

# EFFECTS OF NUCLEOTIDE VARIATION ON THE STRUCTURE AND FUNCTION OF HUMAN ARYLAMINE N-ACETYLTRANSFERASE 1

By

WISDOM ALEMYA AKURUGU

A thesis presented in fulfillment of the requirements of *Magister Scientiae* at the South African National Bioinformatics Institute, University of the Western Cape.

UNIVERSITY of the  
WESTERN CAPE

05/01/2012

Supervisor: Prof. Alan Christoffels



UNIVERSITY of the  
WESTERN CAPE



SANBI  
South African National  
Bioinformatics Institute

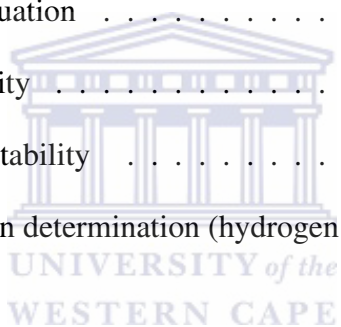
# Contents

<b>TABLE OF CONTENTS</b>	<b>i</b>
<b>LIST OF FIGURES</b>	<b>v</b>
<b>LIST OF TABLES</b>	<b>vii</b>
<b>LIST OF ABBREVIATIONS</b>	<b>viii</b>
<b>KEYWORDS</b>	<b>x</b>
<b>ABSTRACT</b>	<b>xi</b>
<b>DECLARATION</b>	<b>xiii</b>
<b>ACKNOWLEDGEMENTS</b>	<b>xiv</b>
<b>1 INTRODUCTION AND LITERATURE REVIEW:</b>	<b>1</b>
1.1 Arylamine <i>N</i> -acetyltransferases (NATs) . . . . .	1
1.2 Tuberculosis . . . . .	3
1.3 Single nucleotide polymorphisms (SNPs) . . . . .	5
1.4 Polymorphisms of human NAT1 . . . . .	6
1.5 Functions of human NAT1 and NAT2 . . . . .	7
1.6 Protein structures and residue interactions . . . . .	7



1.6.1	Residue interactions . . . . .	9
1.6.2	The secondary structure . . . . .	9
1.6.3	Tertiary and quaternary structures . . . . .	11
1.6.4	Structural features of human NATs . . . . .	11
1.7	Project rationale and objectives . . . . .	14
1.7.1	Rationale . . . . .	14
1.7.2	Aim and Objectives . . . . .	16
<b>2</b>	<b>MATERIALS AND METHODS</b>	<b>17</b>
2.1	MATERIALS . . . . .	17
2.1.1	Sorting Intolerant From Tolerant (SIFT) prediction server . . . . .	17
2.1.2	Polymorphism phenotyping version 2 (POLYPHEN-2) prediction server . . . . .	18
2.1.3	CLUSTALW2 multiple sequence alignment server . . . . .	19
2.1.4	Homology modelling and MODELLER software . . . . .	20
2.1.5	PROCHECK Software . . . . .	21
2.1.6	UCSF CHIMERA . . . . .	21
2.1.7	FOLDX Software . . . . .	22
2.1.8	NACCESS Software . . . . .	22
2.2	METHODS . . . . .	23
2.2.1	Sequence Data Acquisition . . . . .	23
2.2.1.1	SNPs . . . . .	23
2.2.1.2	Proteins . . . . .	25
2.2.2	SIFT and POLYPHEN-2 analysis . . . . .	26
2.2.3	Sequence conservation analysis . . . . .	26
2.2.4	Structural analysis . . . . .	26
2.2.4.1	Homology modelling . . . . .	27

2.2.4.2	Evaluation of modelled structures . . . . .	27
2.2.4.3	Residue interaction determination . . . . .	27
2.2.4.4	Protein structure stability calculation . . . . .	27
2.2.4.5	Solvent accessibility calculation . . . . .	28
<b>3</b>	<b>RESULTS AND DISCUSSION</b>	<b>29</b>
3.1	SIFT and POLYPHEN-2 analysis . . . . .	29
3.2	Sequence conservation profile between human and prokaryotic NAT1 enzymes . . . . .	33
3.3	Structural analysis . . . . .	39
3.3.1	Homology modelling of NAT1 . . . . .	39
3.3.2	PROCHECK evaluation . . . . .	42
3.3.3	Solvent accessibility . . . . .	43
3.3.4	Protein structure stability . . . . .	44
3.3.5	Residue interaction determination (hydrogen bonds and salt bridges) . . . . .	46
<b>4</b>	<b>CONCLUSION</b>	<b>63</b>
	<b>Bibliography</b>	<b>64</b>
	<b>Appendix</b>	<b>78</b>
<b>A</b>	<b>Command line options</b>	<b>78</b>
A.1	UCSF CHIMERA . . . . .	78
A.2	FOLDX . . . . .	78
A.3	NACCESS . . . . .	79
<b>B</b>	Residue interactions of wild-type residue R64 and SNP residue W64. . . . .	<b>80</b>
<b>C</b>	Residue interactions of wild-type residue V149 and SNP residue I149 . . . . .	<b>81</b>





<b>D</b>	Residue interactions of wild-type residue R187 and SNP residue Q187 . . . . .	<b>82</b>
<b>E</b>	Residue interactions of wild-type residue I263 and SNP residue V263 . . . . .	<b>83</b>
<b>F</b>	Residue interactions of wild-type residue M205 and SNP residue V205. . . . .	<b>84</b>
<b>G</b>	Residue interactions of wild-type residue S214 and SNP residue A214. . . . .	<b>85</b>
<b>H</b>	Residue interactions of wild-type residue D251 and SNP residue V251. . . . .	<b>86</b>
<b>I</b>	Residue interactions of wild-type residue E261 and SNP residue K261. . . . .	<b>87</b>



# List of Figures

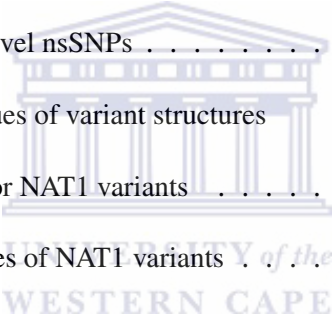
1.1	Protein coding exons of human NAT1 and NAT2 . . . . .	2
1.2	NAT1 active site . . . . .	3
1.3	Estimated incidence of all forms of TB, classified by WHO Regions, 2009 . . . . .	4
1.4	Functionally important residues of NAT1 protein . . . . .	13
2.1	Methodology flow chart . . . . .	25
3.1	Multiple sequence alignment of related NAT1 enzymes from positions 1 to 140 . . . . .	34
3.2	Multiple sequence alignment of related NAT1 enzymes from positions 141 to 288 . . . . .	35
3.3	DOPE plots for published NAT1 variants . . . . .	41
3.4	DOPE plots for novel NAT1 variants . . . . .	42
3.5	Residue interactions involving wild-type residue R242 and SNP residue M242 . . . . .	47
3.6	Residue interactions involving wild-type residue N245 and SNP residue I245 . . . . .	48
3.7	Residue interactions involving wild-type residue S259 and SNP residue R259 . . . . .	49
3.8	Residue interactions involving wild-type residue E264 and SNP residue K264 . . . . .	50
3.9	Residue interactions involving wild-type residue R117 and SNP residue T117 . . . . .	51
3.10	Residue interactions involving wild-type residue R166 and SNP residue T166 . . . . .	52
3.11	Residue interactions involving wild-type residue E167 and SNP residue Q167 . . . . .	53
3.12	Residue interactions involving wild-type residue T193 and SNP residue S193 . . . . .	54
3.13	Residue interactions involving wild-type residue F202 and SNP residue V202 . . . . .	55
3.14	Residue interactions involving wild-type residue Q210 and SNP residue P210 . . . . .	57

3.15	Residue interactions involving wild-type residue D229 and SNP residue H229 . . . . .	58
3.16	Residue interactions involving wild-type residue V231 and SNP residue G231 . . . . .	59
3.17	Residue interactions involving wild-type residue V235 and SNP residue A235 . . . . .	60
3.18	Residue interactions involving wild-type residue T240 and SNP residue S240 . . . . .	61



# List of Tables

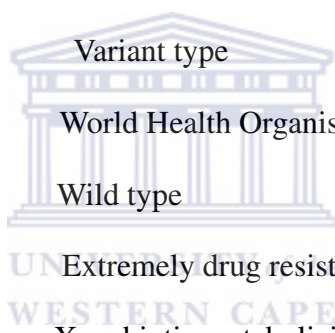
1.1	Experimental functional analysis of human NAT1 SNPs . . . . .	8
2.1	List of nsSNPs and the associated amino acid changes . . . . .	24
3.1	Predicted effects of NAT1 published nsSNPs . . . . .	30
3.2	Predicted effects of NAT1 novel nsSNPs . . . . .	31
3.3	DOPE score and molpdf values of variant structures . . . . .	40
3.4	Ramachandran plot results for NAT1 variants . . . . .	43
3.5	Relative solvent accessibilities of NAT1 variants . . . . .	44



## LIST OF ABBREVIATIONS

CoA	Co-enzyme A
DNA	Deoxyribonucleic acid
DOPE	Discrete optimised potential energy
eSNP	expression Single Nucleotide Polymorphism
H-bond	Hydrogen bond
MDR-TB	Multi Drug Resistant Tuberculosis
mRNA	Messenger Ribonucleic Acid
<i>NAT1</i>	N-Acetyltransferase 1
<i>NAT2</i>	N-Acetyltransferase 2
NATP	N-acetyltransferase pseudogene
NBRF	National biomedical research fund
NMR	Nuclear magnetic resonance
NRF	National research fund
nsSNPs	Non-synonymous Single Nucleotide Polymorphisms
p-ABA	Para-aminobenzoic acid
p-ABAGLU	Para-aminobenzoyl glutamate
p-AS	Para-aminosalicylic acid
PDB	Protein Data Bank
PIR	Protein Information Resource
POLYPHEN-2	Polymorphism phenotyping version 2

PSI-BLAST	Position Specific Iterated Basic Local Alignment Search Tool
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SASA	Solvent Accessible Surface Area
SBP	Substrate binding pocket
SNPs	Single Nucleotide Polymorphisms
SIFT	Sorting Intolerant From Tolerant
TB	Tuberculosis
UCSF	University of California San Francisco
UTRs	3'-untranslated regions
VT	Variant type
WHO	World Health Organisation
WT	Wild type
XDR-TB	Extremely drug resistant tuberculosis
XME	Xenobiotic metabolizing enzyme



## KEYWORDS

Homology Modelling

NAT1

NAT2

N-acetyltransferases

Sequence conservation

Single nucleotide polymorphisms

Tuberculosis

Xenobiotic-metabolizing enzymes



## ABSTRACT

The human arylamine N-acetyltransferase 1 (NAT1) is critical in determining the duration of action and pharmacokinetics of amine-containing drugs such as para-aminosalicylic acid and para-aminobenzoyl glutamate used in clinical therapy of tuberculosis (TB), as well as influencing the balance between detoxification and metabolic activation of these drugs. SNPs in this enzyme are continuously being detected and indicate inter-ethnic and inter-individual variation in the enzyme function. The effect of nsSNPs on the structure and function of proteins are routinely analyzed using SIFT and POLYPHEN-2 prediction algorithms. The false-negative rate of these two algorithms results in as much as 25% of nsSNPs. This study aimed to explore the use of homology modeling including residue interactions, Gibbs free energy change and solvent accessibility as additional evidence for predicting nsSNP effects on enzyme function. This study evaluated the functional effects of 14 nsSNPs identified in a South African mixed ancestry population of which 3 nsSNPs were previously identified in Caucasians. The SNPs were evaluated using structural analysis that included homology modeling, residue interactions, relative solvent accessibility, Gibbs free energy change and sequence conservation in addition to the routinely used nsSNP function prediction algorithms, SIFT and POLYPHEN-2. The structural analysis implemented in this study showed a loss of hydrogen bonds for S259R thereby affecting protein function which contradicts predictions obtained from SIFT and POLYPHEN-2 algorithms. The variant N245I was shown to be neutral but contradicted the predictions from SIFT and POLYPHEN-2. Structural analysis predicted that variant R242M would affect protein stability and therefore NAT1 function in agreement with POLYPHEN-2 predictions but contradicting predictions from SIFT. No structural changes were expected for variant E264K in agreement with predictions obtained from POLYPHEN-2 but contradicting results from SIFT. The functions of the remaining 10 nsSNPs were consistent with those predicted by SIFT and POLYPHEN-2 namely



that four variants R117T, E167Q, T193S and T240S do not affect the NAT1 function whereas R166T, F202V, Q210P, D229H, V231G and V235A could affect the enzyme function.

This study provided the first evaluation of the functional effects of 11 newly characterized nsSNPs on the NAT1 tuberculosis drug-metabolizing enzyme. The six functionally important nsSNPs predicted by all three methods and the four SNPs with contradictory results will be tested experimentally by creating a SNP construct that will be cloned into an expression vector. These combined computational and experimental studies will advance our understanding of NAT1 structure-function relationships and allow us to interpret the NAT1 genetic polymorphisms in individuals who are slow or fast acetylators. The results, albeit a small dataset demonstrate that the routinely used algorithms are not without flaws and that improvements in functional prediction of nsSNPs can be obtained by close scrutiny of the molecular interactions of wild type and variant amino acids.



## DECLARATION

I declare that *Effects of nucleotide variation on the structure and function of human arylamine N-acetyltransferase 1* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.



Wisdom Alemya Akurugu

Sign:

05/01/2013

## ACKNOWLEDGEMENTS

The Good Lord has seen me through this journey in good health and state of mind. To God be the glory and honour now and forever.

I am forever grateful to my supervisor and mentor, Professor Alan Christoffels for his guidance, support, tutorship and the opportunity given me to go through this course of education. My gratitude will forever remain with you Alan, for believing in me and teaching me independence of thought and work. I will jealously protect these values. To Dr. Cedric Werely, Department of Health, Tygerberg Hospital and Stellenbosch University, I appreciate so much your assistance and collaboration especially in providing me with the SNPs. Thanks too for your corrections and comments on this thesis. To Mr. Lucas Amengatigo of the Navrongo Health Research Centre Ghana, I say thank you for your immense contribution to my education. I am appreciative of your mentorship and assistance.

Thanks to the National Research Fund (NRF) of South Africa for providing me the funds to conduct this research in SANBI. The knowledge gained during this training, I believe would one day contribute positively to the development of research in South Africa and Africa as a whole.

Many thanks to Dr Uljana Hesse and Dr Sumir Panji for taking time to go through my thesis. I am most grateful for your comments and corrections. To the SANBI administration and IT staff Ferial Mullins, Maryam Salie, Peter van Huesden, Dale Gibbs, Samantha, Fungi, as well as Junita, thank you all for your assistance. To my friends, Dr. Natasha, Dr. Picone, Mr Mario, Mbandi, Mushal, Darlington, Minah, Fred, Abraham, Emad, Emil, Adugna and Ram for being available in need. Special mention is made of Ruben Cloete for his initial and continuous support for my work. To Dr. Samuel Kojo Kwofie, I say thank you and “ayekoo” for all you have done for me through thin and thick. You were always there when most needed. To all members of the Christofells’ Lab, thank you for your comments and support.

Finally, I am most grateful to my family: wife, Juliet, son, Lord-Michael, mum, Mrs M. A. Akayore Akurugu, late dad, Mr C. Akawologo Akurugu, siblings; Stephen, Robert, Emmanuel, Eden, Lambert, Cecilia and the entire Akabise's family. You have all supported me in diverse ways.



# Chapter 1

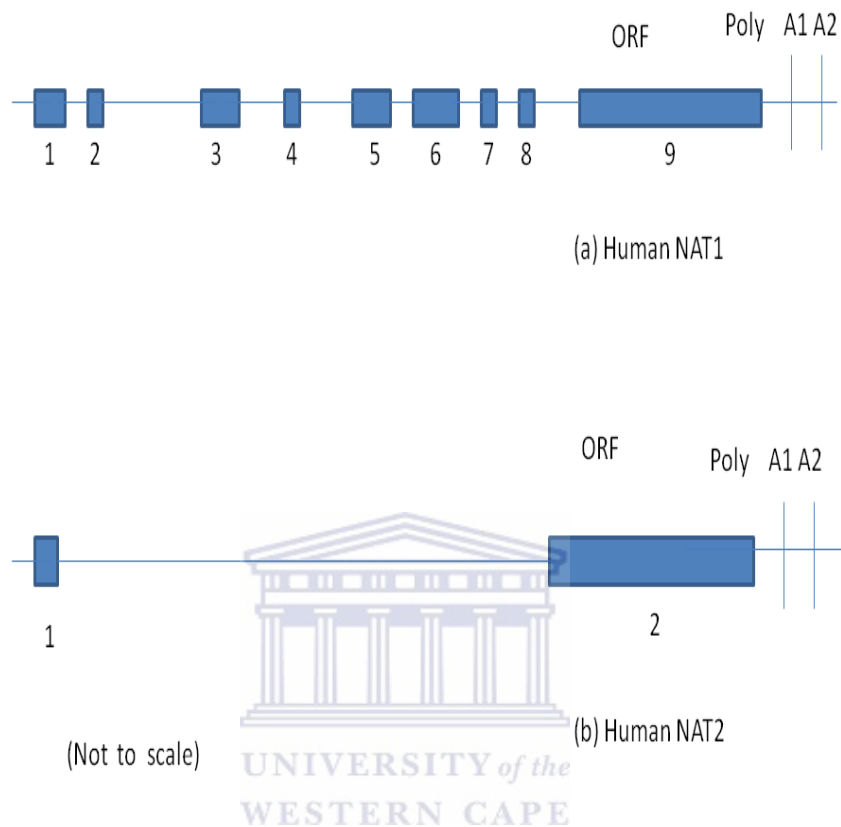
## INTRODUCTION AND LITERATURE REVIEW:

### 1.1 Arylamine *N*-acetyltransferases (NATs)

Arylamine *N*-acetyltransferases (NATs) are xenobiotic metabolizing enzymes (XMEs) that affect the biological activity and toxicity of compounds, including tuberculosis (TB) drugs. These enzymes are found in both prokaryotes and eukaryotes [1]. NATs catalyze either the acetyl-CoA dependent *N*-acetylation of primary aromatic amines and hydrazines (usually deactivating the drug) or the *O*-acetylation of their *N*-hydroxylated metabolites (usually activating the drug). Specifically, NATs transfer acetyl group from acetyl-CoA to the nitrogen or oxygen atom of primary aromatic amines, hydrazines and *N*-hydroxylated metabolites of such compounds [2]. Therefore, NATs play a vital role in detoxification and potential metabolic activation of numerous xenobiotic substances.

Human NAT1 and NAT2, (33kD and 31kD proteins respectively) are products of protein-coding exons of 870-bp open reading frames encoding 290 amino acids [3, 4, 5, 6]. So far NAT genes have been identified in 1674 prokaryotic and 464 eukaryotic genomes [1] and encode between 254 and 332 amino acids [7]. These genes together with a pseudogene, *NATP*, are located on human chromosome 8p22 [3, 8]. The *NAT* loci are separated by 170-360kb and are positioned *NAT1* to *NATP* to *NAT2*, where *NAT1* is located on the centromeric side of marker D8S261 and *NAT2* near marker D8S21 [9]. An 87% nucleotide sequence identity is shared in the coding region between *NAT1* and *NAT2* or an 81% identity at the amino acid level resulting in 55 amino acid differences. The open reading frames of both enzymes are shown in Figure

1.1.



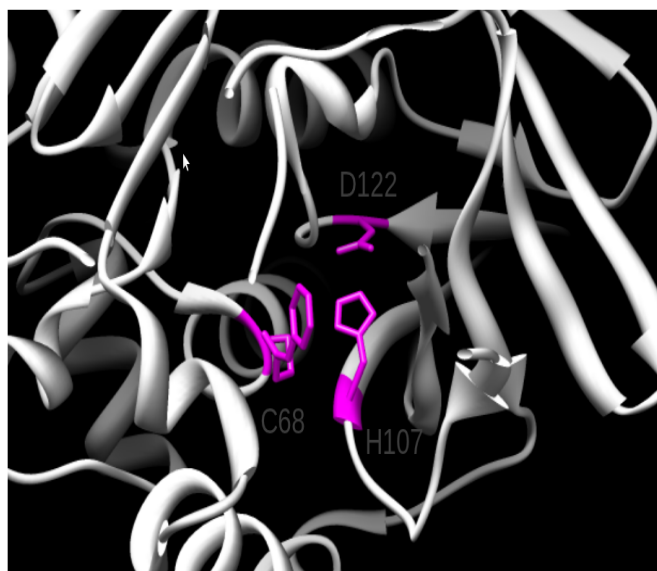
**Figure 1.1** Protein coding exons of human NAT1 and NAT2

ORF- Open reading frame.

*NAT1* protein is encoded by a single exon (9) whereas *NAT2* mRNA is obtained from both the protein-coding exon (2) and a second non-coding exon of 100-bp located about 8 kb upstream of the translation start site (1) [3, 10]

A multiple sequence alignment of NAT proteins indicates that most conserved regions occur at the amino terminus, whereas the carboxyl terminus displays little conservation between the species [7]. All NATs possess the conserved C68/C69, H107 and D122 residues suggested to form a catalytic triad (Figure 1.2). Inhibitor [11, 12] and site-directed mutagenesis studies [13] confirm that the C68 in human proteins is crucial for NAT activity. In humans, the two functional NATs, NAT1 and NAT2, are both expressed in liver [14] while only NAT1 is expressed in mononuclear leukocytes [15]. Human NAT1 catalyses the metabolism of para-aminosalicylic acid [16] used in the treatment of tuberculosis; one of the most serious

lung diseases in the world.



**Figure 1.2** NAT1 active site

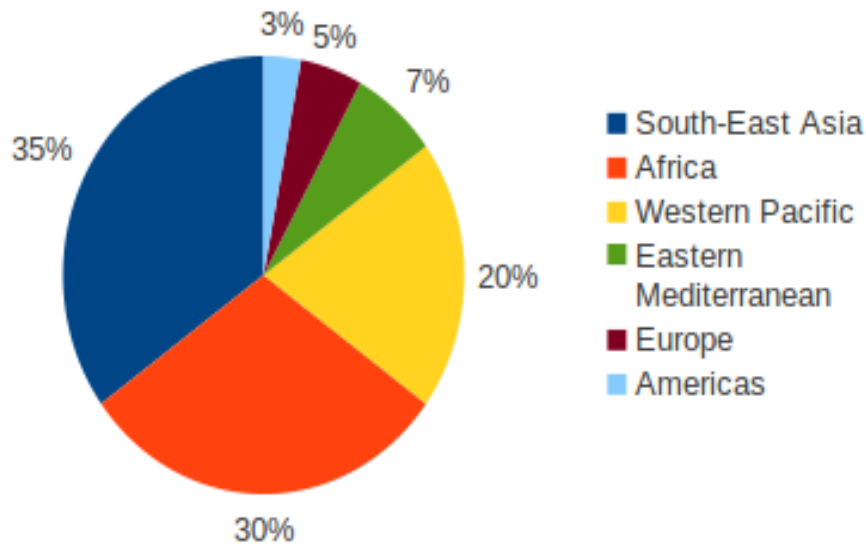
Catalytic triad residues C68, H107 and D122 are highlighted in pink.

## 1.2 Tuberculosis



According to the 2010 World Health Organization (WHO) Report [17], in 2009 there were an estimated 9.4 million incident cases of TB globally, implying 137 cases per 100000 humans. Most incidents were recorded in the South-East Asia, Africa and the Western Pacific (Figure 1.3).

With approximately 0.5 million cases, South Africa is among the five countries with the largest number of incidents. Here, the problem of TB is complicated by HIV/TB co-infections. HIV increases the susceptibility of a patient to infection with tuberculosis and the spread of HIV epidemic in the past decade has been accompanied by up to a fourfold increase in the number of TB cases [18]. It is estimated that up to 50% of new adult cases of TB in South Africa are co-infected with HIV [19]. These statistics indicate that tuberculosis remains a serious health problem affecting the world with devastating economic, social and financial burdens calling for effective measures to identify, prevent, control and treat tuberculosis.



**Figure 1.3** Estimated incidence of all forms of TB, classified by WHO Regions, 2009

Source: Global Tuberculosis Control; WHO Report 2010, World Health Organisation, Geneva 2010. WHO/HTM/2010.7

The treatment of TB falls into two phases: an initial intensive phase and a continuation phase or consolidation phase. WHO recommends an intensive phase of two months with isoniazid, rifampicin, pyrazinamide and ethambutol for all new cases and a continuation phase of four months with isoniazid and rifampicin [17]. In combination with streptomycin these drugs form what is known as the first-line drugs for TB treatment. A drug may be classified as second-line instead of first-line for various reasons, namely it is less effective than the first-line drug (e.g. para-aminosalicylic acid (p-AS)), it has toxic side-effects (e.g. cycloserine) or is unavailable in many developing countries (e.g. fluoroquinolones). Second line drugs include kanamycin, viomycin, ciprofloxacin, prothiomide, p-AS and a host of others. Until 2007, the standard treatment regimen in South Africa consisted of a four-month intensive phase with five anti-TB drugs (kanamycin, ethionamide, pyrazinamide, ofloxacin and cycloserine or ethambutol), followed by a 12-18 months continuation phase with three drugs (ethionamide, ofloxacin and cycloserine or ethambutol) [20]. These drugs were administered five times per week in out-patient clinics and seven times per week in hospitals.

However, an increasing number of TB infections have become resistant to the major anti-tuberculosis drugs [21, 22, 23]. Two types of drug resistance can be identified: multi-drug resistant TB (MDR-TB) and extensively drug resistant TB (XDR-TB). Multi-drug resistant TB is a form of tuberculosis that



fails to respond to standard first-line drugs while extensively drug-resistant TB occurs when resistance to second-line drugs develops in addition to MDR-TB. The identification of XDR-TB in Tugela Ferry, KwaZulu-Natal, South Africa in 2006, highlighted the inadequacies of the National TB Control program and emphasized the importance of infection control [24]. As a result of this outbreak, the treatment guidelines of South Africa were revised in 2007 [20]. Ethambutol was replaced by terizidone for the treatment of MDR-TB. To treat XDR-TB, amikacin/kanamycin, pyrazinamide and ofloxacin were replaced by capreomycin, p-AS and moxifloxacin. These changes were implemented based on the assumption that both capreomycin and p-AS have not been used extensively in South Africa and therefore resistance to these drugs should be rare. Yet, a national survey conducted in 2008 in South Africa revealed that 20.2% of all notified TB cases of that year (13000) showed resistance to isoniazid and nearly half of them (9.6% of all cases) were MDR-TB [25], indicating a 3-fold increase of MDR-TB cases since 2002 (3.1%). Furthermore, recent analyses of *M. tuberculosis* strains revealed mutations in the *inhA* promoter, *katG*, *rpoB*, *embB*, *pncA*, *rrs* and *gyrA* genes that caused resistance to isoniazid, ethionamide, rifampicin, ethambutol, pyrazinamide, streptomycin, amikacin, capreomycin and ofloxacin [26, 27, 28, 29, 30, 31]. These findings implied the emergence of totally drug resistant (TDR)-TB.

Thus far resistance to p-AS has not been reported making p-AS one of the most important anti-TB drugs in South Africa. The primary role of NATs in activating and/or deactivating a large and diverse number of aromatic amines and hydrazine drugs makes them important in clinical pharmacology and toxicology [32]. Human NAT1 catalyses the N-acetylation of p-AS used as part of the TB treatment regime in South Africa. This makes NAT1 an important enzyme of research in the fight to control and treat TB in South Africa.

### **1.3 Single nucleotide polymorphisms (SNPs)**

Research has shown that single nucleotide polymorphisms in genomic DNA of NAT1 can change the amino acid sequence [33, 34, 35, 36, 37]. This can destabilize the enzyme structure and negatively affect the function of NAT1. A point mutation or single base substitution is a mutation that replaces a single nucleotide with another, or leads to deletion or insertion of a base pair. Point mutations can be catego-

alized by the type of replacement or by function. For example purine-purine (or pyrimidine-pyrimidine) replacements are termed transitions. Transversions are replacements of a purine by a pyrimidine and vice versa. Two of three SNPs involve replacement of cytosine (C) with thymine (T) [38]. Functionally, point mutations can be defined as nonsense, non-synonymous and synonymous mutations. Nonsense mutations generate stop codons or a nonsense codon and produce a truncated, usually non-functional peptide chain. Non-synonymous mutations introduce a different amino acid, whereas synonymous mutations code for the same amino acid. Non-synonymous mutations may be either conservative, replacing the amino acid by a closely related one, or non-conservative, in which an amino acid is replaced by another with different properties.

Single nucleotide polymorphisms are point mutations that occur in at least 1% of a population. About 90% of sequence variations in humans involve differences in single nucleotides [38]. They occur every 100 to 300 bases in the human genome. SNPs may occur in coding regions (exons), non-coding regions (introns) or 3' and 5' untranslated regions (UTRs). SNPs outside of protein-coding regions may affect gene splicing, transcription factor binding or the sequence of non-coding RNA [39] and is referred to as an expression SNP (eSNP). Point mutations within the coding region, non-conservative non-synonymous mutations in particular, can affect the stability or folding of a protein and hence may change its function [40].

## **1.4 Polymorphisms of human NAT1**

The NAT1 isozyme was initially considered genetically invariant due to its substrate specificity for p-AS and other drugs. It has now been shown to be subject to polymorphisms [41, 42], which can affect the activity of NAT1 towards p-AS and para-aminobenzoic acid (p-ABA) [6, 14, 15, 43, 44, 45, 46]. Experimental laboratory analysis is the best method for the assessment of functional effects of nsSNPs. The effects of a number of human *NAT1* SNPs in the coding region and 3'-untranslated region have been explored experimentally (Table 1.1). *NAT1*\*4 has historically been designated "wild-type" because it is the most common occurring allele in some but not all ethnic groups (the designation of "wild-type" allele is somewhat arbitrary and is dependent upon the ethnicity of the population studied). Relative to

NAT1\*4, the reference allele, SNPs were found to increase or decrease NAT1 activity resulting in rapid or slow acetylation phenotypes.

## **1.5 Functions of human NAT1 and NAT2**

The human NAT enzymes perform both bioactivation and bioinactivation reactions [55]. Though both catalyse the acetylation of xenobiotics, they have a largely distinct substrate profile that overlap only to some degree. The differences in substrate specificity are due to interactions within the active site cleft [56] and the C-terminal region [48]. Residues 124-129 [16] together with the C-terminal domain are involved in interactions with substrates at the active site cleft (Figure 1.2).

Minchin [57] and Ward [58], demonstrated that human NAT1 participates in the metabolic breakdown of folate by acetylating the folate catabolite para-aminobenzoyl glutamate (p-ABAGLU). Other compounds like p-AS are also metabolised by human NAT1 [59]. Human NAT2 by contrast metabolizes the drugs hydrazide, hydralazine, phenelzine and arylamine drugs such as procainamide and sulphamethazine [60]. Despite differences in substrate selectivities, some substrates such as 2-aminofluorene are metabolised by both NAT1 and NAT2 enzymes [61] and NAT1 has been known to metabolise compounds primarily metabolised by NAT2 in slow acetylators [62].

Clinicians in South Africa are currently using p-AS as part of the treatment regime because of drug resistance in TB. It is important to expand the work on NAT1 given the paucity of data for NAT1 function relative to NAT2 and the fact that NAT1 metabolizes p-AS.

## **1.6 Protein structures and residue interactions**

The function and chemical properties of a protein are determined by its three-dimensional (3D) structure. This 3D-structure begins with a linear arrangement of the amino acids (its primary structure) and progresses through several protein folding steps creating secondary, tertiary and quaternary structures of the protein [63].

**Table 1.1** Experimental functional analysis of human NAT1 SNPs

Allele	Nucleotide change(s)	Amino acid change(s)	Effect on activity relative to NAT1*4	Reference
NAT1*10	1088T>A, 1095C>A	R187Q	2-fold higher	[47]
NAT1*10	1088T>A	R187Q	Slightly elevated	[48]
NAT1*14	560G>A, 1088T>A, 1095C>A	R187Q	Lower, 15 to 20 decrease	[49, 50, 51]
NAT1*11	459G>A; 640T>G	V149I, S214A	Lower	[50]
NAT1*3, NAT1*10	1088T>A, 1095C>A	R187Q	Similar to NAT1*4	[50]
NAT1*14B/NAT1*15	560G>A, 559C>T	R187Q	Lower	[49, 50, 51, 52]
NAT1*17	190C>T	R64W	Lower	[37]
NAT1*22	752A>T	D251V	Lower	[37]
NAT1*11	459G>A, 445G>A, 640T>G	T153T, V149I, S214A	Higher	[53]
NAT1*17, NAT1*22	190C>T, 613A>G	R64W, D251V	Reduced catalytic activity	[54]
NAT1*10	1088T>A, 1095C>A	R187Q	10-fold higher	[54]
NAT1*19, NAT1*15	97C>T, 559C>T	R33Stop, R187Stop	Reduced catalytic activity	[54]

For the column “Nucleotide change(s)” where more than one nucleotide change is seen means that they are known to occur in the corresponding alleles and have the amino acid changes indicated under column “Amino acid change(s)”.

### 1.6.1 Residue interactions

Protein amino acid interactions are mostly non-covalent and include the interactions discussed below. Hydrophobic interactions occur between non-polar residues and give the largest single contribution to protein stability [63]. Hydrogen bonds (H-bonds) also contribute to protein stability. Hydrogen bonds are formed when a positively charged hydrogen atom covalently attached to an electronegative atom (the donor) interacts with another electronegative atom (the acceptor). Repulsion forces appear between electron orbitals of atoms and are defined by their van der Waals radii. CH... $\pi$ -interactions are weak polar interactions involving  $\pi$ -system acceptor groups [64]. They are largely caused by dispersion of charges and partly from charge-transfer and electrostatic forces. CH... $\pi$ -interactions between aromatic groups are found mostly in the interior of the protein. Chi... $\pi$ -interactions between hydrophilic residues are located at the surface of the protein. Aromatic-aromatic interactions ( $\pi$ - $\pi$  stacking) also occur in proteins and are non-covalent interactions between aromatic rings. These interactions are important in protein folding, base stacking of DNA nucleotides and molecular recognition. Besides the above mentioned non-covalent interactions, the structure of proteins are also stabilized by covalent bonds. Disulfide bridges are formed between pairs of cystyl residues. Electrostatic forces occur between residues with net opposite charges, known as salt bridges or between two dipoles, each formed from the asymmetric distribution of electrons within the residues.

### 1.6.2 The secondary structure

Regular arrangement of the linear polypeptide chains (secondary structure) with repeating values for the  $\phi$  and  $\psi$  torsions angles and main-chain hydrogen bonding results in two major secondary structure types,  $\alpha$ -helices and  $\beta$ -strands [63, 65]. The inner part of the  $\alpha$ -helix is formed by the coiled polypeptide main-chain and the surface by the side-chains projecting outwards in a helical arrangement. It is stabilized by hydrogen bonds between the carboxyl group of the amino acid residues of the main-chain and the amino group of the residue located four residues away in the amino acid sequence, with repeated torsion angle values of about  $-60^\circ$  for  $\phi$  and  $-40^\circ$  for  $\psi$  [63, 65]. It is also stabilized by the van der Waals interactions generated by the close packing of the backbone atoms. The various amino acid residues along a polypeptide chain have different tendencies to form  $\alpha$ -helices [63, 66]. For example, A, L, F,

W, M, H and Q stabilize  $\alpha$ -helices, whereas S, I, T, Q, D and G have a destabilizing effect and P and hydroxyproline create sharp bends in the helices that destroy the helices. Other types of helices with hydrogen bonds between residues residing closer together ( $n + 3$ ) or farther apart ( $n + 5$ ) are also known [63]. The former is called the  $3_{10}$ -helix, with 3 residues per turn and 10 atoms between the donor and the acceptor in the hydrogen bond. The dipoles of the  $3_{10}$ -helix are not so well aligned as in the  $\alpha$ -helix, i.e. it is a less stable structure and side chain packing is less favorable. The later ( $n + 5$ ) is denoted as the  $\pi$ -helix [63].

The other major secondary structure is the  $\beta$ -strand. A  $\beta$ -strand is basically made of a 5 to 10 residue unit of the polypeptide whose backbone is almost fully extended, with rotation angle values of about  $-120^\circ$  for  $\phi$  and  $140^\circ$  for  $\psi$  [63]. A  $\beta$ -strand has two residues per turn and a translation of 3.4 Å per residue and is not a stable structure [63].  $\beta$ -strands therefore, tend to interact with other  $\beta$ -strands that either belong to other regions of the same polypeptide chain but are apart in the primary structure or are present in different polypeptide chains. Adjacent  $\beta$ -strands are stabilized by hydrogen bonds formed between the carbonyl groups of one  $\beta$ -strand and the amino groups of another  $\beta$ -strand and vice versa. The corresponding structures contain alternate  $C^\alpha$  atoms lying a little above and a little below the plane of the sheet and are called  $\beta$  pleated sheets. The sheets contain on average 2 to 6  $\beta$ -strands and are designated either parallel, anti-parallel or mixed  $\beta$ -sheets depending on the relative direction of the strands. Similar to  $\alpha$ -helices, some amino acid residues have a higher tendency to form  $\beta$ -strands than others [63, 66]. Thus V, I and T, which contain branched side-chains and the three aromatic residues F, Y and W favour the formation of  $\beta$ -strands. In contrast, E, Q, L, D, N, C and P exhibit low propensity to form  $\beta$ -strands.

A number of non-repetitive well-ordered structures are present in protein chains in addition to the major secondary structures. These non-repetitive structures occur mostly at the surface of proteins and frequently contain G and P residues due to the special conformational properties of these two residues. The non-repetitive structures are mainly turns, connections and loops that allow formation of structural motifs that constitute super-secondary structure. Examples of these motifs are the helix–loop–helix motif, consisting of two  $\alpha$ -helices joined by a loop region and the hairpin  $\beta$  motif, composed of two adjacent antiparallel  $\beta$ -strands connected by a loop.

### 1.6.3 Tertiary and quaternary structures

The tertiary structure is also determined by the packing of the  $\alpha$ -helices and  $\beta$ -strands, which combine to form units called domains [67, 68]. These domains form the fundamental elements of the globular polypeptide chains with regards to function. The polypeptide chain with secondary structures may occur as a folded structure [69, 67, 68]. The tendency of hydrophobic groups to minimize their contact with water and hence occur on the interior while hydrophilic groups occur on the exterior provide the folding force resulting in the lowest energy conformation. This lowest energy conformation is the tertiary structure. The tertiary structure (native conformation) of a protein is its biologically active conformation. The major stabilizing force responsible for the compact three-dimensional form of a protein is hydrophobic interactions between non polar side chains of amino acids. Polar side chains undergoing hydrogen bonding and ionic interactions also stabilize the tertiary structure.

Additionally, Van der Waals interactions and disulphide bonds contribute towards the stability of the native protein structure. Many proteins are made up of two or more polypeptide chains, called subunits or monomers, which represent the quaternary structure. The subunits are folded independently and interact at surfaces complementary in shape and physical properties.

The interactions between two identical subunits (homo-dimers) are either isologous or heterologous. In isologous associations the same surfaces on both subunits are utilized, whereas heterologous interactions involve two different sites [63]. Non-polar interactions occur preferentially at the centre of the interfaces while hydrophilic groups are often located at the surface allowing interaction with polar solvents. As is the case with crystals, the asymmetrical packing of subunits minimizes the free energy in the aggregated form [63].

Thus any amino acid changes that occur in the protein structure due to nucleotide variation can affect residue interactions and stabilise or destabilize the structure and subsequently affect protein function.

### 1.6.4 Structural features of human NATs

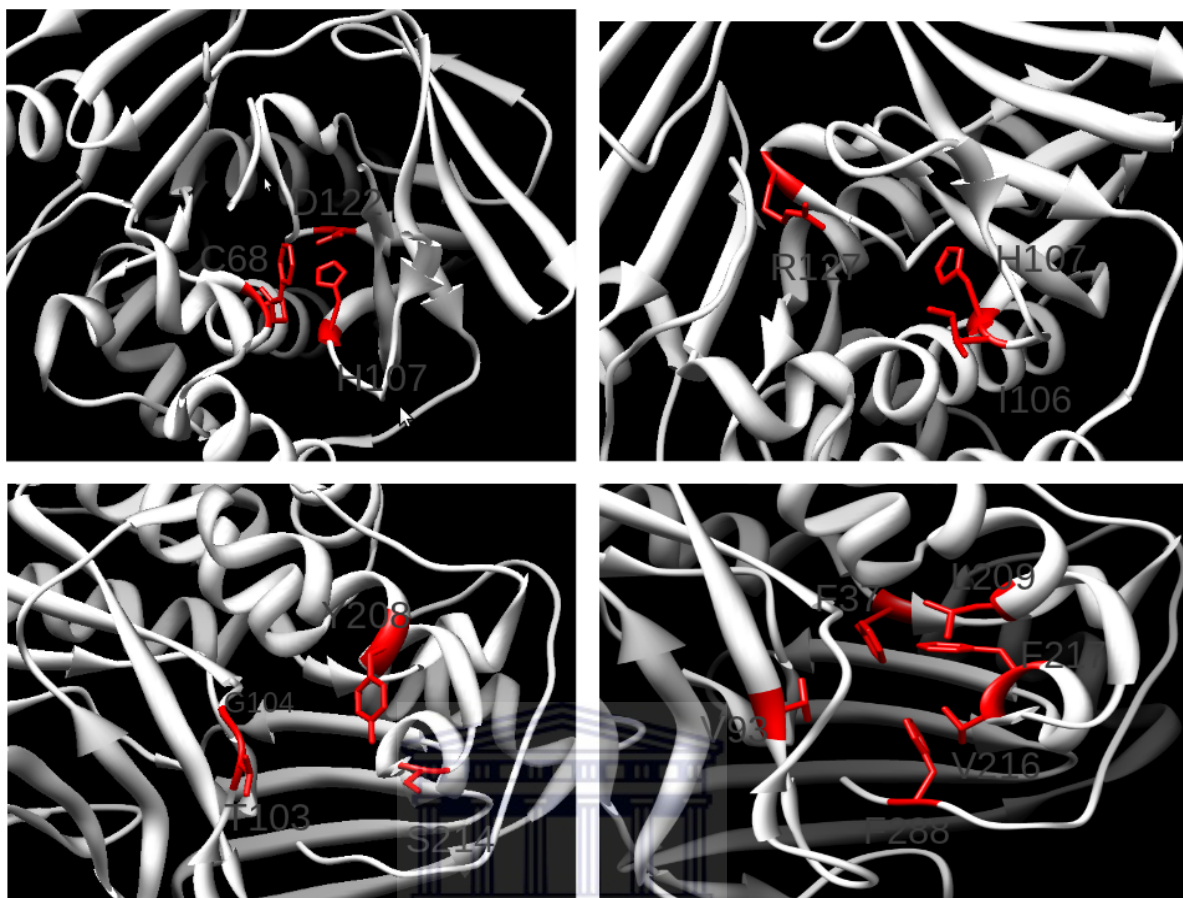
Human NATs like any other protein, have a unique structure that relates to its function. Residues involved in substrate-binding by van der Waals contacts include F37, F93, L209, F217, S216 and L288 (Fig 1.4)



[56]. Hydrogen bonds to CoA involve the amide nitrogens of T103 and G104 in the  $\beta 3$ - $\beta 4$  loop and the hydroxyl groups of S214 (or Y214 in NAT2) from the  $\beta 9$  and the  $\beta 9$ - $\beta 10$  loops respectively. The N6 of the adenine ring of CoA form a single hydrogen bond with side chain of S287. This ring makes contact with the hydrophobic residues F125 and V98. The hydrophobic surface of the substrate-binding site of human NAT1 is formed by the aromatic ring of F125 and the side chain of V93 including the hydrophobic residues I106 and F217 (Fig 1.4). The substrate-binding site in NAT1 is smaller ( $162\text{\AA}$ ) than NAT2 ( $257\text{\AA}$ ) as a result of two important residue substitutions at positions 127 and 129 [56]. In NAT1, these residues are R127 and Y129 while NAT2 carries S127 and S129. Another important substitution to the substrate-binding site occurs at position 93 where NAT1 exhibits valine and NAT2 carries phenylalanine. In NAT2 this amino acid leads to a “lip” in the van der Waals surfaces of the binding site making it more selective for substrates that can fit this feature.







**Figure 1.4** Functionally important residues of NAT1 protein

The catalytic triad residues comprise C68, H107 and D122, which are highlighted in red at the top left corner. The substrate, p-AS binding residues I106, H107 and R127 are displayed in red at the top right corner. The residues critical for CoA binding are displayed in red at the bottom left (T103, G104, Y208 and S214). At the bottom right corner are other substrates binding hydrophobic residues (F37, V93, L209, V216, F217, F288).

A 17-residue insertion (167 to 183) in human NATs represents a striking difference between human and prokaryotic NATs [56]. This insertion is proposed to contribute to the stability of the human enzymes by extending the 4-stranded anti-parallel  $\beta$ -sheets in prokaryotes (domain III) by a short  $\beta$ -strand [56]. It buries the carboxy terminus of the NAT protein and forms hydrogen bonds with residues from  $\beta$ 14 and  $\beta$ 15 strands,  $\beta$ 14- $\beta$ 15,  $\beta$ 9- $\beta$ 10 and  $\alpha$ 6- $\beta$ 14 loops [56]. The conserved catalytic triad of C68, H107 and D122 in human NATs is structurally superimposable with that of prokaryotic NATs [56].

## 1.7 Project rationale and objectives

### 1.7.1 Rationale

A major interest in human genetics is to distinguish between functionally neutral mutations and those that contribute to disease [70] as amino acid substitutions account for approximately half of the known genes responsible for inherited diseases [39]. Several nsSNP variants have been functionally characterised particularly in human NAT1 [35, 49, 50, 53, 54, 71, 72, 73, 74, 75]. However, such experimental studies are expensive and time consuming. Computational analysis can discriminate between neutral SNPs which constitute the majority of genetic variation and SNPs likely to affect protein function [76]. This helps to inform experimentalists on prioritizing SNPs for functional studies and to deepen the understanding of genotype-phenotype relationships.

Computational methods have been developed to predict the impact of amino acid substitution on the structure and function of a protein [70, 77, 78, 79, 80]. These algorithms are based on sequence conservation over an evolutionary period, the physical and chemical properties of the substituted amino acids and/or protein structural domain information. The Sorting Intolerant From Tolerant (SIFT) algorithm is based on sequence identity of related genes and domains over evolutionary time. It also considers the characteristics of the amino acid residues when predicting the effect of the substitution. The Polymorphism phenotyping version 2 (POLYPHEN-2) algorithm takes into account the sequence conservation, the amino acid characteristics and the location of the substitution within functional domains of the protein available in the annotated database of SWISS-PROT [81].

Ng and Henikoff [70], used SIFT to predict the effects of nsSNPs on several proteins in SWISS-PROT or TrEMBL. In their analysis, 69% (3626/5218) of the substitutions known to be involved in disease were predicted as damaging. Furthermore, SIFT predicted that 18 of 22 nsSNPs found in diseased patients affect protein function and that 9 of 10 nsSNPs found in control patients are functionally neutral. Similarly, Xi and colleagues [78] analysed effects of amino acid substitutions on protein activity of DNA repair genes using SIFT and POLYPHEN. For the above study, SIFT classified 226 of 508 (44%) as intolerant while POLYPHEN classified 165 of 489 (34%) as possibly damaging. The results from the two algorithms were in agreement, with concordance of predicted impact observed for ~62% of the vari-

ants [78]. Johnson and colleagues [80] observed a similar high concordance of 73% between SIFT and POLYPHEN predictions for nsSNP effects in genes involved in steroid hormone metabolism and response. Di and colleagues [76] carried out analysis on 247 nsSNPs using SIFT and POLYPHEN. In the above study, scatter graphs showed a negative correlation with a Spearman's rank correlation coefficient of -0.709 ( $p \leq 0.01$ ) illustrating a significant concordance between the prediction scores from SIFT and POLYPHEN algorithms. These studies together with others [77, 79] provide evidence that SIFT and POLYPHEN can differentiate between damaging and neutral nsSNP in two-third cases. Given that SIFT and POLYPHEN-2 can not correctly predict over 25% of nsSNPs [76, 78], it is important to expand the assessment of the functional effects of nsSNPs using other methods such as homology modelling, Gibbs free energy change and residue interactions.

Human arylamine *N*-acetyltransferase 1 (NAT1) is a xenobiotic metabolizing enzyme that affects the biological activity and toxicity of compounds including tuberculosis (TB) drugs. NAT1 catalyses either the acetyl-CoA dependent *N*-acetylation of primary aromatic amines and hydrazines (usually deactivating) or the *O*-acetylation of their *N*-hydroxylated metabolites (usually activating) [7]. NAT1 specifically acts on para-aminosalicylic acid (p-AS) which is used in TB treatment [16], a primary drug in TB treatment in South Africa [20, 82]. Polymorphisms in NAT1 divide individuals into slow, intermediate and rapid acetylators. Rapid acetylators are at risk of not responding to treatment while slow acetylators are at risk of drug toxicity.

Twenty-three SNPs that affect protein function [35, 36, 37, 47, 49, 50, 52, 53, 54, 71, 72], have been detected in NAT1 mostly among Caucasians [49] and more recently have been detected in the South African mixed ancestry population in the Western Cape region. Eleven novel nsSNPs in NAT1 have also been detected in the South African mixed ancestry population in the Western Cape region. The effects of these 11 nsSNPs on the function of NAT1 are not known. Previous studies have shown that functional effects of nsSNPs can be predicted computationally [83, 84].

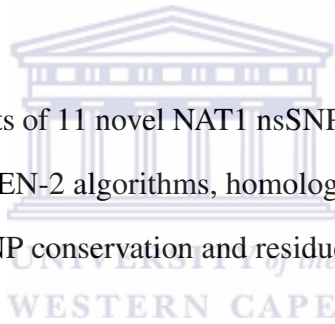
This project proposes to computationally test the effects of 11 novel nsSNPs on the structure and function of NAT1 using structural analysis including homology modelling, residue interactions, stability calculation and solvent accessibility in addition to the routinely used SIFT and POLYPHEN-2 algorithms. Our results will allow us to prioritize the SNPs for experimental analysis.

### 1.7.2 Aim and Objectives

The aim of this study is to assess the value of supplementing the routinely used SIFT and POLYPHEN-2 algorithms with homology modelling, relative solvent accessibility, Gibbs free energy change, SNP conservation and residue interactions, to predict the effects of nsSNPs on the function of a protein (i.e. NAT1).

#### **Objectives:**

1. To assess the functional effects of 11 NAT1 nsSNPs previously identified in Caucasians and recently confirmed in a South African population using SIFT and POLYPHEN-2 algorithms, homology modelling, relative solvent accessibility, Gibbs free energy change, SNP conservation and residue interactions analyses.
2. To assess the functional effects of 11 novel NAT1 nsSNPs identified in a South African population, using the SIFT and POLYPHEN-2 algorithms, homology modelling, relative solvent accessibility, Gibbs free energy change, SNP conservation and residue interactions analyses.



# Chapter 2

## MATERIALS AND METHODS

### 2.1 MATERIALS

The functional effects of nsSNPs are routinely assessed with the SIFT and POLYPHEN-2 algorithms. These can be complemented with structural analysis. Computational analysis involving hydrogen bond or salt bridge determination, solvent accessible surface area calculation and stability changes, would require SNP residues modelled in the structure of the protein. This can be achieved with homology modelling.

#### 2.1.1 Sorting Intolerant From Tolerant (SIFT) prediction server

A useful tool in assessing the functional effects of amino acid substitutions is Sorting Intolerant From Tolerant (SIFT). SIFT aims to predict whether an amino acid substitution will affect the function of a protein considering the amino acid sequence of the protein and physical properties of the amino acids [85]. SIFT also requires homologue information from phylogeny analysis or a sequence alignment. SIFT can use position-specific information obtained from sequence alignments collected through PSI-BLAST. SIFT assumes that conserved amino acids are functionally important. Changes at well-conserved positions tend to be identified as deleterious. Substitutions of residues with amino acids of similar character will generally be accepted as neutral. SIFT is a multi-step procedure:

1. use a group of sequences to search for closely related sequences,

2. then chooses closely related sequences that may have a function similar to the query sequence,
3. generates the alignment of the chosen sequences, and
4. calculates normalized probabilities for all possible substitutions from the alignment.

The reliability of predictions of the functional effects of nsSNPs depends on the diversity of the sequences in the alignment. Closely related sequences will indicate many conserved positions. SIFT will hence predict most substitutions to affect protein function. This may lead to a high proportion of false positive errors where functionally neutral substitutions are predicted to be deleterious. To account for these false errors, SIFT calculates the median conservation value to quantify the diversity of the sequences in the alignment. The degree of conservation is calculated for each position in the alignment and the median of these values is obtained in logarithmic terms. Median conservation values range from 4.32, for a perfectly conserved residue to 0.00, when all 20 amino acids are observed at a position with equal probability. By default, SIFT generates alignments with a median conservation value of 3.0. However, the recommended range is 2.75 to 3.25. Predictions based on sequence alignments with higher median conservation values are less diverse and will have a higher false positive error.

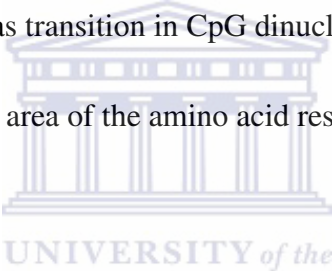
Substitutions at positions with normalized probabilities below 0.05 are predicted to be deleterious (affect protein function) and those greater than or equal to 0.05 are predicted to be tolerated [85].

The databases employed in SIFT analysis include UniRef (<http://www.ebi.ac.uk/uniref/>), UniProt-SwissProt ([http://web.expasy.org/docs/swiss-prot\\_guideline.html](http://web.expasy.org/docs/swiss-prot_guideline.html)), UniProt-TrEMBL (<http://www.ebi.ac.uk/uniprot/>) and NCBI non-redundant database (<http://www.ncbi.nlm.nih.gov/>). The algorithm and instructions for analysis of amino acid substitutions are available at <http://sift.jcvi.org/>.

### **2.1.2 Polymorphism phenotyping version 2 (POLYPHEN-2) prediction server**

Polymorphism phenotyping version 2 (POLYPHEN-2) is a computer programme that attempts to predict the effects of non-synonymous mutations on protein function [86]. POLYPHEN-2 uses sequence, phylogenetic or evolutionary information and works best if structural information is available as well. In addition to conservation scores, it adds physico-chemical differences and structural features of the poly-

morphic variants to enable the prediction of functional effects of amino acid change. It uses the following sequence-based and structure-based predictive features. The predictive features include:

- position-specific independent count (PSIC) score for the wild-type allele,
- differences of PSIC scores,
- number of distinct amino acids observed at the position of the multiple alignment,
- position of the mutation within/outside a protein domain as defined by Pfam,
- congruency of the mutant allele to the multiple alignment,
- sequence identity with the closest homologue deviating from wild-type allele,
- whether the variant occurred as transition in CpG dinucleotide context,
- normalized accessible surface area of the amino acid residue,
- crystallographic  $\beta$ -factor and UNIVERSITY of the
- change in accessible surface area propensity for buried residues.

The input is the amino acid sequence of a protein or the SWALL database ID or accession number, together with sequence position and two amino acid variants characterizing the polymorphism. It predicts a substitution to be damaging i.e. to affect protein function or benign (i.e. most likely lacking any phenotypic effect). The POLYPHEN-2 scores range from 0 to 1. A variant with less than or equal to 0.5 score is considered benign while a variant score above 0.5 means the variant is damaging [86]. The SWALL database is located at <http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+SWALL>. Additional information and instructions for analysis of amino acid substitutions are available at <http://genetics.bwh.harvard.edu/pph2/>.

### 2.1.3 CLUSTALW2 multiple sequence alignment server

CLUSTALW2 is a web-based tool used to align DNA or proteins sequences [87].



Three or more sequences to be aligned are entered directly into the sequence input window. However, a file containing three or more valid sequences in the right format can be uploaded and used as input for the multiple sequence alignment. A sequence type, alignment type, protein weight matrix, gap open, gap extension are all options in CLUSTALW2 to ensure quality alignment. Detailed information on CLUSTALW2 is available at <http://www.ebi.ac.uk/Tools/msa/clustalw2/>.

#### **2.1.4 Homology modelling and MODELLER software**

Homology modelling is a computational method to predict the three-dimensional structure of proteins based on related experimentally elucidated protein structures. Protein folds are more highly conserved than amino acid sequences. Proteins with significant sequence homology are generally assumed to be related and to share a common core structure [88]. This means that the structure of a protein can be inferred from proteins of known structure. Typically, homology modelling requires the following information:

1. the sequence of the protein of unknown structure or the “target sequence”.
2. a coordinate file describing the three-dimensional arrangement of atoms derived experimentally by techniques such as X-ray crystallography or nuclear magnetic resonance spectroscopy or the “template”; and
3. an alignment between the target sequence and the template sequence.

MODELLER is a computer program that generates three dimensional models of proteins by attempting to accommodate spatial restraints. It is commonly used for homology or comparative modelling of protein structures. It allows the user to provide sequence alignment of a sequence to be modelled with known related structures and then calculates a model with all non-hydrogen atoms [89]. MODELLER also performs other tasks such as optimization of the protein structure model with respect to a defined objective function.

The inputs to MODELLER are restraints in the spatial sequence of amino acids and ligands to be modelled. The output is a 3D model that satisfies defined restraints. MODELLER aligns the target sequence with the template structure and builds 3D models of each target/template alignment producing the number of models specified. The best model can be selected with discrete optimized potential energy (DOPE)



and GA341 assessment scores. DOPE is a statistical potential used to assess homology models in protein structure prediction. DOPE is based on a reference state that corresponds to noninteracting atoms in a homogeneous sphere with the radius dependent on a sample native structure. DOPE is implemented in the MODELLER software and is used to assess the energy of the protein model generated by MODELLER. The DOPE profiles of proteins may be plotted with a graphical tool such as GNUPLOT [90] a portable command-line driven graphing utility or plotting engine for Linux, OS/2, MS Windows, OSX, VMS and other platforms.

### **2.1.5 PROCHECK Software**

Protein structures derived from experimental data as well as those obtained from “model building” are subject to many sources of errors. The PROCHECK suite of programs [91] provides a detailed check on the stereochemistry of a protein structure and outputs a number of plots in post script format and a comprehensive residue-by-residue listing with an assessment of the overall quality of the structure as compared to refined structures of the same resolution. It also highlights regions that may need further investigation and categorizes these regions into most favoured regions, additionally allowed regions, generously allowed regions and disallowed regions. A good quality structure would be expected to have over 90% percent of its residues within the most favoured regions (<http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>).

### **2.1.6 UCSF CHIMERA**

The structure of a protein reveals more information about function than the amino acid sequence alone. This structure is, however, determined by countless interactions of amino acid residues with each other. These interactions as mentioned in section 1.6, include hydrogen bonding, hydrophobic interactions, ionic or electrostatic interactions, van der Waals interactions and disulphide bonds, among main chain or side chains of residues. Hence, any changes due to mutations that significantly disrupt these interactions would affect the structure of the protein and consequently its function.

The molecular graphics program UCSF CHIMERA [92] includes a suite of tools for interactive analyses

of sequences and structures. Structures automatically associate with sequences in imported sequence alignments. CHIMERA allows the superimposition of structures without pre-existing sequence alignment. CHIMERA generates structure-based sequence alignments from superpositions of two or more proteins. It has a multi-scale extension, that provides the functionality to visualize large-scale molecular assemblies such as proteins while the multalign viewer, permits multiple sequence alignments and associated structures to be viewed (<http://www.cgl.ucsf.edu/chimera/>).

### **2.1.7 FOLDX Software**

Protein structures are stabilized by non-covalent inter-molecular interactions between amino acid side chains. All biological processes depend on proteins being stable and in the appropriate folded conformation. The impact of amino acid substitutions on the stability of proteins may be estimated by calculating the Gibbs free energy changes associated with such substitutions. Such energy changes indicate stabilization or destabilization of the approximate protein native structure. FOLDX attempts to quantify the impact that substitutions have on the stability of proteins and protein complexes using an empirical force field [93]. Given a full atomic description of the structure of the protein, it calculates free energy change in kcal/mol of the wild-type (WT) and variant type (VT). The different energy terms used by FOLDX for calculating the free energy are weighted using empirical data from protein engineering experiments [93]. The basic analysis performed is the calculation of Gibbs free energy of folding. The following are options : temperature (K), water, pH, ionic-strength, metal and vdWDesign. Amino acid substitutions with Gibbs free energies lower than the zero are inferred to be stabilizing while those with free energies larger than zero are destabilizing [93]. FOLDX attempts to optimize the conformation of residues with bad torsion angles, van der Waals clashes or total energy before calculating the free energy changes (<http://foldx.crg.es/>).

### **2.1.8 NACCESS Software**

NACCESS calculates the accessible surface area of a molecule provided as a PDB format file [94]. It calculates the atomic accessible surface area defined by rolling a probe of given size over a van der Waals

surface. The program accomodates up to 20000 atoms and allows the user to change the probe size and atomic radii. It genrates 3 files: atomic accessibility file (.asa ), residue accessibility file (.rsa ) and a log file (.log). <http://www.bioinf.manchester.ac.uk/naccess/>.

## 2.2 METHODS

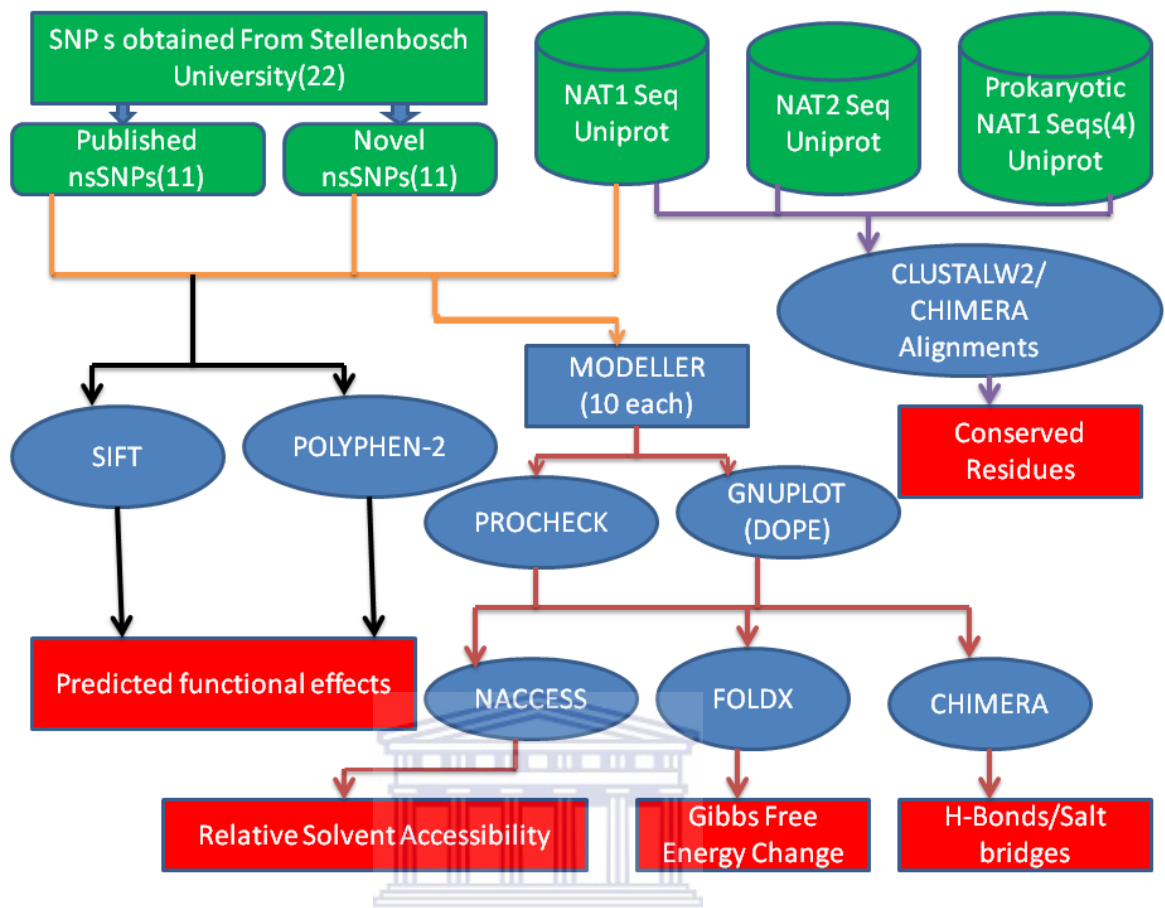
### 2.2.1 Sequence Data Acquisition

#### 2.2.1.1 SNPs

Information for 22 nsSNPs in the coding region of human NAT1 identified in the mixed population in South Africa were provided by Dr. Cedric Werely at Stellenbosch University (Table 2.1). Of the 22 nsSNPs, 11 matched published nsSNPs while the other 11 are novel and are all located near the 3'-untranslated region of the *NAT1*. The effects of eight of the 11 published nsSNPs have previously been analysed structurally by Walraven and colleagues [83]. These 8 nsSNPs were used as an internal control to verify the methodology of Walraven and colleagues and to validate our methodology. The remaining 3 published nsSNPs and the 11 novel nsSNPs were analysed as described in this chapter and Fig 2.1.

**Table 2.1** List of nsSNPs and the associated amino acid changes

SNP Status	NAT1 nsSNPs (Gene Positions)	Amino Acid Substitutions
Published	C190T	R64W
	G350C	R117T
	G445A	V149I
	G497C	R166T
	G498C	E167Q
	G560A	R187Q
	A613G	M205V
	T640G	S214A
	A752T	D251V
	G781A	E261K
	A787G	I263V
Novel	A1017T	T193S
	T1046G	F202V
	A1069C	Q210P
	G1125C	D229H
	T1132G	V231G
	T1144C	V235A
	C1159G	T240S
	G1165T	R242M
	A1174T	N245I
	T1217G	S259R
	G1230A	E264K



**Figure 2.1** Methodology flow chart

The effects of nsSNPs on enzyme function were analysed with SIFT and POLYPHEN-2, homology modelling, relative solvent accessibility, Gibbs free energy change, SNP conservation and residue interactions. The 22 nsSNPs in human NAT1 were homology modelled using MODELLER. The models with the lowest DOPE scores, were evaluated with PROCHECK and their DOPE profiles plotted with GNUPLOT. The evaluated models were used for solvent accessibility calculation with NACCESS, Gibbs free energy change with FOLDX and hydrogen bonds or salt bridge determination with CHIMERA. These calculations provide evidence for the structural and hind functional effects of a nsSNP. Sequences of human NAT1, human NAT2 and four prokaryotic NAT1s were aligned with CLUSTALW2 and optimized with CHIMERA to identify functionally important residues. Green marks data used, blue the tools used and red the results.

### 2.2.1.2 Proteins

The amino acid sequences of human NAT1, NAT2 and NAT1 of *M. marinum*, *M. bovis*, *M. tuberculosis* and *M. smegmatis* were retrieved from the Universal Protein Resource (<http://www.uniprot.org>; Uniprot). The search term “human arylamine *N*-acetyltransferases” was used to retrieve the FASTA format sequences of both human NAT1 and NAT2 while “arylamine *N*-acetyltransferases” was used to search for the other related NATs (i.e. prokaryotic NATs). In this work, the SNP containing sequences were obtained by replacing each wild-type amino acid with the corresponding variant (i.e. the nsSNP)

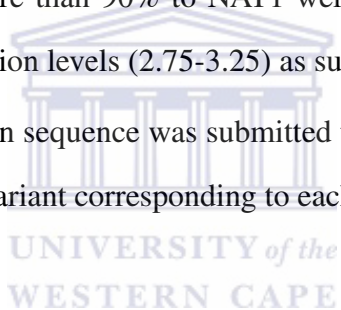
as indicated in Table 2.1. These modified sequences were the “targets” and the PDB crystal structure of human NAT1 was the “template” (hereafter referred to as wild-type structure) in the homology modelling analysis.

### **2.2.2 SIFT and POLYPHEN-2 analysis**

Two on-line servers were used: the SIFT program [85] and POLYPHEN-2 program [95] to predict the functional impact of the nsSNPs.

A FASTA NAT1 protein sequence and a file containing the wild-type and SNP residues were uploaded to the SIFT server. The input NAT1 protein was searched against the UNIREF release April (2011) database with default parameter for median conservation of sequence (see section 2.1.1 for definition). Sequences in UNIREF that had similarity more than 90% to NAT1 were excluded from the SIFT predictions to obtain acceptable median conservation levels (2.75-3.25) as suggested by the authors.

The FASTA formatted NAT1 protein sequence was submitted to the POLYPHEN-2 server. The position of each amino acid wild-type and variant corresponding to each of the 22 NAT1 nsSNPs were specified.



### **2.2.3 Sequence conservation analysis**

Residues critical to the structure and function of proteins are mostly highly conserved [96]. Changes or substitutions of such residues can lead to significant consequences for the structure and hence function of the protein. Thus the conservation of a residue gives information about its likelihood of affecting the structure and function when replaced by a different amino acid. CLUSTALW2 [87] and UCSF CHIMERA 1.5.3 [92] were used to align and optimize sequences and visualize the conserved residues respectively.

### **2.2.4 Structural analysis**

As part of expanding our assessment of the functional effects of NAT1 nsSNPs, our structural analysis included (a) homology modelling, (b) residue interactions analysis, (c) determination of Gibbs free energy changes associated with the nsSNPs and (d) solvent accessibilities calculations.

#### **2.2.4.1 Homology modelling**

The crystal structure of each nsSNP sequence was modelled with MODELLER 9.9 [89] using the wild-type (WT) NAT1 (PDB code: 2PQT) as a template. Each SNP sequence in PIR format was then aligned with the template using the MODELLER align2d() to construct the “target-template” alignment.

The automodel class of MODELLER was then used to generate 10 3D models (hereafter referred to as “variant structures”) of each target-template alignment and selection of the “best” model with the lowest DOPE scores chosen. DOPE scores were obtained using MODELLER “assess\_methods with DOPE and GA341 assessment scores” option. The DOPE potentials for best models and WT were evaluated with MODELLER assess\_dope() command. Four non-physiological residues at the N-terminus of NAT1 PDB crystal structure were removed before calculating the DOPE potential of the WT structure. Energy profiles of all variant (VT) structures are plotted against the WT structure energy profile as DOPE plots using GNUPLOT and structures were analysed visually using the UCSF CHIMERA 1.5.3 [92].

#### **2.2.4.2 Evaluation of modelled structures**

PROCHECK [91] was used to check the stereochemical quality of all structures (WT and VT) using WT resolution of 1.78 Å.

#### **2.2.4.3 Residue interaction determination**

Hydrogen bonds and salt-bridges, were assessed using UCSF CHIMERA 1.5.3, with the “Command Line” option and “FindHBond” function (Appendix A.1).

#### **2.2.4.4 Protein structure stability calculation**

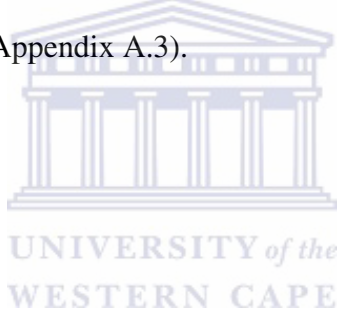
Protein structures are stabilized by non-covalent intra-molecular interactions between amino acid side chains and main chains. The impact of mutations on the stability of proteins may be tentatively quantified by estimating the Gibbs free energy changes associated with such mutations. Such energy changes may reveal destabilization of the native structure of the protein and hence changes in function or cause of disease [97]. Changes to the structure of a protein due to nsSNPs can be evaluated using 2 approaches

[97], namely, determining changes to the 3D-structure predicted from crystallography studies and measuring changes in protein stability by unfolding energy. Functional alterations due to structural changes could potentially be derived from either of the two methods. Estimating protein stability in terms of unfolding energy changes is simpler and less expensive [97]. Gibbs free energy of protein unfolding energy is calculated using the equation  $\Delta\Delta G = \Delta G_{\text{mutant}} - \Delta G_{\text{wild-type}}$  [97] for example as in FOLDX.

To assess the effect of the mutations on the stability of NAT1, FOLDX 3.0  $\beta$  5.1c [93] was used to calculate the change in Gibbs free energy associated with the substitutions (see Appendix A.2).

#### **2.2.4.5 Solvent accessibility calculation**

Intramolecular hydrogen bonding in proteins is suggested to depend on the accessibility of the donors and acceptors to water molecules [98]. NACCESS 2.1.1 [94] was used to provide relative and absolute solvent accessibilities per residue (Appendix A.3).





# Chapter 3

## RESULTS AND DISCUSSION

The human NAT1 enzyme is a xenobiotic metabolizing enzyme which inter alia catalyzes the acetylation of the anti-tuberculosis drugs para-aminosalicylic acid and para-aminobenzoyl glutamate. The identification and characterization of SNPs in genes of drug-metabolizing enzymes such as *NAT1* is critical in understanding differences in drug metabolism, therapeutic efficacy and inherited diseases of individuals. It is therefore important to investigate the functional impact of SNPs on these enzymes.

This study provides a first step in this direction by predicting the effects of 11 novel nsSNPs on the structure and function of human NAT1 using homology modelling, relative solvent accessibility calculation, Gibbs free energy change calculation, SNP conservation and residue interactions to augment the commonly used SIFT and POLYPHEN-2 algorithms.

### 3.1 SIFT and POLYPHEN-2 analysis

Functional consequences of non-synonymous SNPs are routinely assessed using the SIFT and POLYPHEN-2 algorithms. The results for 11 published and 11 novel nsSNPs in NAT1 are listed in Tables 3.1 (published nsSNPs) and 3.2 (novel nsSNPs). SIFT predictions were based on 27 or 28 related sequences and a median conservation scores of 3.04-3.05. A median conservation score of 2.75 to 3.25 is informative (here 3.05) [85].

Both SIFT and POLYPHEN-2 predict the three published variants R64W, R166T and D251V and the six

**Table 3.1** Predicted effects of NAT1 published nsSNPs

SNP	SIFT	Score	POLYPHEN-2	Score	Stability (kcal/mol)	WT <sup>1</sup> Bonds	VT <sup>2</sup> Bonds
R64W	Affect protein function	0.00	Probably damaging	1.00	3.12	7	2
R166T	Affect protein function	0.00	Probably damaging	1.00	2.62	3	1
D251V	Affect protein function	0.01	Probably damaging	1.00	0.21	6	2
R117T	Tolerated	0.15	Benign	0.07	-1.32	2	2
V149I	Tolerated	1.00	Benign	0.00	-3.18	2	3
E167Q	Tolerated	0.40	Benign	0.00	7.09	1	1
M205V	Tolerated	0.70	Benign	0.00	-21.29	1	1
S214A	Tolerated	1.00	Benign	0.00	1.04	1	2
E261K	Tolerated	0.15	Benign	0.01	-20.20	3	1
I263V	Tolerated	1.00	Benign	0.00	-17.79	2	2
R187Q	Tolerated	0.59	Possibly damaging	0.81	-0.68	2	3

WT<sup>1</sup> - Wild-type and VT<sup>2</sup> - Variant type.

SIFT scores range from 0 to 1, with scores below 0.05 indicating deleterious substitutions while a score above 0.05 indicates tolerated variant. A median conservation score of 2.75 to 3.25 is deemed informative (here 3.05) [85]. POLYPHEN-2 scores also range from 0 to 1, with below 0.5 score indicating benign and scores above 0.5 damaging variants [86]. A negative Gibbs free energy change indicates a stabilizing mutation, a positive value a destabilizing mutations [93].

**Table 3.2** Predicted effects of NAT1 novel nsSNPs

SNP	SIFT	Score	POLYPHEN-2	Score	Stability (kcal/mol)	WT <sup>1</sup> Bonds	VT <sup>2</sup> Bonds
F202V	Affect protein function	0.00	probably damaging	1.00	-1.49	2	2
Q210P	Affect protein function	0.00	probably damaging	1.00	-6.20	4	0
D229H	Affect protein function	0.03	possibly damaging	0.60	8.33	3	1
V231G	Affect protein function	0.01	probably damaging	0.98	6.29	2	2
V235A	Affect protein function	0.01	probably damaging	0.97	6.98	2	2
N245I	Affect protein function	0.00	possibly damaging	0.87	-4.80	1	1
T193S	Tolerated	0.07	Benign	0.13	-8.92	1	1
T240S	Tolerated	0.58	Benign	0.05	2.47	3	3
S259R	Tolerated	0.55	Benign	0.02	3.50	5	2
R242M	Tolerated	0.16	possibly damaging	0.91	-7.79	5	2
E264K	Affect protein function	0.01	Benign	0.00	-15.00	1	1

WT<sup>1</sup> - Wild-type and VT<sup>2</sup> - Variant type.

SIFT scores range from 0 to 1, with scores below 0.05 indicating deleterious substitutions while a score above 0.05 indicates tolerated variant. A median conservation score of 2.75 to 3.25 is deemed informative (here 3.05) [85]. POLYPHEN-2 scores also range from 0 to 1, with below 0.5 score indicating benign and scores above 0.5 damaging variants [86]. A negative Gibbs free energy change indicates a stabilizing mutation, a positive value a destabilizing mutations [93].

novel variants F202V, Q210P, D229H, V231G, V235A and N245I to affect the function of NAT1. The variants R117T, V149I, E167Q, M205V, S214A, E261K, I263V (published), T193S, T240S and S259R (novel) were predicted by both algorithms as having no effects on the function of NAT1. For the nsSNPs R187Q (published), R242M and E264K (both novel) the algorithms showed contradictory results. Thus SIFT and POLYPHEN-2 agree in this assessment of nineteen (8 published and 11 novel) nsSNPs or a concordance of 86% for 22 nsSNPs compared to ~62% [78] and 73% [80] in earlier analyses.

Although useful in their potential to predict the functional consequence of nsSNPs, SIFT and POLYPHEN-2 have limitations that affect their prediction accuracy. Firstly, SIFT and POLYPHEN-2 rely on several different databases for SNP information. When too few sequences are available for a particular gene or when the sequences are closely related, a neutral variant may be predicted to affect protein function [85, 99]. Databases with erroneous SNP reports and bias of the data towards disease-related allelic variants are likely to lead to an over prediction of the number of deleterious nsSNPs [86]. For example, Ramensky *et al.* [86] predicted nsSNPs (grouped or categorized) to affect protein function for human genome variation (HGV-base) entries and showed that for the category “Proven” nsSNPs (SNPs confirmed by independent and solid experimental verification) was 28.9%, the overall prediction rate for the category “Suspected” nsSNPs (other SNP candidates) was 31.4% and for another “Proven” nsSNPs from systematic studies on healthy individual was 27.6%. Similarly, Di *et al.* [76] collected data from 63 *in vitro* and *in vivo* studies and found that the false negative predictions for SIFT and POLYPHEN were 43% and 33% respectively. Additionally, programs may identify base differences in a pseudogene as SNPs in the functional protein affecting the accuracy of prediction tools [70]. However, as more genome sequences become available SNP databases, should improve increasing their reliability [100, 101].

Errors may also arise because SIFT not accounting for mutations that affect transcription, translation, splicing and other post-translational alterations [70]. Further, the prediction may appear incorrect due to lack of association with altered phenotypes. SIFT and POLYPHEN-2 may be sensitive to an amino acid substitution and predict it to be damaging to protein function. On the other hand, SIFT and POLYPHEN-2 may predict a deleterious substitution to be “tolerated or benign” which may not be an obvious phenotype. This may be interpreted as a correct prediction although it is a recessive or undiagnosed deleterious substitution. Another limitation of these algorithms is that variants combination are not assessed and

the dependence of functional impact of a variant on genotype of other genes or on exposure risk is not addressed [99]. One final restriction of SIFT and POLYPHEN-2 is that the algorithms are unable to predict the impact of SNPs that occur outside of the coding region, such as promoter and enhancer regions and splice sites that may affect protein levels or protein function.

### **3.2 Sequence conservation profile between human and prokaryotic NAT1 enzymes**

As indicated, residues critical to the structure and function of proteins are mostly well conserved [96]. Substitutions of such residues will affect the structure and hence function of the protein. To assess the conservation of NAT1 nsSNPs in humans and prokaryotes, we aligned 6 wild-type NAT1 proteins using CLUSTALW2 and UCSF CHIMERA 1.5.3. The first 2 nsSNPs positions (64 and 117) are depicted in Figure 3.1, which highlights residues 1 to 140 of NAT1. The remaining 20 nsSNPs are displayed in Figure 3.2, which incorporates residues 141 to 288 (see arrows). Bacterial NATs shared between 26 and 29 % sequence identity to NAT1, while human NAT2 shares 81% identity with NAT1.

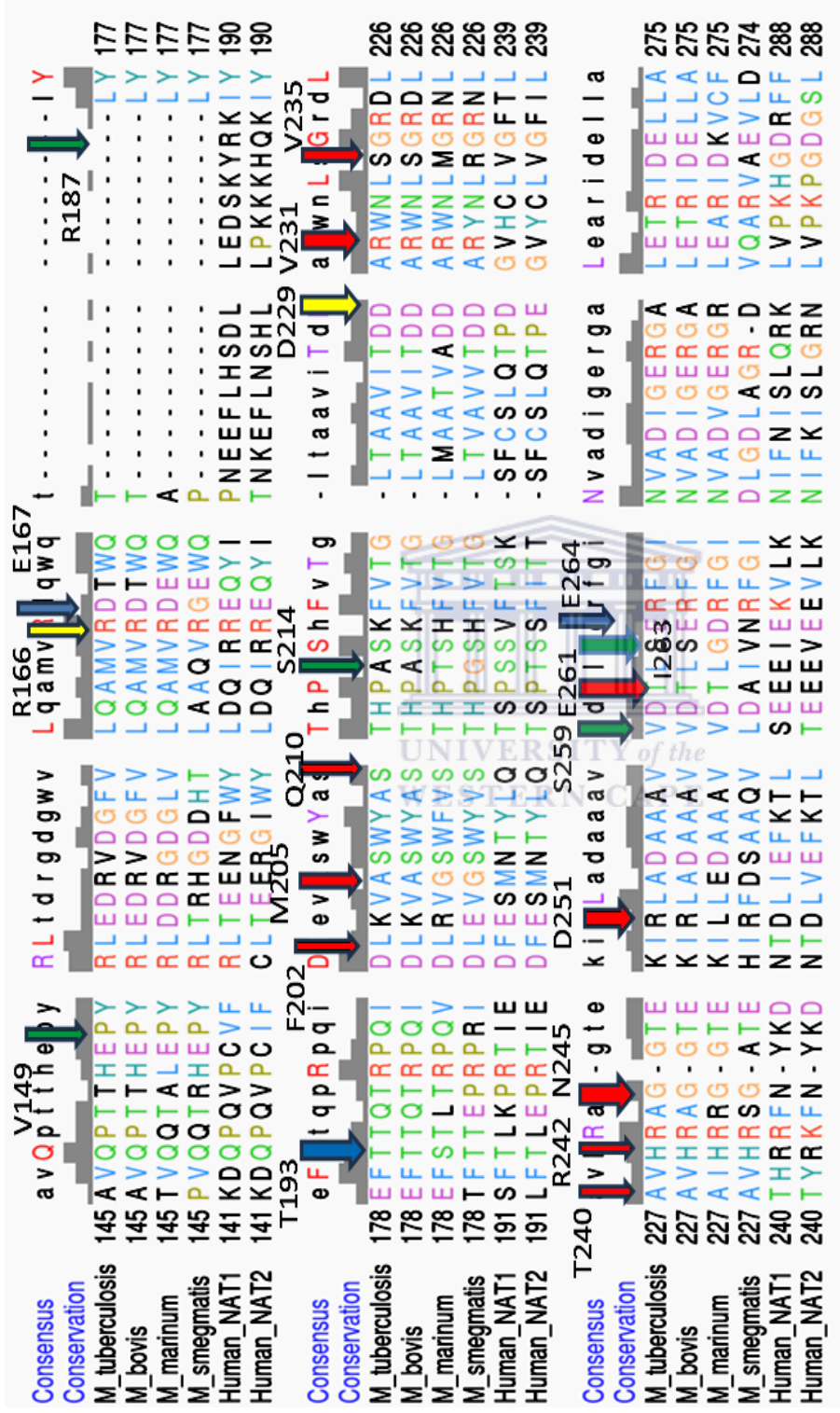
Residues 202, 205, 210, 231, 235, 242, 245, 251 and 261 are conserved in human NATs (red arrows, Figure 3.2) but distinct from prokaryotic sequences. Residues conserved in all sequences include positions 166, 229 (yellow arrows, Figure 3.2) and R64 (black arrow, Figure 3.2). The wild-type residues V149, R187, S214, S259 and I263 only occur in the human NAT1 sequence (green arrows, Figure 3.2). Positions 172 to 188 of the human NATs are not present in prokaryotes. This region corresponds to the 17-residue insert region unique to human NATs (Figure 3.2) and nsSNP position 187 occur in the 17-residue insert region of the human NATs (an R in NAT1 and Q in NAT2).

#### **R64W, R117T and R166T variants**

The same multiple sequence analysis indicates R64 and R166 to be conserved in all six sequences indicating their structural and functional importance. Replacing basic R64 by tryptophan would affect the function of NAT1. Similarly, substituting basic R166 with hydrophilic threonine would affect the function of NAT1. The prediction of both algorithms for variants R64W and R166T are probably correct. The sequence alignment further indicated R117 to be conserved in human NATs but replaced by







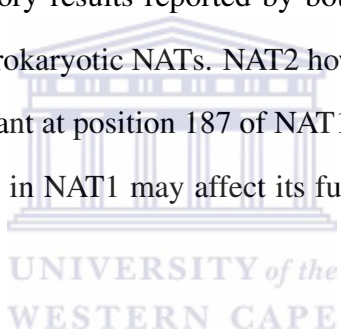
**Figure 3.2** Multiple sequence alignment of related NAT1 enzymes from positions 141 to 288

Multiple sequence alignment of NAT1 proteins of *M. tuberculosis*, *M. bovis*, *M. marinum*, *M. smegmatis*, Human NAT1 and Human NAT2. Highly conserved residues (80% or greater) are capitalized and shown in purple except for completely conserved residues which are indicated in red relative to the consensus header line. In the conservation header line, full bar height indicates complete identity, 2/3 bar height indicates Clustal strong group conservation, and 1/3 bar height indicates Clustal weak group conservation. Individual residues are coloured within a column according to the CLUSTALW2 colouring scheme depending on both residue type and pattern of conservation within the column. Amino acid positions quoted in the legend are relative to the sequence IDs. Wild-type residue positions of 20 published and novel SNPs are indicated by arrows. Residues conserved in only human NATs are indicated by red arrows, those most conserved are shown by yellow arrows, green arrows show residues that vary between the human NATs and other residues are in blue arrows.

glycine in other species. This implies that R117 may be critical for NATs and hence substituting it with threonine (T166) would have some significant functional impairment of NAT1 as predicted by SIFT and POLYPHEN-2.

### **V149I, E167Q and R187Q variants**

V149 only occurs in NAT1, it is replaced by isoleucine in NAT2 and proline in prokaryotic NATs. Substituting I149 for V149 may thus affect minimally the function of NAT1 as indicated by the algorithms. Both human NAT sequences have Glu at position 167 replaced by aspartic acid or glycine in prokaryotic NATs. No significant structural impact is thus expected when substituting by glutamine, an amino acid of similar properties and supports the conclusion of SIFT and POLYPHEN-2. The variant R187Q, is one of the variants with contradictory results reported by both algorithms. NAT1 R187 occurs in the 17-residue insert region absent in prokaryotic NATs. NAT2 however, has Q187. This may indicate that a basic residue is functionally important at position 187 of NAT1 whereas an acidic residue is required for NAT2. Hence substitution of Q187 in NAT1 may affect its function as predicted by the POLYPHEN-2 algorithm.



### **M205V, S214A and D251V variants**

M205 only occurs in human NATs, while *M. marinum* and *M. smegmatis* have a glycine and *M. tuberculosis* and *M. bovis* an alanine at this position. The above residues are all hydrophobic and as such substitution of hydrophobic valine, V205 may not have any effect on the function of NAT1 as predicted by SIFT and POLYPHEN-2. The sequence analysis shows that S214 occurs in only NAT1 while NAT2 has T214. The prokaryotic sequences have alanine, glycine or threonine. This means that the position accepts hydrophilic or hydrophobic residues for the function of the human NAT1. Substitution of S214, by similar size alanine will therefore have no effect on NAT1 function as predicted by SIFT and POLYPHEN-2. The residue D251 only occur in human NATs while 3 prokaryotic NATs possess R251 and one prokaryotic NAT possess L251. This implies charged residues are important in position 251. Substitution of non-charged V251 in NAT1 would therefore affect its function as predicted.



### **T193S, E261K and I263V variants**

The sequence alignment indicates that E261 occur only in the human sequences while three prokaryotic sequences possess threonine and one possess alanine. Substituting acidic E261 with basic K261 could affect the function of NAT1. In position 263, there are four hydrophobic residues; one isoleucine, two valine residues and one alanine residue and two hydroxylic serine residues. Thus substituting hydrophobic I263 for hydrophobic V263 in NAT1 is not expected to substantially affect the function of the protein. At position 193 all sequences possess the hydrophobic residue threonine except *M. marinum* which possess the hydrophobic serine. Thus substituted hydrophobic S193 should have no effect on NAT1 function.

### **F202V, Q210P and D229H variants**

F202 is only conserved in the human NATs while the prokaryotic NATs possess the aliphatic amino acid leucine. This implies that the residue F202 may not be critical for the structure and function within the prokaryotic family but could play important roles in the structure of the human proteins. The substitution of aromatic phenylalanine, F202V with aliphatic valine, V202 is expected to cause an effect on the function of NAT1 due to differences in chemical properties as indicated by the algorithms. The residue Q210 was observed to be conserved in only the human NATs with S210 in the prokaryotic NATs. It could be inferred that Q210 is important for the structure of the eukaryotic NAT. Substituting Q210 with the P210 (with an aliphatic heterocycle) is expected to affect the structure and function of NAT1 and this agrees with the predictions of the algorithms. The alignment revealed that D229 is conserved among five species except for human NAT2 where it is a glutamic acid residue. This is an indication that residue position 229 accommodates acidic residues and could be important for the function of the NAT family and residue D229 particularly for human NAT1. Its substitution by basic H229 would therefore presumably affect the function of NAT1 as predicted by SIFT and POLYPHEN-2.

### **V231G, V235A and T240S variants**

The residue V231 is only conserved in human NATs and may therefore not be critical for the structure and function of the prokaryotic family but could be vital for the human NATs. In the position 231, the prokaryotic NATs possess arginine. The substitution of G231 in NAT1 is therefore expected to affect the

function of the protein as predicted by the algorithms. For wild-type residue V235, the alignment indicates that it only occurs in human NATs while the prokaryotes possess S235, R235 or M235. Substitution of A235 with similar properties to V235 is not expected to affect the function of NAT1 contrary to the prediction of both algorithms. T240, from the sequence alignment, occur in only the human NATs while the prokaryotes possess the hydrophobic residue, A240. Thus substituting hydrophobic T240 by another hydrophobic S240 should not affect NAT1 function as predicted.

### **R242M, N245I, S259R and E264K variants**

The variant R242M had contradictory results from both algorithms. The multiple sequence alignment revealed that R242 is conserved among the human NATs but is a histidine residue in *M. tuberculosis*, *M. bovis*, *M. marinum* and *M. smegmatis*. This implies the position accommodates aromatic or basic residues. Substitution of a hydrophobic residue M242 is thus expected to have an affect on NAT1 protein function as predicted by POLYPHEN-2. The alignment indicates that N245 is conserved only in the human NATs while the bacterial NATs have a glycine residue. The substitution of asparagine with hydrophobic I245 could affect NAT1 function as predicted. The residue position 259 is highly variable according to the multiple sequence alignment. The position is occupied by serine residue in human NAT1, human NAT2 possess a threonine residue, *M. smegmatis* possess lysine residue while *M. tuberculosis*, *M. bovis* and *M. marinum* all possess a valine residue. Hence the sequence alignment suggests that substitution of hydroxylic residues in the human NATs would be accepted without change in NAT1 function. The replacement with basic R259 is expected to affect NAT1 function contrary to the predictions of SIFT and POLYPHEN-2. The variant E264K was one of the variants with contradictory results from the predictions. The residue E264 is conserved among the human NATs, *M. tuberculosis* and *M. bovis*. The other sequences have either D or N residues at position 264. This implies that the position accepts acidic polar residues and hence substituting basic polar K264 would affect the function of NAT1 as predicted by the SIFT algorithm.

### **3.3 Structural analysis**

Variant structures were modelled to determine the solvent accessibility, the Gibbs free energy changes associated with the substitutions and hydrogen bonding or salt bridges associated with the 22 nsSNP residues. The modelling was carried out with MODELLER, Gibbs free energy calculated with FOLDX and the hydrogen bonds determined with CHIMERA.

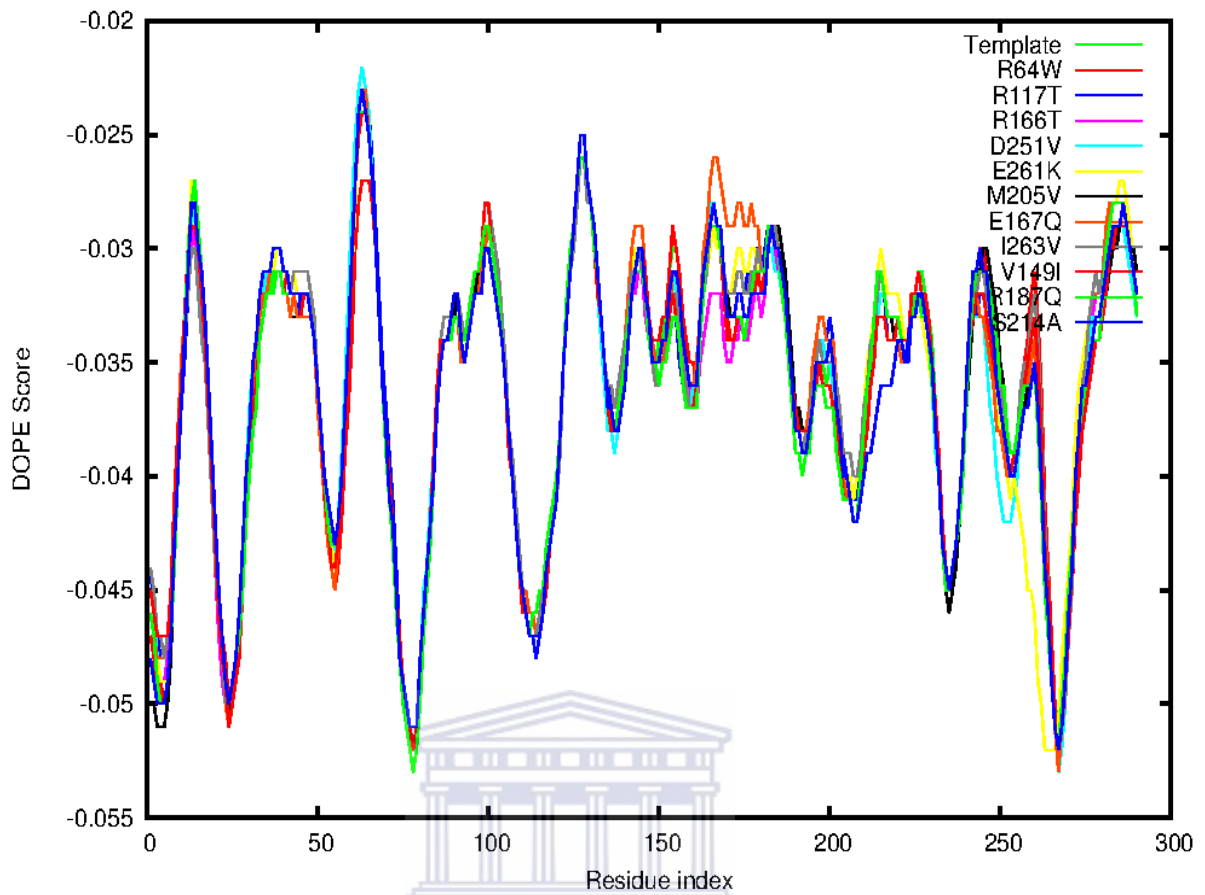
#### **3.3.1 Homology modelling of NAT1**

The native structure of NAT1 has eight  $\alpha$ -helices, 16  $\beta$ -strands and 26 loop regions. We were able to successfully reproduce the native structure of NAT1 with all the eight  $\alpha$ -helices, 16  $\beta$ -strands and 26 loop regions for all the 22 nsSNP substitutions. Structural superimposition of the variant structures with the wild-type did not produce visible structural distortions. The variant structures with associated molpdf, DOPE and GA341 scores are listed in Table 3.3. Near identical profiles were observed between the DOPE plots of the variant and the wild-type structures (Figures 3.3 and 3.4) and suggests that the models reflect the wild-type structure. A general trend can be seen across all plots; variant structures around residues 140 to 215 where most of the nsSNPs are located, show slight variation in energy compared with the wild-type.

**Table 3.3** DOPE score and molpdf values of variant structures

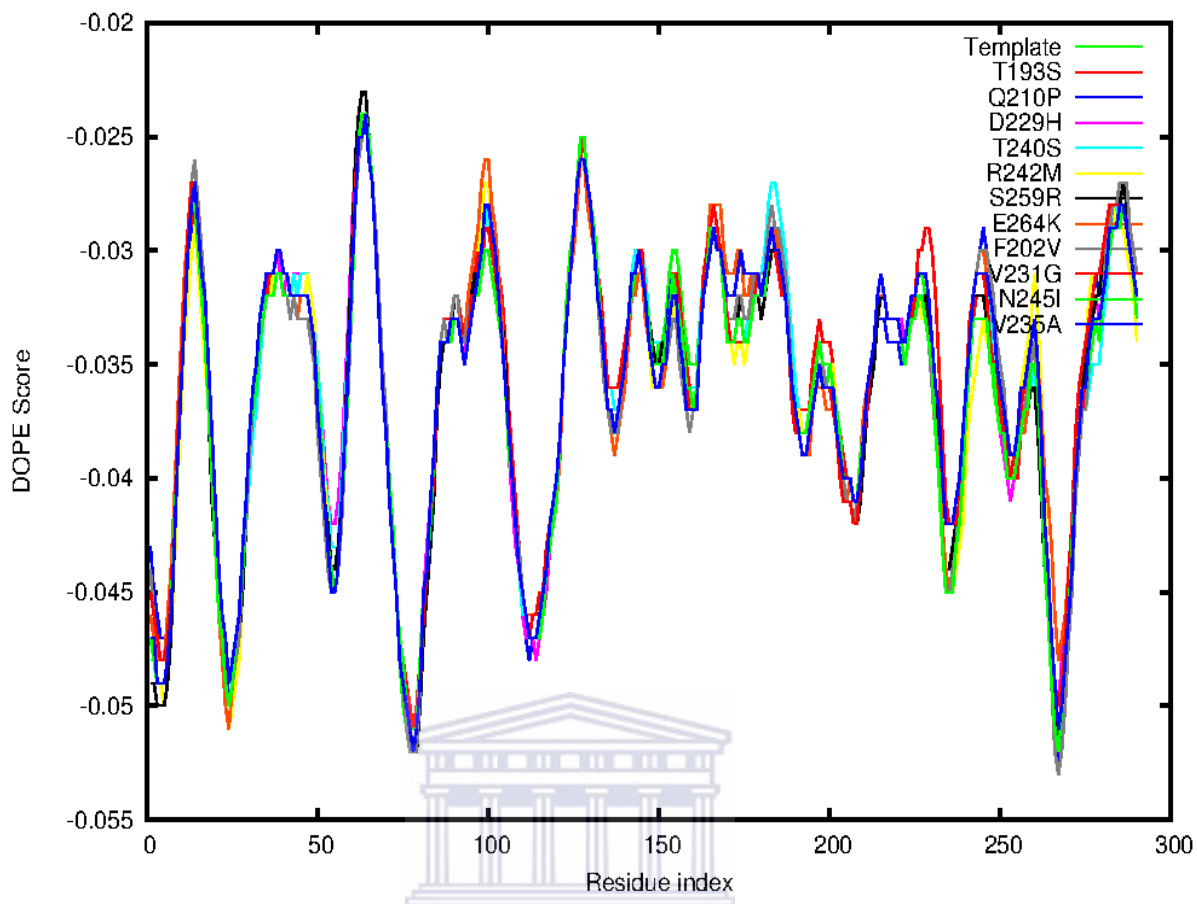
SNP Status	SNP	Filename	molpdf	DOPE Score	GA341 Score
Published	R64W	R64W.pdb	1535.23	-37601.86	1.0
	R117T	R117T.pdb	1612.84	-37319.45	1.0
	V149I	V149I.pdb	1539.37	-37511.91	1.0
	R166T	R166T.pdb	1562.57	-37509.80	1.0
	E167Q	E167Q.pdb	1567.06	-37261.23	1.0
	R187Q	R187Q.pdb	1651.64	-37469.41	1.0
	M205V	M205V.pdb	1667.50	-37590.77	1.0
	S214A	S214A.pdb	1529.88	-37305.33	1.0
	D251V	D251V.pdb	1569.23	-37482.27	1.0
	E261K	E261K.pdb	1546.98	-37559.84	1.0
I263V	I263V.pdb	1608.11	-37254.16	1.0	
Novel	T193S	T193S.pdb	1555.35	-37389.62	1.0
	F202V	F202V.pdb	1564.16	-37272.33	1.0
	Q210P	Q210P.pdb	1588.94	-37355.79	1.0
	D229H	D229H.pdb	1607.88	-37341.77	1.0
	V231G	V231G.pdb	1648.96	-37022.09	1.0
	V235A	V235A.pdb	1569.63	-37132.11	1.0
	T240S	T240S.pdb	1602.17	-37333.45	1.0
	R242M	R242M.pdb	1656.57	-37513.19	1.0
	N245I	N245I.pdb	1680.36	-37460.92	1.0
	S259R	S259R.pdb	1612.42	-37350.60	1.0
	E264K	E264K.pdb	1556.21	-37364.08	1.0

UNIVERSITY of the  
WESTERN CAPE



**Figure 3.3** DOPE plots for published NAT1 variants

Plots of the energy profiles of the variant structures along with the energy profiles of the wild-type structure for published nsSNPs. Each variant energy profile overlaps with the wild-type energy profile. However slight deviations are observed from residues 40 to 50 and from 140 to 270. Each variant is represented by a different colour.



**Figure 3.4** DOPE plots for novel NAT1 variants

Plots of the energy profiles of the variant structures along with the energy profiles of the wild-type structure for published nsSNPs. Each variant energy profile overlaps with the wild-type energy profile. However slight deviations are observed from residues 40 to 50 and from 140 to 270. Each variant is represented by a different colour.

### 3.3.2 PROCHECK evaluation

Modelled structures have many sources of errors including changes in main chain dihedral angles i.e. stereochemical errors. The stereochemical qualities of all variant and wild-type structures were assessed with the PROCHECK suite of programs. All the variant structures of NAT1 (published and novel nsSNPs) had over 90% of the residues in the most favoured region and none in the disallowed regions of the Ramachandran plots (Table 3.4). This is an indication that most of the models are of high quality. The residues in the most favoured region range from 91.6-93.5%.

**Table 3.4** Ramachandran plot results for NAT1 variants

SNP Status	SNP variant	MFR(R) <sup>1</sup>	MFR(%)	AAR (R) <sup>2</sup>	AAR (%)	GAR(R) <sup>3</sup>	GAR(%)	DR(%) <sup>4</sup>
Published	Wild-type	243	93.1	15	5.7	3	1.1	0.0
	R64W	241	92.3	16	6.1	4	1.1	0.0
	R117T	242	92.7	15	5.7	4	1.1	0.0
	V149I	243	93.1	15	5.7	3	1.1	0.0
	R166T	243	93.1	15	5.7	3	1.1	0.0
	E167Q	243	93.1	14	5.4	4	1.5	0.0
	R187Q	242	92.7	16	6.1	3	1.1	0.0
	M205V	241	92.3	18	6.7	2	0.8	0.0
	S214A	243	93.1	15	5.7	3	1.1	0.0
	D251V	241	92.3	17	6.5	3	1.1	0.0
	E261K	242	92.7	16	6.1	3	1.1	0.0
Novel	I263V	242	92.7	16	6.1	3	1.1	0.0
	T193S	242	92.7	16	6.1	3	1.0	0.0
	F202V	242	92.7	15	5.7	4	1.5	0.0
	Q210P	242	93.1	16	6.2	2	0.8	0.0
	D229H	239	91.6	19	7.3	3	1.1	0.0
	V231G	240	92.0	18	6.9	3	1.1	0.0
	V235A	241	92.3	17	6.5	3	1.1	0.0
	T240S	241	92.7	16	6.2	3	1.2	0.0
	R242M	240	92.0	18	6.9	3	1.1	0.0
	N245I	244	93.5	14	5.4	3	1.1	0.0
	S259R	241	92.3	18	6.9	2	0.8	0.0
	E264K	239	91.6	20	7.7	2	0.8	0.0

<sup>1</sup>MFR(R) - Most Favoured Region (Residues), <sup>2</sup>AAR (R) - Additionally Allowed Region (Residues), <sup>3</sup>GAR (R)- Generously Allowed Region (Residues), <sup>4</sup>DR- Disallowed Region.

The variant structure containing the nsSNP E167Q has the least residues while E264K has the greatest residues occurring in the additionally allowed region. No residues were observed in the disallowed regions for any of the variant structures.

### 3.3.3 Solvent accessibility

The accessibility of donor and acceptor atoms to water molecules affect residue interactions. The relative solvent accessibilities of the side chains of the amino acids for residue interactions are therefore vital when considering effects of amino acid substitutions. We used NACCESS to calculate the solvent accessible surface areas of the nsSNP residues (Table 3.5). For most variants we observed only minor changes in solvent accessibilities. However, the published nsSNP I263V and novel nsSNPs D229H,

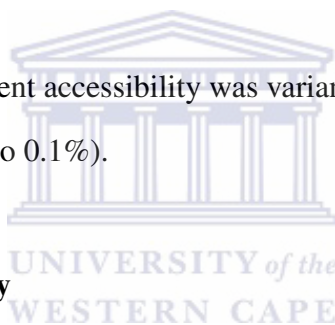
V235G, S259R and E264K (novel) had in excess of 10% change in relative solvent accessibility.

**Table 3.5** Relative solvent accessibilities of NAT1 variants

Published			Novel		
SNP Variant	(WR <sup>1</sup> )/%	(VR <sup>2</sup> )/%	SNP Variant	(WR)/%	(VR)/%
R64W	13.4	8.60	T193S	23.9	22.3
R117T	41.0	45.00	F202V	0.1	3.3
V149I	39.0	34.30	Q210P	12.6	13.4
R166T	22.60	25.30	D229H	67.3	83.3
E167Q	65.0	64.20	V231G	0.7	21.3
R187Q	4.0	4.90	V235A	0.0	0.1
M205V	19.2	15.00	T240S	3.6	3.9
S214A	72.2	72.00	R242M	3.4	5.7
D251V	2.1	1.10	N245I	66.5	67.6
E261K	87.5	88.7	S259R	47.3	80.3
I263V	6.2	18.00	E264K	92.8	78.8

<sup>1</sup>WR - Wild-type residue, <sup>2</sup>VR - Variant residue.

The most significant change in solvent accessibility was variant S259R (47.3% to 80.3%) and the lowest change was V235A variant ( 0.0% to 0.1%).



### 3.3.4 Protein structure stability

The Gibbs free energy change associated with a mutation is a measure of the stability of the protein. A greater than zero Gibbs free energy change is indicative of instability while a negative value implies a stable variant protein structure [93]. The effects of the 22 nsSNPs on NAT1 structure stability was calculated with FOLDX. Stability analysis for NAT1 published nsSNPs indicated that 5 published and 5 novel nsSNPs have Gibbs free energy values greater than zero (Table 3.1).

The negative Gibbs free energy changes associated with published variants R117T, V149I, R187Q, M205V, E261K, I263V and novel variants F202V, Q210P, N245I, T193S, R242M, E264K suggests the structures are relatively stable compared to the wild-type. This means that the above variants should have no stability effects on human NAT1 function. For the above published nsSNPs, all except R187Q were confirmed by both SIFT and POLYPHEN-2 algorithms to have no effects on NAT1 function. Variant R187Q was confirmed by SIFT to have no effect while POLYPHEN-2 indicated that it will affect the function of NAT1. Among the novel nsSNPs with negative Gibbs free energy changes, only T193S was



confirmed by both algorithms to have no functional effects. Variants F202V and Q210P had contrarily predicted functional effects from both algorithms to the Gibbs free energy changes. While the prediction of variant R242M by SIFT confirmed the Gibbs free energy change, POLYPHEN-2 result was in the contrary.

The positive Gibbs free energy associated with the variants R64W, R166T, E167Q, S214A, D251V (published) and D229H, V231G, V235A, T240S, S259R (novel) indicate destabilizing effects to the variant structures. Thus the above variants are expected to affect the function of NAT1. For the published nsSNPs R64W, R166T and D251V, the predictions by both algorithms were in agreement with the Gibbs free energy changes. However, predictions of the functional effects of variants E167Q and S214A contradicted the Gibbs free energy changes. For the novel nsSNPs with positive Gibbs free energy changes, D229H, V231G, V235A were confirmed by both algorithms to have functional effects on NAT1 while T240S and S259R showed contrary results from the Gibbs free energy changes.

However, there is no single threshold in Gibbs free energy change at which we are certain that changes in function have occurred. Bromberg and Burkhard [97] attribute the above to the following three reasons; firstly, the threshold at which a mutation is destabilizing enough to disrupt function by reducing the number of folded active proteins depends on the unfolding energy of the wild-type molecule, which ranges from 3-15 kcal/mol [102]. This indicates that more stable proteins require a larger change to significantly alter the concentration of active molecules. Secondly, without exactly knowing the particular mechanism of protein function, protein destabilization or stabilization events are equally likely to alter function. Finally, destabilizations affecting active sites of the protein may not be manifest in a large  $\Delta G$ , but can still affect function. Bromberg and Burkhard [97] conducted a study to find the threshold energy value that would cause change in function and found that 16 of 19 (84%) stabilizing mutants ( $\Delta G < 0$  kcal/mol; average = -0.73) were functionally disruptive. Similarly, 36 of 45 (80%) destabilizing mutants ( $\Delta G = 1.67$ ) affected function. This implies the direction of stability change alone could not determine functional effect. They [97] demonstrated that in general magnitudes of both destabilizing and stabilizing changes are not very informative. Bromberg and Burkhard [97] concluded that the knowledge of stability is not adequate enough to predict functional effects and that many substitutions that change protein function have no effect on stability.

The variants F202V, Q210P and N245I had stability changes of -1.49, -6.20 and -4.80 kcal/mol respectively i.e. stabilizing but were predicted by both SIFT and POLYPHEN-2 to affect NAT1 function. Similarly T240S and S259R had destabilizing Gibbs free energy changes of 2.47 and 3.50 kcal/mol respectively, yet were predicted as having no effect on function of NAT1. This is similar to the observation by Bromberg and Burkhard [97] that the direction of stability change alone could not determine functional effect. On the other hand D229H, V231G and V235A had destabilizing Gibbs free energy changes (8.33, 6.29, 6.98 kcal/mol respectively) and were predicted to have functional alterations on NAT1. Also T193S had a stabilizing energy change (-8.92 kcal/mol) and predicted by both algorithms as functionally neutral as observed by Bromberg and Burkhard [97]. The nsSNPs R242M and E264K had -7.79 kcal/mol and -15 kcal/mol respectively but both had contradictory results from the two algorithms. While R242M variant was predicted to be neutral by the SIFT algorithm in agreement with the Gibbs free energy change, POLYPHEN-2 predicted it as possibly damaging. Also E264K nsSNP was predicted by SIFT to affect NAT1 function in disagreement with the Gibbs free energy change value but predicted as neutral by POLYPHEN-2. Thus no clear correlation could be seen from the Gibbs free energy changes and functional effects predicted by the two algorithms. Comparing the predictive abilities of both algorithms with each other and/or with the Gibbs free energy changes, SIFT and POLYPHEN-2 perform better than Gibbs free energy changes in discriminating neutral and non-neutral nsSNPs as observed by Bromberg and Burkhard [97].

From the above analysis of the 22 nsSNPs no correlation can be observed between predicted functional effects and the Gibbs free energy (stability) changes calculated.

### **3.3.5 Residue interaction determination (hydrogen bonds and salt bridges)**

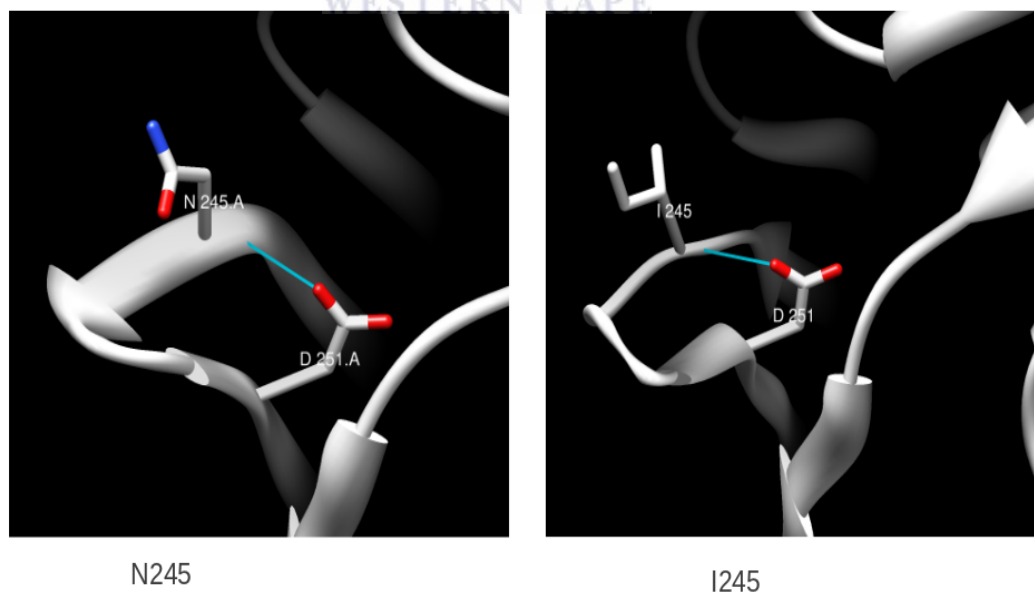
The substitution of one amino acid for another can lead to a loss or gain of residue interactions in proteins. A significant loss of such interactions due to a nsSNP substitution can affect the structure and function of the protein. We used UCSF CHIMERA to calculate the hydrogen bonds or salt bridges for each of the 22 nsSNP residues (Figures 3.5-3.18, Tables 3.1 and 3.2). The eight published nsSNPs that have been structurally analysed by Walraven and colleagues using structural analysis [83] are shown in Appendices B-I. Our analysis of these 8 nsSNPs indicated similar results. However, 3 additional nsSNPs (R117T,



$\beta$ -strands contribute to the protein stability [104]. Replacing the R residue with M residue would result in loss of the side-chain hydrogen bond interactions of R242 except the main-chain hydrogen bonds with basic V231 (V231 N: M242 O, 3.0 Å and M242 N: V231 O, 2.8 Å) (Figure 3.5, right panel). This is expected to affect the dynamics and conformation of the domain II loop, thereby altering NAT1 protein stability and function. This variant is therefore expected to affect NAT1 protein function in agreement with POLYPHEN-2 prediction.

### A1174T (N245I) variant

The residue N245 is situated on  $\beta$ 14 and its main-chain forms a hydrogen bond with D251 which is situated on  $\beta$ 15 (N245 N: D251 OD1, 2.9 Å) and both are in domain III (Figure 3.6, left panel). This interaction is not necessary for the stability of the  $\beta$ -strands [83]. Replacement of polar N residue for hydrophobic I residue maintains this interaction (I245 N: D251 OD1, 2.9 Å) (Figure 3.6, right panel). Hydrophobic interactions are not possible because neighbouring residues are far more than 4.0 Å apart. However placing a hydrophobic residue into a hydrophilic environment will destabilize the protein. Thus the variant N245I is expected to alter the stability of NAT1 as predicted by SIFT and POLYPHEN-2.

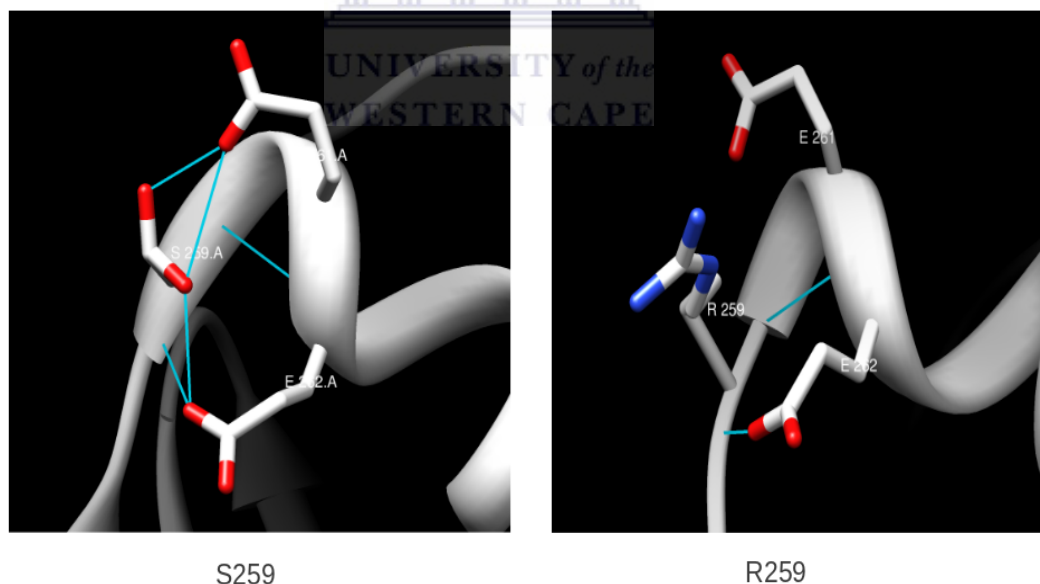


**Figure 3.6** Residue interactions involving wild-type residue N245 and SNP residue I245

A single H-bond formed between wild-type residue N245 and D251 is maintained by the SNP residue I245. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.

### T1217G (S259R) variant

The residue S259 is the first residue on  $\alpha$ 10 of domain III. Its side-chain forms hydrogen bonds with the side-chains of E261 (S259 OG.A: E261 OE1, 2.7 Å and S259 OG.B: E261 OE1, 3.5 Å) and also with E262 (S259 N: E262 OE1, 2.9 Å) (Figure 3.7, left panel). The main-chain of S259 also forms hydrogen bonds with E262 (E262 N: S259 O, 3.1 Å and S259 OG.B: E262 OE1, 2.9 Å). These hydrogen bonds are expected to stabilize  $\alpha$ 10 of domain III. Substitution of the S residue with R residue results in the loss of the side-chain hydrogen bonds except the main-chain hydrogen bonds with E262 (R259 N: E262 OE1, 2.9 Å and E262 N: R259 O, 3.1 Å) (Figure 3.7, right panel). This loss of hydrogen bonds would destabilize the  $\alpha$ 10 of domain III and is reflected in the positive Gibbs free energy change accompanying the substitution (Table 3.2). The replacement of R is expected to contribute to hydrophobic interactions from its side-chain. However, its side-chain is directed away from near-by residues and such interactions may not be possible. The instability due to loss of hydrogen bonds would affect the function of the S259R variant protein. This is in stark contrast to the predictions of SIFT and POLYPHEN-2.

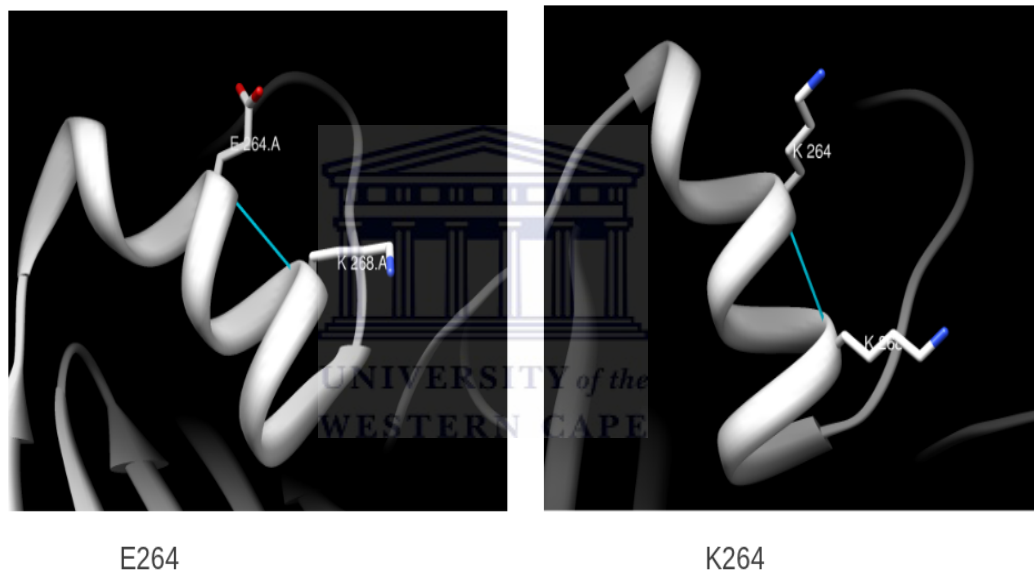


**Figure 3.7** Residue interactions involving wild-type residue S259 and SNP residue R259

Side-chain of wild-type residue S259 forms three salt bridges with E262 and E261 while its main-chain forms one salt bridge and one H-bond with E261 and E262 respectively. Substitution of SNP residue R259 results in the loss of the side-chain salt bridges of S259. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.

### G1230A (E264K) variant

The residue E264 is located on  $\alpha 10$  of the NAT1 crystal structure. The main-chain of E264 forms a hydrogen bond with residue K268 on  $\alpha 10$  (K268 N: E264 O, 3.2 Å) (Figure 3.8, left panel). This hydrogen bond should stabilize the regular secondary structure  $\alpha 10$ . Substitution with residue K results in the hydrogen bond, K268 N: K264 O, 3.2 Å (Figure 3.8, right panel). Since the substitution maintains the interaction between 264 and 268 residues, no structural changes are expected from this substitution. This variant is therefore not expected to have altered function in agreement with POLYPHEN-2 but different from the SIFT prediction.



**Figure 3.8** Residue interactions involving wild-type residue E264 and SNP residue K264

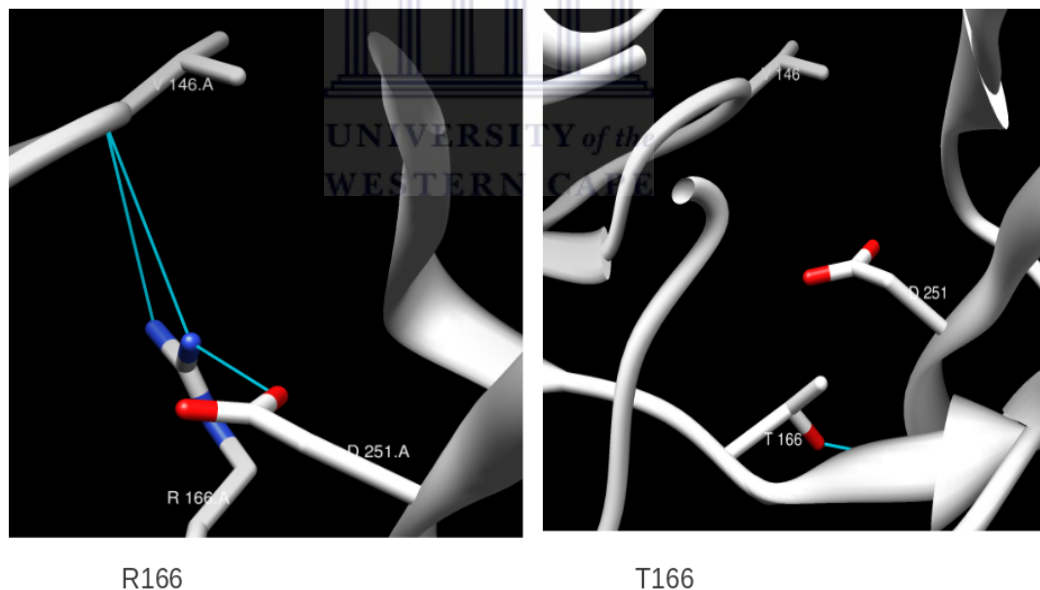
The single main-chain H-bond between wild-type residue E264 and K268 is maintained by SNP residue K264. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.





### G497C (R166T) variant

In the crystal structure of NAT1, R166 is located on the connection between  $\beta 9$  and  $\beta 10$  and forms a salt bridge with D251 (R166 NH1: D251 OD1, 3.0 Å) and two hydrogen bonds with V146 (R166 NH1: V146 O, 3.1 Å and R166 NH2: V146 O, 2.6 Å) in domain II (Figure 3.10 left panel). These bonds should contribute to the stability of the domain loop formed by  $\beta 9$  and  $\beta 10$  strands. However, a replacement with T results in only one bond (T166 OG1: N249 O, 2.6 Å) on  $\beta 15$  in domain III (Figure 3.10 right panel). This loss in hydrogen bonds and the salt bridge due to the substitution is expected to affect the stability of the domain II loop and that of domain III loop. This would subsequently affect the protein structure leading to a functional impairment as predicted. The acetylation status (rapid, intermediate or slow) of this variant has not been reported. From the structural analysis the substitution could affect the structure and is expected to affect the function of the NAT1 protein as revealed by SIFT and POLYPHEN-2 predictions.



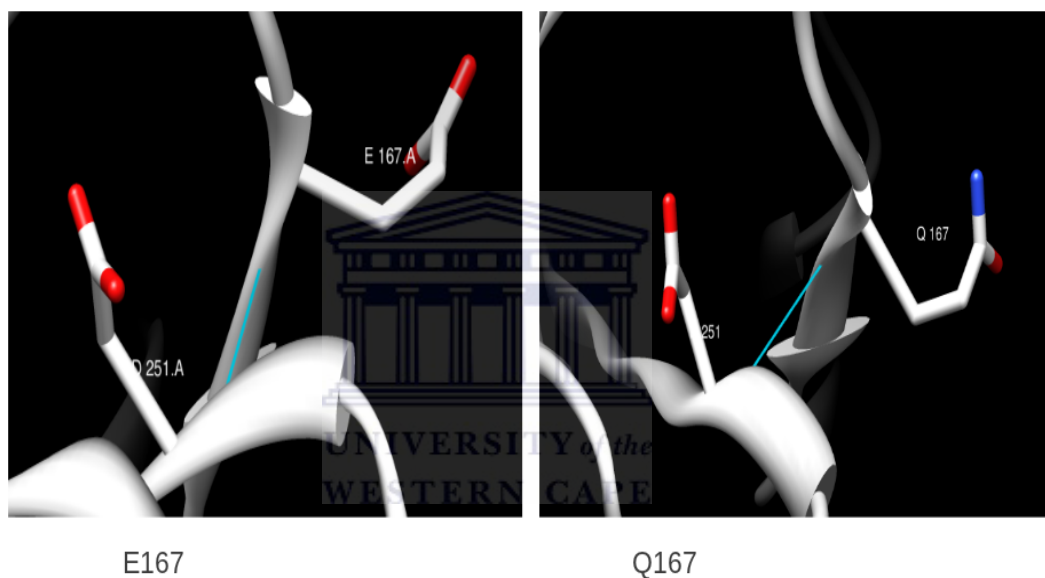
**Figure 3.10** Residue interactions involving wild-type residue R166 and SNP residue T166

The side-chain of wild-type residue R166 forms a salt bridge with main-chain of D251 and two H-bonds with V146. Substitution of SNP residue T166 results in a loss of the three bonds with new H-bond between T166 and N249 (hidden). Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.



### G498C (E167Q) variant

The residue E167 occurs in  $\beta$ 10 and forms a hydrogen bond with D251 on  $\beta$ 15 in domain II (D251 N: E167 O, 2.7 Å) (Figure 3.11 left panel). This bond is expected to maintain the stability of the four anti-parallel  $\beta$ -strands. Replacement with Q maintains the hydrogen bond with D251: (D251 N: Q167 O, 2.7 Å) (Figure 3.11 right panel). The variant structure did not show any observable change in the secondary structures or orientation of the four anti-parallel  $\beta$ -strands. Thus no change in the function of the NAT1 protein is expected and this is in agreement with SIFT and POLYPHEN-2 predictions.

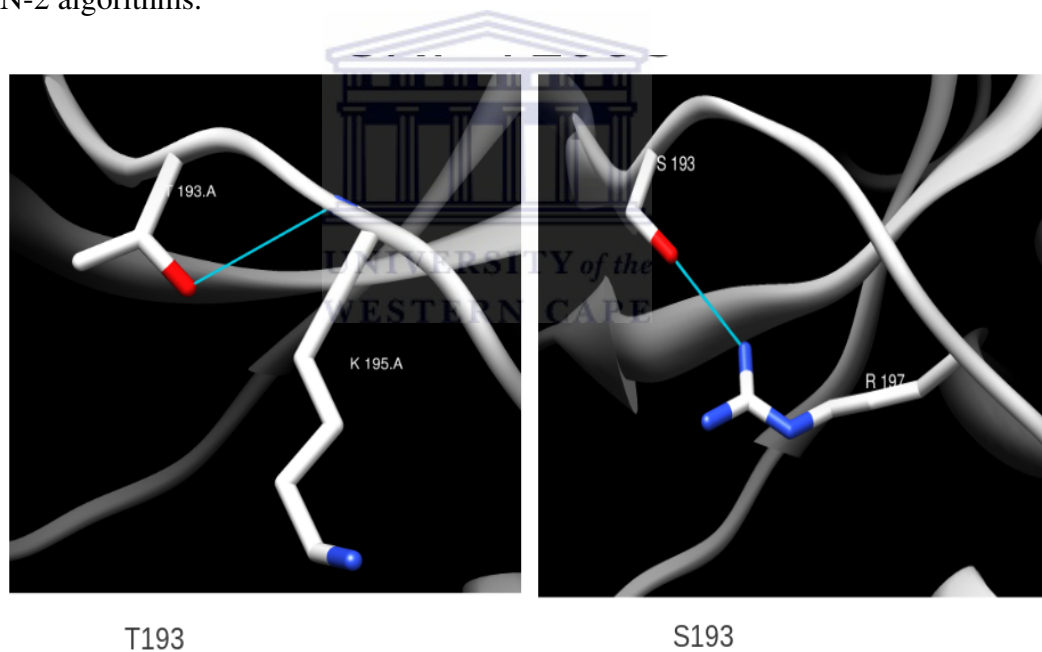


**Figure 3.11** Residue interactions involving wild-type residue E167 and SNP residue Q167

Wild-type residue E167, forms a single H-bond with D251 which is maintained by the replacement of SNP residue Q167. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.

### A1017T (T193S) variant

Oxygen atom (OG1) of T193 which is in the inter-domain region interacts with K195 also in the inter-domain region through the following hydrogen bond: (K195 N: T193 OG1, 3.4 Å) (Figure 3.12 left panel) and in addition forms two hydrogen bonds with solvent molecules. These interactions should contribute to the stability of the inter-domain region. However, the conservative substitution of polar T for polar S results in the following interaction (R197 NH1: S193 OG, 3.1 Å) (Figure 3.12 right panel). The side-chain of residue R197 forms a hydrogen bond with solvent but upon the substitution of T193 with S193 forms the above hydrogen bond instead. Thus the stability of the inter-domain region is expected to remain constant. Thus the integrity of the region is maintained and hence no significant effect to the structure of the NAT1 protein is expected. These results support the predictions of SIFT and POLYPHEN-2 algorithms.

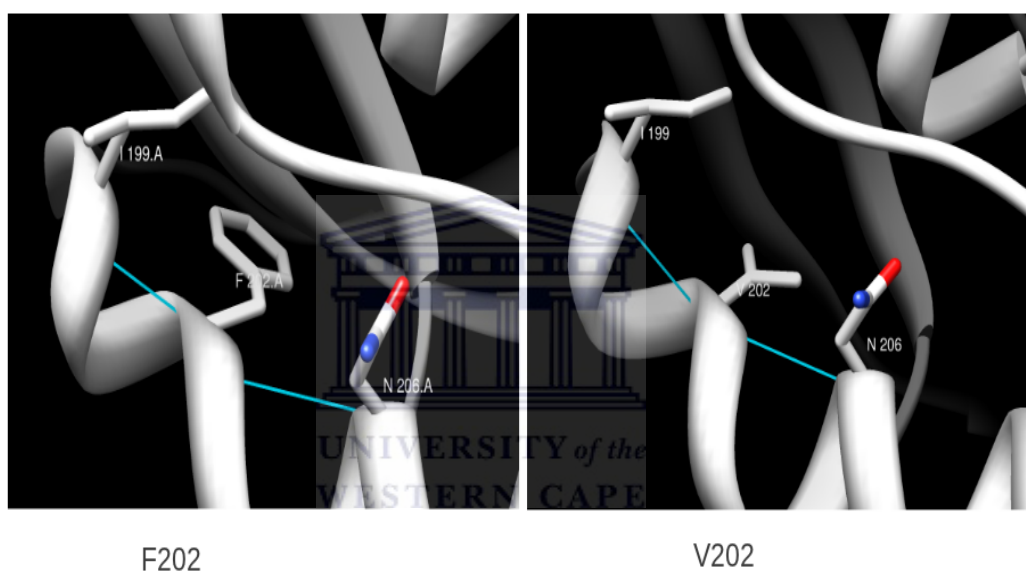


**Figure 3.12** Residue interactions involving wild-type residue T193 and SNP residue S193

The side-chain of wild-type residue T193 forms a H-bond with main-chain of K195. Substitution of S193 leads to a side-chain to side-chain H-bond with residue R197. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.

### T1046G (F202V) variant

The crystal structure of NAT1, shows that residue F202 is located on  $\alpha 8$  with its side-chain oriented towards the core of the protein structure while its main-chain forms a hydrogen bond with I199 in the inter-domain region (F202 N: I199 O, 2.9 Å) and N206 in domain III (N206 N: F202 O, 2.8 Å)(Figure 3.13 left panel), both of which are in  $\alpha 8$  together with the CoA binding residue Y208 (Figure 1.4). Such interactions are maintained by the replacement of F with V (V202 N: I199 O, 2.8 Å and N206 N: V202 O, 2.8 Å) (Figure 3.13 right panel).



**Figure 3.13** Residue interactions involving wild-type residue F202 and SNP residue V202

The wild-type residue F202 forms two main-chain H-bonds with N206 and I199. These main-chain H-bonds are maintained when replaced with SNP residue F202V. The conservative substitution of hydrophobic F with V maintain the hydrophobic interactions within the protein core. However, replacing F with a high propensity for  $\alpha$ -helix formation and therefore stabilises  $\alpha 8$  with V which favours  $\beta$ -strand formation instead [69, 107] resulted in a change in the orientation of the side-chain of N206 and that could affect the orientation of the CoA binding residue Y208 and hence the conformation of the co-factor binding pocket. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.

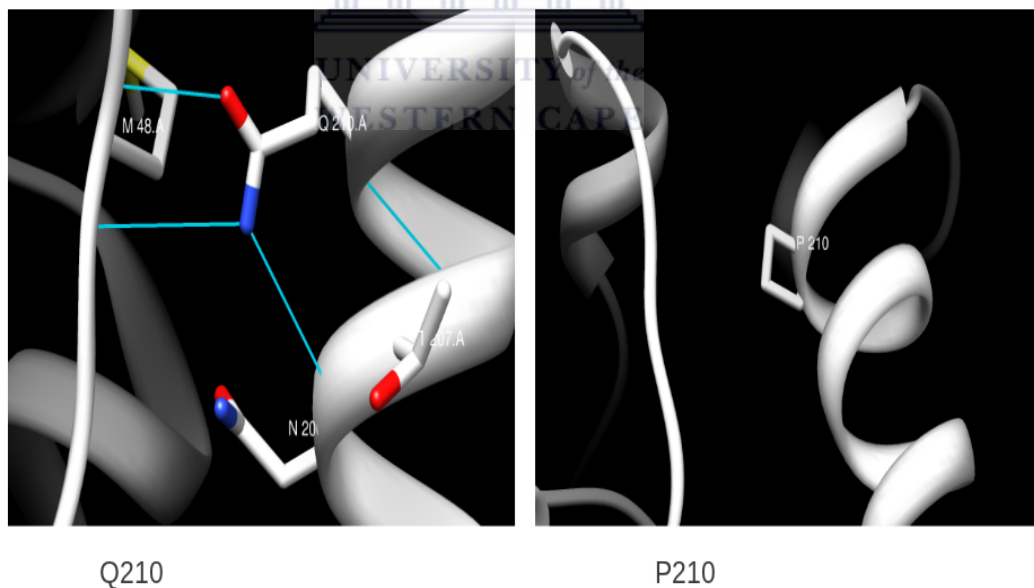
Moreover, the conservative substitution of hydrophobic F with V should maintain the hydrophobic interactions within the protein core. However, replacing F with a larger size and stabilises  $\alpha 8$  with V of smaller size [69, 107] resulted in a change in the orientation of the side-chain of N206 and that could affect the orientation of the CoA binding residue Y208 and hence the conformation of the co-factor binding pocket. This would thus affect the substrate binding and lead to a change in function of NAT1 protein.

Hence this variant is likely to cause a change in the function of NAT1 protein as predicted by SIFT and POLYPHEN-2 analysis.



### A1069C (Q210P) variant

In the crystal structure of NAT1, Q210 is located on  $\alpha 8$  and its side-chain interacts with residue M48 in domain I (M48 N: Q210 OE1, 2.8 Å and Q210 NE2: M48 O, 2.9 Å) and residue T207 in domain III (Q210 N: T207 O, 3.3 Å) which is also in  $\alpha 8$  (Figure 3.14 left panel). The Q210 residue main-chain also interacts with N206 in domain III,  $\alpha 8$  (Q210 NE2: N206 O, 3.0 Å). These interactions should maintain the structural integrity of both domains and the orientation of the CoA binding residue Y208 as well as stabilizing the 5-residue connection between  $\alpha 3$  and  $\alpha 4$ . A non-conservative substitution of the polar Q210 residue which favours  $\alpha$ -helix formation with a hydrophobic cyclic P210 which has a high tendency to form bends or turns [69, 107] results in a complete loss of these interactions (Figure 3.14 right panel). This subsequently would affect the stability of both domains, the connection between  $\alpha 3$  and  $\alpha 4$  and the orientation of the critical residue Y208. This substitution would significantly affect the structure of NAT1 protein and hence its function is expected to be altered as predicted by both SIFT and POLYPHEN-2 algorithms.

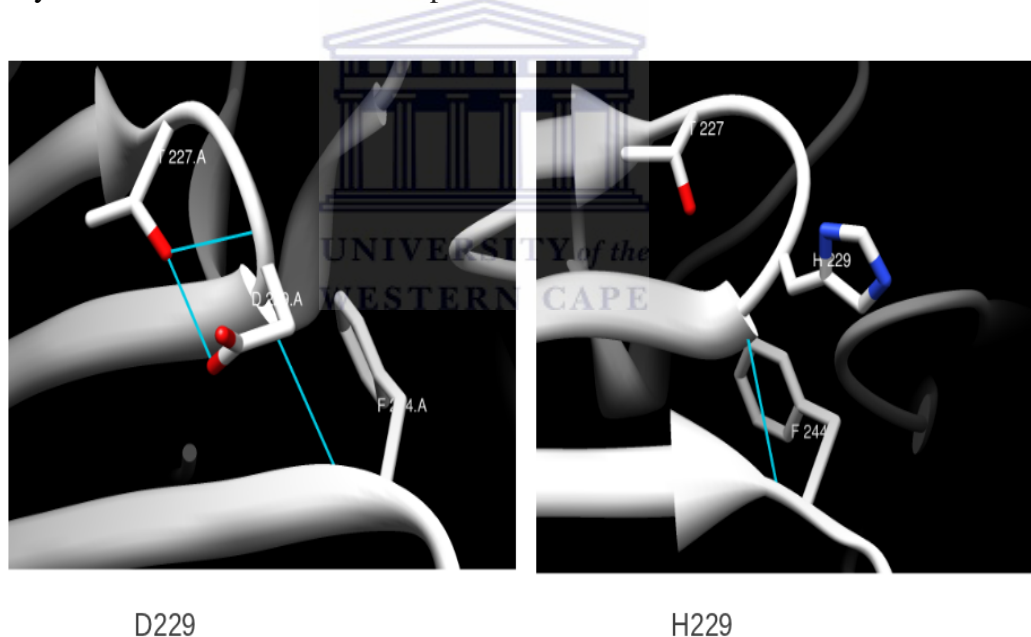


**Figure 3.14** Residue interactions involving wild-type residue Q210 and SNP residue P210

Wild-type residue Q210 forms one main-chain H-bond with T207 and three side-chain H-bonds, one with N206 and two with M48. All these are lost substituting SNP residue P210. Blue denotes nitrogen-atoms, yellow denotes sulphur, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.

### G1125C (D229H) variant

The residue D229 is situated on a loop/turn between anti-parallel  $\beta$ 12 and  $\beta$ 13 strands. D229 forms two hydrogen bonds with residue T227 (D229 N: T227 OG1, 3.3 Å and T227 OG1: D229 OD1, 3.5 Å) on  $\beta$ 12 strand in domain III (Figure 3.15 left panel). It also interacts with residue F244 (F244 N: D229 O, 2.9 Å) on  $\beta$ 14 strand in domain III as well as a solvent molecule. These hydrogen bonds should contribute to the stability of the loop or turn between  $\beta$ 12 and  $\beta$ 13 and stability of the four anti-parallel  $\beta$ -strands of domain III. The substitution of an acidic residue D229 with the basic residue H229 results in a loss of these interactions except for F224 (F244 N: H229 O, 3.0 Å) (Figure 3.15, right panel). The loss of these interactions would affect the stability of the loop or turn between  $\beta$ 12 and  $\beta$ 13 and stability of the four anti-parallel  $\beta$ -strands of domain III. This variant may thus alter the function of NAT1 protein as predicted by the SIFT and POLYPHEN-2 predictions.



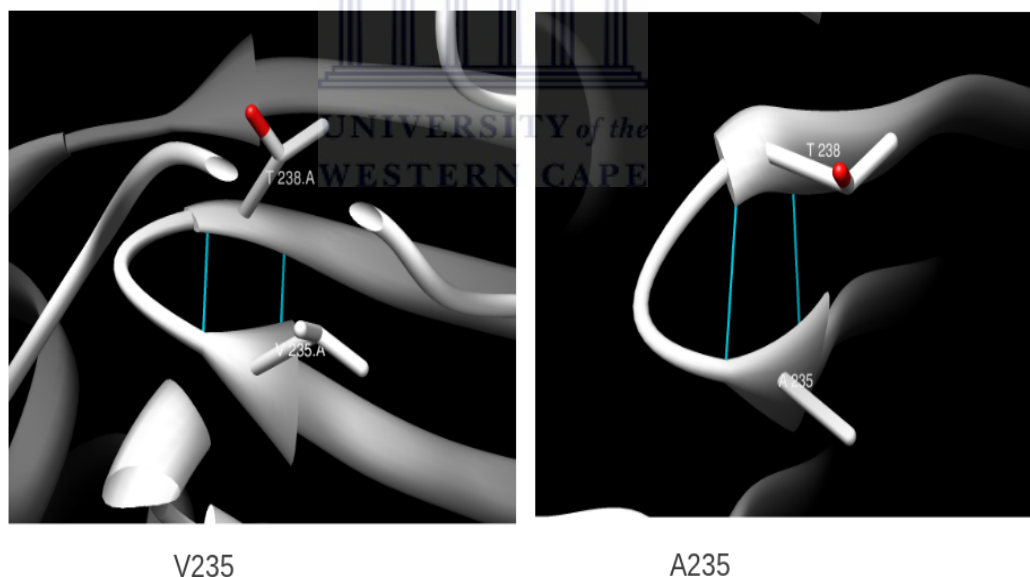
**Figure 3.15** Residue interactions involving wild-type residue D229 and SNP residue H229

D229 forms two H-bonds with T227 and one H-bond with F244. Substituting H229 leads to lost of two bonds except the main-chain H-bond with F244. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.



### T1144C (V235A) variant

The structural analysis indicated that residue V235 located at the end of  $\beta$ 13, interacts with T238 in domain III (V235 N: T238 O, 2.9 Å and T238 N: V235 O, 2.9 Å) (Figure 3.17 left panel) and creates a hydrogen bond with surrounding solvent molecule. Replacement with A residue maintains the bonds with T238 (A235 N: T238 O, 2.8 Å and T238 N: A235 O, 2.8 Å) (Figure 3.17 right panel). The side-chain of V235 is 0.00% exposed (Table 3.5) and oriented to the hydrophobic core contributing to hydrophobic interactions in the core of the protein with close residues L180 and I290. Possible CH... $\pi$ -interaction with H283 may contribute to stabilization of the protein core. Substituted A231 is 0.10% exposed to the hydrophobic core (Table 3.5). This increment in solvent accessible area following the substitution could cause instability in the hydrophobic core by hydrogen bonding with water and affect the orientation of the four anti-parallel  $\beta$ -strands and subsequently affect the CoA residue Y208. This may potentially impact negatively on NAT1 protein function as predicted by both algorithms.



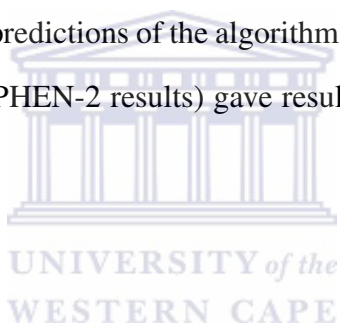
**Figure 3.17** Residue interactions involving wild-type residue V235 and SNP residue A235

Wild-type residue V235 forms two main-chain H-bonds with T238 which are maintained by SNP residue A235. Blue denotes nitrogen-atoms, red represents oxygen-atoms, grey represents main-chains and cyan represents H-bonds.





Biological interpretation of the molecular phenotype in relation to a disease phenotype is a complex matter [108]. It is also difficult to assess how tolerant a specific protein is to structural alteration. Additionally, the natural rigidity of a protein may cause the change in stability that is allowed before severe conformational changes are introduced and on the cellular level biological interpretation is even harder: because it is difficult to predict the role the protein quality control system plays in this tolerance level noting that not all interactions are described at the molecular level [108]. Reumers *et al.* [108], indicated that even if we can predict the molecular effect accurately, this might not necessarily result in a disease phenotype because of functional redundancy of the protein. However, the residue interaction analyses above, have confirmed the functional effects predicted by SIFT and POLYPHEN-2 for the published variants R117T, R166T, E167Q and novel variants T193S, F202V, Q210P, D229H, V231G, V235A, T240S and R242M representing 71% (10/14) of all variants analysed. Analysis of variants N245I and S259R indicated contradictory results to the predictions of the algorithms. Analysis of variants R242M and E264K (with contradictory SIFT and POLYPHEN-2 results) gave results that agreed with the POLYPHEN-2 results.



# Chapter 4

## CONCLUSION

The effects of nsSNPs on the structure and function of proteins are routinely analyzed using SIFT and POLYPHEN-2 prediction algorithms. The false-negative rate of these two algorithms results in as much as 25% of nsSNPs without any evidence of putative impact on protein function. The underlying algorithm implemented in SIFT and POLYPHEN-2 use a combination of sequence features and limited structural information. In the absence of experimental evidence for functional consequences of nsSNPs and the limited accuracy of SIFT and POLYPHEN-2, this study aimed to explore the use of homology modeling including residue interactions, Gibbs free energy changes and solvent accessibility as additional evidence for predicting nsSNP function. An ability to pin-point functionally important nsSNPs in the context of NAT1, an enzyme that metabolizes tuberculosis drugs, will add value to treatment decisions.

This study analyzed 11 nsSNPs previously identified in NAT1 in Caucasians and more recently in a South African mixed ancestry population. However, 8 of these SNPs were recently subjected to structural-functional analysis and provided an internal control for our study methodology. An additional 11 nsSNPs were found specifically in the same South African population group. The structural analysis implemented in this study showed contradictory results for the functions of two nsSNPs compared to the predictions made by SIFT and POLYPHEN-2. The structural analysis also provided additional evidence for two nsSNPs that were in agreement with only one of the two algorithms (POLYPHEN-2). The functions of the remaining 10 nsSNPs were consistent with those predicted by SIFT and POLYPHEN-2.

This study provided the first evaluation of the functional effects of 11 newly characterized nsSNPs on the

NAT1 tuberculosis drug-metabolizing enzyme. The six functionally important nsSNPs will be tested experimentally by creating a SNP construct that will be cloned into an expression vector. These combined computational and experimental studies will advance our understanding of NAT1 structure-function relationships and allow us to interpret the NAT1 genetic polymorphisms in individuals who are slow or fast acetylators.

The results, albeit a small dataset demonstrate that the routinely used algorithms are not without flaws and that improvements in functional prediction of nsSNPs can be obtained by close scrutiny of the molecular interactions of wild type and variant amino acids.



# Bibliography

- [1] E. Vagena, G. Fakis, and S. Boukouvala. Arylamine n-acetyltransferases in prokaryotic and eukaryotic genomes: a survey of public databases. *Current drug metabolism*, 9(7):628–660, September 2008. PMID: 18781915.
- [2] F. Rodrigues-Lima, C. Delomenie, G. H. Goodfellow, D. M. Grant, and J. M. Dupret. Homology modelling and structural analysis of human arylamine n-acetyltransferase NAT1: evidence for the conservation of a cysteine protease catalytic domain and an active-site loop. *Biochemical Journal*, 356(Pt 2):327–334, June 2001. PMID: 11368758 PMCID: PMC1221842.
- [3] M. Blum, D. M. Grant, W. McBride, M. Heim, and U. A. Meyer. Human arylamine n-acetyltransferase genes: isolation, chromosomal localization, and functional expression. *DNA and Cell Biology*, 9(3):193–203, April 1990. PMID: 2340091.
- [4] E. Sim. Arylamine n-acetyltransferases: from structure to function. *Drug metabolism reviews*, 40(3):479, 2008.
- [5] M. Blum, A. Demierre, D. M. Grant, M. Heim, and U. A. Meyer. Molecular mechanism of slow acetylation of drugs and carcinogens in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 88(12):5237–5241, June 1991. PMID: 1675794.
- [6] K. P. Vatsis, K. J. Martell, and W. W. Weber. Diverse point mutations in the human gene for polymorphic n-acetyltransferase. *Proceedings of the National Academy of Sciences of the United States of America*, 88(14):6333–6337, July 1991. PMID: 2068113.

- [7] N. J. Butcher, S. Boukouvala, E. Sim, and R. F. Minchin. Pharmacogenetics of the arylamine n-acetyltransferases. *Pharmacogenomics Journal*, 2(1):30, January 2002.
- [8] D. Hickman, A. Risch, V. Buckle, N. K. Spurr, S. J. Jeremiah, A. McCarthy, and E. Sim. Chromosomal localization of human genes for arylamine n-acetyltransferase. *The Biochemical Journal*, 297 ( Pt 3):441–445, February 1994. PMID: 8110178.
- [9] N. Matas, P. Thygesen, M. Stacey, A. Risch, and E. Sim. Mapping AAC1, AAC2 and AACP, the genes for arylamine n-acetyltransferases, carcinogen metabolising enzymes on human chromosome 8p22, a region frequently deleted in tumours. *Cytogenetics and Cell Genetics*, 77(3-4):290–295, 1997. PMID: 9284941.
- [10] T. Ebisawa and T. Deguchi. Structure and restriction fragment length polymorphism of genes for human liver arylamine n-acetyltransferases. *Biochemical and Biophysical Research Communications*, 177(3):1252–1257, June 1991. PMID: 1676262.
- [11] H. H. Andres, A. J. Klem, L. M. Schopfer, J. K. Harrison, and W. W. Weber. On the active site of liver acetyl-CoA. arylamine n-acetyltransferase from rapid acetylator rabbits (III/J). *J. Biol. Chem.*, 263(16):7521–7527, June 1988.
- [12] D. L. Address, G. A. Howard, and R. S. Birnbaum. Identification of a low molecular weight inhibitor of osteoblast mitogenesis in uremic plasma. *Kidney Int*, 39(5):942–945, May 1991.
- [13] J. M. Dupret and D. M. Grant. Site-directed mutagenesis of recombinant human arylamine n-acetyltransferase expressed in escherichia coli. evidence for direct involvement of cys68 in the catalytic mechanism of polymorphic human NAT2. *The Journal of Biological Chemistry*, 267(11):7381–7385, April 1992. PMID: 1559981.
- [14] D. M. Grant, M. Blum, M. Beer, and U. A. Meyer. Monomorphic and polymorphic human arylamine n-acetyltransferases: a comparison of liver isozymes and expressed products of two cloned genes. *Molecular Pharmacology*, 39(2):184–191, February 1991. PMID: 1996083.
- [15] A. E. Cribb, D. M. Grant, M. A. Miller, and S. P. Spielberg. Expression of monomorphic arylamine

n-acetyltransferase (NAT1) in human leukocytes. *Journal of Pharmacology and Experimental Therapeutics*, 259(3):1241–1246, 1991.

- [16] G. H. Goodfellow, J. M. Dupret, and D. M. Grant. Identification of amino acids imparting acceptor substrate selectivity to human arylamine acetyltransferases NAT1 and NAT2. *The Biochemical Journal*, 348 Pt 1:159–166, May 2000. PMID: 10794727.
- [17] L. Blanc, D. Falzon, C. Fitzpatrick, K. Floyd, I. Garcia, C. Gilpin, P. Glaziou, T. Hiatt, D. Sculier, C. Sismanidis, H. Timimi, and M. Uplekar. Global tuberculosis control 2010, 2010.
- [18] G. Mkele. The role of the pharmacist in TB management. *SA Pharmaceutical Journal*, pages 18–20, March 2010.
- [19] The south african tuberculosis control programme. Technical report, 2004.
- [20] E. M. Streicher, B. Muller, V. Chihota, C. Mlambo, M. Tait, M. Pillay, A. Trollip, K. G. P. Hoek, F. A. Sirgel, N. C. Gey van Pittius, P. D. van Helden, T. C. Victor, and R. M. Warren. Emergence and treatment of multidrug resistant (MDR) and extensively drug-resistant (XDR) tuberculosis in south africa. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(4):686–694, June 2012. PMID: 21839855.
- [21] C. Dye. Tuberculosis 2000-2010: control, but not elimination the comstock lecture. *The International Journal of Tuberculosis and Lung Disease*, 4(12s2):S146–S152, 2000.
- [22] C. Dye and M. A. Espinal. Will tuberculosis become resistant to all antibiotics? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1462):45, 2001.
- [23] A. Pablos-Mendez, M. C. Raviglione, A. Laszlo, N. Binkin, H. L. Rieder, F. Bustreo, D. L. Cohn, C. S. Lambregts van Weesenbeek, S. J. Kim, P. Chaulet, et al. Global surveillance for antituberculosis-drug resistance. *N Engl J Med*, 338:1641–9, 1998.
- [24] A. Van Rie and D. Enarson. XDR tuberculosis: an indicator of public-health negligence. *Lancet*, 368(9547):1554–1556, November 2006. PMID: 17084741.
- [25] Global tuberculosis control: WHO report 2010.

- [26] R. A. Devasia, A. Blackman, C. May, E. Svetlana, T. Smith, N. Hooper, F. Maruri, C. Stratton, A. Shintani, and T. R. Sterling. Fluoroquinolone resistance in mycobacterium tuberculosis: an assessment of MGIT 960, MODS and nitrate reductase assay and fluoroquinolone cross-resistance. *Journal of Antimicrobial Chemotherapy*, 63(6):1173–1178, January 2009.
- [27] K. M. Kam, C. W. Yip, T. L. Cheung, H. S. Tang, O. C. Leung, and M. Y. Chan. Stepwise decrease in moxifloxacin susceptibility amongst clinical isolates of multidrug-resistant mycobacterium tuberculosis: correlation with ofloxacin susceptibility. *Microbial drug resistance (Larchmont, N.Y.)*, 12(1):7–11, 2006. PMID: 16584301.
- [28] G. P. Morlock, B. Metchock, D. Sikes, J. T Crawford, and R. C. Cooksey. ethA, inhA, and katG loci of ethionamide-resistant clinical mycobacterium tuberculosis isolates. *Antimicrobial agents and chemotherapy*, 47(12):3799–3805, December 2003. PMID: 14638486.
- [29] B. Muller, E. M. Streicher, K. G. P. Hoek, M. Tait, A. Trollip, M. E. Bosman, G. J. Coetzee, E. M. Chabula-Nxiweni, E. Hoosain, N. C. Gey van Pittius, T. C. Victor, P. D. van Helden, and R. M. Warren. inhA promoter mutations: a gateway to extensively drug-resistant tuberculosis in south africa? *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease*, 15(3):344–351, March 2011. PMID: 21333101.
- [30] F. A. Sirgel, M. Tait, R. M. Warren, E. M. Streicher, E. C. Bottger, P. D. van Helden, N. C. Gey van Pittius, G. Coetzee, E. Y. Hoosain, M. Chabula-Nxiweni, C. Hayes, T. C. Victor, and A. Trollip. Mutations in the rrs A1401G gene and phenotypic resistance to amikacin and capreomycin in mycobacterium tuberculosis. *Microbial drug resistance (Larchmont, N.Y.)*, 18(2):193–197, April 2012. PMID: 21732736.
- [31] L. E. Via, Sang-Nae Cho, S. Hwang, H. Bang, S. K. Park, H. S. Kang, D. Jeon, S. Y. Min, T. Oh, Y. Kim, Y. M. Kim, V. Rajan, S. Y. Wong, I. C. Shamputa, M. Carroll, L. Goldfeder, S. A. Lee, S. M. Holland, S. Eum, H. Lee, and C. E. Barry. Polymorphisms associated with resistance and cross-resistance to aminoglycosides and capreomycin in mycobacterium tuberculosis



- isolates from south korean patients with drug-resistant tuberculosis. *Journal of clinical microbiology*, 48(2):402–411, February 2010. PMID: 20032248.
- [32] D. A. Evans. N-acetyltransferase. *Pharmacology & Therapeutics*, 42(2):157–234, 1989. PMID: 2664821.
- [33] T. Deguchi. Sequences and expression of alleles of polymorphic arylamine n-acetyltransferase of human liver. *The Journal of Biological Chemistry*, 267(25):18140–18147, September 1992. PMID: 1381364.
- [34] K. P. Vatsis and W. W. Weber. Structural heterogeneity of caucasian n-acetyltransferase at the NAT1 gene locus. *Archives of Biochemistry and Biophysics*, 301(1):71–76, February 1993. PMID: 8442668.
- [35] M. A. Doll, W. Jiang, A. C. Deitz, T. D. Rustan, and D. W. Hein. Identification of a novel allele at the human NAT1 acetyltransferase locus. *Biochemical and Biophysical Research Communications*, 233(3):584–591, April 1997. PMID: 9168895.
- [36] J. H. de Leon, K. P. Vatsis, and W. W. Weber. Characterization of naturally occurring and recombinant HumanN-Acetyltransferase variants encoded byNAT1\*. *Molecular Pharmacology*, 58(2):288–299, 2000.
- [37] H. J. Lin, N. M. Probst-Hensch, N. C. Hughes, G. T. Sakamoto, A. D. Louie, I. H. Kau, B. K. Lin, D. B. Lee, J. Lin, H. D. Frankl, E. R. Lee, S. Hardy, D. M. Grant, and R. W. Haile. Variants of n-acetyltransferase NAT1 and a case-control study of colorectal adenomas. *Pharmacogenetics*, 8(3):269–281, June 1998. PMID: 9682272.
- [38] F. S. Collins, L. D. Brooks, and A. Chakravarti. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, 8(12):1229–1231, December 1998. PMID: 9872978.
- [39] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S.T. Thomas, and D. N. Cooper. The human gene mutation database: 2008 update. 1(1):13–13. PMID: 19348700 PMCID: 2651586.

- [40] S. Sunyaev, V. Ramensky, and P Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in Genetics: TIG*, 16(5):198–200, May 2000. PMID: 10782110.
- [41] D. M. Grant. Detoxification pathways in the liver. *Journal of Inherited Metabolic Disease*, 14(4):421–430, July 1991.
- [42] S. Ozawa, M. Abu-Zeid, Y. Kawakubo, S. Toyama, Y. Yamazoe, and R. Kato. Monomorphic and polymorphic isozymes of arylamine n-acetyltransferases in hamster liver: purification of the isozymes and genetic basis of n-acetylation polymorphism. *Carcinogenesis*, 11(12):2137–2144, December 1990. PMID: 2265466.
- [43] I. B. Glowinski, H. E. Radtke, and W. W. Weber. Genetic variation in n-acetylation of carcinogenic arylamines by human and rabbit liver. *Molecular pharmacology*, 14(5):940–949, September 1978. PMID: 714030.
- [44] A. J. Kilbane, T. Petroff, and W. W. Weber. Kinetics of acetyl CoA: arylamine n-acetyltransferase from rapid and slow acetylator human liver. *Drug metabolism and disposition: the biological fate of chemicals*, 19(2):503–507, April 1991. PMID: 1676662.
- [45] A. Ward, D. Hickman, J. W. Gordon, and E. Sim. Arylamine n-acetyltransferase in human red blood cells. *Biochemical pharmacology*, 44(6):1099–1104, September 1992. PMID: 1329759.
- [46] W. W. Weber and K. P. Vatsis. Individual variability in p-aminobenzoic acid n-acetylation by human n-acetyltransferase (NAT1) of peripheral blood. *Pharmacogenetics*, 3(4):209–212, August 1993. PMID: 8220441.
- [47] D. A. Bell, A. F. Badawi, N. P. Lang, K. F. Ilett, F. F. Kadlubar, and A. Hirvonen. Polymorphism in the n-acetyltransferase 1 (NAT1) polyadenylation signal: association of NAT1\*10 allele with higher n-acetylation activity in bladder and colon tissue. *Cancer Research*, 55(22):5226–5229, November 1995. PMID: 7585580.
- [48] M. A. Payton and E. Sim. Genotyping human arylamine n-acetyltransferase type 1 (NAT1): the

identification of two novel allelic variants. *Biochemical Pharmacology*, 55(3):361–366, February 1998. PMID: 9484803.

- [49] N. C. Hughes, S. A. Janezic, K. L. McQueen, M. A. S. Jewett, T. Castranio, D. A. Bell, and D. M. Grant. Identification and characterization of variant alleles of human acetyltransferase NAT1 with defective function using p-aminosalicylate as an in-vivo and in-vitro probe. *Pharmacogenetics and Genomics*, 8(1):55, 1998.
- [50] C. Bruhn, J. Brockmoller, I. Cascorbi, I. Roots, and H. H. Borchert. Correlation between genotype and phenotype of the human arylamine n-acetyltransferase type 1 (NAT1). *Biochemical Pharmacology*, 58(11), December.
- [51] N. J. Butcher, K. F. Ilett, and R. F. Minchin. Functional polymorphism of the human arylamine n-acetyltransferase type 1 gene caused by C190T and G560A mutations. *Pharmacogenetics*, 8(1):67–72, February 1998. PMID: 9511183.
- [52] C. Bouchardy, K. Mitrunen, H. Wikman, K. Husgafvel-Pursiainen, P. Dayer, S. Benhamou, A. Hirvonen, et al. N-acetyltransferase NAT1 and NAT2 genotypes and lung cancer risk. *Pharmacogenetics*, 8(4):291, 1998.
- [53] Y. Zhu, J. C. States, Y. Wang, and D. W. Hein. Functional effects of genetic polymorphisms in the n-acetyltransferase 1 coding and 3' untranslated regions. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 91(2):77–84, February 2011. PMID: 21290563.
- [54] Y. Zhu and D. W. Hein. Functional effects of single nucleotide polymorphisms in the coding region of human n-acetyltransferase 1. *The pharmacogenomics journal*, 8(5):339–348, October 2008. PMID: 17909564 PMCID: 2575040.
- [55] D. W. Hein, M. A. Doll, T. D. Rustan, K. Gray, Y. Feng, R. J. Ferguson, and D. M. Grant. Metabolic activation and deactivation of arylamine carcinogens by recombinant human NAT1 and polymorphic NAT2 acetyltransferases. *Carcinogenesis*, 14(8):1633–1638, August 1993. PMID: 8353847.
- [56] H. Wu, L. Dombrovsky, W. Tempel, F. Martin, P. Loppnau, G. H. Goodfellow, D. M. Grant, and A. N. Plotnikov. Structural basis of substrate-binding specificity of human arylamine n-

acetyltransferases. *The Journal of Biological Chemistry*, 282(41):30189–30197, October 2007. PMID: 17656365.

- [57] R. F. Minchin. Acetylation of p-aminobenzoylglutamate, a folic acid catabolite, by recombinant human arylamine n-acetyltransferase and u937 cells. *The Biochemical Journal*, 307 ( Pt 1):1–3, April 1995. PMID: 7717963.
- [58] A. Ward, M. J. Summers, and E. Sim. Purification of recombinant human n-acetyltransferase type 1 (NAT1) expressed in e. coli and characterization of its potential role in folate metabolism. *Biochemical Pharmacology*, 49(12):1759–1767, June 1995. PMID: 7598738.
- [59] D. W. Hein, S. Boukouvala, D. M. Grant, R. F. Minchin, and E. Sim. Changes in consensus arylamine n-acetyltransferase gene nomenclature. *Pharmacogenetics and Genomics*, 18(4):367–368, April 2008. PMID: 18334921.
- [60] W. W. Weber and D. W. Hein. N-acetylation pharmacogenetics. *Pharmacological Reviews*, 37(1):25–79, March 1985. PMID: 2860675.
- [61] D. M. Grant, M. Blum, and U. A. Meyer. Polymorphisms of n-acetyltransferase genes. *Xenobiotica*, 22(9-10):1073–1081, January 1992.
- [62] A. E. Cribb, H. Nakamura, D. M. Grant, M. A. Miller, and S. P. Spielberg. Role of polymorphic and monomorphic human arylamine n-acetyltransferases in determining sulfamethoxazole metabolism. *Biochemical Pharmacology*, 45(6):1277–1282, March 1993.
- [63] A. J. Cozzzone. *Proteins: Fundamental chemical properties*. John Wiley & Sons, Ltd., France, 2002.
- [64] M. J. Plevin, D. L. Bryce, and J. Boisbouvier. Direct detection of CH/ $\pi$  interactions in proteins. *Nature Chemistry*, 2(6):466–471, 2010.
- [65] G. D. Rose and R. Wolfenden. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annual review of biophysics and biomolecular structure*, 22:381–415, 1993. PMID: 8347995.

- [66] K.D. Schwenke, D. Mobius, and R. Miller. Proteins: Some principles of classification and structure. In *Proteins at Liquid Interfaces*, volume Volume 7, pages 1–50. Elsevier, 1998.
- [67] Puri. *Textbook Of Biochemistry*. Elsevier India, January 2005.
- [68] B. Mehta and M. Mehta. *Organic Chemistry*. PHI Learning Pvt. Ltd., March 2005.
- [69] A. J. Cozzone. Proteins: Fundamental chemical properties.
- [70] P. C. Ng and S. Henikoff. Accounting for human polymorphisms predicted to affect protein function. *Genome Research*, 12(3):436–446, March 2002. PMID: 11875032 PMCID: 155281.
- [71] N. J. Butcher, A. Arulpragasam, and R. F. Minchin. Proteasomal degradation of n-acetyltransferase 1 is prevented by acetylation of the active site cysteine: a mechanism for the slow acetylator phenotype and substrate-dependent down-regulation. *The Journal of Biological Chemistry*, 279(21):22131–22137, May 2004. PMID: 15039438.
- [72] A. J. Fretland, M. A. Doll, Y. Zhu, L. Smith, M. A. Leff, and D. W. Hein. Effect of nucleotide substitutions in n-acetyltransferase-1 on n-acetylation (deactivation) and o-acetylation (activation) of arylamine carcinogens: implications for cancer predisposition. *Cancer Detection and Prevention*, 26(1):10–14, 2002. PMID: 12088197.
- [73] F. Liu, N. Zhang, X. Zhou, P. E. Hanna, C. R. Wagner, D. M. Koepp, and K. J. Walters. Arylamine n-acetyltransferase aggregation and constitutive ubiquitylation. *Journal of Molecular Biology*, 361(3):482–492, August 2006. PMID: 16857211.
- [74] Y. Zhu, C. States, and D. W. Hein. Functional characterization of single nucleotide polymorphisms of human n- acetyltransferase 1 in mammalian cells. *AACR Meeting Abstracts*, 2004(1):674, March 2004.
- [75] X. Zhangwei, X. Jianming, M. Qiao, and X. Xinhua. N-Acetyltransferase-1 gene polymorphisms and correlation between genotype and its activity in a central chinese han population. *Clinica Chimica Acta*, 371(1-2):85–91, September 2006.

- [76] Y. M. Di, E. Chan, M. Q. Wei, Jun-Ping Liu, and Shu-Feng Zhou. Prediction of deleterious non-synonymous single-nucleotide polymorphisms of human uridine diphosphate glucuronosyl-transferase genes. *The AAPS journal*, 11(3):469–480, September 2009. PMID: 19572200.
- [77] R. J. Livingston, A. von Niederhausen, A. G. Jegga, D. C. Crawford, C. S. Carlson, M. J. Rieder, S. Gowrisankar, B. J. Aronow, R. B. Weiss, and D. A. Nickerson. Pattern of sequence variation across 213 environmental response genes. *Genome research*, 14(10A):1821–1831, October 2004. PMID: 15364900.
- [78] T. Xi, I. M. Jones, and H. W. Mohrenweiser. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*, 83(6):970–979, June 2004. PMID: 15177551.
- [79] E. Mathe, M. Olivier, S. Kato, C. Ishioka, P. Hainaut, and S. V. Tavtigian. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Research*, 34(5):1317–1325, 2006. PMID: 16522644 PMCID: 1390679.
- [80] M. M. Johnson, J. Houck, and C. Chen. Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response. *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 14(5):1326–1329, May 2005. PMID: 15894696.
- [81] Y. L. Yip, H. Scheib, A. V. Diemand, A. Gattiker, L. M. Famiglietti, E. Gasteiger, and A. Bairoch. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Human Mutation*, 23(5):464–470, May 2004. PMID: 15108278.
- [82] South african national tuberculosis management guidelines, 2009.
- [83] J. M. Walraven, J. O. Trent, and D. W. Hein. Structure-function analyses of single nucleotide

- polymorphisms in human n-acetyltransferase 1. *Drug metabolism reviews*, 40(1):169–184, 2008. PMID: 18259988 PMCID: 2265210.
- [84] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe, A. S. Kondrashov, and P. Bork. Prediction of deleterious human alleles. *Human Molecular Genetics*, 10(6):591–597, March 2001. PMID: 11230178.
- [85] P. C. Ng and S. Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, July 2003. PMID: 12824425 PMCID: 168916.
- [86] V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30(17):3894–3900, 2002.
- [87] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–2948, November 2007. PMID: 17846036.
- [88] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991. PMID: 2017436.
- [89] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, December 1993. PMID: 8254673.
- [90] T. Williams and C. Kelley. Gnuplot: an interactive plotting program. <http://gnuplot.sourceforge.net>, March 2010.
- [91] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283–291, April 1993.
- [92] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, October 2004. PMID: 15264254.



- [93] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server issue):W382–388, July 2005. PMID: 15980494.
- [94] S. J. Hubbard and J. M. Thornton. NACCESS. *Department of Biochemistry and Molecular Biology, University College London, 1993.*
- [95] V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30(17):3894–3900, September 2002. PMID: 12202775.
- [96] W. S. J. Valdar. Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics*, 48(2):227–241, 2002.
- [97] Y. Bromberg and B. Rost. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, 35(11):3823–3835, June 2007.
- [98] A. V. Efimov and E. V. Brazhnikov. Relationship between intramolecular hydrogen bonding and solvent accessibility of side-chain donors and acceptors in proteins. *FEBS Letters*, 554(3):389–393, November 2003. PMID: 14623099.
- [99] T. R. Rebbeck, M. Spitz, and X. Wu. Assessing the function of genetic variants in candidate gene association studies. *Nature reviews. Genetics*, 5(8):589–597, August 2004. PMID: 15266341.
- [100] R. Yamada, T. Tanaka, Y. Ohnishi, K. Suematsu, M. Minami, T. Seki, M. Yukioka, A. Maeda, N. Murata, O. Saiki, R. Teshima, O. Kudo, K. Ishikawa, A. Ueyosi, H. Tateishi, M. Inaba, H. Goto, Y. Nishizawa, S. Tohma, T. Ochi, K. Yamamoto, and Y. Nakamura. Identification of 142 single nucleotide polymorphisms in 41 candidate genes for rheumatoid arthritis in the japanese population. *Human genetics*, 106(3):293–297, March 2000. PMID: 10798357.
- [101] M. K. Leabman, C. C. Huang, J. DeYoung, E. J. Carlson, T. R. Taylor, M. de la Cruz, S. J. Johns, D. Stryke, M. Kawamoto, T. J. Urban, D. L. Kroetz, T. E. Ferrin, A. G. Clark, N. Risch, I. Herskowitz, and K. M. Giacomini. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5896–5901, May 2003. PMID: 12719533.



- [102] C. N. Pace, G. R. Grimsley, and J. M. Scholtz. Protein stability.
- [103] J. M. Walraven, J. O. Trent, and D. W. Hein. Structure-function analyses of single nucleotide polymorphisms in human n-acetyltransferase 1. *Drug Metabolism Reviews*, 40(1):169–184, 2008. PMID: 18259988.
- [104] J. M. Walraven, J. O. Trent, and D. W. Hein. Computational and experimental analyses of mammalian arylamine n-acetyltransferase structure and function. *Drug metabolism and disposition: the biological fate of chemicals*, 35(6):1001–1007, June 2007. PMID: 17371801 PMCID: 2085365.
- [105] B. Forood, E. J. Feliciano, and K. P. Nambiar. Stabilization of alpha-helical structures in short peptides via end capping. *Proceedings of the National Academy of Sciences*, 90(3):838–842, January 1993.
- [106] S. Ohsako and T. Deguchi. Cloning and expression of cDNAs for polymorphic and monomorphic arylamine n-acetyltransferases from human liver. *The Journal of Biological Chemistry*, 265(8):4630–4634, March 1990. PMID: 1968463.
- [107] Debajyoti Das. *Biochemistry*. Academic Publishers, 1978.
- [108] J. Reumers, J. Schymkowitz, and F. Rousseau. Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations. *BMC Bioinformatics*, 10(Suppl 8):S9, 2009.

# Appendix A

## Command line options

### A.1 UCSF CHIMERA

By clicking the “Favorites” option in CHIMERA, “Command Line” option was then selected. In the command line, each residue was selected by “select: # of residue” and the side-chains shown from the “Actions” option by selecting “Atoms/Bonds”, then “Show”. The H-bonding of the wild-type residues in the wild-type structure (2PQT of NAT1) and the variants residues in the modeled structures were calculated with the “FindHBond” function by choosing “Structure analysis” from “Tools”. In the “H-bond Parameters” dialog box, “Relax H-bonds constraints” was by 0.4Å and 20.0 degrees. Only find H-bonds with at least one end selected and write information to reply log, were checked. From the reply log file, residues that form H-bonds with the residue of interest were also selected and shown. This then indicated the hydrogen bonding of the residue with its surrounding residues. The best orientations of the images were saved.

### A.2 FOLDX

The “RepairPDB” command was used to repair all residues (the wild-type and of each variant pdb structure) with bad torsion angles, or Vander Waals’ clashes, or total energy at a temperature of 298°C, pH of 7, ionic strength of 0.05 and Van der Waals radius of 2. The above parameters were contained in the

“run.txt” file together with the “RepairPDB” command with “list.txt” containing the PDB file of each variant or wild-type structure. This was run from the command line by typing “foldx”, then selecting “3” and typing “run.txt”. The new PDB structure files obtained were then added to the list2.txt file. The free energy changes were then calculated with the “Stability” command with the above conditions/options.

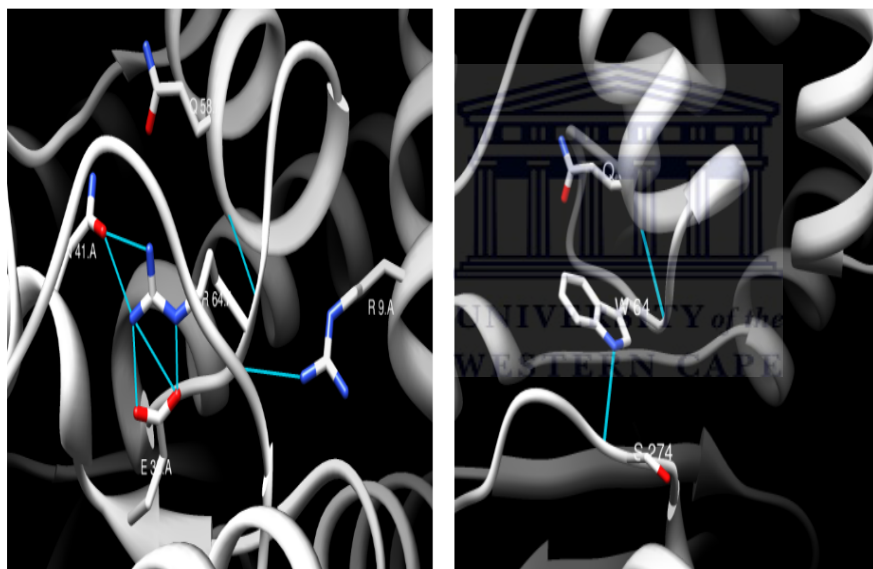
### **A.3 NACCESS**

The NACCESS program was installed locally and used. The program was run from the linux terminal with “naccess variant/wild-type PDBCode.pdb”. This produced a PDBCode.log, PDBCode.rsa and PDBCode.asa files containing the run information, relative accessibilities and absolute solvent accessibilities respectively of the atoms and residues in each structure.



# Appendix B

Residue interactions of wild-type residue R64 and SNP residue W64. ....



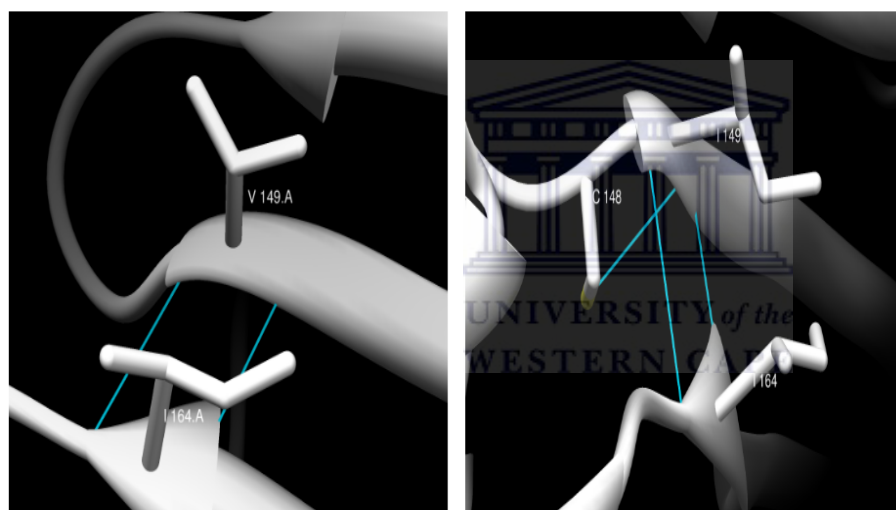
R64

W64

Left panel- wild-type; Right panel- variant. Seven hydrogen bonds with the wild-type results in two bonds substituting tryptophan for arginine.

# Appendix C

Residue interactions of wild-type residue V149 and SNP residue I149 .....



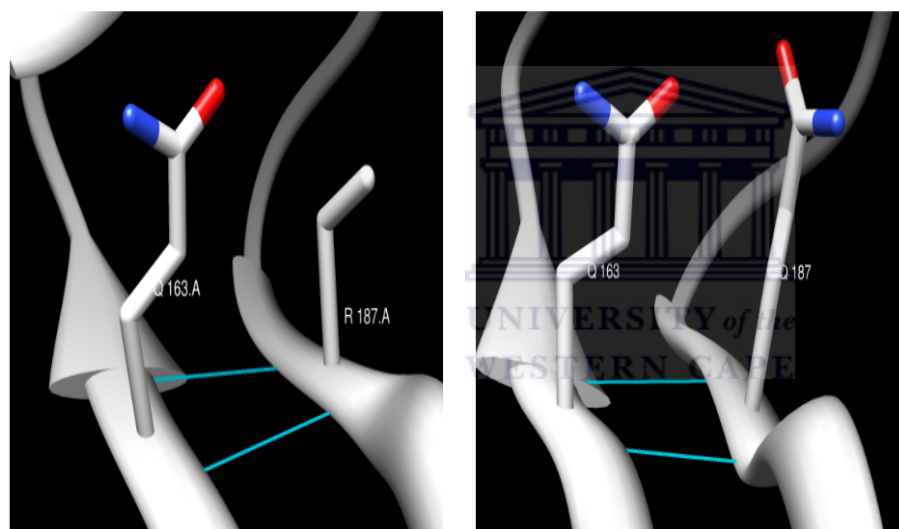
V149

I149

Left panel- wild-type; Right panel- variant. Additional bond formed following isoleucine substitution.

# Appendix D

Residue interactions of wild-type residue R187 and SNP residue Q187 . . . . .



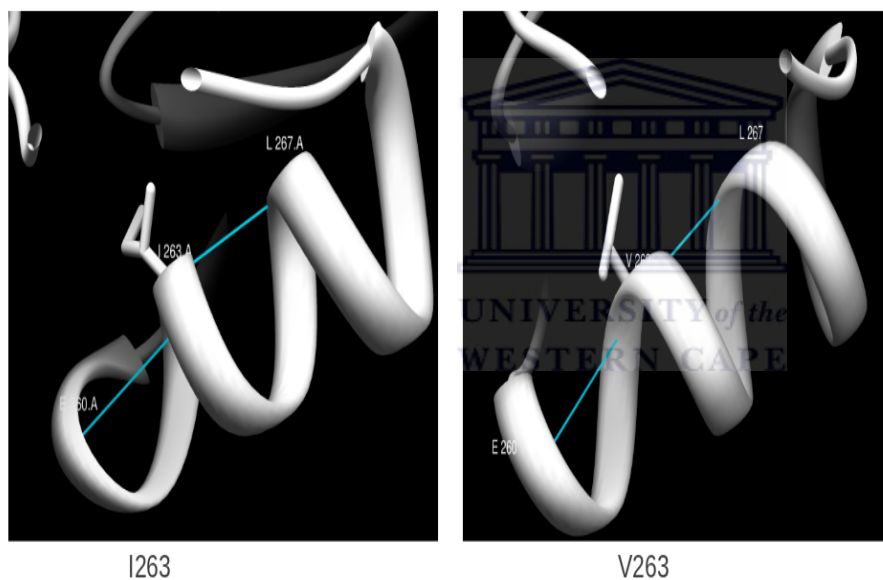
R187

Q187

Left panel- wild-type; Right panel- variant. Main-chain bonds maintained with substituted glutamine

# Appendix E

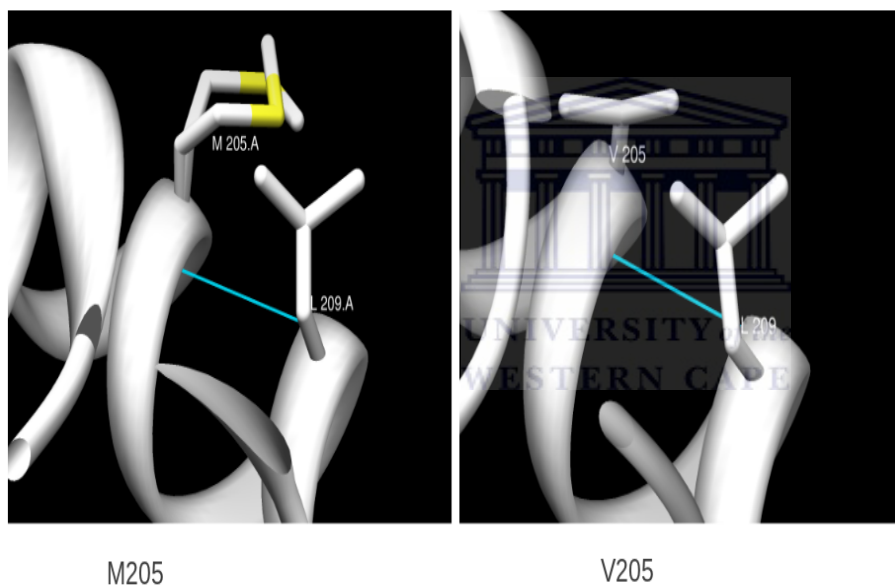
Residue interactions of wild-type residue I263 and SNP residue V263 .....



Left panel- wild-type; Right panel- variant. No change in hydrogen bonds

# Appendix F

Residue interactions of wild-type residue M205 and SNP residue V205. . . . .

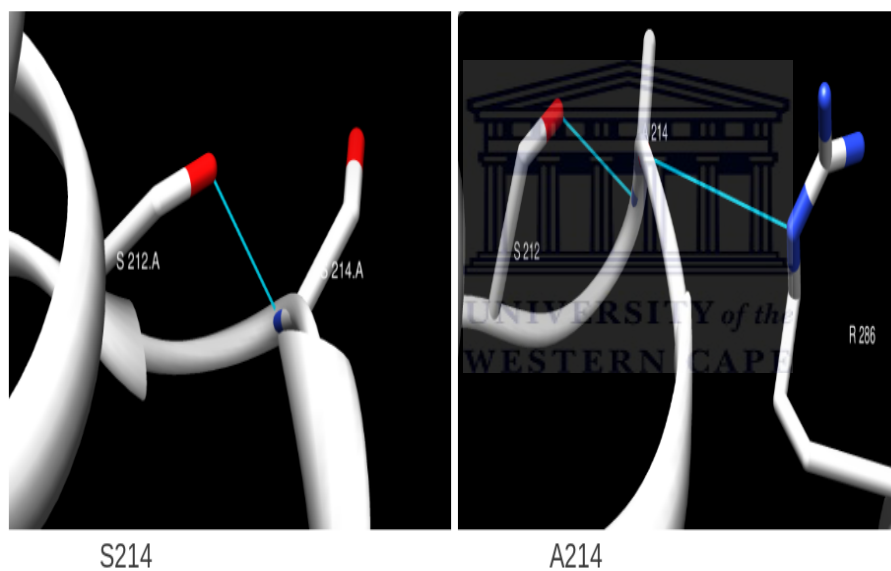


Left panel- wild-type; Right panel- variant. No change in hydrogen bonds



# Appendix G

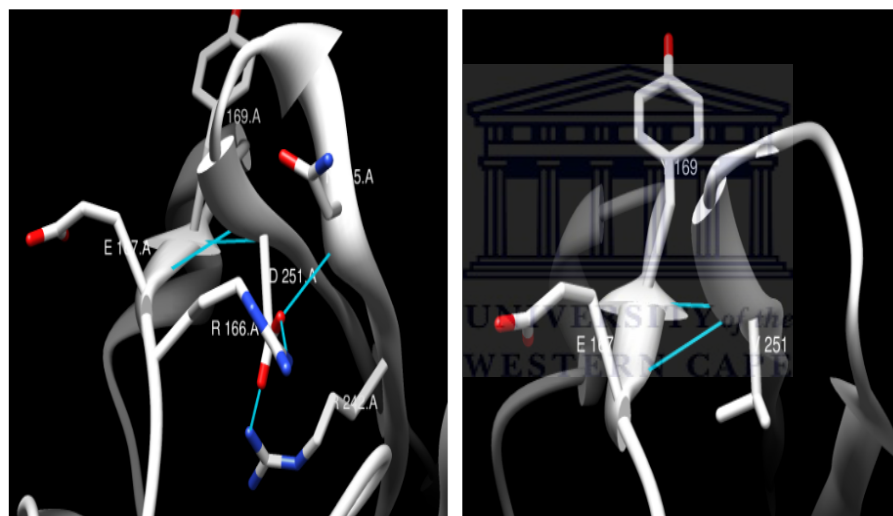
Residue interactions of wild-type residue S214 and SNP residue A214. ....



Left panel- wild-type; Right panel- variant. Additional bond formed between alanine 214 and arginine 256

# Appendix H

Residue interactions of wild-type residue D251 and SNP residue V251. ....



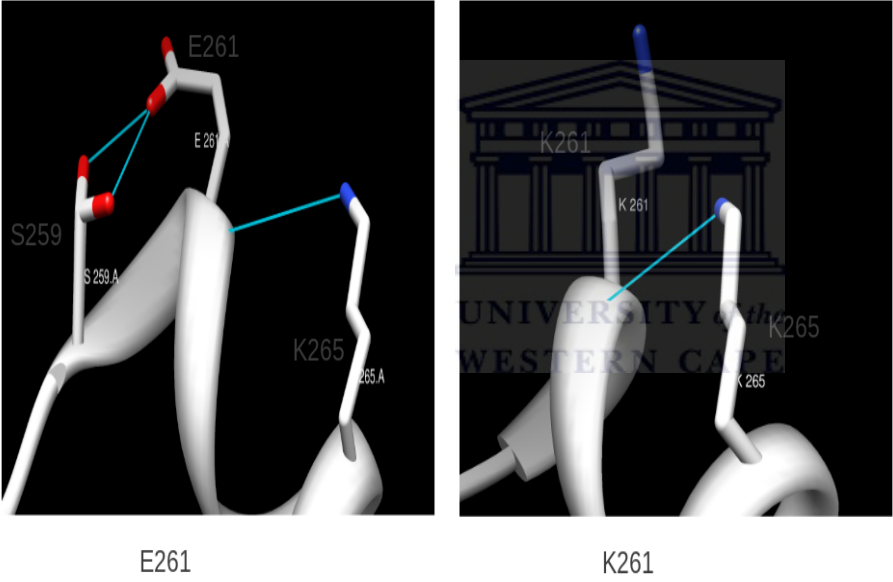
D251

V251

Left panel- wild-type; Right panel- variant. A significant number of hydrogen bonds are lost following substitution of valine.

# Appendix I

Residue interactions of wild-type residue E261 and SNP residue K261. ....



Left panel- wild-type; Right panel- variant. Loss of side-chain bonds with S259 replacing lysine for glutamic acid.