Germline and Somatic Non-Synonymous Single Nucleotide Variations in Drug Binding Sites

By Cheng Yan

B.S. in Pharmacy, June 2012, First Moscow State Medical University I.M. Sechenov

A thesis submitted to

The Faculty of
The Columbian College of Arts & Sciences of The George Washington University in partial fulfillment of the requirements for the degree of Master of Science

May 17, 2015

Thesis directed by

Raja Mazumder
Associate Professor of Biochemistry and Molecular Biology

Acknowledgement

The author would like to thank academic advisor, Dr. Raja Mazumder, for his valuable suggestions on this thesis. He also wants to thank Dr. Nagarajan Pattabiraman for his efforts on protein-drug binding site calculation in this thesis. Lastly, HIVE, The High-performance Integrated Virtual Environment developed by Dr. Raja Mazumder and Dr. Vahan Simonyan, is used for data collection.

Abstract of Thesis

Germline and Somatic Non-Synonymous Single Nucleotide Variations in Drug Binding Sites

Advancements in next-generation sequencing (NGS) technologies is generating huge amount of data. However, the challenge of translating NGS data to actionable clinical interpretation remains. We have combined in a comprehensive manner, germline and somatic non-synonymous single-nucleotide variations (nsSNV) that impact drug binding sites to investigate the prevalence of such variations. The integrated data thus generated in conjunction with exome or whole-genome sequencing data can be used to identify patients who may not respond to a specific drug because of alterations in drug binding efficacy due to nsSNVs in the target protein. To identify the nsSNVs that may affect drug binding, protein-drug complex structures were retrieved from Protein Data Bank (PDB) followed by identification of amino acids in the protein-drug binding sites using an occluded surface method. Then, the germline and somatic mutations were mapped to these amino acids to identify which of them alter amino acid-drug binding sites. Using this method we identified 12,993 amino acid-drug binding sites across 253 unique proteins bound to 235 unique drugs. The integration of amino acid-drug binding sites data with both germline and somatic nsSNVs datasets revealed 3,133 nsSNVs affecting amino acid-drug binding sites. Based on protein similarity and conservation of amino acid-drug binding sites, 81 paralogs were identified as additional proteins that can serve as alternative and potential drug targets. Information on these key drugs, their protein binding sites and prevalence of both somatic and germline nsSNVs that disrupt these binding sites can provide valuable knowledge for personalized medicine treatment. A web portal is available (http://hive.biochemistry.gwu.edu/tools/drugvar) where nsSNVs

from individual patient can be checked by scanning against DrugVar to determine if any of the single-nucleotide variations affect the binding of any drug in the database.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

With the development of massively parallel sequencing, also known as next-generation sequencing (NGS), a vast amount of NGS data is being generated with greater throughput and decreased cost compared with its predecessor technology, Sanger sequencing (1-4). In order to accommodate the rapid increase of data generated by the new generation sequencers, NGS data repositories have been created, such as Short Read Archive (SRA) developed by National Center for Biotechnology Information (NCBI) (5) and CGHub supported by National Cancer Institute. Up to now, short sequence reads amounting to more than three petabytes have been deposited into these archives. How to convert the NGS data to actionable, medically meaningful knowledge is still a challenge which many researchers including our group is attempting to address.

The identification of single-nucleotide variations (SNVs) is one of the most common tasks in NGS data analysis (6). Although most of the SNVs are found within intergenic region, when they exist at more crucial location, such as protein coding genes or regulatory region, they may play a more direct role in causing disease by changing the protein structure or disrupt protein expression level (7,8). Therefore, SNV identification has been widely used in investigating the relationship between sequence variation and disease (9). For example, for cancer, many SNVs found both in oncogene and tumor suppressor genes have been shown to play critical role in tumorigenesis (9). SNVs currently also serve as ideal biomarkers in cancer prediction and diagnosis. Pharmacogenetics and pharmacogenomics studies have shown that SNVs can affect how

patients respond to drugs (10-13). The most direct example is where a SNVs exists within the coding region of a gene coding for target protein and these SNVs alter the amino acid of binding site of the drug resulting in changes to drug binding affinity and consequent therapeutic effect (14-16).

Structure/ligand based drug discovery technology has been routinely used to design biologically active molecules (17,18). The design of HIV protease inhibitors (19-21) is an example of protein structure-based drug discovery. The Protein Data Bank (PDB) (22), one of the largest three-dimensional (3D) structure database, contains three dimensional structure data of proteins complexed with small molecules such as substrates, cofactors, inhibitors and drugs and is used in drug discovery research (23-25). Secondary databases, such as CREDO (26) and FireDB (27) uses data from PDB and through further analysis provides value-added information which can be used for knowledge discovery. CREDO stores atomic interaction of molecular contacts and limited germline amino acid variations mapped onto protein structures. FireDB consists of ligands and annotated functional site residues. The database consists of 15,777 PDB sequences clustered based on 93% sequence similarity. None of the above databases provide comprehensive somatic mutation or polymorphism mapping; neither do they provide drug centric information.

Research has shown that individual's genetic makeup can contribute to differential drug response (10,28-30). PharmGKB mines information from this type of research (31) and at the time of writing of this manuscript contained over 5000 variant annotations in over

900 proteins related to over 600 drugs. The variant-drug-disease relationship is extracted from publications and is then annotated and integrated into PharmGKB by experts.

The project described in this manuscript involves identification and integration of amino acid-drug binding sites from PDB and non-synonymous single-nucleotide variations (nsSNVs) compiled from various sources to create a comprehensive dataset, called DrugVar, which can be used to scan exome or whole genome sequencing data from patients. DrugVar includes both somatic mutations identified by cancer sequencing projects, such as COSMIC (32), TCGA (http://cancergenome.nih.gov/) and ICGC (33) and germline variations from dbSNP (34).

Chapter 2: Methods

Generation of amino acid-drug binding dataset

Amino acid-drug complex structure data was obtained from PDB database (35). The Anatomical Therapeutic Chemical Classification (ATC), a hierarchical representation of drugs was used for identifying cancer and non-cancer related drugs (http://www.rcsb.org/pdb/browse/jbrowse.do?t=18&useMenu=no). Since DrugVar is a drug-target centric database, a series of criteria was set to filter the data obtained from ATC browser. In majority of amino acid-drug PDB entries, in addition to the actual drug molecule other small molecules, such as ions and stabilizer used in crystallization procedures are also present. To separate drugs from other small molecules, we manually curated the data. Only proteins of human origin were selected for our study. For proteins

which have multiple chains, we considered only one chain in the PDB file containing the bound drug. Since the amino acid sequence positions are not always the same between PDB and UniProtKB (36) due to modification and chopping of N and C-terminal amino acids for obtaining good crystals, we mapped the PDB ID and the amino acid sequence positions to UniProtKB accession using ID Mapping (37) followed by pairwise alignment to map the amino acid sequence positions. For each atom of the amino acid residues in the drug binding pocket, occluded surface (buried) area by the drug in its binding pocket was computed using the PDB structure of amino acid-drug complex and the program OS (38). From atom's occluded surface areas in a residue, the total occluded surface area for each residue was calculated and converted to percentage of occlusions for each residue. For each amino acid-drug complex, a list of residues was generated based by rank ordering their percentage of occlusion. In cases where multiple PDB IDs are associated with the same protein all protein-drug pairs were considered.

Generation of nsSNVs dataset

A comprehensive non-redundant dataset of both germline and somatic non-synonymous single nucleotide variations (nsSNVs) dataset from The Cancer Genome Atlas (TCGA))(http://cancergenome.nih.gov/), International Cancer Genome Consortium (ICGC) (33), IntOGen (39), CSR (40) cancer cell lines from NCI-60 panel (41), dbSNP (42) and UniProtKB/Swiss-Prot (43) was generated using methods described earlier (44). The final dataset contains 1,705,286 somatic nsSNVs and 1,132,832 germline nsSNVs and all data was loaded into the DrugVar database.

Integration of amino acid-drug binding dataset, nsSNVs dataset and additional drug related information

In order to identify the nsSNVs that affect amino acid-drug binding sites, amino acid-drug binding site entries and nsSNVs entries were integrated together if an nsSNV was found on amino acid-drug binding site between specific protein and drug. Also, additional drug IDs are retrieved from DrugBank or manually extracted from literature and other relevant databases. DrugBank ID provides detailed pharmacology and pharmaceutical knowledge and drug target information; CAS ID and CID ID can be used to explore the chemical information of relevant drug; PharmaGKB ID provides curated pharmacogenetic and pharmacogenomic data, such as knowledge of drug response on genetic variation and relevant clinical implementation. Figure 1 shows the workflow used to generate the final protein-drug interaction dataset.

Statistical significance survey of amino acid on drug binding site

To investigate the distribution of binding sites between amino acids and drugs, a statistical significance analysis was conducted. Significance of observed versus expected of amino acids in drug binding sites was calculated using Binomial statistic described by Mi et. al. and applied in our previous studies (44,45).

Protein Functional Class Enrichment analysis

Protein Functional Class Enrichment analysis was performed using PANTHER Classification System 9.0 (46).

Paralog identification for drug discovery

In the amino acid-drug binding site dataset, each protein was used to BLAST (47) against the human proteome to identify paralogs (48,49) which might also bind the drug. The binding sites in the paralogs were checked to ensure they are conserved. Paralogs where the binding sites are conserved were included into DrugVar. Homology modelings of these paralogs were carried out using MODELLER (50). During the homology modeling, the bound drug was kept in the binding pocket. These homology models of paralogs were energy minimized using the AMBER force field in MOE (Molecular Operating Environment, Chemical Computing Group, Quebec, Canada)

Structural and mutational modelling for key proteins / drugs

In the amino acid-drug PDB structures, different drugs bound to same protein or single drug bound to multiple proteins were identified. Visualization of the amino acids in the binding site for these amino acid-drug complexes was carried out using UCSF CHIMERA (51). To understand the effect of mutations on a drug binding to a protein target, we chose the x-ray structure of Carbonic anhydrase 2 bound to the drug Lacosamide (PDB ID: 3IEO) (52). Using MOE, we built a mutated protein (N67K, Q92P and F131L) with this drug following the above procedure as in case of paralog modeling. Proteins structural alterations and mutations, such as conservation and nsSNVs, were mapped to see their impact on amino acid-drug conformation change. The wild type amino acid-drug and the mutated amino acid-drug complexes were energy minimized using the AMBER force field in MOE. The binding energies of the same drug were computed for both these complexes.

Chapter 3: RESULTS

Identification of amino acid-drug binding sites and nsSNV sites

A total of 12,993 drug binding sites for 235 unique drugs in 827 PDB structures were identified. The 827 PDB structures were mapped to 253 UniProtKB/Swiss-Prot proteins. For each entry, binding site amino acid position from both PDB and UniProtKB sequences were collected. Out of 253 proteins in amino acid-drug binding dataset, 210 proteins had 1618 drug binding site altering nsSNVs. 1162 binding sites are impacted due to germline variations and 456 binding sites are affected by somatic mutations. Amino acid-drug binding sites and nsSNV sites are available from the DrugVar portal (http://hive.biochemistry.gwu.edu/tools/drugvar) and a snapshot of the web interface is shown in Figure 2. All drugs were classified into 14 groups based on Anatomical Therapeutic Chemical Classification (ATC). Table 1 shows an overview of the dataset based on the ATC Classification.

Antineoplastic drug binding sites and somatic mutation impact

Out of 12,993, 1408 protein binding sites are associated with 25 antineoplastic drugs (inhibiting or preventing the growth and spread of tumors or malignant cells) (ATC Code: L01). Some of them act by disrupting the metabolism of DNA synthesis components. For example, folic acid analogues, such as methotrexate and raltitrexed, inhibit key enzymes that are necessary for synthesis of thymidylate, an essential component of DNA (53,54). Similarly, purine analogues, such as cladribine, and pyrimidine analogues, such as Gemcitabine, inhibit cancer cell growth by disrupting DNA synthesis. Other drugs target proteins which are abnormally expressed due to

somatic mutations, such as nilotinib, that bind to ATP-binding site of BCR-ABL protein, a protein that only appears in cancer cells of chronic myelogenous leukemia patients, and inhibit the tumor growth.

25 unique antineoplastic agent drugs (ATC Code: L01) and 35 unique target proteins were further examined. Figure 3 shows a Circos plot (55) representing the amino acid-drug relationship. Imatinib, a tyrosine-kinase inhibitor used in the treatment of multiple cancers, most notably Philadelphia chromosome-positive chronic myelogenous leukemia, can bind to as many as eight proteins (gene symbols: ABL1, ABL2, DDR1, KIT, LCK, MAPK14, NQO2, SYK) (56-61). As one of the most well known targeted therapy drug for cancer treatment, imatinib blocks BCR-Abl enzyme, a product of translocation between chromosome 9 and chromosome 22 that almost only exists in acute myelogenous leukemia (AML), to cut off the energy source for cancer proliferation, therefore, inhibiting the tumor growth (56,62). It is also known that imatinib can also inhibit mast/stem cell growth factor receptor Kit (KIT) (57) to stop tumor growth. However, the effect of imatinib on other proteins shown in the Figure 3 still is not entirely known.

Sunitinib can bind to six proteins (CDK2, KDR, KIT, PHKG2, ITK, MAPK14). It inhibits cellular signaling by targeting tyrosine kinase including platelet-derived growth factor receptor beta, vascular endothelial growth factor receptor 1, mast/stem cell growth factor receptor Kit (KIT) and vascular endothelial growth factor receptor 2 (63-66). But

further studies are needed for the physiological outcome of its binding to rest of the proteins.

Deoxycytidine kinase can bind to four cancer therapy drugs (cladribine, gemcitabine, clofarabine, cytarabine). All of these drugs belong to antimetabolite that masquerade as purine or pyrimidine to block DNA synthesis. Evidences show that these four antimetabolites are phosphorylated by deoxycytidine kinase to the nucleotide cladribine triphosphate, which is incorporated into DNA in cells that have high level of deoxycytidine kinase (67-70).

In terms of nsSNVs within the antineoplastic drug target proteins, 608 out of 3133 nsSNVs are found on binding sites between 25 antineoplastic drugs and their 36 target proteins. 178 of them are somatic mutations and 429 of them are germline mutations. Gefitinib is primarily used in treatment of non-small cell lung cancer (71). In DrugVar dataset, six lung cancer associated somatic mutations exist within binding sites between epidermal growth factor receptor (EGFR; PDB: 2ITO) and gefitinib on amino acid positions 744, 766, 790, 792, 844 and 855. Also, a total number of seven germline mutations exist within binding sites between epidermal growth factor receptor and gefitinib on amino acid positions 719, 726, 743, 788, 796, 800 and 854. It is possible that these mutations on epidermal growth factor receptor may weaken binding efficacy of gefitinib to target protein in lung cancer patients. Therefore, for targeted therapy, one way to improve the therapy is to sequencing the cancer tissue to identify both germline and

somatic mutations and adjust the drug intake dose or considered the alternative drug based on the SNVs impact results.

It is reasonable to assume that the more proteins a drug binds to, the more pharmacogenetically or pharmacogenomically unstable the drug becomes. imatinib has eight target proteins in DrugVar database. Out of them, seven proteins have a total of 58 amino acid mutations in their binding sites with imatinib. As one of the primary target protein, mast/stem cell growth factor receptor Kit (KIT) is inhibited by imatinib (72) and contains 11 possible variations in its imatinib binding sites. These mutations may affect the drug binding affinity and make imatinib less effective under regular dose intake. Tyrosine-protein kinase ABL1 (ABL1) is another imatinib targeted protein and contains as much as 29 variations on binding sites. Out of 29 mutations, 19 of them are germline mutations and eight of them somatic mutations associated with diffuse large B-cell lymphoma. We report here many such amino acid variations in drug binding sites of proteins where the exact mechanism of action has not yet been elucidated. Such examples include epithelial discoidin domain-containing receptor 1 (DDR1) which is suggested to be a pharmacologically available target for cancer treatment (61,73), tyrosine-protein kinase Lck (LCK), tyrosine-protein kinase SYK (SYK), mitogen-activated protein kinase 14 (MAPK14) and abelson tyrosine-protein kinase 2 (ABL2).

To better understand the binding of the drugs we further analyzed the binding pocket of one of the drugs. Figure 4a shows the superposition of x-ray crystal structures of imatinib binding to its eight target proteins. The drug is located inside the protein pockets of the

protein represented by ribbon structures. The amino acid conservations are shown on the ribbon structures from high to low as the color becomes from red to blue through pink and cyan. Figure 4b shows the side chains which are identified as binding sites amino acids with imatinib which is marked as green color in the pocket of target proteins. The conservation of amino acids is shown the same as Figure 4a. The structural differences could help to modify imatinib to improve the selectivity for a given protein target. Figure 4c shows the amino acids that are reported as mutations for each target protein in the amino acid-drug binding sites The drug is located at the center of the proteins each of which is shown in different color for clarity.

Our study also emphasizes the importance of proteins which may bind to only few drugs, but still contain a large number of mutations on the drug binding sites. Epidermal growth factor receptor (EGFR) is a well-known protein target for antineoplastic drugs (74), especially, some drugs, such as lapatinib and erlotinib (75,76), have been designed as antagonist of epidermal growth factor receptor to prevent its activation required for tumor proliferation (76,77). However, we identified 25 mutations within epidermal growth factor receptor binding sites with lapatinib.

Significance survey of amino acid on antineoplastic drug-protein binding sites

In terms of the amino acid on binding site between antineoplastic drugs and their targets proteins, Table 2 shows the significance of amino acid on antineoplastic drug – protein binding sites. Within 20 amino acids, serine has the most significant over-representation (p value: 2.54E-29) across human proteome, whereas proline has most significantly

under-representation (p value: 9.45E-19) across human proteome. It is interestingly to note that, out of eight amino acids with hydrophobic side chains (alanine, valine, isoleucine, leucine, methionine, phenylalanine, tyrosine and tryptophan), seven of them are over-represented in amino acid-drug binding site on proteome level. Since drugs have to get out of solvent environments to bind to their respective protein targets, a part of the binding sites should be hydrophobic to interact with the hydrophobic part of the binding pocket. The binding of drugs should be tight-enough to perform their biological functions, aromatic residues such as phenylalanine, tryptophan and tyrosine amino acids are required. These aromatic residues will not tightly pack compared to aliphatic amino acids such as leucine and valine.

Gene Ontology analysis of the cancer drug protein targets using PANTHER (45) predictably reveals that all the top three significantly enriched terms show kinase relevance (non-receptor tyrosine protein kinase, protein kinase) which suggests the importance of kinase in anti-neoplastic drug target discovery. As the primary targeted therapy target protein class, a large number of antineoplastic drugs have been developed as kinase inhibitor to deregulate uncontrollable kinase activity, a major contributor to oncogenesis (78).

Identification of paralogs of antineoplastic drug target proteins

Within the 36 unique proteins that bind to 25 antineoplastic drugs, 5 proteins have at least one paralog which have all the binding sites conserved. The paralog search provides additional and potential drug targets. These paralogs can be useful based on the assumption that, they can be alternate drug targets and their expression and mutation

profiles could be evaluated in drug efficacy studies. Although extensive experiments have to be done to structurally modify the drug and to test the overall effect of drug on new proteins, because of the structural similarity between target protein and its paralogs, this approach requires much less time-consuming and expensive drug screening and therefore provides a new starting point and a shortcut for new drug target discovery.

Carbonic anhydrase 5A (CA5A), carbonic anhydrase 5B (CA5B), carbonic anhydrase 7 (CA7) and carbonic anhydrase 13 (CA13) are identified as four paralogs of carbonic anhydrase 2 (CA2) which is a target protein of hydroxyurea, an antineoplastic agent. These paralogs have conserved binding sites. Figure 5 shows a modelling of carbonic anhydrase 2 and its paralogs carbonic anhydrase 13 and carbonic anhydrase 7 binding to hydroxyurea (magenta color). The 3D structure shows the consistent binding between side chains of carbonic anhydrase 2, carbonic anhydrase 13, carbonic anhydrase 7 and hydroxyurea and implicates the potential binding affinity between carbonic anhydrase 5B, Carbonic anhydrase 5A and hydroxyurea. In DrugBank database, the only listed target for hydroxyurea is ribonucleodise-diphosphate reductase large subunit (79). In PharmaGKB, the only available target protein of hydroxyurea is mitogen-activated protein kinase kinase kinase 5. Therefore, the data we provide here is complimentary to what is available on DrugBank, PharmaGKB and other resources.

Non-antineoplastic drug binding sites and germline mutation impact

There are 11,124 amino acid-drug binding sites associated with 200 non-antineoplastic drugs. These 200 non-antineoplastic drugs can be further categorized into 13 subclasses.

Table 1 shows the number of proteins and drugs associated with each group categorized by ATC Classification System.

Out of total 3133 nsSNVs, 2704 are found on 200 non-antineoplastic drug binding sites of target proteins. 2125 of SNVs are germline mutations. One thing to note is that, unlike germline mutations, somatic mutations only exist within the cancer cells, they usually do not exist in cancer free individual. For cancer patients, drug efficacy can only be affected by somatic mutations if a large proportion of drug targeted cell are tumor cells. Therefore, for non-antineoplastic drug users, it is enough to extract easily accessible sample, such as saliva, for genetic test to identify the germline mutations that may affect drug efficacy.

There are seven proteins in our datasets that have seven or more drugs that can bind to them. The proteins are androgen receptor (AR), deoxycytidine kinase (DCK), aldo-keto reductase family 1 member C3 (AKR1C3), aldo-keto reductase family 1 member C2 (AKR1C2), transthyretin (TTR), carbonic anhydrase 2 (CA2) and serum albumin (ALB). Most of them play important role in drug transportation and metabolism pathway. For example, serum albumin (ALB), a major plasma protein functions as drug transporter, binds to 18 drugs and contains 97 amino acid variation sites within the binding sites for these drugs. Consistently, transthyretin (TTR), another major drug transporter, also binds to 11 drugs and contains 39 amino acid variations. Since drug that bind to the plasma protein will remain in the circulation and only serve as the reservoir before it is unbound from plasma protein, the binding affinity between drug and plasma protein directly

influence the biological half-life of drug. When the binding affinity is high than expected, the drug tends to remain inside the organism longer and have higher chance to cause side effect. On the other hand, if the binding affinity is low than expected, the unbound drug may rapidly be metabolized and excreted. In this case, the drug efficacy will be reduced. Also, many proteins that involve in drug metabolism pathway can bind to multiple drugs. Cytochrome P450 3A4 (P08684), the major enzyme that modify and detoxify drugs by oxidation reaction in liver, binds to 5 drugs and contains 28 variations within drug binding sites. Figure 6 shows the binding of different drugs (bromocriptine, erythromycin, metyrapone, ritonavir, progesterone) and cytochrome P450 3A4 (CYP3A4). The drugs shown as magenta color in the protein pocket bind to the amino acid residues which are shown as cyan and blue color. The blue colors mean the amino acids are altered due to nsSNV. The heme of Cytochrome P450 3A4 is marked as yellow color. Panels on the top of Figure 6 show that bromocriptine and erythromycin chemically bind to heme while panels on the bottom show that metyrapone, ritonavir and progesterone occupy the pocket without directly bind to them. Aldo-keto reductase family 1 member C2 (AKR1C2) and aldo-keto reductase family 1 member C3 (AKR1C3), the major enzymes to detoxify drugs by conjugation reaction, can bind to 10 and 8 drugs and contain 15 and 20 mutations within protein binding sites respectively. For these key proteins, we believe that the identified mutations on the drug-binding sites are one of the determinants for drug binding affinity, consequently, the drug efficacy and toxicity. As an example, Figure 7 shows the superposition of energy minimized structures for the wild-type (WT) carbonic anhydrase 2 bound to lacosamide (PDB: 3IEO) and the mutated models (N67K, Q92P and F131L) bound to the same drug. Only the WT (orange

color) and mutated residues (cyan color) with the bound drugs (yellow: WT and green color: mutant) are shown. Due to the mutations, the drug has to shift from the WT location to another location in the binding pocket which caused the binding energy unfavorable by 4 kcals compared to the WT binding. Using the mutational profile, it is possible to understand the difference between a drug bound to WT and the mutant protein. Therefore, the identification of their mutational profile contributes to their pharmacogenetic and pharmacogenomic characterizations.

Significance survey of amino acid on non-antineoplastic drug-protein binding sites

Table 3 provides an overview of the distribution of the amino acids in the non-antineoplastic drug-protein binding sites. Within the 20 amino acids, phenylalanine, tyrosine and tryptophan have the most significant over-representation (8.85E-116, 5.63E-87 and 2.73E-68) across human proteome, whereas proline has most significantly under-representation (4.50E-52) across human proteome. Seven out of eight amino acids with hydrophobic side chains show over-representation in amino acid-drug binding site on proteome level.

Table 4 shows the enrichment analysis of gene ontology terms that are enriched in the protein targets. Figure 8 shows the detailed non-antineoplastic drug target proteins. Reductase and oxidoreductase share a large number of proteins because reductase is a hierarchical child node of oxidoreductase. Proteins which fall into these two protein classes, such as aldo-keto reductase family, cytochrome P450 and glutathione S-transferase, are primarily responsible for xenobiotic metabolism and detoxication.

Transferase shares a small number of proteins with reductase and oxidoreductase because these proteins metabolize and detoxify xenobiotic by removing specific group of xenobiotic. Out of 5 protein classes, nuclear hormone receptor appears to be an independent protein class without overlap with other classes because of their distinct function.

Identification of paralogs of non-antineoplastic drug target proteins

Among 253 unique target proteins, we identified at least one paralog for 42 of them which have the drug binding site conserved. Usually paralogs have different expression levels in different tissues and cell types (80). These properties sometimes allow researchers to structurally modify the drugs to bind to target protein more selectively without causing side effect by binding to others. For example, carbonic anhydrase 2 (CA2) is one paralogs in the carbonic anhydrase family and exists along with several of its family memebrs ubiquitously across various tissues (81). It is also a well-known target for diuretic and glaucoma drugs, such as ethoxzolamide, acetazolamide and methazolamide (82). However, its existence in vascular tissue also causes hypertension as side effect by a number of drugs. Our study shows that carbonic anhydrase 2 (CA2) has paralogs Carbonic anhydrase 13 (CA13), Carbonic anhydrase 5A (CA5A), carbonic anhydrase 5B (CA5B) and Carbonic anhydrase 7 (CA7) which have conserved drug binding sites. Search for its druggable paralogs may lead to structurally modified alternative drugs that target its more tissue-selective paralogs and show similar pharmacological effect but less side effect.

Roflumilast shows antiinflammatory and antimodulatory effects in the pulmonary system by selectively inhibiting phosphodiesterase-4 (PDE-4) (83). Amino acid-drug binding data retrieved from PDB show the binding between cAMP-specific 3',5'-cyclic phosphodiesterase 4D (Q08499) and Roflumilast. Our paralogs survey shows that isoform of cAMP-specific 3',5'-cyclic phosphodiesterase 4D: cAMP-specific 3',5'-cyclic phosphodiesterase 4A (P27815), cAMP-specific 3',5'-cyclic phosphodiesterase 4B (Q07343) and cAMP-specific 3',5'-cyclic phosphodiesterase 4C (Q08493) are potential alternative target proteins as the replacement of cAMP-specific 3',5'-cyclic phosphodiesterase 4D.

DrugVar utility and access

All data from this study is available through DrugVar (http://hive.biochemistry.gwu.edu/tools/drugvar). A pipeline called DrugVar scan has been developed and publically available on DrugVar website. It allows users to upload a standard variation call formatted (VCF) file (http://vcftools.sourceforge.net/specs.html) and see if any of the variations can alter any of the drug binding sites.

In addition to scanning VCF files against DrugVar data the interface also supports both protein based (UniProtKB accession, PDB ID) and drug based (DrugBank ID, CAS ID, CID ID) searches. For each protein/drug of interest, amino acid-drug binding and variation information are shown (Figure 1). Both datasets show UniProtKB accession and PDB ID of protein and drug, the amino acid binding/mutation position of both UniProtKB and PDB (because the amino acid position between PDB and UniProtKB is

not always consistent), the drug name and ATC Classification Code for corresponding drug. In the mutation section, the amino acid variation and its associated cancer are provided and PharmGKB ID of corresponding drug links to The Pharmacogenomics Knowledge Base.

Both datasets show UniProtKB ID and PDB ID of proteins and drugs, the amino acid binding/mutation position of both UniProtKB and PDB (because the amino acid position between PDB and UniProtKB is not always consistent), the binding score, the drug name and ATC Classification Code for corresponding drug. The PDB ID links to Jmol for 3D structure display and ATC Classification Code links to WHO Collaborating Centre for Drug Statistics Methodology for description of drug usage. CID ID links to PubChem Compound Database where physical-chemical property of corresponding compound is provided. In addition, amino acid-drug binding section provides protein paralog which has high sequence similarity and identical amino acids on drug binding sites. In the mutation section, the amino acid variation and its associated cancer are provided.

Future perspectives

Although as comprehensive as the data we collected, limitation exists. There are many more drug-protein interactions than what are available structural databases such as PDB. It is expected that over time because of structural genomics initiatives more structural data will be available. We will update DrugVar at least once every year to capture all such new data. Additionally, since the target protein concentration influences amino acid-drug binding saturation, and further the drug efficacy and toxicity, information regarding

protein structure alteration alone is obviously insufficient for personalized medicine. Our future goal is to integrate specific gene expression data into DrugVar and its effect on drug binding as they become available through projects such as the Illumina Body Map project. Furthermore, non-human origin protein-drug binding structures will be collected from PDB and orthologs will be identified to predict protein-drug interactions.

List of Figures

Figure 1: DrugVar workflow to identify protein-drug binding sites and nsSNVs which affect amino acid-drug binding sites.

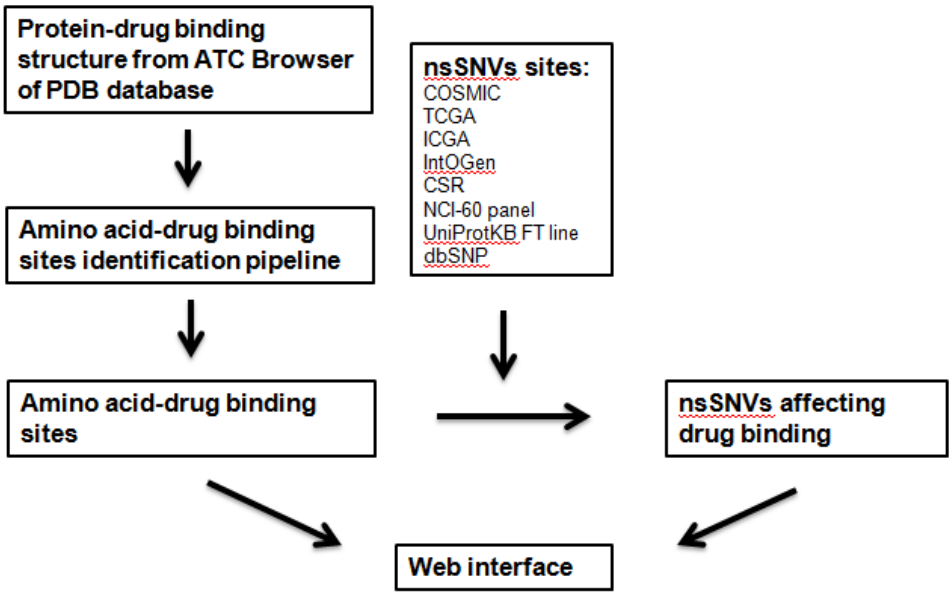
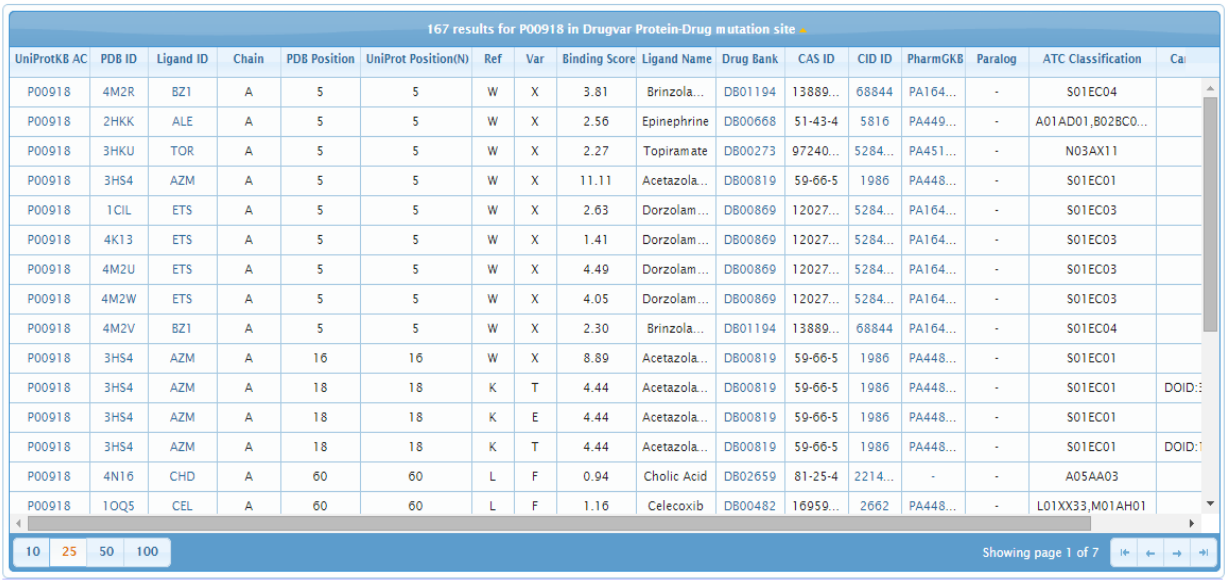


Figure 2: DrugVar database browser interface. The columns below are hyperlinked to source data wherever possible.



167 results for P00918 in Drugvar Protein-Drug mutation site

UniProtKB AC	PDB ID	Ligand ID	Chain	PDB Position	UniProt Position(N)	Ref	Var	Binding Score	Ligand Name	Drug Bank	CAS ID	CID ID	PharmGKB	Paralog	ATC Classification	Ca
P00918	4M2R	BZ1	A	5	5	W	X	3.81	Brinzola...	DB01194	13889...	68844	PA164...	-	S01EC04	
P00918	2HKK	ALE	A	5	5	W	X	2.56	Epinephrine	DB00668	51-43-4	5816	PA449...	-	A01AD01,B02BC0...	
P00918	3HKU	TOR	A	5	5	W	X	2.27	Topiramate	DB00273	97240...	5284...	PA451...	-	N03AX11	
P00918	3HS4	AZM	A	5	5	W	X	11.11	Acetazola...	DB00819	59-66-5	1986	PA448...	-	S01EC01	
P00918	1CIL	ETS	A	5	5	W	X	2.63	Dorzolam...	DB00869	12027...	5284...	PA164...	-	S01EC03	
P00918	4K13	ETS	A	5	5	W	X	1.41	Dorzolam...	DB00869	12027...	5284...	PA164...	-	S01EC03	
P00918	4M2U	ETS	A	5	5	W	X	4.49	Dorzolam...	DB00869	12027...	5284...	PA164...	-	S01EC03	
P00918	4M2W	ETS	A	5	5	W	X	4.05	Dorzolam...	DB00869	12027...	5284...	PA164...	-	S01EC03	
P00918	4M2V	BZ1	A	5	5	W	X	2.30	Brinzola...	DB01194	13889...	68844	PA164...	-	S01EC04	
P00918	3HS4	AZM	A	16	16	W	X	8.89	Acetazola...	DB00819	59-66-5	1986	PA448...	-	S01EC01	
P00918	3HS4	AZM	A	18	18	K	T	4.44	Acetazola...	DB00819	59-66-5	1986	PA448...	-	S01EC01	DOID:3
P00918	3HS4	AZM	A	18	18	K	E	4.44	Acetazola...	DB00819	59-66-5	1986	PA448...	-	S01EC01	
P00918	3HS4	AZM	A	18	18	K	T	4.44	Acetazola...	DB00819	59-66-5	1986	PA448...	-	S01EC01	DOID:1
P00918	4N16	CHD	A	60	60	L	F	0.94	Cholic Acid	DB02659	81-25-4	2214...	-	-	A05AA03	
P00918	1OQ5	CEL	A	60	60	L	F	1.16	Celecoxib	DB00482	16959...	2662	PA448...	-	L01XX33,M01AH01	

10 25 50 100

Showing page 1 of 7

Figure 3: Circus plot representing the binding between antineoplastic drugs and their target proteins. Proteins are presented as UniProKB ID. All antineoplastic agents bind to at least one protein. Each antineoplastic agent is assigned to a unique color and the ribbon width indicates the amount of target protein**.**

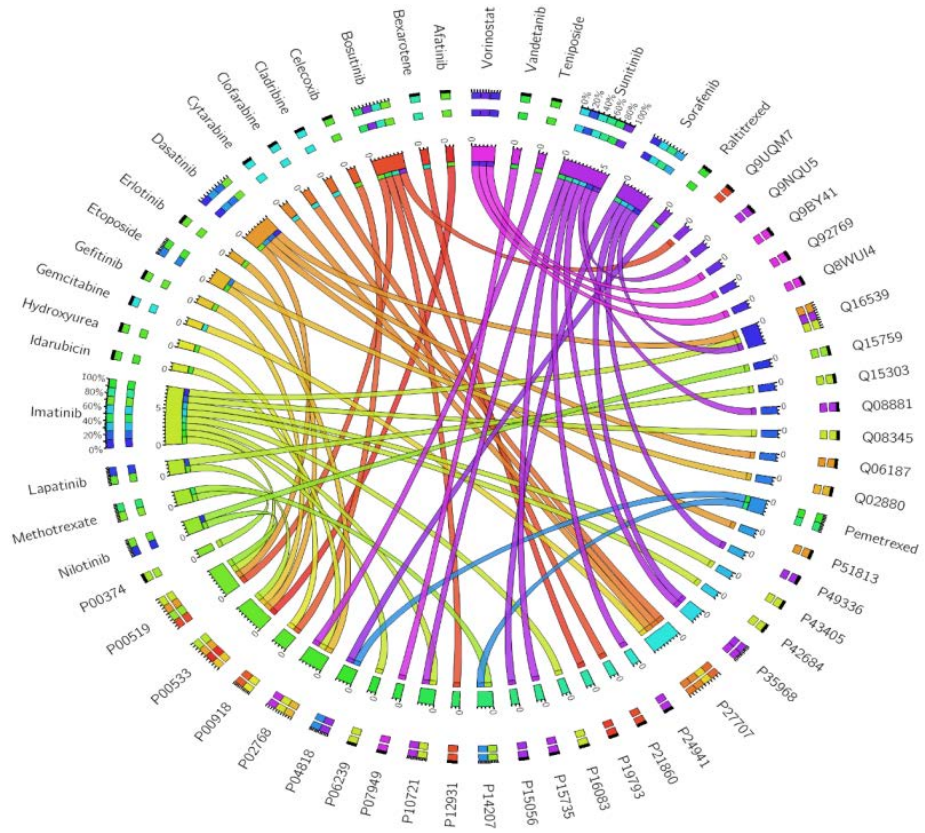


Figure 4: Superposition of x-ray crystal structures of imatinib binding to its target proteins.

4a Shows superposition (c-alpha atoms) of imatinib binding to eight target protein x-ray structures (ABL1, LCK, KIT, NQO2, ABL2, SYK, DDR1 and MAPK14). The superimposed proteins protein structures are shown by color coded ribbon. The blue color shows the lowest amino acid conservation region and the red color shows the highest amino acid conservation region. The ligand is shown bound to protein pockets. **4b** Shows the imatinib binding to the same target proteins as shown in Figure 4a. Only the side chains of binding sites are shown. The colors on protein from red to blue side chain show conservation. **4c** Shows imatinib binding to its target proteins. The side chains of proteins are imatinib binding sites that are mutated.

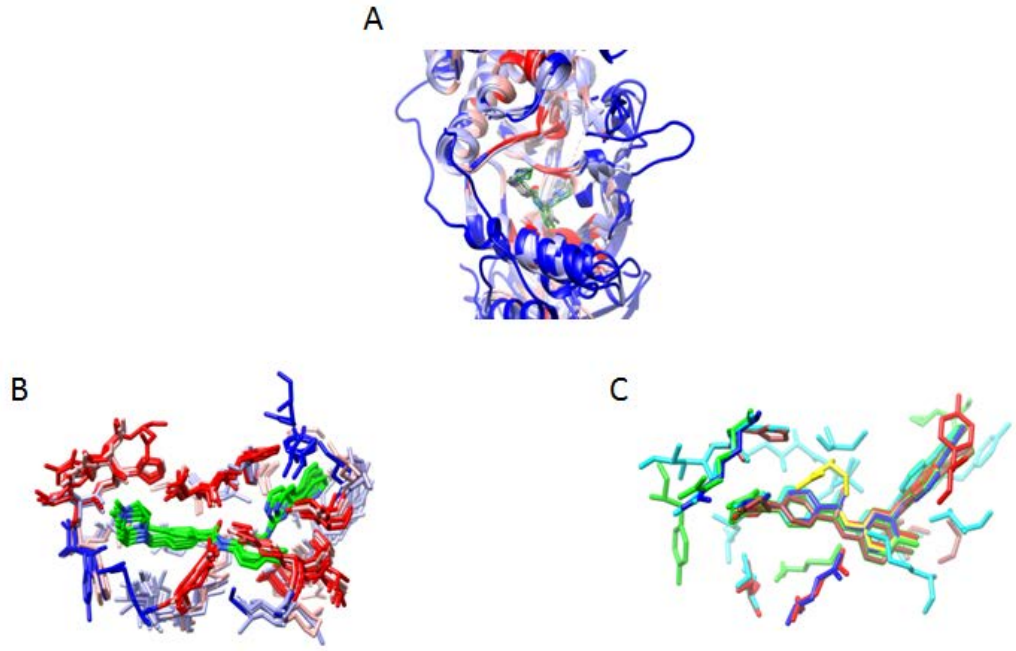


Figure 5: Modelling of carbonic anhydrase 2 and its paralogs (carbonic anhydrase 13 and carbonic anhydrase 7) binding to hydroxyurea.

The carbon structure of P00918 and its paralogs is shown as pink color. The hydroxyurea in the protein pocket shown as magenta color and it bind to amino acid residues which are conserved across P00918 and its paralogs

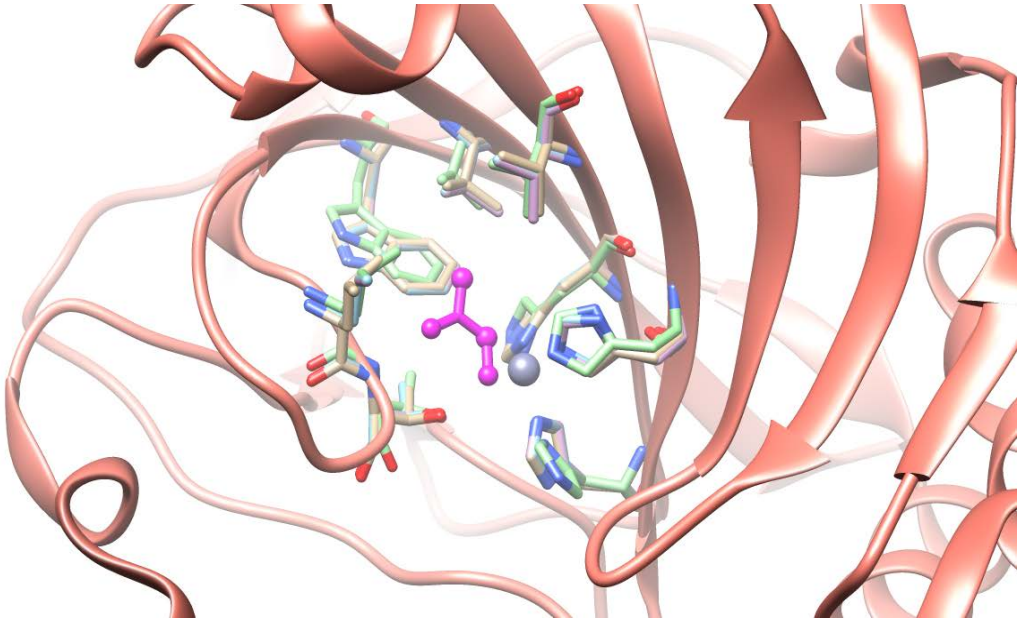


Figure 6: Modelling of cytochrome P450 3A4 binding to its target proteins (bromocriptine, erythromycin, metyrapone, ritonavir, progesterone).The drugs shown as magenta color in the protein pocket bind to the amino acid residues which shown as cyan except that mutated amino acids are marked as blue. The yellow color is the hem of Cytochrome P450 3A4.

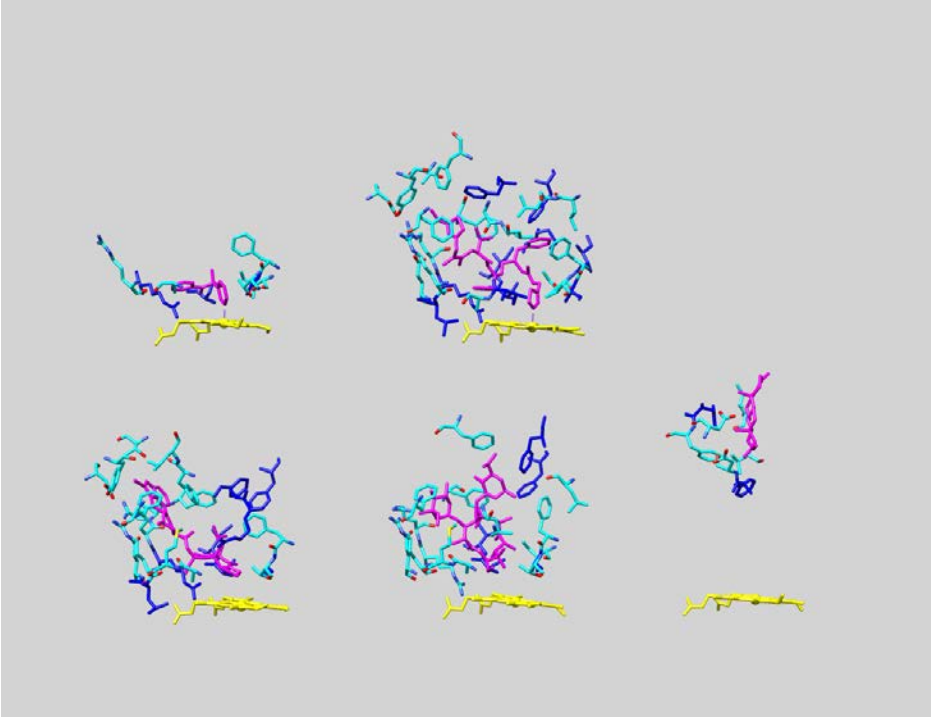


Figure 7: Superposition of energy minimized structures for the WT carbonic anhydrase 2 bound to lacosamide (PDB: 3IEO) and the mutated models (N67K, Q92P and F131L) bound to the same drug.

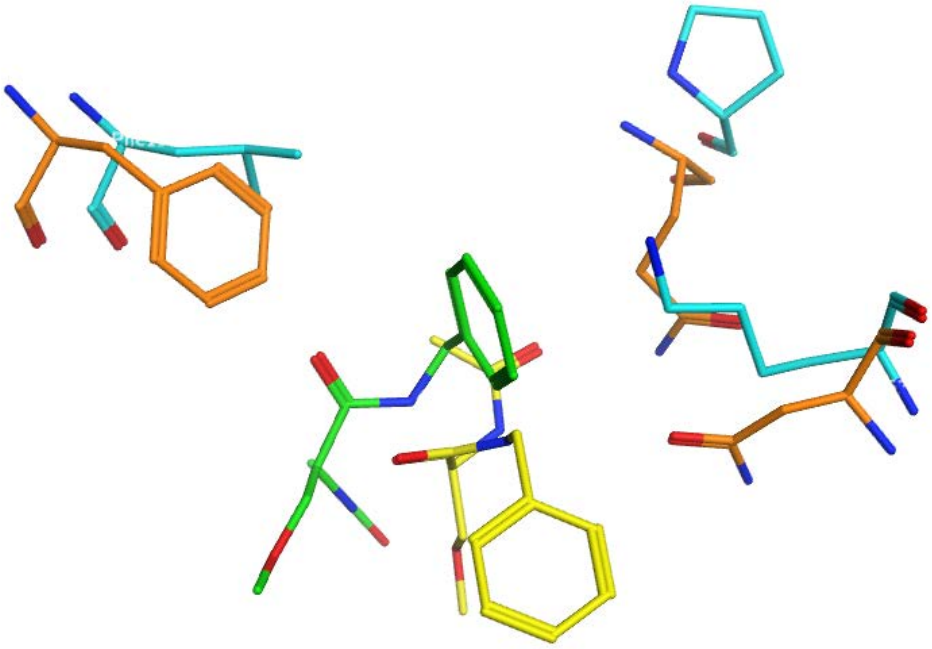
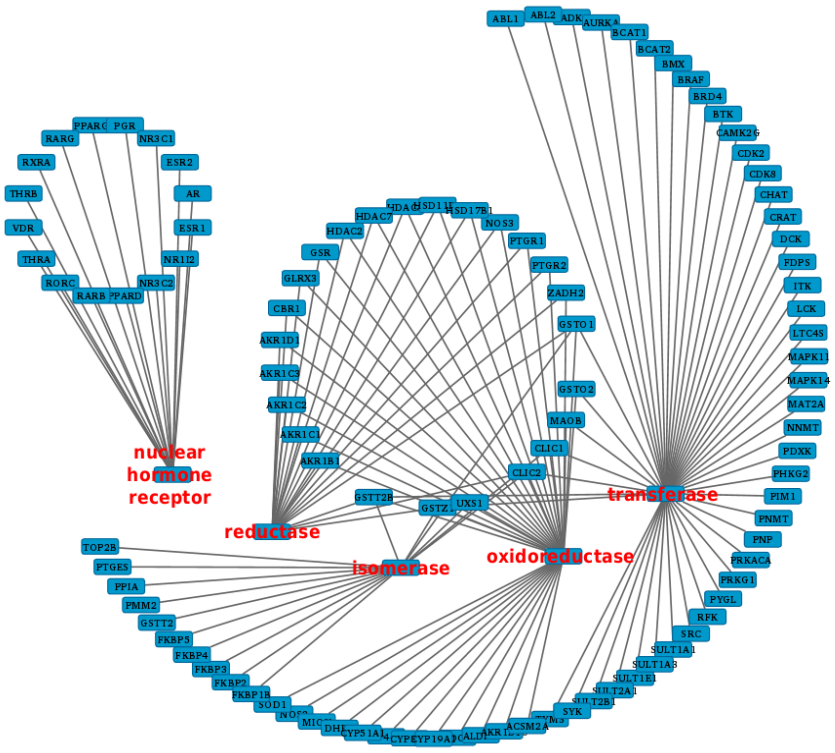


Figure 8: Detailed non-neoplastic drug target proteins classification using gene ontology enrichment analysis

Detailed non-neoplastic drug target proteins classification using gene ontology enrichment analysis which shows non-neoplastic drug target proteins belonging to top 5 protein classes (reductase, nuclear hormone receptor, transferase, oxidoreductase and isomerase) from gene enrichment analysis. Each blue box represents a protein connected to its protein class.



List of Tables

Table 1: Statistical summary of dataset based on the ATC Classification.

ATC Classification	Number of Protein	Number of Drug
Cancer drugs		
Antineoplastic agents	36	25
Non-cancer drugs		
Alimentary tract and metabolism	56	36
Blood and blood forming organs	28	9
Cardiovascular system	43	33
Dermatologicals	23	15
Genito-urinary system and sex hormones	28	17
Systemic hormonal preparations, excluding sex hormones and insulins	8	4
Antiinfectives for systemic use	11	13
Musculo-skeletal system	18	22
Nervous system	26	26
Antiparasitic products, insecticides and repellents	8	9
Respiratory system	13	5
Sensory organs	9	7
Antidote[1]	28	4

[1]Antidote is a subclass of "Various" (ATC Code: V).

Table 2: Survey of significance of amino acid on antineoplastic protein-drug binding site across proteome

Amino acid	Number of AA on binding site	Number of AA on proteome level	Ratio	Expected	Difference	P-value
Ala	89	793727	7.01E-02	98.82	-9.82	1.66E-01
Arg	52	638708	5.64E-02	79.52	-27.52	4.89E-04
Asp	83	536167	4.74E-02	66.75	16.25	2.71E-02
Cys	32	259982	2.30E-02	32.37	-0.37	5.21E-01
Gln	30	539313	4.77E-02	67.14	-37.14	1.78E-07
Glu	77	803284	7.10E-02	100.01	-23.01	7.99E-03
His	53	297674	2.63E-02	37.06	15.94	7.22E-03
Ile	117	491109	4.34E-02	61.14	55.86	4.50E-11
Gly	84	743824	6.57E-02	92.61	-8.61	1.93E-01
Asn	19	406270	3.59E-02	50.58	-31.58	2.29E-07
Leu	198	1127317	9.96E-02	140.35	57.65	6.75E-07
Lys	62	647773	5.72E-02	80.65	-18.65	1.60E-02
Met	72	241049	2.13E-02	30.01	41.99	3.11E-11
Phe	103	413386	3.65E-02	51.47	51.53	6.35E-11
Pro	21	714069	6.31E-02	88.9	-67.9	9.45E-19
Ser	21	941329	8.32E-02	117.2	-96.2	2.54E-29
Thr	62	606366	5.36E-02	75.49	-13.49	5.89E-02
Trp	30	137896	1.22E-02	17.17	12.83	2.95E-03
Tyr	63	301607	2.67E-02	37.55	25.45	7.30E-05
Val	141	675593	5.97E-02	84.11	56.89	2.75E-09

Total number of AA on proteome level is 11317281.

Total number of binding site is 1409

Table 3: Survey of significance of amino acid on non-antineoplastic protein-drug binding site across proteome

Amino acid	Number of AA on binding site	Number of AA on proteome level	Ratio	Expected	Difference	P-value
Ala	335	793727	7.01E-02	475.93	-140.93	1.13E-12
Arg	319	638708	5.64E-02	382.98	-63.98	3.08E-04
Asp	190	536167	4.74E-02	321.49	-131.49	3.59E-16
Cys	157	259982	2.30E-02	155.89	1.11	4.75E-01
Gln	221	539313	4.77E-02	323.38	-102.38	4.32E-10
Glu	258	803284	7.10E-02	481.66	-223.66	6.75E-31
His	328	297674	2.63E-02	178.49	149.51	1.56E-24
Ile	418	491109	4.34E-02	294.48	123.52	2.24E-12
Gly	286	743824	6.57E-02	446.01	-160.01	4.32E-17
Asn	190	406270	3.59E-02	243.61	-53.61	1.66E-04
Leu	845	1127317	9.96E-02	675.96	169.04	1.94E-11
Lys	205	647773	5.72E-02	388.41	-183.41	7.43E-26
Met	322	241049	2.13E-02	144.54	177.46	4.17E-38
Phe	673	413386	3.65E-02	247.87	425.13	8.85E-116
Pro	161	714069	6.31E-02	428.17	-267.17	4.50E-52
Ser	309	941329	8.32E-02	564.43	-255.43	2.29E-34
Thr	308	606366	5.36E-02	363.59	-55.59	1.19E-03
Trp	284	137896	1.22E-02	82.68	201.32	2.73E-68
Tyr	498	301607	2.67E-02	180.85	317.15	5.63E-87
Val	465	675593	5.97E-02	405.1	59.9	1.42E-03

Total number of AA on proteome level is 11317281.

Total number of binding site is 6786

Table 4: Statistical overrepresentation of non-neoplastic drug target protein class

Protein Class	Homo sapiens	Drug Binding Protein	Expected	Actual	P-value
Reductase	211	243	108.42	63	5.85E-18
Nuclear hormone receptor	48	243	2.35	27	1.65E-16
Oxidoreductase	639	243	0.35	16	1.22E-14
Transferase	1313	243	7.12	38	4.46E-12
Isomerase	168	243	14.63	50	2.57E-10

References

1. Venter, J.C., Levy, S., Stockwell, T., Remington, K. and Halpern, A. (2003) Massive parallelism, randomness and genomic advances. *Nature genetics*, **33 Suppl**, 219-227.
2. Zhang, J., Chiodini, R., Badr, A. and Zhang, G. (2011) The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*, **38**, 95-109.
3. Gullapalli, R.R., Lyons-Weiler, M., Petrosko, P., Dhir, R., Becich, M.J. and LaFramboise, W.A. (2012) Clinical integration of next-generation sequencing technology. *Clinics in laboratory medicine*, **32**, 585-599.
4. Bahassi el, M. and Stambrook, P.J. (2014) Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, **29**, 303-310.
5. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, **36**, D13-21.
6. Pavlopoulos, G.A., Oulas, A., Iacucci, E., Sifrim, A., Moreau, Y., Schneider, R., Aerts, J. and Iliopoulos, I. (2013) Unraveling genomic variation from next generation sequencing data. *BioData mining*, **6**, 13.
7. Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic acids research*, **30**, 3894-3900.
8. Nakken, S., Alseth, I. and Rognes, T. (2007) Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. *Neuroscience*, **145**, 1273-1279.
9. Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S. and Sunyaev, S. (2013) Sequencing studies in human genetics: design and interpretation. *Nature reviews. Genetics*, **14**, 460-470.
10. Giacomini, K.M., Brett, C.M., Altman, R.B., Benowitz, N.L., Dolan, M.E., Flockhart, D.A., Johnson, J.A., Hayes, D.F., Klein, T., Krauss, R.M. *et al.* (2007) The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clinical pharmacology and therapeutics*, **81**, 328-345.
11. Evans, W.E. and Relling, M.V. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*, **286**, 487-491.
12. McLeod, H.L. and Yu, J. (2003) Cancer pharmacogenomics: SNPs, chips, and the individual patient. *Cancer investigation*, **21**, 630-640.
13. Shastry, B.S. (2003) SNPs and haplotypes: genetic markers for disease and drug response (review). *International journal of molecular medicine*, **11**, 379-382.
14. Geisler, T., Schaeffeler, E., Gawaz, M. and Schwab, M. (2013) Genetic variation of platelet function and pharmacology: an update of current knowledge. *Thrombosis and haemostasis*, **110**, 876-887.

15. Brandl, E.J., Chowdhury, N.I., Tiwari, A.K., Lett, T.A., Meltzer, H.Y., Kennedy, J.L. and Muller, D.J. (2015) Genetic variation in CYP3A43 is associated with response to antipsychotic medication. *J Neural Transm*, **122**, 29-34.
16. Sun, H.Y., Ji, F.Q., Fu, L.Y., Wang, Z.Y. and Zhang, H.Y. (2013) Structural and energetic analyses of SNPs in drug targets and implications for drug therapy. *Journal of chemical information and modeling*, **53**, 3343-3351.
17. Szelke, M., Leckie, B., Hallett, A., Jones, D.M., Sueiras, J., Atrash, B. and Lever, A.F. (1982) Potent new inhibitors of human renin. *Nature*, **299**, 555-557.
18. Blundell, T., Sibanda, B.L. and Pearl, L. (1983) Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature*, **304**, 273-275.
19. Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J.R., Whittle, P.J., Danley, D.E., Geoghegan, K.F. *et al.* (1989) X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. *Nature*, **342**, 299-302.
20. Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B. and Wlodawer, A. (1989) Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science*, **246**, 1149-1152.
21. Wlodawer, A. (2002) Rational approach to AIDS drug design through structural biology. *Annual review of medicine*, **53**, 595-614.
22. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S. *et al.* (2002) The Protein Data Bank. *Acta crystallographica. Section D, Biological crystallography*, **58**, 899-907.
23. Higueruelo, A.P., Schreyer, A., Bickerton, G.R., Pitt, W.R., Groom, C.R. and Blundell, T.L. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chemical biology & drug design*, **74**, 457-467.
24. Winter, A., Higueruelo, A.P., Marsh, M., Sigurdardottir, A., Pitt, W.R. and Blundell, T.L. (2012) Biophysical and computational fragment-based approaches to targeting protein-protein interactions: applications in structure-guided drug discovery. *Quarterly reviews of biophysics*, **45**, 383-426.
25. Idrees, S. and Ashfaq, U.A. (2014) Discovery and design of cyclic peptides as dengue virus inhibitors through structure-based molecular docking. *Asian Pacific journal of tropical medicine*, **7**, 513-516.
26. Schreyer, A.M. and Blundell, T.L. (2013) CREDO: a structural interactomics database for drug discovery. *Database : the journal of biological databases and curation*, **2013**, bat049.
27. Lopez, G., Valencia, A. and Tress, M. (2007) FireDB--a database of functionally important residues from proteins of known structure. *Nucleic acids research*, **35**, D219-223.
28. Kalow, W., Tang, B.K. and Endrenyi, L. (1998) Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics*, **8**, 283-289.
29. Mango, R., Vecchione, L., Raso, B., Borgiani, P., Brunetti, E., Mehta, J.L., Lauro, R., Romeo, F. and Novelli, G. (2005) Pharmacogenomics in cardiovascular

disease: the role of single nucleotide polymorphisms in improving drug therapy. *Expert opinion on pharmacotherapy*, **6**, 2565-2576.
30. Wang, Z., Wang, J., Tantoso, E., Wang, B., Tai, A.Y., Ooi, L.L., Chong, S.S. and Lee, C.G. (2007) Signatures of recent positive selection at the ATP-binding cassette drug transporter superfamily gene loci. *Human molecular genetics*, **16**, 1367-1380.
31. Thorn, C.F., Klein, T.E. and Altman, R.B. (2010) Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, **11**, 501-505.
32. Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A. *et al.* (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic acids research*, **38**, D652-657.
33. Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993-998.
34. Smigielski, E.M., Sirotkin, K., Ward, M. and Sherry, S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research*, **28**, 352-355.
35. 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 research*, **43**, D345-356.
36. UniProt_Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*, **41**, D43-47.
37. Huang, H., McGarvey, P.B., Suzek, B.E., Mazumder, R., Zhang, J., Chen, Y. and Wu, C.H. (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics*, **27**, 1190-1191.
38. Pattabiraman, N., Ward, K.B. and Fleming, P.J. (1995) Occluded molecular surface: analysis of protein packing. *Journal of molecular recognition : JMR*, **8**, 334-344.
39. Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A. and Lopez-Bigas, N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods*, **10**, 1081-1082.
40. Cole, C., Krampis, K., Karagiannis, K., Almeida, J.S., Faison, W.J., Motwani, M., Wan, Q., Golikov, A., Pan, Y., Simonyan, V. *et al.* (2014) Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data. *BMC bioinformatics*, **15**, 28.
41. Abaan, O.D., Polley, E.C., Davis, S.R., Zhu, Y.J., Bilke, S., Walker, R.L., Pineda, M., Gindin, Y., Jiang, Y., Reinhold, W.C. *et al.* (2013) The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer research*, **73**, 4372-4382.
42. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, **29**, 308-311.

43. UniProt_Consortium. (2015) UniProt: a hub for protein information. *Nucleic acids research*, **43**, D204-212.
44. Wu, T.J., Shamsaddini, A., Pan, Y., Smith, K., Crichton, D.J., Simonyan, V. and Mazumder, R. (2014) A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). *Database : the journal of biological databases and curation*, **2014**, bau022.
45. Mi, H. and Thomas, P. (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*, **563**, 123-140.
46. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, **13**, 2129-2141.
47. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389-3402.
48. Mazumder, R. and Vasudevan, S. (2008) Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function. *PLoS computational biology*, **4**, e1000151.
49. Mazumder, R., Vasudevan, S. and Nikolskaya, A.N. (2008) Protein functional annotation by homology. *Methods Mol Biol*, **484**, 465-490.
50. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U. and Sali, A. (2006) Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 5**, Unit 5 6.
51. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, **25**, 1605-1612.
52. Temperini, C., Innocenti, A., Scozzafava, A., Parkkila, S. and Supuran, C.T. (2010) The coumarin-binding site in carbonic anhydrase accommodates structurally diverse inhibitors: the antiepileptic lacosamide as an example and lead molecule for novel classes of carbonic anhydrase inhibitors. *Journal of medicinal chemistry*, **53**, 850-854.
53. Uga, H., Kuramori, C., Ohta, A., Tsuboi, Y., Tanaka, H., Hatakeyama, M., Yamaguchi, Y., Takahashi, T., Kizaki, M. and Handa, H. (2006) A new mechanism of methotrexate action revealed by target screening with affinity beads. *Molecular pharmacology*, **70**, 1832-1839.
54. Kindler, H.L. (2002) Pemetrexed in pancreatic cancer. *Seminars in oncology*, **29**, 49-53.
55. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome research*, **19**, 1639-1645.
56. Nadal, E. and Olavarria, E. (2004) Imatinib mesylate (Gleevec/Glivec) a molecular-targeted therapy for chronic myeloid leukaemia and other malignancies. *International journal of clinical practice*, **58**, 511-516.

57. Lee, J.L., Kim, J.Y., Ryu, M.H., Kang, H.J., Chang, H.M., Kim, T.W., Lee, H., Park, J.H., Kim, H.C., Kim, J.S. *et al.* (2006) Response to imatinib in KIT- and PDGFRA-wild type gastrointestinal stromal associated with neurofibromatosis type 1. *Digestive diseases and sciences*, **51**, 1043-1046.
58. de Groot, J.W., Plaza Menacho, I., Schepers, H., Drenth-Diephuis, L.J., Osinga, J., Plukker, J.T., Links, T.P., Eggen, B.J. and Hofstra, R.M. (2006) Cellular effects of imatinib on medullary thyroid cancer cells harboring multiple endocrine neoplasia Type 2A and 2B associated RET mutations. *Surgery*, **139**, 806-814.
59. Delbaldo, C. (2007) [Pharmacokinetic-pharmacodynamics relationships of imatinib (Glivec)]. *Therapie*, **62**, 87-90.
60. Dewar, A.L., Farrugia, A.N., Condina, M.R., Bik To, L., Hughes, T.P., Vernon-Roberts, B. and Zannettino, A.C. (2006) Imatinib as a potential antiresorptive therapy for bone disease. *Blood*, **107**, 4334-4337.
61. Xu, L., Tong, R., Cochran, D.M. and Jain, R.K. (2005) Blocking platelet-derived growth factor-D/platelet-derived growth factor receptor beta signaling inhibits human renal cell carcinoma progression in an orthotopic mouse model. *Cancer research*, **65**, 5711-5719.
62. Waller, C.F. (2010) Imatinib mesylate. *Recent results in cancer research. Fortschritte der Krebsforschung. Progres dans les recherches sur le cancer*, **184**, 3-20.
63. Mendel, D.B., Laird, A.D., Xin, X., Louie, S.G., Christensen, J.G., Li, G., Schreck, R.E., Abrams, T.J., Ngai, T.J., Lee, L.B. *et al.* (2003) In vivo antitumor activity of SU11248, a novel tyrosine kinase inhibitor targeting vascular endothelial growth factor and platelet-derived growth factor receptors: determination of a pharmacokinetic/pharmacodynamic relationship. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **9**, 327-337.
64. O'Farrell, A.M., Foran, J.M., Fiedler, W., Serve, H., Paquette, R.L., Cooper, M.A., Yuen, H.A., Louie, S.G., Kim, H., Nicholas, S. *et al.* (2003) An innovative phase I clinical study demonstrates inhibition of FLT3 phosphorylation by SU11248 in acute myeloid leukemia patients. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **9**, 5465-5476.
65. Pietras, K. and Hanahan, D. (2005) A multitargeted, metronomic, and maximum-tolerated dose "chemo-switch" regimen is antiangiogenic, producing objective responses and survival benefit in a mouse model of cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **23**, 939-952.
66. Amino, N., Ideyama, Y., Yamano, M., Kuromitsu, S., Tajinda, K., Samizu, K., Hisamichi, H., Matsuhisa, A., Shirasuna, K., Kudoh, M. *et al.* (2006) YM-359445, an orally bioavailable vascular endothelial growth factor receptor-2 tyrosine kinase inhibitor, has highly potent antitumor activity against established tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **12**, 1630-1638.
67. Lamba, J.K. (2009) Genetic factors influencing cytarabine therapy. *Pharmacogenomics*, **10**, 1657-1674.

68. Larson, M.L. and Venugopal, P. (2009) Clofarabine: a new treatment option for patients with acute myeloid leukemia. *Expert opinion on pharmacotherapy*, **10**, 1353-1357.
69. Wong, A., Soo, R.A., Yong, W.P. and Innocenti, F. (2009) Clinical pharmacology and pharmacogenetics of gemcitabine. *Drug metabolism reviews*, **41**, 77-88.
70. Zhenchuk, A., Lotfi, K., Juliusson, G. and Albertioni, F. (2009) Mechanisms of anti-cancer action and pharmacology of clofarabine. *Biochemical pharmacology*, **78**, 1351-1359.
71. Ciardiello, F., Caputo, R., Bianco, R., Damiano, V., Pomatico, G., De Placido, S., Bianco, A.R. and Tortora, G. (2000) Antitumor effect and potentiation of cytotoxic drugs activity in human cancer cells by ZD-1839 (Iressa), an epidermal growth factor receptor-selective tyrosine kinase inhibitor. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **6**, 2053-2063.
72. De Giorgi, U. (2007) KIT mutations and imatinib dose effects in patients with gastrointestinal stromal tumors. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **25**, 1146-1147; author reply 1147-1148.
73. Gotlib, J., Berube, C., Growney, J.D., Chen, C.C., George, T.I., Williams, C., Kajiguchi, T., Ruan, J., Lilleberg, S.L., Durocher, J.A. *et al.* (2005) Activity of the tyrosine kinase inhibitor PKC412 in a patient with mast cell leukemia with the D816V KIT mutation. *Blood*, **106**, 2865-2870.
74. Langer, C.J. (2004) Emerging role of epidermal growth factor receptor inhibition in therapy for advanced malignancy: focus on NSCLC. *International journal of radiation oncology, biology, physics*, **58**, 991-1002.
75. Medina, P.J. and Goodin, S. (2008) Lapatinib: a dual inhibitor of human epidermal growth factor receptor tyrosine kinases. *Clinical therapeutics*, **30**, 1426-1447.
76. Tevaarwerk, A.J. and Kolesar, J.M. (2009) Lapatinib: a small-molecule inhibitor of epidermal growth factor receptor and human epidermal growth factor receptor-2 tyrosine kinases used in the treatment of breast cancer. *Clinical therapeutics*, **31 Pt 2**, 2332-2348.
77. Bulgaru, A.M., Mani, S., Goel, S. and Perez-Soler, R. (2003) Erlotinib (Tarceva): a promising drug targeting epidermal growth factor receptor tyrosine kinase. *Expert review of anticancer therapy*, **3**, 269-279.
78. Zhang, J., Yang, P.L. and Gray, N.S. (2009) Targeting cancer with small molecule kinase inhibitors. *Nature reviews. Cancer*, **9**, 28-39.
79. Jiang, W., Xie, J., Varano, P.T., Krebs, C. and Bollinger, J.M., Jr. (2010) Two distinct mechanisms of inactivation of the class Ic ribonucleotide reductase from Chlamydia trachomatis by hydroxyurea: implications for the protein gating of intersubunit electron transfer. *Biochemistry*, **49**, 5340-5349.
80. Padawer, T., Leighty, R.E. and Wang, D. (2012) Duplicate gene enrichment and expression pattern diversification in multicellularity. *Nucleic acids research*, **40**, 7597-7605.
81. Supuran, C.T. (2007) Carbonic anhydrases as drug targets--an overview. *Current topics in medicinal chemistry*, **7**, 825-833.

82. Mincione, F., Scozzafava, A. and Supuran, C.T. (2008) The development of topically acting carbonic anhydrase inhibitors as antiglaucoma agents. *Current pharmaceutical design*, **14**, 649-654.
83. Barone, F.C., Barton, M.E., White, R.F., Legos, J.J., Kikkawa, H., Shimamura, M., Kuratani, K. and Kinoshita, M. (2008) Inhibition of phosphodiesterase type 4 decreases stress-induced defecation in rats and mice. *Pharmacology*, **81**, 11-17.