A landscape of population-specific variation effects on ligand binding site in the healthy population

By:

Mo Sun

A Thesis
Submitted to the Faculty
Of

Worcester Polytechnic Institute

In partial fulfillment of the requirements for the

Degree of Master of Science

In

Bioinformatics and Computational Biology

May 2021

APPROVED:

____
Dr. Dmitry Korkin, Major Advisor

____
Dr. Zheyang Wu, Reader

Abstract:

There has been of considerable interest in understanding how the genetic mutations could rewire the macromolecular interaction network mediated by proteins. Protein-ligand interactions, as one of the macromolecular interactions, can be associated with mutations at the ligand-binding sites (LBSs) to influencing protein structure stability, binding affinity with small molecules, hormone regulation, and drug resistance in people. Recent studies have shown that ligand binding residues have a significantly higher mutation rate than other parts of the protein in 16 types of cancers. Our study focuses on LBSs mutation in healthy populations and makes the most comprehensive human LBSs mutation map to date. By integrating BioLip and gnomAD data, we created a comprehensive LBSs-associated mutation map in eight populations and to help exploring the relationship between the variation profiles of different populations and the ligand-binding region corresponding to different types of ligands. Specifically, each LBSs was annotated and grouped into different categories based on the molecular structure and function similarity of the corresponding ligand, which allows looking for common ligand mutation patterns. In our work, we hypothesized that LBSs of proteins are enriched with the population-specific mutations, which means that the frequencies of these mutations are not evenly distributed in every population. Furthermore, we test if the specific categories of LBSs are not equally susceptible to variation across different healthy groups. We observed distinct mutation across different groups of ligands. As a case study, we determined 20 variants in 14 pharmacological genes from PharmGKB (https://www.pharmgkb.org/) VIPs to study the deleterious, neutral, and beneficial effect on the LBSs of nicotine, phenytoin, and other drugs. Meanwhile, further study is needed to determine whether all the genetic variants in our map would damage the normal interaction and what the specific mechanism about on how they would damage

Acknowledgments:

As time goes by, I never imagine that the two-year master's study is coming to an end. Here I want to express my sincere thanks to my teachers, friends who have helped me, and my parents. Thank you for your help in the past two years.

First of all, I would like to thank my advisor Dr. Dmitry Korkin, who was also the first professor I got to know in WPI. It was he who led me into the door of bioinformatics and opened the door to a whole new field for me. In the past two years, his guidance has made me understand that being good at finding problems is as important as solving problems. In the past two years of study, I have been encountering problems constantly, and the door of his office has been open to me to help me solve all problems in the study. I want to thank Dr. Zheyang Wu provided lots of useful suggestions on the content of this thesis paper.

Secondly, I would like to thank all professors in the BCB program. In every seminar in the past two years, the presentation and discussions of the teachers have greatly improved me personally.

I would like to express my fellow BCB students in the lab, especially Ziyang Gao. She is my close friend and she has provided great help in my project and the study and life in the United States in the last two years. Thank her for helping me contribute to the data collection and statistical testing part of the work. Without her guidance and help, I can't finish such a large and important project.

Finally, I want to thank my parents for everything they have done for me! There is nothing I can do in return. I can only work harder and bravely to take the next step in life and make them proud forever.

Thanks to WPI and every teacher here, the two years of study have benefited me a lot and were a meaningful experience for my entire life.

Table of Contents:

List of Figures

List of Tables

1. INTRODUCTION

1.1 Genetic variation determines the functional diversity among the populations of healthy individuals

Next-generation sequencing and RNA-Seq technology have revealed that complex genetic diseases commonly due to pathogenic variations at the genetic posttranscriptional, and epigenetic levels [1-6]. Single nucleotide variations (SNVs) occurring in coding, as well as non-coding areas in genomes, are the most well-document studied class of genetic change and disorders. Nucleotide substitutions that cause an amino acid change are non-synonymous variants.

According to the National Cancer Institute-National Human Genome Research Institute catalog of published GWAS projects, there have already been publish 14876 SNVs collected [7]. At the same time, many genetic variation databases have been developed to collect almost all common and many rare variations in humans. Some well-established human genetic variation databases include 1000 Genomes [8], Database of short Genetic variations (dbSNP) [9]. Other databases such as HGMD [10], OMMIM [11] can provide us with genotype-phenotype information which can be used for functional annotation of mutations and their effects.

The effect of SNVs on molecular function in humans can be benign (no influence or trivial) and impactful function (affect the protein function). The impactful effect, in turn, can be deleterious or beneficial effect [12]. Many recent studies of disease networks have linked nsSNVs with macromolecular interactomes [13-15]. Now there are many functional annotation tools for genetic variation and most of the tools can be applied to the annotation of SNVs, while some tools also cover indels [16-19]. For example, ANNOVAR [20] could be applied at the whole-genome level. Tools like PolyPhen [21] and SIFT [22] are also popular tools to predict the damaging effect of missense mutation on protein function.

Non-synonymous SNVs are abundant in the human genome, and the frequency distribution is different among the population group. Population-specific differences in genetic variation could contribute to some of the observed differences in susceptibility to common diseases, response to drug treatments, and immune response [23-24]. One field, Pharmacogenomics, seeks to relate genetic variability in drug response, and the range of study from candidate gene studies to variation across whole genomes of human populations containing individuals who exhibit a range of response to different drugs [25-26]. Another field genome-wide association studies (GWAS) used in genetics research to associate specific genetic variations with particular diseases. This method involves studying genomes from many different people and is used to predict the presence of a disease in different populations.

1.2 Role of ligand binding in human function

Molecular recognition plays a fundamental role in all biology processes [27]. Many proteins function through binding small ligands. A protein-ligand interaction, whether this ligand is a hormone, fatty acid, drug, or metabolite can be associated with a variety of biological functions such as lipid metabolism, inflammatory processes, and hormonal regulation or lead to catastrophic events, such as adverse drug reaction, allergy or poisoning. The majority of the drug targets are also proteins [28-29].

The way of protein-ligand interaction is through LBSs, so the identification of specific LBS on proteins is often an important step toward understanding the function of protein molecules or the design of new therapeutic compounds to modulate the various functions in human [30]. Some point mutations at LBS may change the binding affinities of the ligands. Potential effects of genetic variation on LBS included metabolizing certain food, susceptibility to the drug, Metabolic abnormalities, physiological differences, and other catastrophic functions [31-32].

1.3 Protein-ligand Interaction Study Progress

In the last few decades, molecular recognition and protein-ligand interaction have been traditionally studied using biophysical methods [33-35]. With the development of genome sequencing technology, people found that molecular mechanisms driving many genetic diseases, so more and more people start to focus on genetic variant disease networks. However, we have yet to get any comparable in its scale insights to how genetic mutations regulate the interactions between macromolecules and small ligands compared with Protein-Protein interactions and Protein-RNA [36] interactions work because of the high-throughput experimental approaches. Besides, most of the ligand-related work's insight just focuses on diseased people. A recent analysis revealed that ligand and binding residues had a significantly higher mutation rate than other parts of the protein across 16 cancer types [37] and built mutLBSgeneDB [38] which included 12000 mutations at 10000 LBSs in cancer and 744 drug-gene. Moreover, mutation-induced molecular modifications in protein-ligand interactions have been identified. Epidermal growth factor receptor mutation in glioblastoma increased ligand binding affinity for EGF [39] and mutation in neuraminidase 1 gene conferred high ligand binding affinity [40].

1.4 Genetic variation affects ligand function in the health population

Although the relationship between genotype-phenotype among diseases is complex than what we thought previously because of gene pleiotropy, the mutation rewiring genetic effects may lead to disease phenotype still convincing. In this work, we focused on the genetic variation in the healthy population and to study whether common genetic variation will affect ligand function. The genetic mutation also exists in healthy people, even if people don't have any disease, so we draw a connection between these genetic variations with ligand functions. Recent evidence suggests that many disease-associated residues on or near the protein-ligand interaction[41-42] because the region are required for the interactions with other proteins or small ligand, so we also comprehensive mapped and extracted the genetic variants across different populations on the protein ligand-binding region to build such mutation ligand map. We collected the variant data from gnomAD [43] which aggregate and harmonize both exome and genome sequencing data from large-scale sequencing projects. We combined the WGS and WES variant frequency of ligand binding genes within specific populations and performed the statistical analysis over the LBS variant frequency with outside LBS region frequency. To study what potential ligand function has been affected in different people by genetic variation, we mapped the phenotype important pharmacological gene haplotypes over the ligand-binding variants and analyze the case studies.

2. METHODS

2.1 Problem Formulation

Our study showed a big interest in how common variants affect normal protein-ligand interaction and the goal of this project is to catalog characterize the potential effects of disease and population-specific genetic variations on protein-ligand interaction function. We created a comprehensive human LBS mutation map to figure out the patterns of protein mutations that happen in the ligand-binding region. Specifically, we hypothesized that LBSs of proteins are enriched with the population-specific mutations including some disease-associated mutations. Besides, we expected that many mutations would occur on hormone receptors. Testing this hypothesis, we provided insights on whether any specific category of LBSs-fatty acid, hormone, drug, or metabolite is equally susceptible to mutations in healthy populations.

To finish our goal, we applied the semi-manual similarity score comparison algorithm to classify the types of ligands and compared the distribution of different ligand type mutations pattern (site, result, and frequency of mutations) in the population. Finally, we found some important genes which involved important biological functions to explore the potential functional effect of the genetic variation on the protein-ligand interaction from previous research.

2.2 Methodology Overview

In our research, we developed a new computational approaches pipeline (Fig.1). First, we used the human ligand binding interaction data extracted from the BioLip [44]. Second, we combined the two versions (V2 and V3) of genetic variation data in gnomAD based on the variant frequency as our genetic variant data. Then, we annotated the type and function of each ligand involved based on the result of their chemical structure cluster. Finally, only the gene with the variants that hit the ligand-binding region has been selected as the ligand gene to do further analysis. At the downstream stage, we used statistical analysis to calculate whether there is a significant difference between the mutations in the LBSs region and the mutations in other regions. Next, we generated population variation distribution and observed each specific ligand variation patterns. Finally, we selected some important pharmacological gene from PharmKGB as the case to study and discuss the effect of this mutation on drug-related functions.

Throughout the whole research process, most of the coding work was finished in R. Some specific process like combined the WGS and WES variant frequency and classified the ligand type was performed in Python. In addition, we also use python API for bulk alignment in Emboss-NEEDLE in residues renumber. Besides, some public database like ChemSpider [45], DrugBank [46] was used to annotate the type and function of each ligand.

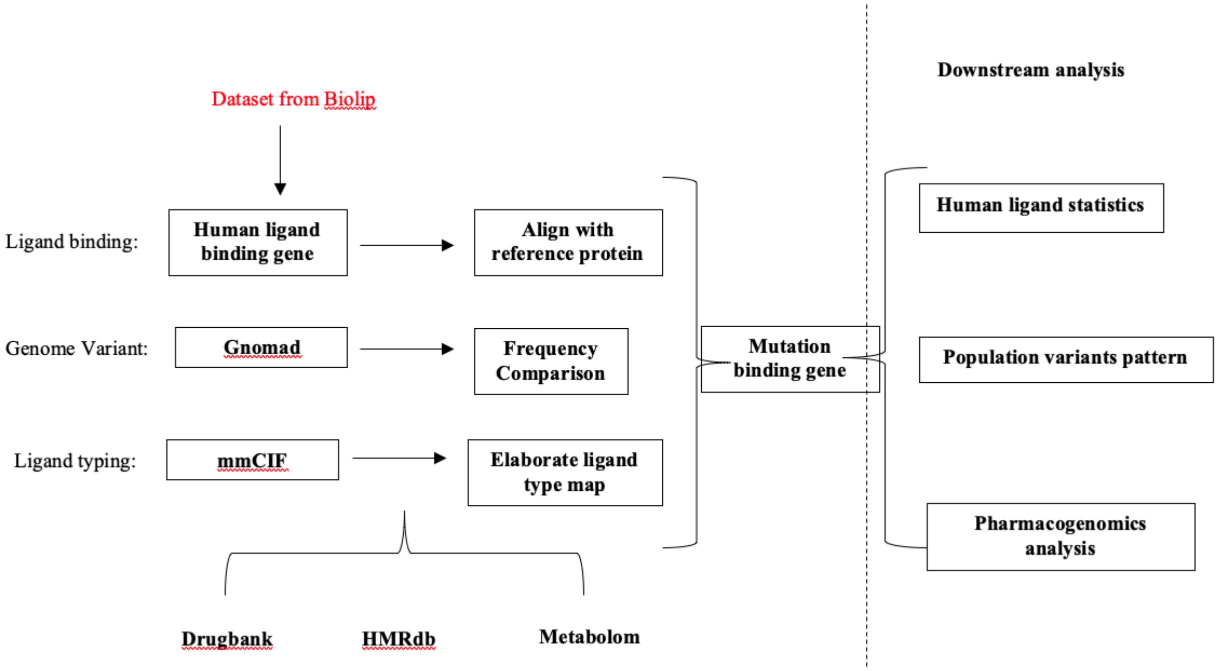


Figure 1. A pipeline of experiment methodology: The original ligand data set was obtained from BioLip, which contains 2000+ human proteins and 12000+ ligand binding pairs. All the misplaced LBSs were renumbered against the reference protein sequence. The genetic variant data was extracted from gnomAD after the combination of two version data.

2.3 Data Collection and Analysis

All ligand binding data were downloaded from BioLip, we used EMBOSS-NEEDLE to align the receptor protein sequence to target reference protein sequences and then extract the original residue position and renumber high-quality LBSs based on the alignment result (coverage rate >0.5). Redundant records removed by leaving the highest resolution, and only human protein receptors with ligands were selected. The annotated variants of the selected genes were imported from gnomAD included missense, frameshift, splice region, stop_gained, synonymous, and others. Ensembl canonical transcripts were chosen to represent the corresponding gene. Matching the binding data with variants, only the ligand receptors that are fully covered by the Ensembl reference protein sequences are left for future analysis (1980 genes). The LBSs were renumbered to fit the ensemble reference protein sequences, then the variants that hit the ligand-binding region were filtered out (1727 genes), some genes with variants do not hit the ligand-binding region were eliminated. For each variation, we define the high frequency of the variant by at least one population frequency > 0.0001.

2.4 Ligand Type Annotation

We grouped and annotated BioLiP ligands into biologically 8 relevant groups (Table1) with the method development by Mona Singh group in Department of Biomedical Informatics, Harvard Medical School [47]. The ligands form co-complex structures with "ion" and "metal" in their full names are assigned to the ion group and with metal. To highlight domain positions comprise metabolically relevant, hormone-relevant, and/or potentially druggable binding pockets, we further categorized small molecules as follows. Any small molecule ligand with a Tanimoto coefficient> 0.85 (Open Babel Package, v2.4.1) [48-49] between its SMILES string (wwPDB's Chemical Component Dictionary, v3.30) and the SMILES strings of endogenous human metabolites (Human Metabolome Database, v3.6) [50], drugs (DrugBank, v5.0.1) [46] and hormone (HormoneBase [51]) is respectively classified as a Metabolite, Druglike, and Hormone group. Other ligand has been grouped into Regular, Peptide, and Nucleic Acid group by BioLip. After grouped each ligand into different categories, we annotated each ligand by their chemical name and functions they involved by two public databased ChemSpider and DrugBank. In these groups, we are more concerned about the result of the variant pattern in Metal, Ions, Hormone, Drug, Metabolite, and Regular ligand group because ligand in those group usually involved in important human biological reactions compared with the Nucleic Acid and Peptide group.

Ligand Type	Reference	E.X
Metal	PDBx/mmCIF dictionary	Fe
Ions	PDBx/mmCIF dictionary	SCN
Hormone like	HMRdb	R18
Drug Like	DrugBank	TDZ
Metabolite	The Human Metabolome Database	TDR
Peptide	PDBx/mmCIF dictionary	III
Nucleic Acid	PDBx/mmCIF dictionary	Cl
Regular	PDBx/mmCIF dictionary	TDU

Table1: List of Ligand type group: the reference column is the source of the ligand molecular structure information we collected. The E.X column is the ligand id example in different ligand type group

2.5 Mutation Mapping

We used EMBOSS-NEEDLE to align the receptor sequence with the protein sequence of hg38 (end-to-end, pairwise) and got renumber LBSs residues in BioLip by the result of sequence alignment. Then, we mapped the comprehensive corresponding variation sites in gnomAD with their variant protein position and extracted the variant sites at the ligand-receptor residues by combined those datasets, so we got all the mutation binding site location and their specific mutation information. Some LBSs residues with multiple non-synonymous variations, which is at high frequency in at least one population were defined as LBS Polymorphism after the repeated level caused by ligand chains were removed.

2.6 Pharmacological Gene Mutation Mapping

After mutation mapping, we explored some drug mutations already been studied before as the case study. The study of Pharmacogenomics provides us the impact of the genetic variations in drug genes on the drug response. We found some important genes which involved in the Pharmacogenomics study (Pharmacological Gene) in our ligand data as a case study to explore the effect of the mutation on LBSs in different population. We extracted 68 important genes in PharmGKB [49] VIPs (Very Important Pharmacogenes) channel and used a similar mapping method to mapped back to our mutation ligand data. VIPs provides an overview of a significant gene in the metabolism of, or response to one or several important drugs to human-like some gene plays the role in the metabolism of many drugs like CYP2D6, or other variant contribute to severe drug response like HLA-B. The source of these 68 genes in VIPs channel was from several authority agencies like the US Food and Drug Administration (FDA) biomarker list, the Clinical Pharmacogenetic Implementation Consortium (CPIC), and others. Besides, the VIPs channel also provides us the background information of each gene include any disease association, as well as in-depth information on the gene's pharmacogenetics, included haplotype, each variant annotation, clinical annotation, and phenotype. We mapped these 68 genes to our mutation LBS data can get the several pharmacological gene mutations map in different populations. The final visualization mutation results are displayed with TBtools [53]. Finally, we studied the effect of each haplotype mutation in a specific population from other public papers to help us predict the potential function and mechanism of the mutation to protein-ligand interaction.

2.7 Statistical Testing

The nvSNV-driven effects on protein-ligand interactions followed by the statistical analysis of the co-localization results. All variant residues in this experiment filtered out into the variant within LBSs or outside the LBSs. To compare the frequency results of the variant in LBSs and explore whether it has a specific amino acid change in the ligand-binding region, we calculated the reference and their alternate amino acid frequency in the heatmap matrix. We explored whether the ligand region is filled with a large number of non-synonymous mutations, so the differences between the overall number of the synonymous variant in ligand gene directly comparing the number of non-synonymous variants and differences between the number of these two variants in LBS or non-LBS region were applied by one-way analysis of variance (ANOVA) or two-way ANOVA. P values of < 0.01 were considered significant. All the statistical analyses were carried out in R version 3.6.3

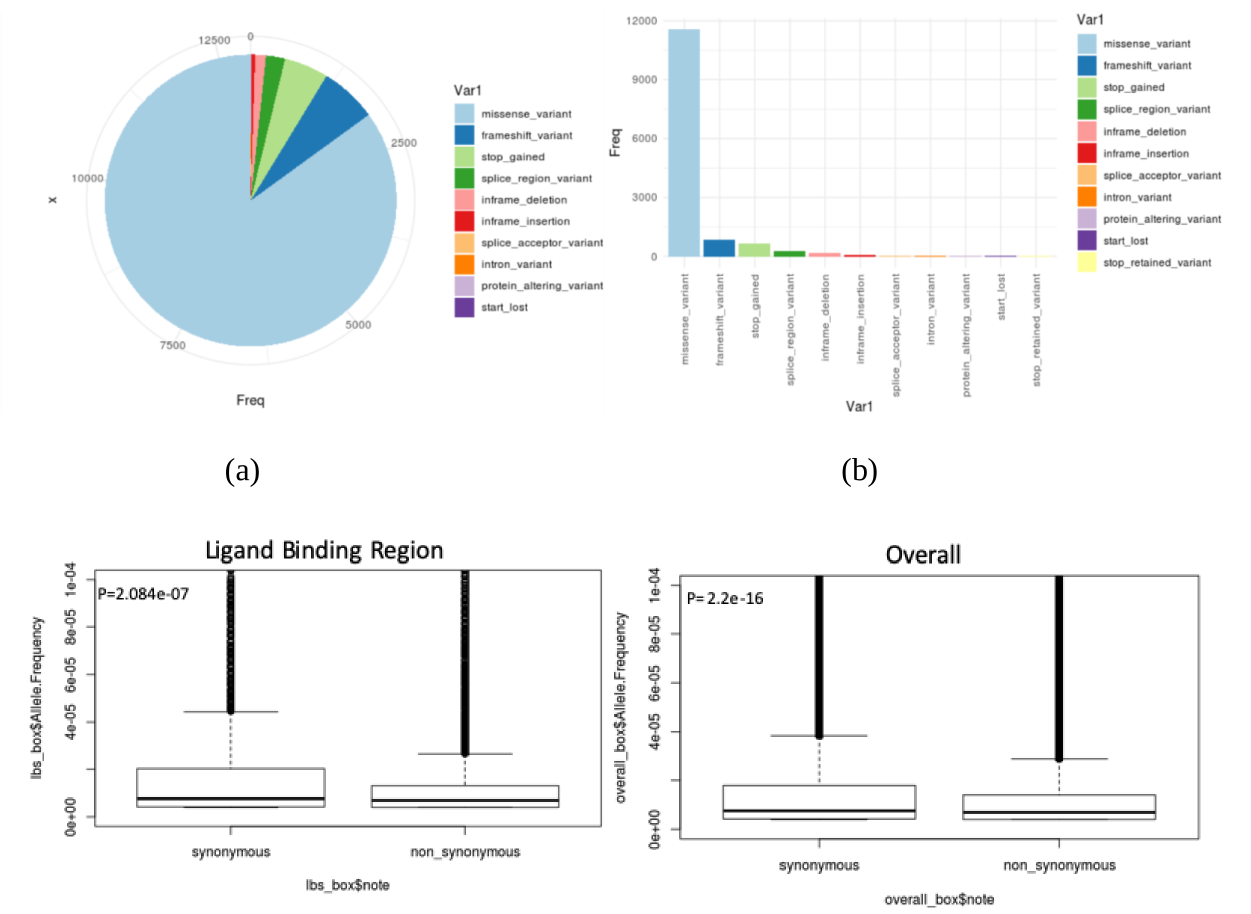
3. RESULTS

Initially, we hypothesized that most of the mutations on the protein-ligand binding receptor are non-synonymous variants and a large percentage within these variants should be associated with hormone receptors based on the previous study that showed that the hormone receptor mutations might be the drivers behind the disease and adverse reactions. Specifically, a mutation of leucine to serine in residue 454 (L454S) of the thyroid hormone leads to severe resistance to thyroid hormone (RTH) [54]. But by mapping variant data in gnomAD into LBSs, the number of hormone mutation binding is much fewer than the metabolize and drug binding.

3.1 Statistics on the data collected and integrated

For the original genetic variation data set collected from gnomAD, most of the variants were encoded by their rsID. Besides, other important information like mutation position, variant type, amino acid change, and allele frequency in each population were also provided.

After combined the two versions of genetic variation data and mapped to the receptor protein, we total got 1980 ligand-receptor genes which included 83653 variants. In these 1980 genes, 1727 gene ligand-binding receptor was affected by the variation. The cases with multiple chains and one chain able to bind to multiple same ligands were also included. Within all variants in ligand binding region record, there are 1 start lost, 595 stop_gained, 10756 missense variants (non-synonymous) (Fig.2a), and 8098 synonymous. Among all these variants, the non-synonymous is what we were more concerned about. The overall number of the non-synonymous variants in the LBSs-containing gene is significantly higher than the synonymous based on the p-value, but the overall frequency of the non-synonymous and synonymous variants in the LBS region don't have significant difference based on the one-way analysis of variance results (Fig.2c), showing that the frequency of the synonymous variants is relatively higher than the non-synonymous. This result is in line with our expectations because our data comes from the healthy population rather than disease-associated people, so the non-synonymous mutation frequency did not show significance in overall mutations.



(a)

(b)

(c)

(d)

Figure.2: Variation frequency and distribution: (a) and (b) are the frequency of different types of mutations in LBSs. (c) is the one-way analysis of variance result about synonymous and non-synonymous frequency in the ligand-binding region. (d) is the one-way analysis of variance result about overall synonymous and non-synonymous frequency

To test the hypothesis about whether any specific mutation is more easily appear on LBSs in healthy populations and help us to predict the protein change pattern, we calculated all the frequency of the amino acid changes in protein caused by these 10756 missense variants (Fig.3). We found that there are some specific higher frequency amino acid changes happened such as the reference amino acid Arg is more likely to become Trp, Gln, His and Cys compared to other amino acids.

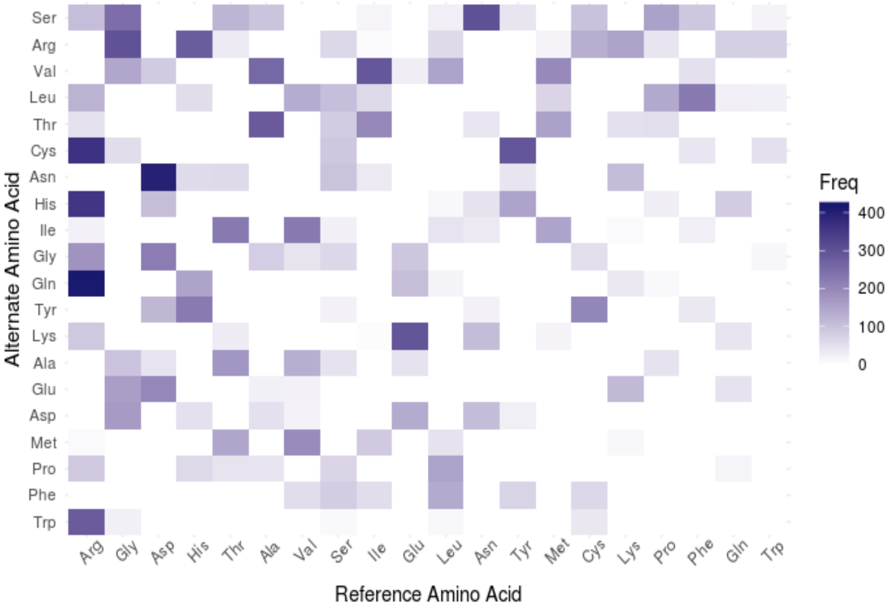


Fig.3: Missense reference and alternate: The darker the color in the heatmap, the higher the frequency of this amino acid change.

We are still far away from complete understanding what the biological functions or diseases will be impacted by the LBS-associated variants due to the limited clinical genomics and protein-ligand interaction structural data. But these preliminary data can provide further directions for in-depth predicting potential mutation patterns in the binding regions by leveraging machine learning and GWAS technology in the future.

3.2 Ligand type-specific variation pattern

After founding the mutation binding residues, we got the map of the overall population variation on the ligand-binding region (Fig.4). The figure result showed that Ashkenazi Jewish and European Finnish populations have a higher overlap of mutation patterns (similar site, result, and frequency of mutations) in the ligand-binding region, and the African and Latino have a similar mutation patterns in this region based on the hierarchical cluster analysis. But to our surprise, it showed that the variation of East Asians is quite different from that of other populations. This result is beyond our expectations because, in several other big human genome project cluster analysis experiments, the results often show that the mutation pattern in African populations is the most different from other populations and has a higher frequency. And according to records, some regional rare genetic diseases often appear in Africa, but the result in our study showed that the East Asian populations have some specific mutations in the ligand-binding region that influence the normal human protein functions.

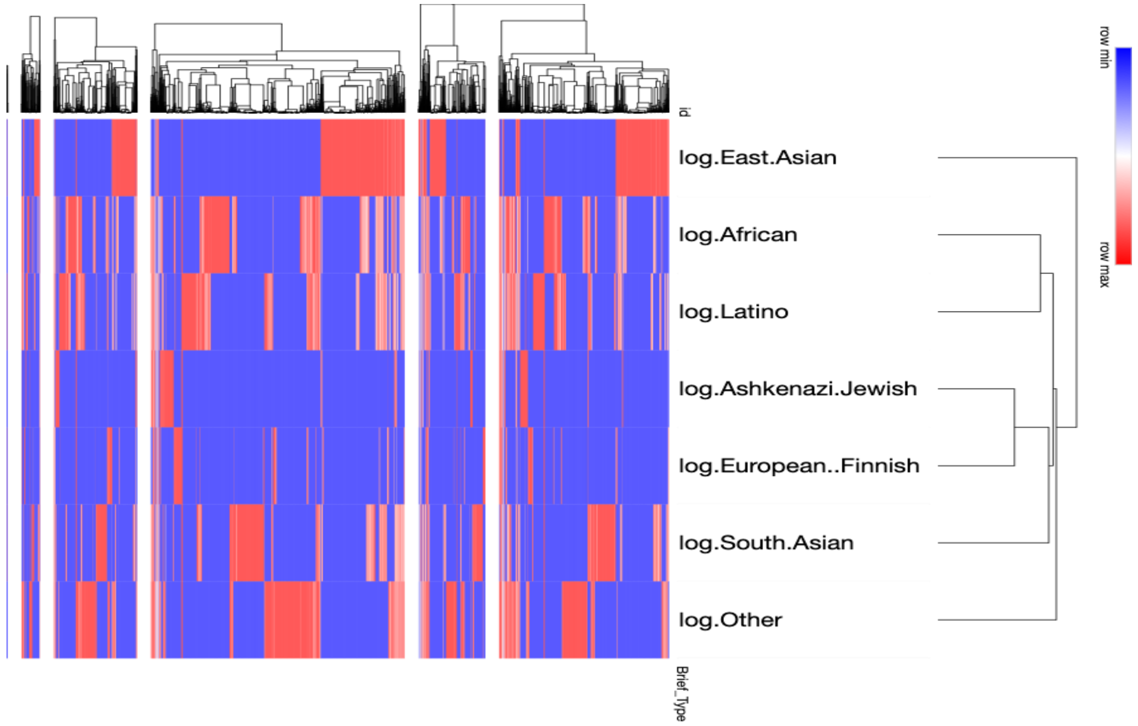
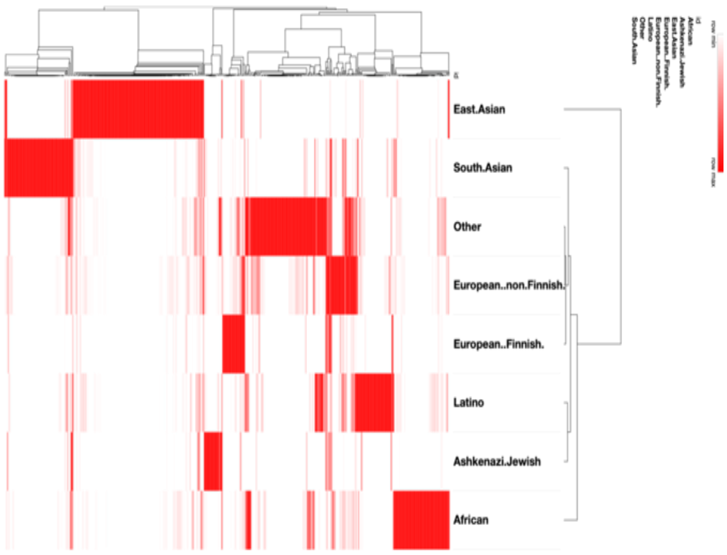
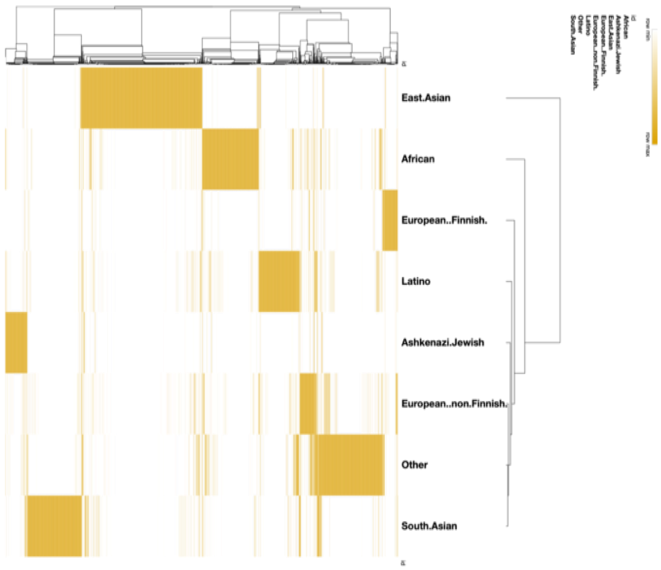


Figure.4: Overall LBSs Variation Pattern: The abscissa represents the gene whose protein has LBSs, and the ordinate represents each population. The darker the color, the higher the frequency of the mutation.

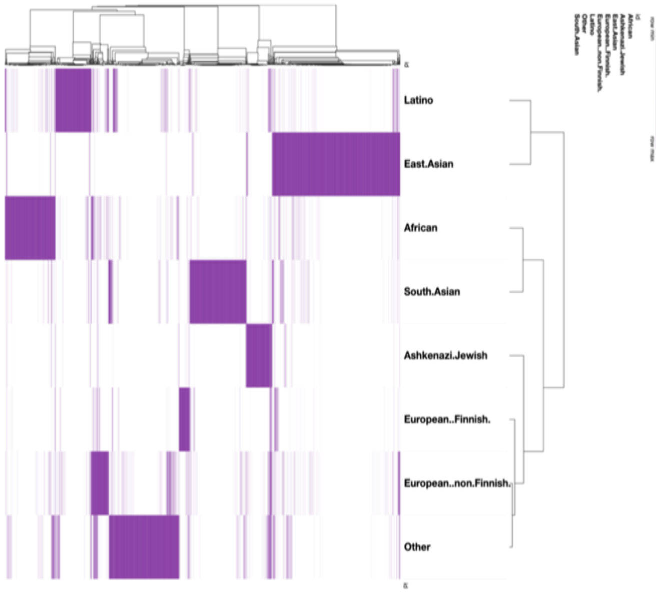
When we specifically focused on the mutations of different ligand-receptor types, the results are different. After got each type of ligand in different groups, we also calculated drug, metabolite ligand-target site variation frequency in different populations and drew their distribution map to validate our hypothesis (Fig.5). When we compared the variation pattern in each ligand type, some mutations are indeed specific in a certain population. For example, we found some mutations only exist in East Asian in the drug group which proves that the fact about genetic variation in the human can reflect different individual responses to the drug and other biological functions. On the other side, we also found that some of the variants exist in different populations and it has many overlapping parts in one column. It showed that such mutations are common mutations although the value of mutation frequency is different, so such result can help us to understand some common genetic-related phenotype or diseases that are widespread in humans not in one specific population such as allergic reaction or drug resistance. Drug group (Fig.5.a) and Metabolites (Fig.5.c) groups result are of special interest because the impact of the mutation on these two groups can be directly related to different life styles, food preferences and drug susceptibilities. For example, mutations in the drug group have a higher chance to relate drug response or side effect, while mutations in the metabolite group may directly relate to consumption of different foods. In the drug group, we found that the Latino and Ashkenazi Jewish populations have similar mutation patterns in the mutation location and mutation result on LBSs. Finnish and European populations also have similar mutation patterns based on the clustering result. But in the metabolite group, Finnish has a big different mutation patterns from other European populations. Mutations in African occur on different LBSs compared with other populations which can prove that Africa has some area-related diseases caused by metabolism disease. The same conclusion in these two groups is the East Asian genetic variation pattern is the most different from other health populations. This result showed the same conclusion as the overall distribution result. It provides novel insight to the other biology or bioinformatic analysis in Asian population ligand-related research such as Evolutionary analysis, GWAS, or other studies.



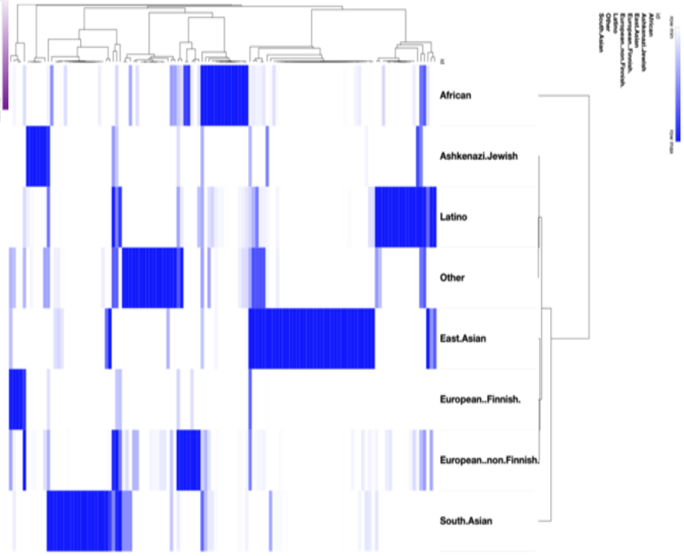
(a)Drug



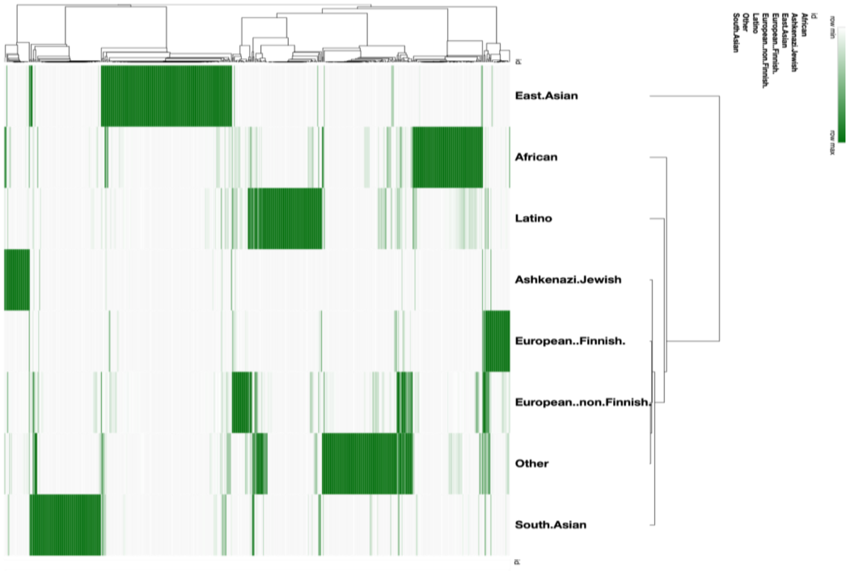
(c) Metabolites



(b)Regular



(d) Ion



(e) Metal

Figure5: Different Ligand Type Variant Pattern: (a) is the Drug LBSs map (b) is Regular LBSs map (c) is Metabolites LBSs map (d) is Ion LBSs map and (e) is Metal LBSs map. The abscissa represents the gene whose protein has LBSs, and the ordinate represents the population. The darker the color, the higher the frequency of the mutation.

3.3 Passive Binding with Human Protein Case

In our drug ligand group, one interesting drug case we found is Abacavir [55-57] which is a passive binding with human ligand case. Abacavir is a power nucleoside analog reverse transcriptase inhibitor used to treat HIV and AIDS. Chemically, it is a synthetic carbocyclic nucleoside and is the enantiomer with 1S, 4R absolute configuration on the cyclopentene ring. However, serious hypersensitivity reactions (HSR) have been associated with abacavir by the fact that 5% of individuals who receive abacavir develop an immune-mediated HSR. There is large evidence to show that people who carry the HLA-B*57:01 variant are at significantly increased risk of developing HSR in different populations and the HLA-B*57:01 allele frequency in worldwide populations is in Table2. One most potential mechanism is the abacavir itself may covalently bind to the normal peptide ligand for HLA-B*57:01, making the immune system begins to mount a defense [58].

Population	Allele count	Allele Number	Number of Homozygotes	Allele Frequency
South Asian	1405	27008	52	0.05202
Ashkenazi Jewish	404	9006	17	0.04489
European (non-Finnish)	4050	119512	82	0.03389
Other	177	6484	1	0.02730
European (Finnish)	238	24604	1	0.009673
Latino/Admixed American	276	28864	11	0.009562
African/African-American	189	24262	3	0.007790
Ease Asian	58	18444	1	0.003145

Table 2: List of HLA-B*5701 allele frequency in worldwide populations: the data collected from gnomAD

3.4 Pharmacogenomic Analysis Information

In total, we have got 20 pharmacological gene variants (Table3) on the ligand-binding region after we mapped the 68 important genes to our LBSs map. We drew the connection between these selected variants in different haplotypes and the function or phenotype of this mutation on protein-ligand interaction.

Gene	Variant	Haplotype	Allele. count	Population
CYP2C9	L361I	*55	2	East.Asian European.non.Finnish
CYP2C9	rs578144976	*66	27	European.non.Finnish South.Asian
CYP2C9	rs762239445	*39	8	East.Asian European.non.Finnish.
CYP2C9	rs767576260	*43	14	European.non.Finnish. Latino East.Asian South.Asian
CYP2C9	rs776908257	*67	12	Latino European.non.Finnish South.Asian
CYP2A13	R101Q	*4	888	European.non.Finnish.
CYP1A2	rs72547517	*8	12	African Latino East.Asian European.non.Finnish. South.Asian
CYP2D6	rs1406719554	*123	1	African
CYP2D6	rs199535154	*20	20	African East.Asian European.non.Finnish.
CYP2D6	rs532668079	*75	4	African
CYP2D6	rs730882251	*62	24	African, Latino Jewish European.non.Finnish.
NAT2	rs56387565	*12F	53	African Latino Other
NAT2	rs72554615	*6D	38	Amish European.Finnish. European.non.Finnish.
CYP2A6	rs143731390	*24A	1238	African Latino, Jewish East.Asian European.Finnish. European.non.Finnish. Other South.Asian
CYP2A6	rs143731390	*24B	1238	African Latino, Jewish East.Asian

				European.Finnish. European.non.Finnish. Other South.Asian
CYP2A6	rs143731390	*35A	1238	African Latino, Jewish East.Asian European.Finnish. European.non.Finnish. Other South.Asian
CYP2A6	rs143731390	*35B	1238	African Latino, Jewish East.Asian European.Finnish. European.non.Finnish. Other South.Asian
CYP2A6	rs143731390	*36	1238	African Latino, Jewish East.Asian European.Finnish. European.non.Finnish. Other South.Asian
CYP2A6	rs143731390	*37	1238	African Latino, Jewish East.Asian European.Finnish. European.non.Finnish. Other South.Asian
CYP2B6	rs139801276	*35	284	African Latino East.Asian European.non.Finnish. Other
CYP2B6	rs36079186	*27	168	African Latino European.non.Finnish. Other South.Asian
CYP2B6	rs36079186	*35	168	African Latino European.non.Finnish. Other South.Asian
CYP2B6	rs564083989	*24	4	East.Asian Other

CYP2C19	rs41291556	*8	225	African Latino Jewish European.Finnish. European.non.Finnish. Other South.Asian
CYP2C19	rs56337013	*5	1	East.Asian European.non.Finnish. South.Asian
TPMT	rs759836180	*42	27	European.Finnish. European.non.Finnish.

Table 3: List of Pharmacological Gene Variant haplotype information

We also got the map of those pharmacological variations on the ligand-binding region. The result (Fig.6) by TBtools [53] suggested that even though such drug LBSs-associated variants have potentially different impacts across different population groups due to the higher frequencies of occurrence in East. Asian and African subpopulations.

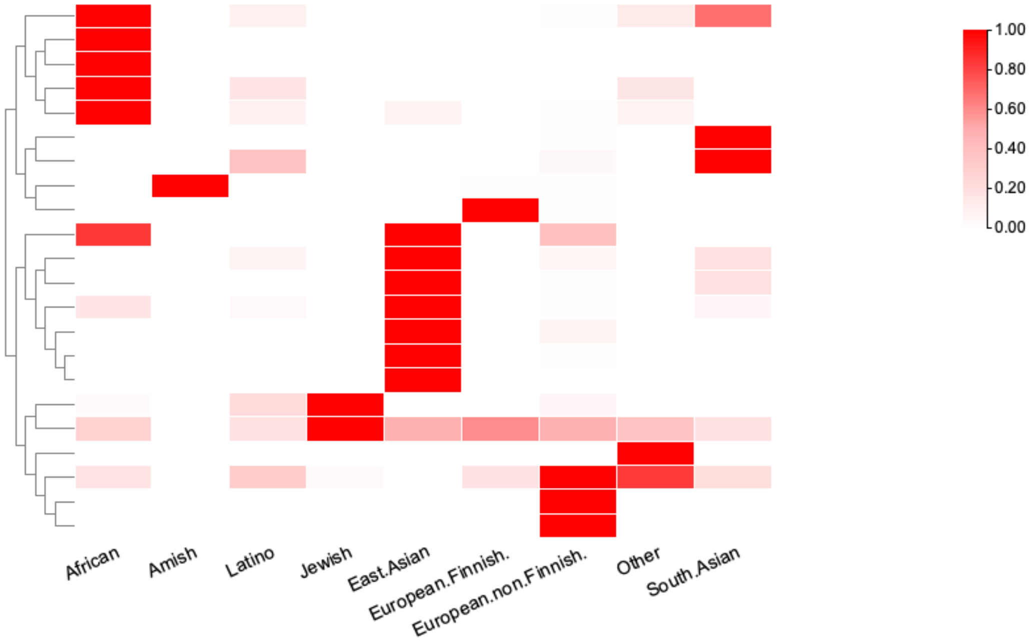


Figure6: Pharmacological Variant Pattern Distribution: The abscissa represents the population. The ordinate represents got 20 pharmacological gene variants we selected. The value of the result is standardized and presented in the form of color shades in the figure

Furthermore, we also explored whether these mutations work synergistically or antagonistically in our pharmacological gene group. In our pharmacogenomics analysis, although most of these medicinal ligand mutations are rare, we do not know the function, mechanism of these mutations on the role of the ligand, but few existing research has shown that nsSNV functions to the ligand-binding interaction also present deleterious, beneficial and neutral effects which similar to study the nsSNV functions on protein-protein interaction.

3.4.1 Deleterious Effect Case

CYP2A6 (Fig.7), an enzyme responsible for the metabolism of a large number of xenobiotic compounds: many drugs metabolized by CYP-450 enzymes involve CYP2A6. The gene is highly polymorphic, with variations resulting in altered enzymatic activity. The novel CYP2A6*24 allele had two non-synonymous variants on Va110 and Asn438 compared to the CYP2A6 wild type. Some study shows that the later one (N438Y) is happened on heme-binding amino acids and also adjacent to two heme-binding sites: Arg437 and Cys439, which alter the active site cavity and change binding affinity. Such mutation caused the nicotine metabolism to decrease and nicotine dependency disorder found in Black African descent [59].

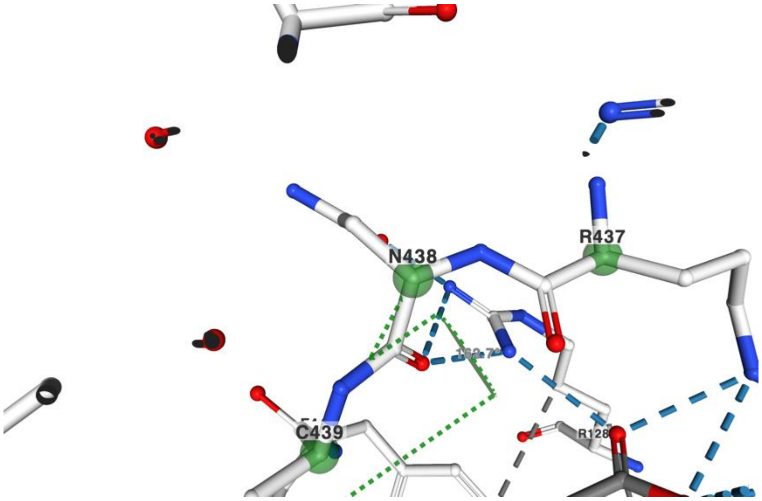


Figure.7: CYP2A6 PDB Structure View (N438): pdb_id:3T3Q

CYP2C9 (Fig.8), is a phase I drug-metabolizing cytochrome P450 (CYP450) enzyme isoform that plays a major role in the oxidation of both xenobiotics and endogenous compounds. The discovered binding pocket shows that it may simultaneously accommodate multiple ligands during its biological function and provides a foundation for understanding complex drug–drug interaction. CYP2C9*66 is one of the novels and rare haplotypes formed by some novel and rare variants which already found in three projects. Leu362Val is a specific variant in the south-Asian and the putative functional variants analyzed shows that even though the mutation Leu362Val present within CYP2C9*66 is predicted to be tolerated/benign, conversion from leucine to valine can affect assess of the drug to the heme group of active site result from Leu362 is present within the hydrophobic substrate-binding pocket of CYP2C9 [60].

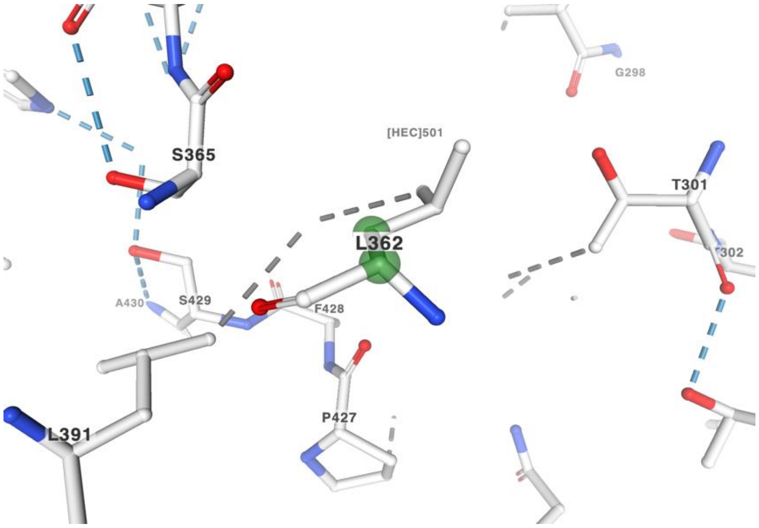


Figure.8: CYP2C9 PDB Structure View (L362): pdb_id:1OG2

3.4.2 Neutral Effect Case

Still, CYP2C9 (Fig.9), the haplotype CYP2C9*43 in our data was described as “almost null catalytic activity” is not associated with expression of CYP2C9 was considered to be a 'tolerated' mutation. The novel coding variant (R124W) had almost null catalytic activity compared to wild-type protein based on the in vitro catalytic activity analysis in the Han Chinese population [61].

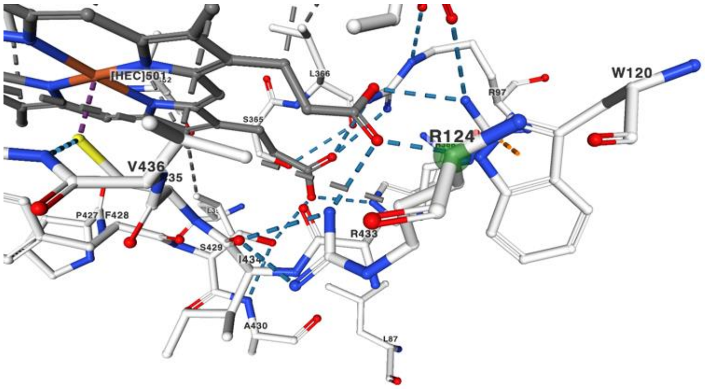


Figure.9: CYP2C9 PDB Structure View(R124): pdb_id:1OG2

3.4.3 Beneficial Effect Case

Genetic variants of NAT2 (Fig.10) have primarily been associated with drug metabolism, response, and toxicity of multiple drugs, most notably, anti-infective agents. Non-synonymous variant rs56387565 in NAT2*12F haplotype causes the amino acid change Tyr208His in CoA binding. Most of the identified SNPs in NAT2 should be innocuous to the function of the NAT2 enzyme but specifically, the Tyr208His would even enhance NAT2 affinity for CoA in Brazil population since it is possible that the positively charged side chain in His208 would better stabilize the negatively charged phosphate groups in that cofactor. This view was supported by the observation that His is also naturally found in the homologous position of the chicken NAT2 sequence [62].

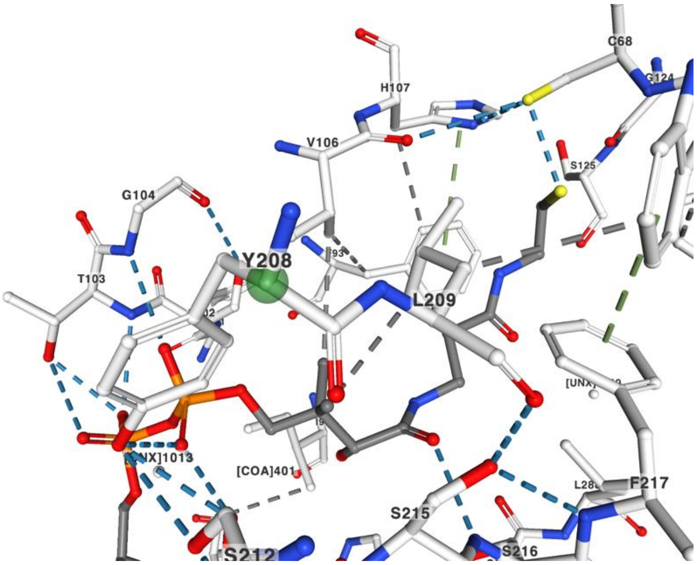


Figure.10: NAT2 PDB Structure View (Y208): pdb_id:2PFR

4. Discussion

4.1 Overview of Results

In this work, we studied the mutations occurring in the ligand binding sites LBSs of human proteins and built the most comprehensive human LBS mutation map. The construction of the first population-specific atlas of SNVs associated with ligand binding sites followed by the analysis of the functional impact of mutations between different race cohorts. Our analysis helps one to determine and characterize the patterns of protein mutations in the ligand-binding regions and explore the profile of different population variants associated with the ligand binding function. After we collected all the genetic variation information, we found that most variants in the ligand-binding genes are the non-synonymous substitutions that alter the amino acid sequence of the protein that may or may not lead to phenotypic changes. Interestingly, we found no apparent difference between the frequencies of synonymous and non-synonymous substitutions located specifically in the ligand-binding sites, which was expected because our data were collected from the healthy population with no obvious disease conditions or abnormal phenotypes reported. We grouped and classified all the ligands in BioLip into eight different types, which leads to a surprising observation that the number of the metabolite binding sites were the group of LBSs most affected by mutations.

The results of the LBSs genetic variation mapping and the cluster analysis show that the Ashkenazi Jewish and European Finnish population have a similar genetic variant pattern in the ligand-binding region and both of their variant allele frequency is lower than the other six populations. The African and Latino have a similar variant pattern and their variant frequency is higher than the Ashkenazi Jewish and European Finnish. To our surprise, the East Asian people mutation frequency in the LBSs region are highest among people and are the most different variant pattern with other populations based on the cluster analysis, so it shows that it must have some specific area or higher frequency ligand binding mutation-related behavior that needs us to explore in East Asian people.

In the drug map analysis, we found that the European non. Finnish and European Finish populations have a close mutation pattern, so it suggested that all European population maintains a certain degree of similarity in the drug response or metabolism of drug function dominated by ligand mutations. In the metabolites group, South Asians start to have a similar variant pattern with the European group, and the African variant pattern has a somewhat greater difference with the European and Latino populations. The drug group and metabolite group maps both suggested that the East Asian population has the most distinct mutation ligand binding pattern compared with other population groups in the world. Therefore, one can conclude that East Asian is a unique and therefore very important population for a ligand-related study with its special variant pattern based on our map result. The above distribution and clustering results support our hypothesis that the ligand region in different populations will be affected by different mutations. Among them, the location and frequency of mutations in East Asians are quite different from those in other populations.

In the pharmacogenomics analysis, we selected an important drug receptor to see if the population variants associated with this gene may have a functional effect on receptor binding based. We found 20 pharmacological gene variants in our map that were mainly distributed among East Asians and Africans. From this pool, four variants in three important genes CYP2A6, CYP2C9, and NAT2 have proved that the different SNVs impact on protein-ligand interaction included deleterious, neutral, and beneficial from previous public paper in nicotine, phenytoin, and other drugs studies which similar to protein-protein interaction rewiring by non-synonymous variants[15]. It has proved that those binding mutations will decrease the metabolism function of the drug or increase the risk of disease when treated with the drug. These results further suggest that SNPs could rewire protein-ligand interaction altering the function mediated by the interaction.

4.2 Limitation of the Research and Future Work

In our research, we collected the genetic variation data from gnomAD database. The main advantage of using this database, compared with other human genome projects like ExAC and the 1000 genomes project, is it is one of the newest ones, with a larger scale and scope, and provides a richer interpretable set of genetic variations. The main limitation is that gnomAD database only has mutation information from 8 populations. Therefore, although our map has the largest number of mutations and annotation information, we do not have the most comprehensive population information enough. If people want to focus on a certain country or even more small population ligand area mutation information, this result cannot show. Besides, there has not been a single large-scale study on the impact of specific mutations on macromolecules and how genetic mutations regulate the interaction. As a new database, most of the rare variants in gnomAD no one has studied before, so this study is lacking resource for studying the function of rare variants. For example, rs759836180 in TPMT in VIP group we don’t know what impact of this mutation and what function it has, so it’s very difficult for us to understand the function of every mutation in our ligand map.

As to the future direction, in spite of the current progress, as well as the significance of the results discussed above, the developed methodologies developed could be extended into several other directions. Since the work on predicting the functional effect of each mutation will be the direction of future study, we plan to integrate the obtained annotation with the previous GWAS analyses to provide a more detailed analysis of the functional impact of mutations. In our research, we have already evaluated the frequency of each amino acid change patterns in the ligand-binding region, so we think this is a useful resource to predict what possible mutation will happen in the future by applying the machine learning method. If it is successful, it will make a great contribution to the diagnosis of some human diseases or abnormal reactions caused by a mutation in LBS in the future. Besides, we found that the mutation will change the binding affinity and affect the normal response of humans, so how much does the mutation change the binding affinity of protein-ligand interaction is the future direction that we need to consider. Recently, supervised learning have been introduced that directly predict the binding affinity of protein-ligand interactions [63-64], while not studying mutations effects, our project lays a foundation to study the semi-supervised regression to estimate the change in the binding affinity between the mutant and wild-type LBS, and it will be the first semi-supervised regression method in the fields of systems biology and genetics. How to get and predict more specific function and mechanism especially about the rare SNPs rewire protein-ligand interaction altering the function, we need to combine more experiment study and computational methods. But based on the results we have obtained so far and the development of machine learning and computational data mining methods, we remain optimistic about getting more accurate and meaningful conclusions.

References:

[1] Raphael, Benjamin J., et al. "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine." *Genome medicine* 6.1 (2014): 1-17.

[2] Ziebarth, Jesse D., Anindya Bhattacharya, and Yan Cui. "CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization." *Nucleic acids research* 41.D1 (2012): D188-D194.

[3] Pylkäs, Katri, et al. "Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network." *PLoS Gene*t 8.6 (2012): e1002734.

[4] Venables, Julian P. "Aberrant and alternative splicing in cancer." *Cancer research* 64.21 (2004): 7647-7654.

[5] Eswaran, Jeyanthy, et al. "RNA sequencing of cancer reveals novel splicing alterations." *Scientific reports* 3.1 (2013): 1-12.

[6] Chen, J., and W. A. Weiss. "Alternative splicing in cancer: implications for biology and therapy." Oncogene 34.1 (2015): 1-14.

[7] Welter, Danielle, et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic acids research* 42.D1 (2014): D1001-D1006.

[8] Sudmant, Peter H., et al. "An integrated map of structural variation in 2,504 human genomes." *Nature* 526.7571 (2015): 75-81.

[9] Sherry, Stephen T., et al. "dbSNP: the NCBI database of genetic variation." *Nucleic acids research* 29.1 (2001): 308-311.

[10] Stenson, Peter D., et al. "Human gene mutation database (HGMD®): 2003 update." *Human mutation* 21.6 (2003): 577-581.

[11] Amberger, Joanna S., et al. "OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders." *Nucleic acids research* 43.D1 (2015): D789-D798.

[12] Cui, Hongzhu, et al. "The variation game: Cracking complex genetic disorders with NGS and omics data." *Methods* 79 (2015): 18-31.

[13] David, Alessia, et al. "Protein–protein interaction sites are hot spots for disease-associated nonsynonymous SNPs." *Human mutation* 33.2 (2012): 359-363.

[14] Wang, Xiujuan, et al. "Three-dimensional reconstruction of protein networks provides insight into human genetic disease." *Nature biotechnology* 30.2 (2012): 159-164.

[15] Sahni, Nidhi, et al. "Widespread macromolecular interaction perturbations in human genetic disorders." *Cell* 161.3 (2015): 647-660.

[16] Pabinger, Stephan, et al. "A survey of tools for variant analysis of next-generation genome sequencing data." *Briefings in bioinformatics* 15.2 (2014): 256-278.

[17] Cooper, Gregory M., and Jay Shendure. "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data*." Nature Reviews Genetics* 12.9 (2011): 628-640.

[18] Ward, Lucas D., and Manolis Kellis. "Interpreting noncoding genetic variation in complex traits and human disease*." Nature biotechnology* 30.11 (2012): 1095-1106.

[19] Raphael, Benjamin J., et al. "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine." *Genome medicine* 6.1 (2014): 1-17.

[20] Wang, Kai, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." *Nucleic acids research* 38.16 (2010): e164-e164.

[21] Adzhubei, Ivan A., et al. "A method and server for predicting damaging missense mutations." *Nature methods* 7.4 (2010): 248-249.

[22] Kumar, Prateek, Steven Henikoff, and Pauline C. Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." *Nature protocols* 4.7 (2009): 1073.

[23] Nédélec, Yohann, et al. "Genetic ancestry and natural selection drive population differences in immune responses to pathogens." *Cell* 167.3 (2016): 657-669.

[24] Zhang, Wei, et al. "Evaluation of genetic variation contributing to differences in gene expression between populations." *The American Journal of Human Genetics* 82.3 (2008): 631-640

[25] Bachtiar, Maulana, and Caroline GL Lee. "Genetics of population differences in drug response." *Current Genetic Medicine Reports* 1.3 (2013): 162-170.

[26] Madian, Ashraf G., et al. "Relating human genetic variation to variation in drug responses." *Trends in genetics* 28.10 (2012): 487-495.

[27] Vogt, Austin D., and Enrico Di Cera. "Conformational selection is a dominant mechanism of ligand binding." *Biochemistry* 52.34 (2013): 5723-5729.

[28] Overington, John P., Bissan Al-Lazikani, and Andrew L. Hopkins. "How many drug targets are there?." *Nature reviews Drug discovery* 5.12 (2006): 993-996.

[29] Rask-Andersen, Mathias, Surendar Masuram, and Helgi B. Schiöth. "The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication." *Annual review of pharmacology and toxicology* 54 (2014): 9-26.

[30] Yang, Jianyi, Ambrish Roy, and Yang Zhang. "Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment." *Bioinformatics* 29.20 (2013): 2588-2595.

[31] Fuselli, Silvia. "Beyond drugs: the evolution of genes involved in human response to medications." *Proceedings of the Royal Society B* 286.1913 (2019): 20191716.

[32] Zanger, Ulrich M., and Matthias Schwab. "Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation." *Pharmacology & therapeutics* 138.1 (2013): 103-141.

[33] Brooijmans, Natasja, and Irwin D. Kuntz. "Molecular recognition and docking algorithms." *Annual review of biophysics and biomolecular structure* 32.1 (2003): 335-373.

[34] Leach, Andrew R., Brian K. Shoichet, and Catherine E. Peishoff. "Prediction of protein−ligand interactions. Docking and scoring: successes and gaps." *Journal of medicinal chemistry* 49.20 (2006): 5851-5855.

[35] Du, Xing, et al. "Insights into protein–ligand interactions: mechanisms, models, and methods." *International journal of molecular sciences* 17.2 (2016): 144.

[36] Betts, Matthew J., et al. "Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions." *Nucleic acids research* 43.2 (2015): e10-e10.

[37] Zhao, Junfei, et al. "Systematic prioritization of druggable mutations in∼ 5000 genomes across 16 cancer types using a structural genomics-based approach." *Molecular & cellular proteomics* 15.2 (2016): 642-656.

[38] Kim, Pora, et al. "mutLBSgeneDB: mutated ligand binding site gene DataBase." *Nucleic acids research* 45.D1 (2017): D256-D263.

[39] Bessman, Nicholas J., et al. "Complex relationship between ligand binding and dimerization in the epidermal growth factor receptor." *Cell reports* 9.4 (2014): 1306-1317.

[40] Ben-Levy, Rachel, et al. "An oncogenic point mutation confers high affinity ligand binding to the neu receptor. Implications for the generation of site heterogeneity." *Journal of Biological Chemistry* 267.24 (1992): 17304-17313.

[41] Zinger, Lotem, et al. "Ligand-binding Domain–activating Mutations of ESR1 Rewire Cellular Metabolism of Breast Cancer Cells." *Clinical Cancer Research* 25.9 (2019): 2900-2914.

[42] Guo, Yu Amanda, et al. "Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers." *Nature communications* 9.1 (2018): 1-14.

[43] Karczewski, Konrad J., et al. "The mutational constraint spectrum quantified from variation in 141,456 humans." *Nature* 581.7809 (2020): 434-443.

[44] Yang, Jianyi, Ambrish Roy, and Yang Zhang. "BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions." *Nucleic acids research* 41.D1 (2012): D1096-D1103.

[45] Pence, Harry E., and Antony Williams. "ChemSpider: an online chemical information resource." (2010): 1123-1124.

[46] Wishart, David S., et al. "DrugBank: a knowledgebase for drugs, drug actions and drug targets." *Nucleic acids research* 36.suppl_1 (2008): D901-D906.

[47] Kobren, Shilpa Nadimpalli, and Mona Singh. "Systematic domain-based aggregation of protein structures highlights DNA-, RNA-and other ligand-binding positions." *Nucleic acids research* 47.2 (2019): 582-593.

[48] Rogers, David J., and Taffee T. Tanimoto. "A computer program for classifying plants." *Science* 132.3434 (1960): 1115-1118.

[49] O'Boyle, Noel M., et al. "Open Babel: An open chemical toolbox." *Journal of cheminformatics* 3.1 (2011): 1-14.

[50] Wishart, David S., et al. "HMDB 3.0—the human metabolome database in 2013." *Nucleic acids research* 41.D1 (2012): D801-D807.

[51] Vitousek, Maren N., et al. "HormoneBase, a population-level database of steroid hormone levels across vertebrates." *Scientific data* 5.1 (2018): 1-7.

[52] Hewett, Micheal, et al. "PharmGKB: the pharmacogenetics knowledge base." *Nucleic acids research* 30.1 (2002): 163-165.

[53] Chen, Chengjie, et al. "TBtools: an integrative toolkit developed for interactive analyses of big biological data." Molecular plant 13.8 (2020): 1194-1202.

[54] Tagami, Tetsuya, et al. "A novel natural mutation in the thyroid hormone receptor defines a dual functional domain that exchanges nuclear receptor corepressors and coactivators." *Molecular Endocrinology* 12.12 (1998): 1888-1902.

[55] Saag, Michael, et al. "High sensitivity of human leukocyte antigen-b* 5701 as a marker for immunologically confirmed abacavir hypersensitivity in white and black patients." *Clinical infectious diseases* 46.7 (2008): 1111-1118.

[56] Hetherington, Seth, et al. "Genetic variations in HLA-B region and hypersensitivity reactions to abacavir." *The Lancet* 359.9312 (2002): 1121-1122.

[57] Mounzer, Karam, et al. "HLA-B* 57: 01 screening and hypersensitivity reaction to abacavir between 1999 and 2016 in the OPERA® observational database: a cohort study." *AIDS research and therapy* 16.1 (2019): 1-9.

[58] Martin, Michael A., and Deanna L. Kroetz. "Abacavir pharmacogenetics–from initial reports to standard of care." Pharmacotherapy: *The Journal of Human Pharmacology and Drug Therapy* 33.7 (2013): 765-775.

[59] Mwenifumbo, Jill C., et al. "Novel and established CYP2A6 alleles impair in vivo nicotine metabolism in a population of Black African descent." *Human mutation* 29.5 (2008): 679-688.

[60] Nizamuddin, Sheikh, et al. "CYP2C9 Variations and Their Pharmacogenetic Implications Among Diverse South Asian Populations." *Pharmacogenomics and Personalized Medicine* 14 (2021): 135.

[61] Dai, D. P., et al. "CYP2C9 polymorphism analysis in Han Chinese populations: building the largest allele frequency database." *The pharmacogenomics journal* 14.1 (2014): 85-92.

[62] Teixeira, Raquel LF, et al. "Sequence analysis of NAT2 gene in Brazilians: identification of undescribed single nucleotide polymorphisms and molecular modeling of the N-acetyltransferase 2 protein structure." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 683.1-2 (2010): 43-49.

[63] Jiménez, José, et al. "K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks." *Journal of chemical information and modeling* 58.2 (2018): 287-296.

[64] Wang, Debby D., et al. "Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods." *Computational and structural biotechnology journal* 18 (2020): 439-454.