

**GENETIC CHARACTERIZATION AND STRUCTURE
PREDICTION OF GAMETOCYTE DEVELOPMENT 1
AND APETALA2-G IN *Plasmodium falciparum* ISOLATES
FROM BARINGO, UASIN GISHU, AND NANDI
COUNTIES, KENYA**

JOSEPHAT KIPSANG BUNGEI

**MASTER OF SCIENCE
(Molecular Biology and Bioinformatics)**

**JOMO KENYATTA UNIVERSITY OF
AGRICULTURE AND TECHNOLOGY**

2021

**Genetic characterization and structure prediction of gametocyte
development 1 and apetala2-g in *Plasmodium falciparum* isolates from
Baringo, Uasin Gishu, and Nandi Counties, Kenya**

Josephat Kipsang Bungei

**A thesis submitted in partial fulfilment of the requirements for
the degree of Master of Science in Molecular Biology and
Bioinformatics of the Jomo Kenyatta
University of Agriculture and Technology**

2021

DECLARATION

This thesis is my original work and has not been presented elsewhere for a degree award

Sign..... Date

Josephat Kipsang Bungei

This thesis has been submitted for examination with our approval as the university supervisors

Sign..... Date.....

Dr. Steven Ger Nyanjom, PhD

Biochemistry Department, School of Biomedical Sciences

JKUAT, Kenya

Sign..... Date.....

Dr. Victor Atunga Mobegi, PhD

Biochemistry Department, School of Medicine

UoN, Kenya

DEDICATION

I dedicate this thesis to my supportive wife, Scolastine Sang, and our children Carson and Claire for their love and joy that gave me the drive to achieve my goals. I would also like to dedicate it to my dear parents, Mr. & Mrs. Bungei, siblings, and friends for their support and encouragement. Therefore, I would like to appreciate your efforts and pray for God's blessings to be upon you all.

ACKNOWLEDGEMENT

Foremost, I am so grateful to God for his love, care, power, and blessings enabled me to accomplish numerous tasks that culminated in the achievement of this great work.

I am also grateful to Jomo Kenyatta University of Agriculture & Technology for providing required laboratory resources, University of Nairobi for bioinformatics analysis, AFRICA-ai-JAPAN PROJECT (JKU/ADM/10B) for funding this research under Innovation Research Fund 2017-2018, and Institutional Research and Ethics Committee (IREC) of Moi Teaching and Referral Hospital for providing ethical approval.

I would like to offer sincere appreciation to my supervisors Dr. Steven Ger Nyanjom and Dr. Victor Atunga Mobegi for their exemplary guidance, handy assistance, invaluable advice, significant patience, and persistent drive that led to successful performance of research and completion of this thesis.

I am extending my gratitude to laboratory technicians and healthcare centers of Baringo County Referral Hospital, Uasin Gishu County Hospital, and Kapsabet County Referral Hospital for aiding me in the collection of blood samples from patients infected with malaria.

I would also wish to recognize my colleagues, Sebastian Musundi, Rebecca Kerubo, Donwillians Omuoyo, Mary Maranga, Brian Musyoka, and John Lukoye, for their moral support and technical assistance in the course of my research.

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF APPENDICES	xii
ABBREVIATIONS AND/OR ACRONYMS	xiii
ABSTRACT	xv
CHAPTER ONE	1
INTRODUCTION	1
1.1 Background.....	1
1.2 Statement of the Problem	4
1.3 Justification of the Study.....	5
1.4 Research Questions	6
1.5 Hypothesis	6
1.6 Objectives.....	7
1.6.1 General Objective.....	7
1.6.2 Specific Objectives.....	7
CHAPTER TWO	8
LITERATURE REVIEW	8
2.1 Epidemiology of Malaria.....	8
2.2 Life Cycle of <i>Plasmodium</i>	8
2.3 Genome Structure of <i>Plasmodium</i>	10
2.4 Vaccine Development	11
2.4.1 Pre-Erythrocytic Vaccines.....	13
2.4.2 Erythrocytic Vaccines	14
2.4.3 Transmission-Blocking Vaccines	15

2.5 SNPs and Targets for Vaccine Development	16
2.6 Gametocytogenesis.....	17
2.7 <i>Plasmodium falciparum</i> Gametocyte Development 1.....	18
2.8 <i>Plasmodium falciparum</i> Apetala2-G.....	18
CHAPTER THREE	20
METHODOLOGY.....	20
3.1 Research Design	20
3.2 Ethical Consideration	20
3.3 Study Sites	21
3.4 Sampling of Participants.....	21
3.5 Sample Collection	23
3.6 Primer Design.....	24
3.7 Extraction of DNA	24
3.8 Target Amplification	25
3.9 Retrieval and Processing of Sequences	27
3.10 Detection of Synonymous and Non-synonymous SNPs	28
3.11 Selection Analysis	28
3.12 Protein Structure Prediction	28
3.13 Protein-Ligand Docking	30
CHAPTER FOUR.....	32
RESULTS.....	32
4.1 Description of Samples.....	32
4.2 Quantity and Quality Extracted Genomic DNA.....	34
4.3 Primers Designed.....	35
4.4 Fragment Sizes of PCR Products	35
4.5 PCR Products of Target Genes.....	36
4.5.1 The First Fragment of <i>Pfgdv1</i>	36
4.5.2 The Second Fragment of <i>Pfgdv1</i>	37
4.5.3 Fragment of <i>Pfap2g</i>	38

4.6 Sequences	39
4.6.1 Secondary Sequences	39
4.6.2 Primary Sequences	40
4.7 Single Nucleotide Polymorphisms	41
4.7.1 SNPs of <i>Pfgdv1</i> Gene	41
4.7.2 SNPs of <i>Pfap2g</i> gene.....	45
4.7.3 Effect of Non-synonymous Substitutions	49
4.8 Selection Analysis	50
4.8.1 Selection Analysis of <i>Pfgdv1</i>	50
4.8.2 Selection Analysis of <i>Pfap2g</i>	52
4.9 Protein Structure Prediction	54
4.9.1 Prediction of <i>Pfgdv1</i> Protein Structure.....	54
4.9.1.1 Primary Structure Analysis of <i>Pfdv1</i>	54
4.9.1.2 Secondary Structure Analysis of <i>Pfdv1</i> protein	56
4.9.1.3 Tertiary Structure Prediction of <i>Pfdv1</i> Protein	57
4.9.2 Prediction of <i>Pfap2g</i> Protein Structure	63
4.9.2.1 Primary Structure Analysis of <i>Pfap2g</i> Protein	63
4.9.2.2 Secondary Structure Analysis of <i>Pfap2g</i> Protein	65
4.9.2.3 Tertiary Structure Prediction of <i>Pfap2g</i>	65
4.10 Protein-Ligand Docking	71
4.10.1 Protein-Ligand Docking of <i>Pfgdv1</i>	71
4.10.2 Protein-Ligand Docking of <i>Pfap2g</i>	73
CHAPTER FIVE.....	74
DISCUSSION	74
5.1 Description of Samples.....	74
5.2 Single Nucleotide Polymorphisms	74
5.3 Protein Structure Prediction	77
5.3 Protein-Ligand Docking	80
CHAPTER SIX.....	82

CONCLUSIONS AND RECOMMENDATIONS 82
 6.1 Conclusions 82
 6.2 Recommendations 83
REFERENCES 84
APPENDICES 101

LIST OF TABLES

Table 4.1: Characteristics of malaria patients and samples collected	32
Table 4.2: Descriptive statistics of the DNA quantity, absorbance, and purity of DNA	34
Table 4.3: Attributes of primers designed.....	35
Table 4.4: Geographical distribution of <i>P. falciparum</i> isolates in PlasmodDB	40
Table 4.5: SNPs identified in <i>Pfgdv1</i> among secondary isolates	42
Table 4.6: SNPs identified in <i>Pfgdv1</i> among primary isolates	43
Table 4.7: SNPs identified in <i>Pfap2g</i> among secondary isolates	46
Table 4.8: SNPs identified in <i>Pfap2g</i> among primary isolates	47
Table 4.9: Gibbs free-energy gaps (ddG) of nsSNPs in <i>Pfgdv1</i>	49
Table 4.10: Gibbs free-energy gaps (ddG) of nsSNPs in <i>Pfap2g</i>	50
Table 4.11: Results from Tajima's neutrality test	51
Table 4.12: Selection analysis outcomes of primary data of <i>Pfgdv1</i>	51
Table 4.13: Selection analysis outcomes of secondary data of <i>Pfgdv1</i>	52
Table 4.14: Results from Tajima's neutrality test of <i>Pfap2g</i>	53
Table 4.15: Selection analysis outcomes of secondary data of <i>Pfap2g</i>	54
Table 4.16: Physicochemical properties of <i>Pfgdv1</i> as predicted by Protparam	55
Table 4.17: The leading 10 threading templates of <i>Pfgdv1</i>	60
Table 4.18: The leading 10 structural analogs of <i>Pfgdv1</i> identified in PDB	61
Table 4.19: Physicochemical properties of <i>Pfap2g</i> as predicted by Protparam.....	64
Table 4.20: The leading 10 threading templates of <i>Pfap2g</i> (1-200 and 1201-2432)	68
Table 4.21: The Leading 10 Structural Analogs of <i>Pfap2g</i> Identified in PDB	69
Table 4.22: Ligands and their respective binding sites	72
Table 4.23: Ligands and their respective binding sites	73

LIST OF FIGURES

Figure 1.1: Mechanism of gametogenesis as regulated by <i>Pfgdv1</i> , heterochromatin protein 1, and <i>Pfapg2</i>	4
Figure 2.1: The life cycle of <i>P. falciparum</i>	9
Figure 2.2: Vaccine target areas in the life cycle of the malaria parasite.	12
Figure 3.1: Kenyan map showing sampling sites.	21
Figure 4.1: The distribution of parasitaemia levels of 30 patients according to study sites and gender (A) and age groups (B)	33
Figure 4.2: PCR products of amplified target regions of <i>P. falciparum</i> genes resolved on 2% agarose gel electrophoresis.....	36
Figure 4.3: Amplicon size of 1306 bp amplified from <i>Pfgdv1</i> gene of <i>P. falciparum</i> resolved on 2% agarose gel electrophoresis	37
Figure 4.4: Amplicon size of 650 bp amplified from <i>Pfgdv1</i> gene of <i>Plasmodium falciparum</i> resolved on 2% agarose gel electrophoresis.	38
Figure 4.5: Amplicon size of 341 bp amplified from <i>Pfap2g</i> gene of <i>P. falciparum</i> resolved on 2% agarose gel electrophoresis..	39
Figure 4.6: MSA visualized using Jalview to highlight four nsSNPs in <i>Pfgdv1</i> sequences.....	44
Figure 4.7: MSA of <i>Pfap2g</i> visualized using Jalview to highlight insertion and four nsSNPs.....	48
Figure 4.8: The normalized B-factor (thermal mobility) of <i>Pfgdv1</i> protein.....	57
Figure 4.9: Prediction of <i>Pfgdv1</i> protein using InterPro showing no family, domain, and gene ontology identified.....	58
Figure 4.10: Structure prediction of the first domain A (1-143) the second domain B (144-599) of <i>Pfgdv1</i> protein using Raptor X.....	59
Figure 4.11: Global and local accuracy estimations of predicted protein structure of <i>Pfgdv1</i> of the selected model.....	62
Figure 4.12: Front view of the globally estimated <i>Pfgdv1</i> protein model (A) and highlighted three domains (B)	63

Figure 4.13: Prediction of <i>Pfap2g</i> using InterPro showing apetalla2 domain	66
Figure 4.14: Prediction of protein location, function, and structure using InterPro showing no identified	67
Figure 4.15: Local and global accuracy of the predicted model of <i>Pfap2g</i> (1-1200 residues).....	70
Figure 4.16: Local and global accuracy of the predicted model of <i>Pfap2g</i> (1201-2432 residues).....	71

LIST OF APPENDICES

Appendix I: Ethical clearance.....	101
Appendix II: Informed consent form.....	102
Appendix III: Samples collected.....	106
Appendix IV: Results of DNA Quantification Using Nanodrop Spectrophotometer ..	109
Appendix V: Sequencing results.....	111
Appendix VI: Protein prediction of <i>Pfgdv1</i>	118
Appendix VII: Protein prediction of <i>Pfap2g</i>	122
Appendix VIII: Protein ligand docking of <i>Pfgdv1</i>	126
Appendix IX: Protein ligand docking of <i>Pfap2g</i>	128
Appendix X: Publication.....	131

ABBREVIATIONS AND/OR ACRONYMS

cAMP	cyclic Adenosine Monophosphate
CDS	Coding Sequence
CellTOS	Cell-traversal protein for ookinetes and sporozoites
cGMP	cyclic Guanosine Monophosphate
COACH	Consensus Approach
ddNTPs	dideoxynucleosides Triphosphates
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleoside Triphosphates
EC	Enzyme Commission
GNP	Gametocyte Non-Producer lines
GO	Gene Ontology
HBsAg	Hepatitis B virus surface antigen
I-TASSER	Iterative Threading Assembly Refinement
LBSs	Ligand Binding Sites
MEGA	Molecular Evolutionary Genetics Analysis
MSA	Multiple Sequence Alignment
MUSCLE	Multiple Sequence Comparison by Log-Expectation
nsSNPs	Non-synonymous SNPs
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
<i>Pfap2g</i>	<i>Plasmodium falciparum</i> apetala2-g transcription factor
<i>Pfgdv1</i>	<i>Plasmodium falciparum</i> gametocyte development 1
<i>Pfge</i>	<i>Plasmodium falciparum</i> gametocytogenesis early genes
<i>PfRH5</i>	Reticulocyte Binding Protein Homologue 5
PKG	cGMP-dependent protein kinase
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA

SDS	Sodium Dodecyl Sulfate
SLAC	Single-Likelihood Ancestor Counting
SNPs	Single Nucleotide Polymorphisms
S-SITE	Sequence Site
TAE	Tris-Acetate-EDTA buffer
TBVs	Transmission-Blocking Vaccines
TM-SITE	Threading Model Site
tRNA	Transfer RNA

ABSTRACT

Plasmodium falciparum, the primary causative pathogen of malaria, relies on gametocytogenesis to transmit from humans to mosquitoes. Gametocyte development 1 (*Pfgdv1*) is an upstream activator and epigenetic controller of gametocytogenesis, whereas apetala2 g-box motif (*Pfap2g*) acts as a master transcription factor that induces gametocytogenesis and determine the transmission rate of malaria. As the gametocyte stage is crucial in the transmission of malaria parasites, characterization of these genes and structure prediction of proteins that mediate gametocytogenesis is essential. This study characterized *Pfgdv1* and *Pfap2g* genes by examining single-nucleotide polymorphisms (SNPs), predicting protein structure, and carrying out protein-ligand docking in isolates of *P. falciparum*. Thirty blood samples were collected from patients infected with *P. falciparum*, genomic DNA was extracted, PCR was done to amplify *Pfgdv1* gene and the most polymorphic region of *Pfap2g* gene with subsequent sequencing. Secondary sequences of *Pfgdv1* and *Pfap2g* genes were retrieved from PlasmoDB for comparative analysis. The processing of raw sequence data was done using ChromasPro. Multiple sequence alignment (MSA) was conducted in MEGA using MUSCLE program. Tajima's D test and SLAC were used to examine evolutionary trends and codon sites under selection pressure, respectively. The effect of non-synonymous SNPs (nsSNPs) on the stability of the protein structure was evaluated using STRUM. Prediction of protein structure was done using RaptorX-Property, RaptorX, and I-TASSER, while protein-ligand docking was performed using I-TASSER and COACH. MSA of primary and secondary data established the existence of a synonymous SNP and four (4) nsSNPs in *Pfgdv1* and a synonymous SNP and eleven (11) snSNPs and synonymous SNP in *Pfap2g*. Tajima's D indicated that both *Pfgdv1* and *Pfap2g* exhibit balancing selection, while P217H in *Pfgdv1* and K984T and K21R in *Pfap2g* are under strong positive selection. Thermodynamics analysis indicated that P217H had a destabilizing effect while R398Q and D497E had a stabilizing effect on *Pfgdv1*. The analysis of nsSNPs in *Pfap2g* revealed that all had stabilizing effects on the predicted structure of the protein. The predicted structure of *Pfgdv1* has 599 amino acid residues (1800 nucleotides), molecular weight of 71.964 kDa, and ordered structure (95%) with c-score of -2.95, TM-score of 0.38±13, and RMSD of 15.1±3.5 Å. The predicted model of *Pfap2g* has 2432 amino acid residues (7299 nucleotides), a molecular weight of 284.064kDa, apetala2 domain (AP2/ERF), and 50% disordered structure. Moreover, the predicted model have c-scores of -0.97 and 0.09, TM-scores of 0.59±14 and 0.73±11, and RMSD of 11.7±4.5Å and 9.2±4.6Å for amino acid residues 1-1200 and 1201-2432, respectively. Docking results show that peptide (III), N-octadecane, and ZCT are reliable ligands that bind to *Pfgdv1*, while flavin mononucleotide and zinc ion are ligands that bind to 1-1200 and 1201-2432 protein residues of *Pfap2g*, respectively. As results demonstrate that both *Pfgdv1* and *Pfap2g* are relatively conserved genes in *P. falciparum* isolates, they are potential targets for drugs or vaccines that block the transmission of malaria from humans to mosquitoes.

CHAPTER ONE

INTRODUCTION

1.1 Background

Plasmodium falciparum, the leading causative species of malaria, is a protozoan parasite transmitted by female *Anopheles* mosquitoes when they bite humans to obtain their blood meal. Globally, malaria transmission remains undeterred with epidemiological data showing that malaria affected about 228 million people with approximately 405,000 deaths (WHO, 2019). Kenya is one of the malaria-endemic countries in sub-Saharan Africa with highland areas in the North Rift region, such as Nandi, Uasin Gishu, and Baringo Counties, experiencing epidemic malaria (Kipruto *et al.*, 2017; Noor *et al.*, 2018). Despite great strides made in the use of insecticides, anti-malarial drugs, and effective healthcare services, prevention and control strategies of malaria are still challenging due to drug resistance, inaccessibility to treatment, insecticide resistance, and residual transmission (Arama & Troye-Blomberg, 2014; Delves *et al.*, 2018; Muduli *et al.*, 2018; Sinden *et al.*, 2012; WHO, 2018). Hence, there is a need to develop new strategies that target the transmission of malaria in both epidemic and endemic regions.

Gametocytogenesis is a critical stage in the human-to-mosquito transmission of *P. falciparum*. In commitment to sexual development, about 10% of parasites undergo gametocytogenesis, leading to the formation of gametocytes (Josling & Llinás, 2015). Transmission occurs when a mosquito bites and ingests gametocytes together with the blood meal. In the mosquito midgut, gametocytes grow into macro- and micro-gametes, which fuse to form a zygote (Campbell *et al.*, 2010; Josling & Llinás, 2015). Thus, gametocytogenesis plays a critical role in the host-vector transmission of malaria parasites.

Evidently, gametocyte induction and development is subject to biochemical and genetic factors. It has been demonstrated that cAMP and phorbol ester are notable chemicals that play a role in gametocyte induction and development in an *in vivo* environment (Campino *et al.*, 2016). McRobert *et al.* (2008) demonstrated that xanthurenic acid and

zaprinas are mosquito-derived chemicals that stimulate gametocytogenesis through cGMP-dependent protein kinase. Heparin is more effective in the elimination of asexual stages in erythrocytes when compared to N-acetylglucosamine because it acts rapidly and prevents hemolysis in *in vitro* environment (Miao *et al.*, 2013). The effects of these chemicals have led to the establishment of genetic factors that influence gametocyte development in *P. falciparum*.

Genetic factors determine the induction of asexual or sexual development in *P. falciparum*. In 3D7 clone of *P. falciparum*, PF3D7_0935400, which is a gene located in the right arm of chromosome nine in the sub-telomeric region, codes for gametocyte development protein 1 (*Pfgdv1*) with the function of gametocyte induction and development (Day *et al.*, 1993; Campino *et al.*, 2016). Deletion of this gene results in the inability to form gametocytes while complementation of this gene results in the restoration of gametocyte formation (Campino *et al.*, 2016). A genome-wide analysis detected differential selection of *Pfgdv1* gene in two *P. falciparum* populations with differing transmission intensity (Mobegi *et al.*, 2014). Thus, further characterization of the gene is essential to provide additional information regarding its polymorphism.

Another gene that induces gametocyte formation and development is PF3D7_1222600, which encodes for apatla2 g-box motif (*Pfap2g*) protein (Yuda *et al.*, 2009). Researchers discovered that *Pfap2g* is a DNA-binding protein and a master regulator of gametocyte development first discovered in protozoan parasites (Kafsack *et al.*, 2014). In essence, *Pfap2g* is a transcriptional regulator of gametocyte development in *P. falciparum*. The gametocyte non-producer (GNP) lines, such as F12 and A4, have lost the capacity to form gametocytes due to mutations in *Pfap2g* gene (Kafsack *et al.*, 2014; Josling & Llinás, 2015; Campino *et al.*, 2016). Moreover, genetic characterization of this gene is necessary to understand its role as a master regulator of gametocytogenesis.

Current strategies of malaria control focus on the development of transmission-blocking interventions targeting proteins associated with gametocytes and gametes in humans

(Sinden *et al.*, 2012; Delves *et al.*, 2018). Vaccine development and drug design target different stages of parasites, which are gametocyte stage, pre-erythrocytic stage, and erythrocytic stage (Sinden *et al.*, 2012; Arama & Troye-Blomberg, 2014; Delves *et al.*, 2018). Therefore, understanding the transmission of malaria parasites plays a significant role in identification of new therapeutic targets for control of malaria.

The elucidation of the mechanism of gametocytogenesis forms the basis of developing transmission-blocking vaccines and drugs. Recent studies have established that *Pfgdv1* induces gametocyte development through the mechanism of antagonizing the silencing effect of heterochromatin protein 1 on *Pfap2g* (Eksi *et al.*, 2012; Josling & Llinás, 2015; Campino *et al.*, 2016; Filarsky *et al.*, 2018; Rea *et al.*, 2018). *Pfgdv1* is an upstream activator that regulates the transcription activity of *Pfap2g* via the mechanism of epigenetic control of heterochromatin protein 1 (HP1) (Figure 1.1). Genetic characterizations of these genes and their respective proteins explain their functions and enhance understanding of transmission-blocking targets and gametocyte development in *P. falciparum* isolates.

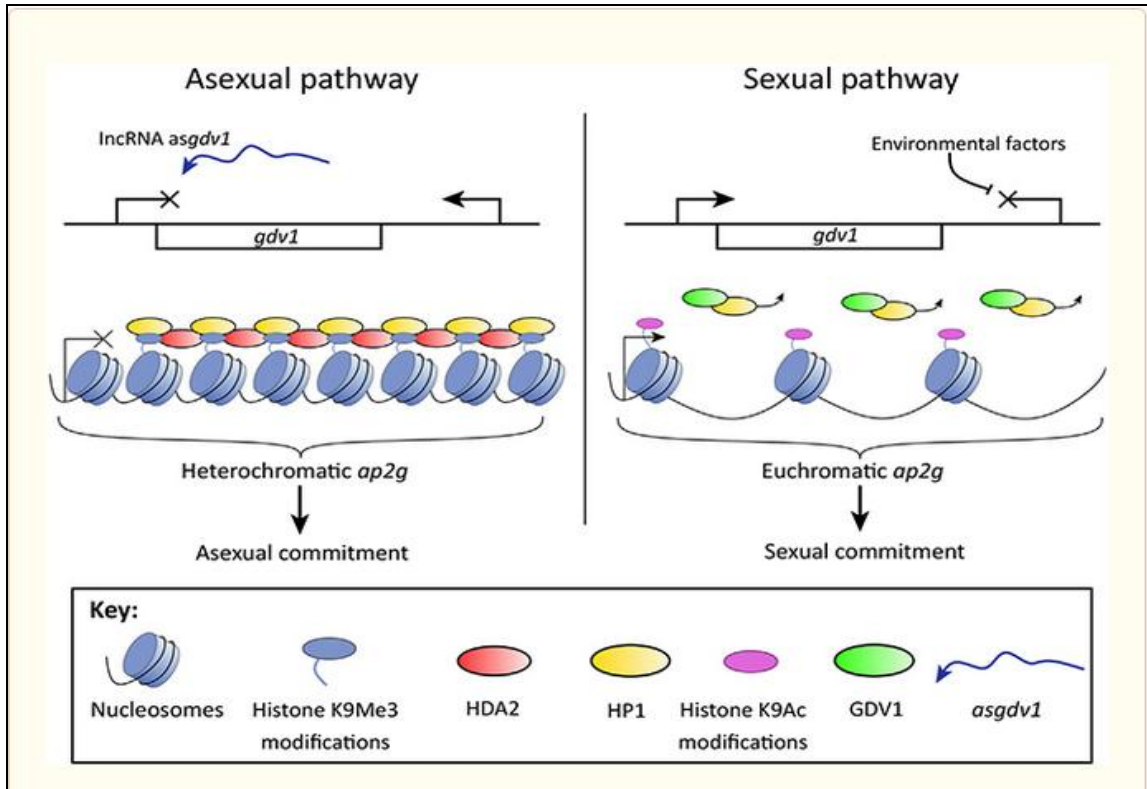


Figure 1.1: Mechanism of gametogenesis as regulated by *Pfgdv1*, heterochromatin protein 1, and *Pfap2g* (Filarsky *et al.*, 2018).

To achieve the objectives of this study, partial primary sequences and complete secondary sequences of *Pfdv1* and *Pfap2g* genes were used. Subsequently, these sequences were utilized in the identification and characterization of single-nucleotide polymorphisms (SNPs) in *Pfgdv1* and *Pfap2g* genes and proteins. Protein structures of these two genes were predicted and their respective potential ligands that could modulate gametocytogenesis were identified.

1.2 Statement of the Problem

Prevention and control strategies of malaria are still ineffective despite great strides made in the use of insecticides and anti-malarial drugs. The malaria report of 2019 shows that malaria is the leading infectious disease because it infects over 200 million people out of which more than 0.4 million of them die yearly (WHO, 2019). Moreover,

increasing trends of anti-malarial resistance pose a significant challenge to the prevention and control of malaria among the population globally. The World Health Organization has recognized the emergence of *Plasmodium* resistance to antimalarial drugs such as artemisinin and has recommended the use of artemisinin-based combination therapies and design new interventions to prevent the transmission of resistant malaria parasites (Muduli *et al.*, 2018). Vaccine development is a current strategy aimed at disrupting the lifecycle of malaria parasite. The development of pre-erythrocytic vaccines has yielded RTS,S (Mosquirix) with the highest efficacy (39%) among vaccines under clinical trials (Olotu *et al.*, 2016; van den Berg *et al.*, 2019). Given that malaria parasites undergo diverse stages of development, they exhibit diverse antigens making them dodge protective immune responses (Arama & Troye-Blomberg, 2014). Overall, prevention and control of malaria remain a challenge despite the use of different efficacious interventions.

1.3 Justification of the Study

Sexual development is critical in the transmission of *P. falciparum* because it leads to the formation of gametes, which enter the mosquito vector, fuse into a zygote, and produce infective sporozoites. Recent studies have established that *Pfgdv1* encoded by PF3D7_0935400 and *Pfap2g* encoded by PF3D7-1222600 are proteins that induce gametocyte development in *P. falciparum* (Eksi *et al.*, 2012; Kafsack *et al.*, 2014; Josling & Llinás, 2015; Campino *et al.*, 2016). *Pfgdv1* is a peri-nuclear protein that stimulates early gametocyte differentiation, while *Pfap2g* is a transcription factor that binds to DNA and stimulates gametocyte development at trophozoite stage. Genetic characterization of these two genes and their respective proteins elucidates their functions and enhances understanding of gametocyte development in *P. falciparum* isolates. In the wake of *Plasmodium* resistance to the current medications, there is a need to identify new drug targets and apply them in the control of malaria.

In the design of transmission-blocking strategies, researchers recommend that the current and future global efforts of eradicating malaria should focus on molecular and

cellular interventions aimed at gametocyte development of *Plasmodium* (Sinden *et al.*, 2012). The current strategies of designing vaccines target different stages of parasites, which are the human-vector transmission, pre-erythrocytic stage, and erythrocytic stage (Arama & Troye-Blomberg, 2014). In this view, gametocyte development falls under the human-vector transmission, and thus, is central in the eradication of malaria transmission. Therefore, there is a need to characterize *Pfgdv1* and *Pfap2g* genes among *P. falciparum* isolates to identify polymorphisms and provide important information required in drug design.

1.4 Research Questions

- i) What are characteristics of single nucleotide polymorphisms that exist in *Pfgdv1* and *Pfap2g* genes of *P. falciparum* isolates?
- ii) What are the predicted protein structures of *Pfgdv1* and *Pfap2g* genes in *P. falciparum* isolates?
- iii) What are the potential protein-ligand interactions of *Pfgdv1* and *Pfap2g* gene products in *P. falciparum* isolates?

1.5 Hypothesis

As conserved genes are appropriate targets for drug discovery and vaccine development, this study hypothesized that non-synonymous mutations driven by evolutionary forces in *Pfgdv1* and *Pfap2g* genes do not have a marked influence on the predicted protein structures *P. falciparum* isolates.

1.6 Objectives

1.6.1 General Objective

- To determine genetic characteristics of *Pfgdv1* and *Pfap2g* genes in *P. falciparum* isolates, predict structure, and identify ligands that dock with their gene products.

1.6.2 Specific Objectives

- To characterize single-nucleotide polymorphisms in *Pfgdv1* and *Pfap2g* genes from *P. falciparum* isolates.
- To predict the structure of *Pfgdv1* and *Pfap2g* gene products in *P. falciparum* isolates.
- To identify potential ligands that dock with *Pfgdv1* and *Pfap2g* gene products in *P. falciparum* isolates.

CHAPTER TWO

LITERATURE REVIEW

2.1 Epidemiology of Malaria

Plasmodium falciparum, a protozoan parasite that causes malaria, is common in sub-Saharan Africa, particularly among vulnerable populations of children aged below five years and pregnant women (Bashir *et al.*, 2019). Other protozoan parasites that cause malaria are *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*, and *Plasmodium knowlesi*. The epidemiological data reveals that malaria is one of the leading causes of death among infectious diseases in tropical and subtropical regions. Epidemiological data show that about 93%, 3.4%, and 2.1% of total cases of malaria occur in Africa, South-East Asia, and Eastern Mediterranean parts of the world, respectively (WHO, 2019). Malaria is prevalent in sub-Saharan Africa because the effective mosquito vector, *Anopheles gambiae*, and the most severe form of malaria, *P. falciparum*, are widespread. Global epidemiological data indicates that malaria affected about 228 million people and caused approximately 405,000 deaths (WHO, 2019). Kenya is one of the malaria-endemic countries in sub-Saharan Africa with highland areas in the North Rift region, such as Nandi, Uasin Gishu, and Baringo Counties, experiencing epidemic malaria (Kipruto *et al.*, 2017; Noor *et al.*, 2018). Although numerous interventions have decreased the prevalence of malaria in endemic regions, the increasing populations coupled with the development of drug-resistant parasites have complicated the control and prevention of malaria.

2.2 Life Cycle of *Plasmodium*

Transmission of malaria is dependent on the ability of the *Plasmodium* species to infect both the mosquito vector and the human host (Figure 2.1). Malaria infection commences when an infected anopheline mosquito bites a human host, obtain a blood meal, and releases sporozoites. The released sporozoites enter into the body and migrate through the blood vessels to the liver where they develop in hepatocytes. After several days, sporozoites mature into merozoites, which exit the liver and invade erythrocytes. In the

erythrocytes, most merozoites undergo a repeated erythrocytic cycle of the asexual development involving erythrocytic invasion, replication, and reinvasion, resulting in the intermittent febrile wave a characteristic symptom of malaria (Josling & Llinás, 2015). Some parasites in asexual phase (less than 10%) differentiate through sexual development and form gametocytes (Josling & Llinás, 2015). Fundamentally, the differentiation of merozoites is an adaptive process that allows malaria parasites to transmit to another human host through the anopheline mosquito bite.

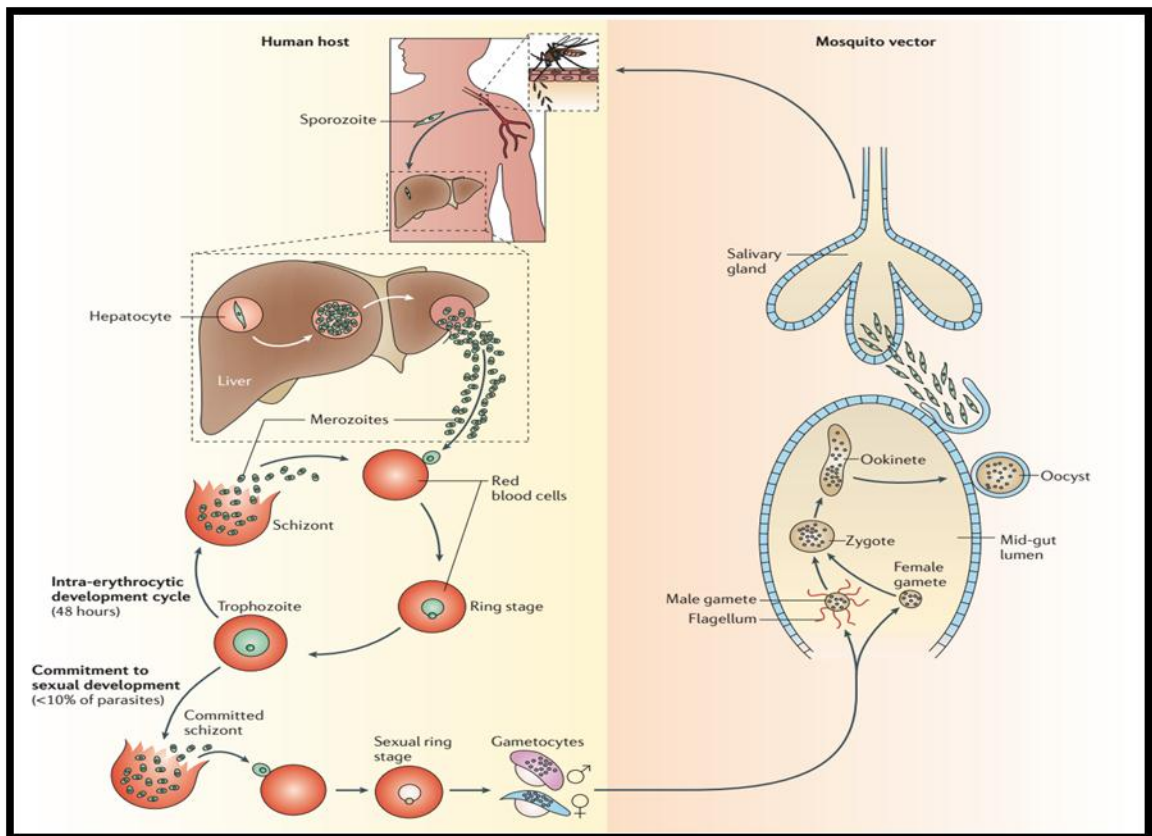


Figure 2.1: The life cycle of *P. falciparum* (Josling & Llinás, 2015)

Once anopheline mosquito bites infected human host, it ingests blood meal together with the gametocytes, which mature into micro- and macro-gametes that subsequently undergo fertilization in the gut of mosquito to form zygote (Josling & Llinás, 2015). The zygote then develops into ookinete, which is an elongated motile zygote that migrates to

the midgut of the anopheline mosquito here it encysts as oocyst (Campbell *et al.*, 2010). When oocyst matures, it ruptures and releases sporozoites that migrate into the salivary glands of the mosquito vector pending infection of the human host through a blood meal bite. Thus, the life cycle indicates that the anopheline mosquito and the human host are critical in the transmission of malaria in the populations.

2.3 Genome Structure of *Plasmodium*

Increased efforts to control and prevent malaria have led to the sequencing of *Plasmodium* genome (Carter *et al.*, 2000). The first sequencing of *P. falciparum* (3D7) using whole chromosome shotgun sequencing strategy revealed that it has 14 chromosomes, which contain 22.8 mega-bases and 5,300 genes (Gardner *et al.*, 2002). The analysis of the genome indicated *P. falciparum* has a gene density of 1/4300 base pairs and biased GC content of 19.4% (Gardner *et al.*, 2002). Subsequent sequencing had enhanced accuracy that led to the discovery of more genomic structure. Next-generation sequencing techniques such as single-molecule real-time sequencing and *de novo* assembly have enhanced genetic characterization of *Plasmodium* species (Vembar *et al.*, 2016; Oyola *et al.*, 2016; Shen *et al.*, 2018). The development of PlasmoDB (<https://plasmodb.org>) allowed integration of experimental and computational data of sequencing that has improved genome assembly and annotation (Bahl *et al.*, 2003).

Owing to improved sequencing techniques, PlasmoDB (2019) shows *P. falciparum* (3D7) has 23.3 megabases of genome and 5777 genes. Other *Plasmodium* species that have been sequenced with their respective genome sizes and total genes (Mb, genes) are *P. berghei* ANKA (18.78, 5254), *P. chabaudi* (18.97, 5364), *P. cynomolgi* strain B (26.18, 5776), *P. falciparum* IT (22.98, 5699), *P. knowlesi* strain H (24.40, 5483), *P. reichenowi* CDC (23.92, 6069), *P. vivax* Sal-1 (27.01, 5626), *P. yoeliiyoelii* 17X (22.76, 6102), *P. gallinaceum* 8AP (16.93, 18) *P. yoeliiyoelii* 17XNL (22.94, 7774), and *P. yoeliiyoelii* YM (22.03, 5833) (PlasmoDB, 2019). Out of these species, *P. falciparum* and *P. vivax* Sal-1 are the only ones with established SNPs in their genomes.

Comparison of the genome sizes indicates that *Plasmodium* species have highly variable genome sizes (16.93-27.01 Mb) and total genes (4951-7774 genes).

Sequencing of *Plasmodium* species has inspired numerous studies aimed at understanding the genomic structure and functions of diverse genes and proteins. Evolutionary studies established that *P. reichenowi* and *P. gaboni*, which are species that infects chimpanzees, are ten times more diverse than those affecting humans (Sundararaman *et al.*, 2016). Development of drugs and vaccines rely on sequences of *Plasmodium* species as molecular targets (Arama & Troye-Blomberg, 2014). The development of vaccines originated from assembled and annotated sequences of *Plasmodium* species (Okie, 2005; Hermsen *et al.*, 2006; Foquet *et al.*, 2014; Ogutu *et al.*, 2009; Gonçaves & Hunziker, 2016). As the study aims to characterize *Pfgdv1* (PF3D7_0935400) and *Pfap2g* (PF3D7_1222600) genes, *P. falciparum* 3D7 is essential because it acts as a reference genome in SNPs analysis, protein structure prediction, and protein-ligand docking. Researchers identified *Pfgdv1* by observing that expression levels of gametocytogenesis early genes of gametocyte-producing lines (3D7.G+) was ten-fold higher than that of gametocyte-deficient lines (3D7.G-) of *Plasmodium* (Eksi *et al.*, 2012). *Pfap2g* was identified by genome-wide analysis, which indicated that F12 and A4 lines of 3D7 have mutations in the DNA-binding domains; hence, making them lose their transcription function (Kafsack *et al.*, 2014; Campino *et al.*, 2016).

2.4 Vaccine Development

The increasing trends of resistance to anti-malarial drugs and advancements in malaria research have led to the search of a vaccine. Fundamentally, vaccine development expands interventions aimed at eradicating malaria as the discovery of novel drugs as continually resulted in the evolution of malaria parasite and development of resistant strains (Lu *et al.*, 2015). The development of vaccine against malaria is a complex process because sporozoites evade immune responses of the body by changing their antigenic properties (Arama & Troye-Blomberg, 2014). For *P. falciparum* to complete

its life cycle, it lives in five varied tissues and undergoes ten morphological changes that affect its antigenic properties (Mackinnon & Marsh, 2010) (Figure 2.2).

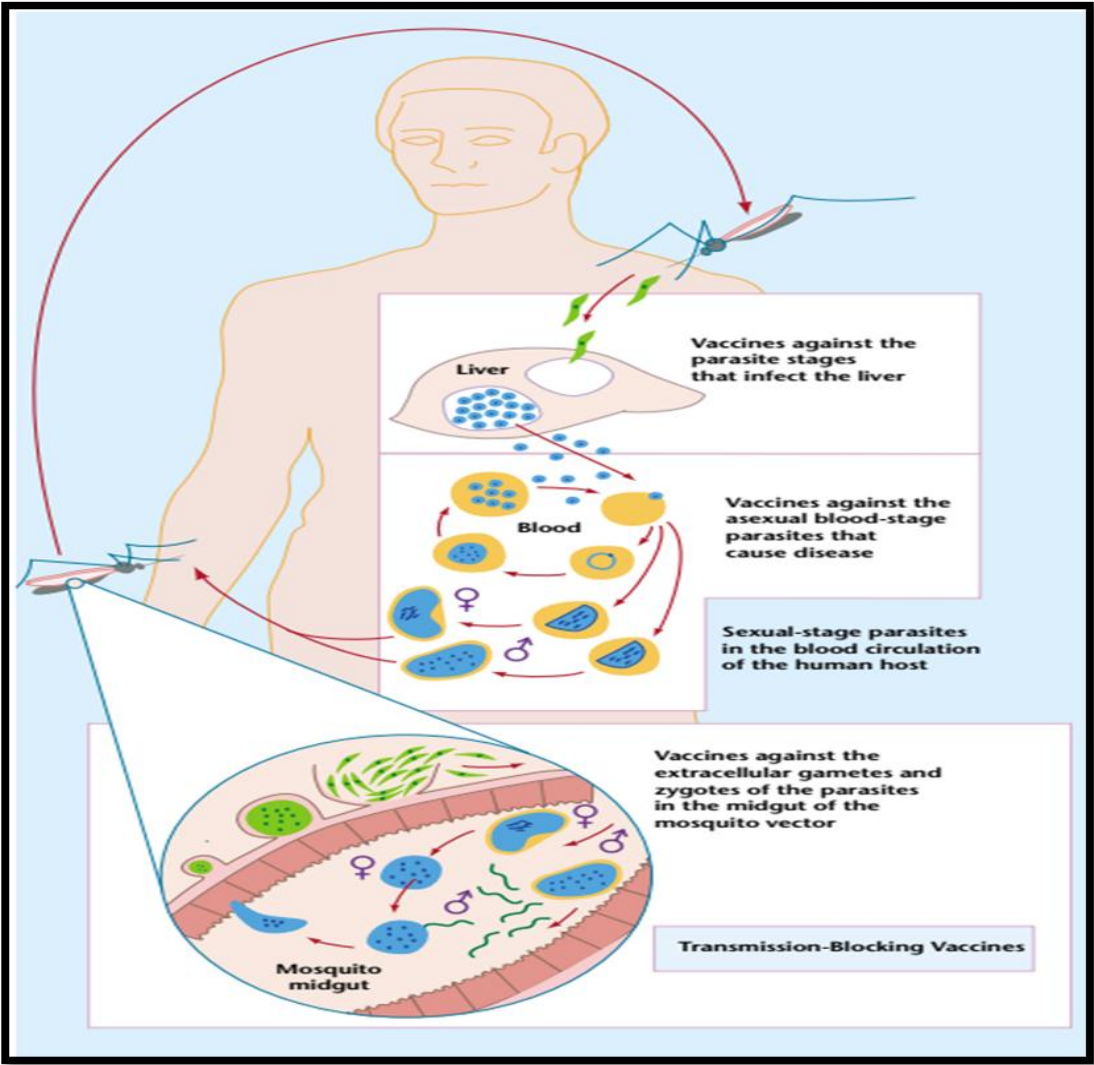


Figure 2.2: Vaccine target areas in the life cycle of the malaria parasite (Carter *et al.*, 2000).

In this view, efforts of vaccine development focus on diverse antigenic properties with a view of ensuring that malaria parasites do not evade the immune responses of the body. Vaccines designed or under design focus on three distinct stages of *Plasmodium*, namely, pre-erythrocytic stage, erythrocytic stage, and sexual stage (Arama & Troye-Blomberg, 2014). These three distinct stages have resulted in the development of pre-erythrocytic vaccines, erythrocytic vaccines, and transmission-blocking vaccines (TBVs) (Figure 2.2).

2.4.1 Pre-Erythrocytic Vaccines

Pre-erythrocytic vaccines target the development of malaria parasites in the hepatocytes. These vaccines target this stage because it is the early stage of development, which is not pathogenic and easy to disrupt. The pre-erythrocytic stage is a robust therapeutic target for both drugs and vaccines because it is the earliest stage of malaria parasites in the human body (March, 2013). In this stage, infected anopheline mosquito bites and injects sporozoites into the liver via the bloodstream. In the liver, sporozoites grow, replicate, and develop into merozoites, which invade erythrocytes and cause febrile waves of malaria. Vaccines that target pre-erythrocytic stage comprise circumsporozoite protein and killed sporozoites (Okie, 2005). Recent advances in the development of pre-erythrocytic vaccines have yielded RTS,S (Mosquirix), which is an approved malaria vaccine made of a portion of circumsporozoite protein combined with hepatitis B virus surface antigen (HBsAg) to form a recombinant protein (Arama & Troye-Blomberg, 2014). As the mechanism of action, RTS,S boosts humoral and cellular immune response by stimulating production of CD4⁺ T-cell against circumsporozoite protein (Foquet *et al.*, 2014). Consequently, the immune response disrupts merozoites in the hepatocytes and thus prevents them from invading erythrocytes. A randomized clinical trial of 447 children between the ages 5-17 months proved that RTS,S three-dose and four dose vaccinations have efficacy of 28% and 36% respectively (Olotu *et al.*, 2016). Thus, RTS,S is effective in supplementing but not replacing standard interventions of treating and controlling malaria.

Further research has led to the discovery of other antigens that have the potential of triggering an immense immune response and boosting immunity against *Plasmodium* parasites. Cell-traversal protein for ookinetes and sporozoites (CelTOS) is a unique protein that is critical for the migration of malaria parasites in mammalian and vector hosts (Kariu *et al.*, 2006). Vaccines that target this protein prevent the migration of sporozoites in the body resulting in the interference of the sporozoite growth and development in hepatocytes. A comparative analysis of infectivity of malaria parasites established that malaria parasites without CelTOS had 200-fold decrease in infectivity in mosquito host and eliminated infectivity in human host (Kariu *et al.*, 2006). Another study performed to determine the effect of CelTOS vaccine on murine models demonstrated that it induces significant protection against sporozoite growth and development in hepatocytes (Bergmann-Leitner *et al.*, 2011).

2.4.2 Erythrocytic Vaccines

Erythrocytic vaccines target the blood stages, which are mainly merozoites invading erythrocytes and causing febrile waves, a major symptom of malaria infection. The design of erythrocytic vaccines aims to elicit immune responses that are anti-invasion and anti-pathogenic (Arama & Troye-Blomberg, 2014). The rationale of the design is that anti-invasion responses hinder merozoites from invading erythrocytes while anti-pathogenic responses prevent merozoites from growing and developing in erythrocytes. Established blood erythrocytic antigens are erythrocyte-binding antigen (EBA-17) (El Sahly *et al.*, 2010) serine repeat antigen 5 (SERA5), apical membrane antigen 1 (AMA 1) (Schussek *et al.*, 2013), merozoite surface protein (MSP) (Esen *et al.*, 2009), and glutamate-rich protein (GLURP) (Hermsen *et al.*, 2007; Esen *et al.*, 2009). These erythrocytic antigens are under clinical trials and the findings are yet to prove the effectiveness of their respective vaccines in preventing growth and development of malaria parasites. Researchers recommend consideration of polymorphism of different antigens in the design of cross-reactive vaccine constructs, which trigger extensive

immune responses that cover wide genetic diversity of malaria parasites (Arama & Troye-Blomberg, 2014).

Far-reaching research on erythrocytic vaccines specific to *P. falciparum* has led to the discovery of novel vaccine candidates. Bustamante *et al.* (2013) demonstrated that reticulocyte binding protein homolog 5 (PfRH5), which is a homolog of reticulocyte binding like proteins (RBLs) is critical in erythrocyte invasion, triggers immune response that generates antibodies against *Plasmodium* parasites expressing these proteins. Fundamentally, the basigin-PfRH5 interaction is feasible in all strains of *P. falciparum*, and thus, a potential candidate of cross-reactive vaccine (Crosnier *et al.*, 2011). Moreover, PfRH5 is an effective erythrocytic vaccine candidate because it induces cross-strain antibodies against *Plasmodium* parasites (Douglas *et al.*, 2011). Further studies also established that rhoptry-associated leucine zipper-like protein 1 (RALP1) as a possible vaccine candidate. RALP1 is essential for erythrocytic survival of *P. falciparum* for it is found in the rhoptries of merozoites and schizonts in the blood sampled in Mali and Thailand (Ito *et al.*, 2013). Therefore, PfRH5 and RALP1 are the current erythrocyte vaccine candidates under investigation.

2.4.3 Transmission-Blocking Vaccines

Transmission-blocking vaccines (TBVs) target human host-mosquito vector transmission of malaria parasites by interfering with surface proteins of sexual stages of malarial development, such as gametocyte, gamete, zygote, and ookinete. In essence, the role of TBVs is to prevent transmission of malaria parasites, and thus, they do not benefit infected individuals but uninfected individuals in a community. The development of TBVs is essential because they are effective in the prevention of malaria transmission in endemic regions (Carter *et al.*, 2000). Surface proteins of gametocytes and gametes include Pfs2400, Pfs48/45, Pfs230, Pfg27, Ps25, Ps28, Ps21, and chitinase are surface proteins of ookinete and zygote (Gonçalves & Hunziker, 2016). In their study to determine immune response to Pfs25 proteins, researchers found out that nasal immunization is effective against oocyst development because it triggers Th1 and Th2

responses (Arakawa *et al.*, 2005). Current studies explore the mechanisms of gametocytogenesis with a view to identify key genes that regulate it and target them as vaccine candidates (Filarsky *et al.*, 2018; Duffy *et al.*, 2018; Usui *et al.*, 2019). Therefore, optimization of these target areas through protein recombination and exploration of other potential targets is essential to improve TBVs coverage of diverse strains of malaria parasites.

2.5 SNPs and Targets for Vaccine Development

Genetic diversity is a critical factor that requires consideration in the development of vaccines. Arama and Troye-Blomberg (2014) explain that a major challenge with the development of vaccines is the diversity of the target antigens to elicit specific and effective immune responses against malaria. Ajibaye *et al.* (2020) add that genetic polymorphism reduces the efficacy of vaccines by altering antigenic epitopes. According to Takala and Plowe (2009), the exploration of target genes is essential to enhance the understanding of genetic diversity for rational vaccine design. The analysis of genes aims to identify conserved genes with constant protein structure and antigenic properties for the immune response to be specific and effective against malaria (Conway, 2015). The characterization of SNPs would determine if a certain target gene has conserved sequences appropriate for vaccine design. Current researchers employ a combination of molecular, epidemiological, evolutionary, population genetic, immunological, and structural prediction tools in identifying polymorphism in target genes. In this case, the study aimed to characterize SNPs, predict protein structures, and identify ligands of *Pfgdv1* and *Pfap2g*.

2.6 Gametocytogenesis

Gametocytogenesis is an integral stage in the life cycle of *Plasmodium* because it determines the formation of gametes through the sexual development and influences the proportion of merozoites that undergo the asexual development. In the erythrocytic cycle, *Plasmodium* parasite encounters adaptive trade-off between the sexual development to transmit to new human hosts and the asexual development to proliferate within the host. Primarily, numerous factors influence sexual and asexual development of merozoites in the human host. Biochemical pathways that lead to the production of cAMP and phorbol esters mediate gametocytogenesis in *P. falciparum* (Campino *et al.*, 2016). In a study on the role of cGMP-dependent protein kinase (PKG), researchers established that mosquito-derived chemicals, namely, xanthurenic acid and zaprinast, stimulate gametocytogenesis (McRobert *et al.*, 2008). In a bid to develop pure cultures of gametocytes, a study identified heparin and N-acetylglucosamine as chemicals that effectively stimulate gametocytogenesis and eliminate asexual development in an *in vitro* environment (Miao *et al.*, 2013). Thus, the literature shows that both *in vivo* and *in vitro* chemical can stimulate or inhibit sexual and asexual development of *P. falciparum*.

Efforts to elucidate mechanisms of stimulation or inhibition of gametocytogenesis have led to the establishment of genetic factors that mediate gametocytogenesis. Genomic analysis has revealed numerous genes involved in the regulating of gametocytogenesis and sex-specific gametocyte development (Josling & Llinás, 2015). Genes that have a significant influence on gametocytogenesis are located on chromosomes nine and twelve. PF3D7_0935400, a gene that codes for gametocyte development protein 1 (*Pfgdv1*), has been implicated in gametogenesis because its deletion results in the loss of this function (Campino *et al.*, 2016). Moreover, PF3D7-1222600, a gene that codes for transcriptional factor (*Pfap2g*), is a master switch that regulates gametocytogenesis epigenetically. Evidently, *Pfgdv1* and *Pfap2g* are genes that stimulate gametocytogenesis in *P. falciparum* (Eksi *et al.*, 2012; Kafsack *et al.*, 2014; Josling &

Llinás, 2015; Campino *et al.*, 2016). Thus, gene products of these genes stimulate gametocytogenesis in *P. falciparum*.

2.7 *Plasmodium falciparum* Gametocyte Development 1

In 3D7 clone of *P. falciparum*, *Pfgdv1* is a gene involved in gametocytogenesis. The location of this gene is the right arm of chromosome nine in the sub-telomeric region (Campino *et al.*, 2016). As this gene codes for *Pfgdv 1* in *P. falciparum*, its deletion results in the inability to form gametocytes while its complementation results in the restoration of gametocyte formation. Comparative analysis of the phenotypic functions of this gene shows that it stimulates gametocytogenesis in *P. falciparum*. Researchers have designated *Pfgdv1* as a peri-nuclear protein that stimulates early sexual development during gametocytogenesis (Eksi *et al.*, 2012). Transcriptional analysis revealed that clone lines without *Pfgdv1* had greater than ten-fold down-regulation of the expression levels in a set of early genes for gametocytogenesis (*Pfge*). The ten-fold decline in expression levels of *Pfge* indicates that *Pfgdv1* has a significant role in the stimulation of gametocytogenesis. In elucidating the mechanism, recent studies established that *Pfgdv1* antagonizes the effect of heterochromatin protein I in silencing the expression of genes, including *Pfap2g* and *Pfge*, resulting in their derepression and gametocytogenesis (Filarsky *et al.*, 2018; Rea *et al.*, 2018). Successive accumulation of *Pfge* during erythrocytic replication of *P. falciparum* amplifies gametogenesis, and thus, promotes transmission of malaria.

2.8 *Plasmodium falciparum* Apetala2-G

Pfap2g (PF3D7_1222600) is another gene of interest involved in gametocytogenesis in *P. falciparum*. This gene encodes for *Pfap2g* protein, which is a DNA-binding protein involved in gametogenesis. Furthermore, researchers describe *Pfap2g* as a DNA-binding protein and a master regulator of gametocyte development first discovered in protozoan parasites (Kafsack *et al.*, 2014). *Pfap2g* is a transcriptional factor that stimulates gametocytogenesis by binding to specific sequences located in the 5'upstream location and triggers transcription process leading to early gametocyte differentiation (Campino

et al., 2016; Bechti & Waters, 2017; Filarsky *et al.*, 2018; Rea *et al.*, 2018). In describing *Pfap2g*, researchers found out that it exhibits genetic diversity among different isolates obtained across the globe, and its expression levels showed a strong correlation with the rate of gametogenesis in *P. falciparum* (Kafsack *et al.*, 2014).

CHAPTER THREE

METHODOLOGY

3.1 Research Design

The study comprised of fieldwork (case-only design), laboratory research (experimental design), and bioinformatics analysis (*in-silico* design). The study employed case-only research design in field investigations since blood samples for the study were collected from individuals with malaria. The case-only design is a relevant approach because it does not require controls, and thus, appropriate in determining polymorphism amongst organisms in their native environment (Hassanzadeh *et al.*, 2012). In the laboratory research, genomic DNA was isolated from blood samples, followed by amplifications of target genes (*Pfgdv1* and *Pfap2g*) of *P. falciparum*, and sequencing of amplicons.

3.2 Ethical Consideration

Ethical clearance was obtained from the Institutional Research and Ethics Committee (IREC) of Moi University College of Health Sciences (MU/CHS) and Moi Teaching & Referral Hospital (MTRH). The ethical approval letter (IREC/2017/41) is in Appendix I. Collection of blood samples was done by qualified technicians and technologists. To comply with ethical requirements, the study informed the patients regarding the objective of the study, allowed voluntary participation, and kept information confidential and private. During sample collection, the study administered written informed consent form (Appendix II) to patients requesting them to donate their blood samples voluntarily without coercion and enticement. The shared data were coded to guarantee confidentiality and privacy, and the principal investigator only had access to sensitive information of patients.

3.3 Study Sites

The collection of malaria blood samples was done in Baringo County Referral Hospital (0.48N, 35.74E), Uasin Gishu County Hospital (0.51N, 35.27E) and Kapsabet County Referral Hospital (0.20N, 35.10E) (Figure 3.1). The laboratory work was carried out at the Biochemistry Laboratory, and Molecular Biology and Biotechnology Laboratory of PAUSTI in JKUAT. Amplicons were shipped to MacroGen Europe at the Netherlands for sequencing using the Sanger technique.

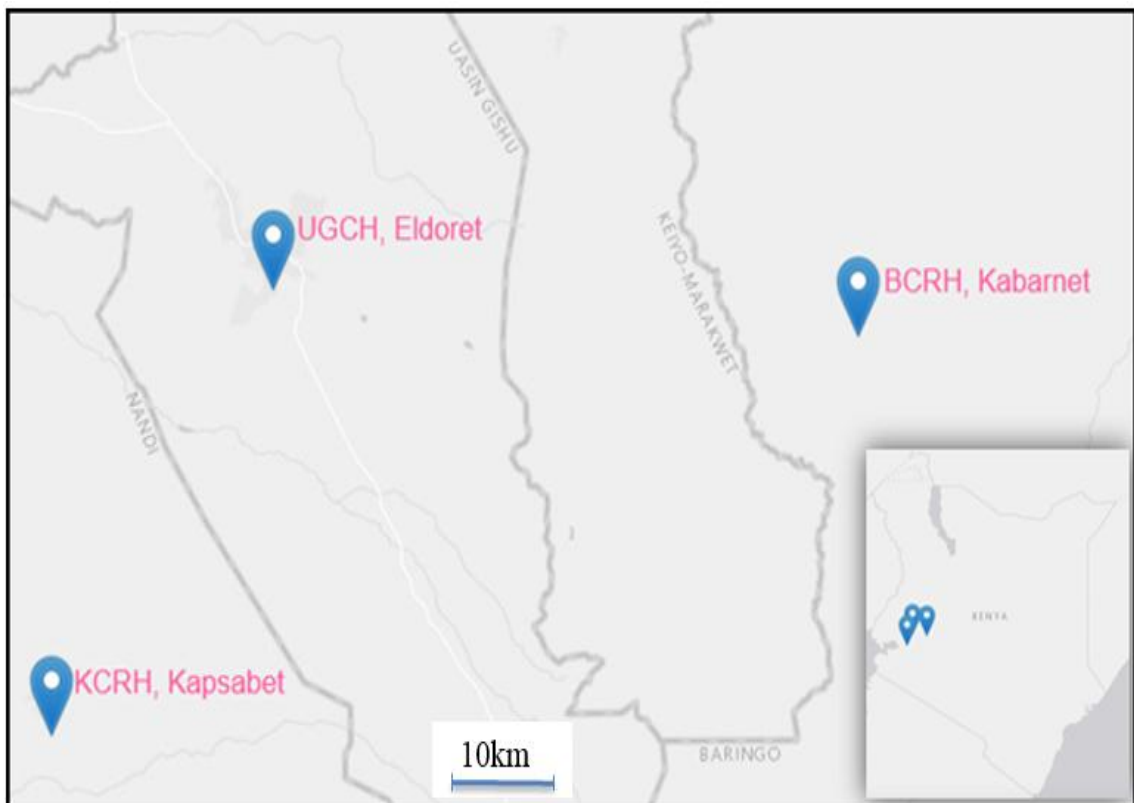


Figure 3.1: Kenyan map showing sampling sites: Kapsabet County Referral Hospital (KCRH), Uasin Gishu County Hospital (UGCH), and Baringo County Referral Hospital (BCRH) (National Geographic Society, 2019).

3.4 Sampling of Participants

The study used the stratified purposeful sampling method to select a representative number of patients with malaria and obtain blood samples from them. As the North Rift comprises an epidemic region of malaria due to human interactions, seasonal rains, warm temperatures, and favorable ecological conditions, the study stratified the sample of 30 patients into three sampling sites, namely, Baringo County, Uasin Gishu County, and Nandi County. The study utilized the following formula in calculating the sample size of target patients with malaria who presented at respective hospitals (Cochran, 2007).

$$\text{Sample size (n)} = \frac{S}{1 + \left(\frac{S-1}{N}\right)}$$

$$\text{Where: } S = \frac{Z^2 PQ}{ME^2} = \frac{1.96^2 \times 0.4 \times 0.6}{0.1^2} = 92.2$$

N = target population of 42 patients in two weeks

Z = Confidence level of Z-score at 95% (1.96)

P = proportion of patients in the age group 20-45 years (0.4)

Q = Proportion of patients not in the target age group (0.6)

ME = Margin of error of 10%

$$\text{Therefore, the sample size, } n = \frac{92.2}{1 + \left(\frac{92.2}{42}\right)} = \frac{92.2 \times 42}{42 + 92.2} = \frac{3872.4}{133.2} = 29.1 \text{ (at least 30)}$$

Based on stratification, the study used selected 10 patients with malaria from each of the three strata. As the study aimed to determine polymorphism among different isolates, blood samples were obtained from patients with malaria who attended Baringo County Referral Hospital, Uasin Gishu County Hospital, and Kapsabet County Referral Hospital. The inclusion criteria of the participants were patients diagnosed with malaria aged between 20 and 45 years. The study excluded pregnant women, people with mental illnesses, and prisoners because they are members of the vulnerable populations.

3.5 Sample Collection

The sampling of participants commenced once qualified medical laboratory technicians and technologists of respective hospitals confirmed the positive diagnosis of malaria among patients. In this view, the researcher recruited medical laboratory technicians and technologists who aided in the diagnosis and determination of parasitaemia using the conventional light microscopy. Moreover, rapid diagnostic kits (Paracheck) were used to confirm positive diagnosis of malaria (Orchid Biomedical Systems, Goa, India) (Proux *et al.*, 2001). The study employed microscopy standards in detection and identification of *Plasmodium* parasites (Giemsa stained thin blood film), as well as quantification (Giemsa stained thick blood film) using assumed 8000 white blood cells per microliter (Adu-Gyasi *et al.*, 2015). Thick smear of blood was prepared and left to dry in air, stained with Giemsa stain for 15 minutes, and rinsed to remove excess stain. Subsequently, examination of the thick smear using microscope and counting the number of parasites in every 200 white blood cells was performed and used in determination of paracetaemia.

After the administration of written informed consent, qualified technicians obtained malaria blood from sampled patients. The blood specimens (2-4 ml) were collected through venipuncture, stored in EDTA vacutainers (Becton Dickinson, Franklin Lakes, NJ), preserved at -4 °C during transport, and stored at -20 °C awaiting DNA extraction. Moreover, duplicate samples were preserved as dried blood spots (DBS) on 903 Whatman card (GE Healthcare, Cardiff, UK), put in zip-lock bags, kept desiccated and transported at ambient temperature, and eventually stored at -20 °C in the laboratory awaiting analysis (Grüner *et al.*, 2015). Appendix C provides details of malaria blood samples collected.

Descriptive analysis was employed to describe blood samples collected. Frequencies were used to examine the distribution of patients based on study sites, gender, and age groups. Measures of dispersion and central tendency were used to describe blood

volume collected and parasitaemia. In this view, tables and graphs were used to present descriptive statistics of samples.

3.6 Primer Design

Primers targeting the two genes (PF3D7_0935400 and PF3D7_1222600) were designed using Primer3Plus (Untergasser *et al.*, 2007). The primer sizes were set between 18 bp and 22 bp with optimum sizes of 20 bp. The melting temperatures of primers were set between 50 °C and 60 °C with optimum temperatures of 55 °C. Since *Plasmodium* genome is AT-rich, GC content of the primers were set at an optimum value of 45% and ranged from 36% to 50%. Suitability of the designed primers was checked with Sequence Manipulation Suite (Stothard, 2000). Based on statistics, primers without hairpin formation, self-annealing, single-base runs, and dinucleotide-base runs, but with substantial GC content, appropriate melting temperatures, and GC clamp were selected. The control primers for detection of *P. falciparum* DNA targeted a highly conserved 18S rRNA gene (M19173) (Mangold *et al.*, 2005). Summary of characteristics of primers used in the study are available on Table 4.3

3.7 Extraction of DNA

DNA was extracted from the 30 malaria blood samples using QIAamp DNA Mini Kit® (Cat No.51304) by following manufacturer's protocol (Qiagen, 2016). Malaria blood samples in EDTA vacutainer tubes that have been stored at freezing conditions were thawed and mixed before pipetting 200 µl of each sample into 1.5 ml Eppendorf tubes. Subsequently, 20 µl of Qiagen protease (Proteinase K) was added into respective Eppendorf tubes and mixed thoroughly by vortexing. To lyse the cells and release the genetic material, 200 µl of lysis buffer (AL) was added into mixture and vortexed to form a homogenous solution and guarantee efficient lysis. The mixtures in Eppendorf tubes were incubated for 15 minutes in a water bath set at 56 °C after which they were centrifuged using MIKRO 220 Model (Hettich, Westphalia, Germany) for 1 minute at 6,000×g to settle droplets in the lid. To precipitate genetic material, 200 µl of absolute ethanol (96-100%) was added to the mixture, vortexed for 15 seconds, and centrifuged

for 1 minute at 6,000×g settle droplets inside the tube. The mixture was then added carefully into spin column in 2ml collecting tube and centrifuged at 6,000×g for 1 minute. The filtrate collected was discarded with the collecting tube and replaced with a clean one. The precipitated DNA in the spin column was purified twice by washing with buffers AW1 and AW2. Specifically, 500 µl of buffer AW1 was added into the spin column and centrifuged at 20,000 g for 1 minute to clear proteins and the filtrate discarded. Next, 500 µl of buffer AW2 was added into the column and centrifuged at 6,000×g for three minutes to clear salts and the filtrate discarded.

To prevent contamination by ethanol in buffer AW2, the spin column was placed in a clean spin column and centrifuged for 1 minute at 20,000×g. The remaining filtrate was discarded and the spin column was placed in Eppendorf tube (1.5 ml) for elution of DNA. To avert the effects of EDTA on amplification reactions, the spin column was incubated with 200 µl of nuclease-free water for 10 minutes to increase DNA yields at room temperature (15-25 °C). Eventually, the DNA was collected by centrifuging incubated spin column for 1 minute at 6,000×g. The quality and quantity of the extracted DNA were checked using Nanodrop spectrophotometer (PCRmax Lambda) (Desjardins & Conklin, 2010) and 1% agarose gel electrophoresis. The extracted DNA samples were stored at -20°C to preserve them as they awaited amplification.

3.8 Target Amplification

The sections of target genes (PF3D7_0935400 and PF3D7_1222600) were amplified using the designed primers and 2X master mix (Thermo Scientific Phusion High-Fidelity PCR Master Mix with GC Buffer) (Thermo Fisher Scientific, 2018). The presence of *Plasmodium* DNA was detected using consensus primers (Mangold *et al.*, 2005), which target 18S rRNA (PL1473F18 [5'-TAACGAACGAGATCTTAA-3' and PL1679R18

[5'-GTTCTCTAAGAAGCTTT-3']). The final volume of reaction (20 µl) constituted of 10 µl of the master mix, 1 µl of forward primer (10µM), 1 µl of reverse primer (10 µM), 6 µl of nuclease-free water, and 2 µl genomic DNA (Mean = 30.27 ng/µl). PCR

conditions for detecting *Plasmodium* DNA were initial denaturation of 98 °C for 30 seconds, denaturation of 98°C for 10 seconds, annealing of 54 °C for 30 seconds, extension of 72 °C for 30 seconds, and final extension of 72 °C for 10 minutes. Following confirmation of the presence *Plasmodium* DNA, target genes were subsequently amplified.

The first fragment of PF3D7_0935400 (1306bp) was amplified with forward primer (PF400_50F [5'-GTAGGCGTCGAAATAGTGCT-3']) and reverse primer (PF400_1355R [5'-TCAGGATGTGTTATGGTATC-3']). To obtain adequate amplicons for qualitative and quantitative analysis, as well as sequencing, 50 µl was the final volume of each reaction. The final volume comprised 25 µl of the master mix, 2.5 µl of forward primer (10 µM), 2.5 µl of reverse primer (10µM), 15 µl of nuclease-free water, and 5 µl genomic DNA (M = 30.27 ng/µl). The PCR conditions commenced with an initial denaturation at 98 °C for 30 seconds, then 35 thermal cycles of denaturation at 98 °C for 10 seconds, annealing at 54.5 °C for 30 seconds, and initial extension at 72 °C for 1 minute, and eventually a final extension at 72 °C for 10 minutes. The second fragment of PF3D7_0935400 (650 bp) was amplified with forward primer (PF400_1081F [5'-GGTATTCCTGTTGTTATGAG-3']) and reverse primer (PF400_1731R [5'-AGAAGATGATGAATGCTGACG-3']). Similar thermal cycling conditions with the annealing temperature of 57.5 °C for 30 seconds were used in amplification of the fragment. A fragment of PF3D7_1222600 was amplified with forward primer (PF600_5662F [5'-GTGTACGGGTAATAAATAAAG-3']) and reverse primer (PF600_6002R [5'-TCGTTGCTGTTATTGTTG-3']) with the same PCR conditions with annealing temperature of 55 °C. Nanodrop-spectrophometer and 2% agarose gel electrophoresis were used to determine the quality and quantity of amplicons. Moreover, the amplicons were stored at -20 °C to preserve them pending purification and shipment for sequencing.

3.9 Purification of Amplicons

PCR products were purified using QIAquick PCR Purification Kit according to manufacturer's protocol (Qiagen, 2015). As the volume of PCR product was 50 μ l, the first washing procedure entailed the addition of 250 μ l of buffer PB (Guanidine hydrochloride, 30% (v/v) isopropanol, and ethanol) and mixing them thoroughly in 1.5 ml Eppendorf tube for effective binding of DNA to the spin-column matrix. The mixture was put into QIAquick spin column setup to bind DNA, and then centrifuged for 1 minute at 17,900 g and the eluent discarded. The second washing procedure comprised addition of 750 μ l of buffer PE (Tris-HCL in 80% ethanol) and centrifuged for 1 minute and the filtrate discarded. To ensure thorough washing, the same procedure in the second wash was repeated and spin-column centrifuged for additional minute to eliminate traces of buffer. DNA in the column was added 35 μ l of nuclease-free water and left to stand for 5 minutes before elution. The purity of the DNA washed was checked using 2% agarose gel electrophoresis. Purified PCR products were put in 96-well plate at 30 μ l volumes and shipped to Macrogen Europe, Amsterdam the Netherlands for Sanger sequencing.

3.9 Retrieval and Processing of Sequences

Secondary sequences were retrieved from PlasmoDB, which has 218 isolates of *P. falciparum* (Aurrecochea *et al.*, 2009; PlasmoDB, 2019). Moreover, in the analysis of data in the database, the search strategy entailed selection of SNPs based on gene identifications and selecting those with over 85% base call, 80% read frequency, and 2% minor allele frequency.

Raw sequence data in the form of chromatograms (ab1 files) were edited and contiguous sequences of reverse and forward sequences, as well as overlapping sequences, were created using ChromasPro (Technelysium Pty Ltd, 2018). A new project was created in ChromasPro, chromatogram files uploaded into the project. Low-quality regions on both ends of chromatograms were trimmed in all sequences. Reference sequence of 3D7 strain was uploaded into ChromasPro to allow it perform contiguous assembly.

Contiguous sequences of forward sequence and reverse-complement sequences for overlapping sequences were created through the sequence assembly. Visual inspection of chromatograms was done to resolve ambiguities in contiguous sequences based on base-call quality scores. Contiguous sequences created were saved in multifasta format for further bioinformatics analysis.

3.10 Detection of Synonymous and Non-synonymous SNPs

Multiple sequence alignment (MSA) was done to detect single nucleotide polymorphisms (SNPs). Nucleotide sequences of primary and secondary data were aligned using MUSCLE in MEGA X, and then visually inspected to identify SNPs (Edgar, 2004; Kumar *et al.*, 2018). Furthermore, the aligned sequences were translated into amino acid sequences using the standard genetic code in MEGA X (Kumar *et al.*, 2018) and visualized in Jalview (Jalview, 2018) to differentiate between synonymous and non-synonymous SNPs, along with conservative and non-conservative mutations. The findings of SNPs were tabulated to indicate CDS positions, major alleles, minor alleles, phenotypes, and effects on amino acids.

3.11 Selection Analysis

In selection analysis, Tajima's D in MEGA X was used to determine if these SNPs exhibit neutrality in their occurrence (Tajima, 1989; Kumar *et al.*, 2018), while SLAC (<http://datamonkey.org/slac>) (Pond & Frost, 2005) was used to calculate the rates of non-synonymous changes (dN) and synonymous changes (dS). Both the primary data and the secondary data of *Pfdv1* and *Pfap2-g* genes were used in the selection analysis.

3.12 Protein Structure Prediction

The primary structure of protein sequences was analyzed using Protparam, which is a web-based program at ExPASy that predicts physical and chemical properties (<https://web.expasy.org/cgi-bin/protparam/protparam>) (Gasteiger *et al.*, 2003). The program predicts molecular weight, theoretical isoelectric point, amino acid

composition, grand average of hydropathicity (GRAVY), estimated half-life, extinction coefficient, instability index, and aliphatic index.

RaptorX-Property (<http://raptorx.uchicago.edu/StructurePropertyPred/predict/>) was used to predict the secondary structure of protein sequences (Wang *et al.*, 2016). The webserver predicts 3-state secondary structure (3SS), 8-state secondary structure (8SS), solvent accessibility (ACC), and disordered regions (DISO).

InterPro version 71 was used to predict the location of domains of primary and secondary protein sequences of *Pfdv1* and *Pfap2g* genes (European Bioinformatics Institute, 2018). InterPro searches integrated databases using InterProScan and offers predictive information such as domains, functional sites, and families (Mitchell *et al.*, 2018). Subsequently, the study used SWISS-MODEL (<https://swissmodel.expasy.org/>) to perform homology modeling to predict 3D protein structure (Swiss Institute of Bioinformatics, 2018; Waterhouse *et al.*, 2018). However, due to low-homology between the target and template sequences, it was not feasible to use SWISS-MODEL in modeling the proteins. Effective prediction of protein structure requires the sequence-template identity of more than 30%, which is a threshold for homology modeling (Xiang, 2006; Peng & Xu, 2011; Sensoy *et al.*, 2017). Evidently, the identity of both amino acid sequences was less than 25%, and thus, not appropriate in modeling tertiary structure of proteins.

To overcome the challenge of low-homology, protein threading was used to predict the tertiary structure of amino acid sequences. Specifically, RaptorX (<http://raptorx.uchicago.edu/StructurePrediction/predict/>) and Iterative Threading Assembly Refinement (I-TASSER), which are server-based software (<https://zhanglab.ccmb.med.umich.edu/I-TASSER/>), were used to model the structures of target amino acid sequences. RaptorX has the capacity to predict both secondary and tertiary structures of proteins using protein-threading approach and advanced probabilistic alignments (Peng & Xu, 2011; Källberg *et al.*, 2012). As the leading protein

structure prediction method in community-wide experiments as evidenced in Critical Assessment of Structure Prediction (CASP) of 2012 (CASP10), 2014 (CASP11), and 2016 (CASP12), I-TASSER is an appropriate method (Yang & Zhang, 2015a, 2015b; Yang *et al.*, 2016). The three major steps of I-TASSER are multiple identifications of structural templates, iterative assembly of structures, and annotation of functions based on structures (Zhang, 2009; Roy *et al.*, 2012; Yang & Zhang, 2015a, 2015b;). I-TASSER uses the first 10 leading templates generated by protein threading and provides the five leading models (Appendices V and VI). The predicted 3-dimensional models of proteins were visualized using Jmol and relevant regions highlighted using console scripts. The global and local accuracy estimations of the model were validated using b-factor profile (BFP) and residue-specific quality (ReSQ) (Yang *et al.*, 2016).

In the analysis of the effect of SNPs on the predicted protein structure, the study employed STRUM. From the perspective of thermodynamics, STRUM determines the stability of a protein using Gibbs-free-energy (G). STRUM first determines the difference between Gibbs-free-energy of folded (G_f) and unfolded (G_u) ($\Delta G = G_u - G_f$), and then the difference between ΔG of wild type and mutant protein structure ($\Delta\Delta G = \Delta G_m - \Delta G_w$) (Quan *et al.*, 2016). The rationale for STRUM is that SNPs affects protein stability by reducing free energy difference between fold and unfold conformations of protein, and thus, $\Delta\Delta G$ less than zero implies that the mutation destabilizes protein structure.

3.13 Protein-Ligand Docking

Protein-ligand docking was performed using the consensus approach called COACH, which is a server-based algorithm. COACH employs both threading models (TM-SITE) and sequence alignments (S-SITE) in predicting ligand-binding sites (LBSs) of proteins (Yang *et al.*, 2013b). TM-SITE and S-SITE derive LBSs from BioLip, which is a highly curated database of established protein-ligand interactions obtained from Protein Data Bank (PDB) (Yang *et al.*, 2013a, 2013b; Wu *et al.*, 2018). Moreover, the study utilized COFACTOR in identifying protein functions based on LBSs, gene ontology (GO), and

enzyme commission (EC) (Roy *et al.*, 2012; Zhang *et al.*, 2017). Therefore, results generated by COACH and COFACTOR (Appendices G and H) provide comprehensive information regarding ligands and binding sites for robust protein-ligand docking.

CHAPTER FOUR
RESULTS

4.1 Description of Samples

The study analyzed data from 30 samples of blood obtained from malaria patients (Females =18, Males = 12), 10 samples from Baringo County Referral Hospital (B1-10), 10 samples from Uasin Gishu County Hospital (U1-10), and 10 samples from Kapsabet County Referral Hospital (N1-10) (Table 4.1). The raw data of samples collected are available in Appendix III.

Table 4.1: Characteristics of malaria patients and samples collected

Statistics	Age	Blood volume drawn (ml)	Parasitaemia (p/μl)
Mean	30.47	2.35	29060.67
Standard Deviation	5.11	0.59	22920.31
Range	19.00	2.50	74960.00
Minimum	21.00	1.50	5040.00
Maximum	40.00	4.00	80000.00
Count (N)	30.00	30.00	30.00
Groups	Frequency	Percent (%)	M±SD (p/μl)
Females	18	60	26434±19020.28
Males	12	40	33000±27317.04
20-25 Years	6	20	37846±26433.69
26-30 Years	8	26.7	20542±9792.89
31-35 Years	10	33.3	23960±19465.42
36-40 Years	6	20	40133±33527.06
Baringo County	10	33.3	39164±21733.83
Uasin Gishu County	10	33.3	20984±19020.28
Nandi County	10	33.3	27034±25848.98

The ages of patients ranged from 21 to 40 years (M = 30.47, SD= 5.11) with the highest proportion of patients (33.3%) in the age group of 31-35 years, followed by 26.7% in age group of 26-30 years and 20% in both age groups of 20-25 years and 36-40 years

(Table 2). Blood volumes withdrawn from patients fluctuated from 1.5 ml to 4 ml ($M = 2.35$, $SD = 0.59$) and parasitaemia varied from 5040 to 80,000 parasites/ μ l ($M = 29,060.67$, $SD = 22,920.31$) (Table 4.1). Differences in the levels of parasitaemia were statistically insignificant by gender, age groups, and study sites ($p > 0.05$).

Age group comparisons of parasitaemia exhibited positive skew in the distribution with the most variation occurring among patients in the group of 36-40 years followed by those in 20-25 years (Figure 4.1). In contrast, patients in the age group of 26-30 years had the lowest variation in the distribution of parasitaemia, while those in the age group of 31-35 years had a moderate variation (Figure 4.1).

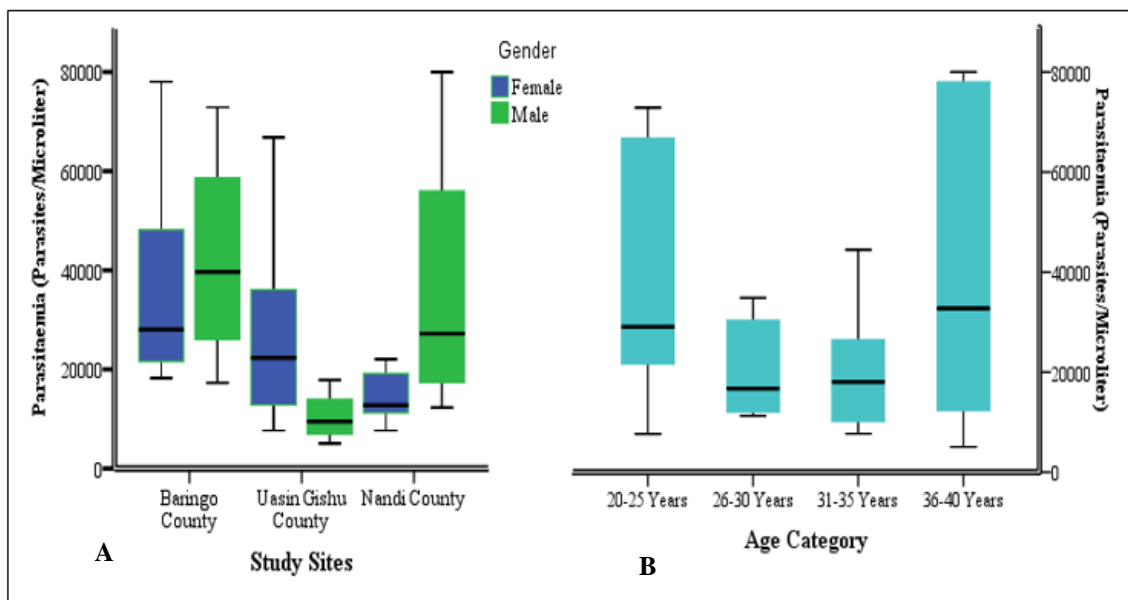


Figure 4.1: The distribution of parasitaemia levels of 30 patients according to study sites and gender (A) and age groups (B)

Comparisons of parasitaemia based on gender and study sites depicted disparities in distributions (Figure 4.1). In samples from Baringo County, the distribution of parasitaemia exhibited positive skew with similar range of variation in both genders, but males had a higher median than females. A considerable difference exists in the distribution of parasitaemia in samples from Uasin Gishu because females had a higher median and variation than males. Contrastingly, in samples from Nandi County, males

had a lower median and variation than females. Overall, the distributions of parasitaemia displayed positive skew with the highest median in Baringo followed by Nandi County and Uasin Gishu.

4.2 Quantity and Quality Extracted Genomic DNA

The quantity and quality of DNA samples (n = 30) extracted from malaria blood samples were recorded. The quantity of extracted DNA samples ranged from 8.206 ng/μl to 61.73 ng/μl with a median of 29.260 ng/μl (M = 30.256, SD = 13.084) (Table 4.2). The analysis of the purity of DNA (260nm/280nm) indicates that the ratio ranged from 1.714 to 2.176 with a median of 1.937 and mode of 2.000 (M = 1.957, SD = 0.132) (Table 4.2). Essentially, the quantity and quality of DNA extracted from blood were adequate and pure for PCR.

Table 4.2: Description of the DNA quantity, absorbance, and purity of DNA

Statistics	DNA Quantity (ng/μl)	Absorbance at 260nm	Absorbance at 280 nm	The purity of DNA (260nm/280nm)
Mean	30.256	0.030	0.015	1.957
Median	29.260	0.029	0.015	1.937
Standard Deviation	13.084	0.013	0.006	0.132
Minimum	8.206	0.008	0.004	1.714
Maximum	61.730	0.061	0.029	2.176
Sum	907.666	0.906	0.462	58.703
Count	30.000	30.000	30.000	30.000
Confidence Level (95.0%)	4.886	0.005	0.002	0.049

4.3 Primers Designed

The designed primers were tabulated to depict genes targeted, identification numbers (ID), primers designed, primer sizes, target regions, and product sizes (Table 4.3).

Table 4.3: Attributes of primers designed

Gene Name	Gene ID	Primers	GC (%)	Tm (°C)	Size (b)	Target Region	Product Sizes (bp)
Gametocyte development protein 1 (GDV1)	>PF3D7_0935400 (1800bp)	>PF400_50F GTAGGCGTCGAAATAGTGCT	50.0	57.1	20	50	1306
		>PF400_1355R TCAGGATGTGTTATGGTATC	40.0	53.8	20	1355	
		>PF400_1081F GGTATTCCTGTTGTTATGAG	40.0	54.0	20	1081	650
		>PF400_1731R AGAAGATGATGAATGCTGACG	43.0	57.9	21	1731	
Transcription factor with AP2 domains (AP2-G)	PF3D7_1222600 (7299bp)	>PF600_5662F GTGTACGGGTAATAAATAAAG	36.4	52.6	22	5662	341
		>PF600_6002R TCGTTGCTGTTATTGTTG	38.9	51.2	18	6002	
Pf 18S rRNA (Positive Control)	M19173.1 (2040bp)	>PL1473F18 TAACGAACGAGATCTTAA	33.3	52.9	18	1473	224
		>PL1679R18 GTTCTCTAAGAAGCTTT	38.9	53.8	18	1679	

4.4 Fragment Sizes of PCR Products

PCR products were resolved on 2% agarose gel electrophoresis and fragment sizes were 1306bp, 650bp, and 341bp fragments as expected (Figure 4.2). *In silico* PCR indicated that the expected sizes of *Pfgdv1* gene are 1306bp fragment for coding region (50 to 1355) and 650bp fragment for the coding region (1081 to 1731), whereas the expected size of the most polymorphic site of *Pfap2g* is 341bp.

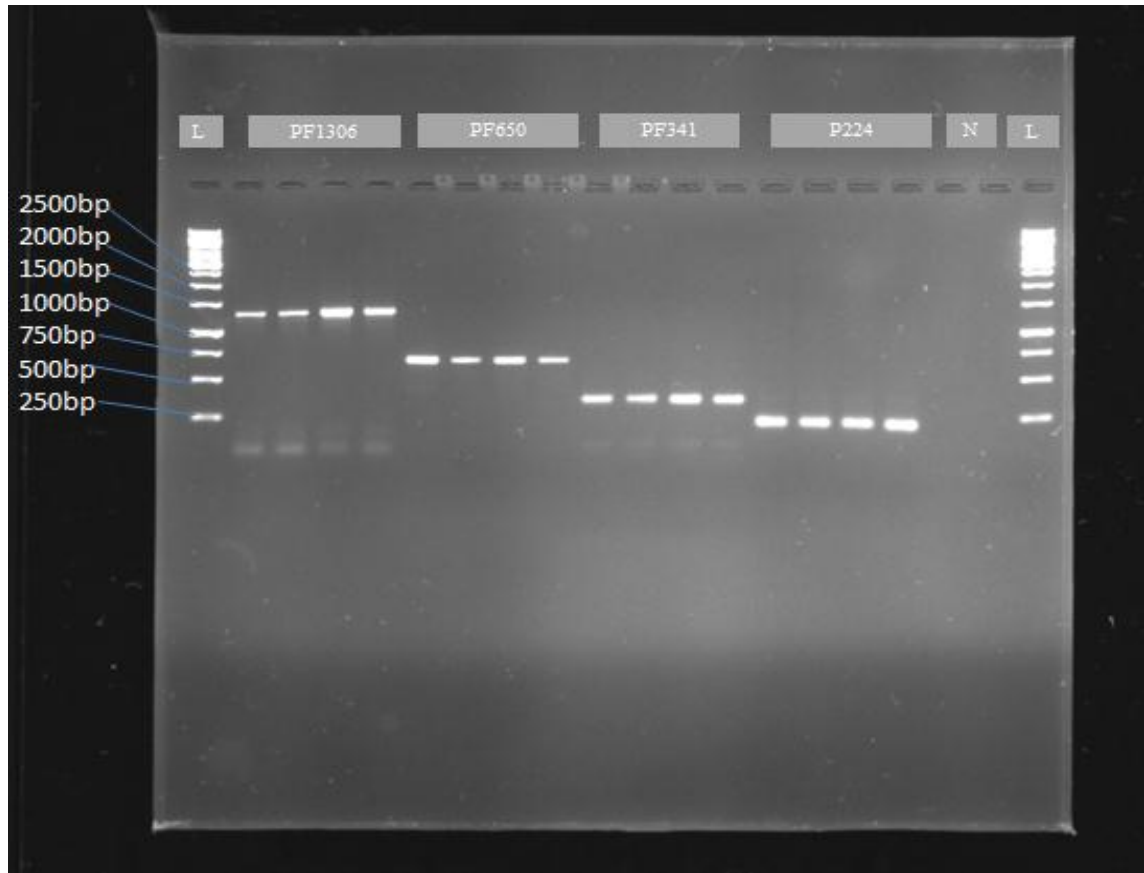


Figure 4.2: PCR products of amplified target regions of *P. falciparum* genes resolved on 2% agarose gel electrophoresis. L: 1kb DNA ladder (O'GeneRuler), PF1306: 1306 bp targeted size of amplicon, PF650: 650bp targeted size of amplicon, PF341: 341 bp targeted size of amplicon, P224: 224bp size of positive control, N: negative control.

4.5 PCR Products of Target Genes

4.5.1 The First Fragment of *Pfgdv1*

The resolution of amplicons on 2% agarose gel electrophoresis gave the expected size of 1306 bp falling between 1000 bp and 1500 bp of 1Kb DNA ladder (Figure 4.3). Since bands of 30 samples and positive controls have clear and uniform sizes, it implies that the primers were specific to the target region of *Pfgdv1*.

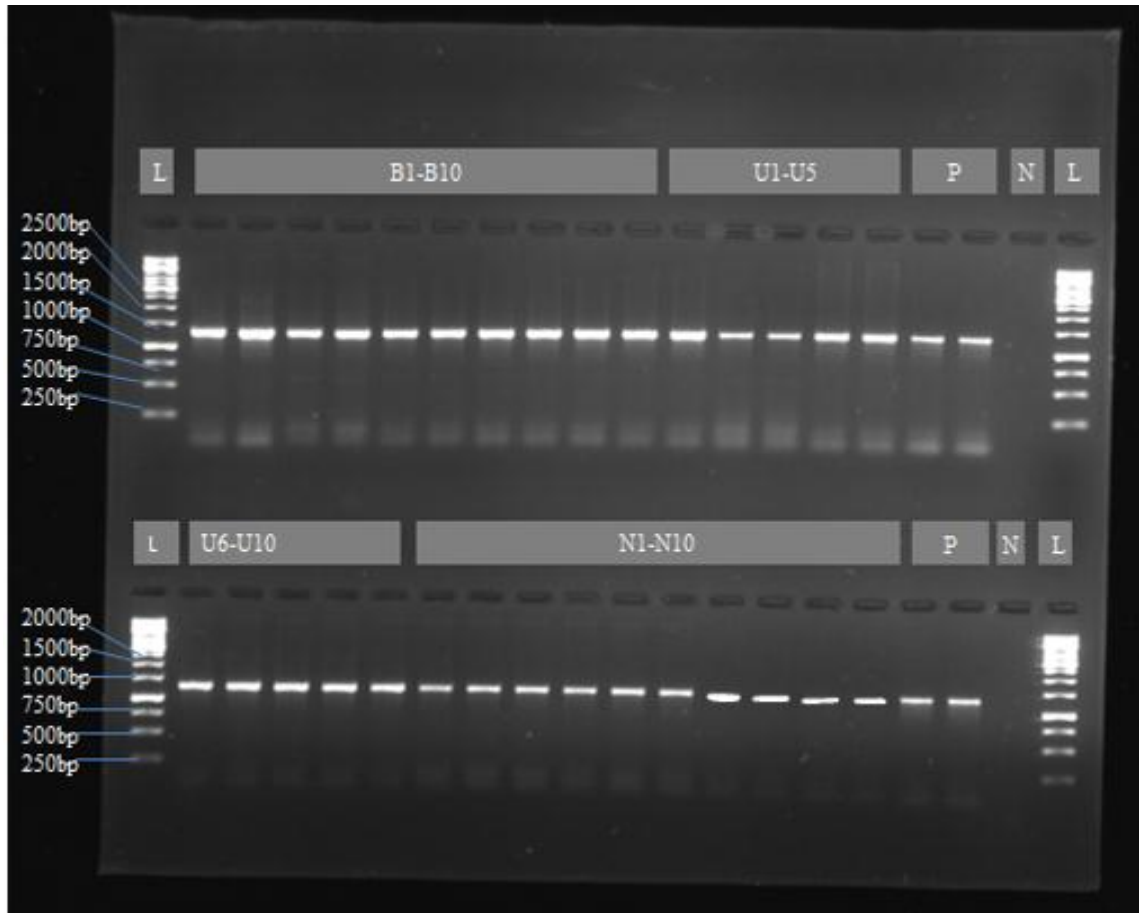


Figure 4.3: Amplicon size of 1306 bp amplified from *Pfgdv1* gene (Position 50-1355) of *P. falciparum* resolved on 2% agarose gel electrophoresis. L: 1kb DNA ladder (O'GeneRuler), Samples: B1-B10, U1-U10, and N1-N10. P: Positive control (3D7), N: Negative control.

4.5.2 The Second Fragment of *Pfgdv1*

The resolution of the second fragment of *Pfgdv1* in 2% agarose gel electrophoresis generated the expected size of 650 bp, which appears between 500 bp and 750 bp bands of 1Kb DNA ladder. The formation of clear and even bands for all the 30 samples and positive control demonstrates that the primers used were specific to the target fragment of *Pfgdv1*(Figure 4.4).

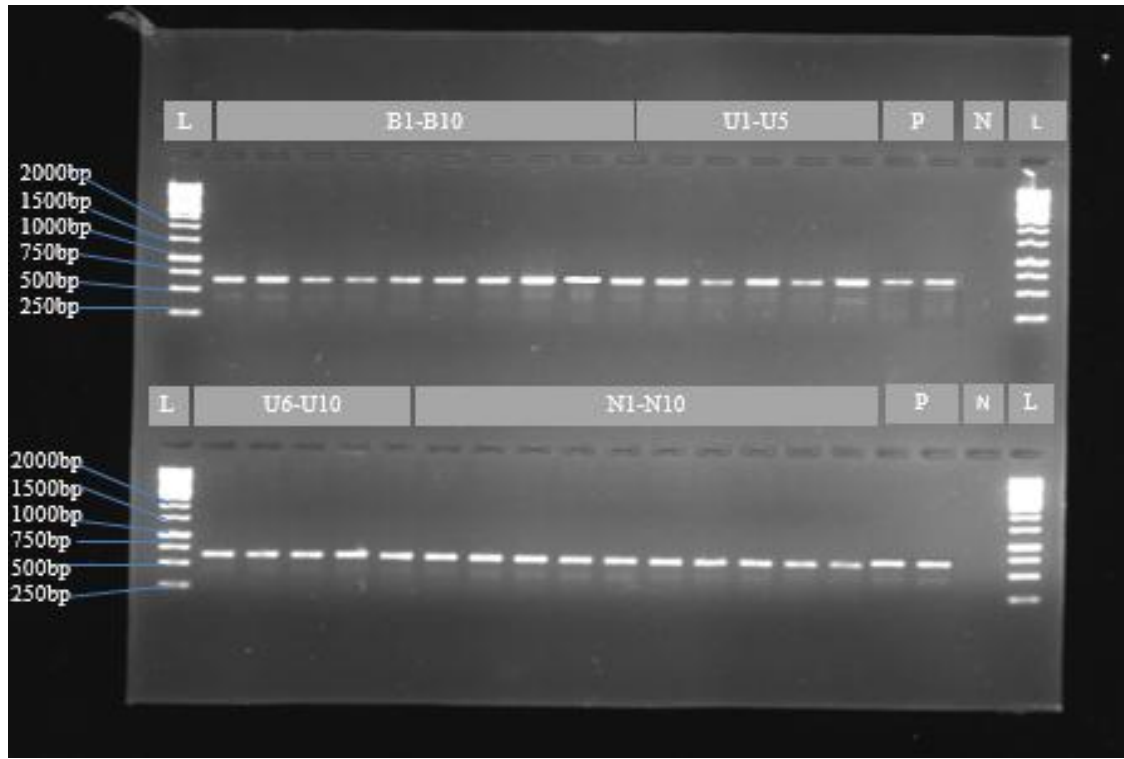


Figure 4.4: Amplicon size of 650 bp amplified from *Pfgdv1* gene (Position 1081-1730) of *Plasmodium falciparum* resolved on 2% agarose gel electrophoresis. Labelling:- L: 1kb DNA ladder (O’GeneRuler), Samples: B1-B10, U1-U10, and N1-N10. P: Positive control (3D7), N: Negative control.

4.5.3 Fragment of *Pfap2g*

PCR products generated from the most polymorphic fragment of *Pfap2* were resolved on 2% agarose gel electrophoresis (Figure 4.5). The analysis of the amplicons shows that they fall in the range of the expected size of 341 bp because they appear in between 250 bp and 500 bp bands of 1 kb DNA ladders. As the generated bands are clear and constant across samples and positive control, it implies that primers used to amplify the target region in *Pfap2g* are specific.



Figure 4.5: Amplicon size of 341 bp amplified from *Pfap2g* gene (Position 5661-6002) of *P. falciparum* resolved on 2% agarose gel electrophoresis. L: 1kb DNA ladder (O'geneRuler), Samples: B1-B10, U1-U10, and N1-10. P: Positive control (3D7), N: Negative control.

4.6 Sequences

4.6.1 Secondary Sequences

The secondary sequences of *Pfgdvl* and *Pfap2* retrieved from PlasmoDB (www.plasmodb.org) are for 216 isolates from various countries across the world. Most sequences were obtained from 70 isolates (32.4%) of Senegal followed by 65 isolates (30.1) of Gambia. Isolates of Mali were 23 (10.6%), while those of French Guiana Region were 22 (10.2%) (Table 4.4). Isolates of Uganda constituted 5.1% (11), while isolates of Brazil and Ghana comprised of 2.3% (5) and 1.4% (3), respectively. Thailand, Laos, Honduras, and Cambodia had 2 (0.9%) isolates each. The remaining isolates were from Viet Nam, Togo, Sudan, Kingdom of the Netherlands, Guinea, El Salvador, Gabon, the Democratic Republic of Congo, and Kenya, which had an isolate (0.5%) each, collectively representing 4.5% of the total population.

Table 4.4: Geographical distribution of *P. falciparum* isolates in PlasmodDB

Country	Number of Isolates	Percent Proportion of Isolates
Senegal	70	32.4%
Gambia	65	30.1%
Mali	23	10.6%
French Guiana Region	22	10.2%
Uganda	11	5.1%
Brazil	5	2.3%
Ghana	3	1.4%
Thailand	2	0.9%
Laos	2	0.9%
Honduras	2	0.9%
Cambodia	2	0.9%
Viet Nam	1	0.5%
Togo	1	0.5%
Sudan	1	0.5%
Kingdom of the Netherlands	1	0.5%
Guinea	1	0.5%
El Salvador	1	0.5%
Gabon	1	0.5%
The Democratic Republic of Congo	1	0.5%
Kenya	1	0.5%
Total	216	100%

4.6.2 Primary Sequences

Sequencing results (Appendix IV) provided raw sequences (forward and reverse sequences) in the form of chromatograms, which were used to generate primary sequences. The processing of chromatograms generated 30 contiguous sequences from forward and reverse raw sequences of two fragments of *Pfdv1*. Moreover, the processing of chromatograms generated 30 contiguous sequences from forward and reverse sequences of *Pfap2g* genes.

4.7 Single Nucleotide Polymorphisms

4.7.1 SNPs of *Pfgdv1* Gene

Multiple Sequence Alignment (MSA) of secondary sequences for *Pfgdv1* established the existence of five SNPs (Table 4.5). Additionally, MSA of primary sequences of *Pfgdv1* corroborated the existence of these five SNPs, namely, C650A, G1193A, C1249A, T1491A, and A1542T (Table 4.6). These SNPs give rise to four non-synonymous mutations (C650A, G1193A, C1249A, and T1491A) (Figure 4.8) and one synonymous substitution (A1542T). The allelic distribution indicates that mutations of C650A, C1249A, and A1542T form major alleles, whereas those of G1193A and T1491A constitute minor alleles. Visualization of multiple sequence alignment shows that P217H and H417N dominated all isolates of *P. falciparum*, while R398Q and D497E were unique to those isolates obtained from Baringo and Uasin Gishu Counties, correspondingly (Figure 4.6).

Table 4.5: SNPs identified in *Pfgdv1* among secondary isolates

CDS Position	Alleles			Protein Position	Products				% Calls	Strain Count
	Reference	Major	Minor		Major	Minor	Phenotype	Effects		
650	C	A (61%)	C (39%)	217	H	P	non- synonymous	non- conservative	87.6	207
1193	G	G (75%)	A (25%)	398	R	Q	non- synonymous	non- conservative	83.9	205
1249	C	A (82%)	C (18%)	417	N	H	non- synonymous	non- conservative	83.5	202
1491	T	T (90%)	A (10%)	497	D	E	non- synonymous	conservative	89	208
1542	A	T (74%)	A (26%)	514	S	Null	synonymous	conservative	90.4	200

Table 4.6: SNPs identified in *Pfgdv1* among primary isolates

CDS Position	Reference	Alleles		Protein Position	Products			Effects	% Calls	Isolate Count
		Major	Minor		Major	Minor	Phenotype			
650	C	A (63%)	C (37%)	217	H	P	non- synonymous	Non- conservative	94	30
1193	G	G (87%)	A (13%)	398	R	Q	non- synonymous	Non- conservative	94	30
1249	C	A (87%)	C (13%)	417	N	H	non- synonymous	Non- conservative	94	30
1491	T	T (90%)	A (10%)	497	D	E	non- synonymous	Conservative	94	30
1542	A	T (80%)	A (20%)	514	S	null	synonymous	Conservative	94	30

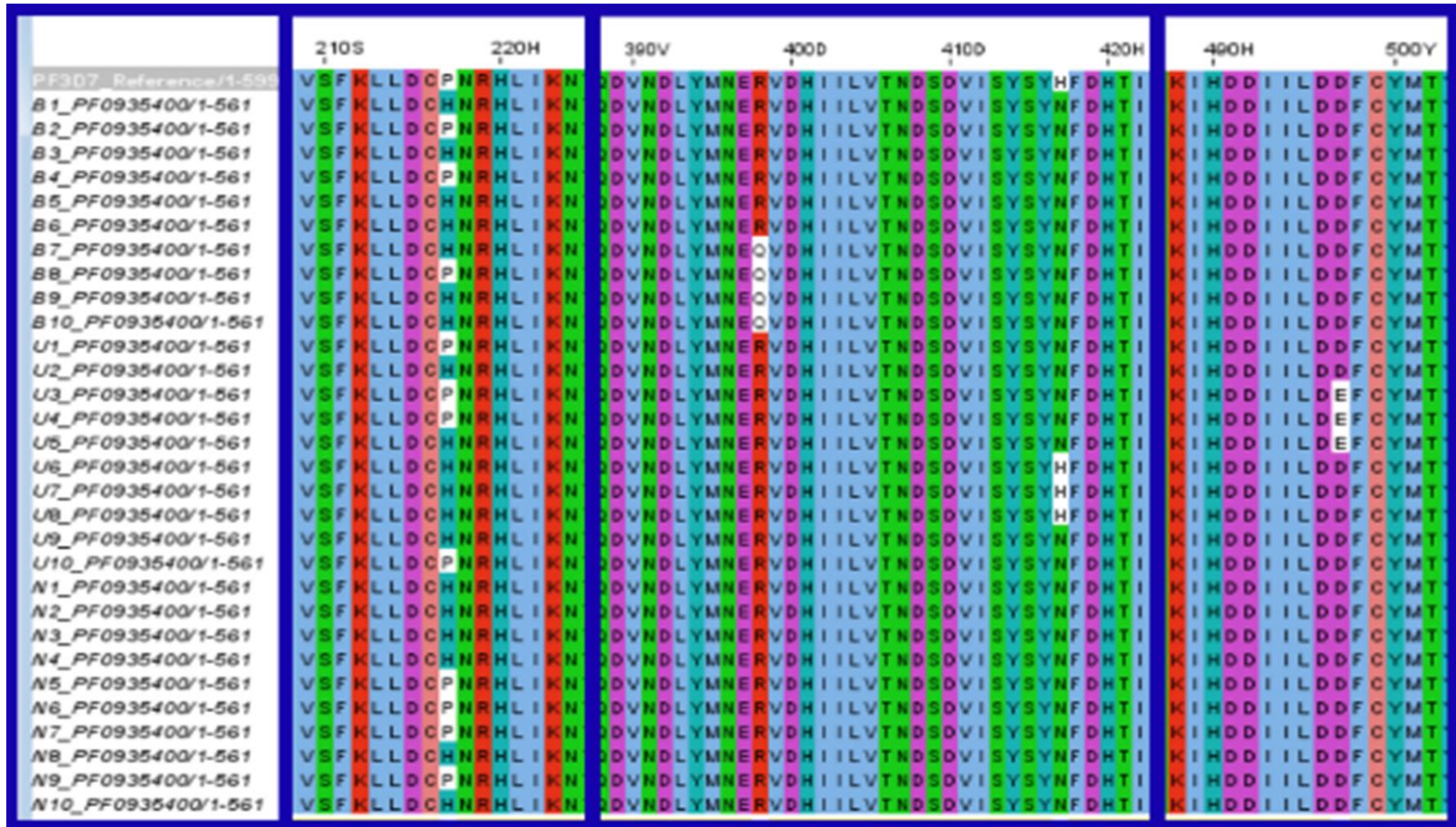


Figure 4.6: MSA visualized using Jalview to highlight four nsSNPs (P217H, R398Q, H417N, and D497E) in *Pfgdv1* sequences generated from 30 primary isolates of *P. falciparum* (B1-B10, U1-U10, N1-N10) aligned to the reference sequence (PF3D7).

4.7.2 SNPs of *Pfap2g* gene

The multiple sequences alignment of the second gene, *Pfap2g*, showed the existence of 12 polymorphic sites of SNPs (Table 4.7). The 12 SNPs comprises 11 non-synonymous and 1 synonymous with variation in the proportion of minor alleles ranging from 7% to 50%. Additionally, SNPs emanate from 7 transition mutations and 5 transversion mutations on 62, 817, 2034, 3539, 4753, 4762, 4970, 5638, 5767, 5922, and 5955 positions on the coding sequence. A summary of MSA outcomes of *Pfap2g* sequences are shown in the next page (Table 4.8)

Table 4.7: SNPs identified in *Pfap2g* among secondary isolates

CDS Position	Alleles			Protein Position	Products		Phenotype	Effects	% Calls	Strain Count
	Reference	Major	Minor		Major	Minor				
62	A	G (74%)	A (26%)	21	R	K	non-synonymous	Non-conservative	91.1	198
817	T	C (89%)	T (11%)	273	H	Y	non-synonymous	Non-conservative	96	198
2034	T	T (92%)	C (8%)	678	C	null	synonymous	Conservative	97	197
2951	A	A (55%)	C (45%)	984	K	T	non-synonymous	Non-conservative	89.1	199
3539	G	G (53%)	A (47%)	1180	G	E	non-synonymous	Non-conservative	86.6	194
4753	G	G (92%)	A (8%)	1585	G	R	non-synonymous	Non-conservative	95	197
4762	G	G (93%)	C (7%)	1588	E	Q	non-synonymous	Non-conservative	99	201
4970	A	A (92%)	T (8%)	1657	N	I	non-synonymous	Non-conservative	96.5	199
5638	G	G (86%)	A (14%)	1880	D	N	non-synonymous	Non-conservative	94.1	193
5767	G	G (90%)	A (10%)	1923	E	K	non-synonymous	Non-conservative	89.6	189
5922	A	T (50%)	A (50%)	1974	N	K	non-synonymous	Non-conservative	87.1	196
5955	C	C (54%)	A (45%)	1985	N	K	non-synonymous	Non-conservative	88.1	199

MSA indicates that the most polymorphic region of *Pfap2g* ranges from 5630 to 6000 nucleotide positions with an insertion and 4 SNPs (Table 4.8). The analysis of primary sequences shows that there is an insertion of codon (GAG) in all sequences and SNPs of A5761G, G5770A, and C5958A (Table 4.8). Visualization of MSA depicts that the insertion (1920E) and N1921D are present in all sequences, E1924K are evenly distributed in sampling sites, while K1975N and N1986K are dominant in isolates from Baringo County (Figure 4.7).

Table 4.8: SNPs identified in *Pfap2g* among primary isolates

CDS Position	Alleles			Protein Position	Products				Isolate Count
	Reference	Major	Minor		Major	Minor	Phenotype	Effects	
5758	-	G	-						
5759	-	A	-	1920	E		null	Non-conservative	30
5760	-	G	-						
5761	A	G (97%)	A (3%)	1921	D	N	non-syn	Non-conservative	30
5770	G	G (90%)	A (27%)	1924	E	K	non-syn	Non-conservative	30
5925	A	T (83%)	A (17%)	1975	N	K	non-syn	Non-conservative	30
5958	C	A (80%)	C (20%)	1986	N	K	non-syn	Conservative	30

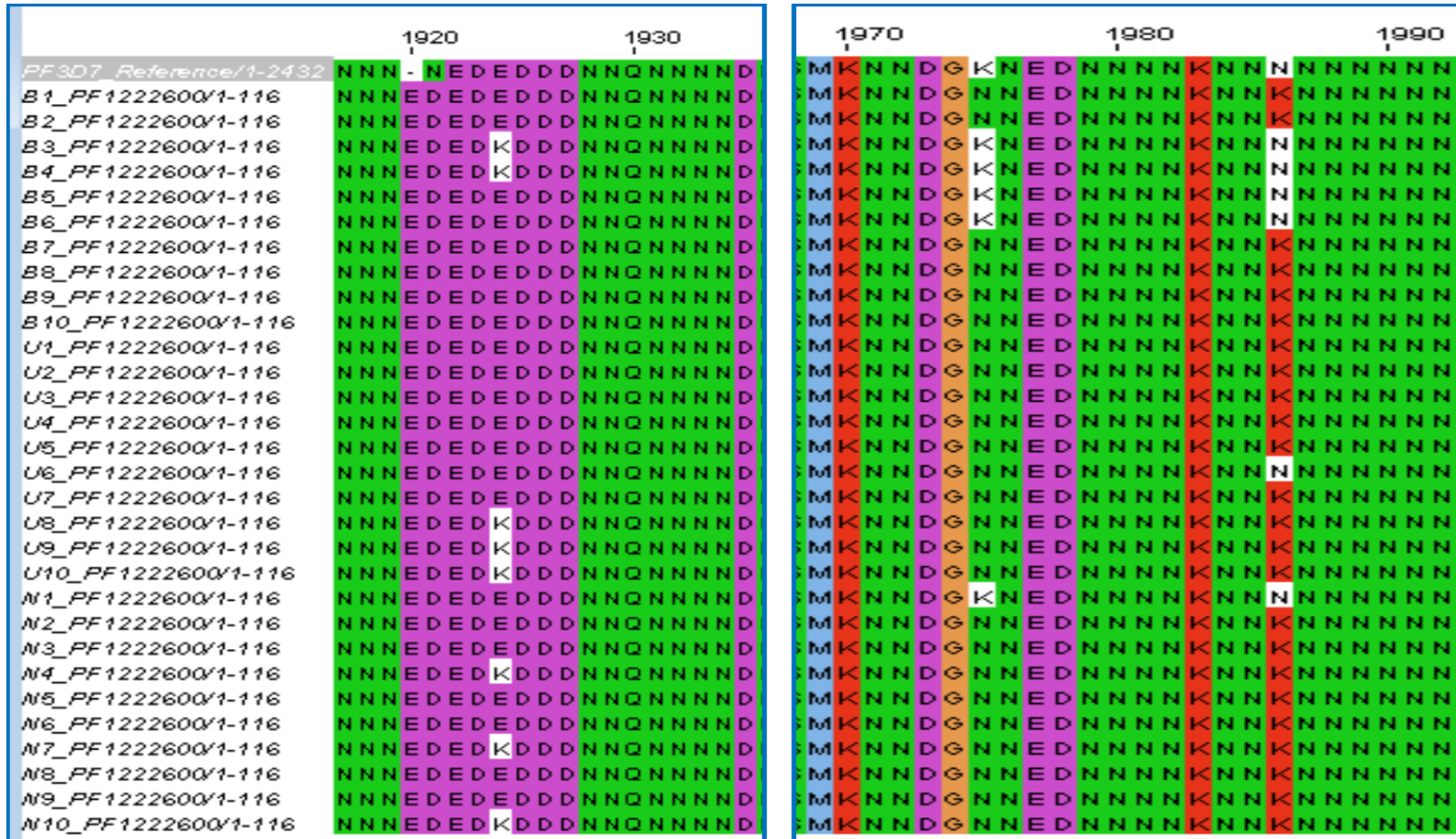


Figure 4.7: MSA of *Pfap2g* visualized using Jalview to highlight insertion of 1920E, and four nsSNPs (N1921D, E1924K, K1975N, and N1986K) in sequences generated from 30 primary isolates of *P. falciparum* (B1-B10, U1-U10, N1-N10) aligned to the reference sequence (PF3D7).

4.7.3 Effect of Non-synonymous Substitutions

The prediction of protein stability changes revealed that three out of four nsSNPs influence the stability of *Pfgdv1* protein structure (Table 4.9). While P217H (ddG = -2.22) has a destabilizing effect, D497E (ddG=1.62) and H417N (ddG = 0.92) have a stabilizing effect on the protein structure. Since its Gibbs free-energy gap is close to zero, R398Q (ddG = 0.07) has no considerable effect on the stability of protein structure. Stabilizing effect means that a nsSNP does not influence commitment to gametocytes, while destabilizing effect implies that a nsSNP has a significant influence on gametocyte development.

Table 4.9: Gibbs free-energy gaps (ddG) of nsSNPs in *Pfgdv1*

Position	Wild Type	Mutant Type	ddG
217	P	H	-2.22
398	R	Q	0.07
417	H	N	0.92
497	D	E	1.62

The analysis of energy changes due to 11 nsSNP on *Pfap2g* protein shows that all have stabilizing effects since their Gibbs free-energy gaps are positive. Overall, Gibbs free energy gaps ranged from 0.55 to 3.60 (Table 4.10).

Table 4.10: Gibbs free-energy gaps (ddG) of nsSNPs in *Pfap2g*

Position	Wild Type	Mutant Type	ddG
21	K	R	2.97
273	Y	H	2.71
984	K	T	2.51
1180	G	E	2.86
1585	G	R	3.60
1588	E	Q	0.55
1657	N	I	2.62
1880	D	N	2.10
1923	E	K	1.85
1974	K	N	0.62
1985	N	K	1.49

4.8 Selection Analysis

4.8.1 Selection Analysis of *Pfgdv1*

The analysis of the four-polymorphic sites indicated that nsSNPs in *Pfgdv1* are under significant selection ($p = 0.007$), leading to the rejection of the hypothesis of neutrality (Table 4.11). Results for the primary data ($n = 30$) showed a positive Tajima's D (0.209) since observed variation ($\pi = 0.0019$) is more than expected variation ($\Theta = 0.0018$), which is statistically significant ($p = 0.007$). Additionally, results obtained from the secondary data ($n = 184$) revealed a similar trend due to a positive Tajima's D (0.929) with observed variation ($\pi = 0.0018$) that is greater than expected variation ($\Theta = 0.0012$).

Table 4.11: Results from Tajima's neutrality test

M	S	p_s	Θ	Π	D
30	4	0.007	0.0018	0.0019	0.209
184	4	0.007	0.0012	0.0018	0.929

Abbreviations: m = number of sequences, S = number of polymorphic sites, p_s = p-value (S/n), Θ = expected variation (p_s/a_1), π = observed variation (nucleotide diversity), and D = the Tajima test statistic

SLAC shows that p.P217H is under the strongest positive selection in both the primary (dN-dS = 361.68) and secondary data (dN-dS = 177.70) based on statistically significant model of nucleotide substitution (The General Time Reversible) and stringent significance level ($p < 0.05$) (Tables 4.12 and 4.13). Moreover, SLAC indicates that R398Q and D497E are other nsSNPs that exhibit positive selection but not statistically significant ($p > 0.3$). S514S is an nsynonymous SNP that is under negative selection (dN-dS = -214.70).

Table 4.12: Selection analysis outcomes of primary data of *Pfgdv1*

Local Isolates (n = 30)								
Partition	Sites	Branches		Branch Length			Selected at $p \leq 0.05$	
		Tested	Total	Tested	% of total	Total	Positive	Negative
1	561	15	15	0.00233	50.0%	0.00466	1	0
DN/DS Analysis								
Codon Site	ES	EN	S	N	dS	dN	dN-dS	
217	0.63	2.38	0.00	4.00	0.00	1.68	361.68	
398	1.65	1.35	0.00	1.00	0.00	0.74	158.98	
497	0.02	2.98	0.00	1.00	0.00	0.34	72.17	
417	0.00	3.00	0.00	1.00	0.00	0.33	0.00	
514	1.00	1.69	1.00	0.00	1.00	0.00	-214.70	

Selection analysis of five SNPs in 30 primary sequences showing a positively selected nsSNP at codon site 217 based on a p-value of less than 0.05 (dN-dS = 361.68)

Table 4.13: Selection analysis outcomes of secondary data of *Pfgdv1*

Global Isolates (n = 184)								
Partition	Sites	Branches		Branch Length		Selected at p≤0.05		
		Tested	Total	Tested	% of total	Total	Positive	Negative
1	599	49	49	0.00333	35.30%	0.00945	1	1

DN/DS Analysis								
Site	ES	EN	S	N	dS	dN	dN-dS	
217	0.62	2.38	0.00	4.00	0.00	1.68	177.70	
497	0.02	2.98	0.00	2.00	0.00	0.67	70.94	
398	0.87	2.13	0.00	1.00	0.00	0.47	49.72	
417	0.00	3.00	0.00	5.00	0.00	1.67	0.00	
514	1.00	1.62	5.00	0.00	5.00	0.00	-529.25	

Selection analysis of five SNPs in 184 secondary sequences of *Pfgdv1* showing a positively selected nsSNP at codon site 217 (dN-dS = 177.70) and negatively selected synonymous SNP at codon site 514 (dN-dS = -529.25) based on a p-value of less than 0.05

4.8.2 Selection Analysis of *Pfap2g*

The selection analysis using Tajima's D test of the 11-polymorphic sites of *Pfap2g* rejects the hypothesis of neutrality (p = 0.0049) (Table 4.14). Outcomes of 172 sequences revealed that a positive Tajima's D (1.426) because the observed variation ($\pi = 0.0014$) is greater than the expected variation ($\Theta = 0.0009$).

Table 4.14: Results from Tajima's neutrality test of *Pfap2g*

M	S	p_s	Θ	Π	D
172	11	0.0049	0.0009	0.0014	1.426

Abbreviations: m = number of sequences, S = number of polymorphic sites, p_s = p-value (S/n), Θ = expected variation (p_s/a₁), π = observed variation (nucleotide diversity), and D = the Tajima test statistic

SLAC indicated that K984T (dN-dS = 747.19) is under the strongest positive selection followed by K21R (dN-dS = 702.84). These two selection sites are statistically significant at p-value that is less than 0.05. Other nsSNPs that are statistically insignificant since their p-values are greater than 0.05 are Y273H (dN-dS = 397.88), G1180E (dN-dS = 362.61), and K1974N (dN-dS = 335.78) based on the substitution model. However, a synonymous SNP of C678C is under a negative selection (dN-dS = -323.42), which is insignificant (p > 0.05).

Table 4.15: Selection analysis outcomes of secondary data of *Pfap2g*

Global Isolates (n = 172)								
Partition	Sites	Branches		Branch Length			Selected at $p \leq 0.05$	
		Tested	Total	Tested	% of total	Total	Positive	Negative
1	2432	83	83	0.00325	52.2%	0.00624	2	0
DN/DS Analysis								
Codon Site	ES	EN	S	N	dS	dN	dN-dS	
21	0.76	2.05	0.00	9.00	0.00	4.38	702.84	
273	0.50	2.42	0.00	6.00	0.00	2.48	397.88	
678	0.50	2.00	1.00	0.00	2.02	0.00	-323.42	
984	0.98	1.93	0.00	9.00	0.00	4.66	747.19	
1180	0.79	2.21	0.00	5.00	0.00	2.26	362.61	
1585	0.96	2.00	0.00	2.00	0.00	1.00	160.37	
1588	0.53	2.45	0.00	1.00	0.00	0.41	65.35	
1657	0.55	2.45	0.00	2.00	0.00	0.82	131.15	
1880	0.37	2.63	0.00	3.00	0.00	1.14	182.82	
1923	0.53	2.45	0.00	1.00	0.00	0.41	65.51	
1974	0.51	2.39	0.00	5.00	0.00	2.09	335.78	
1985	0.44	2.48	0.00	2.00	0.00	0.81	129.37	

Selection analysis of 12 SNPs in 172 secondary sequences of *Pfap2g* showing positively selected nsSNPs at codon sites 21 (dN-dS = 702.84) and 984 (dN-dS = 747.19) based on a p-value of less than 0.05

4.9 Protein Structure Prediction

4.9.1 Prediction of *Pfgdv1* Protein Structure

4.9.1.1 Primary Structure Analysis of *Pfdv1*

Predicted physicochemical properties shows that *Pfgdv1* has 599 amino acid residues, molecular weight of 71.964 kDa, theoretical isoelectric point of 8.84, 10080 atoms, 72 negatively charged amino acid residues (Asp +Glu), 85 positively charged amino acid residues (Arg + Lys), aliphatic index of 86.99, grand average of hydropathicity of -0.596. *Pfgdv1* also have extinction efficiencies of 74330 and 73580 when cysteine

residues are oxidized and reduced, respectively (Table 4.16). The half-life of *Pfgdv1* is less than 30 hours with instability index of 42.52 (unstable). Amino acid composition indicates that is rich in asparagine (12%), lysine (10.2%), isoleucine (9.3%), and leucine (9.2%), but poor in tryptophan (0.3%), alanine (1.2%), glutamine (1.7%), and glycine (1.8%).

Table 4.16: Physicochemical properties of *Pfgdv1* as predicted by Protparam

Physicochemical Properties		Values
Number of amino acids		599
Molecular formula		$C_{3238}H_{5015}N_{867}O_{928}S_{32}$
Total number of atoms		10080
Molecular weight		71.964 kDa
Theoretical isoelectric point		8.84
Total number of negatively charged residues	(Asp + Glu)	72
Total number of positively charged residues	(Arg + Lys)	85
Extinction coefficients are in units of $M^{-1}cm^{-1}$, at 280 nm measured in water.		
Extinction coefficient (Absorbance 0.1% (1/l))	74330	1.033, assuming all pairs of cysteine residues form cystines
Extinction coefficient (Absorbance 0.1% (1/l))	73580	1.022, assuming all cysteine residues are reduced
Estimated half-life	30 hours (mammalian reticulocytes) >20 hours (yeast, in vivo) > 10 hours (<i>Escherichia coli</i> , in vivo)	
Aliphatic index	86.99	
Instability index	42.52	(Classifies protein as unstable)
Grand average of hydropathicity (GRAVY)	-0.596	
Amino acid composition	Ala (A)	7(1.2%)
	Arg (R)	24(4%)
	Asn (N)	77(12.9%)
	Asp (D)	40(6.7%)

Cys (C)	13(2.2%)
Gln (Q)	10(1.7%)
Glu (E)	32(5.3%)
Gly (G)	11(1.8%)
His (H)	23(3.8%)
Ile (I)	56(9.3%)
Leu (L)	55(9.2%)
Lys (K)	61(10.2%)
Met (M)	19(3.2%)
Phe (F)	30(5.0%)
Pro (P)	14(2.3%)
Ser (S)	31(5.2%)
Thr (T)	24(4.0%)
Trp (W)	2(0.3%)
Tyr (Y)	42(7.0%)
Val (V)	28(4.7%)

4.9.1.2 Secondary Structure Analysis of *Pfdv1* protein

The prediction of the secondary structure using RaptorX-Property indicates the 3-state secondary structure (SS3), 8-state secondary structure (SS8), solvent accessibility (ACC), and disorder regions (DISO) of *Pfdv1* protein. In the prediction of the SS3, *Pfdv1* constitutes 33% of α -helix, 16% of beta-sheet, and 49% of coil structures. The SS8 prediction shows that coil, α -helix, and extended strand in beta ladder dominate the secondary structure. In the aspect of ACC, buried (<10%), medium (10-40%), and exposed (>40%) formed 31%, 27%, and 40%, correspondingly. B-factor profile shows that thermal mobility of *Pfdv1* fluctuates between negative and positive values, indicating instability of the protein (Figure 4.8). DISO shows that 5% of the secondary protein structure is disordered, leaving the remaining 95% percent ordered.

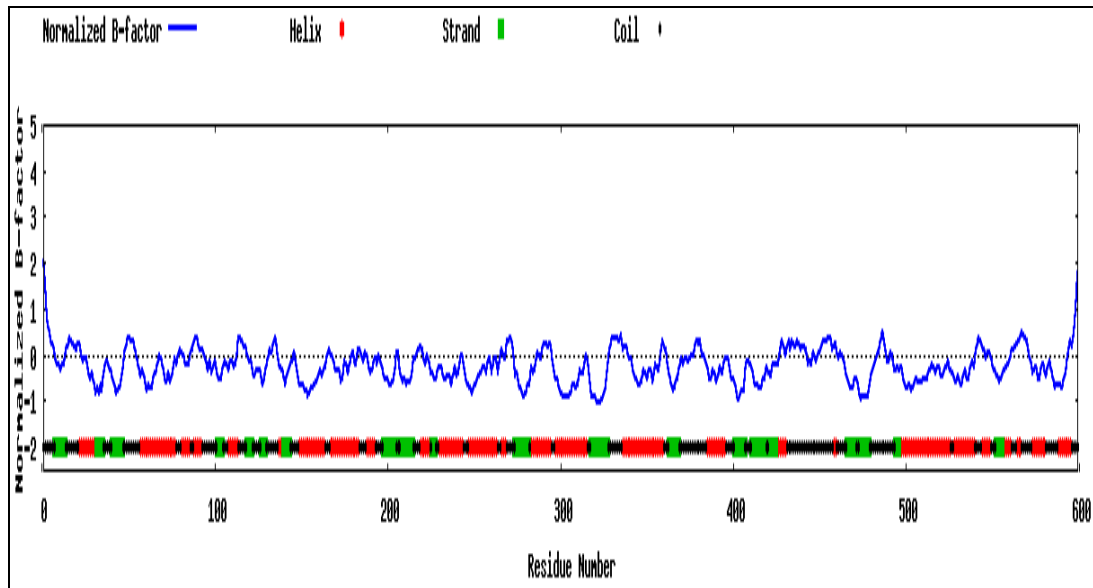


Figure 4.8: The normalized B-factor (thermal mobility) of *Pfgdv1* protein. Red, green, and black colors depicts trends of alpha helix, beta strand and coil structures of proteins, respectively, while line graph shows how thermal stability of protein varies along its residues.

4.9.1.3 Tertiary Structure Prediction of *Pfdv1* Protein

Predictive information obtained from InterPro shows that *Pfgdv1* has no established or predicted domains, family, and functional sites. Moreover, gene ontology shows that *Pfgdv1* has no predicated information regarding biological process, molecular function, and cellular component. However, IntroPro predicts the presence of a coil between 340 and 360 residues of the protein sequence (Figure 4.9).

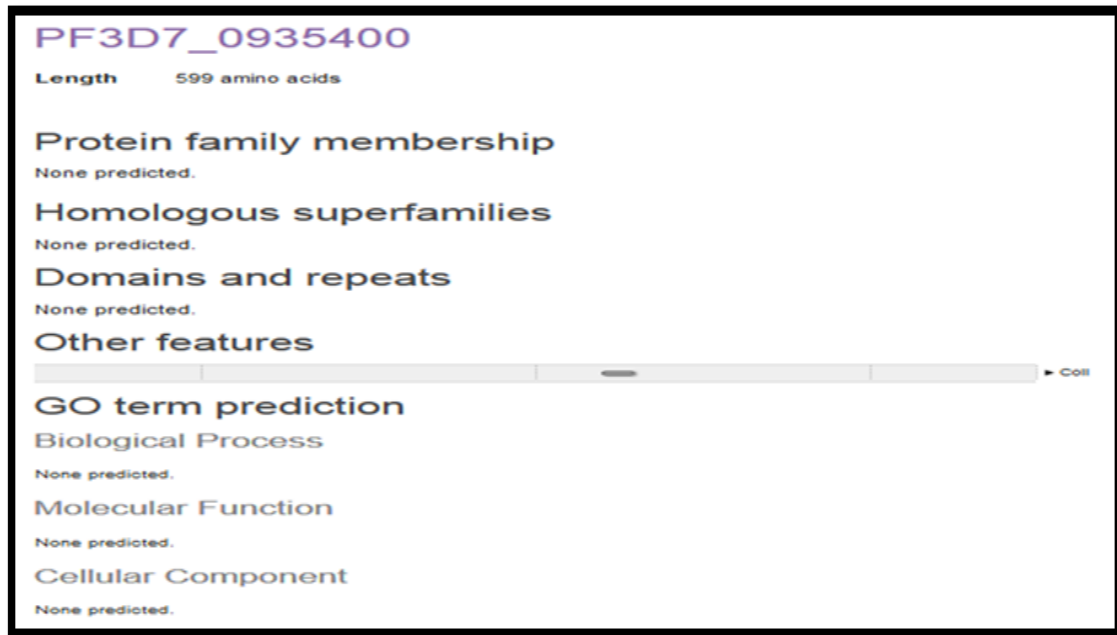


Figure 4.9: Prediction of *Pfgdv1* protein using InterPro showing no family, domain, and gene ontology identified

In predicting the tertiary structure, RaptorX indicates that *Pfgdv1* protein has two domains. The first domain (A) covers 1-143 amino acid residues with the global distance test of 19, domain score of 21, and p-value of 1.97e-02. 3fvzA, 3soqA, 1ijqA, 3v64C, and 3u4yA are the templates that RaptorX used in modelling the structure of the first domain (Figure 4.12). The second domain (B) covers 144-599 amino acid residues with the global distance test of 12, domain score of 81, and p-value of 2.45e-03. Templates used in modelling the protein structure of the second domain are 4f9iA, 5ur2A, and 4o8aA (Figure 4.10).

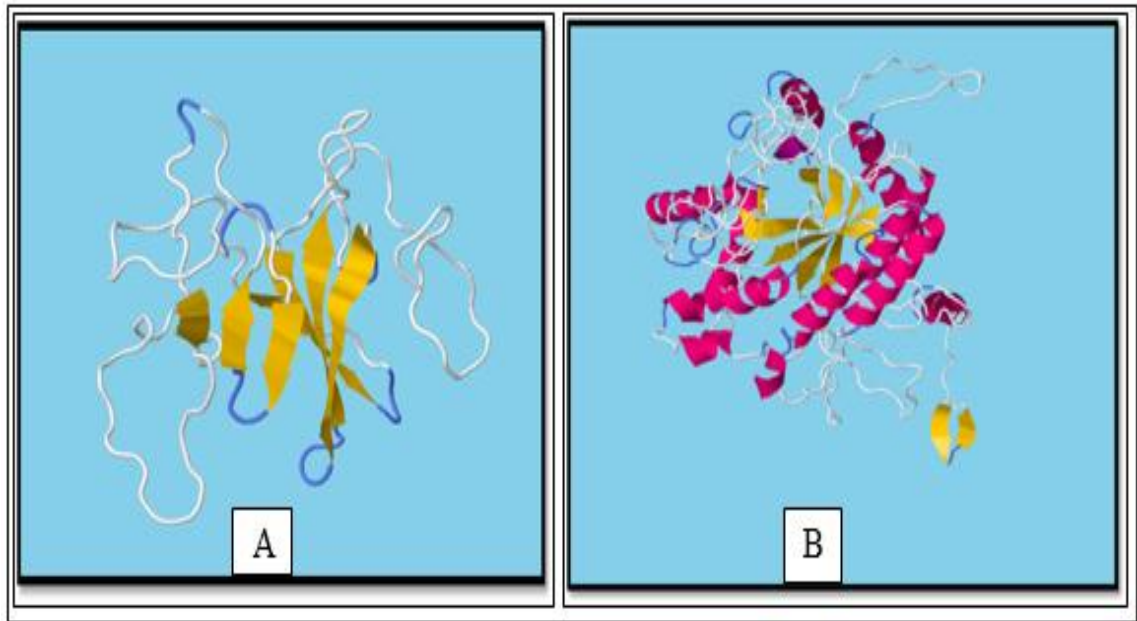


Figure 4.10: Structure prediction of the first domain A (1-143) the second domain B (144-599) of *Pfgdv1* protein using Raptor X

A further prediction of the protein structure using I-TASSER provided additional information about *Pfgdv1* protein (Appendix V). Table 4.17 below shows PDB hit, local identity, global identity, coverage, and normalized Z score of the ten threading templates used by I-TASSER to model *Pfgdv1* 3D structure. The leading threading templates are pitrilysin (1q21A), apoptosome (1vt4), Cmr2 unit (4w8yA), Gtr1p-Gtr2p complex (3r7wA), endonuclease (5wtjA), TubZ (4ei7A), Cryo-EM (5yfpE), ABC transporters (4pevA), human patched1 (6d4hA), and *Legionella pneumophila* effector RidL (5ot4A).

Table 4.17: The leading 10 threading templates of *Pfgdv1*

Rank ^a	PDB Hit ^b	Identity 1 ^c (%)	Identity 2 ^d (%)	Coverage ^e (%)	Normalized Z-Score ^f
1	1q21A	0.14	0.17	0.82	1.21
2	1vt4	0.19	0.07	0.73	1.02
3	4w8yA	0.09	0.20	0.95	1.62
4	3r7wA	0.18	0.11	0.38	1.10
5	5wtjA	0.10	0.21	0.98	1.62
6	4ei7A	0.13	0.13	0.47	1.10
7	5yfpE	0.09	0.20	0.91	1.62
8	4pevA	0.14	0.10	0.52	1.10
9	6d4hA	0.07	0.16	0.93	1.58
10	5ot4A	0.10	0.17	0.87	1.56

Ranked numbers of templates (a), PDB hit ID (b), percent local identity (c), percent global identity (d), percent coverage of *Pfgdv1* (e), and Z-score greater than 0.5 rated as good threading alignment

Outcomes of I-TASSER indicate structural analogs of *Pfgdv1* in PDB and their respective template modeling score (TM-score), root-mean-square-deviation (RMSD), local identity, and coverage (Table 4.18). The analogs were pitrilysin (1q21A), IDE-bradykinin complex (3cwwB), falcilysin (3s5hA), presequence protease PreP (2fgeA), human presequence protease (4l3tA), insulin-degrading enzyme (2g47A1), ferredoxin protease FusC (6b03A), Pyrroloquinoline quinone (5cioA), mitochondrial cytochrome bc1 complex (110nA), and cytochrome BC1 Complex (3cwbA).

Table 4.18: The leading 10 structural analogs of *Pfgdv1* identified in PDB

Rank ^a	PDB Hit ^b	TM-Score ^c	RMSD ^d	Identity ^e	Coverage ^f
1	1q2lA	0.785	2.83	0.098	0.851
2	3cwwB	0.628	5.13	0.077	0.783
3	3s5hA	0.553	5.78	0.078	0.738
4	2fgeA	0.533	5.73	0.067	0.713
5	4l3tA	0.531	5.69	0.082	0.705
6	2g47A1	0.529	3.43	0.101	0.596
7	6b03A	0.515	6.37	0.042	0.726
8	5cioA	0.510	5.45	0.055	0.658
9	1l0nA	0.450	4.69	0.054	0.553
10	3cwbA	0.449	4.62	0.063	0.549

Ranked numbers of analogs (a), PDB hit ID (b), threading template score greater than 0.5 rated as good (c), RMSD less than 5 regarded as reliable (d), percent identity (e) and coverage of *Pfgdv1* (f)

I-TASSER also predicted global and local accuracy estimations of the most accurate model of *Pfgdv1* (Figure 4.11). The global accuracy estimation indicates that the predicted structure has a confidence score (c-score) of -2.95, TM-score of 0.38 ± 13 , and RMSD of 15.1 ± 3.5 Å. The estimated local accuracy of the model (plot) shows that most regions are accurate with an estimated mean distance of 6 Å from the native structure. Figure 4.12 shows the predicted 3D model by I-TASSER with three apparent domains.

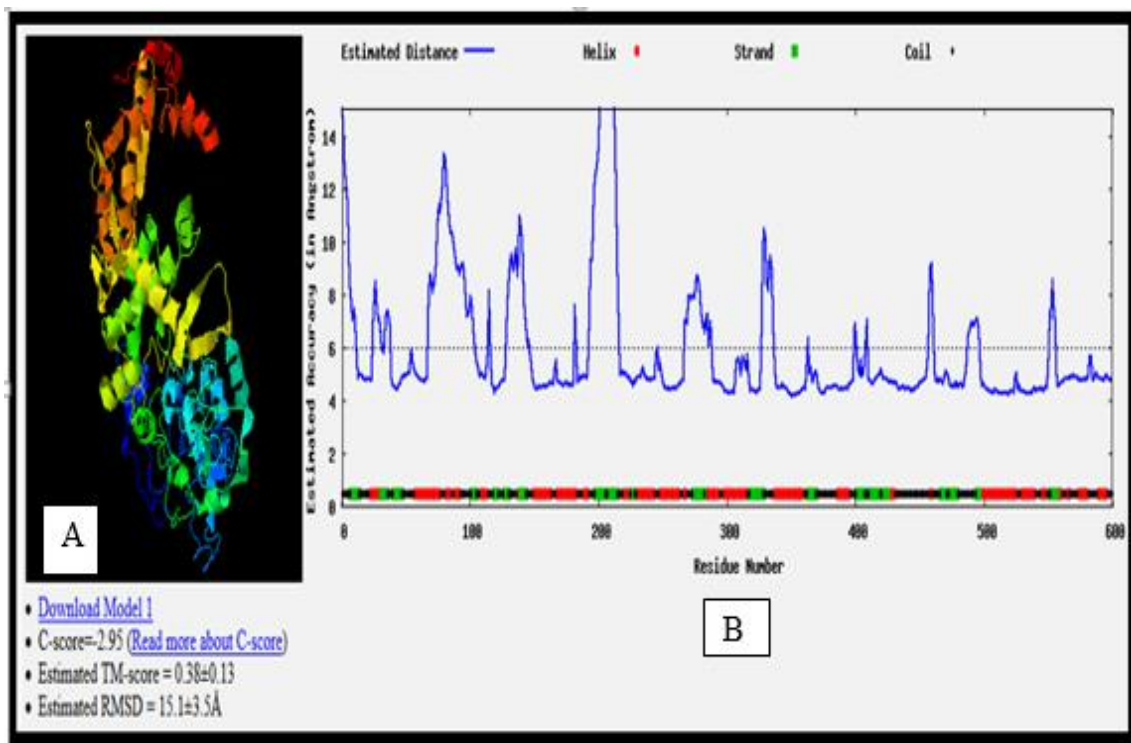


Figure 4.11: The predicted model of *Pfgdv1* protein structure showing the global and local accuracy estimations. **A:** Showing global accuracy estimation of the *Pfgdv1* model in C-score (-2.95), TM-score (0.38±0.13), and RMSD (15±3.5 Å) and **B:** Showing the local accuracy estimation of the *Pfgdv1* model in Angstroms with average value of 6 Å with variations in estimated distance of each residue number and trends of helix (red), strand (green), and coil (black) structures

According to Figure 4.13, the predicted structure of *Pfgdv1* has three apparent domains, which cover 1-277, 278-437, and 438-599 segments of the amino acid sequences.

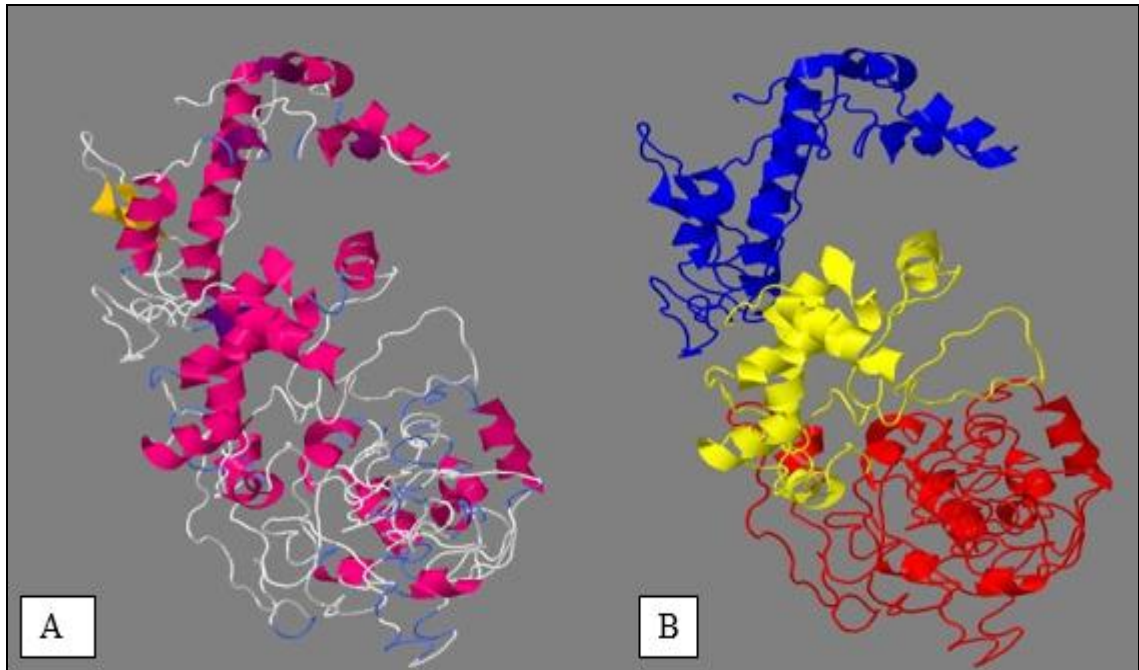


Figure 4.12: Front view of the globally estimated *Pfgdv1* protein model. A: Depicts the predicted protein model of *Pfgdv1* with c-score of -2.95, TM-score of 0.38 ± 13 , and RMSD of $15.1 \pm 3.5 \text{ \AA}$, and B: Indicates the orientation of the model with highlighted three domains (B) (1-277: Red, 278-437: Yellow, 438-599: Blue).

4.9.2 Prediction of *Pfap2g* Protein Structure

4.9.2.1 Primary Structure Analysis of *Pfap2g* Protein

The analysis outcomes of the primary structure of *Pfap2g* protein shows that it has 2432 amino acid residues, molecular weight of 284.064 kDa, theoretical isoelectric point of 8.72, and 38495 atoms (Table 4.19). Moreover, it has 246 negatively charged amino acid residues (Asp +Glu), 282 positively charged amino acid residues (Arg + Lys), aliphatic index of 56.05, and grand average of hydropathicity of 1.382. *Pfap2g* also has extinction efficiencies of 214385 and 212260 when cysteine residues are oxidized and reduced, correspondingly. The half-life of *Pfap2g* is less than 30 hours with instability index of 39.93. The amino acid composition indicates that it is rich in asparagine (29.9%), but poor in tryptophan (0.2%) and proline (1.6%).

Table 4.19: Physicochemical properties of *Pfap2g* as predicted by Protparam

Physicochemical Properties	Values	
Number of amino acids	2432	
Molecular formula	$C_{11968}H_{18585}N_{3751}O_{4067}S_{124}$	
Total number of atoms	38495	
Molecular weight	284.064 kDa	
Theoretical isoelectric point	8.72	
Total number of negatively charged residues	(Asp + Glu)	246
Total number of positively charged residues	(Arg + Lys)	282
Extinction coefficients are in units of M/cm, at 280 nm measured in water		
Extinction coefficient (Absorbance 0.1% (1/l))	214385	0.758, assuming all pairs of cysteine residues form cystines
Extinction coefficient (Absorbance 0.1% (1/l))	212260	0.747, assuming all cysteine residues are reduced
Estimated half-life	30 hours (mammalian reticulocytes) >20 hours (yeast, in vivo) > 10 hours (Escherichia coli, in vivo)	
Aliphatic index	56.05	(Classifies protein as stable)
Instability index	39.93	
Grand average of hydropathicity (GRAVY)	-1.382	
Amino acid composition	Ala (A)	34(1.4%)
	Arg (R)	56(2.3%)
	Asn (N)	726(29.9%)
	Asp (D)	146(6.0%)
	Cys (C)	51(2.1%)
	Gln (Q)	54(2.2%)
	Glu (E)	100(4.1%)
	Gly (G)	66(2.7%)
	His (H)	70(2.9%)
	Ile (I)	165(6.8%)
	Leu (L)	123(5.1%)
	Lys (K)	226(9.3%)
	Met (M)	73(3.0%)
	Phe (F)	66(2.7%)
	Pro (P)	38(1.6%)
	Ser (S)	144(5.9%)
	Thr (T)	94(3.9%)
	Trp (W)	5(0.2%)
	Tyr (Y)	124(5.1%)
	Val (V)	71(2.9%)

4.9.2.2 Secondary Structure Analysis of *Pfap2g* Protein

The prediction of the secondary structure using RaptorX-Property shows the 3-class secondary structure (SS3), 8-class secondary structure (SS8), solvent accessibility (ACC), and disorder regions (DISO) of *Pfap2g*. The prediction of SS3 of *Pfap2g* reveals that it mainly comprises of 77% coils with α -helix and β -sheet forming 13% and 8%, respectively. The prediction of SS8 structure indicates that coil, α -helix, and extended strand in the beta ladder are major secondary structures. Regarding ACC, prediction shows that exposed region (>40%) forms the major part (63%), while buried (<10%) region and medium (10-40%) region forms 19% and 17% of secondary structure respectively. The prediction of DISO depicts that half of the protein (50%) is disordered while the remaining 50% is ordered.

4.9.2.3 Tertiary Structure Prediction of *Pfap2g*

Functional prediction of *Pfap2g* indicates that it has neither predicted protein family membership nor homologous superfamilies. As a protein sequence with 2432 amino acid residues, *Pfap2g* has an apetala2 domain (AP2/ERF) made of 60 amino acid residues (2,160-2210) (Figure 4.13). This domain, named ethylene-responsive element (ERE), interacts with the GCC-box and regulate transcription.

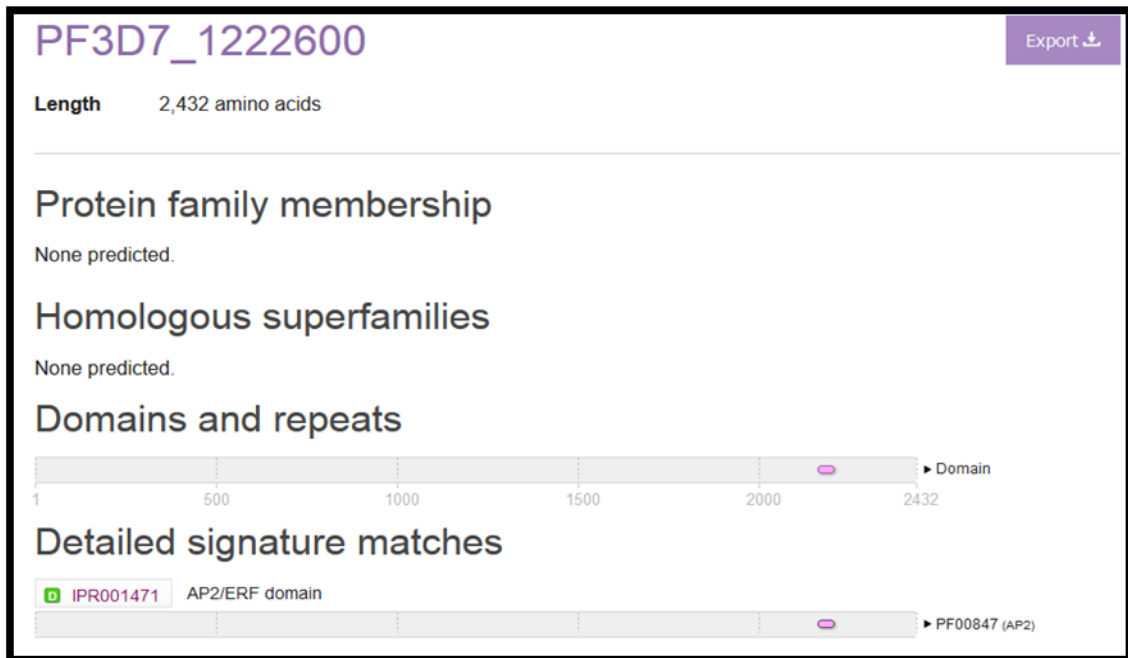


Figure 4.13: Prediction of *Pfap2g* using InterPro showing *apetalla2* domain without any known protein family and homologous superfamilies

Other features of *Pfap2g* indicate that it has cytoplasmic domain at 236-254 amino acid residues (Figure 4.14). It has four coils at positions 502-522, 1329-1358, 1656-1676, and 2242-2262. The non-cytoplasmic domain forms the most of the protein because it covers 1-215 and 275-2432 positions of the protein sequence. Transmembrane region covers 216-235 and 255-274 protein sequences. Disorder prediction using indicates that 424-552, 722-758, 880-915, 1330-1375, 1425-1445, 1509-1627, 1668-1748, 1898-2008, and 2249-2292 are regions with orders have no stable structures or conformations. Overall, gene ontology shows that *Pfap2g* regulates the process of transcription by acting as a DNA-binding transcription factor at a molecular level.

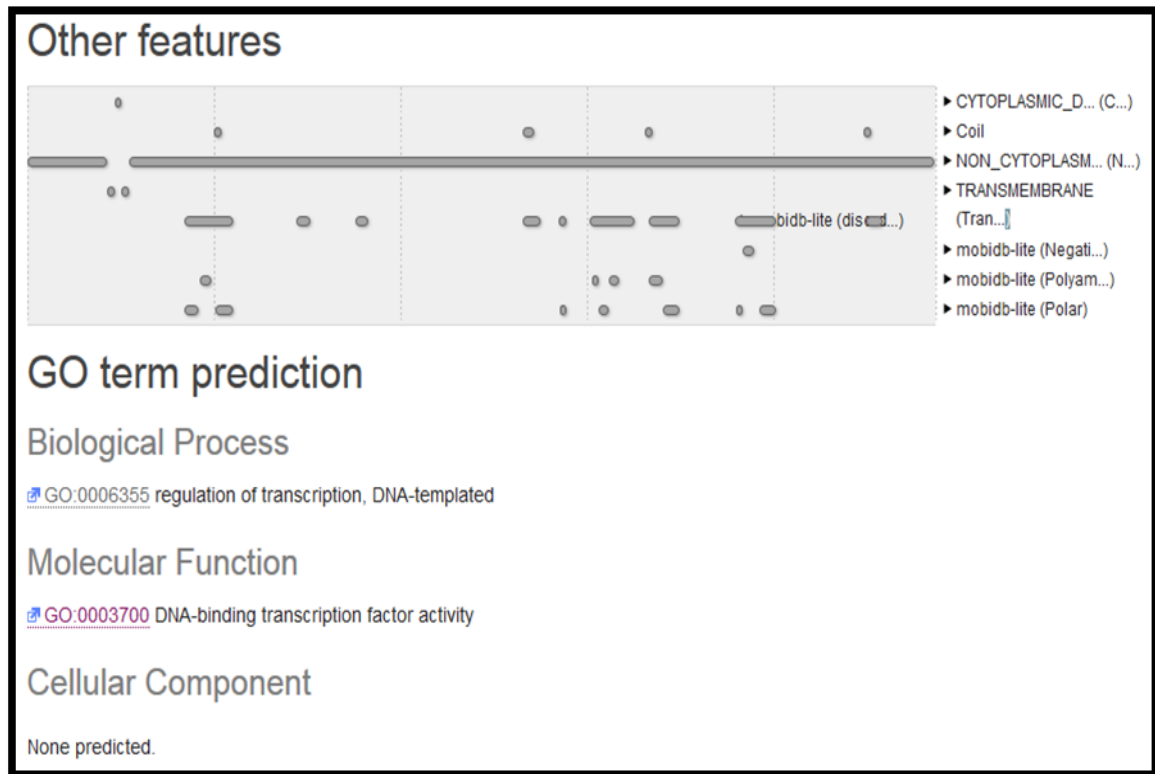


Figure 4.14: Prediction outcomes of *Pfap2g* using InterPro identified it as a DNA-binding protein that regulates transcription with specific gene ontology terms

The prediction of the protein structure using I-TASSER provided additional information about *Pfap2g* (Appendix VI). I-TASSER also showed PDB hit, local identity, global identity, coverage, and normalized Z score of the ten threading templates of the modelled 3D-structure of *Pfap2g* (Table 4.20). The leading templates are CRISPR-associated endonuclease Cas12a (6i1kA), yeast fatty acid synthase (2pffA), apoptosome (1vt4), erythrocyte membrane protein (2yk0A), toxin B (6ar6A), piezo-type mechanosensitive ion channel component 1 (6ar6A), 60S ribosomal protein L2-A (5jcsS), PF14_0633 protein (3igm), ywmB protein (2fpnA), ywmB protein (1n7dA), RNA-dependent RNA polymerase (3ja4A), (3javA), and nvTRPM2 channel (6co7A).

Table 4.20: The leading 10 threading templates of *Pfap2g* (1-200 and 1201-2432)

Rank ^a	PDB Hit ^b	Identity 1 ^c (%)	Identity 2 ^d (%)	Coverage ^e (%)	Normalized Z-Score ^f
1	6i1kA	0.15	0.19	0.95	1.00
2	2pffA	0.07	0.06	0.96	0.78
3	6i1kA	0.09	0.19	0.92	0.87
4	2pffA	0.14	0.03	0.20	1.61
5	2pffA	0.14	0.03	0.20	1.43
6	6i1kA	0.11	0.19	0.94	0.28
7	1vt4	0.17	0.09	0.03	2.04
8	2yk0A	0.12	0.12	0.41	0.92
9	6i1kA	0.16	0.19	0.90	0.63
10	6i1kA	0.09	0.16	0.78	1.99
1	6ar6A	0.09	0.18	1.00	2.74
2	6b3rA	0.07	0.14	0.25	0.71
3	5jcsS	0.06	0.16	1.00	4.06
4	3igm	0.17	0.01	0.05	2.11
5	2fpnA	0.08	0.02	0.11	0.58
6	1n7dA	0.07	0.08	0.48	2.23
7	3ja4A	0.09	0.13	0.85	1.92
8	3javA	0.07	0.12	0.78	4.28
9	3imB	0.21	0.01	0.03	0.77
10	6co7A	0.06	0.13	0.84	1.64

Ranked numbers of templates (a), PDB hit ID (b), percent local identity (c), percent global identity (d), percent coverage of *Pfap2g* (e), and Z-score greater than 0.5 rated as good threading alignment

Further analysis identified structural analogs of *Pfap2g*, 1-1200 and 1201-2432 amino acid sequences, as identified in PDB and their respective threading-model scores (TM-score), root-mean-square-deviation (RMSD), local identity, and coverage (Table 4.21). The leading structural analogs for the first sequence are fatty acid synthase (2pffA), phosphatidylinositol 4-kinase III alpha (6bq1A), DNA helicase I (5n8oA), apoptosome (1vt41), CRISPR-associated endonuclease Cpf1 (5b43A), glutamate receptor ionotropic, (6iraA), RNA polymerase sigma-A holoenzyme (6fedC), Pol2CORE-M644G (6fwkA), and CRISPR-associated endonuclease Cas12a (6i1kA). The leading structural analogs for the second sequence are toxin B (6ar6A), toxin A (4r04A), chromatin modification-related protein EAF1 (5y81A), tripartite Tc toxin (1vw1A), human mediator subunit

MED23 (6h02A), glycogen debranching enzyme (5d06A), spliceosome (3jb9A1), human spliceosome (5xjcA), and separin (5u1sA).

Table 4.21: The Leading 10 Structural Analogs of *Pfap2g* Identified in PDB

Rank ^a	PDB Hit ^b	TM-Score ^c	RMSD ^d	Identity ^e	Coverage ^f
1	2pffA	0.262	9.90	0.019	0.422
2	6bq1A	0.246	9.55	0.029	0.388
3	6ifoA	0.246	9.71	0.034	0.385
4	5n8oA	0.240	9.24	0.022	0.363
5	1vt41	0.240	9.87	0.008	0.388
6	5b43A	0.229	9.77	0.036	0.362
7	6iraA	0.220	7.98	0.030	0.306
8	6fedC	0.217	9.85	0.022	0.347
9	6fwkA	0.208	9.70	0.021	0.329
10	6i1kA	0.201	9.60	0.030	0.312
1	6ar6A	0.989	1.40	0.083	1.000
2	4r04A	0.858	4.90	0.075	0.980
3	6c0bA	0.359	3.17	0.078	0.383
4	5y81A	0.295	10.30	0.038	0.481
5	1vw1A	0.292	9.72	0.028	0.455
6	6h02A	0.289	9.80	0.035	0.458
7	5d06A	0.287	9.62	0.028	0.444
8	3jb9A1	0.280	9.97	0.027	0.445
9	5xjcA	0.279	9.87	0.040	0.439
10	5u1sA	0.279	9.63	0.041	0.433

Ranked numbers of analogs (a), PDB hit ID (b), threading template score greater than 0.5 rated as good (c), RMSD less than 5 regarded as reliable (d), percent identity (d) and coverage of *Pfap2g* (f)

I-TASSER results depict global and local accuracy estimations of the most accurate models of *Pfap2g* as predicted by I-TASSER (Figures 4.15 and 4.16). The global accuracy estimation indicates that the predicted structure has a c-score of -0.97 and 0.09 for amino acid residues 1-1200 and 1201-2432 respectively. Moreover, 1-1200 amino acid residues have TM-score of 0.59 ± 14 and RMSD of $11.7 \pm 4.5 \text{ \AA}$, while 1201-2432 amino acid residues have TM-score of 0.73 ± 11 and RMSD of $9.2 \pm 4.6 \text{ \AA}$. The estimated local accuracy of the model shows that the first 400 amino acid residues deviated from

the native structure, whereas subsequent sequences are accurate with an estimated mean distance of about 4Å from the native structure.

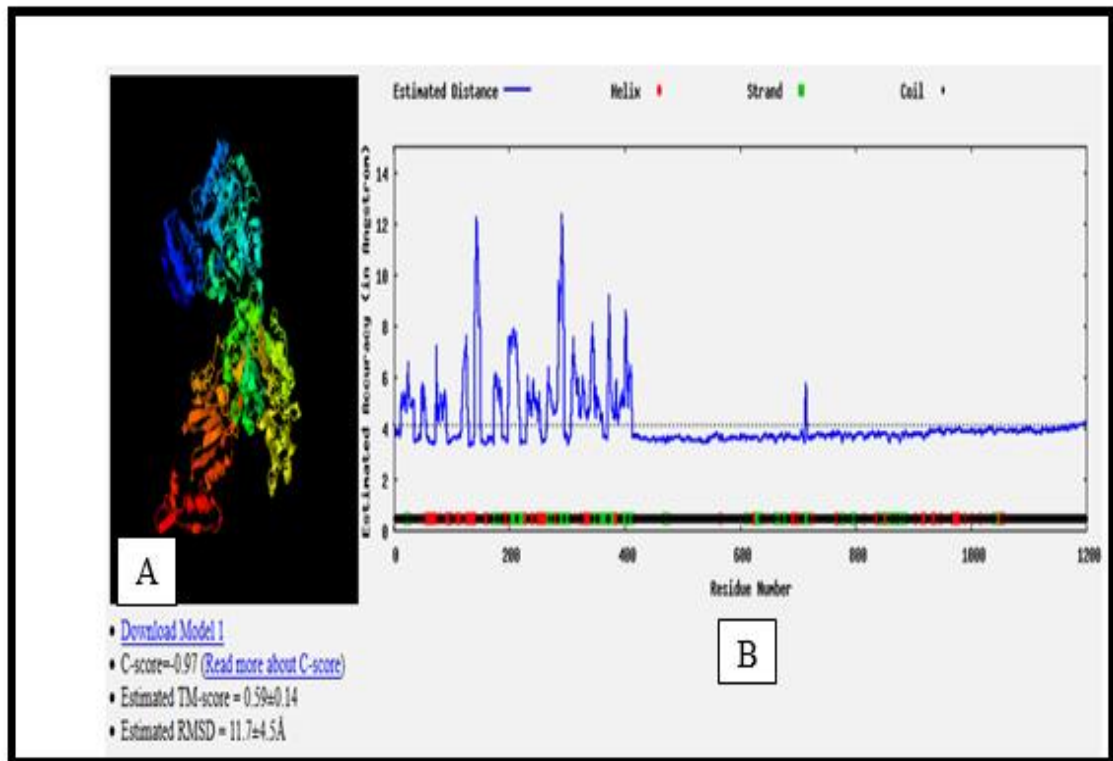


Figure 4.15: Local and global accuracy of the predicted model of *Pfab2g* (1-1200 residues). A: Showing global accuracy estimation of the *Pfgdv1* model in C-score (-0.97), TM-score (0.59±0.14), and RMSD (11.7±4.5 Å) and B: Showing the local accuracy estimation of the *Pfab2g* model in Angstroms with average value of about 4 Å with variations in estimated distance of each residue number and trends of helix (red), strand (green), and coil (black) structures

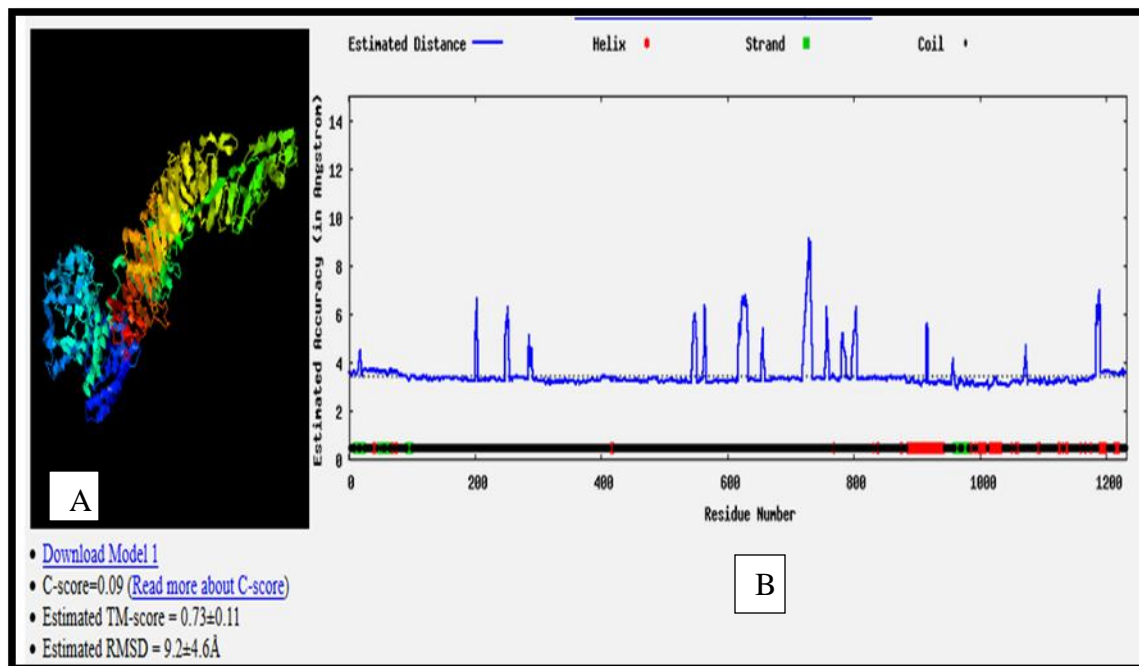


Figure 4.16: Local and global accuracy of the predicted model of *Pfgdv1* (1201-2432 residues). A: Showing global accuracy estimation of the *Pfgdv1* model in C-score (-0.09), TM-score (0.73±0.11), and RMSD (9.2±4.6 Å) and B: Showing the local accuracy estimation of the *Pfgdv1* model in Angstroms with average value of about 4 Å with variations in estimated distance of each residue number and trends of helix (red), strand (green), and coil (black) structures

4.10 Protein-Ligand Docking

4.10.1 Protein-Ligand Docking of *Pfgdv1*

I-TASSER and COACH were used to predict ligands and binding sites of *Pfgdv1* based on BioLip database (Appendix VII). I-TASSER gives three peptide (III), N-octadecane, calcium ion (Ca^{2+}), and magnesium ion (Mg^{2+}), while COACH generates ZCT, pyridoxal phosphate, N-acetyl-D-Glucosamine (NAG), and chlorophyll A, Manganese Ion (Mn^{2+}), Xenon (XE), Nitritriacetic Acid (NTA), and Lanthanum Ion (La^{3+}) as additional ligands (Table 4.22). The confidence scores for ligands ranged from 0.09 to 0.02 with multiple binding sites.

Table 4.22: Ligands and their respective binding sites

Server ^a	Ligand Name ^b	PDB Hit ^c	C-Score ^d	Cluster Size ^c	Residue Numbers ^d
I-TASSER	Peptide (III)	3e4aA	0.09	4	G15, M27, Y32, L33
	Peptide (III)	2wk3A	0.05	2	H62, K65, D121, L122, L123, S124, F141, K180, Y227
	N-Octadecane	5d91A	0.05	2	F512, Q513
	Calcium Ion (Ca ²⁺)	4wb2A	0.02	1	K351, D355
	Magnesium ion (Mg ²⁺)	3d6aA	0.02	1	S344, Y345
COACH	ZCT	3b6aD	0.07	4	L546, K547
	Pyridoxal Phosphate (PLP)	4a0gD	0.04	2	I412, S413
	N-Acetyl-D-Glucosamine (NAG)	2qmjA	0.04	2	K5, N16, S22, A23, F173
	Chlorophyll A (CLA)	3bz1H	0.04	2	Y585, I588
	Magnesium Ion (Mg ²⁺)	1jgtB	0.04	2	D301, D410
	Manganese Ion (Mn ²⁺)	1khwB	0.04	2	Y284, D408
	Xenon (XE)	2z8yM	0.02	1	Y129, L190, L213, N229, A230, I233
	Nitrilotriacetic Acid (NTA)	3ufkA	0.02	1	C8, K10, N165, K166
Lanthanum Ion (La ³⁺)	1djgB	0.02	1	I494, D496, N526	

Servers used to dock *Pfgdv1* protein are I-TASSER and COACH (a), ligand names from BioLip database (b), PDB hit ID (c), C-score range from 0 to 1, the number of templates in a cluster (d), and amino acid residue numbers (e)

4.10.2 Protein-Ligand Docking of *Pfap2g*

Ligands and binding sites were identified using I-TASSER, which rely on BioLip database (Appendix VIII). I-TASSER generated four ligands for 1-1200 protein sequence and five ligands for 1201-2432 protein sequence. Predicted ligands are Flavin Mononucleotide (FMN), Manganese (Mn^{2+}), Nicotinamide Adenine Dinucleotide Phosphate (NAP), Magnesium Ion (Mg^{2+}), Zinc Ion (Zn^{2+}), Cyclic AMP (CMP), Zinc (Zn^{2+}), Calcium ion (Ca^{2+}), and Magnesium Ion (Mg^{2+}) (Table 4.23). The confidence scores range from 0.15 to 0.04 with multiple binding sites.

Table 4.23: Ligands and their respective binding sites

Server	Ligand Name	PDB Hit	C-Score	Cluster Size	Residue Numbers
I-TASSER 1-1200 Residues	Flavin Mononucleotide (FMN)	2uv8G	0.15	8	495-499, 548, 550, 579, 603, 629, 632, 633, 663, 693, 694, 697, 937, 942
	Manganese (Mn^{2+})	3bsnA	0.04	2	498, 520, 523, 550, 555, 634, 635, 741, 824, 893, 895, 896, 914, 915, 917
	Nicotinamide Adenine Dinucleotide Phosphate (NAP)	2uvdcl	0.04	2	41, 189
	Magnesium Ion (Mg^{2+})	1bofA	0.02	1	389, 390
I-TASSER 1201-2432	Zinc Ion (Zn^{2+})	3jpyA	0.03	4	389, 414
	Cyclic AMP (CMP)	3r6sA	0.03	2	410, 407
	Zinc (Zn^{2+})	5ij5A	0.02	1	400, 407
	Calcium ion (Ca^{2+})	2qejC	0.02	1	1188, 1193
	Magnesium Ion (Mg^{2+})	4gbfA	0.02	1	517, 522

Server used to dock *Pfgdv1* protein is I-TASSER (a), ligand names from BioLip database (b), PDB hit ID (c), C-score range from 0 to 1, the number of templates in a cluster (d), and amino acid residue numbers (e)

CHAPTER FIVE

DISCUSSION

5.1 Description of Samples

Gametocytogenesis is a critical step in the lifecycle of *P. falciparum* because it determines the rate of transmission, and it is under the epigenetic control of *Pfgdv1* and *Pfap2g* genes. To achieve the objectives of this study, molecular procedures and bioinformatics analyses were employed. In sample collection, this study obtained blood samples (n = 30) from patients with high levels of parasitaemia as determined by microscopy using the assumed white blood cell (WBC) count of 8000/ul (Adu-Gyasi *et al.*, 2015). Comparative analysis revealed that parasitaemia levels did not vary statistically significantly based on gender, age group, and study sites. The isolation of genomic DNA, amplification target genes, and sequencing of PCR products generated valid primary sequence data of *Pfgdv1* and *Pfap2g* genes from field isolates of *P. falciparum*, while PlasmoDB (<https://plasmodb.org>) provided secondary data for the study.

5.2 Single Nucleotide Polymorphisms

Multiple sequence alignment of primary and secondary data established the existence of SNPs in *Pfgdv1* and *Pfap2g* based on the threshold of 1% of the population (Karki *et al.*, 2015). *Pfgdv1* has five SNPs comprising of four non-synonymous mutations, namely C650A (P217H), G1193A (R398Q), C1249A (H417N), and T1491A (D497E) and a synonymous substitution of A1542T (S514S) with variation in the proportion of minor alleles from 10% to 39% in secondary sequences and 10% to 37% in primary isolates. The presence of SNPs in high frequencies supports the earlier findings that *Pfgdv1* is the molecular marker of gametocytogenesis and transmission level that exhibits geographical divergence (Mobegi *et al.*, 2014; Duffy *et al.*, 2018). Overall, the analysis of SNPs indicates that mutations of C650A, C1249A, and A1542T formed major alleles, whereas those of G1193A and T1491A constituted minor alleles. In a genome-wide

analysis study, this C650A (P217H) is one of the five SNPs that accumulate within the 15kb region of chromosome 9 with the highest value of fixation index (Mobegi *et al.*, 2014). In their study of determining the rate of gametocyte conversion, Usui *et al.* (2019) noted that P217H increases the effectiveness of gametocyte differentiation in *P. falciparum*.

MSA of *Pfap2g* showed the existence of 12 polymorphic sites of SNPs, which comprise 11 non-synonymous and 1 synonymous with variation in the proportion of minor alleles ranging from 7% to 50%. Additionally, these SNPs emanates from 7 transition mutations and 5 transversion mutations on A62G(K21R), T817C(Y273H), T2034C (C678C), A2951C(K984T) G3539A(G1180E), G4753A (G1585R), G4762C (E1588Q), A4970T(N1657I), G5638A(D1880N), G5767A(E1923K), A5922T(K1974N), and C5955A(N1985K) positions on the coding sequence. In nature, transitions are more common than transversions owing to minimal structural changes required (Robert & Pelletier, 2018). MSA indicates that the most polymorphic region of *Pfap2g* ranges from 5630 to 6000 nucleotide positions. The analysis of primary sequences shows that there is an insertion of codon (GAG) in all sequences, which codes for glutamate (-1920E). Although SNPs explain most variations in genetic diversity, insertions and deletions (indels) constitute other markers that account for significant variations in genomic diversity in *P. falciparum* (Miles *et al.*, 2016; Duffy *et al.*, 2018). This insertion is a non-conservative mutation because it introduces glutamate, which is a negatively charged amino acid into *Pfap2g*.

Comparison of primary and secondary sequences shows that both *Pfgdv1* and *Pfap2g* have common non-synonymous SNPs and relatively the same distributions of minor allele frequencies. Both primary and secondary sequences of *Pfgdv1* have a synonymous and four nsSNPs. Similarly, *Pfap2g* has a synonymous and 11 nsSNPs in both primary and secondary sequences. The similarity in the distribution of synonymous and non-synonymous SNPs implies that these two genes have conserved sequences that enable them to regulate gametocytogenesis effectively. Comparative analysis shows that the

distribution of synonymous and nsSNPs were random since they existed in isolates from different regions. Shen *et al.*, (2017) hold that conserved sequences are appropriate target genes for drug design and vaccine development because they have definite protein structures. However, the primary sequences of *Pfap2g* have an insertion of GAG codon, which codes for glutamate. Miles *et al.* (2018) explain that insertions comprise adaptive drivers of evolution that results in changes in protein structure and function. In this case, the existence of an insertion suggest that local isolates of *Plasmodium falciparum* have an adaptive insertion in their *Pfap2g*.

In selection analysis, the positive values of Tajima's D obtained from both primary and secondary data imply that SNPs on *Pfgdv1* and *Pfap2g* exhibit balancing selection. Tajima's D values of SNPs on *Pfgdv1* and *Pfap2g* were not only positive for both primary isolates and secondary isolates but also statistically significant ($p < 0.05$). Since the observed variations were higher than the expected variations, it shows that both *Pfgdv1* and *Pfap2g* evolve through the process of balancing selection (Tajima, 1989). According to Amambua-Ngwa (2012), genes that exhibit balancing selection in their evolution are target candidates for elucidating immune mechanisms and vaccine development. Balancing selection ensures that SNPs on *Pfgdv1* and *Pfap2g* remain constant and genetic drift does not occur in the population. SLAC analysis pointed out that P217H is the only SNP under the strongest positive selection in *Pfgdv1*. In the genome-wide analysis, Mobegi *et al.* (2014) established that P217H has a fixation index of 0.3, which differentiates between Gambian and Guinean isolates. These findings suggest that the positive selection of P217H discriminates *P. falciparum* isolates in high and low transmission regions. SLAC analysis identified K984T is under the strongest positive selection followed by K21R. Genes associated with strong selection pressure are signatures for transmission, drug resistance, and immune evasion (Shen *et al.*, 2017). Thus, further analysis of these nsSNPs would indicate their role in the transmission rate of malaria from humans to mosquitoes.

The analysis of the effects of nsSNPs on the stability of the predicted protein using STRUM revealed that both *Pfgdv1* and *Pfap2g* have relatively conserved structures. Out of four nsSNP, three of them have a considerable effect on the stability of the predicted protein structure of *Pfgdv1*. A nsSNP causes energy changes in the fold and unfolded conformations of proteins, which have destabilizing or stabilizing effects on protein structure (Quan *et al.*, 2016). D497E has no considerable effect on the thermal stability of the protein structure for its Gibbs free-energy gap is close to zero. R398Q and H417N are two nsSPNs that stabilize the protein structure of *Pfgdv1*. Nevertheless, P217H is a snSNP that destabilizes the protein structure because it has a substantial negative free-energy gap ($\Delta\Delta G = -2.22$). In addition, the SNP is a non-synonymous mutation that causes a non-conservative change in amino acid from polar proline (P) to positively charged histidine (H). The predicted destabilization of the protein structure supports earlier findings that this mutation increases the induction of sexual differentiation and determines transmission rates (Mobegi *et al.*, 2014; Usui *et al.*, 2019). In the analysis of *Pfap2g*, the analysis of energy changes indicates that 11 snSNPs increase the stability of protein. In this case, K21R, Y273H, K954T, G1180E, G1585R, N1657I, and D1880N have free-energy gaps that are greater than two, while E1588Q, K1974N, N1984K, and N1923K have free energy gaps less than two. Therefore, the analysis of the effects of snSNPs indicates that they stabilize protein structure and they do not occur in the ERE/ap2 domain, suggesting that *Pfap2g* protein is conserved despite having numerous mutations.

5.3 Protein Structure Prediction

The prediction of physicochemical properties shows that *Pfgdv1* has 599 amino acid residues, molecular weight of 71.964 kDa, and theoretical isoelectric point of 8.84. These properties reveal that *Pfgdv1* is an average protein in size with a basic zwitterionic molecule. The higher proportions of positively charged amino acid residues (arginine and lysine) than negatively charged ones (aspartate and glutamate) means that the protein forms ionic interactions and hydrogen bonding on its surfaces (Sokalingam *et*

al., 2012). Moreover, as the distribution of amino acids shows that *Pfgdv1* is an asparagine-rich protein, it explains its interaction with cations, such as Mg^{2+} , Ca^{2+} , La^{3+} , and Mn^{2+} , via ionic and hydrogen bond molecular forces in docking results. In a cloning and expression experiment, the size of *Pfgdv1* was estimated to be 67kDa in SDS-PAGE (Su *et al.*, 2016). The prediction of the 3-state secondary structure indicated that coil structure (49%) predominates followed by α -helix (33%), and beta-sheet (16%). The dominance of coils implies that *Pfgdv1* has a definite structure because coiled-coils act as molecular spacers and pillars that situate active sites on α -helices and beta-sheets in their optimal conformations (Truebestein & Leonard, 2016). The analysis of solvent accessibility surface areas shows that most of the protein (40%) is exposed, while about equal proportions remain buried (31%) and moderately exposed (27%). The higher the proportion of surface area exposed means that the protein is hydrophilic as indicated by the hydrophilic index (Wang *et al.*, 2016). The fluctuation of the B-factor profile shows that the predicted protein is unstable (Yang, Wang, *et al.*, 2016). Given that a significant proportion of the protein is ordered (95%), its structure is predictable and stable (Oldfield & Dunker, 2014). InterPro failed to classify *Pfgdv1* because it lacks established or predicted domains, family, and functional sites (European Bioinformatics Institute, 2018). However, InterPro predicts the presence of coil between 340 and 360 residues of the protein sequence.

The prediction of the tertiary protein structure using RaptorX reveals that *Pfgdv1* has two domains covering 1-143 and 144-599 amino acid residues. These domains formed the basis of predicting the tertiary structure of the protein. Protein structure prediction using I-TASSER generated a three-dimensional structure based on leading threading templates and structural analogs. The threading templates have good alignment (Z-score >1), significant structural similarity (TM-score > 0.5), and moderate structural resolution (RMSD = 2-5 Å) (Zhang, 2009; Roy *et al.*, 2010; Yang & Zhang, 2015a). These findings imply that the best templates were used in the prediction of the protein. Although I-TASSER generated five models, the most accurate model had a c-score of -

2.95, TM-score of 0.38 ± 13 , and RMSD of 15.1 ± 3.5 Å. The predicted model had a lower quality because its C-score is less than -1.5, a weak structural correlation that is less than 0.5, and low resolution that is greater than 5 Å (Zhang, 2009; Roy *et al.*, 2010; Yang & Zhang, 2015b). The estimated local accuracy of the model shows that most regions are accurate with an estimated mean distance of 6 Å from the native structure. However, the local accuracy of the model has a moderate resolution as indicated by residue-specific quality (RSQ) (Yang, Wang, *et al.*, 2016).

The prediction of the primary structure of *Pfap2g* shows that it has 2432 amino acid residues, a molecular weight of 284.064kDa, and theoretical isoelectric point of 8.72. These properties are critical in Western blotting because it shows that *Pfap2g* is a large molecular protein with a basic isoelectric pH (Bass *et al.*, 2017). The protein has a net positive charge since positively charged residues (arginine and lysine) are more than negatively charged residues (aspartate and glutamate). Under physiological conditions, arginine and lysine are basic amino acids situated on protein surfaces to form ionic interactions and hydrogen bonds (Sokalingam *et al.*, 2012). As asparagine-rich protein, *Pfap2g* forms hydrogen bonding and ionic interactions with cations, such as Zn^{2+} , Mn^{2+} , Ca^{2+} , and Mg^{2+} , as shown in docking results. *Pfap2g* has a low aliphatic index (56.05), which reduces thermal stability and hydrophobic property (1.382). Since *Pfap2g* is rich in asparagine (30%), they have functional importance because asparagine repeats act as tRNA sponges and promote aggregation of proteins (Muralidharan & Goldberg, 2013). The half-life of 30 hours and the instability index of 39.93 shows that *Pfap2g* conformation is stable protein under both *in-vitro* and *in-vivo* conditions.

The prediction of the 3-state secondary structure using RaptorX-Property indicates that of *Pfap2g* mainly comprises of 77% coils with α -helix and β -sheet forming 13% and 8% respectively. The high proportion of coiled-coils in the protein matches the function of *Pfap2g* in gene regulation because it is a transcription factor (Truebestein & Leonard, 2016). Solvent accessibility shows that *Pfap2g* is an exposed protein because the majority part (63%) has more than 40% exposure proportion. The high exposure level

means that the protein has a hydrophilic property that enables it to bind on DNA and regulation transcription and gametocytogenesis. Additionally, being a transmembrane protein underscores the role of *Pgap2g* in as a DNA-binding protein and transcriptional regulator of gametocytogenesis. The protein is hard to predict because 50% of its structure is disordered.

The prediction of domains using InterPro reveals that *Pfap2g* has no defined protein family membership or belong to homologous superfamilies. However, *Pfap2g* has an apetela2 domain (AP2/ERF) made of 60 amino acid residues (2,160-2210). This domain, named ethylene-responsive element (ERE), interacts with the GCC-box and regulate transcription (Modrzynska *et al.*, 2017). Z-scores of I-TASSER show that the threading templates had good alignment because more than half were greater than 1, structural analogs had weak correlation because most were less than 0.5, poor resolution because RMSD was about 10 Å (Roy *et al.*, 2010; Yang & Zhang, 2015a, 2015b).

The global accuracy estimation indicates that the predicted structure of *Pfap2g* has a c-score of -0.97 and 0.09, TM-score of 0.59 ± 14 and TM-score of 0.73 ± 11 , and RMSD of $11.7 \pm 4.5 \text{Å}$ and RMSD of $9.2 \pm 4.6 \text{Å}$ for amino acid residues 1-1200 and 1201-2432 respectively. Since C-score is higher than -1.5, TM-score is greater than 0.5, and an estimated mean distance of about 4Å from the native structure, it implies the predicted model is quality and has a significant level of correlation. (Yang, Wang, *et al.*, 2016). Therefore, poor alignment of templates, RMSD of about 10 Å, a high proportion of disordered region (50%) and low quality of local accuracy are factors that affect accurate prediction of the *Pfap2g* protein.

5.3 Protein-Ligand Docking

Protein-ligand docking identified numerous ligands in BioLip database with over 100,000-curated proteins that interact with *Pfgdv1* with reliable degree of prediction (Yang *et al.*, 2013a, Yang *et al.*, 2013b). Consideration of c-score shows that peptide (III) (3e4aA, 2wk3A), N-octadecane (5d91A), and ZCT (3b6aD) are the most reliable

ligands of *Pfgdv1* with values of 0.09, 0.05, 0.05, and 0.07, respectively. According to PDB hit, both peptide (III) ligands constitute an inhibitor of insulin-degrading enzyme (IDE) in humans (3e4aA) and IDE-amyloid-beta (1-42) (2wk3A) (Guo *et al.*, 2010; Leissring *et al.*, 2010; RCSB PDB, 2019). N-octadecane is a ligand of phosphatidylinositolphosphate (PIP) synthase that is responsible for linking proteins and carbohydrates to outer membranes of cells (Clarke *et al.*, 2015), whereas ZCT is actinorhodin antibiotic produced by *Streptomyces coelicolor* (Willems *et al.*, 2008). Moreover, other ligands that are unreliable are calcium ion (Ca^{2+}), and magnesium ion (Mg^{2+}), pyridoxal phosphate, N-acetyl-D-Glucosamine (NAG), chlorophyll A, Manganese Ion (Mn^{2+}), Xenon (XE), Nitritotriacetic Acid (NTA), and Lanthanum ion (La^{3+}) because they have low c-scores with values ranging from 0.04 to 0.02. The commonality of cations as potential ligands being scores by two predictors stems from the primary structure of these proteins where aspartate, glutamate, tyrosine, and asparagine confer negative charges to the predicted structures.

Protein-ligand docking identified 2uv8G (Flavin mononucleotide) and 3jpyA (Zn^{2+}) as two reliable ligands for 1-1200 and 1201-2432 protein residues, respectively. Flavin mononucleotide had a c-score of 0.15, while Zn^{2+} had a c-score of 0.07. In a range of 0-1, a higher c-score shows that ligands are reliably predicted (Wu *et al.*, 2018; Yang *et al.*, 2013a, 2013b). Flavin mononucleotide is a coenzyme derived from vitamin B2 (riboflavin), which form part of flavoproteins involved in redox metabolism and gene expression via FMN riboswitch (García-Angulo, 2017). The existence of a positive correlation between FMN and parasitaemia, as well as the occurrence of deficiency of FMN in malaria patients suggest the role of FMN in pathophysiology of malaria (Traunmüller *et al.*, 2003). Zn^{2+} is another significant ligand with four clusters that bind to 1201-2432 protein sequence of *Pfap2g*. Blood stage *P. falciparum* parasites accumulate zinc ions in erythrocytes for their growth and development in asexual cycle (Marvin *et al.*, 2012). The prediction of zinc ion as a ligand of *Pfap2g* shows that it plays a role in gametocytogenesis.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

In this study, the genetic characterization of *Pfgdv1* and *Pfap2g* genes reported the existence of SNPs comprising non-synonymous substitutions and synonymous substitutions. The dominance of non-synonymous SNPs implies that *Pfgdv1* and *Pfap2g* gene exhibit a small degree of diversity among *Plasmodium* species. Furthermore, these non-synonymous SNPs have non-conservative changes on amino acids, which have marked influence on protein structure and function, and thus, suggesting a possible effect on the process of gametocytogenesis. Tajima's D indicated that both *Pfgdv1* and *Pfap2g* exhibit balancing selection, whereas SLAC shows the existence of codon sites with strong positive selection. The balancing selection explains the constant proportion of nsSNPs in both local and global isolates of *P. falciparum*. Thermodynamic analysis of the changes in Gibbs free energy demonstrated that nsSNPs have more of stabilizing effect than destabilizing effect on *Pfgdv1* and *Pfap2g* proteins, indicating a net conservative effect.

The prediction of the structure revealed that both *Pfgdv1* and *Pfap2g* have low c-score values, average TM-scores, and poor resolution, which affects the quality and validity of the predicted models. The predicted structure of *Pfgdv1* has 599 amino acid residues (1800 nucleotides), a molecular weight of 71.964 kDa, and ordered structure (95%) with c-score of -2.95, TM-score of 0.38 ± 13 , and RMSD of 15.1 ± 3.5 Å. The predicted model of *Pfap2g* has 2432 amino acid residues (7299 nucleotides), a molecular weight of 284.064 kDa, ap2 domain (AP2/ERF), and 50% disordered structure. Moreover, the predicted model has a c-score of -0.97 and 0.09, TM-score of 0.59 ± 14 and TM-score of 0.73 ± 11 , and RMSD of 11.7 ± 4.5 Å and RMSD of 9.2 ± 4.6 Å for amino acid residues 1-1200 and 1201-2432, respectively.

In docking, results show that peptide (III) (3e4aA, 2wk3A), N-octadecane (5d91A), and ZCT (3b6aD) are reliable ligands that bind to *Pfgdv1*. Moreover, docking results show

that Flavin mononucleotide (2uv8G) and zinc ion (3jpyA) are ligands that bind to 1-1200 and 1201-2432 protein residues of *Pfgap2g*, respectively. Peptide (III), N-octadecane, ZCT, flavin mononucleotide, and zinc ions are potential ligands that modulate the functions of *Pfgdv1* and *Pfap2g* in inducing gametocytogenesis. Thus, these findings form the basis of drug design studies to determine the protein structure, functional analysis of SNPs, and evaluation of ligands experimentally.

6.2 Recommendations

- Since these findings indicated that *Pfgdv1* and *Pfgap2g* are relatively conserved proteins, the study suggests the experimental determination of protein structure to target it in drug design studies.
- Functional analysis of various SNPs should be done to establish the role of non-synonymous SNPs of both *Pfgdv1* and *Pfap2g* on gametocytogenesis and transmission of malaria in both endemic and epidemic regions.
- Experimental evaluation of the identified ligands should be performed to determine their effects on gametocytogenesis.

REFERENCES

- Adu-Gyasi, D., Asante, K. P., Newton, S., Amoako, S., Dosoo, D., Ankrah, L., ... Owusu-Agyei, S. (2015). Malaria parasite density estimated with white blood cells count reference value agrees with density estimated with absolute in children less than 5 years in central Ghana. *Malaria Research and Treatment*, 2015(923674), 1-8. <https://doi.org/10.1155/2015/923674>
- Ajibaye, O., Osuntoki, A. A., Balogun, E. O., Olukosi, Y. A., Iwalokun, B. A., Oyebola, K. M., ... & Amambua-Ngwa, A. (2020). Genetic polymorphisms in malaria vaccine candidate *Plasmodium falciparum* reticulocyte-binding protein homologue-5 among populations in Lagos, Nigeria. *Malaria journal*, 19(1), 1-12. <https://doi.org/10.1186/s12936-019-3096-0>
- Amambua-Ngwa, A., Tetteh, K. K. A., Manske, M., Gomez-Escobar, N., Stewart, L. B., Deerhake, M. E., Cheeseman, I. H., Newbold, C. I., Holder, A. A., Knuepfer, E., Janha, O., Jallow, M., Campino, S., MacInnis, B., Kwiatkowski, D. P., & Conway, D. J. (2012). Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genetics*, 8(11), 1-14. <https://doi.org/10.1371/journal.pgen.1002992>
- Arakawa, T., Komesu, A., Otsuki, H., Sattabongkot, J., Udomsangpetch, R., Matsumoto, Y., ... Tsuboi, T. (2005). Nasal immunization with a malaria transmission-blocking vaccine candidate, Pfs25, induces complete protective immunity in mice against field isolates of *Plasmodium falciparum*. *Infection and Immunity*, 73(11), 7375-7380. <https://doi.org/10.1128/IAI.73.11.7375-7380.2005>
- Arama, C., & Troye-Blomberg, M. (2014). The path of malaria vaccine development: Challenges and perspectives. *Journal of Internal Medicine*, 275(5), 456-466. <https://doi.org/10.1111/joim.12223>

- Aurrecochea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., ... Wang, H. (2009). PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Research*, 37(1), D539-D543. <https://doi.org/10.1093/nar/gkn814>
- Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M. J., Gajria, B., Grant, G. R., ... Whetzel, P. (2003). PlasmoDB: The Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Research*, 31(1), 212-215. <https://doi.org/10.1093/nar/gkg081>
- Bashir, I. M., Nyakoe, N. & van der Sande, M. (2019). Targeting remaining pockets of malaria transmission in Kenya to hasten progress towards national elimination goals: an assessment of prevalence and risk factors in children from the Lake endemic region. *Malaria Journal*, 18(233), 1-10. <https://doi.org/10.1186/s12936-019-2876-x>
- Bass, J. J., Wilkinson, D. J., Rankin, D., Phillips, B. E., Szewczyk, N. J., Smith, K., & Atherton, P. J. (2017). An overview of technical considerations for Western blotting applications to physiological research. *Scandinavian Journal of Medicine and Science in Sports*, 27(1), 4-25. <https://doi.org/10.1111/sms.12702>
- Bechtsi, D. P., & Waters, A. P. (2017). Genomics and epigenetics of sexual commitment in Plasmodium. *International Journal for Parasitology*, 47(7), 425-434. <https://doi.org/10.1016/j.ijpara.2017.03.002>
- Bergmann-Leitner, E. S., Legler, P. M., Savranskaya, T., Ockenhouse, C. F., & Angov, E. (2011). Cellular and humoral immune effector mechanisms required for sterile protection against sporozoite challenge induced with the novel malaria vaccine candidate CelTOS. *Vaccine*, 29(35), 5940-5949. <https://doi.org/10.1016/j.vaccine.2011.06.053>
- Brancucci, N. M. B., Bertschi, N. L., Zhu, L., Niederwieser, I., Chin, W. H., Wampfler,

- R., ... Voss, T. S. (2014). Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host and Microbe*, 16(2), 165-176. <https://doi.org/10.1016/j.chom.2014.07.004>
- Campbell, T. L., de Silva, E. K., Olszewski, K. L., Elemento, O., & Llinás, M. (2010). Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathogens*, 6(10), 1-15. <https://doi.org/10.1371/journal.ppat.1001165>
- Campino, S., Benavente, E. D., Assefa, S., Thompson, E., Drought, L. G., Taylor, C. J., ... Clark, T. G. (2016). Genomic variation in two gametocyte non-producing *Plasmodium falciparum* clonal lines. *Malaria Journal*, 15(1), 1-10. <https://doi.org/10.1186/s12936-016-1254-1>
- Carter, R., Mendis, K. N., Miller, L. H., Molineaux, L., & Saul, A. (2000). Malaria transmission-blocking vaccines - How can their development be supported? *Nature Medicine*, 6(3), 241-244. <https://doi.org/10.1038/73062>
- Chan, S., Ch'ng, J. H., Wahlgren, M., & Thutkawkorapin, J. (2017). Frequent GU wobble pairings reduce translation efficiency in *Plasmodium falciparum*. *Scientific Reports*, 7(723), 1-14. <https://doi.org/10.1038/s41598-017-00801-9>
- Clarke, O. B., Tomasek, D., Jorge, C. D., Dufrisne, M. B., Kim, M., Banerjee, S., ... Mancia, F. (2015). Structural basis for phosphatidylinositol-phosphate biosynthesis. *Nature Communications*, 6(1-11). <https://doi.org/10.1038/ncomms9505>
- Conway, D. J. (2015). Paths to a malaria vaccine illuminated by parasite genomics. *Trends Genetics*, 31(2), 97-107. doi:10.1016/j.tig.2014.12.005
- Crosnier, C., Mboup, S., Theron, M., Bei, A. K., Rayner, J. C., Kwiatkowski, D. P., ... Bustamante, L. Y. (2011). Basigin is a receptor essential for erythrocyte invasion

by *Plasmodium falciparum*. *Nature*, 480(7378), 534-537.
<https://doi.org/10.1038/nature10606>

Day, K. P., Karamalis, F., Thompson, J., Barnes, D. A., Peterson, C., Brown, H., ... Kemp, D. J. (1993). Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. *Proc Natl Acad Sci U S A*, 90(17), 8292-8296.
<https://doi.org/10.1016/j.jnutbio.2010.11.006>

Delves, M. J., Angrisano, F., & Blagborough, A. M. (2018). Antimalarial Transmission-Blocking Interventions: Past, Present, and Future. *Trends in Parasitology*, 34(9), 735-746. <https://doi.org/10.1016/j.pt.2018.07.001>

Desjardins, P., & Conklin, D. (2010). Nanodrop microvolume quantification of nucleic acids. *Journal of Visualized Experiments*, 45(2565), 1-4.
<https://doi.org/10.3791/2565>

Douglas, A. D., Wright, G. J., Long, C. A., Goodman, A. L., Williams, A. R., Wyllie, D. H., ... Kamuyu, G. (2011). The blood-stage malaria antigen PfRH5 is susceptible to vaccine-inducible cross-strain neutralizing antibody. *Nature Communications*, 2(1), 1-8. <https://doi.org/10.1038/ncomms1615>

Duffy, C. W., Amambua-Ngwa, A., Ahouidi, A. D., Diakite, M., Awandare, G. A., Ba, H., ... Conway, D. J. (2018). Multi-population genomic analysis of malaria parasites indicates local selection and differentiation at the *gdv1* locus regulating sexual development. *Scientific Reports*, 8(1), 1-12.
<https://doi.org/10.1038/s41598-018-34078-3>

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(13), 1-19.
<https://doi.org/10.1186/1471-2105-5-113>

- Eksi, S., Morahan, B. J., Haile, Y., Furuya, T., Jiang, H., Ali, O., ... Williamson, K. C. (2012). *Plasmodium falciparum* gametocyte development 1 (Pfgdv1) and gametocytogenesis early gene identification and commitment to sexual development. *PLoS Pathogens*, 8(10), 1-14. <https://doi.org/10.1371/journal.ppat.1002964>
- El Sahly, H. M., Keitel, W. A., Sim, B. K. L., Atmar, R. L., Thompson, D., Long, C., ... Dube, T. (2010). Safety and Immunogenicity of a Recombinant Nonglycosylated Erythrocyte Binding Antigen 175 Region II Malaria Vaccine in Healthy Adults Living in an Area Where Malaria Is Not Endemic. *Clinical and Vaccine Immunology*, 17(10), 1552-1559. <https://doi.org/10.1128/cvi.00082-10>
- Esen, M., Schleucher, R., Mordmüller, B., Schumm, M., Knobloch, J., Leroy, O., ... Jepsen, S. (2009). Safety and immunogenicity of GMZ2-a MSP3-GLURP fusion protein malaria vaccine candidate. *Vaccine*, 27(49), 6862-6868. <https://doi.org/10.1016/j.vaccine.2009.09.011>
- European Bioinformatics Institute. (2018). *InterPro: protein sequence analysis & classification*. Retrieved from <https://www.ebi.ac.uk/interpro/>
- Filarsky, M., Fraschka, S. A., Niederwieser, I., Brancucci, N. M. B., Carrington, E., Carrió, E., ... Voss, T. S. (2018). GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science*, 359(6381), 1259-1263. <https://doi.org/10.1126/science.aan6042>
- Foquet, L., Hermsen, C. C., Van Gemert, G. J., Van Braeckel, E., Weening, K. E., Sauerwein, R., ... Leroux-Roels, G. (2014). Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent *Plasmodium falciparum* infection. *Journal of Clinical Investigation*, 124(1), 140-144. <https://doi.org/10.1172/JCI70349>
- García-Angulo, V. A. (2017). Overlapping riboflavin supply pathways in bacteria.

Critical Reviews in Microbiology, 43(2), 196-209.
<https://doi.org/10.1080/1040841X.2016.1192578>

- Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., ... Barrell, B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906), 498-511. <https://doi.org/10.1038/nature01097>
- Gasteiger, E., Bairoch, A., Sanchez, J.-C., Wilkins, M. R., Appel, R. D., Hochstrasser, D. F., ... Williams, K. L. (2003). Protein Identification and Analysis Tools in the ExPASy Server. <https://doi.org/10.1385/1-59259-584-7:531>
- Gonçalves, D., & Hunziker, P. (2016). Transmission-blocking strategies: The roadmap from laboratory bench to the community. *Malaria Journal*, 15(1), 1-13. <https://doi.org/10.1186/s12936-016-1163-3>
- Grüner, N., Stambouli, O., & Ross, R. S. (2015). Dried Blood Spots - Preparing and Processing for Use in Immunoassays and in Molecular Techniques. *Journal of Visualized Experiments*, 52619(97), 1-9. <https://doi.org/10.3791/52619>
- Guo, Q., Manolopoulou, M., Bian, Y., Schilling, A. B., & Tang, W. J. (2010). Molecular Basis for the Recognition and Cleavages of IGF-II, TGF- α , and Amylin by Human Insulin-Degrading Enzyme. *Journal of Molecular Biology*, 395(2), 430-443. <https://doi.org/10.1016/j.jmb.2009.10.072>
- Hassanzadeh, J., Moradzadeh, R., Rajaeefard, A., Tahmasebi, S., & Golmohammadi, P. (2012). A Comparison of Case-control and Case-only Designs to Investigate Gene-environment Interactions using Breast Cancer Data. *Iranian Journal of Medical Sciences*, 37(2), 112-118.
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1-8. <https://doi.org/10.1016/j.ygeno.2015.11.003>

- Hermesen, C. C., Verhage, D. F., Teelen, K., Theisen, M., Bolad, A., Corradin, G., ... Berzins, K. (2006). Glutamate-rich protein (GLURP) induces antibodies that inhibit in vitro growth of *Plasmodium falciparum* in a phase 1 malaria vaccine trial. *Vaccine*, 25(15), 2930-2940. <https://doi.org/10.1016/j.vaccine.2006.06.081>
- Ito, D., Hasegawa, T., Miura, K., Yamasaki, T., Arumugam, T. U., Thongkukiatkul, A., ... Tsuboia, T. (2013). RALP1 Is a rhoptry neck erythrocyte-binding protein of *Plasmodium falciparum* merozoites and a potential blood-stage vaccine candidate antigen. *Infection and Immunity*, 81(11), 4289-4298. <https://doi.org/10.1128/IAI.00690-13>
- Jalview. (2018). *Jalview* 2.10.5. Retrieved from <http://www.jalview.org/development/release-history/Jalview-2105>
- Josling, G. A., & Llinás, M. (2015). Sexual development in Plasmodium parasites: Knowing when it's time to commit. *Nature Reviews Microbiology*, 13(9), 573-587. <https://doi.org/10.1038/nrmicro3519>
- Kafsack, B. F. C., Rovira-Graells, N., Clark, T. G., Bancells, C., Crowley, V. M., Campino, S. G., ... Llinás, M. (2014). A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature*, 507(7491), 248-252. <https://doi.org/10.1038/nature12920>
- Källberg, M., Wang, H., Lu, H., Xu, J., Peng, J., Wang, S., & Wang, Z. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, 7(8), 1511-1522. <https://doi.org/10.1038/nprot.2012.085>
- Karger, B. L., & Guttman, A. (2009). DNA sequencing by CE. *Electrophoresis*, 30(1), 196-202. <https://doi.org/10.1002/elps.200900218>
- Kariu, T., Ishino, T., Yano, K., Chinzei, Y., & Yuda, M. (2006). CelTOS, a novel malarial protein that mediates transmission to mosquito and vertebrate hosts.

- Molecular Microbiology*, 59(5), 1369-1379. <https://doi.org/10.1111/j.1365-2958.2005.05024.x>
- Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Medical Genomics*, 8(37), 1-7. <https://doi.org/10.1186/s12920-015-0115-z>
- Kipruto, E. K., Ochieng, A. O., Anyona, D. N., Mbalanya, M., Mutua, E. N., Onguru, D., ... Estambale, B. B. A. (2017). Effect of climatic variability on malaria trends in Baringo County, Kenya. *Malaria Journal*, 16(1), 220. <https://doi.org/10.1186/s12936-017-1848-2>
- Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology Evolution*, 35(6), 1547-1549. <https://doi.org/10.1093/molbev/msy096>
- Leissring, M. A., Malito, E., Hedouin, S., Reinstatler, L., Sahara, T., Abdul-Hay, S. O., ... Selkoe, D. J. (2010). Designed inhibitors of insulin-degrading enzyme regulate the catabolism and activity of insulin. *PLoS ONE*, 5(5), 1-13. <https://doi.org/10.1371/journal.pone.0010504>
- Lu, X. M., Bunnik, E. M., Pokhriyal, N., Nasser, S., Lonardi, S., & Le Roch, K. G. (2015). Analysis of nucleosome positioning landscapes enables gene discovery in the human malaria parasite *Plasmodium falciparum*. *BMC Genomics*. <https://doi.org/10.1186/s12864-015-2214-9>
- Mackinnon, M. J., & Marsh, K. (2010). The selection landscape of malaria parasites. *Science*, 328(5980), 866-871. <https://doi.org/10.1126/science.1185410>
- Mangold, K. A., Manson, R. U., Koay, E. S. C., Stephens, L., Regner, M. A., Thomson, R. B., ... Kaul, K. L. (2005). Real-time PCR for detection and identification of *Plasmodium* spp. *Journal of Clinical Microbiology*, 43(5), 2435-2440.

<https://doi.org/10.1128/JCM.43.5.2435-2440.2005>

- Marvin, R. G., Wolford, J. L., Kidd, M. J., Murphy, S., Ward, J., Que, E. L., ... O'Halloran, T. V. (2012). Fluxes in “free” and total zinc are essential for progression of intraerythrocytic stages of *Plasmodium falciparum*. *Chemistry and Biology*, *19*(6), 731-741. <https://doi.org/10.1016/j.chembiol.2012.04.013>
- McRobert, L., Taylor, C. J., Deng, W., Fivelman, Q. L., Cummings, R. M., Polley, S. D., ... Baker, D. A. (2008). Gametogenesis in malaria parasites is mediated by the cGMP-dependent protein kinase. *PLoS Biology*, *6*(6), 1243-1252. <https://doi.org/10.1371/journal.pbio.0060139>
- Miao, J., Wang, Z., Liu, M., Parker, D., Li, X., Chen, X., & Cui, L. (2013). *Plasmodium falciparum*: Generation of pure gametocyte culture by heparin treatment. *Experimental Parasitology*, *135*(3), 541-545. <https://doi.org/10.1016/j.exppara.2013.09.010>
- Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., ... Kwiatkowski, D. (2016). Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Research*, *26*(9), 1288-1299. <https://doi.org/10.1101/gr.203711.115>
- Mitchell, A. L., Potter, S. C., Necci, M., Salazar, G. A., Mi, H., Letunic, I., ... Luciani, A. (2018). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351-D360. <https://doi.org/10.1093/nar/gky1100>
- Mobegi, V. A., Duffy, C. W., Amambua-Ngwa, A., Loua, K. M., Laman, E., Nwakanma, D. C., ... Conway, D. J. (2014). Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Molecular Biology and Evolution*, *31*(6), 1490-1499. <https://doi.org/10.1093/molbev/msu106>

- Modrzynska, K., Pfander, C., Chappell, L., Yu, L., Suarez, C., Dundas, K., ... Billker, O. (2017). A Knockout Screen of ApiAP2 Genes Reveals Networks of Interacting Transcriptional Regulators Controlling the Plasmodium Life Cycle. *Cell Host and Microbe*, 21(1), 11-22. <https://doi.org/10.1016/j.chom.2016.12.003>
- Muduli, S., Pandey, P., Devatha, G., Babar, R., Thripuranthaka, M., Kothari, D. C., ... Ogale, S. (2018). Photoluminescence Quenching in Self-Assembled CsPbBr₃ Quantum Dots on Few-Layer Black Phosphorus Sheets. *Angewandte Chemie - International Edition*, 57(26), 7682-7686. <https://doi.org/10.1002/anie.201712608>
- Muralidharan, V., & Goldberg, D. E. (2013). Asparagine Repeats in *Plasmodium falciparum* Proteins: Good for Nothing? *PLoS Pathogens*, 9(8), 1-4. <https://doi.org/10.1371/journal.ppat.1003488>
- National Geographic Society. (2019). MapMaker interactive. Retrieved from <https://mapmaker.nationalgeographic.org>
- Noor, A. M., Kiptui, R., Waqo, E., Snow, R. W., Macharia, P. M., Giorgi, E., & Okiro, E. A. (2018). Spatio-temporal analysis of *Plasmodium falciparum* prevalence to understand the past and chart the future of malaria control in Kenya. *Malaria Journal*, 17(340), 1-13. <https://doi.org/10.1186/s12936-018-2489-9>
- Ogotu, B. R., Apollo, O. J., Soisson, L. A., Diggs, C., Angov, E., Milman, J. B., ... Heppner, D. G. (2009). Blood Stage Malaria Vaccine Eliciting High Antigen-Specific Antibody Concentrations Confers No Protection to Young Children in Western Kenya. *PLoS ONE*, 4(3), 1-11. <https://doi.org/10.1371/journal.pone.0004708>
- Okie, S. (2005). Betting on a Malaria Vaccine. *New England Journal of Medicine*, 353(18), 1877-1881. <https://doi.org/10.1056/nejmp058217>

- Oldfield, C. J., & Dunker, A. K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annual Review of Biochemistry*, 83(1), 553-584. <https://doi.org/10.1146/annurev-biochem-072711-164947>
- Olotu, A., Lievens, M., Njuguna, P., Fegan, G., Leach, A., Kaslow, D. C., ... Bejon, P. (2016). Seven-Year Efficacy of RTS,S/AS01 Malaria Vaccine among Young African Children. *New England Journal of Medicine*, 374(26), 2519-2529. <https://doi.org/10.1056/nejmoa1515257>
- Oyola, S. O., Ariani, C. V., Hamilton, W. L., Kekre, M., Amenga-Etego, L. N., Ghansah, A., ... & Kwiatkowski, D. P. (2016). Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification. *Malaria Journal*, 15(1), 1-12. <https://doi.org/10.1186/s12936-016-1641-7>
- Peng, J., & Xu, J. (2011). Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function and Bioinformatics*, 79(10), 161-171. <https://doi.org/10.1002/prot.23175>
- PlasmoDB. (2019). *Plasmodium genomic resource*. Retrieved from <http://plasmodb.org/plasmo/>
- Pond, S. L. K., & Frost, S. D. W. (2005). Frost, not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22(5), 1208-1222. <https://doi.org/10.1093/molbev/msi105>
- Proux, S., Hkirijareon, L., Ngamngonkiri, C., McConnell, S., & Nosten, F. (2001). Short communication: Paracheck-Pf®: A new, inexpensive and reliable rapid test for *P. falciparum* malaria. *Tropical Medicine and International Health*, 6(2), 99-101. <https://doi.org/10.1046/j.1365-3156.2001.00694.x>
- Qiagen. (2015). QIAquick® Spin Handbook. Retrieved from

<https://www.qiagen.com/us/shop/sample-technologies/dna/dna-clean-up/qiaquick-pcr-purification-kit/#resources>

Qiagen. (2016). QIAamp DNA Mini and Blood Mini Handbook. Retrieved from <https://www.qiagen.com/be/resources/download.aspx?id=62a200d6-faf4-469b-b50f-2b59cf738962&lang=en>

Quan, L., Lv, Q., & Zhang, Y. (2016). STRUM: Structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, *32*(19), 2936-2946. <https://doi.org/10.1093/bioinformatics/btw361>

RCSB PDB. (2019). *Structures*. Retrieved from <https://www.rcsb.org>

Rea, E., Le Roch, K. G., & Tewari, R. (2018). Sex in *Plasmodium falciparum*: Silence Play between GDV1 and HP1. *Trends in Parasitology*, *34*(6), 450-452. <https://doi.org/10.1016/j.pt.2018.04.006>

Robert, F., & Pelletier, J. (2018). Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. *Frontiers in Genetics*, *9*(507), 1-11. <https://doi.org/10.3389/fgene.2018.00507>

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725-738. <https://doi.org/10.1038/nprot.2010.5>

Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, *40*(W1), W471-W477. <https://doi.org/10.1093/nar/gks372>

Schussek, S., Trieu, A., Apte, S. H., Sidney, J., Sette, A., & Doolana, D. L. (2013). Immunization with apical membrane antigen 1 confers sterile infection-blocking immunity against plasmodium sporozoite challenge in a rodent model. *Infection*

and Immunity, 81(10), 3586-3599. <https://doi.org/10.1128/IAI.00544-13>

Sensoy, O., Almeida, J. G., Shabbir, J., Moreira, I. S., & Morra, G. (2017). Computational studies of G protein-coupled receptor complexes: Structure and dynamics. *Methods in Cell Biology*, 142, 205-245. <https://doi.org/10.1016/bs.mcb.2017.07.011>

Shen, H., Chen, S., Wang, Y., Xu, B., Abe, E. M., & Chen, J. (2017). Genome-wide scans for the identification of *Plasmodium vivax* genes under positive selection. *Malaria Journal*, 16(238), 1-12. <https://doi.org/10.1186/s12936-017-1882-0>.

Shen, H. M., Chen, S. B., Cui, Y. B., Xu, B., Kassegne, K., Abe, E. M., ... & Chen, J. H. (2018). Whole-genome sequencing and analysis of *Plasmodium falciparum* isolates from China-Myanmar border area. *Infectious diseases of poverty*, 7(1), 1-7.

Sinden, R. E., Carter, R., Drakeley, C., & Leroy, D. (2012). The biology of sexual development of Plasmodium: The design and implementation of transmission-blocking strategies. *Malaria Journal*, 11(70), 1-11. <https://doi.org/10.1186/1475-2875-11-70>

Sokalingam, S., Raghunathan, G., Soundrarajan, N., & Lee, S. G. (2012). A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein. *PLoS ONE*, 7(7), 1-12. <https://doi.org/10.1371/journal.pone.0040410>

Srisutham, S., Saralamba, N., Sriprawat, K., Mayxay, M., Smithuis, F., Nosten, F., ... Imwong, M. (2018). Genetic diversity of three surface protein genes in *Plasmodium malariae* from three Asian countries. *Malaria Journal*, 17(24), 1-10. <https://doi.org/10.1186/s12936-018-2176-x>

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing

- and formatting protein and DNA sequences. *BioTechniques*, 28(6), 1102-1104.
<https://doi.org/10.2144/00286ir01>
- Su, P. P., Meng, L. W., Li, J. Y., Tao, Z. Y., Chen, Y., Qiao, J. C., ... Xia, H. (2016). Cloning, expression and identification of gametocyte specific protein PFGDVL of *Plasmodium falciparum*. *Chinese Journal of Schistosomiasis Control*, 28(1), 34-38. <https://doi.org/10.16250/j.32.1374.2015168>
- Sundararaman, S. A., Plenderleith, L. J., Li, Y., Ayouba, A., Rayner, J. C., Brisson, D., ... Speede, S. (2016). Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nature Communications*, 7(1), 1-15. <https://doi.org/10.1038/ncomms11078>
- Swiss Institute of Bioinformatics. (2018). *ExPASy: SWISS-MODEL*. Retrieved from <https://swissmodel.expasy.org/>
- Tajima F. (1989). Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585-595.
- Takala, S. L., & Plowe, C. V. (2009). Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. *Parasite immunology*, 31(9), 560-573. <https://doi.org/10.1111/j.1365-3024.2009.01138.x>
- Technelysium Pty Ltd. (2018). *ChromasPro version 2.1.8*. Retrieved from <https://technelysium.com.au/wp/chromaspro/>
- Thermo Fisher Scientific. (2018). Phusion High-Fidelity PCR Master Mix with GC Buffer. Retrieved from <https://www.thermofisher.com/order/catalog/product/F532S>
- Traunmüller, F., Ramharter, M., Lagler, H., Thalhammer, F., Kremsner, P. G.,

- Graninger, W., & Winkler, S. (2003). Normal riboflavin status in malaria patients in Gabon. *American Journal of Tropical Medicine and Hygiene*, *68*(2), 182-185.
- Truebestein, L., & Leonard, T. A. (2016). Coiled-coils: The long and short of it. *BioEssays*, *38*(9), 903-916. <https://doi.org/10.1002/bies.201600062>
- Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. M. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, *35*(2), 71-74. <https://doi.org/10.1093/nar/gkm306>
- Usui, M., Prajapati, S. K., Ayanful-Torgby, R., Acquah, F. K., Cudjoe, E., Kakaney, C., ... Williamson, K. C. (2019). *Plasmodium falciparum* sexual differentiation in malaria patients is associated with host factors and GDV1-dependent genes. *Nature Communications*, *10*(1), 1-15. <https://doi.org/10.1038/s41467-019-10172-6>
- van den Berg, M., Ogutu, B., Sewankambo, N.K., Biller-Andorno, N., & Tanner, M. (2019). RTS,S malaria vaccine pilot studies: Addressing the human realities in large-scale clinical trials. *Trials*, *20*(316), 1-4. <https://doi.org/10.1186/s13063-019-3391-7>
- Vembar, S. S., Seetin, M., Scherf, A., Baybayan, P., Nattestad, M., Lambert, C., ... Smith, M. L. (2016). Complete telomere-to-telomere de novo assembly of the *Plasmodium falciparum* genome through long-read (>11 kb), single molecule, real-time sequencing. *DNA Research*, *23*(4), 339-351. <https://doi.org/10.1093/dnares/dsw022>
- Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, *44*(1), 430-435. <https://doi.org/10.1093/nar/gkw306>




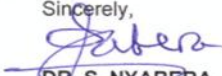
- Waterhouse, A., Bienert, S., Heer, F. T., de Beer, T. A. P., Lepore, R., Rempfer, C., ... Studer, G. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(1), 296-303. <https://doi.org/10.1093/nar/gky427>
- WHO. (2018). *Responding to antimalarial drug resistance*. Retrieved from https://www.who.int/malaria/areas/drug_resistance/overview/en/
- WHO. (2019). *World malaria report 2019*. Retrieved from <https://www.who.int/malaria/publications/world-malaria-report-2019/report/en/>
- Willems, A. R., Tahlan, K., Taguchi, T., Zhang, K., Lee, Z. Z., Ichinose, K., ... Nodwell, J. R. (2008). Crystal Structures of the Streptomyces coelicolor TetR-Like Protein ActR Alone and in Complex with Actinorhodin or the Actinorhodin Biosynthetic Precursor (S)-DNPA. *Journal of Molecular Biology*, *376*(5), 1377-1387. <https://doi.org/10.1016/j.jmb.2007.12.061>
- Wu, Q., Peng, Z., Zhang, Y., & Yang, J. (2018). COACH-D: Improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Research*, *46*(W1), W438-W442. <https://doi.org/10.1093/nar/gky439>
- Xiang, Z. (2006). Advances in Homology Protein Structure Modeling. *Current Protein & Peptide Science*, *7*(3), 217-227. <https://doi.org/10.2174/138920306777452312>
- Yang, J., Roy, A., & Zhang, Y. (2013a). BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, *41*(D1), D1096-D1103. <https://doi.org/10.1093/nar/gks966>
- Yang, J., Roy, A., & Zhang, Y. (2013b). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, *29*(20), 2588-2595.

<https://doi.org/10.1093/bioinformatics/btt447>



- Yang, J., Wang, Y., & Zhang, Y. (2016). ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. *Journal of Molecular Biology*, 428(4), 693-701. <https://doi.org/10.1016/j.jmb.2015.09.024>
- Yang, J., Zhang, W., He, B., Walker, S. E., Zhang, H., Govindarajoo, B., ... Zhang, Y. (2016). Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. *Proteins*, 84, 233-246. <https://doi.org/10.1002/prot.24918>
- Yang, J., & Zhang, Y. (2015a). I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research*, 43(1), 174-181. <https://doi.org/10.1093/nar/gkv342>
- Yang, J., & Zhang, Y. (2015b). Protein Structure and Function Prediction Using I-TASSER. *Current Protocols in Bioinformatics*, 52(5.8), 1-15. <https://doi.org/10.1002/0471250953.bi0508s52>
- Yuda, M., Iwanaga, S., Shigenobu, S., Mair, G. R., Janse, C. J., Waters, A. P., ... Kaneko, I. (2009). Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Molecular Microbiology*, 71(6), 1402-1414. <https://doi.org/10.1111/j.1365-2958.2009.06609.x>
- Zhang, C., Freddolino, P. L., & Zhang, Y. (2017). COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Research*, 45(1), 291-299. <https://doi.org/10.1093/nar/gkx366>
- Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins: Structure, Function and Bioinformatics*, 77(9), 100-113. <https://doi.org/10.1002/prot.22588>

APPENDICES

Appendix I: Ethical clearance

	
INSTITUTIONAL RESEARCH AND ETHICS COMMITTEE (IREC)	
MOI TEACHING AND REFERRAL HOSPITAL P.O. BOX 3 ELDORET Tel: 334711/2/3	MOI UNIVERSITY COLLEGE OF HEALTH SCIENCES P.O. BOX 4606 ELDORET
Reference: IREC/2017/41 Approval Number: 0001927	27 th July, 2017
Josephat Kipsang Bungei, Jomo Kenyatta University of Agriculture & Technology, P.O. Box 62000-00200, <u>NAIROBI-KENYA.</u>	
Dear Mr. Bungei,	
<u>RE: FORMAL APPROVAL</u>	
The Institutional Research and Ethics Committee has reviewed your research proposal titled:- <i>"Genetic Characterization of Gametocyte Development Protein I and PFAP2-G in Plasmodium Falciparum Isolates".</i>	
Your proposal has been granted a Formal Approval Number: FAN: IREC 1927 on 27 th July, 2017. You are therefore permitted to begin your investigations.	
Note that this approval is for 1 year; it will thus expire on 26 th July, 2018. If it is necessary to continue with this research beyond the expiry date, a request for continuation should be made in writing to IREC Secretariat two months prior to the expiry date.	
You are required to submit progress report(s) regularly as dictated by your proposal. Furthermore, you must notify the Committee of any proposal change (s) or amendment (s), serious or unexpected outcomes related to the conduct of the study, or study termination for any reason. The Committee expects to receive a final report at the end of the study.	
Sincerely,  <u>DR. S. NYABERA</u> DEPUTY-CHAIRMAN <u>INSTITUTIONAL RESEARCH AND ETHICS COMMITTEE</u>	
cc CEO - MTRH Dean - SOP Dean - SOM Principal - CHS Dean - SON Dean - SOD	

Appendix II: Informed consent form

Genetic Characterization of Gametocyte Development Protein I and PfAP2-G in <i>Plasmodium falciparum</i> Isolates	
	
MOI UNIVERSITY COLLEGE OF HEALTH SCIENCES / MOI TEACHING AND REFERRAL HOSPITAL INSTITUTIONAL RESEARCH AND ETHICS COMMITTEE (IREC) INFORMED CONSENT FORM (ICF)	
Study Title: Genetic Characterization of Gametocyte Development Protein I and PfAP2-G in <i>Plasmodium falciparum</i> Isolates	
Name of Principal Investigator(s): Mr. Josephat Kipsang Bungei	
Co Investigators: Dr. Steven Ger Nyanjom and Dr. Victor Mobegi	
Name of Organization: JKUAT, P.O. BOX 62000-00200, Nairobi. Tel. 067-5870001	
Name of Sponsor: None	
Informed Consent Form for: Malaria Patients	
This Informed Consent Form has two parts: <ul style="list-style-type: none">• Information Sheet (to share information about the study with you)• Certificate of Consent (for signatures if you choose to participate)	
Part I: Information Sheet	
Introduction: <p>You are being asked to take part in a research study. This information is provided to tell you about the study. Please read this form carefully. You will be given a chance to ask questions. If you decide to be in the study, you will be given a copy of this consent form for your records.</p>	
<p>Taking part in this research study is voluntary. You may choose not to take part in the study. You could still receive other treatments. Saying no will not affect your rights to health care or services. You are also free to withdraw from this study at any time. If after data collection you choose to quit, you can request that the information provided by you be destroyed under supervision- and thus not used in the research study. You will be notified if new information becomes available about the risks or benefits of this research. Then you can decide if you want to stay in the study</p>	
Purpose of the study: <p>The purpose of the study is to find out whether genes involved in transmission of malaria parasite are diverse or the same and use the outcome in searching for molecules that can act as drugs to prevent transmission of these parasites from human to mosquito.</p>	
Type of Research Project/Intervention: <p>This is a laboratory study that will entail collection of blood samples, extraction, and amplification of DNA, and subsequent sequencing of DNA amplicons. As a research participant, you are required to provide 2ml of blood sample through venipuncture.</p>	
<hr/>	
Subject's Initial ----- Code	Page 1 of 4

Genetic Characterization of Gametocyte Development Protein I and PfAP2-G in *Plasmodium falciparum* Isolates

Why have I been identified to Participate in this study?

As the inclusion criteria of the participants, the study identified you as a participant because you have been diagnosed with malaria and you fall in the age group of 20-45 years. Moreover, the study included you because you are not pregnant, do not have a mental illness, and are not a prisoner.

How long will the study last?

You will be in this study for a short period less than 30 minutes for you to give informed consent and permit collection of blood sample

What will happen to me during the study?

A. Provide a brief introduction to the format of the research study.

We are asking you to help us learn more about transmission of malaria. If you accept, you will be asked to donate blood sample for study of malaria. The procedure is simple one which entails venipuncture to collect blood sample for study. Your blood will be used in the study of malaria only because the malarial DNA will be extracted and used in molecular studies aimed at understanding transmission proteins of malaria.

B. Explain the type of questions that the participants will answer

The participants will only be asked demographic questions to identify their name, gender, age, place of residence, and contact information

What side effects or risks I can expect from being in the study?

There are no likely side effects or risks posed by the study because qualified technicians and technologists will collect blood samples.

Are there benefits to taking part in the study?

- a) The possible benefits to you from this study are... **or**
- b) You may not benefit personally from this study...

There are no direct benefits you will get in participating in the study

- c) The possible benefits to society may include...

Prevention of malaria transmission and eradication of malaria

Reimbursements:

The study does not make any reimbursements to participants

Who do I call if I have questions about the study?

Questions about the study: PI Contact Info...

Name: Josephat Kipsang Bungei

Mobile No. 0722638262

Email: josephatbungei@gmail.com

Questions about your rights as a research subject: You may contact Institutional Review Ethics Committee (IREC) 053 33471 Ext.3008. IREC is a group of people that reviews studies for safety and to protect the rights of study subjects.

Will the information I provide be kept private?

All reasonable efforts will be made to keep your protected information (private and confidential). Protected Information is information that is, or has been, collected or maintained and can be linked back to you. Using or sharing ("disclosure") of such information must follow National privacy guidelines. By signing the consent document for this study, you are giving permission ("authorization") for the uses and disclosures of your personal information. A decision to take part in this research means that you agree to

Subject's Initial ----- Code

Genetic Characterization of Gametocyte Development Protein I and PfAP2-G in *Plasmodium falciparum* Isolates

let the research team use and share your Protected Information as described below.

As part of the study, [Josephat Kipsang Bungei] and his study team may share the results of your [plasmodium DNA, DNA sequences of target genes, polymorphic data, predicated protein structures, drug targets, and ligands). These may be study or non-study related. They may also share portions of your medical record, with the groups named below:

- The National Bioethics. Committee,
- The Institutional Review and Ethics Committee
- Jomo Kenyatta University of Agriculture & Technology

National privacy regulations may not apply to these groups; however, they have their own policies and guidelines to assure that all reasonable efforts will be made to keep your personal information private and confidential.

[OPTIONAL: The sponsor may give your personal health information, not containing your name, to others or use it for research purposes other than those listed in this form. In handling your personal information, the sponsor, [PI] and associated staff will keep your information in strict confidence, and shall comply with any and all applicable laws regarding the confidentiality of such information.]

The study results will be retained in your research record for at least six years after the study is completed. At that time, the research information not already in your medical record will be discarded. Any research information entered into your medical record will be kept indefinitely.

Unless otherwise indicated, this permission to use or share your Personal Information does not have an expiration date. If you decide to withdraw your permission, we ask that you contact [Josephat Kipsang Bungei] in writing and let him know that you are withdrawing your permission. The mailing address is [josephatbungei@gmail.com]. At that time, we will stop further collection of any information about you. However, the health information collected before this withdrawal may continue to be used for the purposes of reporting and research quality.

[OPTIONAL: You have the right to see and copy your personal information related to the research study for as long as the study doctor or research institution holds this information. However, to ensure the scientific quality of the research study, you will not be able to review some of your research information until after the research study has been completed.]

Your treatment, payment, or enrollment in any health plans or eligibility for benefits will not be affected if you decide not to take part. You will receive a copy of this form after it is signed.

Subject's Initial ----- Code

Genetic Characterization of Gametocyte Development Protein I and PfAP2-G in *Plasmodium falciparum* Isolates

Part II: Consent of Subject:

I have read or have had read to me the description of the research study. The investigator or his/her representative has explained the study to me and has answered all of the questions I have at this time. I have been told of the potential risks, discomforts and side effects as well as the possible benefits (if any) of the study. I freely volunteer to take part in this study of malaria research. The study entails drawing of blood sample (2 ml) through venipuncture and used in the extraction of plasmodium DNA for research objectives of determining polymorphism, predicting protein structures, and identification drug targets and ligands. In this view, I will participate in the study by donating my blood sample and allow its use in stated research objectives.

Name of Participant (Witness to print if the subject is unable to write)	Signature of subject/thumbprint	Date & Time
--	---------------------------------	-------------

Name of Representative/Witness	Relationship to Subject
--------------------------------	-------------------------

Name of person Obtaining Consent	Signature of person Obtaining Consent	Date
----------------------------------	--	------

Josephat Kipsang_ Bungei_ Printed name of Investigator	Signature of Investigator	Date
---	---------------------------	------

Subject's Initial ----- Code

Appendix III: Samples collected

Blood samples with plasmodium parasites were collected from adults and the information recorded in the following tables.

Details of samples collected from the study site of Baringo County Referral Hospital

Patient ID (Coded)	Gender	Age	Date Collected	Blood volume drawn (ml)	Parasitaemia (p/ul)	Sample number	Selected Samples
DT01	F	28	22/09/2017	2	6400	KBT1	
WL02	F	32	27/9/2017	2.5	18240	KBT2	B1
CJ03	F	34	27/9/2017	1.5	26440	KBT3	B2
KK04	M	21	2/10/2017	3	72880	KBT4	B3
SC05	F	23	3/10/2017	2.5	21640	KBT5	B4
GO06	F	27	13/10/2017	2	29680	KBT6	B5
JC07	F	38	13/10/2017	2	78040	KBT7	B6
AK08	M	29	16/10/2017	2	34840	KBT8	B7
WT09	M	31	16/10/2017	2	44440	KBT9	B8
DK10	M	40	4/11/2017	1.5	17280	KBT10	B9
AA11	F	40	6/11/2017	2.5	48160	KBT11	B10

Details of samples collected from the study site of Uasin Gishu County Hospital

Patient ID (Coded)	Gender	Age	Date Collected	Blood volume drawn (ml)	Parasitaemia (p/ul)	Sample number	Selected Samples
SN01	F	24	22/09/2017	2	36080	UG1	U1
RA02	M	33	29/09/2017	2	10160	UG2	U2
EC03	F	27	10/10/2017	3	12800	UG3	U3
CJ04	F	28	11/10/2017	2	31040	UG4	U4
SR05	M	31	14/10/2017	2	17840	UG5	U5
MO06	F	25	15/10/2017	1.5	66800	UG6	U6
MA07	F	32	22/10/2017	2	13560	UG7	U7
SM08	M	36	2/11/2017	2.5	5040	UG8	U8
SO09	M	34	3/11/2017	2	8840	UG9	U9
HW10	F	32	3/11/2017	3	7680	UG10	U10
CK11	F	22	5/11/2017	1.5	48200	UG11	
GB12	F	27	6/11/2017	2	70080	UG12	

Details of samples collected from the study site of Kapsabet County Referral Hospital

Patient ID (Coded)	Gender	Age	Date Collected	Blood volume drawn (ml)	Parasitaemia (p/ul)	Sample number	Selected Samples
DK01	M	32	17/11/2017	3	5240	KPS1	
KB02	M	25	27/12/2017	3	10440	KPS2	
VO03	M	28	28/122017	4	4800	KPS3	
SC04	F	26	30/12/2017	3	19200	KPS4	N1
PK05	M	36	2/1/2018	2	12280	KPS5	N2
ER06	F	24	8/1/2018	3	22040	KPS6	N3
AO07	M	33	8/1/2018	2	70000	KPS7	N4
TO08	M	34	11/1/2018	3	22400	KPS8	N5
JA09	F	25	22/1/2018	4	7640	KPS9	N6
LA10	F	28	22/1/2018	3	14200	KPS10	N7
MJ11	F	29	22/2/2018	2	11260	KPS11	N8
AA12	F	26	6/3/2018	3	11320	KPS12	N9
LA13	M	36	22/3/2018	2	80000	KPS13	N10

Appendix IV: Results of DNA Quantification Using Nanodrop Spectrophotometer

Sample Number	DNA Quantity (ng/ul)	Absorbance at 260nm	Absorbance at 280 nm	260nm/280nm	Selected Samples
KBT1	43.31	0.043	0.026	1.65	
KBT2	8.206	0.008	0.004	2.00	B1
KBT3	19.54	0.02	0.011	1.82	B2
KBT4	40.85	0.041	0.019	2.16	B3
KBT5	61.73	0.061	0.029	2.10	B4
KBT6	14.58	0.015	0.007	2.14	B5
KBT7	11.97	0.012	0.006	2.00	B6
KBT8	24.84	0.025	0.013	1.92	B7
KBT9	15.69	0.016	0.009	1.78	B8
KBT10	12.35	0.012	0.007	1.71	B9
KBT11	29.61	0.03	0.014	2.14	B10
UG1	42.81	0.043	0.022	1.95	U1
UG2	29.04	0.029	0.015	1.93	U2
UG3	27.55	0.028	0.015	1.87	U3
UG4	37.2	0.037	0.017	2.18	U4
UG5	22.41	0.022	0.012	1.83	U5
UG6	33.24	0.033	0.018	1.83	U6
UG7	32.59	0.033	0.016	2.06	U7
UG8	43.4	0.043	0.023	1.87	U8
UG9	53.28	0.053	0.027	1.96	U9

Sample Number	DNA Quantity (ng/ul)	Absorbance at 260nm	Absorbance at 280 nm	260nm/280nm	Selected Samples
UG10	35.87	0.036	0.02	1.80	U10
UG11	45.87	0.046	0.022	2.09	
UG12	28.66	0.029	0.017	1.71	
KPS1	57.19	0.057	0.029	1.97	
KPS2	16.7	0.017	0.009	1.89	
KPS3	15	0.015	0.008	1.88	
KPS4	18.51	0.017	0.009	1.89	N1
KPS5	23.45	0.023	0.012	1.92	N2
KPS6	32.56	0.033	0.017	1.94	N3
KPS7	42.58	0.043	0.021	2.05	N4
KPS8	29.48	0.029	0.016	1.81	N5
KPS9	27.52	0.028	0.013	2.15	N6
KPS10	19.34	0.019	0.009	2.11	N7
KPS11	44.65	0.045	0.025	1.80	N8
KPS12	49.47	0.049	0.024	2.04	N9
KPS13	23.35	0.023	0.012	1.92	N10

Appendix V: Sequencing results

Reaction Number	Label	Read Length			GC%
		Normal	QV \geq 16	QV \geq 20	
1	B1 PF4001F	1245	1071	1071	25.0
2	B2 PF4001F	1311	1098	1099	24.0
3	B3 PF4001F	1213	816	817	25.0
4	B4 PF4001F	1269	1078	1076	25.0
5	B5 PF4001F	1338	1152	1150	27.0
6	B6 PF4001F	831	72	60	33.0
7	B7 PF4001F	1242	789	702	28.0
8	B8 PF4001F	1238	936	773	25.0
9	B9 PF4001F	1238	1025	1026	24.0
10	B10 PF4001F	1304	1003	1005	26.0
11	U1 PF4001F	1259	519	424	28.0
12	U2 PF4001F	1254	1029	1029	25.0
13	U3 PF4001F	1350	770	769	27.0
14	U4 PF4001F	1150	367	367	29.0
15	U5 PF4001F	1358	755	751	28.0
16	U6 PF4001F	1276	951	884	27.0
17	U7 PF4001F	1323	682	670	30.0
18	U8 PF4001F	1322	960	943	26.0
19	U9 PF4001F	1163	773	776	26.0
20	U10 PF4001F	1268	1077	1077	26.0
21	N1 PF4001F	1152	401	399	28.0
22	N2 PF4001F	1154	453	454	27.0
23	N3 PF4001F	1262	1044	1044	25.0
24	N4 PF4001F	1268	976	976	25.0
25	N5 PF4001F	568	465	451	48.0
26	N6 PF4001F	1323	515	515	32.0
27	N7 PF4001F	2224	206	210	35.0
28	N8 PF4001F	1318	1096	1087	24.0

Reaction Number	Label	Read Length			GC%
29	N9 PF4001F	840	830	712	25.0
30	N10 PF4001F	1094	840	740	25.0
31	B11 PF4002F	556	555	555	25.0
32	B12 PF4002F	579	543	535	25.0
33	B13 PF4002F	600	599	587	25.0
34	B14 PF4002F	592	591	591	25.0
35	B15 PF4002F	587	586	586	25.0
36	B16 PF4002F	600	122	40	30.0
37	B17 PF4002F	598	597	597	25.0
38	B18 PF4002F	583	565	567	25.0
39	B19 PF4002F	610	603	603	25.0
40	B20 PF4002F	573	549	370	27.0
41	U11 PF4002F	602	601	601	25.0
42	U12 PF4002F	578	555	556	25.0
43	U13 PF4002F	593	592	583	25.0
44	U14 PF4002F	582	564	561	25.0
45	U15 PF4002F	591	578	578	25.0
46	U16 PF4002F	591	590	590	25.0
47	U17 PF4002F	591	63	49	30.0
48	U18 PF4002F	586	585	585	26.0
49	U19 PF4002F	582	581	557	25.0
50	U20 PF4002F	601	600	600	26.0
51	N11 PF4002F	572	571	571	26.0
52	N12 PF4002F	594	593	593	25.0
53	N13 PF4002F	592	591	591	25.0
54	N14 PF4002F	589	575	351	25.0
55	N15 PF4002F	578	577	577	25.0
56	N16 PF4002F	608	596	596	25.0
57	N17 PF4002F	576	575	575	25.0
58	N18 PF4002F	589	581	581	25.0
59	N19 PF4002F	588	560	560	25.0
60	N20 PF4002F	586	585	585	25.0

Reaction Number	Label	Read Length			GC%
61	<u>B21 PF600F</u>	527	263	261	46.0
62	<u>B22 PF600F</u>	285	263	260	19.0
63	<u>B23 PF600F</u>	600	147	135	26.0
64	<u>B24 PF600F</u>	579	51	21	34.0
65	<u>B25 PF600F</u>	313	312	312	19.0
66	<u>B26 PF600F</u>	301	300	287	18.0
67	<u>B27 PF600F</u>	288	287	287	19.0
68	<u>B28 PF600F</u>	285	261	263	19.0
69	<u>B29 PF600F</u>	271	270	257	20.0
70	<u>B30 PF600F</u>	264	263	253	20.0
71	<u>U21 PF600F</u>	554	133	32	32.0
72	<u>U22 PF600F</u>	555	155	82	27.0
73	<u>U23 PF600F</u>	536	65	19	35.0
74	<u>U24 PF600F</u>	516	84	28	36.0
75	<u>U25 PF600F</u>	294	245	239	25.0
76	<u>U26 PF600F</u>	256	220	124	20.0
77	<u>U27 PF600F</u>	168	167	167	21.0
78	<u>U28 PF600F</u>	534	221	178	37.0
79	<u>U29 PF600F</u>	217	216	216	19.0
80	<u>U30 PF600F</u>	225	213	210	19.0
81	<u>N21 PF600F</u>	478	53	41	36.0
82	<u>N22 PF600F</u>	521	60	36	33.0
83	<u>N23 PF600F</u>	297	169	77	21.0
84	<u>N24 PF600F</u>	295	269	269	18.0
85	<u>N25 PF600F</u>	291	252	237	20.0
86	<u>N26 PF600F</u>	355	234	234	29.0
87	<u>N27 PF600F</u>	231	219	218	19.0
88	<u>N28 PF600F</u>	285	272	257	19.0
89	<u>N29 PF600F</u>	238	237	237	19.0
90	<u>N30 PF600F</u>	298	285	262	19.0
91	<u>Control1 PF4001F</u>	1180	980	950	24.0
92	<u>Control2 PF4001F</u>	1230	1186	1055	20.0

Reaction Number	Label	Read Length			GC%
93	Control3 PF4002F	602	580	16	31.0
94	Control4 PF4002F	589	523	44	36.0
95	Control5 PF600F	299	276	20	32.0
96	Control6 PF600F	287	264	25	31.0
97	B1 PF4001R	1277	1095	1095	25.0
98	B2 PF4001R	1280	1058	1058	25.0
99	B3 PF4001R	1282	1054	968	25.0
100	B4 PF4001R	1357	1059	973	26.0
101	B5 PF4001R	1261	1095	831	25.0
102	B6 PF4001R	1377	1093	1023	26.0
103	B7 PF4001R	1534	715	715	28.0
104	B8 PF4001R	1281	1060	957	24.0
105	B9 PF4001R	1217	1106	1106	23.0
106	B10 PF4001R	1293	1158	1081	26.0
107	U1 PF4001R	1577	355	355	27.0
108	U2 PF4001R	1277	1108	1108	24.0
109	U3 PF4001R	1290	1112	1035	24.0
110	U4 PF4001R	1205	782	670	26.0
111	U5 PF4001R	1322	1133	1005	26.0
112	U6 PF4001R	1289	1093	1015	25.0
113	U7 PF4001R	1358	1022	934	28.0
114	U8 PF4001R	1291	898	892	25.0
115	U9 PF4001R	1269	1095	1021	24.0
116	U10 PF4001R	1252	945	946	25.0
117	N1 PF4001R	1273	829	816	24.0
118	N2 PF4001R	2157	221	147	30.0
119	N3 PF4001R	1346	1098	1005	27.0
120	N4 PF4001R	1336	1183	1108	26.0
121	N5 PF4001R	1078	346	347	35.0
122	N6 PF4001R	1287	833	831	26.0
123	N7 PF4001R	1282	906	905	26.0
124	N8 PF4001R	1269	1111	1111	23.0

Reaction Number	Label	Read Length			GC%
125	<u>N9 PF4001R</u>	1754	346	346	29.0
126	<u>N10 PF4001R</u>	1096	955	878	25.0
127	<u>B11 PF4002R</u>	608	592	592	24.0
128	<u>B12 PF4002R</u>	595	562	559	25.0
129	<u>B13 PF4002R</u>	595	594	594	25.0
130	<u>B14 PF4002R</u>	602	601	592	25.0
131	<u>B15 PF4002R</u>	613	600	596	25.0
132	<u>B16 PF4002R</u>	609	608	608	24.0
133	<u>B17 PF4002R</u>	603	602	602	24.0
134	<u>B18 PF4002R</u>	601	590	590	25.0
135	<u>B19 PF4002R</u>	589	588	588	25.0
136	<u>B20 PF4002R</u>	1100	254	91	30.0
137	<u>U11 PF4002R</u>	601	600	600	24.0
138	<u>U12 PF4002R</u>	585	584	584	25.0
139	<u>U13 PF4002R</u>	613	603	593	25.0
140	<u>U14 PF4002R</u>	598	597	597	25.0
141	<u>U15 PF4002R</u>	599	598	598	25.0
142	<u>U16 PF4002R</u>	606	605	605	25.0
143	<u>U17 PF4002R</u>	594	593	593	24.0
144	<u>U18 PF4002R</u>	601	600	582	25.0
145	<u>U19 PF4002R</u>	598	597	591	24.0
146	<u>U20 PF4002R</u>	603	602	602	25.0
147	<u>N11 PF4002R</u>	599	598	598	25.0
148	<u>N12 PF4002R</u>	601	600	600	25.0
149	<u>N13 PF4002R</u>	599	598	598	25.0
150	<u>N14 PF4002R</u>	609	608	608	26.0
151	<u>N15 PF4002R</u>	592	591	591	24.0
152	<u>N16 PF4002R</u>	595	594	594	24.0
153	<u>N17 PF4002R</u>	590	540	540	24.0
154	<u>N18 PF4002R</u>	592	559	559	25.0
155	<u>N19 PF4002R</u>	593	592	592	24.0
156	<u>N20 PF4002R</u>	602	601	601	25.0

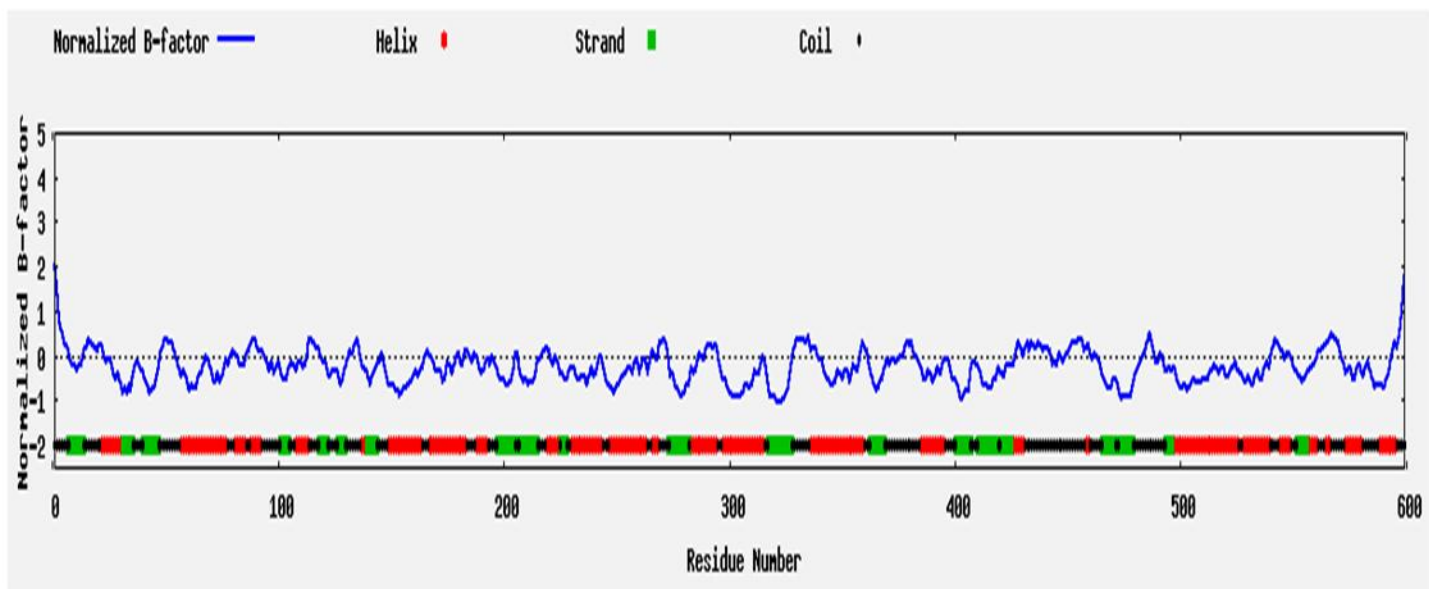
Reaction Number	Label	Read Length			GC%
157	<u>B21 PF600R</u>	262	261	261	19.0
158	<u>B22 PF600R</u>	430	267	231	28.0
159	<u>B23 PF600R</u>	256	216	218	23.0
160	<u>B24 PF600R</u>	209	175	175	17.0
161	<u>B25 PF600R</u>	305	304	304	18.0
162	<u>B26 PF600R</u>	306	282	271	18.0
163	<u>B27 PF600R</u>	264	228	220	19.0
164	<u>B28 PF600R</u>	298	231	231	18.0
165	<u>B29 PF600R</u>	240	230	227	20.0
166	<u>B30 PF600R</u>	262	243	233	20.0
167	<u>U21 PF600R</u>	620	197	56	28.0
168	<u>U22 PF600R</u>	576	258	247	45.0
169	<u>U23 PF600R</u>	223	175	175	18.0
170	<u>U24 PF600R</u>	262	214	211	21.0
171	<u>U25 PF600R</u>	244	219	219	21.0
172	<u>U26 PF600R</u>	246	235	213	19.0
173	<u>U27 PF600R</u>	274	40	32	29.0
174	<u>U28 PF600R</u>	224	223	223	19.0
175	<u>U29 PF600R</u>	220	211	211	19.0
176	<u>U30 PF600R</u>	242	226	223	19.0
177	<u>N21 PF600R</u>	210	175	175	17.0
178	<u>N22 PF600R</u>	228	181	181	18.0
179	<u>N23 PF600R</u>	671	252	217	46.0
180	<u>N24 PF600R</u>	256	232	232	19.0
181	<u>N25 PF600R</u>	285	236	236	19.0
182	<u>N26 PF600R</u>	246	231	231	21.0
183	<u>N27 PF600R</u>	230	229	229	19.0
184	<u>N28 PF600R</u>	859	227	228	33.0
185	<u>N29 PF600R</u>	246	231	230	19.0
186	<u>N30 PF600R</u>	245	233	229	20.0
187	<u>Contro1 PF4001R</u>	1275	1189	1125	27.0
188	<u>Contro2 PF4001R</u>	1290	1019	1113	25.0

Reaction Number	Label	Read Length			GC%
189	<u>Contro3 PF4002R</u>	561	586	522	25.0
190	<u>Contro4 PF4002R</u>	605	557	507	25.0
191	<u>Contro5 PF600R</u>	325	275	216	26.0
192	<u>Contro6 PF600R</u>	310	284	256	39.0

Appendix VI: Protein prediction of *Pfgdv1*

Predicted normalized B-factor

(B-factor is a value to indicate the extent of the inherent thermal mobility of residues/atoms in proteins. In I-TASSER, this value is deduced from threading template proteins from the PDB in combination with the sequence profiles derived from sequence databases. The reported B-factor profile in the figure below corresponds to the normalized B-factor of the target protein, defined by $B = (B' - u) / s$, where B' is the raw B-factor value, u and s are respectively the mean and standard deviation of the raw B-factors along the sequence. [Click here to read more about predicted normalized B-factor](#))



Top 10 threading templates used by I-TASSER

(I-TASSER modeling starts from the structure templates identified by LOMETS from the PDB library. LOMETS is a meta-server threading approach containing multiple threading programs, where each threading program can generate tens of thousands of template alignments. I-TASSER only uses the templates of the highest significance in the threading alignments, the significance of which are measured by the Z-score, i.e. the difference between the raw and average scores in the unit of standard deviation. The templates in this section are the 10 best templates selected from the LOMETS threading programs. Usually, one template of the highest Z-score is selected from each threading program, where the threading programs are sorted by the average performance in the large-scale benchmark test experiments.)

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.	20	40	60	80	100
							20	40	60	80	100
							Sec.Str				
							Seq				
1	1q2lA	0.14	0.17	0.82	1.21	Download	-----ETGWQ-PIQETIRKSDKDNR-----QYQAILDNAVKSLSALVVPVGSLEDEAYQGLAAYLEHMS-----LMGSKK--YPQADLAEYLKM-HG				
2	1vt4	0.19	0.07	0.73	1.01	Download	-----MDFETGEHQYQYKDNFDKDVQDMPSEEIIMSKDAVSGTLRL----FWTLLKQEMVQKFEV				
3	4w8yA	0.09	0.20	0.95	1.62	Download	MVLKIEENHEERSKILSSGNIQDKVQADALSSKTQRFIIREPVIDFLGRFHVGPVLRGSRNRRGERFVNEFLERVSKLEGDVLKEVFEASNKFKGEESKQWALKEGVKEFAKSELK				
4	3r7wA	0.18	0.11	0.38	1.10	Download	-----TTVNTD-----DFSRHAKAEELDLELITFFASTNELNKIFSRNINNDENKKILNSKIKIIRDLDTNFIKFTKIGTNERNRILHAIKERDLQGTQDDYNKVINIIQNLIKISD				
5	5wtjA	0.10	0.21	0.98	1.62	Download	-----SLYELDPKWKLLKTDNDFLGGTLVNEFVQELSKDHRNDVLIDANTKNLPTNEKQDRQYFSEQVATQEVHSENVIKLSKDLHTLLTFDKLDRRLTYIQSVELIRRYNDFYSMGKS				
6	4ei7A	0.13	0.13	0.47	1.10	Download	-----GNFSEIESQGNKETQIK--NKNYP				
7	5yfpE	0.09	0.20	0.91	1.61	Download	-----LKAANLETNVEELWVEVGRVSRRELYTRQKIGEEAMFNPQLMIQTPKEEGANVLTTEALLQHLDSALQASR---VHVYMYNRQWKEHLQCYKSGELITETGYMDQIIEYLYPCLII				
8	4pevA	0.14	0.10	0.52	1.10	Download	-----FL-SREVVDALEERVEKLEQEAARKGFDSYV-----QSLSHNALLAKKNGLESTTAAGFKNSLDEPY---KTYLPESEWERAQGVLGARYLQAVLSSGTQ				
9	6d4hA	0.07	0.16	0.93	1.58	Download					
10	5ot4A	0.10	0.17	0.87	1.56	Download					

- (a) All the residues are colored in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in color. Coloring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are colored in dark shade. ([more about the colors used](#))
- (b) Rank of templates represents the top ten threading templates used by I-TASSER.
- (c) Iden1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.
- (d) Iden2 is the percentage sequence identity of the whole template chains with query sequence.
- (e) Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein.
- (f) Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.
- (g) Download Align. provides the 3D structure of the aligned regions of the threading templates.
- (h) The top 10 alignments reported above (in order of their ranking) are from the following threading programs:
 1: pGenTHREADER 2: HHSEARCH1 3: PROSPECT2 4: pGenTHREADER 5: PROSPECT2 6: pGenTHREADER 7: PROSPECT2 8: pGenTHREADER 9: PROSPECT2 10: PROSPECT2

Proteins structurally close to the target in the PDB (as identified by [TM-align](#))

(After the structure assembly simulation, I-TASSER uses the TM-align structural alignment program to match the first I-TASSER model to all structures in the PDB library. This section reports the top 10 proteins from the PDB that have the closest structural similarity, i.e. the highest [TM-score](#), to the predicted I-TASSER model. Due to the structural similarity, these proteins often have similar function to the target. However, users are encouraged to use the data in the next section 'Predicted function using COACH' to infer the function of the target protein, since COACH has been extensively trained to derive biological functions from multi-source of sequence and structure features which has on average a higher accuracy than the function annotations derived only from the global structure comparison.)



Top 10 Identified structural analogs in PDB

Click to view	Rank	PDB Hit	TM-score	RMSD ^a	IDEN ^d	Cov	Alignment
<input type="radio"/>	1	1q2IA	0.785	2.83	0.098	0.851	Download
<input type="radio"/>	2	3cwwB	0.628	5.13	0.077	0.783	Download
<input type="radio"/>	3	3s5hA	0.553	5.78	0.078	0.738	Download
<input type="radio"/>	4	2fgeA	0.533	5.73	0.067	0.713	Download
<input type="radio"/>	5	4I3tA	0.531	5.69	0.082	0.705	Download
<input type="radio"/>	6	2q47A1	0.529	3.43	0.101	0.596	Download
<input type="radio"/>	7	6b03A	0.515	6.37	0.042	0.726	Download
<input type="radio"/>	8	5cioA	0.510	5.45	0.055	0.658	Download
<input type="radio"/>	9	1l0nA	0.450	4.69	0.054	0.553	Download
<input type="radio"/>	10	3cwbA	0.449	4.62	0.063	0.549	Download

(a) Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.

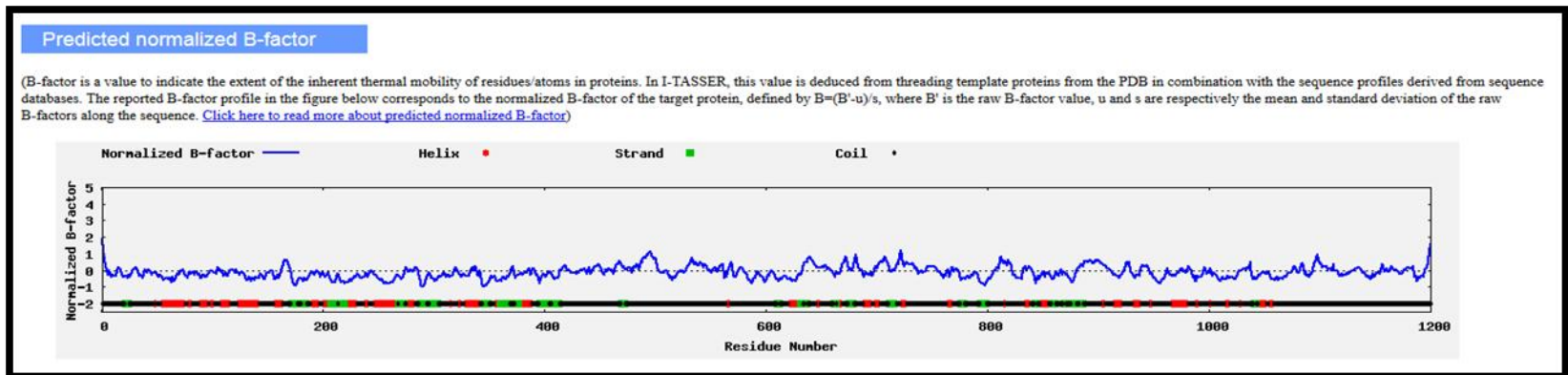
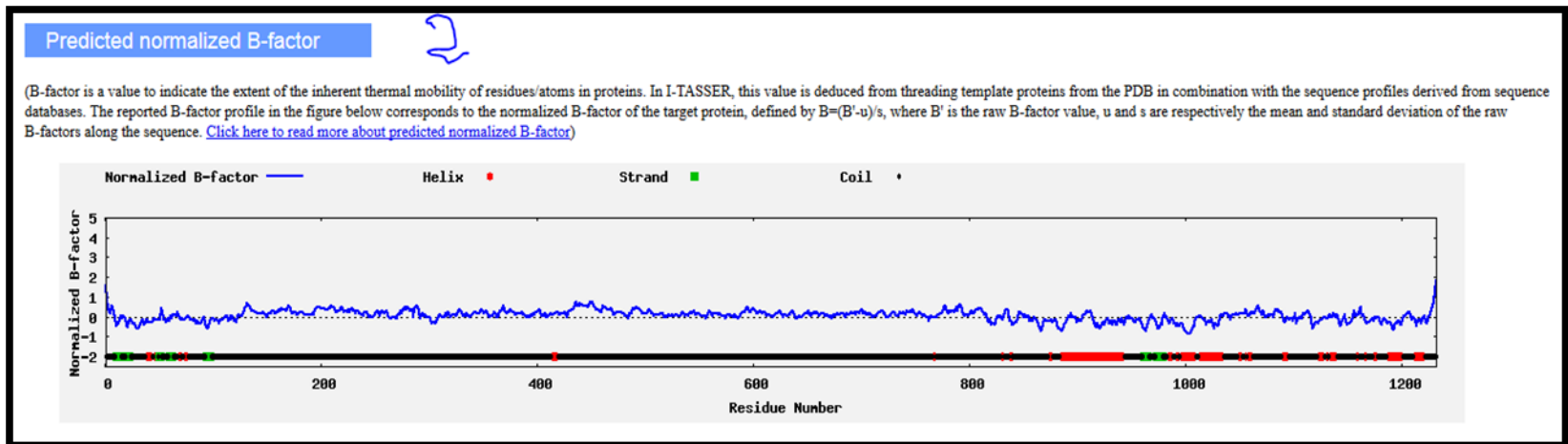
(b) Ranking of proteins is based on TM-score of the structural alignment between the query structure and known structures in the PDB library.

(c) RMSD^a is the RMSD between residues that are structurally aligned by TM-align.

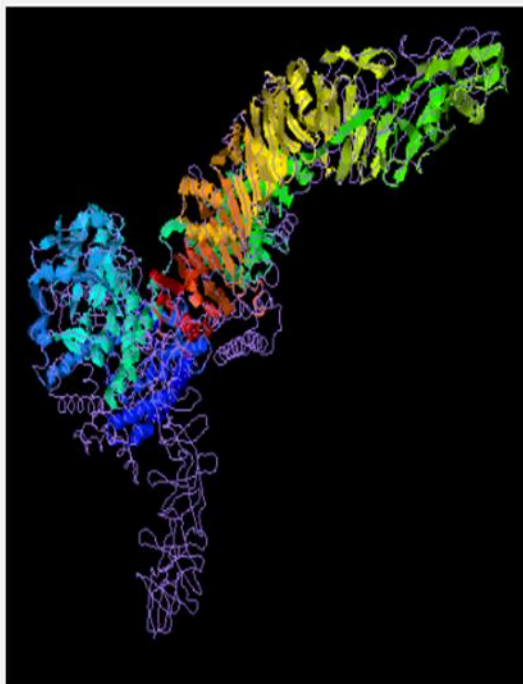
(d) IDEN^d is the percentage sequence identity in the structurally aligned region.

(e) Cov represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

Appendix VII: Protein prediction of *Pfap2g*



1201-2432



Top 10 Identified structural analogs in PDB

Click to view	Rank	PDB Hit	TM-score	RMSD ^a	IDEN ^a	Cov	Alignment
<input type="radio"/>	1	6ar6A	0.989	1.40	0.083	1.000	Download
<input type="radio"/>	2	4r04A	0.858	4.90	0.075	0.980	Download
<input type="radio"/>	3	6c0bA	0.359	3.17	0.078	0.383	Download
<input type="radio"/>	4	5v81A	0.295	10.30	0.038	0.481	Download
<input type="radio"/>	5	1w1A	0.292	9.72	0.028	0.455	Download
<input type="radio"/>	6	6h02A	0.289	9.80	0.035	0.458	Download
<input type="radio"/>	7	5d06A	0.287	9.62	0.028	0.444	Download
<input type="radio"/>	8	3ib9A1	0.280	9.97	0.027	0.445	Download
<input type="radio"/>	9	5xicA	0.279	9.87	0.040	0.439	Download
<input type="radio"/>	10	5u1sA	0.279	9.63	0.041	0.433	Download

(a) Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.

(b) Ranking of proteins is based on TM-score of the structural alignment between the query structure and known structures in the PDB library.

(c) RMSD^a is the RMSD between residues that are structurally aligned by TM-align.

(d) IDEN^a is the percentage sequence identity in the structurally aligned region.

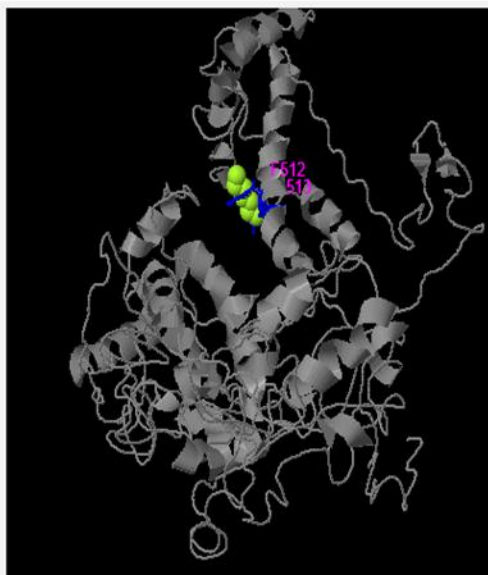
(e) Cov represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

Appendix VIII: Protein ligand docking of *Pfgdv1*

Predicted function using [COFACTOR](#) and [COACH](#)

(This section reports biological annotations of the target protein by COFACTOR and COACH based on the I-TASSER structure prediction. While COFACTOR deduces protein functions (ligand-binding sites, EC and GO) using structure comparison and protein-protein networks, COACH is a meta-server approach that combines multiple function annotation results (on ligand-binding sites) from the COFACTOR, TM-SITE and S-SITE programs.)

Ligand binding sites



Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Ligand Binding Site Residues
<input type="radio"/>	1	0.09	4	3e4aA	PEPTIDE	Rep. Mult.	15,27,32,33
<input type="radio"/>	2	0.05	2	2wk3A	PEPTIDE	Rep. Mult.	62,65,121,122,123,124,141,180,227
<input checked="" type="radio"/>	3	0.05	2	5d91A	8K6	Rep. Mult.	512,513
<input type="radio"/>	4	0.02	1	4wb2A	CA	Rep. Mult.	351,355
<input type="radio"/>	5	0.02	1	3d6aA	MG	Rep. Mult.	344,345

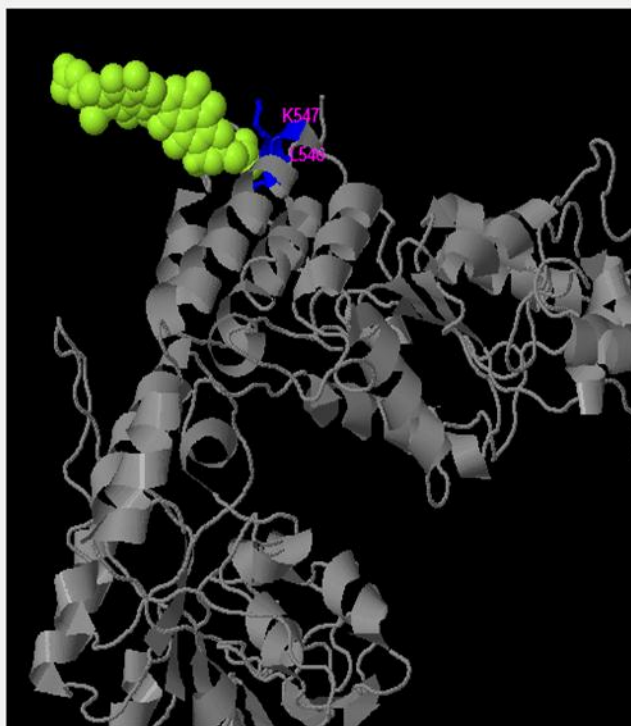
[Download](#) the residue-specific ligand binding probability, which is estimated by SVM.

[Download](#) the all possible binding ligands and detailed prediction summary.

[Download](#) the templates clustering results.

- (a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.
- (b) **Cluster size** is the total number of templates in a cluster.
- (c) **Lig Name** is name of possible binding ligand. Click the name to view its information in [the BioLiP database](#).
- (d) **Rep** is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the **Lig Name** column.
Mult is the complex structures with all potential binding ligands in the cluster.

COACH Results



Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Consensus Binding Residues
<input checked="" type="radio"/>	1	0.07	4	3b6aD	ZCT	Rep. Mult	546,547
<input type="radio"/>	2	0.04	2	4a0gD	PLP	Rep. Mult	412,413
<input type="radio"/>	3	0.04	2	2qmjA	NAG	Rep. Mult	5,16,22,23,173
<input type="radio"/>	4	0.04	2	3bz1H	CLA	Rep. Mult	585,588
<input type="radio"/>	5	0.04	2	1jgtB	MG	N/A	301,410
<input type="radio"/>	6	0.04	2	1khwB	MN	N/A	284,408
<input type="radio"/>	7	0.02	1	2z8yM	XE	Rep. Mult	129,190,213,229,230,233
<input type="radio"/>	8	0.02	1	3ufkA	NTA	Rep. Mult	8,10,165,166
<input type="radio"/>	9	0.02	1	1digB	LA	N/A	494,496,526
<input type="radio"/>	10	0.02	1	N/A	N/A	N/A	199,236,239,242,243,244,384,405,407,566,570,575,578

[Download](#) the residue-specific binding probability, which is estimated by SVM.

[Download](#) the predicted bound ligands and detailed prediction summary.

[Download](#) the templates clustering results.

- (a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.
- (b) **Cluster size** is the total number of templates in a cluster.
- (c) **Lig Name** is the name of possible binding ligand. Click the ligand name to view its information in [the BioLiP database](#).
- (d) **Rep** is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the **Lig Name** column. **Mult** is the complex structures with all potential binding ligands in the cluster.

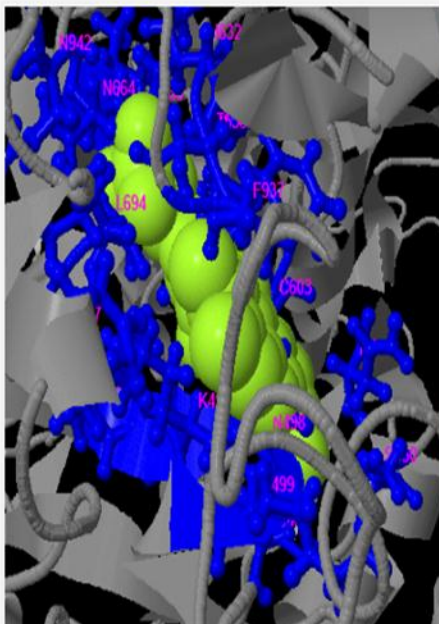
Appendix IX: Protein ligand docking of *Pfap2g*

Predicted function using COFACTOR and COACH

1-1200

(This section reports biological annotations of the target protein by COFACTOR and COACH based on the I-TASSER structure prediction. While COFACTOR deduces protein functions (ligand-binding sites, EC and GO) using structure comparison and protein-protein networks, COACH is a meta-server approach that combines multiple function annotation results (on ligand-binding sites) from the COFACTOR, TM-SITE and S-SITE programs.)

Ligand binding sites



Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Ligand Binding Site Residues
<input checked="" type="radio"/>	1	0.15	8	2uv8G	FMN	Rep. Mult	495,496,497,498,499,548,550,579,603,629,632,633,663,664,693,694,697,937,942
<input type="radio"/>	2	0.04	2	3bsnA	MN	N/A	201,202,275
<input type="radio"/>	3	0.04	2	2uvcl	NAP	Rep. Mult	498,520,523,550,555,634,635,741,824,893,895,896,914,915,917
<input type="radio"/>	4	0.02	1	N/A	N/A	N/A	596,598,621,986,987,988,993,1014,1017,1018,1021,1022,1024,1029,1032,1033,1035
<input type="radio"/>	5	0.02	1	1bofA	MG	N/A	41,189

[Download](#) the residue-specific ligand binding probability, which is estimated by SVM.

[Download](#) the all possible binding ligands and detailed prediction summary.

[Download](#) the templates clustering results.

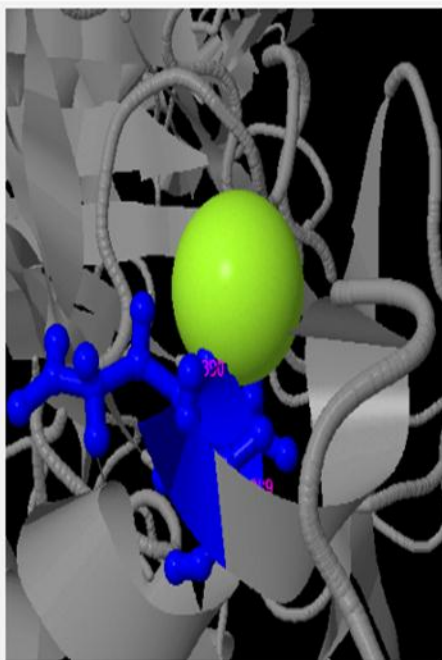
- (a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.
 (b) **Cluster size** is the total number of templates in a cluster.
 (c) **Lig Name** is name of possible binding ligand. Click the name to view its information in [the BiOLIP database](#).
 (d) **Rep** is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the **Lig Name** column.
Mult is the complex structures with all potential binding ligands in the cluster.

Predicted function using [COFACTOR](#) and [COACH](#)

1201-2432

(This section reports biological annotations of the target protein by COFACTOR and COACH based on the I-TASSER structure prediction. While COFACTOR deduces protein functions (ligand-binding sites, EC and GO) using structure comparison and protein-protein networks, COACH is a meta-server approach that combines multiple function annotation results (on ligand-binding sites) from the COFACTOR, TM-SITE and S-SITE programs.)

Ligand binding sites



Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Ligand Binding Site Residues
<input checked="" type="radio"/>	1	0.03	2	3ipvA	ZN	Rep. Mult.	389,390
<input type="radio"/>	2	0.03	2	3r6sA	CMP	Rep. Mult.	410,414
<input type="radio"/>	3	0.02	1	5ij5A	ZN	Rep. Mult.	400,407
<input type="radio"/>	4	0.02	1	2qeIC	CA	N/A	1188,1193
<input type="radio"/>	5	0.02	1	4qbFA	MG	Rep. Mult.	517,522

[Download](#) the residue-specific ligand binding probability, which is estimated by SVM.

[Download](#) the all possible binding ligands and detailed prediction summary.

[Download](#) the templates clustering results.

- (a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.
- (b) **Cluster size** is the total number of templates in a cluster.
- (c) **Lig Name** is name of possible binding ligand. Click the name to view its information in [the BioLiP database](#).
- (d) **Rep** is a single complex structure with the most representative ligand in the cluster; i.e., the one listed in the **Lig Name** column.
Mult is the complex structures with all potential binding ligands in the cluster.

Appendix X: Publication

Heliyon 6 (2020) e03453

CellPress

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Single-nucleotide polymorphism characterization of gametocyte development 1 gene in *Plasmodium falciparum* isolates from Baringo, Uasin Gishu, and Nandi Counties, Kenya



Josephat K. Bungei^{a,b}, Victor A. Mobegi^{b,*}, Steven G. Nyanjom^a

^a Department of Biochemistry, JKUAT, Kenya

^b Department of Biochemistry, School of Medicine, University of Nairobi, Kenya

ARTICLE INFO

Keywords:

Bioinformatics
DNA sequencing
Gene mutation
Biochemistry
Molecular biology
Parasitology
Plasmodium falciparum
Pfgdv1
Gametocytogenesis
Single-nucleotide polymorphism
Selection analysis

ABSTRACT

Introduction: *Plasmodium falciparum* relies on gametocytogenesis to transmit from humans to mosquitoes. Gametocyte development 1 (*Pfgdv1*) is an upstream activator and epigenetic controller of gametocytogenesis. The emergence of drug resistance is a major public health concern and this requires the development of new strategies that target the transmission of malaria. As a putative drug target, *Pfgdv1* has not been characterized to identify its polymorphisms and alleles under selection and how such polymorphisms influence protein structure.

Methods: This study characterized single-nucleotide polymorphisms (SNPs) in primary sequences ($n = 30$) of *Pfgdv1* gene generated from thirty blood samples collected from patients infected with *P. falciparum* and secondary sequences ($n = 216$) retrieved from PlasmoDB. ChromasPro, MUSCLE, Tajima's D statistic, SLAC, and STRUM were used in editing raw sequences, performing multiple sequence alignment (MSA), identifying signatures of selection, detecting codon sites under selection pressure, and determining the effect of SNPs, respectively.

Results: MSA of primary and secondary sequences established the existence of five SNPs, consisting of four non-synonymous substitutions (nsSNPs) (p.P217H, p.R398Q, p.H417N, and p.D497E), and a synonymous substitution (p.S514S). The analysis of amino acid changes reveals that p.P217H, p.R398Q, and p.H417N comprise non-conservative changes. Tajima's D statistic showed that these SNPs were under balancing selection, while SLAC analysis identified p.P217H to be under the strongest positive selection. Further analysis based on thermodynamics indicated that p.P217H has a destabilizing effect, while p.R398Q and p.D497E have stabilizing effects on the protein structure.

Conclusions: The existence of four nsSNPs implies that *Pfgdv1* has a minimal diversity in the encoded protein. Selection analysis demonstrates that these nsSNPs are under balancing selection in both local and global populations. However, p.P217H exhibits positive directional selection consistent with previous reports where it showed differential selection of *P. falciparum* in low and high transmission regions. Therefore, *in-silico* prediction and experimental determination of protein structure are necessary to evaluate *Pfgdv1* as a target candidate for drug design and development.

1. Introduction

Plasmodium falciparum, the leading causative species of malaria, is a protozoan parasite transmitted by female *Anopheles* mosquitoes when they bite humans to obtain their blood meal. Globally, malaria transmission remains undeterred with epidemiological data showing that malaria affected about 228 million people with approximately 405,000 deaths in 2018 (WHO, 2019). Kenya is one of the malaria-endemic countries in sub-Saharan Africa with highland areas in the North Rift

region, such as Nandi, Uasin Gishu, and Baringo Counties, experiencing epidemic malaria (Kipruto et al., 2017; Noor et al., 2018). Despite great strides made in the use of insecticides, anti-malarial drugs, and effective healthcare services, prevention and control strategies of malaria are still challenging due to the emergence of drug resistance (Arama and Troye-Blomberg, 2014; Delves et al., 2018; Muduli et al., 2018; Sinden et al., 2012; WHO, 2018). Hence, there is a need to develop new strategies that target the transmission of malaria in both epidemic and endemic regions.

* Corresponding author.

E-mail address: vatunga@uonbi.ac.ke (V.A. Mobegi).

<https://doi.org/10.1016/j.heliyon.2020.e03453>

Received 24 September 2019; Received in revised form 18 January 2020; Accepted 17 February 2020

2405-8440/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).