Special Issue
"Protein-Ligand Interactions"

Vol. 13, pp. 157–163 (2016)
doi: 10.2142/biophysico.13.0_157

Biophysics and Physicobiology

https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Structural characterization of single nucleotide variants at ligand binding sites and enzyme active sites of human proteins

Kazunori D. Yamada[1,*], Hafumi Nishi[1,*], Junichi Nakata[2] and Kengo Kinoshita[1,2,3]

[1]*Graduate School of Information Sciences, Tohoku University, Miyagi 980-8597, Japan*
[2]*Tohoku Medical Megabank Organization, Tohoku University, Miyagi 980-8573, Japan*
[3]*Institute of Development, Aging, and Cancer, Tohoku University, Miyagi 980-8575, Japan*

Received January 7, 2016; accepted April 10, 2016

Functional sites on proteins play an important role in various molecular interactions and reactions between proteins and other molecules. Thus, mutations in functional sites can severely affect the overall phenotype. Progress of genome sequencing projects has yielded a wealth of information on single nucleotide variants (SNVs), especially those with less than 1% minor allele frequency (rare variants). To understand the functional influence of genetic variants at a protein level, we investigated the relationship between SNVs and protein functional sites in terms of minor allele frequency and the structural position of variants. As a result, we observed that SNVs were less abundant at ligand binding sites, which is consistent with a previous study on SNVs and protein interaction sites. Additionally, we found that non-rare variants tended to be located slightly apart from enzyme active sites. Examination of non-rare variants revealed that most of the mutations resulted in moderate changes of the physico-chemical properties of amino acids, suggesting the existence of functional constraints. In conclusion, this study shows that the mapping of genetic variants on protein structures could be a powerful approach to evaluate the functional impact of rare genetic variations.

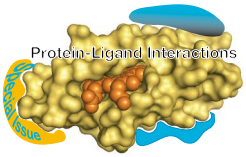
Key words: rare variant, non-synonymous mutation, protein-ligand interaction, 3D structure

Molecular interactions between proteins and other molecules play an essential role in numerous cellular processes. Functional sites on protein surfaces, such as protein-ligand binding sites and enzyme active sites, are the most important elements of these interactions. Mutations in the functional sites will change the affinity of protein-ligand binding and the protein's catalytic activity, which can result in a negative phenotypic outcome. It is known that even mutations with no significant effect under normal circumstances can change the strength of protein-drug interactions [1], stressing the importance of understanding mutations in terms of protein-ligand binding and enzymatic activity. In general, linking information about specific mutations to biological sequence data and/or protein structural data would be helpful in studying the effect of mutations on protein function [2–4]. More-

* These authors contributed equally to this work.
Corresponding author: Kengo Kinoshita, Graduate School of Information Sciences, Tohoku University, 6-3-09 Aramaki-asa-Aoba, Aoba-ku, Sendai, Miyagi 980-8597, Japan. e-mail: kengo@ecei.tohoku.ac.jp

◀ ***Significance*** ▶

Functional sites on proteins play a key part in molecular interactions and reactions. Thus, mutations in functional sites can severely affect the overall phenotype. To understand the influence of genetic variants at a protein level, we investigated the relationship between single nucleotide variants and functional sites in terms of minor allele frequency and the structural position of variants. We found that variants were less abundant at ligand binding sites, and non-rare variants tended to be located slightly apart from enzyme active sites. Most of non-rare mutations were moderate changes of the physico-chemical properties, suggesting the existence of functional constraints.



over, locating mutations on known protein 3D structures is an established approach for assessing the relevance of mutations [5–7].

Advancements in genome sequencing projects have yielded a wealth of information on genetic variants among human populations, such as single nucleotide variants (SNVs). Population size, in particular, has seen a great increase with every study, allowing the identification of rare variants with less than 1% minor allele frequency (MAF) [8,9]. Rare variants are thought to bear a greater responsibility for severe diseases than do common variants [10], and it was reported that rare variants were more abundant among ligand binding sites and enzyme active sites compared to other factors related with structural stability [9]. However, rare variants are problematic when attempting to elucidate the association between variants and phenotypes in conventional statistical analyses due to their low frequency. Therefore, in order to gain a deeper understanding of variation between individual genomes, it is crucial to characterize the molecular details of rare variants and their relationship to a particular phenotype.

As the number of protein structures deposited in the Protein Data Bank (PDB) has reached more than a hundred thousand macromolecular structures [11], a considerable number of computational studies have reported on protein structural features at mutation sites and the altered stability and dynamics caused by non-synonymous mutations [12–15]. However, most studies have focused only on differences between disease mutations and non-disease mutations. As a result, our knowledge of the effect of genetic variants on proteins is still insufficient. Moreover, only a few attempts have been made to directly link known variants, regardless of their severity or deleteriousness, to protein structures [16,17].

Recently, we investigated the relationship between SNVs and protein-protein interaction sites in terms of MAF and the structural positions of variants. We showed that common variants followed functional constraints, whereas rare variants were found at random positions along DNA sequences [17]. In the present study, we expanded our investigations to interactions between proteins and other molecules, with the aim of defining a more comprehensive picture of the impact genetic variants have on protein structures. Here we focused on the association of SNVs with protein-ligand interaction sites and enzyme active sites, in order to evaluate the impact of natural variants on such protein functions. Non-synonymous variants were obtained from the NHLBI Exome Sequencing Project [18] and then mapped onto protein 3D structures. Upon integrating variant data with functional site information, variants were further classified into different categories based on MAF. We examined the occurrence of variants at functional sites, as well as the spatial relationship between enzyme active sites and variants with different MAFs. Individual cases of variants with higher MAF also gave us some clues for the general characteristics of variants at functional sites.

Materials and Methods

Integration of variant data with functional site data

Single nucleotide variants were retrieved from the NHLBI Exome Sequence Project website [18] and mapped onto protein structures as previously described [17]. In brief, of the 1,936,451 variants, 1,074,023 mutations were mapped to RefSeq [19] protein sequences, and the corresponding position was searched for in the PDB [11] using BLAST [20]. The threshold for sequence identity between the query sequence and the structure was set to 90% for the protein-ligand dataset and 50% for the active site dataset due to the limited annotations of enzyme active sites. Additionally we removed structures with a resolution worse than 3 Å from the protein-ligand dataset.

Functional sites on protein structures were identified by transferring annotations from BioLiP [21] for ligand binding sites and UniProt [22] for active sites. BioLiP is a semi-manually curated database that includes biologically relevant protein-ligand binding interactions. It compiles protein-ligand binding structural data from PDB, based on literature search and referencing to other, more specific databases. This version of BioLiP (updated on May 30, 2015) contained information on 315,241 protein-ligand binding sites, derived from 67,803 protein structures. In BioLiP, a residue is identified as a ligand-binding site residue, if the closest distance between the residue and the target ligand is less than 0.5 Å in addition to the sum of van der Waals radii of the atoms. We classified the residues of proteins into internal and surface residues, based on their relative accessible surface area (rASA) and used ligand-binding residues on protein surface (rASA>25%). A sequence identity threshold of 40% was employed in order to eliminate any redundancy among examined protein structures. The clustering of the proteins was performed using kClust [23]. In total, we analyzed 230 variants at ligand binding sites (16,390 for entire proteins) among 977 structures in PDB. For active sites on enzymes, we referred to the "active site" annotation in UniProt, and removed the redundancy with the same criterion as for the ligand binding site dataset. However, we employed both, the non-redundant and the original dataset (the dataset before removing redundancy) for active site analyses, due to the small size of the dataset and the nature of catalytic sites. Catalytic site residues tend to be conserved even among divergent proteins. The original dataset contained 49 variants at the enzyme active sites (8,117 for entire proteins) among 670 structures in PDB. For analyses of the non-redundant dataset, 39 variants (6,008 for the entire proteins) among 513 structures were used.

Finally, variants in the datasets were classified into three categories according to their MAF: rare (MAF$\leq$1%), intermediate (1%<MAF$\leq$5%), and common (MAF>5%).

Results and Discussion

Single nucleotide variants at ligand binding sites

We mapped single nucleotide variants on protein structures and classified them into surface, interior, and ligand binding sites, as described in the materials and methods section. Table 1 shows the statistics relative to the number of variants and non-variants in the classification. In this calculation, we analyzed 977 protein structures consisting of 291,913 residues in total. These were further divided into 16,390 variant and 275,523 non-variant sites. First, we focused on the types of binding ligands for all ligand binding sites and 220 binding site where rare variants were mapped. Table 2 shows the top 10 most frequently observed binding ligands for all binding sites and the rare-variant mapped binding sites. As a result, both types of binding sites had the majority of ligands in common and there was no clear difference on the orders of the ligands. In addition, the tendency was the same when the whole lists of ligands were compared. The Spearman rank-correlation coefficient was 0.702 and it was statistically significant (1.12×10^{-12}, according to an association test). Therefore, we concluded that there was no difference on binding ligand types between all binding sites and rare-variant mapped sites at least in the present dataset. However, since there exist a slight tendency of more weighted or larger ligands such as DNA/RNA and peptides to be highly ranked among the rare-variant mapped sites, it might be possible to find some relationship between frequencies of mutations and sizes of binding ligand types with more data available in the future. Next, we focused on the location of mutated residues. The number of mutated residues at ligand binding sites was only 230 (1.4%). Mutation rates of residues at binding sites and protein surface were 5.44% and 6.51%, respectively, indicating that mutations at ligand binding sites occurred less frequently than those on protein surfaces (p-value=0.00475 with Fisher's exact test). The ratio of rare variants at ligand binding sites to those in whole proteins was about 1.4% (220/15813). This result coincided with the previous study [9] which reported the ratio was 1.7%. Table 1 also includes a breakdown of the mutations, which were categorized into three classes, depending on their allele frequency. Accordingly, almost all mutations were identified as rare variants and there were only ten common variants (Table 3). This finding seemed reasonable, because mutations at functional sites would be eliminated by natural selection and their frequency would not increase,

Table 1 Statistics of variants and non-variants at ligand binding sites and other different locations of proteins

	Binding site	Non-binding surface	Interior	Total
All variants	230	9,070	7,090	16,390
Rare	220	8,709	6,884	15,813
Intermediate	4	187	106	297
Common	6	174	100	280
Non-variants	3,995	130,181	141,347	275,523
Total	4,225	139,251	148,437	291,913

Table 2 Top 10 ligands for all binding sites and the sites where rare variants were mapped. All ligands except peptides and DNA/RNA are represented by the PDB three-letter codes

All binding sites			Rare-variant mapped sites		
Rank	Ligand type	Count	Rank	Ligand type	Count
1	Peptide	90	1	DNA/RNA	42
1	CA	90	2	Peptide	22
3	ZN	85	3	CA	12
4	DNA/RNA	52	3	NAP	12
5	MG	41	5	ADP	8
6	ADP	26	6	FAD	7
7	SAH	18	7	SAM	6
8	NAP	17	8	NDP	5
9	FAD	15	8	NAD	5
10	ANP	13	10	ZN	4
10	NAD	13	10	HEM	4

Table 3 Intermediate and common variants at ligand binding sites. MAF (EA): minor allele frequency in the European American population, MAF (AA): minor allele frequency among African Americans, SAS: sulfasalazine, PNP: 4-nitrophenyl hydrogen methylphosphonate, CA: calcium ion, FAD: flavin adenine dinucleotide

Mutation	Protein	Ligand	MAF (EA)	MAF (AA)	MAF (All)	RS number
I105V	GSTP1	SAS	33.27	41.96	36.08	rs1695
D101N	HLA-A	peptide	31.37	26.85	29.84	rs1136688
L273M	ALPPL2	PNP	25.31	31.24	27.34	rs17416141
K186E	KLK1	CA	26.86	19.63	24.41	rs5517
E750D	LPHN1	peptide	1.116	18.23	6.912	rs41276898
T117S	CYB5R3	FAD	0.06980	27.49	9.357	rs1800457
T97I	HLA-A	peptide	3.716	7.461	4.985	rs1136688
D197N	DOK7	peptide	0.05810	6.446	2.222	rs16844422
A114T	PABPC3	peptide	1.535	0.4539	1.169	rs117014540
Y89H	HLA-DRB1	peptide	1.086	0.8700	1.013	rs17882583

unless there were other reasons. To examine why some common variants were found at ligand binding sites, we focused on some exceptional cases.

Examination of intermediate and common variants at ligand binding sites

Of ten exceptional mutations, three cases (D101N, Y97I, and Y89H) had occurred on the human leukocyte antigen (HLA), which is known to be highly polymorphic [24–26]. The occurrence of these mutations could be explained by the immunological need for genetic variation. Six mutations were found to result in similar amino acids [27]: I105V on glutathione S-transferase pi 1 (GSTP1), L273M on alkaline phosphatase placental like 2 (ALPPL2), T117S on cytochrome b5 reductase 3 (CYB5R3), D197N on docking protein 7 (DOK7), A114T on poly(A) binding protein, cytoplasmic 3 (PABPC3) and E750D on latrophilin 1 (LPHN1). Mutations between similar amino acids will have a reduced functional impact and could spread among the population, even when they are located at ligand binding sites, as in the case of ALPPL2 (Fig. 1A). Another illustrative case of substitutions between similar amino acids was the T117S mutation on CYB5R3, which was characterized by a large difference of MAF among populations, as discussed later (Fig. 1B). The last case was that of the K186E mutation on kallikrein 1 (KLK1), where a basic amino acid (Lys) was replaced with an acidic amino acid (Glu) (Fig. 1C).

We chose to examine further the T117S mutation at the flavin adenine dinucleotide (FAD) binding site of CYB5R3 (Fig. 1B). CYB5R3 is an NADH-dependent enzyme that catalyzes several redox reactions, such as the conversion of methemoglobin to hemoglobin, using FAD as a cofactor [28]. This mutation was originally found to be a high frequency polymorphism specific to the African American population [29]. Subsequent studies reported that the mutation was not associated with any significant changes to protein function and structure [30–32]. The result made sense when looking at protein structure, because the residue was on a loop region and it interacted with FAD through a hydrogen bond between the oxygen atom of side chain of threonine and the nitrogen atom of FAD. Interestingly, this case showed a large difference between MAF of African American (AA) and European American (EA). The MAF of AA and EA were 27.5% and 0.0698%, respectively. To identify the possible cause of this difference, we examined the rs1800457 mutation from the HapMap project [33]. We found this allele to be heterozygous only in Yoruba Africans (YRI), while homozygous in other populations, such as European (CEU), Japanese Tokyo (JPT), and Han Chinese (CHB). YRI are the people residing in Nigeria, of which a large population had moved to the U.S. in the past [34]. Based on this observation, we assume that the mutation originated from a bottleneck effect [35]. The allele might be derived from a small population, who had transferred from the region to the U.S. in the past and gradually expanded from there. Since the mutation does not have any harmful effect on the protein, the mutation continues to exist without being replaced by the dominant allele. Therefore, we concluded that the biased distribution of the T117S mutation derived from a specific population and it was maintained because it did not adversely affect protein function.

Additionally, we examined the K186E mutation, as an example of a radical change of amino acid properties (Fig. 1C). The mutation occurred at the calcium ion (Ca) binding site of KLK1 protein. The protein is a serine protease and its active site consists of H65, D120, and S214. The mutation site was far from these residues, such that the distances between the Cα carbon atom of the mutation site and each active site residue were 32.8 Å, 17.0 Å, and 20.6 Å, respectively. The mutation site is located on the edge of the protein, as shown (Fig. 1C). In spite of several reports about the mutation, it remains unclear whether the mutation has any effect on protein function [36,37]. The Ca ion interacts with the protein through the main chain atoms of lysine. In

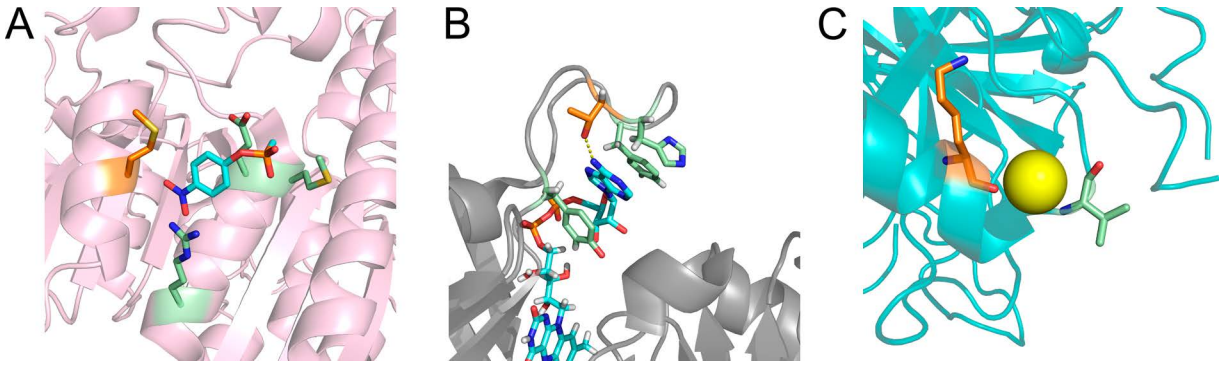


Figure 1 Examples of non-rare variants at ligand binding sites. A: alkaline phosphatase (colored in pink) with 4-nitrophenyl hydrogen methylphosphonate (cyan) (PDBID: 1zed), B: cytochrome b5 reductase 3 (gray) with flavin adenine dinucleotide (cyan) (PDBID: 3w5h), C: kallikrein 1 (blue) with a calcium ion (yellow) (PDBID: 1spj). Variant sites are shown with an orange stick model, and other ligand binding residues are shown as a light green stick. Note that not all ligand binding residues are presented in the figure.

general, this kind of interaction would be biologically non-specific. While solving the protein structure of KLK1 [38], it was reported that the Ca ion mediated an interaction between the protein molecule and the symmetry-related molecule. That is, the Ca ion promotes crystallization, even though this binding site has been defined as a biologically specific site by the BioLiP database. We nevertheless concluded that the interaction was not specific. The apparent discrepancy could be attributed to the fact that the BioLiP database considered all metal ions, except sodium, as potential biologically relevant ligands, unless specified otherwise in the literature [21]. Since metal ions often play an important role in protein function, such as activation of polymerases, they should not be excluded a priori as non-specific ligands. In fact, it would be desirable that the mode of interaction between ligands and proteins is examined in more detail in order to exclude misannotation in databases.

In summary, irregular mutations that do not occur on a residue essential for protein-ligand binding, become fixed in a population. We could not find any unexpected case, such as a mutation with a highly radical change of amino acid properties, regardless of extremely high MAF. Such an example had been reported in our previous study on protein-protein interaction sites [17].

Single nucleotide variants and enzyme active sites

As many ligand-binding processes include enzyme-substrate interactions, we also performed the above analyses on enzyme active sites to investigate the potential effect of genetic variants on catalytic activity and substrate binding. Forty-nine out of 8,117 variants on enzymes were identified at active sites, indicating that the overall ratio of variants among active sites was about 5% (4.89% in the original and 5.08% in the non-redundant set, respectively). This value was slightly lower than that reported for other functional sites (protein-ligand binding sites: 5.44%, protein-protein interaction sites: 5.54% [17]). In addition, 48 of 49 active site variants were rare variants, whereas no common variants were observed in our dataset (Table 4). These results suggest that enzyme activity is crucial for biological processes, to the point that any loss or reduction in catalytic activity or binding affinity may have negative phenotypic consequences, untenable for the population. Next, we examined the spatial distances and hence the influence of variants on the corresponding active sites. Figure 2 shows the distribution of Cα distances between rare/non-rare variants and their closest active sites, with zero indicating that variants were directly mapped on the active sites. The data clearly illustrate that non-rare (common and intermediate) variants tend to be located farther from the active sites, as compared to the rare variants (p-value: 2.5×10^{-4} using the Wilcoxon rank-sum test). This trend could be partially explained by the bias of location of rare variants and common variants, since it is known that common variants tend to be abundant on protein surface and rare variants are basically distributed on a protein evenly. As shown in Figure 2, the distribution of distances between rare variants and active sites overlapped with that of distances of any protein interior residues from active sites, whereas the spatial distances of non-rare variants have intermediate characteristics between those of protein interior and surface.

Table 4 Statistics of variants and non-variants at enzyme active sites. Numbers corresponding to the non-redundant dataset are shown in parenthesis

All variants				Non-variants	Total
Rare	Intermediate	Common	Total		
48 (39)	1 (0)	0 (0)	49 (39)	953 (728)	1,002 (767)

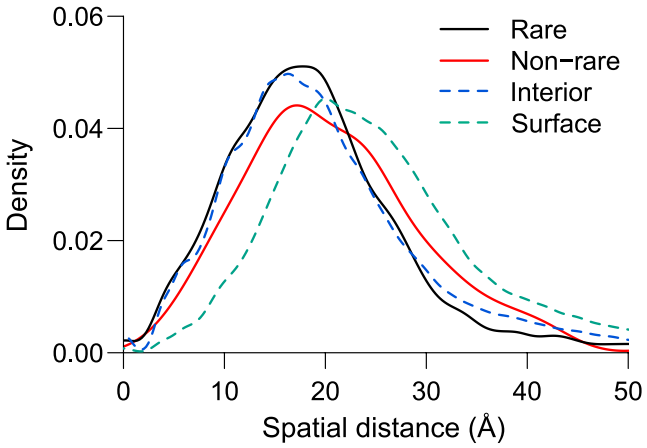


Figure 2 The spatial distances of variants and all residues from active sites in the non-redundant set. Black: rare variants, red: non-rare (intermediate and common) variants, blue: protein interior residues, green: protein surface residues. Note that the total numbers of rare and non-rare variants are 5,817 and 191, respectively.

Variants at enzyme active sites - Case study

In our dataset of variants surrounding active sites, we found a non-rare variant: H66R from chymase (rsid: rs5247). Chymase is a chymotrypsin-like serine protease and is known as a major protease secreted by mast cells. It is thought that chymase is involved in vasoactive peptide generation, extracellular matrix degradation, and regulation of gland secretion. The targets of chymase are many; however, the functional details of this enzyme are still unclear [39]. The histidine residue of chymase is one of the catalytic residues, along with Asp110 and Ser203 (Fig. 3). MAF of the H66R variant was 2.1% among the European American population, and 0.34% among the African American population. His66 acts as a general base in the catalytic triad, implying that the substitution from His to Arg will disrupt the catalytic activity of chymase. Although several genome-wide association studies attempted to elucidate the relationship between

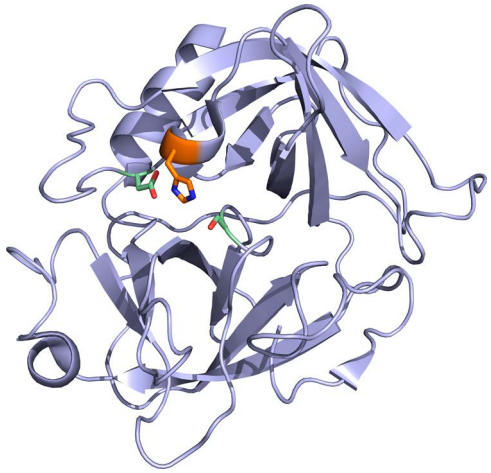


Figure 3 Example of non-rare variants in the active sites of chymase (PDBID: 4afq). The active sites of chymase (Ser62, His66, and Asp89) are shown as sticks. The His66 variant site is colored in orange, non-variant sites Ser62 and Asp89 are shown in light green.

certain diseases and chymase variants [40, 41], no clinical or biochemical association with the H66R variant has been reported. Further experimental investigation is needed to elucidate the biological meaning and the phenotypic outcome of this variant.

Conclusion

In this study, we investigated the relation between single nucleotide variants and functional sites. We showed that SNVs were less abundant at protein-ligand binding sites, and non-rare variants tended to be located slightly apart from enzyme active sites. Investigation of intermediate and common variants at functional sites revealed that the amino acid substitutions of most non-rare variants resulted in only moderate changes of physico-chemical properties, making these mutations neutral to the population. We observed that the trends among the variants on protein-ligand binding sites and enzyme active sites were similar to those of protein-protein interaction sites, whereas the ratios of variants among functional sites were slightly different. As a result, our study shed new light on the role of genetic variants with different MAFs on protein structures. This may lead to a better understanding of the functional impact of rare variants at the protein level.

Acknowledgments

The research was partially supported by the Platform Project for Supporting in Drug Discovery and Life Science Research (Platform for Drug Discovery, Informatics, and Structural Life Science) from Japan Agency for Medical Research and Development (AMED). K. K. was also supported by JSPS KAKENHI Grant Number 15H02773. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Author Contribution

K. K. conceived this study. K. D. Y., J. N., and K. K. prepared the datasets. K. D. Y., H. N., and K. K. performed the statistical analyses and manual investigations. K. D. Y., H. N., and K. K. co-wrote the manuscript.

References

[1] Ma, Q. & Lu, A. Y. Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacol Rev.* **63**, 437–459 (2011).
[2] Reeves, G. A., Thornton, J. M. & BioSapiens Network of Excellence. Integrating biological data through the genome. *Hum. Mol. Genet.* **15**, R81–87 (2006).
[3] Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., *et al.* Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **101**, 10205–10210 (2004).
[4] Oresic, M., Clish, C. B., Davidov, E. J., Verheil, E., Vogels, J., Havekes, L. M., *et al.* Phenotype characterisation using integrated gene transcript, protein and metabolite profiling. *Appl. Bioinformatics* **3**, 205–217 (2004).
[5] Pallan, P. S., Lei, L., Wang, C., Waterman, M. R., Guengerich, F. P. & Egli, M. Research Resource: Correlating Human Cytochrome P450 21A2 Crystal Structure and Phenotypes of Mutations in Congenital Adrenal Hyperplasia. *Mol. Endocrinol.* **29**, 1375–1384 (2015).
[6] Okawa, T., Yoshida, M., Usui, T., Kudou, T., Iwasaki, Y., Fukuoka, K., *et al.* A novel loss-of-function mutation of *GATA3* (p.R299Q) in a Japanese family with Hypoparathyroidism, Deafness, and Renal Dysplasia (HDR) syndrome. *BMC Endocr. Disord.* **15**, 66 (2015).
[7] Morra, S., Maurelli, S., Chiesa, M., Mulder, D. W., Ratzloff, M. W., Giamello, E., *et al.* The effect of a C298D mutation in CaHydA [FeFe]-hydrogenase: Insights into the protein-metal cluster interaction by EPR and FTIR spectroscopic investigation. *Biochim. Biophys. Acta* **1857**, 98–106 (2016).
[8] Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16 (2015).
[9] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
[10] Manolio, T. A., Collins, F, S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
[11] Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **43**, D345–356 (2015).
[12] Doss, C. G. & Nagasundaram, N. Investigating the structural impacts of I64T and P311S mutations in APE1-DNA com-

plex: a molecular dynamics approach. *PLoS ONE.* **7**, e31677 (2012).
[13] Kumar, A. & Purohit, R. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLoS Comput. Biol.* **10**, e1003318 (2014).
[14] David, A. & Sternberg, M. J. The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease. *J. Mol. Biol.* **427**, 2886–2898 (2015).
[15] Lu, H. C., Chung, S. S., Fornili, A. & Fraternali, F. Anatomy of protein disorder, flexibility and disease-related mutations. *Front. Mol. Biosci.* **2**, 47 (2015).
[16] Mueller, S. C., Backes, C., Kalinina, O. V., Meder, B., Stöckel, D., Lenhof, H. P., *et al.* BALL-SNP: combining genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms. *Genome Med.* **7**, 65 (2015).
[17] Nishi, H., Nakata, J. & Kinoshita, K. Distribution of single-nucleotide variants on protein-protein interaction sites and its relationship with minor allele frequency. *Protein Sci.* **25**, 316–321 (2015).
[18] *Exome Variant Server*. NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, (URL: http://evs.gs.washington.edu/EVS/) [Aug, 2013 accessed].
[19] Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashym, A., Ermolaeva, O., *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–763 (2014).
[20] Altschul, S. F., Madden, T. L., Shäffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
[21] Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096–1103 (2013).
[22] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–212 (2015).
[23] Hauser, M., Mayer, C. E. & Soding, J. kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics* **14**, 248 (2013).
[24] Reimann, J. & Miller, R. G. Polymorphism and MHC gene function. *Dev. Comp. Immunol.* **7**, 403–412 (1983).
[25] Bergstrom, T. & Gyllensten, U. Evolution of Mhc class II polymorphism: the rise and fall of class II gene function in primates. *Immunol. Rev.* **143**, 13–31 (1995).
[26] Glass, E. J. Genetic variation and responses to vaccines. *Anim. Health Res. Rev.* **5**, 197–208 (2004).
[27] Taylor, W. R. The classification of amino acid conservation. *J. Theor. Biol.* **119**, 205–218 (1986).
[28] Elahian, F., Sepehrizadeh, Z., Moghimi, B. & Mirzaei, S. A. Human cytochrome b5 reductase: structure, function, and potential applications. *Crit. Rev. Biotechnol.* **34**, 134–143 (2014).
[29] Jenkins, M. M. & Prchal, J. T. A novel mutation found in the 3' domain of NADH-cytochrome B5 reductase in an African-American family with type I congenital methemoglobinemia. *Blood* **87**, 2993–2999 (1996).
[30] Jenkins, M. M. & Prchal, J. T. A high-frequency polymorphism of NADH-cytochrome b5 reductase in African-Americans. *Hum. Genet.* **99**, 248–250 (1997).
[31] Sacco, J. C. & Trepanier, L. A. Cytochrome b5 and NADH cytochrome b5 reductase: genotype-phenotype correlations for hydroxylamine reduction. *Pharmacogenet. Genomics* **20**, 26–37 (2010).
[32] Roma, T. Systematic analysis of structure-function relationships of conserved sequence motifs in the NADH-binding lobe of cytochrome b5 reductase. (*Graduate These, University of South Florida*, 2008).
[33] The International HapMap 3 Consortium, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
[34] Lovejoy, P. The impact of the atlantic slave trade on Africa: a review of the literature. *J. Afr. Hist.* **30**, 365–394 (1989).
[35] Nei, M. Bottlenecks, genetic polymorphism and speciation. *Genetics* **170**, 1–4 (2005).
[36] Baker, A. R. & Shine, J. Human kidney kallikrein: cDNA cloning and sequence analysis. *DNA* **4**, 445–450 (1985).
[37] Fukushima, D., Kitamura, N. & Nakanishi, S. Nucleotide sequence of cloned cDNA for human pancreatic kallikrein. *Biochemistry* **24**, 8037–8043 (1985).
[38] Laxmikanthan, G., Blaber, S. I., Bernett, M. J., Scarisbrick, I. A., Juliano, M. A. & Blaber, M. 1.70 Å X-ray structure of human *apo* kallikrein 1: structural changes upon peptide inhibitor/substrate binding. *Proteins* **58**, 802–814 (2005).
[39] de Souza Junior, D. A., Santana, A. C., da Silva, E. Z., Oliver, C. & Jamur, M. C. The Role of Mast Cell Specific Chymases and Tryptases in Tumor Angiogenesis. *Biomed Res. Int.* **2015**, 142359 (2015).
[40] Orlowska-Baranowska, E., Gora, J., Baranowski, R., Stoklosa, P., vel Betka, L. G., Pedzich-Placha, E., *et al.* Association of the common genetic polymorphisms and haplotypes of the chymase gene with left ventricular mass in male patients with symptomatic aortic stenosis. *PLoS ONE* **9**, e96306 (2014).
[41] Weidinger, S., Rümmler, L., Klopp, N., Wagenpfeil, S., Baurecht, H. J., Fischer, G., *et al.* Association study of mast cell chymase polymorphisms with atopy. *Allergy* **60**, 1256–1261 (2005).