F.*, et al.* (2015). Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. Nat Commun *6*, 6605.

Janne, P.A. (2008). Challenges of detecting EGFR T790M in gefitinib/erlotinib-resistant tumours. Lung Cancer *60 Suppl 2*, S3-9.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database (Oxford) *2011*, bar049.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A.*, et al.* (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A.*, et al.* (2016). PubChem Substance and Compound databases. Nucleic Acids Res *44*, D1202-1213.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.

Kobayashi, S., Boggon, T.J., Dayaram, T., Janne, P.A., Kocher, O., Meyerson, M., Johnson, B.E., Eck, M.J., Tenen, D.G., and Halmos, B. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med *352*, 786-792.

Kosaka, T., Yatabe, Y., Endoh, H., Yoshida, K., Hida, T., Tsuboi, M., Tada, H., Kuwano, H., and Mitsudomi, T. (2006). Analysis of epidermal growth factor receptor gene mutation in patients with non-small cell lung cancer and acquired resistance to gefitinib. Clin Cancer Res *12*, 5764-5769.

Kumar, A., Rajendran, V., Sethumadhavan, R., and Purohit, R. (2013). Molecular dynamic simulation reveals damaging impact of RAC1 F28L mutation in the switch I region. PLoS One *8*, e77453.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc *4*, 1073-1081.

Kumar, S., Clarke, D., and Gerstein, M. (2016). Localized structural frustration for evaluating the impact of sequence variants. Nucleic Acids Res *44*, 10062-10073.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. (1982). A geometric approach to macromolecule-ligand interactions. J Mol Biol *161*, 269-288.

Kunzmann, V., Bauer, E., and Wilhelm, M. (1999). Gamma/delta T-cell stimulation by pamidronate. N Engl J Med *340*, 737-738.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J Mol Biol *157*, 105-132.

Laing, R.E., Hess, P., Shen, Y., Wang, J., and Hu, S.X. (2011). The role and impact of SNPs in pharmacogenomics and personalized medicine. Curr Drug Metab *12*, 460-

486.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res *42*, D980-985.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Levitt, M., and Warshel, A. (1975). Computer simulation of protein folding. Nature *253*, 694-698.

Liu, Y.L., Lindert, S., Zhu, W., Wang, K., McCammon, J.A., and Oldfield, E. (2014). Taxodione and arenarone inhibit farnesyl diphosphate synthase by binding to the isopentenyl diphosphate site. Proc Natl Acad Sci U S A *111*, E2530-2539.

Madian, A.G., Wheeler, H.E., Jones, R.B., and Dolan, M.E. (2012). Relating human genetic variation to variation in drug responses. Trends Genet *28*, 487-495.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Martin, M.B., Grimley, J.S., Lewis, J.C., Heath, H.T., 3rd, Bailey, B.N., Kendrick, H., Yardley, V., Caldera, A., Lira, R., Urbina, J.A.*, et al.* (2001). Bisphosphonates inhibit the growth of Trypanosoma brucei, Trypanosoma cruzi, Leishmania donovani, Toxoplasma gondii, and Plasmodium falciparum: a potential route to chemotherapy. J Med Chem *44*, 909-916.

McCammon, J.A., Gelin, B.R., and Karplus, M. (1977). Dynamics of folded proteins. Nature *267*, 585-590.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics *10*, 213.

Meyer, M.J., Beltran, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods.

Meyer, U.A., Zanger, U.M., and Schwab, M. (2013). Omics and drug response. Annu Rev Pharmacol Toxicol *53*, 475-502.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., and Olson, A.J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem *30*, 2785-2791.

Pao, W., Miller, V.A., Politi, K.A., Riely, G.J., Somwar, R., Zakowski, M.F., Kris, M.G., and Varmus, H. (2005). Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med *2*, e73.

Pedregosa, F.a.V., G. and Gramfort, A. and Michel, V., and Thirion, B.a.G., O. and Blondel, M. and Prettenhofer, P., and Weiss, R.a.D., V. and Vanderplas, J. and Passos, A. and, and Cournapeau, D.a.B., M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825--2830.

Peterson, T.A., Doughty, E., and Kann, M.G. (2013). Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. Journal of Molecular Biology *425*, 4047-4063.

Pires, D.E., Blundell, T.L., and Ascher, D.B. (2015). Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res *43*, D387-391.

Raphael, B.J., Dobson, J.R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med *6*, 5.

Reichold, M., Zdebik, A.A., Lieberer, E., Rapedius, M., Schmidt, K., Bandulik, S., Sterner, C., Tegtmeier, I., Penton, D., Baukrowitz, T.*, et al.* (2010). KCNJ10 gene mutations causing EAST syndrome (epilepsy, ataxia, sensorineural deafness, and tubulopathy) disrupt channel function. Proc Natl Acad Sci U S A *107*, 14490-14495.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol *383*, 66-93.

Rose, P.W., Prlic, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J.*, et al.* (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res *43*, D345-356.

Russell, R.G. (2011). Bisphosphonates: the first 40 years. Bone *49*, 2-19.

Sethi, A., Clarke, D., Chen, J., Kumar, S., Galeev, T.R., Regan, L., and Gerstein, M. (2015). Reads meet rotamers: structural biology in the age of deep sequencing. Curr Opin Struct Biol *35*, 125-134.

Shipman, C.M., Croucher, P.I., Russell, R.G., Helfrich, M.H., and Rogers, M.J. (1998). The bisphosphonate incadronate (YM175) causes apoptosis of human myeloma cells in vitro by inhibiting the mevalonate pathway. Cancer Res *58*, 5294-5297.

Smith, R.D., Damm-Ganamet, K.L., Dunbar, J.B., Jr., Ahmed, A., Chinnaswamy, K., Delproposto, J.E., Kubish, G.M., Tinberg, C.E., Khare, S.D., Dou, J.*, et al.* (2016). CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking, and Scoring/Ranking Challenge. J Chem Inf Model *56*, 1022-1031.

Spear, B.B., Heath-Chiozzi, M., and Huff, J. (2001). Clinical application of pharmacogenetics. Trends Mol Med *7*, 201-204.

Stefl, S., Nishi, H., Petukh, M., Panchenko, A.R., and Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. J Mol Biol *425*, 3919-3936.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). 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.

Szpiech, Z.A., Strauli, N.B., White, K.A., Ruiz, D.G., Jacobson, M.P., Barber, D.L., and Hernandez, R.D. (2017). Prominent features of the amino acid mutation landscape in cancer. PLoS One *12*, e0183273.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G.*, et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64-69.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem *31*, 455-461.

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S., and Hou, T. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys *18*, 12964-12975.

Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Protein Sci *86*, 291-2937.

Wilkinson, G.R. (2005). Drug metabolism and variability among patients in drug response. N Engl J Med *352*, 2211-2221.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z.*, et al.* (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46*, D1074-D1082.

Wood, J., Bonjean, K., Ruetz, S., Bellahcene, A., Devy, L., Foidart, J.M., Castronovo, V., and Green, J.R. (2002). Novel antiangiogenic effects of the bisphosphonate compound zoledronic acid. J Pharmacol Exp Ther *302*, 1055-1061.

Yan, C., Grinter, S.Z., Merideth, B.R., Ma, Z., and Zou, X. (2016). Iterative Knowledge-Based Scoring Functions Derived from Rigid and Flexible Decoy Structures: Evaluation with the 2013 and 2014 CSAR Benchmarks. J Chem Inf Model *56*, 1013-1021.

Zhan, Y., and Guo, S. (2015). Three-dimensional (3D) structure prediction and function analysis of the chitin-binding domain 3 protein HD73_3189 from Bacillus thuringiensis HD73. Biomed Mater Eng *26 Suppl 1*, S2019-2024.

Zhang, Z., Miteva, M.A., Wang, L., and Alexov, E. (2012). Analyzing effects of naturally occurring missense mutations. Comput Math Methods Med *2012*, 805827.

Zhang, Z., Norris, J., Kalscheuer, V., Wood, T., Wang, L., Schwartz, C., Alexov, E., and Van Esch, H. (2013). A Y328C missense mutation in spermine synthase causes a mild form of Snyder-Robinson syndrome. Hum Mol Genet *22*, 3789-3797.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A *111*, E455-464.

Main Figure Captions and main Figures

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

(a) A flowchart for collecting 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). SNVs are mapped with protein drug co-crystal structures to form each SNV-Structure-Ligand entry in our database. We then calculate the binding affinity change for each mutation to construct the pseudo gold standard for further statistical learning model.

(b) Illustration of protein-ligand binding affinity change calculations. For each native co-crystal structure in our dataset, we generate a mutant structure using Modeller. For each of the native and mutated structure pair, we calculate the binding affinity using AutoDock, respectively, in order to obtain the binding affinity change (ΔBA) upon the point mutation. ΔBA for each SNV serves as the pseudo gold standard set for later on classification model.

(c) Construction of the random forest model to predict the direction of protein-ligand binding affinity change ($\Delta BA > 0 \; or \; \Delta BA \leq 0$). With feature engineering and exploration, several SNV features (i.e. SIFT, GERP, Polyphen-2), drug ligand features (i.e. molecular weight, hydrogen bond donor/acceptor count), and structure features (i.e. binding site, side chain volume and hydropathy change) are combined to predict the direction of protein-ligand binding affinity change. GenoDock program suite is trained and tested with a rigorous cross-validation, model selection and evaluation manner, with four application models available. We employ Platinum dataset as our real gold standard providing about 100 records of experimentally measured binding affinity changes for human protein mutations. By constructing docking calculations for binding affinity changes, we have calculated binding affinity change for each of the ~10,000 mutations in GenoDock database, making it possible to train a supervised learning model with confidence. With the trained GenoDock model, it is then possible to screen large scale datasets for drug ligands (e.g. ~2.7k from DrugBank database), human protein structures (~30,000 from RCSB PDB website with resolution higher than 3.0 Å), and exonic SNVs (~1,000,000 sequenced exonic SNVs; ~175,000 SNVs mapped with at least one human protein structure with 2.8Å or higher resolution from RCSB PDB database).

Figure 2. Boxplot of ligand binding affinity changes for different types of SNVs in GenoDock

An overall comparison of common, rare, passenger and driver SNVs in terms of binding affinity change from GenoDock data source. SNVs with $\Delta BA > 0$ are plotted in order to compare the extent of destabilization towards ligand binding activities by each SNV 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, where driver SNVs 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 ($\Delta BA < 0\ or\ \Delta BA = 0$; 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 SNVs have a greater probability to result in a positive binding affinity change compared with the other three groups.

Fig. 3. Boxplot distribution between disruptive SNVs (positive binding affinity shift) and non-disruptive SNVs (non-positive binding affinity shift) regarding different features

groups:

(a) PolyPhen-2, SIFT and GERP score as SNV 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. SNVs that disrupt ligand protein binding have a higher mean Polyphen-2 score (mean Polyphen-2 value: 0.665 and 0.516 for disruptive and non-disruptive SNVs, respectively) and a lower SIFT score (mean SIFT value: 0.101 and 0.149 for disruptive and non-disruptive SNVs, respectively), both indicating a more deleterious role of disruptive SNVs on protein function. In terms of GERP score, SNVs 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 disruptive and non-disruptive SNVs, respectively).

(b) Side-chain volume and hydropathy change as protein structure features; distance between ligand and mutated residue when co-crystal structure is present. Amino acid side chain volume and hydropathy change before and after mutation directly affect interaction of protein residue with ligand. We observe that the mean value of both side chain volume and hydropathy are statistically significant. On average, SNVs 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 disruptive and non-disruptive SNVs, respectively). For side chain hydropathy change, there is also a significant difference between the two classes of SNVs (mean hydropathy change: 0.6306 and 0.3562 for disruptive and non-disruptive SNVs, 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 SNVs 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 disruptive and non-disruptive SNVs, respectively).

(c) Polar surface area and molecular weight as ligand features. Within the context of protein drug ligand interaction, physicochemical features of drug molecules play vital roles to interpret SNV implications. We observe that SNVs that disrupt binding affinity, the drug ligands tend to have a significant smaller average polar surface area that those corresponded with SNVs in the other class (mean ligand polar surface area: 94.62Å^2 and 105.5Å^2 for disruptive and non-disruptive

SNVs, respectively). We also observe that the average molecular weight of drug ligands interacting with disruptive SNVs is significantly higher than those corresponded with the other class (mean molecular weight of ligand: 361.0g/mol and 341.2g/mol for disruptive and non-disruptive SNVs, respectively).

Figure 4. Performance and implementation of GenoDock 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 + Structure” and “SNV + ligand” model). Our classifier achieved AUC of 0.73 (SNV only), 0.91 (SNV + Structure), 0.96 (SNV + Ligand), and 0.97 (SNV + Structure + Ligand), respectively. For “SNV + Structure” and “SNV + Ligand” models, we train the model including binding site information, and we test the data with original binding site information of each single SNV.

(b) ROC plots for four GenoDock models with different input feature groups (with “Binding Site” feature excluded during training process in “SNV + Structure” and “SNV + Ligand” model). Our classifier achieved AUC of 0.73 (SNV only), 0.74 (SNV + Structure), 0.79 (SNV + Ligand), and 0.97 (SNV + Structure + Ligand), respectively. For “SNV + PDB” and “SNV + Ligand” models, we train and test the model without “Binding Site” feature to predict the influence of SNVs 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.

Figure 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 SNV (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.

Figure 6. An example of GenoDock application on large scale dataset

We apply GenoDock ("SNV + Ligand" model) on a pool of 10 drug ligands on a set of 290,515 somatic exonic variants from original TCGA dataset to estimate how vulnerable each drug ligand is to be disrupted by individual carried variants. We assume that each variant is associated with amino acids locating in the binding pocket in order to estimate the maximum probability of this SNV to disrupt protein-ligand binding. Each line stands for density distribution of these ~0.3 million variants in terms of probability of being disruptive to a certain drug ligand. Variants with a probability higher than 0.5 is highly likely to impair the binding. The less disruptive SNVs a drug is associated with, the more likely the drug will be in terms of retaining its efficacy for individuals carrying a variety of variants. Within the 10 drug ligands selected, imatinib (STI) has the most number (65%) of SNVs which are likely to disrupt its binding with proteins; relatively, acetazolamide (AZM) has least portion of disruptive SNVs (7%) compared with other drug ligands.