Single nucleotide polymorphism induces divergent dynamic patterns in CYP3A5: a microsecond scale biomolecular simulation of variants identified in Sub-Saharan African populations

Houcemeddine Othman[a,*], Jorge da Rocha[a,b], Scott Hazelhurst[a,c]

[a] *Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.*
[b] *Division of Human Genetics, National Health Laboratory Service, and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.*
[c] *School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa.*

Abstract

Pharmacogenomics aims to reveal variants associated with drug response phenotypes. Genes whose roles involve the absorption, distribution, metabolism, and excretion of drugs, are highly polymorphic between populations. High coverage whole genome sequencing showed that a large proportion of the variants for these genes are rare in African populations. This study investigates the impact of such variants on protein structure to assess their functional importance. We use genetic data of *CYP3A5* from 458 individuals from sub-Saharan Africa to conduct a structural bioinformatics analysis. Five missense variants were modeled and microsecond scale molecular dynamics simulations were conducted for each, as well as for the CYP3A5 wildtype, and the Y53C variant, which has a known deleterious impact on enzyme activity. The binding of ritonavir and artemether to CYP3A5 variant structures was also evaluated. Our results showed different conformational characteristics between all the variants. No significant structural changes were noticed. However, the genetic variability acts on the plasticity of the protein. The impact on drug binding may be drug dependant. We conclude that rare variants hold relevance in determining the pharmacogenomics properties of populations. This could have a significant impact on precision medicine applications in sub-Saharan Africa.

Keywords: Pharmacogenomics, molecular dynamics, CYP3A5, Sub-saharan Africa

1. Introduction

The fate of drug molecules are determined by a set of biological processes controlling the Absorption, Distribution, Metabolism, and Excretion (ADME) in the organism.

*Corresponding author

Upon administration, drugs can induce adverse effects which include toxicity. The balance between the induction of the therapeutic effect and undesirable outcomes depends on the set of proteins encoding the ADME functions. Among these proteins, the CYP P450 superfamily is an important group of enzymes which play vital roles in Phase I metabolism of multiple drug types. Most bio-catalyzation reactions are processed by the CYP3A subfamily, with an estimated fraction of 24% of the total number of chemicals [27]. In particular, *CYP3A5* is a highly polymorphic gene with 29 total variants belonging to 9 star alleles (`https://www.pharmvar.org/gene/CYP3A5`) [28]. *CYP3A5*3* results in a defective mRNA and is perhaps the most relevant allele for phenotype classification [15]. Other variants of clinical importance include *CYP3A5*6* and *CYP3A5*7* which cause a protein truncation and a reading frame shift respectively (*idem*). The frequencies of these three major alleles are population dependant, particularly on ethnicity or ancestry. Allele frequencies range between 4-81%, 5-25% and 0-21% for *CYP3A5*3*, *CYP3A5*3* and *CYP3A5*7* respectively in Africa [3].

More than 280 drugs are known to interact with CYP3A5 including artemether for the treatment of uncomplicated *Plasmodium falciparum* infections [11] and ritonavir as a protease inhibitor for HIV [10]. Our recent analysis of the ADME genes from the genomes of 458 individuals from sub-Saharan Africa showed a high level of diversity even between populations residing at close geographical proximity [7]. Moreover, we also found that a large proportion of rare variants might be determinant for diversifying drug-response in African individuals. From these observations, we coined the term *"genetic diversity bottleneck in precision medicine"* to express the difficulty in accounting for low levels of granularity imposed by the genetic diversity when setting up targeted healthcare strategies.

Conceptual views of the functional implication of genetic variants have been long perceived as a binary categorization to be either deleterious or neutral. In particular for pharmacogenomic studies, this simplistic view offers limited insight into molecular mechanisms that control drug processing function. So far, the significant impact of rare variants in pharmacogenomics was only theorized in the context of genomic studies. Sequencing more African individuals will enrich the repertoire of rare variants with potentially high impact. The extent of their effect at the gene product level is poorly studied and such information is necessary to establish prioritization plans of variant actionability that is not based solely on the prevalence in the population. In addition, considering a single pharmacogene, the evaluations of the generic effect of rare variants on different drugs is a valuable information in defining the spectrum of action in diagnostic and treatment.

To evaluate the extent of rare variants genetic diversity in affecting the gene product functionality, we extended our analysis by focusing on *CYP3A5* gene based on the data

from high coverage whole genome sequencing from individuals from sub-Saharan Africa (H3Africa data), combined with 1000 Genome Project data. We then used molecular modelling and molecular dynamics to evaluate the effect of single amino acid variants on the conformational properties of protein structure. We also studied the effect on the binding of artemether and ritonavir as proof-of-concept.

2. Materials and Methods

2.1. Samples and genome analysis

The genetic characterisation of *CYP3A5* gene was derived from our previous study, which included the analysis of high coverage whole genome sequencing data from 458 samples from 15 African countries. These data were collected by the H3Africa/GSK ADME consortium. This analysis also includes data from 507 African samples from the *1000 Genomes* initiative. For the sake of brevity, here we only give the outlines of the genome analysis method but a detailed description is extensively exhibited by da Rocha *et al* [7] (Paper accepted in frontiers in pharmacology and is still under production process but can be accessed in its preprint version). The GrCh37 assembly was used as a reference for the genome coordinates and the gene transcripts. The variants of *CYP3A5* gene were annotated using the 'g_miner' workflow (`(https://github.com/hothman/PGx-Tools/tree/master/workflows/g_miner`). Variant Effect Predictor (VEP) [20] was used for the functional annotation of the missense variants that combines different annotation and prediction algorithms. We have also integrated the predictive approach by Zhou *et al* [39] developed specifically for the annotation of pharmacogenes. Allele frequencies of *CYP3A5* were calculated using PLINK version 1.9 [5].

2.2. System preparation

We have simulated three different systems of CYP3A5 protein. These include: a substrate free enzyme (PDB code: 6MJM) [12], the complex with ritonavir (PDB code 5VEU) [13] and the CYP3A5/artemether complex generated by docking. The latter also uses the PDB entry 5VEU for the structure of the enzyme. Missing loops (segments 260-270, 280-288 in the substrate free structure and segment 282-288 in the bound structure) were generated with the loop modelling routine from MODELLER software version 9.22 [30]. Five hundred conformations were generated and the structure with the best DOPE score [33] was retained for the further steps. We estimated the ionization states at pH of 7.4 of all the charged groups of CYP3A5 using PROPKA3 [26] according to which we neutralized residues Asp^{182}, Asp^{269} and Glu^{294}.

2.3. Force field parameters for heme group, ritonavir and artemether

Parameters for the CYP3A5 amino acid residues are derived from the AMBER ff14SB force field which is in better agreement with experimental results to simulate cytochrome P450 proteins [24]. General Amber Force Field (GAFF) [38] atom types were assigned to the heme group and parameters from Shahrokh *et al*, [32] were used to account for the bonded configuration between the Cys^{441} sulphur atom and the Fe metal ion as well as the particular electronic state of the heme group. For the substrate-free enzyme, we have used the parameters of compound I characterised by the formation of a covalent bond between an oxygen atom and the heme iron as it was seen in the crystal structure [23]. For the bound configuration with ritonavir and artemether, we used the parameters of the penta-coordinate ferric high-spin configuration of the heme group. Parameters from the GAFF force field were assigned to both ritanavir and artemether ligands and the partial charges were calculated by the AM1-BCC method implemented in the antechamber tool [37].

2.4. Artemether docking to CYP3A5

Autodock Vina [36] was used to predict the complex between artemether and CYP3A5. The partial charges calculated by antechamber were assigned to the structure of the ligand in 'PDBQT' format. Partial charges calculated by Shahrokh *et al*, were assigned to the heme group atoms *et al* [32] as part of the receptor structure. This method has been shown by the authors, to significantly improve the performance of the docking. Gasteiger charges were assigned to all amino acids of CYP3A5. The docking was conducted in a grid box that covers the entire catalytic pocket of the receptor. We executed five runs of docking starting from different seeds to ensure the consistency of the results. The degree of exhaustiveness was set to 200 while keeping active all the rotatable bonds of artemether. From the nine binding modes that were returned, we selected those that interact directly with the heme atoms, after which we ran 100 ns of molecular dynamics simulation. The choice of a reasonable binding model is based on five properties that include the fraction of native contacts, the distance of the center of mass to the iron atom of the heme group, and the Root Mean Square Deviation of all atoms in the ligand compared to the docking pose.

2.5. Generating the structures for the variants

We have generated five structures incorporating the variants from the combined H3Africa and 1000 Genomes project data as well as for the Y53C variant (corresponding to the *CYP3A5*11* allele) for which there is strong biochemical evidence about its role in reducing the catalytic capacity of the enzyme [17]. Structures of the variants

were generated by removing all uncommon atoms that have different atom names of the side chains with the reference position. The amino acid's name is then changed to its corresponding variant. We then used the software 'leap' from AMBER 18 [4] to build the missing amino acid atoms. Initial refining included removal of severe steric clashes by running two *in vacuo* energy minimization cycles corresponding to 250 steps of the steepest descent stage followed by 400 steps of conjugate gradient minimization. During the run, a restraint force constant of 2 kcal/mol/Å^2 were applied on all the atoms of the system except those belonging to the variant position. The same routine was also applied to the reference structure (Wild type form).

2.6. Molecular dynamics

Molecular dynamics experiments involve the *in silico* simulation of a protein structure over a time based trajectory. The molecular dynamics simulation aims to sample conformations of highly populated states of a protein. The sampling constructs a trajectory of conformations that are representative of an energy landscape. The latter is altered between the CYP3A5 variants, subjects to this analysis. Interpretations of variant impact can be made by assessing structural changes throughout the trajectory. These can be used to better our understanding of variant impact on structural stability and ligand interaction, which may have implications for drug response. Prior to running the simulation, the system was first solvated with a TIP3P octahedron box. Chloride counter-ions were added conveniently in order to neutralize the charge of the system. The distance between the edges of the simulation box and the solute atoms was set to a minimum of 15 Å. Before the execution of the molecular dynamics simulation, two stages of energy minimization were run to remove the steric clashes resulting from adding the solvent molecules to the system. In the first stage, we run a steepest descent integrator for 500 steps followed by 9500 conjugate gradient minimization while restraining all the atoms of the solute with a force of 500 kcal/mol/Å^2. In the second stage, we run two consecutive steepest descent and conjugate gradient minimizations for a 1000 and 6000 steps respectively. The refinement was run under implicit solvent conditions by activating the Hawkins, Cramer, Truhlar Generalized Born model.

The molecular dynamics simulation begins with heating the system from 50 K to the desired temperature of 300 K. We have fixed the random seed number (ig) to 96465. Restraining forces were applied on all the atoms of the solute at a value of 10 kcal/mol/Å^2. A Langevin thermostat with a collision frequency of 5 ps^{-1} was applied to control the temperature fluctuation around 300 K. The heating stage was run for a total time of 20 ps after which we run an Isothermal–isobaric (NPT) simulation. At this stage, we used a restraining force of 10 kcal/mol/Å^2 applied on all atoms of the solute. Then, the force was gradually lifted by a decrement of 2 kcal/mol/Å^2 until the system is equilibrated. Each

of the stages is run for a total time 80 ps. The final production phase was operated in the canonical ensemble (NVT). We generated a continuous trajectory spanning a simulation time of 1.5 µs for the unbound CYP3A5 structure. During all the simulation stages, periodic boundary conditions were applied. The Particle Mesh Ewald approximation was used for efficient calculation of the electrostatic interactions. We also used a cutoff value of 12 Å for the non-bonded-interactions. The SHAKE algorithm was applied to allow for an integration time of 2 ps and snapshots are collected at intervals of 10 ps.
We have used a slightly different protocol for the molecular dynamics simulation of the CYP3A5/artemether and CYP3A5/ritonavir complexes by running three independent simulations each with a total of 200 ns production time. The random seed number for the three trajectories were set to 18917, 74269, and 96465.

2.7. Analysis of molecular dynamics simulation

An in house script was employed to analyse the data from the molecular trajectory simulations. The code is available at `https://github.com/hothman/MolSim/blob/master/molecular_dynamics/traj_analysis.py`. The library is written in python and uses mdtraj [19] and pytraj [29] modules for the development of routines and functions.

2.7.1. Essential dynamics analysis

Extracting relevant information from an ensemble of molecular dynamics conformations is facilitated by essential dynamics methods. These aim to reduce the dimensionality of the dataset and to detect biologically relevant patterns of collective motions. In doing so, we applied principal component analysis for the conformations at the equilibrated phase of the trajectory. The ensemble of snapshots was first fitted to the crystal structure of CYP3A5 to remove the degrees of freedom related to transitions and rotations. The low dimension components were calculated by the 'pca' method from 'pytraj' to return the corresponding eigenvalues and eigenvectors as well as the projection of the atomic coordinates into the lower-dimensional subspaces. Only C_α atoms were considered for the calculations of the Principal Components (PCs).

The Root Mean Square Inner Product (RMSIP) [1] was calculated to evaluate the overlap between the wild type and the variant conformational ensembles according to the following formula :

$$RMSIP = \sqrt{\frac{1}{20}\sum_{i=1}^{20}\sum_{j=1}^{20}(\mu_i \cdot \nu_i)^2}$$

Here, μ_i and ν_i correspond, respectively, to the first 20 eigenvectors calculated from the wild type and the variant trajectories.

Elements of the dynamic cross-correlation matrix were calculated according to the following formula.

$$DCC(i,j) = \frac{\langle \Delta \mathbf{r}_i(t) \cdot \Delta \mathbf{r}_j(t) \rangle}{\sqrt{\langle \Delta \mathbf{r}_i(t) \cdot \Delta \mathbf{r}_i(t) \rangle \langle \Delta \mathbf{r}_i(t) \cdot \Delta \mathbf{r}_j(t) \rangle}}.$$

$\Delta \mathbf{r}_i(t)$ and $\Delta \mathbf{r}_j(t)$ are displacement vectors at time t calculated by subtracting the coordinates vector of an atom i or j from the average coordinates calculated over the entire ensemble.

2.7.2. Free energy landscape

The Free Energy Landscape (FEL) can be established using the density data of reaction coordinates to explore the energy property of a protein in a 3-dimensional space. FEL was obtained according to the following equation:

$$\Delta G = -k_b T \ln P(X)$$

where k_b corresponds to the Boltzmann constant, T is the absolute temperature and P(X) is the probability of the reaction coordinate X obtained by binning the distribution of the data. In our study we have used the projections of C_α atom coordinates over PC1 and PC2 as reaction coordinates.

2.7.3. Binding free energy calculation

The binding energy between a ligand and a receptor was estimated according to the formalism established by Duan *et al,* [9]. The method deviates slightly from the original scheme [35] by stating that the binding energy is the outcome of a linear combination between a gas term (ΔG_{gas}) and a solvation term (ΔG_{solv}).

$$G_{\text{binding}} = \Delta G_{\text{gas}} + \Delta G_{\text{solv}}$$

The entropic contribution makes part of the gas term according to the following equation.

$$\Delta G_{\text{gas}} = \langle E_{\text{pl}}^{\text{int}} \rangle - T\Delta S$$

$\langle E_{\text{pl}}^{\text{int}} \rangle$ is calculated by averaging the interaction energy between the ligand and the protein over the conformations constituting the ensemble of molecular dynamics trajectory, here:

$$-T\Delta S = k_b T \quad \ln[\langle e^{\beta \Delta E_{\text{pl}}^{\text{int}}} \rangle]$$

k_b is the Boltzmann constant, T is the absolute Temperature and $\beta = 1/(k_b T)$. ΔG_{gas}, ΔG_{sol} and $\langle E_{\text{pl}}^{\text{int}} \rangle$ can be calculated by the MMPBSA.py program [22] from AMBER molecular dynamics package.

2.7.4. Cavity volume and tunnels geometry calculations

To compute the volume of the catalytic binding site of CYP3A5 we employed the program 'mdpocket'[31]. Additionally, we used caver 3.0.3 [6] to calculate the geometry of the tunnels within the enzyme used by the drugs to enter to the catalytic binding site. The starting point coordinates were set to the position of the iron atom in the heme group. The probe radius, the shell radius, and the shell depth were set to values of 0.9, 5, and 4 respectively and we used a threshold of 3 for the hierarchical clustering of the tunnel's geometries within CYP3A5.

3. Results

3.1. Genetic characterization of CYP3A5 variants from WGS

The genome analysis and variant annotation of the combined High Coverage African ADME dataset (HAAD) and 1000 Genomes Project (KGP) datasets identified 524 variants of which a large proportion corresponds to intronic and non-coding regions. A frameshift variant (rs41303343-A), caused by a deletion, has a frequency of 0.113 in African populations. The variant is identified in *CYP3A5*1*, *CYP3A5*3*, *CYP3A5*6*, and *CYP3A5*7*, and is associated with defective metabolism of many drugs including tacrolimus and ritonavir. Only 5 missense variants were identified in the source African populations of this study, all are rare with frequencies below 0.006 (Table 1). None of these variants have been reported by ClinVar.

These variants were mapped to the protein structure of CYP3A5 (Figure 1). Except for amino acid substitutions R28C and V238A, all the other variants are located in the buried core of the protein (Figure 1). V238A belongs to the F/G loop of CYP3A5 that controls the access to the catalytic binding site [18].

3.2. Molecular dynamics simulation of CYP3A5 variants

Molecular dynamics simulation was run for the wild type and the variant forms of CYP3A5 in the absence of the substrate (Unbound structure) for a total production time of 1.5 µs.

Table 1: Frequency and amino acid substitution of missense variants characterized for *CYP3A5* gene from combined HAAD and KGP data.

RS ID	Amino acid variant	Frequency in HAAD+KGP
rs55817950	R28C	0.0021
rs56244447	L82R	0.0054
rs142823108	I149T	0.0043
rs542523237	V238A	0.001
rs145774441	I276T	0.0021
rs72552791	Y53C	0.000

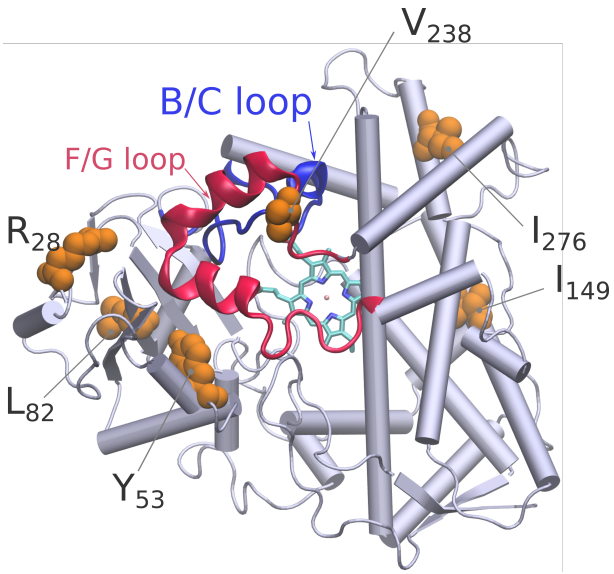

Figure 1: Position of the amino acids corresponding to the missense variants on the 3D structure of CYP3A5. The figure was generated using the Visual Molecular Dynamics program [14].

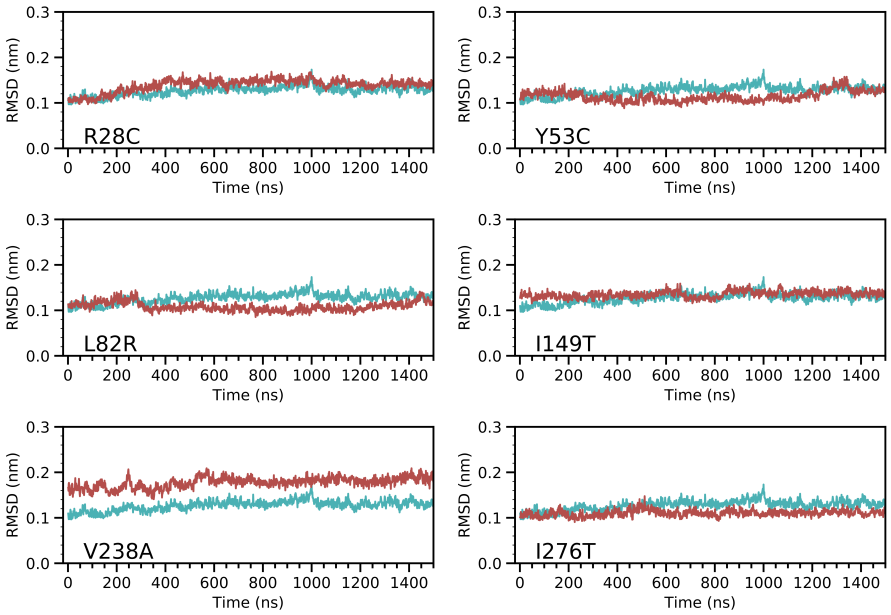

Figure 2: Analysis of the structural deviation CYP3A5 variants (Red plots) compared to the wild type form (Cyan plot). All the trajectory frames were first fitted to the crystal structure of the enzyme and the calculation was made using the coordinates of the backbone atoms. Loops 260-270 and 280-288 were omitted from the calculation.

3.2.1. Overall effect of the variants on CYP3A5 structure

The convergence of molecular dynamics trajectories was achieved (supplementary material 1). We calculated the Root Mean Square Deviation (RMSD) along the entire production time. We excluded the loop segments that were missing in the crystal the structure of CYP3A5 from the RMSD calculation because their high flexibility may prevent the detection of any relevant structural divergence located on other parts of the protein. The RMSD calculation considered the unbound crystal structure of CYP3A5 (PDB code: 6MJM) as a reference. The trajectory of the wild type structure was equilibrated at around 200 ns and shows a stable behaviour along the rest of the simulation time with short plateau phases that indicate the sampling of different local energy minima (Figure 2). Unless mentioned explicitly, all further analysis uses the snapshots from the equilibrated phase of the molecular dynamics trajectory. All the variant structures show little deviations from the WT RMSD profile and remain below 0.2 nm. In particular, the Y53C variant which was found to decrease the activity of CYP3A5 [17] has no significant deviation compared to the WT backbone atoms. The largest deviation was observed for the V238A variant which averages 0.5 Å along the full simulation time.

To assess the local fluctuation of CYP3A5 segments, we calculated the Root Mean

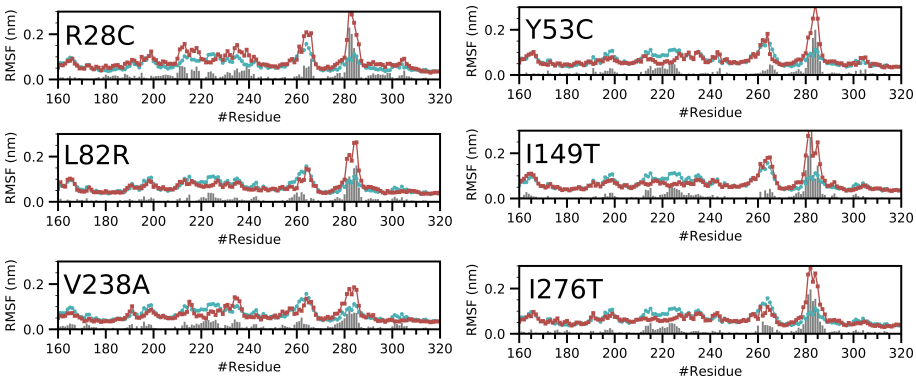

Figure 3: The fluctuation of the amino acid residues for CYP3A5 variants compared to the wild type form (Blue plots) and the variant forms (red plots). The bars indicate the difference of the fluctuation between the equivalent amino acids of the wild type and the variant. We only show the segment corresponding to amino acids 160-320. The RMSF profiles for the entire protein can be found in the supplementary material 2.

Square Fluctuation (RMSF) for the C_α atoms (Figure 3). The highest fluctuation values correspond to the loop regions 260-270, 280-288 which is expected because the electron density map in the protein structure was not solved for these segments. All other parts of CYP3A5 do not show a major deviation from the wild type structure except for the F/G loop segment spanning the 210-240 amino acids. The divergent fluctuation pattern for these residues differs between the variants. L82R seems to be the least affected given that the local differences in fluctuations are low compared to the other variants. The extend of the divergent fluctuation at the F/G loop is limited to segment 210-230 for Y53C, I149T, I276T variants while it is much wider for R28C and V238A.

3.3. Analysis of the collective motions within CYP3A5

Dynamic Cross-Correlation Matrices detect blocks of residues that move collectively in the 3D space. A collective motion can be either correlated or anticorrelated. We focused our analysis on pairs of residues that show correlation values of more than 0.5 or less than -0.5. We noticed that variants R28C, Y53C, L82R, and V238C display differences in the DCCM compared to the wild type form (Figure 4) characterized by an increase in the intensity of some zones in the DCCM and a dense network of connectivities between the C_α atoms. The analysis revealed four regions of CYP3A5 that are most affected by the single amino acid substitution: The F/G loop, the F and G helices, the N-terminus domain, and the I and G helices. The R28C variant shows a dense network at the N-terminus domain. The motion of the region 291-302 of the I-helix is correlated with the N-terminus part of the G helix spanning residues 243-252. The DCCM analysis shows also correlated dynamics between the entirety of the F helix and the first three N-

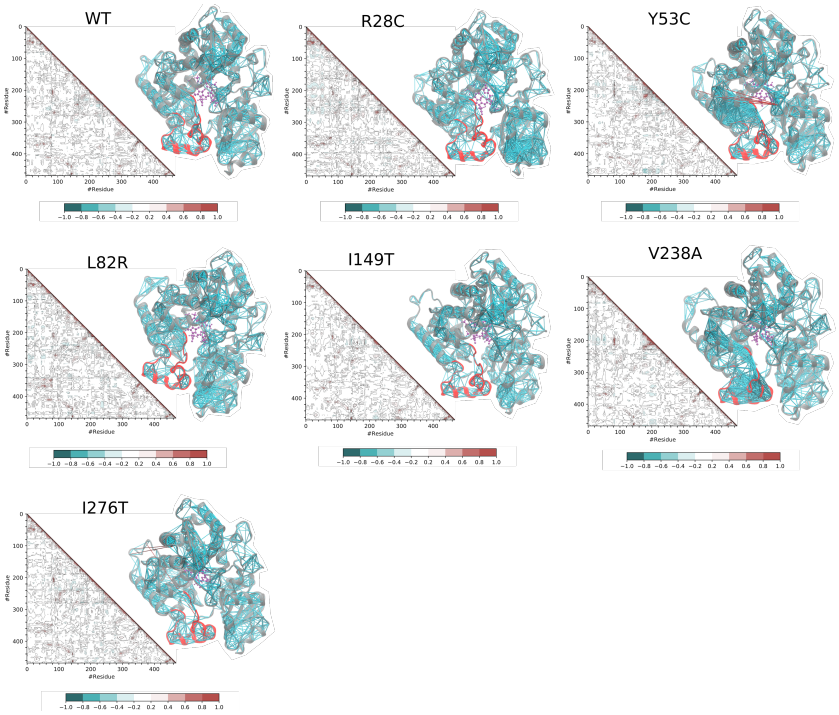

Figure 4: Dynamic Cross-Correlation Matrix (DCCM) analysis for the wild type and the variant forms. The data from each DCCM were mapped on the structure as a link between the pairs of residues established if $DCC(i, j)$ is above 0.5 or below -0.5. The F/G loop is marked in red colour on the structure. The index of the residues started at 0 and the reader should adjust with an offset of 25 to get the correct position in the sequence.

terminus turns of the G helix. The same type of collective dynamic was also observed in L82R, however not as much as important as for Y53C. V238A variant shows an increase in the network density for the F/G loop, and between the F and G loops. Moreover, with five residues belonging to the B/C loop, the correlation network established with the F/G loop is extensive compared to the wild type form which involves only three residues.

3.4. Essential dynamics of CYP3A5

The analysis, up to this point, has shown local differences between the different variants and the wild type forms. We then asked the question of whether the single variation in the amino acid sequence of CYP3A5 can lead to a significant effect on the essential conformational modes. The C_α atom coordinates of each snapshot of the trajectories were projected onto PC1, PC2, PC3, and PC4 subspaces. Because the flexible loops of CYP3A5 can be dominant in the calculated modes, we choose to discard them from the PCA analysis. We noticed that the first 7 PCs are different between the

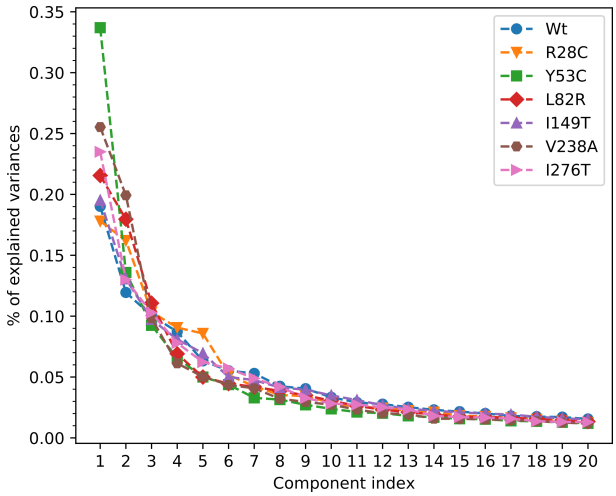

Figure 5: Scree plot showing the proportion of explained variance as a function of the component rank.

analyzed variants as shown from their divergent weights in the explained variance. The first four PCs account for 50% of the explained variance for the wild type form (Figure 5). The value increases to 63% and 61% for Y53C and V238A variants. It remains below 57% for all the other variants. Most of the differences are found for the first PC (PC1). For the wild type form, it contributes to 19% of the explained variance. I149T and R28C show close values to the wild type form. Y53C and V238A show the highest differences with values of 34% and 25% respectively for PC1.

To estimate the overlap between the macromolecular modes, we calculated the RMSIP between the first 20 eigenvectors of the wild type form and each of the 6 variants. The RMSIP values can be rounded to 0.70-0.8 for all the variants which shows a significant overlap between the trajectory of the wild type form and the other variants. The PC1 vs PC2 plot shows however some disparities between the 2 major dynamic modes (Figure 6). The molecular dynamics snapshots of R28C, Y53C, L82R, V238A, and I276T are extended to a wider range in the 2D phase space compared to the wild type form. In general, the variance is more important for PC1. On the other hand, PC3 vs PC4 plot does not differ widely between the wild type and the mutant (supplementary material 3).

The amplitude and the direction of movement of each C_α atoms were estimated by projecting the coordinates on the first major eigenvector that explains the variance of the dynamics. The porcupine plot (Figure 7) shows that residues belonging to the segments controlling the access to the catalytic site are the most major contributors in the protein motion for all the variants. However, the amplitude and the direction of these amino

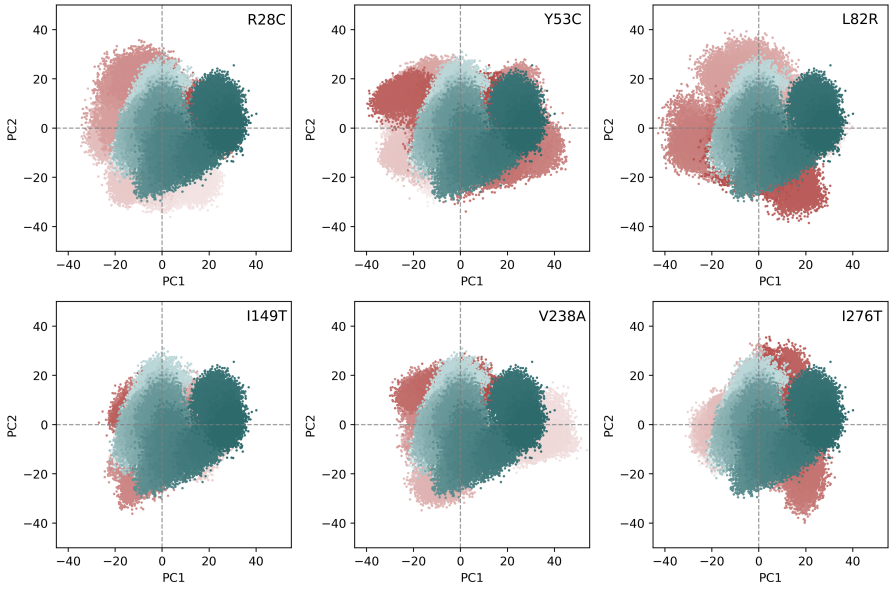

Figure 6: Projection of the C_α atoms coordinates on the first and second principal components. The blue and red colours indicate the snapshots of the wild type and the variant trajectories respectively. The calculation included only the snapshots collected after the 200 ns time.

acids differ significantly from one variant to another and compared to the wild type form. Both Y53C and V238A variants showed a wider divergence from the wild type form. The amplitude of displacement is important, and mainly affects the F/G loop segment.

3.5. Analysis of the free energy landscape

Next, we asked the question of whether the variation in the amino acid sequence of CYP3A5 can affect the energy property of folding which would be best assessed at the global minimum of the Free Energy Landscape (FEL). To proceed, we choose PC1 and PC2 as the reaction coordinates of the FEL to ensure better discrimination between the different protein states. We noticed that the free energy surface differs significantly in terms of shape between the different assessed variants. Compared to the wild type, Y53C, I149T, V238A, and I276T have many local energy wells that have a depth close to the global energy (supplementary material 4). To appreciate the differences between the conformation at the global minimum of each trajectory, we extracted them from each FEL, then projected their coordinates onto the subspaces described by the first and the second eigenvectors of the wild type form. In figure 8, we show the FEL of the wild type form in 3D representation and its corresponding 2D projection. We noticed that all the conformations at the global minimum of each variant are located away from the

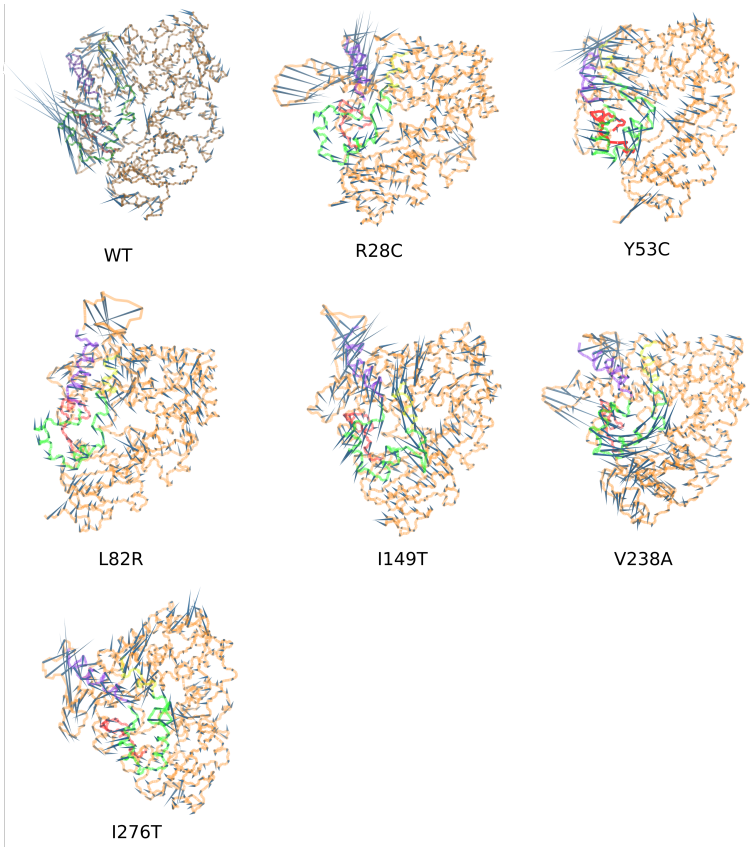

Figure 7: Projection of the C_α atoms coordinates on the first and second principal components. The blue and red colours indicate the snapshots of the wild type and the variant trajectories respectively. The calculation included only the snapshots collected after the 200 ns time.

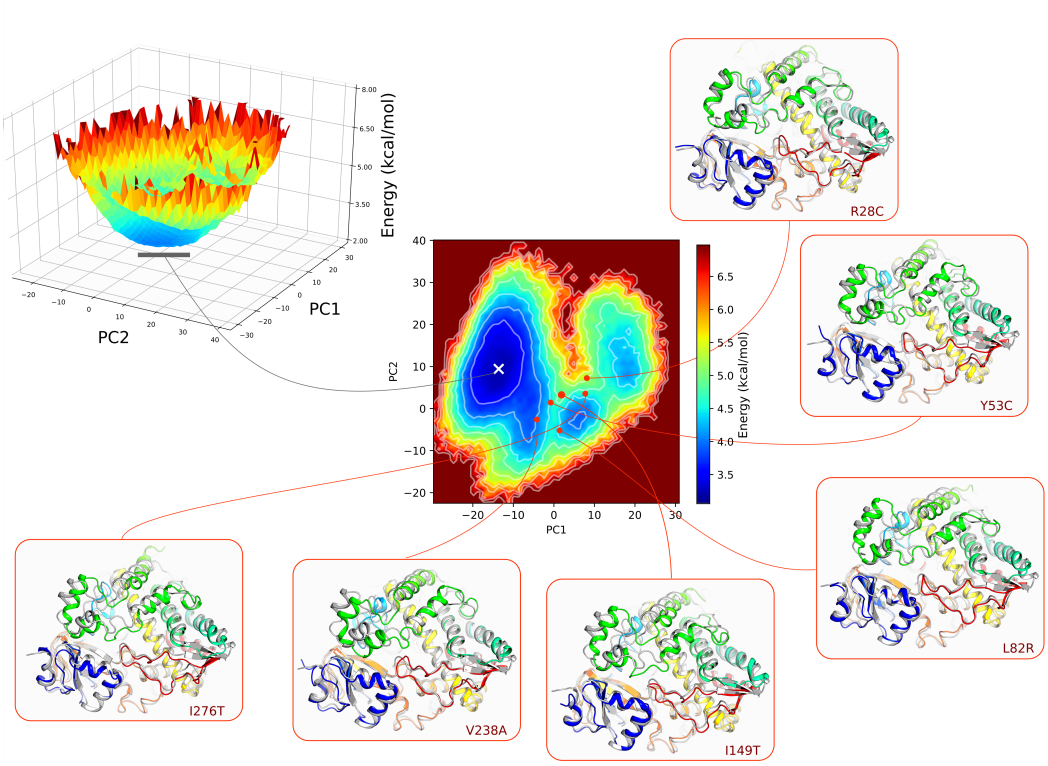

Figure 8: Free energy landscape of CYP3A5. PC1 and PC2 were calculated for all the C_α atoms excluding the highly flexible loops. The 3D representation of the FEL was projected in a 2D heatmap. Contour plots were also included to differentiate between the different levels of the energy. The minimum and the maximum energy values are 3.05 kcal mol^{-1} and 6.97 kcal mol^{-1}. each level in the contour plot corresponds to $\pm$3.05 kcal mol^{-1}. The position of the conformation at the global minimum of the wild type form is marked by the white cross. Structural fitting was performed between the conformations of the variants at their corresponding global minimum (Rainbow colour) and the conformation of the wild type form (Gray colour). The segments spanning the F helix, G helix and F/G loops are coloured in green and the B/C loop corresponds to the blue cartoon.

position of the global energy well of the wild type form. Some of the conformations are even positioned at higher energy levels. Structural fitting to the wild type conformation at the global minimum shows a large deviation mainly at the F/G loop, the F and G helices and the B/C loop.

3.6. Catalytic pocket volume calculation

We desired to verify the impact of the genetic diversity on the geometry of the catalytic pocket of CYP3A5 (Figure 9A). All the simulated variants show differences from the wild type form with similar values of variance (Figure 9B). The Y53C variant shows the least difference with an average volume of 2019.04 Å^3 which is slightly larger than the

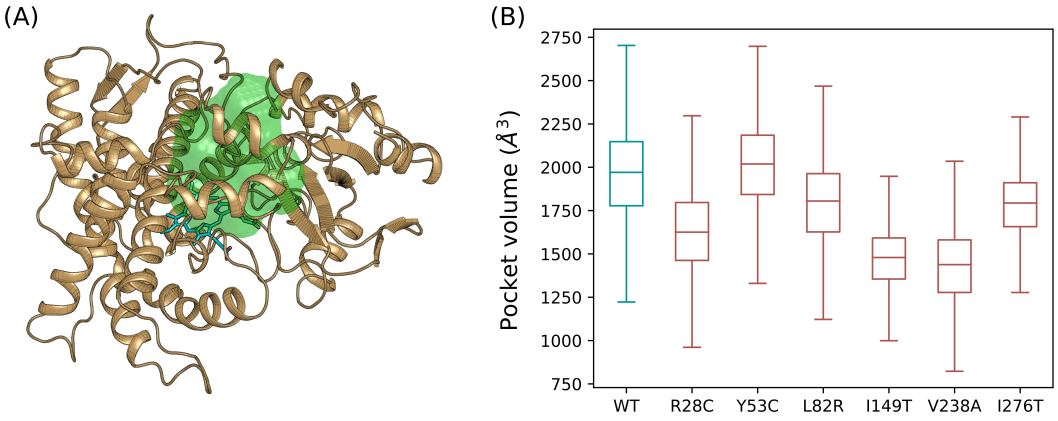

Figure 9: Calculation of the catalytic pocket volume of CYP3A5. (A) Location of the catalytic pocket calculated for the wild type form relative to the heme group. (B) The catalytic pocket volume was calculated for an ensemble of molecular dynamics snapshots at the equilibrium phase represented in boxplots.

volume of the wild type pocket measuring 1970.70 $Å^3$. L82R and I276T show close values of 1805.27 $Å^3$ and 1793.63 $Å^3$ respectively. Moreover, R28C, I149T, and V238A have the smallest catalytic pocket volume with values of 1625.61 $Å^3$, 1479.37 $Å^3$, and 1438.22 $Å^3$ respectively which represent a shrinkage of 17%, 24%, and 27% respectively relative to the volume calculated for the wild type form.

3.7. Binding free energy estimation of CYP3A5/ritonavir and CYP3A5/artemether complexes

To assess the effect of CYP3A5 variants on its binding capacity to drug molecules, we studied the dynamics of the enzyme in its bound form with ritonavir and artemether having 50 and 21 heavy atoms respectively. They were selected to test the effect on the binding of small and large size ligands to the catalytic binding pocket. We have obtained a reasonable solution for the CYP3A5/artemether predicted by molecular docking. The complexes were used to run 3 independent molecular dynamics simulations for a accumulated time of 600 ns for each variant. The binding free energy included the snapshots collected from the last 100 ns for each independent trajectory (a total time of 300 ns). We have verified the convergence of the calculation using this approach and we have put more details in supplementary material 5.

According to the MM-GBSA energy (Table 2) estimated for CYP3A5/ritonavir complex, the drug binds to all the forms with favourable energies except for V238A where $\Delta G_{\text{binding}} = 1.44$ kcal/mol. Y53C and I149T also show weak binding energy values while the affinity of the drug is better for R28C, L82R, and I276T compared to the

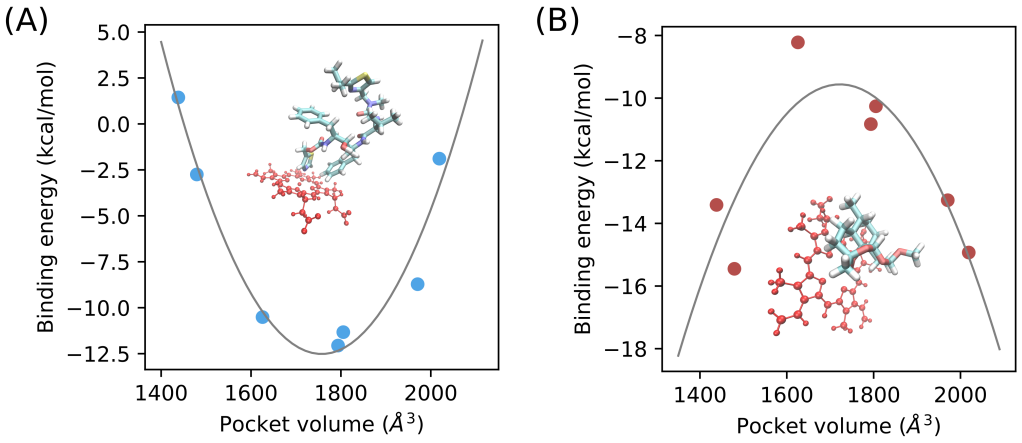

Figure 10: MM-GBSA energy as a function of the catalytic pocket volume for ritonavir (A) and artemether (B). The curves were fitted using a polynomial model of the form $y = (x-a)(x-b)(x-c)$ where a= $1.33.10^{-4}$, b=-0.468, c=$3.98.10^{2}$ for ritonavir and a= $-6.25.10^{-5}$, b=0.215, c=$-1.95.10^{2}$ for artemether. We also show the ligand spatial configuration relative to the heme group of CYP3A5.

wild-type form. In general, the variance of the ΔG_{sol} term is less important compared to the entropy. V238A showed the least favourable $\langle E_{\text{pl}}^{\text{int}} \rangle$ among all the other variants at $-55.97\,\text{kcal mol}^{-1}$. Y53C and V238A show the least favourable entropy terms evaluated at 21.08 kcal/mol and 19.45 kcal/mol respectively.

Figure 10A shows the variation of the estimated MM-GBSA energy as a function of the cavity volume calculated from the previous analysis. Relative to the range of the calculated binding energies, extreme values correspond to either small or large pocket volume. We have fitted these data to a polynomial regression model of the form $y = (x-a)(x-b)(x-c)$ where a= $1.33.10^{-4}$, b=-0.468, c=$3.98.10^{2}$ for ritonavir and a= $-6.25.10^{-5}$, b=0.215, c=$-1.95.10^{2}$ for artemether. A strong strong coefficient of determination (R^2) of 0.94 was noticed for ritonavir. Artemether binding to CYP3A5 is less favourable to all variants ranging from $-15.45\,\text{kcal mol}^{-1}$ for I149T to $-8.22\,\text{kcal mol}^{-1}$ for R28C 2). The variance of $\Delta G_{\text{binding}}$ is less important compared to the values estimated for ritonavir. The variance is even lower considering $\langle E_{\text{pl}}^{\text{int}} \rangle$ energy term which ranges from -29.31 to $-33.16\,\text{kcal mol}^{-1}$. Also, $-T\Delta S$ and ΔG_{solv} energy terms equally contribute to the unfavourable energy penalty in all variants. For instance, we estimated similar values of 12.08 kcal/mol and 12.85 kcal/mol respectively for $-T\Delta S$ and ΔG_{solv} energy terms for the R28C variant. Moreover, the fitting of a polynomial regression between the catalytic pocket volume and $\Delta G_{\text{binding}}$ for artemether shows a weaker R^2 of 0.73 compared to ritonavir (Figure 10B).

Table 2: Binding free energy and corresponding energy terms calculated for CYP3A5/ritonavir and CYP3A5/artemether complexes.

Drug	Variant	$\Delta G_{binding}$ (kcal/mol)	$< E_{pl}^{int} >$ (kcal/mol)	$-T\Delta S$ (kcal/mol)	ΔG_{solv} (kcal/mol)
ritonavir	WT	-8.72	-58.46	11.01	38.72
	R28C	-10.51	-65.18	14.01	40.66
	Y53C	-1.89	-64.09	21.08	41.11
	L82R	-11.33	-60.87	12.69	36.84
	I149T	-2.75	-57.84	15.29	39.79
	V238A	1.44	-55.97	19.45	37.96
	I276T	-12.06	-60.85	13.58	35.20
artemether	WT	-13.26	-29.57	7.75	8.55
	R28C	-8.22	-33.16	12.08	12.85
	Y53C	-14.93	-29.31	6.13	8.24
	L82R	-10.26	-30.71	10.36	10.08
	I149T	-15.45	-31.0	7.66	7.88
	V238A	-13.41	-29.59	8.43	7.74
	I276T	-10.83	-29.80	9.47	9.49

3.8. Analysis of the tunnels

Analysis of CYP3A5 tunnels allows for the evaluation of the effect of the dynamics on the drug pathways that facilitate the transition of the drug to the catalytic site. Figure 11 shows first four clusters identified by Caver 3.0 according to their priority values – the higher the priority the more important is the tunnel. Indeed, all the simulated systems show a distinguished set of tunnels calculated from the microsecond level simulation. In general, the opening gates to the catalytic site are located either in the B/C loop side or between the F/G loop, the 477-483 segment, and the 49-57 segment (upper entry). The distribution of the tunnels differs considerably between the variants. For instance, while the wild type form has a single gate, G1, located near the B/C loop, variants R28C, Y53C, L82R I149T, and I276T have multiple gates. Moreover, Y53C has a single gate at the upper entry contrary to the wild type and other variants that possess more than one gate. At the quantitative level, we also noticed that there is no favourite common gate between the evaluated variants. For instance, the highest priority P_{C1} is equal to 0.64 and corresponds to the G4 gate in the wild type form, while P_{C1} is equal to 0.65 in I149T and corresponds to G5 gate located nearby the B/C loop.

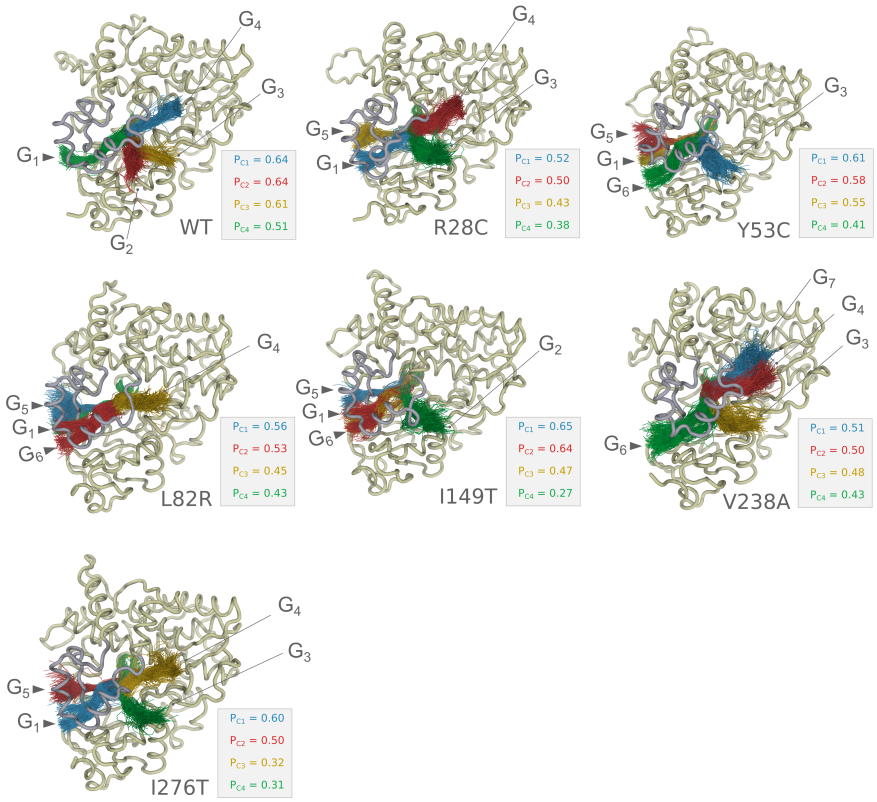

Figure 11: Analysis of tunnels for CYP3A5 variants. for each variant, we show tunnels with the best priority values encoded by the colours as follows: blue for the first rank, red for the second rank, yellow for the third rank, and green for the fourth rank. We also show the position of the gates for each variant.

4. Discussion

Understanding the pharmacogenetic landscape within a particular gene is a key step for establishing targeted clinical interventions in precision medicine. Variant interpretation, however, is still a limiting factor to acquire a better insight into the genotype-phenotype relationship and its implication in an individual's responses to drugs. This is mainly the cause of experimental difficulties in collecting data at the molecular level of the gene product. Thus, multiscale modelling is a valuable tool to predict the molecular impact of genetic diversity at the protein level and to understand the molecular mechanisms of the phenotype expression.

None of the studied variants induces a large structural drift compared to the wild-type form even after 1.5 μs of simulation including Y53C which has been shown to reduce the activity of CYP3A5. This is unexpected since the mapping of the variants on the protein structure suggests that some of the variable amino acids might indeed have a destabilising effect such as the case of L82R representing a substitution of a buried hydrophobic residue with a charged amino acid. Indeed, this is consistent with previous results suggesting the high level of structural conservation within the Cytochromes P450 superfamily [21]. It appeared however that most of the divergence between the variants is mainly affected by the B/C loop and F/G loop dynamics as shown by the local fluctuations.

We, therefore, hypothesized that the main impact of the genetic variability affects the plasticity of different structural elements that are critical for different steps of the catalysis process. These include the controlling of the entry of the drug to the channel that leads to the catalytic binding site, the binding to the catalytic pocket, and the transport process within the enzyme.

The evidence presented here shows that the genetic variability of CYP3A5 affects the structural plasticity of the protein specifically the region spanning the B/C helices, the F/G helices, and the upper roof segments relative to the heme group. Different modes of dynamics for the F/G loop might have an impact on the way each form of CYP3A5 responds to the triggering of the catalysis process. Diversification in the low-frequency dynamic modes revealed by essential dynamics analysis might be the consequence of the different observed correlated motion patterns. These outcomes result in a significant effect to shift the dynamic equilibrium of CYP3A5. This was supported by the establishment of the free energy landscape using the two first major PCs as reaction coordinates. The major conformational events revolve around the structure at the global minimum. Therefore, the different topologies of the FEL, as well as the divergent conformations of the structure at the global minima, revealed that for each of the CYP3A5 variants, different levels of energy barriers have to be crossed by the protein to fulfill its function. Considering all that, we also assume that the conformational properties, encoded by the

genetic information [25] could be responsible for diversifying the repertoire of drugs that can be processed by CYP3A5.

The genetic diversity affects the binding of the enzyme to different drugs as shown from the estimation of the binding free energy of ritonavir and artemether. In such regard, we believe that we have obtained reliable results based on previous enzyme assays where the K_i value was estimated at 0.6-0.8 μM for ritonavir/CYP3A5 [34]. The equivalence in Gibbs free energy is -8.4 to $-8.3\,\mathrm{kcal\,mol^{-1}}$ which is very close to what we have obtained for the wild type variant ($-8.72\,\mathrm{kcal\,mol^{-1}}$). Moreover, in addition to the genetic determinant, the binding capacity of CYP3A5 appears to be drug-dependent. The strong polynomial correlation that we have noticed for ritonavir between the estimated free energy of binding and the pocket volume might be explained as follows. When the pocket volume is very low to optimally fit the ligand, atomic clashes may occur which renders the interaction less favourable. At specific ranges of pocket volume, shape complementarity is more favourable and ritonavir would fit properly, therefore the steric clashes are less severe and the interaction energy is more favourable. When the volume increases too much, even with the presence of the drug, the Van der Waal's interactions would not be satisfied which would penalise the interaction energy. Such a model could explain the behaviour of bulky ligands where the stability depends on the flexible segments surrounding the catalytic pocket including the upper roof of the heme group and F/G loop. Artemether however does not seem to be impacted by the flexibility of these segments when interacting with the heme group because the drug is less bulky and the steric penalty would be less relevant.

The investigation of the major active tunnels for each variant has revealed the potential implication of the genetic variability in determining the access to the catalytic site. Although the access points to the tunnels on the surface of the protein can be broadly positioned at two sites, i.e, the upper roof segments nearby the F/G loop and the B/C loop, which are almost noticed for all the variants, they, however, implicate different sets of amino acids at each gate provided with different chemical properties. Therefore, it affects the selectivity of the drug and the kinetic of association to the protein. Similar conclusions were obtained previously from molecular dynamics simulations of CYP2D6 variants [2].

R28C and Y53C variants have been characterised functionally using enzymatic assays following expression in *Escherichia coli* and purification protocols. They were found to decrease CYP3A5 catalytic activities for testosterone and Nifedipine respectively [16, 17]. Nevertheless, we have not detected an obvious structural deviation during the 1.5 µs simulation but only significant alterations of the dynamics. More specifically for R28C, we have detected a significant reduction of catalytic pocket volume. This suggests that

the functional implication of this variant might also be related to its capacity of inducing long-distance structural modulation of CYP3A5 unrelated to the membrane-binding function as suggested previously [16]. The R28 side chain is involved in stabilizing the local structure by establishing contacts with T27 and Q77 which would now be unreachable with the substitution to cysteine due to the short side chain.

The L82R variant has not been successfully expressed in *Escherichia coli* and the authors suggested that the lack of stability as the main reason for that [16]. However, we were not able to detect any major structural re-arrangement but a significant variation of the major modes of dynamics. It might be possible that the folding process of CYP3A5 in the expression system misbehaved and such phenomena are often observed in *E.coli* [8]. L82R might seem to be a highly deleterious variant resulting from the burial of a charged residue in the core of the protein. However, a frequency of 2% has been detected in ethnically diverse samples from the USA that cause the same missense substitution L82R [16] which suggests a possible fixation in agreement with our findings.

Additionally, V238A (corresponding to variant rs542523237) appears to have the most significant effect on CYP3A5 because of its location on the F/G loop segment. The variant seems to affect the allostery of the protein, the pocket volume, and ritonavir binding.

Conclusions

We have showed, in our work, that the genetic variability of *CYP3A5* result in a significant impact on the functional properties of the protein that include both plasticity-related and drug-binding related characteristics. It is, therefore, inconvenient to perceive the clinical impact as associated with deleterious or non-deleterious genotype in the context of pharmacogenes. Our study highlights also the drug dependant phenotype relationship rather than the global effect related to the genotype. All the analyzed variants are rare, yet the heterogeneity of conformational properties revealed a significant effect. At least at the level of CYP3A5, we underlined the importance of the rare variants in determining the pharmacogenomics properties in populations from sub-Saharan Africa. It is also expected for other ADME proteins, particularly those in the cytochrome P450 family, to demonstrate the same properties. Consequently, it seems that the "genetic diversity bottleneck for precision medicine" is indeed a standing issue. It is still unclear, however, if the genetic variability will manifest into a significant clinical effect and the presented results for this paper should motivate for a complete functional analysis of these variants. In the upcoming years, the expected drop in sequencing cost will lead to a better characterisation of rare variants for ADME genes. As a result, it will be

difficult to overlook the clinical implication of rare variants as key determinants of the pharmacogenomic landscape.

Author contributions

HO coordinated and designed the study, developed the code related to this study, contributed to the analysis, and led the writing of the paper. JEBDR contributed to the analysis of the molecular dynamics simulation. SH acquired the funding for this study. All authors contributed to writing the manuscript, and read and agreed to the published version of the manuscript.

funding

This work was primarily funded through a grant by GlaxoSmithKline Research & Development Ltd to the Wits Health Consortium. JEBDR was partially funded by the South African National Research Foundation (SFH170626244782).

Informed consent

No new data was generated specifically for this project – this is a secondary analysis of data that had been generated and studied for other purposes. Genomic data were obtained from different collaborators. All participants have provided their written consent and ethical approval was obtained from qualified local committees of countries involved in this study. More details about the ethical approval are provided by da Rocha *et al.* [7] `https://doi.org/10.3389/fphar.2021.634016`.

Data Availability

All the raw data and the code used to make the analysis and figures of this paper are available online from the Zenodo repository under the DOI 10.5281/zenodo.4548257.

Acknowledgments

The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project.

Conflicts of interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] A. Amadei, A. B. Linssen, and H. J. Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–425, Dec 1993.

[2] Y. V. Ariza M?rquez, I. Brice?o, F. Aristiz?bal, L. F. Ni?o, and J. Yosa Reyes. Dynamic Effects of CYP2D6 Genetic Variants in a Set of Poor Metaboliser Patients with Infiltrating Ductal Cancer Under Treatment with Tamoxifen. *Sci Rep*, 9(1):2521, 02 2019.

[3] R. K. Bains. African variation at Cytochrome P450 genes: Evolutionary aspects and the implications for the treatment of infectious diseases. *Evol Med Public Health*, 2013(1):118–134, Jan 2013.

[4] D. A. Case, I. Ben-Shalom, S. Brozell, D. Cerutti, I. T.E. Cheatham, V. Cruzeiro, T. Darden, R. Duke, D. Ghoreishi, M. Gilson, H. Gohlke, A. Goetz, D. Greene, R. Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R. Walker, J. Wang, H. Wei, R. Wolf, X. Wu, L. Xiao, D. York, and P. Kollman. *AMBER 2018*. University of California, San Francisco, 2018.

[5] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, 2015.

[6] E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor, and J. Damborsky. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol*, 8(10):e1002708, 2012.

[7] J. E. B. da Rocha, H. Othman, G. Botha, L. Cottino, D. Twesigomwe, S. Ahmed, B. I. Drögemöller, F. M. Fadlelmola, P. Machanick, M. Mbiyavanga, S. Panji, G. E. B. Wright, C. Adebamowo, M. Matshaba, M. Ramsay, G. Simo, M. C. Simuunza, C. T. Tiemessen, S. Baldwin, M. Chiano, C. Cox, A. S. Gross, P. Thomas, F.-J. Gamo, and S. Hazelhurst. The extent and impact of variation in adme genes in sub-saharan african populations. *Frontiers in Pharmacology*, 12:366, 2021.

[8] A. de Marco, L. Vigh, S. Diamant, and P. Goloubinoff. Native folding of aggregation-prone recombinant proteins in Escherichia coli by osmolytes, plasmid- or benzyl alcohol-overexpressed molecular chaperones. *Cell Stress Chaperones*, 10(4):329–339, 2005.

[9] L. Duan, X. Liu, and J. Z. Zhang. Interaction entropy: a new paradigm for highly efficient and reliable computation of protein–ligand binding free energy. *Journal of the American Chemical Society*, 138(17):5722–5728, 2016.

[10] C. S. Ernest, S. D. Hall, and D. R. Jones. Mechanism-based inactivation of CYP3A by HIV protease inhibitors. *J Pharmacol Exp Ther*, 312(2):583–591, Feb 2005.

[11] E. B. Esu, E. E. Effa, O. N. Opie, and M. M. Meremikwu. Artemether for severe malaria. *Cochrane Database Syst Rev*, 6:CD010678, 06 2019.

[12] M. H. Hsu and E. F. Johnson. Active-site differences between substrate-free and ritonavir-bound cytochrome P450 (CYP) 3A5 reveal plasticity differences between CYP3A5 and CYP3A4. *J. Biol. Chem.*, 294(20):8015–8022, 05 2019.

[13] M. H. Hsu, U. Savas, and E. F. Johnson. The X-Ray Crystal Structure of the Human Mono-Oxygenase Cytochrome P450 3A5-Ritonavir Complex Reveals Active Site Differences between P450s 3A4 and 3A5. *Mol. Pharmacol.*, 93(1):14–24, 01 2018.

[14] W. Humphrey, A. Dalke, K. Schulten, et al. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[15] J. Lamba, J. M. Hebert, E. G. Schuetz, T. E. Klein, and R. B. Altman. PharmGKB summary: very important pharmacogene information for CYP3A5. *Pharmacogenet Genomics*, 22(7):555–558, Jul 2012.

[16] S. J. Lee, K. A. Usmani, B. Chanas, B. Ghanayem, T. Xi, E. Hodgson, H. W. Mohrenweiser, and J. A. Goldstein. Genetic findings and functional studies of human CYP3A5 single nucleotide polymorphisms in different ethnic groups. *Pharmacogenetics*, 13(8):461–472, Aug 2003.

[17] S. J. Lee, I. P. van der Heiden, J. A. Goldstein, and R. H. van Schaik. A new CYP3A5 variant, CYP3A5*11, is shown to be defective in nifedipine metabolism in a recombinant cDNA expression system. *Drug Metab Dispos*, 35(1):67–71, Jan 2007.

[18] Y. T. Lee, R. F. Wilson, I. Rupniewski, and D. B. Goodin. P450cam visits an open conformation in the absence of substrate. *Biochemistry*, 49(16):3412–3419, Apr 2010.

[19] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015.

[20] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The Ensembl Variant Effect Predictor. *Genome Biol.*, 17(1):122, 06 2016.

[21] J. Mestres. Structure conservation in cytochromes P450. *Proteins*, 58(3):596–609, Feb 2005.

[22] B. R. Miller III, T. D. McGee Jr, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. Mmpbsa.py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation*, 8(9):3314–3321, 2012.

[23] P. C. E. Moody and E. L. Raven. The Nature and Reactivity of Ferryl Heme in Compounds I and II. *Acc Chem Res*, 51(2):427–435, 02 2018.

[24] G. Mustafa, P. P. Nandekar, G. Mukherjee, N. J. Bruce, and R. C. Wade. The Effect of Force-Field Parameters on Cytochrome P450-Membrane Interactions: Structure and Dynamics. *Sci Rep*, 10(1):7284, 04 2020.

[25] R. Nussinov, H. Jang, C. J. Tsai, and F. Cheng. Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. *PLoS Comput Biol*, 15(3):e1006658, 03 2019.

[26] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen. Propka3: Consistent treatment of internal and surface residues in empirical pka predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, 2011.

[27] S. Rendic and F. P. Guengerich. Survey of Human Oxidoreductases and Cytochrome P450 Enzymes Involved in the Metabolism of Xenobiotic and Natural Chemicals. *Chem Res Toxicol*, 28(1):38–42, Jan 2015.

[28] J. D. Robarge, L. Li, Z. Desta, A. Nguyen, and D. A. Flockhart. The star-allele nomenclature: retooling for translational genomics. *Clin Pharmacol Ther*, 82(3):244–248, Sep 2007.

[29] D. R. Roe and T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput*, 9(7):3084–3095, Jul 2013.

[30] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234(3):779–815, Dec 1993.

[31] P. Schmidtke, A. Bidon-Chanal, F. J. Luque, and X. Barril. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, 27(23):3276–3285, Dec

2011.

[32] K. Shahrokh, A. Orendt, G. S. Yost, and T. E. Cheatham. Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J Comput Chem*, 33(2):119–133, Jan 2012.

[33] M. Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, 15(11):2507–2524, Nov 2006.

[34] M. G. Soars, K. Grime, and R. J. Riley. Comparative analysis of substrate and inhibitor interactions with CYP3A4 and CYP3A5. *Xenobiotica*, 36(4):287–299, Apr 2006.

[35] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case. Continuum solvent studies of the stability of dna, rna, and phosphoramidate- dna helices. *Journal of the American Chemical Society*, 120(37):9401–9409, 1998.

[36] O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, Jan 2010.

[37] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*, 25(2):247–260, Oct 2006.

[38] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–1174, Jul 2004.

[39] Y. Zhou, S. Mkrtchian, M. Kumondai, M. Hiratsuka, and V. M. Lauschke. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics J.*, 19(2):115–126, 04 2019.