

**PREDICTION OF GENETIC VARIANTS ASSOCIATED WITH  
ANTIMALARIAL DRUG RESISTANCE USING SET COVERING  
MACHINE**

**APATA OLUWABUKOLA RACHEAL  
19PBF02173**

**MAY, 2022**

**PREDICTION OF GENETIC VARIANTS ASSOCIATED WITH  
ANTIMALARIAL DRUG RESISTANCE USING SET COVERING  
MACHINE**

**BY**

**APATA OLUWABUKOLA RACHEAL  
(19PBF02173)  
B.Sc (Hons) Microbiology, Obafemi Awolowo University**

**A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE  
STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE  
AWARD OF MASTER OF SCIENCE (M.Sc) DEGREE IN BIOINFORMATICS  
IN THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES,  
COLLEGE OF SCIENCE AND TECHNOLOGY COVENANT UNIVERSITY,  
OTA, OGUN STATE, NIGERIA**

**MAY, 2022**

## ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfilment of the requirements for the award of the Master's degree in **Bioinformatics** in the department of **Computer and Information Sciences**, College of Science and Technology, Covenant University, Ota.

**Mr. Taiwo B. Erewumi**  
(Secretary, School of Postgraduate Studies)

**Signature & Date**

**Prof. Akan B. Williams**  
(Dean, School of Postgraduate Studies)

**Signature & Date**

## **DECLARATION**

I, **APATA OLUWABUKOLA RACHEAL (19PBF02173)**, declare that this research was carried out by me under the supervision of Dr. Isewon I.M of Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. I attest that this thesis has not been presented either wholly or partly for the award of any degree elsewhere. All sources of data and scholarly information used in this thesis are duly acknowledged.

**APATA OLUWABUKOLA RACHEAL**

Signature & Date

## CERTIFICATION

We certify that the thesis titled “**PREDICTION OF GENETIC VARIANTS ASSOCIATED WITH ANTIMALARIAL DRUG RESISTANCE USING SET COVERING MACHINE**” is an original work carried out by **APATA OLUWABUKOLA RACHEAL, (19PBF02173)**, in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria, under the supervision of Dr. Isewon I.M. We have examined and found that the work acceptable for the award of Master’s degree in Bioinformatics.

**Dr. Itunuoluwa M. Isewon**  
(Supervisor)

Signature & Date

**Prof. Olufunke O. Oladipupo**  
(Head of Department)

Signature & Date

**Dr. Victor T. Odumuyiwa**  
(External Examiner)

Signature & Date

**Prof. Akan B. Williams**  
(Dean, School of Postgraduate Studies)

Signature & Date

## **DEDICATION**

I dedicate this thesis to God almighty and the lovelies that hold a special place in my heart - my mother, and siblings.

## ACKNOWLEDGEMENTS

I give all praise and glory to God the Father, the Son, and the Holy spirit for showers of blessings throughout my research work and grace for successful completion. I want to express my heartfelt gratitude to Dr. Itunuola Isewon, my research supervisor, for her selfless commitment in giving invaluable guidance during this research. Her dynamism, motivation, sincerity, and vision profoundly inspire me. I thank the management of Covenant University, CUBRe, and CApIC-ACE for the funding and support on this work. A profound thank you to the HoD, Associate Professor Oladipupo Grace, the PG coordinator, Dr. Oni and other faculty of Computer and Information Science depart for academic supports and contributions.

I am incredibly grateful to my lovely mother, my gold - Mrs. Apata Mobolanle J., for her love, prayers, care, and sacrifices to educate and prepare me for adulthood. Thank you, mummy. I cannot love you less. I appreciate my best half - Fabelurin Omololu O. for his love, understanding, prayers, and continuing support in the course of this research work. I express my appreciation to my brother – Olusegun Apata, sisters – Mobolaji, Afolakemi, Temitope and Oluwafunmilayo Apata, brothers-in-law, sisters-in-law and other relatives for their unwavering support and prayers. I extend my gratitude to my friends, co-research assistants and colleagues – Owolabi Jesujoba, Oluwamuyiwa Fesobi, Bisi-Adeniyi Titilayo, Molo Mbaso, Vingi Patrick, Enoma David, Oladejo David, Suraj Sadeeq for their constant support and companionship. I thank Prof. Jelili Oyelade for his support and words of encouragement. My Sincere appreciation also goes to the Admin staffs of CUBRe and CApIC-ACE – Mr. Babajide, Miss Helen, Mr. Obaoye. Finally, I want to express my gratitude to everyone who has helped me complete the research work, whether directly or indirectly. God bless you all.

## TABLE OF CONTENTS

CONTENTS	PAGES
<b>COVER PAGE</b>	i
<b>TITLE PAGE</b>	ii
<b>ACCEPTANCE</b>	iii
<b>DECLARATION</b>	iv
<b>CERTIFICATION</b>	v
<b>DEDICATION</b>	vi
<b>ACKNOWLEDGEMENTS</b>	vii
<b>TABLE OF CONTENTS</b>	viii
<b>LIST OF TABLES</b>	xi
<b>LIST OF FIGURES</b>	xii
<b>ABBREVIATIONS</b>	xv
<b>ABSTRACT</b>	xvii
<b>CHAPTER ONE: INTRODUCTION</b>	
1.1 Background Information	1
1.2 Statement of the Problem	5
1.3 Research Questions	8
1.4 Aim and Objectives of the Study	8
1.5 Research Methodology	9
1.6 Significance of the Study	13
1.7 Scope of the Study	13
1.8 Limitation of the Study	14
1.9 Organization of the dissertation	14
<b>CHAPTER TWO: LITERATURE REVIEW</b>	
2.1 Introduction	15
2.2 Malaria Life Cycle	15
2.2.1 Parasite	15
2.2.2 Malaria Vector	16
2.2.3 Human Host	17
2.3 Antimalarial drug action mechanisms	19
2.3.1 Antimalarial Drugs Resistance Development	20
2.3.2 Molecular mechanism of resistance	22
2.4 Assessing Antimalarial Drug Efficacy and Resistance	23
2.4.1 In vivo testing	23
2.4.2 In vitro testing	25
2.4.3 Identification of Molecular/Genetic Markers of Antimalarial Drug Resistance	26
2.5 Genome-Wide Association Study	27
2.5.1 Microbial Genome-Wide Association Study	27
2.5.2 Association study of plasmodium falciparum antimalarial drug resistance	28



2.6	Types of variation in microbes	31
2.6.1	INDELs and SNPs	31
2.6.2	Gene Insertion and Deletion	31
2.6.3	Copy Number Variation and Sequence Inversions	32
2.7	Phenotype	32
2.8	Machine learning-based phenotype prediction	33
2.8.1	Why kmer-based method?	34
2.8.2	Advantages kmer	34
2.8.3	Algorithms for k-mer-based prediction	36
2.8.4	Benefits of biomarkers identification	37
2.9	Related Works	38
<b>CHAPTER THREE: METHODOLOGY</b>		
3.1	Introduction	41
3.2	Data	41
3.3	Data Preprocessing	43
3.3.1	Quality Control and Assessment	43
3.3.2	Genome Assembly	43
3.3.3	Data encoding and class label curation	44
3.4	Model Selection	46
3.5	Methods	46
3.5.1	Model	46
3.6	Cross-Validation	50
3.7	Performance Evaluation	51
3.7.1	ROC Curve	51
3.7.2	Area Under the ROC Curve	51
3.7.3	F1 Score	51
3.8	Interpreting Significant K-mers	52
3.9	Study Workflow	52
<b>CHAPTER FOUR: RESULTS</b>		
4.1	Data Exploration and Preprocessing	54
4.2	Evaluation of predictive performances of k-mer-based phenotype prediction tools on <i>Klebsiella pneumonia</i> data	61
4.2.1	Implementation on <i>Klebsiella pneumoniae</i> data	61
4.3	Implementing k-mer-based phenotype prediction algorithm on <i>Plasmodium falciparum</i> data	65
4.3.1	Chloroquine Resistance	65
4.3.2	Dihydroartemisinin Resistance	68
4.3.3	Lumafantrine Resistance	71
4.3.4	Mefloquine Resistance	72
4.3.5	Primaquine Resistance	73
4.3.6	Pyrimethamine Resistance	75
4.4	Biological Interpretation of the K-mer Variants Predicted to be Linked to <i>Plasmodium falciparum</i>	80
4.4.1	Analysis and interpretation of k-mers associated with <i>Plasmodium</i>	

<i>falciparum</i> chloroquine resistance	81
4.4.2 Analysis and interpretation of k-mers associated with <i>Plasmodium falciparum</i> dihydroartemisinin resistance	85
4.4.3 Analysis and interpretation of k-mers associated with <i>Plasmodium falciparum</i> pyrimethamine resistance	87
<b>CHAPTER FIVE: CONCLUSION AND RECOMMENDATION</b>	
5.1 Summary	94
5.2 Conclusion	94
5.3 Contribution to knowledge	95
5.4 Recommendation	96
<b>REFERENCES</b>	97
<b>APPENDIX</b>	108

## LIST OF TABLES

<b>TABLE</b>	<b>TITLE OF TABLES</b>	<b>PAGES</b>
1.1	Summary Of Objectives And Methodology	12
2.1	Major categories of antimalarial drugs based on the mode of action and chemical constituents	19
2.2	Widely used derivatives of artemisinin	22
3.1	List of genomic and pheotypic data used	42
3.2	Resistance threshold used for conversion of continuous to binary resistance phenotype	45
4.1	Important K-mers reported in Kover	77
4.2	Identified known genes and variants associated with antimalarial drug resistance	88
4.3	Identified novel genes and variants associated with antimalarial drug resistance	89

## LIST OF FIGURES

<b>FIGURE</b>	<b>TITLE OF FIGURES</b>	<b>PAGES</b>
2.1	Life Cycle of Malaria Parasite	18
2.2	Drug efficacy and resistance monitoring approaches	24
2.3	Illustration of how k-mers are generated from genomic sequences of individual parasite	35
3.1	Genome Assembly Pipeline	44
3.2	A pipeline for SV/k-mer based phenotype prediction	53
4.1	Bar plot of phenotype availability for the different drugs	55
4.2	Venn diagram quantifying the number of instances of co-occurrence of resistance between drugs.	56
4.3	Per base sequence quality report visualization before trimming and cleanup	57
4.4	Per sequence quality score visualization before trimming and cleanup	58
4.5	Per base sequence quality report visualization after preprocessing	59
4.6	Per sequence quality score visualization after preprocessing	60
4.7	Confusion matrix of pyseer, phenotypeseeker, kover polymyxin resistance prediction	62
4.8	Bar chart of performance metric scores of model on bacteria data ( <i>Klebsiella pneumoniae</i> )	63
4.9	Comparison of performance of Kover, PhenotypeSeeker, and Pyseer in predicting polymyxin resistance and susceptibility from <i>Klebsiella pneumonia</i> whole genome sequence. A. AUROC curve B. Precision-recall curve.	64
4.10	Bar chart of performance metric scores of models on <i>Plasmodium falciparum</i> data (Chloroquine Resistance)	66
4.11	Comparison of performance of Kover, PhenotypeSeeker in predicting chloroquine resistance and susceptibility from <i>Plasmodium falciparum</i> whole genome sequence. AUROC curve (left), Precision-recall curve (right)	67

4.12	Bar chart of performance metric scores of model cross validation and testing on Plasmodium falciparum data (Dihydroartemisinin Resistance)	69
4.13	Comparison of performance of Kover, PhenotypeSeeker in predicting dihydroartemisinin resistance and susceptibility from Plasmodium falciparum whole genome sequence. AUROC curve(left), Precision-recall curve(right)	70
4.14	Comparison of performance of Kover, PhenotypeSeeker in predicting lumefantrine resistance and susceptibility from Plasmodium falciparum whole genome sequence. AUROC curve (left), Precision-recall curve (right)	71
4.15	Comparison of performance of Kover, PhenotypeSeeker in predicting mefloquine resistance and susceptibility from Plasmodium falciparum whole genome sequence. AUROC curve (left), Precision-recall curve (right)	72
4.16	Comparison of performance of Kover, PhenotypeSeeker in predicting primaquine resistance and susceptibility from Plasmodium falciparum whole genome sequence. AUROC curve (left), Precision-recall curve (right)	74
4.17	Bar chart of performance metric scores of model cross validation and testing on Plasmodium falciparum data (Pyrimethamine Resistance)	75
4.18	Comparison of performance of Kover, PhenotypeSeeker in predicting pyrimethamine resistance and susceptibility from Plasmodium falciparum whole genome sequence. AUROC curve (left), Precision-recall curve (right)	76
4.19	Comparison of performance of Kover in the prediction of Plasmodium falciparum resistance and susceptibility to the six antimalarial drugs	78
4.20	Comparison of performance of Kover in the prediction of Plasmodium falciparum resistance and susceptibility to the six antimalarial drugs. AUROC curve (left), Precision-recall curve (right)	79

4.21	Visualization of the alignment of the consensus k-mers sequence to the reference genome ( <i>Plasmodium falciparum</i> 3D7 ASM276v2 genome sequences)	80
4.22	Visualization of high-scoring segment pair distribution on the 14 chromosomes of reference genome	81
4.23	Hits generated from each query consensus sequence	82
4.24	Alignment with the <i>cg1</i> gene region	84
4.25	Alignment with nonsynonymous SNPs in the <i>var</i> gene region	84
4.26	Mismatches (SNPs) in alignment with <i>Plasmodium falciparum</i> erythrocyte protein (PfEMP) 1 gene	84
4.27	Alignment with the erythrocyte membrane protein gene region showing various haplotypes	86

## ABBREVIATIONS

ABBREVIATION	MEANING
ACT	Artemisinin-based Combination Therapy
ADR	Antimicrobial Drug Resistance
AMR	Antimalarial Resistance
BP	BasePair
CNV	Copy-Number Variation
CQ	Chloroquine
DHA	Dihydroartemisinin
DHFR	Dihydrofolate reductase-thymidylate synthase
DHPS	Dihydropteroate synthase
DNA	Deoxyribonucleic Acid
ENA	European Nucleotides Archive
GLM	Generalized Linear Model
GWAS	Genome-Wide Association Study
IC50	50% inhibitory concentration
ICEs	Integrative and Conjugative Elements
INDELs	Insertions and Deletions
ISs	Insertion Sequences
Kb	Kilobases
LUM	Lumefantrine
MDR	Multidrug Resistance
MGEs	Mobile Genetic Elements
MQ	Mefloquine

MTB	<i>Mycobacterium tuberculosis</i>
NCBI	National Center for Biotechnology Information
PFCRT	Plasmodium falciparum chloroquine resistance transporter gene
PFMDR1	Plasmodium Falciparum Multidrug Resistance
Pgh1	P-glycoprotein homologue 1
PQ	Primaquine
PVL	Panton-Valentine Leucocidin
PYR	Pyremethamine
SCM	Set Covering Machine
SIs	Sequence Inversions
SNP	Single Nucleotide Polymorphism
TESs	Therapeutic Efficacy Studies
WHO	World Health Organization



## ABSTRACT

Antimalarial resistance (AMR) has become a major issue in malaria-endemic countries, and novel methods for identifying strains resistant or susceptible to specific medications are critical in the fight against antimalarial-resistant *Plasmodium* parasites. The growing availability of genetic information has enabled the application of computational methods in surveying resistance patterns. K-mer-based machine learning approaches have shown considerable potential as a diagnostic and research tool. In this work, Set Covering Machine (SCM) algorithm was applied to predict antimalarial drug response outcomes and their genetic determinants. The model predicted six antimalarial drugs (Chloroquine, Dihydroartemisinin, Lumafantrine, Primaquine, Pyrimethamine, and Mefloquine) response phenotype in *Plasmodium falciparum*. The model used the most compact set of k-mers generated from the genomes of the parasite isolates to learn and predict binary drug response outcomes. To avoid model overfitting, ten-fold cross-validation was conducted on the training set to choose the optimal hyperparameter values. Regardless of the resistance mechanism, whether acquired resistance or point mutations in the chromosome, the training accuracy (mean cross-validation score) and testing accuracy of SCM prediction of the six antimalarial drug resistance was above 85%. The model significantly classified the resistant isolates from the sensitive isolates of the parasite and could be used as potential tools in antimalarial resistance surveillance and clinical studies. A number of sequence k-mers associated with antimalarial drug resistance were identified. We identified several already known genes and loci associated with the six drugs, including those containing *pfcr* and *pfdhfr*. Novel genes and loci were also discovered. Of particular interest are the variant regions on the var genes on chromosomes 6, 8, 10, and 13 containing the *Plasmodium falciparum* erythrocyte membrane protein 1 (*PfEMP1*). The *PfEMP1* variant k-mers were found to be associated with chloroquine, dihydroartemisinin, and pyrimethamine resistance. The var genes encode PfEMP1. The genes have extreme variability and are a principal virulence factor of malaria parasite with extreme antigenic variability. The variations in these var genes were found to play a role in antimalarial drug resistance in *P. falciparum*.

**Keywords:** *Machine learning, Malaria, Plasmodium falciparum, Genome-Wide Association Study, Phenotype prediction*