Pharmacogenomics of CYP2D6: A Study of CYP2D6 and its Allelic Variants using *in silico* Approaches

---

A thesis submitted in fulfilment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

Department of Biochemistry, Microbiology and Bioinformatics

Rhodes University

By

Chiratidzo Respina Chamboko

ORCID ID: https://orcid.org/0000-0002-7631-4474

March 2025

Abstract

Over the past few years, pharmacogenomics research has gained more attention due to the growing interest in personalized medicine. Polymorphisms in pharmacogenes have greatly affected drug therapies and this either causes low drug efficacy or adverse drug reactions in humans. Specifically, polymorphisms in pharmacogenes that encode drug metabolizing enzymes from the first three families of human cytochrome P450 (CYP) enzyme enzymes have been major contributors, as these enzymes metabolize numerous xenobiotics and drugs on the market today. Cytochrome P450 2D6 (CYP2D6), a CYP2 family enzyme, accounts for at least 25% of drug metabolism, making it one of the key enzymes for pharmacogenomic research. Studies on this enzyme will essentially lead to personalized medical therapies for a wide range of drugs. In the research conducted here, investigations were conducted to advance 3D pharmacogenomic research on CYP2D6, with an emphasis on African populations, which have been underrepresented in this field. The first aim of this study was to characterize enzymes within CYP families 1, 2 and 3, so as to uncover structural and sequential characteristics that may be shared among all enzymes within the three families, and identify those that may be unique to each family, and particularly, unique to CYP2D6. *In silico* processes such as multiple sequence alignment, motif analysis and phylogenetic tree construction were employed to analyze sequences from the three families. Through analysis, it was observed that the three enzyme families contain defined conservation in regions close to or in contact with the heme group, however unique patterns were detected within each CYP enzyme family, and these regions were located towards the helices responsible for substrate specificity and active site access. This indicated relationships within families were stronger, and phylogenetic analysis highlighted this with family-wise clustering of sequences observed. With the results obtained from this research, it is evident that similarities within the three families do exist, but each family contains unique sequence and structural features that possibly relate to the different drugs they metabolize. Interestingly, CYP2D6 was noted to have fewer unique regions in comparison to other enzymes within the CYP2 family, which may explain its ability to metabolize a wider variety of drugs than other enzymes. The next aim was to generate a potentially more effective way of assessing CYP2D6 reduced function, by investigating the possible variant-induced mechanisms leading to the reduced activity of CYP2D6, then thereafter

apply these mechanisms to Afrocentric alleles with unknown function at the time of this research, to observe if any likely harbored this activity. Several computational techniques and tools were utilized to achieve this aim. The analysis was conducted with CYP2D6 alleles *2, *10, *14, *17, *27, *29, *33, *34, *39, *45, *48, *49, *50, *53, *54, and *55 with known normal or reduced function, and Afrocentric alleles *149, *152, *153, *154, *155, *157, *159, *160, *162 and *163 with unknown clinical activity. Molecular dynamics (MD) simulations and traditional post-MD analyses, including RMSD and Rg, were used to assess structural stability changes in CYP2D6 alleles. Geometric clustering was applied to evaluate allele conformations after stabilization and secondary structure alterations were analyzed with the DSSP tool and ΔRMSF. Hydrogen bonding throughout the different systems was also conducted to observe changes in hydrogen bonding due to variation presence. We examined heme interactions by measuring the center of mass between the heme and the active site, hydrogen bonding, and contacts with surrounding residues, then tunnel analysis of whole protein systems was conducted to observe dominant tunnel opening frequencies. Finally, DRN metrics were employed to detect variation-induced changes in allele communication. The key findings within this study prompted the proposal of four mechanism that potentially lead to reduced activity within CYP2D6 alleles, which are: 1) instability and conformational changes of protein; 2) secondary structure variation and changes in flexibility of ≥3 highly important protein regions; 3) more frequently closed/ more than one frequently open or partially open channel simultaneously; 4) unique changes in inter-residue communications within the B' helix and/or β1 sheet region. We also proposed from data analyzed, that when predicting reduced functionality on alleles with unknown activity, the allele must have at least three of the mechanisms occurring. This then led to the potential prediction of reduced activity among some of the Afrocentric alleles studied here. Overall, this study contributes to the pharmacogenomic research on CYP2D6 enzymes. It unveiled the effects of variations in CYP2D6 alleles and identified potential mechanisms that may bring about reduced activity of these alleles. This study also contributed to the knowledge of functionality of novel Afrocentric alleles with no clinical activity as yet, that could potentially be useful in personalized medicine.

Declaration

I, Chiratidzo Respina Chamboko, declare that the thesis submitted to Rhodes University is my own work and has not previously been submitted for a degree or diploma at this or any other institution.

Signature:  **Date:** .....06/03/2025……

Acknowledgements

I would wholeheartedly like to express my deepest gratitude to my supervisor, Prof. Özlem Tastan Bishop, for her unwavering support and guidance throughout my PhD journey. Her expertise, thoughtful advice, and encouragement were invaluable at every stage of this research. Thank you for believing in me; I truly appreciate it.

To my colleagues in the research unit in bioinformatics (RUBi) team, thank you for all your kind assistance, friendship, and support during the course of this research. I appreciate all the stimulating conversations and happy memories we have shared.

A special thank you to my parents, Dr. Tafireyi Leo Chamboko and Mrs. Gloria Miriam Changundega Chamboko. Your continued support throughout my educational journey has been indispensable, and I am deeply grateful for your unwavering encouragement and belief in me. Truly you have sacrificed a lot for me to become the person I am today, and for this, I am eternally grateful. Thank you to my loving siblings, Kudakwashe Leo Tafireyi Chamboko and Tanyaradzwa Michelle Chamboko, for always supporting me. The unconditional love was greatly felt, and I appreciate everything you have done to brighten my days. Thank you to the members of the Chamboko family at large who have also supported and cheered me on through this journey, I really appreciate. To my friends Samantha Soweto, Tarisai Tafirenyika, Cecilia Munamati, Ruvimbo Zishiri, Laura Zenenga, Sinalo Gqunu, and my colleagues turned friends, Dr Jessica Okeke and Dr Tendai Muroonzi, I appreciate the checkups, support, and love you have shown me during this time, thank you.

This research was funded by the National Research Foundation of South Africa (Grant Number: MND200623535654) and Project Africa Gradient (Genomic Research Approach for Diversity and Optimizing Therapeutics) (SAMRC-RFA-GSK/NVS 01-2021). Thank you for granting me the opportunity to study further. A thank you to the Center for High Performance Computing (CHPC), South Africa, for providing the large computational resources required for this research.

To Almighty God above, none of this would have been possible without You.

“For I know the plans I have for you," declares the LORD, "plans to prosper you and not to harm you, plans to give you hope and a future.” ~Jeremiah 29:11.

Dedication

I dedicate this to my loving parents

Dr. Tafireyi. L. Chamboko

Mrs. Gloria. M.C Chamboko

Thank you eternally for everything.

“In the eyes of a child, parents are the superheroes who make everything better.”

Table of Contents

Abstract ..... **i**

Declaration ..... **iii**

Acknowledgements ..... **iv**

Dedication ..... **v**

Table of Contents ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xxi**

List of Supplementary Figures and Tables ..... **xxii**

List of Abbreviations ..... **xxvi**

List of Amino Acids ..... **xxix**

Research Outputs ..... **xxx**

Chapter 1 ..... **1**

Literature Review **1**

1.1 Introduction 2

1.2 Pharmacogenes and Variation Effects 4

1.2.1 Drug transporters 4

1.2.2 Drug targets 5

1.2.3 Drug metabolizing enzymes 6

1.3 Human Cytochrome P450 Enzymes 7

1.3.1 Nomenclature of human CYP enzymes 8

1.3.2 Sequence, structure and mechanism of action 10

1.3.3 Classes of human CYP enzymes and their redox partners 18

1.3.4 Existing structural information on human CYP enzymes within CYP families 1, 2 and 3 19

1.3.5 Genetic variations in human CYP enzymes with a particular focus on CYP2D6 polymorphisms 25
1.4 Problem Statement and Motivation 32
1.5 Aims and Objectives 33
Chapter 2 35
Characterization of Human Cytochrome P450 Enzymes in Families 1, 2 and 3 with respect to CYP2D6 and its variants 35
2.1 Introduction 36
2.1.1 Sequence and structure retrieval 36
2.1.2 Multiple sequence alignment 38
2.1.3 Motif discovery 40
2.1.4 Phylogenetic analysis 41
2.2 Methodology 43
2.2.1 Sequence retrieval 43
2.2.2 Multiple sequence alignment 43
2.2.3 Phylogenetic analysis 44
2.2.4 Motif discovery and mapping 45
2.2.5 CPY2D6 variant selection and analysis with motifs identified 46
2.3 Results and Discussions 47
2.3.1 Multiple sequence analysis 47
2.3.2 Phylogenetic analysis 52
2.3.3 Motif analysis 54
2.3.4 Motif and variation analysis 60
2.4 Conclusion 66
Chapter 3 68

Unraveling Mechanisms of Variant-induced Reduced Activity in Cytochrome P450 2D6 Alleles 68

3.1 Introduction 69

3.1.1 CYP2D6 genetic and structural information 69

3.1.2 Polymorphism in CYP2D6 72

3.1.3 PharmVar and PharmGKB 73

3.1.4 Metal ion parameterization and Molecular Dynamic simulations 76

3.1.5 Dynamic Residue Network Analysis 79

3.2 Methodology 83

3.2.1 CPY2D6 allele selection 83

3.2.2 Retrieval of reference and allele CYP2D6 structures 83

3.2.3 Parameterization and preparation of reference and allele CYP2D6 structures for MD simulations 84

3.2.4 Molecular Dynamic simulations and Trajectory analysis 85

3.2.5 Heme region analysis 86

3.2.6 Tunnel analysis 87

3.2.7 Dynamic Residue Network Analysis 88

3.2.8 Delta calculations 89

3.3 Results and Discussions 91

3.3.1 Identification of CYP2D6 alleles with known clinical function, 3D structure generation, and MD simulations 91

3.3.2 Instability and conformational changes in CYP2D6 alleles 93

3.3.3 Assessing changes in secondary structure, flexibility and hydrogen bonding in CYP2D6 alleles with known clinical activity 97

3.3.4 Heme region analysis 114

3.3.5 Substrate channel analysis 117

3.3.6 DRN analysis 123
3.4 Conclusion 133
Chapter 4 139
Predicting Reduced Functional Consequence of Afrocentric/ Sub-Saharan African CYP2D6 Alleles: Insights from Mechanistic Findings 139
4.1 Introduction 140
4.2 Methodology 142
4.2.1 Afrocentric CPY2D6 allele selection 142
4.3 Results and Discussions 143
4.3.1 Selection of Afrocentric alleles with unknown clinical function, 3D structure generation, and MD simulations 144
4.3.2 RMSD and Rg calculations to assess stability, conformational variation and compactness of Afrocentric alleles 146
4.3.3 Evaluation of changes in secondary structure, flexibility and hydrogen bonding in functionally important regions of novel Afrocentric CYP2D6 alleles 148
4.3.4 Heme region analysis in Afrocentric CYP2D6 alleles 166
4.3.5 Substrate channel analysis 169
4.3.6 DRN analysis 172
4.3.7 Potential functional consequence of CYP2D6 Afrocentric Alleles 177
4.4 Conclusion 181
Chapter 5 184
Concluding Remarks and Future Perspectives 184
5.1 Overall Conclusion 185
5.2 Future work 190
Supplementary data 191
Supplementary Figures 191

Supplementary Tables 207

References 229

List of Figures

Figure 1-1: Differing drug responses in patients. Image depicting patients who were prescribed the same clinical drug and dosage using a "one-size-fits-all" approach, yet experienced varying responses to the treatment. Patient responses can be categorized as a desired response (green), partial response (orange), adverse drug reactions (red), or no response (grey). Modified from Gill *et al*., 2021 [13]. 3

Figure 1-2: Structure of nomenclature used to name different human CYP enzymes. The nomenclature used to designate human CYP enzymes follows the structure "CYPxyz" and for alleles, the format is “CYPxyz*ij”. Used CYP2D6 as an example. 9

Figure 1-3: (A) General structure of CYP enzymes. CYP2C8 (PDB ID: 2NNJ) is used to map the major helices denoted A-L in different colours. Substrate entrance site is highlighted in red and an arrow pointing towards the entrance is in black. (B) Conserved catalytically important residues mapped onto the CYP structure, and the cysteine that acts as a proximal axial thiolate ligand for the heme iron is highlighted. Threonine in the conserved motif (A/G)XXT is coloured in yellow, while glutamic acid and arginine in the EXXR motif are coloured in warm pink. Phenylalanine, glycine and cysteine in the conserved motif FXXGXXXCXG are all coloured in purple. (C) CYP2C8 enzyme residue numbers are denoted, as an example, for the helices observed in CYP enzymes. Adapted from Chamboko et al., 2023 [3]. 13

Figure 1-4: Mechanism of CYP enzymes in a seven-step catalytic cycle. Enzyme open, semi-open and closed conformations at each stage denoted by a grey shape. The substrate, hydrogens, electrons and oxygen are denoted by yellow, purple, green and dark blue circles respectively. Water is shown by a blue drop. Adapted from Chamboko *et al*., 2023 [3]. 16

Figure 1-5: Open and closed conformations of CYP2D6. Open conformation of CYP2D6 structure in cyan from PDB (3TBG) superimposed onto closed conformation structure in grey (3QM4). Direction of displacement of F’ and Pre-A regions to form the open structure is shown in dotted arrows. Adapted from Supplementary Data, Chamboko *et al*., 2023 [3]. 24

Figure 1-6: Map showing prevalent CYP2D6 alleles in the different biogeographical groups. CYP2D6 alleles (highlighted in red) and the location where they are highly frequent shown in the world map. There are nine biogeographical groups as described by PharmGKB and CPIC however the African American/Afro-Caribbean and Latino groups could not be shown as they encompass populations with considerable gene flow between different geographic regions due to post-colonization and post-Diaspora events. The biogeographical groups colour coded according to the key in the figure. Modified from PharmGKB [122, 123]. 28

Figure 2-1: Representation of all CYP enzymes from families 1, 2 and 3 studied in this chapter. Different families shown in different colours and different subfamilies within the families shown in a lighter version of the family colour..... 37

Figure 2-2: Multiple sequence alignment of CYP enzymes in families 1, 2 and 3. The multiple sequence alignment of CYP sequences was produced by T-COFFEE alignment program and results viewed in Jalview. The conserved regions were observed using the Jalview Clustal X Colour Scheme. The PPGPXPXP is highlighted by a yellow box. The (A/G)XXT region is denoted with a green box, while the EXXR region is shown in the purple box. The FXXGXXXCXG conserved domains is in a blue box. The red dots depict the most conserved and important catalytic residues in CYP enzymes. The unique regions (as identified by motif analysis) to CYP 1 family denoted by pink boxes, while the unique regions found in CYP 2 family in red boxes. CYP 3 family unique regions are marked by cyan boxes. Helix regions for the different helices within CYP enzymes shown by grey bars, and CYP 2C8 enzyme sequence was used to map the helix regions according to [58] findings. 51

Figure 2-3: Sequence analysis of CYP 1, 2 and 3 families. A) Phylogenetic tree produced using MEGAX and the maximum likelihood method was applied using the LG+G+I protein model with 100% gap deletion. The bootstrap values on branches are percentages of 1000 bootstraps on the branch nodes. B) All versus all pairwise sequence alignment results for CYP 1, 2, 3 family sequences presented as a heat map produced using Python and MATLAB. Conservation values in heat map is increasing from blue to red. 53

Figure 2-4: Motif analysis. A) Heat map showing the frequency of each significant motif in the CYP 1, 2 and 3 family sequences sampled. Matplotlib and in-house Python scripts were used to create heat map and conservation of motifs across all CYP sequences increased from blue to red. Numbering of motifs and their respective logos was based on the results generated by MEME. B) Motif logos showing amino acid conservation of each residue within each motif. 56

Figure 2-5: Motifs identified mapped to CYP structures. A) Conserved motifs mapped to CYP2D6 structure (PDB ID 3TDA) with highly conserved regions (AGXXT, PPGPXPXP, EXXR and the FXXGXXXCXG domains) highlighted by their motif logos and their location within the motif shown with a black bracket. B), C) and D) are unique motifs mapped to the representative CYP enzyme family structures. B) CYP 1 unique motifs 16, 19 and 31; C) CYP 2 unique motifs 13, 15, 16, 18, 19, 20, 21, 24 and 25; D) CYP 3 unique motifs 22, 23, 26, 27, 28, 29 and 30. PDB structures 2HI4, 4GQS and 1TQN are used respectively as representatives for CYP 1, 2 and 3 families. Motif 17 unique to CYP2 and CYP3 families could not be mapped to structures due to the structures not containing this region. 58

Figure 2-6: Motif and variation analysis. A) Motifs identified within CYP families 1, 2 and 3 sequence dataset. Matplotlib and in-house Python script were used to create heat map. A total of 23 sequences observed to have 31 significant motifs discovered. Conservation of motifs across all CYP sequences increased from blue to red. Motifs occurring in CYP2D6 (highlighted in green) shown by a black box. B) All conserved and unique motifs specific to CYP2D6 mapped to its structure and location of CYP2D6 allele variations within motifs shown in the structure as red spheres. 64

Figure 3-1: Location of the CYP2D6 gene on a human chromosome 22. Modified from Alvarado *et al*., 2021 [175]. 70

Figure 3-2: CYP2D6 structural characteristics. A) Structure of CYP2D6 with allele variations mapped and showing their positions in the different helices. Variations are shown as red spheres. Helices in CYP2D6 are denoted A-L and differently coloured in the structure. A colourmap showing the helix names and corresponding colours is at the bottom of the structures. β – sheets are coloured in black and the active site cavity is coloured in magenta. B) The active site region

magnified and showing a “right boot” shaped cavity and heme (coloured orange) with an iron center (coloured salmon). C) Location of SRSs in CYP2D6 structure coloured in deep olive and channels identified in literature to be present in CYP2D6 with and without membrane present labelled. The channel labels are at their entry sites in the structure. 71

Figure 3-3: A) RMSD vs time line plots for the whole 500 ns MD simulation of all alleles. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue. B) Different stages of the MD simulation (0-500ns, 50-500ns, 100-500ns, 150-500ns, 200-500ns) shown as RMSD violin plots, with the same colouring key as line plots. C) All vs all RMSD heatmaps for whole MD simulation (500 ns). The x and y axes show the frames at different time stamps and the degree of variation is from low to high (coloured from white, through yellow and red, to black). 95

Figure 3-4: CYP2D6 reference and allele Rg analysis. A) timeline plots for the whole 500 ns MD simulation. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue. B) Rg violin plots for different stages of references and allele MD simulations (0-500ns and 200-500ns). Violin plots coloured using the same colouring key used for line plots. 96

Figure 3-5: Conformational clustering analysis. A) REF1 (grey) and REF2 (light orange) median cluster structures superimposed. The B’ and F-G regions where deviations were observed in alleles, were highlighted in these structures. B) Median cluster structures for normal and reduced function alleles superimposed to REF1. Regions with changes in secondary structure or conformation as compared to REF1 were shown with red circles and arrows. Alleles *10, *14, *17, *29, *34 *39, *48, *50, *54, *55 with more deviating conformations as compared to REF1, or with secondary structure changes noted were highlighted in grey boxes. Cluster structures for the different alleles in B) are coloured differently to differentiate from REF1 structure and each other. 100

Figure 3-6: Secondary structure profiles of references and alleles with differing regions highlighted. A) Secondary structure preference per residue (for both CYP2D6 references and alleles) within the last 300 ns of the MD simulations. Bar graphs represent probability of each residue adopting a specific secondary structure in the last 300ns of each trajectory. The regions

where differences are noted within the secondary structure profiles relative to references are highlighted with black boxes, and a colour key for the different secondary structure assignments (Alpha-helix, beta-bridge, beta-ladder, 3-10 helix, pi helix, poly-proline II or kappa-helix, bend, hydrogen bonded turn, coil/loop) provided next to the profiles. B) Regions in alleles noted to be different as compared to references mapped to CYP2D6 REF1 structure and colours of regions are matched to areas highlighted at the bottom of the bar graphs. 105

Figure 3-7: A) Residues with top and bottom 3% ΔRMSF values compiled into a heatmap, and significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Values within top and bottom 3% ΔRMSF values and not significant are marked with a "-". Regions exhibiting significant changes in flexibility, which correlate with the observations from DSSP analysis, are highlighted in boxes and colour-coded to match the DSSP analysis. B) Residues with top and bottom 3% ΔRMSF values for each system mapped to reference and allele conformational cluster structures and shown as sticks. Residues with significant changes in ΔRMSF corresponding to changes noted in DSSP analysis marked with black circles. Variation positions for each allele shown in spheres and coloured in yellow. In A) and B) ΔRMSF values range from positive to negative. Residues with a loss in flexibility (becoming more rigid) within the systems are coloured in blue, and residues with gain in flexibility in systems are coloured in red. 108

Figure 3-8: Hydrogen bond analysis. A) Residue pairs with top and bottom 3% ΔH-bond values compiled into a heatmap, with significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Values within top and bottom 3% ΔH-bond values and not significant are marked with a "-". B) Significant changes in hydrogen bonding noted between residue pairs for alleles *10, *17, *27, *29, *45, *50, *54 and *55. Variations in each allele shown by yellow spheres. For both A) and B), ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive). 113

Figure 3-9: Analysis of heme region in CYP2D6 references and alleles. A) Line plots showing COM distances between the active site cavity and the heme B) Line plots displaying hydrogen bonds formed between heme and protein residues. In A) and B), REF1 is shown in grey, REF2 is

shown in orange, normal activity alleles are shown in green and decreased activity alleles in blue. C) Contact analysis of heme with surrounding residues ≤0.4 nm from it. The degree of contact of each residue with the heme is shown by the colour scale from 0-1. Last 300 ns trajectories used to analyze heme environments. 116

Figure 3-10: CYP2D6 tunnel analysis. A) Spatial representation of identified channels shown in CYP2D6 REF1. Channels are shown as van der Waals representations and named using nomenclature defined by Cojocaru *et al*., [61]. Channel 2b is coloured blue, channel 2c is coloured red, channel 2e is coloured green, channel 2f is coloured purple and solvent channel is coloured orange. Location of SRSs is also shown and coloured in teal. B) Heatmaps showing time evolution of channel bottleneck radius for channel 2b, 2f, 2e, 2c and the solvent channel. Snapshots were taken after every 2ns for the 500ns MD simulation run for all systems, and the total snapshots produced for analysis were 1001 including snapshot at 0ns. The colour map ranges from 0.9 Å to 1.5 Å or greater, with the lowest radii coloured in red and the highest radii coloured in green. Grey shows no channel with a bottleneck radius ≥0.9 Å was identified for that specific time in each instance. 122

Figure 3-11: REF_AV and Δ*BC* analysis. A) Residues with top 6% averaged *BC* values for REF_AV mapped to a reference structure, shown as spheres and coloured in red, and shown in a bar plot below the structure. B) Residues with top and bottom 3% Δ*BC* values compiled into a heatmap. Significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Top and bottom 3% ΔBC values that are not significant are marked with a "-". C) Reduced function alleles *10, *14, *17, *29, *50, *54 and *55, and normal function allele *39 with significantly changing averaged *BC* values for residues within the B' helix and/or the loop between β1-1 and β1-2 shown. REF1 and REF2 also shown with no changes in averaged *BC* values occurring in the regions marked in the reduced function alleles. Changes in communication within the B' helix shown with green labels and changes noted in the loop between β1-1 and β1-2 shown with light blue labels. Variation positions for alleles shown in spheres and coloured in yellow. For both B) and C), the values range from positive to negative. Residues with a loss in averaged *BC* within the alleles are coloured in blue, and residues with gain in averaged *BC* in alleles are coloured in red. 126

Figure 3-12: REF_AV and Δ*EC* analysis. A) Residues with top 6% averaged *EC* values for REF_AV were mapped to a reference structure, shown as spheres and coloured in red, and shown in a bar plot beside the structure. B) Residues with top and bottom 3% Δ*EC* values compiled into a heatmap. Significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Top and bottom 3% Δ*EC* values that are not significant are marked with a "-". Regions with larger changes in averaged *EC* within reduced function alleles highlighted with black brackets. C) Reduced function alleles *10, *14, *17, and *55 with more significantly changing averaged *EC* shown. REF1 and REF2 also shown for comparison. Variation positions for alleles shown in spheres and coloured in yellow. For both B) and C), the values range from positive to negative. Residues with a loss in averaged *EC* within the alleles are coloured in blue, and residues with gain in averaged *EC* in alleles are coloured in red..... 130

Figure 3-13: REF_AV and Δ*CC* analysis. A) Residues with top 6% averaged *CC* values for REF_AV mapped to a reference structure, shown as spheres and coloured in red, and shown in a bar plot below the structure. Residues with top and bottom 3% Δ*CC* values represented in a heatmap (B) and mapped onto their corresponding structure obtained from conformational clustering program and shown as spheres (C). In B), significant values (outliers not within 3SD of the mean of the reference data) are marked with an "X" in the heatmap. Top and bottom 3% ΔBC values that are not significant are marked with a "-". In B) and C) the values range from positive to negative. Residues with a loss in averaged *CC* within the alleles are coloured in blue, and residues with gain in averaged *CC* in alleles are coloured in red. REF1 and REF2 also shown for comparison and variation positions for alleles shown in spheres and coloured in yellow. 132

Figure 4-1: Geographic origin of Afrocentric alleles used in this study..... 141

Figure 4-2: Summary of variant-induced mechanisms identified to potentially cause reduced activity in CYP2D6 alleles represented in a flow chart. 143

Figure 4-3: Variations in novel CYP2D6 Afrocentric alleles. A) Table showing all Afrocentric alleles studied in this article with their respective backbone allele and variations. Variation written in red and underlined signify new variation added to backbone allele to form Afrocentric allele. B)

Distribution of variations in a CYP2D6 reference structure with different helices denoted by the different colours with a key at the bottom. 145

Figure 4-4: CYP2D6 reference and Afrocentric alleles' RMSD analysis. A) RMSD vs time-line plots for the whole MD simulation (500ns) for references and alleles. B) All vs all RMSD heatmaps for whole MD simulations. The x and y axes show the frames at different time stamps and the degree of variation from low to high, are coloured from white, through yellow and red, to black. C) RMSD violin plots for stages of MD simulation (0-500ns/ whole trajectory, 50-500ns, 100-500ns, 150-500ns, 200-500ns). For A) and C), REF1 is shown in blue, REF2 is shown in orange and CYP2D6 Afrocentric alleles are shown in green. 147

Figure 4-5: Rg analysis. A) Rg line plots for the whole 500 ns trajectory for references and Afrocentric alleles. B) Violin plots for the whole MD simulation (500 ns) for references and Afrocentric alleles. In both A) and B), REF1 is shown in blue, REF2 is shown in orange, and CYP2D6 Afrocentric alleles are shown in green. 148

Figure 4-6: Analysis of conformational cluster structures for each reference and Afrocentric allele obtained through geometric clustering algorithm. A) Cluster structures for REF1 and REF2 superimposed and coloured grey and light orange respectively. B) All cluster structures for references and alleles superimposed, with deviating regions that were noted in alleles as compared to references highlighted by a red dashed circle and labels. Substrate entrance region shown by a black arrow. C) Each Afrocentric allele superimposed to REF1. Cluster structures for different alleles are shown in different colours and areas with varying conformations and secondary structures are highlighted with red circles and arrows..... 150

Figure 4-7: DSSP analysis. The bar graphs illustrate the likelihood of each residue (in references or Afrocentric alleles) adopting a specific secondary structure during the final 300 ns of each MD simulation. Areas where secondary structure assignments differ from the references are highlighted with red boxes and the regions are labelled and colour coded according to regions identified in DSSP analysis within chapter 3. A colour key for various secondary structures (including alpha-helix, beta-bridge, beta-ladder, 3-10 helix, pi helix, poly-proline II or kappa-helix, bend, hydrogen-bonded turn, and coil/loop) is provided in the figure. 152

Figure 4-8: ΔRMSF A) Heatmap showing residues with top and bottom 3% ΔRMSF values. Significant values (outliers not within 3SD of the mean of the reference data) are marked with an "X" in the heatmap and values within top and bottom 3% ΔRMSF values and not significant are marked with a "-". Areas with changes in flexibility, which correlate with DSSP analysis are highlighted and colour-coded to match the DSSP analysis. B) Residues with top and bottom 3% ΔRMSF values for each system are mapped to reference and allele cluster structures and shown as sticks, and the residues with significant changes in ΔRMSF corresponding to changes noted in DSSP analysis are marked with black circles. Yellow spheres show the variation positions in each allele. In both A) and B) ΔRMSF values range from positive to negative. Residues with a loss in flexibility (becoming more rigid) within the systems are coloured in blue, and residues with gain in flexibility in systems are coloured in red. 158

Figure 4-9: Hydrogen bond analysis. A) Residue pairs with top and bottom 3% ΔH-bond values shown in a heatmap, with significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. The values within top and bottom 3% ΔH-bond values and not significant are marked with a "-". B) Significant changes in hydrogen bonding noted between residue pairs for Afrocentric alleles. Yellow spheres show variation positions within Afrocentric alleles. For both A) and B), ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive). 164

Figure 4-10: CYP2D6 Afrocentric allele heme analysis. Analysis of heme environment in CYP2D6 references and Afrocentric alleles for the last 300 ns of the MD simulation. A) Line plots showing center of mass (COM) distances between the heme and the active site cavity for references and alleles. B) Stick representation of heme position in CYP2D6 and residues anchoring it in place. Heme is coloured in green and residues coloured in deep teal. C) Line plots showing hydrogen bond formation between heme and surrounding residues. For both A) and C), REF1 is shown in blue, REF2 is shown in orange and CYP2D6 Afrocentric alleles are shown in green. D) Heme contact/non-covalent interaction analysis with residues ≤4 Å from the heme. The probability of contact of each residue with the heme is shown by the colour scale from 0-100%. 168

Figure 4-11: Channel analysis of CYP2D6 Afrocentric alleles. The heatmaps illustrate time evolution of the channel bottleneck radius for channels 2b, 2f, 2e, 2c, and the solvent channel, for the 500 ns MD simulation. A total of 1001 snapshots were used in the analysis, including the snapshot at 0 ns. A colour map ranging from 0.9 Å to 1.5 Å (and above) is shown below heatmaps, with the smallest radii represented in red and the largest in green. Grey indicates that no channel with a bottleneck radius of ≥0.9 Å was identified at that specific time point..... 170

Figure 4-12: Δ*BC* analysis of Afrocentric alleles. A) Residues with top and bottom 3% Δ*BC* values shown in a heatmap with significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap and top and bottom 3% Δ*BC* values that are not significant marked with a "-". B) Alleles *152, *153 and *159 with significantly changing averaged *BC* values for residues within the B' helix shown (labelled in green labels). REF1 and REF2 also shown with no significant changes in averaged *BC* values in this region. C) Allele *157 with residue L399 shown in a red label with loss in averaged *BC*, while there was a pathway of gain in averaged *BC* formed by residues V370, P371, S400, L403 and R365 from the heme to the K helix. For B) and C), variation positions for alleles shown in spheres and coloured in yellow. For all (A, B and C) the *ΔBC* values range from positive to negative, with positive indicating a gain in averaged *BC* within the alleles (shown in red colour) and negative indicating a loss in averaged *BC* in alleles (shown in blue colour). 175

List of Tables

Table 1-1: 3D structures of enzymes from CYP 1, 2 and 3 enzyme families that are available on PDB with their respective resolutions and co-crystalized ligands as of November 2022. Adapted from Supplementary Data, Chamboko *et al*., 2023 [3]. 20

Table 1-2: CYP2D6 alleles studied and their variations and population frequencies across the nine biogeographical groups as described by CPIC and PharmGKB [122, 123]. Highest frequencies for each allele are highlighted in bold..... 29

Table 2-1: Human Cytochrome P450 enzyme families 1, 2 and 3 UniProt sequence attributes.. 48

Table 2-2: Regular expressions and E-values of significant motifs identified in CYP families 1, 2 and 3 and used for further analysis. 54

Table 2-3: Conserved and unique motifs identified to be occurring in CYP2D6, with their regular expressions, web logos, motif residue range and all variations occurring in each motif. 61

Table 3-1: CYP2D6 alleles studied with their respective functional consequence, variations, combinations and location of variations in the structure. Variations added to alleles to form new alleles are highlighted in bold. 92

Table 3-2: Prominent substrate channels identified within CYP2D6 references and alleles. 118

Table 3-3: Summary of mechanisms potentially leading to reduced CYP2D6 activity due to genetic variations. 138

Table 4-1: Summary of identified mechanisms potentially leading to reduced activity in the Afrocentric alleles. 178

List of Supplementary Figures and Tables

Figure S 1: MSA of CYP family 1, 2 and 3. A, B and C showing MSA of CYP families 1, 2 and 3 respectively. Alignments were produced by T-COFFEE alignment program and results viewed in Jalview. 191

Figure S 2: Bootstrap tree for the phylogenetic tree chosen. Bootstrap tree was produced using MEGAX. 192

Figure S 3: All motifs identified within CYP families 1, 2 and 3 represented in a heatmap. Conservation of motifs across all CYP sequences increased from blue to red. 193

Figure S 4: RMSD line plots of the whole trajectory (500 ns) (A) and the last 300 ns (B) selected and used for subsequent analysis of references, normal activity alleles and reduced activity alleles. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue. 194

Figure S 5: RMSF line plots and heatmaps of the last 300 ns of reference and allele (normal and reduced) MD simulations. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue. 195

Figure S 6: Effects of variation R296C on hydrogen bonding between residue position 296 and surrounding residues. In alleles *2, *14, *17, *29, *34 and *55, there was a significant loss in hydrogen bonding noted between residue 296 and residues 252 and 293. A significant gain in hydrogen bonding was then noted between residue 293 and 115 only in alleles *2 and *14. Variations in each allele shown by yellow spheres. ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive). 196

Figure S 7: Calculated average bottleneck radii for each channel in each system represented in bar graphs. Bars within the bar graphs are coloured according to the following: REF1 in grey, REF2 in orange, normal activity alleles in green and reduced activity alleles in blue. Red line indicating the cut-off bottleneck radii used to determine whether a channel is open or closed (≥ 1.4 Å)... 197

Figure S 8: Δ*BC* analysis. Residues with top and bottom 3% Δ*BC* values mapped to reference and allele conformational clustering structures and shown as spheres. Normal alleles *2, *27, *33, *34, *45, *48 and *53, and reduced function allele *49 with no significantly changing averaged *BC* values for residues within the B' helix and/or the loop between β1-1 and β1-2 shown. Δ*BC* values range from positive to negative. Residues with a loss in averaged *BC* within the alleles are coloured in blue, and residues with gain in averaged *BC* in alleles are coloured in red. Variation positions for alleles shown in spheres and coloured in yellow..... 198

Figure S 9: Δ*EC* analysis. Residues with top and bottom 3% Δ*EC* values mapped to reference and allele conformational clustering structures and shown as spheres. Δ*EC* values range from positive to negative. Residues with a loss in averaged *EC* within the alleles are coloured in blue, and residues with gain in averaged *EC* in alleles are coloured in red. Variation positions for alleles shown in spheres and coloured in yellow. 199

Figure S 10: A) RMSD line plots of the whole trajectory (500 ns). B) The last 300 ns selected and used for subsequent analysis of references, and Afrocentric alleles. REF1 line plot is in blue, REF2 line plot is in orange and Afrocentric allele line plots are in green. 200

Figure S 11: RMSF line plots and heatmaps of the last 300 ns of reference and Afrocentric alleles. REF1 line plot is in blue, REF2 line plot is in orange and Afrocentric allele line plots are in green. 201

Figure S 12: Hydrogen bond analysis for Afrocentric alleles showing effects of variation R296C on hydrogen bonding between residue position 296 and surrounding residues. In alleles *154, *155, *157, *159, *160, *162, *163, there was a significant loss in hydrogen bonding noted between residue 296 and residues 252 and 293. Variations in each allele shown by yellow spheres. ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive). 202

Figure S 13: Average bottleneck radii for each channel in each system calculated and represented in bar graphs. Bars are coloured as follows: REF1 is in blue, REF2 in orange and Afrocentric

alleles in green. Dark green line indicates the cut-off bottleneck radii used to determine whether a channel is open or closed (≥ 1.4 Å). 203

Figure S 14: Δ*BC* analysis of Afrocentric alleles. Residues with top and bottom 3% Δ*BC* values mapped to reference and Afrocentric allele conformational clustering structures and shown as spheres. Alleles *149, *154, *155, *157, *160, *162 and *163 with no significantly changing averaged *BC* values for residues within the B' helix and/or the loop between β1-1 and β1-2 shown. In these alleles, Δ*BC* values range from positive to negative. Residues with a loss in averaged *BC* within the alleles are coloured in blue, and residues with gain in averaged *BC* in alleles are coloured in red. Variation positions for Afrocentric alleles shown in spheres and coloured in yellow. 204

Figure S 15: Δ*EC* analysis of Afrocentric alleles. Residues with top and bottom 3% Δ*EC* values shown in a heatmap (A) and mapped to their corresponding conformational cluster structures (B). In the heatmap, significant values (outliers not within 3SD of the mean of the reference data) are marked with an "X" and top and bottom 3% Δ*EC* values that are not significant are marked with a "-". For B), variation positions for alleles are shown in spheres and coloured in yellow. For both A) and B), the *ΔEC* values range from positive to negative, with positive indicating a gain in averaged *EC* within the alleles (shown in red colour) and negative indicating a loss in averaged *EC* in alleles (shown in blue colour). 205

Figure S 16: Δ*CC* analysis of Afrocentric alleles. Residues with top and bottom 3% Δ*CC* values shown in a heatmap (A) and mapped to each systems' corresponding conformational cluster structures (B). The heatmap shows significant top and bottom 3% Δ*CC* values (outliers not within 3SD of the mean of the reference data) with an "X" and values that are not significant with a "-". For B), variation positions for alleles are shown in spheres and coloured in yellow. For both A) and B), the *ΔCC* values range from positive to negative, with positive showing a gain in averaged *CC* within the alleles (in red) and negative showing a loss in averaged *CC* in alleles (in blue). 206

Table S 1: Cytochrome P450 enzymes from families 1, 2 and 3 associated with metabolism of antimalarial and antituberculosis drugs with their alleles and amino acid variation positions. Adapted from Supplementary data, Chamboko *et al.*, 2023 [3]. 207

Table S 2: Human Cytochrome P450 enzyme families 1, 2 and 3 UniProt sequences. 217

Table S 3: Best-fit substitution models for the construction of nine phylogenetic trees selected based on the best (lowest) BIC score. 222

Table S 4: The significant motifs generated with MEME program, their positions in CYP enzymes within families 1, 2 and 3, and their respective E-values. 223

Table S 5: Conformational clustering information for references and function assigned alleles for the last 300 ns of each trajectory..... 225

Table S 6: Frequency in snapshots of tunnel presence in CYP2D6 references and function assigned alleles. 226

Table S 7: Conformational clustering attributes of references and Afrocentric alleles for the last 300 ns of each trajectory. 227

Table S 8: Snapshot frequencies of tunnel presence in CYP2D6 references and Afrocentric alleles. 228

List of Abbreviations

3D	Three-dimensional
ABCB1	ATP binding cassette subfamily B member 1
ACPYPE	AnteChamber PYthon Parser interface
ACT	Artemisinin-based combination therapy
ADRs	Adverse drug reactions
ADRs	Adverse drug reactions
AIC	Akaike information criterion
AMBER	Assisted Model Building with Energy Refinement
ATP	Adenosine tri-phosphate
BC	Betweenness centrality
BIC	Bayesian information criterion
CC	Closeness centrality
CFTR	Cystic fibrosis transmembrane conductance regulator
CHARMM	Chemistry at HARvard Macromolecular Mechanics
COM	Center of mass
CPIC	Clinical Pharmacogenetics Implementation Consortium
CYP	Cytochrome P450
CYP2D6	Cytochrome P450 2D6
DRN	Dynamic residue network analysis
DSSP	Dictionary of Secondary Structure in Proteins
EC	Eigenvector centrality
EM	Extensive metabolizers

FAD	Flavin adenine dinucleotide
FMN	Flavin mononucleotide
GROMACS	GROningen MAchine for Chemical Simulations
GROMOS	GROningen MOlecular Simulation
HER2	Human epidermal growth factor receptor 2
HS	High-spin
IM	Intermediate metabolizers
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
LOE	Level of evidence
LS	Low-spin
MCPB	Metal centre parameter builder
MD	Molecular dynamics
MDR-TB	Multidrug-resistant tuberculosis
MEGA	Molecular Evolutionary Genetic Analysis
MEME	Multiple Expectation Maximization for Motif Elicitation
MM	Molecular mechanics
MSA	Multiple Sequence Alignment
NADPH	Nicotinamide adenine dinucleotide phosphate
NAMD	Nanoscale Molecular Dynamics
NMR	Nuclear magnetic resonance
NIH	National Institutes of Health
OPLS	Optimized Potential for Liquid Simulations
P-gp	P-glycoprotein 1

PDB	Protein Data Bank
PharmGKB	Pharmacogenomics Knowledge Base
PharmVar	Pharmacogene Variation Consortium database
PGRN	Pharmacogenetics Research Network
PIR	Protein Information Resource
PM	Poor metabolizers
POR	NADPH-cytochrome P450 oxidoreductase
QM	Quantum mechanics
Rg	Ratio of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SNPs	Single nucleotide polymorphisms
SRS	Substrate recognition site
SSA	Sub-Saharan Africa
TB	Tuberculosis
TIP3P	Three-Point Transferable Intermolecular Potentials
UM	Ultrarapid metabolizers
VMD	Visual Molecular Dynamics

List of Amino Acids

Amino Acid	Three letter code	One letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Research Outputs

Publications

1. Chamboko, C.R., Veldman, W., Tata, R.B., Schoeberl, B. and Tastan Bishop, Ö., 2023. Human cytochrome P450 1, 2, 3 families as pharmacogenes with emphases on their antimalarial and antituberculosis drugs and prevalent African alleles. International Journal of Molecular Sciences, 24(4), p.3383. doi.org/10.3390/ijms24043383.

Conference Presentations

1. Chiratidzo R. Chamboko, Wayde Veldman, Taremekedzwa Allan Sanyanga and Özlem Tastan Bishop. Understanding the Impact of Variations in Decreased Function Cytochrome P450 2D6 Alleles. The Centre for High-Performance Computing (CHPC) National Conference (2024). Port Elizabeth, South Africa.

Awards

1. The Centre for High-Performance Computing (CHPC) National Conference PhD Student Poster Presentation Second Place Award (2024). Port Elizabeth, South Africa.

Scientific trips

1. International trip to A trip to Uppsala University in Sweden for the Inaugural meeting of the Cluster of Research Excellence: Addressing Global and African Challenges through Methods from Artificial Intelligence, Data Science and Theoretical and Computational Thinking (CoRE AI). This is one of the seventeen research clusters between the African

Research Universities Alliance (ARUA) and The Guild of European Research-Intensive Universities (The Guild) (November 2023).

2. Local trip to the South African Medical Research Council (SAMRC) in Cape Town for the Project Africa GRADIENT workshop (February 2024).

Chapter 1

Literature Review

Chapter overview

This is an introductory chapter that presents precision medicine and pharmacogenomics, then delves deeper into cytochrome P450 (CYP) enzymes which are a focal point in scientific research that will aid in the achievement of individualized therapy.

1.1 Introduction

An interesting short article entitled “New Era of Personalized Medicine: Targeting Drugs For Each Unique Genetic Profile” appeared in *The Wall Street Journal* on the 16th of April 1999, introducing a new concept, “personalized medicine” [1]. Since this publication, the idea of individualized therapy has increasingly become a topic of interest that has brought pharmaceutical companies and academic research institutions together. Their main goal was to one day create a world where medical decisions and interventions are tailored to the individual based on their genetic makeup, thereby personalizing pharmacotherapy.

For many years, doctors have used a "one-size-fits-all" approach when prescribing medicine to patients, and although this has presented positive effects in therapy of many diseases, not all patients have responded well to treatments. There are some negative effects associated with this concept, such as drug toxicity or adverse drug reactions (ADRs) that may lead to illness, or in severe cases, death (Figure 1-1). An observational study on individuals in sub-Saharan Africa (SSA) reported 16% deaths and 8.4% of hospital admissions to be due to ADRs [2]. To begin tackling this problem, personalized medicine, also known as precision medicine, is an innovative approach in the medical field, which involves scientists and doctors, that uses an individual’s genetic information to tailor their treatment, for the best and most favorable therapeutic outcome [3–5]. Many factors affect a patient’s response to medicine, such as their sex, age and the environment [3], but their genetic makeup is directly associated with the breakdown and clearance of medicines from the body [6–8]. The field by which this phenomenon is investigated is called

pharmacogenomics, a term that was created for the integration of two fields; pharmaceuticals and genetics [3, 9]. The genes within the human genome that code for proteins which partake in the clearance of drugs are called pharmacogenes [10], and variations within these pharmacogenes, such as insertions and deletions, single nucleotide polymorphisms (SNPs) or copy number variations, can result in irregularly functioning proteins, which may affect drug metabolism and efficacy [10–12].

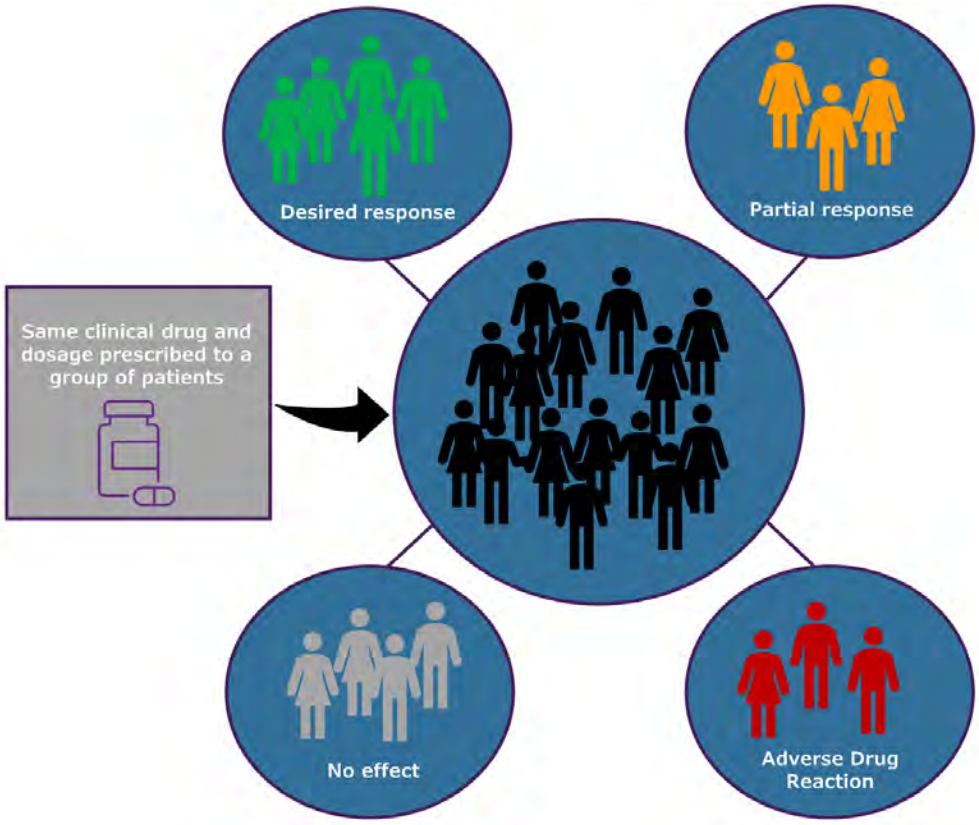


Figure 1-1: Differing drug responses in patients. Image depicting patients who were prescribed the same clinical drug and dosage using a "one-size-fits-all" approach, yet experienced varying responses to the treatment. Patient responses can be categorized as a desired response (green), partial response (orange), adverse drug reactions (red), or no response (grey). Modified from Gill *et al*., 2021 [13].

1.2 Pharmacogenes and Variation Effects

Pharmacogenomics research encompasses many pharmacogenes which can be divided based on their function, into three main classes: **(1)** drug metabolizers [14] (e.g., phase 1 metabolizers such as cytochrome P450 enzymes and phase 2 metabolizers such as thiopurine methyltransferase (TPMT) [15, 16]); **(2)** drug transporters [17] (e.g., influx transporters such as SLCO1B1, efflux transporters such as ATP binding cassette subfamily B member 1 (ABCB1), and P-glycoprotein 1 (P-gp) [18–21]); and **(3)** drug targets [22] (e.g., cystic fibrosis transmembrane conductance regulator (CFTR) and human epidermal growth factor receptor 2 (HER2) [23, 24]). All three types of pharmacogenes have separate roles they play in ensuring therapeutic effects of drugs, however, variations within these genes can potentially result in abnormal functioning of resultant proteins which can change how we as humans react to drugs.

1.2.1 Drug transporters

Drug transporters are proteins located on cell surfaces, that facilitate transportation of drugs/xenobiotics through the cell membrane [17]. These proteins are important for the absorption, distribution, and excretion of drugs, and can be categorized into two major classes: influx and efflux transporters [17]. Influx transporters control the movement of substrates into cells, while efflux transporters aid in the movement of substrates out of cells [17, 25]. Drug transporters are specific, i.e., there are different substrates for each transporter, thus their distribution throughout a human body can determine which chemicals may reach certain areas within the body, and the

concentrations of chemicals at each area [26]. The introduction of variations within drug transporters can have detrimental effects. It could lead to increased or decreased transporter function, which would essentially result in unwanted higher or lower concentrations of chemicals at drug sites of action, thereby increasing risk of drug toxicities or hindering full efficacy of drugs [6]. Because of the variations in this type of pharmacogene, differences in drug dosages would be required to accommodate for the difference in transporter activity, to achieve the desired therapeutic effect [6].

1.2.2 Drug targets

The second type of pharmacogenes, drug targets, can be enzymes or proteins within a vital pathway, that a drug is designed to affect, to deliver a desired therapy [22]. The three main types of drug targets are described as either signal transduction cascades or downstream proteins, direct drug protein targets, or proteins in pathogenic microbes causing disease [22]. The drug target types found in humans are the focus of pharmacogenomic research. Variations occurring in these pharmacogenes can alter the amount of drug target by either increasing or decreasing them, thus affecting the efficacy of a drug [27]. An example is the use of the drug enalapril in the treatment of high blood pressure, heart failure and diabetic kidney disease. Its inhibition of a normal angiotensin-converting enzyme (ACE) produces a desired therapeutic effect, however, variations occurring within the ACE drug target could alter its availability, and possibly cause drug toxicities or adverse drug reactions [22]. Personalized medication to cater for the high or low amounts of ACE drug target would be required, for the right therapeutic effect to be displayed.

1.2.3 Drug metabolizing enzymes

Drug metabolizing enzymes are the third class of pharmacogenes that have been of interest in pharmacogenomics research for many years [6, 14, 28]. They are mainly responsible for the biotransformation of xenobiotics into more polar and water-soluble compounds that can ultimately be excreted from the body [14]. Metabolism to clear xenobiotics occurs in two main steps: phase I and phase II. Phase I metabolism involves altering a drugs' chemical structure through reactions such as oxidation, reduction, cyclization or hydrolysis, to form an activated xenobiotic metabolite that becomes a substrate for phase II metabolism [29]. The main enzymes partaking in phase I reactions are CYP enzymes, known to be part of the liver microsomal mixed function oxidase system [14, 29]. Phase II metabolism, catalyzed by a large group of transferases such as glutathione S-transferases, methyltransferases, sulfotransferases, N-acetyltransferases and UDP-glucuronosyltransferases, involves conjugation reactions where the activated xenobiotic metabolite with either a nucleophilic or electrophilic group, is coupled with another molecule in mechanisms such as glutathione conjugation, methylation, sulphation, acetylation and glucuronidation [29]. These reactions form more polar, physiologically inactive and water-soluble products that can be excreted from the body easily through bile or urine [29, 30]. Variations introduced to pharmacogenes that code for enzymes within phase I and phase II metabolism can increase or decrease enzyme function, which will in turn alter the chemical processes that clear drugs from the human body, thereby subjecting patients to either drug toxicities from lack of clearance of drug compounds, or low efficacy of drugs [14, 28].

Drug metabolizing enzymes, therefore, are important determinants of drug action, and research encompassing variations in these proteins is of great interest and will directly aid in the goal of one day having personalized medication for all. Here, we delve deeper into understanding the drug metabolizing enzymes within the cytochrome P450 enzyme system, and their related variations.

1.3 Human Cytochrome P450 Enzymes

The cytochrome P450 (CYP) enzyme superfamily is composed of heme-thiolate proteins, better known as CYP enzymes that are present in all kingdoms of life and partake in many different processes throughout the different kingdoms [31–34]. In humans, CYPs primarily function as monooxygenases, oxidizing fatty acids, steroids, and xenobiotic [35, 36], and can be found mainly in the mitochondria and endoplasmic reticulum of hepatic cells [3]. They do however exist in other parts of the human body, including the lungs, small intestine, kidneys and placenta [3]. The human genome has been discovered to carry about 57 CYP genes, with each CYP enzyme being a product of a distinct CYP gene [3, 37]. These CYP enzymes are classified into 18 families, with the CYP2, CYP3, and CYP4 families having more genes than the other 15 families [3, 37]. To aid in organization and classification of these very similar enzymes based on their genetic similarities and function [38, 39], a nomenclature was required. In 1962, the term “cytochrome P450” was first introduced as a provisional name for a coloured substance within cells that, when reduced and bound to carbon monoxide, exhibited an unusual absorption peak at 450 nm [38]. Following the

identification of this substance as an enzyme, this name was retained, and the nomenclature for these enzymes has evolved and become more specific [38, 39].

1.3.1 Nomenclature of human CYP enzymes

With the increasing number of CYPs isolated in different laboratories over the years and the excess different names becoming more and more confusing, especially for those outside the field, a standardized nomenclature for CYP enzymes was finally established in the year 1987 [39]. This newly discovered nomenclature at the time was built with only 65 P450 genes, however the final nomenclature was based on 221 P450 genes [40]. Before this naming system was implemented, CYP enzymes were grouped according to their amino acid sequence identities. If they shared ≤36% sequence identity, they were classified within the same family, and if they shared ≥70% sequence identity, they were grouped into the same subfamily [41]. However, this classification system and nomenclature were revised as more data on CYP enzymes became available, since sequence identity values could change with the addition of new data [39–42]. Currently, sequence identity and phylogenetic tree clustering is used for classification [3, 43, 44], and the nomenclature used to name CYP enzymes in the present day is seen in Figure 1-2.

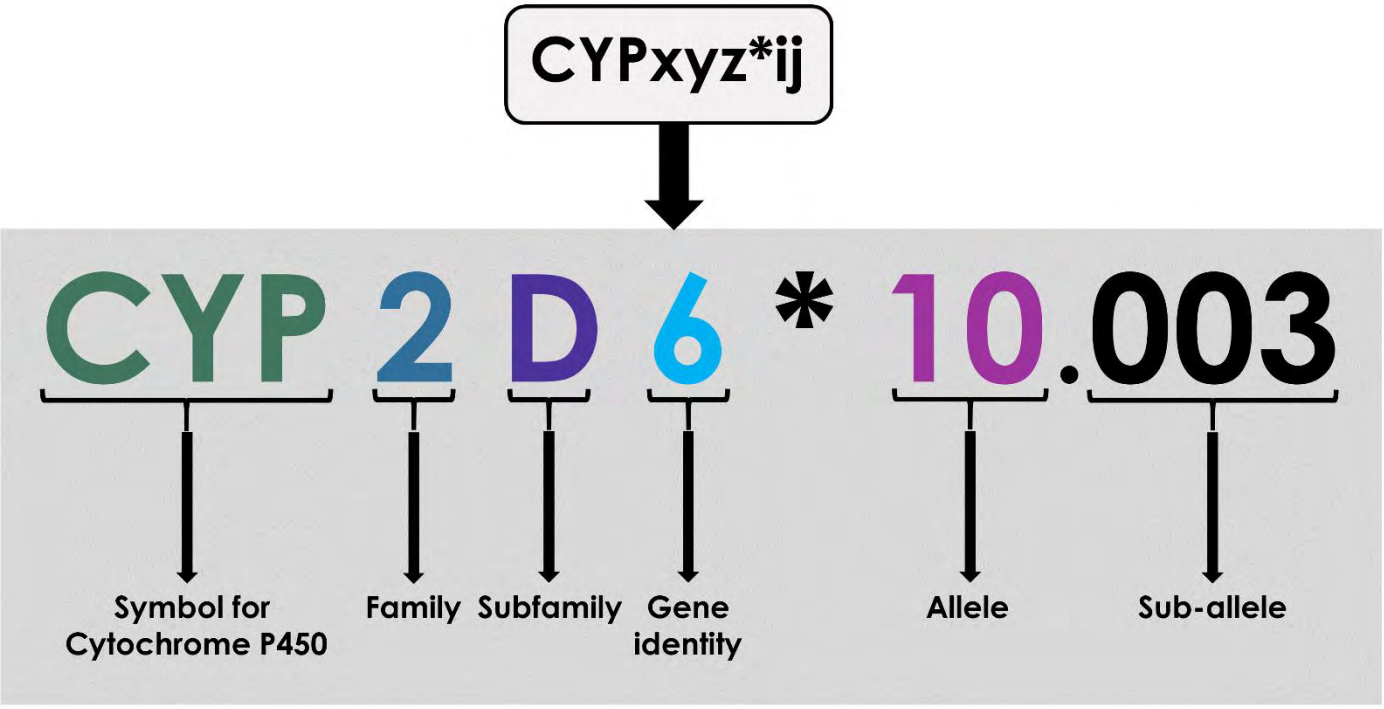


Figure 1-2: Structure of nomenclature used to name different human CYP enzymes**.** The nomenclature used to designate human CYP enzymes follows the structure "CYPxyz" and for alleles, the format is “CYPxyz*ij”. Used CYP2D6 as an example.

As shown in Figure 1-2, the nomenclature for CYP enzymes is structured as "CYPxyz," where "CYP" denotes Cytochrome P450, "x" represents the family (Arabic numeral), "y" corresponds to the subfamily (capital letter if multiple subfamilies exist), and "z" refers to the specific gene name (also in Arabic numerals) [3, 38, 40, 41, 45]. In specific instances, if the gene is a pseudogene, the letter "P" can be appended to the end of this nomenclature [40].

Over time, it became apparent that variations exist within the CYP genes, and thus an allele naming system was required to better categorize these genetic variants. This included the already established nomenclature for the gene level (“CYPxyz”), however, for each CYP allele, an asterisk is added after the gene name followed by an allele number in Arabic numerals ("i") [46, 47]. If applicable, a capital letter may be added to indicate silent mutations within the gene ("j"), forming sub-alleles [46, 47]. Overall, CYP alleles follow the nomenclature structure “CYPxyz*ij” (Figure 1-2). As seen on the Pharmacogene Variation Consortium database (PharmVar) however, updates to the nomenclature for sub-alleles (alleles with silent mutations) have been implemented, such that the “j” represents three decimal point digits after the allele name [48]. Sub-allele names that use the older naming system are listed under "legacy labels" on PharmVar, while the updated nomenclature appears under "allele name" [48, 49].

1.3.2 Sequence, structure and mechanism of action

1.3.2.1 Sequences of human CYP enzymes

Human CYP enzymes have been known to have highly similar structures and function, however, their amino acid sequences can exhibit considerable variability. Each human CYP enzyme has a sequence length of about 400-500 amino acids, and according to the updated nomenclature and classification system for enzymes within this superfamily, sequences sharing about 40% sequence identity are classified within the same family, while those with more than 55% sequence identity are grouped into the same subfamily [3, 50]. There are four highly conserved short regions documented to be found in CYP enzymes [51, 52]. The proline-rich

PPGPXPXP region is the first conserved region located near the N-terminal of CYP enzymes. It is crucial for the correct folding of CYP enzymes and for their anchoring to the membrane. [52–55]. The second conserved region is the (A/G)XXXT motif located within helix I [51, 56], critical for oxygen binding and activation, and containing a highly conserved threonine essential for the third step of the CYP enzyme catalytic cycle [51]. The EXXR motif is the third most conserved region in CYP enzymes, which features conserved glutamic acid and arginine residues that establish salt-bridge interactions crucial for the enzyme's final tertiary structure [51]. Finally, the fourth most conserved region in the CYP enzyme superfamily is the FXXGXXXCXG motif [57]. This heme-binding domain includes phenylalanine, glycine, and cysteine—three highly conserved residues within the CYP superfamily, with cysteine serving as the axial ligand to the heme [51, 57].

1.3.2.2 Structure of human CYP enzymes

CYP enzyme structures have been documented to be highly similar throughout the superfamily; however, eukaryotic CYPs are membrane proteins, in contrast to prokaryotic CYPs [3]. They are specifically bitopic membrane proteins, meaning they span the lipid bilayer once with an N-terminal helical anchor traversing the membrane bilayer. This transmembrane portion of CYP enzymes is then attached to a large globular domain that is located on the matrix side of the inner mitochondrial membrane or the cytoplasmic side of the endoplasmic reticulum, which contains the C-terminal end of the enzyme [3]. The CYP globular domain exhibits a highly conserved fold structure across all human CYP enzymes [56, 58]. As shown in Figure 1-3A, the enzymes consist of 14 helices and loops denoted A-L (A, B, B′, C, D, E, F, G, H, I, J, K, K′, L),

however, other helices such as A′, B″, F′, G′, J′, K″, and L′ are present in some CYP enzyme structures [58, 59]. CYP enzyme structures also contain four β-sheets—β1 (5 strands), β2 (2 strands), β3 (3 strands), β4 (2 strands), and like helices, there are two β-sheets (β5 and β6), that only occur in some CYP enzymes [50, 58, 59].

Similar to all other CYP enzymes in the superfamily, human CYPs have an active site located in the enzyme's hydrophobic core, which is connected to the outside of the enzyme by channels [60–63]. Water, substrates, products and O_2 can move through these channels to access the active site [61, 64–66]. Even though the overall structure of human CYP enzymes is highly conserved, the active sites have shown great variability [67]. The cavities differ in shape and size, and their differences stem from discrepancies within the F-G and B' helices [67]. The B–C, F–G and L helix regions are the most variable within human CYP enzymes, and play a key role in substrate specificity, particularly the B′ helix [67, 68]. Additionally, these regions undergo conformational changes upon substrate entry and binding [3]. It is because of these dissimilar regions that these CYP enzymes have variable substrate preferences [3]. For instance, the CYP1A2 enzyme from the CYP family 1 preferentially binds small heterocyclic amines or planar aromatic substrates [68]. This specificity is due to polar residues, such as Thr118, Ser122, and Thr124, located in the B'–C helix region facing the active site, along with Thr223 (in the F helix) and Asp320 (in the I helix), which are positioned on the "roof" of the active cavity [68]. In contrast, CYP2A6 from CYP family 2, which has a smaller active site pocket, favors the metabolism of small aromatic substrates, aided by the substrate-orienting residue Asn297 [69, 70].

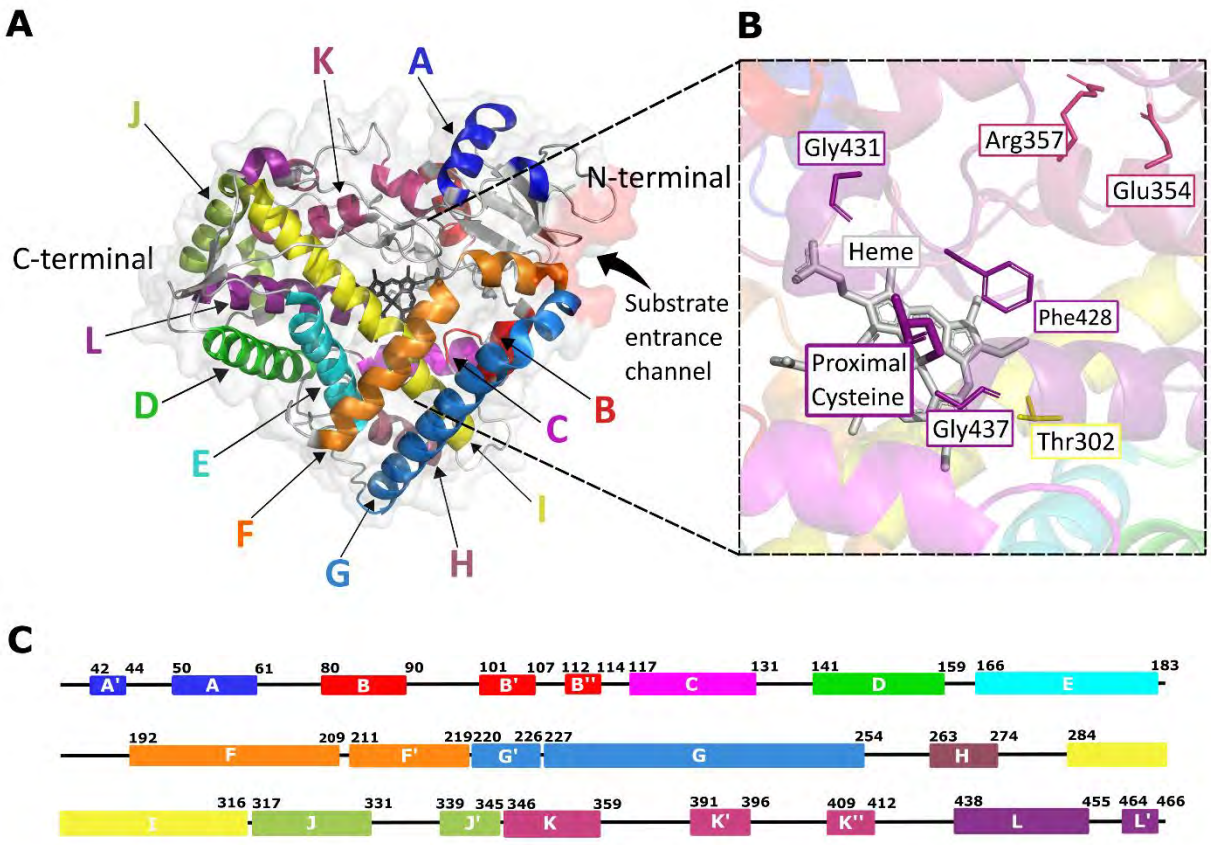


Figure 1-3: (A) General structure of CYP enzymes. CYP2C8 (PDB ID: 2NNJ) is used to map the major helices denoted A-L in different colours. Substrate entrance site is highlighted in red and an arrow pointing towards the entrance is in black. (B) Conserved catalytically important residues mapped onto the CYP structure, and the cysteine that acts as a proximal axial thiolate ligand for the heme iron is highlighted. Threonine in the conserved motif (A/G)XXT is coloured in yellow, while glutamic acid and arginine in the EXXR motif are coloured in warm pink. Phenylalanine, glycine and cysteine in the conserved motif FXXGXXXCXG are all coloured in purple. (C) CYP2C8 enzyme residue numbers are denoted, as an example, for the helices observed in CYP enzymes. Adapted from Chamboko et al., 2023 [3].

CYP enzymes contain a prosthetic heme group that forms the "floor" of the active site and plays a crucial role in enzyme function during metabolism [67, 71]. This heme group is situated between helices I and L, and is covalently bound to the enzyme through a sulfur atom of a conserved cysteine residue which serves as the proximal axial thiolate ligand for the heme iron [3, 67, 72, 73] (Figure 1-3B). The I helix in close proximity to the heme houses a conserved acid/alcohol pair that is critical for the CYP active cycle, while the cysteine acting as the proximal axial thiolate ligand is situated on the loop before the L helix [73]. The highest structural conservation of CYPs is most prominent around this heme–thiolate oxygen activation chemistry [67]. The primary role of the heme in CYP enzymes is to serve as a reactive center, enabling oxygen activation and substrate oxidation [71, 73], and water movement is facilitated by coordinated water gates, which are regulated by the propionate side chains of the heme and their salt bridge interactions [67].

The oxidative capabilities of CYP enzymes are attributed to the formation of a coupled high-valent iron (IV)–oxo porphyrin π-cation radical species (Compound I), which involves the enzyme's heme iron and dioxygen, along with two reducing and two proton equivalents, mainly supplied during the catalytic cycle of the enzyme [64, 74, 75]. The β-bulge region (FXXGXXXCXG), which contains the cysteine axial thiolate ligand, is highly conserved and rigidly structured to maintain a hydrogen-bonding distance from two peptide NH groups—present in all CYPs—even though the geometry only allows for one hydrogen bond [67]. These hydrogen bonds play a key role in regulating the redox potential of the heme iron, as without them, the potential would be too low for reduction by redox partners [76, 77]. To maintain a physiologically

relevant redox potential, the enzyme must create a suitable electrostatic environment around the cysteine ligand. While Compound I is considered the ultimate oxidant in CYP enzymes, their oxidative versatility is likely due to their structural flexibility, particularly within the active site, and the ability of the catalytic cycle to be initiated by the binding of an oxidizable substrate [64]. Furthermore, specific features in the active sites and entry channels of CYP enzymes—often resulting from mutations—can influence the stereo- and regioselectivity of substrate oxidation, leading to different oxidized products from the same substrate [64, 78, 79].

1.3.2.3 Mechanism of action of human CYP enzymes

As previously mentioned, human CYP enzymes are involved in phase 1 metabolism [29, 80] where they act as monooxygenases, i.e., they catalyze the addition of one atom of molecular oxygen into a substrate, while the other atom is reduced to water [35, 36, 81–83]. They have been observed to metabolize different environmental chemicals/pollutants, drugs, toxins and xenobiotics [84–86]. The first three families, CYP1, CYP2, and CYP3, include CYP enzymes responsible for the metabolism and clearance of approximately 70-80% of pharmaceutical drugs, and within these families, the enzymes CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C18, CYP2C19, CYP2D6, CYP2E1, CYP3A4, and CYP3A5 account for the majority of that metabolism [7, 50].

CYP enzymes metabolize drugs through oxidation reactions such as epoxidation, hydroxylation, sulfoxidation, desaturation and C–C bond cleavage [87–93]. They carry out this

metabolism in a sequence of steps, with a reaction occurring at each step, that forms a transient intermediate and product which can be utilized as substrates in the next step (for the next reaction). As the sequence of reactions take place in metabolism of substrates with CYP enzymes, the heme iron changes spin states because of substrate binding, as the outer orbital has one or more unpaired electrons, which gives it a higher energy state [94]. Figure 1-4 displays an overview of the typical seven- step catalytic cycle CYP enzymes partake in [3, 64, 78, 79].

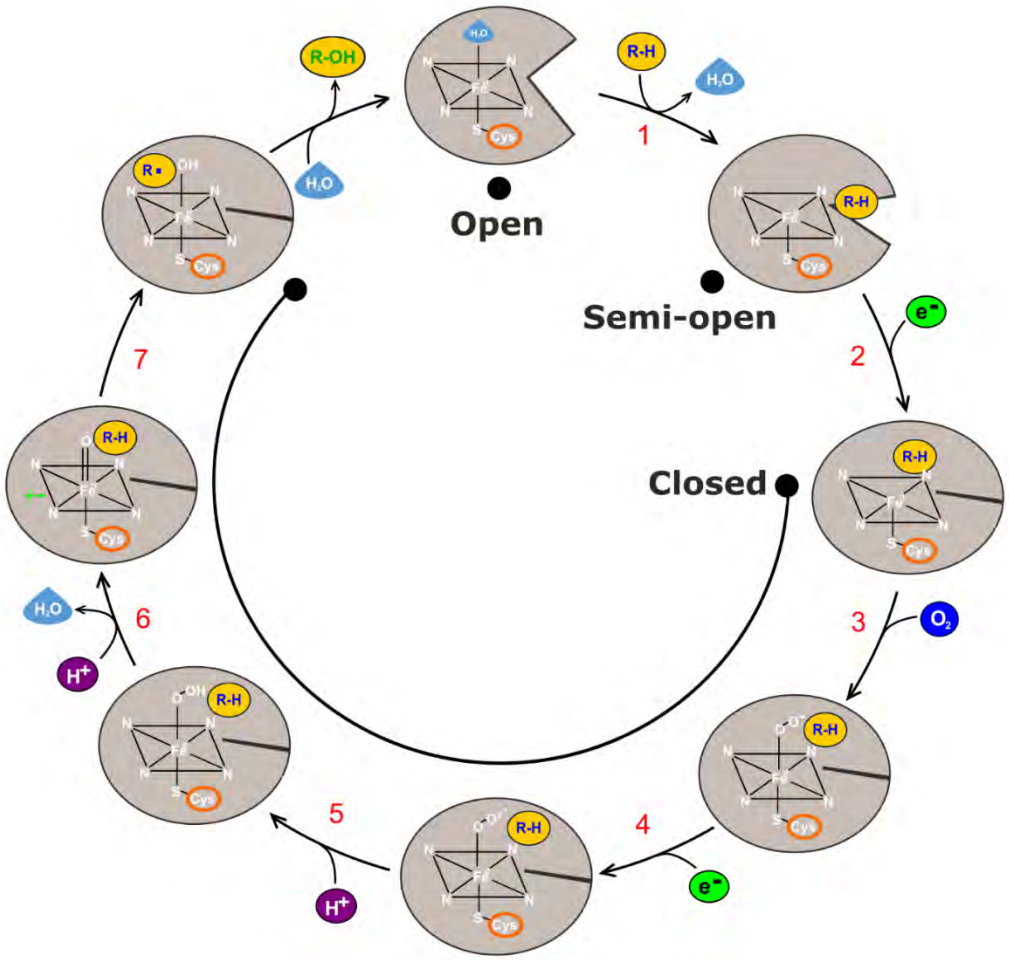


Figure 1-4: Mechanism of CYP enzymes in a seven-step catalytic cycle. Enzyme open, semi-open and closed conformations at each stage denoted by a grey shape. The substrate, hydrogens, electrons and oxygen are denoted by yellow, purple, green and dark blue circles, respectively. Water is shown by a blue drop. Adapted from Chamboko *et al*., 2023 [3].

Prior to substrate binding, the heme iron within CYP enzymes is coordinated to a water molecule and exhibits a low-spin ferric resting state [3]. The P450 catalytic cycle then begins with the substrate (RH) binding to the CYP enzyme, initiating an active site conformational change that displaces the water molecule from the distal axil coordination position of the heme iron (Figure 1-4). A spin state change of the heme iron also occurs, from a low-spin (LS) to a high-spin (HS) [93, 95]. In the second step of the cycle, the first reduction occurs, where the Fe^{3+} ion in the complex is reduced by an electron that is transferred from NAD(P)H through an electron transfer chain to its more ferrous state Fe^{2+} [95–97]. An oxygen molecule (O_2) then binds rapidly to the Fe^{2+} in step 3, forming O_2- Fe^{2+}- RH [98]. This oxygen binding leads to a slow rearrangement of the O_2- Fe^{2+} -RH complex, forming an oxy-P450 complex (O - - O – Fe^{3+} - RH) [98]. In the cycle, this is the last relatively stable intermediate as compared with other intermediates.

In step 4, a second reduction occurs, and this has been determined to be the rate-determining step of the cycle [3, 95]. A second electron is supplied to the complex, forming a peroxo-ferric intermediate (O_2- - O – Fe^{3+} - RH). This complex is then rapidly protonated by the surrounding amino acid side chains, to form a hydroperoxo-ferric intermediate in step 5 (Compound 0 (Cpd 0)) [64]. A second protonation also occurs in step 6, at the distal oxygen atom, and the heterolysis of an O-O bond occurs, leading to the release of a mole of water and the production of Compound I (Cpd I), which is a ferryl (Fe^{IV})-oxo-π porphyrin cation radical and the ultimate oxidant [3, 98]. In step 7 of the catalytic cycle, oxygen is inserted into the substrate, forming a hydroxylated form of the substrate (Figure 1-4). Depending on the substrate, it can be activated by either having a hydrogen or an electron removed from its structure, leaving a carbon with a spare electron, and

making it a reactive radical that is much more likely to react with a hydroxyl group. The carbon radical and the hydroxyl group react, forming an alcohol [98]. The product has now undergone chemical and structural changes, transforming into a metabolite that cannot remain bound to the CYP active site [95]. It is then released, and the enzyme returns to its resting/ferric state, with a water molecule occupying the distal coordination position of the iron nucleus. The enzyme is now ready for a new substrate to bind, and the cycle may begin again [3, 98].

1.3.3 Classes of human CYP enzymes and their redox partners

Studies have revealed that CYP enzymes cannot function on their own and require a redox partner to bind and facilitate the electron transfer necessary for ligand metabolism [67]. During this electron transport, the transfer of reducing equivalents from nicotinamide adenine dinucleotide phosphate (NADPH) to molecular oxygen occurs [99]. Although this is a very important step in the metabolism of substrates with CYP enzymes, studies with CYPs and redox partners in complex form are challenging, as the complexes are difficult to crystallize. This results in a limited number of protein-protein complexes with redox partners and CYPs in the Protein Data Bank (PDB) [3].

There are different redox partners for human CYP enzymes within the different locations CYPs can be found and because of this, CYPs are grouped into two main classes (class I and class II), depending on cellular location and redox partners [3]. In class I, there are CYP enzymes found within the mitochondria, that bind two redox partners, which are ferredoxin/adrenodoxin reductase (a reductase containing flavin) and ferredoxin/adrenodoxin (an iron–sulfur protein). Class II CYP

enzymes are located in the endoplasmic reticulum and bind to a redox partner called NADPH-cytochrome P450 oxidoreductase (POR) [3]. POR is a single polypeptide containing one molecule of flavin mononucleotide (FMN) and one molecule of flavin adenine dinucleotide (FAD) [67]. It transfers electrons to the CYP heme group from the two-electron donor NADPH, with FMN acting as an electron carrier and FAD serving as a dehydrogenase flavin. The CYP enzymes studied in this work are from class II.

1.3.4 Existing structural information on human CYP enzymes within CYP 1, 2 and 3 families

This thesis focuses primarily on enzymes within the first three CYP families, making it essential to obtain structural information for the *in silico* experiments. Therefore, it was important to determine what structural data on these enzymes were available. When the PDB [100] was analyzed in November 2022 by Chamboko *et al*., [3], it contained a total of 177 structures for members of the CYP families 1, 2, and 3. CYP1A2, CYP2D6, and CYP3A4 were among the first enzymes in the CYP family to have their structures deposited, with the first submission occurring in 2003 [3]. Over a decade later, structures had become available for many of the 23 human CYP enzymes from the first three families, however, certain enzymes, including CYP2A7, CYP2C18, CYP2F1, CYP2J2, CYP2S1, CYP2U1, CYP2W1, and CYP3A43, still lacked structural data [3]. Of the enzymes with structures, CYP3A4 had the most structural data, with 87 available structures in the PDB [3, 100]. CYP2A6 and CYP2D6 followed closely, each with 14 structures [3, 100]. The PDB also includes variants of these enzymes, such as CYP2C9 alleles *2, *3, *8, and *30 [3].

A summary of the most representative structures for each enzyme in these families is provided in Table 1-1.

Table 1-1: 3D structures of enzymes from CYP 1, 2 and 3 enzyme families that are available on PDB with their respective resolutions and co-crystalized ligands as of November 2022. Adapted from Supplementary Data, Chamboko *et al*., 2023 [3].

Family	Enzyme name	PDB IDs	Resolution (Å)	Co-crystalised ligand
CYP 1	CYP 1A1	4I8V	2.60	Alpha-naphthylflavone
		6DWM	2.85	Bergamottin
		6UDL	2.85	Duocarmycin Prodrug (S) ICT-2700
		6DWN	3	Erlotinib
		6UDM	3.07	Duocarmycin Prodrug (S) ICT-2700
		6O5Y	3.17	5-amino-N-{5-[(4R,5R)-4-amino-5-fluoroazepan-1-yl]-1-methyl-1H-pyrazol-4-yl}-2-(2,6-difluorophenyl)-1,3-thiazole-4-carboxamide
	CYP 1A2	2HI4	1.95	Alpha-naphthylflavone
	CYP 1B1	3PM0	2.70	Alpha-naphthylflavone
		6OYU	2.95	Alpha-naphthylflavone
		6OYV	3.10	Estradiol
		6IQ5	3.70	2-(cis-4-azidocyclohexyl)-4H-naphtho[1,2-b]pyran-4-one
CYP 2	CYP 2A6[#]	2FDV	1.65	N-Methyl(5-(pyridin-3-yl)furan-2-yl)methanamine
		2FDU	1.85	N,N-Dimethyl(5-(pyridin-3-yl)furan-2-yl)methanamine
		1Z10	1.90	Coumarin
		2FDY	1.95	Adrithiol
		2PG5	1.95	-
		2FDW	2.05	5-(Pyridin-3-yl)furan-2-yl)methanamine
		1Z11	2.05	
		3T3Q	2.10	Methoxsalen
		3EBS	2.15	Pilocarpine
		4EJJ	2.30	Phenacetin
				Nicotine
	CYP 2A7	-	-	-
	CYP 2A13	4EJI	2.10	4-(methylnitrosamino)-1-(3-puridyl)-1-butanone

	2P85	2.35	Indole
	4EJH	2.35	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone
	4EJG	2.50	Nicotine
	3T3S	3.00	Pilocarpine
CYP 2B6[#]	5UFG	1.76	Myrtenyl bromide
	5UDA	1.93	Camphane
	4I91	2.00	Alpha-pinene
	3IBD	2.00	4-(4-chlorophenyl)imidazole
	5UAP	2.03	Bornyl bromide
	4RQL	2.10	Sabinene
	3QOA	2.10	4-Benzylpyridine
	4RRT	2.20	(+)-3-carene
	4ZV8	2.24	Alpha-pinene
	5UEC	2.27	Myrtenyl bromide
CYP 2C8	2NNJ	2.28	Felodipine
	2NNH	2.60	9-cis-retinoic acid
	1PQ2	2.70	-
	2VN0	2.70	Troglitazone
	2NNI	2.80	Montelukast
CYP 2C9[#]	5X23	2.00	Losartan
	5W0C	2.00	Ethyl {2-[([1,3]thiazolo[4,5-c]pyridine-2-carbonyl)amino]thiophene-3-carbonyl}carbamate
	1R9O	2.00	Flurbiprofen
	5A5I	2.00	N-[4-(3-chloranyl-4-cyano-phenoxy)cyclohexyl]-1,1,1-tris(fluoranyl)methanesulfonamide
	7RL2	2.23	Losartan
	5XXI	2.30	Losartan
	5K7K	2.30	4-[4-chloranyl-2-(1~{H}-pyrazol-4-yl)phenoxy]-3-cyano-~{N}-(1,3-thiazol-2-yl)benzenesulfonamide
	4NZ2	2.45	(2R)-N-{4-[(3-bromophenyl)sulfonyl]-2-chlorophenyl}-3,3,3-trifluoro-2-hydroxy-2-methylpropanamide
	5X24	2.48	Losartan
	1OG5	2.55	S-warfarin
CYP 2C18	-	-	-
CYP 2C19	4GQS	2.87	(4-hydroxy-3,5-dimethylphenyl)(2-methyl-1-benzofuran-3-yl)methanone
CYP 2D6[#]	3TBG	2.10	Thioridazine
	4WNU	2.26	Quinidine
	4WNV	2.35	Quinine

		6CSD	2.39	Prinomastat
		6CSB	2.39	Thioridazine
		4XRZ	2.40	(4aR,6R,8aS)-8a-(2,4-difluorophenyl)-6-(1H-pyrazol-4-yl)-4,4a,5,6,8,8a-hexahydropyrano[3,4-d][1,3]thiazin-2-amine
		4XRY	2.50	(4aR,6R,8aS)-8a-(2,4-difluorophenyl)-6-(1-methyl-1H-pyrazol-4-yl)-4,4a,5,6,8,8a-hexahydropyrano[3,4-d][1,3]thiazin-2-amine
		4WNT	2.60	
		3TDA	2.67	
		5TFT	2.71	Ajmalicine Prinomastat (4S)-4-[2,4-difluoro-5-({[1-(trifluoromethyl) cyclopropyl]amino}methyl)phenyl]-4-methyl-5,6-dihydro-4H-1,3-thiazin-2-amine
	CYP 2E1	3E6I	2.20	1H-indazole
		3T3Z	2.35	Pilocarpine
		3E4E	2.60	4-methylpyrazole
		3GPH	2.70	Omega-imidazolyl-decanoic acid
		3KOH	2.90	Omega-imidazolyl octanoic acid
		3LC4	3.10	Omega-imidazolyl-dodecanoic Acid
	CYP 2F1	-	-	-
	CYP 2J2	-	-	-
	CYP 2R1	3CZH	2.30	Vitamin D2
		3DL9	2.72	1-alpha-hydroxy-vitamin D2
		3C6G	2.80	Vitamin D3
	CYP 2S1	-	-	-
	CYP 2U1	-	-	-
	CYP2W1	-	-	-
CYP 3	CYP 3A4[#]	5VCC	1.70	-
		6MA8	1.83	Phenylmethylsulfonyl fluoride
		4D6Z	1.93	Tert-butyl {6-oxo-6-[(pyridin-3-ylmethyl)amino]hexyl}carbamate and Imidazole
		5VCD	1.95	-
		3NXU	2.00	Ritonavir
		1TQN	2.05	-
		6MA7	2.09	Fluconazole
		3UA1	2.15	Bromoergocryptine
		6DAA	2.15	Tert-butyl [(2S)-1-{[(2R)-1-oxo-3-phenyl-1-{[(pyridin-3-yl)methyl]amino}propan-2-
		6MA6	2.18	

			yl]sulfanyl}-3-phenylpropan-2-yl]carbamate Metyrapone
CYP 3A5	6MJM	2.20	-
	7SV2	2.46	Azamulin
	7LAD	2.65	Clobetasol propionate
	5VEU	2.91	Ritonavir
CYP 3A7	7MK8	2.15	Dithiothreitol
CYP 3A43	-	-	-

[#] CYP enzymes with more than 10 structures on PDB (for the table a cut-off of 10 structures with the best resolution was used).

CYP enzyme structures in PDB can be found in either open, partially open, or closed conformations and typically, the open conformation corresponds to ligand-free proteins, while the closed conformation is mainly associated with a substrate bound structure [3, 101]. There are however, instances of closed ligand-free structures and partially open ligand-bound forms identified [101, 102]. Differences in areas such as the pre-helix A, the B–C and F–G regions are responsible for the various conformations observed in human CYP enzymes [3, 60, 101–104] and as an example, the open and closed conformations of CYP2D6 are depicted in Figure 1-5. In the closed state, residue G42 occurring in the pre-helix A is 14.3 Å from the E222 residue in helix F', whereas in the open state, this distance increases by 0.6 Å to 14.9 Å, which helps open the substrate channel and facilitates access to the active site [3, 102]. This structural change is followed by further alterations in adjacent regions, including the F and G helices and the B-C loop with helix B', contributing to the open conformation of CYP2D6 [3, 102].

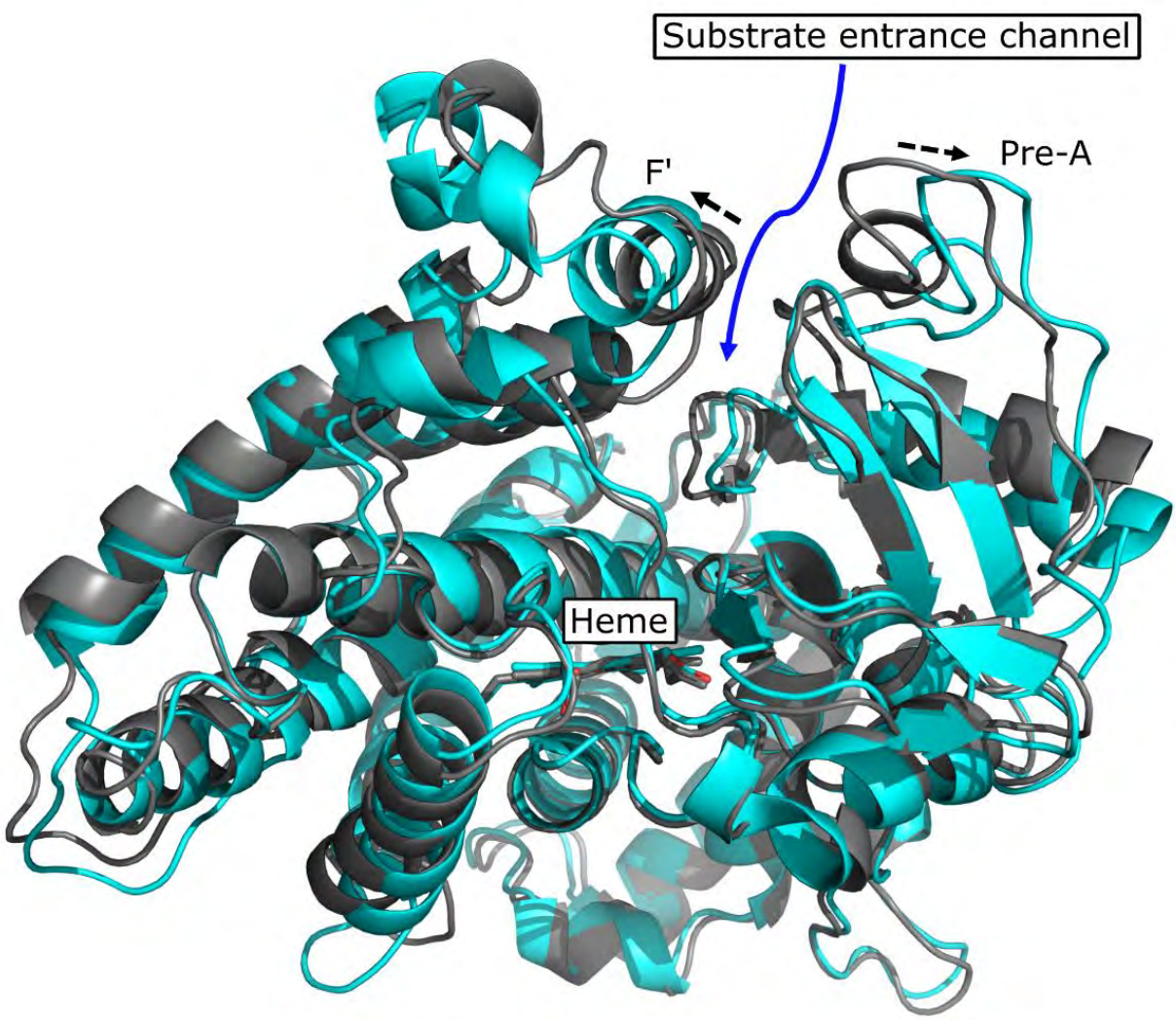


Figure 1-5: Open and closed conformations of CYP2D6. Open conformation of CYP2D6 structure in cyan from PDB (3TBG) superimposed onto closed conformation structure in grey (3QM4). Direction of displacement of F' and Pre-A regions to form the open structure is shown in dotted arrows. Adapted from Supplementary Data, Chamboko *et al*., 2023 [3].

1.3.5 Genetic variations in human CYP enzymes with a particular focus on CYP2D6 polymorphisms

Genes coding for CYP enzymes have been noted to be highly polymorphic, with variations in these pharmacogenes leading to enzymes with different activities which in turn impact the efficacies of drugs they metabolize [3]. The genetic variations may cause phenotypes such as poor, intermediate, extensive and ultra-rapid metabolism [46, 105, 106]. Extensive metabolism is regarded as the normal activity of CYP enzymes, and extensive metabolizers contain two alleles with normal activity, which are generally referred to as the *1 or reference allele [3]. The reference/consensus allele typically constitutes the major proportion of the population with normal activity, whereas polymorphisms that lead to variant alleles should be present in at least 1% of the population [46, 107]. Other metabolizers such as ultra-rapid metabolizers may possess multiple copies of the same allele, but this is not always the case, and intermediate and poor metabolizers usually have one and two defective alleles respectively [108, 109].

Among all the human CYP families, families 2 and 3 are the most significantly affected by variations in pharmacogenes, and this impacts Africa greatly, as it highly affects the metabolism of drugs used to treat two major diseases causing death in Africa: malaria and tuberculosis (TB). Examples of human CYP enzymes within the first three families and the antimalarial or antituberculosis drugs that they metabolize can be viewed in Table S1. A study by Yusof *et al*., [110] uncovered that there were higher adverse effects in patients who were administered the antimalarial prodrug artesunate. These patients carried the CYP2A6*1B allele, which is associated with ultra-rapid metabolism. This suggested that the patients may have had an accumulation of an

active metabolite (dihydroartemisinin), which caused the adverse effects. CYP2B6 and CYP2C8 are known to metabolize drugs used in the artemisinin-based combination therapy (ACT) for malaria patients. Alleles that are more common, especially within the African population, such as CYP2B6*6 and CYP2C8*2, as well as CYP2C8*3, which is more common in Caucasians, exhibit reduced activity, potentially affecting patient response to the malarial drugs [111–113]. Chlorproguanil and proguanil are metabolized by CYP2C19, and its alleles CYP2C19*2 and CYP2C19*3 are primarily linked to poor metabolism of these antimalarial drugs, while allele CYP2C19*17 is associated with an increase in metabolism [114]. CYP2D6 is involved in the metabolism of primaquine, and reduced activity alleles (such as CYP2D6*10 and CYP2D6*17), have been linked to decreased primaquine antimalarial activity, as insufficient production of active metabolites may occur [115]. Lastly, with regards to CYP3A4 and CYP3A5, there is evidence suggesting alleles CYP3A4*22, CYP3A5*3, CYP3A5*6, and CYP3A5*7 exhibit poor-metabolizer phenotypes and affect the metabolism of the malarial drug quinine [20, 52, 116].

In relation to TB, although the antituberculosis drug bedaquiline is metabolized by CYP3A4, a population pharmacokinetic model showed that clearance of this drug was 52% faster in Africans, possibly due to higher expression of CYP3A5 in this population compared to others [117, 118]. This overlap in substrate specificity is quite often observed between CYP3A4 and CYP3A5. However, in some studies, the allele CYP3A5*3 that had been noted to be non-functional, was related to slow bedaquiline clearance in South Africans that were treated for multidrug-resistant tuberculosis (MDR-TB) [119, 120].

A considerable number of CYP alleles have been identified and cataloged in databases such as PharmVar [48, 49], ClinVar [121] and the Pharmacogenomics Knowledge Base (PharmGKB) [122, 123], however, information on the clinical function of most of the alleles discovered remains limited, which, in turn, restricts our comprehension of their effects on certain clinical drugs. Of all human CYPs, CYP2D6 is highly polymorphic, resulting in a large number of genetic variants that present differing metabolic activities among individuals [124]. It is also responsible for the metabolism of ~25% of clinical drugs, making it one of the most significant CYP enzymes in drug metabolism [125–127]. Consequently, genetic variations within this enzyme can significantly influence the metabolism and safety of a wide range of therapeutic agents such as certain antidepressants, malarial drugs (Table S1), pain medication and antipsychotics [128–131]. CYP2D6 alleles are widespread globally; however, for the purposes of this study, only a select few were included to meet the research requirements. As described by PharmGKB and the Clinical Pharmacogenetics Implementation Consortium (CPIC), there are nine biogeographical groups and the retained CYP2D6 alleles, and their biogeographical prevalence are detailed in Table 1-2 and shown in Figure 1-6 [122, 123].

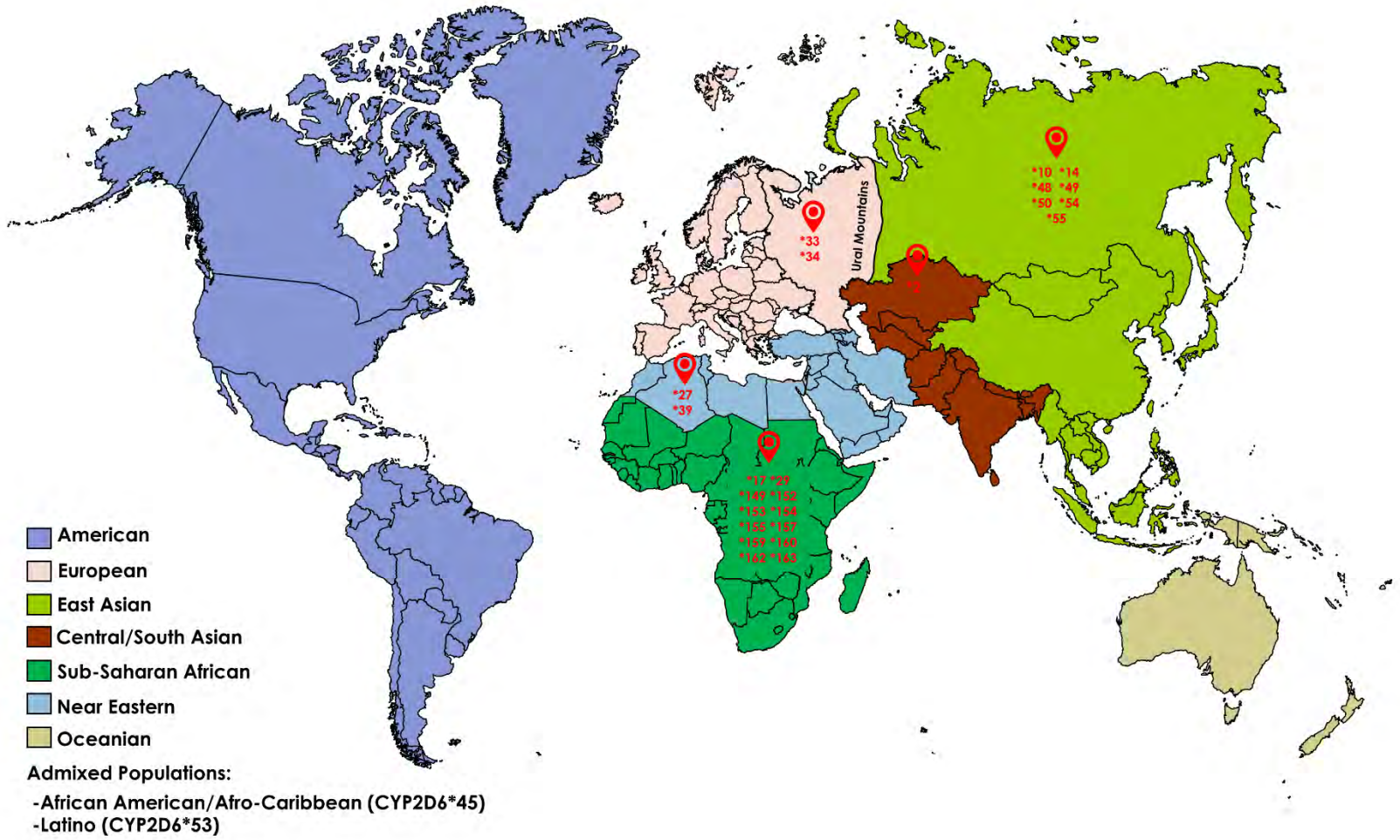


Figure 1-6: Map showing prevalent CYP2D6 alleles in the different biogeographical groups. CYP2D6 alleles (highlighted in red) and the location where they are highly frequent shown in the world map. There are nine biogeographical groups as described by PharmGKB and CPIC however the African American/Afro-Caribbean and Latino groups could not be shown as they encompass populations with considerable gene flow between different geographic regions due to post-colonization and post-Diaspora events. The biogeographical groups colour coded according to the key in the figure. Modified from PharmGKB [122, 123].

Table 1-2: CYP2D6 alleles studied and their variations and population frequencies across the nine biogeographical groups as described by CPIC and PharmGKB [122, 123]. Highest frequencies for each allele are highlighted in bold.

CYP 2D6 Allele	Variations	Population frequencies (%)[#]								
		African American/Afro-Caribbean	**American**	**Central/South Asian**	**East Asian**	**European**	**Latino**	**Near Eastern**	**Oceanian**	**Sub-Saharan African**
***2**	R296C, S486T	15.47	21.74	**27.45**	11.92	18.54	22.68	19.00	6.11	17.41
***10**	P34S, S486T	3.82	1.45	7.56	**42.84**	1.57	2.63	6.77	5.71	4.87
***14**	G169R, R296C, S486T	0.00	0.00	0.00	**0.47**	0.00	0.00	-	0.00	0.00
***17**	T107I, R296C, S486T	16.88	0.53	0.05	0.01	0.39	2.33	3.10	0.10	**19.36**
***27**	E410K	0.30	0.30	0.00	0.06	0.00	-	**0.70**	-	0.34
***29**	V136I, R296C, V338M, S486T	8.74	0.20	0.16	0.01	0.10	1.52	0.78	0.00	**10.83**
***33**	A237S	0.00	0.22	0.20	0.02	**0.95**	-	0.00	-	0.00
***34**	R296C	0.00	0.00	0.00	0.89	**1.07**	0.05	-	0.18	0.10
***39**	S486T	2.01	0.10	0.34	0.56	1.40	0.80	**2.68**	0.63	0.00
***45**	E155K, R296C, S486T	**5.01**	0.30	0.00	0.00	0.00	-	0.00	-	1.00
***48**	A90V	-	-	-	**0.02**	-	-	0.00	-	-
***49**	P34S, F120I, S486T	0.00	0.00	0.00	**0.99**	0.00	-	0.00	-	0.00
***50**	E156A	-	-	-	**0.04**	0.00	-	0.00	-	0.00
***53**	F120I, A122S	-	-	-	0.02	-	**0.50**	0.00	-	0.00

***54**	P34S, T261I, S486T	-	-	-	**0.04**	0.00	-	0.00	-	-
***55**	R296C, K404Q, S486T	-	-	-	**0.02**	-	-	0.00	-	-
***149**	D97E, V136I, R296C, V338M, S486T	0.00	0.00	0.00	0.00	0.00	-	-	-	**15.02**
***152**	A226V	-	-	-	-	-	-	-	-	**0.05**
***153**	E215K	-	-	-	-	-	-	-	-	**0.05**
***154**	T107I, R296C, R441H, S486T	-	-	-	-	-	-	-	-	**0.05**
***155**	V136I, R296C, V338M, M451I, S486T	-	-	-	-	-	-	-	-	**0.10**
***157**	V136I, A165V, R296C, V338M, S486T	-	-	-	-	-	-	-	-	**0.05**
***159**	P41L, R296C, S486T	-	-	-	-	-	-	-	-	**0.05**
***160**	L91M, H94R, R296C, S486T	-	-	-	-	-	-	-	-	**0.05**
***162**	R296C, E410K, S486T	-	-	-	-	-	-	-	-	**0.05**
***163**	W128R, R296C, S486T	-	-	-	-	-	-	-	-	**0.05**

[#]Calculated by CPIC and PharmGKB
-No data available

According to Table 1-2, CYP2D6*2 is more frequently found in the Central/South Asian population, while CYP2D6*45 is more frequently observed in the African American/Afro-Caribbean population. The Latino population showed a higher frequency of the CYP2D6*53 allele, while the CYP2D6 alleles *10, *14, *48, *49, *50, *54, and *55 were more commonly found in the East Asian population. In Europe, CYP2D6*33 and CYP2D6*34 were more frequently observed in the population, while alleles CYP2D6*27 and CYP2D6*39 were highly prevalent within the Near Eastern population. Populations in Sub-Saharan Africa were noted to have a higher frequency of the well characterized CYP2D6*17 and CYP2D6*29, as well as the recently identified alleles *149, *152, *153, *154, *155, *157, *159, *160, *162 and *163.

1.4 Problem Statement and Motivation

For decades, a "one-size-fits-all" approach to prescribing medication has been the standard treatment method for patients, and although this method has proven effective for many, there are others for whom it has not yielded favorable outcomes. There have been numerous reports outlining different drugs that have caused toxicity in patients, or ADRs that led to severe illness and in extreme cases, death. Many factors contribute to these negative effects, such as age, gender and the environment [3], however, a patient's genetic makeup has been identified as having a direct influence on how drugs are metabolized and cleared from the body [6–8]. Specifically, enzymes belonging to the CYP enzyme superfamily partake in phase I drug metabolism [14, 29], where they convert drugs into activated metabolites that can then be utilized further in the body [29]. Enzymes from CYP families 1, 2, and 3 are particularly important, as they are responsible for metabolizing most foreign substances in the human body, including 70–80% of all clinical drugs currently used [3, 50]. The genes coding for these CYP enzymes are part of the pharmacogenes within the human genome [10], and variations within these pharmacogenes, such as deletions, insertions, copy number variations or SNPs can result in an enzyme with altered activity, that will in turn affect the breakdown of drugs in the body and drug efficacy [3]. Thus, research on the enzymes within the first three CYP families is of importance as it may aid in understanding these polymorphisms and the resultant enzyme activities.

A CYP family 2 enzyme, CYP2D6, is the most polymorphic member of the cytochrome P450 family, that is responsible for the metabolism of a large number of clinical drugs (~25%)

[125–127]. As such, polymorphisms in this enzyme can have a substantial impact on the metabolism and safety of various therapeutic agents. The alleles exhibit a range of functional consequences, including null function, reduced function, normal function, or increased function. Although *in vitro* studies have explored the impact of genetic polymorphisms on CYP2D6 activity, the growing number of allelic variants necessitates the development of more efficient approaches to assess enzyme function, and in the current study, we aim to advance the development of these methods. Furthermore, there is a lack of pharmacogenomic studies at the 3D level, both with and without the drugs they metabolize, and research focusing on African variants is even more neglected.Thus, this study also aims to expand pharmacogenomic research at a 3D level and within the African context. The findings will contribute to the furtherment of personalized medicine such that in a future time, healthcare providers will be able to select the most effective therapies and drug doses for each patient, thereby improving treatment outcomes and minimizing adverse effects in all populations.

1.5 Aims and Objectives

The main aim of this research was to advance 3D pharmacogenomic research on a CYP2 family enzyme, CYP2D6, with an emphasis on African populations. To achieve this aim, extensive *in silico* analyses was carried out with the following objectives:

1. Characterization of Cytochrome P450 enzymes from CYP families 1, 2, and 3, to explore and comprehend the structural and functional relationships between enzymes within these three families, specifically noting how CYP2D6 connects to them.
2. Identify CYP2D6 variant alleles with known normal and reduced function, as well as Afrocentric alleles with no known clinical function, then select CYP2D6 reference structure and model variant allele structures.
3. Perform Molecular Dynamic simulations on reference and allele structures, and utilize different *in silico* tools to analyze and deduce possible mechanisms by which polymorphisms lead to reduced activity from alleles with known functions.
4. With the mechanisms identified, the next step was to apply them for the prediction of reduced functionality, on recently identified Sub-Saharan African/Afrocentric alleles without any clinical function assigned to them, to advance pharmacogenomic research on CYP2D6 alleles in African populations.

Chapter 2

Characterization of Human Cytochrome P450 Enzymes in Families 1, 2 and 3 with respect to CYP2D6 and its variants

Chapter overview

This chapter presents studies that cover sequence, motif, and phylogenetic analysis of Cytochrome P450 enzymes in families 1, 2, and 3 to characterize and identify unique sequence and structural features crucial for the functioning of these enzymes in metabolizing various compounds. Additionally, we investigate the locations of variations in CYP2D6 alleles, a member of the CYP 2 family, to establish a basis for studying how these variations affect the protein structure and potential functional consequences.

2.1 Introduction

The CYP superfamily includes numerous families and subfamilies, each with specific substrate affinities, that have been extensively studied across various organisms [61, 78, 132–134]. Research has explored their sequence variations and structural characteristics, however, while studying the entire superfamily is valuable, focusing on the CYP1, CYP2, and CYP3 families is particularly important for precision medicine, as these families contain CYP enzymes (mainly CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C18, CYP2C19, CYP2D6, CYP2E1, CYP3A4, and CYP3A5) involved in the metabolism of approximately 70-80% of all clinical drugs [7, 50, 135]. Here, we investigate enzymes from the first three families of the CYP superfamily to characterize and identify regions essential for their proper functioning in metabolizing various compounds. A representation of all CYP enzymes from families 1, 2 and 3 studied in this chapter can be seen in Figure 2-1. There are several techniques and tools available for the *in-silico* characterization of proteins, including multiple sequence analysis (MSA), motif analysis, and phylogenetic tree analysis.

2.1.1 Sequence and structure retrieval

Prior to all characterization techniques, sequence and structural information of proteins to be studied must be obtained from reliable sources. There are many different databases that have been developed over the years that harbor sequence and structural data, to aid in the research of organisms. Such databases include UniProt [136], Swiss-Prot [137], RefSeq [138], Protein Information Resource (PIR) [139] and the Protein Data Bank (PDB) [100].

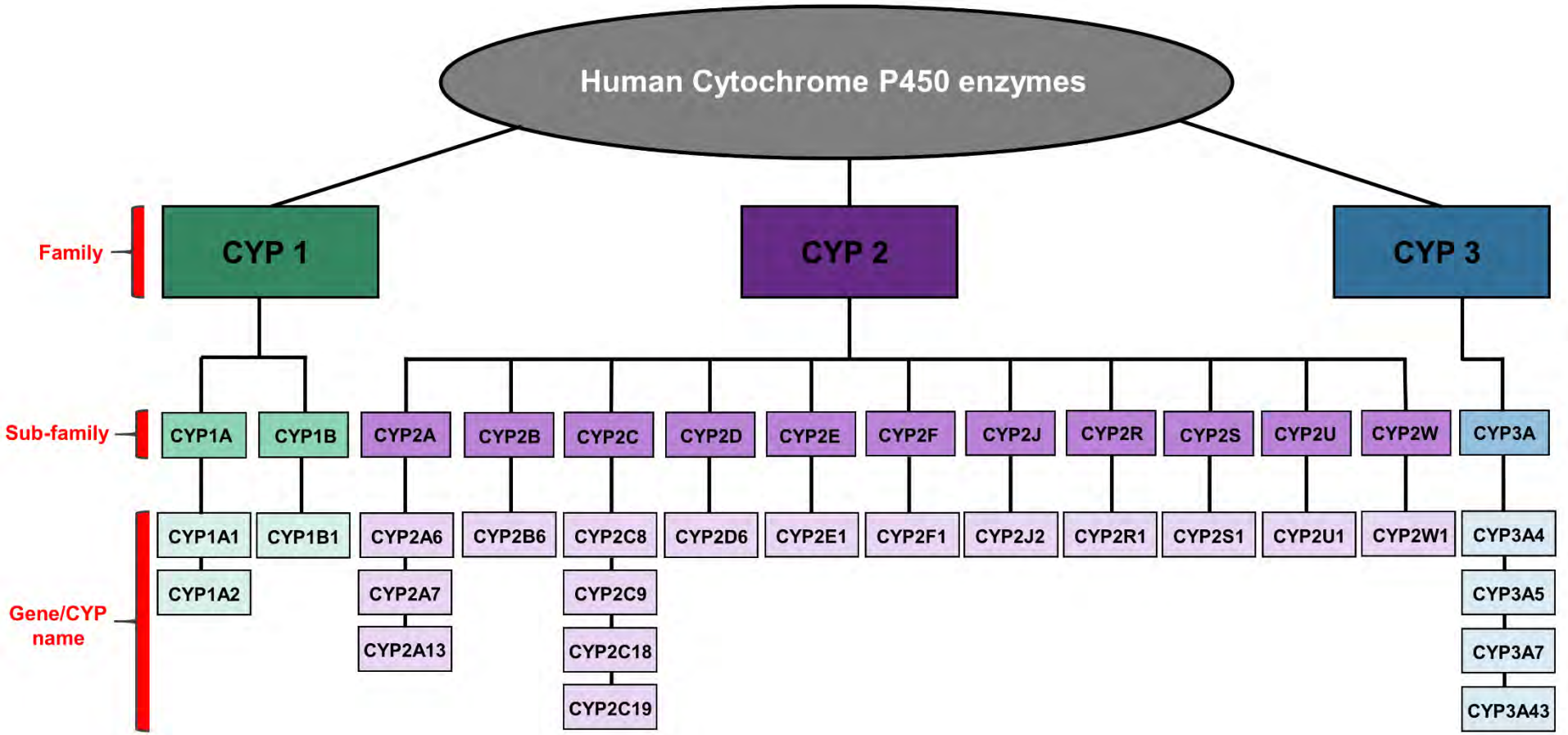


Figure 2-1: Representation of all CYP enzymes from families 1, 2 and 3 studied in this chapter. Different families shown in different colours and different subfamilies within the families shown in a lighter version of the family colour.

In the research presented here, sequences were retrieved from the UniProt database, and structural data was obtained from the PDB. The UniProt database is a freely available resource that provides comprehensive, high quality and curated protein sequence and functional information to researchers [136]. The expert curation of protein information displayed in the UniProt database involves a critical review of experimental data obtained from literature, computational predictions, and interpretations of results from sequence analysis tools by an expert biocuration team [136]. This ensures the data obtained by researchers is reliable for use in their different studies. Structural information of proteins is provided by the PDB, which is a global repository for 3D structural data of proteins and nucleic acids [100]. It was established in 1971 and originally contained 7 structures, however, it has grown tremendously, surpassing 200000 structures in 2023 [140]. The structures available on the PDB are experimentally determined, through techniques such as nuclear magnetic resonance (NMR) spectroscopy, electron microscopy (3DEM) and X-ray crystallography [100]. In the database, the 3D structure is stored in a standardized text file format (.pdb), which contains atomic coordinates and other structural data, and it is assigned a PDB identifier so a structure can be easily located when needed [100]. The file format is a user-friendly format that allows researchers to easily utilize in their different studies or visualize and analyze the structures of macromolecules using various tools such as Chimera or PyMOL.

2.1.2 Multiple sequence alignment

Multiple sequence alignments (MSA) are a key part of *in silico* protein studies as they provide valuable insights into functional regions of related protein sequences. By definition, MSA is a bioinformatics technique that aligns three or more sequences which share evolutionary history,

so as to identify similarities and differences between them [141]. They form the basis of a number of methods, including phylogenetic tree calculations for evolutionary inference, that have also been conducted in this research. There are different approaches and scoring functions that have been developed for the attainment of accurate MSAs, but the heuristic approaches which are divided into three main categories: iterative, progressive and block-based approaches, are not computationally demanding and thus more preferred [142]. In this work, the progressive alignment approach was utilized in programs such as T-COFFEE and PROMALS3D. This is a method that begins by aligning two of the most similar sequences within a given set of sequences, then progressively combines more pairwise sequence alignments, to obtain a final MSA [141]. The most similar sequences are aligned based on a heuristic pairwise alignment method of the Needleman–Wunsch global alignment [143]. In the first stage of this MSA method, a guide tree is constructed to represent the relationships between all sequences in the dataset, then it is used as a reference when sequences are added to the growing MSA in the second stage [141]. Gaps are introduced and extended in sequences during the alignment process to optimize the results. The success of this method, however, depends greatly on the input sequences. Errors within the starting sequences remain fixed, cannot be corrected and are carried through to the end, therefore a careful inspection of the input sequences is required. Nonetheless, the progressive alignment approach is a quick and effective method for aligning large sequence datasets, and it is commonly accessible through various publicly available web servers.

Another heuristic approach used in this work is the iterative method, by programs such as MAFFT and MUSCLE (the two programs both utilize progressive and iterative methods [144,

145]). In this approach, a global alignment is performed initially, followed by the realignment of sequence subsets iteratively to refine the alignment step by step so as to improve its accuracy [146]. This method is based on a concept that an optimal solution can be achieved by iteratively refining existing suboptimal solutions. A more accurate MSA can be produced with this approach, however, it can only align a few hundred sequences in a dataset, which limits researchers working with much larger sequence numbers [146]. In this study, our analysis encompasses less than a hundred sequences, therefore this limitation is mitigated. The results of all MSAs can be viewed using tools such as Bioedit, Cinema or Jalview.

2.1.3 Motif discovery

Motifs are short sequence patterns that are repeated across a given set of sequences, which reveal functional and structural information associated with the sequences [147]. They are evolutionary conserved and unique patterns, however they can further highlight important unique residues that are required for structural stability and for retaining enzyme function [148, 149]. The most popularly used tool to identify motifs is the Multiple Expectation Maximization for Motif Elicitation (MEME) tool [147], which is mainly based on a probabilistic approach [147]. It utilizes statistical techniques to identify *de novo* motifs within a set of unaligned sequences, and results from this method show motif sequence, motif position within the sequences and the length of the motif [147]. The motifs are also displayed as colourful motif logos, which are a representation of the position-specific probability matrices that indicate the likelihood/probability of each possible residue at every position within the motif.

2.1.4 Phylogenetic analysis

Since nature is in a constant state of evolution which results in the creation of many different organisms, the study of ancestral origin and evolutionary divergence within species is important, as it helps researchers better understand present-day organisms. To do this, phylogenetic analysis is carried out, which is the study of the evolutionary relationships among a group of organisms or genes [150]. The evolutionary relationships are usually represented graphically as phylogenetic trees, which show the similarity or divergence of genes or organisms and the level of evolutionary distance between them [151]. Construction of phylogenetic trees begins with an accurately aligned set of sequences (MSAs), and based on sequence homology, amino acid substitution models that best fit the dataset are created and used to build the trees [150]. The amino acid substitution models classified as empirical models are used for phylogenetic inference of amino acid/protein sequences and these include the Whelan and Goldman (WAG) model [152], the Le and Gascuel (LG) model [153], the Jones, Taylor and Thornton (JTT) model [154], and the Dayhoff model [155], just to name a few.

A root is usually the start of a phylogenetic tree, with branches that extend from it and show evolutionary relationships between sequences/genes. Depending on whether an ancestral root is available or unavailable, these trees can be rooted or unrooted, however it is more common to provide unrooted trees [150]. Branches within the phylogenetic trees have different lengths, and this signifies the evolutionary time [150].

Different methods can be utilized to construct phylogenetic trees, but the most commonly employed methods which are regarded as more robust, are the maximum parsimony, maximum likelihood and minimum distance methods [150]. In this study, the maximum likelihood method was used in the Molecular Evolutionary Genetics Analysis (MEGA) program, and although this approach is considered to be quite slow and computationally costly, it is however more accurate [156]. This method is mainly based on the idea that the evolution of each residue in a sequence is independent, meaning the analysis of phylogenetic relationships at each site is enabled [156]. For every residue position in a sequence, the maximum likelihood method applies algorithms that estimate the probability of that residue at a given position, based on whether the ancestral sequences contain that residue [156]. A bifurcating tree starts to emerge after these probabilities are calculated, ultimately leading to the construction of a phylogenetic tree [156].

2.2 Methodology

2.2.1 Sequence retrieval

Curated protein sequences for human Cytochrome P450 enzymes from families 1, 2, and 3 were obtained from UniProt [136]. Specifically, three sequences represented the CYP1 family (CYP1A1, CYP1A2 and CYP1B1), sixteen sequences represented the CYP2 family (CYP2A6, CYP2A7, CYP2A13, CYP2B6, CYP2C8, CYP2C9, CYP2C18, CYP2C19, CYP2D6, CYP2E1, CYP2F1, CYP2J2, CYP2R1, CYP2S1, CYP2U1 and CYP2W1), and four sequences represented the CYP3 family (CYP3A4, CYP3A5, CYP3A7 and CYP3A43). In total, 23 sequences were saved in FASTA format for use in Multiple Sequence Alignment (MSA) and motif discovery. The CYP names and accession numbers are listed in Table 2-1 and actual sequences can be seen in Table S2.

2.2.2 Multiple sequence alignment

The MSA tools, MAFFT [144], T-COFFEE [157] and PROMALS3D [158] were used to carry out MSA on all sequences from the first three CYP families. The 23 sequences were saved in a FASTA file and submitted to all tools. With T-COFFEE, the alignment was produced using default parameters and the alignment from MAFFT was produced using the E-INS-i algorithm. To produce the alignment with PROMALS3D, CYP enzyme PDB IDs were added, to include structural information for the alignment. The most accurate alignment among the three tools was

achieved using T-COFFEE, due to its capability in aligning residues effectively, with a particular emphasis on preserving known conserved functional residues, thus, MSA on individual CYP families was then carried out using T-COFFEE. A FASTA file containing CYP 1 sequences only was created and submitted to T-COFFEE, with an alignment being produced using the default parameters. This was replicated for CYP families 2 and 3 to produce three alignments for the individual CYP families. A total of four alignments were produced in this analysis and they were viewed in Jalview [159]. The conserved regions were observed using the Jalview Clustal X Colour Scheme.

2.2.3 Phylogenetic analysis

Construction of phylogenetic trees for CYP families 1, 2 and 3 was performed using Molecular Evolutionary Genetic Analysis (MEGA) vX tool [160, 161]. The T-COFFEE sequence alignment result with all three families aligned together was used, and the maximum likelihood statistical methods were employed to deduce evolutionary relationships. Best-fit substitution models for the data were initially searched for, using a Neighbor-joining tree and a strong branch swap filter. Gaps/missing data treatments used were complete deletion and partial deletion (with site coverage cutoff percentages set to 95% and 90%). The best three models from each of the three different missing data treatments used were chosen based on the lowest Akaike information criterion (AIC) scores and Bayesian information criterion (BIC) (Table S3). The nine models were then used to construct nine phylogenetic trees, using the bootstrap method as the test of phylogeny and 1000 bootstrap replications. The resultant trees were compared to their respective bootstrap consensus trees to ensure branching patterns were consistent and accurate. The tree observed to

have the best branching patterns, that were more accurate and consistent than the rest, was chosen, and this was a tree from the 100% gap deletion data treatment, with the model being Le Gascuel (LG) model with Gamma Distributed with invariant sites (G+I) 27. The bootstrap tree for the phylogenetic tree chosen for this study is shown in Figure S2. Statistical support provided by the bootstrap values indicated that observed phylogenetic branches were reliable. The branches had a bootstrap value over 70, and only seven were below this number, with the lowest being 33. Considering the other trees constructed, this tree exhibited a more reliable set of bootstrap values and branching patterns. An all vs all sequence heatmap was then produced using Python and MATLAB scripting languages to further display the sequence relationships shown on the phylogenetic tree.

2.2.4 Motif discovery and mapping

A FASTA file containing 23 unaligned CYP sequences from families 1, 2, and 3 was analyzed using MEME v4.11 to identify motifs within these families [147]. The analysis used default settings with a few adjustments: motif width was restricted to 6-20 residues, at least 100 motifs were required to be identified, and motifs were allowed to appear multiple times within each sequence. Pearson correlation calculations within the MEME suite were used to detect recurring motifs. The identified motifs were then mapped to representative structures from each CYP family using PyMOL [162], and structures sourced from the PDB database (PDB IDs: CYP 1-2HI4, CYP 2-4GQS, CYP 3-1TQN) were used [100].

2.2.5 CPY2D6 variant selection and analysis with motifs identified

The Pharmacogene Variation Consortium (PharmVar) [48], which provides variation information for CYP enzymes, was utilized to compile a list of CYP2D6 alleles. The alleles with clinical functions assigned by the Clinical Pharmacogenetics Implementation Consortium (CPIC) were selected and this resulted in a set of 16 CYP2D6 alleles classified by CPIC as having either normal or reduced enzyme activity. Ten Afrocentric alleles recently deposited in PharmVar with no clinically assigned activity were also compiled. A full list of all the alleles is provided in Table 1-2 (see chapter 1). Their variation information was extracted from PharmVar and they were then mapped to the motifs identified in this section, to determine if they occurred in any of them.

2.3 Results and Discussions

The first three families within cytochrome P450 superfamily specify human cytochrome P450 enzymes that metabolize ~70-80% of clinical drugs on the market thus far, and although enzymes within these three families are all closely related, it was important to characterize the enzymes, to better understand similarities that result in shared function and possibly highlight distinguishable factors within each group that make them unique. Here, we characterize human CYP enzymes from families 1, 2 and 3, using MSA, motif and phylogenetic analysis tools.

2.3.1 Multiple sequence analysis

We retrieved the curated protein sequences for human CYPs of families, 1, 2 and 3 from UniProt [136], with the dataset including three CYP1, 16 CYP2, and four CYP3 family sequences (Table 2-1 and Table S2). Multiple sequence alignment of the 23 CYP sequences was conducted using MAFFT, T-COFFEE, and PROMALS3D, and the T-COFFEE alignment, that best preserved residues especially those of functional significance, was selected (Figure 2-2). The alignment of the four highly conserved regions was included in the assessment of the accuracy of the MSA, and overall, the alignment of all 23 sequences with T-COFFEE was effective in aligning the important residues in these regions.

Table 2-1: Human Cytochrome P450 enzyme families 1, 2 and 3 UniProt sequence attributes.

Enzyme family	Enzyme/Gene name	Sequence length [aa]	UniProt sequence accession number
CYP 1	CYP 1A1	512	P04798
	CYP 1A2	516	P05177
	CYP 1B1	543	Q16678
CYP 2	CYP 2A6	494	P11509
	CYP 2A7	494	P20853
	CYP 2A13	494	Q16696
	CYP 2B6	491	P20813
	CYP 2C8	490	P10632
	CYP 2C9	490	P11712
	CYP 2C18	490	P33260
	CYP 2C19	490	P33261
	CYP 2D6	497	P10635
	CYP 2E1	493	P05181
	CYP 2F1	491	P24903
	CYP 2J2	502	P51589
	CYP 2R1	501	Q6VVX0
	CYP 2S1	504	Q96SQ9
	CYP 2U1	544	Q7Z449
	CYP2W1	490	Q8TAV3
CYP 3	CYP 3A4	503	P08684
	CYP 3A5	502	P20815
	CYP 3A7	503	P24462
	CYP 3A43	503	Q9HB55

[aa] Amino acid sequence

Upon observing the T-COFFEE alignment of all 23 human CYP sequences in Figure 2-2, it is evident the conservation throughout the sequences is low, with most conservation focused on the AGXXT, PPGPXPXP, EXXR and the FXXGXXXCXG highly conserved domains, and residues in the B, J and K'' helices. More sequence variations were noted in the alignment with all three families together, indicating a level of uniqueness present. It has been defined that CYP enzymes in the superfamily with more than 40% sequence identity are grouped into the same families, and those with more than 55% sequence identity are assigned into the same subfamily [35]. Therefore, the higher diversity observed with the alignment of the sequences from CYP 1, 2 and 3 agrees with this notion, as there was a higher number of families included in the alignment, meaning the sequences share less than 40% sequence identity, and at least 60% differing sequence regions.

A separate MSA was carried out on each of the first three CYP families and can be observed in Figure S1. In the alignments for CYP 1 and 3 families, there was high conservation observed, however, in the CYP 2 family alignment, the conservation throughout sequences in the family decreased. The difference in conservation with the CYP2 family alignment may be attributed to the different subfamilies present in analysis in the family, and as stated earlier, this meant the sequences in CYP2 family only share at least 40% sequence identity, whereas CYP1 and CYP3, the percentage identity may be higher. Although there was less conservation observed within the CYP2 family than the other two families, it was higher than what was observed in the alignment with all three families combined. This highlighted that CYP sequences within different subfamilies

are more closely related as they share a higher sequence identity, indicating a high degree of functional similarity.

Overall, MSA revealed that there are regions of high conservation throughout the three CYP families altogether, however, there is some uniqueness to each CYP family, that may be related to the type of drugs they metabolize, as different CYP enzymes do metabolize different drugs [163–165].

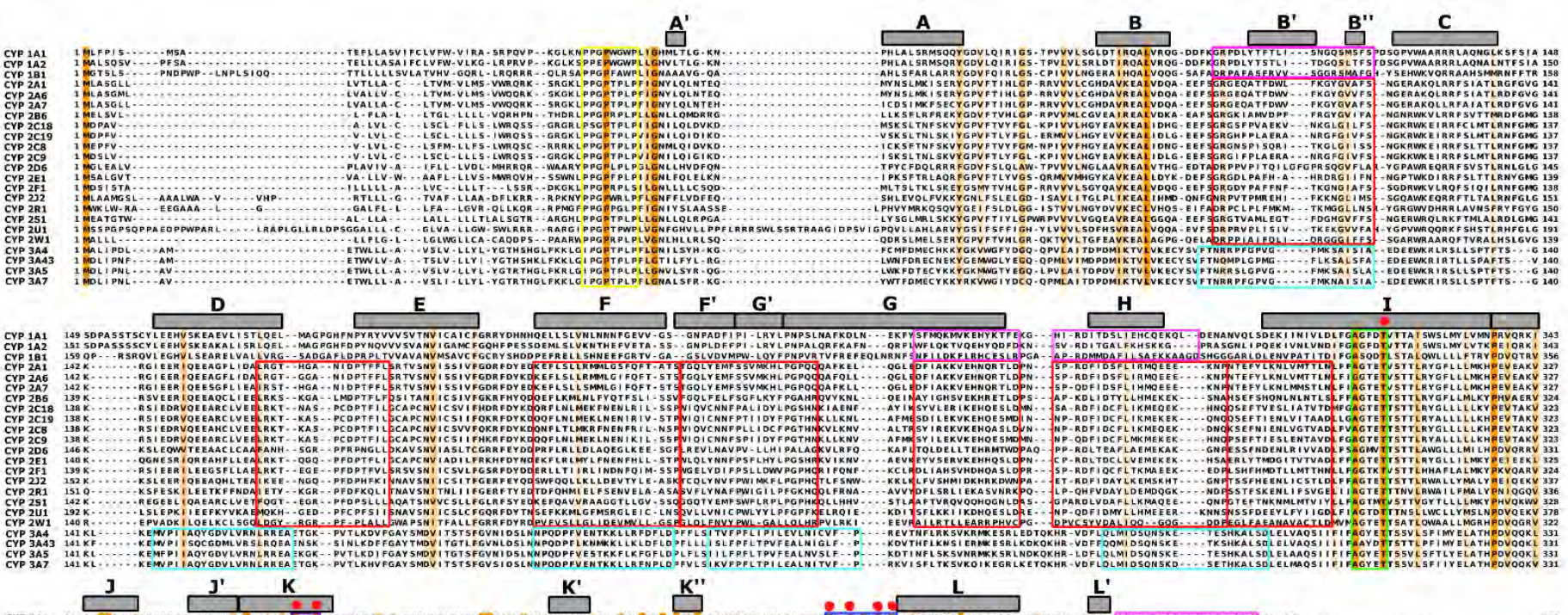


Figure 2-2: Multiple sequence alignment of CYP enzymes in families 1, 2 and 3. The multiple sequence alignment of CYP sequences was produced by T-COFFEE alignment program and results viewed in Jalview. The conserved regions were observed using the Jalview Clustal X Colour Scheme. The PPGPXPXP is highlighted by a yellow box. The (A/G)XXT region is denoted with a green box, while the EXXR region is shown in the purple box. The FXXGXXXCXG conserved domains is in a blue box. The red dots depict the most conserved and important catalytic residues in CYP enzymes. The unique regions (as identified by motif analysis) to CYP 1 family denoted by pink boxes, while the unique regions found in CYP 2 family in red boxes. CYP 3 family unique regions are marked by cyan boxes. Helix regions for the different helices within CYP enzymes shown by grey bars, and CYP 2C8 enzyme sequence was used to map the helix regions according to [58] findings.

2.3.2 Phylogenetic analysis

We further analyzed the sequences via phylogenetic tree calculations using Molecular Evolutionary Genetic Analysis (MEGA) vX tool [160] (Figure 2-3A), and all versus all sequence identity calculations using Python and MATLAB scripts to further display the sequence relationships shown on the phylogenetic tree (Figure 2-3B) [166, 167]. In the phylogenetic tree calculations, nine models were used to construct nine phylogenetic trees using the bootstrap method as the test of phylogeny, and the tree that was chosen for further analysis had the most consistent and accurate branching patterns observed when comparing to its respective bootstrap consensus tree. The bootstrap tree for the phylogenetic tree chosen for this study is shown in Figure S2.

On conducting analysis, as expected, the CYP families clustered together, and as seen in the sequence identity heatmap, the highest sequence identity was observed within the 2A, 2C, 3A, and 1A sub-families (mostly red boxes observed). This indicated higher relationships and shared ancestry within families, and more so within subfamilies, of CYP enzymes within the first three human CYP families. These results revealed that the observed high conservation within families and subfamilies, as seen with MSA analysis, is attributed to the shared ancestry of sequences from similar origins. Interestingly, from the phylogenetic tree, CYP 2 family was divided into two main groups, with CYP 2A, B, C, E, F and S forming one group, and CYP 2D, W, J, R and U forming the second group. This observation was confirmed by the pairwise sequence identity heat map presented in Figure 2-3B. The CYP 2A, B, C, E, F and S sub-families clustering together indicated that they share higher sequence identity with each other (as seen by the red, orange, and cyan

blocks in the top-left corner of the heat map) and are more closely related to each other as compared to the CYP 2D, W, J, R sub-families (with mostly blue boxes in the heatmap). Overall, the phylogenetic analysis demonstrated the close relationships and shared ancestry among CYP enzymes within each family, with even closer relationships observed within subfamilies. These relationships are closely linked to the conservation patterns revealed in the MSA analysis.

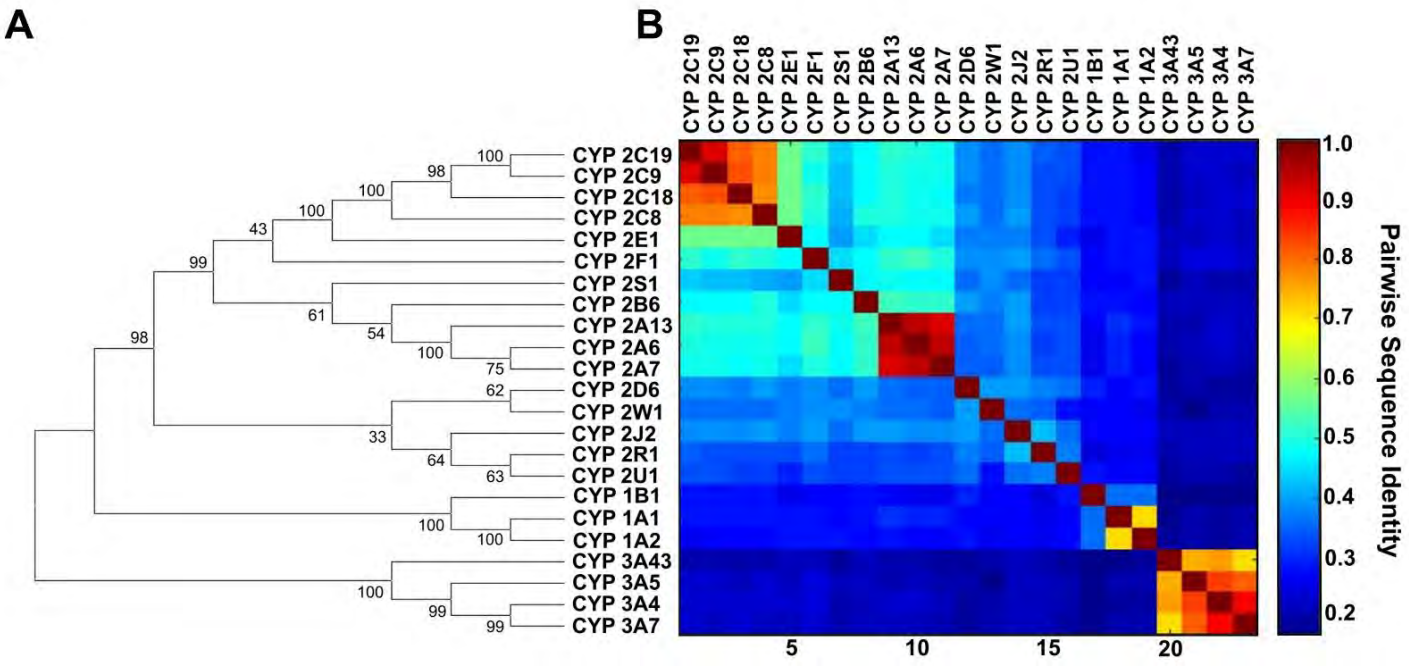


Figure 2-3: Sequence analysis of CYP 1, 2 and 3 families. A) Phylogenetic tree produced using MEGAX and the maximum likelihood method was applied using the LG+G+I protein model with 100% gap deletion. The bootstrap values on branches are percentages of 1000 bootstraps on the branch nodes. B) All versus all pairwise sequence alignment results for CYP 1, 2, 3 family sequences presented as a heat map produced using Python and MATLAB. Conservation values in heat map is increasing from blue to red.

2.3.3 Motif analysis

Next, to uncover regions of important functionality within human CYP families 1, 2 and 3, motif analysis was carried out. 23 curated sequences from CYP families 1, 2 and 3 as seen in Table S2, were submitted to MEME and 100 motifs were identified (Figure S3), but only 31 significant motifs with E-values less than 0.1 were retained and selected for further analysis in this study (Table 2-2). All the 31 motifs with their positions in each CYP enzyme are outlined in Table S4. In Figure 2-4, we represented a heatmap with conservation of the 31 significant motifs decreasing across all CYP families and is reflected in the colour change where motif conservation decreases from red to blue. It was observed that there were 11 motifs conserved in all CYP families, and motif 12 and 14 occurring in all except CYP3A5, and CYP1B1 and CYP2W1 respectively. We also observed motifs that were less conserved, and more unique to each family, with CYP1 family of enzymes having the least number of unique motifs identified.

Table 2-2: Regular expressions and E-values of significant motifs identified in CYP families 1, 2 and 3 and used for further analysis.

Motif number	E-Value	Motif regular expressions
1	4.6e-236	PF[SG][AI]GKRNC[IL]GE[GR]LARMELF
2	7.6e-204	[HY]P[ED]VQA[KR]V[QH]EE[IL][DE]RV[IL]GRNR
3	1.1e-203	DRA[HQK][ML]PY[TL][DE]AV[IV]HE[VIT][QL]R[FL][GIS]D
4	1.1e-179	[LH][HR]DPK[EY][WF]PNPEEF[ND]PE[HR]FLD
5	1.3e-175	L[FI][FG]AGT[ED]T[TV]STTL[RS][YW][GA][LF]LL[LM]
6	3.2e-171	CA[VP][SC]N[VI]ICSI[LV]FGKRFDYK[DN]
7	2.7e-162	[HR]A[VT][TK][KR]D[TV][KE][FL][RN]G[YF][FL][IL]PKGT[TE][VI]
8	8.9e-161	[RK]GKLPPGPTPLP[FL][IL]GN[IL]LQL

9	9.7e-151	[VL]V[LI][HT][GD]Y[DE]A[VI][KR]EAL[VI]DQ[GA]EEF
10	2.0e-142	[NS]G[EP]RWK[EQ][IL]RR[FL][SA]LX[TA][LF]R[NS]FG
11	2.4e-124	[MT][REK][LF]S[KE][KRV]YG[PD]VFT[ILV][YH]LGP[QR][PR]V
12	3.5e-107	[MS]GKRS[IL]E[ED]R[IV]QEEA[KR][CF]L[VI][ED][AE]
13	2.9e-105	[SN]P[RQ]D[FL]ID[CA][FY]L[IL][KE]ME[KEQ]E[KE]KN[PQ]
14	4.8e-094	LF[LF]TT[ILV][LM]QNF[SN][FL]K[SP]PX[DE]PKD
15	9.1e-078	[WT][GI]QL[YC][NE][NM]F[PS]S[FILV][ILM][DK][YH][LF]PG[PT][HQ][QN]
16	2.2e-055	GR[GP]XX[PA][TL]F[ED]X[IV][FT][KR]GYG[IV]VFS
17	4.7e-046	[VT][AVW]L[LV][LA][CA][LV][ST][LVC][LM][LV]L[LMY][SL][LVY][WG][RT][QR][SR][HKS]
18	1.8e-045	[TS]P[KV][HV][NV]GF[AG][ST][IV]P[PR]NY[QT][LM][CS]F[IL]P
19	4.2e-042	[SD][FY]I[AL][KE]KV[KE][EH][HN]Q[ER][ST]LD
20	6.2e-042	LR[KG]T[KH][GA][AS][PN][CIF]DPTF[IFL]L
21	3.4e-038	NG[QN]FKKS[ED][AY]F[MV]
22	2.3e-028	Q[LQ]MIDSQNSK[ED][TS][EK][ST]HKALSD
23	1.0e-023	M[VF]PII[AS]Q[YC]GD[VM]LVR[NS]LR[RQ]EA
24	1.5e-020	[QK]EFL[NS]L[ML][ER][LKM][FM][NL][EG]NF[QR][IFL][LT]S[ST][PS]
25	7.3e-019	N[ST][ES]F[TY][IL][EK]NL[VI]M[TS][TV][LA][DN]
26	9.8e-016	FTN[RQ][RM][PS][FL]GP[VM]GF[ML]K[SN]A[IL]S[IFL]A
27	1.7e-013	NPQDPF[VL][EK][NS][TM]KK[LF]L[KR][FL][DGN][FP]LD
28	3.9e-011	L[QL][PT]EKP[IV]VLK[VA][EDH][SL]RD[GE][TI][VLT]SG
29	3.9e-010	[NK]KD[NS]ID[PL]Y[IR]Y[TI]
30	7.3e-007	I[IKST][LV]FPFL[TI]P[IV][FL]E[AV]LN[IV][CGST][LV]FP
31	1.5e-005	D[LM]TPIYGLTMKHA[CR]CEH[FV]Q

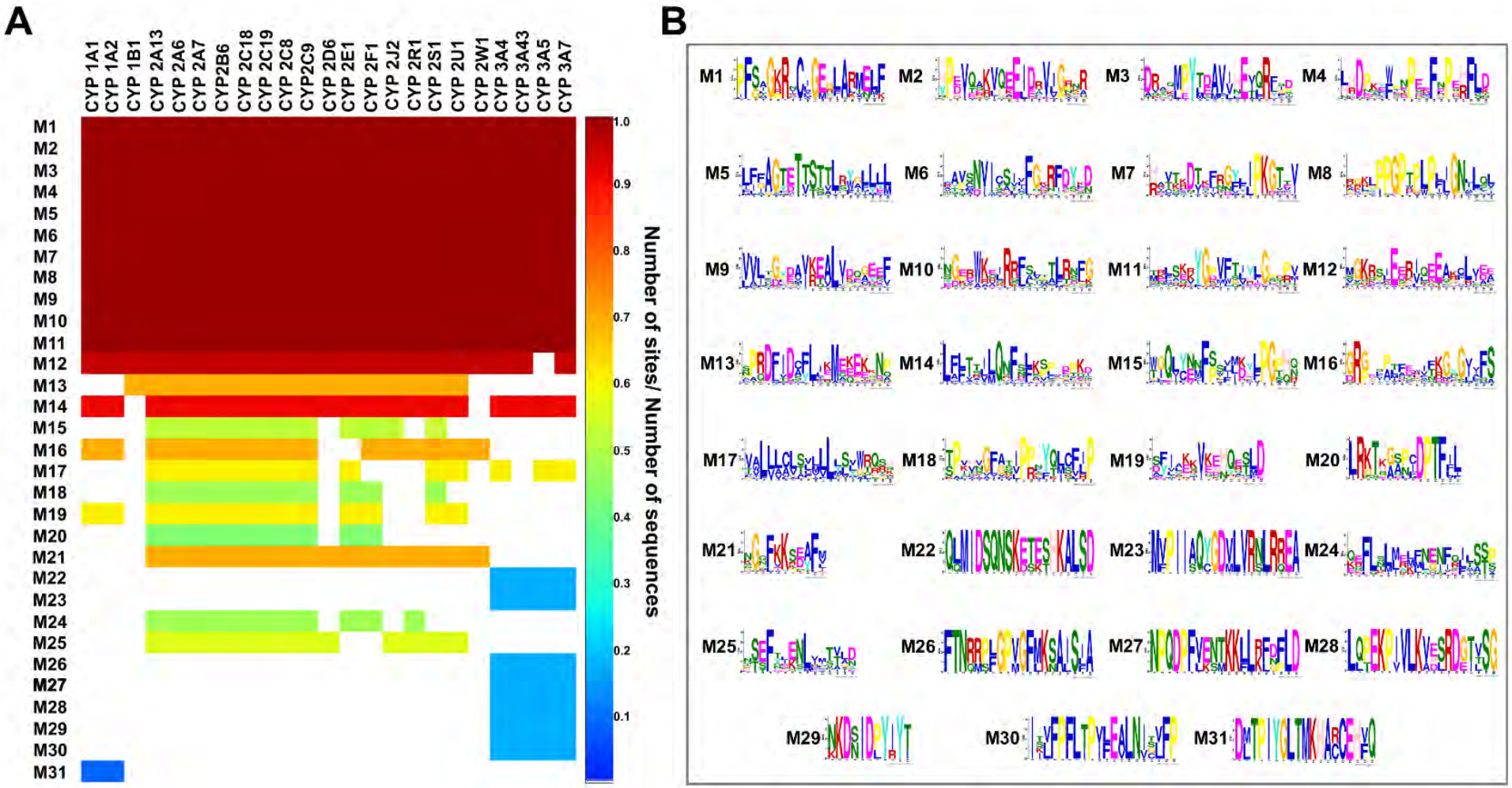


Figure 2-4: Motif analysis. A) Heat map showing the frequency of each significant motif in the CYP 1, 2 and 3 family sequences sampled. Matplotlib and in-house Python scripts were used to create heat map and conservation of motifs across all CYP sequences increased from blue to red. Numbering of motifs and their respective logos was based on the results generated by MEME. B) Motif logos showing amino acid conservation of each residue within each motif.

The conserved motifs identified to be present in all CYP families were mapped to a CYP2D6 structure using PyMOL as seen in Figure 2-5A. The conserved motifs mostly mapped to the pre-A loop, A', A, B, C, D, E, C-terminal half of I, J', J, K, K'-K'' loop and L helices. They also mapped partly or in all residues of β1-1, β1-2, β1-3, β2-1, β2-2 and β3-3. Previous research on CYP enzyme families has shown that the most conserved regions in the superfamily are located within the pre-A loop and the A, I, K and L helices, which are known to either be in direct contact with the heme group present in all CYP enzymes, or partake in the structure formation and stability [51, 53, 59, 67, 168]. Motif 1 ("PF[SG][AI]GKRNC[IL]GE[GR]LARMELF") partially overlapped onto helix L, and contained the FXXGXXXCXG, one of the four most conserved regions mentioned earlier (Chapter 1), revealed in literature as a signature for CYP enzymes and a heme-binding domain. Motif 3 ("DRA[HQK][ML]PY[TL][DE]AV[IV]HE[VIT][QL]R[FL][GIS]D"), mapping to the K helix, motif 5 ("L[FI][FG]AGT[ED]T[TV]STTL[RS][YW][GA][LF]LL[LM]"), mapping to the C-terminal side of the I helix, and motif 8 ("[RK]GKLPPGPTPLP[FL][IL]GN[IL]LQL") mapping to the pre-A loop contained the other three highly conserved regions; EXXR, (A/G)XXXT and PPGPXPXP respectively. Additional to these known conserved regions, our motif analysis has revealed other conserved regions within CYP enzymes not reported before, and as conserved motifs are a sign of shared ancestry among a group of sequences for enzymes that share very similar functionality, and highlight important regions for functionality within enzymes, we determine these conserved regions to be highly important for functionality of the CYP enzymes.

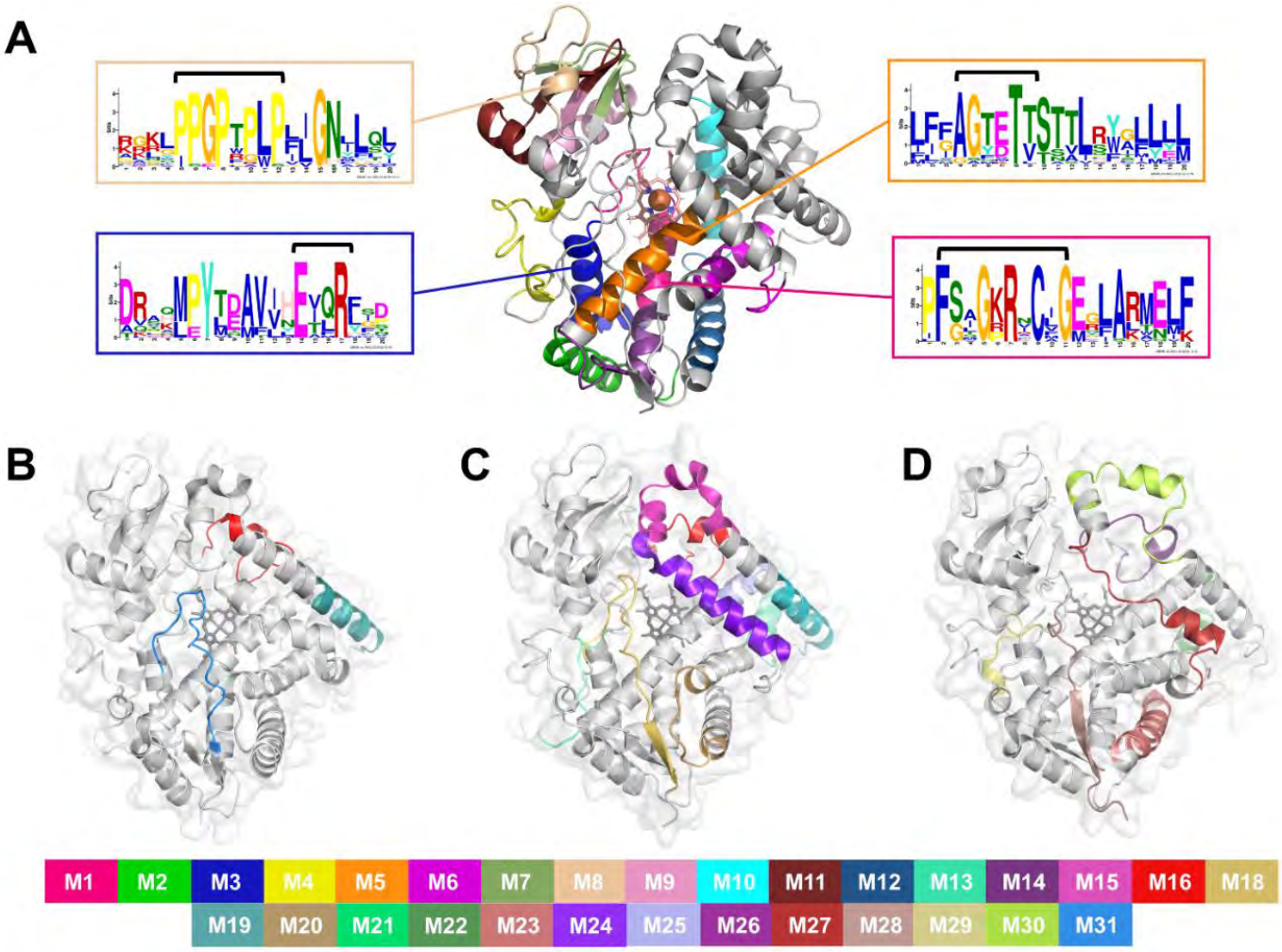


Figure 2-5: Motifs identified mapped to CYP structures. A) Conserved motifs mapped to CYP2D6 structure (PDB ID 3TDA) with highly conserved regions (AGXXT, PPGPXPXP, EXXR and the FXXGXXXCXG domains) highlighted by their motif logos and their location within the motif shown with a black bracket. B), C) and D) are unique motifs mapped to the representative CYP enzyme family structures. B) CYP 1 unique motifs 16, 19 and 31; C) CYP 2 unique motifs 13, 15, 16, 18, 19, 20, 21, 24 and 25; D) CYP 3 unique motifs 22, 23, 26, 27, 28, 29 and 30. PDB structures 2HI4, 4GQS and 1TQN are used respectively as representatives for CYP 1, 2 and 3 families. Motif 17 unique to CYP2 and CYP3 families could not be mapped to structures due to the structures not containing this region.

Identification of motifs unique to each enzyme can provide more insights into the key regions that specify its enzymatic function. Here, identified unique motifs to each CYP family were mapped to their sequence regions in the MSA conducted earlier as seen in Figure 2-2, and to their representative structures (CYP 1- 2HI4 [169], CYP 2- 4GQS [63], CYP 3- 1TQN [170]) using PyMOL, as seen in Figure 2-5B-D. The motifs identified to be unique to each of the three CYP families were noted to be within the varying regions identified in MSA. The conservation in these regions was low. Structurally, motifs unique to CYP 1 family were mapping mostly to the B' helix, B-B' and B'-C loops (motif 16), part of the G helix (motif 19), and β4-1 and β4-2 sheet region (motif 31). The unique motifs for CYP 2 family (motifs 13, 15, 16, 18, 19, 20, 21, 24 and 25) were mostly mapping to the B', D, F, F', G, H and N-terminal half of I helices, the B-B', B'-C, D-E, E-F, G-H, H-I and K''-L loops, and the β4-1 and β4-2 sheet regions. Motifs identified uniquely for CYP 3 family (motifs 22, 23, 26, 27, 28, 29 and 30) were mapping mostly on the B', D, F, F', G, G', H and K'' helices, the G'-G and K''-L loops and the β3-2, β4-1 and β4-2 sheet regions. It has been discerned that within the CYP superfamily, the variable regions are within the B, C, F, G, L and N-terminal half of the I helix regions, and these areas also control substrate access and specificity within each CYP enzyme [58, 67, 68, 132, 171]. CYP enzymes can be specific in the type of xenobiotic that they metabolize [163, 172] however, some CYP enzymes may metabolize an array of drugs [163, 172], and the flexibility in the types of drugs the CYP enzymes metabolize is accredited to the differences in the regions mentioned above [171]. It was interesting to note that the enzyme of focus in this thesis, CYP2D6, a member of the CYP2 family, only had 3 unique motifs identified (motifs 13, 21 and 25). They were mapped to the H helix, the G-H, H-I and K''-L loop regions, and the N-terminal end of the I helix, that contain substrate recognition sites 4 and 5. Motifs 15, 16, 18, 19, 20 and 24, unique to most CYP 2 family enzymes, were not observed in

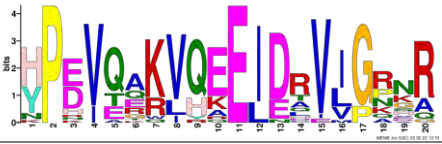
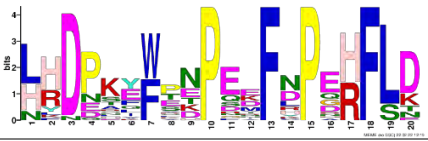
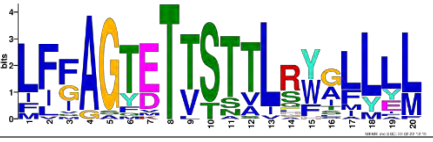
CYP2D6. This shows less variation in these regions for CYP2D6 and may be the reason for the high number of substrates CYP2D6 is able to metabolize unlike other CYP enzymes (CYP2D6 metabolizes ~25% of clinical drugs [125–127]). From these results, we can deduce that the motif identification calculation has identified the highly conserved regions for structure and function in CYP enzymes within CYP families 1, 2 and 3, however, it also highlighted unique varying regions specific to each family that may control substrate recognition and access.

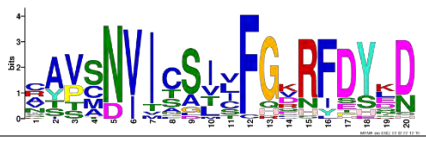
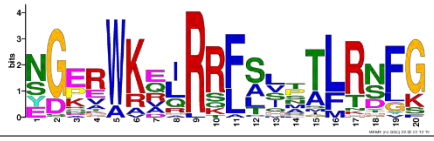
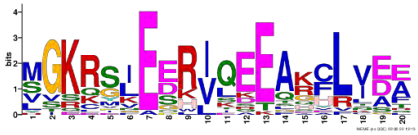
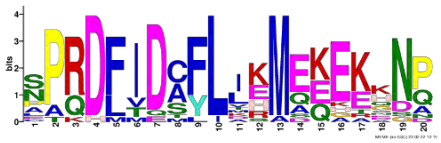
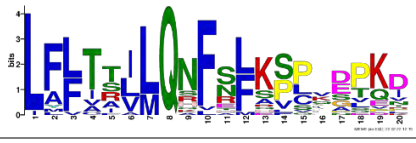
2.3.4 Motif and variation analysis

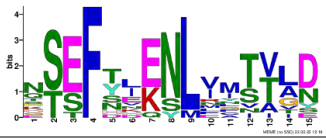
Cytochrome P450 2D6 is the main enzyme of focus in this thesis, mainly because of its highly polymorphic characteristic and poorly understood allele behavior. Therefore, it was of our interest in this chapter to observe whether the variations making up CYP2D6 alleles (Table 1-2) were occurring in any of the motifs identified. This was carried out by identifying the variation residue number in the allele, then checking if this number was within the range of residue numbers making up a motif. Table 2-3 shows all conserved and unique motifs found in CYP2D6, with the motif regular expressions, web logos (which are representations of the amino acid sequence within each motif, and the conservation of each amino acid in that motif sequence), residue numbers making up the motif, and variations occurring within the motif (location of all variations in motifs were also shown in Figure 2-6B). For Table 2-3, it is important to note that some variations mapped to motifs were different amino acids to those that appear in the regular expressions of the motifs, and this is because at each position within the motif, there may be more than one type of amino acid found in all sequences belonging to CYP 1, 2 and 3 family enzymes, and conservations of the different types identified at each position may differ. Thus, web logos were provided, as they list

the different amino acids that occupy each position in a motif, and their conservation (the bigger the letter in the logo, the more conserved that residue is across all sequences) so if they do not appear on the regular expression, they can be identified in this web logo. Such variations in Table 2-3 are H94R, D97E, V136I and R296C, K404Q and E410K. They contain amino acids not appearing in the regular expressions of their respective motifs, meaning these residues at these positions are less conserved and differ to those that appear in most of the other CYP enzymes in families 1, 2 and 3. The differences in amino acids at these positions in CYP2D6 could be an additional feature to the enzymes uniquenss.

Table 2-3: Conserved and unique motifs identified to be occurring in CYP2D6, with their regular expressions, web logos, motif residue range and all variations occurring in each motif.

Motif number	Motif regular expressions	Web logo	Motif residue numbers	Variation in motif
1	PF[SG][AI]GK RNC[IL]GE[GR]LARMELF		435 – 454	R441H, M451I
2	[HY]P[ED]VQ A[KR]V[QH]E E[IL][DE]RV[I L]GRNR		324 – 343	V338M
3	DRA[HQK][ML]PY[TL][DE]A V[IV]HE[VIT][QL]R[FL][GIS] D		349 – 368	-
4	[LH][HR]DPK[EY][WF]PNPE EF[ND]PE[HR] FLD		403 – 422	K404Q ;E410K
5	L[FI][FG]AGT[ED]T[TV]STTL [RS][YW][GA][LF]LL[LM]		302 – 321	-

6	CA[VP][SC]N[VI]ICSI[LV]FGKRFDYK[DN]		180 – 199	-
7	[HR]A[VT][TK][KR]D[TV][KE][FL][RN]G[YF][FL][IL]PKGT[TE][VI]		376 – 395	-
8	[RK]GKLPPGPTPLP[FL][IL]GN[IL]LQL		30 – 49	P34S, P41L
9	[VL]V[LI][HT][GD]Y[DE]A[VI][KR]EAL[VI]DQ[GA]EEF		79 – 98	A90V, L91M, H94R, D97E
10	[NS]G[EP]RWK[EQ][IL]RR[FL][SA]LX[TA][LF]R[NS]FG		124 – 143	W128R, V136I
11	[MT][REK][LF]S[KE][KRV]YG[PD]VFT[ILV][YH]LGP[QR][PR]V		59 – 78	-
12	[MS]GKRS[IL]E[ED]R[IV]QEEA[KR][CF]L[VI][ED][AE]		144 – 163	E155K; E156A
13	[SN]P[RQ]D[FL]ID[CA][FY]L[IL][KE]ME[KEQ]E[KE]KN[PQ]		267 – 286	-
14	LF[LF]TT[ILV][LM]QNF[SN][FL]K[SP]PX[DE]PKD		455 – 474	-
21	NG[QN]FKKS[ED][AY]F[MV]		424 – 434	-

25	N[ST][ES]F[TY][IL][EK]NL[VI]M[TS][TV][LA][DN]		287 – 301	R296C

Red indicates conserved motifs and green indicates unique motifs to CYP2D6.
Residues underlined and coloured in red in regular expression is the position the variation is located in the motif.

Variations P34S, P41L, A90V, L91M, H94R, D97E, W128R, V136I, E155K, E156A, V338M, K404Q, E410K, R441H and M451I were located in conserved motifs identified, and from these, it was noted that initial residues in variations P34S, P41L, A90V, L91M, W128R, E156A, V338M, R441H and M451I were highly conserved in their positions within the motifs. This could suggest substitutions at these positions to have detremental effects as highly conserved residues in motif positions indicate its importance in the correct functioning of the protein [173]. However, variations such as P34S, E156A and V338M occur in alleles *10, *50 and *29, which are known to have reduced activity [48, 49], therefore the variations in these highly conserved regions may not necessarily diminish activity, but reduce enzyme functionality. Additional to this observation, the valine residue in variation V338M was positioned in conserved motif 2 which is located inbetween motif 3 and motif 5 that harbour the important EXXR and (A/G)XXXT conserved regions. Residues arginine and methionine in variations R441H and M451I respectively were located within motif 1 with the important FXXGXXXCXG conserved region, and a few residues from this motif, the lysine and glutamate residues in variations K404Q and E410K respectively, were occuring in motif 4. The occurrence of these variations near the highly conserved regions important for structure formation and functioning of the enzyme, suggests that they may affect how the protein behaves. The prolines in variations P34S and P41L are in conserved motif 8 and are the first and eighth residues in the proline rich PPGPXPXP region found within this motif. As

mentioned earlier in this thesis the 'PPGPXPXP' region near the N-terminus of CYP2D6 is crucial for the enzyme's correct folding and membrane anchoring, therefore the substitution of these prolines in this region could greatly affect the protein structure (section 3.3.3.1), and binding to the membrane [52–55].

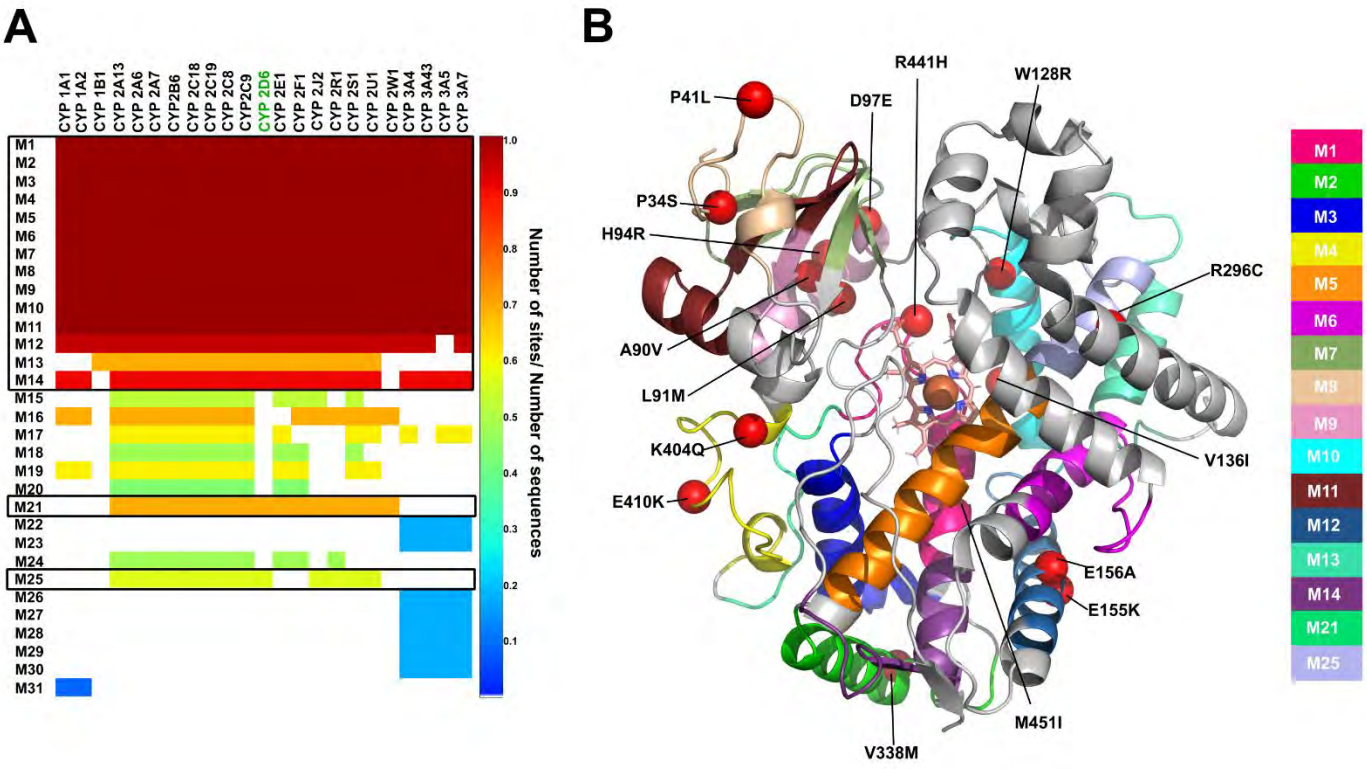


Figure 2-6: Motif and variation analysis. A) Motifs identified within CYP families 1, 2 and 3 sequence dataset. Matplotlib and in-house Python script were used to create heat map. A total of 23 sequences observed to have 31 significant motifs discovered. Conservation of motifs across all CYP sequences increased from blue to red. Motifs occurring in CYP2D6 (highlighted in green) shown by a black box. B) All conserved and unique motifs specific to CYP2D6 mapped to its structure and location of CYP2D6 allele variations within motifs shown in the structure as red spheres.

Variations H94R and D97E in motif 9, as well as V136I in motif 10, had residues with low conservation at their respective positions within the motifs. This, combined with the fact that the

amino acid properties were retained in all three variations (i.e., changes did not alter the chemical properties at those positions—basic histidine was replaced by basic arginine at position 94, acidic aspartate by acidic glutamate at position 97, and non-polar valine by non-polar isoleucine at position 136), suggests that these variations are unlikely to have detrimental effects. Although variation A90V was mentioned earlier to have the alanine residue highly conserved in motif 9, the change in amino acids to a valine may also unlikely be detrimental as chemical property did not change. Conserved motif 12 contained glutamate residues in variations E155K and E156A. Though this motif has no known function, a substitution of the highly conserved glutamate residue to a smaller and non-polar alanine residue as dictated by the variation E156A, will increase hydrophobicity, and remove the ability of salt bridges being formed by glutamate in this region, which may indirectly alter the structure and functioning of CYP2D6. The change of a glutamate residue to a lysine in variation E155K, results in a change in overall charge of the ionized residue at this position, which may affect overall protein structure and function. Only variation R296C had the arginine residue occurring in a unique motif (motif 25), and this motif lies on the N terminal side of the I helix that has been established as a varying region in CYP2 families which aids in substrate access [132]. The location of this variation could suggest a possible effect on the substrate access in alleles with which it occurs in. This was confirmed to be true for alleles *14, *17, *29 and *55 (chapter 3), and alleles *155, *159, *160 and *162 (Chapter 4) where opening frequencies of substrate access channels within these alleles were altered. To better understand all the effects of these variations on CYP2D6, further investigations using *in silico* tools were utilized, and we report our findings in this thesis.

2.4 Conclusion

The primary pharmacogenes involved in drug metabolism in humans are the Cytochrome P450 enzymes. Among these, the first three families—CYP1, CYP2, and CYP3—play key roles in the phase I metabolism of most clinical drugs. Gaining a deeper understanding of the enzymes within these families could advance research in personalized medicine. Consequently, this chapter aimed to characterize the enzymes from the CYP 1, 2 and 3 families in both a sequential and structural context, using MSA, phylogenetic tree and motif analysis tools.

A total of 23 sequences that represented CYP enzymes from the first three families were obtained and utilized throughout the analyses, and some interesting information about these CYPs were revealed. Multiple sequence alignment and motif analysis highlighted important conserved regions in the pre-A loop, A', A, B, C, D, E, C-terminal half of I, J', J, K, K'-K'' loop and L helices, and also in beta sheets β1-1, β1-2, β1-3, β2-1, β2-2 and β3-3, necessary for enzyme function and structural integrity, while phylogenetic analysis portrayed a family-wise clustering of the sequences. Furthermore, unique motifs distinguishing the different families were identified – CYP1: in the B' helix, B-B' and B'-C loops, part of the G helix, and β4-1 and β4-2 sheet region; CYP2: in the B', D, F, F', G, H and N-terminal half of I helices, the B-B', B'-C, D-E, E-F, G-H, H-I and K''-L loops, and the β4-1 and β4-2 sheet regions; and CYP3: in the B', D, F, F', G, G', H and K'' helices, the G'-G and K''-L loops and the β3-2, β4-1 and β4-2 sheet regions. These unique motifs account for the specificity of drug types metabolized by these enzymes. Interestingly, the CYP enzyme of focus in this thesis, CYP2D6 from the CYP2 family, had only

three unique motifs identified (motifs 13, 21 and 25). These motifs mapped to the H helix, the G-H and H-I loop regions, and the N-terminal end of the I helix. This observation may clarify why CYP2D6 can metabolize a wider variety of clinical drugs than CYP enzymes from other families. Variations of CYP2D6 to be studied in this thesis were examined to see if they occurred within any of the identified motifs. Interestingly, many of these variations were found within the conserved motifs, and additionally, variations P34S, P41L, A90V, L91M , W128R, E156A, V338M, R441H and M451I were highly conserved in their positions within the motifs, suggesting that they might impact the structure and function of CYP2D6. From clinical data noted thus far, this is true for alleles *10, *50 and *29, with variations P34S, E156A and V338M respectively, as they harbor reduced functionality. However, to explore variation effects for the other alleles (while also confirming information from literature), the next chapter (Chapter 3), investigates the effects of these variations in functionally assigned CYP2D6 alleles, to possibly determine mechanisms for reduced functionality, then in Chapter 4 we examine variation effects in Afrocentric alleles, to identify if there are any behavioral patterns similar to that of the reduced activity alleles (application of mechanisms identified in Chapter 3).

Chapter 3

Unraveling Mechanisms of Variant-induced Reduced Activity in Cytochrome P450 2D6 Alleles

Chapter overview

In this chapter, several computational tools and approaches are utilized to investigate variation effects on CYP2D6 alleles with normal and reduced clinical function. From the different observations, different mechanisms that potentially result in reduced activity of CYP2D6 are deduced.

3.1 Introduction

As mentioned in Chapter 1, cytochrome P450 2D6 (CYP2D6) is an important enzyme belonging to the human CYP 2 family that metabolizes ~25% of clinical drugs on the market [125–127] , therefore, studies on this enzyme would greatly contribute towards precision medicine.

3.1.1 CYP2D6 genetic and structural information

The pharmacogene that codes for CYP2D6 enzyme is part of a supergene family that encodes the cytochrome P450 (CYP) enzymes within the human body, and is located on the long arm of chromosome 22 (22q13.1) (Figure 3-1) [174–177]. The CYP2D6 gene is on a gene cluster (CYP2D) that also contains two pseudogenes, CYP2D7 and CYP2D8, which are highly homologous and share over 90% similarity [176, 178]. It consists of nine exons with 4383 base pairs that code for a protein that has 497 amino acids [174–176] (Figure 3-1).

CYP2D6 structure is highly similar to other CYP enzymes within the CYP superfamily, comprising of an N and C-terminal, with helices denoted from A-L and 4 β-sheets (Figure 3-2A) [3]. The active sites of CYP2D6, though similarly located within the core of the protein like all other CYP enzymes, differs significantly in shape and size [3]. As seen in Figure 3-2B, the shape of CYP2D6 active site can be likened to a “right boot”, with the heel located just above a heme group, and an arch running over active site residues and variation site F120 [179, 180]. The heme group within this active site contains an iron at the center that is held in the protein by the S−Fe

bridge formed between a protein cysteine residue and the heme iron [3]. Its main function is to take up an O_2 molecule that can be utilized in the oxidation of substrates, a reaction CYP2D6 catalyzes [3, 95, 181, 182]. This reaction is biologically important in transforming xenobiotics entering the human body, into more polar compounds that can be excreted [3, 183, 184].

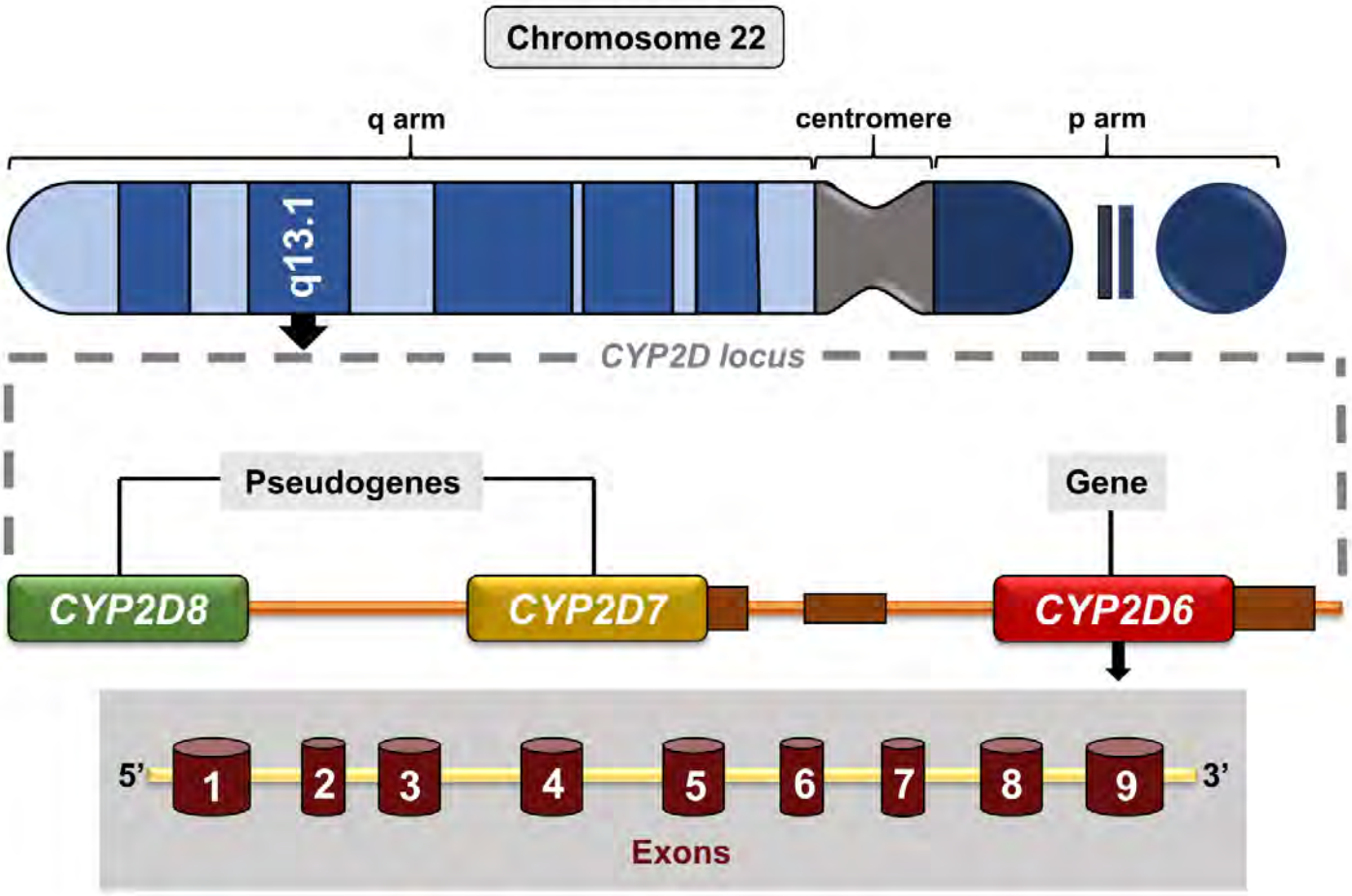


Figure 3-1: Location of the CYP2D6 gene on a human chromosome 22. Modified from Alvarado *et al*., 2021 [175].

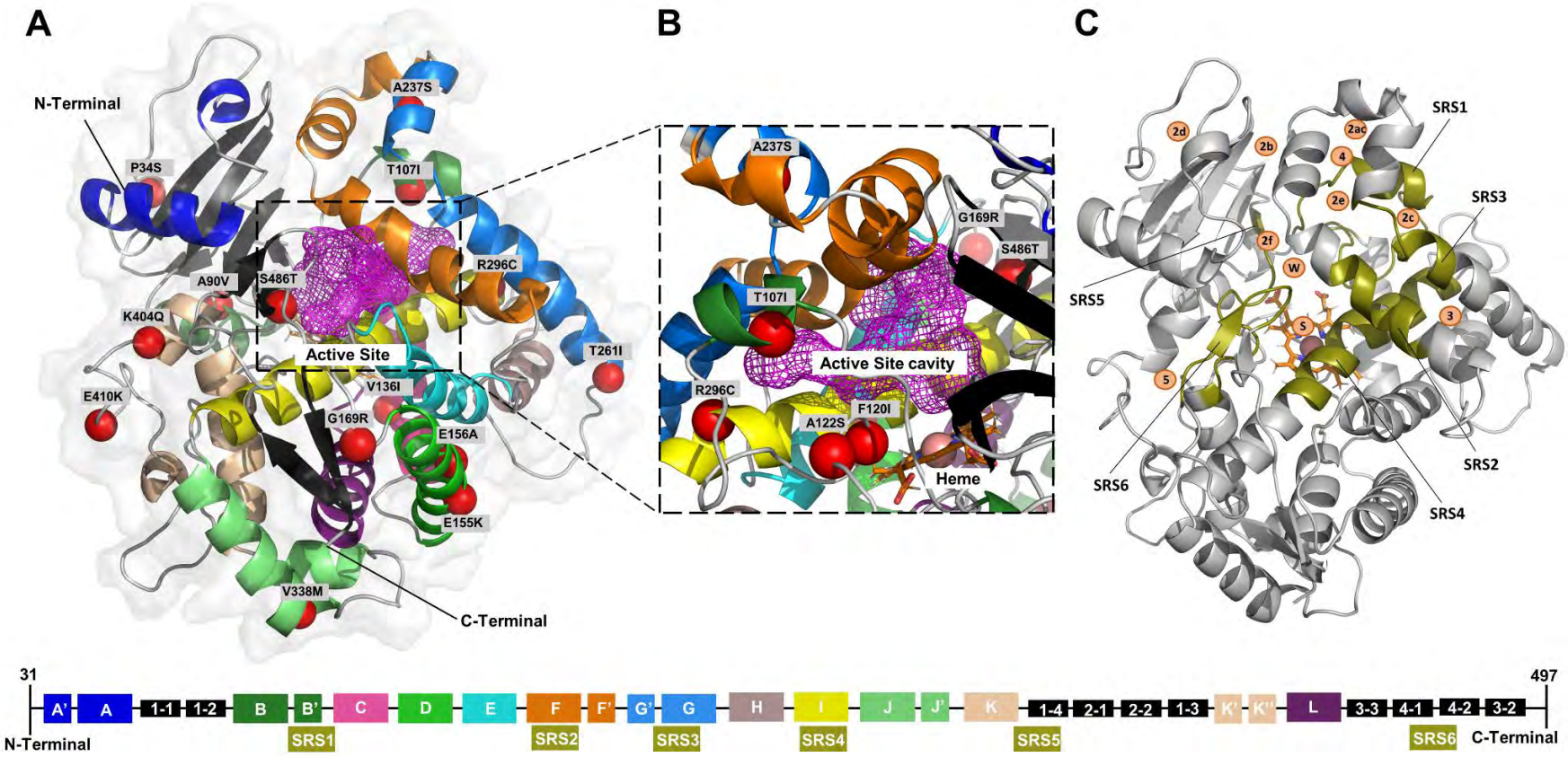


Figure 3-2: CYP2D6 structural characteristics. A) Structure of CYP2D6 with allele variations mapped and showing their positions in the different helices. Variations are shown as red spheres. Helices in CYP2D6 are denoted A-L and differently coloured in the structure. A colourmap showing the helix names and corresponding colours is at the bottom of the structures. β – sheets are coloured in black and the active site cavity is coloured in magenta. B) The active site region magnified and showing a "right boot" shaped cavity and heme (coloured orange) with an iron centre (coloured salmon). C) Location of SRSs in CYP2D6 structure coloured in deep olive and channels identified in literature to be present in CYP2D6 with and without membrane present labelled. The channel labels are at their entry sites in the structure.

CYP2D6, similar to other cytochrome P450 enzymes, possesses channels that that connect the buried active site with the external enzyme environment [61, 185]. Their entrance sites can be seen in Figure 3-2C. The more prominent channels—designated as 2b, 2f, and the solvent channel—have been identified as key pathways for the movement of substrates into and out of CYP2D6 [61, 185, 186]. These channels play a critical role in the enzyme's function and substrate accessibility to the catalytic site. Additional channels, such as 2c, 2d, 2e, 2ac, 3, 4, 5, and W, have also been identified. Channels 2c and 2e were observed in CYP2D6 in the absence of the membrane [186], while channels 2d, 2ac, 3, 4, 5, and W were detected in the presence of a membrane [187, 188]. However, all of these channels were considered too small to permit the passage of ligands into and out of the enzyme [61, 185, 186, 188]. Certain residues lining the channels in CYP2D6 have been shown to play a critical role in ligand selectivity, recognition, binding, and translocation to the active site [185]. Six regions containing these key residues have been identified in CYP2D6, and referred to as substrate recognition sites (SRSs) (Figure 3-2C) [132, 189]. SRS1 is situated on the B-B' loop, as well as the B' helix and the B'-C loop. SRS2 is located on the F-F' loop, while SRS3 is found on the G'-G loop and a portion of the G helix. SRS4 is positioned on the I helix, and SRS5 resides in the loop between the K helix and the β1-4 sheet. SRS6 is located on the β4-1 and β4-2 sheets [132, 189].

3.1.2 Polymorphism in CYP2D6

Unfortunately, CYP2D6 is a highly polymorphic enzyme, with over 140 “star” (*) alleles listed by the Pharmacogene Variation Consortium (PharmVar) [48, 190] and because of these genetic variations, the alleles can be classified as having either no function, reduced function,

normal function or increased function [48]. Depending on the combination of alleles an individual has (diplotype), CYP2D6 metabolizing phenotype can then be assigned as either poor (PM), intermediate (IM), extensive (EM, also normal activity of CYP2D6) or ultrarapid metabolizers (UM) [3, 46]. The location of variations studied in this chapter are shown in Figure 3-2A. With the exponential growth of identified CYP2D6 alleles that possibly have varying functional consequences, there is a need for the development of more efficient ways to assess CYP2D6 function, to avoid adverse drug effects or drug toxicities. Therefore, in this study, the main aim was to elucidate the mechanisms by which polymorphisms/variations impact CYP2D6, leading to reduced enzymatic activity. To achieve this aim, different *in silico* approaches, tools and databases were utilized. Pharmvar and PharmGKB were initially used to obtain variation and clinical information for alleles used in this research, then Molecular Dynamic (MD) simulations were performed on CYP2D6 reference and all alleles, to study their dynamic behavior and interactions at an atomic level, to provide insights into variation effects on CYP2D6, that can be related to their functionality. Different post-MD analyses were also performed, including Dynamic Residue Network (DRN) analysis, developed by the Research Unit in Bioinformatics (RUBi), which explores per-residue behavior during simulations to determine if any modifications to residue interaction networks occurred.

3.1.3 PharmVar and PharmGKB

The Pharmacogenomics Knowledgebase (PharmGKB) and the Pharmacogene Variation Consortium (PharmVar) are important resources for research conducted in the pharmacogenomics field of study. PharmGKB ensures the provision of information on how human genetic variations

affect drug responses, while PharmVar is a central repository that harbors haplotype structure and allele variation information.

3.1.3.1 PharmGKB

The knowledgebase was established in the year 2000 as part of the National Institutes of Health (NIH)-funded Pharmacogenetics Research Network (PGRN) [122]. It is a publicly available resource that was originally created to be a PGRN's genotype and phenotype data repository, however, over the years it has grown considerably to include gene-drug associations that are supported by literature which has been peer-reviewed [122]. PharmGKB is a collection of manually curated and disseminated knowledge on how human genetic variations affect patient response to medications, essentially providing a comprehensive resource on genotype-phenotype relationships. The most popular resource on this knowledgebase are the clinical annotations, that provide users with a summary of the association between drug-variant pairs. These annotations show related findings from curated peer-reviewed literature and are each assigned a level of evidence (LOE) that indicates the amount of literature which supports the annotation[122].

3.1.3.2 PharmVar

PharmVar is a consortium that was originally the Human Cytochrome P450 (CYP) Allele Nomenclature Database ("Nomenclature website" for short) before is transition. In the 1990s, a group of researchers came together to establish a systematic way of naming CYP2D6 alleles that were accumulating at the time [48]. They proposed a star (*) nomenclature system that was

essentially a success [48]. Due to the discovery and exponential growth of alleles from other CYP genes in the superfamily that metabolize different drugs, Ingelman-Sundberg *et al*., [191] proposed to apply the same star (*) nomenclature system for CYP2D6 gene to the other CYP genes. This then led to the establishment of the Nomenclature website in 2002, that harbored systematically categorized CYP alleles with this naming system, and since its inception, it had been under the leadership of Magnus Ingelman-Sundberg, and maintained by him and his team [48].

When Magnus Ingelman-Sundberg and colleagues retired, a committee was formed to maintain and sustain this resource, and thus PharmVar was established, and was to serve as a "Next-Generation" pharmacogene data repository [48]. This newly formed database contains allele information with a revised nomenclature system and standardized submission protocol and criteria for new alleles identified [48, 49]. It is widely used by researchers, PharmGKB and the Clinical Pharmacogenetic Implementation Consortium (CPIC). CPIC utilizes the allele designations as defined by PharmVar to gather peer-reviewed literature on the alleles so as to assign function that may lead to a phenotype assignment and optimize medication dosages [48, 122]. Researchers use this resource to gather information about alleles they are studying, and any findings will be relatable to the community at large with the systematic naming system. PharmGKB mainly utilizes PharmVar during the curation of data. Alleles from data or articles are mapped to PharmVar core alleles, to determine whether they do exist [48].

3.1.4 Metal ion parameterization and Molecular Dynamic simulations

Molecular Dynamic (MD) simulation is a computational method used to simulate molecular motion of many large biomolecules such as proteins or nucleic acids, through an iterative application of Newtons law of motion [192, 193]. This method models the behavior and interactions of atoms and molecules in a system within a three-dimensional (3D) space over time, in an effort to reproduce the natural behavior of a system. The interactions between atoms (including bond angles, lengths and dihedrals) in a system are usually described by molecular mechanical force fields, and many force fields have been developed to calculate the potential energy from positions of atoms or atomic coordinates [194–196]. Many of them calculate the potential energy using Equation 3-1 and Equation 3-2 for bonded (intramolecular, internal) and non-bonded (intermolecular, external) terms, respectively [197]. In Equation 3-1, the $K_{\phi,n}$ represents the dihedral amplitude, and the K_θ, K_b and K_φ symbolize force constants for the angle, bond, and improper dihedral respectively. The n and δ_n represent the dihedral multiplicity and dihedral phase, while the b_0, θ_0 and φ_0 indicate the bond length, angle and improper dihedral angle. In Equation 3-2, the D symbolizes the dielectric constant, and the $q_i q_j$ represent the partial charges. The *Rmin*, *ij* and ε_{ij} in the equation indicate the Lennard-Jones radius and well depth for calculation of the van der Waals interactions [194, 197]. Examples of force fields used in simulations include the Chemistry at HARvard Macromolecular Mechanics (CHARMM) [198], the Assisted Model Building with Energy Refinement (AMBER) [199, 200], the GROningen MOlecular Simulation (GROMOS) [201] and the Optimized Potential for Liquid Simulations (OPLS) [202].

$$E_{bonded} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{\substack{improper \\ dihedrals}} K_\varphi(\varphi - \varphi_0)^2$$

$$+ \sum_{dihedrals} \sum_{n=1}^{6} K_{\phi,n}(1 + \cos(n\phi - \delta_n))$$

Equation 3-1: Bonded (intramolecular, internal) terms.

$$E_{nonbonded} = \sum_{\substack{nonbonded \\ pairs\ ij}} \frac{q_i q_j}{4\pi D r_{ij}} + \sum_{\substack{nonbonded \\ pairs\ ij}} \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6} \right]$$

Equation 3-2: Nonbonded (intermolecular, external) terms.

Many enzymes contain different metals in their catalytic sites, such as Mg^{2+}, Zn^{2+}, Co^{2+} and Fe^{2+}, which are usually fixed in the enzyme core through a metal complex cofactor (such as a heme group) or can be coordinated by specific amino acid residues. These enzymes are generally called metalloenzymes, and they contain metal ions either to aid catalysis or maintain enzyme structure [203]. CYP2D6, like all other CYP enzymes in the superfamily, is a metalloenzyme, as it contains a metal ion in its prosthetic heme group. Specifically, it contains Fe^{2+} metal that is coordinated by a cysteine residue, which act as catalytic center where oxidation takes place during catalysis [3, 67, 72, 73]. AMBER force fields, such as the one utilized in this study, does not include parameters for metal ions and as such, this necessitates the generation of these parameters for simulations to

mimic natural biological system. In this research, the parameters for the penta-coordinate ferric heme state (with a charge of +2 for the iron) were generated (in RUBi) to provide an accurate and comprehensive description of the metal ion, its properties and interactions with residues surrounding it for MD simulations.

Generation of force field parameters is usually based on molecular mechanics (MM), quantum mechanics (QM) or a combination of both (QM/MM). There are three models that have been developed for metal ion parameterization, and they are the bonded, non-bonded and cationic dummy models [204]. The non-bonded model uses the formal charge of the metal ion and description of the coordination environment of metals is by their van der Waals and electrostatic interactions [204–206]. The cationic dummy model involves the insertion of dummy atoms around the metal, and covalent bonds are formed between the metal and atoms. The bonds are determined by predefined coordination site geometry [207]. These two model approaches are however constrained by the nature of electrostatic forces (they are long-range) and may cause the metal to adopt an unusual geometry or escape its coordination center altogether [204, 208]. The limitations of these models are addressed though, in the bonded model, as the metal is maintained within its coordination environment through the formation of covalent bonds to the metal [209]. These covalent bonds have similar characteristics to typical covalent bonds (including their bond angles, lengths, and dihedrals).

MD simulations can be performed using different software packages such as GROningen Machine for Chemical Simulations (GROMACS) [193, 210–212], AMBER [213–215],

CHARMM [216], Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [217], Nanoscale Molecular Dynamics (NAMD) [218] and (OpenMM) [219, 220]. Each program has unique features, force field compatibility, and optimizations designed to achieve sufficiently fast simulations for systems with atom numbers ranging from thousands to millions. AMBER is quite advantageous however, as it contains a tool, AnteChamber PYthon Parser interface (ACPYPE) [221], that converts topology and coordinate files generated with AMBER, to GROMACS files for MD simulations.

3.1.5 Dynamic Residue Network Analysis

Dynamic residue network (DRN) analysis is an approach developed by RUBi [222, 223], to analyze multi-layered data generated from MD simulations of proteins. It explores residue behavior during simulations, through analysis of modifications of residue interaction networks at specific time frames in an MD simulation [223]. In a protein system, residues form a complex network through which information flows, enabling residues to communicate efficiently, to carry out all protein functions, and each residue is regarded as a node, with the non-covalent interactions between nodes are referred to as edges [224, 225]. All nodes are key in maintaining a normal communication pathway within a protein network, however, if changes are imposed on one or more residues in a protein, normal network patterns are rearranged, possibly changing the communication pathway that was established for protein function. The importance of a node in a network is deduced from a given network topology, and loss of such important nodes within a network could diminish the function of the protein. The calculations for DRN are carried out using MD-TASK suite [222, 223], and in this study, three metrices were used:

1. *Betweenness centrality (BC)*
2. *Closeness centrality* (*CC*)
3. *Eigenvector centrality* (*EC*)

Utilizing DRN metrices in this research is particularly useful in determining the effects of variations on the interactions established for normal protein function. The presence of variations may disrupt interactions and communication pathways that are important for the stability and functioning of the protein. On the other hand, proteins may change the interaction network to compensate for variant presence. All this can be observed with this approach.

3.1.5.1 Betweenness centrality (BC)

Betweenness centrality is mainly used to analyze the communication networks within an enzyme. It has been identified as the most informative metric from past research, having been able to identify ligand and mutation-induced residue network changes in different proteins, and highlighted key allosteric residues [226–229]. This method calculates the importance of a residue (node) for protein communication by computing its usage in varied shortest pathways when communication is occurring involving all other nodes. The higher the averaged *BC* value for a particular node, the more important that node is in protein communication. *BC* metric is computed using the following equation:

$$\overline{BC}(v) = \frac{1}{m}\sum_{i=1}^{m}\sum_{u=1}^{n-1}\frac{\delta(s_i, t_i|v_i)}{\delta(s_i,t_i)}$$

Equation 3-3: Averaged *BC* calculation.

In Equation 3-3, *m* symbolizes the number of frames, while *v* stands for the entire set of nodes. The $\delta(s,t)$ represent number of shortest paths that link nodes *s* and *t*, then $\delta(s,t|v)$ represents the number of these paths passing through another node, *v*. The *i* depicts the number of frames [222].

3.1.5.2 *Closeness centrality (CC)*

Closeness centrality mainly identifies the central nodes that are closest to other nodes within a protein network. For a node, the *closeness centrality* is inversely proportional to total distance from other nodes. Therefore, the higher the averaged *CC* values are for particular nodes, the more centrally located the nodes are in the network. To calculate the *CC* metric, the following equation is utilized:

$$\overline{CC}(v) = \frac{n-1}{m}\sum_{i=1}^{m}\sum_{u=1}^{n-1} d(v,u)$$

Equation 3-4: Averaged *CC* calculation.

In Equation 3-4, the *d(v, u)* symbolizes the shortest path between nodes *v* and *u*. The *n* represents the number of nodes in the graph [222].

3.1.5.3 *Eigenvector centrality* (*EC*)

Eigenvector centrality identifies highly influential nodes within a protein system, that are highly connected. The higher the averaged *EC* value is for a particular node, the more influential that node is within a group of other nodes in a network. The *EC* metric is computed using the following equation:

$$A \cdot \overrightarrow{EC} = \lambda \cdot \overrightarrow{EC} \quad \text{(i)}$$

$$\overline{EC}(i) = \frac{1}{m}\sum_{k=1}^{m} EC_{ik} \quad \text{(ii)}$$

Equation 3-5: Averaged *EC* calculation.

In Equation 3-5, (i) Lambda (λ) is the eigenvalue for the eigen decomposition of adjacency matrix *A*, and *EC* is the eigenvector. This is obtained by power iteration in NetworkX. In (ii) of Equation 3-5, the averaged *EC* is computed for i^{th} residue through computing the vector for each trajectory frame and then using this to obtain an average [222].

3.2 Methodology

3.2.1 CPY2D6 allele selection

The Pharmacogene Variation Consortium (PharmVar) [48] that harbors variation information for CYP enzymes was used to compile a list of alleles for CYP2D6. From the list, alleles with clinical function assigned by the Clinical Pharmacogenetics Implementation Consortium (CPIC) were retained. There were 16 CYP2D6 alleles with either normal (*2, *27, *33, *34, *39, *45, *48, and *53) or reduced (*10, *14, *17, *29, *49, *50, *54, and *55) enzyme activity as assigned by CPIC.

3.2.2 Retrieval of reference and allele CYP2D6 structures

A 3D structure for CYP2D6 reference was obtained from the RCSB Protein Data Bank (PDB) [100] (PDB ID: 3TDA) and cleaned by removing all co-crystalized ligands and waters. Out of the 14 structures available for CYP2D6 within the database, the chosen structure had no missing residues within the protein (only missing 8 residues at the N-terminal and four residues at the C-terminal), had a good PDB validation and resolution of 2.67 Å, and had no mutations within the structure. Alleles with normal and reduced activity selected were created by using the retrieved CYP2D6 reference protein as a template and introducing their respective variations with the mutate option in Discovery Studio. The rotamers with the least clashes were chosen.

3.2.3 Parameterization and preparation of reference and allele CYP2D6 structures for MD simulations

CYP2D6 is an enzyme that contains a heme cofactor and metal iron in the center, which is crucial for its metabolic function. The heme cofactor is coordinated by a cysteine residue, forming the S−Fe bridge between the protein and heme. Parameters for the aspects of this cofactor during metabolism of CYP2D6 are however not available in the Amberff14SB forcefield to be used in this research, therefore they had to be generated, so interactions of protein residues can be governed, and the cofactor can be prevented from escaping the active site during MD simulations. Metal parameterization was performed using the cleaned CYP2D6 reference structure as a template. All hydrogens were removed from the structure using *reduce* with AmberTools19, then all titratable residues were protonated on the H++ server to a neutral pH of 7 using an internal and external dielectric of 10 and 80, respectively, and a salinity of 0.15 M. An implicit solvent topology (**.top*) file and a coordinate file (**.crd*) were downloaded and an Amber tool, *ambpdb*, was utilized to combine the files and create a PDB file with coordinates of the protonated CYP2D6 structure. Validation of the protonated structure was conducted to ensure titratable residues were in their correct protonation state. The Metal center parameter builder (MCPB) bundle with AmberTools19 was then used on the protonated system to generate input files for Gaussian09 quantum mechanical (QM) calculations, utilizing 2.8 Å as a cutoff distance to describe interactions within the metal site. The QM calculations were carried out with the B3LYP/6-31G* density functional theory on 192 CPU cores at the Center for High Performance Computing (CHPC) cluster, Cape Town South Africa. The forcefield parameters to be used (including bond lengths, angles, dihedrals and force constants between heme group and coordinating atoms) were calculated

using the seminario method and MCPB. Parameters were generated by Dr Allan T Sanyanga. The parameters were for the penta-coordinate ferric heme state with a charge of +2 for the iron were generated and allelic variant structures were also prepared by applying this complete parameter set to all alleles before molecular dynamic (MD) simulations were performed.

3.2.4 Molecular Dynamic simulations and Trajectory analysis

To ascertain the effects of variations on CYP2D6, 500 ns MD simulations of reference and alleles were carried out. Prior to the generation of allele MD simulations, duplicate MD simulations for references were performed, to validate the parameter set and confirm its behavior and reproducibility. All systems were taken through *tleap* [213] for solvation using a TIP3P water model and 0.15 M NaCl to neutralize the system. The AMBER ff14SB force field was used [200], and a coordinate and topology file for each structure was then produced. The AMBER tool, *AnteChamber PYthon Parser interface* (*ACPYPE*) [221] was finally used to generate files compatible with GROMACS v2016.1 [210] for subsequent MD simulations steps. Steepest descent minimization was performed without position restraints and with a minimization step size of 0.01, and once the systems converged, equilibration under NVT (canonical) conditions were carried out. This NVT equilibration utilized a modified Berendsen thermostat for temperature coupling and the reference temperature was set to 300 K. Next, NPT/isothermal-isobaric equilibration was performed with a Berendsen barostat, however the Parrinello-Rahman barostat was used for production MD simulations and data collection. Once temperature and pressure equilibrations were completed, production runs of 500ns each were performed for reference and alleles, using the same 300K reference temperature and a time step of 0.002 fs. During simulation,

coordinates and energies were saved after every 10 ps. All MD simulations were performed at the CHPC located in Cape Town, South Africa. When all simulations were completed, the *gmx trjconv* tool was utilized to remove periodic boundary conditions and jumps, and to center and fit the trajectories to the reference structure. The trajectories were then viewed using the visual molecular dynamics tool (VMD) [230] to observe any clear structural conformational changes. Next, the stability and global structural changes were assessed using GROMACS built-in tools (*gmx rms*, *gmx rmsf*, *gmx gyrate*, *gmx cluster* and *gmx dssp*). These tools calculated RMSD, RMSF, Rg, geometric clustering and secondary structure of proteins. All versus all RMSD calculations using Cα atoms [226] were also performed to compare the conformational variability throughout each simulation. Hydrogen bonding analysis for all systems was carried out using *CPPTRAJ* available in AMBER tools. Delta RMSF (ΔRMSF) and delta hydrogen bonding (ΔH-bond) were computed using the methodology in section 3.2.8.

3.2.5 Heme region analysis

The heme cofactor within the core of CYP2D6 is crucial for its functionality as a monooxygenase, therefore analysis was conducted on the heme region to determine whether variations affected this area. The GROMACS tools, *gmx dist* and *gmx hbond*, were initially used in analysis. The *gmx dist* tool was used to calculate the distance between the center of mass (COM) of the heme and the COM of the active site in all systems for the duration of the equilibrated MD simulation (last 300 ns). The *gmx hbond* was utilized to calculate the number of hydrogen bonds being formed between the heme and surrounding protein residues. The heme cofactor, however, partakes in different non-covalent interactions with surrounding residues, in order to stabilize it

within the CYP2D6 core. This ensures optimum functionality of the cofactor within the protein. Due to most of these interactions occurring within ~4 Å, an ad hoc python script was generated to monitor the interactions between the heme and residues around it, throughout the last 300 ns MD simulations for all systems. The script used a trajectory (*.*xtc*) and a coordinate (**pdb*) file as inputs, then selected all heavy atoms of the heme. The distances between the heavy atoms of the heme and heavy atoms of protein residues were calculated after every 100ps in the trajectory, and utilizing a cutoff distance of 4 Å, every time the distance between a heavy atom of the protein and a heavy atom of the heme was within 4 Å, it was considered a contact between the residue atom and the heme atom. The contact range for each protein atom with a heme atom varied between 0 and 1. A contact value of 1 signified full interaction between a residue atom and a heme atom, while a value of 0 indicated no interaction. The contacts of atoms from the same residue were then added to form total contact of each residue with the heme, and a heatmap was generated to show the total contact of surrounding residues with the heme for all systems.

3.2.6 Tunnel analysis

The analysis of possible entrance and egress tunnels in CYP2D6 reference and alleles was carried out using CAVER 3.0 [231]. For each reference and allele in this study, 1001 snapshots were obtained and aligned from their respective 500ns MD simulations, at 2ns intervals, with the snapshot at 0ns included. These snapshots were then used as input in CAVER, and two active site residues, E216 and D301, as well as the heme, were used for the starting point. Tunnel analysis was then performed using a probe radius of 0.9 Å and a clustering threshold of 4.5. All other parameters were left on default. Using the acquired data, we generated heatmaps showing tunnel

opening frequencies for references and alleles. In the heatmaps, the change in bottleneck radii for each tunnel throughout the trajectories was represented. We considered a tunnel to be open if its bottleneck radius was ≥ 1.4 Å for the greater part of the simulation. Tunnels were visualized in PyMOL.

3.2.7 Dynamic Residue Network Analysis

Using the MDM-TASK-web service [222], Dynamic Residue Network (DRN) analysis was carried out to identify the impact of variations on the information flow within the CYP2D6 protein residue network. With this method, the Cβ atoms (Cα for glycine) of residues within a protein network are regarded as nodes, while the edges between nodes signify unbonded interactions or linkages with a cutoff distance of ≤ 6.7 [232]. Using the last 300 ns of MD simulations for the references and alleles, DRN metric calculations were performed (averaged *betweenness centrality* (*BC*), averaged *closeness centrality* (*CC*), and averaged *eigenvector centrality* (*EC*)) to detect any modifications to the CYP2D6 communication networks caused by variations. The importance of a residue in a protein network is measured by *BC*, which counts how often a node appears in the protein's shortest pathways [233]. *CC* is a metric that measures how fast information flows through a node, by measuring how close each node in a network is to every other node, and the *CC* of each node is calculated by averaging the shortest pathways between that node and every other node in the network [233, 234]. Lastly *EC* calculates a node's level of influence within a network, by iteratively taking into account the significance of a node's neighbors [234, 235]. The delta *BC* (Δ*BC*), *CC* (Δ*CC*) and *EC* (Δ*EC*) were then calculated according to methodology in section 3.2.8.

3.2.8 Delta calculations

Delta calculations are informative in quantifying differences when comparing systems, therefore delta values for RMSF, hydrogen bonding and DRN metrices BC, CC and EC were computed. For each of the results obtained from calculations for the different metrices, the following was carried out: 1) The values for the two references were extracted, and averaged per residue, for comparison with the other systems; 2) The averaged values for the references were then subtracted from the allelic metric values (allele - reference), per residue, to obtain ΔRMSF, ΔH-bond, Δ*BC*, Δ*CC* and Δ*EC*; 3) For each system, the delta metric values obtained were then sorted in descending order, and the residues with the top and bottom 3% delta metric values were obtained. For hydrogen bonding, the residue pairs with the top and bottom 3% delta metric values were obtained. The residues with the top and bottom 3% delta values per system were then compiled into a heatmap, with corresponding residues per system that did not fall within the residues with the top and bottom 3% delta values added for comparison. The residues with top and bottom 3% delta metric values were also mapped to their corresponding structures obtained from conformational clustering analysis; 4) Significance of the delta metric values was computed. Using the raw metric values for the two references, standard deviation was computed per residue, then the delta values in the corresponding residues of allelic system were considered significant if they were outliers not within 3SD of the mean of the reference data (significant values were noted to be within the top and bottom 3% delta metric values extracted). In the heatmaps, significant values within the top and bottom 3% delta values per system are marked with an “X” and values in the top and bottom 3% delta values and not significant are marked with a “-“.

Additionally, for the DRN metrics, to highlight regions with high metric values within the references (without delta calculations), the top 6% of the averaged reference values for each metric were obtained and mapped to a CYP2D6 reference structure and shown in a bar plot. Scripts for these analyses were generated by Dr Wayde Veldman.

3.3 Results and Discussions

3.3.1 Identification of CYP2D6 alleles with known clinical function, 3D structure generation, and MD simulations

CYP2D6 alleles classified as having normal or reduced functional consequence were retrieved from the PharmVar website [48], and as can been seen in Table 3-1, each allele contained either one variation, or multiple variations occurring simultaneously. The variations found in the alleles were spread out, from the N-terminal to the C-terminal (variations P34S, A90V, T107N, F120I, A122S, V136I, E155K, E156K, G169R, A237S, T261I, R296C, V338M, K404Q, E410K and S486T) (Figure 3-2A). Variation S486T located on the edge of the active site is common in most alleles that exhibit reduced functional consequence, except for CYP2D6*50. This variation is also found at a lesser extent in the alleles with normal clinical function (CYP2D6*2, CYP2D6*39 and CYP2D6*45). Another variation, R296C, on the edge of the active site, is commonly found in alleles CYP2D6*2, *34, *45, *14, *17, *29, and *55 which either have normal or reduced functional consequence.

A reference CYP2D6 structure was then obtained from PDB (PDB ID: 3TDA), and all structures for CYP2D6 alleles listed in Table 3-1 were created by mutating amino acids at their respective positions on the reference structure. Molecular dynamic (MD) simulations (500 ns) were conducted on the reference and all 16 alleles, and different *in silico* tools were used to analyze the results. Duplicate MD simulations were conducted for the reference (control), to confirm its

behavior and thereby note that any differences observed in alleles are due to variations. We refer to the first simulation of the reference as REF1 and the duplicate as REF2. The results of REF1 and REF2 were found to be highly consistent in all analysis performed, and any areas where discrepancies between them were observed, were excluded from analysis.

Table 3-1: CYP2D6 alleles studied with their respective functional consequence, variations, combinations and location of variations in the structure. Variations added to alleles to form new alleles are highlighted in bold.

Enzyme	Functional Consequence	Alleles	Variations	Location in enzyme	Combinations
CYP2D6	Normal (EM)	CYP2D6*2	R296C	I helix and SRS4	CYP2D6*34 + CYP2D6*39
			S486T	β4-2 sheet	
		CYP2D6*27	E410K	K'-K'' loop	-
		CYP2D6*33	A237S	G' helix	-
		CYP2D6*34	R296C	I helix and SRS4	-
		CYP2D6*39	S486T	β4-2 sheet	-
		CYP2D6*45	E155K	D helix	CYP2D6*2 + **E155K**
			R296C	I helix and SRS4	
			S486T	β4-2 sheet	
		CYP2D6*48	A90V	B helix	-
		CYP2D6*53	F120I	B'-C loop, SRS1 and active site residue	-
			A122S	B'-C loop, SRS1 and active site residue	
	Reduced (IM)	CYP2D6*10	P34S	Pre-A loop	CYP2D6*39 + **P34S**
			S486T	β4-2 sheet	

		CYP2D6*14	G169R;	D-E loop	CYP2D6*2 + **G169R**
			R296C	I helix and SRS4	
			S486T	β4-2 sheet	
		CYP2D6*17	T107I	B' helix and SRS1	CYP2D6*2 + **T107I**
			R296C	I helix and SRS4	
			S486T	β4-2 sheet	
		CYP2D6*29	V136I	C helix	CYP2D6*2 + **V136I** + **V338M**
			R296C	I helix and SRS4	
			V338M	J helix	
			S486T	β4-2 sheet	
		CYP2D6*49	P34S	Pre-A loop	CYP2D6*10 + **F120I**
			F120I	B'-C loop, SRS1 and active site residue	
			S486T	β4-2 sheet	
		CYP2D6*50	E156A	D helix	-
		CYP2D6*54	P34S	Pre-A loop	CYP2D6*10 + **T261I**
			T261I	G helix	
			S486T	β4-2 sheet	
		CYP2D6*55	R296C	I helix and SRS4	CYP2D6*2 + **K404Q**
			K404Q	K' helix	
			S486T	β4-2 sheet	

3.3.2 Instability and conformational changes in CYP2D6 alleles

The stability and conformational changes of CYP2D6 references and all alleles were analyzed throughout the simulation using RMSD and Rg. Results for both RMSD and Rg were visualized using the more common timeline plots and violin plots, and RMSD data was further analyzed using all vs all heatmaps. As seen in the line plots in Figure 3-3A, the normal functioning

alleles *2, *27, *33, *34, *39, *45, *48, and *53 were exhibiting behavior similar to both reference systems, reaching a more stable state earlier in the trajectory. Contrary to this, reduced activity alleles *10, *17, *29, *50 and *55 showed slightly higher RMSD values compared to the references and stabilized between 150 ns and 200 ns (Figure S4).

To identify conformational differences between the references and all 16 alleles, violin plots and all vs all heatmaps were created using the RMSD values and shown in Figure 3-3B and Figures 3-3C respectively. When observing the whole trajectory (0-500ns) RMSD violin plots, the results distinctly differentiated the references and normal functioning systems from the reduced functioning systems. The alleles with normal functionality had unimodal distribution observed, suggesting they maintained one conformation throughout the simulation (Figure 3-3B). However, for alleles *10, *14, *17, *29, *50, *54, and *55 with reduced functionality, bimodal distribution was observed, indicating they visited at least two conformations in the simulation. After ~200 ns, these reduced activity systems stabilized and were noted to exhibit one conformation for the rest of the trajectory, as seen in the violin plots in Figure 3-3B. These findings were also observed in the all vs all RMSD heatmaps (Figure 3-3C).

The compactness of references and alleles throughout the simulation was evaluated using Rg calculations and displayed as line and violin plots in Figure 3-4. After analyzing the results at different time points in the simulation, only minor differences in compactness were observed between references and all alleles.

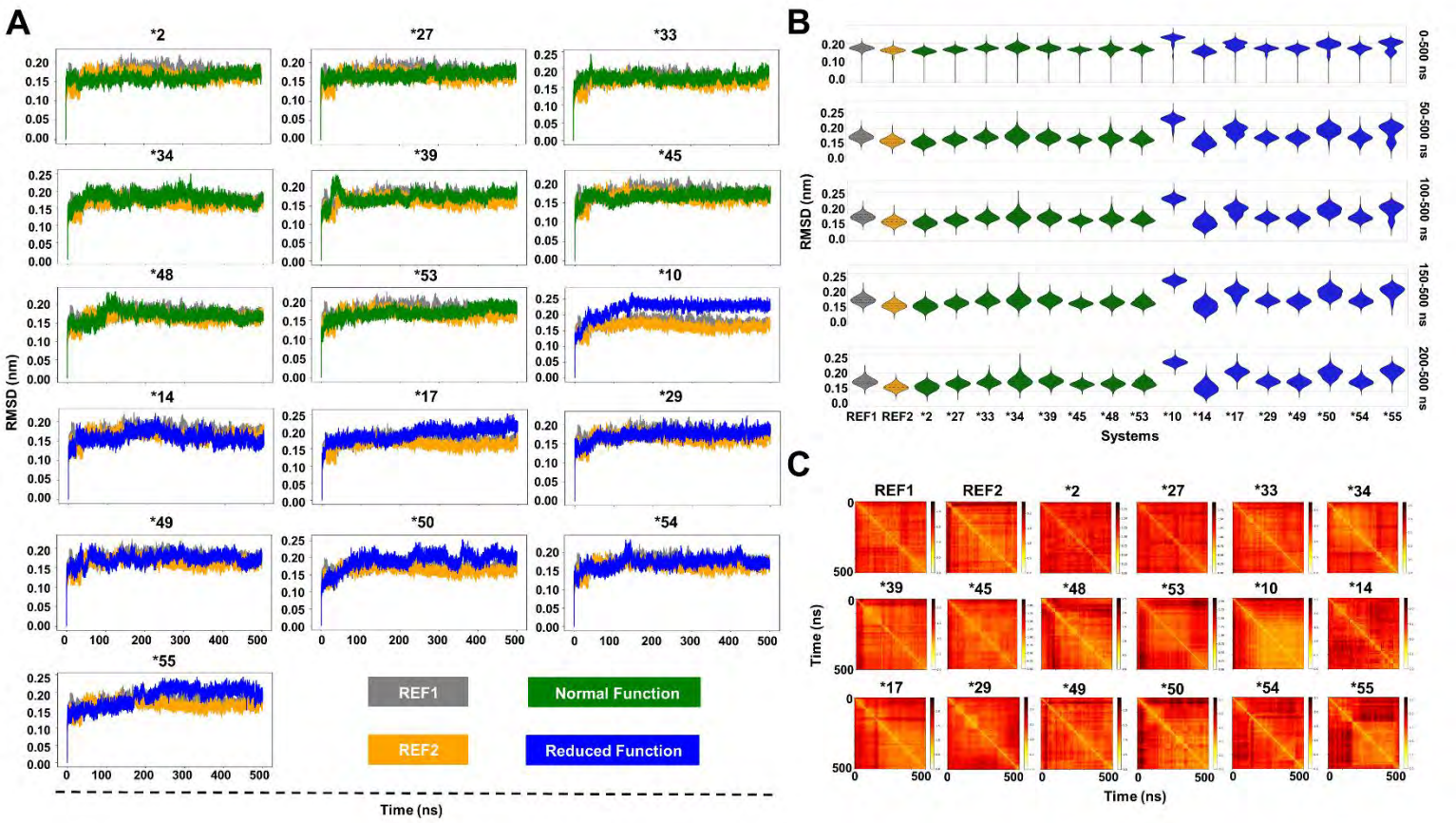


Figure 3-3: A) RMSD vs timeline plots for the whole 500 ns MD simulation of all alleles. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue. B) Different stages of the MD simulation (0-500ns, 50-500ns, 100-500ns, 150-500ns, 200-500ns) shown as RMSD violin plots, with the same colouring key as line plots. C) All vs all RMSD heatmaps for whole MD simulation (500 ns). The x and y axes show the frames at different time stamps and the degree of variation is from low to high (coloured from white, through yellow and red, to black).

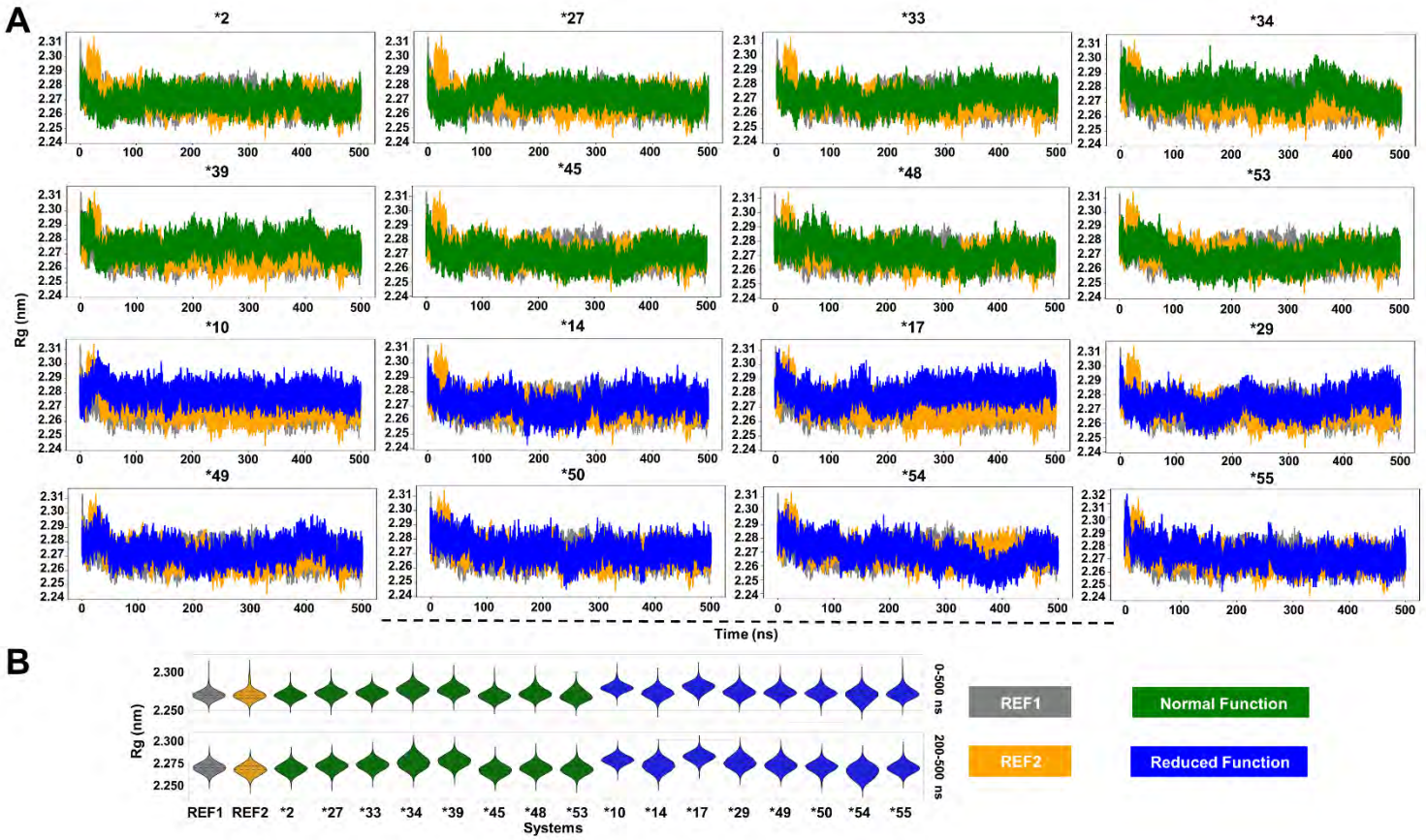


Figure 3-4: CYP2D6 reference and allele Rg analysis. A) timeline plots for the whole 500 ns MD simulation. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue. B) Rg violin plots for different stages of references and allele MD simulations (0-500ns and 200-500ns). Violin plots coloured using the same colouring key used for line plots.

From the findings presented here, it was evident that many reduced functioning alleles reached stability considerably later in their simulations as compared with references and normal function alleles, stabilizing between 150 ns and 200 ns. During this equilibration, structural changes are inevitable [236], and thus it was crucial to determine whether any changes in structure occurred during the convergence of the systems. We examined this in the next section.

3.3.3 Assessing changes in secondary structure, flexibility and hydrogen bonding in CYP2D6 alleles with known clinical activity

The secondary structure and interaction changes in allele systems were analyzed using four methods in this section. These are: conformational clustering, secondary structure profiling, RMSF and hydrogen bond analysis. The more stable part of all simulations (last 300 ns) was used for all calculations conducted.

3.3.3.1 Conformational clustering analysis

Highly similar structures in the MD simulations of all alleles and references were grouped together using the GROMACS geometric clustering program. This method produced a median structure as a representative of each cluster/group. Upon analysis of the results, it was noted that the trajectories for all alleles and references contained one cluster with structures having an RMSD < 0.2 nm (Table S5), indicating there was one conformation sampled throughout the last 300 ns of each MD simulation. These findings were consistent with what was observed in the RMSD violin plots for this portion of the simulation.

The two reference median structures obtained from conformational clustering were then superimposed on each other, and it was noted that they were highly similar, with no major differing regions in conformation (Figure 3-5A). Because of this high similarity between the references, either of the median reference structures could be used to compare with allele structures, and as

such, the median structure of each allele was superimposed to REF1 structure and shown in Figure 3-5B.

Discrepancies in structural conformation were noted in all alleles, with differences observed mainly in the B' and F-G regions (Figure 3-5B). A highly deviating conformation relative to REF1 was however observed in reduced function alleles*10, *17, *29, *50, *54 and *55, while allele *14 however, was noted to have a more open conformation. Additionally, secondary structure changes were also detected in alleles *10, *14, *17, *29 and *55 (Figure 3-5C). In allele *10, a bend within the F' helix was lost, resulting in a straighter helix and a slight closure of the structure. Its G' helix also displaced, leaving residues 235-237 in a helical structure, and the beginning of the G helix (residues 242-245) in a loop structure. As mentioned in Chapter 2 (Section 2.3.4), allele *10 contains the P34S variation in the proline-rich PPGPXPXP region, which is responsible for the enzyme's correct folding and anchoring to the membrane. We hypothesized that the presence of this variation within this region would affect the structure of CYP2D6, and the major discrepancies in structure observed in the *10 allele confirm the effect of this variation on the structure of CYP2D6. Residues 234-237 of the G' helix in allele *14 turned into a loop, while residues 214-216 (F helix) and 242-244 (G helix) in allele *17 also changed to loop structures. In allele *29, the helix segment after the bend in the F' helix (residues 228-232) fully dissociated into a loop structure, and residues 242-243 in the initial portion of the G helix also became a loop. The G' helix in allele *54 (residues 234-240) was observed to have dissociated, with a loop forming in its place.

Interestingly, normal activity allele *39 was found to be in a more closed conformation as compared to the reference (resembling reduced activity allele structures), while alleles *34 and *48 contained secondary structure changes that were mainly noted in reduced activity alleles. Both *34 and *48 had residue 189 located in the F' helix and residues 234 and 235 within the G' helix, respectively, forming a loop instead of the helical structure observed in both references.

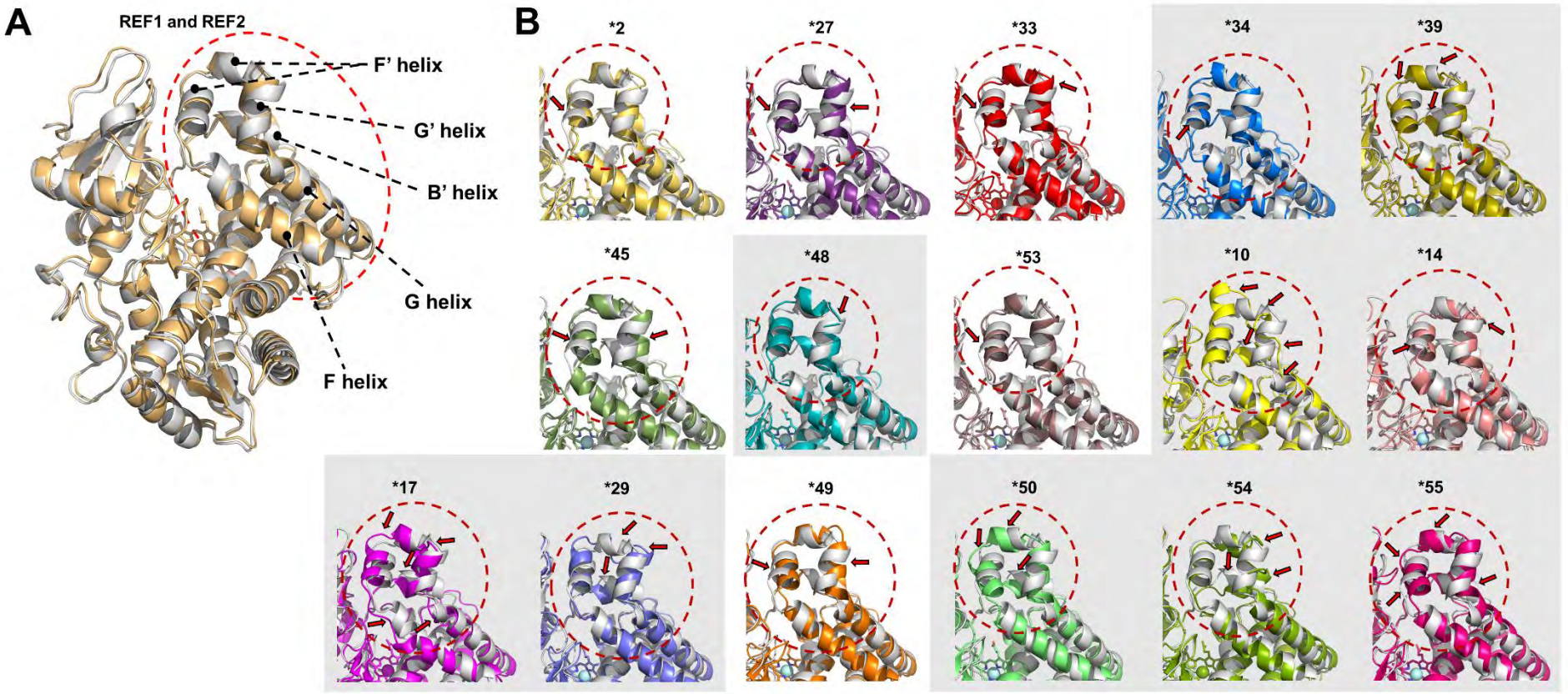


Figure 3-5: Conformational clustering analysis. A) REF1 (grey) and REF2 (light orange) median cluster structures superimposed. The B' and F-G regions where deviations were observed in alleles, were highlighted in these structures. B) Median cluster structures for normal and reduced function alleles superimposed to REF1. Regions with changes in secondary structure or conformation as compared to REF1 were shown with red circles and arrows. Alleles *10, *14, *17, *29, *34 *39, *48, *50, *54, *55 with more deviating conformations as compared to REF1, or with secondary structure changes noted were highlighted in grey boxes. Cluster structures for the different alleles in B) are coloured differently to differentiate from REF1 structure and each other.

3.3.3.2 Secondary structure profiles

Due to the divergence in secondary structure observed for allele in conformational clustering analysis, as a next step, it was important to analyze the secondary structure profiles of the references and alleles. To do this, the DSSP program in GROMACS was employed. It calculates the secondary structure preference per residue over a simulation. The more stable part of the MD simulation (last 300 ns) of each reference and allele system was used, and the resultant secondary structure profiles are seen in Figure 3-6. Overall, the normal functioning alleles were highly similar to both references, with minimum secondary structural changes noted in the A'-A loop (*34), the B' helix region (*2, *27, *33, and *45), the F helix region (*27, *33, *45 and *53) and the β1 sheet region (*34 and *48). A more divergent secondary structure profile was observed for allele *39, however, with differences mainly in the B' and F helices, and β4 sheet regions (Figure 3-6A). This finding was contradictory to the pattern that was observed in all other alleles with normal activity (where only one or two regions differed in secondary structure).

Reduced activity alleles presented significant deviations in secondary structure profile as compared with references, with many differences in the A'-A and E-F loops, the β1 sheet region and the B', E, F, F', G, and G' helices (Figure 3-6A). Each of the alleles (*10, *14, *17, *29, *50, *54 and *55) was noted to have three or more regions differing as compared with both references.

As seen in Figure 3-6A, alleles *10, *14, *17, *29, *50, *54 and *55 were observed to have changes in the B' helix region (residues 103-110), with the 3-10 helix secondary structure shifting

to an alpha helix. Alpha helices are more stable secondary structures over 3-10 or pi-helices, mainly due to the hydrogen bonding between a carbonyl oxygen of one amino acid residue and an amide hydrogen of the fourth residue from this point in the peptide, that stabilize the helical structure and aids in maintaining protein structure [237, 238]. Alpha helices are also more compact, with a tightly coiled structure that allows for efficient packing of residues within the protein, to maintain the overall folding and structure of proteins [237]. Therefore, a change in the B' helix, that shifts to allow for free substrate entry and product exit [179], from a less stable 3-10 helix to a more stable and compact alpha helix, may reduce the mobility of this region, affecting substrate entry. Additional to this observation, simulations of alleles *10, *14, *17, *54 and *55 were noted to contain more flexible loop secondary structures within the β1 sheet region (β -1 and β1-2 sheet regions with residues 72-80). The flexible loops substituted the more stable hydrogen-bonded turn and beta-ladder secondary structures noted in reference simulations. A channel 2b gating residue (W75) that helps shuttle substrates to the active site [185] is situated in this β1 sheet region, and increased flexibility of this area could affect the function of W75, thereby reducing entrance of substrates into the active site for catalysis (discussed further in section 3.3.5).

A kappa-helix appeared in the E-F loop region of alleles *17 and *50 (residues 194, 195 and 196), indicating an interchange of two unstable secondary structure designations within this region, possibly increasing its flexibility. This region has specific mobility (acts as a switch for the regulation of the movements within the F-G regions [239, 240]) and increasing the flexibility of this region could potentially destabilize it and reduce its functionality (this was confirmed with ΔRMSF analysis later in the chapter). Also in allele *50, part of the E helix (residues 187-189), a

key region for structural integrity of the CYP2D6 core [241–243], was observed to sample the less stable pi helix for a portion of the MD simulation, in place of an alpha helix secondary structure, likely reducing the stability of this region. Another part of the E helix (residues 175-177) was similarly destabilized in alleles *29 and *55, with the alpha helix being replaced by a less stable hydrogen bonded turn secondary structure. In allele *14, there was an increase in sampling of the 3-10 helix secondary structure within the A'-A loop region (residues 51-55), meaning the area increasingly sampled different secondary structures interchangeably throughout the simulation, leading to a possible destabilization of the region. The A'-A region is paramount for the correct folding of CYP2D6 and its anchoring to the membrane [52, 53], and changes in flexibility .

The F-G region (residues 200-261) is an integral part of the CYP2D6 enzyme, as it facilitates the entry and exit of substrates, by displacements and movements of certain areas within these helices and loops [32, 60]. In most reduced activity alleles studied here, there were numerous secondary structure changes noted within this region as compared to the references, and these changes may hinder its proper functioning. In the *10 allele, residues 212-215 and 226-232 within the F and F' helices respectively, were noted to have a reduced sampling of the 3-10 helix secondary structure in the MD simulation, while the sampling of the alpha helix increased, suggesting a decrease in mobility of these helices. Contrary to this, residues within the G and G' region (233-243) showed an increased sampling of the 3-10 helix structure, which may render the region more mobile. In the F and F' helices of alleles *14 and *29 (residues 212-219 and 212-225 respectively), there was an increase in the sampling of the 3-10 helix secondary structure, likely also increasing mobility within these areas. However, there was a loss of 3-10 helix secondary

structure sampling within these same regions, similar to allele *10 (residues 212-215), in alleles *50 and *55, reducing their mobility in these alleles.

In allele *54, there was an interchange of secondary structure assignments noted within the F'-G region (residues 231-243) throughout the simulation. The interchange involved a pi helix, a 3-10 helix, a coil/loop and a hydrogen bonded turn. This consistent change between structurally unstable secondary structures would likely make the region less stable and highly flexible, as compared with the references that sampled the more compact and stable alpha helix in this region. Changes observed thus far in secondary structure profiles of the F-G region correspond to what was observed conformational clustering and delta RMSF analysis (which will be discussed in the next section).

Interestingly, only one region in the secondary structure profile of reduced activity allele *49 was visibly different from references, and this was contradictory to the pattern observed for reduced function alleles. This difference was noted in the F helix (residues 212-215), and as can be seen in Figure 3-6A, the region sampled more of the alpha helix secondary structure, making this region more stable. The minimal changes seen in this allele aligned with what was observed for normal activity alleles, and all other results from analysis so far (i.e., stability and conformation analysis with RSMD and Rg, and conformational clustering) have been consistent with findings from references and normal function alleles.

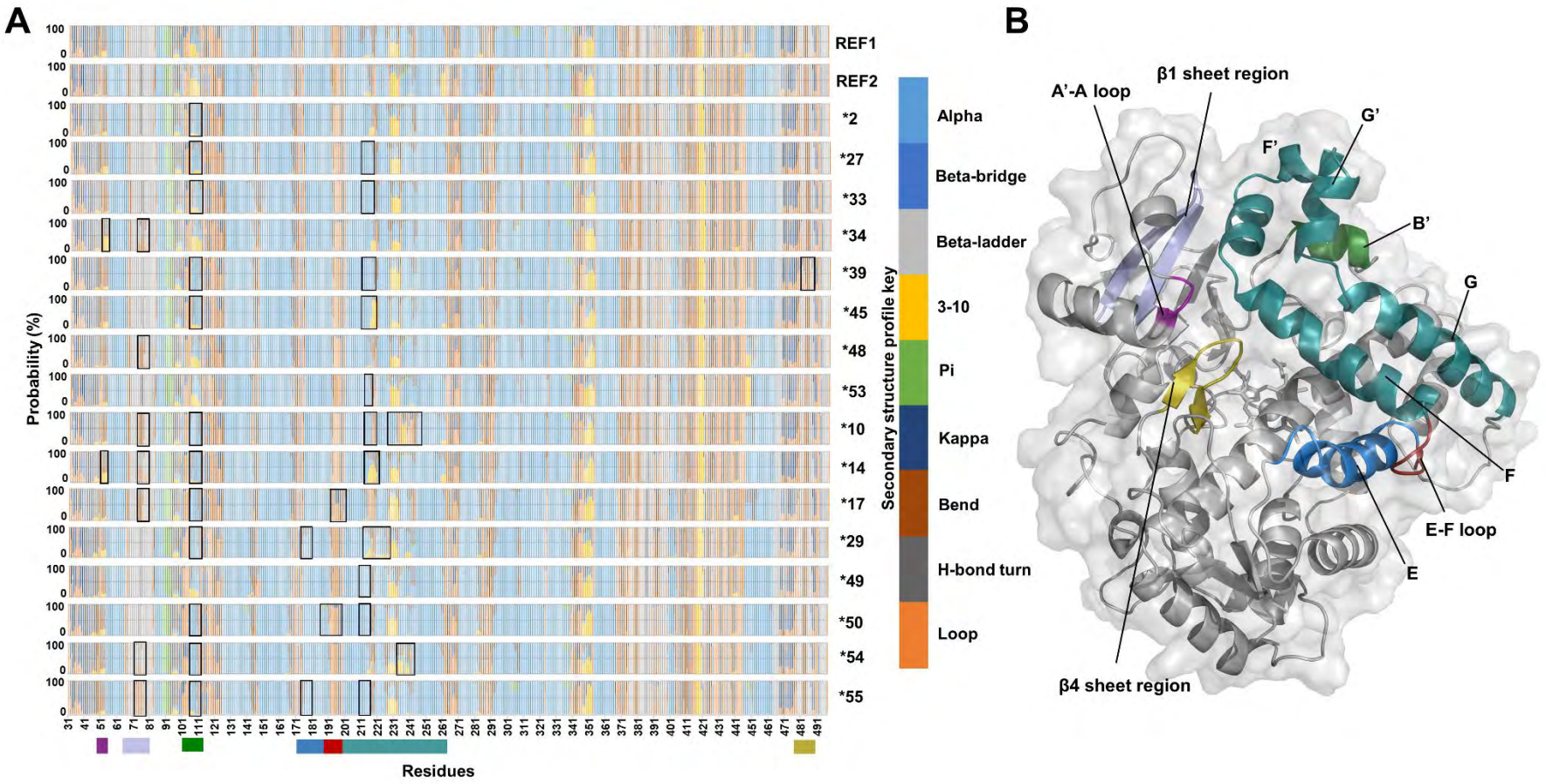


Figure 3-6: Secondary structure profiles of references and alleles with differing regions highlighted. A) Secondary structure preference per residue (for both CYP2D6 references and alleles) within the last 300 ns of the MD simulations. Bar graphs represent probability of each residue adopting a specific secondary structure in the last 300ns of each trajectory. The regions where differences are noted within the secondary structure profiles relative to references are highlighted with black boxes, and a colour key for the different secondary structure assignments (Alpha-helix, beta-bridge, beta-ladder, 3-10 helix, pi helix, poly-proline II or kappa-helix, bend, hydrogen bonded turn, coil/loop) provided next to the profiles. B) Regions in alleles noted to be different as compared to references mapped to CYP2D6 REF1 structure and colours of regions are matched to areas highlighted at the bottom of the bar graphs.

3.3.3.3 RMSF analysis

To validate our findings in secondary structure profile analysis that suggested certain regions in the alleles had changes in flexibility and stability, RMSF calculations were conducted using the same trajectory region as DSSP calculations (las 300 ns). As seen in Figure S5, there were changes in flexibility within alleles as compared to references, however, to better conceptualize these differences in respect to the references, ΔRMSF was computed. Results were obtained and the residues with the top and bottom 3% ΔRMSF values were obtained and plotted in a heatmap and in structures as seen in Figure 3-7. Residues exhibiting significant changes in RMSF within the top and bottom 3% (outliers not within 3SD of the mean of the reference data) were identified and are marked with an "X" on the heatmap.

We observed a correlation between DSSP and ΔRMSF results. In alleles *2, *10, *14, *17, *27, *29, *39, *45, *50, and *55, the MD simulations showed reduced sampling of the 3-10 secondary structure at residues 103-115 (B' helix region), favoring the more stable alpha helix. This shift correlated with a decrease in flexibility in these alleles, as indicated by ΔRMSF (Figure 3-7), with significant changes observed in specific residues within this region for alleles *2, *10, *14, *17, *27, *39, *45, and *50 (Figure 3-7). Allele *17 and *50 had a significant increase in flexibility within the E-F loop (residues 194-199 and 190-200 respectively in each allele) as seen in Figure 3-7, and similar results were observed in DSSP analysis where the region interchanged between the less stable kappa helix and coil/loop secondary structures. A significant increase in flexibility was also observed within this region for alleles *2 and *27; however, no changes in secondary structure were detected based on DSSP analysis. In the A'-A loop region of alleles *14

and *34, we observed an increased sampling of a 3-10 helix secondary structure between residues 51 and 55, indicating destabilization of this region. This destabilization was further supported by the ΔRMSF analysis, which revealed a significant increase in the flexibility of this region Figure 3-7.

Numerous changes in flexibility were observed across all alleles within the F-G region, with more significant alterations noted in some alleles associated with reduced functionality (Figure 3-7). In allele *10, a significant increase in flexibility was observed in the G' and G helices (residues 235-259) compared to both references. This observation was consistent with DSSP analysis, which indicated a higher prevalence of coil/loop secondary structures being sampled in this region throughout the simulation. In allele *29, a significant increase in flexibility was observed in the F'-G' helix region (residues 227-242), with similar results noted in DSSP analysis, where an increased sampling of the less stable 3-10 helix secondary structure was detected throughout the simulation. Additionally, within the F'-G' helix region, allele *54 exhibited a significant increase in flexibility in residues 225-240 (Figure 3-7). This finding was also consistent with DSSP analysis, which revealed that this region increasingly sampled the less stable 3-10 helix conformation throughout the simulation, as compared to references. Regardless of functionality however, alleles*10, *27, *33, *49, *53, and *55, were noted to have reduced 3-10 helix sampling in the F helix (residues 212-215) in DSSP analysis, and a corresponding loss in flexibility based on ΔRMSF.

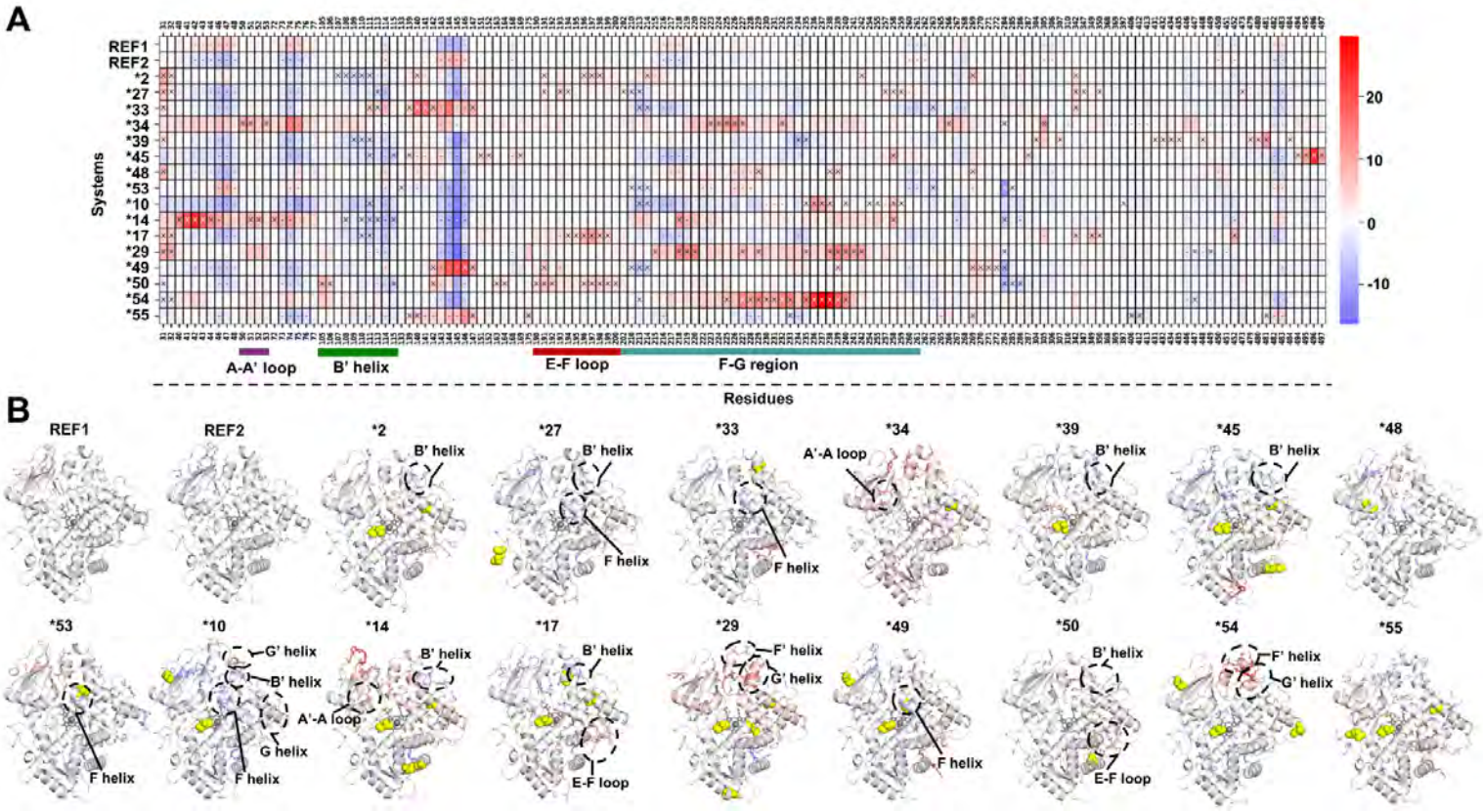


Figure 3-7: A) Residues with top and bottom 3% ΔRMSF values compiled into a heatmap, and significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Values within top and bottom 3% ΔRMSF values and not significant are marked with a "-". Regions exhibiting significant changes in flexibility, which correlate with the observations from DSSP analysis, are highlighted in boxes and colour-coded to match the DSSP analysis. B) Residues with top and bottom 3% ΔRMSF values for each system mapped to reference and allele conformational cluster structures and shown as sticks. Residues with significant changes in ΔRMSF corresponding to changes noted in DSSP analysis marked with black circles. Variation positions for each allele shown in spheres and coloured in yellow. In A) and B) ΔRMSF values range from positive to negative. Residues with a loss in flexibility (becoming more rigid) within the systems are coloured in blue, and residues with gain in flexibility in systems are coloured in red.

3.3.3.4 Hydrogen bond analysis

Delta hydrogen bond (ΔH-bond) calculations were carried out, so as to examine the variations in hydrogen bond interactions across all system as compared to the references, following the observed changes in conformation, secondary structure and flexibility noted with various tools in this chapter. Using the same trajectories used for conformational clustering, DSSP and ΔRMSF, the residue pairs with the top and bottom 3% ΔH-bond values per system were obtained and shown in a heatmap as seen in Figure 3-8A. Notable changes in hydrogen bonding between residue pairs observed in the different alleles were then mapped to corresponding allelic structures and shown in Figure 3-8B. In the heatmap, significant changes in hydrogen bonding between two residues as compared to the references were marked with an "X". Any residue pairs with deviations between REF1 and REF2 (as seen in the heatmap in Figure 3-8A) were not included in analysis.

Generally, we did not observe any definite patterns of hydrogen bonding differentiating normal activity alleles from reduced activity alleles, however, there were some significant changes in hydrogen bonding noted in some alleles at different regions of the enzymes. There was a significant increase in hydrogen bonding between residue pairs 215-242, 216-239, 222-75, and 225-106 in *10 allele, while in *55 allele, residue 221 showed a notable gain in hydrogen bonding with residue 244. In *29 allele, variations caused a considerable increase in hydrogen bonding in residue pairs 102 – 104, 222-225 and 221-240 (Figure 3-8B). Some residues mentioned in the three alleles – *10, *29 and *55 – with the increase in hydrogen bonding, are located either in the B' or the F'-G regions, which have previously been observed to undergo conformational and secondary structure changes. The formation of these hydrogen bonds in these regions could stabilize this

region in the highly different conformation that may potentially restrict substrate access through channels with entrance sites in this area (see section 3.3.5). Additionally, the F-G region became more rigid in alleles *10 and *55, as indicated by the ΔRMSF results, though the change was less pronounced in the *55 allele.

Other hydrogen bond changes in alleles were observed that were not directly related to the conformational and structural alterations previously noted, but may have been an immediate consequence of variations. The *27 allele showed to have a significant loss in hydrogen bonding between residue pairs 62-405 and 62-67 that are adjacent to variation E410K, however, to counter this variation effect and possibly stabilize this region, the residue 62 gained hydrogen bonding significantly with residue 59 (Figure 3-8B). In *45 allele, there was a notable loss in hydrogen bonding at variation E155K. The change from glutamate to a lysine amino acid significantly reduced the hydrogen bonding in this position with residue 344, however, this residue 344 immediately gained hydrogen bonding significantly with residue 150. Additionally, in *45 allele, variation R296C caused a significant loss in hydrogen bonding at this position, with residues 252 and 293 (Figure 3-8), and this observation was also noted in alleles *2, *14, *17, *29, *34 and *55 where this variation occurred (Figure S6). Only in alleles *2, *14 and *45 did we observe a gain in hydrogen bonding between the residue 293 and 115, with the change only being significant in *45.

As seen in Figure 3-8B, *17 allele was observed to have a significant loss in hydrogen bonding at the variation T107I position (residue pair 104-107), but there was a gain in this this

interaction between residues 102-104, although not significant. We then observed a further decrease in hydrogen bonding near variation R296C, between residue pairs 156-197, 184-195 and 190-272 which have residues from the conserved D and E helix (containing motif 6 identified in Chapter 2) and the E-F loop. Similar results were observed for *50 allele, where we noted a considerable loss in hydrogen bonding at the variation E156A site. Residue 156 lost hydrogen bonding with residues 184, 194 and 197 which are also located within the E helix and E-F loop. The E helix also contained another significant loss in hydrogen bonding between residues 190 and 272. These substantial losses in interactions within this region for both *17 and *50 alleles, caused the E-F loop region to gain flexibility significantly, as observed with ΔRMSF in section 3.3.3.3. Interestingly in *54 allele, there was a chain of hydrogen bond changes effected by the variation T261I (Figure 3-8B) and continuing to the B' and F'-G' region. Significant changes in hydrogen bonding within the chain included residue pairs 292-296, 293-296, 293-115 and 104-107. In the B' helix, the significant loss in hydrogen bonding between residues 104 and 107 caused this region increase in flexibility as seen with ΔRMSF results.

Overall, from our findings within this section, we can deduce that the secondary structure and flexibility of the alleles with normal activity were quite comparable to those of the two references; however, in most reduced activity alleles, several areas critical to the correct functioning of CYP2D6 had altered conformation, secondary structure and flexibility. These areas include the A-A' region, the B' region, the E-F region, F-G region and the β1 sheet region that are located in and around the substrate entrance channels. We also noted through hydrogen bond analysis, that the different variations were causing changes in hydrogen bonding networks within

the different alleles to change, which may have resulted in the structural and flexibility differences observed. These changes especially in regions where substrate entrance sites are located could be highly detrimental.

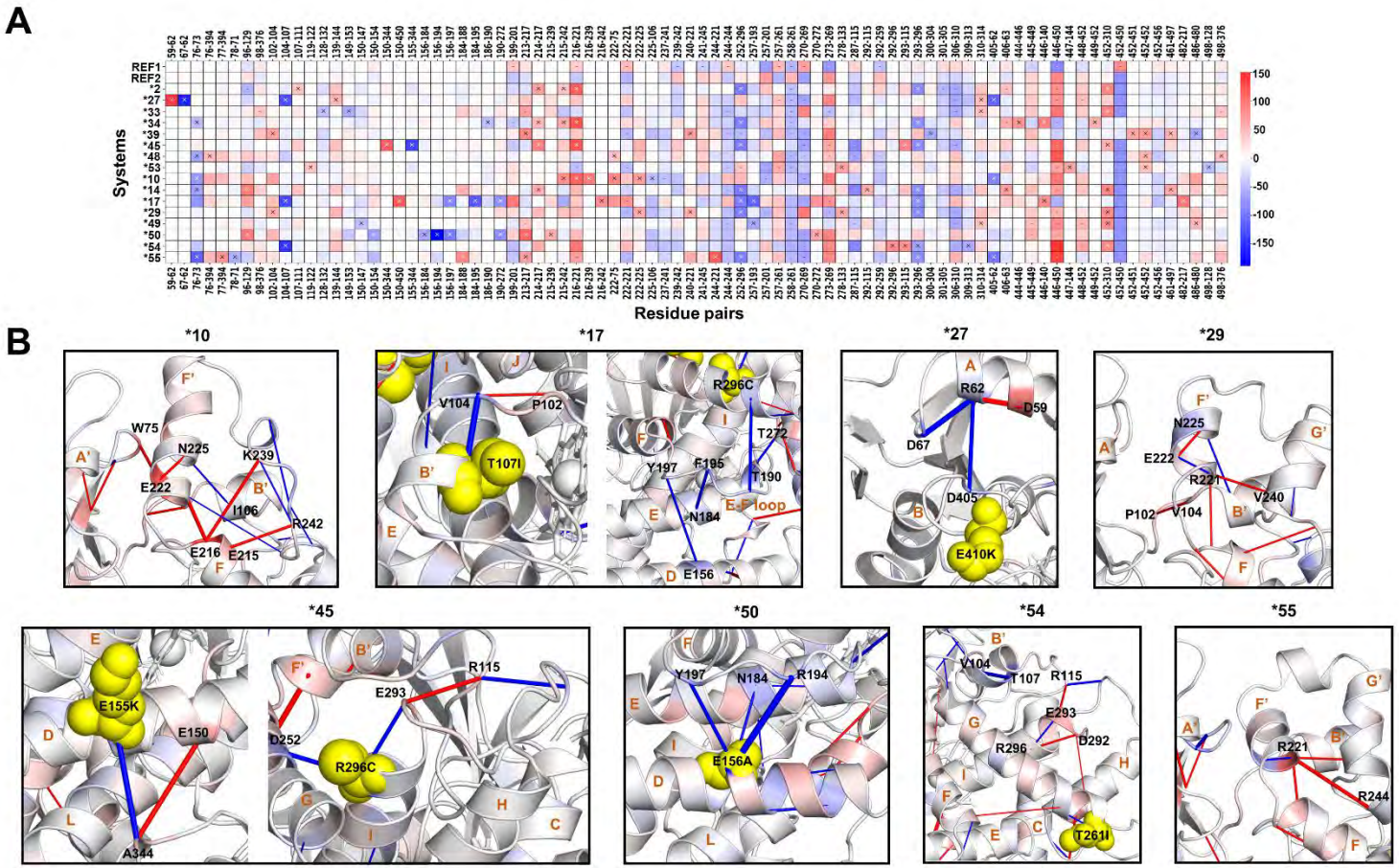


Figure 3-8: Hydrogen bond analysis. A) Residue pairs with top and bottom 3% ΔH-bond values compiled into a heatmap, with significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Values within top and bottom 3% ΔH-bond values and not significant are marked with a "-". B) Significant changes in hydrogen bonding noted between residue pairs for alleles *10, *17, *27, *29, *45, *50, *54 and *55. Variations in each allele shown by yellow spheres. For both A) and B), ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive).

3.3.4 Heme region analysis

The most vital part of CYP2D6 has been thought to be the heme area within the active site, as this is the site of catalysis [244], and studies thus far have shown no major structural changes or instability in this area for any of the CYP2D6 alleles. It was therefore important to conduct additional studies on this region, to support our hypothesis, that for the CYP2D6 to function as oxidases, whether with reduced or increased activity, the stability of the heme area is crucial. Here we aimed to analyze the heme and protein interactions in the presence and absence of variations, using the more stable part of the MD simulation (last 300ns).

We began with analyzing the hydrogen bonds between the heme group and the residues surrounding it (within 3.5 Å), in the references and alleles (Figure 3-9A). Overall, the number of hydrogen bonds formed in the references and all alleles were similar, with no major changes observed. Next, we analyzed how the heme and active site positioning was changing over the last 300 ns of the MD simulations, in both references and alleles, through measurements of the distance between the center of mass (COM) of the heme and the COM of the active site cavity (Figure 3-9B). The active site cavity residues 106, 110, 112, 117-122, 213, 216-217, 220, 244, 247-248, 297, 300-301, 304-305, 308-309, 370, 373-374, 483 and 484 were utilized in this calculation. In all alleles, the distance between COM of active site and heme was generally consistent with both reference distances, and any deviations present were minor. Finally, due to the presence of various non-covalent interactions (ionic and hydrophobic) between heme groups and protein residues, we analyzed these interactions of the heme with all residues within ~4 Å of it, using an ad hoc python script. As seen in Figure 3-9C, there were no major changes in contacts between the heme and

protein residues for references and alleles. In summary, all our analysis and findings of the heme region aligned with those of Skopalík *et al*., [245], who determined the heme binding cores of CYP2A6, CYP2C9 and CYP3A4 to be relatively rigid.

Structurally, there were no deviations noted within the heme area of alleles as shown by the heme studies here. However, our previous investigations using conformational clustering, DSSP and ΔRMSF showed reduced activity alleles to have substantial deviations in structure and stability/flexibility of the B' helix, the F-G region and the β1 sheet region that are near the substrate entrance tunnels. Thus, it was important to investigate these channels, and results of this analysis are discussed in the next section.

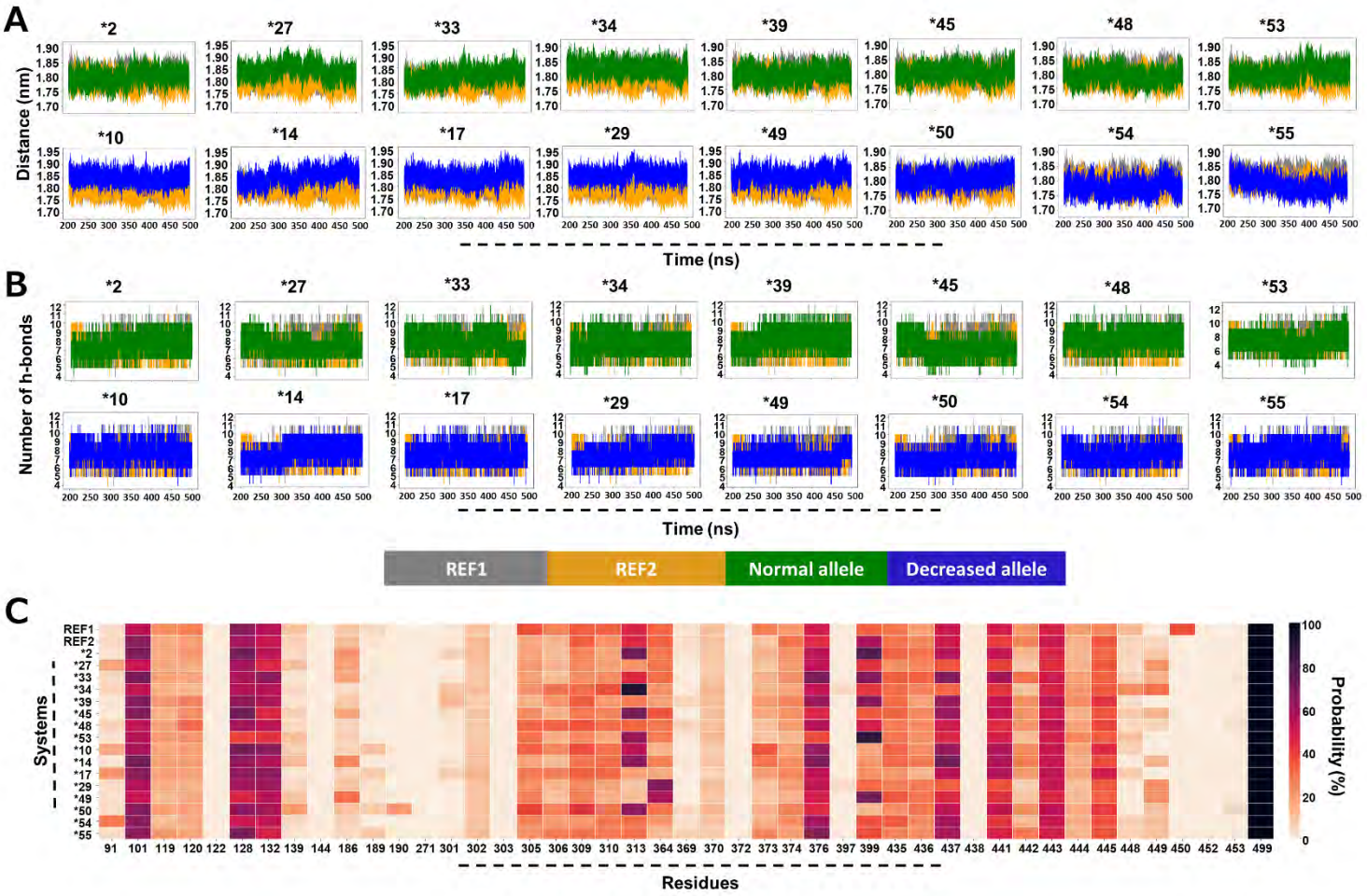


Figure 3-9: Analysis of heme region in CYP2D6 references and alleles. A) Line plots showing COM distances between the active site cavity and the heme B) Line plots displaying hydrogen bonds formed between heme and protein residues. In A) and B), REF1 is shown in grey, REF2 is shown in orange, normal activity alleles are shown in green and decreased activity alleles in blue. C) Contact analysis of heme with surrounding residues ≤0.4 nm from it. The degree of contact of each residue with the heme is shown by the colour scale from 0-1. Last 300 ns trajectories used to analyse heme environments.

3.3.5 Substrate channel analysis

Similar to other CYP enzymes, CYP2D6 active site is connected to the enzyme's surrounding environment by substrate access tunnels, that serve as pathways for substrates to enter the active site and interact with its catalytic center [61, 246, 247]. The entrance sites of the most prominent tunnels known in CYP2D6 are in the areas where secondary structure and conformational changes were noted with other analysis within this chapter, especially in reduced activity alleles. Therefore, it was crucial to decipher whether these changes affected tunnel structures in any way. A channel analysis tool, CAVER 3.0, was thus employed, to first identify major tunnels within the CYP2D6 structure used in this research (that connects the active site to the enzyme's peripheral), then measure the bottleneck radii, which is the narrowest part of an enzyme tunnel, to assess whether they were in a more open or closed state. This analysis was carried out using snapshots of the whole MD simulation (500 ns), so as to observe all changes of tunnels throughout the simulations.

The most prominent channels identified in references and alleles were channels 2b, 2c, 2e, 2f and solvent channel (S), using the tunnel nomenclature from Cojocaru *et al*., [61]. Their location are described in Table 3-2 and shown in Figure 3-10A. The entrance sites to the identified tunnels were close to CYP2D6 SRS that are known to recognize and bind substrates specific to CYP2D6 [132, 189]. Entrance sites of channels 2b, 2c, and 2e are located near SRS1 which is within the B' region and its flanking area. Additionally, channels 2c and 2e are near SRS3 and SRS4, located at the N-terminal ends of the G and I helices respectively. The solvent channel and channel 2f entrances are in close proximity to SRS2 and SRS6, which are found at the C-terminal end of the F helix and the β4 sheet turn (Figure 3-10A).

Table 3-2: Prominent substrate channels identified within CYP2D6 references and alleles.

Channel name	Location of channel within the enzymes	Substrate recognition site close to channel
2b	Between the B–B′ loop and the β1-1 and β1-2 sheets.	SRS1 (B' and flanking area).
2c	Between the G and I helices and the B′ helix.	SRS1 (B' and flanking area), SRS3 (N-terminal end of Helix G) and SRS4 (N-terminal half of Helix I) .
2e	Between the B-B’ loop, B’ helix and the B’-C loop.	SRS1 (B' and flanking area), SRS3 (N-terminal end of Helix G) and SRS4 (N-terminal half of Helix I) .
2f	Between the F′ helix/F–G loop, and the β4 sheet	SRS2 (C-terminal end of Helix F) and SRS-6 (turn of β4 sheet).
Solvent (S)	Between the F and I helices and the β4 sheet region.	SRS2 (C-terminal end of Helix F) and SRS6 (turn of β4 sheet).

When analyzing tunnel presence in each snapshot of each MD simulation (a tunnel was present in a snapshot if its bottleneck radii was at least 0.9 Å), it was noted that channel 2b was the most frequently present channel in snapshots of references and most alleles, with the exception of alleles *10, *17, *48 and *55, that had channel 2f with high frequency in snapshots (Table S6). This suggests channel 2b to be the main substrate channel for references and alleles *2, *27, *33, *34, *39, *45, *53, *14, *29, *49, *50 and *54, and likely preferred for entry and exit of substrates, while alleles *10, *17, *48 and *55 preferred channel 2f. In studies conducted previously, channel 2b and 2f were determined to be primary substrate access and egress channels within CYP2D6 and its alleles, and our findings aligned with these observations [61, 185, 186].

The bottleneck radii for all systems throughout the simulations was then measured and represented as heatmaps (Figure 3-10B). A channel was considered open if its bottleneck radius in each snapshot was ≥ 1.4 Å for the greater part of the MD simulation. Therefore, the heatmaps in Figure 3-10B are a representation of the opening frequencies of tunnels in each system throughout the trajectories. It was observed that both references and most normal function alleles (*2, *27, *33, *34, *45 and *53) had their main preferred tunnel identified (channel 2b), predominantly open throughout the simulation, with an average bottleneck radius above 1.67 Å for the tunnel in all alleles (Figure S7). All other tunnels within these systems were more frequently closed throughout the simulation, with an average bottleneck radius of the tunnels in the simulation below 1.4 Å (Figure S7). These findings suggest channel 2b to be functional in both references and normal function alleles *2, *27, *33, *34, *45 and *53, with an open state to allow potential substrates in and out of the active site, with no restrictions.

For reduced activity alleles *10, *17 and *55, all tunnels were observed to be more frequently closed especially in the more equilibrated part of the MD simulations (last 300 ns) (Figure 3-10B), including the channel identified as the main substrate pathway within these alleles (channel 2f). This suggests access to the active site may be compromised in these alleles through any of the identified tunnels. In alleles *14, *29, *50 and *54, their main substrate pathway (channel 2b) was observed to be more frequently open, however, channel 2f was also noted to be more frequently open in allele *14, or partially open in alleles *29, *50 and *54 throughout the MD simulations. Additionally, allele *50 had a third channel partially open throughout the MD

simulation (solvent channel). These findings suggest alleles *14, *29, *50 and *54 adopted a more open conformation throughout their simulations, with more than one tunnel more frequently open or partially open simultaneously, differing from what was observed in references and normal activity alleles. Wang *et al*., [102] conducted studies that showed the ability of CYP2D6 in binding two substrates synchronously when it is in an open conformation. In their study, one substrate bound to the active site, while the other bound to an initial binding site within channel 2b identified to bind substrates upon entering the enzyme (residue E222) [102]. For catalysis of a substrate to occur, CYP2D6 must be in a closed conformation, which is achieved when a substrate binds to the active site and triggers a rearrangement of residues and their side chains (including the side chain of residue E222) [3, 102]. Therefore, if a substrate is bound to E222, the enzyme may not adopt a closed conformation, and catalysis may not occur, potentially resulting in reduced enzyme activity. CYP2D6 alleles *14, *17, *29, and *55, which are reported here to have altered opening frequencies, also share the R296C variation occurring in the unique motif 25 (located on the N-terminal half of the I helix). This confirms what was discussed in chapter 2 of this thesis, that the presence of this variation in CYP2D6 can affect substrate access to the active site.

It was interesting to note normal activity alleles *39 and *48 having all tunnels more frequently closed throughout the simulation, resembling those of reduced functionality, and reduced function allele *49 to have its more dominant channel more frequently open and all others more frequently closed, resembling that of the references and normal activity alleles. This was expected as we did note significant conformational and secondary structure changes within alleles

*39 and *48 in regions where substrate entrance sites were located, and minimal changes in allele *49.

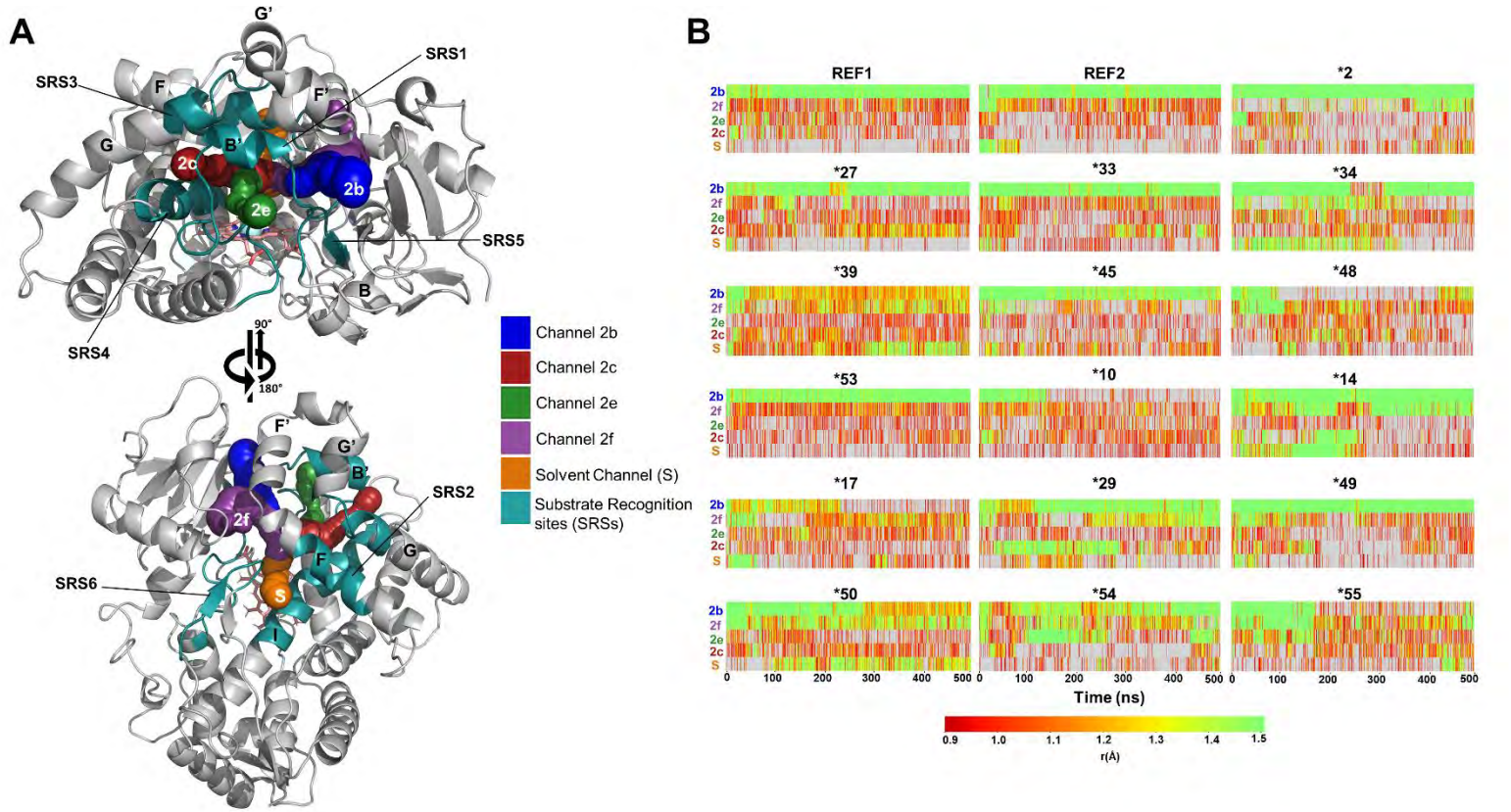


Figure 3-10: CYP2D6 tunnel analysis. A) Spatial representation of identified channels shown in CYP2D6 REF1. Channels are shown as van der Waals representations and named using nomenclature defined by Cojocaru *et al*., [61]. Channel 2b is coloured blue, channel 2c is coloured red, channel 2e is coloured green, channel 2f is coloured purple and solvent channel is coloured orange. Location of SRSs is also shown and coloured in teal. B) Heatmaps showing time evolution of channel bottleneck radius for channel 2b, 2f, 2e, 2c and the solvent channel. Snapshots were taken after every 2ns for the 500ns MD simulation run for all systems, and the total snapshots produced for analysis were 1001 including snapshot at 0ns. The colour map ranges from 0.9 Å to 1.5 Å or greater, with the lowest radii coloured in red and the highest radii coloured in green. Grey shows no channel with a bottleneck radius ≥0.9 Å was identified for that specific time in each instance.

Overall, through channel analysis, we deduced that alleles with reduced function had altered tunnel opening frequencies compared to the reference and most normally functioning alleles. In the reference systems, one main tunnel remained frequently open throughout the simulation, while all others were more frequently closed. This observation was reflected in normal activity alleles. However, variations in the alleles with reduced function caused the main substrate channels to either be more frequently closed or have multiple channels more frequently open or partially open simultaneously, resulting in a more open protein conformation that may be less favorable for substrate metabolism.

3.3.6 DRN analysis

Structurally, variation effects have been observed in this study, however, to further investigate these variation effects on CYP2D6 at a local level, DRN analysis was utilized. This method determines how variations can affect communication within a protein. Here, we employed the DRN metrices: *Betweenness centrality* (*BC*), *Closeness centrality* (*CC*) and *Eigenvector centrality* (*EC*) to identify any changes within communication networks of CYP2D6 effected by variations, on the more equilibrated part of reference and allele MD simulations (last 300 ns).

In this study, we calculated these different DRN metrics for all references and alleles, using MDM-TASK-web service [222], then thereafter, with each metric computed, we averaged the values of the two references, and subtracted this from the DRN metric values of the references and alleles from the average reference values obtained, to acquire ΔBC, ΔCC and ΔEC. The top and

bottom 3% Δ*BC*, Δ*CC* and Δ*EC* were then obtained and from this, and significantly changing values were noted (marked by an "X" in the respective heatmaps) and used in the analysis to follow. To highlight the regions with high metric values within the references, for each metric, the top 6% of the averaged reference values was obtained and mapped to a CYP2D6 reference structure, and we refer to this as REF_AV in this study.

3.3.6.1 *Betweenness centrality*

Betweenness centrality was the first metric used to analyze the communication networks within CYP2D6. The residues with the top 6% averaged *BC* values for REF_AV were mapped to a reference structure in Figure 3-11A, and residues in references and alleles with top and bottom 3% Δ*BC* values were shown in a heatmap (Figure 3-11B) and mapped onto their corresponding structures obtained from conformational clustering in Figure 3-11C and Figure S8.

In the REF_AV structure, we observed residues with high averaged *BC* to be mapping to the A, C, E, K, K', L helices, β4 sheet region, C-terminal I helix, K''-L loop and loop regions before β1-4 and after β1-3 (Figure 3-11A). Notably, some of these residues were occurring in important motifs identified in Chapter 2. Residues 435, 437 and 443 in the K''-L loop and residues 452 and 453 in the L helix were all located within motif 1, while residues 360 and 363 in the K helix were within motif 3. Residues 55 (A helix) and 135 (C helix) were occurring in motif 11 and motif 10 respectively, and 185 and 186 were within motif 6 motif. The residues 306, 307, 308, 309, 310, 312, 313 and 315 on the C-terminal side of the I helix were on motif 5. All these motifs were

conserved throughout the first three CYP families as seen in Chapter 2, with motif 1 containing the highly conserved and important heme-binding domain, motif 3 with residues that build important salt bridges which form the final tertiary structure of the enzyme, and motif 5 containing the oxygen binding and activation motif. High averaged *BC* within these residues signifies high usage of these residues in inter-protein communication, confirming their importance in CYP2D6.

To then determine variation-induced effects on the communication of residues within alleles as compared to the references, Δ*BC* was employed. We did note significant changes in averaged *BC* between the references and alleles in many regions within the proteins, including residues 185, 306-308, 310, 312, 360, 363, 443, 452 and 453 noted to have high averaged *BC* within REF_AV, signifying a change in inter-residue communication within these areas caused by variations in the alleles. But interestingly, in reduced function alleles only, there were additional significant changes in inter-residue communication within the B' helix and the loop between β1-1 and β1-2 important for substrate access (Figure 3-11B and C).

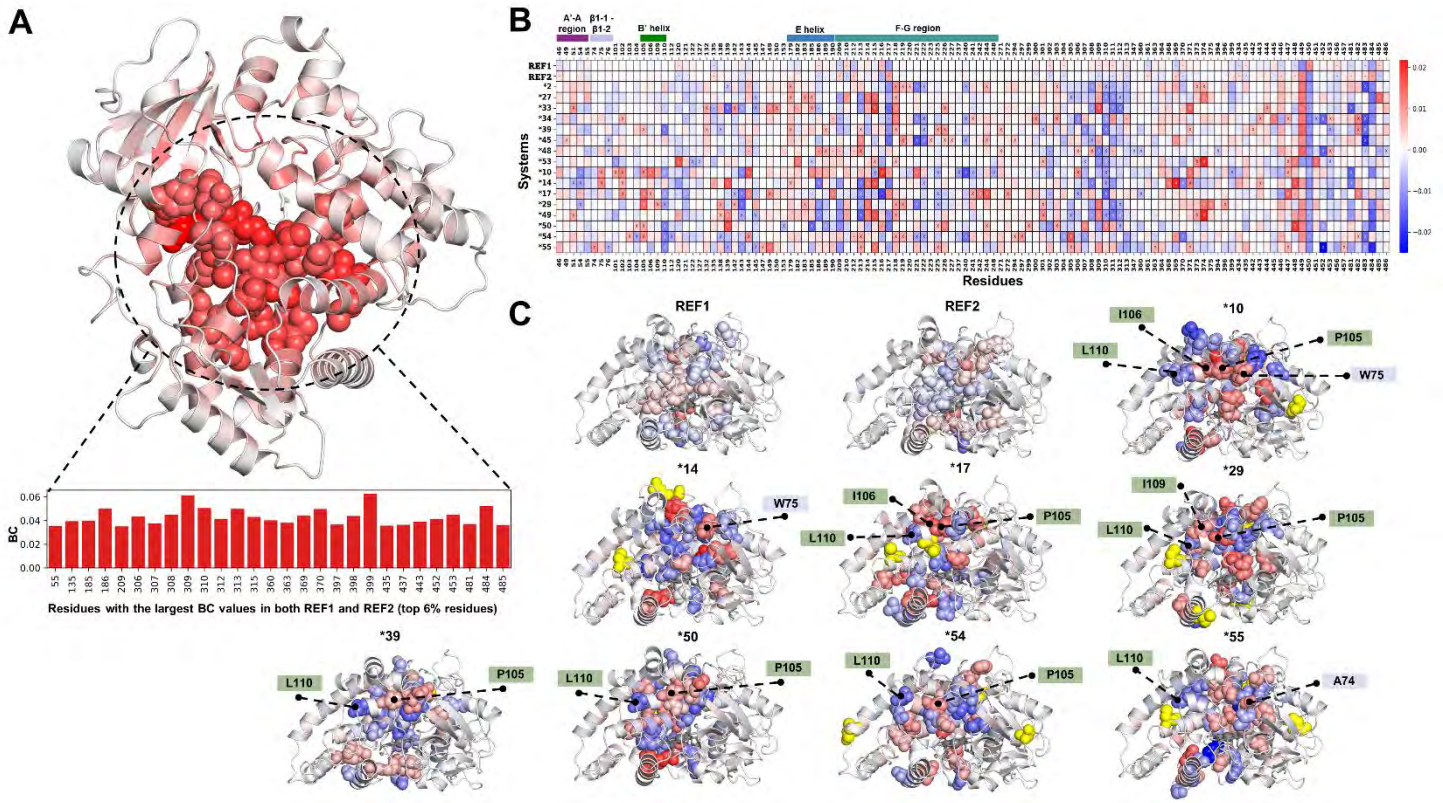


Figure 3-11: REF_AV and Δ*BC* analysis. A) Residues with top 6% averaged *BC* values for REF_AV mapped to a reference structure, shown as spheres and coloured in red, and shown in a bar plot below the structure. B) Residues with top and bottom 3% Δ*BC* values compiled into a heatmap. Significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Top and bottom 3% ΔBC values that are not significant are marked with a "-". C) Reduced function alleles *10, *14, *17, *29, *50, *54 and *55, and normal function allele *39 with significantly changing averaged *BC* values for residues within the B' helix and/or the loop between β1-1 and β1-2 shown. REF1 and REF2 also shown with no changes in averaged *BC* values occurring in the regions marked in the reduced function alleles. Changes in communication within the B' helix shown with green labels and changes noted in the loop between β1-1 and β1-2 shown with light blue labels. Variation positions for alleles shown in spheres and coloured in yellow. For both B) and C), the values range from positive to negative. Residues with a loss in averaged *BC* within the alleles are coloured in blue, and residues with gain in averaged *BC* in alleles are coloured in red.

In reduced activity alleles *10, *17, *29, *50, *54 and *55, there was a significant change in averaged *BC* as compared to the references in residues 105 and 110 within the B' helix (only residue 110 in *55) (Figure 3-11B and C). Additional to these residues, *10 and *17 each had a significant change in averaged *BC* within residue 106, and *29 in residue 109, both of which also occur in the B' helix. All alleles mentioned thus far with inter-residue communication changes within the B' helix contain the variation S486T (except *50), and when observing this variation in *39, where it occurs alone, significant changes in inter-residue communication within the B' helix (residues 105, and 110) was noted. Furthermore, in reduced activity alleles *10, *14 and *55 all sharing the variation S486T, a significant change in averaged *BC* was observed in residues 74 (*55) and 75 (*10 and *14) within the loop between β1-1 and β1-2 , signifying this variation not only affects inter-residue communication changes within the B' helix, but also this β1-1 and β1-2 loop region. This finding was not observed in normal functioning alleles. The results obtained here can be aligned to the structural changes we observed in secondary structure analysis, where we noted more significant changes in secondary structure and flexibility within reduced function alleles, in regions such as the B' helix, the F-G region and the β1 sheet region (β1-1 and β1-2).

3.3.6.2 *Eigenvector centrality*

The next DRN metric to be explored in the network analysis of CYP2D6 references and alleles is *Eigenvector centrality*. The residues with top 6% averaged *EC* values for REF_AV are shown in Figure 3-12A, while residues with the top and bottom 3% Δ*EC* values in alleles were represented in a heatmap shown in Figure 3-12B, and regions with the highest changes in averaged *EC* within alleles as compared to both references mapped to their corresponding structure obtained

from conformational clustering program in Figure 3-12C. All other alleles with less notable changes in averaged *EC* were mapped to their corresponding conformational cluster structures and shown in Figure S9.

As seen in Figure 3-12A, residues with the top 6% averaged *EC* values for REF_AV were mapping to helices B, K' and L, and the β-sheet rich region in CYP2D6 (β1-1, β1-2, β1-4, β2-1 and β1-3). Motif analysis in Chapter 2 also identified these regions as highly conserved in CYP2D6, with the presence of conserved motifs 7, 9, 11 and 14 in the regions. The presence of both high averaged *EC* values and highly conserved motifs in these regions indicates their relevance/importance in the functionality of the enzyme. In most of the alleles, there was an increase in averaged *EC* within these regions noted (except in alleles *27, *50 and *53 that had a decrease in averaged *EC*). However, in reduced functioning alleles *10, *14, *17, and *55, the increase in averaged *EC* was more significant, indicating a more notable change of information flow in this group of residues, as compared with the references and most alleles with normal activity (Figure 3-12B and C).

Other regions identified to have high average *EC* in REF_AV (even though they are not included in the top 6% of the reference data) are the C-terminal half of I helix (residues 308-316), the J-J' loop (residue 345), the K helix (residues 360 and 364), and the L helix (451-459) (Figure 3-12A). These regions were identified to be highly conserved in CYP2D6, as indicated in Chapter 2, with the presence of motif 5 in the C-terminal half of I helix, motif 3 in the K helix and motif 1 and 14 in the L helix (motif 1, 3 and 5 residues were also noted to have high averaged *BC*). A

decrease in averaged *EC* was observed within these regions within most alleles (except alleles *27, *50 and *53 that were observed to have an increase in averaged *EC*). A significant decrease in average *EC* in these areas was however noted in reduced functioning alleles *10, *14, *17 and *55, indicating a shift in importance of these region in these alleles (Figure 3-12B and C). Through all the analysis carried out in this study, these regions were stable, with minimal changes noted in all alleles, however, this notable change in information flow in the alleles *10, *14, *17 and *55, could have been triggered by the major structural changes noted in these alleles.

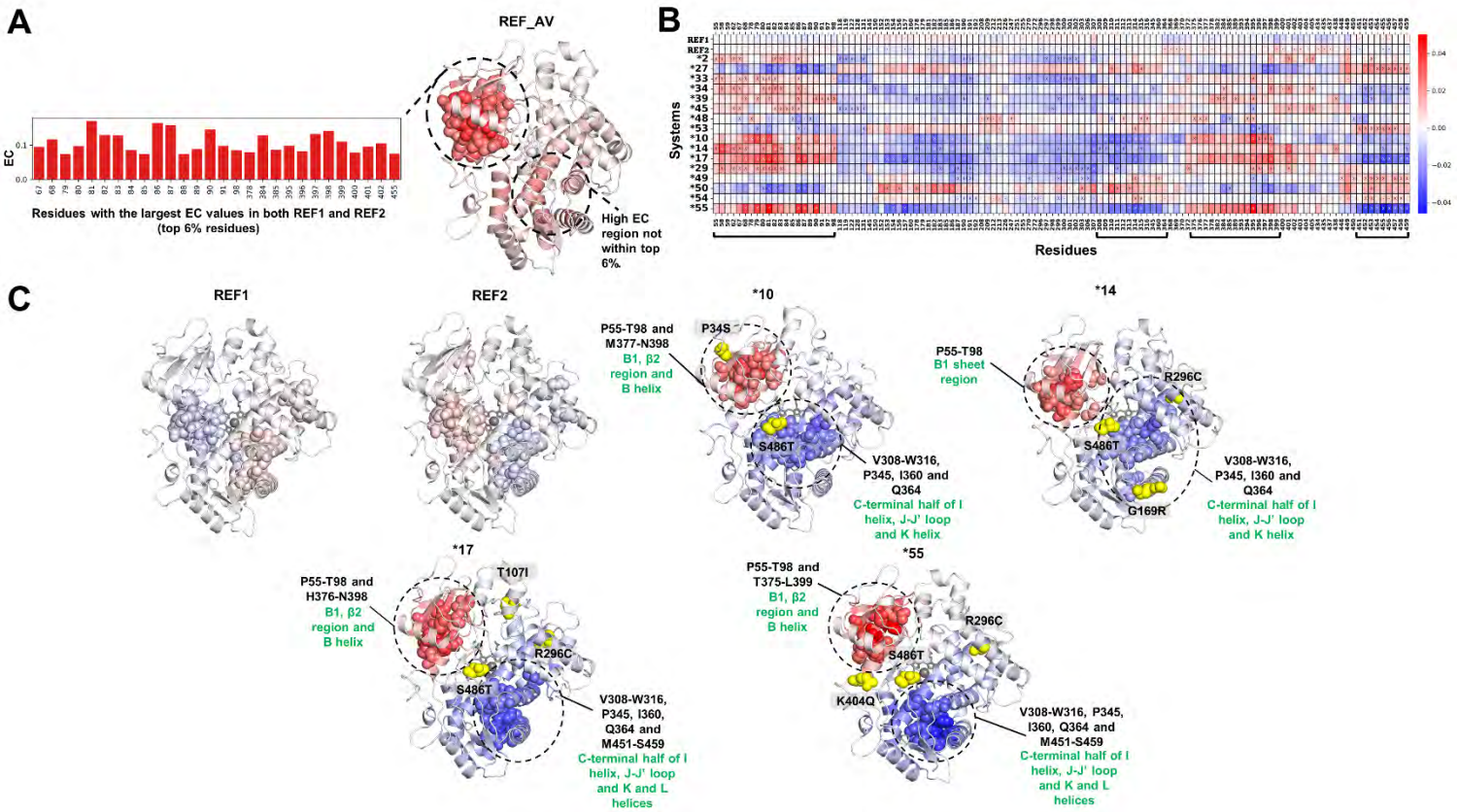


Figure 3-12: REF_AV and Δ*EC* analysis. A) Residues with top 6% averaged *EC* values for REF_AV were mapped to a reference structure, shown as spheres and coloured in red, and shown in a bar plot beside the structure. B) Residues with top and bottom 3% Δ*EC* values compiled into a heatmap. Significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. Top and bottom 3% Δ*EC* values that are not significant are marked with a "-". Regions with larger changes in averaged *EC* within reduced function alleles highlighted with black brackets. C) Reduced function alleles *10, *14, *17, and *55 with more significantly changing averaged *EC* shown. REF1 and REF2 also shown for comparison. Variation positions for alleles shown in spheres and coloured in yellow. For both B) and C), the values range from positive to negative. Residues with a loss in averaged *EC* within the alleles are coloured in blue, and residues with gain in averaged *EC* in alleles are coloured in red.

3.3.6.3 *Closeness centrality*

The final DRN metric used to analyze references and alleles in this study is *Closeness centrality*. The residues with the top and bottom 3% Δ*CC* values were represented in a heatmap shown in Figure 3-13B and mapped onto their corresponding structure obtained from conformational clustering program in Figure 3-13C. Also, residues with the top 6% averaged *CC* values for REF_AV were mapped to a reference structure in Figure 3-13A and shown in a bar plot below the structure. In our previous studies, we showed that high averaged *CC* values occur mainly in the core of the protein, and results in this study demonstrate a similar pattern [226, 227, 248]. Residues with high averaged *CC* were mapping towards the core of the protein where the heme group (catalytic site) is located (Figure 3-13A). Because substrate oxidation takes place in this critical region of CYP2D6, averaged *CC* has highlighted the centrality of this region in the protein function. High *closeness centrality* at this site also implies a closely bound network of active site residues in CYP2D6, which may influence protein activity. When we analyzed the Δ*CC* results, no major differences were noted between the references and alleles in the core of the protein (Figure 3-13B and C), indicating no major change in centrality of this highly important core of CYP2D6 for function, regardless of any variations introduced. This correlated with our heme analysis, were we did not find any major structural changes immediately around the heme area. It was interesting to note however, minimal changes in averaged *CC* on the outside of the core of the protein for allele *10 (significant increase and decrease of averaged *CC* in residues 229-232, 235, 237, 238, 240 and 241).

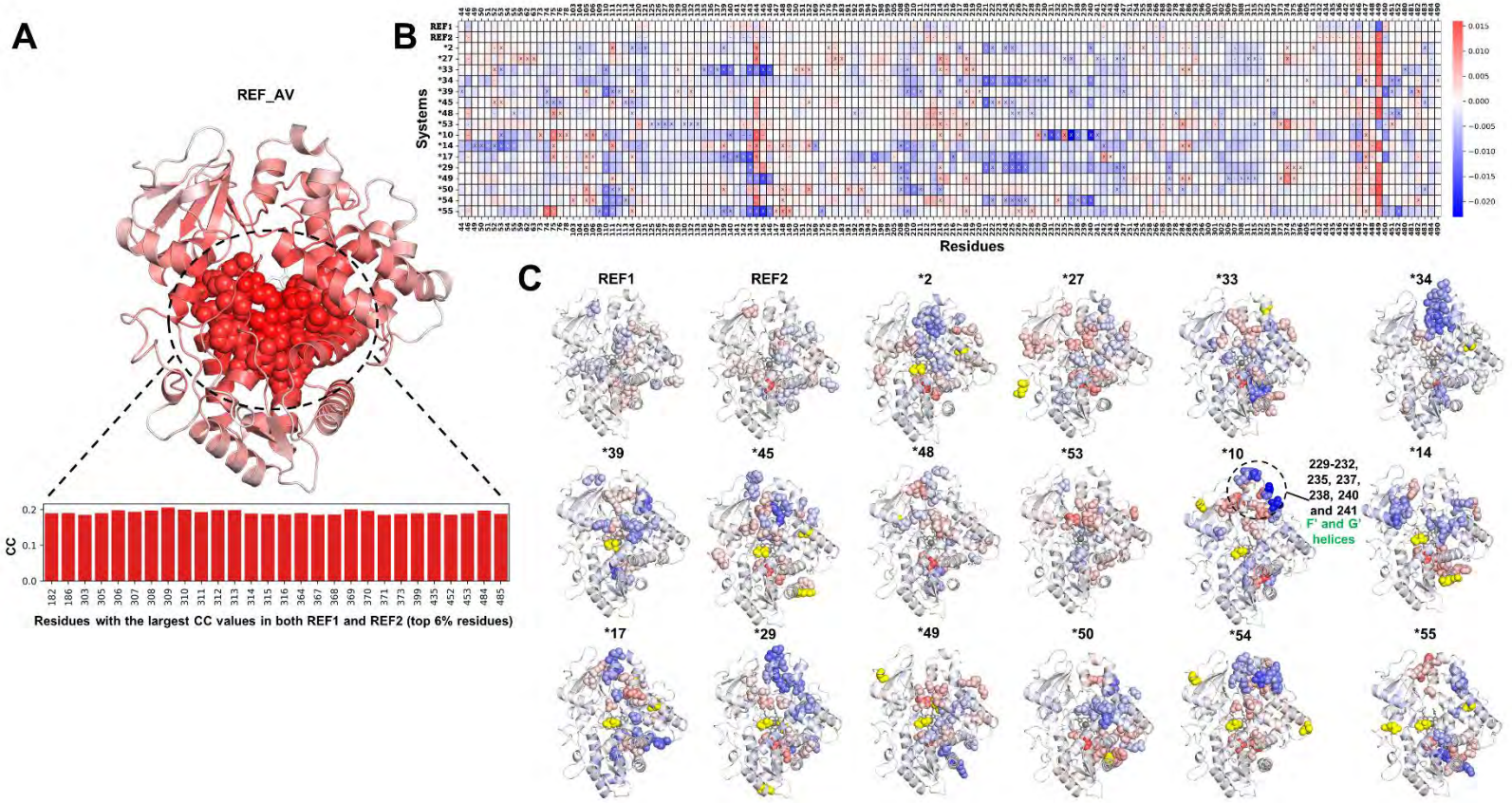


Figure 3-13: REF_AV and Δ*CC* analysis. A) Residues with top 6% averaged *CC* values for REF_AV mapped to a reference structure, shown as spheres and coloured in red, and shown in a bar plot below the structure. Residues with top and bottom 3% Δ*CC* values represented in a heatmap (B) and mapped onto their corresponding structure obtained from conformational clustering program and shown as spheres (C). In B), significant values (outliers not within 3SD of the mean of the reference data) are marked with an "X" in the heatmap. Top and bottom 3% ΔBC values that are not significant are marked with a "-". In B) and C) the values range from positive to negative. Residues with a loss in averaged *CC* within the alleles are coloured in blue, and residues with gain in averaged *CC* in alleles are coloured in red. REF1 and REF2 also shown for comparison and variation positions for alleles shown in spheres and coloured in yellow.

3.4 Conclusion

Polymorphisms within drug metabolizing enzymes have subsequently affected drug response in humans, and this has raised many complications in eradicating and treating many diseases in the world today, as the drugs either have little to no therapeutic effect or cause adverse drug reactions in patients. The alleles of these drug metabolizing enzymes are poorly understood, thus research on them was essential to forward the cause of personalized treatment that will essentially bring about optimum drug response in all patients. In this study, we mainly aimed to decipher variation mechanisms within a CYP2 family drug metabolizing enzyme, CYP2D6, thereby identifying variation effects that ideally lead to reduced activity of this enzyme, and potentially utilize these observations to predict reduced function of other CYP2D6 alleles with unknown activity in future studies.

We performed comprehensive computational analysis of CYP2D6 and its alleles with known activity, taking note to include both normal (*2, *27, *33, *34, *39, *45, *48 and *53) and reduced (*10, *14, *17, *29, *49, *50, *54, *55) function alleles. Molecular dynamic simulations were then run, and traditional post MD analysis including RMSD and Rg, were utilized to observe any structural stability changes due to variations in alleles, and the geometric clustering module employed to check conformations of alleles once stable/equilibrated. Changes in secondary structure were checked using the DSSP tool and ΔRMSF, and changes in hydrogen bonding throughout the enzymes were checked using ΔH-bond analysis. The heme of all references and alleles was analyzed by measuring the COM between the heme and active site, measuring the

hydrogen bonding around the heme, and calculating the number of contacts the heme had with the surrounding protein residues. Due to the structural changes observed in close proximity to the known substrate entrance sites of CYP2D6, tunnel analysis was carried out using CAVER 3.0 [231], to identify channels leading to the active site of CYP2D6, and identify any changes with opening frequencies of those tunnels that may have been affected by structural changes in regions around them. Finally, we employed DRN metric analysis [222, 248], to observe any changes within the communication of CYP2D6 alleles that may have been induced by variations.

Through the analysis, it was determined via RMSD analysis that several alleles with reduced functionality reached a more stable state significantly later in the MD simulation (between 150 and 200 ns) than references and normally functioning alleles, indicating reduced activity alleles to be more unstable. Once reaching equilibration, the reduced function alleles were observed to be in a more divergent conformation, as identified by conformational clustering, and contained more secondary structure and flexibility changes, determined by DSSP and ΔRMSF analysis, in the B' and E helix, the E-F loop, the F-G region, the A'-A region and β1 sheet region. Hydrogen bond analysis revealed that variations were causing changes in hydrogen bonding networks within the different alleles, which consequently led to structural and flexibility differences. The areas with structural changes were noted to be occurring in and around the substrate entrance sites reported in previous studies to occur in CYP enzymes [61], therefore tunnel analysis was carried out. Upon observation, the reduced activity alleles noted to have structural changes within the B' and E helix, the E-F loop, the F-G region, the A'-A region or the β1 sheet region, either had their dominant channels for substrate entry more frequently closed or multiple channels were more frequently

open or partially open simultaneously throughout the simulation, creating a highly open protein that is less favorable for metabolism of substrates.

It was evident the addition of variations to CYP2D6 was causing structural changes within the highly important regions for substrate entry especially within reduced activity alleles, and inevitably, protein communication networks may be affected. So, to fully understand the change in networks within these proteins, investigations at a local level using DRN metrics – averaged *BC*, *EC* and *CC* – were performed. Overall, DRN metrics, more specifically averaged *BC* and *EC*, demonstrated a shared relationship of identifying highly functional regions within CYP2D6 that could be related to motif analysis conducted in Chapter 2. The highly conserved motifs identified were regions observed to have high centrality, highlighting the importance of those regions in the function of CYP2D6. Δ*BC* and Δ*EC* then showed changed centrality within regions important for structural stability and substrate access and egress. With Δ*BC*, for all alleles, we noted a change in inter-residue communication throughout most regions within the enzymes, however in reduced function alleles only, there were further changes in the B' helix and the loop between β1-1 and β1-2. With Δ*EC*, we observed a substantial increase in averaged *EC* within the β-sheet rich region that included β1 sheets, located close to substrate access channels, and a substantial loss in averaged *EC* within the more conserved C-terminal half of I helix, the J-J' loop, the K helix, and the L helix mainly in reduced activity alleles, signifying a shift in communication/information flow from one region of CYP2D6 to another. Finally, with Δ*CC*, we noted minimal changes within the core of the protein (harboring the heme group critical in the oxidation of substrates) where we had observed high *CC* in the reference structures, signifying no major changes in centrality of this

region in protein functionality. This observation was aligning with no structural changes observed when heme analysis was carried out.

Interestingly in all analysis of allele *39, we noted it to be quite unstable with many secondary structure changes, imitating the behavior of reduced function alleles, while *49 had consistently shown to be as stable as the references in all analysis. We assigned function to the alleles in this study based on CPIC records provided in PharmVar [48], and upon observation of the clinical annotations of alleles *39, *48 and *49, they each had a low level of evidence (with a score of 3) that suggests the assignment of function was based on a single study noted, or several studies that failed to replicate the associated function. A study carried out by Stern *et al*., [249] also reported on limited data availability for *49, and suggested the reduced activity assignment of *49 may either be because of catalytic incompetence or reduced availability of functional protein.

In summary, we propose that variations within CYP2D6 consequently change hydrogen bonding networks within the alleles, destabilizing them, and ultimately causing structural and flexibility changes in and around substrate channel regions. The instability and changes in structure and flexibility are more pronounced in reduced activity alleles, as they result in altered channel opening frequencies that potentially limit access of substrates to the active site to undergo catalysis. Residue communication patterns within the substrate access channel regions are also affected as a result, with significant changes noted mainly in reduced activity alleles. Through these observations, we deduce four mechanisms that can be utilized to identify CYP2D6 alleles with

reduced functionality. We outline these four mechanisms in Table 3-3, and according to our observations using the normal and reduced activity alleles, we propose that if an allele passes at least three of the four mechanisms, the likelihood of reduced functionality is high.

All mechanisms identified through research conducted within this chapter have been through *in silico* experiments, and although the predictions may be good, it is important to state some of the limitations of this approach. Biological systems are complex in nature, and *in silico* experiments may not fully capture and accurately represent the complexities within the system. For instance, CYP2D6 is membrane bound, and binds NADPH-cytochrome P450 oxidoreductase (POR), to supply/transfer electrons to the CYP catalytic site during metabolism of a drug. These variables were not incorporated into the present study; therefore, the resulting predicted mechanisms should be interpreted with caution, as their validity cannot be fully ascertained.

Table 3-3: Summary of mechanisms potentially leading to reduced CYP2D6 activity due to genetic variations.

Function/ Activity of allele as assigned by PharmVar	Alleles	Instability and conformational changes of protein as measured by RMSD and Rg		Secondary structure variation and changes in flexibility of ≥3 protein regions (DSSP, conformational clustering and RMSF)								More frequently closed/ more than one frequently open or partially open channel simultaneously (Caver3.0)	Unique changes in inter-residue communication s within the B' helix and/or β1 sheet region (ΔBC)	
		Instability (equilibrated after ~200 ns)	More than one conformation visited in 500ns	β4 sheet	F-G region	E-F loop	E helix	B' helix	β1 sheet	A'-A loop			β1 sheet	B' helix
Normal	***2**							√						
	***27**				√			√						
	***33**				√			√						
	***34**									√				
	***39+**			√	√			√				√		√
	***45**				√	√		√						
	***48**								√		√	√		
	***53**				√									
Reduced	***10+**	√	√		√	√		√	√	√		√	√	√
	***14+**	√	√		√			√	√	√			√	
	***17+**	√	√			√		√	√	√		√	√	√
	***29+**	√	√		√		√	√	√			√	√	√
	***49**				√	√								
	***50+**	√	√		√	√		√	√	√		√		√
	***54+**	√	√		√			√	√	√		√		√
	***55+**	√	√		√		√	√	√	√		√	√	√

[+]Alleles observed to have three or more mechanisms identified causing reduced activity. Red boxes highlight the alleles where the mechanism occur and √ shows exact changes noted in allele.

Chapter 4

Predicting Reduced Functional Consequence of Afrocentric/ Sub-Saharan African CYP2D6 Alleles: Insights from Mechanistic Findings

Chapter overview

In this chapter, a range of computational tools were used to investigate Afrocentric CYP2D6 alleles from Sub-Saharan Africa, specifically identified in individuals from the Niger–Congo language speaking African region. At the time of this research, these alleles had no clinical functionality assigned to them. Therefore, utilizing the mechanisms identified to potentially result in reduced activity of CYP2D6 (chapter 3), reduced functional consequence was predicted for some of the Afrocentric alleles.

4.1 Introduction

The involvement of CYP2D6 in the biotransformation of numerous substrates/medications has positioned it as a crucial enzyme in drug metabolism, and variations within this enzyme could greatly impact the metabolism of many drugs. It is therefore important to understand effects of variations within this enzyme in all populations of the world. So far, pharmacogenomic studies to characterize its variants and star alleles, as well as determination of function related to each allele identified has been widespread in many populations around the world, however, African populations have yet to be sufficiently represented in these studies [250–253].

Compared to other superpopulations, Africa exhibits greater genetic diversity, and a lack of pharmacogenomic and pharmacogenetics research may cause significant implications for drug safety and efficacy in Africa [252, 253]. This is particularly concerning as Africa is endemic to some of the world's most lethal diseases, such as malaria and TB, both of which have high mortality rates within the continent. Reduced drug efficacy, increased toxicities, and adverse drug reactions may amplify the situation. Therefore, in this study, to further pharmacogenomic research for personalized medicine in Africa, we aimed to predict if any of the novel Afrocentric CYP2D6 alleles, with no clinical function assigned to them at the time of this research, had reduced activity. The mechanisms identified for reduced activity alleles in Chapter 3 were utilized, and the Afrocentric alleles (*149, *152, *153, *154, *155, *157, *159, *160, *162 and *163) were identified from PharmVar. In an effort to continue the work by Twesigomwe *et al*., [251], all Afrocentric alleles studied here were fully characterized by Wang *et al*., [252], using samples from

the 1000 Genomes Project [254] and genomic DNA from cell lines obtained from Coriell Institute for Medical Research (Camden, NJ, USA). The samples were from Niger–Congo language speaking African participants, specifically the Esan and Yoruba from Nigeria, Mende from Sierra Leone, Luhya from Kenya, and the Gambian Mandinka (Figure 4-1) [252]. Niger–Congo language predominates in the Sub-Saharan region of Africa (Figure 4-1), spanning from Dakar, Senegal in the west, to Mombasa, Kenya in the east, and southward to Cape Town, South Africa.

Figure 4-1: Geographic origin of Afrocentric alleles used in this study.

4.2 Methodology

Similar methodology to chapter 3 (section 3.2) was utilized in the research presented in this chapter, except for those stated here. In all analyses, the reference data from the previous study (chapter 3) was used for comparison with the data produced here, for the Afrocentric allele systems.

4.2.1 Afrocentric CPY2D6 allele selection

The Pharmacogene Variation Consortium (PharmVar) [48] which provides information on variations in CYP enzymes, was utilized to compile a list of recently identified novel Afrocentric CYP2D6 alleles for which the Clinical Pharmacogenetics Implementation Consortium (CPIC) has not yet assigned functional annotations. In total, 10 Afrocentric alleles (*149, *152, *153, *154, *155, *157, *159, *160, *162 and *163) were included for investigation, and their variation details are summarized in Figure 4-3.

4.3 Results and Discussions

In Chapter 3, mechanisms by which variations cause reduced activity in CYP2D6 alleles were identified, and in this present study, the main aim is to utilize these identified patterns to infer possible reduced functional consequence to newly detected CYP2D6 alleles. Mechanisms identified previously are summarized in Figure 4-2. It was observed that references and normal function alleles were observed to have less than 3 of the mechanisms identified to be occurring in them, and the heme area was stable, indicating normal activity can also be inferred to Afrocentric alleles that behave similarly Figure 4-2.

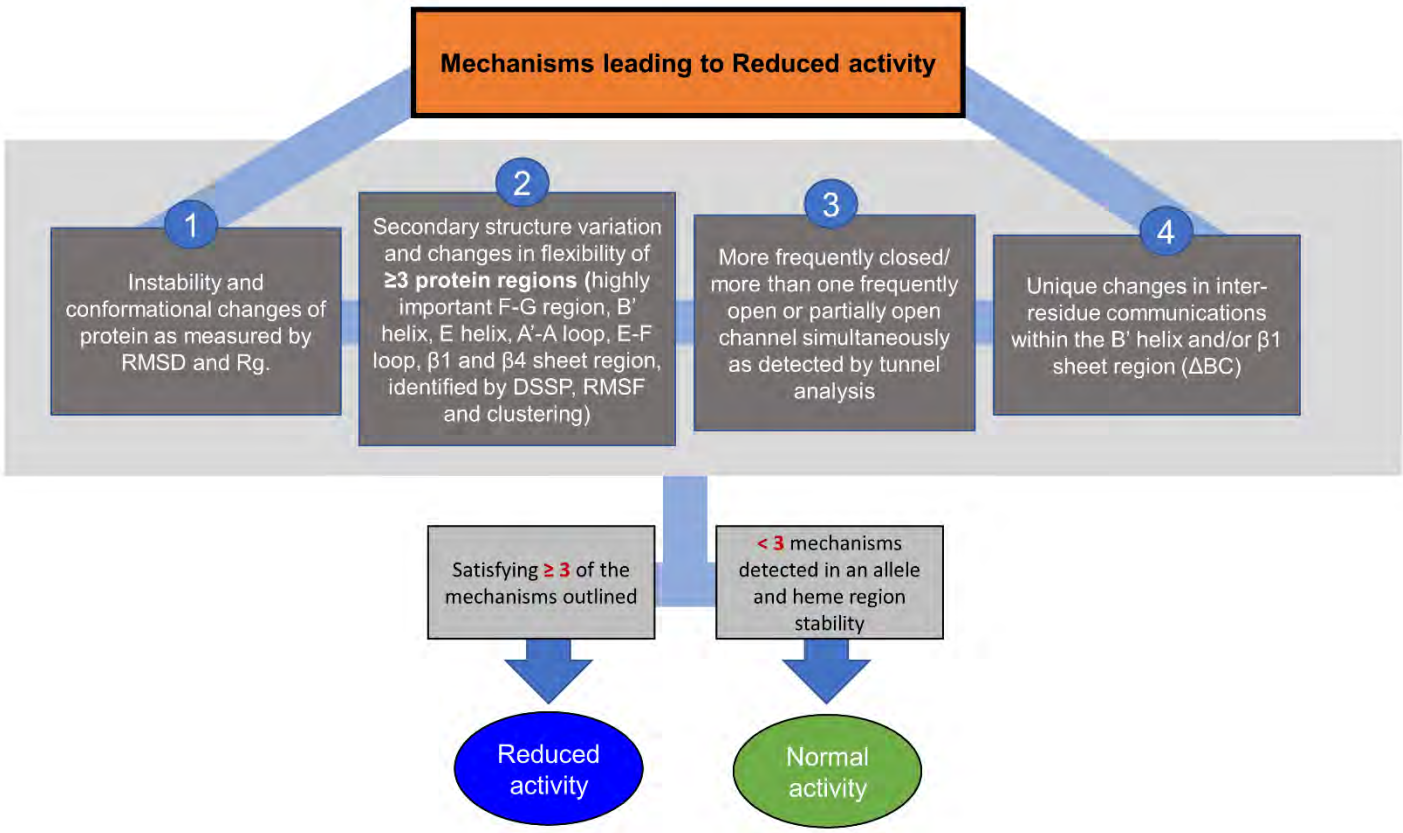


Figure 4-2: Summary of variant-induced mechanisms identified to potentially cause reduced activity in CYP2D6 alleles represented in a flow chart.

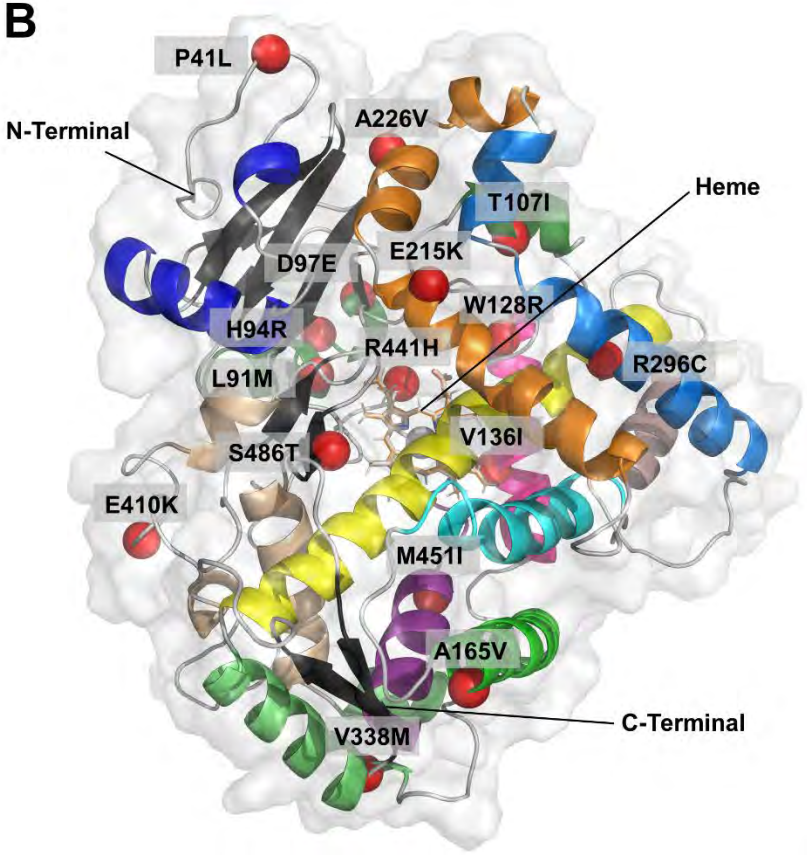
4.3.1 Selection of Afrocentric alleles with unknown clinical function, 3D structure generation, and MD simulations

Recently discovered Afrocentric CYP2D6 alleles with no known clinical function were retrieved from the PharmVar consortium [48] (Figure 4-3A), and it was noted that each allele was formed by the addition of a variation to an already existing allele with known function (This is referred to as a backbone allele in this chapter). Alleles CYP2D6*159, *160, *162, and *163 all have the normal function CYP2D6*2 backbone allele. Variations P41L, E410K, and W128R were added to form *159, *162, and *163, respectively, while variations L91M and H94R were both added to form *160. Alleles CYP2D6*149, *155, and *157 were formed by adding variations D97E, M451I, and A165V, respectively, to the decreased activity backbone allele CYP2D6*29. Allele CYP2D6*154 was formed by adding the variation R441H to the decreased activity allele CYP2D6*17. The only alleles with the reference CYP2D6 as a backbone were noted to be CYP2D6*152 and *153. The distribution of these variations throughout CYP2D6 is shown in Figure 4-3B. All alleles were created using the reference CYP2D6 structure and subjected to 500 ns MD simulations. Similar *in silico* tools from previous work in chapter 3 were used for analysis.

A

Enzyme	Backbone	Novel Afrocentric allele	Variations
CYP2D6	*1 (WT)	*152	A226V
		*153	E215K
	*2	*159	P41L, R296C, S486T
		*160	L91M, H94R, R296C, S486T
		*162	R296C, E410K, S486T
		*163	W128R, R296C, S486T
	*17	*154	T107I, R296C, R441H, S486T
	*29	*149	D97E, V136I, R296C, V338M, S486T
		*155	V136I, R296C, V338M, M451I, S486T
		*157	V136I, A165V, R296C, V338M, S486T

B



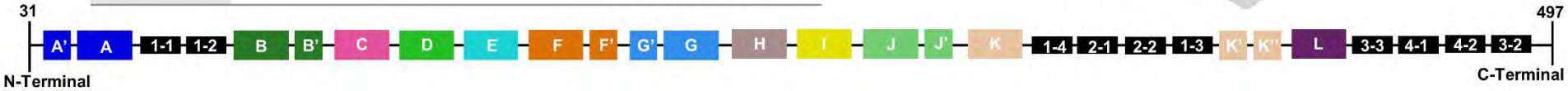


Figure 4-3: Variations in novel CYP2D6 Afrocentric alleles. A) Table showing all Afrocentric alleles studied in this article with their respective backbone allele and variations. Variation written in red and underlined signify new variation added to backbone allele to form Afrocentric allele. B) Distribution of variations in a CYP2D6 reference structure with different helices denoted by the different colours with a key at the bottom.

4.3.2 RMSD and Rg calculations to assess stability, conformational variation and compactness of Afrocentric alleles

Instability, compactness and overall conformational changes in the alleles throughout the simulations were measured using RMSD and Rg. The resulting data for RMSD and Rg was presented using line plots, violin plots, and additionally for the RMSD data, all-vs-all heatmaps. RMSD calculations were undertaken to evaluate the stability of CYP2D6 allele structures and compare different conformations visited during MD simulations. As seen in Figure 4-4, CYP2D6 alleles *149 and *163 reached equilibrium rapidly, similar to the reference systems, and displayed greater stability than all other alleles as they showed to exhibit unimodal distribution. This indicated alleles *149 and *163 both sampled one main conformation throughout the MD simulation. In contrast, alleles *152, *153, *154, *155, *157, *159, *160, and *162 exhibited greater conformational diversity, as evidenced by the bimodal distribution of RMSD values shown in Figure 4-4C, indicating at least two conformations were visited during the MD simulation. These observations were also illustrated in the all-vs-all RMSD heatmaps presented in Figure 4-4B. These alleles then reached a more stable state at ~ 200 ns, which was much later than the reference systems (Figure S10).

The compactness of the CYP2D6 Afrocentric alleles was assessed throughout MD simulations with Rg calculations, to evaluate whether variations influenced the protein's overall compactness. As shown in Figure 4-5, the compactness of all alleles did not greatly differ from the reference alleles, indicating that the variations did not lead to major changes in the protein's overall compactness.

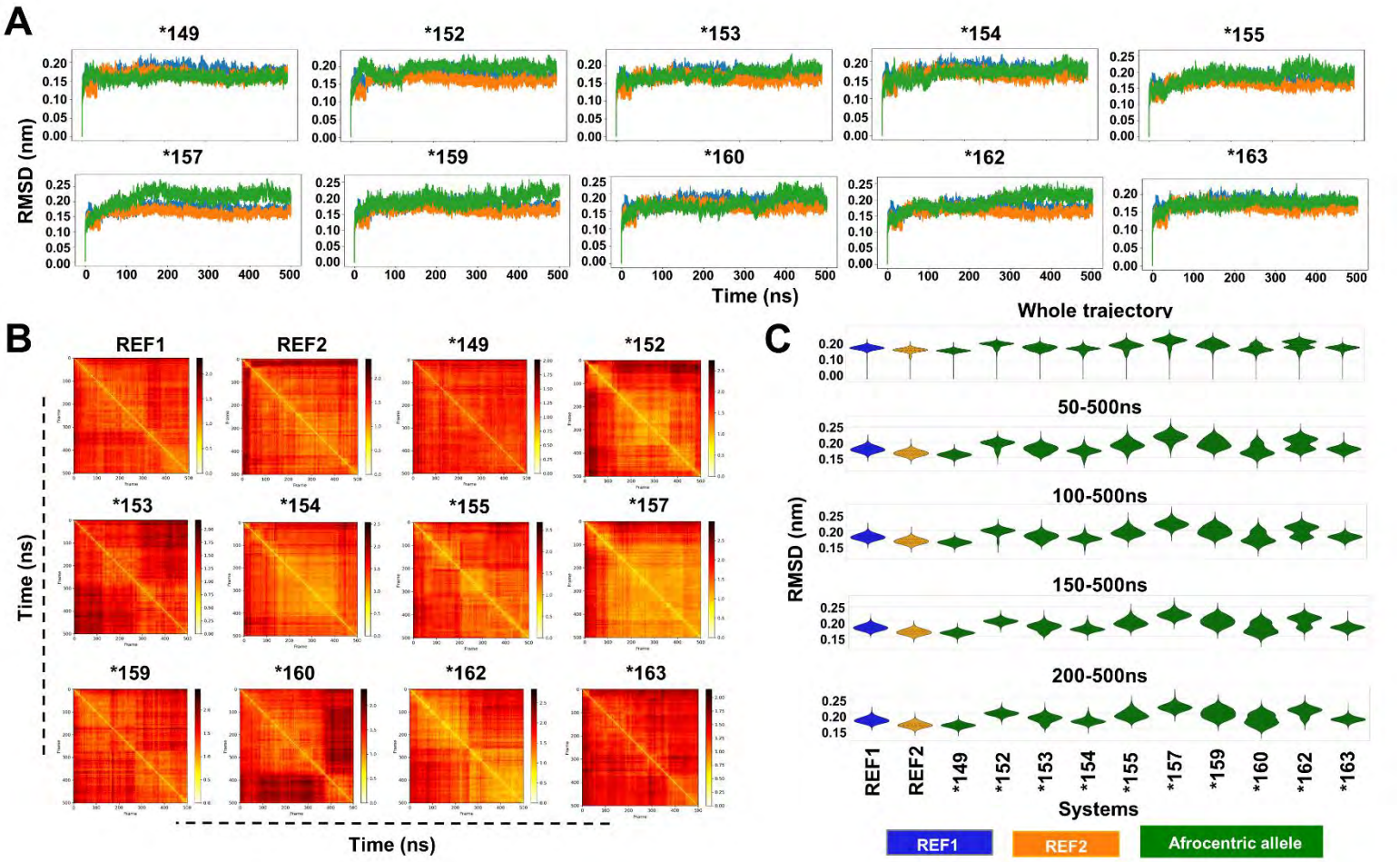


Figure 4-4: CYP2D6 reference and Afrocentric alleles' RMSD analysis. A) RMSD vs time-line plots for the whole MD simulation (500ns) for references and alleles. B) All vs all RMSD heatmaps for whole MD simulations. The x and y axes show the frames at different time stamps and the degree of variation from low to high, are coloured from white, through yellow and red, to black. C) RMSD violin plots for stages of MD simulation (0-500ns/ whole trajectory, 50-500ns, 100-500ns, 150-500ns, 200-500ns). For A) and C), REF1 is shown in blue, REF2 is shown in orange and CYP2D6 Afrocentric alleles are shown in green.

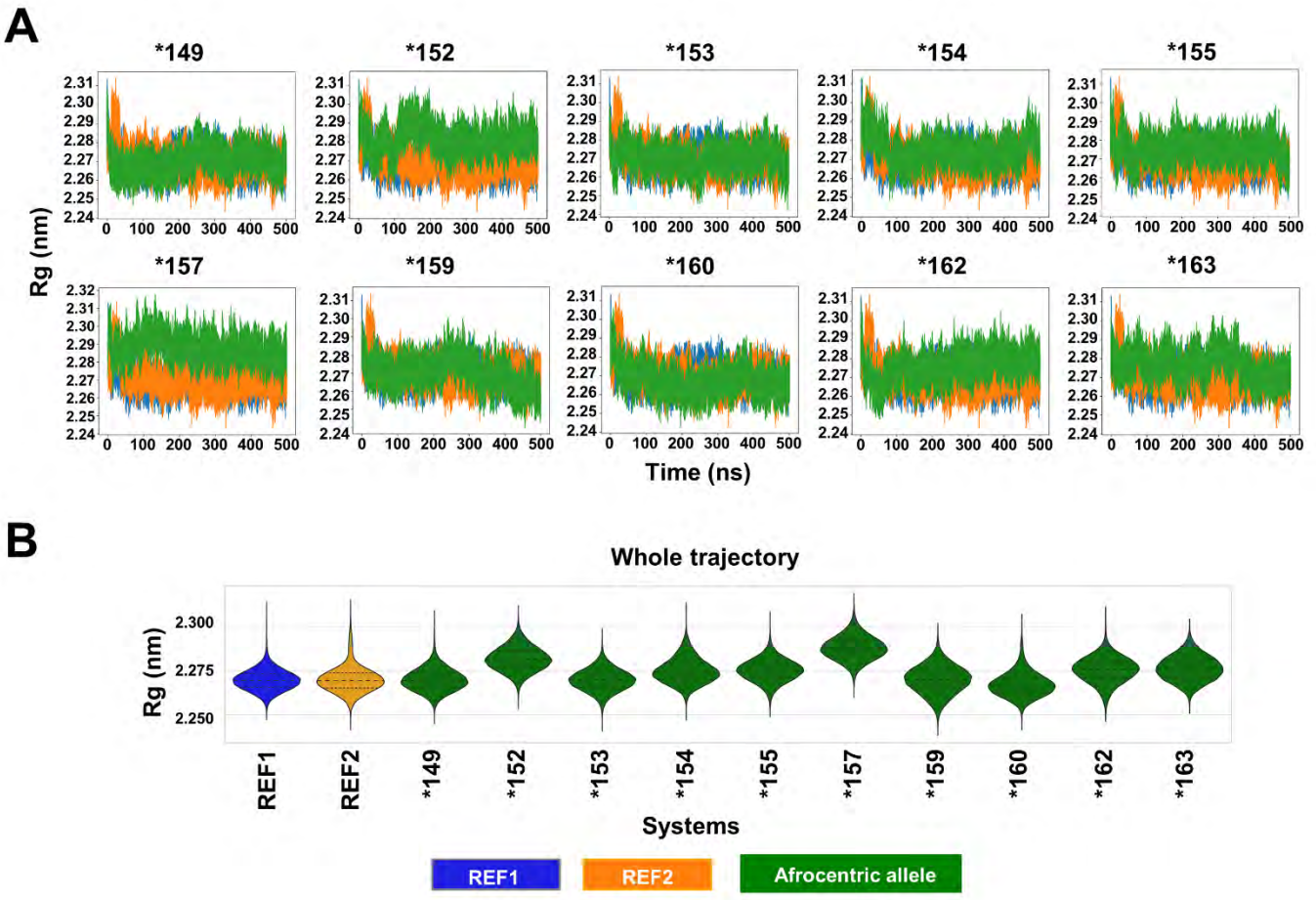


Figure 4-5: Rg analysis. A) Rg line plots for the whole 500 ns trajectory for references and Afrocentric alleles. B) Violin plots for the whole MD simulation (500 ns) for references and Afrocentric alleles. In both A) and B), REF1 is shown in blue, REF2 is shown in orange, and CYP2D6 Afrocentric alleles are shown in green.

4.3.3 Evaluation of changes in secondary structure, flexibility, and hydrogen bonding in functionally important regions of novel Afrocentric CYP2D6 alleles

4.3.3.1 Conformational clustering analysis

Next, it was essential to analyze the conformations of the alleles compared to the references after they reached stability (within the last 300 ns of the MD simulations), to observe if they had

highly similar conformations to references and assess if there were any secondary structure changes noted. To achieve this, geometric clustering program provided by GROMACS was used. The results demonstrated that all systems sampled a single conformation for the remainder of the simulation, highlighting their stability in this period. All structures maintained an RMSD of less than 0.2 nm, as shown in Table S7. As both REF1 and REF2 cluster structures were highly similar (as determined in chapter 3 and shown in Figure 4-6A), all cluster structures for the Afrocentric alleles were then superimposed to REF1 for comparison. Figure 4-6B, shows all Afrocentric allele cluster structures superimposed onto each other and REF1, and it was observed the main deviating regions were the F-G region and the B' helix region. Each Afrocentric allele cluster structure was then superimposed to REF1 separately to highlight the differences for each allele Figure 4-6C.

As seen in Figure 4-6C, alleles *149, *154, *160 and *163 had minimal changes in conformation and secondary structure as compared to REF1. The changes were mainly localized on the C-terminal end of the F helix and the N-terminal end of the F' helix. Major changes in structural conformation were however noted in alleles *152, *153, *155, *157, *159 and *162, and the differences in these alleles occurred mainly within the F-G region, and alleles *152, *153 and *159 having additional changes within the B' helix. Alleles *155 and *159 also exhibited more open conformations. Additionally, alleles *152, *157 and *162 were noted to have slight secondary structure changes as compared to the references in their conformational clustering structures. Allele *152 had an F' helix region turn to a loop region (residues 219 and 220), and alleles *157 and *162 both had residue 218 in a helix form rather than a loop, and residue 219 as a loop rather than a helix.

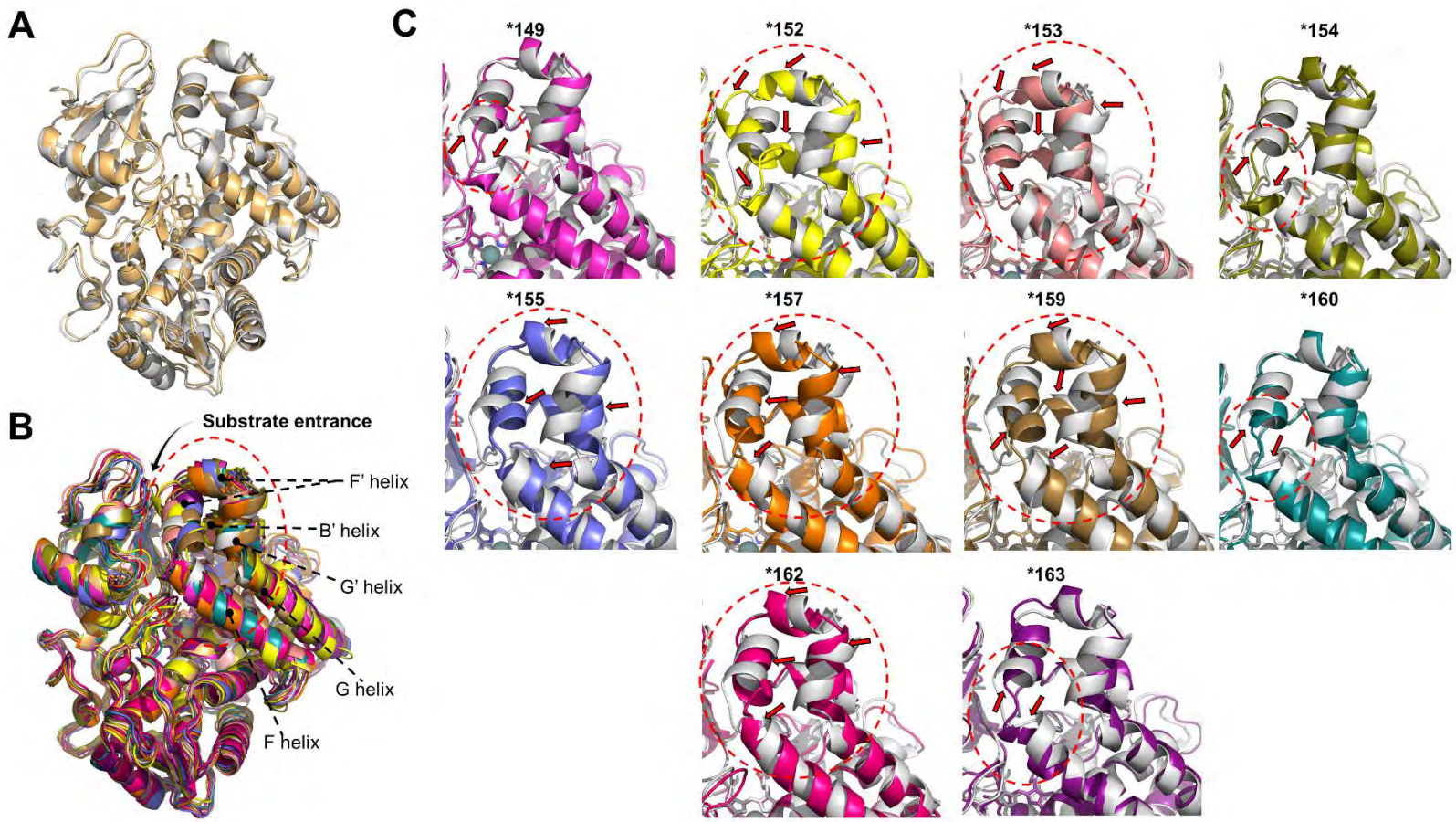


Figure 4-6: Analysis of conformational cluster structures for each reference and Afrocentric allele obtained through geometric clustering algorithm. A) Cluster structures for REF1 and REF2 superimposed and coloured grey and light orange respectively. B) All cluster structures for references and alleles superimposed, with deviating regions that were noted in alleles as compared to references highlighted by a red dashed circle and labels. Substrate entrance region shown by a black arrow. C) Each Afrocentric allele superimposed to REF1. Cluster structures for different alleles are shown in different colours and areas with varying conformations and secondary structures are highlighted with red circles and arrows.

The secondary structure changes though minimal, highlighted the existence of secondary structure changes in some of the Afrocentric alleles throughout their MD simulations that may all not be identified in one cluster structure. Therefore, it was paramount to run calculations that would aid us in fully understanding the secondary structure profiles of each allele in the simulations, and therefore compare with the reference systems to observe any similarities or deviations caused by variations.

4.3.3.2 Secondary structure profiles

Define Secondary Structure of protein (DSSP) metric in GROMACS was employed to calculate the secondary structure preference per residue over a simulation, for all reference and Afrocentric alleles. The last 300 ns of each MD simulation belonging to the references and alleles was used, as this region exhibited to be the most equilibrated. Results obtained were then represented as bar plots in Figure 4-7, and as can be seen in the figure, the Afrocentric alleles showed to differ from the references in the A' -A, B', D and E helix regions (residues 40-55, 103-115, 147-153 and 175-177 respectively), part of the F-F' region (residues 200-228) , and the β1 sheet region (β -1 and β1-2 sheet regions with residues 72-80). These regions are critical for the proper functioning of CYP2D6: the B' helix and the F-G region shift to facilitate substrate entry and exit [32, 179, 247]; the β1 sheet region contains the residue W75 that helps shuttle substrates to the active site [185]; the A'-A helix region is essential for the correct folding of CYP2D6 structure and its anchoring to the membrane [52, 53]; and the D and E helix are part of the highly conserved four-helix bundle important for structural integrity of the CYP2D6 core [241–243].

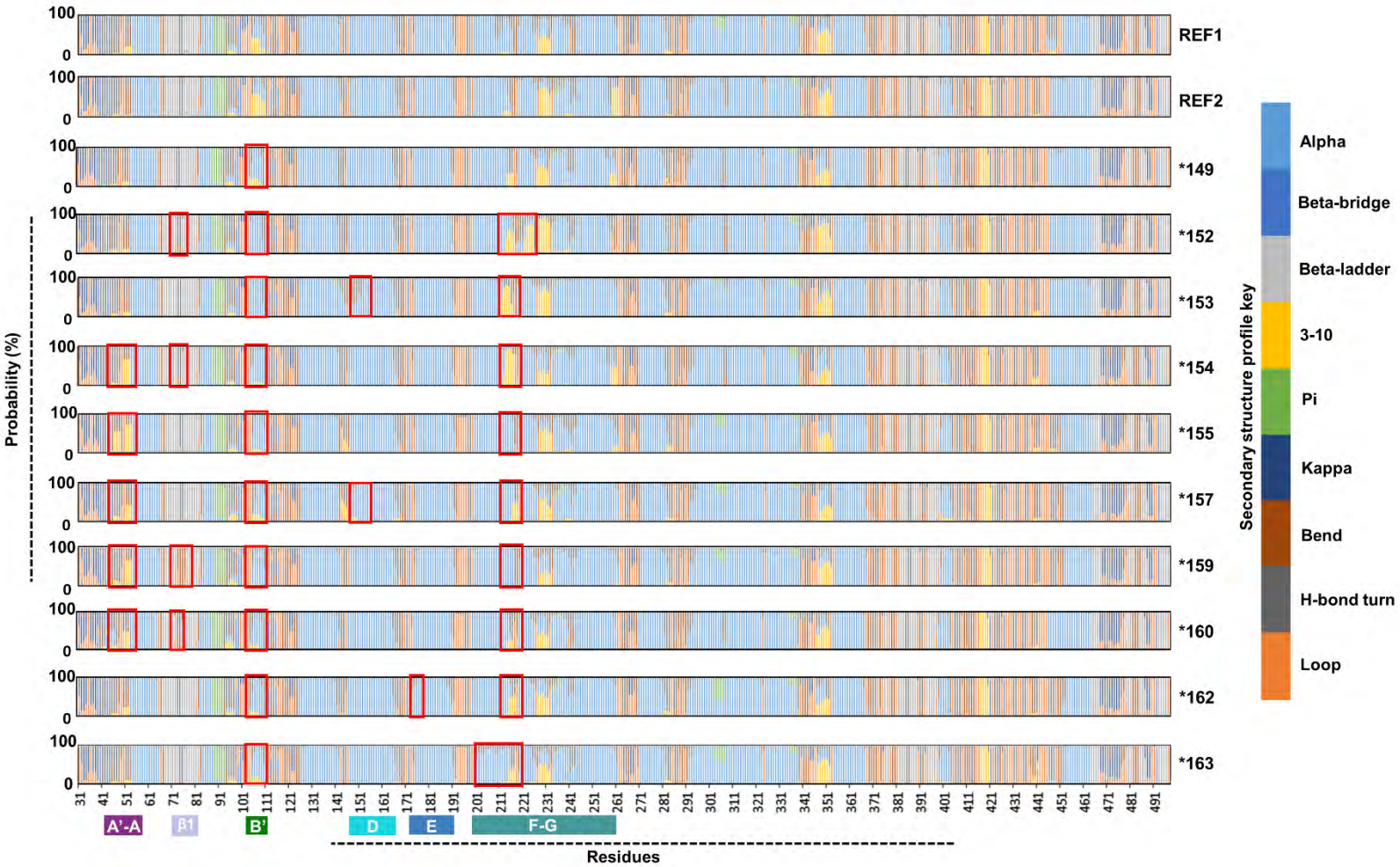


Figure 4-7: DSSP analysis. The bar graphs illustrate the likelihood of each residue (in references or Afrocentric alleles) adopting a specific secondary structure during the final 300 ns of each MD simulation. Areas where secondary structure assignments differ from the references are highlighted with red boxes and the regions are labelled and colour coded according to regions identified in DSSP analysis within chapter 3. A colour key for various secondary structures (including alpha-helix, beta-bridge, beta-ladder, 3-10 helix, pi helix, poly-proline II or kappa-helix, bend, hydrogen-bonded turn, and coil/loop) is provided in the figure.

All Afrocentric alleles were noted to have a decrease in sampling of the 3-10 helix secondary structure in the B' helix region, with this region preferring more of the alpha helix secondary structure, potentially making this region less flexible and stable in these alleles for a period of the simulation. Also, all alleles except for *149, displayed changes in secondary structure within the F-F' region, where alleles *155 and *159 had reduced sampling of the 3-10 helix and an increase in the sampling of the alpha helix secondary structure, leading to possible increased stability in this region during the simulation. Conversely, alleles *152, *153, *154, *157, *160, *162, and *163 exhibited a greater sampling of the less stable 3-10 helix structure, in a region which is normally stabilized by the alpha helix. This suggests the region became more unstable and flexible within these alleles throughout the simulation. Notably, allele *163 also developed a hydrogen-bonded turn structure during the simulation in this region, indicating a possible shift in secondary structures that could further destabilize the area.

Additional to these observations, alleles *154, *155, *157, *159 and *160 were showing an increase in the 3-10 helix secondary structure within the A' and A helix region, increasing the interchanging of unstable secondary structures in this area and thus possibly making the region less stable. Alleles *152, *154, *159, and *160 also showed a decrease in stability of the β1 sheet region by the introduction of unstable loop and bend secondar structures in a region with mainly stable hydrogen bonded turn and beta-ladder secondary structures. In alleles *153 and *157, there was an increased sampling of a hydrogen bonded turn and a 3-10 helix secondary structure, respectively, within the D helix, possibly increasing flexibility of the region. Allele *162 exhibited an additional change in the E helix, with the increased sampling of a less stable hydrogen bonded

turn secondary structure in place of the more stable alpha helix for a portion of the simulation, potentially increasing flexibility in this region.

These regions in Afrocentric alleles, which deviated from the references, were observed to be the same regions that were previously identified as differing in studies with the functionally characterized alleles (chapter 3). However, a new region identified to be differing between references and alleles *153 and *157 (D helix) was observed. It is part of a very conserved four-helix bundle that stabilizes the core of the protein, thus regarded as a key region for functionality of CYP2D6. As seen in Figure 4-2, we proposed that reduced activity alleles harbour at least three regions with changed secondary structure profiles as compared to references, in highly important regions within the enzyme, and as determined with the DSSP analysis here, alleles *152, 153, *154, *155, *157, *159, *160 and *162 all contained ≥3 regions with differing secondary structure profiles.

4.3.3.3 RMSF analysis

For support to our findings from DSSP analysis, as was done in our previous research, we conducted ΔRMSF calculations to observe if the changes in secondary structure profiles increased or decreased flexibility of regions throughout the simulations. Traditional RMSF values were computed per system as can be seen in Figure S11, however, to fully quantify the flexibility changes, ΔRMSF was computed. From this, the residues with the top and bottom 3% ΔRMSF values were obtained and plotted in a heatmap as seen in Figure 4-8A and onto conformational cluster structures per system in Figure 4-8B. Residues with significant changes in RMSF within

the top and bottom 3% (outliers that fall outside 3SD of the mean) were recognized and are indicated with an "X" on the heatmap Figure 4-8A.

In the DSSP analysis, all Afrocentric alleles were observed to have a reduction in the sampling of the 3-10 secondary structure in the B' helix region (residues 103-115), and an increase in sampling of the more stable alpha helix throughout the trajectories, suggested this region to be more stable and less flexible in these alleles. Indeed, this was observed with ΔRMSF analysis (Figure 4-8), as there was a decrease in flexibility for alleles *149, *152, *153, *154, *155, *159 and *162 within this region, with significant decrease noted for some residues in alleles *153, *155 and *159. Notably, alleles *160 and *163 had an increase in flexibility within this region, with allele *163 having a more significant increase. The B' helix region lies between the B and C helix regions (in the B-C loop), and both *160 and *163 contain variations within these helices; with allele *160 having variations L91M and H94R in the B helix, while allele *163 has variation W128R within the C helix. It was observed that these variations caused some residues within this B-C region to lose hydrogen bonding (section 4.3.3.4), and although DSSP analysis showed this region to increase in sampling of the more stable alpha helix secondary structure, the loss in hydrogen bonding would ultimately render this region unstable/more flexible, as seen here. In allele *162, there was an increase in sampling of a less stable hydrogen bonded turn secondary structure in place of a stable alpha helix within the E helix region (residues 175-177), possibly increasing flexibility of the region. With ΔRMSF analysis, we determined this region to increase in flexibility, with increase more significant in residue 175.

The A'-A helix region (residues 40-55) of alleles *154, *155, *157, *159, *160 was observed to have an increased sampling of a 3-10 helix secondary structure, indicating a higher interchange of secondary structures in this region throughout the simulations, which may suggest instability of the region. As seen with ΔRMSF analysis, we observed a gain in flexibility within this region in these alleles, with a more significant gain in flexibility in alleles *159 and *160 Figure 4-8. The higher flexibility observed in alleles *159 and *160 could have been due to variations and significant loss in hydrogen bonding near the region (discussed further in hydrogen bond analysis, section 4.3.3.4).

In Afrocentric alleles *152, *153, *154, *157, *160, *162, and *163, there was an increased sampling of the 3-10 helix secondary structure within the F-F' region (residues 200-228), with allele *163 having an additional hydrogen-bonded turn secondary structure sampled in the region throughout the simulations. These changes possibly destabilized/increased flexibility of this region in these alleles. Through ΔRMSF analysis, this was confirmed, as there was increase in flexibility within the overall F-G region of these alleles (*152, *153, *154, *157, *160, *162, and *163), with some residues in this region increasing in flexibility significantly (Figure 4-8). There were changes in hydrogen bonding within some of these alleles noted in this region, that could have contributed to the increased flexibility, and this is expanded upon in section 4.3.3.4. Allele *155 and *159 conversely sampled more of the stable alpha helix secondary structure in the F-F' helix region throughout the simulation, suggesting this region gained stability within these alleles. However, it was revealed with ΔRMSF analysis that this region had in fact gained flexibility

significantly within the overall F-G region, and this may be due to the significant loss in hydrogen bonding noted within these alleles in this region (section 4.3.3.4).

Alleles *153 and *157 were additionally observed to have changes within the **D helix region** (residues 147-153) through DSSP analysis. In *153, there was an increased sampling of a hydrogen bonded turn secondary structure within the region, while in allele *157, an increase in the sampling of a 3-10 helix secondary structure occurred throughout the trajectory. This region originally sampled the more stable alpha helix throughout the simulations, as seen in reference secondary structure profiles, and an increase in less stable secondary structures as shown in the two alleles could render the region unstable and more flexible. With ΔRMSF analysis, this region was observed to have gained flexibility significantly (Figure 4-8).

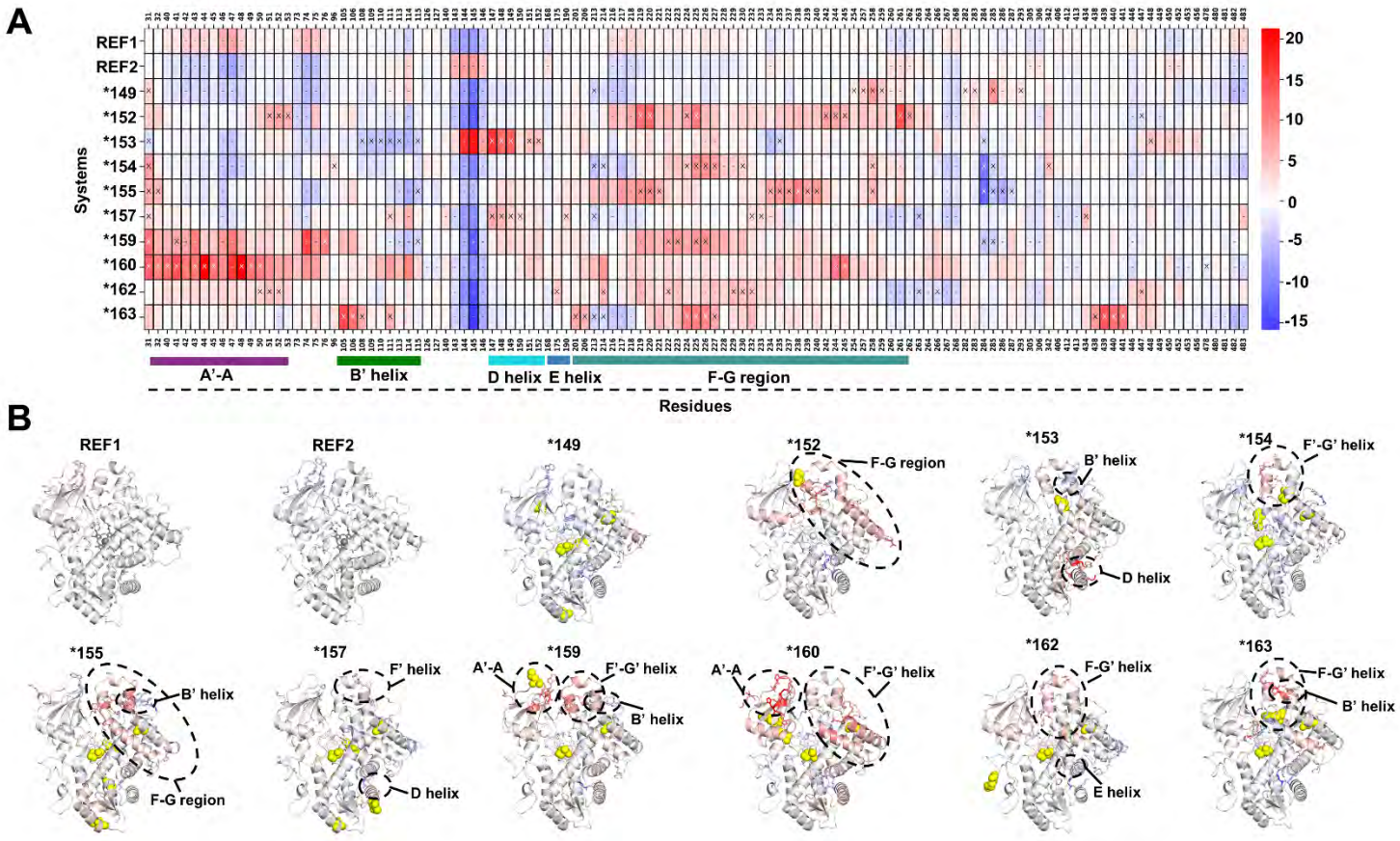


Figure 4-8: ΔRMSF A) Heatmap showing residues with top and bottom 3% ΔRMSF values. Significant values (outliers not within 3SD of the mean of the reference data) are marked with an "X" in the heatmap and values within top and bottom 3% ΔRMSF values and not significant are marked with a "-". Areas with changes in flexibility, which correlate with DSSP analysis are highlighted and colour-coded to match the DSSP analysis. B) Residues with top and bottom 3% ΔRMSF values for each system are mapped to reference and allele cluster structures and shown as sticks, and the residues with significant changes in ΔRMSF corresponding to changes noted in DSSP analysis are marked with black circles. Yellow spheres show the variation positions in each allele. In both A) and B) ΔRMSF values range from positive to negative. Residues with a loss in flexibility (becoming more rigid) within the systems are coloured in blue, and residues with gain in flexibility in systems are coloured in red.

4.3.3.4 Hydrogen bond analysis

Previously, delta hydrogen bonding (ΔH-bond) calculations revealed that hydrogen bonding networks were changed by variation presence in CYP2D6 alleles (chapter 3). Although there were no definite patterns defined by this analysis to differentiate normal from reduced activity alleles, there were changes noted in hydrogen bonding between residues throughout simulations. Therefore, in this study, we performed ΔH-bond calculations to observe how variations in each Afrocentric allele affect the hydrogen bond interactions within each system, which may have influenced the changes in conformation, secondary structure, and flexibility noted in previous studies within this chapter. The last 300 ns trajectories for each system were used for calculations, and residue pairs with the top and bottom 3% ΔH-bond values per system were obtained and represented in a heatmap shown in Figure 4-9A. Notable changes observed in hydrogen bonding between pairs were then mapped to corresponding allelic structures and shown in Figure 4-9B. Residue pairs showing hydrogen bonding differences between REF1 and REF2, as seen in the heatmap in Figure 4-9A, were not included in the analysis.

In allele *149, there was a loss in hydrogen bonding at the variation site R296C on the I helix, with residue 296 significantly losing hydrogen bonding with residues 252 (G helix) and 293 (I helix) Figure 4-9. This finding was also observed in Afrocentric alleles *154, *155, *157, *159, *160, *162, *163 Figure S12 (and function assigned alleles in chapter 3), where this variation was also present. Only in Afrocentric allele *149 was there a significant gain in hydrogen bonding of residue 293 with residue 115 noted Figure 4-9, an observation also noted in normal activity allele *45 (chapter 3). The loss in hydrogen bonding of a G helix residue (252) with a residue within the

more stable and highly conserved I helix (296) due to variation presence may contribute to the increased flexibility within the F-G region (confirmed in section 4.3.3.3 with ΔRMSF analysis).

In allele *152, there was a variation A226V in the F' region, and immediately next to it there was a significant gain in hydrogen bonding between residues 222 and 225 also in the F' helix. Additionally, at the base of the F-G region, several significant changes in hydrogen bonding were noted. There was a notable gain in hydrogen bonding between residues 259 (G helix) and 270 (G-H loop), but a distinguishable loss in hydrogen bonding of G helix residues 257 with 193 (E-F loop). On the E helix just before the E-F loop, residue 190 significantly lost hydrogen bonding with residue 272 (H helix) (Figure 4-9). Similar findings at the base of the F-G region in alleles *155 and *162 were also noted. In allele *162, there was a significant gain in hydrogen bonding between residues 258 (G helix) and 272 (H helix), however, there was a notable loss in hydrogen bonding between residues 270 (G-H loop) and 273 (H helix). Residue 190 on the E helix just before the E-F loop significantly lost hydrogen bonding with residues 186 (E helix) and 272 (H helix), In allele *155, the G helix residues 257 lost hydrogen bonding with residue 193 on the E-F loop (Figure 4-9). The substantial loss of hydrogen bonding within this F-G base region in alleles *152, *155 and *162, especially with residues in the E helix (important for structural integrity) and E-F loop just before the F helix, may destabilize the F-G region. An increase in flexibility noted in the F-G region within these alleles as analyzed by ΔRMSF confirmed this (section 4.3.3.3).

Additional to these observations in allele *155, there was a significant reduction in hydrogen bonding noted in residue pair 45-48 (A' helix region), and a notable increase in hydrogen bonding within residue pair 51-54 close to residues 45 and 48 (in the same A' region), possibly to counter the reduced hydrogen bonding within the region and maintain stability. The stability of those residues was confirmed as seen with ΔRMSF, as they had no changes in flexibility. Interestingly, next to variation M451I in the L helix, that was added to backbone allele *29 to form this allele, residue 450 significantly gained hydrogen bonding with residue 364.

In allele *153, variation E215K in the F helix was comprised of a glutamate changing to a larger lysine residue. This change resulted in a significant loss of hydrogen bonding between residue 215 and residue 242 in the G helix, as seen in Figure 4-9. Loss in hydrogen bonding at this site changed the flexibility of the F-F' region in this allele (as seen with ΔRMSF analysis), although not as significantly. Additionally, a change in the smaller amino acid to a larger lysine residue could have changed structure of substrate access tunnels (such as channel 2b, 2e and 2c) that passes through this variation site, essentially closing channel (section 4.3.5). Other significant hydrogen bond changes not directly linked to residue 215 where variation occurred were noted in this allele. Next to the variation, there was an increase in hydrogen bonding between residues 482 and 484 on the loop between β 4-1 and β 4-2. In addition, on the D helix, residue 150 significantly lost hydrogen bonding with residue 147 (D helix), however notably gained hydrogen bonding with residue 148 (D helix) and itself (intramolecular hydrogen bonding). In DSSP analysis (section 4.3.3.2), we noted allele *153 to have an increased sampling of a hydrogen bonded turn secondary

structure within this region, and the gain in hydrogen bonding observed here could be due to this secondary structure appearance within the region.

Allele *154 was observed to have a significant loss in hydrogen bonding at the variation T107I position in B' helix (residue pair 104-107), similar to that of its backbone allele *17, however there was no gain in interactions between residues 102-104 as seen in *17. Furthermore, a notable loss in hydrogen bonding within allele *154 was observed at the variation R441H site (Figure 4-9), with residue 441 losing hydrogen bonding with residues 99 (B-B' loop) and 122 (B'-C loop). Histidine is a smaller and less interactive amino acid as compared to arginine [255], therefore the loss in hydrogen bonding at this variation site was expected. It was interesting to note residue 486 (with variation S486T) in this allele significantly gaining hydrogen bonding with residue 480. This observation was not seen in any other Afrocentric allele or function assigned allele where this variation was occurring.

There was a significant increase in flexibility at the N-terminal of alleles *159 and *160 (as mentioned in the ΔRMSF section 4.3.3.3). The N-terminal region with the noted changes included the pre-A loop, A' and A helix region, β1 sheet region (β 1-1, β 1-2) and the B helix. This increased flexibility could have been due to the substantial loss in hydrogen bonding within this region in both alleles, as well as secondary structure changes. In allele *159, there was a significant loss in hydrogen bonding between residues 73 and 76 (β 1-1, β 1-2 sheet region) located next to variation P41I (pre-A loop). In *160, we noted a substantial loss in hydrogen bonding within residue pair 89-93 (B helix) located in between variations L91M and H94R also in the B helix. Further up from

the variations in the pre-A and A' helix regions, there was a notable gain in hydrogen bonding between residues 40 and 45, however a significant loss in hydrogen bonding occurred in residue pairs 40-44 and 45-72 (Figure 4-9).

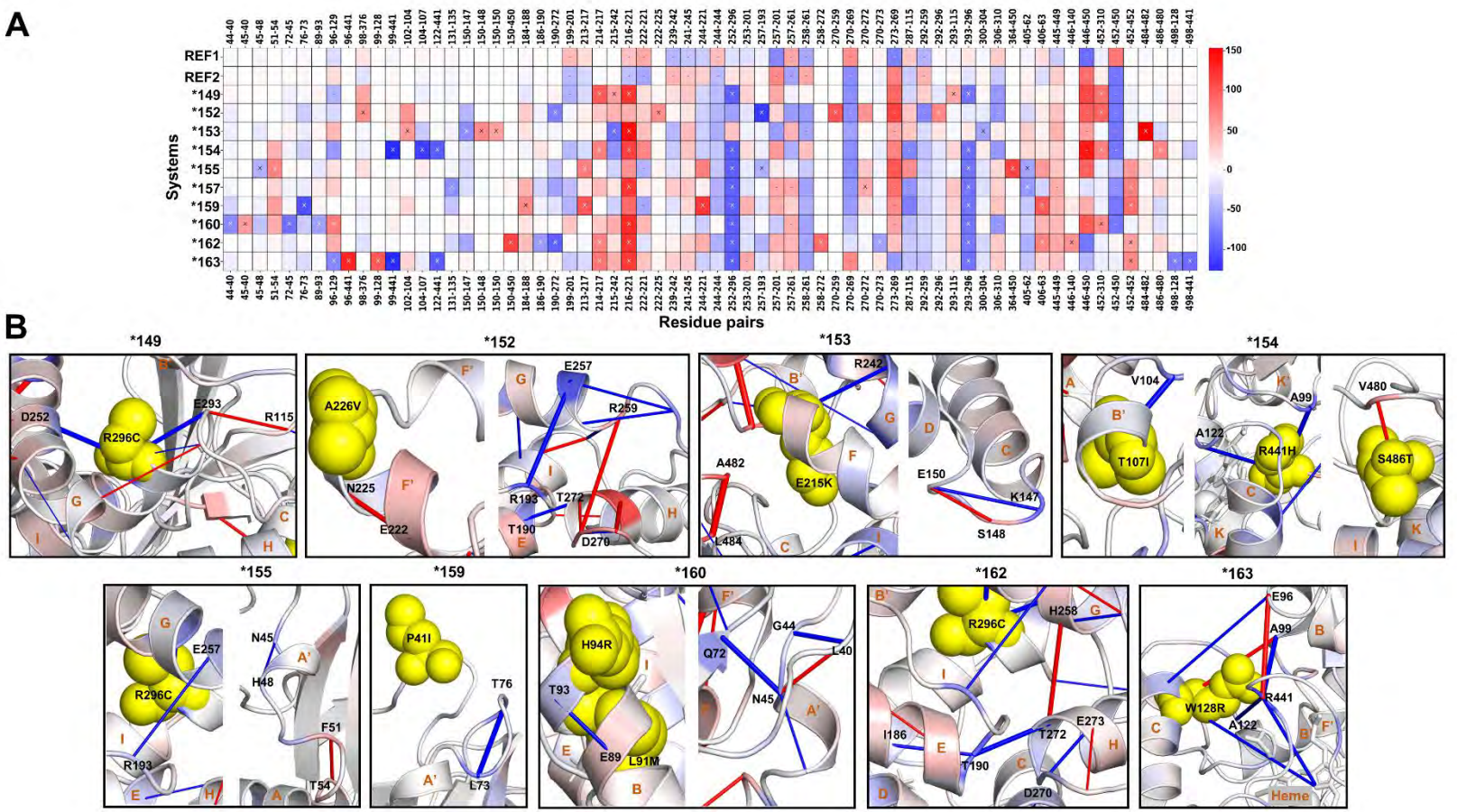


Figure 4-9: Hydrogen bond analysis. A) Residue pairs with top and bottom 3% ΔH-bond values shown in a heatmap, with significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap. The values within top and bottom 3% ΔH-bond values and not significant are marked with a "-". B) Significant changes in hydrogen bonding noted between residue pairs for Afrocentric alleles. Yellow spheres show variation positions within Afrocentric alleles. For both A) and B), ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive).

Allele *163 was noted to have a significant reduction in hydrogen bonding between the heme and residues 128 (variation W128R site in the C helix) and 441 (K''-L loop) (Figure 4-9). This observation was also noted in heme analysis (section 4.3.4). These residues then gained hydrogen bonding significantly with other residues. Residue 128 gained hydrogen bonding with residue 99 (B helix), while residue 441 gained hydrogen bonding with B helix residue 96 (which ultimately then lost hydrogen bonding with residue 129 from the C helix). This suggested residues 128 and 441 may have been stabilized by the formation of these new hydrogen bonds after losing bonding with the heme, however, residue 441 was noted to have further hydrogen bonds lost significantly with residues 99 and 122. The substantial loss of hydrogen bonding noted for residue 441 with other residues in the structure could suggest instability in the region this residue is located. ΔRMSF showed a significant increase in flexibility within residues 438-441, confirming the instability of the region from loss of hydrogen bonding.

Overall, the evaluated changes in conformation, secondary structure, and flexibility within this section were able to assist in determining which Afrocentric alleles were similar to reduced function alleles, thereby aiding in the prediction of whether they have reduced activity. Additionally, through hydrogen bond analysis, we were able to note variation-induced changes in hydrogen bonding networks within Afrocentric alleles and identify changes in hydrogen bonding that may have contributed to the changes observed in secondary structure conformation and flexibility.

4.3.4 Heme region analysis in Afrocentric CYP2D6 alleles

The heme within CYP2D6 is the catalytic center of the enzyme, where oxygen required for catalysis binds [256, 257]. It is therefore a crucial part of the protein for substrate metabolism, and any changes in this region may affect how CYP2D6 metabolizes substrates. Various tools to analyze this region in CYP2D6 references and Afrocentric alleles on the more equilibrated region of their MD simulations (last 300 ns) were employed, as in chapter 3. These tools assessed the number of hydrogen bonds between the heme cofactor and adjacent protein residues, the contacts/non-covalent interactions of the heme with nearby residues (within ~4 Å), and the distance between the heme's center of mass (COM) and the active site COM.

Our previous studies examining this heme region in both normal and reduced function alleles, showed minimal changes, and this finding was consistent across the Afrocentric alleles *149, *150, *152, *153, *154, *155, *159, *160, and *162 analyzed in this research (Figure 4-10). Alleles *157 and *163 however, exhibited changes within this region. There was a substantial increase in the distance between COM of the active site and heme within allele *157 throughout the simulation. Furthermore, this allele showed to have a substantial decrease in contacts between the heme and residues 313 and 399. Variation A165V within allele *157, occurs on the D helix, which has been observed to be bisected by the plane of the heme in some CYP enzymes [258], and is part of a highly conserved four-helix bundle (with helices E, I and L) that is paramount for the structural integrity of CYP2D6 enzyme core [241–243]. The findings presented here suggest the addition of variation A165V (located on the highly conserved D helix within the four-helix bundle) to the backbone allele *29, creating the Afrocentric allele *157, induced the loss of important non-

covalent interactions of heme with residues 313 and 399, destabilized the heme region in this newly formed allele, as seen by the changed distance between the COM of the heme and the active site.

In allele *163, there was a considerable decrease in hydrogen bonding exhibited between the heme and surrounding residues, and when observing the contacts with each residue closely, most of the contacts were reduced with residues 128 ad 441 (Figure 4-10C and D). The reduction in interactions of the heme with residue 128 could be due to variation W128R occurring within this allele. Tryptophan (W) is the largest amino acid when compared with the other 19 amino acids found in human proteins, as it is the only one to contain a side chain with two rings, named an indole, that is aromatic with a binuclear ring structure [259, 260]. It engages in hydrogen or π–cation bonds, and the amino acid change from the large, globular, hydrophobic tryptophan amino acid, for a smaller basic arginine residue, distorts these interactions with the heme. Furthermore, additional to the highly conserved four-helix bundle for the structural integrity of the CYP2D6 core, the heme is stabilized by surrounding amino acids, such as the cysteine at residue number 443 (primary anchor), and the side chains of residues 101, 128, 132, 376, 437, and 441 (Figure 4-10B) [261]. Distortion or reduction in the interactions between these residues and the heme could indicate possible instability in the core of the enzyme, as seen here in the allele *163.

The changes in the heme region in Afrocentric alleles *157 and *163 was not observed in alleles with normal or reduced activity, suggesting that these alleles may not have these functional consequences. Additionally, a study conducted by Klein et al., [261] revealed that a substitution of amino acids anchoring or stabilizing the heme resulted in complete loss of function.

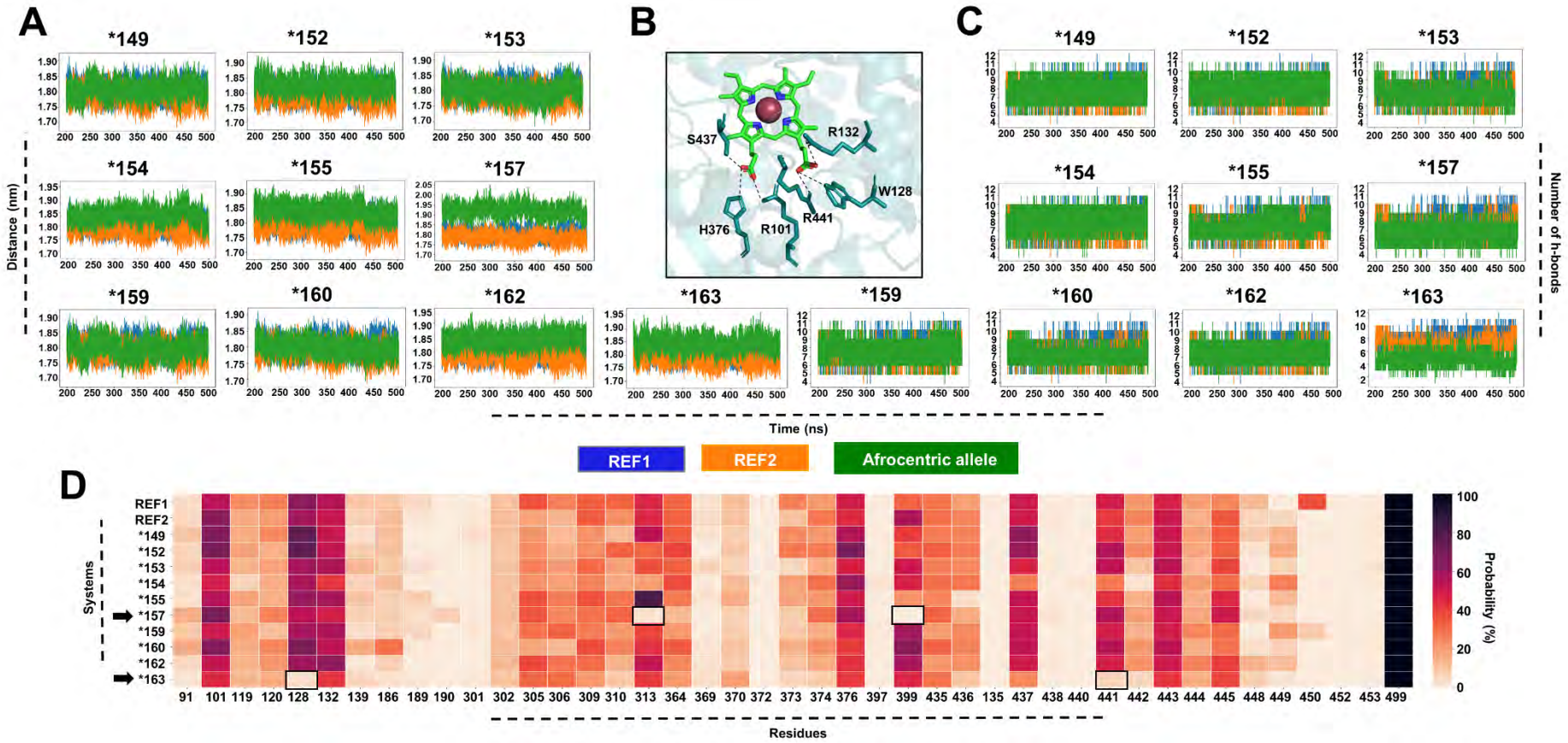


Figure 4-10: CYP2D6 Afrocentric allele heme analysis. Analysis of heme environment in CYP2D6 references and Afrocentric alleles for the last 300 ns of the MD simulation. A) Line plots showing center of mass (COM) distances between the heme and the active site cavity for references and alleles. B) Stick representation of heme position in CYP2D6 and residues anchoring it in place. Heme is coloured in green and residues coloured in deep teal. C) Line plots showing hydrogen bond formation between heme and surrounding residues. For both A) and C), REF1 is shown in blue, REF2 is shown in orange and CYP2D6 Afrocentric alleles are shown in green. D) Heme contact/non-covalent interaction analysis with residues ≤4 Å from the heme. The probability of contact of each residue with the heme is shown by the colour scale from 0-100%.

4.3.5 Substrate channel analysis

Analysis of the tunnels in CYP2D6 Afrocentric alleles was then carried out, using CAVER 3.0 and snapshots of the whole MD simulation (500 ns). The identified tunnels in the previous work in chapter 3 for CYP2D6 alleles with functional consequence, were also observed here. These include channel 2b, 2f, 2e, 2c and the solvent channel, using the channel nomenclature from Cojocaru *et al*., [61]. Channel 2b was noted as the possible main substrate access channel in all the CYP2D6 Afrocentric alleles, as this channel appeared in at least 800 snapshots of the respective MD simulations, the highest frequencies observed as compared with the other channels (Table S8). This channel was noted to be the main channel in the reference systems, indicating no shift in preference for this tunnel for substrate access to the enzyme in the Afrocentric alleles.

Tunnel opening frequencies were then computed, to observe which of the tunnels were more open and which were closed. According to Figure 4-2, reduced function alleles either had all channels more frequently closed, or there was more than one channel more frequently open or partially open simultaneously throughout the simulation, creating a more open protein that may be unfavorable for substrate metabolism. For each of the reference and Afrocentric allele, the bottleneck radii for each tunnel throughout the 500 ns MD simulation were plotted in heatmaps, as seen in Figure 4-11, and a tunnel was regarded open if most of the bottleneck radii recorded for the tunnel were above ≥1.4Å.

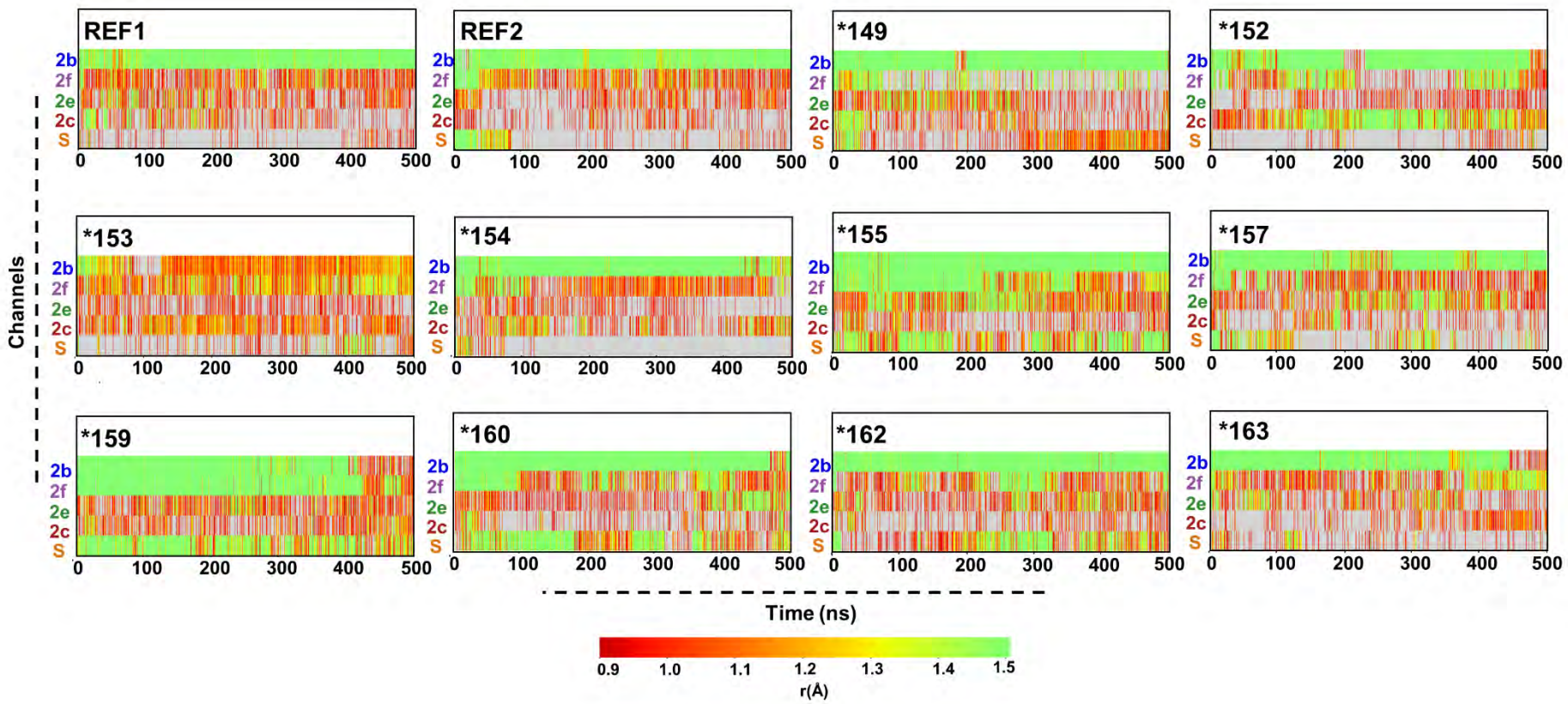


Figure 4-11: Channel analysis of CYP2D6 Afrocentric alleles. The heatmaps illustrate time evolution of the channel bottleneck radius for channels 2b, 2f, 2e, 2c, and the solvent channel, for the 500 ns MD simulation. A total of 1001 snapshots were used in the analysis, including the snapshot at 0 ns. A colour map ranging from 0.9 Å to 1.5 Å (and above) is shown below heatmaps, with the smallest radii represented in red and the largest in green. Grey indicates that no channel with a bottleneck radius of ≥0.9 Å was identified at that specific time point.

From Figure 4-11, it was noted that alleles *149, *154, *157 and *163 had channel opening frequencies that resembled that of the two CYP2D6 references. Their main channel 2b was more frequently open, with an average bottleneck radius above 1.4 Å (Figure S13), and all other channels more frequently closed (average bottleneck radii less than 1.4 Å as seen in Figure S13), indicating the possible movement of substrates to the active site to be similar to the reference systems, via channel 2b.

Alleles *152, *153, *155, *159, *160 and *162 however, behaved similarly to the reduced function alleles from our previous study. The allele *153 was found to have its main substrate access channel and all other channels more frequently closed, suggesting that the movement to and from the active site is impaired. As mentioned earlier (section 4.3.3.4), the variation E215K is a change in amino acid residue glutamate for a larger amino acid lysine that may block channels as this variation occurs where most of the channels pass through. Alleles *152, *155, *159, *160 and *162 were observed to have more than one channel more frequently open or partially open throughout the simulation, especially within the equilibrated regions. All these alleles had their main pathway, channel 2b, more frequently open, however, in allele *152, channel 2c was noted to be partially open, while in alleles *160 and *162, the solvent channel was partially open throughout the equilibrated regions of the simulations. Alleles *155 and *159 had channel 2f more frequently open for most part of their simulations, simultaneously with channel 2b, and additionally also had the solvent channel partially open. These findings suggest alleles *152, *155, *159, *160 and *162 may be in a more open conformation that could allow for simultaneous binding of substrates which does not favor substrate metabolism. Channel opening frequencies

may have been altered by the differences in structure and flexibility noted earlier within these alleles. Additionally, in alleles *155, *159, *160 and *162, the occurrence of variation R296C in the unique motif 25 could have influenced the discrepancies in tunnel opening frequencies, as discussed in chapter 2.

4.3.6 DRN analysis

The final mechanism identified leading to the determination of reduced function alleles as stated in Figure 4-2 was brought about by analyzing the CYP2D6 enzyme communication networks. To achieve this in our previous research, we employed DRN analysis metrices; *BC*, *EC,* and *CC*, and determined *BC* to be the most informative in distinguishing communication networks between normal functioning CYP2D6 enzymes and those with reduced function. Specifically, Δ*BC* showed that the reduced function alleles had major changes in inter-residue communications around the enzyme, similar to those observed in the normal function alleles. However, additional unique changes were noted in the B' helix and/or the β1 sheet region (loop between β1-1 and β1-2) important for substrate access. In this current study, the same DRN metrices were utilized and Δ*BC*, Δ*CC* and Δ*EC* were computed to determine whether such patterns observed in the function assigned CYP2D6 alleles were observed in the Afrocentric alleles, so as to contribute to the potential prediction of reduced function of these Afrocentric alleles. The DRN metric values for both references and their averages (REF_AV) were retained from our previous study and used in this current research.

4.3.6.1 Betweenness Centrality

After obtaining the Δ*BC* values for all references and Afrocentric alleles, the residues with top and bottom 3% Δ*BC* values were determined and shown in a heatmap (Figure 4-12A) and mapped onto their respective conformational clustering structures for each system, as seen in Figure 4-12B and Figure S14. From the results, we noted alleles *149, *154, *155, *157, *160, *162 and *163 to contain changes in averaged *BC* as compared to the references all around the enzymes, however no significant changes were observed within the B' helix or within the loop between the β1-1 and β1-2 sheets, indicating there was no inter-residue communication changes observed in these alleles that may resemble that of reduced activity alleles (Figure S14). In contrast to this, alleles *152, *153 and *159 were observed to have inter-residue communication changes around the protein, and either in the B' helix or the β1 sheet region (loop between the β1-1 and β1-2 sheets), resembling that of reduced function alleles (Figure 4-12A and B). Allele *152 had significant changes in averaged *BC* as compared to the references in residues 105, 106, 109 and 110, while *153 had significant changes in residues 105, 106 and 110, all within the B' helix. In allele *159, a significant change in averaged *BC* was only noted in residue 110 (B' helix).

For allele *157 that was noted to have changes around the heme area in section 4.3.4 of this chapter, further unique changes were noted that were not observed in reduced function alleles or any Afrocentric alleles studied here. In this allele, residue 399 was identified as one of the residues that lost contact with the heme group, and through analysis of the Δ*BC* results for this residue, we noted a significant decrease in averaged *BC* as compared with the references, suggesting a disruption in inter-residue communication at this point within the enzyme (Figure 4-12A and C).

Interestingly, immediate surrounding residues to 399 (residues 365, 370, 371, 400 and 403) had a significantly increased averaged *BC*, and formed a pathway from the heme to the more stable and conserved K helix, as seen in Figure 4-12C, suggesting a shift in inter-residue communication within this allele, between the heme and surrounding residues.

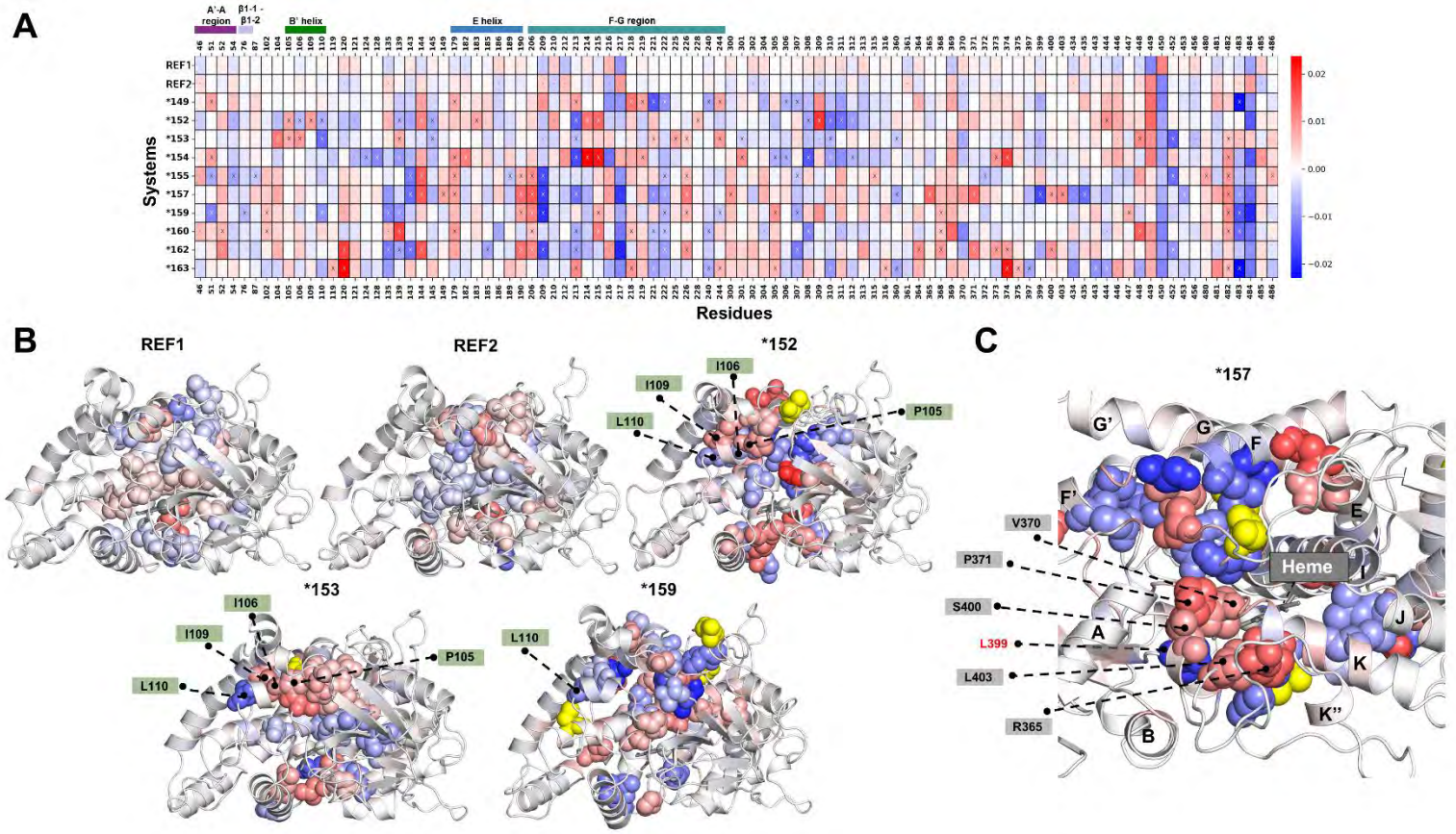


Figure 4-12: Δ*BC* analysis of Afrocentric alleles. A) Residues with top and bottom 3% Δ*BC* values shown in a heatmap with significant values (outliers not within 3SD of the mean of the reference data) marked with an "X" in the heatmap and top and bottom 3% Δ*BC* values that are not significant marked with a "-". B) Alleles *152, *153 and *159 with significantly changing averaged *BC* values for residues within the B' helix shown (labelled in green labels). REF1 and REF2 also shown with no significant changes in averaged *BC* values in this region. C) Allele *157 with residue L399 shown in a red label with loss in averaged *BC*, while there was a pathway of gain in averaged *BC* formed by residues V370, P371, S400, L403 and R365 from the heme to the K helix. For B) and C), variation positions for alleles shown in spheres and coloured in yellow. For all (A, B and C) the *ΔBC* values range from positive to negative, with positive indicating a gain in averaged *BC* within the alleles (shown in red colour) and negative indicating a loss in averaged *BC* in alleles (shown in blue colour).

4.3.6.2 *Eigenvector centrality* and *Closeness Centrality*

Closeness centrality and *Eigenvector centrality* were the next metrices used to analyze the communication patterns within the Afrocentric alleles. The residues with top and bottom 3% Δ*EC* and Δ*CC* values were identified for each system and represented in heatmaps as seen in Figures S15A and S16A, and illustrated as spheres on corresponding conformational cluster structures for each system, as depicted in Figure S15B and Figure S16B, respectively. Although both metrics were not utilized when defining the different mechanisms for reduced activity in CYP2D6, it was interesting to analyze the results so as to observe if similar patterns noted in some reduced function alleles were present, or if there were any new findings that shown unique behavior different from either normal or reduced function alleles.

According to Figure S15, similar changes in averaged *EC* were noted in Afrocentric alleles *149, *152, *153, *154, *155, *159, *160, *162 and *163, as those noted in normal activity alleles. These alleles were noted to have a less significant increase in averaged *EC* in the helices B, K' and L, and the β-sheet rich region in CYP2D6 (β1-1, β1-2, β1-4, β2-1 and β1-3), while a less notable decrease was noted in the C-terminal half of I helix, the J-J' loop, and the K and L helices. Allele *157 however, was noted to have a significant increase and decrease in averaged *EC* within the same region (the region with helices B, K' and L, and the β-sheet rich area (β1-1, β1-2, β1-4, β2-1 and β1-3)) as seen in Figure S15. This indicated that residues within this area became less influential within the communication system, however other residues in this same region gained influence. This observation was not noted in any of the normal or reduced function alleles.

Closeness centrality primarily identifies the central residues that are nearest to other residues in a protein network, and usually regions with high *CC* map to the core of a protein [226, 227, 248]. In the previous study with function assigned alleles, residues with high averaged *CC* were noted to be within the core of the protein where the heme and active site are located. As mentioned before, this is where substrate oxidation occurs, and high averaged *CC* in this region highlights the closely bound network of this site, which influences enzyme activity. Here, Δ*CC* results shown in Figure S16 revealed minimal to no changes in the residues identified to have high CC (as seen in section 3.3.6.3 in chapter 3), indicating no substantial change in centrality of the highly functional catalytic center of CYP2D6, regardless of any variation presence.

4.3.7 Potential functional consequence of CYP2D6 Afrocentric Alleles

Through all analysis carried out in this chapter, the main aim was to possibly predict if any of the CYP2D6 Afrocentric alleles from Sub-Saharan Africa had reduced activity. As seen in Figure 4-2, there are four mechanisms identified (from chapter 3) that potentially lead to reduced activity in CYP2D6 alleles. We proposed that if an allele were to pass at least three of the four mechanisms, the possibility of that allele harboring reduced activity was high. In Table 4-1 we summarize our findings for CYP2D6 Afrocentric alleles, by outlining the four mechanisms for reduced activity, and then identifying which alleles passed each mechanism.

Table 4-1: Summary of identified mechanisms potentially leading to reduced activity in the Afrocentric alleles.

Afrocentric allele	Instability and conformational changes of protein as measured by RMSD and Rg		Secondary structure variation and changes in flexibility of ≥3 highly important protein regions (DSSP, conformational clustering and RMSF)								More frequently closed/ more than one frequently open or partially open channel simultaneously (Caver3.0)	Unique changes in inter-residue communications within the B' helix and/or β1 sheet region (ΔBC)	
	Instability (equilibrated after ~200 ns)	More than one conformation visited in 500ns	A' -A region	β1 sheet	B' helix	E helix	E-F loop	F-G region	β4 sheet	D helix		β1 sheet	B' helix
***149**					√								
***152⁺**	√	√		√	√			√			√		√
***153⁺**	√	√			√			√		√	√		√
***154**	√	√	√	√	√			√					
***155⁺**	√	√	√		√			√			√		
***157**	√	√	√		√			√		√			
***159⁺**	√	√	√	√	√			√			√		√
***160⁺**	√	√	√	√	√			√			√		
***162⁺**	√	√			√	√		√			√		
***163**					√			√					

⁺Afrocentric alleles observed to have three or more mechanisms that potentially lead to reduced activity identified. Red boxes highlight the alleles where the mechanism occur and √ shows exact changes noted in allele.

According to Table 4-1, alleles *152, *153 and *159 passed all four mechanisms that have been identified to potentially induce reduced activity within CYP2D6. In these alleles, we observed instability and conformational diversity, where each allele contained bimodal distribution of data throughout the trajectories and only stabilized after ~200 ns. There were also significant changes in secondary structure and flexibility noted in highly important regions within CYP2D6, such as the A'-A region, the β1 sheet region, the E-F loop, the F-G region and the B', D and E helices. We also noted allele *153 to have its main channel more frequently closed throughout the trajectory, while alleles *152 and *159 had more than one channel frequently open tunnel, which may be unfavorable for substrate metabolism. Uniquely identified changes in inter-residue communications within the B' helix and/or β1 sheet region in reduced activity alleles were also noted for these alleles. Through these findings, reduced function for these alleles may be inferred. Alleles *155, *160, and *162 indeed had three mechanisms for reduced activity identified, as notable instability, conformational and secondary structure changes were detected, within these systems. However, there were no unique changes in inter-residue communications within the B' helix and/or β1 sheet region, which noted in all reduced activity alleles (chapter 3). Therefore, reduced activity may be predicted for these alleles, though further analysis may be required to fully support this prediction.

Allele *154,*157 and *163 had less than three mechanisms identified that potentially lead to reduced activity. Instability, conformational changes and significant changes in secondary structure were noted for all alleles, however they had channel opening frequencies highly similar to references and normal function alleles rather than reduced activity alleles and did not contain

any inter-residue communication changes within the B' helix and/or β1 sheet region. Additionally, alleles *157 and *163 both contained changes in the heme region that were not observed in previous studies with function assigned alleles (chapter 3). In allele *157, residues 313 and 399 lost interactions with the heme, and the distance between the COM of the heme and the active site notably increased, while in allele *163, residues 128 and 441 lost hydrogen bond interactions significantly with the heme. This therefore indicates that alleles *154, *157 and *163 cannot be classified as having either normal or reduced activity in the context of this study, and further investigations may be required to predict functionality.

In all analysis carried out, allele *149 behaved highly similar to references, with no significant changes noted in stability, secondary structure, conformation, hydrogen bonding, channel opening frequencies and communication networks. Additionally, the heme region was as stable as what was observed in reference systems. It is therefore highly probable this allele may harbor normal functionality. All predicted functionality in this research was derived from in silico experiments. Although the predictions may be robust, in vitro experiments are required to validate the claims presented here regarding the novel Afrocentric alleles.

4.4 Conclusion

Human CYP2D6 is greatly involved in the biotransformation of numerous xenobiotics/drugs, making it one of the key enzymes in drug metabolism within the CYP superfamily. It is also a highly polymorphic enzyme among all other CYPs, indicating metabolism of many drugs may be implicated by the existence of different variants of this enzyme. Pharmacogenomic studies on CYP2D6 have been carried out to characterize its variants and star alleles across many different populations, however, African populations are yet to be adequately represented in this area of research. Therefore, in this chapter, we aimed to further pharmacogenomic studies in Africa, by potentially predicting functionality of novel Afrocentric alleles recently identified in individuals from Sub-Saharan Africa (specifically in the Niger–Congo language speaking African region) with unknown clinical activity data at present. In the previous chapter, we managed to identify mechanisms that possibly lead to reduced activity of this enzyme, and also noted how references and other normal activity enzymes behave. It was determined that if an allele passes ≥3 mechanisms outlined; reduced functionality would be highly probable. In the study presented within this chapter, we then utilized these mechanisms to potentially predict reduced activity among Afrocentric CYP2D6 alleles.

An in-depth computational analysis was conducted on Afrocentric alleles *149, *152, *153, *154, *155, *157, *159, *160, *162 and *163, employing *in silico* tools utilized within chapter 3 to uncover any patterns in their behavior that may be comparable with those identified for reduced activity alleles. Through analysis, we were able to determine alleles *152, *153 and *159 as having

highly similar behavior to reduced activity alleles. They passed all four mechanisms that could bring about reduced activity identified, indicating they may harbor this functionality. Alleles *155, *160, and *162 passed at least three mechanisms for reduced activity, suggesting they may possess this functionality, as outlined in the conditions set as part of the mechanisms in the previous study. However, no specific alterations in inter-residue communications within the B' helix and/or β1 sheet region were noted in these alleles, which were present in all reduced activity alleles (chapter 3). Consequently, while reduced functionality is a plausible prediction for these alleles, further analysis may be necessary to substantiate this conclusion.

Alleles *154, *157 and *163 failed to pass ≥3 mechanisms, as they did not contain the distinct changes in inter-residue communication within the specified regions and their channel opening frequencies resembled that of the references and normal activity alleles rather than reduced activity alleles. This indicated they may possess normal functionality, however, in alleles *157 and *163, there were additional changes noted in the heme region which were not identified in any of the functionally assigned alleles. Due to these observations, more analysis would be essential to either support normal functionality of these alleles, or specifically in alleles *157 and *163, support the null functionality, as studies conducted by Klein et al., [261] revealed that instability within the heme region may result in complete loss of function. The only Afrocentric allele with behavior highly similar to references and all other normal functioning alleles through all analysis performed (i.e., failed to pass any of the outlined mechanisms for reduced function), was allele *149, and due to this consistency, we propose this allele bears normal activity.

As mentioned in chapter 3, the results obtained in this research are solely based on *in silico* experiments. We have identified potential functional consequences of the different Afrocentric alleles, however, to substantiate these findings, complementary *in vitro* experiments can be conducted to confirm the observed functional effects.

Chapter 5

Concluding Remarks and Future Perspectives

Chapter overview

This chapter provides an overall conclusion of the work presented in this thesis and provides additional future work that could be carried out to further support findings outlined.

5.1 Overall Conclusion

The effectiveness and outcomes of drug therapies has been greatly impacted over the years by many different factors, however the human genetic makeup has been a major influence in how individuals metabolize or respond to certain drugs. Specifically, pharmacogenes coding for CYP enzymes in families 1, 2 and 3, that metabolize about 70–80% of all clinical drugs currently used, may carry variations that can potentially result in enzymes with altered activity, subsequently influencing the metabolism and clearance of drugs. Research on these enzymes is therefore crucial, as it may contribute to a better understanding of CYP polymorphisms and their impact on enzyme activities, which will aid in the formation of personalized drug therapies that will enhance drug efficacy and reduce adverse effects. A CYP family 2 enzyme, CYP2D6, is of particular importance, as it is the only CYP enzyme that metabolizes ~25% of the clinical drugs. This indicates that variations within this enzyme can influence the metabolism and clearance of a broad spectrum of medications from the body, and as such, was selected as the primary focus for the research conducted in this thesis.

In chapter 2, we aimed to characterize enzymes from the CYP 1, 2 and 3 families, to gain deeper understanding of the enzymes within these families, reveal the shared relationships of these enzymes with CYP2D6, and unmask the unique structural and sequential characteristics of CYP2D6. 23 sequences belonging to the CYP enzymes in families 1, 2 and 3 were identified, and bioinformatics tools for MSA, phylogenetic tree and motif analysis were employed for analysis.

Through MSA and motif analysis, we were able to identify and highlight highly conserved regions in CYPs within the first three families of the CYP superfamily, that are necessary for structural integrity and enzymatic function. These included the pre-A loop, A', A, B, C, D, E, C-terminal half of I, J', J, K, K'-K'' loop and L helices, and also the beta sheets β1-1, β1-2, β1-3, β2-1, β2-2 and β3-3. Additionally, unique regions within the different families were also uncovered. In CYP1 family, the unique regions were within the B' helix, B-B' and B'-C loops, part of the G helix, and β4-1 and β4-2 sheet regions. In CYP2 family, unique regions were noted in the B', D, F, F', G, H and N-terminal half of I helices, the B-B', B'-C, D-E, E-F, G-H, H-I and K''-L loops, and the β4-1 and β4-2 sheets, and finally in CYP3 family, the B', D, F, F', G, G', H and K'' helices, the G'-G and K''-L loops and the β3-2, β4-1 and β4-2 sheets contained their unique regions. These findings indicated that CYP enzymes within CYP families 1, 2 and 3 all contain particular regions in common, however they do have unique regions within each family, that may cater to the differences in the drugs they metabolize. Phylogenetic analysis further emphasized this, as there was family-wise clustering of sequences, which highlighted the strong relationships within families.

CYP2D6 however, differed from other CYP2 family enzymes, as it only contained three unique regions/motifs identified, that were in the H helix, the G-H and H-I loop regions, and the N-terminal end of the I helix. This finding may elucidate why CYP2D6 is able to metabolize a greater diversity of clinical drugs than CYP enzymes from other families. The variations in CYP2D6 alleles to be studied in this thesis were then checked to observe if any occurred within the discovered conserved or unique regions identified. Interestingly, most of the variations mapped

to conserved motifs, which may suggest that they could impact overall structure and functionality of these regions. However only one variation (R296C) mapped to the unique motif 25 (a region known to be important in substrate access to the active site), suggesting it may affect substrate channels. This was confirmed with substrate access channel analysis for alleles *14, *17, *29 and *55 in chapter 3, and alleles *155, *159, *160 and *162 in chapter 4. To fully understand variation effects, a more comprehensive investigation would be required, and this was carried out in subsequent chapters of the thesis.

In chapter 3 we then aimed to understand the effects of the variations in functionally assigned CYP2D6, so as to potentially decipher variation-induced mechanisms that lead to reduced activity of this enzyme. A thorough computational analysis of CYP2D6 reference and alleles with known clinical function was conducted, including normal activity alleles *2, *27, *33, *34, *39, *45, *48 and *53, and reduced activity alleles *10, *14, *17, *29, *49, *50, *54 and *55. These alleles along with the CYP2D6 reference were subjected to 500 ns MD simulations, and different *in silico* tools were employed for analysis. It was determined through RMSD and Rg analysis that multiple reduced activity alleles stabilized significantly later in the MD simulations as compared with references and normal activity alleles, with most reaching stability ~200 ns. This suggested reduced function alleles to be more unstable. Once equilibration was reached, the alleles with decreased activity adopted a more distinctly different conformation around the F-G and B' helix regions. Additionally, more significant secondary structure and flexibility changes were noted, particularly in the B' and E helix, the E-F loop, the F-G region, the A'-A region and β1 sheet region, for the duration of the equilibrated region of the trajectories. Hydrogen bond analysis

further revealed that variations in the alleles were altering hydrogen bonding networks, which, in turn, contributed to structural and flexibility differences. The areas with notable structure and flexibility changes were occurring around channels vital for substrate entrance and exit, therefore tunnel analysis was carried out. Ultimately, reduced activity alleles were noted to have altered channel opening frequencies as compared to references and normal function alleles. They either had all channels more frequently closed throughout the simulation, or multiple channels were more frequently open or partially open simultaneously throughout the simulation, resulting in highly open enzymes that are less conducive to substrate metabolism.

Due to the changes in structure and flexibility noted in alleles, protein communication networks may have also been affected and therefore investigations at a local level using DRN metrics were conducted. The averaged *BC* metric was more informative, as through Δ*BC* calculations, highlighted the unique changes in inter-residue communication within reduced function alleles, that were not observed in normal functioning alleles. In all analysis carried out, the heme region remained stable in all alleles. Overall, from the findings in this chapter, we noted variations in CYP2D6 to have allosteric effects, as most changes within the different alleles were occurring further from the variations and more in the substrate access regions. We also deduced four potential mechanisms that lead to reduced activity in CYP2D6. These were: 1) instability and conformational changes of protein; 2) secondary structure variation and changes in flexibility of ≥3 highly important protein regions; 3) more frequently closed/ more than one frequently open or partially open channel simultaneously; 4) unique changes in inter-residue communications within the B' helix and/or β1 sheet region. We also proposed from results observed in this chapter, that

when inferring reduced functionality on alleles with unknown activity, the allele must pass at least three of the four mechanisms, as in our studies, all reduced function alleles had passed at least three of the mechanisms identified.

Finally, in chapter 4, we aimed to utilize the identified mechanisms that may possibly lead to reduced activity, to potentially predict reduced functionality in Afrocentric CYP2D6 alleles. These Afrocentric alleles were specifically selected so as to advance the pharmacogenomic research on alleles occurring in African populations, as they were understudied. A comprehensive analysis using similar methods from chapter 3 was conducted on Afrocentric alleles *149, *152, *153, *154, *155, *157, *159, *160, *162 and *163. Through our findings, the behavior of alleles *152, *153 and *159 was highly similar to that of the reduced activity alleles, as all mechanisms outlined were passed. This suggested that reduced activity in these alleles is highly likely. Alleles *155, *160 and *162 were observed to have at least three mechanisms occurring, indicating they may harbor reduced activity. However, there were no specific changes in inter-residue communications within the B' helix and/or β1 sheet region observed in these alleles, that were consistently present in all reduced activity alleles. Therefore, more analysis may be required to confirm reduced function in these alleles. The behavior of Afrocentric alleles *154, *157 and *163 did not manage to pass at least three of the mechanisms identified for reduced functionality and additionally, in alleles *157 and *163, further heme region discrepancies were noted that were inconsistent with observations made for normal and reduced activity alleles. These findings imply neither normal nor reduced activity can be predicted for these alleles and further investigations are

necessary. Lastly, allele *149 consistently behaved similarly to references and normal function alleles, suggesting this allele may have normal functionality.

In conclusion, the studies presented in this thesis have provided novel insights on CYP2D6 pharmacogenomics that may advance the progress of personalized medicine. A potentially more efficient way of assessing CYP2D6 reduced enzyme function was developed and applied on Afrocentric CYP2D6 alleles with no clinical function assigned to them at the time of this study.

5.2 Future work

Research presented here was more focused on determination of potential mechanisms that lead to reduced activity of CYP2D6 enzymes, however, variations within this enzyme can cause a range of functional consequences such as null function, reduced function, normal function, or increased function. Further experiments to decipher potential mechanisms that lead to the different activities not studied here could be highly beneficial, especially for Afrocentric CYP2D6 alleles *157 and *163 that demonstrated a behavior not observed in either normal or reduced function alleles. Furthermore, studies including the N-termina transmembrane anchor of CYP2D6 [262, 263] and the membrane could be conducted, to more accurately replicate the cellular environment of CYP2D6 and yield more precise results. *In vitro* experiments involving Afrocentric CYP2D6 alleles studied here, with known CYP2D6 substrates could be conducted to potentially validate the predicted functionalities, therefore increasing the confidence of the assignment.

Thank you for reading

Supplementary data

Supplementary Figures

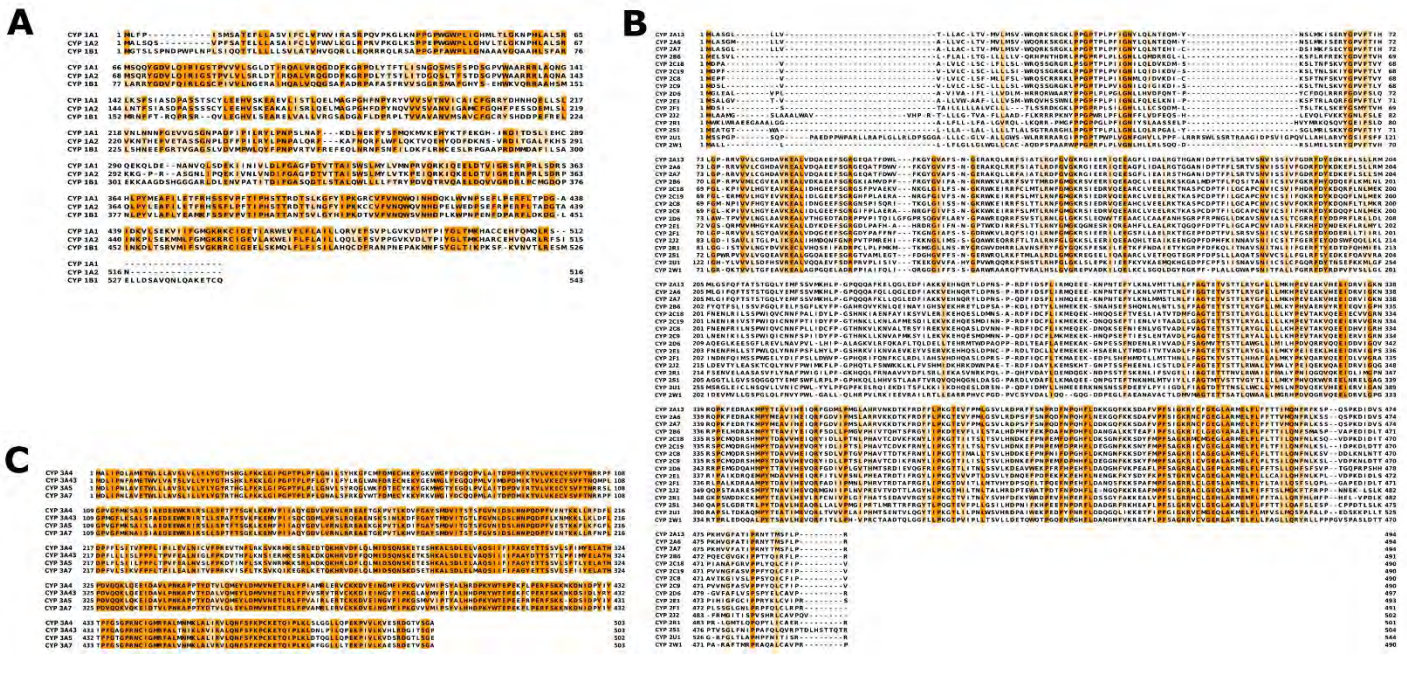


Figure S 1: MSA of CYP family 1, 2 and 3. A, B and C showing MSA of CYP families 1, 2 and 3 respectively. Alignments were produced by T-COFFEE alignment program and results viewed in Jalview.

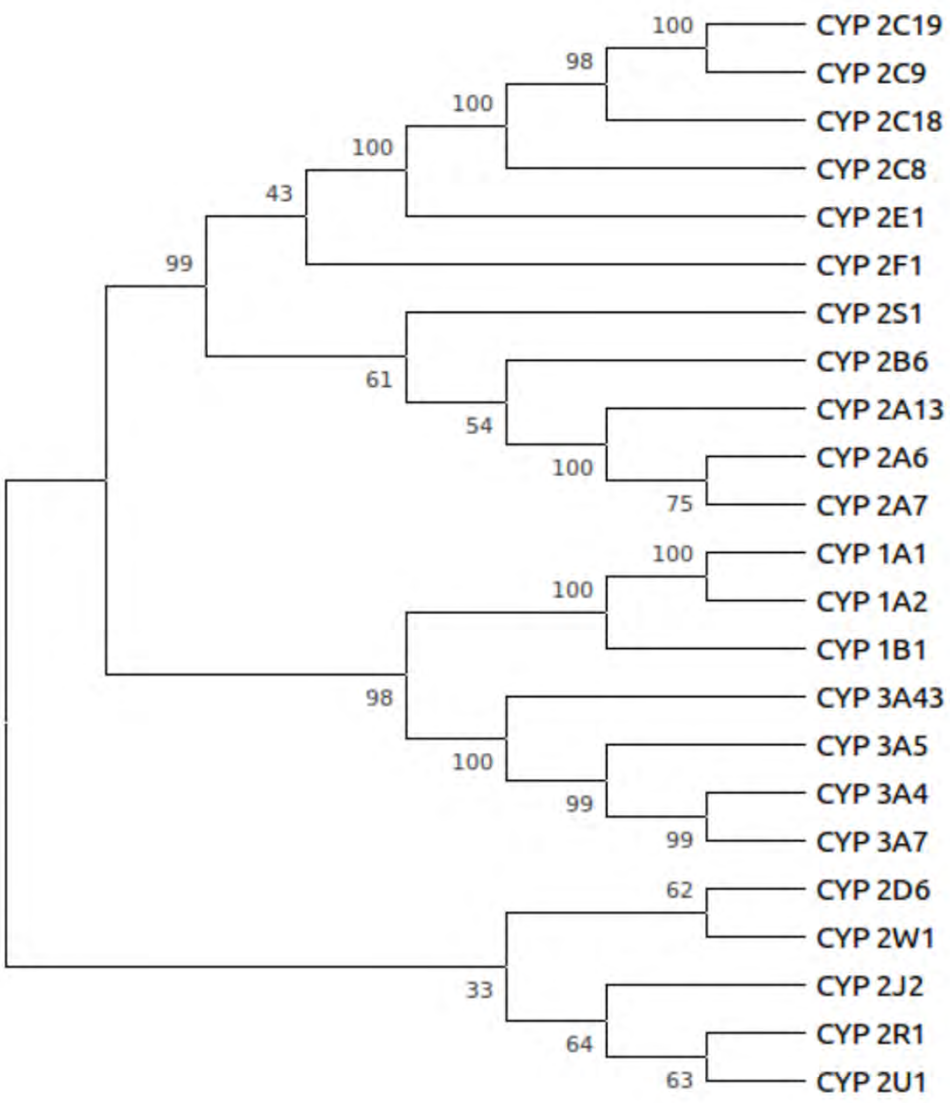


Figure S 2: Bootstrap tree for the phylogenetic tree chosen. Bootstrap tree was produced using MEGAX.

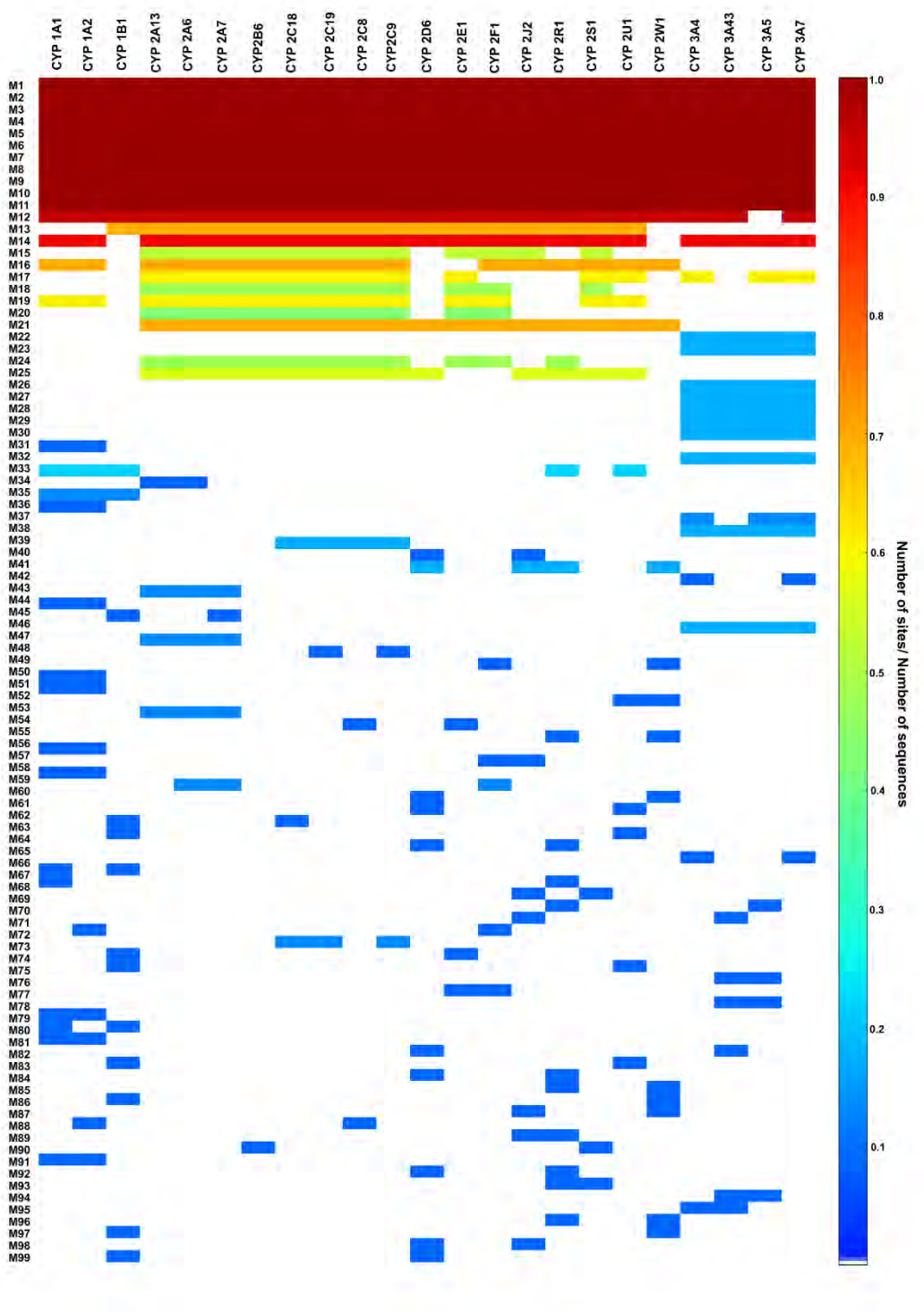


Figure S 3: All motifs identified within CYP families 1, 2 and 3 represented in a heatmap. Conservation of motifs across all CYP sequences increased from blue to red.

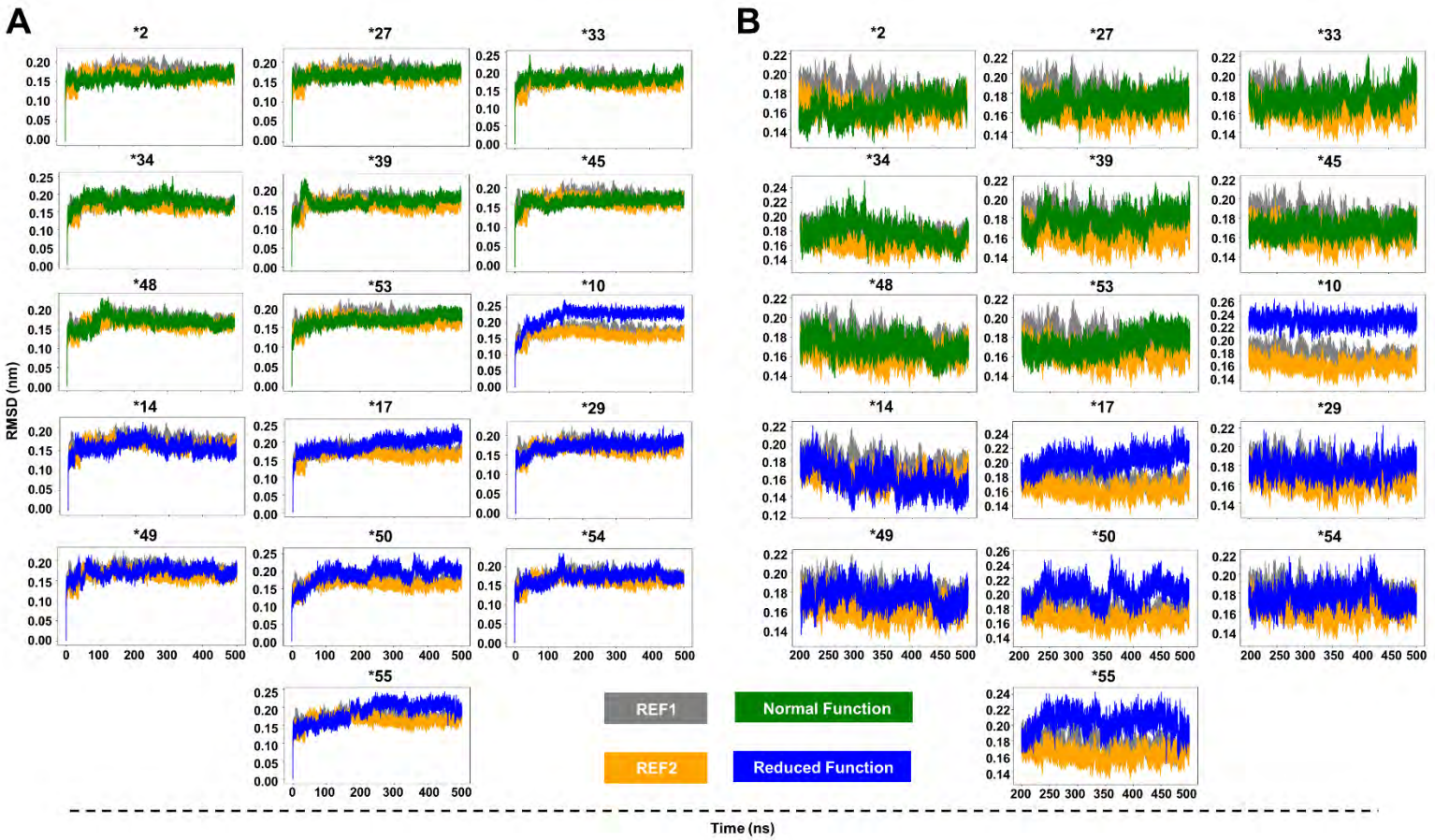


Figure S 4: RMSD line plots of the whole trajectory (500 ns) (A) and the last 300 ns (B) selected and used for subsequent analysis of references, normal activity alleles and reduced activity alleles. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue.

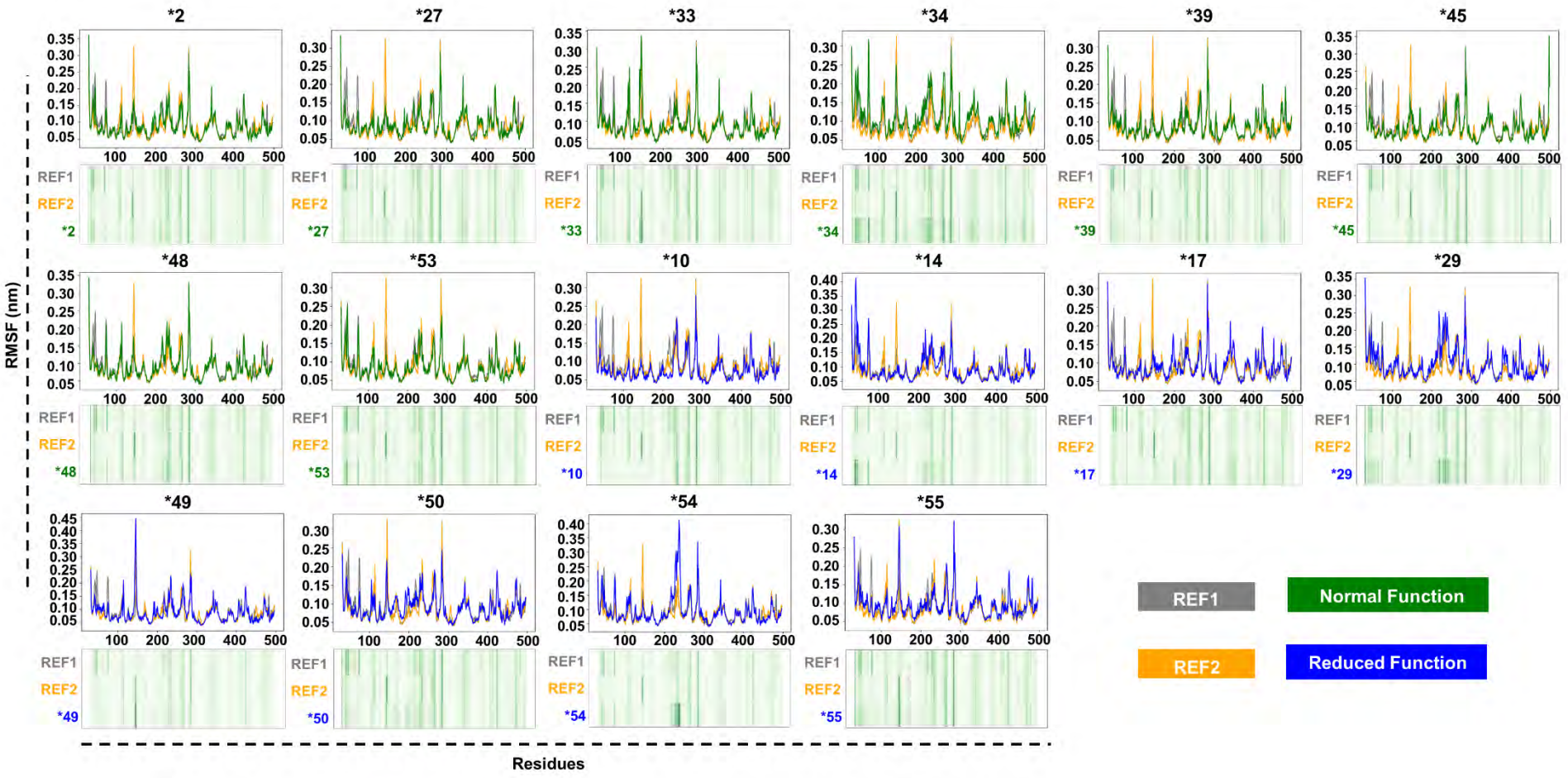


Figure S 5: RMSF line plots and heatmaps of the last 300 ns of reference and allele (normal and reduced) MD simulations. REF1 line plot is in grey, REF2 line plot is in orange, normal activity allele line plots are in green and reduced activity allele line plots are shown in blue.

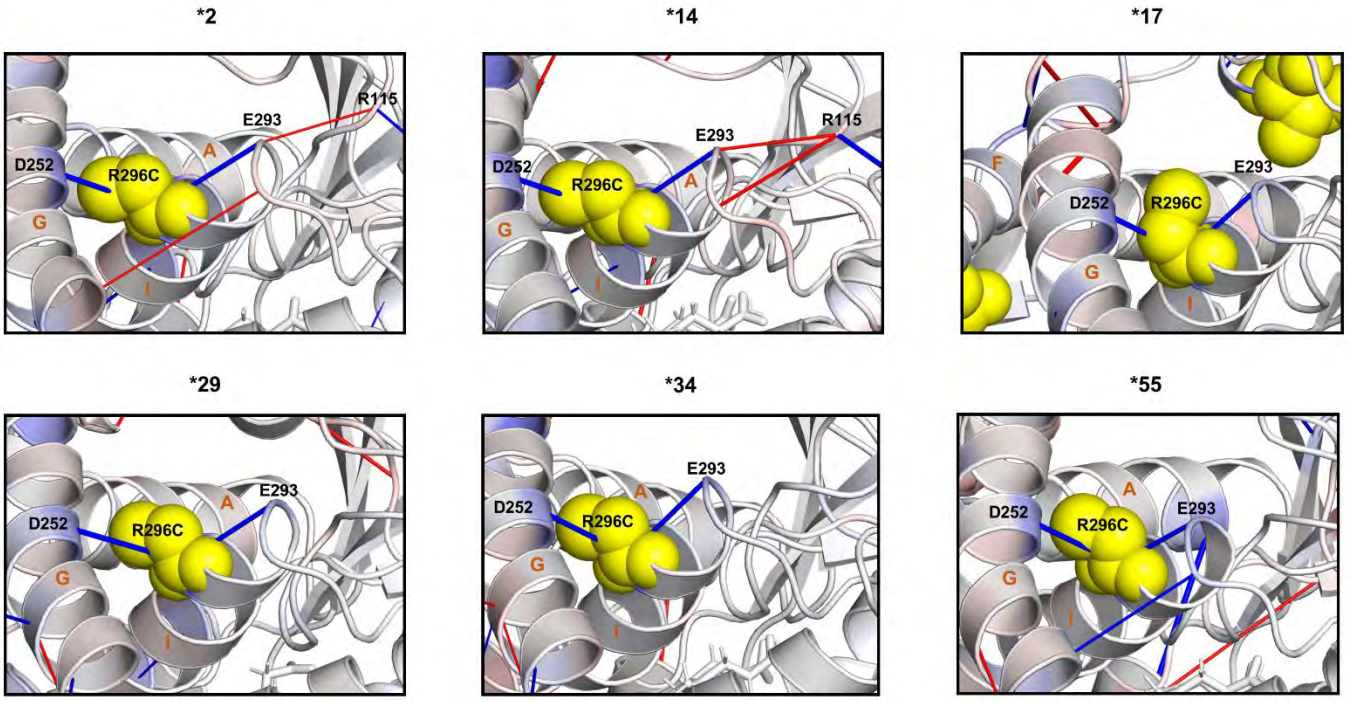


Figure S 6: Effects of variation R296C on hydrogen bonding between residue position 296 and surrounding residues. In alleles *2, *14, *17, *29, *34 and *55, there was a significant loss in hydrogen bonding noted between residue 296 and residues 252 and 293. A significant gain in hydrogen bonding was then noted between residue 293 and 115 only in alleles *2 and *14. Variations in each allele shown by yellow spheres. ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive).

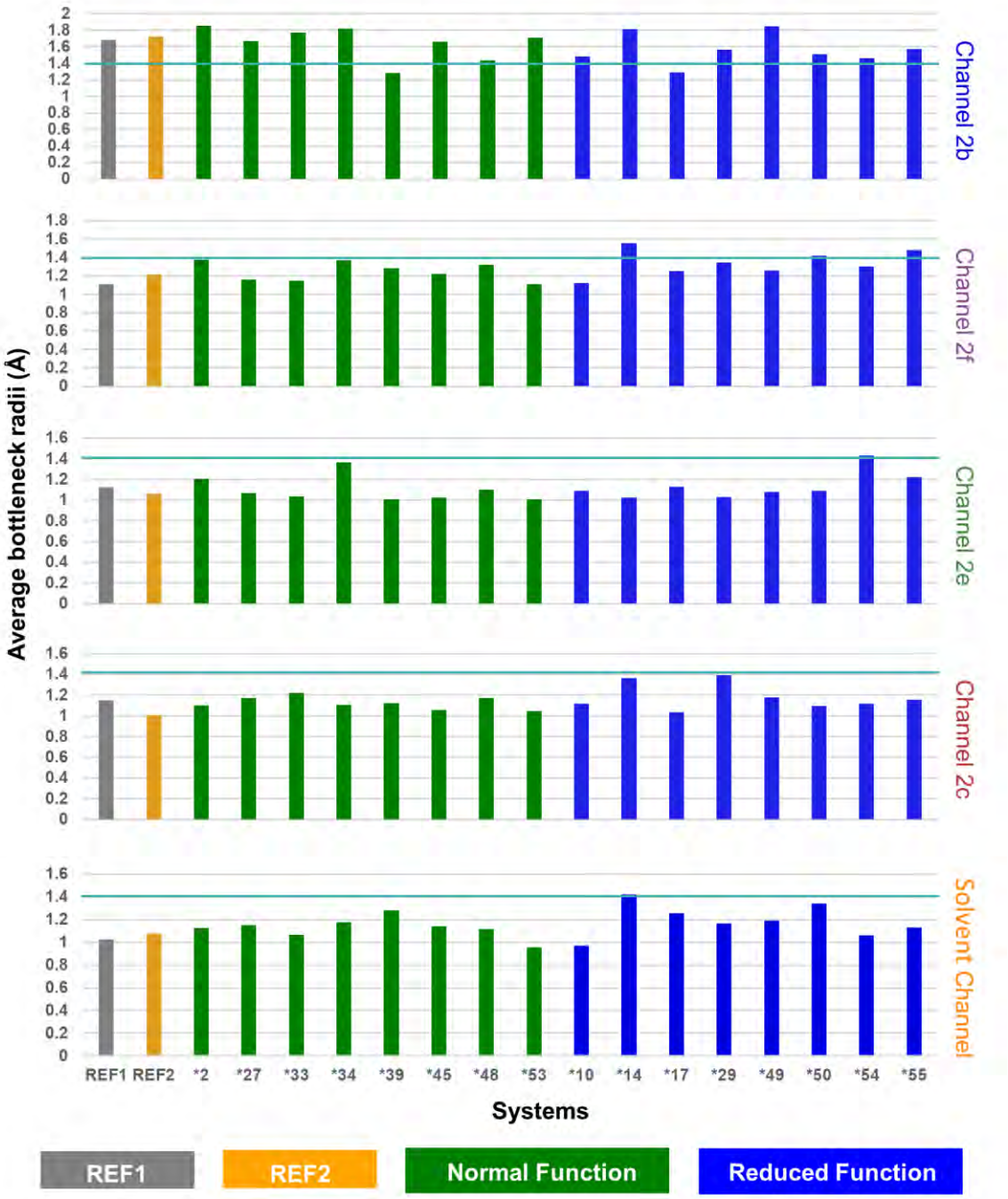


Figure S 7: Calculated average bottleneck radii for each channel in each system represented in bar graphs. Bars within the bar graphs are coloured according to the following: REF1 in grey, REF2 in orange, normal activity alleles in green and reduced activity alleles in blue. Red line indicating the cut-off bottleneck radii used to determine whether a channel is open or closed (≥ 1.4 Å).

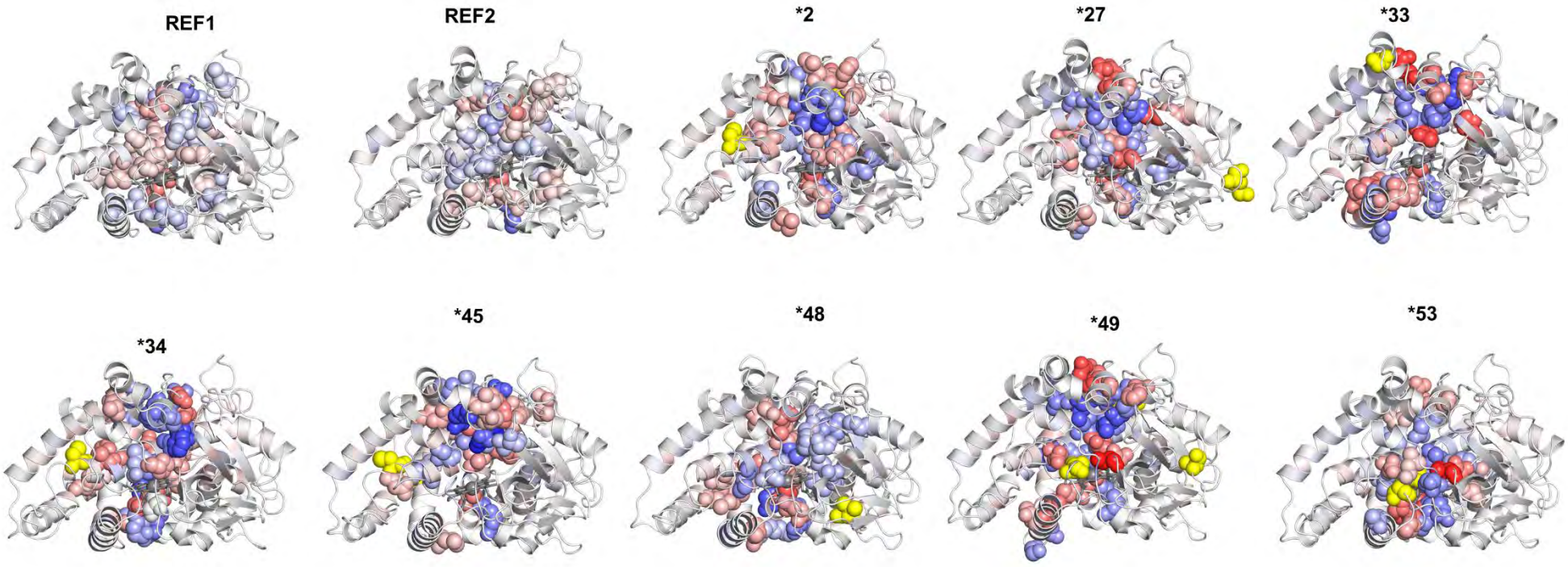


Figure S 8: Δ*BC* analysis. Residues with top and bottom 3% Δ*BC* values mapped to reference and allele conformational clustering structures and shown as spheres. Normal alleles *2, *27, *33, *34, *45, *48 and *53, and reduced function allele *49 with no significantly changing averaged *BC* values for residues within the B' helix and/or the loop between β1-1 and β1-2 shown. Δ*BC* values range from positive to negative. Residues with a loss in averaged *BC* within the alleles are coloured in blue, and residues with gain in averaged *BC* in alleles are coloured in red. Variation positions for alleles shown in spheres and coloured in yellow.

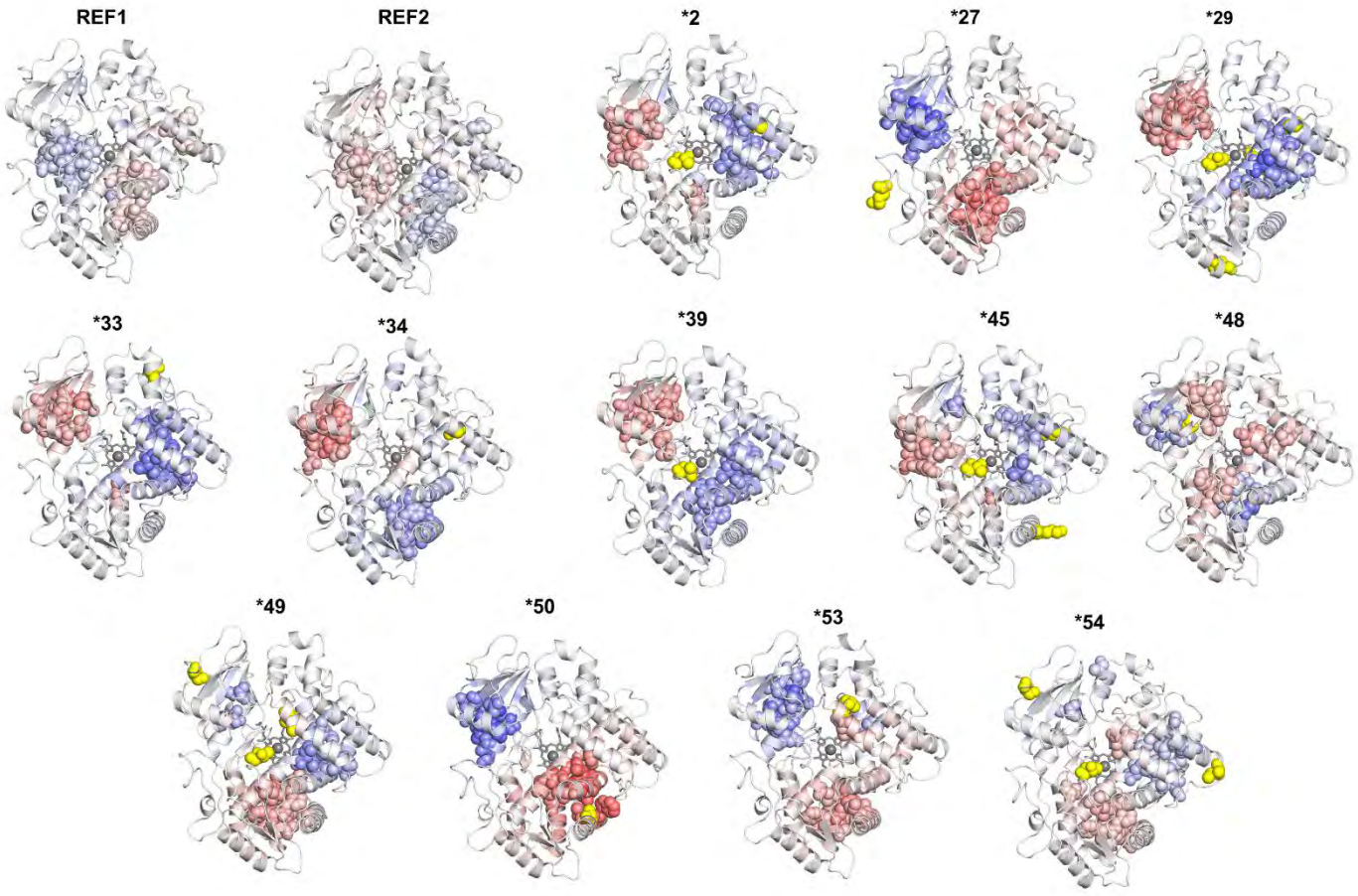


Figure S 9: Δ*EC* analysis. Residues with top and bottom 3% Δ*EC* values mapped to reference and allele conformational clustering structures and shown as spheres. Δ*EC* values range from positive to negative. Residues with a loss in averaged *EC* within the alleles are coloured in blue, and residues with gain in averaged *EC* in alleles are coloured in red. Variation positions for alleles shown in spheres and coloured in yellow.

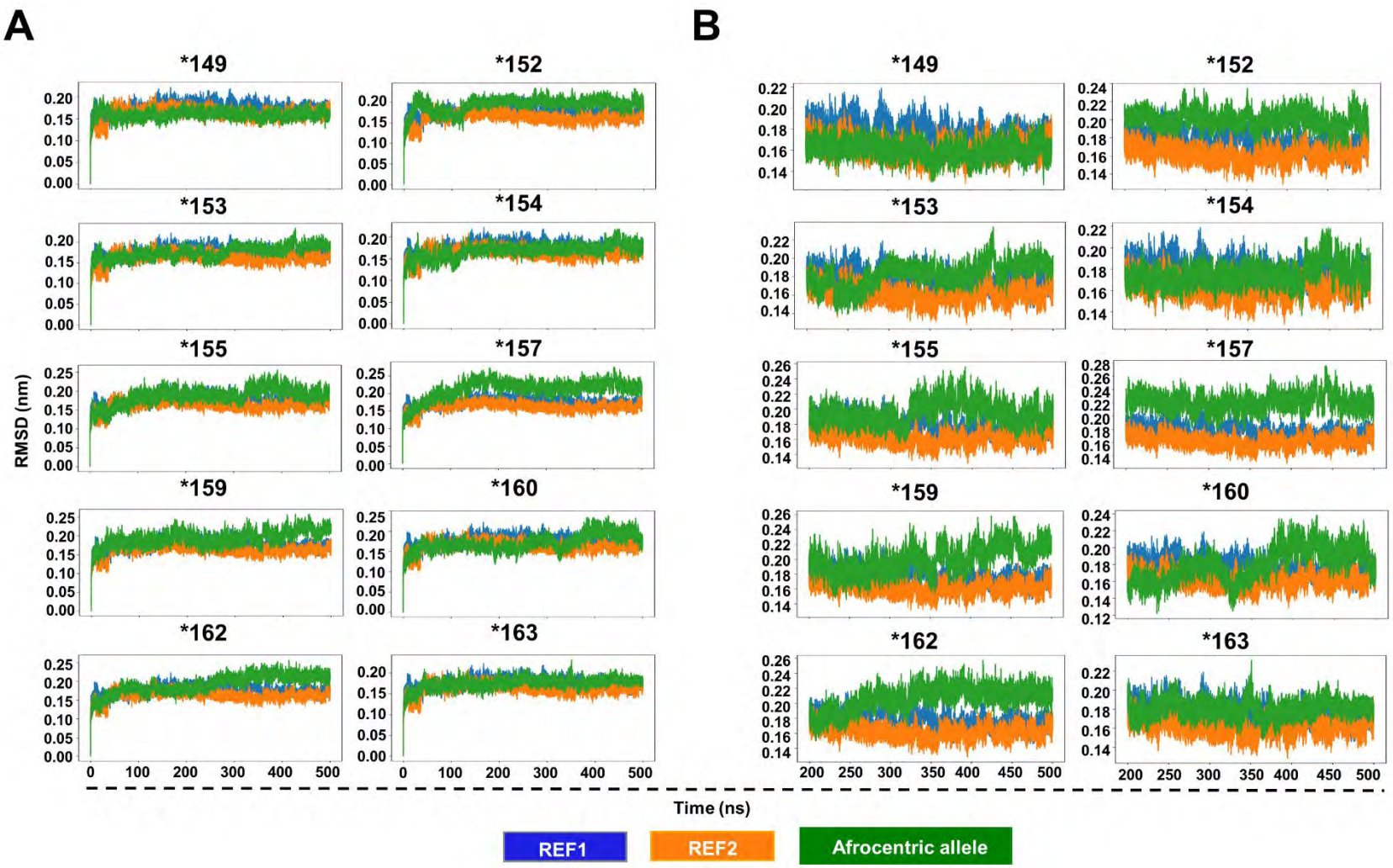


Figure S 10: A) RMSD line plots of the whole trajectory (500 ns). B) The last 300 ns selected and used for subsequent analysis of references, and Afrocentric alleles. REF1 line plot is in blue, REF2 line plot is in orange and Afrocentric allele line plots are in green.

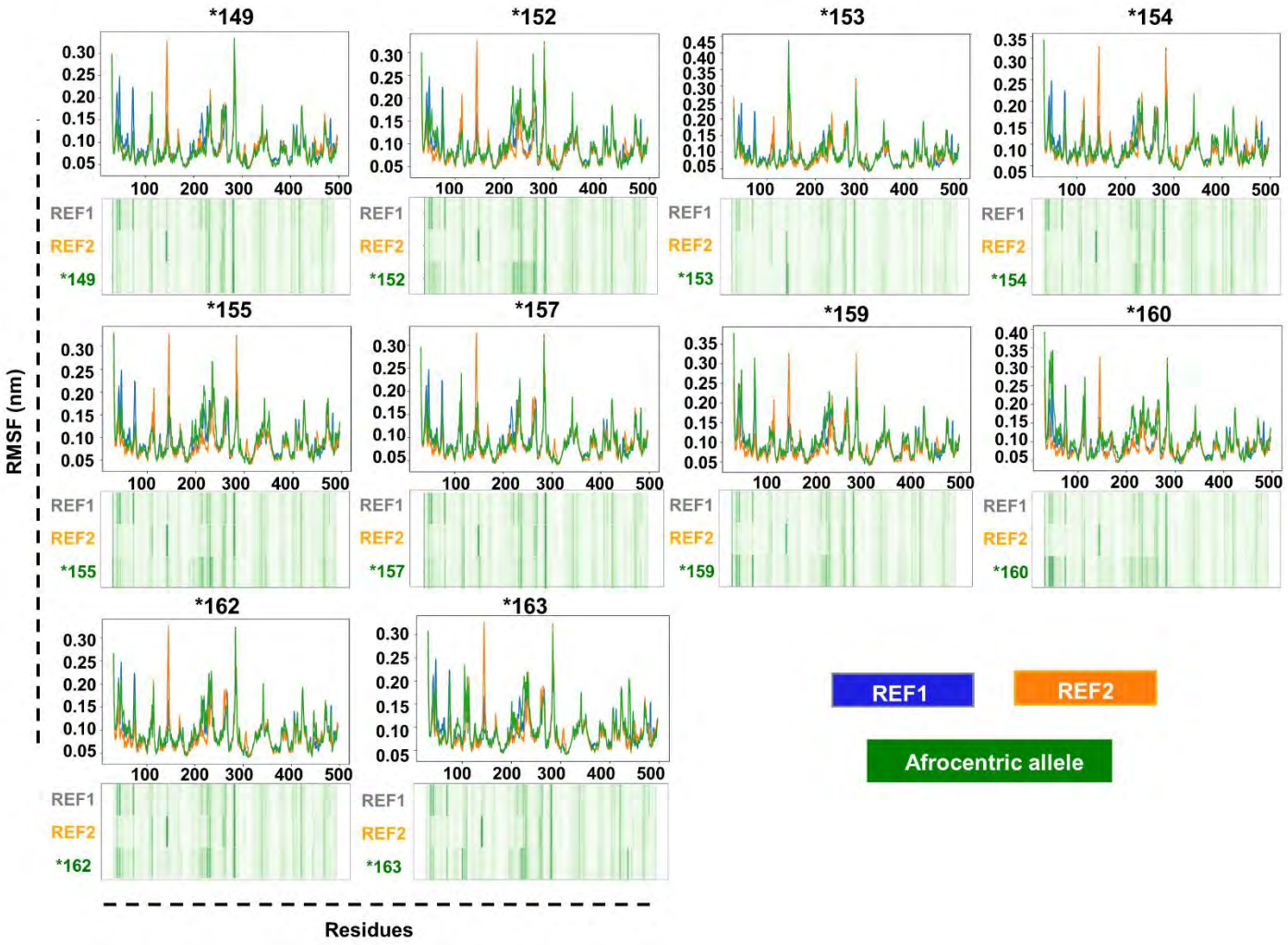


Figure S 11: RMSF line plots and heatmaps of the last 300 ns of reference and Afrocentric alleles. REF1 line plot is in blue, REF2 line plot is in orange and Afrocentric allele line plots are in green.

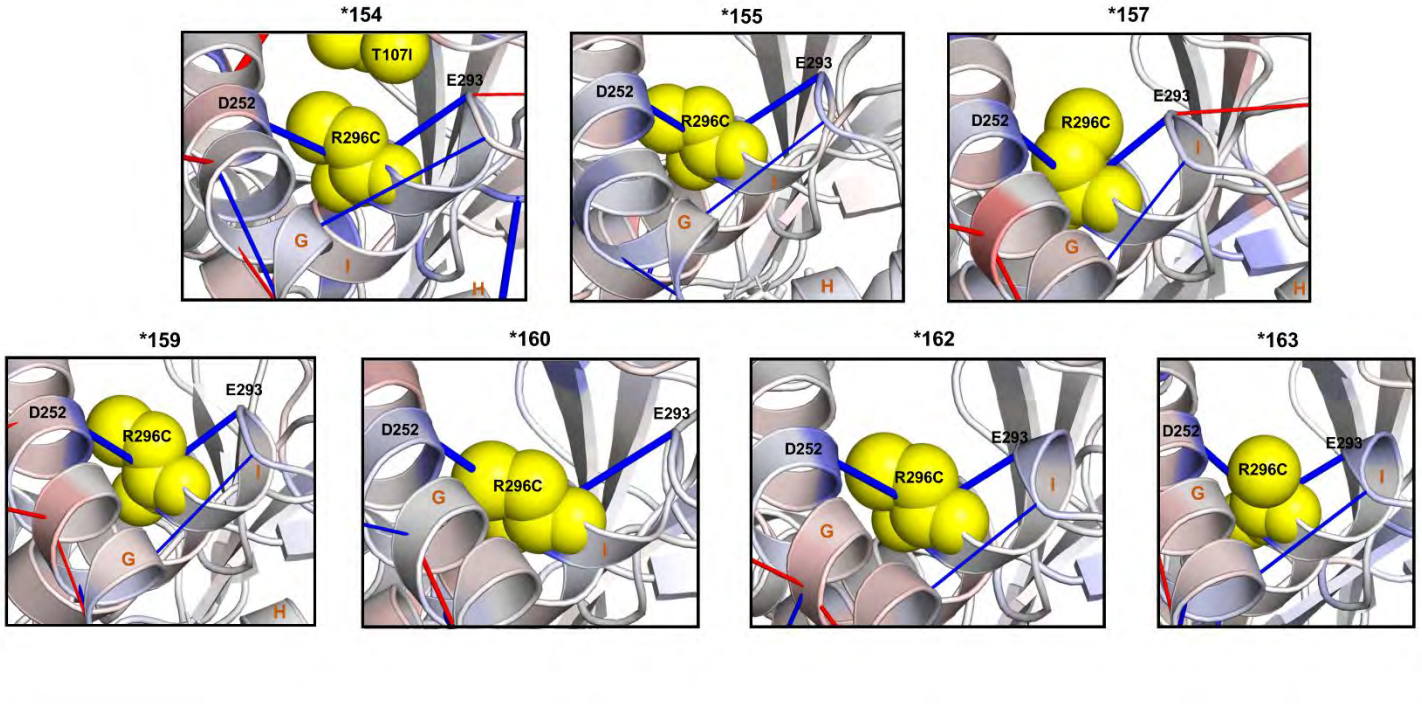


Figure S 12: Hydrogen bond analysis for Afrocentric alleles showing effects of variation R296C on hydrogen bonding between residue position 296 and surrounding residues. In alleles *154, *155, *157, *159, *160, *162, *163, there was a significant loss in hydrogen bonding noted between residue 296 and residues 252 and 293. Variations in each allele shown by yellow spheres. ΔH-bond values range from positive to negative. Residue pairs with loss in hydrogen bonding within the systems are coloured in blue (negative), and residue pairs with gain in hydrogen bonding are coloured in red (positive).

Figure S 13: Average bottleneck radii for each channel in each system calculated and represented in bar graphs. Bars are coloured as follows: REF1 is in blue, REF2 in orange and Afrocentric alleles in green. Dark green line indicates the cut-off bottleneck radii used to determine whether a channel is open or closed (≥ 1.4 Å).

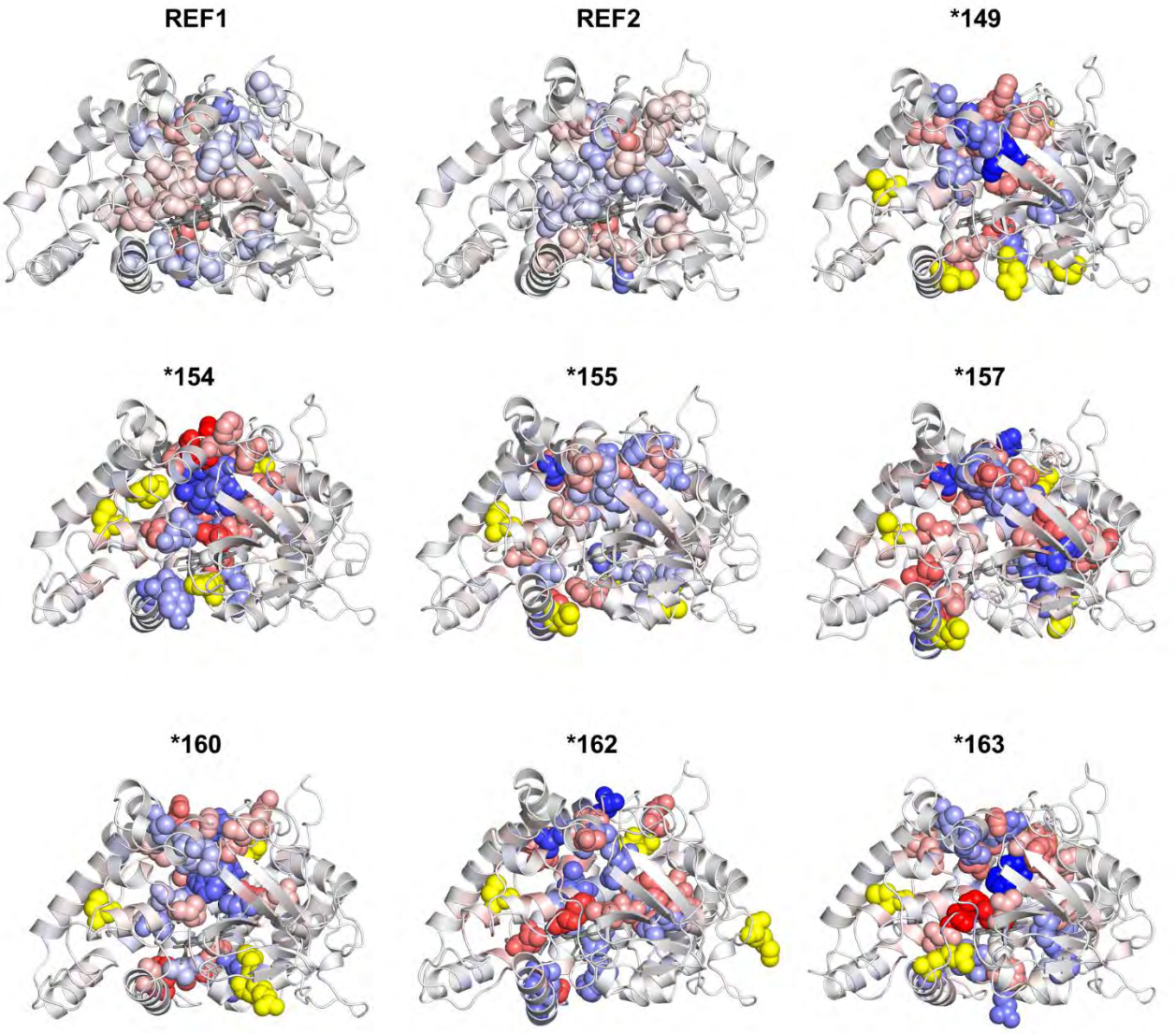


Figure S 14: Δ*BC* analysis of Afrocentric alleles. Residues with top and bottom 3% Δ*BC* values mapped to reference and Afrocentric allele conformational clustering structures and shown as spheres. Alleles *149, *154, *155, *157, *160, *162 and *163 with no significantly changing averaged *BC* values for residues within the B' helix and/or the loop between β1-1 and β1-2 shown. In these alleles, Δ*BC* values range from positive to negative. Residues with a loss in averaged *BC* within the alleles are coloured in blue, and residues with gain in averaged *BC* in alleles are coloured in red. Variation positions for Afrocentric alleles shown in spheres and coloured in yellow.

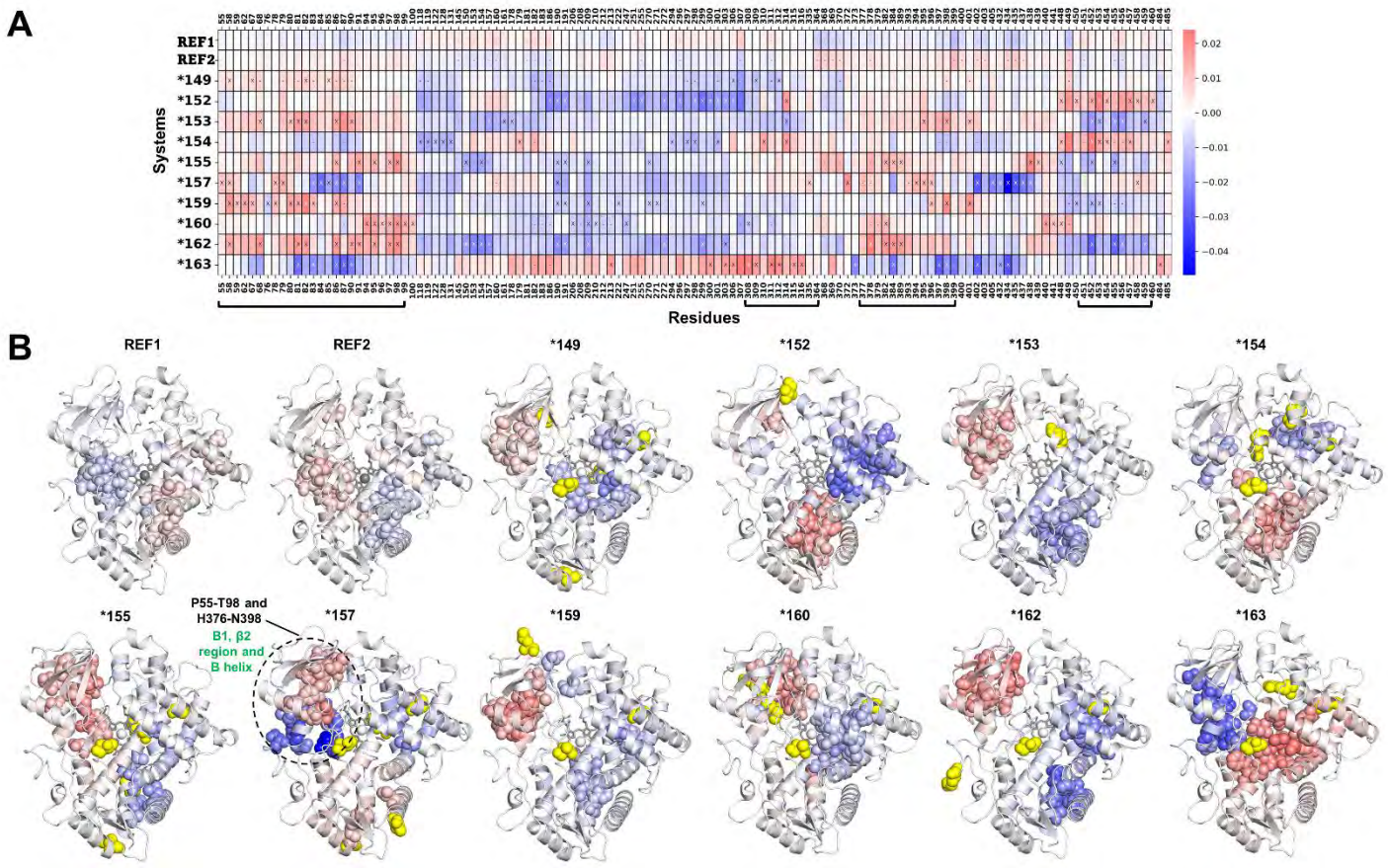


Figure S 15: Δ*EC* analysis of Afrocentric alleles. Residues with top and bottom 3% Δ*EC* values shown in a heatmap (A) and mapped to their corresponding conformational cluster structures (B). In the heatmap, significant values (outliers not within 3SD of the mean of the reference data) are marked with an "X" and top and bottom 3% Δ*EC* values that are not significant are marked with a "-". For B), variation positions for alleles are shown in spheres and coloured in yellow. For both A) and B), the *ΔEC* values range from positive to negative, with positive indicating a gain in averaged *EC* within the alleles (shown in red colour) and negative indicating a loss in averaged *EC* in alleles (shown in blue colour).

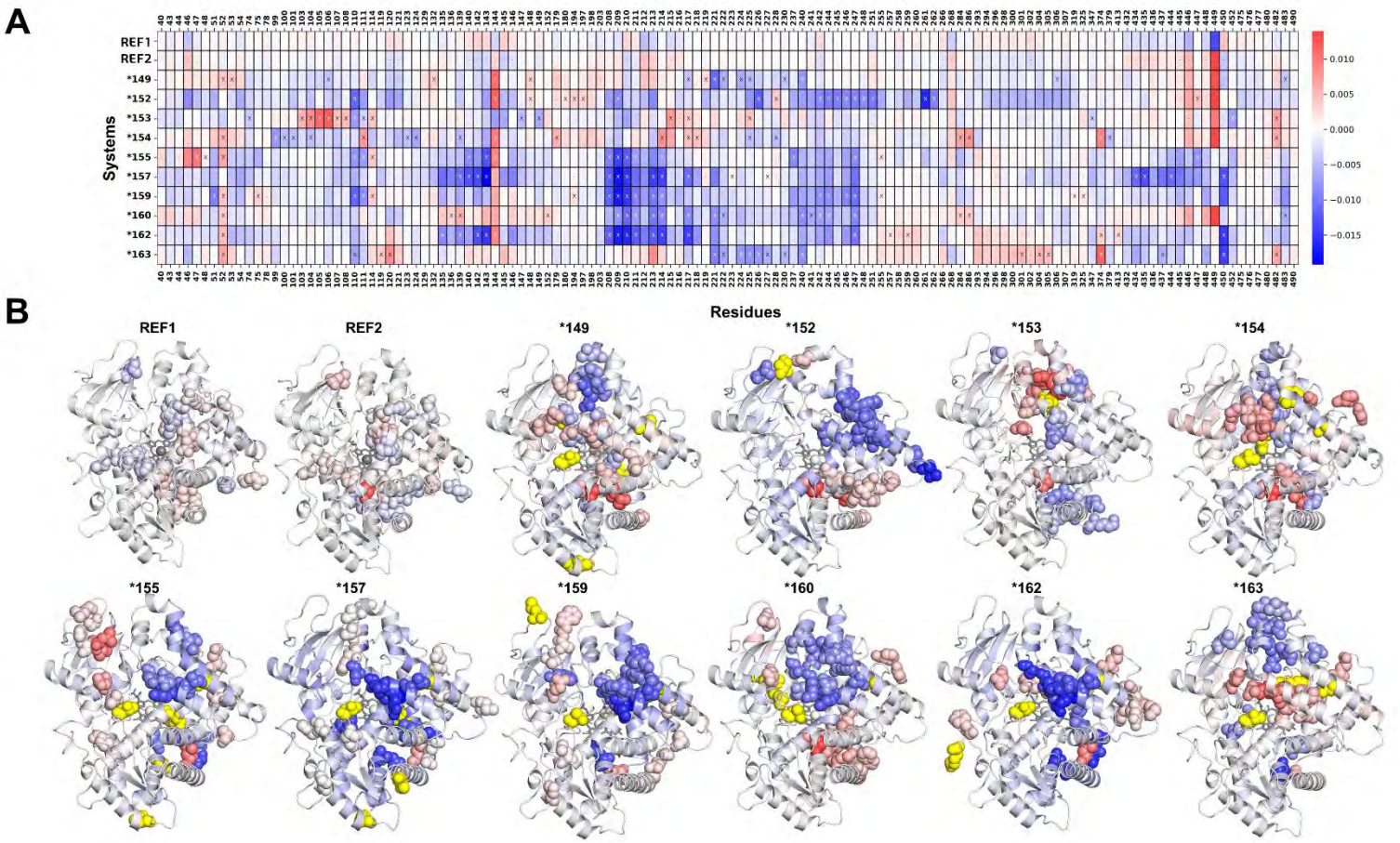


Figure S 16: Δ*CC* analysis of Afrocentric alleles. Residues with top and bottom 3% Δ*CC* values shown in a heatmap (A) and mapped to each systems' corresponding conformational cluster structures (B). The heatmap shows significant top and bottom 3% Δ*CC* values (outliers not within 3SD of the mean of the reference data) with an "X" and values that are not significant with a "-". For B), variation positions for alleles are shown in spheres and coloured in yellow. For both A) and B), the *ΔCC* values range from positive to negative, with positive showing a gain in averaged *CC* within the alleles (in red) and negative showing a loss in averaged *CC* in alleles (in blue).

Supplementary Tables

Table S 1: Cytochrome P450 enzymes from families 1, 2 and 3 associated with metabolism of antimalarial and antituberculosis drugs with their alleles and amino acid variation positions. Adapted from Supplementary data, Chamboko *et al.*, 2023 [3].

CYP family	CYP enzyme	Antimalarial/antituberculosis drug affected	Allele name	Amino acid mutation position
CYP 1	CYP1A1	Bedaquiline	CYP1A1*4	T461N
			CYP1A1*5	R464S
			CYP1A1*6	M331I
			CYP1A1*8	I448N
			CYP1A1*9	R464C
			CYP1A1*10	R477W
			CYP1A1*11	P492R
			CYP1A1*12	A62P
			CYP1A1*13	G45D
	CYP1A2	Mefloquine, primaquine	CYP1A2*2	F21L
			CYP1A2*3	D348N
			CYP1A2*4	I386F
			CYP1A2*5	C406Y
			CYP1A2*6	R431W
			CYP1A2*8	R456H
			CYP1A2*9	T83M
			CYP1A2*10	E168Q
			CYP1A2*11	F186L
			CYP1A2*12	S212C
			CYP1A2*13	G299S
			CYP1A2*14	T438I
			CYP1A2*15	P42R
			CYP1A2*16	R377Q
			CYP1A2*17	T395M
			CYP1A2*18	N397H
			CYP1A2*19	R510Q
			CYP1A2*20	D436N
	CYP1B1	Linezolid	CYP1B1*2	R48G; A119S
			CYP1B1*3	L432V
			CYP1B1*4	N453S
			CYP1B1*5	R48G; L432V
			CYP1B1*6	R48G; A119S; L432V

			CYP1B1*7	R48G; A119S; L432V; A443G
			CYP1B1*8	L432V; D441H
			CYP1B1*11	W57C
			CYP1B1*12	G61E
			CYP1B1*18	G365W
			CYP1B1*19	P379L
			CYP1B1*20	E387K
			CYP1B1*21	R390H
			CYP1B1*23	P437L
			CYP1B1*25	R469W
CYP 2	CYP2A6	Artemisinin, artesunate	CYP2A6*2	L160H
			CYP2A6*5	G479V
			CYP2A6*6	R128Q
			CYP2A6*7	I471T
			CYP2A6*8	R485L
			CYP2A6*10	I471T; R485L
			CYP2A6*11	S224P
			CYP2A6*13	G5R
			CYP2A6*14	S29N
			CYP2A6*15	K194E
			CYP2A6*16	R203S
			CYP2A6*17	V365M
			CYP2A6*18	Y392F
			CYP2A6*19	Y392F; I471T
			CYP2A6*23	R203C
			CYP2A6*24	V110L; N438Y
			CYP2A6*25	F118L
			CYP2A6*26	F118L; R128L; S131A
			CYP2A6*28	N418D; E419D
			CYP2A6*31	M6L
			CYP2A6*34	M6L; V13A; K30R; Q53H; M54I; Y55C; N56D; L58I; I61F; R64C; V117A; R128L; S131A; A153S; D158E; L160I; G162S; G164H; K194E; R203S; M204V
			CYP2A6*35	N438Y
			CYP2A6*36	N438Y; I471T
			CYP2A6*37	N438Y; I471T; R485L
			CYP2A6*38	Y351H
			CYP2A6*39	V68M
			CYP2A6*40	I149M

			CYP2A6*41	R265Q
			CYP2A6*42	I268T
			CYP2A6*43	T303I
			CYP2A6*44	E390K; N418D; E419D
			CYP2A6*45	L462P
	CYP2B6	Artemisinin, artesunate, arteether	CYP2B6*2	R22C
			CYP2B6*3	S259R
			CYP2B6*4	K262R
			CYP2B6*5	R487C
			CYP2B6*6	Q172H; K262R
			CYP2B6*7	Q172H; K262R; R487C
			CYP2B6*8	K139E
			CYP2B6*9	Q172H
			CYP2B6*10	Q21L; R22C
			CYP2B6*11	M46V
			CYP2B6*12	G99E
			CYP2B6*13	K139E; Q172H; K262R
			CYP2B6*14	R140Q
			CYP2B6*15	I391N
			CYP2B6*16	K262R; I328T
			CYP2B6*17	T26S; D28G; R29T
			CYP2B6*18	I328T
			CYP2B6*19	Q172H; K262R; R336C
			CYP2B6*20	T168I; Q172H; K262R
			CYP2B6*21	P428T
			CYP2B6*23	M459V
			CYP2B6*24	G476D
			CYP2B6*25	Q485L
			CYP2B6*26	P167A; Q172H; K262R
			CYP2B6*27	M198T
			CYP2B6*31	L313I
			CYP2B6*32	A407T
			CYP2B6*33	R487S
			CYP2B6*34	Q172H; K262R; R487S
			CYP2B6*35	G110V; I114T; E148D; M198T; A279P
			CYP2B6*36	Q172H; K262R
			CYP2B6*37	Q172H; V183G; K262R
	CYP2C8	Chloroquine, amodiaquine, bedaquiline	CYP2C8*2	I269F
			CYP2C8*3	R139K; K399R
			CYP2C8*4	I264M
			CYP2C8*6	G171S
			CYP2C8*8	R186G

			CYP2C8*9	K247R
			CYP2C8*10	K383N
			CYP2C8*13	I223M
			CYP2C8*14	A238P
			CYP2C8*15	V181I
			CYP2C8*16	I331T
			CYP2C8*17	I244V
			CYP2C8*18	L361F
	CYP2C9	Dapsone	CYP2C9*2	R144C
			CYP2C9*3	I359L
			CYP2C9*4	I359L
			CYP2C9*5	D360E
			CYP2C9*7	L19I
			CYP2C9*8	R150H
			CYP2C9*9	H251R
			CYP2C9*10	E272G
			CYP2C9*11	R335W
			CYP2C9*12	P489S
			CYP2C9*13	L90P
			CYP2C9*14	R125H
			CYP2C9*16	T299A
			CYP2C9*17	P382S
			CYP2C9*18	I359L; D397A
			CYP2C9*19	Q454H
			CYP2C9*20	G70R
			CYP2C9*21	P30L
			CYP2C9*22	N41D
			CYP2C9*23	V76M
			CYP2C9*24	E354K
			CYP2C9*26	T130R
			CYP2C9*27	R150L
			CYP2C9*28	Q214L
			CYP2C9*29	P279T
			CYP2C9*30	A477T
			CYP2C9*31	I327T
			CYP2C9*32	V490F
			CYP2C9*33	R132Q
			CYP2C9*34	R335Q
			CYP2C9*35	R125L; R144C
			CYP2C9*36	M1V
			CYP2C9*37	D49G
			CYP2C9*38	G96A
			CYP2C9*39	G98V

			CYP2C9*40	F110S
			CYP2C9*41	K119R
			CYP2C9*42	R124Q
			CYP2C9*43	R124W
			CYP2C9*44	T130M
			CYP2C9*45	R132W
			CYP2C9*46	A149T
			CYP2C9*47	P163L
			CYP2C9*48	I207T
			CYP2C9*49	I222V
			CYP2C9*50	P227S
			CYP2C9*51	I284V
			CYP2C9*52	T299R
			CYP2C9*53	P317S
			CYP2C9*54	S343R
			CYP2C9*55	L361I
			CYP2C9*56	I387V
			CYP2C9*57	N204H
			CYP2C9*58	P337T
			CYP2C9*59	I434F
			CYP2C9*60	L467P
			CYP2C9*61	R144C; N457S
			CYP2C9*62	R125C
			CYP2C9*63	R144H
			CYP2C9*64	S280C
			CYP2C9*65	E326D
			CYP2C9*66	L362V
			CYP2C9*67	R433W
			CYP2C9*69	D191G
			CYP2C9*70	G442V
			CYP2C9*71	E272G; P489S
			CYP2C9*72	A149V
			CYP2C9*73	R150C
			CYP2C9*74	Q214H
			CYP2C9*75	N418T
			CYP2C9*76	G40R
			CYP2C9*77	G96R
			CYP2C9*78	L128R
			CYP2C9*79	P346L
			CYP2C9*80	C13G
			CYP2C9*81	E300V
			CYP2C9*82	P227L
			CYP2C9*83	L36F

			CYP2C9*84	G431R
	CYP2C18	Bedaquiline	-	-
	CYP2C19	Proguanil, chlorproguanil quinine, dapsone	CYP2C19*1	I331V
			CYP2C19*4	M1V; I331V
			CYP2C19*5	R433W
			CYP2C19*6	R132Q; I331V
			CYP2C19*8	W120R
			CYP2C19*9	R144H; I331V
			CYP2C19*10	P227L; I331V
			CYP2C19*11	R150H; I331V
			CYP2C19*13	I331V; R410C
			CYP2C19*14	L17P; I331V
			CYP2C19*15	I19L; I331V
			CYP2C19*16	R442C
			CYP2C19*17	I331V
			CYP2C19*18	R329H; I331V
			CYP2C19*19	S51G; I331V
			CYP2C19*22	R186P; I331V
			CYP2C19*23	G91R; I331V
			CYP2C19*24	I331V; R335Q
			CYP2C19*25	I331V; F448L
			CYP2C19*26	D256N; I331V
			CYP2C19*28	I19L; I331V; V374I
			CYP2C19*29	K28I; I331V
			CYP2C19*30	R73C
			CYP2C19*31	H78Y; I331V
			CYP2C19*32	H99R; I331V
			CYP2C19*33	D188N; I331V
			CYP2C19*34	P3S; F4L
			CYP2C19*39	I19L; E122A; I331V
	CYP2D6	Chloroquine, primaquine	CYP2D6*2	R296C; S486T
			CYP2D6*7	H324P
			CYP2D6*10	P34S; S486T
			CYP2D6*12	G42R; R296C; S486T
			CYP2D6*14	G169R; R296C; S486T
			CYP2D6*17	T107I; R296C; S486T
			CYP2D6*22	R28C
			CYP2D6*23	A85V
			CYP2D6*24	I297L
			CYP2D6*25	R343G
			CYP2D6*26	I369T
			CYP2D6*27	E410K
			CYP2D6*28	V7M; Q151E; R296C; S486T

			CYP2D6*29	V136I; R296C; V338M; S486T
			CYP2D6*31	R296C; R440H; S486T
			CYP2D6*33	A237S
			CYP2D6*34	R296C
			CYP2D6*35	V11M; R296C; S486T
			CYP2D6*36	P34S; P469A; T470A; H478S; G479R; F481V; A482S; S486T
			CYP2D6*37	P34S; R201H; S486T
			CYP2D6*39	S486T
			CYP2D6*43	R26H
			CYP2D6*45	E155K; R296C; S486T
			CYP2D6*46	R26H; E155K; R296C; S486T
			CYP2D6*47	R25W; P34S; S486T
			CYP2D6*48	A90V
			CYP2D6*49	P34S; F120I; S486T
			CYP2D6*50	E156A
			CYP2D6*51	R296C; E334A; S486T
			CYP2D6*52	P34S; E418K; S486T
			CYP2D6*53	F120I; A122S
			CYP2D6*54	P34S; T261I; S486T
			CYP2D6*55	R296C; K404Q; S486T
			CYP2D6*57	P34S; R62W; P469A; T470A; H478S; G479R; F481V; A482S; S486T
			CYP2D6*62	R441C
			CYP2D6*64	P34S; T107I; S486T
			CYP2D6*65	P34S; R296C; S486T
			CYP2D6*70	V119M; V136I; V338M; S486T
			CYP2D6*71	G42E
			CYP2D6*72	P34S; E383K; S486T
			CYP2D6*73	V104M; R296C; S486T
			CYP2D6*74	L91M
			CYP2D6*75	R441H
			CYP2D6*82	L91M; H94R; V104A; T107Y; I109V
			CYP2D6*83	P469A; T470A; H478S; G479R; F481V; A482S
			CYP2D6*84	P267H; R296C; S486T
			CYP2D6*85	R296C; H478Q; S486T
			CYP2D6*86	E278K; M279K

			CYP2D6*87	A5V; P34S; S486T
			CYP2D6*88	V104A; S486T
			CYP2D6*89	L142S
			CYP2D6*90	K147R
			CYP2D6*93	T249P
			CYP2D6*94	P34S; D337G; S486T
			CYP2D6*95	P34S; R388H; S486T
			CYP2D6*97	F457L
			CYP2D6*98	R296C; H463D; S486T
			CYP2D6*99	P34S; R88P; S486T
			CYP2D6*102	A90V; R296C; S486T
			CYP2D6*103	A90V; N166D; R296C; S486T
			CYP2D6*104	E156V; R296C; S486T
			CYP2D6*105	R296C; F366S; S486T
			CYP2D6*106	E418K
			CYP2D6*107	V136M
			CYP2D6*108	H352R; Y355C
			CYP2D6*110	G445R
			CYP2D6*111	G111S; R296C; S486T
			CYP2D6*112	S488F
			CYP2D6*113	G439D
			CYP2D6*114	P34S; G169R; R296C; S486T
			CYP2D6*116	R330P
			CYP2D6*117	R296C; D337N; S486T
			CYP2D6*118	T310A
			CYP2D6*121	N166D; H167Q; S168A; R296C; S486T
			CYP2D6*122	V370I
			CYP2D6*125	R296C; R450H; S486T
			CYP2D6*126	S135F; R296C; S486T
			CYP2D6*127	Y355C; R365H
			CYP2D6*128	V68G; R296C; S486T
			CYP2D6*130	C191L
			CYP2D6*131	L231P
			CYP2D6*132	P34S; N285S; S486T
			CYP2D6*133	R296C; G340R; S486T
			CYP2D6*134	R344Q
			CYP2D6*135	R296C; R414C; S486T
			CYP2D6*136	R296C; P430L; S486T
			CYP2D6*137	R497H
			CYP2D6*139	R365H
			CYP2D6*140	L230F

			CYP2D6*141	T107I; R296C; E410K; P469A; T470A; H478S; G479R; F481V; A482S; S486T
			CYP2D6*142	R28C; P34S; S486T
			CYP2D6*146	R26H; R296C; S486T
			CYP2D6*148	R26H; R296C; E410K; S486T
			CYP2D6*149	D97E; V136I; R296C; V338M; S486T
			CYP2D6*150	L213P; R296C; S486T
			CYP2D6*151	N175S
			CYP2D6*152	A226V
			CYP2D6*153	E215K
			CYP2D6*154	T107I; R296C; R441H; S486T
			CYP2D6*155	V136I; R296C; V338M; M451I; S486T
			CYP2D6*157	V136I; A165V; R296C; V338M; S486T
			CYP2D6*159	P41L; R296C; S486T
			CYP2D6*160	L91M; H94R; R296C; S486T
			CYP2D6*162	R296C; 410K; S486T
			CYP2D6*163	W128R; R296C; S486T
			CYP2D6*164	V136I; S486T
			CYP2D6*165	V136I; R173H; R296C; V338M; S486T
			CYP2D6*167	Q117R
			CYP2D6*168	G42E; E196K
			CYP2D6*169	E402L
			CYP2D6*170	P430L
			CYP2D6*171	V119M; V136I; R296C; V338M; S486T
	CYP2J2	Linezolid	CYP2J2*2	T143A
			CYP2J2*3	R158C
			CYP2J2*4	I192N
			CYP2J2*5	D342N
			CYP2J2*6	N404Y
			CYP2J2*8	G312R
			CYP2J2*9	P351L
			CYP2J2*10	P115L
CYP 3	CYP3A4	Quinine, quinidine, chloroquine,	CYP3A4*2	S222P
			CYP3A4*3	M445T

		mefloquine, primaquine, halofantrine, lumefantrine, dapsone, artemisinin, artemether, arteether, artelinic acid, bedaquiline, delaminid	CYP3A4*4	I118V
			CYP3A4*5	P218R
			CYP3A4*7	G56D
			CYP3A4*8	R130Q
			CYP3A4*9	V170I
			CYP3A4*10	D174H
			CYP3A4*11	T363M
			CYP3A4*12	L373F
			CYP3A4*13	P416L
			CYP3A4*14	L15P
			CYP3A4*15	R162Q
			CYP3A4*16	T185S
			CYP3A4*17	F189S
			CYP3A4*18	L293P
			CYP3A4*19	P467S
			CYP3A4*21	Y319C
			CYP3A4*23	R162W
			CYP3A4*24	Q200H
			CYP3A4*28	L22V
			CYP3A4*29	F113I
			CYP3A4*31	H324Q
			CYP3A4*32	I335T
			CYP3A4*33	A370S
			CYP3A4*34	I427V
			CYP3A4*35	L3V
	CYP3A5	Chloroquine, halofantrine, lumefantrine, arteether, artelinic acid, quinine	CYP3A5*8	R28C
			CYP3A5*9	A337T

Table S 2: Human Cytochrome P450 enzyme families 1, 2 and 3 UniProt sequences.

CYP enzyme family	CYP enzyme name	UniProt sequence
CYP 1	CYP 1A1	MLFPISMSATEFLLASVIFCLVFWVIRASRPQVPKGLKNPPGPWGWPLIGHM LTLGKNPHLALSRMSQQYGDVLQIRIGSTPVVVLSGLDTIRQALVRQGDDFK GRPDLYTFTLISNGQSMSFSPDSGPVWAARRRLAQNGLKSFSIASDPASSTSC YLEEHVSKEAEVLISTLQELMAGPGHFNPYRYVVVSVTNVICAICFGRRYDH NHQELLSLVNLNNNFGEVVGSGNPADFIPILRYLPNPSLNAFKDLNEKFYSF MQKMVKEHYKTFEKGHIRDITDSLIEHCQEKQLDENANVQLSDEKIINIVLD LFGAGFDTVTTAISWSLMYLVMNPRVQRKIQEELDTVIGRSRRPRLSDRSHL PYMEAFILETFRHSSFVPFTIPHSTTRDTSLKGFYIPKGRCVFVNQWQINHDQ KLWVNPSEFLPERFLTPDGAIDKVLSEKVIIFGMGKRKCIGETIARWEVFLFL AILLQRVEFSVPLGVKVDMTPIYGLTMKHACCEHFQMQLRS
	CYP 1A2	MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLG HVLTLGKNPHLALSRMSQRYGDVLQIRIGSTPVLVLSRLDTIRQALVRQGDD FKGRPDLYTSTLITDGQSLTFSTDSGPVWAARRRLAQNALNTFSIASDPASSS SCYLEEHVSKEAKALISRLQELMAGPGHFDPYNQVVVSVANVIGAMCFGQH FPESSDEMLSLVKNTHEFVETASSGNPLDFFPILRYLPNPALQRFKAFNQRFL WFLQKTVQEHYQDFDKNSVRDITGALFKHSKKGPRASGNLIPQEKIVNLVN DIFGAGFDTVTTAISWSLMYLVTKPEIQRKIQKELDTVIGRERRPRLSDRPQL PYLEAFILETFRHSSFLPFTIPHSTTRDTTLNGFYIPKKCCVFVNQWQVNHDPE LWEDPSEFRPERFLTADGTAINKPLSEKMMLFGMGKRRCIGEVLAKWEIFLF LAILLQQLEFSVPPGVKVDLTPIYGLTMKHARCEHVQARLRFSIN
	CYP 1B1	MGTSLSPNDPWPLNPLSIQQTTLLLLLSVLATVHVGQRLLRQRRRQLRSAPP GPFAWPLIGNAAAVGQAAHLSFARLARRYGDVFQIRLGSCPIVVLNGERAIH QALVQQGSAFADRPAFASFRVVSGGRSMAFGHYSEHWKVQRRAAHSMMR NFFTRQPRSRQVLEGHVLSEARELVALLVRGSADGAFLDPRPLTVVAVANV MSAVCFGCRYSHDDPEFRELLSHNEEFGRTVGAGSLVDVMPWLQYFPNPVR TVFREFEQLNRNFSNFILDKFLRHCESLRPGAAPRDMMDAFILSAEKKAAGD SHGGGARLDLENVPATITDIFGASQDTLSTALQWLLLLFTRYPDVQTRVQAE LDQVVGRDRLPCMGDQPNLPYVLAFLYEAMRFSSFVPVTIPHATTANTSVL GYHIPKDTVVFVNQWSVNHDPLKWPNPENFDPARFLDKDGLINKDLTSRV MIFSVGKRRCIGEELSKMQLFLFISILAHQCDFRANPNEPAKMNFSYGLTIKP KSFKVNVTLRESMELLDSAVQNLQAKETCQ
CYP 2	CYP 2A6	MLASGMLLVALLVCLTVMVLMSVWQQRKSKGKLPPGPTPLPFIGNYLQLN TEQMYNSLMKISERYGPVFTIHLGPRRVVVLCGHDAVREALVDQAEEFSGR GEQATFDWVFKGYGVVFSNGERAKQLRRFSIATLRDFGVGKRGIEERIQEEA GFLIDALRGTGGANIDPTFFLSRTVSNVISSIVFGDRFDYKDKEFLSLLRMML GIFQFTSTSTGQLYEMFSSVMKHLPGPQQQAFQLLQGLEDFIAKKVEHNQRT LDPNSPRDFIDSFLIRMQEEEKNPNTEFYLKNLVMTTLNLFIGGTETVSTTLR YGFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDRAKMPYMEAVIHEIQRF GDVIPMSLARRVKKDTKFRDFFLPKGTEVYPMLGSVLRDPSFFSNPQDFNPQ HFLNEKGQFKKSDAFVPFSIGKRNCFGEGLARMELFLFFTTVMQNFRLKSSQ SPKDIDVSPKHVGFATIPRNYTMSFLPR
	CYP 2A7	MLASGLLLVALLACLTVMVLMSVWQQRKSRGKLPPGPTPLPFIGNYLQLNT EHICDSIMKFSECYGPVFTIHLGPRRVVVLCGHDAVREALVDQAEEFSGRGE QATFDWVFKGYGVAFSNGERAKQLLRFAIATLRDFGVGKRGIEERIQEESGF LIEAIRSTHGANIDPTFFLSRTVSNVISSIVFGDRFDYEDKEFLSLLSMMLGIFQ FTSTSTGQLYEMFSSVMKHLPGPQQQAFKLLQGLEDFIAKKVEHNQRTLDP NSPQDFIDSFLIHMQEEEKNPNTEFYLKNLMMSTLNLFIAGTETVSTTLRYGF LLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDRTKMPYMEAVIHEIQRFGDVI PMSLARRVKKDTKFRDFFLPKGTEVFPMLGSVLRDPSFFSNPQDFNPQHFLD DKGQFKKSDAFVPFSIGKRNCFGEGLARMELFLFFTTVMQNFRLKSSQSPKD IDVSPKHVVFATIPRNYTMSFLPR

CYP 2A13	MLASGLLLVTLLACLTVMVLMSVWRQRKSRGKLPPGPTPLPFIGNYLQLNT EQMYNSLMKISERYGPVFTIHLGPRRVVVLCGHDAVKEALVDQAEEFSGRG EQATFDWLFKGYGVAFSNGERAKQLRRFSIATLRGFGVGKRGIEERIQEEAG FLIDALRGTHGANIDPTFFLSRTVSNVISSIVFGDRFDYEDKEFLSLLRMMLGS FQFTATSTGQLYEMFSSVMKHLPGPQQQAFKELQGLEDFIAKKVEHNQRTL DPNSPRDFIDSFLIRMQEEEKNPNTEFYLKNLVMTTLNLFFAGTETVSTTLRY GFLLLMKHPEVEAKVHEEIDRVIGKNRQPKFEDRAKMPYTEAVIHEIQRFGD MLPMGLAHRVNKDTKFRDFFLPKGTEVFPMLGSVLRDPRFFSNPRDFNPQH FLDKKGQFKKSDAFVPFSIGKRYCFGEGLARMELFLFFTTIMQNFRFKSPQSP KDIDVSPKHVGFATIPRNYTMSFLPR
CYP 2B6	MELSVLLFLALLTGLLLLLVQRHPNTHDRLPPGPRPLPLLGNLLQMDRRGLL KSFLRFREKYGDVFTVHLGPRPVVMLCGVEAIREALVDKAEAFSGRGKIAM VDPFFRGYGVIFANGNRWKVLRRFSVTTMRDFGMGKRSVEERIQEEAQCLI EELRKSKGALMDPTFLFQSITANIICSIVFGKRFHYQDQEFLKMLNLFYQTFS LISSVFGQLFELFSGFLKYFPGAHRQVYKNLQEINAYIGHSVEKHRETLDPSA PKDLIDTYLLHMEKEKSNAHSEFSHQNLNLNTLSLFFAGTETTSTTLRYGFLL MLKYPHVAERVYREIEQVIGPHRPPELHDRAKMPYTEAVIYEIQRFSDLLPM GVPHIVTQHTSFRGYIIPKDTEVFLILSTALHDPHYFEKPDAFNPDHFLDANG ALKKTEAFIPFSLGKRICLGEGIARAELFLFFTTILQNFSMASPVAPEDIDLTPQ ECGVGKIPPTYQIRFLPR
CYP 2C8	MEPFVVLVLCLSFMLLFSLWRQSCRRRKLPPGPTPLPIIGNMLQIDVKDICKS FTNFSKVYGPVFTVYFGMNPIVVFHGYEAVKEALIDNGEEFSGRGNSPISQRI TKGLGIISSNGKRWKEIRRFSLTTLRNFGMGKRSIEDRVQEEAHCLVEELRKT KASPCDPTFILGCAPCNVICSVVFQKRFDYKDQNFLTLMKRFNENFRILNSP WIQVCNNFPLLIDCFPGTHNKVLKNVALTRSYIREKVKEHQASLDVNNPRDF IDCFLIKMEQEKDNQKSEFNIENLVGTVADLFVAGTETTSTTLRYGLLLLLK HPEVTAKVQEEIDHVIGRHRSPCMQDRSHMPYTDAVVHEIQRYSDLVPTGV PHAVTTDTKFRNYLIPKGTTIMALLTSVLHDDKEFPNPNIFDPGHFLDKNGNF KKSDYFMPFSAGKRICAGEGLARMELFLFLTTILQNFNLKSVDDLKNLNTTA VTKGIVSLPPSYQICFIPV
CYP 2C9	MDSLVVLVLCLSCLLLLSLWRQSSGRGKLPPGPTPLPVIGNILQIGIKDISKSL TNLSKVYGPVFTLYFGLKPIVVLHGYEAVKEALIDLGEEFSGRGIFPLAERAN RGFGIVFSNGKKWKEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKT KASPCDPTFILGCAPCNVICSIIFHKRFDYKDQQFLNLMEKLNENIKILSSPWI QICNNFSPIIDYFPGTHNKLLKNVAFMKSYILEKVKEHQESMDMNNPQDFID CFLMKMEKEKHNQPSEFTIESLENTAVDLFGAGTETTSTTLRYALLLLLKHP EVTAKVQEEIERVIGRNRSPCMQDRSHMPYTDAVVHEVQRYIDLLPTSLPHA VTCDIKFRNYLIPKGTTILISLTSVLHDNKEFPNPEMFDPHHFLDEGGNFKKS KYFMPFSAGKRICVGEALAGMELFLFLTSILQNFNLKSLVDPKNLDTTPVVN GFASVPPFYQLCFIPV
CYP 2C18	MDPAVALVLCLSCLFLLSLWRQSSGRGRLPSGPTPLPIIGNILQLDVKDMSKS LTNFSKVYGPVFTVYFGLKPIVVLHGYEAVKEALIDHGEEFSGRGSFPVAEK VNKGLGILFSNGKRWKEIRRFCLMTLRNFGMGKRSIEDRVQEEARCLVEEL RKTNASPCDPTFILGCAPCNVICSVIFHDRFDYKDQRFLNLMEKFNENLRILS SPWIQVCNNFPALIDYLPGSHNKIAENFAYIKSYVLERIKEHQESLDMNSARD FIDCFLIKMEQEKHNQQSEFTVESLIATVTDMFGAGTETTSTTLRYGLLLLLK YPEVTAKVQEEIECVVGRNRSPCMQDRSHMPYTDAVVHEIQRYIDLLPTNLP HAVTCDVKFKNYLIPKGTTIITSLTSVLHNDKEFPNPEMFDPGHFLDKSGNFK KSDYFMPFSAGKRMCMGEGLARMELFLFLTTILQNFNLKSQVDPKDIDITPI ANAFGRVPPLYQLCFIPV
CYP 2C19	MDPFVVLVLCLSCLLLLSIWRQSSGRGKLPPGPTPLPVIGNILQIDIKDVSKSL TNLSKIYGPVFTLYFGLERMVVLHGYEVVKEALIDLGEEFSGRGHFPLAERA NRGFGIVFSNGKRWKEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRK TKASPCDPTFILGCAPCNVICSIIFQKRFDYKDQQFLNLMEKLNENIRIVSTPW IQICNNFPTIIDYFPGTHNKLLKNLAFMESDILEKVKEHQESMDINNPRDFIDC FLIKMEKEKQNQQSEFTIENLVITAADLLGAGTETTSTTLRYALLLLLKHPEV TAKVQEEIERVVGRNRSPCMQDRGHMPYTDAVVHEVQRYIDLIPTSLPHAV TCDVKFRNYLIPKGTTILTSLTSVLHDNKEFPNPEMFDPRHFLDEGGNFKKSN

	YFMPFSAGKRICVGEGLARMELFLFLTFILQNFNLKSLIDPKDLDTTPVVNGF ASVPPFYQLCFIPV
CYP 2D6	MGLEALVPLAVIVAIFLLLVDLMHRRQRWAARYPPGPLPLPGLGNLLHVDF QNTPYCFDQLRRRFGDVFSLQLAWTPVVVLNGLAAVREALVTHGEDTADR PPVPITQILGFGPRSQGVFLARYGPAWREQRRFSVSTLRNLGLGKKSLEQWV TEEAACLCAAFANHSGRPFRPNGLLDKAVSNVIASLTCGRRFEYDDPRFLRL LDLAQEGLKEESGFLREVLNAVPVLLHIPALAGKVLRFQKAFLTQLDELLTE HRMTWDPAQPPRDLTEAFLAEMEKAKGNPESSFNDENLRIVVADLFSAGM VTTSTTLAWGLLLMILHPDVQRRVQQEIDDVIGQVRRPEMGDQAHMPYTTA VIHEVQRFGDIVPLGVTHMTSRDIEVQGFRIPKGTTLITNLSSVLKDEAVWEK PFRFHPEHFLDAQGHFVKPEAFLPFSAGRRACLGEPLARMELFLFFTSLLQHF SFSVPTGQPRPSHHGVFAFLVSPSPYELCAVPR
CYP 2E1	MSALGVTVALLVWAAFLLLVSMWRQVHSSWNLPPGPFPLPIIGNLFQLELK NIPKSFTRLAQRFGPVFTLYVGSQRMVVMHGYKAVKEALLDYKDEFSGRG DLPAFHAHRDRGIIFNNGPTWKDIRRFSLTTLRNYGMGKQGNESRIQREAHF LLEALRKTQGQPFDPTFLIGCAPCNVIADILFRKHFDYNDEKFLRLMYLFNEN FHLLSTPWLQLYNNFPSFLHYLPGSHRKVIKNVAEVKEYVSERVKEHHQSL DPNCPRDLTDCLLVEMEKEKHSAERLYTMDGITVTVADLFFAGTETTSTTLR YGLLILMKYPEIEEKLHEEIDRVIGPSRIPAIKDRQEMPYMDAVVHEIQRFITL VPSNLPHEATRDTIFRGYLIPKGTVVVPTLDSVLYDNQEFPDPEKFKPEHFLN ENGKFKYSDYFKPFSTGKRVCAGEGLARMELFLLLCAILQHFNLKPLVDPK DIDLSPIHIGFGCIPPRYKLCVIPRS
CYP 2F1	MDSISTAILLLLLALVCLLLTLSSRDKGKLPPGPRPLSILGNLLLLCSQDMLTS LTKLSKEYGSMYTVHLGPRRVVVLSGYQAVKEALVDQGEEFSGRGDYPAF FNFTKGNGIAFSSGDRWKVLRQFSIQILRNFGMGKRSIEERILEEGSFLLAELR KTEGEPFDPTFVLSRSVSNIICSVLFGSRFDYDDERLLTIIRLINDNFQIMSSPW GELYDIFPSLLDWVPGPHQRIFQNFKCLRDLIAHSVHDHQASLDPRSPRDFIQ CFLTKMAEEKEDPLSHFHMDTLLMTTHNLLFGGTKTVSTTLHHAFLALMKY PKVQARVQEEIDLVVGRARLPALKDRAAMPYTDAVIHEVQRFADIIPMNLP HRVTRDTAFRGFLIPKGTDVITLLNTVHYDPSQFLTPQEFNPEHFLDANQSFK KSPAFMPFSAGRRLCLGESLARMELFLYLTAILQSFSLQPLGAPEDIDLTPLSS GLGNLPRPFQLCLRPR
CYP 2J2	MLAAMGSLAAALWAVVHPRTLLLGTVAFLLAADFLKRRRPKNYPPGPWRL PFLGNFFLVDFEQSHLEVQLFVKKYGNLFSLELGDISAVLITGLPLIKEALIHM DQNFGNRPVTPMREHIFKKNGLIMSSGQAWKEQRRFTLTALRNFGLGKKSL EERIQEEAQHLTEAIKEENGQPFDPHFKINNAVSNIICSITFGERFEYQDSWFQ QLLKLLDEVTYLEASKTCQLYNVFPWIMKFLPGPHQTLFSNWKKLKLFVSH MIDKHRKDWNPAETRDFIDAYLKEMSKHTGNPTSSFHEENLICSTLDLFFAG TETTSTTLRWALLYMALYPEIQEKVQAEIDRVIGQGQQPSTAARESMPYTNA VIHEVQRMGNIIPLNVPREVTVDTTLAGYHLPKGTMILTNLTALHRDPTEWA TPDTFNPDHFLENGQFKKREAFMPFSIGKRACLGEQLARTELFIFFTSLMQKF TFRPPNNEKLSLKFRMGITISPVSHRLCAVPQV
CYP 2R1	MWKLWRAEEGAAALGGALFLLLFALGVRQLLKQRRPMGFPPGPPGLPFIGN IYSLAASSELPHVYMRKQSQVYGEIFSLDLGGISTVVLNGYDVVKECLVHQS EIFADRPCLPLFMKMTKMGGLLNSRYGRGWVDHRRLAVNSFRYFGYGQKS FESKILEETKFFNDAIETYKGRPFDFKQLITNAVSNITNLIIFGERFTYEDTDFQ HMIELFSENVELAASASVFLYNAFPWIGILPFGKHQQLFRNAAVVYDFLSRLI EKASVNRKPQLPQHFVDAYLDEMDQGKNDPSSTFSKENLIFSVGELIIAGTET TTNVLRWAILFMALYPNIQGQVQKEIDLIMGPNGKPSWDDKCKMPYTEAVL HEVLRFCNIVPLGIFHATSEDAVVRGYSIPKGTTVITNLYSVHFDEKYWRDPE VFHPERFLDSSGYFAKKEALVPFSLGRRHCLGEHLARMEMFLFFTALLQRFH LHFPHELVPDLKPRLGMTLQPQPYLICAERR
CYP 2S1	MEATGTWALLLALALLLLLTLALSGTRARGHLPPGPTPLPLLGNLLQLRPGA LYSGLMRLSKKYGPVFTIYLGPWRPVVVLVGQEAVREALGGQAEEFSGRGT VAMLEGTFDGHGVFFSNGERWRQLRKFTMLALRDLGMGKREGEELIQAEA RCLVETFQGTEGRPFDPSLLLAQATSNVVCSLLFGLRFSYEDKEFQAVVRAA GGTLLGVSSQGGQTYEMFSWFLRPLPGPHKQLLHHVSTLAAFTVRQVQQH QGNLDASGPARDLVDAFLLKMAQEEQNPGTEFTNKNMLMTVIYLLFAGTM

		TVSTTVGYTLLLLMKYPHVQKWVREELNRELGAGQAPSLGDRTRLPYTDA VLHEAQRLLALVPMGIPRTLMRTTRFRGYTLPQGTEVFPLLGSILHDPNIFKH PEEFNPDRFLDADGRFRKHEAFLPFSLGKRVCLGEGLAKAELFLFFTTILQAF SLESPCPPDTLSLKPTVSGLFNIPPAFQLQVRPTDLHSTTQTR
	CYP 2U1	MSSPGPSQPPAEDPPWPARLLRAPLGLLRLDPSGGALLLCGLVALLGWSWL RRRRARGIPPGPTPWPLVGNFGHVLLPPFLRRRSWLSSRTRAAGIDPSVIGPQ VLLAHLARVYGSIFSFFIGHYLVVVLSDFHSVREALVQQAEVFSDRPRVPLISI VTKEKGVVFAHYGPVWRQQRKFSHSTLRHFGLGKLSLEPKIIEEFKYVKAE MQKHGEDPFCPFSIISNAVSNIICSLCFGQRFDYTNSEFKKMLGFMSRGLEICL NSQVLLVNICPWLYYLPFGPFKELRQIEKDITSFLKKIIKDHQESLDRENPQDF IDMYLLHMEEERKNNSNSSFDEEYLFYIIGDLFIAGTDTTTNSLLWCLLYMSL NPDVQEKVHEEIERVIGANRAPSLTDKAQMPYTEATIMEVQRLTVVVPLAIP HMTSENTVLQGYTIPKGTLILPNLWSVHRDPAIWEKPEDFYPNRFLDDQGQL IKKETFIPFGIGKRVCMGEQLAKMELFLMFVSLMQSFAFALPEDSKKPLLTG RFGLTLAPHPFNITISRR
	CYP2W1	MALLLLLFLGLLGLWGLLCACAQDPSPAARWPPGPRPLPLVGNLHLLRLSQ QDRSLMELSERYGPVFTVHLGRQKTVVLTGFEAVKEALAGPGQELADRPPI AIFQLIQRGGGIFFSSGARWRAARQFTVRALHSLGVGREPVADKILQELKCLS GQLDGYRGRPFPLALLGWAPSNITFALLFGRRFDYRDPVFVSLLGLIDEVMV LLGSPGLQLFNVYPWLGALLQLHRPVLRKIEEVRAILRTLLEARRPHVCPGD PVCSYVDALIQQGQGDDPEGLFAEANAVACTLDMVMAGTETTSATLQWAA LLMGRHPDVQGRVQEELDRVLGPGRTPRLEDQQALPYTSAVLHEVQRFITL LPHVPRCTAADTQLGGFLLPKGTPVIPLLTSVLLDETQWQTPGQFNPGHFLD ANGHFVKREAFLPFSAGRRVCVGERLARTELFLLFAGLLQRYRLLPPPGVSP ASLDTTPARAFTMRPRAQALCAVPRP
CYP 3	CYP 3A4	MALIPDLAMETWLLLAVSLVLLYLYGTHSHGLFKKLGIPGPTPLPFLGNILSY HKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDMIKTVLVKECYSVFT NRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKEMVPIIAQYGDVL VRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSLNNPQDPFVENTK KLLRFDFLDPFFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESRLED TQKHRVDFLQLMIDSQNSKETESHKALSDLELVAQSIIFIFAGYETTSSVLSFI MYELATHPDVQQKLQEEIDAVLPNKAPPTYDTVLQMEYLDMVVNETLRLF PIAMRLERVCKKDVEINGMFIPKGVVVMIPSYALHRDPKYWTEPEKFLPERF SKKNKDNIDPYIYTPFGSGPRNCIGMRFALMNMKLALIRVLQNFSFKPCKET QIPLKLSLGGLLQPEKPVVLKVESRDGTVSGA
	CYP 3A5	MDLIPNLAVETWLLLAVSLVLLYLYGTRTHGLFKRLGIPGPTPLPLLGNVLS YRQGLWKFDTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVF TNRRSLGPVGFMKSAISLAEDEEWKRIRSLLSPTFTSGKLKEMFPIIAQYGDV LVRNLRREAEKGKPVTLKDIFGAYSMDVITGTSFGVNIDSLNNPQDPFVEST KKFLKFGFLDPLFLSIILFPFLTPVFEALNVSLFPKDTINFLSKSVNRMKKSRL NDKQKHRLDFLQLMIDSQNSKETESHKALSDLELAAQSIIFIFAGYETTSSVL SFTLYELATHPDVQQKLQKEIDAVLPNKAPPTYDAVVQMEYLDMVVNETL RLFPVAIRLERTCKKDVEINGVFIPKGSMVVIPTYALHHDPKYWTEPEEFRPE RFSKKKDSIDPYIYTPFGTGPRNCIGMRFALMNMKLALIRVLQNFSFKPCKET QIPLKLDTQGLLQPEKPIVLKVDSRDGTLSGE
	CYP 3A7	MDLIPNLAVETWLLLAVSLILLYLYGTRTHGLFKKLGIPGPTPLPFLGNALSF RKGYWTFDMECYKKYRKVWGIYDCQQPMLAITDPDMIKTVLVKECYSVFT NRRPFGPVGFMKNAISIAEDEEWKRIRSLLSPTFTSGKLKEMVPIIAQYGDVL VRNLRREAETGKPVTLKHVFGAYSMDVITSTSFGVSIDSLNNPQDPFVENTK KLLRFNPLDPFVLSIKVFPFLTPILEALNITVFPRKVISFLTKSVKQIKEGRLKE TQKHRVDFLQLMIDSQNSKDSETHKALSDLELMAQSIIFIFAGYETTSSVLSFI IYELATHPDVQQKVQKEIDTVLPNKAPPTYDTVLQLEYLDMVVNETLRLFP VAMRLERVCKKDVEINGMFIPKGVVVMIPSYVLHHDPKYWTEPEKFLPERF SKKNKDNIDPYIYTPFGSGPRNCIGMRFALVNMKLALVRVLQNFSFKPCKET QIPLKLRFGGLLLTEKPIVLKAESRDETVSGA
	CYP 3A43	MDLIPNFAMETWVLVATSLVLLYIYGTHSHKLFKKLGIPGPTPLPFLGTILFY LRGLWNFDRECNEKYGEMWGLYEGQQPMLVIMDPDMIKTVLVKECYSVFT NQMPLGPMGFLKSALSFAEDEEWKRIRTLLSPAFTSVKFKEMVPIISQCGDM

LVRSLRQEAENSKSINLKDFFGAYTMDVITGTLFGVNLDSLNNPQDPFLKNM
KKLLKLDFLDPFLLLISLFPFLTPVFEALNIGLFPKDVTHFLKNSIERMKESRL
KDKQKHRVDFFQQMIDSQNSKETKSHKALSDLELVAQSIIIIFAAYDTTSTTL
PFIMYELATHPDVQQKLQEEIDAVLPNKAPVTYDALVQMEYLDMVVNETL
RLFPVVSRVTRVCKKDIEINGVFIPKGLAVMVPIYALHHDPKYWTEPEKFCP
ERFSKKNKDSIDLYRYIPFGAGPRNCIGMRFALTNIKLAVIRALQNFSFKPCK
ETQIPLKLDNLPILQPEKPIVLKVHLRDGITSGP

Table S 3: Best-fit substitution models for the construction of nine phylogenetic trees selected based on the best (lowest) BIC score.

Missing data treatment	Model	BIC score
Complete deletion	LG+G+I	30978.1996
	LG+G	30980.9851
	LG+G+I+F	31059.3459
Partial deletion (95%)	LG+G+I	31884.2744
	LG+G	31887.6735
	LG+G+I+F	31978.3684
Partial deletion (90%)	LG+G+I	31959.6167
	LG+G	31963.0989
	LG+G+I+F	32052.9764

Table S 4: The significant motifs generated with MEME program, their positions in CYP enzymes within families 1, 2 and 3, and their respective E-values.

		CYP enzyme and residue positions of motifs																				
Motif	**E-value**	**1A1**	**1A2**	**1B1**	**2A6**	**2A7**	**2A13**	**2B6**	**2C8**	**2C9**	**2C18**	**2C19**	**2D6**	**2E1**	**2F1**	**2J2**	**2R1**	**2S1**	**2U1**	**2W1**	**3A4**	**3A5**	**3A7**	**3A43**
1	4.6e-236	449-468	450-469	462-481	431-450	431-450	431-450	428-447	427-446	427-446	427-446	427-446	435-454	429-448	428-447	440-459	440-459	432-451	482-501	425-444	434-453	433-452	434-453	434-453
2	7.6e-204	336-355	336-355	349-368	320-339	320-339	320-339	317-336	316-335	316-335	316-335	316-335	324-343	318-337	317-336	330-349	329-348	321-340	371-390	315-334	324-343	324-343	324-343	324-343
3	1.1e-203	361-380	361-380	374-393	345-364	345-364	345-364	342-361	341-360	341-360	341-360	341-360	349-368	343-362	342-361	355-374	354-373	346-365	396-415	340-359	349-368	349-368	349-368	349-368
4	1.1e-179	415-434	415-434	428-447	399-418	399-418	399-418	396-415	395-414	395-414	395-414	395-414	403-422	397-416	396-415	409-428	408-427	400-419	450-469	393-412	402-421	402-421	402-421	402-421
5	1.3e-175	314-333	314-333	327-346	298-317	298-317	298-317	295-314	294-313	294-313	294-313	294-313	302-321	296-315	295-314	308-327	307-326	299-318	349-368	293-312	302-321	302-321	302-321	302-321
6	3.2e-171	192-211	194-213	199-218	176-195	176-195	176-195	173-192	172-191	172-191	172-191	172-191	180-199	174-193	173-192	186-205	185-204	176-195	226-245	173-192	178-197	178-197	178-197	178-197
7	2.7e-162	388-407	388-407	401-420	372-391	372-391	372-391	369-388	368-387	368-387	368-387	368-387	376-395	370-389	369-388	383-401	381-400	373-392	423-442	366-385	375-394	375-394	375-394	375-394
8	8.9e-161	36-55	38-57	47-66	30-49	30-49	30-49	27-46	26-45	26-45	26-45	26-45	30-49	29-48	27-46	40-59	36-55	29-48	56-75	28-47	34-53	34-53	34-53	34-53
9	9.7e-151	84-103	86-105	95-114	79-98	79-98	79-98	76-95	75-94	75-94	75-94	75-94	79-98	78-97	76-95	89-108	87-106	79-98	128-147	77-96	82-101	82-101	82-101	82-101
10	2.0e-142	127-146	129-148	137-156	120-139	120-139	120-139	117-136	116-135	116-135	116-135	116-135	124-143	118-137	117-136	130-149	129-148	120-139	170-189	118-137	122-141	122-141	122-141	122-141
11	2.4e-124	64-83	66-85	75-94	59-78	59-78	59-78	56-75	55-74	55-74	55-74	55-74	59-78	58-77	56-75	69-88	67-86	58-77	108-127	57-76	62-81	62-81	62-81	62-81
12	3.5e-107	154-173	156-175	161-180	140-159	140-159	140-159	137-156	136-155	136-155	136-155	136-155	144-163	138-157	137-156	150-169	149-168	140-159	190-209	138-157	252-271	-	252-271	252-271
13	2.9e-105	-	-	288-307	263-282	263-282	263-282	260-279	259-278	259-278	259-278	259-278	267-286	261-280	260-279	273-292	272-291	264-283	313-332	-	-	-	-	-
14	4.8e-094	469-488	470-489	-	451-470	451-470	451-470	448-467	447-466	447-466	447-466	447-466	455-474	449-468	448-467	460-479	460-479	452-471	502-521	-	454-473	453-472	454-473	454-473

15	9.1e-078	-	-	-	216-235	216-235	216-235	213-232	212-231	212-231	212-231	212-231	-	214-233	213-232	226-245	-	216-235	-	-	-	-	-	-
16	2.2e-055	105-124	107-126	-	100-119	100-119	100-119	97-116	96-115	96-115	96-115	96-115	-	-	97-116	110-129	108-127	100-119	149-168	98-117	-	-	-	-
17	4.7e-046	-	-	-	9-28	9-28	9-28	6-25	5-24	5-24	5-24	5-24	-	8-27	-	-	-	7-26	35-54	-	11-30	11-30	11-30	-
18	1.8e-045	-	-	-	474-493	474-493	474-493	471-490	470-489	470-489	470-489	470-489	-	472-491	471-490	-	-	475-494	-	-	-	-	-	-
19	4.2e-042	260-274	262-276	-	246-260	246-260	246-260	243-257	242-256	242-256	242-256	242-256	-	244-258	243-257	-	-	246-260	296-310	-	-	-	-	-
20	6.2e-042	-	-	-	160-174	160-174	160-174	157-171	156-170	156-170	156-170	156-170	-	158-172	157-171	-	-	-	-	-	-	-	-	-
21	3.4e-038	-	-	-	420-430	420-430	420-430	417-427	416-426	416-426	416-426	416-426	424-434	418-428	417-427	429-439	429-439	421-431	471-481	414-424	-	-	-	-
22	2.3e-028	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	273-292	273-292	273-292	273-292
23	1.0e-023	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	145-164	145-164	145-164	145-164
24	1.5e-020	-	-	-	196-215	196-215	196-215	193-212	192-211	192-211	192-211	192-211	-	194-213	193-212	-	205-224	-	-	-	-	-	-	-
25	7.3e-019	-	-	-	283-297	283-297	283-297	280-294	279-293	279-293	279-293	279-293	287-301	-	-	293-307	292-306	284-298	334-348	-	-	-	-	-
26	9.8e-016	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	102-121	102-121	102-121	102-121
27	1.7e-013	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	198-217	198-217	198-217	198-217
28	3.9e-011	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	483-502	482-501	483-502	483-502
29	3.9e-010	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	423-433	422-432	423-433	423-433
30	7.3e-007	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	223-242	223-242	223-242	223-242
31	1.5e-005	489-507	490-508	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table S 5: Conformational clustering information for references and function assigned alleles for the last 300 ns of each trajectory.

Systems	Number of clusters	Number of structures in cluster	Average RMSD (nm)
REF 1	1	3001	0.122
REF 2	1	3001	0.166
***2**	1	3001	0.117
***27**	1	3001	0.115
***33**	1	3001	0.120
***34**	1	3001	0.142
***39**	1	3001	0.117
***45**	1	3001	0.110
***48**	1	3001	0.120
***53**	1	3001	0.115
***10**	1	3001	0.102
***14**	1	3001	0.135
***17**	1	3001	0.112
***29**	1	3001	0.136
***49**	1	3001	0.123
***50**	1	3001	0.119
***54**	1	3001	0.135
***55**	1	3001	0.125

Green signifies normal activity alleles and blue signifies decreased activity alleles.

Table S 6: Frequency in snapshots of tunnel presence in CYP2D6 references and function assigned alleles.

Systems	Channel 2b [a]	Channel 2f [a]	Channel 2e [a]	Channel 2c [a]	Solvent (S) [a]
REF1	1000	703	550	391	130
REF2	983	735	415	247	239
***2**	1000	456	390	296	404
***27**	1001	459	542	751	117
***33**	998	677	415	560	322
***34**	901	545	500	344	487
***39**	994	855	470	645	856
***45**	1001	441	350	299	549
***48**	421	811	466	520	281
***53**	1001	756	388	437	198
***10**	390	585	452	586	302
***14**	1001	935	325	448	636
***17**	573	769	654	321	443
***29**	1000	744	377	784	330
***49**	998	652	405	448	207
***50**	996	976	465	602	765
***54**	966	768	708	258	282
***55**	570	899	733	569	318

[a] Number of snapshots a tunnel was present out of the 1001 snapshots used. A tunnel was present in a snapshot if its bottleneck radii was at least 0.9 Å.

Green signifies normal activity alleles and blue signifies decreased activity alleles.

Table S 7: Conformational clustering attributes of references and Afrocentric alleles for the last 300 ns of each trajectory.

Systems	**Number of clusters**	**Number of structures in cluster**	**Average RMSD (nm)**
REF1	1	3001	0.122
REF2	1	3001	0.166
*149	1	3001	0.112
*152	1	3001	0.131
*153	1	3001	0.125
*154	1	3001	0.116
*155	1	3001	0.126
*157	1	3001	0.119
*159	1	3001	0.133
*160	1	3001	0.145
*162	1	3001	0.125
*163	1	3001	0.126

Table S 8: Snapshot frequencies of tunnel presence in CYP2D6 references and Afrocentric alleles.

Systems	Channel 2b [a]	Channel 2f [a]	Channel 2e [a]	Channel 2c [a]	Solvent (S) [a]
REF1	1000	703	550	391	130
REF2	983	735	415	247	239
***149**	985	399	572	406	543
***152**	936	617	352	895	127
***153**	895	875	387	722	279
***154**	991	743	299	514	87
***155**	1000	931	864	384	809
***157**	978	669	687	299	376
***159**	977	531	667	168	544
***160**	989	796	587	302	779
***162**	998	688	612	283	689
***163**	935	712	473	335	198

[a] Number of snapshots a tunnel was present out of the 1001 snapshots used. A tunnel was present in a snapshot if its bottleneck radii was at least 0.9 Å.

References

1. Langreth R, Waldholz M (1999) New Era of Personalized Medicine: Targeting Drugs For Each Unique Genetic Profile. The Oncologist 4:426–427. https://doi.org/10.1634/theoncologist.4-5-426

2. Mouton JP (2022) The burden of serious adverse drug reactions in South Africa

3. Chamboko CR, Veldman W, Tata RB, et al (2023) Human Cytochrome P450 1, 2, 3 Families as Pharmacogenes with Emphases on Their Antimalarial and Antituberculosis Drugs and Prevalent African Alleles. International Journal of Molecular Sciences 24:3383. https://doi.org/10.3390/ijms24043383

4. Ginsburg GS, Willard HF (2009) Genomic and personalized medicine: foundations and applications. Translational Research 154:277–287. https://doi.org/10.1016/j.trsl.2009.09.005

5. Goetz LH, Schork NJ (2018) Personalized Medicine: Motivation, Challenges and Progress. Fertil Steril 109:952–963. https://doi.org/10.1016/j.fertnstert.2018.05.006

6. Ahmed S, Zhou Z, Zhou J, Chen S-Q (2016) Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. Genomics Proteomics Bioinformatics 14:298–313. https://doi.org/10.1016/j.gpb.2016.03.008

7. Zanger UM, Schwab M (2013) Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacology & Therapeutics 138:103–141. https://doi.org/10.1016/J.PHARMTHERA.2012.12.007

8. Liu X, Ren S, Zhang J, et al (2021) The association between cytochrome P450 polymorphisms and anti-tuberculosis drug-induced liver injury: a systematic review and meta-analysis. Annals of Palliative Medicine 10:6518534–6516534. https://doi.org/10.21037/APM-21-1224

9. T P A, M SS, Jose A, et al (2009) Pharmacogenomics: The Right Drug to the Right Person. J Clin Med Res 1:191–194. https://doi.org/10.4021/jocmr2009.08.1255

10. Katara P, Yadav A (2019) Pharmacogenes (PGx-genes): Current understanding and future directions. Gene 718:144050. https://doi.org/10.1016/j.gene.2019.144050

11. Hassan R, Allali I, Agamah FE, et al (2021) Drug response in association with pharmacogenomics and pharmacomicrobiomics: towards a better personalized medicine. Briefings in bioinformatics 22:. https://doi.org/10.1093/bib/bbaa292

12. Johnson JA (2008) Ethnic differences in cardiovascular drug response: potential contribution of pharmacogenetics. Circulation 118:1383–1393. https://doi.org/10.1161/CIRCULATIONAHA.107.704023

13. Gill P, Yu F, Porter-Gill P, et al (2021) Implementing Pharmacogenomics Testing: Single Center Experience at Arkansas Children's Hospital. JPM 11:394. https://doi.org/10.3390/jpm11050394

14. Ingelman-Sundberg M, Rodriguez-Antona C (2005) Pharmacogenetics of drug-metabolizing enzymes: implications for a safer and more effective drug therapy. Philosophical Transactions of the Royal Society B: Biological Sciences 360:1563–1570. https://doi.org/10.1098/rstb.2005.1685

15. Zöllner A, Buchheit D, Meyer MR, et al (2010) Production of human phase 1 and 2 metabolites by whole-cell biotransformation with recombinant microbes. Bioanalysis 2:1277–1290. https://doi.org/10.4155/bio.10.80

16. Lennard L (2014) Implementation of TPMT testing. Br J Clin Pharmacol 77:704–714. https://doi.org/10.1111/bcp.12226

17. Niemi M, Pasanen MK, Neuvonen PJ (2011) Organic Anion Transporting Polypeptide 1B1: a Genetically Polymorphic Transporter of Major Importance for Hepatic Drug Uptake. Pharmacol Rev 63:157–181. https://doi.org/10.1124/pr.110.002857

18. Kalliokoski A, Neuvonen PJ, Niemi M (2010) SLCO1B1 Polymorphism and Oral Antidiabetic Drugs. Basic & Clinical Pharmacology & Toxicology 107:775–781. https://doi.org/10.1111/j.1742-7843.2010.00581.x

19. Bruckmueller H, Cascorbi I (2021) ABCB1, ABCG2, ABCC1, ABCC2, and ABCC3 drug transporter polymorphisms and their impact on drug bioavailability: what is our current understanding? Expert Opin Drug Metab Toxicol 17:369–396. https://doi.org/10.1080/17425255.2021.1876661

20. Mukonzo JK, Waako P, Ogwal-Okeng J, et al (2010) Genetic variations in ABCB1 and CYP3A5 as well as sex influence quinine disposition among Ugandans. Ther Drug Monit 32:346–352. https://doi.org/10.1097/FTD.0b013e3181da79d6

21. Martinez M, Modric S, Sharkey M, et al (2008) The pharmacogenomics of P-glycoprotein and its role in veterinary medicine. J Vet Pharmacol Ther 31:285–300. https://doi.org/10.1111/j.1365-2885.2008.00964.x

22. Johnson JA (2001) Drug target pharmacogenomics: an overview. Am J Pharmacogenomics 1:271–281. https://doi.org/10.2165/00129785-200101040-00004

23. Rusnati M, D'Ursi P, Pedemonte N, et al (2020) Recent Strategic Advances in CFTR Drug Discovery: An Overview. Int J Mol Sci 21:2407. https://doi.org/10.3390/ijms21072407

24. Bignucolo A, Mattia ED, Cecchin E, et al (2017) Pharmacogenomics of Targeted Agents for Personalization of Colorectal Cancer Treatment. International Journal of Molecular Sciences 18:. https://doi.org/10.3390/ijms18071522

25. Dhananjay P, Ramya Krishna V, Aswani Dutt V, Mitra AK (2013) 2 - Biology of ocular transporters: efflux and influx transporters in the eye. In: Mitra AK (ed) Ocular Transporters and Receptors. Woodhead Publishing, pp 37–84

26. Diallinas G (2014) Understanding transporter specificity and the discrete appearance of channel-like gating domains in transporters. Front Pharmacol 5:207. https://doi.org/10.3389/fphar.2014.00207

27. Surendiran A, Pradhan SC, Adithan C (2008) Role of pharmacogenomics in drug discovery and development. Indian J Pharmacol 40:137–143. https://doi.org/10.4103/0253-7613.43158

28. Topić E, Štefanović M, Primorac D, et al (2023) Pharmacogenomics of Drug-Metabolizing Enzymes. In: Primorac D, Höppner W, Bach-Rojecky L (eds) Pharmacogenomics in Clinical Practice. Springer International Publishing, Cham, pp 35–60

29. Iyanagi T (2007) Molecular mechanism of phase I and phase II drug-metabolizing enzymes: implications for detoxification. International review of cytology 260:35–112. https://doi.org/10.1016/S0074-7696(06)60002-8

30. Mohsin N ul A, Farrukh M, Shahzadi S, et al (2024) Drug Metabolism: Phase I and Phase II Metabolic Pathways. In: Drug Metabolism and Pharmacokinetics. IntechOpen

31. Nelson DR, Nebert DW (2018) Cytochrome P450 ( ) Gene Superfamily. In: eLS. John Wiley & Sons, Ltd, pp 1–19

32. Sevrioukova IF, Poulos TL (2012) UNDERSTANDING THE MECHANISM OF CYTOCHROME P450 3A4: RECENT ADVANCES AND REMAINING PROBLEMS. Dalton transactions (Cambridge, England : 2003) 42:3116. https://doi.org/10.1039/c2dt31833d

33. Omura T (2005) Heme-thiolate proteins. Biochem Biophys Res Commun 338:404–409. https://doi.org/10.1016/j.bbrc.2005.08.267

34. Martínez FMC, López DP, Pavía PDS, et al (2020) 6. Cytochrome P450. The Dioxygen-Activating Heme Thiolate. In: Sosa Torres M, Kroneck P (eds) Transition Metals and Sulfur – A Strong Relationship for Life. De Gruyter, pp 165–198

35. Esteves F, Rueff J, Kranendonk M (2021) The Central Role of Cytochrome P450 in Xenobiotic Metabolism—A Brief Review on a Fascinating Enzyme Family. Journal of Xenobiotics 11:94–114. https://doi.org/10.3390/jox11030007

36. Monostory K, Dvorak Z (2011) Steroid regulation of drug-metabolizing cytochromes P450. Curr Drug Metab 12:154–172. https://doi.org/10.2174/138920011795016854

37. Nebert DW, Wikvall K, Miller WL (2013) Human cytochromes P450 in health and disease. Philosophical Transactions of the Royal Society B: Biological Sciences 368:20120431. https://doi.org/10.1098/rstb.2012.0431

38. McKinnon RA, Sorich MJ, Ward MB (2008) Cytochrome P450 Part 1: Multiplicity and Function. Journal of Pharmacy Practice and Research 38:55–57. https://doi.org/10.1002/j.2055-2335.2008.tb00798.x

39. Nebert ' Milton Adesnik Daniel W, Coon MJ, Estabrook RW, et al (1987) The P450 Gene Superfamily: Recommended Nomenclature. Mary Ann Liebert, Inc., Publishers Pp

40. Nelson DR, Kamataki T, Waxman DJ, et al (1993) The P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. Mary Ann Liebert, Inc., Publishers

41. Nebert DW, Nelson DR, Adesnik M, et al (1989) The P450 superfamily: updated listing of all genes and recommended nomenclature for the chromosomal loci. DNA 8:1–13. https://doi.org/10.1089/dna.1.1989.8.1

42. Nebert DW, NELSONt MINOR J COONÎ RONALD W ESTABROOK DR, Feyereisen R, et al (1991) The P450 Superfamily: Update on New Sequences, Gene Mapping, and Recommended Nomenclature. Mary Ann Liebert, Inc., Publishers Pp

43. Nelson DR, Zeldin DC, Hoffman SMG, et al (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14:1–18. https://doi.org/10.1097/00008571-200401000-00001

44. Hansen CC, Nelson DR, Møller BL, Werck-Reichhart D (2021) Plant cytochrome P450 plasticity and evolution. Mol Plant 14:1244–1265. https://doi.org/10.1016/j.molp.2021.06.028

45. Cupp MJ, Tracy TS (1998) Cytochrome P450: New Nomenclature and Clinical Implications. afp 57:107–116

46. Sim SC, Ingelman-Sundberg M (2010) The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. Human genomics 4:278–281. https://doi.org/10.1186/1479-7364-4-4-278

47. Daly AK, Brockmöller J, Broly F, et al (1996) Nomenclature for human CYP2D6 alleles. Pharmacogenetics 6:193–200. https://doi.org/10.1097/00008571-199606000-00001

48. Gaedigk A, Ingelman-Sundberg M, Miller NA, et al (2018) The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele

Nomenclature Database. Clinical Pharmacology and Therapeutics 103:399–401. https://doi.org/10.1002/cpt.910

49. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, et al (2018) The Evolution of PharmVar. Clinical Pharmacology and Therapeutics 105:29. https://doi.org/10.1002/cpt.1275

50. Zhao M, Ma J, Li M, et al (2021) Cytochrome P450 Enzymes and Drug Metabolism in Humans. Int J Mol Sci 22:12808–12808. https://doi.org/10.3390/ijms222312808

51. Nguyen KT, Nguyen NL, Tung NV, et al (2020) A novel thermostable cytochrome p450 from sequence-based metagenomics of binh chau hot spring as a promising catalyst for testosterone conversion. Catalysts 10:1–17. https://doi.org/10.3390/catal10091083

52. Zhou S-F (2009) Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. Clin Pharmacokinet 48:689–723. https://doi.org/10.2165/11318030-000000000-00000

53. Yamazaki S, Sato K, Suhara K, et al (1993) Importance of the Proline-Rich Region Following Signal-Anchor Sequence in the Formation of Correct Conformation of Microsomal Cytochrome P-450s1

54. Tsuzuki D, Hichiya H, Okuda Y, et al (2003) Alteration in Catalytic Properties of Human CYP2D6 Caused by Substitution of Glycine-42 with Arginine, Lysine and Glutamic Acid. Drug Metabolism and Pharmacokinetics 18:79–85. https://doi.org/10.2133/dmpk.18.79

55. Kusano K, Sakaguchi M, Kagawa N, et al (2001) Microsomal p450s use specific proline-rich sequences for efficient folding, but not for maintenance of the folded structure. J Biochem 129:259–269. https://doi.org/10.1093/oxfordjournals.jbchem.a002853

56. P.B. Danielson BSP (2002) The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Current drug metabolism 3:561–597. https://doi.org/10.2174/1389200023337054

57. Sarparast M, Dattmore D, Alan J, Lee KSS (2020) Cytochrome P450 Metabolism of Polyunsaturated Fatty Acids and Neurodegeneration. Nutrients 2020, Vol 12, Page 3523 12:3523–3523. https://doi.org/10.3390/NU12113523

58. Midlik A, Navrátilová V, Moturu TR, et al (2021) Uncovering of cytochrome P450 anatomy by SecStrAnnotator. Scientific Reports 11:. https://doi.org/10.1038/s41598-021-91494-8

59. Meunier B, de Visser SP, Shaik S (2004) Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. Chemical Reviews 104:3947–3980. https://doi.org/10.1021/CR020443G

60. Urban P, Lautier T, Pompon D, Truan G (2018) Ligand Access Channels in Cytochrome P450 Enzymes: A Review. International Journal of Molecular Sciences 19:1617. https://doi.org/10.3390/ijms19061617

61. Cojocaru V, Winn PJ, Wade RC (2007) The ins and outs of cytochrome P450s. Biochim Biophys Acta 1770:390–401. https://doi.org/10.1016/j.bbagen.2006.07.005

62. Fishelovitch D, Shaik S, Wolfson HJ, Nussinov R (2009) Theoretical Characterization of Substrate Access/Exit Channels in the Human Cytochrome P450 3A4 Enzyme: Involvement of Phenylalanine Residues in the Gating Mechanism. J Phys Chem B 113:13018–13025. https://doi.org/10.1021/jp810386z

63. Reynald RL, Sansen S, Stout CD, Johnson EF (2012) Structural Characterization of Human Cytochrome P450 2C19: ACTIVE SITE DIFFERENCES BETWEEN P450s 2C8, 2C9, AND 2C19*♦. Journal of Biological Chemistry 287:44581–44591. https://doi.org/10.1074/jbc.M112.424895

64. Dubey KD, Shaik S (2019) Cytochrome P450—The Wonderful Nanomachine Revealed through Dynamic Simulations of the Catalytic Cycle. Acc Chem Res 52:389–399. https://doi.org/10.1021/acs.accounts.8b00467

65. Oprea TI, Hummer G, García AE (1997) Identification of a functional water channel in cytochrome P450 enzymes. Proceedings of the National Academy of Sciences of the United States of America 94:2133. https://doi.org/10.1073/pnas.94.6.2133

66. Benkaidali L, André F, Moroy G, et al (2019) Four Major Channels Detected in the Cytochrome P450 3A4: A Step toward Understanding Its Multispecificity. International Journal of Molecular Sciences 20:987. https://doi.org/10.3390/ijms20040987

67. Poulos TL, Johnson EF (2015) Structures of Cytochrome P450 Enzymes. Cytochrome P450: Structure, Mechanism, and Biochemistry, Fourth Edition 3–32. https://doi.org/10.1007/978-3-319-12108-6_1

68. Sansen S, Yano JK, Reynald RL, et al (2007) Adaptations for the Oxidation of Polycyclic Aromatic Hydrocarbons Exhibited by the Structure of Human P450 1A2*♦. Journal of Biological Chemistry 282:14348–14355. https://doi.org/10.1074/jbc.M611692200

69. Huttunen KM, Mahonen N, Raunio H, Rautio J Cytochrome P450-Activated Prodrugs: Targeted Drug Delivery. Current Medicinal Chemistry 15:2346–2365

70. Yano JK, Hsu M-H, Griffin KJ, et al (2005) Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen. Nat Struct Mol Biol 12:822–823. https://doi.org/10.1038/nsmb971

71. Kumar N, Chugh H, Sood D, et al (2018) Biology of Heme: Drug Interactions and Adverse Drug Reactions with CYP450. Current Topics in Medicinal Chemistry 18:2042–2055. https://doi.org/10.2174/1568026619666181129124638

72. Denisov IG, Makris TM, Sligar SG, Schlichting I (2005) Structure and Chemistry of Cytochrome P450. Chem Rev 105:2253–2278. https://doi.org/10.1021/cr0307143

73. Hu Q, Hartmann RW (2014) The Renaissance of CYP17 Inhibitors for the Treatment of Prostate Cancer. In: Cancer Drug Design and Discovery. Elsevier, pp 319–356

74. Wu J, Guan X, Dai Z, et al (2021) Molecular probes for human cytochrome P450 enzymes: Recent progress and future perspectives. Coordination Chemistry Reviews 427:213600. https://doi.org/10.1016/j.ccr.2020.213600

75. Wang B, Li C, Dubey KD, Shaik S (2015) Quantum Mechanical/Molecular Mechanical Calculated Reactivity Networks Reveal How Cytochrome P450cam and Its T252A Mutant Select Their Oxidation Pathways. J Am Chem Soc 137:7379–7390. https://doi.org/10.1021/jacs.5b02800

76. Ueyama N, Terakawa T, Nakata M, Nakamura A (1983) Positive shift of redox potential of [Fe4S4(Z-cys-Gly-Ala-OMe)4]2- in dichloromethane. J Am Chem Soc 105:7098–7102. https://doi.org/10.1021/ja00362a013

77. Ueyama N, Nishikawa N, Yamada Y, et al (1996) Cytochrome P-450 Model (Porphinato)(thiolato)iron(III) Complexes with Single and Double NH···S Hydrogen Bonds at the Thiolate Site. https://doi.org/10.1021/ja9622970.s001

78. Lee D-S, Yamada A, Sugimoto H, et al (2003) Substrate recognition and molecular mechanism of fatty acid hydroxylation by cytochrome P450 from Bacillus subtilis. Crystallographic, spectroscopic, and mutational studies. J Biol Chem 278:9761–9767. https://doi.org/10.1074/jbc.M211575200

79. Matsunaga I, Sumimoto T, Ueda A, et al (2000) Fatty acid-specific, regiospecific, and stereospecific hydroxylation by cytochrome P450 (CYP152B1) from Sphingomonas paucimobilis: substrate structure required for alpha-hydroxylation. Lipids 35:365–371. https://doi.org/10.1007/s11745-000-533-y

80. Rendic S, Guengerich FP (2020) Metabolism and Interactions of Chloroquine and Hydroxychloroquine with Human Cytochrome P450 Enzymes and Drug Transporters. Current drug metabolism 21:1127–1127. https://doi.org/10.2174/1389200221999201208211537

81. Urlacher VB, Girhard M (2012) Cytochrome P450 monooxygenases: an update on perspectives for synthetic application. Trends in Biotechnology 30:26–36. https://doi.org/10.1016/J.TIBTECH.2011.06.012

82. Hayaishi O (2005) An odyssey with oxygen. Biochem Biophys Res Commun 338:2–6. https://doi.org/10.1016/j.bbrc.2005.09.019

83. Waterman MR (2005) Professor Howard Mason and oxygen activation. Biochem Biophys Res Commun 338:7–11. https://doi.org/10.1016/j.bbrc.2005.08.120

84. Vasav AP, Barvkar VT (2019) Phylogenomic analysis of cytochrome P450 multigene family and their differential expression analysis in Solanum lycopersicum L. suggested tissue specific promoters. BMC Genomics 20:. https://doi.org/10.1186/s12864-019-5483-x

85. Gilani B, Cassagnol M (2024) Biochemistry, Cytochrome P450. In: StatPearls. StatPearls Publishing, Treasure Island (FL)

86. Furge LL, Guengerich FP (2006) Cytochrome P450 enzymes in drug metabolism and chemical toxicology: An introduction. Biochemistry and Molecular Biology Education 34:66–74. https://doi.org/10.1002/bmb.2006.49403402066

87. Shaik S, Cohen S, Wang Y, et al (2010) P450 Enzymes: Their Structure, Reactivity, and Selectivity—Modeled by QM/MM Calculations. Chem Rev 110:949–1017. https://doi.org/10.1021/cr900121s

88. Shaik S, Kumar D, de Visser SP, et al (2005) Theoretical Perspective on the Structure and Mechanism of Cytochrome P450 Enzymes. Chem Rev 105:2279–2328. https://doi.org/10.1021/cr030722j

89. Guengerich FP (2018) Mechanisms of Cytochrome P450-Catalyzed Oxidations HHS Public Access. ACS Catal 8:10964–10976. https://doi.org/10.1021/acscatal.8b03401

90. Henry HL (1996) Cytochrome P450 Structure, Mechanism, and Biochemistry, 2nd ed Edited by Paul R. Oritz de Montellano (University of California, San Francisco). Plenum: New York. 1995. xi + 631 pp. $125.00. ISBN 0-306-45141-7. J Am Chem Soc 118:10945–10945. https://doi.org/10.1021/ja965506p

91. Sono M, Roach MP, Coulter ED, Dawson JH (1996) Heme-Containing Oxygenases. Chem Rev 96:2841–2888. https://doi.org/10.1021/cr9500500

92. Poulos TL (2014) Heme Enzyme Structure and Function. Chem Rev 114:3919–3962. https://doi.org/10.1021/cr400415k

93. Hui Bon Hoa G, McLean MA, Sligar SG (2002) High pressure, a tool for exploring heme protein active sites. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1595:297–308. https://doi.org/10.1016/S0167-4838(01)00352-1

94. Mowat CG, Gazur B, Campbell LP, Chapman SK (2010) Flavin-containing heme enzymes. Arch Biochem Biophys 493:37–52. https://doi.org/10.1016/j.abb.2009.10.005

95. Guengerich FP (2007) Mechanisms of cytochrome P450 substrate oxidation: MiniReview. Journal of biochemical and molecular toxicology 21:163–168. https://doi.org/10.1002/JBT.20174

96. Sligar SG, Gunsalus IC (1976) A thermodynamic model of regulation: modulation of redox equilibria in camphor monoxygenase. Proc Natl Acad Sci U S A 73:1078–1082. https://doi.org/10.1073/pnas.73.4.1078

97. Govindaraj S, Poulos TL (1997) The Domain Architecture of Cytochrome P450BM-3*. Journal of Biological Chemistry 272:7915–7921. https://doi.org/10.1074/jbc.272.12.7915

98. Baj-Rossi C, Micheli GD, Carrara S (2011) P450-Based Nano-Bio-Sensors for Personalized Medicine. Biosensors - Emerging Materials and Applications. https://doi.org/10.5772/16328

99. Guengerich FP (2008) Cytochrome p450 and chemical toxicology. Chem Res Toxicol 21:70–83. https://doi.org/10.1021/tx700079z

100. Berman HM, Westbrook J, Feng Z, et al (2000) The Protein Data Bank

101. Yu X, Cojocaru V, Wade RC (2013) Conformational diversity and ligand tunnels of mammalian cytochrome P450s. Biotechnol Appl Biochem 60:134–145. https://doi.org/10.1002/bab.1074

102. Wang A, Stout CD, Zhang Q, Johnson EF (2015) Contributions of Ionic Interactions and Protein Dynamics to Cytochrome P450 2D6 (CYP2D6) Substrate and Inhibitor Binding*. Journal of Biological Chemistry 290:5092–5104. https://doi.org/10.1074/jbc.M114.627661

103. Walsh AA, Szklarz GD, Scott EE (2013) Human Cytochrome P450 1A1 Structure and Utility in Understanding Drug and Xenobiotic Metabolism. J Biol Chem 288:12932–12943. https://doi.org/10.1074/jbc.M113.452953

104. Wang A, Savas U, Stout CD, Johnson EF (2011) Structural characterization of the complex between alpha-naphthoflavone and human cytochrome P450 1B1. J Biol Chem 286:5736–5743. https://doi.org/10.1074/jbc.M110.204420

105. Wilke RA, Lin DW, Roden DM, et al (2007) Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. Nature reviews Drug discovery 6:904–904. https://doi.org/10.1038/NRD2423

106. Weinshilboum R (2003) Inheritance and drug response. The New England journal of medicine 348:529–537. https://doi.org/10.1056/NEJMRA020021

107. Preissner SC, Hoffmann MF, Preissner R, et al (2013) Polymorphic Cytochrome P450 Enzymes (CYPs) and Their Role in Personalized Therapy. PLoS ONE 8:82562–82562. https://doi.org/10.1371/JOURNAL.PONE.0082562

108. Bertilsson L, Dahl M-L, Dalén P, Al-Shurbaji A (2002) Molecular genetics of CYP2D6: Clinical relevance with focus on psychotropic drugs. Br J Clin Pharmacol 53:111–122. https://doi.org/10.1046/j.0306-5251.2001.01548.x

109. Jarvis JP, Peter AP, Shaman JA (2019) Consequences of CYP2D6 Copy-Number Variation for Pharmacogenomics in Psychiatry. Frontiers in Psychiatry 10:

110. Yusof W, Hua GS (2012) Gene, ethnic and gender influences predisposition of adverse drug reactions to artesunate among Malaysians. http://dx.doi.org/103109/153765162011623331 22:184–192. https://doi.org/10.3109/15376516.2011.623331

111. Elewa H, Wilby KJ (2017) A Review of Pharmacogenetics of Antimalarials and Associated Clinical Implications. European Journal of Drug Metabolism and Pharmacokinetics 42:745–756. https://doi.org/10.1007/S13318-016-0399-1/TABLES/4

112. Adehin A, Igbinoba SI, Soyinka JO, et al (2019) Pharmacokinetic Parameters of Quinine in Healthy Subjects and in Patients with Uncomplicated Malaria in Nigeria: Analysis of Data using a Population Approach. Current Therapeutic Research 91:33–38. https://doi.org/10.1016/j.curtheres.2019.100567

113. Pernaute-Lau L, Camara M, Nóbrega de Sousa T, et al (2022) An update on pharmacogenetic factors influencing the metabolism and toxicity of artemisinin-based combination therapy in the treatment of malaria. https://doi.org/101080/1742525520222049235 18:39–59. https://doi.org/10.1080/17425255.2022.2049235

114. Kerb R, Fux R, Mörike K, et al (2009) Pharmacogenetics of antimalarial drugs: effect on metabolism and transport. Lancet Infect Dis 9:760–774. https://doi.org/10.1016/S1473-3099(09)70320-2

115. Potter BMJ, Xie LH, Vuong C, et al (2015) Differential CYP 2D6 Metabolism Alters Primaquine Pharmacokinetics. Antimicrob Agents Chemother 59:2380–2387. https://doi.org/10.1128/AAC.00015-15

116. Mirghani RA, Sayi J, Aklillu E, et al (2006) CYP3A5 genotype has significant effect on quinine 3-hydroxylation in Tanzanians, who have lower total CYP3A activity than a Swedish population. Pharmacogenet Genomics 16:637–645. https://doi.org/10.1097/01.fpc.0000230411.89973.1b

117. McLeay SC, Vis P, van Heeswijk RPG, Green B (2014) Population pharmacokinetics of bedaquiline (TMC207), a novel antituberculosis drug. Antimicrob Agents Chemother 58:5315–5324. https://doi.org/10.1128/AAC.01418-13

118. Haas DW, Abdelwahab MT, van Beek SW, et al (2022) Pharmacogenetics of Between-Individual Variability in Plasma Clearance of Bedaquiline and Clofazimine in South Africa. J Infect Dis 226:147–156. https://doi.org/10.1093/infdis/jiac024

119. Hustert E, Haberl M, Burk O, et al (2001) The genetic determinants of the CYP3A5 polymorphism. Pharmacogenetics 11:773–779. https://doi.org/10.1097/00008571-200112000-00005

120. Lamba J, Hebert JM, Schuetz EG, et al (2012) PharmGKB summary: very important pharmacogene information for CYP3A5. Pharmacogenet Genomics 22:555–558. https://doi.org/10.1097/FPC.0b013e328351d47f

121. Landrum MJ, Lee JM, Benson M, et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46:D1062–D1067. https://doi.org/10.1093/nar/gkx1153

122. Whirl-Carrillo M, Huddart R, Gong L, et al (2021) An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. Clin Pharmacol Ther 110:563–572. https://doi.org/10.1002/cpt.2350

123. Whirl-Carrillo M, McDonagh EM, Hebert JM, et al (2012) Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 92:414–417. https://doi.org/10.1038/clpt.2012.96

124. Kane M (2021) CYP2D6 Overview: Allele and Phenotype Frequencies. In: Medical Genetics Summaries [Internet]. National Center for Biotechnology Information (US)

125. Shannon MW (2007) Chapter 5 - Drug Interactions. In: Shannon MW, Borron SW, Burns MJ (eds) Haddad and Winchester's Clinical Management of Poisoning and Drug Overdose (Fourth Edition). W.B. Saunders, Philadelphia, pp 97–104

126. Kamenski G, Ayazseven S, Berndt A, et al (2020) Clinical Relevance of CYP2D6 Polymorphisms in Patients of an Austrian Medical Practice: A Family Practice-Based Observational Study. Drugs - Real World Outcomes 7:63–73. https://doi.org/10.1007/s40801-019-00177-4

127. Gaedigk A (2013) Complexities of CYP2D6 gene analysis and interpretation. International Review of Psychiatry 25:534–553. https://doi.org/10.3109/09540261.2013.825581

128. Bijl MJ, Visser LE, Hofman A, et al (2008) Influence of the CYP2D6*4 polymorphism on dose, switching and discontinuation of antidepressants. Br J Clin Pharmacol 65:558–564. https://doi.org/10.1111/j.1365-2125.2007.03052.x

129. Mehlotra RK, Gaedigk A, Howes RE, et al (2021) CYP2D6 Genetic Variation and Its Implication for Vivax Malaria Treatment in Madagascar. Front Pharmacol 12:654054. https://doi.org/10.3389/fphar.2021.654054

130. Wannasuphoprasit Y, Andersen SE, Arranz MJ, et al (2022) CYP2D6 Genetic Variation and Antipsychotic-Induced Weight Gain: A Systematic Review and Meta-Analysis. Front Psychol 12:768748. https://doi.org/10.3389/fpsyg.2021.768748

131. Deodhar M, Turgeon J, Michaud V (2021) Contribution of CYP2D6 Functional Activity to Oxycodone Efficacy in Pain Management: Genetic Polymorphisms, Phenoconversion,

and Tissue-Selective Metabolism. Pharmaceutics 13:1466. https://doi.org/10.3390/pharmaceutics13091466

132. Gotoh O (1992) Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. J Biol Chem 267:83–90

133. Bezirtzoglou EEV (2012) Intestinal cytochromes P450 regulating the intestinal microbiota and its probiotic profile. Microb Ecol Health Dis 23:10.3402/mehd.v23i0.18370. https://doi.org/10.3402/mehd.v23i0.18370

134. Stavropoulou E, Pircalabioru GG, Bezirtzoglou E (2018) The Role of Cytochromes P450 in Infection. Frontiers in Immunology 9:. https://doi.org/10.3389/fimmu.2018.00089

135. Ohkawa H, Inui H (2015) Metabolism of agrochemicals and related environmental chemicals based on cytochrome P450s in mammals and plants. Pest Management Science 71:824–828. https://doi.org/10.1002/ps.3871

136. Bateman A, Martin MJ, Orchard S, et al (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49:D480–D489. https://doi.org/10.1093/NAR/GKAA1100

137. Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res 24:21–25. https://doi.org/10.1093/nar/24.1.21

138. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61-65. https://doi.org/10.1093/nar/gkl842

139. Barker WC, Garavelli JS, Huang H, et al (2000) The protein information resource (PIR). Nucleic Acids Res 28:41–44. https://doi.org/10.1093/nar/28.1.41

140. Bittrich S, Bhikadiya C, Bi C, et al (2023) RCSB Protein Data Bank: Efficient Searching and Simultaneous Access to One Million Computed Structure Models Alongside the PDB Structures Enabled by Architectural Advances. Journal of Molecular Biology 435:167994. https://doi.org/10.1016/j.jmb.2023.167994

141. Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881–10890. https://doi.org/10.1093/nar/16.22.10881

142. Chowdhury B, Garai G (2017) A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 109:419–431. https://doi.org/10.1016/j.ygeno.2017.06.007

143. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453. https://doi.org/10.1016/0022-2836(70)90057-4

144. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010

145. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113

146. Daugelaite J, O′ Driscoll A, Sleator RD (2013) An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. International Scholarly Research Notices 2013:615630. https://doi.org/10.1155/2013/615630

147. Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Research 43:39–49. https://doi.org/10.1093/nar/gkv416

148. Zheng F, Grigoryan G (2017) Sequence statistics of tertiary structural motifs reflect protein stability. PLOS ONE 12:e0178272. https://doi.org/10.1371/journal.pone.0178272

149. Mackenzie CO, Grigoryan G (2017) Protein Structural Motifs in Prediction and Design. Curr Opin Struct Biol 44:161–167. https://doi.org/10.1016/j.sbi.2017.03.012

150. Liao B, Shan X, Zhu W, Li R (2006) Phylogenetic tree construction based on 2D graphical representation. Chemical Physics Letters 422:282–288. https://doi.org/10.1016/j.cplett.2006.02.081

151. Sarma H, Pradhan S, Mattaparthi VSK, Kaushik S (2019) Phylogenetic Analysis: Early Evolution of Life. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp 938–952

152. Whelan S, Goldman N (2001) A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution 18:691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

153. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320. https://doi.org/10.1093/molbev/msn067

154. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282. https://doi.org/10.1093/bioinformatics/8.3.275

155. M O (1978) 22 A Model of Evolutionary Change in Proteins

156. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376. https://doi.org/10.1007/BF01734359

157. Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton. Journal of Molecular Biology 302:205–217. https://doi.org/10.1006/jmbi.2000.4042

158. Pei J, Kim B-H, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295–2300. https://doi.org/10.1093/nar/gkn072

159. Waterhouse AM, Procter JB, Martin DMA, et al (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191. https://doi.org/10.1093/BIOINFORMATICS/BTP033

160. Kumar S, Tamura K, Bioinformatics MN-, 1994 undefined (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. academic.oup.com 10:189–191

161. Kumar S, Stecher G, Li M, et al (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096

162. Schrödinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8

163. Lu H (2007) Stereoselectivity in drug metabolism. Expert opinion on drug metabolism & toxicology 3:149–158. https://doi.org/10.1517/17425255.3.2.149

164. Dong D, Wu B, Chow D, Hu M (2012) Substrate selectivity of drug-metabolizing cytochrome P450s predicted from crystal structures and in silico modeling. Drug Metabolism Reviews 44:192–208. https://doi.org/10.3109/03602532.2011.645580

165. Juvonen RO, Ahinko M, Jokinen EM, et al (2021) Substrate Selectivity of Coumarin Derivatives by Human CYP1 Enzymes: In Vitro Enzyme Kinetics and In Silico Modeling. ACS Omega 6:11286–11296. https://doi.org/10.1021/acsomega.1c00123

166. Hatherley R, Clitheroe C-L, Faya N, Tastan Bishop Ö (2015) Plasmodium falciparum Hop: Detailed analysis on complex formation with Hsp70 and Hsp90. Biochemical and Biophysical Research Communications 456:440–445. https://doi.org/10.1016/j.bbrc.2014.11.103

167. Faya N, Penkler DL, Tastan Bishop Ö (2015) Human, vector and parasite Hsp90 proteins: A comparative bioinformatics analysis. FEBS Open Bio 5:916–927. https://doi.org/10.1016/j.fob.2015.11.003

168. Zhou S-F, Liu J-P, Chowbay B (2009) Polymorphism of human cytochrome P450 enzymes and its clinical impact. Drug Metab Rev 41:89–295. https://doi.org/10.1080/03602530902843483

169. Couture J-F, Collazo E, Trievel RC (2006) Molecular recognition of histone H3 by the WD40 protein WDR5. Nat Struct Mol Biol 13:698–703. https://doi.org/10.1038/nsmb1116

170. Yano JK, Wester MR, Schoch GA, et al (2004) The Structure of Human Microsomal Cytochrome P450 3A4 Determined by X-ray Crystallography to 2.05-Å Resolution*. Journal of Biological Chemistry 279:38091–38094. https://doi.org/10.1074/jbc.C400293200

171. Gay SC, Roberts AG, Halpert JR (2010) Structural features of cytochromes P450 and ligands that affect drug metabolism as revealed by x-ray crystallography and NMR. Future Medicinal Chemistry 2:1451–1468. https://doi.org/10.4155/fmc.10.229

172. Guengerich FP (1992) Characterization of human cytochrome P450 enzymes. The FASEB Journal 6:745–748. https://doi.org/10.1096/FASEBJ.6.2.1537465/FORMAT/PDF

173. Sitbon E, Pietrokovski S (2007) Occurrence of protein structure elements in conserved sequence regions. BMC Struct Biol 7:3. https://doi.org/10.1186/1472-6807-7-3

174. Lam YWF (2019) Chapter 1 - Principles of Pharmacogenomics: Pharmacokinetic, Pharmacodynamic, and Clinical Implications. In: Lam YWF, Scott SA (eds) Pharmacogenomics (Second Edition). Academic Press, pp 1–53

175. Alvarado AT, Ybañez-Julca R, Muñoz AM, et al (2021) Frequency of CYP2D6*3 and *4 and metabolizer phenotypes in three mestizo Peruvian populations. Pharmacia 68:891–898. https://doi.org/10.3897/pharmacia.68.e75165

176. Lee HK, Tsongalis GJ (2009) Chapter 30 - Pharmacogenomics and Personalized Medicine in the Treatment of Human Diseases. In: Coleman WB, Tsongalis GJ (eds) Molecular Pathology. Academic Press, San Diego, pp 613–622

177. ur Rasheed MS, Mishra AK, Singh MP (2017) Cytochrome P450 2D6 and Parkinson's Disease: Polymorphism, Metabolic Role, Risk and Protection. Neurochem Res 42:3353–3361. https://doi.org/10.1007/s11064-017-2384-8

178. Puaprasert K, Chu C, Saralamba N, et al (2018) Real time PCR detection of common CYP2D6 genetic variants and its application in a Karen population study. Malaria Journal 17:427. https://doi.org/10.1186/s12936-018-2579-8

179. Rowland P, Blaney FE, Smyth MG, et al (2006) Crystal Structure of Human Cytochrome P450 2D6*. Journal of Biological Chemistry 281:7614–7622. https://doi.org/10.1074/jbc.M511232200

180. Wang A, Savas U, Hsu M-H, et al (2012) Crystal Structure of Human Cytochrome P450 2D6 with Prinomastat Bound. J Biol Chem 287:10834–10843. https://doi.org/10.1074/jbc.M111.307918

181. Guengerich FP (2018) Mechanisms of Cytochrome P450-Catalyzed Oxidations. ACS catalysis 8:10964. https://doi.org/10.1021/acscatal.8b03401

182. Meunier B, de Visser SP, Shaik S (2004) Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. Chem Rev 104:3947–3980. https://doi.org/10.1021/cr020443g

183. Phang-Lyn S, Llerena VA (2024) Biochemistry, Biotransformation. In: StatPearls. StatPearls Publishing, Treasure Island (FL)

184. Iacopetta D, Ceramella J, Catalano A, et al (2023) Impact of Cytochrome P450 Enzymes on the Phase I Metabolism of Drugs. Applied Sciences 13:6045. https://doi.org/10.3390/app13106045

185. McCarty KD, Ratliff SA, Furge KA, Furge LL (2021) Tryptophan-75 Is a Low-Energy Channel-Gating Residue that Facilitates Substrate Egress/Access in Cytochrome P450 2D6. Drug Metab Dispos 49:179–187. https://doi.org/10.1124/dmd.120.000274

186. Waal PW de, Sunden KF, Furge LL (2014) Molecular Dynamics of CYP2D6 Polymorphisms in the Absence and Presence of a Mechanism-Based Inactivator Reveals Changes in Local Flexibility and Dominant Substrate Access Channels. PLOS ONE 9:e108607. https://doi.org/10.1371/journal.pone.0108607

187. Berka K, Paloncýová M, Anzenbacher P, Otyepka M (2013) Behavior of human cytochromes P450 on lipid membranes. J Phys Chem B 117:11556–11564. https://doi.org/10.1021/jp4059559

188. Fischer A, Don CG, Smieško M (2018) Molecular Dynamics Simulations Reveal Structural Differences among Allelic Variants of Membrane-Anchored Cytochrome P450 2D6. J Chem Inf Model 58:1962–1975. https://doi.org/10.1021/acs.jcim.8b00080

189. Dong AN, Ahemad N, Pan Y, et al (2019) Functional and structural characterisation of common cytochrome P450 2D6 allelic variants-roles of Pro34 and Thr107 in catalysis and inhibition. Naunyn Schmiedebergs Arch Pharmacol 392:1015–1029. https://doi.org/10.1007/s00210-019-01651-0

190. Alali M, Ismail Al-khalil W, Rijjal S, et al (2022) Frequencies of CYP2D6 genetic polymorphisms in Arab populations. Human Genomics 16:6. https://doi.org/10.1186/s40246-022-00378-z

191. Ingelman-Sundberg M, Daly AK, Oscarson M, Nebert DW (2000) Human cytochrome P450 (CYP) genes: recommendations for the nomenclature of alleles. Pharmacogenetics and Genomics 10:91

192. Takemura K, Kitao A (2023) Chapter 27 - Molecular dynamics. In: Suetsugu S (ed) Plasma Membrane Shaping. Academic Press, pp 431–443

193. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317. https://doi.org/10.1007/s008940100045

194. Ponder JW, Case DA (2003) Force Fields for Protein Simulations. In: Advances in Protein Chemistry. Academic Press, pp 27–85

195. Hollingsworth SA, Dror RO (2018) Molecular dynamics simulation for all. Neuron 99:1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

196. Lopes PEM, Guvench O, MacKerell AD (2015) Current Status of Protein Force Fields for Molecular Dynamics. Methods Mol Biol 1215:47–71. https://doi.org/10.1007/978-1-4939-1465-4_3

197. Vanommeslaeghe K, Guvench O, MacKerell AD (2014) Molecular Mechanics. Curr Pharm Des 20:3281–3292

198. Best RB, Zhu X, Shim J, et al (2012) Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. J Chem Theory Comput 8:3257–3273. https://doi.org/10.1021/ct300400x

199. Cornell WD, Cieplak P, Bayly CI, et al (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc 117:5179–5197. https://doi.org/10.1021/ja00124a002

200. Maier JA, Martinez C, Kasavajhala K, et al (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 11:3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

201. Scott WRP, Hünenberger PH, Tironi IG, et al (1999) The GROMOS Biomolecular Simulation Program Package. J Phys Chem A 103:3596–3607. https://doi.org/10.1021/jp984217f

202. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J Am Chem Soc 118:11225–11236. https://doi.org/10.1021/ja9621760

203. Matsuo T, Miyake T, Hirota S (2019) Recent developments on creation of artificial metalloenzymes. Tetrahedron Letters 60:151226. https://doi.org/10.1016/j.tetlet.2019.151226

204. Hu L, Ryde U (2011) Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. J Chem Theory Comput 7:2452–2463. https://doi.org/10.1021/ct100725a

205. Stote RH, Karplus M (1995) Zinc binding in proteins and solution: A simple but accurate nonbonded representation. Proteins: Structure, Function, and Bioinformatics 23:12–31. https://doi.org/10.1002/prot.340230104

206. Li P, Song LF, Merz KMJr (2015) Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. J Chem Theory Comput 11:1645–1657. https://doi.org/10.1021/ct500918t

207. Pang YP, Xu K, Yazal JE, Prendergas FG (2000) Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. Protein Sci 9:1857–1865

208. Kollman PA, Massova I, Reyes C, et al (2000) Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. Acc Chem Res 33:889–897. https://doi.org/10.1021/ar000033j

209. Hoops SC, Anderson KW, Merz KMJr (1991) Force field design for metalloproteins. J Am Chem Soc 113:8262–8270. https://doi.org/10.1021/ja00022a010

210. Abraham MJ, Murtola T, Schulz R, et al (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2:19–25. https://doi.org/10.1016/j.softx.2015.06.001

211. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput 4:435–447. https://doi.org/10.1021/ct700301q

212. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. Computer Physics Communications 91:43–56. https://doi.org/10.1016/0010-4655(95)00042-E

213. CASE DA, CHEATHAM TE, DARDEN T, et al (2005) The Amber Biomolecular Simulation Programs. J Comput Chem 26:1668–1688. https://doi.org/10.1002/jcc.20290

214. Case DA, Aktulga HM, Belfon K, et al (2023) AmberTools. J Chem Inf Model 63:6183–6191. https://doi.org/10.1021/acs.jcim.3c01153

215. Salomon-Ferrer R, Case DA, Walker RC (2013) An overview of the Amber biomolecular simulation package. WIREs Computational Molecular Science 3:198–210. https://doi.org/10.1002/wcms.1121

216. Brooks BR, Brooks III CL, Mackerell Jr. AD, et al (2009) CHARMM: The biomolecular simulation program. Journal of Computational Chemistry 30:1545–1614. https://doi.org/10.1002/jcc.21287

217. Thompson AP, Aktulga HM, Berger R, et al (2022) LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales.

Computer Physics Communications 271:108171. https://doi.org/10.1016/j.cpc.2021.108171

218. Phillips JC, Braun R, Wang W, et al (2005) Scalable molecular dynamics with NAMD. Journal of Computational Chemistry 26:1781–1802. https://doi.org/10.1002/jcc.20289

219. Eastman P, Swails J, Chodera JD, et al (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology 13:e1005659. https://doi.org/10.1371/journal.pcbi.1005659

220. Eastman P, Galvelis R, Peláez RP, et al (2024) OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. J Phys Chem B 128:109–116. https://doi.org/10.1021/acs.jpcb.3c06662

221. Sousa da Silva AW, Vranken WF (2012) ACPYPE - AnteChamber PYthon Parser interfacE. BMC Research Notes 5:367. https://doi.org/10.1186/1756-0500-5-367

222. Sheik Amamuddy O, Glenister M, Tshabalala T, Tastan Bishop Ö (2021) MDM-TASK-web: MD-TASK and MODE-TASK web server for analyzing protein dynamics. Computational and Structural Biotechnology Journal 19:5059–5071. https://doi.org/10.1016/j.csbj.2021.08.043

223. Brown DK, Penkler DL, Amamuddy OS, et al (2017) MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics (Oxford, England) 33:2768–2771. https://doi.org/10.1093/BIOINFORMATICS/BTX349

224. Vendruscolo M, Dokholyan NV, Paci E, Karplus M (2002) Small-world view of the amino acids that play a key role in protein folding. Phys Rev E Stat Nonlin Soft Matter Phys 65:061910. https://doi.org/10.1103/PhysRevE.65.061910

225. Saldaño TE, Tosatto SCE, Parisi G, Fernandez-Alberti S (2019) Network analysis of dynamically important residues in protein structures mediating ligand-binding conformational changes. Eur Biophys J 48:559–568. https://doi.org/10.1007/s00249-019-01384-1

226. Barozi V, Musyoka TM, Sheik Amamuddy O, Tastan Bishop Ö (2022) Deciphering Isoniazid Drug Resistance Mechanisms on Dimeric Mycobacterium tuberculosis KatG via Post-molecular Dynamics Analyses Including Combined Dynamic Residue Network Metrics. ACS Omega acsomega.2c01036-acsomega.2c01036. https://doi.org/10.1021/acsomega.2c01036

227. Okeke CJ, Musyoka TM, Sheik Amamuddy O, et al (2021) Allosteric pockets and dynamic residue network hubs of falcipain 2 in mutations including those linked to artemisinin resistance. Comput Struct Biotechnol J 19:5647–5666. https://doi.org/10.1016/j.csbj.2021.10.011

228. Brown DK, Sheik Amamuddy O, Tastan Bishop Ö (2017) Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex. Global Heart 12:121–132. https://doi.org/10.1016/j.gheart.2017.01.006

229. Penkler DL, Atilgan C, Tastan Bishop Ö (2018) Allosteric Modulation of Human Hsp90α Conformational Dynamics. J Chem Inf Model 58:383–404. https://doi.org/10.1021/acs.jcim.7b00630

230. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38, 27–28. https://doi.org/10.1016/0263-7855(96)00018-5

231. Chovancova E, Pavelka A, Benes P, et al (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 8:e1002708. https://doi.org/10.1371/journal.pcbi.1002708

232. Atilgan AR, Akan P, Baysal C (2004) Small-world communication of residues and significance for protein dynamics. Biophys J 86:85–91. https://doi.org/10.1016/S0006-3495(04)74086-2

233. Liang Z, Verkhivker GM, Hu G (2020) Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications. Brief Bioinform 21:815–835. https://doi.org/10.1093/bib/bbz029

234. Golbeck J (2013) Analyzing the social web, First edition. Morgan Kaufmann is an imprint of Elsevier, Waltham, MA

235. Zhan J, Gurung S, Parsa SPK (2017) Identification of top-K nodes in large networks using Katz centrality. Journal of Big Data 4:16. https://doi.org/10.1186/s40537-017-0076-5

236. Ormeño F, General IJ (2024) Convergence and equilibrium in molecular dynamics simulations. Commun Chem 7:1–11. https://doi.org/10.1038/s42004-024-01114-5

237. Vieira-Pires RS, Morais-Cabral JH (2010) 310 helices in channels and other membrane proteins. Journal of General Physiology 136:585–592. https://doi.org/10.1085/jgp.201010508

238. Toniolo C, Benedetti E (1991) The polypeptide 310-helix. Trends Biochem Sci 16:350–353. https://doi.org/10.1016/0968-0004(91)90142-i

239. Oezguen N, Kumar S, Hindupur A, et al (2008) Identification and Analysis of Conserved Sequence Motifs in Cytochrome P450 Family 2. J Biol Chem 283:21808–21816. https://doi.org/10.1074/jbc.M708582200

240. Oezguen N, Kumar S (2011) Analysis of Cytochrome P450 Conserved Sequence Motifs between Helices E and H: Prediction of Critical Motifs and Residues in Enzyme Functions. J Drug Metab Toxicol 2:1000110

241. Peterson JA, Graham SE (1998) A close family resemblance: the importance of structure in understanding cytochromes P450. Structure 6:1079–1085. https://doi.org/10.1016/S0969-2126(98)00109-9

242. Sirim D, Widmann M, Wagner F, Pleiss J (2010) Prediction and analysis of the modular structure of cytochrome P450 monooxygenases. BMC Structural Biology 10:34. https://doi.org/10.1186/1472-6807-10-34

243. Mestres J (2005) Structure conservation in cytochromes P450. Proteins: Structure, Function, and Bioinformatics 58:596–609. https://doi.org/10.1002/prot.20354

244. Gilardi G, Di Nardo G (2017) Heme iron centers in cytochrome P450: structure and catalytic activity. Rend Fis Acc Lincei 28:159–167. https://doi.org/10.1007/s12210-016-0565-z

245. Skopalík J, Anzenbacher P, Otyepka M (2008) Flexibility of Human Cytochromes P450: Molecular Dynamics Reveals Differences between CYPs 3A4, 2C9, and 2A6, which Correlate with Their Substrate Preferences. J Phys Chem B 112:8165–8173. https://doi.org/10.1021/jp800311c

246. Jeřábek P, Florián J, Martínek V (2016) Lipid molecules can induce an opening of membrane-facing tunnels in cytochrome P450 1A2. Phys Chem Phys 18:30344–30356. https://doi.org/10.1039/C6CP03692A

247. Urban P, Lautier T, Pompon D, Truan G (2018) Ligand Access Channels in Cytochrome P450 Enzymes: A Review. Int J Mol Sci 19:1617. https://doi.org/10.3390/ijms19061617

248. Sheik Amamuddy O, Afriyie Boateng R, Barozi V, et al (2021) Novel dynamic residue network analysis approaches to study allosteric modulation: SARS-CoV-2 Mpro and its evolutionary mutations as a case study. Comput Struct Biotechnol J 19:6431–6455. https://doi.org/10.1016/j.csbj.2021.11.016

249. Stern S, Hyland PL, Pacanowski M, Schuck RN (2024) Leveraging in Vitro Models for Clinically Relevant Rare CYP2D6 Variants in Pharmacogenomics. Drug Metab Dispos 52:159–170. https://doi.org/10.1124/dmd.123.001512

250. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, et al (2017) Prediction of CYP2D6 phenotype from genotype across world populations. Genetics in Medicine 19:69–76. https://doi.org/10.1038/gim.2016.80

251. Twesigomwe D, Drögemöller BI, Wright GEB, et al (2022) Characterization of CYP2D6 Pharmacogenetic Variation in Sub-Saharan African Populations. Clin Pharmacol Ther 113:643–659. https://doi.org/10.1002/cpt.2749

252. Wang WY, Twesigomwe D, Nofziger C, et al (2022) Characterization of Novel CYP2D6 Alleles across Sub-Saharan African Populations. J Pers Med 12:1575. https://doi.org/10.3390/jpm12101575

253. da Rocha JEB, Othman H, Botha G, et al (2021) The Extent and Impact of Variation in ADME Genes in Sub-Saharan African Populations. Front Pharmacol 12:. https://doi.org/10.3389/fphar.2021.634016

254. Byrska-Bishop M, Evani US, Zhao X, et al (2022) High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 185:3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

255. Muller L, Jackson SN, Woods AS (2019) Histidine, the Less Interactive Cousin of Arginine. Eur J Mass Spectrom (Chichester) 25:212–218. https://doi.org/10.1177/1469066718791793

256. Rajakumara E, Saniya D, Bajaj P, et al (2022) Hijacking Chemical Reactions of P450 Enzymes for Altered Chemical Reactions and Asymmetric Synthesis. Int J Mol Sci 24:214. https://doi.org/10.3390/ijms24010214

257. Munro AW, McLean KJ, Grant JL, Makris TM (2018) Structure and function of the cytochrome P450 peroxygenase enzymes. Biochem Soc Trans 46:183–196. https://doi.org/10.1042/BST20170218

258. Pochapsky TC, Kazanis S, Dang M (2010) Conformational Plasticity and Structure/Function Relationships in Cytochromes P450. Antioxid Redox Signal 13:1273–1296. https://doi.org/10.1089/ars.2010.3109

259. Barik S (2020) The Uniqueness of Tryptophan in Biology: Properties, Metabolism, Interactions and Localization in Proteins. Int J Mol Sci 21:8776. https://doi.org/10.3390/ijms21228776

260. Khemaissa S, Sagan S, Walrant A (2021) Tryptophan, an Amino-Acid Endowed with Unique Properties and Its Many Roles in Membrane Proteins. Crystals 11:1032. https://doi.org/10.3390/cryst11091032

261. Klein K, Tatzel S, Raimundo S, et al (2007) A natural variant of the heme-binding signature (R441C) resulting in complete loss of function of CYP2D6. Drug Metab Dispos 35:1247–1250. https://doi.org/10.1124/dmd.107.015149

262. Šrejber M, Navrátilová V, Paloncýová M, et al (2018) Membrane-attached mammalian cytochromes P450: An overview of the membrane's effects on structure, drug binding, and interactions with redox partners. Journal of Inorganic Biochemistry 183:117–136. https://doi.org/10.1016/j.jinorgbio.2018.03.002

263. Hasemann CA, Kurumbail RG, Boddupalli SS, et al (1995) Structure and function of cytochromes P450:a comparative analysis of three crystal structures. Structure 3:41–62. https://doi.org/10.1016/S0969-2126(01)00134-4