In silico analysis of the effects of non-synonymous single nucleotide polymorphisms on the Human Macrophage Migration Inhibitory Factor gene and their possible role in Human African Trypanosomiasis susceptibility

A mini-thesis submitted in partial fulfillment of the requirements of a degree

of

Master of Science

in

Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

RHODES UNIVERSITY, SOUTH AFRICA

Department of Biochemistry and Microbiology

Faculty of Science

by

MAGAMBO PHILLIP KIMUDA

15k8798

January 2016

ABSTRACT

Human African trypanosomiasis (HAT) is a public health problem in sub-Saharan Africa, with approximately 10,000 cases being reported per year. The Macrophage Migration Inhibitory Factor (MIF) which is encoded by a functionally polymorphic gene is important in both innate and adaptive immune responses, and has been implicated in affecting the outcome and processes of several inflammatory conditions. A recent study in mice to that effect showed that MIF deficient and anti-MIF antibody treated mice showed lowered inflammatory responses, liver damage and anaemia than the wild type mice when experimentally challenged with Trypanosomes. These findings could mean that the transcript levels and/or polymorphisms in this gene can possibly affect individual risk to trypanosomiasis. This is especially of interest because there have been reports of spontaneous recovery i.e self-cure/resistance in some HAT cases in West Africa. Prior to this discovery the general paradigm was that trypanosomiasis is fatal if left untreated.

The aim of this study was to gain insights into how human genetic variation in forms of non-synonymous SNPs affects the MIF structure and function and possibly HAT susceptibility. NsSNPs in the *mif* gene were obtained from dbSNP. Through homology modeling, SNP prediction tools, protein interface analysis, alanine scanning, changes in free energy of folding, protein interactions calculator (PIC), and molecular dynamics simulations, SNP effects on the protein structure and function were studied. The study cohort comprised of human genome sequence data from 50 North Western Uganda Lugbara endemic individuals of whom 20 were cases (previous HAT patients) and 30 were controls (HAT free individuals).

None of the 26 nsSNPs retrieved from dbSNP (July 2015) were present in the *mif* gene region in the study cohort. Out of the eight variants called in the *mif* coding region there was only one missense variant rs36065127 whose clinical significance is unknown. It was not possible to test for association of this variant with HAT due to its low global MAF that was less than 0.05.

Alanine scanning provided a fast and computationally cheap means of quickly assessing nsSNPs of importance. NsSNPs that were interface residues were more likely to be hotspots (important in protein stability). Assessment of possible compensatory mutations using PIC analysis showed that some nsSNP sites were interacting with others, but this requires further experimentation. Analysis of changes in free energy using FOLDX was not enough to predict which nsSNPs would adversely affect protein structure, function and kinetics. The MD simulations were unfortunately too short to glean any meaningful inferences. This was the first genetic study carried out on the people of Lugbara ethnicity from North Western Uganda.

DECLARATION

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from 15 July 2015 to 15 December 2015 under the supervision of Prof Özlem Taştan Bishop, Prof Nicola Mulder and Prof Enock Matovu.

I, **MAGAMBO PHILLIP KIMUDA**, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature; ..... Date; ….....

Supervisors;

Associate Professor Özlem TAŞTAN BISHOP (PhD)
Director of Research Unit in Bioinformatics (RUBi)
Department of Biochemistry and Microbiology
Rhodes University
P.O. Box 94, Grahamstown, 6140, South Africa

Professor Nicola Mulder (PhD)
Computational Biology Group
Institute of Infectious Disease and Molecular Medicine
UCT Faculty of Health Sciences
Observatory 7925
South Africa
University of Cape Town

Associate Professor Enock Matovu (PhD)
Senior Lecturer,
College of Veterinary Medicine, Animal Resources and Biosecurity,
School of Biosecurity, Biotechnical and Laboratory Sciences,
Department of Biotechnical and Diagnostic Sciences,
Makerere University,
P.O.BOX 7062,
Kampala, Uganda.

ACKNOWLEDGEMENTS

My deep felt thanks go out to the entire team at RUBi and the entire Biological Sciences Department of Rhodes University. This work was only possible thanks to the diligent guidance of my supervisors for whom I am grateful for. This research work was funded by the Trypanogen project (www.trypanogen.net/) and carried out at the Research Unit in Bioinformatics (RUBi) at Rhodes University. Some computations in this study were performed using facilities provided by the Computational Biology (CBIO) group, University of Cape Town's and the ICTS High Performance Computing team: http://hpc.uct.ac.za. Called VCF files used in this study were provided by the Trypanogen Bioinformatician Harry Noyes based at the University of Liverpool. Special thanks go out to the team at Trypanogen and CBIO.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACD	Anaemia of Chronic Diseases
BD	Becton, Dickinson and Company
bp	Base pair
BWA	Burrows Wheeler Aligner
CATT	Card Agglutination Test for Trypanosomiasis
CBIO	Computational Biology Group
CGAS	Candidate Gene Association Study
Chr	Chromosome
CNS	Central Nervous System
CSF	Cerebral Spinal Fluid
CSV	Comma Separated Values
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic Acid
EDTA	Ethyl Diamine Tetra-acetic Acid
GATK	Genome Analysis Tool Kit
GDS	Genome Data Structure
GPI	Glyophosphatidylinositol
GRCh37.p13	Genome Reference Consortium Human Build 37 patch release 13
GROMACS	Groningen Machine for Chemical Simulations
GWAS	Genome Wide Association Study
HAT	Human African Trypanosomiasis
HGMD	Human Gene Mutation Database
HGVbase	Human Sequence Variation Database
IFN-γ	Interferon Gamma

IL	Interleukin
K	Kelvin
Kda	Kilo daltons
MAF	Minor Allele Frequency
MD	Molecular Dynamics
MIF	Macrophage Migration Inhibitory Factor
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
nm	Nano Meter
nsSNP	Non-Synonymous Single Nucleotide Polymorphism
ns	Nano Second
OMIM	Online Mendelian Inheritance in Man
PCA	Principal Component Analysis
PDB	Protein Data Bank
pH	Potential of Hydrogen
PIC	Protein Interactions Calculator
ps	Pico Second
RBC	Red Blood Cells
Rg	Radius of Gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SLE	Systematic Lupus erythematosus
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
SOAP	Short Oligonucleotide Analysis Package
SSE	Systematic Sequencing Errors

VEP	Variant Effect Predictor
VCF	Variant Call Format
VSG	Variant Surface Glyco-protein
WHO	World Health Organisation
3D	Three Dimensional

LIST OF WEB SERVERS USED

MutPred	http://mutpred.mutdb.org/
nsSNPAnalyzer	http://snpanalyzer.uthsc.edu/
PolyPhen 2.0	http://genetics.bwh.harvard.edu/pph2 /
Protein Interface Calculator	http://pic.mbu.iisc.ernet.in/
Variant Effect Predictor	http://grch37.ensembl.org/Homo_sapiens/Tools/VEP

CHAPTER ONE

LITERATURE REVIEW

1.1 Chapter Overview

The purpose of this chapter is to introduce Human African Trypanosomiasis (HAT) and the role of the Macrophage Migration Inhibitory Factor (MIF) in the pathology of the disease. As will be discussed in this chapter, MIF has been implicated in playing a role of several diseases however the main focus will be on African Trypanosomiasis. This chapter will also attempt to make a case for why MIF is a valid target for the study.

1.2 Human African Trypanosomiasis (HAT)

African trypanosomes are devastating human and animal pathogens that cause trypanosomiasis which result in significant human and livestock mortality, morbidity and limits economic development in sub-Saharan Africa. The disease is caused by extracellular hemo-flagellated protozoans called Trypanosomes and transmitted by the tsetse fly (*Glossina* species) to its mammalian hosts. In humans the disease is commonly called sleeping sickness or Human African Trypanosomiasis (HAT), and in livestock the disease, it is called Nagana [1–3].

HAT is caused by two subspecies of the *Trypanosoma brucei* genus namely; *Trypanosoma brucei rhodesiense* that causes an acute form of the diseases and *Trypanosoma brucei gambiense* that causes a chronic form of the disease. *Trypanosoma brucei brucei,* a member of the same genus, is not human infective but is of veterinary importance because it causes Nagana in cattle.

Uganda is the only country with foci of both forms of HAT, where they pose extensive problems due to the risk of geographical overlapping as the acute form of the disease is spreading northwards [4]. This is likely to have an impact on control and treatment strategies.

Reports of new HAT cases per year have dropped below 10,000 but it continues to be a public health concern. There are several reports of recent HAT cases in endemic regions such as Angola, Chad, Southern Sudan, with Central African Republic and the Democratic Republic of Congo being the most severely affected. It serves to note that Central African Republic and the Democratic Republic of Congo are currently conflict regions. Due to this conflict there is limited access to health facilities especially in the rural areas which leads to the number of new HAT cases being severely under-reported [5, 6].

1.2.1 HAT Life Cycle

The genus *T. brucei* belongs to the family *Trypanosomatidae*, which covers a large group of unicellular protozoan parasitic organisms, under the order of *Kinetoplastida*. It is a spindle-shaped cell (20 to 30 by 1.5 to 3.5 μm) with a single flagellum that emerges from the posterior end as shown in Figure 1.1. There are two observed main stages in the trypanosome life cycle: trypomastigote and epimastigote. The blood stream trypomastigote form is observed in the mammalian blood and tissue fluids. While the epimastigote form is observed in the gut of tsetse fly vector (*Glossina* spp.) and its salivary glands [2, 7, 8].

When an infected tsetse fly (genus *Glossina*) has a blood meal, it injects metacyclic trypomastigote into the sub-dermal tissue of the mammalian host. The parasites then form a chancre and via the lymphatic system pass into the bloodstream. Once inside the bloodstream of the mammalian host the parasites transform into bloodstream slender and stumpy forms as shown in Figure 1.1. During this stage the parasites are carried to several sites all over the body, eventually entering other body fluids such as lymph and spinal fluid replicating by binary fission. The trypanosome is an extracellular hemo-flagellate so when a tsetse fly bites an infected mammalian host to feed, the parasites are taken in the blood meal and make their way to the insect's mid-gut. Here the parasites transform into procyclic forms and rapidly multiply by binary fission. The procyclic forms then migrate to the tsetse fly salivary glands where they transform into epimastigotes where they multiply by binary fission. The epimastigotes eventually transform into metacyclic trypomastigotes as shown in step eight of Figure 1.1 which are the infective stage of the insect parasites to the mammalian host. The entire insect stage of the parasite takes about 3 weeks [8]. While cattle and wild game animals play a more significant role as reservoirs of *T. b. rhodesiense*, humans are the main reservoir for *T. b. gambiense* [2, 9].

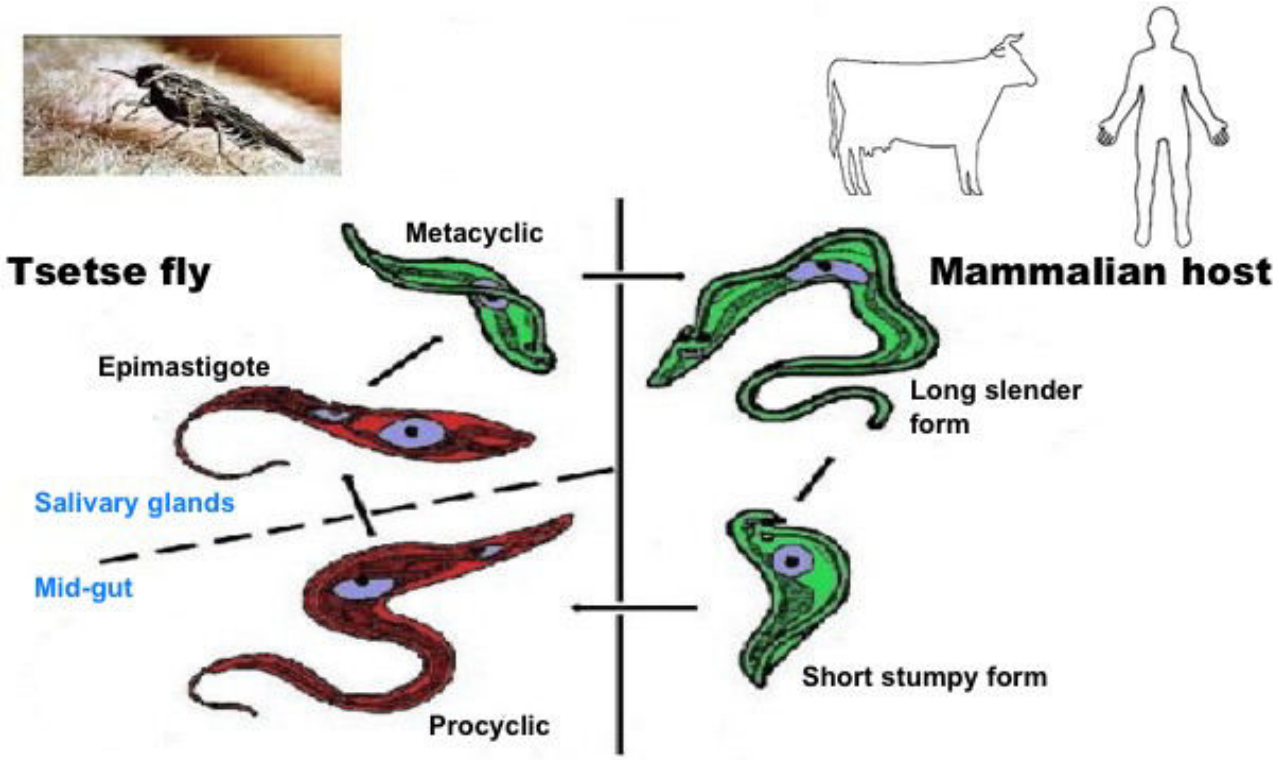

Figure 1.1: Diagrammatic representation of the life cycle of Trypanosoma brucei in the mammalian host and the tsetse fly.

1.2.2 Pathogenesis

HAT clinical pathogenesis is divided into two stages, namely the early hemo-lymphatic stage where the parasites have not yet crossed the blood-brain barrier, followed by the meningo-encephalitic stage where the parasites cross the blood brain barrier and invade the central nervous system (CNS) finally settling in the cerebrospinal fluid (CSF) [10].

The African trypanosome is an extracellular parasite and in the early hemo-lymphatic stages in the mammalian host can be found in blood, lymph and tissue fluids. During this stage the parasite is constantly exposed to the host's immune system. To survive the trypanosome is covered with a dense coat of approximately 10^7 variant surface glycoproteins (VSGs) attached to the plasma membrane by glycophosphatidylinositol (GPI) anchors. The VSG coat protects the parasite from immune attack by constantly switching as such hiding the parasite from the host's immune attacks, this process is called antigenic variation [2, 11]. The host is capable of mounting an effective immune response often eradicating some parasites. However there are still other parasites whose VSG coats it can not recognise. These parasites form the next wave of infection prolonging infection and transmission to other hosts by the tsetse fly vector [2, 12]. The symptoms of this stage

include fever, general malaise, joint pains and a chancre associated with the site of the tsetse fly bite. When the parasites cross the blood brain barrier several neurological symptoms are observed typical of the meningo-encephalitic stage. These include tremors, general motor weaknesses, irritability, confusion, poor coordination, and aggressive behaviour. Disruption of the body's natural circadian sleep/wake rhythm is a key defining feature of the later stages of the meningo-encephalitic stage hence the name 'sleeping sickness'. The general consensus was death follows if no treatment is given. But this is no longer true given that there are cases of spontaneous parasite clearing in some West African individuals [13–17].

Studies have shown that the major cause of pathogenicity in trypano-susceptible animals such as bovine to be anaemia, which is the leading cause of death due to the disease [18]. It is postulated that breeds such as the N'Dama and West African Shorthorns cattle are able to survive infection (mitigate morbidity) by controlling the development of anaemia. Murine studies suggest that a strong pro-inflammatory (type 1) immune response is necessary for the initial control of the growth of trypanosomes, this involves classically activated myeloid cells in particular macrophages (M1). A very strong inflammatory response especially if prolonged leads anaemia and increased pathogenicity. The over stimulation and subsequent activation of myeloid cells has been suggested to be the cause of extra-vascular destruction of red blood cells (RBCs) by the host's spleen and liver M1 cells [18–20] resulting in the characteristic trypanosomiasis associated anaemia which is similar to anaemia of chronic disease (ACD) that is common in chronic infections and sterile inflammations [21, 22]. Pathogenic features of the uncontrolled anaemia and M1 cell over stimulation include cachexia and liver injury.

1.2.3 Current Treatment Strategies

There are currently no effective vaccine remedies for HAT, as this is made difficult by antigenic variation. The current treatment strategy is chemotherapies however the drugs are few, toxic, limited in effectiveness, difficult to administer and prone to emerging resistance [3, 23–28]. The current drugs for HAT treatment are: melarsoprol, pentamidine, nifurtimox–eflornithine, suramin, and eflornithine (World Health Organisation’s list of essential medicines in 2009).

Anti-disease approaches such as the use of IL-10 which is an anti-inflammatory cytokine to mitigate the pathogenic features gives hope to therapies aimed at modulating the host's pro- and anti-inflammatory signals during the disease state which could help in the reduction of tissue injury [29]. Other anti-disease approaches include the use GPI-anchor of the VSG which has been shown to have M1-activating potential to treat animals where it showed reduction in trypanosomiasis associated liver damage, cachexia, anaemia and prolonged host survival. This was as a result of modulation of the myeloid cell activation state (M1 to M2 or vice-versa) [30].

1.3 Macrophage Migration Inhibitory Factor (MIF)

In humans the single *mif* gene lies on chromosome 22q11.2 [31–34]. It is composed of three exons of 205, 173, and 183 bp and two introns of 189 and 95 bp [31, 32, 35, 36] and is regulated by two polymorphic sites in the promoter region [33, 37].

The gene codes for a 12.5kDa polypeptide, consisting of 115 amino acids forming a homotrimer [38, 39]. Each monomer (Figure 1.2) is made up of two anti-parallel alpha-helices that are packed against a four stranded beta-sheet. In total three beta-sheets, and six alpha helices form the homotrimer, that appears in the form of a circular protein with an anterior traversing channel in its center. At the N-terminus of each monomer there's a proline residue [40] which is important in the keto-enole tautomerisation of pyruvoyl moiety [41, 42]. The active form of the protein as revealed by its crystal structure, shows a 37.5kDa homotrimer [38] as shown in Figure 1.2.

MIF is expressed in many cells including; macrophages, monocytes, 2 vascular smooth muscle cells, and cardiomyocytes [43–46]. MIF by means of a CD74 extracellular domain binds to cells in order to initiate ERK-1/2 activation [18, 47]. Once secreted it can activate T-cells and macrophages

Figure 1.2: Cartoon representation of homotrimeric MIF protein (PDB id: 3DJH). Figure is prepared by PyMOL. It is colored by chain.

to produce pro-inflammatory cytokines such as interleukin- (IL-) 1β, tumor necrosis factor alpha (TNF-α), IL-2, IL-6, IL-8, IL-12 and interferon gamma (IFN-γ) [18, 43, 48, 49]. As such it plays a major role in innate and adaptive immune responses [18, 43].

1.3.1 Human MIF as a drug target for HAT

MIF is a cytokine that is important in both innate and adaptive immunity [18, 19, 43, 50] and is a key player in the induction of systematic inflammation and has been implicated in many inflammatory diseases [18, 51, 52]. Its major role in inflammation reactions is the recruitment of myeloid cells to inflammation sites [18, 53]. This is achieved by inducing the differentiation toward M1 cells that secrete TNF [18, 54], anti-inflammatory actions of glucocorticoids [55, 56] and suppression of p53-dependent apoptosis of inflammatory cells [37].

While there are no studies directly linking MIF to HAT, there are murine studies that have shown that in MIF-knock out mice there is an overall reduction in the production of monocyte/macrophage derived pro-inflammatory cytokines such as IL-1b, IL-12, and TNF-a. This shows that MIF plays a role as a mediator of the inflammation cascade which is a key feature in trypanosomiasis-associated associated pathology [18, 57, 58]. MIF deficient mice featured limited anaemia, increased iron bio-availability, improved erythropoiesis and a marked reduction in RBC clearance during chronic stages of infection [18, 58, 59] when the mice were experimentally challenged with *Trypanosoma brucei brucei*. Serves to note the same is not true when MIF deficient mice are experimentally challenged with *Trypanosoma cruzi* where they showed enhanced susceptibility, higher morbidity (severe heart and skeletal muscle immuno-pathology) and mortality [60, 61].

Functional polymorphisms in the *mif* gene, for example in the promoter have been linked with autoimmune diseases such as scleroderma [62], tuberculosis [63], rheumatic arthritis [64], juvenile inflammatory arthritis [65], and systemic lupus erythematosus (SLE) [66]. There are several polymorphisms that have been reported in the human *mif* gene such as a single G/C nucleotide polymorphism at position -173 (rs755622), and a CATT tetra-nucleotide repeat (position -794, rs5844572) both of which have been shown to interfere with the transcriptional activity of the MIF promoter [64]. Other SNPs reported include position +254 (rs2096525), and position +656 (rs2070766) which are located in introns [67, 68].

This data suggests that MIF could possibly promote the most prominent pathological features of trypanosomiasis in an experimental setting (liver and spleen injury). This has implications in human subjects especially now that there have been reports of spontaneous recovery i.e self-cure/resistance in some cases in West Africa [13–17, 69].

1.4 Project Motivation

HAT continues to affect millions of people in Africa, and the drugs available are few, toxic, ineffective, and associated with resistance [24, 28]. Further vaccine options are not very promising [21]. There is an urgent need for new and innovative approaches to tackle this neglected disease. Understanding individual risk to trypanosomiasis is crucial to control and prevention strategies, especially in endemic regions. Host genetics studies are a key starting point and this study is one such attempt.

1.5 Knowledge Gap

To date there is no study that has been carried out to show the association of MIF polymorphisms with HAT susceptibility in a human African endemic population from North Western Uganda. This study will also contribute in furthering the understanding the effects of non-synonymous single nucleotide polymorphisms on the structure and function of the MIF protein which is currently lacking.

1.6 Problem Statement

Trypanosomiasis is a neglected tropical disease of public health concern that affects up-to 10,000 people per year [70]. The drugs available for treatment are few and unfortunately rather toxic and reports of rising drug resistance are also of concern [24, 28]. Reports of spontaneous recovery or human trypano-tolerance gives us drive to investigate host related genetic factors that are key to understanding this observed phenomenon [13–17]. Recent murine studies have shown that MIF plays an important role in Trypanosome pathogenicity [58]. There is a need to map how nsSNPs in the human *mif* gene affect the structure and function of the MIF protein. This information on host genetics can help us further understand HAT epidemiology, pathology and possibly help in identifying new drug and anti-disease targets.

1.7 Broad Objective

To identify host genetic factors in the *mif* gene, that could play a role in African Trypanosomiasis susceptibility

1.8 Specific Objectives

- Determine which nsSNPs are relevant in the structure and function of the MIF protein
- Determine which nsSNPs in the *mif* gene might be associated with HAT susceptibility

1.9 Research Hypothesis

Polymorphisms in the *mif* gene in form of nsSNPs affect the MIF protein structure, function, and as a result HAT susceptibility.

CHAPTER TWO

CANDIDATE GENE ASSOCIATION STUDY

2.1 Chapter Overview

In this chapter we will discuss how the samples were obtained for the study cohort, sequenced, and the variant caller pipeline used to generate the variant call file (VCF) that contains the SNPs called from the samples. We also discuss how VCF files were generated for use in the checking for population stratification, annotation, prioritisation of nsSNPs, and for carrying out the candidate association study.

2.2 Introduction

2.2.1 Variant Calling

Getting an accurate and true picture of variations from NGS data analyses can be difficult given that true variation has to be carefully separated from various machine Artefact, such as false variants due to sequencing errors. When searching for mutations Next Generation Sequencing (NGS) is a powerful tool. However there are many technical challenges involved in getting an accurate representation of sequence variation and eventually turning the raw genome sequence data into information with biological meaning. Several steps, techniques, and capabilities are required to ensure a complete accurate analysis of NGS data in order to gain information on variation and to handle the large amounts of data [71–73]. This usually involves aligning the raw reads to a reference human genome, followed by identifying variants such as short insertions and deletions (indels) or single nucleotide variants (SNVs) that may be of interest for the phenotype under study [74].

2.2.1.1 Quality Control of Raw reads

In this study, the initial steps involved pre-processing of raw reads which involved, adaptor trimming, quality trimming using, the removal of very short reads, and de-duplication using Trimmomatic [75].

2.2.1.2 Alignment

This step, which is considered to be the most important and computationally demanding involves the mapping of the reads to a reference genome [76, 77]. In this step a number of errors that are likely to be passed down to subsequent steps in Variant calling make it very important. Case in

point, because it involves the aligning of each read independently to the reference genome there is a tendency for reads spanning indels to be misaligned as there is no reference. Errors such as Misaligned reads and unreliable base quality scores lead to artefacts that can lead to unreliable variant calls and, errors in genotyping [78–80]. As sequencing technologies evolve, many alignment programs have been developed to map each read to its corresponding location in the reference genome, some of these include; Novoalign (www.novocraft.com), Bowtie [76], BWA [81], SOAP [82], and MAQ [83, 84].

Bowtie

Bowtie [76] operates based on an index built with the Burrows-Wheeler Transformation [85, 86]. Its popularity as an aligner is mainly because it is fast, and has a small memory usage footprint (for example for an entire human genome it uses approximately 1.3 Gigabytes) [87]. However its speed comes at a cost in terms of accuracy. It has been shown to fail to align reads with valid mappings more so when configured for maximum speeds. It also does not guarantee the highest quality read mapping where no exact matches exist [82].

Bwa

Burrows-Wheeler Aligner (BWA) [77, 88] also uses the Burrows Wheeler Transformation [85, 86]. Where it differs is that it provides an added advantage in the form of a meaningful quality score that can be used to discard any mappings that are not very well supported [87]. Bwa and Bowtie utilise an FM-index method that uses a back-tracking strategy in their search for matches that are inexact.

Novoalign

Novoalign by Novocraft (http://www.novocraft.com/) builds an index with a hash table and utilizes an alignment scoring system based on the Needleman-Wunsch algorithm. It has emerged as a popular aligner because of its accuracy and it allows up to eight mismatches per read for single end mappings [87]. Comparison of these aligners on real and simulated data reveals that the alignment programs perform similarly well for reads that have relatively good quality or were pre-processed to trim off any low quality bases [82]. However Novoalign is shown to be more sensitive to any improvement of data quality.

2.2.1.3 Artefact Removal

Artefact removal involves the use of the GATK IndelRealigner [89] which entails local alignment around indels, base quality score recalibration (GATKBQSR), variant calling (GATK haplotype caller) [90, 91] and finally statistical filtering (GATK variant quality score recalibration).

Local realignment

Alignment algorithms align each read independently to a reference genome which often results in errors in alignment around reads spanning indels. These alignment artefacts in the form of wrongly

mapped SNPs, insertions and deletions confounded further by sequencing errors eventually result in false positive variant detection (more so in apparent heterozygous positions) [92]. The local realignment process is designed to locally realign reads in-order to minimise the number of mismatching bases across all reads thus reducing the amount of false positive variants. The GATK IndelRealigner [93] serves to transform regions of misalignment due to indels into clean reads composed of consensus indels thus minimising the number of false positive variants. It utilises the full alignment context to determine if the appropriate alternate references (indels) exist.

2.2.1.4 Base Quality Score Recalibration

Understanding systematic sequencing errors (SSE) and sequence platform biases which are problematic at high sequence depths is important in dealing with whole genome data [94, 95]. The causes of SSE are many, not well understood, and batch run specific, and compensating for them is necessary [96]. GATK BQSR [89, 93, 96] carries out a recalibration of quality scores for bases in reads in-order to make them more accurate (closer to actual probability of mismatching the reference genome). The tool also tries to correct for variations in quality in a read group, machine cycle, base quality score, dinucleotide and sequence context, providing more accurate, and more widely dispersed quality scores. The system works on BAM files coming from several sequencing platforms (SOLiD, Pacific Biosciences, Illumina, 454, Complete Genomics etc). Improvements in GATK 2.0 also allow for recalibration of unknown base insertion and base deletion quality scores (http://gatkforums.broadinstitute.org/discussion/44/base-quality- score-recalibration-bqsr).

2.2.1.5 Calling of Variants

Processing data through next generation sequence pipelines to the point of high quality variant calls still remains a challenge [97] and the performance of many pipelines boils down to the kind of calling strategy and variant callers used [73]. Variant calling is broken down into two basic steps: genotype assignment and variant identification.

GATK UnifiedGenotyper

GATK UnifiedGenotyper [93] is a Bayesian caller that uses a Bayesian genotype likelihood model to simultaneously estimate the most likely allele frequency and genotypes in a population of N samples, producing a genotype for each sample. It separately calls indels and SNPs by considering each variant locus independently (GATK documentation, https://www.broadinstitute.org/gatk/gatkdocs/). GATK recommends the use of its HaplotypeCaller for calling variants but where it is not possible to do so, for example in instances when dealing with a large number of samples, pooled samples, or working with non-diploid organisms the UnifiedGenotyper is recommended.

GATK HaplotypeCaller

GATK HaplotypeCaller [93], is a more developed Bayesian haplotype caller derived from GATK, which is able to provide local assembly in regions spanning variants. It calls SNPs and indels simultaneously via local *de-novo* assembly of haplotypes in an active region i.e re- maps and reassembles reads in that region (GATK documentation, https://www.broadinstitute.org/gatk/gatkdocs/) making it more accurate at calling traditionally difficult regions (for example different types of variants in close proximity) and indels.

2.2.2 Principal Component Analysis (PCA)

Population stratification results from systemic ancestry differences. It influences allele frequencies between cases and controls in association studies as such it is a major confounding factor often leading to spurious associations [98]. PCA refers to a statistical method that utilises an orthogonal transformation to convert the observations to linearly uncorrelated variables also known as principal components. This method is useful in the detection of population stratification. SNPRelate a Bioconductor package was used for carrying out the PCA in this study. its main advantage being that it utilises the Genomic Data Structure (GDS) data format. This format is efficient because it reduces the data to integers with two bits which accelerates computing speed [99].

2.2.3 Variant Annotation

Variant annotation refers to the process by which functional information is assigned to DNA variants. This includes information such as frequency, measures of conservation, variant type (missense or indels for example), predictions of the possible effects of the variant and function [100–103].

2.2.4 Candidate Gene Association Study

There are two main research approaches for population based genetic association studies, both of which are based on genotyping of Single Nucleotide Polymorphisms (SNPs) namely; Candidate Gene Association Studies (CGAS) and Genome-Wide Association Studies (GWAS) [104].

CGAS involves an *a priori* hypothesis that specific genes are associated with disease susceptibility or risk, and is a deductive approach. In this regard it differs from GWAS where association analyses are conducted without prior hypotheses and cover the entire genome [105]. The biggest impediment in CGAS is the selection of suitable candidate genes as this requires knowledge of the biological pathway of the genes that might be suitable potential candidate genes [106]. However there is a growing number of bioinformatics resources available to assist in pathway selection and prioritization of putative disease related genes [107–109]. Coupled with the identification of novel candidate genes for diseases this can be an iterative process [104].

CGAS represents a cost-effective approach for well defined disease questions. However like GWAS, it has limited power to detect all associations between susceptibility genes and the disease. It is also prone to poor reproducibility of results. That is why it is important that CGAS studies be repeated with different cohorts (different data sets) to determine if the findings are reproducible. Every year next generation sequencing technologies are improving, evolving and dropping in cost allowing for more cost effective approaches to high throughput sequencing of human genomes, which is essential for furthering life sciences research [110]. This is most true in the study of entire human genomes to understand the role of variation in human diseases and human genetic diversity [111–114]. NGS also presents an attractive technology for CGAS.

2.3 Chapter Objectives

The objectives of this chapter are as follows:

- To detect any population stratification within the sample cohort
- To annotate the variants in the MIF coding region and prioritise nsSNPs
- To carry out an association study of the nsSNPs with HAT

2.4 Methodology

During this study I was involved in the initial sample collection, processing, DNA extraction, and quantification. The samples were sequenced at the University of Liverpool. Variant calling was done by the Trypanogen project bioinformatician Harry Noyes and the files shipped to the University of Cape Town on hard disk while others were downloaded directly onto their cluster. The VCF files were used to carry out a Principal Component Analysis (PCA) to detect population stratification. This was followed by variant annotation and a candidate gene association study.

2.4.1 Sample Collection

Ugandan subjects from the West Nile region of the Arua district of Lugbara ethnicity were recruited at government run health centres with the help of clinicians and community health volunteers. 20 cases were identified by means of treatment cards, and hospital/clinic records of diagnosis and treatment for HAT. 30 Controls were recruited as matched pairs of the cases in terms of age, sex, ethnicity and how close they stayed to their paired cases. Exclusion criteria applied to any individuals below the age of 18 or any individuals where it was not possible to obtain consent or blood samples. Venous blood was then drawn from them by vene-puncture and collected in EDTA/heparin vacutainer tubes (BD). Buffy coats were then prepared in field laboratories and stored in liquid nitrogen in preparation for DNA extraction that was carried out at the Molecular Biology Laboratory, College of Veterinary Medicine, Animal Resources and Biosecurity, Makerere University, Kampala, Uganda. The DNA was quantified using a QubitTM (Life Technologies),

using a broad range DNA dye (Life Technologies). The samples were then shipped to the University of Liverpool, United Kingdom for whole genome sequencing using Illumina HiSeq. All participants in this study were required to provide informed consent using a consent form administered in their local language (Lugbara) that was approved by local authorities as well as by the local ethical committee (Uganda National Council for Science and Technology).

2.4.2 Variant Calling Pipeline

Samples were sequenced on a HiSeq 2500 machine, and processed using the analysis pipeline displayed in Figure 2.1.

Figure 2.1: Flow-chart showing a summary of the steps taken in the variant calling pipeline

2.4.3 Data Retrieval

The raw sequence reads were passed through a similar variant calling pipeline shown in Figure 2.1 by Harry Noyes (Trypanogen Project Bioinformatician) at the University of Liverpool. The files were transferred in two batches to the Computational Biology (CBIO) group, University of Cape Town, Hex cluster (http://hex.uct.ac.za).

For this study two VCF files were needed, one for carrying out the preliminary PCA and another for the candidate gene association study. The chromosome 1 (Chr1) and MIF coding regions in both

batches were isolated using Vcftools. The Genome Reference Consortium Human Build 37 patch release 13 (GRCh37.p13) was used for getting the locations of the MIF coding region and Chr1. The MIF and Chr1 vcf-subsets were isolated using Vcftools-subset [115], the pseudo-code is shown below.

For variants in the MIF coding region:

```
/vcftools_0.1.13/bin/vcftools --gzvcf ../trypanogen/TrypanogenBatch1.vcf.gz
--chr 22 --from-bp 24236565 --to-bp 24237409 --recode --recode-INFO-all

/vcftools_0.1.13/bin/vcftools    --gzvcf    ../trypanogen/TrypanogenBatch2.vcf.gz
--chr 22 --from-bp 24236565 --to-bp 24237409 --recode --recode-INFO-all
```

For variants in the Chr1 region:

```
/vcftools_0.1.13/bin/vcftools –gzvcf ../trypanogen/TrypanogenBatch1.vcf.gz --chr
1 --from-bp 1000000 --to-bp 2000000 --recode –recode-INFO-all

vcftools_0.1.13/bin/vcftools –gzvcf ../trypanogen/TrypanogenBatch2.vcf.gz --chr
1 --from-bp 1000000 --to-bp 2000000 --recode --recode-INFO-all
```

The two seperate batches were then combined using GATK CombineVariants [93] using the pseudo-code below.

```
java       -jar       /GenomeAnalysisTK.jar       -T       CombineVariants       -R
../bundle/b37/human_g1k_v37_decoy.fasta   --variant   ../OUT/vcfsubset_batch1.vcf
--variant ../OUT/vcfsubset_batch2.vcf -o ../OUT/merged.vcf -genotypeMergeOptions
UNIQUIFY
```

The merged files still contained samples from other cohorts, to isolate samples only specific to the Uganda cohort under study, the sample identities were acquired from the list using bash scripting. The sample list was used in Vcftools to create the final vcf files that only contained the study cohort.

To create the sample identity list:

```
cat Trypanogen_merged.vcf | head -n 1000 | grep "CHROM" | tr "\t" "\n" | grep
".UGA." > sample_ID_Uganda_cohort_only.txt
```

To create the final vcf files with only the samples in the Uganda cohort sample list:

```
cat Trypanogen_merged.vcf | ../vcftools_0.1.13/bin/vcf-subset -c
sample_ID_Uganda_cohort_only.txt > merged_UGA_only.vcf
```

In total there were 20 cases and 30 controls. Two vcf files were created one with variants in the MIF coding region that contained eight variants and another with 12950 variants in Chr1.

2.4.3 Principal Component Analysis (PCA)

The PCA was run using SNPRelate in R as shown in Appendix 1. The first step involved setting the working directory and loading the R packages: gdsfmt and SNPRelate. This was followed by reading in the VCF file. The VCF file was then converted to GDS format, and the output written to the same directory. The GDS file was then read in and used to run the PCA, before the results were plotted the sample identities and population codes were read into the data-frame which was then used to plot the PCA.

2.4.4 Annotation

The MIF vcf-subset VCF was uploaded to the Variant Effect Predictor specifically on the grch37 website (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP) and the results downloaded as CSV files that were imported into excel for further analysis [116].

2.4.5 Candidate Association Study

To run the candidate gene association study, the first step was converting the MIF vcf-subset VCF into the PED format and in the process filtering for rare variants using Plink [117]. The pseudo code is shown below:

```
../bin/plink --vcf MIF_vcfsubset_UGA_only.vcf --maf 0.05 --recode --make-bed
--out MIF_vcfsubset_UGA_only
```

The '.ped' and '.map' files generated were then used to carry out a Fisher's exact text using the Plink '--assoc' command to compare the frequency of the variants in the cases (former HAT patients) and controls (paired controls, who have never suffered from HAT but live in the same endemic areas). The pseudo code is shown below:

```
../bin/plink --ped MIF_vcfsubset_UGA_only.ped --map TMIF_vcfsubset_UGA_only.map
--maf 0.05 --assoc
```

2.5 Results and Discussion

2.5.1 PCA

VCFS from 50 samples, 20 cases and 30 controls, were used in a principal component analysis. Analysis of the first two principal components of 12950 SNPs located on chromosome 1 revealed

that there was no obvious population stratification detected as shown in Figure 2.2.

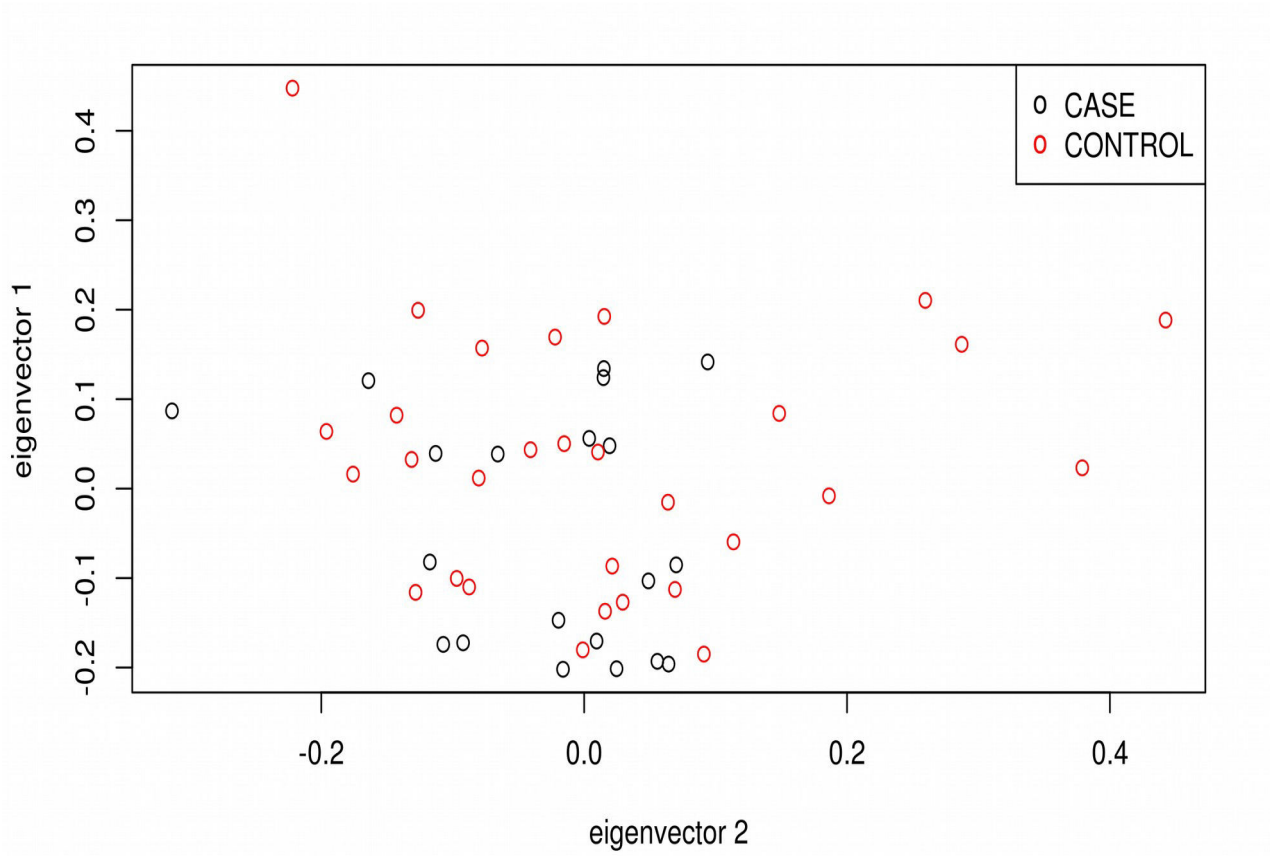

Figure 2.2: The first two principal components from analysing the Chr1 vcf sub-set consisting of 20 cases (coloured black) and 30 controls (coloured red).

From the plot in Figure 2.2 there aren't any significant clusters indicative of differences in ethnicity between the cases and controls. The sample cohort in this study was obtained from individuals of the Lugbara ethnicity of Northern Uganda. Using a questionnaire tool care was taken to ascertain ancestry. For example, language spoken by contemporary ancestors such as fathers, mothers and grand parents of the individuals were used as indicators of ethnicity. Pairing of cases with controls that had no blood relations also ensured that relatedness was avoided. Population stratification is of importance because it leads to spurious association if not properly corrected for [118].

2.5.2 Variant Annotation

The SNPs associated with the *MIF* gene were annotated using the Variant Effect Predictor. Figure 2.3 shows that there were a total of eight variants identified in the MIF coding region, with only a single one being a missense variant (rs36065127, located on Chr 22: 24237348). Most of the variants were intronic or non-coding, however, since the study was interested in functional and structural impact, we focussed on the missense variant.

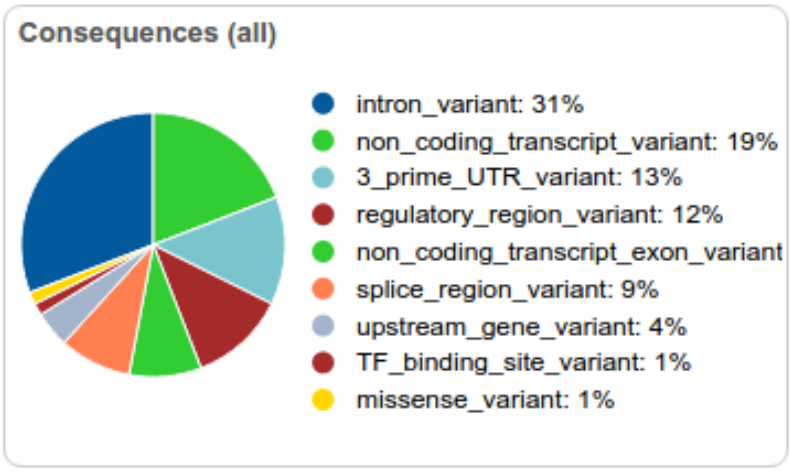

Figure 2.3: Summary of the consequences of the variants

This rs36065127 variant overlaps 10 transcripts, one regulatory feature and has 2599 sample genotypes. The ancestral allele for rs3606512 is 'G' with 'A' and 'T' as alternative alleles. The Variant Effect Predictor tool was used to draw pie charts representing the rs3606512 allelic frequencies in the different 1000 genomes populations. Figure 2.4 shows the allelic frequencies for rs3606512 in the main 1000 genomes populations. The overall allelic frequency for the total 1000 genomes was 2% for allele T, 0% for allele A, and 98% for allele G. The European, East Asian and South Asian populations had no instances of allele T. Only the African and African American had instances of allele T (6% and 1 % respectively).

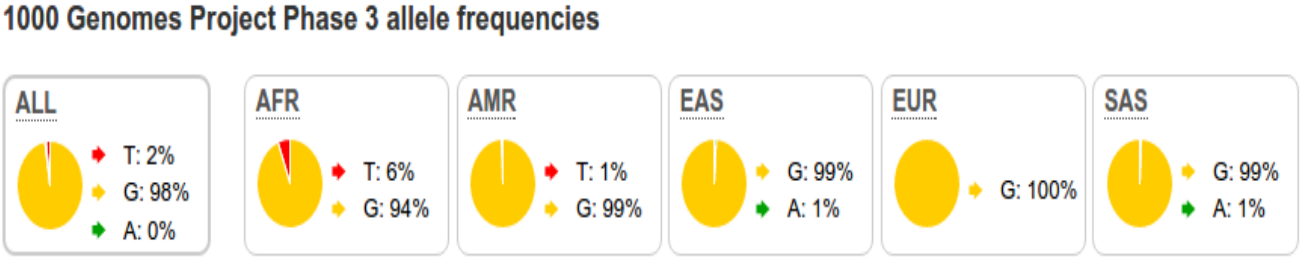

Figure 2.4: Allelic frequencies for the main 1000 Genomes Populations. AFR-African, AMR-American, EAS- East Asian, EUR-European and SAS-South Asian.

Figure 2.5 shows the allelic frequencies for the variant in African sub-populations. Where the highest frequency for the T allele was in the Esan in Nigeria and Mandinka in The Gambia (11% and 8% respectively). Both Nigeria and Gambia have had reported cases of HAT. The variant's low T allelic frequency in other main continental populations is likely due to the fact that the sampled population for this study is of African origin and likely subject to different selection pressures.

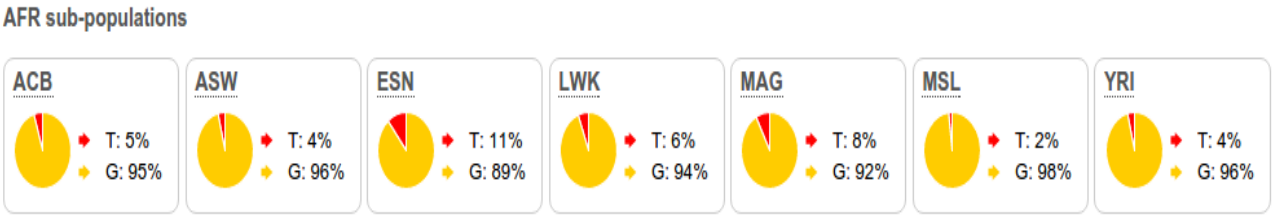

Figure 2.5: Allelic frequencies for the African Sub-Populations. ACB-African Caribbeans in Barbados, ASW-Americans of African Ancestry in SW USA, ESN- Esan in Nigeria, LWK-Luhya in Webuye, Kenya, MAG-Mandinka in The Gambia, MSL- Mende Sierra Leone, and YRI- Yoruba in Ibadan, Nigeria.

2.5.3 Candidate Gene Association Studies

The rs36065127 variant however had a low global minor allele frequency (MAF) of 0.0166 which as a result ended up in it being filtered out during the initial Plink [117] filtering steps. Rare variants with MAF of less than 0.05 are removed before any association study is carried out. This is considered good practice which is in line with the 'common disease common variant' hypothesis [119]. While recent empirical evidence shows that rare variants may play a role in complex diseases [120–122] without a bigger sample size or family-based samples any associations generated at this point would likely be spurious. The rs36065127 variant had a frequency of 0.05 in the cases and 0.033 in the controls, with an odds ratio of 1.526 but these results are likely spurious given the sample size and the low minor allele frequency of the variant.

2.6 Conclusion

The sample cohort composed of 20 cases and 30 controls showed no population stratification. A total of eight variants were called in the MIF coding region. Only one missense variant was called, namely rs36065127 (clinical significance unknown). Testing for association of this variant with HAT was not possible due its low global MAF of 0.0166. However this can be remedied by studying a larger sample size. It is interesting to note that only one of the nsSNPs present in the MIF coding region from dbSNP [123] at the time of sampling appeared in this small sample. This gives more reason for additional African populations not yet sequenced to be sampled as this will broaden our view of variation.

This is the first genetic study carried out on people of Lugbara ethnicity from North Western Uganda. The uniqueness of this population is evidenced by the fact that only one of the nsSNPs isolated from dbSNP at the time of querying the database was present in this population.

CHAPTER THREE

PROTEIN STRUCTURE ANALYSIS

3.1 Chapter Overview

The purpose of this chapter was to investigate the effects of nsSNPs that were retrieved from dbSNP on the MIF protein structure and function. This was done through a combination of several methods and approaches. The final goal was to understand the effects of each of the nsSNPs on the structure and function of the MIF protein. This involved the use of *in silico* SNP effect prediction tools such as MutPred, SNPAnalyser and Polyphen. This was followed by Alanine scanning to identify which nsSNP sites were potential hotspots. A PyMOL script and Protein Interaction Calculator (PIC) analysis were used to predict interface residues. PIC analysis was used to also identify any nsSNPs that may be subject to compensatory mutations. Homology modelling was used to generate MIF protein mutants that contained single, double or triple chain mutations to study how each of them affected the overall protein structure and stability. This was further assessed using calculation of changes in energy by FOLDX and YASARA, and eventually a 1ns molecular dynamics simulation.

3.2 Introduction

3.2.1 *In silico* non-model based prediction of nsSNP effects

Single nucleotide polymorphisms (SNPs) are the most frequent type of genetic variation in humans, and play a crucial role in human diseases and other phenotypic traits. They account for over 85% of mutations associated with specific disease [124]. They are capable of affecting protein function for example in extreme cases by introducing stop codons, which result in truncated proteins that are unable to function. In other cases they result in single amino acid substitutions within protein sequences that may affect the structure of protein, stability, folding and their ability to bind ligands or catalyse reactions.

As the data collected from sequencing technologies increases so have the amount of novel SNPs being discovered [125, 126], and so have the computational approaches to analyse the effects of these SNPs on protein structure and function [127, 128]. About 300,000 novel single nucleotide variants (SNVs) are generated by every newly sequenced genome [129]. Predicting the effects of the nsSNPs on the translated protein structure and function is still a largely unsolved problem. This is an important issue because nsSNPs can influence chemo-therapeutics especially in the era of personalised medicine [130, 131]. SNP effect prediction programs typically classify nsSNPs as

either deleterious or of no consequence/neutral. Some tools use conservation based measures [102]. Others use a combination of both conservation based methods as well as structural features in conjunction with machine-learning approaches such as support-vector machines or neural networks [100, 132–134]. The following freely available web-servers based tools were used in this study to analyse the nsSNPs and predict whether they would be benign or deleterious: Polyphen 2.0 [100], nsSNPAnalyser [135] and Mutpred [136]. In regard to how the tools function, nsSNPAnalyser [135] extracts evolutionary and structural information from a query nsSNP and relies on a Random Forest machine learning approach to predict nsSNP effects. Mutpred [137] and Polyphen 2.0 [100] are probabilistic machine learning classifiers. Unlike Polyphen 2.0 or nsSNPAnalyser that give categories such as probably damaging, benign, neutral or disease, Mutpred returns its results in form of probability known as the MutPred score (that is a figure between 0 and 1). A score > 0.5 is considered as harmful while a score > 0.75 is considered harmful but with a high confidence in the prediction.

3.2.2 Hotspot Prediction Using Alanine Scanning by FOLDX

This refers to the study of relative free energy change (ΔΔG) that occurs when individual residues are mutated to alanine. To predict potential hot spot residues, Alanine scanning was done on each nsSNP position in the MIF structure using a FOLDX [138] plug-in in the YASARA [139, 140] graphical interface program. FOLDX is an empirical force field formulated by analysing a thousand point mutations from eighty two protein-protein complexes. In the calculation of free energy it accounts for several thermodynamic terms known to be of importance to protein stability. These include: solvation effects, Van der Waal's interactions, water bridges, hydrogen bonds, electrostatic and entropy effects for the backbone and the side-chain. This allows for the prediction of hotspot residues, that is to say residues most likely to affect the general stability of the structure [141].

3.2.3 Protein Interface Calculator (PIC) Analysis

Compensatory mutations refer to mutations that occur to correct a loss in fitness. For example if a nsSNP has a deleterious effect, it can be corrected by another mutation to lessen its effect [142]. In this study Protein Interactions Calculator (PIC) analysis [143] was used to identify nsSNP sites that had weak or strong interactions with other nsSNP sites. These interactions included: hydrogen bonds, disulphide bonds, ionic interactions, aromatic-aromatic interactions, cation-n interactions and aromatic sulphur interactions for example between proteins in the complex.

3.2.4 Homology Modelling of MIF Mutants

Homology modelling, also known as comparative modelling or template based modelling refers to

the use of a known protein 3D structure (preferably of high resolution e.g 1 Å or so) to predict an unknown protein structure basing on its protein sequence *in silico*, in a manner that is accurate enough and comparable to results that were achieved experimentally [144–146]. It generally consists of four main steps [147]: (a) Template identification; (b)Multiple sequence alignment to identify true homologs and alignment of the target sequence with the template sequence; (c) model building and refinement; and (d) validation/assessment of the structures [147]. It is especially useful in instances where protein structures by X-ray crystallography or Nuclear Magnetic resonance are unavailable. Protein mutation studies are often expensive, time consuming and laborious. This technique offers a means of carrying out high through-put mutation studies cheaply and quickly, aiding in studies on protein structure, function, and rational drug discovery/design [148–151]. MODELLER [152] is used for homology modelling and involves the use of an alignment composed of a sequence with a known structure and sequence of protein with no structure. MODELLER creates a structure by automatically calculating a model containing all non-hydrogen atoms, with satisfaction of spatial restraints [153]. MODELLER also has the capability to perform tasks such as *de novo* modelling of loops, loop refinement, multiple alignments, clustering, structural comparisons and querying sequence databases [147, 154].

3.2.5 Changes in Free energy

One of the most important characteristics that can be related to a protein's function and structure is its folding free energy. It is therefore one of the means the effects of an amino acid point mutation can be assessed [155–157]. In this study FOLDX (an empirical forcefield) was used to assess changes in free energy of folding [158]. FOLDX was calibrated using experimental mutational free energy changes from a collection of more than 1000 point mutations, covering XX proteins [159] and yielded a correlation of 0.81 with a standard deviation of 0.46 kcal/mol between calculated and experimental ΔΔGs in its current release. To get the most out of FOLDX it is advised to compare relative energies, i.e compare known structures as its absolute energies are not precise (http://FOLDX.crg.es/examples.jsp). It is also worth noting that FOLDX assumes a fixed back bone as a result it does not accurately predict ΔΔG values that may result in other conformational changes likely to affect the protein function [160]. In spite of this FOLDX is still suitable for evaluating changes in stability due to point mutations in protein structures [159, 161, 162].

3.2.6 Protein Interface Network Analysis

Understanding interactions within protein structures and complexes is essential to elucidating their assembly, stability and function. Network analysis of loss and gain of protein interface residues, was done using a PyMOL [163] script available at

http://www.PyMOLwiki.org/index.php/InterfaceResidues shared under the 'GNU Free Documentation License 1.2'. This is important because even though free energy changes maybe negligible a nsSNP can adversely affect protein function by changing it's protein interface interaction network which can change a protein's function and kinetics [164].

3.2.7 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulations have also been used in the analysis of the effects of nsSNPs on protein function and structure, this allows for further understanding on how a single point mutation can affect the overall protein structure. This is done by subjecting the mutated protein to long MD simulations with in order to study the time evolution and time averaged values of structural properties in functionally important regions [165, 166].

3.3 Chapter Objectives

The objectives of this chapter are as follows:

- To determine effects of nsSNPs using web based nsSNP analysis tools
- To determine which nsSNP sites are hotspots
- To determine if any nsSNPs are under compensatory mutation
- To investigate the effects of the nsSNPs on the protein interface of MIF mutants
- To investigate changes in energy due to single, double, and triple chain mutations

3.4 Methodology

The overall methodology followed in this Chapter is summarised as a work flow shown in Figure 3.1.

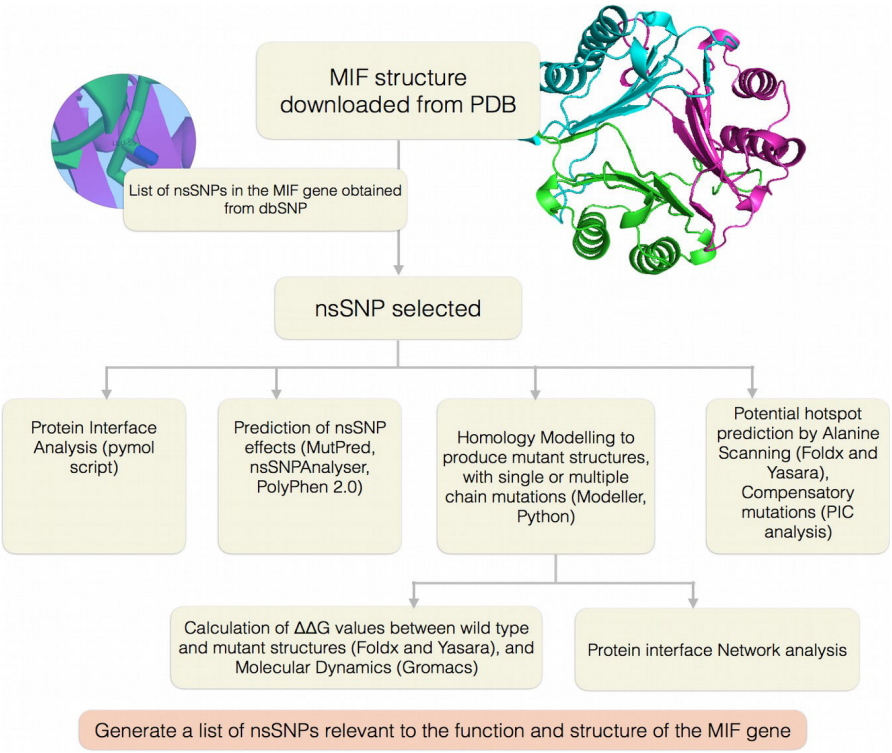

Figure 3.1: Protein structure analysis work-flow

3.4.1 Data Retrieval

For this study, 26 nsSNPs were retrieved from dbSNP [123] as of July 2015. The identified SNPs, then, were validated by 1000 Genome project [167]. Further PyMOL script (see Appendix 1) is used to determine the position of each SNP in the secondary structure prediction.

3.4.2 nsSNP Analysis Via Web-servers

The MIF 3DJH [39] crystal structure was uploaded to the following freely available web-servers: MutPred (http://mutpred.mutdb.org/), Polyphen 2.0 (http://genetics.bwh.harvard.edu/pph2/), and nsSNPAnalyser (http://snpanalyzer.uthsc.edu/). This was followed by the amino acid substitutions that were submitted individually or in lists where a substitution was denoted as A#Y (where A was the original amino acid, # the position and B the mutated amino acid) to the web-servers. The results were then imported into excel spreadsheets for further analysis.

3.4.3 Hotspot Prediction Using Alanine Scanning and Protein Interface Prediction

Computational alanine scanning mutagenesis was carried out using YASARA [139] utilizing a FOLDX plug-in [138, 140] in order to identify potential hot spots of protein-protein interaction. It involved uploading the MIF wild type PDB structure (3DJH), followed by alanine scanning calculation of the relative free energy change (ΔΔG) that occurs when individual residues are mutated to alanine. Default settings were: Temperature 298K, pH 7.0, ionic strength 0.05M and Van der Waals design 2. Any residues giving rise to ≥1.5 kcal/mol were identified as potential hot spots [168].

A PyMOL script available at http://www.PyMOLwiki.org/index.php/InterfaceResidues shared under the 'GNU Free Documentation License 1.2' was used to identify interface residues between the three MIF monomers. The script found interface residues between the monomer chains by taking the area of the complex then splitting it into two pieces (one for each chain) and calculating the chain-only surface area finally taking the difference between the complex based areas and the chain only based areas. When the value is greater than the supplied cut off, it is called an interface residue. The MIF wild-type structure was loaded into PyMOL followed by the following pseudo code in the PyMOL terminal:

```
interfaceResidue complexName[, cA=firstChainName[, cB=secondChainName[,
cutoff=dAsaCutoff[, selName=selectionNameToReturn ]]]]
```

where,

complexName- The name of the complex. cA and cB must be within in this complex

cA- The name of the first chain to investigate

cB- The name of the 2nd chain to investigate

cutoff- The dASA cut-off, in square Angstroms

selName- Name of the selection to return.

This was followed by iteration of the selection name to get the residues present in the specified protein interface. From this list a Python script was used to pull out the nsSNP sites in-order to determine which were interface residues and which weren't depending on the specific chain interface.

3.4.4 Protein Interface Calculator (PIC) Analysis

The native MIF structure was submitted to the Protein Interactions Calculator (PIC) web-based server for intra-protein interaction function (http://pic.mbu.iisc.ernet.in/) [143]. Default settings were used. This was done to identify the interactions around nsSNPs with other nsSNPs or residues.

3.4.5 Homology Modelling of MIF Mutants

3.4.5.1 Template Identification

The MIF coding sequence was retrieved from NCBI RefSeq (NC_000022.11) [169] and used to query Protein Data Bank [170].

3.4.5.2 Target-Template, Alignment, Model Building and Model Evaluation

A Python script (Appendix 2) was used to create protein sequences containing one or more of the amino acid substitutions as a result of the nsSNPs. Another Python script (Appendix 3) was used to to align these mutated sequences to the 3DJH protein sequence using MUSCLE [171], and create the PIR format alignment files. Lastly another Python script (Appendix 4) was used to carry out homology modelling using MODELLER [152] on the in-house cluster. The script generated 100 models per run and from those selected the one with best DOPE-Z score [153]. Evaluation and validation of the structures were done using manual inspection in PyMOL (overlapping the structures) and by using PROCHECK [172].

3.4.6 Changes in Free Energy

The mutant PDB structures created by homology modelling were loaded into YASARA followed by the repair object command in FOLDX [139, 140, 156]. This command results in minimisation of the protein structure by rearranging amino acid side chains in order lower the free energy of the protein. It is an important step before free energy calculations using FOLDX. This was followed by the stability of object command that calculates the difference in free energy between the folded and unfolded state of the protein structure (a lower energy structure is generally considered to be more stable).

3.4.7 Protein Interface Network Analysis

A PyMOL script available at http://www.pymolwiki.org/index.php/InterfaceResidues shared under the 'GNU Free Documentation License 1.2' (described earlier in section 3.4.4) was used to identify interface residues between the three MIF monomers in the triple mutant MIF structures. A Python script was then used to compare the interface residues in the mutants to those in the native MIF structure in order to identify new or lost interface residues.

3.4.8 Molecular Dynamics Simulations

GROMACS (Groningen Machine for Chemical Simulations) [173] was used to run Molecular Dynamics 1000ps simulations for six nsSNPs predicted to be deleterious by all three predictors and three nsSNPs predicted to be inconsequential or not likely to cause an effect by all three predictors. The molecular dynamics simulations were run using in-house scripts and the '.mdp' parameter files were acquired from the GROMACS manual (http://manual.gromacs.org/online/mdp.html). The MD work flow is summarised in the Figure 3.2 below and a sample of the bash script used to submit the MD jobs on the RUBi cluster available in Appendix 5.

Using consensus from the predicting programs Table 3.3, three predicted non-consequential nsSNPs (rs200394994,rs201060788 and rs376184469) and six predicted deleterious nsSNPs (rs11548056, rs199714772, rs200005486, rs200500959, rs202066662,and rs372052137) were selected for this assay. Analyses were carried out using GROMACS 4.5.7 software package using the amber03 force field, changes in RMSD and Rg between native and mutant proteins were the main subject of interest [173].

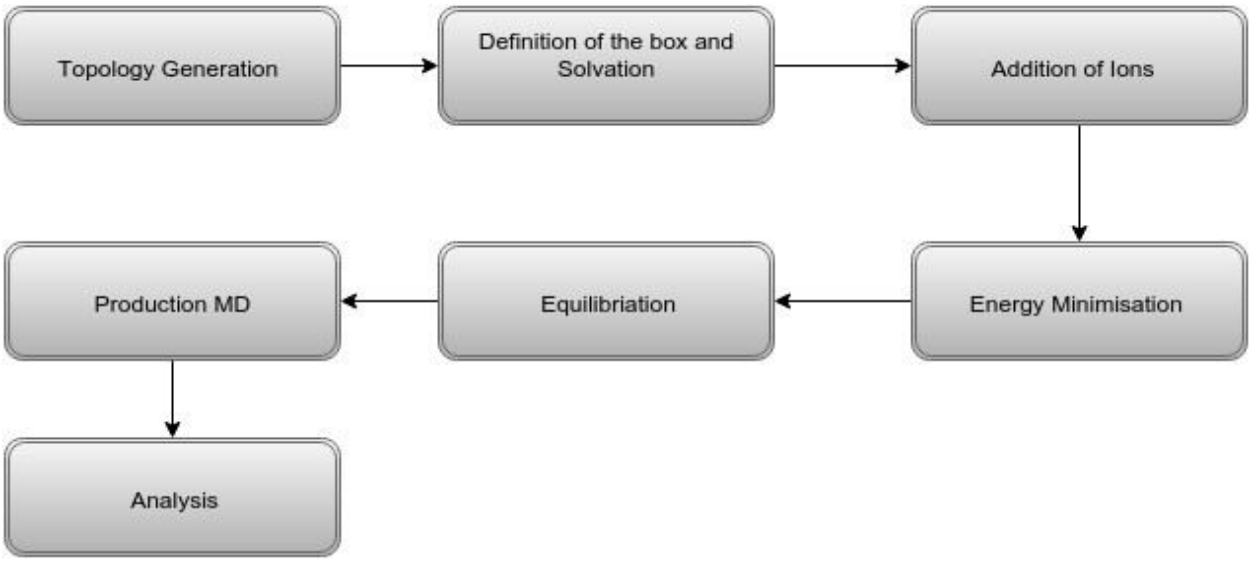

Figure 3.2: Summary of MD work-flow

3.5 Results and Discussion

3.5.1 nsSNP List

Table 3.1 shows a list of 26 nsSNPs that were retrieved from dbSNP [123] as of July 2015 and shown to be located within the *mif* gene. Of these 10 (38%) were not validated, while 16 (62%) were validated. To date there are several databases to get useful information on nsSNPs, these include : Online Mendelian Inheritance in Man (OMIM) database, the UniProt/Swiss-Prot database, the Human Genome Variation database (HGVbase), the Human Gene Mutation Database (HGMD), and single nucleotide polymorphism database (dbSNP). In this study we mainly focused on dbSNP because it's comprehensive, receives the largest number of submissions (open submission policy), has NCBI genomic information, gives the validation status for each SNP, and is the primary source for many of the other curated SNP databases. Of the nsSNPs validated by 1000 Genomes [167] rs139210892, rs1803976 and rs182012324 had a global minor allele frequency (MAF) of 0.0004. The remaining three namely rs201631604, rs201862457, rs372575900 had a global MAF of 0.0002. The purpose of the global MAF is to help in distinguishing between common and rare variants. The nsSNPs in the database as the one found in the CGAS cohort are all rare variants as their global MAFs are less than 0.05.

Table 3.1: List of nsSNPs in the mif gene showing their secondary structure prediction using PyMOL, validation status and location

SNP_ID	Amino acid mutation	Structural Prediction (Pymol)	Status	Location in GRCh37.p13 Primary Assembly
rs1049829	L59F	Sheet	Validated (by cluster)	24237025
rs11548056	I68T	Loop	Validated (by cluster)	24237053
rs11548059	P44Q	Loop	Validated (by cluster)	24236981
rs139210892	T113I	Loop	Validated (by 1000G,by cluster,by frequency)	24237283
rs1803976	N106S	Loop	Validated (by 1000G,by frequency)	24237262
rs182012324	S75F	Helix	Validated (by 1000G,by cluster,by frequency)	24237074
rs199714772	V15M	Loop	Not Validated (No info)	24236704
rs201742529	P92S	Helix	Not Validated (No info)	24237124
rs199774339	P92R	Helix	Not Validated (No info)	24237125
rs199980863	P35L	Helix	Not Validated (No info)	24236765
rs202066662	H41Y	Sheet	Not Validated (No info)	24236971
rs200005486	H41P	Sheet	Validated (by cluster)	24236972
rs200286358	Y99C	Sheet	Validated (by cluster)	24237241
rs200329745	A115V	Loop	Not Validated (No info)	24237289
rs200394994	A71T	Helix	Validated (by cluster)	24237061
rs200500959	P2R	Loop	Validated (by cluster)	24236666
rs200995600	T24S	Helix	Validated (by cluster)	24236731
rs201060788	E86K	Helix	Not Validated (No info)	24237106
rs201307782	S64G	Sheet	Not Validated (No info)	24237040
rs201631604	P34T	Loop	Validated (by 1000G,by cluster)	24236761
rs372052137	I97T	Sheet	Not Validated (No info)	24237235
rs201792625	I97V	Sheet	Not Validated (No info)	24237234
rs201862457	P16Q	Loop	Validated (by 1000G,by cluster)	24236708
rs201465617	I5M	Sheet	Validated (by cluster)	24236676
rs372575900	M48L	Sheet	Validated (by 1000G,by cluster)	24236992
rs376184469	R87H	Helix	Validated (by cluster)	24237110

3.5.2 MIF Crystal Structures

A total of five structures were retrieved that were free of any structural mutations as shown in Table 3.2. The crystal structure utilized in this study, 3DJH [39] was selected because of its high sequence identity, supporting publication, and high resolution.

Table 3.2: MIF crystal structures obtained from the Protein Data Bank (PDB)

PDB ID	Description	Taxonomy	Aligned Protein	Aligned Residues	Sequence Identity
3DJH	Macrophage Migration Inhibitory Factor (MIF) at 1.25 Å Resolution	*Homo Sapiens*	3	341	100%
3IJJ	Ternary Complex of Macrophage Migration Inhibitory Factor (MIF) Bound Both to 4-hydroxyphenylpyruvate and to the Allosteric	*Homo Sapiens*	3	341	100%
1GDO	HUMAN MACROPHAGE MIGRATION INHIBITORY FACTOR (MIF)	*Homo Sapiens*	3	341	100%
4F2K	Macrophage Migration Inhibitory Factor covalently complexed with	*Homo Sapiens*	3	341	100%
4K9G	1.55 Å Crystal Structure of Macrophage Migration Inhibitory Factor bound to ISO-66 and a related compound	*Homo Sapiens*	3	341	100%

3.5.3 *In silico* Non-Model Based Prediction of NsSNP Effects

As shown in Table 3.3 for the given dataset of 26 nsSNPs there was ~62% concordance in prediction across all the three. In spite of the tremendous progress in developing fast and accurate approaches that predict the effects of nsSNPs on protein function and structure there are still no methods that are as good as wet-lab mutation analyses [174]. Instances where the predictions of the tools did not agree could be likely to differences in their classification and analytical algorithms. This could also be confounded further by training data set bias.

Table 3.3: List showing the predicted effects of nsSNPs on MIF using Polyphen2.0, Mutpred, and nsSNPAnalyser. Key: Yellow- where all three predict the same effects.

SNP_ID	Polyphen Predicted Phenotype	NsSNPAnalyser Predicted Phenotype	MutPred probability of deleterious mutation
rs1049829	probably damaging	Neutral	0.631
rs11548056	probably damaging	Disease	0.895
rs11548059	probably damaging	Neutral	0.612
rs139210892	probably damaging	Disease	0.792
rs1803976	probably damaging	Neutral	0.527
rs182012324	benign	Disease	0.452
rs199714772	possibly damaging	Disease	0.739
rs199774339	probably damaging	Neutral	0.524
rs199980863	benign	Neutral	0.474
rs200005486	probably damaging	Disease	0.609
rs200286358	benign	Disease	0.875
rs200329745	possibly damaging	Neutral	0.655
rs200394994	benign	Neutral	0.462
rs200500959	probably damaging	Disease	0.840
rs200995600	possibly damaging	Neutral	0.670
rs201060788	benign	Neutral	0.527
rs201307782	probably damaging	Disease	0.961
rs201631604	possibly damaging	Disease	0.858
rs201742529	possibly damaging	Neutral	0.390
rs201792625	benign	Neutral	0.736
rs201862457	probably damaging	Neutral	0.710
rs202066662	probably damaging	Disease	0.537
rs201465617	benign	Neutral	0.418
rs372575900	benign	Neutral	0.725
rs376184469	benign	Neutral	0.570
rs372052137	probably damaging	Disease	0.736

3.5.4 Hotspot Prediction Using Alanine Scanning and Protein Interface Prediction

The result of mutating each residue (with the exception of Alanine or Glycine) was given in terms of an energy difference (ΔΔG, in kcal mol−1) between the mutant and unmodified protein (wild type), decomposed into the FOLDX energy terms, a cut off of ≥1.5 kcal/mol was used to determine the potential hot spots and < 1.5 kcal/mol for non-potential hot spots [168]. FOLDX mutates a single residue at a time, as a result Figure 3.3 shows the effects of that mutations in different positions in the three chains of the MIF homotrimer. As can be seen from the graph there was variation in terms of the overall effect on the stability of the protein depending on which chain was mutated. This was an indication that perhaps some of the nsSNP positions may not be symmetrical or at protein-protein interfaces.

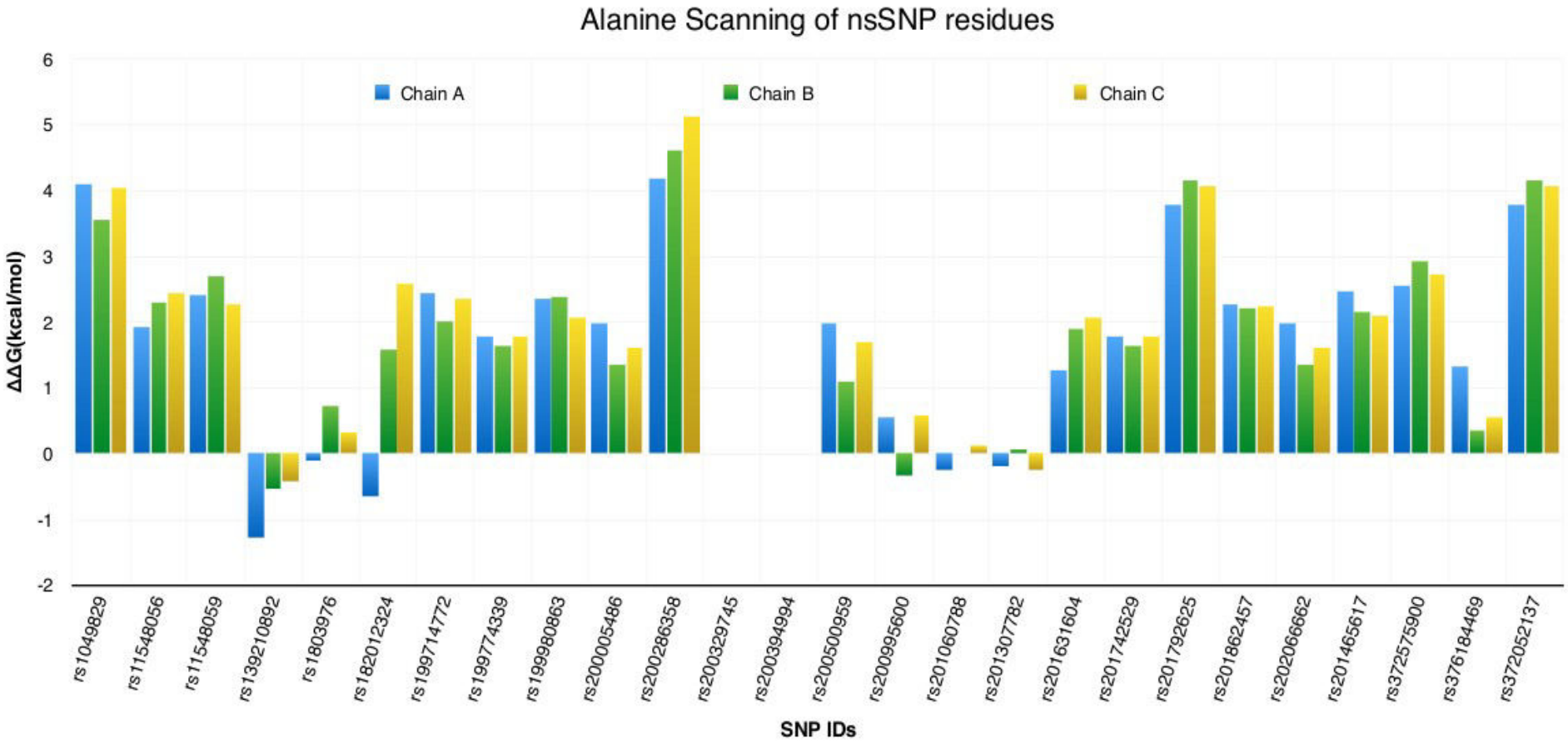

Figure 3.3: Bar-graph showing the results of Alanine Scanning of nsSNP positions on the MIF protein Structure

To further investigate, a PyMOL script was used to ascertain which of the nsSNP positions were at a protein-protein interface in the native MIF protein. The results of these predictions are shown in Table 3.4a and in combination with alanine scanning in Table 3.4b. As can be seen in Table 3.4 amino acid positions P44 and V15 which correspond to nsSNPs rs11548059 and rs199714772 respectively were only interface residues in only two out of the total three chains.

Table 3.4a: Summary of interface residue prediction using a PyMOL script. For each amino acid and position '[]' indicates in which chain the SNP site is.

Amino Acid Position	**Protein-Protein Interface**			**nsSNPs Associated with site**
	Chain A and B	**Chain B and C**	**Chain A and C**	
L59	L59[A]	L59[B]	L59[C]	rs1049829
I68	I68[A]	I68[B]	I68[C]	rs11548056
P44	P44[B]	P44[C]	-	rs11548059
T113	I113[B]	I113[C]	I113[A]	rs139210892
N106	N106[B]	N106[C]	N106[A]	rs1803976
V15	-	V15[C]	V15[A]	rs199714772
P92	P92[A]	P92[B]	P92[C]	rs199774339, rs201742529
P35	P35[B]	P35[C]	P35[A]	rs199980863
H41	H41[B]	H41[C]	H41[A]	rs200005486, rs20206662
Y99	Y99[A]	Y99[B]	Y99[C]	rs200286358
A115	A115[B]	A115[C]	A115[A]	rs200329745
T24	T24[B]	T24[C]	T24[A]	rs200995600
I5	I5[B]	I5[C]	I5[A]	rs201465617
M48	M48[A]	M48[B]	M48[C]	rs372575900
I97	I97[A]	I97[B]	I97[C]	rs372052137, rs201792625

Table 3.4b: Summary of results of alanine scanning and interface residue prediction.

SNP_ID	Chain A ΔΔG(kcal/mol)	Chain B ΔΔG(kcal/mol)	Chain C ΔΔG(kcal/mol)	Potential Hot Spot	Interface Residue (Pymol)
rs1049829	4.11534	3.56665	4.05239	+	+
rs11548056	1.9177	2.2887	2.43543	+	+
rs11548059	2.40483	2.70859	2.27611	+	+
rs139210892	-1.262	-0.535734	-0.428216	-	+
rs1803976	-0.113836	0.730512	0.329332	-	+
rs182012324	-0.635657	1.58745	2.57971	+	-
rs199714772	2.45397	2.01942	2.34551	+	+
rs199774339	1.79648	1.63505	1.78849	+	+
rs199980863	2.36648	2.37635	2.05998	+	+
rs200005486	1.9778	1.35613	1.62601	+	+
rs200286358	4.19305	4.61283	5.13114	+	+
rs200329745	2.04636E-12	5.45697E-12	4.09273E-12	-	+
rs200394994	6.82121E-13	3.86535E-12	3.29692E-12	-	-
rs200500959	1.98369	1.1054	1.69732	+	-
rs200995600	0.556819	-0.336089	0.596481	-	+
rs201060788	-0.242262	0.00394707	0.134711	-	-
rs201307782	-0.188059	0.0543872	-0.241089	-	-
rs201631604	1.28236	1.91272	2.07308	+	-
rs201742529	1.79648	1.63505	1.78849	+	+
rs201792625	3.79561	4.16264	4.06215	+	+
rs201862457	2.27718	2.20669	2.24542	+	-
rs202066662	1.9778	1.35613	1.62601	+	+
rs201465617	2.48033	2.16714	2.09106	+	+
rs372575900	2.56565	2.94069	2.72259	+	+
rs376184469	1.31326	0.341132	0.563377	-	-
rs372052137	3.79561	4.16264	4.06215	+	+

To determine whether the likelihood of a nsSNP being a hot spot from Alanine Scanning was dependent on whether it was an interface residue the Fisher's exact test (since not all the expected frequencies were greater than 5) was used. In R, a contingency table was used to structure the data and carry out the Fishers exact test. Since the test p-value (0.1972) was greater than 0.05 we rejected the null hypothesis and concluded that these data provided sufficient evidence at the 5% level of significance that there is a relationship between Alanine Scanning prediction results and Interface prediction results. This was especially of importance because nsSNPs at protein interface regions have been shown to interfere with ligand binding [175].

The Alanine scanning functionality in YASARA mutates one Alanine residue at a time. A script was used to extract the ΔΔG (kcal mol−1) from each nsSNP position per chain as shown in Table 3.4. In each instance the ΔΔG (kcal mol−1) was acquired for chain specific nsSNP sites. This was done sequentially starting with chain A, then B and C. As can be seen in Table 3.4a there are specified nsSNP sites that were not interface residues in those particular specified chains and this explains why nsSNP sites such as rs139210892, rs1803976, and rs200995600 did not cause significant changes in ΔΔG (kcal mol−1). A better approach for this assay would be to first identify the particular interface residues, and then individually mutate those residues to alanine and then acquire ΔΔG (kcal mol−1). NsSNP sites such as rs200329745 did not show any change in energy because it was already an alanine residue. NsSNP site rs182012324 (S75) is not an interface residue, however it was a hotspot because from its structural prediction its located in the helix and serine residues are known to stabilize bending angles in helices [176].

3.5.5 PIC Analysis

The MIF wild type structure, 3DJH was uploaded to the PIC web-server (http://pic.mbu.iisc.ernet.in/). The analysis was divided into intra-protein interactions i.e interactions between residues within the same chain and protein-protein interactions i.e between residues in different chains.

3.5.5.1 Intra-Protein Interactions

There were 13 intra-protein Hydrophobic interactions within 5 Angstroms between nsSNP sites, shown in Table 3.5a.

Table 3.5a: MIF Intra-protein Hydrophobic interactions within 5 Angstroms

SNP ID	Position	Residue	Chain	SNP ID	Position	Residue	Chain
rs199714772	15	VAL	A	rs201862457	16	PRO	A
rs201631604	34	PRO	A	rs199980863	35	PRO	A
rs11548056	68	ILE	A	rs200286358	99	TYR	A
rs372052137,rs201792625	97	ILE	A	rs200286358	99	TYR	A
rs199714772	130	VAL	B	rs11548059	159	PRO	B
rs372575900	163	MET	B	rs1049829	174	LEU	B
rs11548056	183	ILE	B	rs200286358	214	TYR	B
rs372052137,rs201792625	212	ILE	B	rs200286358	214	TYR	B
rs199714772	245	VAL	C	rs11548059	274	PRO	C
rs201631604	264	PRO	C	rs199980863	265	PRO	C
rs372575900	278	MET	C	rs1049829	289	LEU	C
rs11548056	298	ILE	C	rs200286358	329	TYR	C
rs201742529	327	ILE	C	rs200286358	329	TYR	C

Hydrophobic contacts especially in interface residues are important indicators of thermal stability, and along with hydrogen bonds play a significant role in protein stability [177]. Majority of the MIF homotrimer hydrophobic interactions were intra-protein, with the exception of rs201631604 the rest were sites located within protein interface regions.

There were 5 main chain-main chain hydrogen bonds between nsSNP sites identified and they are shown in Table 3.5b.

Table 3.5b: Intra-protein Main chain-Main chain hydrogen bonds. Key: Dd-a = Distance between donor and acceptor, Dh-a = Distance between Hydrogen and acceptor, A(d-H-N) = Angle Between donor-H-N and A(a-O=C) = Angle between acceptor-O=C

SNP ID	Acceptor Position	Chain	Residue	Atom	SNP ID	Donor Position	Chain	Residue	Atom	Dd-a	Dh-a	A(d-H-N)	A(a-O=C)
rs182012324	75	A	S	N	rs200394994	71	A	A	O	3.13	2.24	148.76	152.35
rs200329745	115	A	A	N	rs139210892	113	A	T	O	3.49	3.39	87.83	68.09
rs182012324	190	B	S	N	rs200394994	186	B	A	O	3.10	2.23	148.16	152.98
rs200329745	230	B	A	N	rs139210892	228	B	T	O	3.49	3.50	81.02	69.02
rs182012324	305	C	S	N	rs200394994	301	C	A	O	3.14	2.25	149.05	152.84

3.5.5.2 Protein-Protein Interactions

All nsSNP sites shown to be interacting in this analysis were interface residues and therefore likely important in overall protein stability. There were 2 protein-protein Hydrophobic interactions within 5 Angstroms between nsSNP sites, shown in Table 3.6a.

Table 3.6a: MIF protein-protein Hydrophobic interactions within 5 Angstroms

SNP ID	Position	Residue	Chain	SNP ID	Position	Residue	Chain
rs20146517	5	ILE	A	rs1049829	289	LEU	C
rs1049829	174	LEU	B	rs20146517	235	ILE	C

There were 2 main chain-main chain hydrogen bonds between nsSNP sites identified and they are shown in Table 3.6b.

Table 3.6b: Protein-protein Main chain-Main chain hydrogen bonds. Key: Dd-a = Distance between donor and acceptor, Dh-a = Distance between Hydrogen and acceptor, A(d-H-N) = Angle Between donor-H-N and A(a-O=C) = Angle between acceptor-O=C

SNP ID	Acceptor Position	Chain	Residue	Atom	SNP ID	Donor Position	Chain	Residue	Atom	Dd-a	Dh-a	A(d-H-N)	A(a-O=C)
rs200286358	214	B	TYR	N	rs1803976	336	C	ASN	O	2.82	1.98	142.63	157.29
rs200286358	329	C	TYR	N	rs1803976	106	A	ASN	O	2.89	1.99	151.38	146.72

There was also an instance of protein-protein Side chain-Side chain hydrogen bonds identified between acceptor position 156 nsSNP site for rs200005486) and donor position 48 (nsSNP sire for rs372575900) as shown in Table 3.6c.

Table 3.6c: Protein-protein Side chain-Side chain hydrogen bonds. Key: Dd-a = Distance between donor and acceptor, Dh-a = Distance between Hydrogen and acceptor, A(d-H-N) = Angle Between donor-H-N and A(a-O=C) = Angle between acceptor-O=C

SNP ID	Acceptor Position	Chain	Residue	Atom	SNP ID	Donor Position	Chain	Residue	Atom	Dd-a	Dh-a	A(d-H-N)	A(a-O=C)
rs200005486	156	B	HIS	ND1	rs372575900	48	A	MET	SD	3.42	3.33	87.84	999.99

PIC analysis revealed nsSNP sites that are likely subject to compensatory mutations but this requires experimental confirmation. This can be done by identifying possible compensatory nsSNPs occurring within the same subjects, as this clustering is expected due to their importance in the overall stability of the protein [178]. Homology modelling can also be used to assess the overall effect of the two compensating mutations within the same structure. This is of importance because studies have shown that coding compensatory mutations do not occur randomly over the gene sequence [142]. Compensatory mutations are essential for improving fitness in cases where mutations prove to be deleterious especially in critical regions such as hydrophobic cores, ligand binding sites and enzyme active sites.

3.5.6 Homology Modelling of MIF Mutants

100 models were generated for each nsSNP and in each instance the Python script would select the one with the best DOPE-Z score. A total of 78 models were generated, ranging from one chain mutation, two chain mutations to three chain mutations in the homotrimer structure. Validations were done using PROCHECK [172] and I-TASER [179]. The MIF 3DJH structure was of high resolution (1.25 Å) and the mutants only had a few point mutations (ranging from 1 to 3) introduced, the resulting structures were of good quality and there was no need for structure refinement.

3.5.7 Changes in Free energy

Most globular proteins are stable under physiological conditions, with their overall thermodynamic stability (ΔG folding) in the range of −5 to −15 kcal/mol [160], however this value is normally not considered absolute as it can have a large error [138]. Large positive values are however generally considered indicative of problems with the protein architecture and it is advised that the structures be scrutinised to check for any errors. As can be seen in Figure 3.4, there were some unusually high ΔΔG in some structures, but this can be explained due to a decrease in side hydrogen bonds in the FOLDX energy terms. The extremely high values included structures from: rs199774339, rs199980863, rs200500959, rs200995600, rs201742529, rs201862457, rs202066662 and rs372052137 (of which with the exception of rs200995600 were all predicted hotspot interface residues). Extremely low values included: rs11548056, rs182012324 both of which are predicted hot spot residues. However there was no clear distinction based on changes in free energy which nsSNPs were likely to be deleterious or not. This is of consequence because changes in free energy alone are not enough to predict changes in a proteins function and kinetics [164].

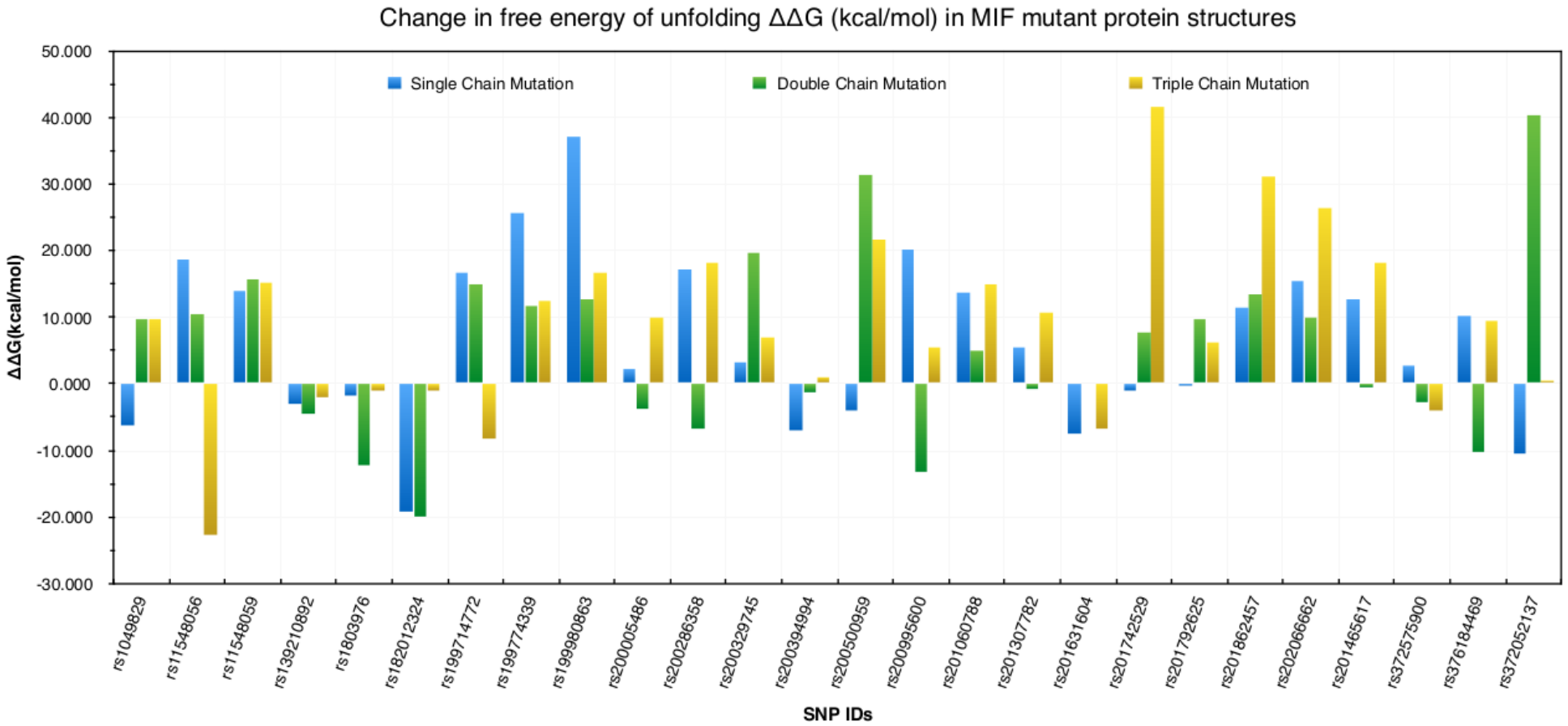

Figure 3.4: Bar-chart showing the Changes in free energy of unfolding in the MIF mutant protein structures

3.5.8 Protein Interface Network Analysis

This was done to further understand why the ΔΔGs were so large, and if changes in the protein interface could be a factor. It was discovered through network analysis that given the close location of the nsSNPs, that they were affecting similar interface residues some of which had been predicted to be hot spots with alanine scanning, these include: PHE 4, ARG 94, PHE 119 and ARG 209, as shown in Table 3.7. As can be seen in Table 3.7 rs201631604, did not change the interface residue network, which correlates with its low ΔΔG value. Further studies in docking [180] are necessary to further understand the impact of residues on function in ligand binding. This is important because protein interface interaction networks affect a protein's function and kinetics [164].

Table 3.7: List of Residues lost and Gained at the interface of Triple Chain MIF Mutants

SNP ID	Lost Protein interface residues	New Protein interface residues
rs1049829	CYS 60,PHE 4,ARG 209,ARG 94,ASP 101,PHE 119,ALA 230	PHE 59,ASP 45,PHE 174,SER 169,PRO 44,ASP 275,SER 284,PHE 289,CYS 175,ASP 216,GLY 82,VAL 130
rs11548056	CYS 60,PHE 4,CYS 290,ARG 209,VAL 245,ARG 94,PHE 119	SER 54,CYS 175,PRO 44,THR 298,THR 183,PRO 232,THR 68,GLY 82,PRO 117,VAL 130
rs11548059	ARG 94,GLY 197,PHE 4,CYS 290,GLY 312,ARG 209,PHE 344,ASP 101,PHE 119	GLY 82,ASP 216,PHE 19,ASP 275,GLN 274,VAL 130,GLN 159
rs139210892	ARG 94, ARG 209, PHE 4,ALA 115,CYS 290,GLY 312,PHE 119	ASP 45,SER 169,PRO 44,ILE 113,ASP 331,CYS 175,PRO 232,ILE 343,SER 54,GLY 82,PRO 117,ILE 228
rs1803976	ILE 68,ILE 183,PHE 4,VAL 15,CYS 290,ILE 298,ARG 209,ILE 235,VAL 245,GLY 69,ARG 94,ASP 101,PHE 119	GLY 82, SER 336,PRO 44,SER 106,ASP 275,VAL 130,SER 221
rs182012324	ARG 94,ARG 209,PHE 4,CYS 290,ILE 235,PHE 119	SER 54,SER 169,PRO 44,CYS 175,GLY 82,VAL 130
rs199714772	ARG 94, ARG 209,PHE 4,PHE 119	GLY 82,CYS 175,MET 15,PRO 44,ASP 331,SER 176,PRO 232,MET 245,MET 130
rs199774339	CYS 60,GLY 197,PHE 4,GLY 312,ARG 209,ARG 94,PHE 119,PHE 229	ARG 92,CYS 175,PRO 44,SER 284,ARG 322,ARG 207,ASP 216,PRO 232,VAL 130
rs199980863	ARG 94,GLY 197,PHE 4,ARG 209,PHE 119	,SER 54, ,CYS 175, ,LEU 35,PRO 44,ASP 216,LEU 265,GLY 82,VAL 130,LEU 150
rs200005486	ASN 7,ASN 122,PHE 4,ASN 237,CYS 290,GLY 312,ARG 209,ILE 235,VAL 245,CYS 60,GLY 69,ARG 94,ASP 101,PHE 119,ARG 127,PHE 229	CYS 57,SER 169, PHE 19,PRO 41,ASP 331,ASP 216,PRO 271,GLY 82,VAL 130,PRO 156
rs200286358	ARG 94,GLY 184,PHE 4,ARG 12,CYS 290,ARG 209,ASP 101,PHE 119	SER 54,SER 169,PRO 44,CYS 329,CYS 175,CYS 214,GLY 82,CYS 99,VAL 130
rs200329745	GLY 69,ARG 209,PHE 4,VAL 15,CYS 290,PHE 344,ARG 94,ASP 101,PHE 119	ASP 45,VAL 345,PRO 44,VAL 115,GLY 82,VAL 130,VAL 230
rs200394994	CYS 60,GLY 197,PHE 4,CYS 290,ARG 209,ARG 94,PHE 229	ASP 45,SER 169,PRO 44,SER 284,ASP 331,CYS 175,PRO 232,SER 54,GLY 82,VAL 130
rs200500959	CYS 60,GLY 197,PHE 4,ARG 209,VAL 245,PHE 344,ARG 94,PHE 119	ASP 45, ASP 216, PRO 44,SER 284,ASP 331,GLY 82,ARG 117,VAL 130
rs200995600	CYS 60,ARG 209,PHE 4,CYS 290,ARG 94,PHE 119	VAL 130,SER 169,SER 24,PRO 44,ASP 331,CYS 175,PRO 232,PHE 249,SER 254,SER 139
rs201060788	CYS 60,ARG 209,PHE 4,CYS 290,GLY 312,ARG 94,ASP 101,PHE 119,ARG 127	GLY 82,CYS 175,PRO 44,SER 284,PHE 234,VAL 130
rs201307782	CYS 60,ARG 209,PHE 4,CYS 290,VAL 245,ARG 94,ASP 101	ASP 45,CYS 175 ,ASP 331,PRO 232
rs201631604		VAL 130
rs201742529	CYS 60,GLY 197, PHE 4,CYS 290,ARG 209,ARG 94,ASP 101,PHE 119	GLY 82,SER 169,PRO 44,SER 322,SER 207,SER 92,VAL 130
rs201792625	ARG 94, ARG 209,PHE 4,VAL 15,CYS 290,VAL 245,PHE 119,SER 136	,GLY 82 ,CYS 175 ,VAL 327,SER 176,VAL 212,PRO 232,VAL 97,PRO 117,VAL 130,PHE 134
rs201862457	ARG 94,GLY 197,PHE 4,PHE 114,CYS 290,ARG 209,ASP 101,PHE 119	ASP 45,ASP 216,PRO 44,ASP 275,ASP 331,GLY 82,VAL 130
rs202066662	CYS 60,GLY 184,VAL 15,VAL 43,CYS 290,GLY 197,ARG 209,ARG 94,PHE 119	GLY 82,SER 169,TYR 41,PRO 44,ASP 331,CYS 175,ASP 216,PRO 232,TYR 271,VAL 130,TYR 156
rs201465617	CYS 60,ARG 209,PHE 4,GLY 69,ARG 94,PHE 119,ALA 230	ASP 45,SER 169,MET 5,PRO 44,CYS 175,SER 176,ASP 216,PRO 232,MET 235,MET 120
rs372575900	PHE 4,GLY 312,GLY 197,ARG 209,ARG 94,ASP 101,PHE 119	LEU 48,LEU 163,PRO 44,LEU 278,SER 284,SER 169,CYS 175,ASP 216,GLY 82,VAL 130
rs376184469	CYS 60,ARG 209,PHE 4,CYS 290,GLY 312,ARG 94,PHE 119,ALA 230	VAL 130,PRO 232,PRO 44,ASP 331,PHE 249
rs372052137	ARG 94,GLY 197,PHE 4,CYS 290,GLY 312,ARG 209,ASP 101,PHE 119	GLY 82 ,VAL 158,PRO 2,PRO 44,THR 327,CYS 175,THR 212,PRO 232,THR 97,VAL 130

3.5.9 Molecular Dynamics Simulations

To gain a better insight into the effects of the nsSNPs on the stability and folding behaviour of the MIF structure, we performed 1000ps molecular dynamics simulations using three structures for each nsSNP (single, double and triple chain mutations respectively). As shown in Figure 3.5, native and mutant structures showed a similar patterns of deviation for the duration of the 1000ps run from their starting structures with few exceptions, but this is confounded by the fact that this is a very short run. The average backbone RMSD ranged from approximately 0.0125 to 0.16 (nm) during the simulation, it is not possible to conclude when the structures would retain their maximum deviation given the short length of the simulation run. Given that MIF is a homotrimer, symmetry plays a role in its overall stability which can be seen as some nsSNP structures such as rs20206662 and rs200394994 varied more widely from the wild type structure in cases where there was a single mutation as compared to a triple mutation, this requires further investigation.

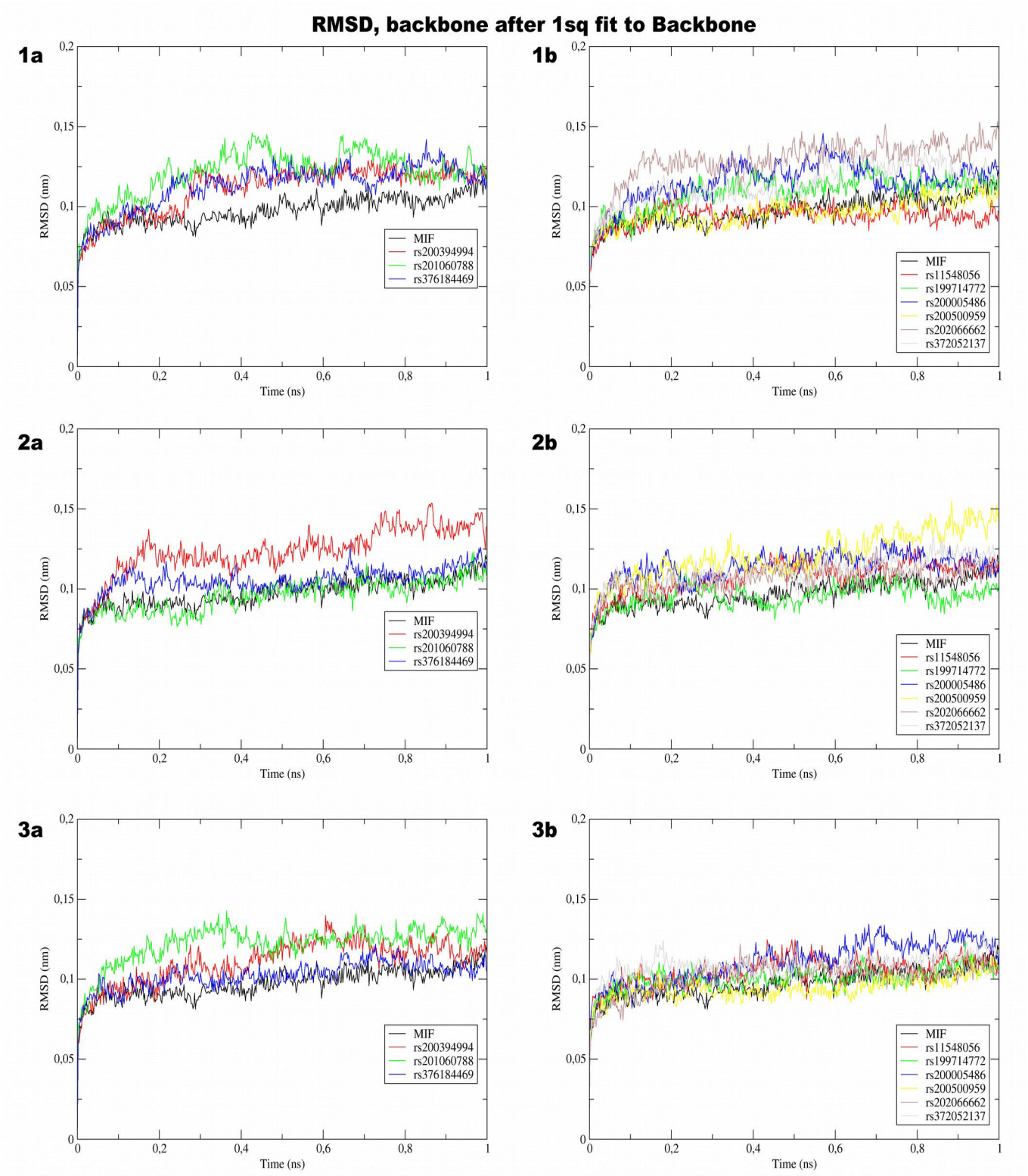

Figure 3.5: Backbone RMSD of side chains of wild type and mutant structures during the simulation. 1a-Predicted non-consequential nsSNP modelled structures with single chain mutation, 1b- Predicted deleterious nsSNP modelled structures with single chain mutation, 2a-Predicted non-consequential nsSNP modelled structures with double chain mutation, 2b-Predicted deleterious nsSNP modelled structures with double chain mutation, 3a- Predicted non-consequential nsSNP modelled structures with triple chain mutation, and 3b-Predicted deleterious nsSNP modelled structures with triple chain mutation.

The radius of gyration which refers to the mass-weight root mean-square distance of a collection of atoms from their common center of mass allowing for the analysis of the overall dimensions of the protein. Radius of gyration plot for Cα atoms of the wild type and mutant protein versus time at 1000 ps is shown in Figure 3.6. The native MIF structure and Mutant structures all showed an Rg value of ~1.86 nm at 0 ps, with the wild type MIF structure having an Rg value of ~1.87 nm at 1000 ps, the mutant structures at 1000ps ranged from ~1.88 to ~1.90 as shown in Table 3.8a and Table 3.8b.

Table 3.8a: Radius of Gyration Results for Non-deleterious nsSNP modelled Structures

SNP ID	**Number of Chain Mutations**	**Radius of Gyration (nm)**		
		0 (ps)	**500 (ps)**	**1000 (ps)**
rs200394994	Single	1.86	1.89	1.87
	Double	1.86	1.88	1.89
	Triple	1.86	1.89	1.89
rs201060788	Single	1.86	1.88	1.89
	Double	1.86	1.88	1.88
	Triple	1.86	1.87	1.89
rs376184469	Single	1.86	1.87	1.88
	Double	1.86	1.88	1.89
	Triple	1.86	1.87	1.88

Table 3.8b: Radius of Gyration Results for deleterious nsSNP modelled Structures

SNP ID	**Number of Chain Mutations**	**Radius of Gyration (nm)**		
		0 (ps)	**500 (ps)**	**1000 (ps)**
rs11548056	Single	1.86	1.87	1.88
	Double	1.86	1.88	1.87
	Triple	1.86	1.88	1.89
rs199714772	Single	1.86	1.88	1.88
	Double	1.86	1.87	1.87
	Triple	1.86	1.89	1.89
rs200005486	Single	1.86	1.88	1.88
	Double	1.86	1.87	1.87
	Triple	1.86	1.88	1.89
rs200500959	Single	1.86	1.87	1.88
	Double	1.86	1.89	1.89
	Triple	1.86	1.89	1.88
rs202066662	Single	1.86	1.90	1.90
	Double	1.860	1.89	1.90
	Triple	1.86	1.88	1.87
rs372052137	Single	1.86	1.88	1.88
	Double	1.86	1.88	1.89
	Triple	1.86	1.87	1.86

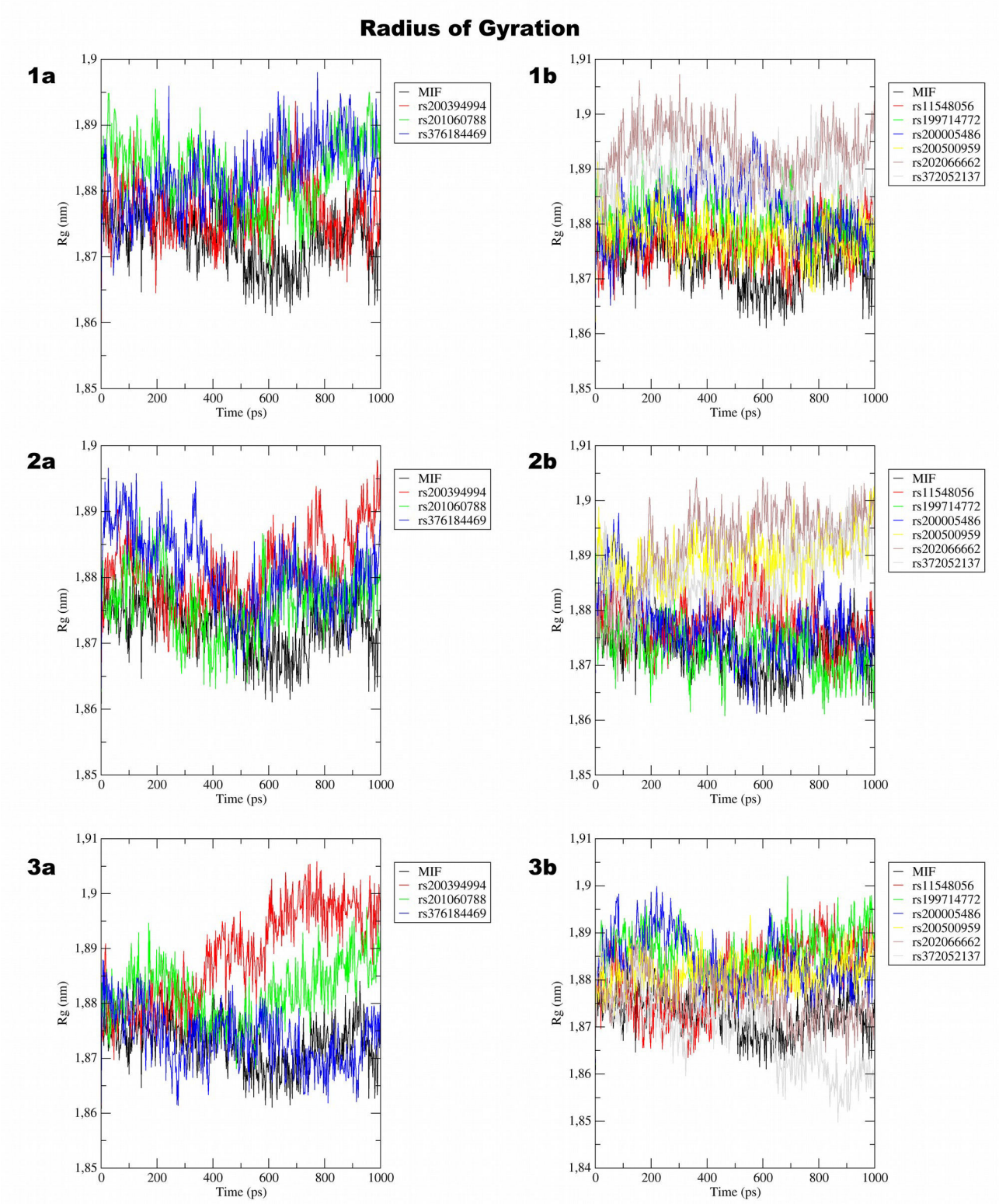

Figure 3.6: Radius of gyration of wild type and mutant structures during the simulation. 1a- Predicted non-consequential nsSNP modelled structures with single chain mutation, 1b- Predicted deleterious nsSNP modelled structures with single chain mutation, 2a-Predicted non-consequential nsSNP modelled structures with double chain mutation, 2b-Predicted deleterious nsSNP modelled structures with double chain mutation, 3a- Predicted non-consequential nsSNP modelled structures with triple chain mutation, and 3b-Predicted deleterious nsSNP modelled structures with triple chain mutation.

Molecular Dynamics simulations are a valuable approach in determining the consequences of mutations on the overall protein structure and kinetics [165, 166, 174, 181]. However to observe more substantial trends in RMSD, Rg, RMSF, solvent accessibility surface area, as a result of point mutations it is necessary to perform long simulation runs, 1000ps is not sufficient to draw in conclusive findings. The molecular dynamics simulations were too short, for more information it is necessary to perform longer runs, for example for 20ns.

3.6 Conclusion

A total of 26 nsSNPs were retrieved from dbSNP (as of July 2015). Of these 10 (38%) were not validated, while 16 (62%) were validated. Of the nsSNPs validated by 1000 Genomes [167] rs139210892, rs1803976 and rs182012324 had a global minor allele frequency (MAF) of 0.0004. The remaining three namely rs201631604, rs201862457, rs372575900 had a global MAF of 0.0002. While these were technically rare variants, many were shown to be possibly deleterious to the structure and function of MIF. Through consensus of the three SNP effect prediction tools, the following nsSNPs were shown to be likely deleterious: rs11548056, rs139210892, rs199714772, rs200005486, rs200500959, rs201307782, rs201631604, rs202066662, and rs372052137. The following were predicted to be likely benign: rs199980863, rs200394994, rs201060788, rs201792625, rs201465617, rs372575900, and rs376184469. The remaining 10 had conflicting predictions .i.e in some cases one tool predicted deleterious while the other two predicted no effect.

Alanine scanning showed that some of the nsSNP sites were indeed hotspots and more likely to be hotspots if they were interface residues. However it is worth noting that nsSNP site rs182012324 (S75) was not an interface residue, but was a hotspot because the serine residue at that site is important in the helix secondary structure [176]. Overall alanine scanning provides a computationally cheap and quick way of scanning which sites in the native protein are likely to be of importance in structural stability as such help in the selection of nsSNPs of importance.

PIC analysis showed that some of the nsSNP sites were interacting with other's which could be a basis for the detection of compensatory mutations but this requires further experimentation. Co-occurance of the different nsSNPs within the same regions MIF regions of multiple individuals would go a long way in shedding more light on this phenomenon. It serves to note that the PIC analysis algorithm does have a few errors such as the assignment of acceptor and hydrogen bond donor groups being interchanged in the main-chain peptide bond atoms. It's distance and bond angle estimations are also prone to errors.

Changes in free energy of folding using FOLDX alone were not enough to predict which nsSNPs

would adversely affect MIF's function and kinetics. Rosetta has been shown to perform better than FOLDX when it comes to assessing changes in protein stability due to point mutations [157]. FOLDX was however used in this study because of its robustness.

Protein interface network analysis revealed that even a point mutation can change the protein interface residue network, and this may not be dependent on whether or not the residue is an interface residue itself. This assay in conjunction with ligand-docking could help shed more light on nsSNPs of importance.

The molecular dynamics simulations were very too short for any meaningful inferences to be made about the trends in RMSD, Rg, RMSF, and solvent accessibility surface area. However from looking at the RMSDs and Rgs of rs372052137 when compared between single, double or triple chain mutations there is a slight difference. This brings a question relating to protein symmetry and its effect on stability.

CHAPTER FOUR

CONCLUSIONS AND RECOMMENDATIONS

4.1 Conclusions

Out of the eight variants called in the mif coding region there was only one missense variant rs36065127 (clinical significance unknown). It was not possible to test for association of this variant with HAT due to its low global MAF that was less than 0.05. This study was the first genetic study carried out on the people of Lugbara ethnicity from North Western Uganda. The nsSNP (rs36065127) though a rare variant has to be further characterised and it's role in the MIF protein structure and function assessed.

As sequencing technologies advance and the cost for sequencing drops, more and more data is being generated on human variation [182]. This allows for better understanding of human variation and how it affects disease states, susceptibility, prevention and control [183–185]. While the initial hope of this study was to discover novel variants and then predict their impact on the MIF protein function and structure, it has become evident that understanding how nsSNPs affect protein structure and function *in silico* is an area in need of further development. No single published tool is capable of accurately predicting the effects of nsSNPs on protein function and stability. More exhaustive methods such as docking followed by molecular dynamics are needed to have a more accurate picture of how nsSNPs and point mutations affect binding residues, and interface networks, however this takes time is computationally costly. There is need to come up with simple fast algorithms that can do preliminary screening of the effects nsSNPs.

4.2 Recommendations

To get a clearer picture of the effect of nsSNPs on protein structure, function, kinetics and overall stability it is necessary to carry out further investigations. Ligand docking followed by longer MD simulations will help to study how the nsSNP affect the interface residues, binding energy and the protein's flexibility. While the use of FOLDX to asses the changes in free energy of folding was not able to distinguish between deleterious and non-deleterious nsSNPs, more sensitive tools like Rosetta may better characterise the data. Further genomic studies are also needed to gain a better insight into the phenomenon of compensatory mutations. The rs36065127 variant in spite of its low minor allelic frequency requires further study especially with a bigger sample size, as it may only appear rare simply because there aren't enough previously un-sequenced populations present in the databases.

REFERENCES

1. Barrett MP, Burchmore RJS, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, Krishna S: **The trypanosomiases**. In *Lancet. Volume 362*; 2003:1469–1480.

2. Franco JR, Simarro PP, Diarra A, Jannin JG: **Epidemiology of human African trypanosomiasis.** *Clin Epidemiol* 2014, **6**:257–75.

3. Faria J, Moraes CB, Song R, Pascoalino BS, Lee N, Siqueira-Neto JL, Cruz DJM, Parkinson T, Ioset J-R, Cordeiro-da-Silva A, Freitas-Junior LH: **Drug Discovery for Human African Trypanosomiasis: Identification of Novel Scaffolds by the Newly Developed HTS SYBR Green Assay for Trypanosoma brucei**. *J Biomol Screen* 2015, **20**:70–81.

4. **Report of the first WHO stakeholders meeting on gambiense human African trypanosomiasis elimination**. 2014(March):25–27.

5. Simarro PP, Diarra A, Ruiz Postigo JA, Franco JR, Jannin JG, Postigo JAR, Franco JR, Jannin JG: **The Human African trypanosomiasis control and surveillance programme of the World Health Organization 2000-2009: the way forward.** *PLoS Negl Trop Dis* 2011, **5**:e1007.

6. Lumbala C, Simarro PP, Cecchi G, Paone M, Franco JR, Kande Betu Ku Mesu V, Makabuza J, Diarra A, Chansy S, Priotto G, Mattioli RC, Jannin JG: **Human African trypanosomiasis in the Democratic Republic of the Congo: disease distribution and risk.** *Int J Health Geogr* 2015, **14**:20.

7. Barry JD, McCulloch R: **Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite.** *Adv Parasitol* 2001, **49**:1–70.

8. Dyer N a., Rose C, Ejeh NO, Acosta-Serrano A: **Flying tryps: Survival and maturation of trypanosomes in tsetse flies**. *Trends Parasitol* 2013, **29**:188–196.

9. Chappuis F, Loutan L, Simarro P, Lejon V, Buscher P: **Options for field diagnosis of human african trypanosomiasis**. *ClinMicrobiolRev* 2005, **18**(0893-8512 (Print)):133–146.

10. Babokhov P, Sanyaolu AO, Oyibo W a, Fagbenro-Beyioku AF, Iriemenam NC: **A current analysis of chemotherapy strategies for the treatment of human African trypanosomiasis.** *Pathog Glob Health* 2013, **107**:242–52.

11. Stuart K, Brun R, Croft S, Fairlamb A, Gürtler RE, McKerrow J, Reed S, Tarleton R: **Kinetoplastids: related protozoan pathogens, different diseases.** *J Clin Invest* 2008, **118**:1301–10.

12. Baral TN: **Immunobiology of African trypanosomes: need of alternative interventions.** *J Biomed Biotechnol* 2010, **2010**:389153.

13. Bucheton B, Macleod A, Jamonneau V: **Human host determinants influencing the outcome of Trypanosoma brucei gambiense infections**. *Parasite Immunology* 2011:438–447.

14. Checchi F, Filipe JAN, Barrett MP, Chandramohan D: **The natural progression of Gambiense sleeping sickness: what is the evidence?** *PLoS Negl Trop Dis* 2008, **2**:e303.

15. Garcia A, Courtin D, Solano P, Koffi M, Jamonneau V: **Human African trypanosomiasis: connecting parasite and host genetics**. *Trends Parasitol* 2006, **22**:405–409.

16. Jamonneau V, Ilboudo H, Kaboré J, Kaba D, Koffi M, Solano P, Garcia A, Courtin D, Laveissière C, Lingue K, Büscher P, Bucheton B: **Untreated human infections by trypanosoma brucei gambiense are not 100% fatal**. *PLoS Negl Trop Dis* 2012, **6**.

17. Sternberg JM, Maclean L: **A spectrum of disease in human African trypanosomiasis: the host and parasite genetics of virulence.** *Parasitology* 2010, **137**:2007–2015.

18. Stijlemans B, Beschin A, Magez S, Van Ginderachter JA, De Baetselier P: **Iron Homeostasis and Trypanosoma brucei Associated Immunopathogenicity Development: A Battle/Quest for Iron**. *Biomed Res Int* 2015, **2015**:1–15.

19. Stijlemans B, Vankrunkelsven A, Caljon G, Bockstal V, Guilliams M, Bosschaerts T, Beschin A, Raes G, Magez S, De Baetselier P: **The central role of macrophages in trypanosomiasis-associated anemia: rationale for therapeutical approaches.** *Endocr Metab Immune Disord Drug Targets* 2010, **10**:71–82.

20. Naessens J: **Bovine trypanotolerance: A natural ability to prevent severe anaemia and haemophagocytic syndrome?** *International Journal for Parasitology* 2006:521–528.

21. Magez S, Caljon G, Tran T, Stijlemans B, Radwanska M: **Current status of vaccination against African trypanosomiasis.** *Parasitology* 2010, **137**:2017–2027.

22. Weiss G, Goodnough LT: **Anemia of chronic disease.** *N Engl J Med* 2005, **352**:1011–1023.

23. Barrett MP, Boykin DW, Brun R, Tidwell RR: **Human African trypanosomiasis: pharmacological re-engagement with a neglected disease.** *Br J Pharmacol* 2007, **152**:1155–71.

24. Matovu E, Enyaru JCK, Legros D, Schmid C, Seebeck T, Kaminsky R: **Melarsoprol refractory T. b. gambiense from Omugo, north-western Uganda**. *Trop Med Int Heal* 2001, **6**:407–411.

25. Matovu E, Seebeck T, Enyaru JCK, Kaminsky R: **Drug resistance in Trypanosoma brucei spp., the causative agents of sleeping sickness in man and nagana in cattle**. *Microbes Infect* 2001, **3**:763–770.

26. Worthen C, Jensen BC, Parsons M: **Diverse effects on mitochondrial and nuclear functions elicited by drugs and genetic knockdowns in bloodstream stage Trypanosoma brucei.** *PLoS Negl Trop Dis* 2010, **4**:e678.

27. Steinmann P, Stone CM, Sutherland CS, Tanner M, Tediosi F: **Contemporary and emerging strategies for eliminating human African trypanosomiasis due to Trypanosoma brucei gambiense: review**. *Trop Med Int Heal* 2015, **20**:707 – 718.

28. Keating J, Yukich JO, Sutherland CS, Woods G, Tediosi F: **Human African trypanosomiasis prevention, treatment and control costs: A systematic review.** *Acta Trop* 2015, **150**(JUNE):4–13.

29. Bosschaerts T, Guilliams M, Stijlemans B, De Baetselier P, Beschin A: **Understanding the role of monocytic cells in liver inflammation using parasite infection as a model**. *Immunobiology*

2009:737–747.

30. Stijlemans B, Baral TN, Guilliams M, Brys L, Korf J, Drennan M, Van Den Abbeele J, De Baetselier P, Magez S: **A glycosylphosphatidylinositol-based treatment alleviates trypanosomiasis-associated immunopathology.** *J Immunol* 2007, **179**:4003–4014.

31. Paralkar V, Wistow G: **Cloning the human gene for macrophage migration inhibitory factor (MIF).** *Genomics* 1994, **19**:48–51.

32. Kozak CA, Adamson MC, Buckler CE, Segovia L, Paralkar V, Wistow G: **Genomic cloning of mouse MIF (macrophage inhibitory factor) and genetic mapping of the human and mouse expressed gene and nine mouse pseudogenes.** *Genomics* 1995, **27**:405–411.

33. Babu SN, Chetal G, Kumar S: **Macrophage migration inhibitory factor: a potential marker for cancer diagnosis and therapy.** *Asian Pac J Cancer Prev* 2012, **13**:1737–44.

34. Israelson A, Ditsworth D, Sun S, Song S, Liang J, Hruska-Plochan M, McAlonis-Downes M, Abu-Hamad S, Zoltsman G, Shani T, Maldonado M, Bui A, Navarro M, Zhou H, Marsala M, Kaspar BK, Da Cruz S, Cleveland DW: **Macrophage Migration Inhibitory Factor as a Chaperone Inhibiting Accumulation of Misfolded SOD1.** *Neuron* 2015, **86**:218–32.

35. Weiser WY, Weiser WY, Temple PA, Temple PA, Witek-Giannotti JS, Witek-Giannotti JS, Remold HG, Remold HG, Clark SC, Clark SC, David JR, David JR: **Molecular cloning of a cDNA encoding a human macrophage migration inhibitory factor**. *Proc Natl Acad Sci U S A* 1989, **86**:7522–7526.

36. Bozza M, Kolakowski LF, Jenkins NA, Gilbert DJ, Copeland NG, David JR, Gerard C: **Structural characterization and chromosomal location of the mouse macrophage migration inhibitory factor gene and pseudogenes.** *Genomics* 1995, **27**:412–419.

37. Mitchell RA, Liao H, Chesney J, Fingerle-Rowson G, Baugh J, David J, Bucala R: **Macrophage migration inhibitory factor (MIF) sustains macrophage proinflammatory function by inhibiting p53: regulatory role in the innate immune response.** *Proc Natl Acad Sci U S A* 2002, **99**:345–350.

38. Sun HW, Bernhagen J, Bucala R, Lolis E: **Crystal structure at 2.6-A resolution of human macrophage migration inhibitory factor.** *Proc Natl Acad Sci U S A* 1996, **93**:5191–5196.

39. Crichlow G V, Lubetsky JB, Leng L, Bucala R, Lolis EJ: **Structural and kinetic analyses of macrophage migration inhibitory factor active site interactions.** *Biochemistry* 2009, **48**:132–9.

40. Swope M, Sun HW, Blake PR, Lolis E: **Direct link between cytokine activity and a catalytic site for macrophage migration inhibitory factor.** *EMBO J* 1998, **17**:3534–3541.

41. Whitman CP: **The 4-oxalocrotonate tautomerase family of enzymes: How nature makes new enzymes using a ??-??-?? structural motif**. *Arch Biochem Biophys* 2002, **402**:1–13.

42. Sarkar S, Siddiqui AA, Mazumder S, De R, Saha SJ, Banerjee C, Iqbal MS, Adhikari S, Alam A, Roy S, Bandyopadhyay U: **Ellagic Acid, a Dietary Polyphenol, Inhibits Tautomerase Activity of Human Macrophage Migration Inhibitory Factor and Its Pro-inflammatory Responses in**

Human Peripheral Blood Mononuclear Cells. *J Agric Food Chem* 2015, **63**:4988–98.

43. Calandra T, Roger T: **Macrophage migration inhibitory factor: a regulator of innate immunity.** *Nat Rev Immunol* 2003, **3**:791–800.

44. Burger-Kentischer A, Goebel H, Seiler R, Fraedrich G, Schaefer HE, Dimmeler S, Kleemann R, Bernhagen J, Ihling C: **Expression of macrophage migration inhibitory factor in different stages of human atherosclerosis**. *Circulation* 2002, **105**(1524-4539 (Electronic)):1561–1566.

45. Willis MS, Carlson DL, Dimaio JM, White MD, White DJ, Adams GA th, Horton JW, Giroir BP: **Macrophage migration inhibitory factor mediates late cardiac dysfunction after burn injury**. *Am J Physiol Hear Circ Physiol* 2005, **288**:H795–804.

46. Jovanović Krivokuća M, Stefanoska I, Abu Rabi T, Al-Abed Y, Stošić-Grujičić S, Vićovac L: **Pharmacological inhibition of MIF interferes with trophoblast cell migration and invasiveness.** *Placenta* 2015, **36**:150–9.

47. Leng L, Bucala R: **Insight into the biology of Macrophage Migration Inhibitory Factor (MIF) revealed by the cloning of its cell surface receptor**. *Cell Res* 2006, **16**:162–168.

48. Morand EF, Leech M, Weedon H, Metz C, Bucala R, Smith MD: **Macrophage migration inhibitory factor in rheumatoid arthritis: clinical correlations.** *Rheumatology (Oxford)* 2002, **41**:558–562.

49. Stijlemans B, Vankrunkelsven A, Brys L, Raes G, Magez S, De Baetselier P: **Scrutinizing the mechanisms underlying the induction of anemia of inflammation through GPI-mediated modulation of macrophage activation in a model of African trypanosomiasis**. *Microbes Infect* 2010, **12**:389–399.

50. Larson DF, Horak K: **Macrophage migration inhibitory factor: controller of systemic inflammation.** *Crit Care* 2006, **10**:138.

51. Ayoub S, Hickey MJ, Morand EF: **Mechanisms of disease: macrophage migration inhibitory factor in SLE, RA and atherosclerosis.** *Nat Clin Pract Rheumatol* 2008, **4**:98–105.

52. Rosado JDD, Rodriguez-Sosa M: **Macrophage migration inhibitory factor (MIF): A Key player in protozoan infections**. *Int J Biol Sci* 2011, **7**:1239–1256.

53. Bernhagen J, Krohn R, Lue H, Gregory JL, Zernecke A, Koenen RR, Dewor M, Georgiev I, Schober A, Leng L, Kooistra T, Fingerle-Rowson G, Ghezzi P, Kleemann R, McColl SR, Bucala R, Hickey MJ, Weber C: **MIF is a noncognate ligand of CXC chemokine receptors in inflammatory and atherogenic cell recruitment.** *Nat Med* 2007, **13**:587–596.

54. Calandra T, Bernhagen J, Metz CN, Spiegel LA, Bacher M, Donnelly T, Cerami A, Bucala R: **MIF as a glucocorticoid-induced modulator of cytokine production.** *Nature* 1995, **377**:68–71.

55. Flaster H, Bernhagen J, Calandra T, Bucala R: **The macrophage migration inhibitory factor-glucocorticoid dyad: regulation of inflammation and immunity.** *Mol Endocrinol* 2007, **21**:1267–1280.

56. Wang F-F, Huang X-F, Shen N, Leng L, Bucala R, Chen S-L, Lu L-J: **A genetic role for macrophage migration inhibitory factor (MIF) in adult-onset Still's disease.** *Arthritis Res Ther* 2013, **15**:R65.

57. Greven D, Leng L, Bucala R: **Autoimmune diseases: MIF as a therapeutic target.** *Expert Opin Ther Targets* 2010, **14**:253–264.

58. Stijlemans B, Leng L, Brys L, Sparkes A, Vansintjan L, Caljon G, Raes G, Van Den Abbeele J, Van Ginderachter J a., Beschin A, Bucala R, De Baetselier P: **MIF Contributes to Trypanosoma brucei Associated Immunopathogenicity Development**. *PLoS Pathog* 2014, **10**:e1004414.

59. Liu G, Xu J, Wu H, Sun D, Zhang X, Zhu X, Magez S, Shi M: **IL-27 Signaling Is Crucial for Survival of Mice Infected with African Trypanosomes via Preventing Lethal Effects of CD4+ T Cells and IFN-γ**. *PLOS Pathog* 2015, **11**:e1005065.

60. Reyes JL, Terrazas LI, Espinoza B, Cruz-Robles D, Soto V, Rivera-Montoya I, Gómez-García L, Snider H, Satoskar AR, Rodríguez-Sosa M: **Macrophage migration inhibitory factor contributes to host defense against acute Trypanosoma cruzi Infection**. *Infect Immun* 2006, **74**:3170–3179.

61. Terrazas CA, Huitron E, Vazquez A, Juarez I, Camacho GM, Calleja EA, Rodriguez-Sosa M: **MIF synergizes with trypanosoma cruzi antigens to promote efficient dendritic cell maturation and IL-12 production via p38 MAPK**. *Int J Biol Sci* 2011, **7**:1298–1310.

62. Wu SP, Leng L, Feng Z, Liu N, Zhao H, McDonald C, Lee A, Arnett FC, Gregersen PK, Mayes MD, Bucala R: **Macrophage migration inhibitory factor promoter polymorphisms and the clinical expression of scleroderma**. *Arthritis Rheum* 2006, **54**:3661–3669.

63. Das R, Koo M-S, Kim BH, Jacob ST, Subbian S, Yao J, Leng L, Levy R, Murchison C, Burman WJ, Moore CC, Scheld WM, David JR, Kaplan G, MacMicking JD, Bucala R: **Macrophage migration inhibitory factor (MIF) is a critical mediator of the innate immune response to Mycobacterium tuberculosis.** *Proc Natl Acad Sci U S A* 2013, **110**:E2997–3006.

64. Baugh JA, Chitnis S, Donnelly SC, Monteiro J, Lin X, Plant BJ, Wolfe F, Gregersen PK, Bucala R: **A functional promoter polymorphism in the macrophage migration inhibitory factor (MIF) gene associated with disease severity in rheumatoid arthritis.** *Genes Immun* 2002, **3**:170–176.

65. Donn R, Alourfi Z, Zeggini E, Lamb R, Jury F, Lunt M, Meazza C, De Benedetti F, Thomson W, Ray D, Abinun M, Becker M, Bell A, Craft A, Crawley E, David J, Foster H, Gardener-Medwin J, Griffin J, Hall A, Hall M, Herrick A, Hollingworth P, Holt L, Jones S, Pountain G, Ryder C, Southwood T, Stewart I, Venning H, et al.: **A Functional Promoter Haplotype of Macrophage Migration Inhibitory Factor Is Linked and Associated with Juvenile Idiopathic Arthritis**. *Arthritis Rheum* 2004, **50**:1604–1610.

66. Sreih A, Ezzeddine R, Leng L, Lachance A, Yu G, Mizue Y, Subrahmanyan L, Pons-Estel BA, Abelson AK, Gunnarsson I, Svenungsson E, Cavett J, Glenn S, Zhang L, Montgomery R, Perl A, Salmon J, Alarcón-Riquelme ME, Harley JB, Bucala R: **Dual effect of the macrophage migration inhibitory factor gene on the development and severity of human systemic lupus**

erythematosus. *Arthritis Rheum* 2011, **63**:3942–3951.

67. Donn RP, Shelley E, Ollier WE, Thomson W: **A novel 5'-flanking region polymorphism of macrophage migration inhibitory factor is associated with systemic-onset juvenile idiopathic arthritis.** *Arthritis Rheum* 2001, **44**:1782–1785.

68. Donn R, Alourfi Z, De Benedetti F, Meazza C, Zeggini E, Lunt M, Stevens A, Shelley E, Lamb R, Ollier WER, Thomson W, Ray D, Abinun M, Becker M, Bell a., Craft a., Crawley E, David J, Foster H, Gardener-Medwin J, Griffin J, Hall a., Hall M, Herrick a., Hollingworth P, Holt L, Jones S, Pountain G, Ryder C, Southwood T, et al.: **Mutation screening of the macrophage migration inhibitory factor gene: Positive association of a functional polymorphism of macrophage migration inhibitory factor with juvenile idiopathic arthritis**. *Arthritis Rheum* 2002, **46**:2402–2409.

69. Ilboudo H, Bras-Gonçalves R, Camara M, Flori L, Camara O, Sakande H, Leno M, Petitdidier E, Jamonneau V, Bucheton B: **Unravelling human trypanotolerance: IL8 is associated with infection control whereas IL10 and TNFα are associated with subsequent disease development.** *PLoS Pathog* 2014, **10**:e1004469.

70. **Trypanosomiasis, Human African (sleeping sickness).**

71. Johansen Taber K a, Dickinson BD, Wilson M: **The promise and challenges of next-generation genome sequencing for clinical care.** *JAMA Intern Med* 2014, **174**:275–80.

72. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR: **A comparative analysis of exome capture**. *Genome Biol* 2011, **12**:R97.

73. Wang Z, Liu X, Yang B-Z, Gelernter J: **The role and challenges of exome sequencing in studies of human diseases.** *Front Genet* 2013, **4**(August):160.

74. Liu X, Han S, Wang Z, Gelernter J, Yang BZ: **Variant Callers for Next-Generation Sequencing Data: A Comparison Study**. *PLoS One* 2013, **8**:1–11.

75. Bolger a. M, Lohse M, Usadel B: **Trimmomatic: A flexible read trimming tool for Illumina NGS data**. *Bioinformatics* 2014, **30**:2114–2120.

76. Langmead B: **Alignment with Bowtie**. 2011, **48**:1–24.

77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–2079.

78. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB: **Quality scores and SNP detection in sequencing-by-synthesis systems.** *Genome Res* 2008, **18**:763–70.

79. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**:186–194.

80. Li M, Nordborg M, Li LM: **Adjust quality scores from alignment and improve sequencing**

accuracy. *Nucleic Acids Res* 2004, **32**:5183–5191.

81. Li H: **Aligning new-sequencing reads by BWA BWA : Burrows-Wheeler Aligner**. 2010.

82. Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, Adams MD, Sun S: **How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?** *BioData Min* 2012, **5**:6.

83. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.

84. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851–8.

85. Burrows M, Wheeler D: **A block-sorting lossless data compression algorithm**. *Algorithm, Data Compression* 1994:18.

86. Ferragina P, Manzini G: **Opportunistic data structures with applications**. *Proceeding FOCS '00 Proc 41st Annu Symp Found Comput Sci FOCS '00 Proc 41st Annu Symp Found Comput Sci* 2000:390–398.

87. Ruffalo M, LaFramboise T, Koyuturk M: **Comparative analysis of algorithms for next-generation sequencing read alignment**. *Bioinformatics* 2011, **27**:2790–2796.

88. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing**. *Brief Bioinform* 2010, **11**:473–483.

89. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**:1297–303.

90. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing**. *arXiv Prepr arXiv12073907* 2012:9.

91. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G: **Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications**. *Nat Genet* 2014, **46**:912–918.

92. Homer N, Nelson SF: **Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA**. *Genome Biol* 2010, **11**:R99.

93. Van der Auwera G, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V., Altshuler D, Gabriel S, DePristo M: *From fastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. 2013.

94. DePristo M, Banks E, Poplin RE, Garimella K V., Maguire JR, Hartl C, Philippakis, Del Angel G, Rivas M, Hanna M, McKenna, Fennell TJ, Kernytsky M, Sivachenko Y, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ, Schmidt S, Van der Auwera G, Carneiro MO, Hartl C, Poplin RE, Del

Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E: **Measuring absorptive capacity**. *Nat Genet* 2002, **11**:11.10.1–11.10.33.

95. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L: **Identification and correction of systematic error in high-throughput sequence data**. *BMC Bioinformatics* 2011, **12**:451.

96. Zook JM, Samarov D, McDaniel J, Sen SK, Salit M: **Synthetic Spike-in Standards Improve Run-Specific Systematic Error Analysis for DNA and RNA Sequencing**. *PLoS One* 2012, **7**:e41356.

97. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ: **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.** *Genome Med* 2013, **5**:28.

98. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN: **Demonstrating stratification in a European American population.** *Nat Genet* 2005, **37**:868–872.

99. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, Flicek P, Gabriel SB, Gibbs R a., Knoppers BM, Lander ES, Lehrach H, Mardis ER, McVean G a., Nickerson D a., Peltonen L, Schafer AJ, Sherry ST, Wang J, Wilson RK, Gibbs R a., Deiros D, Metzker M, Muzny D, Reid J, et al.: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061–1073.

100. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2**. In *Current Protocols in Human Genetics*. *Volume Chapter 7*; 2013(January):7.20.1–7.20.41.

101. Cooper GM, Stone E a., Asimenos G, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence**. *Genome Res* 2005, **15**:901–913.

102. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4**:1073–1081.

103. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease-causing potential of sequence alterations.** *Nature methods* 2010:575–576.

104. Jorgensen TJ, Ruczinski I, Kessing B, Smith MW, Shugart YY, Alberg AJ: **Hypothesis-driven candidate gene association studies: Practical design and analytical considerations**. *Am J Epidemiol* 2009, **170**:986–993.

105. Zondervan KT, Cardon LR: **Designing candidate gene and genome-wide case-control association studies.** *Nat Protoc* 2007, **2**:2492–2501.

106. Patnala R, Clements J, Batra J: **Candidate gene association studies: a comprehensive guide to useful in silico tools.** *BMC Genet* 2013, **14**:39.

107. Adriaens ME, Jaillard M, Waagmeester A, Coort SLM, Pico AR, Evelo CTA: **The public road to high-quality curated biological pathways**. *Drug Discovery Today* 2008:856–862.

108. Baxevanis AD: **The importance of biological databases in biological discovery**. *Curr Protoc Bioinforma* 2011(SUPPL. 34):1–6.

109. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics* 2002, **18 Suppl 2**:S110–S115.

110. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5**:16–18.

111. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho Y-J, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, et al.: **The landscape of somatic copy-number alteration across human cancers.** *Nature* 2010, **463**:899–905.

112. Conrad DF, Keebler JEM, DePristo M, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella K V, Zilversmit M, Cartwright R, Rouleau G, Daly M, Stone E, Hurles ME, Awadalla P: **Variation in genome-wide mutation rates within and between human families.** *Nat Genet* 2011, **43**:712–714.

113. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley N a, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, et al.: **A comprehensive catalogue of somatic mutations from a human cancer genome.** *Nature* 2010, **463**:191–6.

114. Roach JC, Glusman G, Smit AF a, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science* 2010, **328**:636–639.

115. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**:2156–8.

116. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant analysis of next-generation genome sequencing data**. *Brief Bioinform* 2014, **15**:256–278.

117. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559–75.

118. Price AL, Zaitlen N a, Reich D, Patterson N: **New approaches to population stratification in genome-wide association studies.** *Nat Rev Genet* 2010, **11**:459–463.

119. Schork NJ, Murray SS, Frazer K a., Topol EJ: **Common vs. rare allele hypotheses for complex diseases**. *Curr Opin Genet Dev* 2009, **19**:212–219.

120. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnusdottir DN, Johannsdottir H, Helgadottir HT, Stacey SN, Jonasdottir A, Olafsdottir SB, Thorleifsson G, Jonasson JG, Tryggvadottir L, Navarrete S, Fuertes F, Helfand BT, Hu Q, Csiki IE, Mates IN, Jinga V, Aben KK, van Oort IM, Vermeulen SH, Donovan JL, Hamdy FC: **A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer**. *Nat Genet* 2012, **44**:1326–1329.

121. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson P V., Bjornsson S, Stefansson H, Sulem P, Gudbjartsson D, Maloney J, Hoyte K, Gustafson A, Liu Y, Lu Y, Bhangale T, Graham RR, Huttenlocher J, Bjornsdottir G, Andreassen O a., Jönsson EG, Palotie A, Behrens TW, Magnusson OT, Kong A, Thorsteinsdottir U, Watts RJ, Stefansson K: **A mutation in APP protects against Alzheimer's disease and age-related cognitive decline**. *Nature* 2012, **488**:96–99.

122. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagacé C, Neale B, Lo KS, Schumm P, Törkvist L, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S: **Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.** *Nat Genet* 2011, **43**:1066–73.

123. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**:308–311.

124. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: **Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.** *Proc Natl Acad Sci U S A* 2009, **106**:19096–19101.

125. Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, Bairoch A: **Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase**. *Human Mutation* 2008, **29**:361–366.

126. Stenson PD, Mort M, Ball E V., Shaw K, Phillips AD, Cooper DN: **The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine**. *Human Genetics* 2014:1–9.

127. Mueller SC, Backes C, Kalinina O V., Meder B, Stöckel D, Lenhof H-P, Meese E, Keller A: **BALL-SNP: combining genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms**. *Genome Med* 2015, **7**:65.

128. Thusberg J, Olatubosun A, Vihinen M: **Performance of mutation pathogenicity prediction methods on missense variants**. *Hum Mutat* 2011, **32**:358–368.

129. Collins FS, Guyer MS, Chakravarti A: **Variations on a theme: cataloging human DNA sequence variation.** *Science* 1997, **278**:1580–1581.

130. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB: **Bioinformatics challenges for personalized medicine**. *Bioinformatics* 2011, **27**:1741–1748.

131. Giacomini KM, Brett CM, Altman RB, Benowitz NL, Dolan ME, Flockhart D a, Johnson J a, Hayes DF, Klein T, Krauss RM, Kroetz DL, McLeod HL, Nguyen a T, Ratain MJ, Relling M V, Reus V, Roden DM, Schaefer C a, Shuldiner a R, Skaar T, Tantisira K, Tyndale RF, Wang L, Weinshilboum RM, Weiss ST, Zineh I: **The pharmacogenetics research network: from SNP discovery to clinical drug response.** *Clin Pharmacol Ther* 2007, **81**:328–345.

132. Bromberg Y, Rost B: **SNAP: Predict effect of non-synonymous polymorphisms on function**. *Nucleic Acids Res* 2007, **35**:3823–3835.

133. Sunyaev S, Ramensky V, Bork P: **Towards a structural basis of human non-synonymous single nucleotide polymorphisms**. *Trends Genet* 2000, **16**:15–17.

134. Venselaar H, Te Beek T a H, Kuipers RKP, Hekkelman ML, Vriend G: **Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces.** *BMC Bioinformatics* 2010, **11**:548.

135. Bao L, Zhou M, Cui Y: **nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms**. *Nucleic Acids Res* 2005, **33**(SUPPL. 2).

136. Mort M, Sterne-Weiler T, Li B, Ball E V, Cooper DN, Radivojac P, Sanford JR, Mooney SD: **MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing.** *Genome Biol* 2014, **15**:R19.

137. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **MutPred: Automated inference of molecular mechanisms of disease from amino acid substitutions**. *Bioinformatics* 2009, **25**:2744–2750.

138. Serrano L, Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F: **The FoldX web server: an online force field**. *Nucl Acids Res* 2005, **33**:W382–388.

139. Krieger E, Vriend G: **YASARA View - molecular graphics for all devices - from smartphones to workstations.** *Bioinformatics* 2014:1–2.

140. van Durme J, Delgado J, Stricher F, Serrano L, Schymkowitz J, Rousseau F: **A graphical interface for the FoldX forcefield**. *Bioinformatics* 2011, **27**:1711–1712.

141. Morrison KL, Weiss G a.: **Combinatorial alanine-scanning**. *Curr Opin Chem Biol* 2001, **5**:302–307.

142. Davis BH, Poon AFY, Whitlock MC: **Compensatory mutations are repeatable and clustered within proteins.** *Proc Biol Sci* 2009, **276**(February):1823–1827.

143. Tina KG, Bhadra R, Srinivasan N: **PIC: Protein Interactions Calculator**. *Nucleic Acids Res* 2007, **35**:473–476.

144. Sánchez R, Pieper U, Melo F, Eswar N, Martí-Renom M a, Madhusudhan MS, Mirković N, Sali a: **Protein structure modeling for structural genomics.** *Nat Struct Biol* 2000, **7 Suppl**(november):986–990.

145. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science (80-)* 2001,

294:93–6.

146. Chothia C, Lesk M: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5**:823–826.

147. Marti-Renom M, Stuart C, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes**. *Annu Rev Biophys Biomol Struct* 2000, **29**:291–325.

148. Dixit A, Verkhivker GM: **Structure-functional prediction and analysis of cancer mutation effects in protein kinases.** *Comput Math Methods Med* 2014, **2014**:653487.

149. Oyer J a, Huang X, Zheng Y, Shim J, Ezponda T, Carpenter Z, Allegretta M, Okot-Kotber CI, Patel JP, Melnick a, Levine RL, Ferrando a, Mackerell a D, Kelleher NL, Licht JD, Popovic R: **Point mutation E1099K in MMSET/NSD2 enhances its methyltranferase activity and leads to altered global chromatin methylation in lymphoid malignancies.** *Leukemia* 2014, **28**(Cll):198–201.

150. Sonawane P, Patel K, Vishwakarma RK, Singh S, Khan BM: **in Silico mutagenesis and docking studies of active site residues suggest altered substrate specificity and possible physiological role of Cinnamoyl CoA Reductase 1 (Ll-CCRH1).** *Bioinformation* 2013, **9**:224–32.

151. Vyas V, Ukawala R, Chintha C, Ghate M: **Homology modeling a fast tool for drug discovery: Current perspectives**. *Indian Journal of Pharmaceutical Sciences* 2012:1.

152. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-Y, Pieper U, Sali A: **Comparative protein structure modeling using MODELLER.** *Curr Protoc Protein Sci* 2007, **Chapter 2**:Unit 2.9.

153. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J Mol Biol* 1993, **234**:779–815.

154. Fiser A, Sali A: **ModLoop: Automated modeling of loops in protein structures**. *Bioinformatics* 2003, **19**:2500–2501.

155. Rohl CA, Strauss CEM, Misura KMS, Baker D: **Protein structure prediction using Rosetta.** *Methods Enzymol* 2004, **383**:66–93.

156. Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L: **Prediction of water and metal binding sites and their affinities by using the Fold-X force field.** *Proc Natl Acad Sci U S A* 2005, **102**:10147–10152.

157. Thiltgen G, Goldstein R a: **Assessing predictors of changes in protein stability upon mutation using self-consistency.** *PLoS One* 2012, **7**:e46084.

158. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: An online force field**. *Nucleic Acids Res* 2005, **33**(SUPPL. 2).

159. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations**. *J Mol Biol* 2002, **320**:369–387.

160. Tokuriki N, Tawfik DS: **Stability effects of mutations and protein evolvability**. *Curr Opin*

Struct Biol 2009, **19**:596–604.

161. Matsuura Y, Takehira M, Sawano M, Ogasahara K, Tanaka T, Yamamoto H, Kunishima N, Katoh E, Yutani K: **Role of charged residues in stabilization of *Pyrococcus horikoshii* CutA1, which has a denaturation temperature of nearly 150 °c**. *FEBS J* 2012, **279**:78–90.

162. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F: **SNPeffect: A database mapping molecular phenotypic effects of human non-synonymous coding SNPs**. *Nucleic Acids Res* 2005, **33**(DATABASE ISS.):527–532.

163. DeLano WL: **The PyMOL Molecular Graphics System**. *Schrödinger LLC wwwpymolorg* 2002, **Version 1.**:http://www.pymol.org.

164. Tenge VR, Gounder AP, Wiens ME, Lu W, Smith JG: **Delineation of interfaces on human alpha-defensins critical for human adenovirus and human papillomavirus inhibition.** *PLoS Pathog* 2014, **10**:e1004360.

165. Achary MS, Reddy ABM, Chakrabarti S, Panicker SG, Mandal AK, Ahmed N, Balasubramanian D, Hasnain SE, Nagarajaram HA: **Disease-causing mutations in proteins: structural analysis of the CYP1B1 mutations causing primary congenital glaucoma in humans.** *Biophys J* 2006, **91**:4329–39.

166. Offman MN, Krol M, Silman I, Sussman JL, Futerman AH: **Molecular basis of reduced glucosylceramidase activity in the most common Gaucher disease mutant, N370S**. *J Biol Chem* 2010, **285**:42105–42114.

167. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**:56–65.

168. Nantasenamat C, Prachayasittikul V, Bulow L: **Molecular modeling of the human hemoglobin-haptoglobin complex sheds light on the protective mechanisms of haptoglobin.** *PLoS One* 2013, **8**:e62996.

169. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary N a., Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, Dicuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: **RefSeq: An update on mammalian reference sequences**. *Nucleic Acids Res* 2014, **42**:756–763.

170. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235–242.

171. Edgar RC: **MUSCLE: Multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792–1797.

172. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM: **AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.** *J Biomol NMR* 1996, **8**:477–486.

173. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC: **GROMACS: Fast, flexible, and free**. *Journal of Computational Chemistry* 2005:1701–1718.

174. Kamaraj B, Purohit R: **In silico screening and molecular dynamics simulation of disease-associated nsSNP in TYRP1 gene and its structural consequences in OCA3.** *Biomed Res Int* 2013, **2013**:697051.

175. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A: **LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources.** *Bioinformatics* 2005, **21**:2814–20.

176. Ballesteros J a, Deupi X, Olivella M, Haaksma EE, Pardo L: **Serine and threonine residues bend alpha-helices in the chi(1) = g(-) conformation.** *Biophys J* 2000, **79**:2754–2760.

177. Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley B a., Hendricks MM, Iimura S, Gajiwala K, Scholtz JM, Grimsley GR: **Contribution of hydrophobic interactions to protein stability**. *J Mol Biol* 2011, **408**:514–528.

178. Chikenji G, Fujitsuka Y, Takada S: **Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study.** *Proc Natl Acad Sci U S A* 2006, **103**:3141–3146.

179. Roy A, Kucukural A, Zhang Y: **I-TASSER: a unified platform for automated protein structure and function prediction.** *Nat Protoc* 2010, **5**:725–738.

180. Jacob RB, Andersen T, McDougal OM: **Accessible high-throughput virtual screening molecular docking software for students and educators**. *PLoS Comput Biol* 2012, **8**.

181. Cregut D, Serrano L: **Molecular dynamics as a tool to detect protein foldability. A mutant of domain B1 of protein G with non-native secondary structure propensities**. *Protein Sci* 2008, **8**:271–282.

182. Hayden EC: **The $1,000 genome**. *Nature* 2014, **507**:295.

183. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe a, Warrington J, Lipshutz R, Daley GQ, Lander ES: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.** *Nat Genet* 1999, **22**:231–238.

184. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation**. *Genome Res* 1998, **8**:1229–1231.

185. George Priya Doss C, Nagasundaram N, Chakraborty C, Chen L, Zhu H: **Extrapolating the effect of deleterious nsSNPs in the binding adaptability of flavopiridol with CDK7 protein: a molecular dynamics approach**. *Hum Genomics* 2013, **7**:10.

APPENDICES

Appendix 1: R-code for Running a Principal Component Analysis Using SNPRelate

```
##Load the R packages: gdsfmt and SNPRelate
library(gdsfmt)
library(SNPRelate)
##setting the working directoy
getwd()
setwd("/home/phillip/Desktop/Running_PCA/OUT")
##Reading the vcf file into R
vcf.fn <- "Trypanogen_merged_chr1_UO_only.vcf"
##Converting VCF to GDS format
snpgdsVCF2GDS(vcf.fn, "test.gds", method="biallelic.only")
### To get a summary of the data (number of individuals and variants filtered)
snpgdsSummary("test.gds")
## Reading the the GDS file
genofile <- snpgdsOpen("test.gds")
## Running the PCA
pca <- snpgdsPCA(genofile)
pc.percent <- 100 * pca$eigenval[1:16]/sum(pca$eigenval) #first 16 PCA's
###To get the sample ids
sample.id <- read.gdsn(index.gdsn(genofile, "sample.id"))
### Assigning the population codes
pop_code <- scan("pop2.txt", what=character())
## Creating a Data Frame for the results
tab <- data.frame(sample.id = pca$sample.id,
  pop = factor(pop_code)[match(pca$sample.id, sample.id)],
  EV1 = pca$eigenvect[,1],  # the first eigenvector
  EV2 = pca$eigenvect[,2],  # the second eigenvector
  stringsAsFactors = FALSE)

## Plotting the results and generating a high resolution image
png(filename ="Chr1_PCA_UO_trypanogen.png",width = 8, height = 5, units = 'in',
res=300)
plot(tab$EV2, tab$EV1, col=as.integer(tab$pop),
  xlab="eigenvector 2", ylab="eigenvector 1")
legend("topright", legend=levels(tab$pop), pch="o", col=1:nlevels(tab$pop))
dev.off()
```

Appendix 2: Python Scripts for Introducing Point Mutations

point_mutation.py

```
### this script takes a fasta file and a list of nsSNPs as the two arguments
### the goal being for it to generate several fasta files with the introduced mutations
###-------------importing the data-------------
import sys
import textwrap
import os
import os.path
from sys import argv
filename=sys.argv[1]
filename2=sys.argv[2]
###----function for parsing fasta file---------
def read_fasta(filename):
  name = None
  name2seq = {}
  for line in open(filename):
    if line.startswith(">"):
      if name:
        name2seq[name]=seq
      name=line[1:].rstrip()
      seq=""
    else:
      seq+=line.rstrip()
  name2seq[name]=seq
  return name2seq
###----function for parsing the nsSNP file--the format is a the wildtype amino acid, then the position then the mutant amino acid
def read_nsSNP_list(filename2):
  f=open(filename2,"r")
  lines = f.readlines()
  f.close()
  return lines
###---- dictionary data structure for the fasta file-----
fasta_dic=read_fasta(filename)
for key, value in fasta_dic.items() :
  wild_type_seq = value #the variable wild_type_seq is the protein sequence
  header = key
```

```
###----- building the dictionary data structure for the nsSNP_list----
lines=read_nsSNP_list(filename2)
nsSNP_list_dic={}
for i in range(len(lines)):
  snp_ID = lines[i].split()[0]
  mutation = lines[i].split()[1]
  nsSNP_list_dic[snp_ID] = mutation
SNP_ID_list=[]
for key, value in nsSNP_list_dic.items():
  SNP_ID_list+= [key] #contains the SNP IDs
###----data structure for mutated fasta sequences-----
adjusted_mutation_position = 0
mutants_dic={}
seq_as_list=[]
mutated_seq_list=[]
for j in range(len(SNP_ID_list)):
  mutation_position=int(nsSNP_list_dic[SNP_ID_list[j]][1:-1])
   adjusted_mutation_position = mutation_position - 1
if nsSNP_list_dic[SNP_ID_list[j]][0]==wild_type_seq[adjusted_mutation_position]:
    seq_as_list = list(wild_type_seq)
    seq_as_list[adjusted_mutation_position]=nsSNP_list_dic[SNP_ID_list[j]][-1]
    mutated_seq_list = seq_as_list
    mutants_dic[SNP_ID_list[j]]=''.join(mutated_seq_list)
  else:
    print('\nERROR could not find amino acid '+ (nsSNP_list_dic[SNP_ID_list[j]]
[0]) +' at position '+str(mutation_position) +' in the specified fasta file')
####----writing out the new fasta files-------
counter=0
fasta_query=input("\nTo write the new mutant fasta files in the current
directory Please Enter Y or hit ENTER to exit: ")
if fasta_query.upper() == "Y":
  #filename="mutated_fasta_files/" + SNP_ID_list[j] + '.fa' #so that they go to
that directory
  file_name=None
  while fasta_query!="":
    for j in range(len(SNP_ID_list)):
      file_name = SNP_ID_list[j] + '.fa' #generating the file names
       if os.path.exists(file_name)==True: #to check if the file already exists
to prevent overwriting
        print('\n****WARNING file directory already exists***')
```

```
      else:
        fasta_file=open(file_name,'a')
                                fasta_file.writelines("\n".join(textwrap.wrap(">{0}|
{1}".format(header,SNP_ID_list[j]))))
        fasta_file.writelines("\n"+"\n".join(textwrap.wrap(mutants_dic[SNP_ID_li
st[j]],70)).upper())
         fasta_file.writelines("\n") #trouble concatenating later, introduced to
help
        fasta_file.close()
        counter += 1
    fasta_query=input("\nPress ENTER to exit: ")
###-------------summary---------
os.system('clear')
print('-----------------------------------------------------------------------')
print('\nWildtype fasta file enterred \n{0} \n{1}'.format(header,wild_type_seq))
print('-----------------------------------------------------------------------')
print('\n{0} mutant fasta files written to current directory'.format(counter))
print('-----------------------------------------------------------------------')
```

Appendix 3: Python Scripts for carrying out Muscle Alignment and conversion to PIR format

muscle_alignment.py

```
import os
import os.path
fasta_mut_dir = "/home/phillip/Desktop/research_work/fasta_mut"
structure_sequence                                                    =
"/home/phillip/Desktop/research_work/fasta_sequences_templates/3DJH.fasta.txt"
output_path = "/home/phillip/Desktop/research_work/alignment_files/fasta_format"
msa_output_path                                                       =
"/home/phillip/Desktop/research_work/alignment_files/muscle_msa_output"
list_dir = []
list_dir = os.listdir(fasta_mut_dir)
list_dir2= []
list_dir2 = os.listdir(output_path)
counter1=0
counter2=0
for i in range(len(list_dir)):
  name = list_dir[i][0:-3]
  fasta_file = fasta_mut_dir+"/"+name+".fa "
  output_name = output_path + "/"+name + "_mfa_seq.fa"
  os.system("cat "+ fasta_file + structure_sequence + " > " + output_name)
  counter1+=1
```

```
for j in range(len(list_dir2)):
  seqs_fa = output_path + "/" + list_dir2[j]
      msa_output_name   =   msa_output_path   +   "/"+   list_dir2[j][0:-11]
+'_3djh_aligned.afa'
  os.system('muscle -in ' + seqs_fa + ' -out ' + msa_output_name)
  counter2+=1
print(str(counter1) + " files concatenated and " + str(counter2) + " msa files
written                                                                    to
/home/phillip/Desktop/research_work/alignment_files/muscle_msa_output")
```

fasta_to_pir.py

```
#Script to convert fasta to .pir format for the modelling
import os
output_path_mut_1                                                          =
"/home/phillip/Desktop/research_work/alignment_files/pir_format_1_mut"
output_path_mut_2                                                          =
"/home/phillip/Desktop/research_work/alignment_files/pir_format_2_mut"
output_path_mut_3                                                          =
"/home/phillip/Desktop/research_work/alignment_files/pir_format_3_mut"
muscle_msa_path                                                            =
"/home/phillip/Desktop/research_work/alignment_files/muscle_msa_output"

wildtype_seq="""MPMFIVNTNVPRASVPDGFLSELTQQLAQATGKPPQYIAVHVVPDQLMAFGGSSEPCALCSLHS
IGKIGG AQNRSYSKLLCGLLAERLRISPDRVYINYYDMNAANVGWNNSTFA"""

structure_x = """>P1;3DJH.pdb
structureX:3DJH.pdb: 1 :A: 115:C:::2.00:0.25
-PMFIVNTNVPRASVPDGFLSELTQQLAQATGKPPQYIAVHVVPDQLMAFGGSSEPCALCSLHSIGKIGGAQ
NRSYSKLLCGLLAERLRISPDRVYINYYDMNAANVGWNNSTFA
/
-PMFIVNTNVPRASVPDGFLSELTQQLAQATGKPPQYIAVHVVPDQLMAFGGSSEPCALCSLHSIGKIGGAQ
NRSYSKLLCGLLAERLRISPDRVYINYYDMNAANVGWNNSTFA
/
-PMFIVNTNVPRASVPDGFLSELTQQLAQATGKPPQYIAVHVVPDQLMAFGGSSEPCALCSLHSIGKIGGAQ
NRSYSKLLCGLLAERLRISPDRVYINYYDMNAANVGWNNSTFA*"""

###-------------importing the data------------------
def readfastamsa(filename):
  f=open(filename,"r")
  lines = f.readlines()
```

```
  f.close()
  return lines
#---------------loop ----------------------
list_dir = []
list_dir = os.listdir(muscle_msa_path)
counter=0

for i in range(len(list_dir)):

  name = muscle_msa_path+"/"+list_dir[i]
  name2 = list_dir[i][0:-17]
  lines=readfastamsa(name)
  file_name1 = output_path_mut_1+"/"+name2+"_1mut.pir"
  file_name2 = output_path_mut_2+"/"+name2+"_2mut.pir"
  file_name3 = output_path_mut_3+"/"+name2+"_3mut.pir"
  one_mut_file=open(file_name1,'w')
  one_mut_file.writelines(">P1;"+ name2)
  one_mut_file.writelines("\nsequence:"+name2+":1:A:345:C::::")
  one_mut_file.writelines("\n"+ (lines[1]+lines[2]).replace("\n",""))
  one_mut_file.writelines("\n/"+"\n"+ wildtype_seq)
  one_mut_file.writelines("\n/"+"\n"+ wildtype_seq + "*" )
  one_mut_file.writelines("\n"+ structure_x)
  one_mut_file.close()
  two_mut_file=open(file_name2,'w')
  two_mut_file.writelines(">P1;"+ name2)
  two_mut_file.writelines("\nsequence:"+name2+":1:A:345:C::::")
  two_mut_file.writelines("\n"+ (lines[1]+lines[2]).replace("\n",""))
  two_mut_file.writelines("\n/"+ "\n" + (lines[1]+lines[2]).replace("\n",""))
  two_mut_file.writelines("\n/"+"\n"+ wildtype_seq + "*" )
  two_mut_file.writelines("\n"+ structure_x)
  two_mut_file.close()
  three_mut_file=open(file_name3,'w')
  three_mut_file.writelines(">P1;"+ name2)
  three_mut_file.writelines("\nsequence:"+name2+":1:A:345:C::::")
  three_mut_file.writelines("\n"+ (lines[1]+lines[2]).replace("\n",""))
  three_mut_file.writelines("\n/"+ "\n"+ (lines[1]+lines[2]).replace("\n",""))
   three_mut_file.writelines("\n/"+ "\n"+ (lines[1]+lines[2]).replace("\n","") +
"*" )
  three_mut_file.writelines("\n"+ structure_x)
```

```
  three_mut_file.close()
  counter+=1
print(str(counter*3)+" files written")
```

Appendix 4: Python Script used for modelling MIF mutants

Model.py

```
#This was the Python script used for modelling adapted from Sali labs
# Homology modelling by the automodel class
from modeller import *
from modeller.automodel import *           # Load the automodel class
log.verbose()                                          # request verbose output
env = environ()                                        #  create  a  new  MODELLER
environment to build this model in
import sys
import os
from sys import argv
pir_file_name = sys.argv[1]
sequence_name = sys.argv[2]
top_models_dir = sys.argv[3]
#top_models_dir = sys.argv[3]
# directories for input atom files
env.io.atom_files_directory = ['/jabba/JMS/users/phillip/modelling_tasks']
a = automodel(env,
       alnfile = pir_file_name, #'1is8_1wur.pir',                      #
alignment filename
       knowns  = '3DJH.pdb', #('1IS8_A', '1WUR_A'),               # codes of the
templates
           sequence = sequence_name, assess_methods=(assess.DOPE, assess.GA341))
#'Pfalciparum', assess_methods=(assess.DOPE, assess.GA341))        # code of the
target
a.starting_model= 1                                 # index of the first model
a.ending_model = 100                                       # index of the last model
                  # (determines how many models to calculate)
a.final_malign3d = True # generate superimposed templatesand model (*_fit.pdb
files)


a.md_level = None          # No refinement of model


a.make()          # do the actual homology modelling
# Get a list of all successfully built models from a.outputs
```

```
ok_models = [x for x in a.outputs if x['failure'] is None]


# Rank the models by DOPE score
key = 'DOPE score'
if sys.version_info[:2] == (2,3):
  # Python 2.3's sort doesn't have a 'key' argument
  ok_models.sort(lambda a,b: cmp(a[key], b[key]))
else:
  ok_models.sort(key=lambda a: a[key])

# Get top model
m = ok_models[0]
print("Top model: {0} (DOPE score {1})".format(m['name'], m[key]))
os.system("cp {0} {1}".format(m['name'], top_models_dir)) #copy the top model to
that file directory
```

Modelling_job_submission.py

```
##the purpose of this script was to automate the modelling jobs on the RUBi
cluster
import os
pir_path                                                                    =
"/jabba/JMS/users/phillip/modelling_tasks/pir_files/pir_format_1_mut/"
model_script = "/jabba/JMS/users/phillip/modelling_tasks/scripts/model.py"
output_path  =  "/jabba/JMS/users/phillip/modelling_tasks/output/mut1_run"  #run
job from here
job_files_output_path = "/jabba/JMS/users/phillip/modelling_tasks/job_files"
top_models_dir                                                              =
"/jabba/JMS/users/phillip/modelling_tasks/output/top_mut_models_1_muts/"
error_files_path = "/jabba/JMS/users/phillip/modelling_tasks/error_files"
list_dir = []
list_dir = os.listdir(pir_path)
for i in range(len(list_dir)):
      pir_filename = list_dir[i]
      sequence_name = list_dir[i][0:-9]
      jobname = job_files_output_path + "/"+ sequence_name + "1_mut" + ".job"
      jobtext = open(jobname,"w")
      jobtext.writelines("#!/bin/sh" + "\n")
      jobtext.writelines("#PBS -N Modelling" + str(i)+"\n")
      jobtext.writelines("#PBS -l nodes=1:ppn=1,walltime=99:00:00" + "\n")
      jobtext.writelines("#PBS -q opteron" + "\n")
```

```
      jobtext.writelines("#PBS -e localhost:" + error_files_path + "\n")
      jobtext.writelines(". /software/nfs/Modules/default/init/sh" + "\n")
      jobtext.writelines("module load modeller/9.15"+"\n")
      #jobtext.writelines("#PBS -m abe" + "\n")
      jobtext.writelines("modpy.sh Python " + model_script + " " + pir_path + pir_filename + " " + sequence_name + " " + top_models_dir + "\n")
      jobtext.writelines("\n")
      jobtext.writelines("\n")
      jobtext.close()
      os.system("qsub " + jobname )
```

Appendix 5 Bash Script used for running MD Simulations

run_MD.sh (adapted from a tutorial available here http://www.biocode.it/tutorial-mds-8.php)

```
#!/bin/sh
#PBS -N Energy_minimisation_phillip
#PBS -S /bin/bash
#PBS -q throughput
#PBS -l nodes=1:ppn=1,walltime=99:00:00
#PBS -d /jabba/JMS/users/phillip/energy_minimisation/
#PBS -e localhost:/jabba/JMS/users/phillip/modelling_tasks/error_files

in_MIF=/jabba/JMS/users/phillip/modelling_tasks/output/MIF_remodelled
out=/jabba/JMS/users/phillip/energy_minimisation/MIF_minimisation
softwarePATH=/software/nfs/gromacs/4.5.7/bin/

umask 0037

##step 1 file conversion
$softwarePATH/pdb2gmx -f $in_MIF/MIF.B99990074.pdb -o $out/MIF_processed.gro -water spce -ff amber03
##step 2 creating a simulation box
$softwarePATH/editconf -f MIF_processed.gro -o $out/MIF_newbox.gro -c -d 1.0 -bt cubic

##step 3 solvation
$softwarePATH/genbox -cp MIF_newbox.gro -cs spc216 -o MIF_solv.gro -p $out/topol.top
 ##step 4 Neutralising the system
$softwarePATH/grompp -f ions.mdp -c MIF_solv.gro -p topol.top -o ions.tpr
```

```
echo 13 | $softwarePATH/genion -s ions.tpr -o MIF_solv_ions.gro -p topol.top -pname NA -nname CL -nn 1 #for amber03 forcefield"
```

##step 5 Energy minimisation

```
$softwarePATH/grompp -f minim.mdp -c MIF_solv_ions.gro -p topol.top -o em.tpr
```

```
$softwarePATH/mdrun -v -deffnm em
```

```
$softwarePATH/mdrun -v -s em.tpr -o em.trr -e em.edr -c em.gro -g em.log
```

##step 6 running the evaluation

```
echo 10 0 |$softwarePATH/g_energy -f em.edr -o potential.xvg
```

xmgrace potential.xvg

##step 7 Equilibration

```
$softwarePATH/grompp -f equilib-NVT.mdp -c em.gro -p topol.top -o eq.tpr
```

```
$softwarePATH/mdrun -v -s eq.tpr -o eq.trr -e eq.edr -c eq.gro -g eq.log -cpo eq.cpt #run takes about 40mins
```

##step 8

```
echo 15 0 | $softwarePATH/g_energy -f eq.edr -o temperature.xvg
```

xmgrace temperature.xvg

first simulation 100 ps MD run with the equilibration set to the pressure of the system under the canonical NPT ensamble (Constant Number of particles, Pressure and Temperature)

```
$softwarePATH/grompp -f equilib-NPT.mdp -c eq.gro -p topol.top -o eq.tpr
```

```
$softwarePATH/mdrun -v -s eq.tpr -o eq.trr -e eq.edr -c eq.gro -g eq.log -cpi eq.cpt -cpo eq.cpt
```

#step 9 more evaluation

```
echo 15 0 | $softwarePATH/g_energy -f eq.edr -o pressure.xvg
```

##xmgrace pressure.xvg

##step 10 Production dynamics

###The system is finally ready. We can now release the position restraints and start MD for data collection. The procedure is quite similar to the previous step but, with an exception, the time necessary to obtain the final result will be much longer now (almost 35 hours on a dual-processor AMD Opteron 2Ghz with Linux CentOS). We want to run 1ns MD simulation in total but we spent 200ps for the two previous equilibration steps. The "real" MD that we can use for data collection will be 800ps.

```
$softwarePATH/grompp -f md.mdp -c eq.gro -p topol.top -o md_01.tpr
```

```
$softwarePATH/mdrun -v -s md_01.tpr -o md_01.trr -e md_01.edr -c md_01.gro -g md_01.log -cpi eq2.cpt -cpo md_01.cpt -x md_01.xtc
```

###We have a new file .xtc this time, what is this? Trajectory files (the coordinates over time), are written initially by mdrun to contain atomic positions, velocities, and forces. The type of data and the period with which they are written are controlled with .mdp file options. Gromacs will write full-precision portable-binary data in a format known as a .trr file and, optionally, a reduced-precision format for positions only, known as an .xtc file, useful for the analysis.

#step 11 analysis

##To check if the simulation finished properly, the internal tool "gmxcheck" allows you to verify if the simulation ran for the established time:

```
$softwarePATH/gmxcheck -f md_01.trr
```

##The program "trjconv" is normaly used as the first post-processing tool in order to correct a trajectory for periodicity, or to extract specific frames from a trajectory for analysis. Let's say to the program to generate a "corrected" trajectory file which can be .xtc or .pdb:

```
echo 0 | $softwarePATH/trjconv -s md_01.tpr -f md_01.xtc -o md_01-trajectory.xtc -pbc mol -ur compact
```

```
echo 0 | $softwarePATH/trjconv -s md_01.tpr -f md_01.xtc -o md_01-trajectory.pdb -pbc mol -ur compact
```

###Calculate the Convergence of the System (RMSD)

###We can now evaluate the structural stability of MD simulations using the Root Mean Square Deviation (RMSD) as an indicator of convergence of the structure towards an equilibrium state. Typing the following command:

```
echo 4 4 | $softwarePATH/g_rms -s md_01.tpr -f md_01-trajectory.xtc -o rmsd.xvg -tu ns
```

###Calculate the Radius of Gyration of the Protein (Rg)

##The radius of gyration of the protein gives an indication of the shape (compactness) of the molecule at each time. If a protein is folded, it will maintain a relatively steady value of Rg. If a protein unfolds, its Rg will change over time.

```
echo 4 | $softwarePATH/g_gyrate -s md_01.tpr -f md_01-trajectory.xtc -o gyrate.xvg
```