Functional Genome Wide Association Study in Susceptibility and Resistance of Malaria

A Masters Dissertation

By

Etienne Ntumba Kabongo

Division of Human Genetics

Faculty of Health Sciences (FHS)

kbnntu002@myuct.ac.za

Supervisor: Prof. Emile R. Chimusa

Email Address: emile.chimusa@uct.ac.za

February 2021

Contents

1 INTRODUCTION **4**

1.1 **Historical and Overview** . . . 4

1.2 Life Cycle of Malaria Parasite . . . 5

1.3 **Genetic Susceptibility and Resistance of Malaria** . . . 6

1.3.1 Overview . . . 6

1.3.2 Gene mutations involved in susceptibility and resistance to *Plasmodium falciparum Malaria* . . . 7

1.4 **Main Concepts in Genetic Variation** . . . 10

1.4.1 Mutations . . . 10

1.4.2 Genetic Distance (F_{st}) . . . 11

1.4.3 Fixation Index (F_{ST}) . . . 13

1.4.4 Single Nucleotide Polymorphism (SNP) . . . 13

1.4.5 Gene Mapping . . . 14

1.4.6 Linkage Disequilibrium (LD) . . . 14

1.4.7 Minor allele frequency (MAF) . . . 15

1.4.8 TagSNP . . . 15

1.5 **Genome Wide Association Study (GWAS) in Malaria** . . . 16

1.5.1 **Overview of Genome Wide Association Study (GWAS)** . . . 16

1.5.2 Application of GWAS in Severe Malaria of African's Population . . . 17

1.6 **Advantage and limitations of Genome Wide Association Study (GWAS)** . . . 22

1.7 Overview of Post-GWAS . . . 22

1.7.1 Overview of Functional Genome Wide Association Study (FGWAS) . . . 23

1.7.2 Motivation . . . 23

1.7.3 **Objectives** . . . 24

2 Mathematical Approaches used in Functional Genome wide Association (FGWAS) 25

2.1 Overview . . . 25
2.2 Linear Mixed Model for Genetic Association . . . 26
2.3 Bayesian Lasso GWAS Model . . . 32
2.4 Mathematical Approach applied in Functional Genome wide association (FGWAS) 34
2.4.1 Bayes Factors . . . 34
2.4.2 Hierarchical Model . . . 35
2.5 Meta-Analyses of Genome-Wide Association Studies . . . 37
2.5.1 Introduction . . . 37
2.5.2 Aim of Using GWAS Meta-Analysis . . . 37
2.5.3 Fixed Effect (FE) Method . . . 38
2.5.4 Random Effect Method . . . 40
2.5.5 Heterogeneity . . . 41
2.5.6 Binary Effect Method . . . 42
2.5.7 Estimate M-value . . . 42
2.5.8 Tools for Conducting GWAS Meta-Analysis . . . 44
2.6 SKAT-Analysis(Sequence Kernel Association Tests) for the Combined Effect of Rare and Common Variants . . . 49
2.6.1 Fisher's Combination Method . . . 51

3 Materials and Methods 52

3.1 Genome Wide Association Study Analysis . . . 54
3.1.1 Quality Control (QC) . . . 54
3.2 Cross Populations Meta-Analysis . . . 55
3.3 Functional Genome Wide Association Study (FGWAS) Analysis . . . 56
3.3.1 Functional Genomics . . . 56
3.3.2 Explanation of Analysis . . . 57
3.4 Functional mapping and annotation of genetic associations (FUMA) . . . 59
3.5 GeneSet and Enrichment Analysis Using GENEMANIA . . . 59
3.6 Data Analysis for Rare and Common Variants . . . 60
3.7 Polygenic Risk score Analysis (PRS) . . . 61
3.7.1 Base Phenotype Dataset . . . 62
3.7.2 Target Phenotype Data set . . . 62

4 Results **65**

4.1 GWAS Results . . . 65

4.1.1 Quality Control (QC) . . . 65

4.1.2 Biological Functions of the identified variants . . . 70

4.2 Results from Cross Meta-Analysis . . . 72

4.2.1 Metal . . . 72

4.3 Results from Gene and Pathway-based Association using PASCAL tool . . . 76

4.4 Result from FGWAS Analysis . . . 79

4.5 Result from Pathway Analysis, Enrichment Analysis and Functional Mapping and Annotation using FUMA . . . 84

4.6 Results from Rare Variant Association Analysis using SKAT . . . 90

4.6.1 Pathway Analysis and Enrichment Analysis . . . 93

4.7 PRS Result . . . 98

4.7.1 High-Resolution Polygenic Risk Scoring . . . 98

5 General Discussion and Conclusion **105**

5.1 Discussion . . . 105

5.2 Potential Impact of study . . . 109

5.3 Limitations and future work . . . 110

List of Figures

1.1 The life cycle of malaria parasites is presented as follows: First, sporozoites enter the bloodstream, and migrate to the liver. They infect liver cells, where they multiply into merozoites, rupture the liver cells, and return to the bloodstream. The merozoites infect red blood cells, where they develop into ring forms, trophozoites and schizonts that in turn produce further merozoites. Sexual forms are also produced, which, if taken up by a mosquito, infect the insect and continue the life cycle. *"Life Cycle of the Malaria* curtoisie of *"National Institutes of Health (NIH)"* [1]. . 6

1.2 This illustres the **Mutation** during the cellular division 11

1.3 This illustrate Genetic distance between european population, the European genetic structure (based on 273,464 SNPs). Three levels of structure as revealed by PC analysis are shown. https://www.reddit.com/genetic_distance 11

1.4 Represent Single Nucleotide Polymorphism (SNP) plays a role in a wide variety of diseases such as sickle cell anemia and cystic fibrosis Nucleotide 13

1.5 Genetic maps have been used successfully to find the gene responsible for relatively rare, single-gene inherited disorders such as cystic fibrosis and Duchenne muscular dystrophy. Genetic maps are also useful in guiding scientists to the many genes that are believed to play a role in the development of more common disorders such as asthma, heart disease, diabetes, cancer, and psychiatric conditions [2]. 14

1.6 LD plot of SNPs with top-ranked bayes factors in CHB of 1000 Genome Phase I. The colors indicate the strength of pairwise LD according to r2 metrics. The SNPs marked with asterisks represent independent strong associations. Tag SNPs are shadowed in pink) 16

1.7 https://journals.plos.org/plosgenetics/article/figure?id=10.1371/journal.pgen.1007172.t001 20

3.1 Workflow of Analysis . . . 53

4.1 From left to right, we have qqplot of Kenya, Gambia and Malawi based on population stratification . . . 66
4.2 Only 5 variants SNPs have passed the cut-off of p-value < 5e-07 . . . 68
4.3 Manhattan plot of Gambia GWAS based on his Summary Statistic . . . 68
4.4 Kenya Summary statistics Manhattan plot . . . 70
4.5 Relevant information of *"C4orf19" and "PAM"* in cell type tissues . . . 71
4.6 Heatmap of expression for *"C4orf19"* and *"PAM"* . . . 72
4.7 Mahanatthan plot for Meta-Analysis find from Metal . . . 72
4.8 The above figures represent the forestplot 3 top SNPs found in GWAS Meta-Analysis. We have: **rs1479764** mapped to the gene **C4orf19**, **rs12700120** mapped to the gene *IQCE* and rs2059278 mapped to the gene *CDH13* describes . . . 74
4.9 Gene expression in different cell types and tissues type specificity of each gene. . . . 78
4.10 Forestplot of each annotation . . . 80
4.11 Network analysis of genes from based on fGWAS result . . . 85
4.12 Gene expression heatmap of our 14 genes. . . . 86
4.13 Differentially expressed genes of tissues specificity, Pathway scores obtained from pascal analysis. Significant Pathways at Bonferroni corrected P-value $\leq$ 0.05 are coloured in red . . . 87
4.14 Curate gene set . . . 87
4.15 Chemical and Genetic perturbation of our gene set . . . 88
4.16 GWAS Catalogue based on our genes . . . 88
4.17 BarGraph result from fGWAS results based on GWAS Catalog 2019 . . . 89
4.18 Genes ordering in term of p-values related to trait . . . 90
4.19 QQplot . . . 92
4.20 Genes mapped by rare variants and their pathways. . . . 95
4.21 The tissue specificity. . . . 97
4.22 Gene expression heatmap. . . . 98
4.23 Summary of Barplot, recorded High resolution of Risk predictive of Severe Malaria, the best PRS model fit has R^2=0.00443458 . . . 99
4.24 This plot summarizes High-resolution PRS for 'Host having resistance to Malaria' predicting Severe status. The high-resolution best-fit PRS is 0.00443458 at P_T = 0.00165005. . . . 100

4.25 This PRS Strata plot provide the information of each polygenic score interval and the odds ratio for score on Severe Malaria. 100
4.26 Summary of Barplot, recorded High resolution of Risk predictive of Severe Malaria, the best PRS model fit has r^2=0.784666 101
4.27 This plot summarizes High-resolution PRS for 'Host having resistance to Malaria' predicting Severe status. The high-resolution best-fit PRS is 8.16271e-158 at P_T = 1 102
4.28 This PRS Strata plot provide the information of each polygenic score interval and the odds ratio for score on Severe Malaria 102
4.29 Summary of Barplot, recorded High resolution of Risk predictive of Severe Malaria, the best PRS model fit has r^2=0.145282 103
4.30 This plot summarizes High-resolution PRS for 'Host having resistance to Malaria' predicting Severe status. The high-resolution best-fit PRS is 1.51515e-55 at P_T = 1. . 104
4.31 This PRS Strata plot provide the information of each polygenic score interval and the odds ratio for score on Severe Malaria. 104

List of Tables

1.1 Genetic mutations involved in susceptibility -resistance to *Plasmodium falciparum Malaria*, with SMA: severe malaria anaemia; CM: cerebral Malaria; Hb: Haemoglobin; HbAS: Haemoglobin AS or sickle cell trait; G6PD: glucose-6-phosphate dehydrogenase; SM: Severe Malaria, UM : Uncomplicated Malaria 8

1.2 Tanzania Study Population https://journals.plos.org/plosgenetics/article/figure?id=10.1371/journal.pgen.1007172.t001 19

1.3 GWAS Result 19

1.4 Reporting Genes find out across 11 African populations [3] 21

2.1 Key formulae for both approaches [4] 45

2.2 Summary of Meta-Analysis tool with their specificity [5]. 47

3.1 Represent the raw-data collected from European Phenome Genome Archive 53

3.2 QC Result after filtering 55

3.3 Represent the raw data collected from MalariaGen, used as Target phenotype data set for each PRS analysis 63

3.4 Represent the raw data collected from MalarianGen, used as Target phenotype data set for each PRS analysis 63

4.1 Summary Statistics of significant SNPs 67

4.2 SNPs to Genes based on Biocard 67

4.3 We find out 9 SNP variants passed the threshold 69

4.4 Significant variants mapped onto the Gene-level 69

4.5 We got 3 SNPs variants have passed the threshold of 5e-07 72

4.6 We identify only one gene *C4orf19* mapped by these SNP based on significant SNP threshold 5e-07 73

4.7 We find out 20 SNP variants passed p-value threshold 5e-07 and m-value >0.8 . . . 75
4.8 We find out 22 genes variants passed the threshold with p-value <5e-0.8 77
4.9 This table represents the likelihood of each annotation among 17 annotations selected (without *tss_dist* and Brain Microvascular) 79
4.10 Represent each annotation with the respect effect. 80
4.11 Table represented the likelihood of each annotation among 17 annotations selected (without *tss_dist* and Brain Microvascular) 81
4.12 In this table the values of cross validation likelihood penalized by 0.05,0.1,0.15,...,0.65 using C as annotation. 82
4.13 We have 29 significant SNPs with PPA > 0.9. 83
4.14 Top 7 genes having high score. 84
4.15 Summary statistics of our genes and phenotype related. 89
4.16 We have filtered SKAT output with p-values less then 0.05 91
4.17 Different functions and their FDR. 93
4.18 Significant Pathways for these genes. 95
4.19 Top score of 6 genes strongly associated. 96

Acknowledgments

Foremost, I'm going to express my sincere thanks to my supervisor Prof. **Emile R Chimusa** for the continuous support of this thesis for his motivation, patience, enthusiasm, immense knowledge involved and availability. I couldn't have imagined this kind of passion and this better supervision during this long journey.

Besides my supervisor, my sincere thanks are adressed to my mentor Dr. **Anita GHANSAH** for the encouragement, insightful comments, correction and hard questions.

My sincere thanks also goes to Dr.**Delesa Damena** for his strong technical assistance, guidance, serious correction and mental support,I'll never forget your willingness to help me, to encourage me, to believe in my ability.

I thank my fellow group mates in Emile's Group for the stimulating discussions, corrections, the sleepless nights we were working together, before deadlines, and for all the fun we have had during these 2 years.

My sincere thanks goes to the Developing Excellence in Leadership and Genetics Training for Malaria Elimination in sub-Saharan Africa (**DELGEME-AFRICA**) for the financial support and several training.

My sincere thanks to my division of **Human Genetics** for this high collaboration, training and supporting during this long journey.

Many thanks for the encouragements, spiritual support during this long journey

Abstract

Background: More than century, malaria is qualified as a mortal infectious disease, worldwide causing high morbidity and mortality. The World Health Organization (WHO) has shown that, Distribution of Malaria in Africa takes a major part, it's accounting for 95% (about 229 million) and 67% (about 274000) of reported cases and death respectively. One of solutions for reducing this threat is to find drugs or to develop vaccines which can resist and adapt to populations. Unfortunately, despite several efforts, malaria parasites are still developing resistance to the frontline antimalarials.

Objectives: Our aim in this project is to conduct a systematic Meta-analysis and various functional analysis across three study populations in Africa (Kenya, Malawi and Gambia).

Method and Materials: Our first analysis is directed to the Genome Wide Association Study (GWAS) of three study populations (Kenya, Malawi and Gambia) using the Emmax tool to identify the genetic variants associated with severe malaria. We then conducted GWAS based meta-analysis on the summary statistics from the three studies using Metasoft and Metal. Further, we implemented Functional GWAS (FGWAS) to re-weight the GWAS meta-analysis using functional genomic information software (fgwas-tool). Using results from fgwas-tool, we performed biological interpretation using Functional Mapping (FUMA) tool. We mapped the significant SNPs to the genes, and elucidated their functions and their associated cell types. We then performed pathway analysis and enrichment analysis of the genes using Genemania and Enrichr.
Additionally, we performed a polygenic risk score for individuals in each study population using PRSice, and evaluated the level of risk exposure for each individual based on the best predictive threshold. Finally, we filtered the rare variants from each study, and performed SKAT analysis to aggregate the effect of the rare variants

Results: We identified 29 significant SNPs (14 replicates and 15 novels) reweighted from FGWAS

based on GWAS Meta-Analysis. The SNPs mapped to 15 genes *(HBB, HBD, ATP2B4, ABO, CBLB, EYA2, HERPUDI, IQCJ, MPP7, NAVI, NUP210, SAMD5 , TCERG1L ,TMEM229B, C4orf19)* at gene level. Five of these genes *(HBB, HBD, ATP2B4, ABO, CBLB)* had been reported by different studies to be associated with malaria. In the PRS analysis we have shown the best prediction based on the best threshold estimated of each population. We found best-fit prediction best-fit PRS for Gambia is 0.00443458 at P_T = 0.00165005, for Kenya is 8.4666e-158 at P_T= 1 and for Malawi is 1.5151e-55 at P_T = 1 predict the risk of an infectious disease like severe malaria. However, the prediction rate is very low and may fail to distinguish the cases from the controls.

Conclusion: The functional analysis based on fgwas result have shown that 5 genes *(ATP2B4, ABO, HBD, HBB, CBLB)* are highly associated to malaria across these 3 studies populations (Gambia, Malawi and Kenya) and 10 candidate novel genes, including high number of mutations in the gene C4orf19 which will constitute one of the future major studies. Also, we have shown the best prediction based on the best threshold estimated of each population. The results have shown that the prediction rate is very low and may fail to distinguish the cases from the controls.

Chapter 1

INTRODUCTION

1.1 Historical and Overview

More than a century, malaria has been qualified as a mortal infectious diseases , the major factor leading to the causality of this disease is the infection of protozoan parasites including the genus *Plasmodium*. In 1880, the parasite in the blood of malaria patients was discovered by Alphonse Laveran, hence the study of malaria became clear and relevant [6]. While, in 1897 the birds infected by *Plasmodium relictum* and the cycle of transmission in culicine mosquitoes were elucidated by Ronald Ross [7].

In 1898, researchers such as Giovanni Battista Grassi and other malarialogists discovered that mosquitoes (especially Anophelines) can be vectors of human malaria [8]. In 1948, Cyril Garnam and Henry Shortt showed that the malaria parasite is initially developed in the liver and finally enters the bloodstream [9]. However, Wojciech Krotoski has shown that the last stage of life cycle is the dormant stages in the liver [10]. The wide spread and damages of malaria are claimed as a serious global health attack. More than a million of human lives are exposed to malaria [11]. According to the World Health Organization(WHO), the african region represents the largest part of malaria cases in the world. Recent statistics (2019) estimated 229 million cases of malaria worldwide, where 94% of malaria cases and deaths in Africa. Therefore, malaria is by far the leading cause of death in sub-Saharan African countries. Children under 5 years are highly exposed to malaria; they accounted for 67% (274 000) for global malaria deaths, and also malaria claims the lives of many pregnant women [12]. Due to the increasing death toll caused by malaria,there are attempts to reduce this burden. However, malaria parasites keep on developing resistance to the antimalaria drugs [13, 14, 15].

To identify mutations that are responsible for drug resistance in malaria parasites, research has proven that genetic mapping is one of the most powerful tools. Moreover, with a large number of SNPs(single nucleotide polymorphisms) based on high throughput genotyping, a method was developed and is available: genome wide association studies(GWASs) [16]. As stated by [16], malaria in human beings is caused by *Plasmodium parasites* such as:

- *Plasmodium vivax*
- *Plasmodium malariae*
- *Plasmodium knowlesi*
- *Plasmodium ovale*
- *Plasmodium falciparum*

Where *Plasmodium vivax* and *Plasmodium falciparum* are the most widespread and the deadliest respectively[17, 18, 19].

1.2 Life Cycle of Malaria Parasite

The life cycle of malaria parasites is observed in the host and Anopheles mosquitoes. Each infection produces thousands of antigens(Proteins) in the human immune system, which explains its large capacity and complexity [1]. The focus of this study is based on life cycle stages.
During the life cycle of *Plasmodium*, Anopheles (primary host) acts as a vector by transmitting sporozoite to secondary host (human). The transmitted *sporozoite* moves through blood vessels to hepatocytes which are liver cells, where thousands of *merozoites* are produced asexually [20]. The *merozoite* initiates a serie of asexual multiplication cycles to yield 8 to 24 new infective merozoites resulting in cells explosion to begin a new infective cycle [20].
Immature gametocytes are developped by merozoites. During a mosquito's bite to an infected person, immature gametocytes in the blood mature in the mosquito's gut and the male and female gametocytes interact to form *ookinete* [21].
New sporozoites are formed from *ookinete* and they move to the mosquito's salivary glands therefore, available for a new infection [21].
Somewhere, the malaria parasite is uncommonly shared by transfusion. . We can summarize these steps of life cycle malaria parasites on the Figure 1.1 here below.

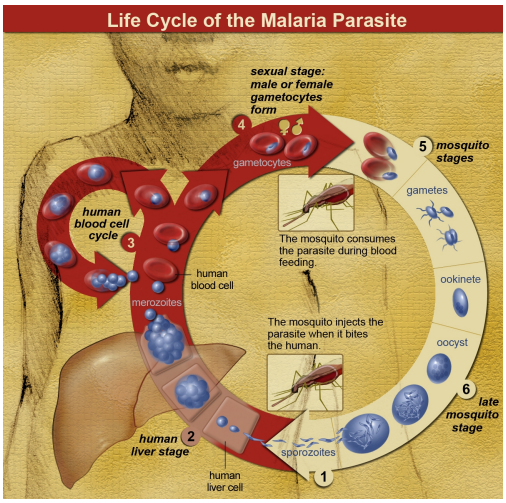


Figure 1.1: *The life cycle of malaria parasites is presented as follows: First, sporozoites enter the bloodstream, and migrate to the liver. They infect liver cells, where they multiply into merozoites, rupture the liver cells, and return to the bloodstream. The merozoites infect red blood cells, where they develop into ring forms, trophozoites and schizonts that in turn produce further merozoites. Sexual forms are also produced, which, if taken up by a mosquito, infect the insect and continue the life cycle. "Life Cycle of the Malaria curtoisie of "National Institutes of Health (NIH)" [1].*

1.3 Genetic Susceptibility and Resistance of Malaria

1.3.1 Overview

Malaria represents a serious burden in Africa, the infected person could be asymptomatic, normally observed without complication and severe [22]. The effectiveness of malaria is influenced by the lack of Key immune system element regulation and parasites in the small brain blood vessels. The molecular regulatory mechanisms and cellular directing the pathogenesis of disease are certainly, not really understood. The available knowledge is highly related to genetics. It has been shown that genetic factors can address the clear understanding of the relevant role of severity and outcome of the disease. Many epidemiological studies have been elucidated for genetic control of human malaria caused by *Plasmodium falciparum*. The idea is to investigate the genes associate with susceptibility or resistance to malaria, this will be a major option for the curative treatment or vaccine [23].

The prior knowledges for host genetics susceptibility under *Plasmodium falciparum* malaria were studied over decades, this brought the human genome endemic regions stratifications imposed

by malaria parasite [24]. Many genes have been identified to malaria phenotype, and factors leading to the severity of malaria [25]. The investigation of genetic variants associated with malaria will result in a deeper understanding of malaria pathophysiology and in the elaboration of new treatment [26]. Also, the understanding of environmental effects due to several endemic regions can help to the new investigations of therapy development and management against malaria [26].

The human genetic polymorphisms play a major role for susceptibility and resistance to malaria [27]. Familial and epidemiological studies showed that malaria protection and/or susceptibility have substantial genetic components with estimated heritability of 25% [27]. However, only few of these associations were replicated in different populations and several conflicting findings were reported [28, 29, 30]. Some variants that are associated with protections in one population were found neutral in another population. In some instances, polymorphisms that were associated with protection against severe malaria in the initial population were found to be associated with increased risk in the second population when replication is attempted [28]. For instance, in a recent multi-center case-control association study conducted in Malaria endemic countries, forty percent of the previously reported loci, failed to replicate. Even though there could be many explanations for these inconsistencies, it is plausible to argue that the root cause is largely the limitations of the study approaches [31].

1.3.2 Gene mutations involved in susceptibility and resistance to *Plasmodium falciparum Malaria*

The different disease aetiology of Malaria is variable and unstable, this can be caused by the parasite virulence, host genetics and environmental factors. The mutation leading severity of *Plasmodium falciparum* infections take in account the phenotypes such as hyper or asymptomatic parasitaemia, cerebral Malaria and severe Malaria anaemia [32]. Polymorphisms and gene mutations in the human, present major benefit, also over generation their frequencies have increased due to the natural selection, this is case of *HbAS*, thalassaemias, glucose-6-phosphate dehydrogenase (*G6PD*) deficiency and haemoglobinopathies [33]. In the last decade, The human genome project and other institutions have identified most common loci affected malaria susceptibility are observed directly or indirectly by modulating immune response and interfering with host-parasite interactions [32, 33]. The intensity of severe malaria (SM) is characterized by interaction between populations and between individuals. Different prior discovers addressing gene mutation inherited in the severe malaria has been reported . However, these gene mutations have been discovered to be

erythrocytes(in Hb variants) including haemoglobin (Hb) variants, haptoglobin and Nitric Oxide metabolism. This is the case for HbAS (sickle cell trait) known as gene responsible for protection against severe malaria in the region highly affected by malaria due to the natural selection, this has been known over centuries.

Many studies reveal that the catabolism of free haem in the body is due to the insufficient rate of the enzyme haem oxygenase I, this has a major function in the pathologies like (Malaria, Sickle cell, Haemolytic disease, ...). Also, the co-inheritance of alpha-thalassaemia can lead the heterozygote protection against malaria caused by absence of sickle cell. Then it's summarized on the figure 1.1.

Table 1.1: *Genetic mutations involved in susceptibility -resistance to Plasmodium falciparum Malaria, with SMA: severe malaria anaemia; CM: cerebral Malaria; Hb: Haemoglobin; HbAS: Haemoglobin AS or sickle cell trait; G6PD: glucose-6-phosphate dehydrogenase; SM: Severe Malaria, UM : Uncomplicated Malaria*

Gene (Symbol)	Phenotype	Proposed protective mechanisms	References
Haemoglobin-C(*HbC*)	UM and SM	Decreased cyto-adherence of erythrocytes that are infected	[34]
Haemoglobin E (*HbE*)	SM parasitaemia	Reduction of invasion of erythrocytes by merozoites, decreased development of intra-erythrocytic parasites and increased phagocytosis of infected erythrocytes	[35]
Haemoglobin S (*HbS*)	UM and SM	Selective sicking of erythrocytes with the infected sickle trait contributes to improved spleen clearance. Reduced invasion of erythrocytes, early phagocytosis and suppressed development of parasites in venous micro vessels under oxygen tension. Innate and acquired immune enhancement	[36]

$\alpha - thalassaemia(\alpha - thal)$	SM and SMA	Decreased reset. The increased number of micro-erythrocytes in homozygotes lowers the amount of haemoglobin lost due to the density of the parasite, thereby defending against extreme anaemia.	[37]
$\beta - thalassaemia(\beta - thal)$	SM		[38]
Glucose-6-Phosphate dehydrogenase (*G6PD*)	UM and SM	The increased susceptibility to oxidant stress of the *G6PD* deficient erythrocyte triggers its defense against parasitization	[39]
Pyruvate kinase (*PKLR*)	parasitaemia	Erythrocyte invasion defect and ring-stage-infected erythrocyte preferential macrophage clearance.	[40]
Ovalocytosis (*SLC4A1*)	SM and CM	Inhibition of the entrance of merozoites into the red cell, inhibition of development of the intracellular parasite and prevention of erythrocyte lysis that occurs with maturation of the parasite, resulting in the release of merozoites into the bloodstream	[41]
Elliptocytosis	SM		[35]

Glycophorins A (GYPABC)	SM		[42]
Blood Groups(*ABO*)	SM	Reduced *Plasmodium falciparum* rosetting	[43]
Haptoglobin (HP)	SM	In HP polymorphic individuals, oxidative damage to uninfected cells can be more marked as HP proteins bind less effectively to Hb, increasing premature erythrocyte degradation and promoting the release of cytokines by these circulating cells	[44]
Nitric oxide synthase 2 (*NOS2*)	SM	The enhanced development of NO induces Th1 cytokines that activate macrophages and may therefore be a mechanism of antimalarial resistance.	[45]
haem oxygenase I (HO-1)	CM	Free haem is produced in the blood stream	[46]

1.4 Main Concepts in Genetic Variation

1.4.1 Mutations

DNA sequence is in permanent alteration, we define the mutations in the DNA as modification, variation or change of single or several basis. This can have some advantages and disadvantages into the organism (have effect or not effect) (https://www.ebi.ac.uk/training/online/course/), illustration in Figure1.2 . The Mutation is limited for modifying the allele frequencies, but able to improve the new alleles in population genetics. This leads to genetic drift, population bottleneck, founder effect, etc ...

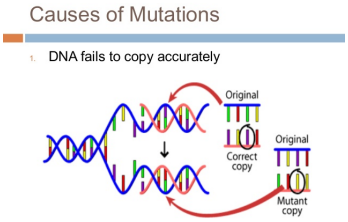


Figure 1.2: *This illustres the **Mutation** during the cellular division*
http://evolution.berkeley.edu/evolibrary/article/evo_20

1.4.2 Genetic Distance (F_{st})

Population genetics provide several information about different layers of populations, this last may be due to population ancestry or the different mutations occurring over time. Genetic distance is defined as a metric explaining the difference between species or populations, maybe the distance measures time from the same ancestry or degree of differentiation [47].

We can illustrate this definition by Figure 1.3 , evidence from genetic distance between european populations based on their ancestors.

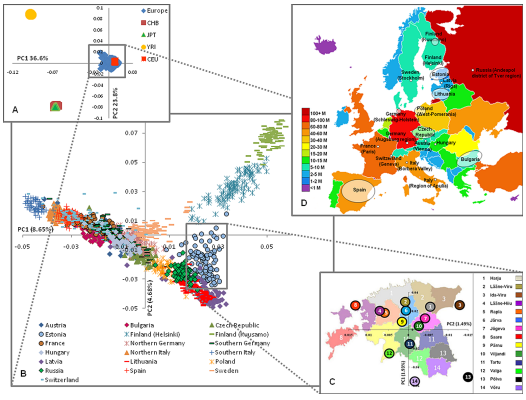


Figure 1.3: *This illustrate Genetic distance between european population, the European genetic structure (based on 273,464 SNPs). Three levels of structure as revealed by PC analysis are shown.* https://www.reddit.com/genetic_distance

The general insight of the origin of biodiversity can be explained by Genetic distance(for example breeds of domesticated animals).

Let define X,Y two randomly mating diploids populations (with segregation of multiple alleles at locus). We consider f_i ($resp.$ g_i) a frequency of i^{th} alleles in X (resp. Y), also we define the identity probability to choose two genes randomly $j_X = \sum_i {f_i}^2$ ($resp.$ $j_Y = \sum_i {g_i}^2$) in population X (resp. Y), also the probability of a gene identity from X and Y is $j_{XY} = \sum_i f_i g_i$

If there is no selection and each allele came from a single mutation in an ancestral generation, then the expected values of j_X and j_Y are equal to the Coefficient of Wright in inbreeding of X and Y [48], also j_{XY} is defined as a coefficient of Malecot of the Kinship [49].
So, We can normalise identity of genes between X and Y to this locus by:

$$I_j = \frac{j_{XY}}{\sqrt{j_X j_Y}} \tag{1.1}$$

$$I_j = \begin{cases} 1, & \text{for X, Y having the same alleles identical frequencies} \\ 0, & \text{if no common alleles} \end{cases}$$

Thus, we can normalize the genes of identical genes from X,Y with respect all loci by :

$$I = \frac{M_{XY}}{\sqrt{M_X M_Y}} \tag{1.2}$$

with $M_X = \frac{1}{n}\sum j_X$, $M_Y = \frac{1}{n}\sum j_Y$ and $M_{XY} = \frac{1}{n}\sum j_X j_Y$, given n loci studied.

By theory, it is easy to find the arithmetic mean of I_J rather than equation 1.2 , but the explanation genetic of the arithmetic mean remains a challenge despite the closeness of numerical values of these two quantities. Therefore we define genetic distance of X and Y by:

$$D = ln(\frac{1}{I}) \tag{1.3}$$

The equation 1.3, can make sense if for all loci the rate of gene substitution remains the same per locus. But in case of different rates for all loci, D underestimates the number of gene substitutions per locus. When the rate of substitution for genes varies with locus and all I_j are large. Therefore, we calculate the genetic distance by the geometrical mean [50], given by:

$$G = \frac{1}{n}\sum_{j=1}^{n} h_j \tag{1.4}$$

where h_j is the value of $-ln(I_j)$ in j^{th} locus , and n represent the number of loci studied.

From 1.2, The minimum genetic difference will be given by :

$$D_m = \frac{(M_X + M_Y)}{2} - M_{XY} \tag{1.5}$$

If I_j = 0, D is estimated by :

$$D = 2\alpha t \tag{1.6}$$

with α is the rate of gene substitution per locus per year and t number of the years.

1.4.3 Fixation Index (F_{ST})

We define Fixation Index as a measure of difference in the population caused by genetic structure. It's based on SNP estimation. It is given by :

$$F_{ST} = \frac{{\sigma_S}^2}{\bar{p} * (1 - \bar{p})} \tag{1.7}$$

$\bar{p}$ = The average frequency of an allele in the global population.

${\sigma_S}^2$ = The Variance two allele frequencies of allele between two subpopulations weighted by sizes of populations. Also we express F_{ST} by [51] :

$$F_{ST} = \frac{\bar{p}(1 - \bar{p}) - \sum_i a_i p_i (1 - p_i)}{\bar{p}(1 - \bar{p})} \tag{1.8}$$

a_i = Relative size of the i^{th} population.

p_i = Allele frequency of i^{th} population

1.4.4 Single Nucleotide Polymorphism (SNP)

SNP is a genetic variation of length 1 (1 base), Figure 3 is an example of SNP 1.4.

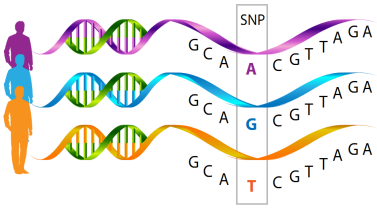


Figure 1.4: *Represent Single Nucleotide Polymorphism (SNP) plays a role in a wide variety of diseases such as sickle cell anemia and cystic fibrosis Nucleotide*

https://bioviva-science.com/blog/

It's important to know that SNPs are normally occurring between individuals . It's estimated to be appeared once within every 1,000 nucleotides. A SNP is identified by *rs number*, attributed from specifical database. So, a SNP defining a risk locus identified through GWAS(Genome wide association study) is called *Lead single-nucleotide polymorphism (LeadSNP)*

1.4.5 Gene Mapping

The main goal of gene mapping is to find the association from genes related to the traits, so the need is to use a tool able to find a gene and precise its region. Gene mapping has been defined as a tool having the role to localize and identify a specific gene region into the chromosome and investigate the distance between genes.
So,In Gene mapping there is two types of mapping such as linkage mapping and physical mapping. The linkage mapping use to find a distance between genes (tight or large distance) in the chromosome, physical mapping used to precise the location of the gene in the chromosome, find illustration in Figure 1.5 here below.

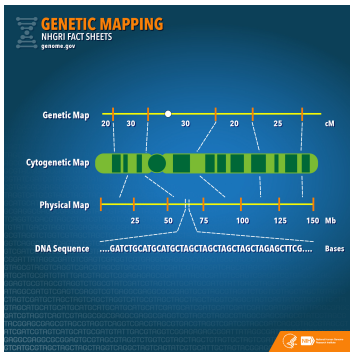


Figure 1.5: *Genetic maps have been used successfully to find the gene responsible for relatively rare, single-gene inherited disorders such as cystic fibrosis and Duchenne muscular dystrophy. Genetic maps are also useful in guiding scientists to the many genes that are believed to play a role in the development of more common disorders such as asthma, heart disease, diabetes, cancer, and psychiatric conditions [2].*

1.4.6 Linkage Disequilibrium (LD)

The genetic variation of different loci is structurally correlated. From genotype information in one locus, we can predict the genotype information in locus nearby, this linkage requires nonrandom mating [52]. The aim of LD is to investigate the related genes associated with the disease. Also, to define the difference between observed frequency and expected frequency for random effect of 2

alleles to loci. [52].

Let A,B two alleles occur with the respectives frequencies P_A, P_B and we define the haplotype frequency of AB (The frequency in the same gamete of both A and B occurs together) P_{AB} [53]. The level of LD or Coefficient of linkage disequilibrium is given by :

$$D = P_{AB} - P_A P_B \tag{1.9}$$

* A and B are in LE (linkage equilibrium) if $D_{AB} = 0$.
* Otherwise , A and B are in LD . Which means A and B are nonrandomly associated.

Generally, Linkage disequilibrium is used to identify the non-random association between two alleles at two specifical loci on a chromosome in a natural breeding population. Let Consider an SNP marker with alleles A and a, where P_A denotes the allele frequency of A. If this SNP marker is not in LD with second marker having alleles B and b with allele frequency P_B for allele B, the frequency P_{AB} of the haplotype AB equals $P_A P_B$. If there is LD, haplotype frequencies are disorder recombination by D that is the covariance between the two SNP markers. Then we need the LD probability[54].

$$\lambda = \frac{D}{P_A(1 - P_B)} \tag{1.10}$$

λ represent a LD probability between markers.

If $P_{AB} = P_A P_B$, then $\lambda = 0$

1.4.7 Minor allele frequency (MAF)

Analysis of disease in genetics is based on understanding of variants from their characteristic, proportion, and so all given in the population study. Intuitively, among the variants , some are common and some not. *MAF* takes the second frequency of most common allele occurs in a given population.

The reason for using MAF Minor allele frequency is due to its high capacity of providing information which can differentiate common and rare variants in population genetics.

1.4.8 TagSNP

Generally, the searching of association region to complex diseases (such as trisomy21, diabetes, ...) is a challenge. Millions of DNA variations may allow the fine dissection of associations. Unfortunately, these studies seeking disease associations become limited, due to SNP genotyping costs. The idea of taking or defining a subset of informative SNPs called (tag SNPs) becomes huge,

used as representatives of the others in LD (Linkage disequilibrium) [55].
Therefore, TagSNP Is defined as a SNP representative of a particular genomic region in high linkage disequilibrium. This portion of SNPs represented by TagSNP is called *haplotype*, find the illustration in Figure 1.6 here below.

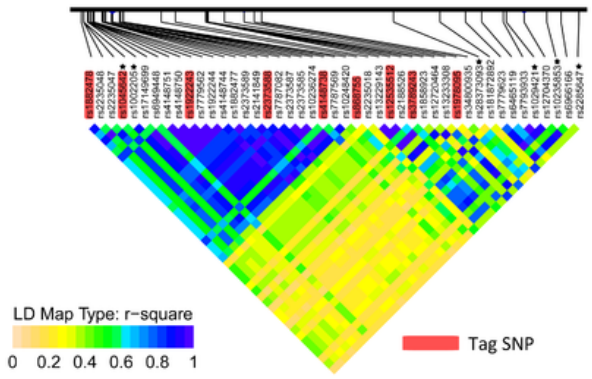


Figure 1.6: *LD plot of SNPs with top-ranked bayes factors in CHB of 1000 Genome Phase I. The colors indicate the strength of pairwise LD according to r2 metrics. The SNPs marked with asterisks represent independent strong associations. Tag SNPs are shadowed in pink)*

https://en.wikipedia.org/wiki/Tag_SNP

1.5 Genome Wide Association Study (GWAS) in Malaria

1.5.1 Overview of Genome Wide Association Study (GWAS)

The genetics is developing with a high level of technology, most human diseases have genetic components,these genetic components vary by disease [56], Some can be determined by the complete genome,whereas most of the common diseases are determined by interaction of many genes, environment and randomly. The understanding of disease susceptibility and resistance contributed by the genome may provide the relevant information which can help for developing an efficient therapeutic. To discover these genetic factors or components, different new technologies were implemented to accurately and improve the qualities of precision, these technologies based on study designs and analytical tools were established to identify the genetic risk factors.
The statistical approach of ***Genome Wide Association Study (GWAS)*** is explained as genetic variants set of individuals having variant associated to the traits wikipedia.org/GWAS, the ultimate goal after identifying new genetic associations is to use these information for developing the different strategies for detecting a better treatment and prevent the disease [57].

Its use linkage disequilibrium, such significant associations are due to the migration of small segments of chromosomes from the same shared ancestor into the population,these segments will carry identical alleles or haplotype without recombination. In case QTL(Quantitative trait loci) in segment, we will observe the identical quantitative trait loci alleles. The analysis of hundred of thousand SNPs and linking them to clinical analysis and observable characteristics GWAS is facilitated by high-throughput genotyping technologies.
More than a decade, nearly 100 loci for some common diseases have been identified , many in unknown genes associated with the disease not previously suspected , and also some particular genomic regions including unknown genes. This shows many advantages of using GWAS in terms of genetic variants discovering, But despite these advantages GWA studies have some challenges, such as their capacity for false positive and false negative, error due genotyping mechanism and biases due to choice of study [58]. GWAS has several applications, this new technology can be applied not only in human genetics but in the field such as criminology, biomedical, phytology, psychology ...

1.5.2 Application of GWAS in Severe Malaria of African's Population

The resistance of Malaria is one of the challenges facing by the world, especially in Africa, the research is still pursuing for reducing the risk of this disease in the community. Some of the factors of genetics can lead to reduce the risk of developing severe malaria. The most common variant associated with severe malaria is sickle cell *HbS variant*, this last was not enough to study and provide more information about resistance of Malaria, It was relevant to improve the researches and find the new strategies, Many technologies and advanced studies was performed to investigate the different variant causal severe resistance of Malaria[59].
The aim of using GWAS in malaria susceptibility study is to extend the limitation of the conventional approaches and bring an explanation elucidated by genetics in terms of genome wide scale, fundamentals consideration in planning to evaluate the influence of genetics in the trait interested . [60], The deep understanding of genetic resistance and susceptibility for severe malaria can lead to relevant information for molecular mechanisms between host-parasites. In large context,these information will build new insights to human physiology and genetics susceptibility for other disease. GWASs have recently been introduced in malaria endemic regions as the underlying genetics, the findings were efficient due to replication of known variants like HbS and ABO blood group. Unfortunately, this achievement has established just a few novel variants; indicating the attenuation of some of the real association signals. This may be the resulted from different

confounding factors (Large population genetic variation in malaria-endemic regions, variations in defensive allele frequencies and effect sizes, Among other aspects, through populations and the intrinsic shortcomings of approaches to GWAS).

The observation shows that some variants are underpowered for association based on GWAS threshold;it brings the idea of polygenic effects. This comes out several questions related to the genetic architecture of malaria protection, polygenic effects, contribution of distribution of heritability. This improves the amount of novel genetic variants, which can help to identify the novel inherited factors of risk that rely upon high-resolution.

The discovery of some gene mutation has been identified as a sickle cell trait that can decrease the risk of severe malaria , and this has modified the allele frequency over generations through natural selection.

However, the study aims to discover the new genetic mutations. GWAS and fine-scale molecular genotyping tools have simplified the analysis for identification.

In this paragraph we are going to illustrate the first GWAS analysis of severe malaria performed in the Tanzanian population. The study recorded a sample size of 914 individuals, the analysis is based on association and identifies the new gene targets in terms of immunological pathways. The method has shown the potential of using GWAS to classify unidentified susceptibility genes for malaria phenotypes in endemic populations.

Many studies have shown that the genetic factors of host and sickle cell trait have been associated with effectiveness to decrease the developing risk of severe malaria.

GWAS results of severe malaria performed in this population (with n = 914 and 15.2 million SNPs) have described Table 1.2 as very relevant. Despite the well-known association of sickle cell HbS variant, they identified the association to protective host from two genes interleukin receptors *IL-23R* and *IL-12RBR2*, also they find *KLHL3* known as kelch-like protein all with p-value ($P < 10^{-6}$).

Their analysis has identified *SYNJ2BP*,*GCLC* and *MHC* as potential loci detected under positive selection, these are detected based on extra haplotype homozygote. from whole genome sequencing of this population with Tanzanian cohort (family study n=247),They verified the allele frequencies of the underlying associations of common polymorphisms, as well as the range and existence of several structural variants that could be related to these SNPs.

From the imputed structural variants in a chromosome 4 known as the region encompassed by glycophorin genes, more than 50 unusual variants were characterized, and no clear evidence of associations with extreme malaria was found individually in our primary dataset *(P>0.3)*. Their technique reveals the promise of a joint genotyping-sequencing method to classify unknown

susceptibility loci in an African population with well-characterized phenotypes of malaria. The areas containing these locations are potential priorities for the implementation of much-needed steps to prevent or cure malaria disease. The protective associations were established as genes interleukin receptors *IL-23R* and *IL-12RBR2*, also *KLHL3* known as kelch-like protein, as well as near-significant effects on haplotypes of the Main Histocompatibility Complex(MHC). We observed prolonged homozygosity of the haplotype, identified *SYNJ2BP, GCLC, and MHC. USP38, FREM3, glycophorins, gypA/B/E, DDC, MARVELD3 and ATP2B4* are novel polymorphisms that have been identified.

Table 1.2: *Tanzania Study Population* https://journals.plos.org/plosgenetics/article/figure?id=10.1371/journal.pgen.1007172.t001

	Controls (n = 465)		Cases (n = 449)		Difference P-value
Age* (median, range)	2.8	0.9–10.9	1.7	0.2–10.0	$<2.2_{x10}^{-16}$
Female	252	54.2%	205	45.7%	0.012
Ethnicity**					0.52
Mzigua	151	32.5%	146	32.5%	
Wasambaa	142	30.5%	135	30.1%	
Wabondei	83	17.8%	86	19.2%	
Mmbena	26	5.6%	23	5.1%	
Mngoni	17	3.7%	18	4.0%	
Pare	11	2.4%	8	1.8%	
Mmakonde	11	2.4%	8	1.8%	
Mgogo	7	1.5%	8	1.8%	
Chagga	9	1.9%	7	1.6%	
Other	8	1.7%	10	2.2%	
Mixed Ethnicity***	150	32.3%	172	38.2%	0.065
Hyperlactatemia/acidosis	-	-	256	57.0%	-
Severe Malarial Anaemia	-	-	221	49.2%	-
Respiratory Distress	-	-	124	27.6%	-
Cerebral Malaria	-	-	120	26.7%	-

* months
** based on paternal ethnicity
*** if parental ethnicities were different

https://doi.org/10.1371/journal.pgen.1007172.t001

The summary result found in 1.3 GWAS Analysis is described as:

Table 1.3: *GWAS Result*

SNP	Gene	n SNPs	Location	Minimum P	Conditional P	Subtype P
rs334	HbS (in *HBB*)	40	11:5248232	HET: 8.59E-13	-	HL: 1.81e-09
rs9296359	*TREML4*	1	6:41205690	HET: 1.21E-07	HET: 4.42e-07	SMA: 3.29e-07
rs149085856	Intergenic (LINC00670)	1	17:12399526	ADD: 2.15E-07	ADD: 1.06e-06	HL: 2.81e-07
rs113449872	Intergenic	20	5:43909343	HET: 2.17E-07	HET: 2.93e-07	SMA: 2.92e-05
rs11335470	*LINC00944*	3	12:127237620	HET: 2.52E-07	HET: 1.86e-06	HL: 9.04e-05
rs73832816	*C4orf17*	1	4:100429757	REC: 3.75E-07	REC: 9.48e-07	CM: 1.02e-06
rs17624383	Intergenic	3	7:53676837	ADD: 5.62E-07	ADD: 3.28e-06	RD: 4.61e-07
rs2967790	*KLHL3, MYOT*	13	5:137011761	ADD: 5.85E-07	ADD: 2.46e-06	HL: 8.65e-06
rs144312179	*FAM155A*	6	13:108228013	ADD: 6.24E-07	ADD: 2.92e-06	HL: 1.35e-06
rs114169033	AF146191.4–004 (lincRNA)	3	4:190717704	ADD: 6.67E-07	ADD: 1.30e-06	RD: 5.62e-07
rs6682413	*IL23R, IL12RB2*	7	1:67731614	REC: 7.98E-07	REC: 1.03e-06	SMA: 1.23e-04
rs73505850	*CSMD1*	5	8:4754838	ADD: 7.98E-07	ADD: 1.42e-06	SMA: 1.20e-05
rs8109875	*ZNF536*	1	19:31069639	REC: 8.69E-07	REC: 3.57e-06	SMA: 2.80e-05
rs1878468	AC108142.1 (antisense)	1	4:182822332	HET: 8.98E-07	HET: 1.19e-06	HL: 8.10e-07
rs3133394	Intergenic	4	11:130417522	ADD: 9.41E-07	ADD: 1.08e-06	CM: 9.49e-06

Allele models: ADD Additive, HET Heterozygous, DOM Dominant, REC Recessive. Subtype significances: HL Hyperlactatemia; SMA Severe Malarial Anaemia; RD Respiratory Distress; CM Cerebral Malaria. Locations correspond to the GRCh37 reference genome. Minimum P indicates the most significant P for SNPs in the locus within the case-control GWAS, whilst Conditional and Subtype Ps indicate the most significant P value for those SNPs when controlling for rs334 status, or considering the severe malarial subtypes.

https://doi.org/10.1371/journal.pgen.1007172.t002

The result can be visualized in Manhattan plot in the figure 1.7 as:

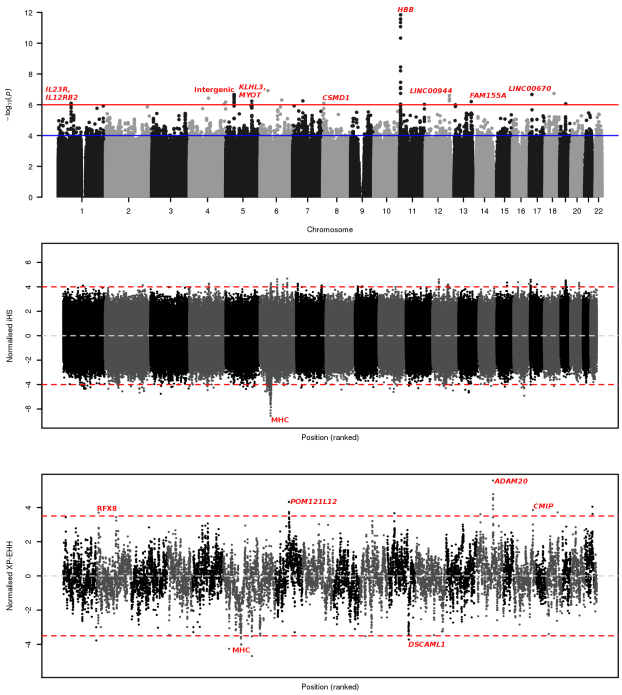


Figure 1.7: *https://journals.plos.org/plosgenetics/article/figure?id=10.1371/journal.pgen.1007172.t001*

The genome-wide association analysis was recorded in 17,000 extreme malaria cases and population controls from 11 countries, notified by family triple sequencing and direct typing of candidate loci in an additional 15,000 samples. [3]. The study identified 5 replicable interactions with genome-wide proof levels, including a new allele on chromosome 6 that was newly involved. Such variations account for around 0.1 of severe malaria heritability, which was calculated to be 23 percent using genome-wide genotypes [3].

They discovered an erythroid-specific transcription start site underlying the known relationship in ATP2B4 after interrogating available functional evidence, but can not establish a possible causal mechanism at the locus of chromosome 6. HLA correlations previously recorded do not replicate in these samples. They have identified a new locus on chromosome 6 the genes *MAP3K7* and *EPHA7*, also ABO, HBB, ATP2B4, the glycophorin region on chromosome 4 in a broad dataset that will provide a base for more studies on the genetic determinants of malaria resistance in different populations in Africa, we can see in Table 1.4 here below:

Table 1.4: *Reporting Genes find out across 11 African populations [3]*

Rsid	Position	Alleles	Log10 BFavg	−Log10 Padd	Nearest gene(s)	Linked phenotype	Frequency in controls/cases	Mode	Subtype effects	Population effects	Log10 BFrep
rs334	11:5,248,232	T/**A**	69.3	54.9/44.9	*HBB*	Sickle trait	7.6%/2.6%	het .8	fixed 1.0	cor/str 1.0	42.3
rs8176719	9:136,132,908	**T**/TC	18.1	18.9/18.0	*ABO*	O bld. grp.	69.8%/64.4%	rec .9	fixed 1.0	Afr. only .9	9.6
rs4951377	1:203,658,471	**A**/G	7.7	7.0/5.3	*ATP2B4*	MCHC	31.8%/29.3%	rec 1.0	fixed 1.0	fixed .7	0.3
rs567544458	4:144,513,361	T/**G**	7.6	10.2/7.3	*FREM3*	Dantu bld. grp.	2.7%/1.5%	add .7 dom .3	fixed 1.0	fixed .6	5.6
rs116423146	3:160,396,863	C/T	6.1	7.8/7.3	*ARL14 (1kb)*		92.1%/90.1%	rec .4 het .5	fixed .8	fixed .7 cor/str .3	−1.0
rs62418762	6:93,218,698	**C**/T	5.7	3.9/8.5	*EPHA7 (731kb)*		94.8%/93.6%	add 1.0	variable 1.0	fixed 1.0	1.0
rs57032711	9:129,250,119	G/**A**	5.3	7.9/7.0	*MVB12B*		13.7%/11.9%	add .6 dom .4	fixed .9	fixed .8	−1.5
rs79124314	11:79,337,866	**G**/A	4.3	2.5/5.1	*TENM4 (186kb)*		96.2%/95.3%	het .9	variable 1.0	fixed 1.0	
rs2523650	6:31,449,022	T/**C**	4.3	5.5/4.6	*HCG26 (9kb)*		75.8%/72.9%	rec .8	fixed 1.0	fixed .4 cor/str .6	−0.7
rs116782507	2:5,435,704	A/**G**	4.2	6.8/4.9	*SOX11 (397kb)*		3.6%/2.7%	add .7	fixed 1.0	fixed .8	−0.9
rs74806154	10:620,483	**G**/GGCAC	4.2	7.1/6.4	*DIP2C*		73.4%/70.8%	add .7 rec .3	fixed .9	fixed .8	
rs73289758	12:24,535,213	C/T	4.0	1.3/6.6	*SOX5*		92.5%/92.1%	add .7	variable 1.0	fixed 1.0	

1.6 Advantage and limitations of Genome Wide Association Study (GWAS)

Genome-wide association studies (GWAS) have soon become a common methodology for the detection of disease genes. It includes screening thousands of samples (Case-Control), using hundreds of thousands of SNP markers found in the human genome, and comparing the frequencies between disease and control cohorts of single SNP alleles, genotypes, or multimarker haplotypes to distinguish these loci by applying statistical methods.
One of the advantages of using GWAS settles on investigation of the entire whole human genome and genotyping the single nucleotide polymorphism at cheapest cost.
Also, genome wide association study approaches present the possibility of testing a very large number of SNPs at the same time, from the result,It has identified a significant number of genetic positions associated with different diseases.
As a limitation, GWAS present huge cases of limitation, Among them we have Lack of well-defined case and control groups, inadequate sample size, multi-test control, and population stratification control are common concerns.
As we know that Genome wide association study provides the information about the identifying of SNP with different traits(we find only association between SNPs and trait), but not accurately the really causal variant of phenotype (disease, trait, ...).
Most GWAS also established genes that only clarify small amounts of the genetic variation that exists with certain characteristics.
GWAS is focused on a simplistic study of the genotype - phenotype, unable to provide substantial knowledge of the biological and biochemical roles of essential genetic variants needed for therapeutic applications.

1.7 Overview of Post-GWAS

GWAS is performed to find the association between genetic variants and traits or disease. But this study was not enough for analysing this association between disease or trait and variant, Since the aim is to recognise the particular genetic variant(s) from a risk-related locus that expresses phenotypic variations depending on the functional biology modulated by it. This expectation target is to detect the real variant causing the trait and investigate function of elements targeted by genetic risk variants[61]. So we need to improve the enrichment power of SNP by increasing

the effect size and try to find the leadSNPs. The heterogeneity presents at risk-associated loci, which is not probably the true causal genetic variations underlying the interaction, is captured by these leadSNPs. Therefore, each risk-related locus consists of a set of genetic variants, all likely causal, related to the initial leadSNP in the linkage disequilibrium (LD). Also, any genetic variant in strong LD with LeadSNP has much probability to influence the different phenotype. Therefore, it becomes very important after GWAS to target the specific genetic variants from a risk-associated locus to observe phenotype differences based on the functional biology, this is what Post-GWAS study involves[62].

1.7.1 Overview of Functional Genome Wide Association Study (FGWAS)

Given the success of research from the Genome-wide association (GWASs) for identifying risk loci of common diseases and complex traits, the approach remains limited due to the limitations evoked on 1.6 , the need is to improve power for these variants with low effect and find the optimal strategy safeguard against false positive associations.

The progress of researches have shown by different genomics that certain categories of variants enriched for disease heritability based on functional genomics data information improve the statistical power [63], This approach conducts the incorporation of prior functional genomic information into association analyses which can potentially lead to increase GWAS power. However, these functional genomic information are related to functional annotation of gene structures and regulatory elements, which means the description of statistical models using association statistics computed across the genome is to identify classes of genomic elements that are enriched or depleted for loci able to influence a trait. This approach becomes one of popular methods used in case of GWAS limitations, it has been implemented in **FINDOR-tool** (functionally informed novel discovery of risk loci) [63], FGWAS-tool (Functional Genome wide association study) of JOSEPH PICKRELL [64].

1.7.2 Motivation

The deep understanding of the genetic basis of resistance and susceptibility to severe malaria could shed lights into molecular mechanisms of pathogenesis and protection that will inform development of treatments and vaccines. GWA studies have been proven to be powerful tools to investigate the genetic architecture of human diseases including Malaria. However, the method is challenged by:

1. The lack of translation into relevant biological theories of related loci [65].

2. The well-known question of heritability lacking [65].

3. The lack of understanding of how many modestly related loci interact within genes, among others, to affect a phenotype.

Those challenges bring out the limitation or weakness to use GWAS for studying the causal and association variant in the population. It becomes more sufficient to perform novel methods studying after GWAS called post-GWAS, but those latter are several. But, the recent advances in capturing polygenicity through post-GWAS, methods for risk prediction, detecting new risk loci,from GWAS summary statistics and LD information under specific reference panel of population studying imputing untyped associated variants and fine-mapping for causal variants [66], are also playing an increasingly critical role in genomic studies. Given a rich source of information on both Malaria GWAS and reference panels, it is now an intriguing time to investigate the genetic architecture of Malaria resistance and susceptibility using advanced tools and leveraging GWAS summary statistics in efficiently carrying out whole-genome analysis of functional phenotypes such as Malaria.

1.7.3 Objectives

We hypothesize that full-genome functional analysis will greatly improve the diagnostic ability to identify major genetic variants impacting malaria tolerance or susceptibility. The main objective of this project is to conduct a systematic meta-analysis across three Africans-specific Malaria GWAS datasets (Kenya, Malawi and Gambia) and perform functional analysis of GWAS summary statistics. To achieve this objective, the specific objectives have to be met.

- We propose to access all Malaria genome-wide association studies (GWAS) data in African populations (Kenya, Malawi and Gambia) from MalariaGEN.
- Conduct Meta-analysis across Africa-specific Malaria genome-wide association study (GWAS) predict global Malaria risk/resistance and genetic heterogeneity in order to understand Malaria-specific genetic architecture.
- To carry out the functional study of GWAS summary statistics, we will apply techniques based on the gene list enrichment principle, According to this, as a proxy, a broad over-representation of genes candidates associated to biological pathway can be used to infer overrepresentation of biological pathway of candidate SNPs [29].

Chapter 2

Mathematical Approaches used in Functional Genome wide Association (FGWAS)

2.1 Overview

The strategies of GWAS are focused on associations and FGWAS ultimately analyzes phenotype genetic regulation by incorporating biological concepts through mathematical and computational bridges into the GWAS system [67], Which can answer a variety of basic issues, such as genetic control trends over growth, the time of genetic effects,the causal of evolution. Also, functional Genome Wide Association Study impacts increased power for gene detection by capitalizing on cumulative phenotypic variation[67].
The objectives of this section is to provide some specific statistical methods useful in FGWAS :

- The detection of genetic markers that are significantly associated with the functional phenotype.
- The detection of the sub-region(s) (or compact set(s)) of the functional phenotype that are significantly associated with some genetic marker(s).

It was precise and observed by different researchers involved in GWAS methods that most of causal SNPs or multi-factorial causing traits fall on non coding part of Genome i.e outside of protein-coding exon [64]. This issue brings the researchers to provide a new genetics catalogue

based on those non coding part, which is the case of "Histone Modification", "Transcription factor binding", in "Regulation mechanism part", this latter was generated by ENCODE project.The idea of combining the different resources rich information of functional genomic data can provide and improve the necessary information leading to the causal variants of trait [64]. Therefore, one of the appropriate methods is "Enrichment", this consists of examining the most probable associated variant from GWAS, and proceeding to the different tests, trying to check if they fall disproportionately in specific genomic regions [68]. Therefore, we will explain:

- Linear Mixed Model for Genetic Association
- Bayesian GWAS Model
- Hierarchical Model
- Bayes factor for GWAS
- Cross-Validation

2.2 Linear Mixed Model for Genetic Association

In order to allow both fixed and random results, an expansion of simple linear models is linear mixed models, and are particularly used in case of non independence between data, which could arise from a hierarchical structure, longitudinal data[69].
The nonindependence of observations may result from serial correlation or clustering of the observations, cluster correlation is presented when the observations are grouped in various ways[70]. In genome wide association studies, the use of linear mixed models (LMMs) becomes common because of the high capacity of methods for correcting the confoundings created by genetic relatedness, such as population structure, familial relatedness. Linear mixed model is more generalized by the expression here below:

$$Y = X\beta + \sum_{i=1}^{h} Z_i u_i + \epsilon \tag{2.1}$$

Where Y is column vector of output, response or phenotype of $N \times 1$, and $\forall\ y_i$ component of Y. X represent predictor variables matrix or covariate matrix with N $\times$ P dimensions.
β is a column vector for fixed effects regression coefficients with P $\times$ 1 dimension, with $\beta \sim N(\mu, \sigma^2)$.
Let consider Z_i a matrix for random effects of $N\times$ q, (q random effect), u_i is column vector of random effects, in $q \times 1$ dimension with $u_i \sim N(0, {\sigma_i}^2)$ from equation 2.1, we can calculate E(Y)

and cov(Y,Y). Infact,

$$\begin{aligned} E(Y) &= E(X\beta + \sum_{i=1}^{h} Z_i u_i + \epsilon) \\ &= E(X\beta) + E(\sum_{i=1}^{h} Z_i u_i + \epsilon) \;\; by\ linearity\ of\ Expectation \\ &= X\beta + E(\sum_{i=1}^{h} Z_i u_i) + E(\epsilon) \\ &= X\beta + \sum_{i=1}^{h} Z_i E(u_i) + E(\epsilon) \end{aligned}$$

As $u_i \sim N(0, {\sigma_i}^2)$ and $\epsilon \sim (0, \sigma^2)$, then we have:

$$E(Y) = X\beta$$

Same procedure for cov(Y,Y), and using Auto-covariance matrix of real random vectors notion, we have:

$$cov(Y, Y) = cov(X\beta + \sum_{i=1}^{h} Z_i u_i + \epsilon)$$

Because $X\beta$ is a fixed effect, then $cov(X\beta) = 0$, then we have

$$\begin{aligned} cov(Y, Y) &= cov(\sum_{i=1}^{h} Z_i u_i + \epsilon) \\ &= \sum_{i=1}^{h} cov(Z_i u_i) + \sum_{i \neq j}^{h} cov(Z_i u_i, Z_j u_j) + \sum_{i=1}^{h} cov(Z_i u_i, \epsilon) + \sum_{i=1}^{h} cov(\epsilon, Z_i u_i) + cov(\epsilon) \\ &= \sum_{i=1}^{h} Z_i cov(u_i) {Z_i}' + \sum_{i \neq j}^{h} Z_i cov(u_i, u_j) {Z_j}' + \sum_{i=1}^{h} Z_i cov(u_i, \epsilon) + \sum_{i=1}^{h} cov(\epsilon, u_i) {Z_i}' + cov(\epsilon) \\ &= \sum_{i=1}^{h} Z_i {Z_i}' {\sigma_i}^2 + \sigma^2 I_n \end{aligned}$$

Therefore, cov(Y,Y) and E(Y) is given by :

$$cov(Y, Y) = \sum_{i=1}^{h} Z_i {Z_i}' {\sigma_i}^2 + \sigma^2 I_n \quad and \quad E(Y) = X\beta$$

We realize that only X provides the mean of Y. Also the covariance of Y is structured by Z, i.e Z is only entered into the covariance structure [71].

Mixed model analysis provides a general, flexible approach in linear model, he allows a wide variety of correlation patterns (or variance and covariance structures) to be explicitly modeled, Is defined by:

$$Y = X\beta + U\gamma + \epsilon \tag{2.2}$$

with $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m$, $X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mp} \end{pmatrix} \in \mathbb{R}^m$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p$,

$\varphi = \begin{pmatrix} D & & \\ & \ddots & \\ & & D \end{pmatrix}$, $U = \begin{pmatrix} \mathcal{U}_1 & O_{n_1 \times q} \dots & O_{n_1 \times q} \\ \vdots & & \vdots \\ O_{n_m \times q} & \dots & \mathcal{U}_m \end{pmatrix} \in \mathbb{R}^{n \times (mq)}$, $\gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_m \end{pmatrix}$,

$\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix}$, $R = \begin{pmatrix} \Sigma_1 & \dots & O \\ \vdots & & \vdots \\ O & \dots & \Sigma_n \end{pmatrix} \in \mathbb{R}^{n \times n}$

And,
$\begin{pmatrix} \gamma \\ \epsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} O \\ O \end{pmatrix}, \begin{pmatrix} \varphi & O_{mq \times n} \\ O_{n \times mq} & R \end{pmatrix} \right)$,

We can use Hierarchical models to write.

$$Y|\gamma \sim \mathcal{N}_n\left(X\beta + \gamma U, R\right) \quad and \quad \gamma \sim \mathcal{N}_{mq}\left(O, R\right)$$

if We call $\epsilon^* = U\gamma + \epsilon$, then we get :

$$\epsilon^* = \begin{pmatrix} U & I_{n \times n} \end{pmatrix} \begin{pmatrix} \gamma \\ \epsilon \end{pmatrix} \tag{2.3}$$

$$A = \begin{pmatrix} U & I_{n \times n} \end{pmatrix} \tag{2.4}$$

Then we have, $\epsilon^* \sim \mathcal{N}\left(O, V\right)$, with:

$$\begin{aligned} V &= A \begin{pmatrix} \varphi & O \\ O & \mathcal{R} \end{pmatrix} A^t \\ &= \begin{pmatrix} U & I_{n \times n} \end{pmatrix} \begin{pmatrix} \varphi & O \\ O & \mathcal{R} \end{pmatrix} \begin{pmatrix} U^t \\ I_{n \times n} \end{pmatrix} \\ &= U\varphi U^t + \mathcal{R} \end{aligned}$$

This model is best written under matrix form:

$$Y = X\beta + Z\mathcal{U} + e \tag{2.5}$$

Where: $Y = (Y_1, ..., Y_n)^t$ is a response, trait or phenotype vector, $\beta = (\beta_1, ..., \beta_p)$ is a fixed effects, X is a $n \times p$ covariates or predictors matrix, with random effects $\mathcal{U} = (u_1, ..., u_q) \sim \mathcal{N}(0, \tau I_q)$ and then $n \times q$ random effects matrix Z. Also the residual error vector $e = (e_1, ..., e_n)^t \sim \mathcal{N}(0, \sigma^2 I_n)$. So to simplify this model we can write:

$$Y = X\beta + \omega + e \tag{2.6}$$

Therefore, $\omega = Z\mathcal{U}$, thus $\omega \sim \mathcal{N}(0, \tau K)$ with $K = ZZ^t$. From 2.5, Y is multilinear, Y follows normal distribution, so our expectation is to define mean and variance, therefore we have $Y \sim \mathcal{N}(X\beta, \tau K + \sigma^2 I_n)$

To conduct genetic association using linear mixed model can be useful:

- In the case of SNPs, the number of SNPs greater than the sample of individuals.
- In the case of population structure.
- In case of related of individuals.

Let assume the additive model:

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I_q {\sigma_e}^2) \tag{2.7}$$

with $\beta = \{\beta_i : i \in \{1, ..., M\}\}$ set of SNPs effect sizes. From multi-linear regression, we can define the conditional probability of Y knowing $X,\ \beta\ \ and\ \ \sigma_e$.

$$P(Y|X, \beta, {\sigma_e}^2) = \prod_{i=1}^{M} f(Y; \beta_i, \sigma_e)$$

Then we have:

$$P(Y|X, \beta, {\sigma_e}^2) = \frac{1}{\sqrt[n]{2\pi{\sigma_e}^2}} \exp \frac{(Y - X\beta)^t (Y - X\beta)}{2{\sigma_e}^2} \tag{2.8}$$

The estimation $\hat{\beta}$ is given by:

$$\ln P(Y|X, \beta, {\sigma_e}^2) = -\frac{N}{2} \ln(2\pi{\sigma_e}^2) - \frac{(Y - X\beta)^t (Y - X\beta)}{2{\sigma_e}^2}$$

Therefore, to optimize $\hat{\beta}$,

$$\begin{aligned} \frac{\partial \ln \beta}{\partial \beta} &= 0 \\ \iff \frac{\partial((Y - X\beta)^t (Y - X\beta))}{\partial \beta} &= 0 \\ \iff Y^t X &= (X^t X)\beta \\ \iff \beta &= (X^t X)^{-1} Y^t X \end{aligned}$$

$M \gg N$

* The case of $M \gg N$, we can describe the test SNP marginally . and for each i SNP, we get

$$\beta_i = ({x_i}^t x_i)^{-1}({x_i}^t Y)$$
$${\chi_i}^2 \approx \rho(x_i, Y)^2 * N$$

Where N is a sample size.

Population Structure

Let's apply linear mixed model in population structure [72] :

$$Y = P\lambda + X\beta + \epsilon \tag{2.9}$$

Consider the population structure.

- P = $\{P_i$, i number of population $\}$
- λ specifies the phenotypic mean of each population .
- A SNP with different means between 2 populations will be associated with Y

We can use PCA to estimate P and include it as a fixed effect in our regression. if P alters phenotypic variant then variance is a function [72] e.g error follows $\mathcal{N}(O, f(P))$, therefore the linear regression and principal components will not produce a uniform p-value distribution.

Related Individuals

Our natural assumption predicts that $\epsilon \sim \mathcal{N}(0, I{\sigma_e}^2)$, unfortunately this assumption is insufficient for providing more information because of environmental effect and family relatedness [72]. Therefore we have

$$Y = X\beta + \epsilon\ ; \quad \epsilon \sim \mathcal{N}(0, F{\sigma_F}^2 + I{\sigma_e}^2) \tag{2.10}$$

With F a matrix of family members and ${\sigma_F}^2$ is the covariance of shared environmental effect on Y. Also $({Z_j}^T Z_j)^{-1}({Z_j}^T Y)^2$ will be inflated.
To increase power by joint modelling all variants and account non i.i.d due to structure, we consider the following phenotypic model [72].

$$Y = X_i\beta_i + \epsilon; \ \ \epsilon \sim \mathcal{N}(O, \Omega), \ \ With \ \ \ \Omega = H{\sigma_g}^2 + I{\sigma_e}^2$$

Then we have estimation likelihood and log likelihood:

$$\begin{aligned}\mathcal{L}(\lambda, {\sigma_g}^2, \beta_i) &= \mathcal{N}(Y, X_i\beta_i, {\sigma_g}^2(T+\lambda I)) \;\; with \;\; \lambda = \frac{{\sigma_g}^2}{{\sigma_e}^2} \\ &= -\frac{N}{2}\log 2\pi{\sigma_g}^2(H+\lambda I) - \frac{1}{2{\sigma_g}^2}(Y-\beta_i X_i)^T(H+\lambda I)^{-1}(Y-X_i\beta_i)\end{aligned}$$

From this model, the score test ${\chi_i}^2$ is :

$${\chi_i}^2 = ({X_i}^T\Omega^{-1}X_i)^{-1}({X_i}^T\Omega^{-1}Y)^2$$

In Linear mixed model, we have 3 important components :

- Defining and Building H
- Estimating parameters $(\beta_i, {\sigma_g}^2, {\sigma_e}^2)$
- Test statistics:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

The choice of H is the major problem in linear mixed models, the condition is to find a coefficient of ${\sigma_e}^2$.

Let's consider the case of accounting polygenicity. We want to gain the power by choosing H. So in case, we define linear mixed models by:

$$Y = X\beta + g_i\beta_i + Zb + \epsilon \;\; b \sim \mathcal{N}(0, \frac{{\sigma_g}^2}{M}I) \;\; , \epsilon \sim (0, {\sigma_e}^2)$$

X is a covariate matrix including intercept term Z centered and scaled G. The effects size b for SNPs in Z are random effects, thus:

$$\begin{aligned} Zb &\sim \mathcal{N}(0, var(Zb)) \\ \implies &\sim \mathcal{N}(0, ZE[bb^T]Z^T) \\ \implies &\sim \mathcal{N}(0, \frac{{\sigma_g}^2}{M}ZZ^T)\end{aligned}$$

Therefore, $H = \dfrac{ZZ^T}{M}$

Consider the case of Relatedness. In this case the goal is to determine the relatedness between individuals sharing the same environment. We assume a linear mixed model define as :

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, F{\sigma_F}^2 + I{\sigma_e}^2) \tag{2.11}$$

F represents a family matrix membership. So, F[i,j] is a fraction of the genome recently shared IBD between individual i and j.

2.3 Bayesian Lasso GWAS Model

Bayesian in probability typically reflects a degree of confidence in a case.The degree of confidence in any outcome may be dependent on historical experience about the event, such as the outcomes of past studies, or on personal assumptions about the event [73, 74]. From statistical Bayesian, The Lasso estimation for linear regression parameters has independent Laplace priors when regression parameters have described as a Bayesian posterior approximation [74]. Gibbs sampling from this posterior is possible using an advanced hierarchy of conjugate normal priors on the regression parameters and distinct exponential priors on their variances. Tractable complete conditional distributions are given by a relationship with the inverse-Gaussian distribution. Interval forecasts (Bayesian credible intervals) are given by the Bayesian Lasso that can direct variable collection. In addition, for choosing the Lasso parameter, the hierarchical form model uses Bayesian and probability approaches [74].

Our interest about this notion is to define prediction according to current information that we have [74]. Using the Bayesian model in GWAS will help us to make predictions and analyze the significant SNPs.

We define a Linear regression:

$$Y = Xb + \epsilon \tag{2.12}$$

with X a $n \times p$ matrix, a regression coefficient $b = (b_1, ..., b_p)^T$ and ϵ is n-vector normal distribution $\mathcal{N}(0, \sigma^2 I_n)$. Linear regression coefficients are calculated by L1-constrained least squares. To approximate regression parameters, the Lasso is commonly used $\beta = (\beta_1, ..., \beta_n)$ in the model

$$Y = \mu + X\beta + \epsilon \tag{2.13}$$

To make the prediction in our model, it's important to reduce matrix X such that the estimated regressions coefficients exceed a threshold λ absolutely positive.

$$X' \ a \ submatrix \ of \ X, \ such \ that \ |\hat{b}_j| > \lambda \tag{2.14}$$

Lasso projections are also interpreted as estimates of L1-penalized least squares. They achieve

$$min(\hat{Y} - X\beta)^T(\hat{Y} - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

To find λ we can use cross-validation (we will talk about it later). And then we would define the estimator of Y by the model, called *Preconditioning Model*, for a phenotypical analysis and genotypes information.

Preconditioning Model :

$$\hat{Y} = \mu + {X_i}^T\alpha + {Z_i}^T\beta + {\gamma_i}^T a + {\omega_i}^T d + \epsilon_i \quad for \;\; i \in \{1, ..., n\} \tag{2.15}$$

with μ = mean X_i = i-column vector of covariate within X means $X_i \in \mathbb{N}^d$ for i phenotype. $\alpha \in \mathbb{R}^d$ and $Z_i \in \mathbb{R}^s, \beta_i \in \mathbb{R}^s$ also and d are respectively additive effects of SNP for phenotype i and dominant effects of SNPs for phenotype i. and ϵ_i represents the residual error follows normal distribution centered to 0 and σ^2. We define the indicator γ_i and ξ_i for additive effects and dominant effects of SNPs [73, 74] . Therefore, we have:

$$\gamma_{ik} = \begin{cases} 1, & \text{if SNP k genotype is AA} \\ 0, & \text{if SNP k genotype is Aa} \\ -1, & \text{if SNP k genotype is aa} \end{cases} \tag{2.16}$$

Also ξ_{ik} is defined as :

$$\xi_{ik} = \begin{cases} 1, & \text{if SNP k genotype is AA} \\ 0, & \text{if SNP k genotype is Aa or aa} \end{cases} \tag{2.17}$$

Although SNPs number is greater than sample size in association study, common of that the regression coefficient presents unstable effects in the phenotype i.e either weak or strong. Thus, to target those SNPs responsible of effects and enhance the prediction performance, we integrate lasso penalties, having the same dimension that ξ and γ_i. Then we have:

$$\sum_{k=1}^{p} |a_k| \leqslant m \quad \sum_{k=1}^{p} |d_k| \leqslant n, \quad with \;\; \forall m, n \in \mathbb{R}_+ \tag{2.18}$$

m and n are selected to penalize the dominant and additive effects. Therefore, we can apply estimate the parameters in 2.15 using the Ordinary least square, then we have:

$$\frac{1}{2}\|\hat{Y} - \mu - {X_i}^T\alpha - {Z_i}^T\beta - {\gamma_i}^T a - {\omega_i}^T d\|^2 + \Psi \sum_{k=1}^{p} |a_k| + \Psi^* \sum_{k=1}^{p} |d_k| \tag{2.19}$$

with $\Psi \;\; and \;\; \Psi^*$ are lasso parameters able to control the shrinkage degree in the estimation of genetic effects.

2.4 Mathematical Approach applied in Functional Genome wide association (FGWAS)

The relevant information used in functional genome wide association study is resulted from two notions which is GWAS and Genomic annotation using the hierarchical model. It consists of splitting a genome into individual blocks, so that the blocks in the population are greater than the size of the linkage disequilibrium. [64]. In each block we have either a causal SNP or not, Thus we are going to use Bayes's approach by defining the prior probability as probability of any given block containing an association and the conditional prior probability as probability of having at least one causal given SNP in the block.
Those probabilities can change depending on the functional annotations.

2.4.1 Bayes Factors

Generally, For the ranking of associations, the Bayes factor is defined as a summary measure which provides an alternative to the P-value [75].
Let's consider a linear regression model. We define an additive model

$$\mathbb{E}[y_i] = \beta g_i + \alpha \tag{2.20}$$

with y_i i-phenotype, component of $\vec{y}$ *and* $i \in \{1, ..., N\}$. $\vec{g}$ is a genotype matrix in $P \times N$ dimensions. β is the effect size. The model is tested by:

$$H_0 : \beta = 0 \tag{2.21}$$

$$H_1 : \beta \neq 0 \tag{2.22}$$

To compare those models 2.21 and 2.22 we use bayes factor, defined :

$$B = \frac{\int P(\vec{y}|\vec{g}, H_0)}{\int P(\vec{y}|\vec{g}, H_1)} \tag{2.23}$$

From Wakefield approach [76], the approximation of Bayes factor consist to find the estimator of maximum-likelihood β called $\hat{\beta}$ with standard error ($se = \sqrt{V}$). This estimator of maximum likelihood is given by:

$$\hat{\beta} \sim \mathcal{N}(\beta, Se) \tag{2.24}$$

from Wakefield [76], β becomes a normal prior with $\beta \sim \mathcal{N}(0, T)$, then we get:

$$B = \frac{\sqrt{1-K}}{\exp[\frac{-Z^2}{2}K]} \quad with \quad K = \frac{T}{V+T} \quad and \quad Z = \frac{\hat{\beta}}{\sqrt{V}} \tag{2.25}$$

Z represents a standard Z-score, T is a prior variance of β. T is determined using prior specification, using relative risk of interest and *MAF*. There are 2 steps to define T.

* Effect-MAF independence

We define variance T independent of MAF, and we define the prior distribution of relative risk (R), we consider relative risk upper (R_u) such that this probability occurs with lower value. Then T is defined as :

$$T = (\log(\frac{RR_u}{\Phi^{-1}(1-q)}))^2 \tag{2.26}$$

* Effect-MAF dependence

We know that large genetic effects is associated to smaller MAFs [77]., T will get the form:

$$T = \vartheta \exp(-\iota M) \tag{2.27}$$

With M a MAF, ι and ϑ are positive parameters chosen in advance.
Also, we define variance of β by a given expression V :

$$V = \frac{n + n^{'}}{n.n^{'}[(1-M)^2 x_0{}^2 + 2M(1-M)x_1 + M^2 x_2{}^2 - ((1-M)^2 x_0 + 2M(1-M)x_1 + M^2 x_2)^2]} \tag{2.28}$$

From 2.25, we realize that from Z-score our V and T, we can use one of them to obtain a Bayes factor measuring the SNPs associated or not to the traits. Also x_i , $i = \{0, 1, 2\}$ depend on the model.

2.4.2 Hierarchical Model

The General idea of Hierarchical model in GWAS is to build a model able to define the correct characteristic of causal SNPs or responsible for traits [64]. Unfortunately, in the genome there are some SNPs strongly correlated according to their Linkage Disequilibrium (LD). It becomes more relevant to split the genome by different blocks and to construct the different site containing the true association [64].
Let's consider the M set of SNPs genotyped from N individuals in GWAS. We split M by t parts having K sizes, $t = E[\frac{M}{K}]$, means we have t blocks of size K.Therefor e, Interpretation by Bayesian becomes more relevant to define the probability of data [64].

$$P(\hat{y}) = \prod_{k=1}^{t} [(1 - C_k) R_k{}^0 + C_k R_k{}^1] \tag{2.29}$$

${R_k}^0$ represents the data probability of block k in case of no SNPs associated .

${R_k}^1$ represents data probability block k for one SNP association.

Ck represents a prior probability of block k containing a causal associated SNP, is expressed by :

$$R_k = \sum_{i \in H_k} C_{ik} R_{ik} \tag{2.30}$$

with H_k = SNPs set in block k

C_{ik} = of SNP i Prior probability in block k being the causal in the region.

R_{ik} is a probability that SNP i in block k being associated with the trait.

Therefore, we can join two prior probabilities such that a single association in block k exist.

Use data set remain inefficient to precise, so we will integrate the genomic annotation. Therefore, we have novel model according to [67] , we have a regional probability:

$$ln(\frac{C_k}{1 - C_k}) = d + \sum_{j=1}^{A} \lambda_j I_{kj} \tag{2.31}$$

And

$I_{ik} = \begin{cases} 1, & \text{region k is annotated by j} \\ 0, & \text{region k is not annotated by j} \end{cases}$ where

A = The size of region-level annotations from the model.

λ_j = Effect associated with annotation j. Therefore, from 2.4.2, 2.31 and 2.30 we can find a likelihood of data fitting model:

$$L(\hat{y}|\theta) = \prod_{k=1}^{t} [(1 - C_k){R_k}^0 + C_k \sum_{i=1}^{K} C_{ik} {R_{ik}}^1] \tag{2.32}$$

Because we are going to compare the models defining the causal SNPs, we have to compare the models using the Bayes factor. Thus, we get:

$$B_i = \frac{{P_{ik}}^1}{{P_k}^0} \tag{2.33}$$

$$L(\hat{y}|\theta) = \prod_{k=1}^{t} {R_k}^0 [(1 - C_k) + C_k \sum_{i=1}^{K} C_{ik} B_i] \tag{2.34}$$

θ represent all parameters of the model.

2.5 Meta-Analyses of Genome-Wide Association Studies

2.5.1 Introduction

An increasingly common method for detecting associations between SNPs and phenotypic characteristics is genome-wide association studies (GWAS). [78] . This method is commonly used within the different disciplines, very useful in social sciences, pharmacogenomics, microbiology, etc . However, despite the celebrity and importance of the method in several fields, the challenge of finding out the large number of variants and true positive remains a problem [78]. Find a significant variant according to GWAS method we use two parameters (P-value and effect size), more your sample is small, more the possibility of finding significant variants is weak, capacity to detect variant with small effect is low. These lacks are maybe dues to several reasons:

- Environmental diversity.
- Genetic factors.
- Genotyping errors
- Participant consent has the issues of sharing genotype data.

To improve these effects and find the significant variants, it's more important to enlarge the sample size. Unfortunately, It's not always obvious due to the cost, the possibility for having access sharing to different raw data is limited. One of new strategies that we can use to figure out into this challenge is to combine different GWAS studies from several population studies based on the same phenotype, the idea is to aggregate the effects of variant across all studies and pooling in single study, this strategy is called "GWAS Meta-Analyses" [79].

2.5.2 Aim of Using GWAS Meta-Analysis

GWAS Meta-analysis is a mathematical method used to integrate the outcomes of various experiments in order to see whether there is a major overall effect. The aim of this method is:

- To increase the statistical power by enlarging the sample size.
- To estimate the average effect size for total samples
- To aggregate the effects from different and independent GWAS studies.
- To assess the heterogeneity of studies.

Due to its broad potential to integrate numerous genome-wide interaction studies in a single study, this methodology becomes common instruments, leading to the identification of association

of SNPs having previously small effect sizes, also the capacity to decrease false positives and improve statistical power of true positives causal [80]. These effect sizes between populations in a meta-analysis may differ, these are due to many factors (geographical positions, ancestry, environment, etc ..) will conduct the *"Heterogeneity"*, if the heterogeneity is observed during this GWAS meta-analysis studies, the understanding of its causal is relevant. Because, the better understanding of correct interpretation can conduct the better understanding of the phenotype (disease, trait, etc ..) and a more accurate to the replication study.

It's quite tricky to interpret the heterogeneous result, That the conventional approach to analyzing the studies' correlation p-values does not accurately predict whether in each sample the impact or effects occurs. We would suggest a structure in this section that promotes the understanding of the meta-analysis process. Our method would be under new metric reflecting the posterior probability in each sample the effect occurs, which is calculated using cross-study results. We will apply the real data in our approach, the studies predicted not to have an impact and the uncertain studies that are under-powered, which will essentially segregate the studies predicted to have an effect. This method will increase the power of significant SNPs under study from GWAS studies.

2.5.3 Fixed Effect (FE) Method

Fixed-effect model is a common approach used in statistics. Assume that for both experiments, the real result is the same (meaning that, in FE method the extent of the effect size remains the same or fixed within the studies), and in case of variations, the estimates between studies are caused by sampling errors.
The simplest GWAS Meta-analysis is based on Fisher's method. He tries to combine P-values from different populations across all studies. Let assume, we have n studies, we define:
$\forall\ P_1 ... P_n$ of each study. then, we get:

$$X^2 = -2 \sum_{i=1}^{n} Log(P_i) \tag{2.35}$$

where X^2 follows $\chi^2(2n)$ (means chi-squared distribution with 2n degrees of freedom). This approach is powerful and simple, but it presents some limitations, such that:

- The weights of all studies are equal, this will conduct high sub-optimal when we will combine GWAS studies with different sample sizes.
- The direction of effect in each study is not considered.

- The differences observed are due to chance.
- Study size of each population is not taken into account.

To improve this method, we propose two types of GWAS Meta-Analysis approaches, which are able to consider study sample size, we have:

a) Z-test Meta-Analysis
b) Inverse Variance Meta-Analysis

To use these above methods we need information of each study i and for each SNP.

a) Z-test Meta-Analyses

To perform Z-test for GWAS Meta-analysis, we need for each study i and each SNP:

- Study population size will use to define the weight
- Significance and sign of effect : Z_i

We use this approach when the effect estimates values can not be combined.
Therefore, we have: $\forall i \in \{1, ..., N\}$,

- Study population size : n_i, (i=1,...,N).
- Z-test (with sign indicating effect direction) Z_i (i=1,...,N)

We weight each study by:

$$W_i = \sqrt{n_i} \tag{2.36}$$

We pooled test statistics:

$$Z = \frac{\sum_{i=1}^{N} W_i Z_i}{\sqrt{\sum_{i=1}^{N} {W_i}^2}} \tag{2.37}$$

$$Z_i = \Phi^{-1}(1 - \frac{P_i}{2}) * e_i \tag{2.38}$$

Z has a standard normal distribution
e_i = effect direction for study i
P_i is the P-value for the i^{th} study

b) Inverse-variance weighted in GWAS Meta-Analysis

Z statistic is calculated as the inverse normal of a p-value from any appropriate statistical test and given the direction sign of the association. The Inverse-variance for GWAS Meta-analysis needed for each study i (with i = 1,...,N) and each SNP:

- Effect size estimate of each study : β_i $(i \in 1, ..., N)$.
- Standard errors of estimates effect size β_i : S_i

In the fixed-effects model, the inverse-variance combines regression results as a weighted sum rather than test statistics. The weights are the opposite of the variance of the impact predictions for this model. Then we have, $\forall$ i $\in \{1, ..., N\}$

- Weight is given by :
$$W_i = \frac{1}{{S_i}^2} \tag{2.39}$$
- Pooled estimate of effect:
$$\beta = \frac{\sum_{i=1}^{N} W_i \beta_i}{\sum_{i=1}^{N} W_i} \tag{2.40}$$
If your traits are dichotomous β_i is given by $\beta_i = log(O_i)$
- Pooled estimation of standard errors:
$$S^2 = \frac{1}{\sum_{i=1}^{N} W_i} \tag{2.41}$$
Then we can combine pooled β and S, pooled test statistics:
$$T^2 = \frac{\beta^2}{S^2} \sim {\chi_1}^2 \tag{2.42}$$

2.5.4 Random Effect Method

The idea of combining different GWAS studies and aggregate them in single analysis remains a major goal of GWAS meta-analysis. This method increases our understanding of variant power improvement of GWAS study with small effects. Despite these advantages, it's important to notice that combining GWAS studies may solve problems of sample size but it will conduct heterogeneity issues because of respective effect sizes from different GWAS studies. Therefore, finding out a new approach that takes account of the heterogeneity issue is very relevant. The random effect method in Meta-analysis comes with a new approach considering the heterogeneity as one of the factors responsible for genetic effects incapacity in the variants.
The objective is to approximate the mean and variance of effect sizes for the underlying population. The traditional way for proceeding the analysis in Random Effect has 2 steps:

1. Estimate the scale of the impact and its confidence interval by taking account of heterogeneity.
2. The magnitude of the effect size is normalized to a z-score that is translated to a p-value (Equivalent under the null hypothesis of heterogeneity).

Here, it is important to know that in each sample the mean result of each SNP is different from μ (mean).

2.5.5 Heterogeneity

Heterogeneity is the key of decision between fixed model and random effect, this phenomenon is due factors, such as sampling error, genetic environmental which can lead to the genetic effect of each study. Some experiments have shown that where the effect size is the same, variability may still exist, but the linkage disequilibrium structures within studies are different [81].

To examine the heterogeneity test we assume:

- H_o : No differences (or by chance) of genetic effects existing between studies using cochran's Q test.
- H_1 : There is variation of genetic effects between different studies based on cochran's Q test.

This Cochran's Q test is a non-parametric statistical test to verify whether inter-studies analysis have identical effects or not [82].This Cochran Q significance test is determined by summing the squared deviations of the estimation of each sample from the overall estimate and then comparing it with the distribution of chi-squared with degrees of freedom of k-1 (df) (where k is the number of studies).' [83, 84], we use:

$$Q = \sum_{i=1}^{N} W_i(\hat{\beta}_i - \hat{\beta}_{avg})^2 \tag{2.43}$$

$$with \ \hat{\beta}_{avg} = \frac{\sum_{i=1}^{N} W_i\hat{\beta}_i}{\sum_i W_i} \tag{2.44}$$

$$and \ \hat{\beta}_i = log(OR_i), \ also \ W_i = \frac{1}{{S_i}^2} \tag{2.45}$$

Then we have the weight of each analysis integrating the variance of heterogeneity between studies, is given by:

$$\lambda^2 = \frac{Q-(k-1)}{M} \tag{2.46}$$

$$with \ M = \sum_i W_i - (\frac{\sum_i {W_i}^2}{\sum_{i=1} W_i}) \tag{2.47}$$

Therefore, the weight for the random effects model is calculated by:

$${W_i}^r = \frac{1}{(\frac{1}{W_i} + \lambda^2)} \tag{2.48}$$

$$Q^r = \sum_{i=1}^{N} {W_i}^r(\hat{\beta}_i - \hat{\beta}_{avg})^2 \tag{2.49}$$

We use Q^r to check heterogeneity, from this test if P-value <0.010 indicates the presence of heterogeneity.

Remark Cochran's Q test is reliable if we have a large number of studies included in our GWAS meta-analysis.
In the case of a small number of studies, Q test becomes under-power to assess heterogeneity. Then, In addition to sampling error I^2, we have the percentage of variance in impact estimates across studies due to heterogeneity, given by:

$$I^2 = max(0, \frac{100 * (Q - (k-1))}{Q}) \tag{2.50}$$

This I^2 is calculated by SNP. If I^2 is close to 0, effect sizes across studies for this SNP are reasonably consistent, it means we have low heterogeneity. If I^2 is close to 100% then there is high heterogeneity between study in that SNP, we have:

$$I^2 \ \ is \ \ in \ \begin{cases} [0\%, 25\%[, \text{ We have low effects} \\ [25\%, 50\%[, \text{ The effect is small} \\ [50\%, 75\%[, \text{ The effect is moderated} \\ [75\% - 100\%], \text{ We have high effect} \end{cases} \tag{2.51}$$

2.5.6 Binary Effect Method

Binary effect method is one of the popular methods used to perform GWAS Meta-analysis. This approach is based on posterior probability using Bayesian approach to predict the effect of variants in our study called *"M-value"* [85]. The methodology is a weighted sum of the z-scores formula that applies a higher weight to the studies that are supposed to have an impact and a lower weight to the studies that are expected to have no effect.[85].
Heterogeneity remains challenge in GWAS Meta-analysis among studies, to come out with this issue this present method propose two assumptions [85]

1. The effect is either present in the studies or missing.
2. If the effect occurs, then the effect sizes between experiments are identical.

We try to emphasize in the first assumption that in some studies the effect sizes are often found to be much smaller than in others. This tendency can be caused by multiple factors such as Gene-environmental interactions, sampling error,etc ..

2.5.7 Estimate M-value

In each meta-analysis study, the M-value is a posterior probability that the effect occurs [85]. Let K_i is an effect size of study i, with $i \in \{1, ..., N\}$ and η_i variance of K_i. Notice that K_i is normally

distributed because of large sample size. Let Consider K = $\{K_i\}$ as observed data, then we have:

$$P(K_i|E) = \begin{cases} \mathcal{N}(K_i; 0, \eta_i), & \text{if E = No effect} \\ \mathcal{N}(K_i; \mu, \eta_i), & \text{E = There is effect} \end{cases} \tag{2.52}$$

Bayesian approaches require prior probability to perform posterior probability [85]. So, the prior for effect size is:

$$\mu = \mathcal{N}(0, \sigma^2) \tag{2.53}$$

With $\sigma^2 \leq 0.2$ for small effect and ≥ 0.4 for large effect.

Let $E_i \in \{0, 1\}$:

$$E_i = \begin{cases} 1, & \text{There is effect} \\ 0, & \text{There is no effect} \end{cases} \tag{2.54}$$

Now we can define the prior probability of study i having effect:

$$P(E_i = 1) = \varpi, \quad with \quad i \in \{1, ..., N\} \quad and \quad \varpi \sim \mathcal{B}(\alpha, \beta) \tag{2.55}$$

As we have N studies, then the possible number of E_i for N studies is 2^N. Let Assume $r = 2^N$, we have a new set of possible values of E_i designed by $L = \{y_1, ..., y_r\}$.

Therefore, we can estimate the m-value of study i having effect using Bayes' Theorem.

$$m_i = P(E_i = 1|K) = \frac{P(K|E_i = 1)P(E_i = 1)}{P(K|E_i = 0)P(E_i = 0) + P(K|E_i = 1)P(E_i = 1)} \tag{2.56}$$

If we consider $G \subset L$, with G set of effect value, then our m-value becomes:

$$m_i = \frac{\sum_{y \in G} P(K|E = y)P(E = y)}{\sum_{y \in L} P(K|E = y)P(E = y)} \tag{2.57}$$

Therefore, we just need to know the posterior probability of E for each, then we have:

$$g(y) = P(K|E = y) * P(E = y) \propto P(E = y|K) \tag{2.58}$$

Then the prior of E :

$$\begin{aligned} P(E = y) &= \int_{-\infty}^{\infty} P(E = y)P(\varpi)d\varpi \\ &= \int_{-\infty}^{\infty} \varpi^{|y|}(1 - \varpi)^{N-|y|} \frac{1}{\mathcal{B}(\alpha, \beta)} p(\varpi)d\varpi \\ &= \int_{-\infty}^{\infty} \varpi^{|y|}(1 - \varpi)^{N-|y|} \frac{1}{\mathcal{B}(\alpha, \beta)} \varpi^{\alpha-1}(1 - \varpi)^{\beta-1}d\varpi \\ &= \frac{1}{\mathcal{B}(\alpha, \beta)} \int_{-\infty}^{\infty} \varpi^{|y|+\alpha-1}(1 - \varpi)^{N-|y|+\beta-1}d\varpi \\ &= \frac{\mathcal{B}(|y| + \alpha, N - |y| + \beta)}{\mathcal{B}(\alpha, \beta)} \end{aligned}$$

We can also calculate P(K | E=y):

$$\begin{aligned} P(K|E=y) &= \prod_{i\in y_o} \mathcal{N}(K_i;0,\eta_i) \int_{-\infty}^{\infty} \prod_{i\in y_1} \mathcal{N}(K_i;\mu,\eta_i) p(u) du \\ &= \overline{D}.\mathcal{N}(\overline{K};0,\overline{V}+\sigma^2) \end{aligned}$$

with D scaling factor, given by:

$$D = \frac{1}{(\sqrt{2\pi})^{N-1}} \sqrt{\frac{\prod_i W_i}{\sum_i W_i}} e^{\{-\frac{1}{2}(\sum_i W_i {K_i}^2 - \frac{\sum_i W_i {K_i}^2}{\sum_i W_i})\}} \tag{2.59}$$

Now, if our assumptions are held. Then, we can weight sum of z-score method based on m-values into the weight

$$S_{be} = \frac{\sum_i m_i Z_i}{\sqrt{\sum_i {m_i}^2 W_i}} \quad , with \quad Z_i = \frac{\beta_i}{\sqrt{\eta_i}} \tag{2.60}$$

Also $\sqrt{W_i} \approx \sqrt{Np(1-p)}$, p is a minor allele frequency.

2.5.8 Tools for Conducting GWAS Meta-Analysis

a) METAL tool (Meta-Analysis)

Goncalo Abecasis, Yun Li and Cristen Willer [4] developed METAL, the most common GWAS Meta-Analysis, and the first version was produced in 2007. It has since become a very common instrument for the study of GWAS scans. This method is particularly suitable because results from individual research can not be evaluated together because of variations in race, distribution of phenotypes, gender or sharing restrictions placed on individual level data [86]. Either test statistics and standard errors may be mixed, or p-values through studies (Consideration of sample size and direction of effect.), the approach used is Fixed effect Method [84]. The method implemented in METAL was based on two strategies:

- A weighted method of Z-score in each study based on sample size, P-value and direction of effect.
- An effect-size dependent methodology weighted by the standard error unique to the sample.

This technique is computationally effective for meta-analysis of genome-wide association scans, a widely used method to facilitate gene mapping studies of power-complex traits. In order to facilitate the study of very broad data sets and to accommodate a range of input file formats, it provides a rich scripting interface and implements effective memory management [4]. Asymptotically, where the distribution of the trait is similar across samples, the two methods are equal (such

that standard errors are a predictable function of sample size). The table below 2.1 is a summary of approaches to analytics.

Table 2.1: *Key formulae for both approaches [4]*

	Analytical strategy	
	Sample size based	Inverse variance based
Inputs	N_i - sample size for study i $P_i - P$-value for study i Δ_i - direction of effect for study i	β_i- effect size estimate for study i se_i - standard error for study i
Intermediate Statistics	$Z_i = \Phi^{-1}(P_i/2) * \text{sign}(\Delta_i)$ $w_i = \sqrt{N_i}$	$w_i = 1/\text{SE}_i^2$ $se = \sqrt{1/\sum_i w_i}$ $\beta = \sum_i \beta_i w_i / \sum_i w_i$
Overall Z-Score	$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$	$Z = \beta/SE$
Overall P-value	$P = 2\Phi(\lvert -Z \rvert)$	

Despite its famous popularity, these approaches have some disadvantages in terms of effects, one of them is that the tool assumes that the effects across the studies are the same, this assumption remains a challenge because of heterogeneity issue between-study.

b) GWAMA (Genome-Wide Association Meta Analysis)

Despite the success of GWAS for detecting the loci associated with the trait, the low effect is still a challenge. Meta-analysis of studies from the same population is one way to boost the ability to find more novel loci, increasing the sample size over each particular study. [5]. Despite proliferation of different tools implemented for GWAS Meta-Analysis, most of them presented the challenges here below:

- Memory efficient data manipulation [5].
- Due to population structure, and between sample differences, over-dispersion of GWA test data, all of which must be accounted for in the meta-analysis [5].
- Computational problems that can be aligned to separate strains in merging findings obtained using different GWA genotyping materials [5].

To overcome some of these issues, a new GWAS Meta-Analysis software was developed for answering these challenges, the software was called *"GWAMA"*. Where appropriate, it has implemented tools to align studies with the same reference strand, regardless of the GWA genotyping result, and optionally carries out genomic control based on summary statistics to correct the population structure for each study and possible difference between [5, 87] studies and potential variation within studies. Some of advantages of using this tool based on method implemented is that, With both direct genotyped and imputed SNPs, fixed effect and random effect meta-analyses are conducted using allelic odds ratio figures of 95% confidence intervals for binary characteristics, allelic effect size estimates, and standard error for quantitative phenotypes. [5]. This tool present several advantages than others:

- The ability to exchange supplementary scripts with software to allow study summary statistical files created by commonly used GWA analysis tools to be pre-processed, also visualize graphical summaries results of the meta-analysis
- The possibility of calculation two metrics of heterogeneity $(I^2 \ \ and \ \ Q)$ of allelic effects between studies.
- Capacity for performing random effects meta-analysis in case of heterogeneity
- Allow the population structure, genomic control correction of the correlation findings of each sample, and the overall meta-analysis.

The only challenge with this tool is about space.

c) MetABEL

For the meta-analysis of genome-wide correlation scans between quantitative or binary traits and SNPs, MetABEL is a R package, extracted from the GenABEL package. It has been developed by Maksim Struchalin and Yurii Aulchenko. This package uses a fixed effect method to analyse GWAS Meta-Analysis [88] .

d) Metasoft (Meta-analysis Software)

METASOFT is a genome-wide association analysis meta-analysis software, designed to effectively perform a variety of simple and advanced meta-analytical approaches, is able to analyse huge amounts of studies.
This tool is more advanced than many GWAS Meta-analysis software [85] , its has been developed based on :

- Fixed Effect Method (based on inverse-variance-weighted effect size)

- Random Effects model (RE)
- The binary Effects model (The new model of random effects is optimized to recognize associations whether the studies have effect and some not)

In this tool, Estimates of summary effect size Beta and standard error for both models (fixed effects and random effects). Also, from heterogeneity it expected to estimate Cochran's Q statistic, p-value, and I^2.
The tool prevents the effect by integrating the Bayesian notion, the assume M-values as Posterior probability designing the existing effect for each study.
The decisions are made according to m-values results. So, we have:

- If m-value < 0.1, there is no predicted effect in the study.
- If m-value > 0.9): The effect is predicted in the study.
- Otherwise, the prediction effect becomes ambiguous in the study. To visualize the results they use ForestPMPlot, this plot can include the component such as p-value, study name, log odds ratio, standard error and summary statistic [85].

We can summarize these GWAS Meta-analysis software information and methods in this Table below

Table 2.2: *Summary of Meta-Analysis tool with their specificity [5].*

Software Pack	METAL	MetABEL	METASOFT	GWAMA
Pre-processing of GWA analysis files	No	*ABEL	SNPTEST	SNPTEST, PLINK
Strand flipping for aligning effect directions	Yes	Yes	Yes	Yes
Fixed effect analysis	Yes	Yes	Yes	Yes
Random effect analysis	No	No	Yes	Yes
Heterogeneity statistics (Cochran's Q statistic, I^2)	Q	Q	Q, I^2	Q, I^2

Automated genomic control for population structure	Yes	Yes	Yes	Yes
Graphical visualisation of meta-analysis results	No	Forest-plot	Forest-plot M	Separate scripts for Manhattan and QQ plots

2.6 SKAT-Analysis(Sequence Kernel Association Tests) for the Combined Effect of Rare and Common Variants

Over the decades, genome-wide association studies (GWASs) have led to the identification of a large scale of thousand genetic variants associated with several risks of complex traits. However, for a given trait, certain variants have typically explained only a small to moderate portion of the predicted heritability. For many traits, it has been shown that a large proportion of heritability could be described cumulatively by several common variants with limited effects. [89].
To challenge this issue, the researchers have committed new statistical approaches based on assessment of each variant individually with univariate test statistics (case of Cochran-Armitage test trend). After performing the univariate test in each variant, we recorded variants under power for rare variants. This conduct to the new groupwise association test called "SKAT" based on improvement of genetic variant effect for rare variants by using burden and variance component [89]. It has been shown that for any disease-related gene, the various influences of uncommon and common variants are not established a priori, and such a weighting scheme may lead to loss of power when common variants are often associated with disease in an area under investigation [90].

It is better to screen for the cumulative influence of uncommon and common variants using a mixed statistical test that enables rare variants and common variants to completely contribute to the overall test statistics in order to eliminate the unclear outcome of the overall purpose of defining genes comprising disease risk variants, either from rare or common variants.
Combining rare variant and common variants together is the strategy required to first divide in 2 subgroups which are group identifies rare variant and group for common variants, then combining results from association tests these variants resulted in these groups, we will use the combining multivariate collapsing (CMC).
In the CMC method, rare variants (i.e variants with minor allele frequency **MAF < 0.01**) are collapsed together, therefore each common variant forms a separate group [91]. But, it becomes quite difficult to accept the minor allele frequency of variant less than 0.01 to be the threshold of rare variant, means the hypothesis makes sense if the sample size of population is small (e.g 200 individuals), the variants with maf < 0.01 have defined as rare variant, but if we have a large sample size of population (e.g 300.000 individuals) the variants with maf less than 0.01 are not rare. To avoid this problem, the theory has defined a new approach of T-test by: [92].

$$T = \frac{1}{\sqrt{2n}} \tag{2.61}$$

and

$$e_i = MAF_i - T \tag{2.62}$$

if $e_i \geqslant 0$, then the i variant is a rare variant. Otherwise, it is a common variant. with MAF_i is the minor allele frequency of each variant in study and n is the sample size. There are several potential ways to run a test for the overall influence of uncommon and common variants, but the simplest one is based on Fisher's method of integrating p-values from the rare and common variant analyses. Alternative methods are based on using weighted-sum statistics to combine the test statistics explicitly. Let's consider k variants for n subjects sequenced in the genomic region. We partitioned k variants by 2 groups of variants of dimension k_1 (resp. k_2), we consider X as a genotype matrix with $n \times k$ dimensions. Then, we define a regression model by:

$$f(E[Y_i]) = \alpha_0 + \begin{pmatrix} c_{11} & \dots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{k1} & \dots & c_{kp} \end{pmatrix} * \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} + \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{k1} & \dots & x_{kp} \end{pmatrix} * \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \tag{2.63}$$

with Y_i is a phenotype value (cont. or dichotomous) with $i \in \{1, ..., k\}$, $C_i = (c_{i1}, ..., c_{ip})$ is covariate vector, $\alpha_i \in \{1, ..., p\}$ coefficient of covariate and X is a genotype matrix $(x_{ij})_{1 \leqslant i \leqslant k, 1 \leqslant j \leqslant p}$ and $\beta_i \in \{1, ..., k\}$ regression coefficient for genetic variants. This β_i follows normal distribution:

$$\beta_i \sim \mathcal{N}(0, {w_j}^2 \tau), \quad cov(\beta_s, \beta_z) = \rho \quad with \quad s \neq z$$

$H_o : \beta = 0$. The variance component score is given by:

$$Q_\rho = (Y - \hat{m}_0)^{'} F_\rho (Y - \hat{m}_0), \quad with \quad F_\rho = XWT_\rho WX^{'} \tag{2.64}$$

and $T_\rho = (1 - \rho)I + \rho I^{'}$, we define the matrix of weight $W = diag(w_1, ..., w_k)$

Therefore, we can calculate SKAT and Burden by:

- If $\rho = 1$, we have Burden test which is given by:

$$Q_{\rho=1} = \sum_{j=1}^{k} {w_j}^2 (\sum_{i=1}^{k} (Y_i - \hat{m}_{i,0}) x_{ij})^2 \tag{2.65}$$

- If $\rho = 0$, we have SKAT test which is given by:

$$Q_{\rho=0} = \sum_{j=1}^{k} {w_j}^2 (\sum_{i=1}^{k} (Y_i - \hat{m}_{i,0}) x_{ij})^2 \tag{2.66}$$

So, the weights of rare and common variants are given by:

$$w_j = \beta(\hat{MAF}_j, 1, 25) \tag{2.67}$$

with :

- $\hat{MAF}_j$ = Estimated based on the all subject for variant j.
- w_j = Weight of variant j.
- $\hat{m}_0$ represents estimate probabilities vector of Y under the null model.

Therefore, we can join the effects of rare and common variants in genomic regions, therefore the score test statistics for uncommon and common variants are combined as a weighted sum, we define the test as convexity of Q under ϕ,by:.

$$Q = \phi Q_{common} + (1-\phi)Q_{rare} \tag{2.68}$$

$$with\ \phi = \frac{\sigma_{rare}}{\sigma_{rare}+\sigma_{common}} \tag{2.69}$$

2.6.1 Fisher's Combination Method

The present approach conduct the combination of P-values from rare variant (P_{rare}) and common variants (P_{common}) test instead of using test statistics
Let's Consider two p-values (P_{rare}, P_{common}), we have:

$$Q_{F,\rho_1,\rho_2} = -2ln(P_{rare}) - 2ln(P_{common}) \tag{2.70}$$

Under null hypothesis $-2ln(P_{rare}), -2ln(P_{common})$ follow $\chi^2(2)$ and Q_{F,ρ_1,ρ_2} follows $\chi^2(4)$, then we have:

$$E[Q_{F,\rho_1,\rho_2}] = 4 \tag{2.71}$$

$$var(Q_{F,\rho_1,\rho_2}) = 4 + 2cov(-2ln(P_{rare}), -2ln(P_{common})) \tag{2.72}$$

In case cov is quadrature function.

$$cov(-2ln(P_{rare}), -2ln(P_{common})) = \begin{cases} r(3.25+0.75r), & 0 \leqslant r \leqslant 1 \\ r(3.27+0.71r), & -0.5 \leqslant r \leqslant 0 \end{cases}$$ The distribution can be approximated by

$$Q_{F,\rho_1,\rho_2} \sim c{\chi^2}_f \tag{2.73}$$

$$with\ \ c = \frac{Q_{F,\rho_1,\rho_2}}{2 * E[Q_{F,\rho_1,\rho_2}]}\ \ and\ \ f = \frac{2E[Q_{F,\rho_1,\rho_2}]}{var[Q_{F,\rho_1,\rho_2}]} \tag{2.74}$$

- If $\rho_1 = \rho_2 = 1$, we have burden-F (Burden fisher's test)

- If $\rho_1 = \rho_2 = 0$, we have SKAT-F (SKAT fisher's test)

Chapter 3

Materials and Methods

This section is based on methods and materials used during our analysis. From the European Phenome Genome Archive (EGA) described in Table 3.1, Following the basic data access protocols outlined in the standard data access protocols. From three African populations (including Kenya, Gambia and Malawi), we handled a severe Malaria GWAS dataset (N = 11,000) [93]. From international protocol defined by WHO, it has been decided as follows:

- Children with severe cases of malaria were admitted to hospital using WHO protocol definitions for cerebral malaria (Blantyre coma score < 3 for children or Glasgow coma score < 11 for adults)[94].
- Severe malaria anaemia (haemoglobin < 5 g/100 mL or haematocrit < 15%) and other symptoms associated with malaria [94].

So,the control cohorts were taken from representative cases of the ethnic groups or from certain local population research sites [93]. The Illumina Omni 2.5M array and QC samples were genotyped and filtered as defined in Table 3.1 as :

Table 3.1: *Represent the raw-data collected from European Phenome Genome Archive*

S/N	Description	Sample size	EGA Dataset ID
1	Kenya	1944 cases, 1708 controls, 180 parents & 33 other	EGAD00010000904.
2	Gambia	2807 cases, 2786 controls & 1 parents	**EGAD00010000902**
3	Malawi	1194 are cases and 1322 are controls	**EGAD00010000903**

So, the workflow of our work is presented in Figure 3.1 as following:

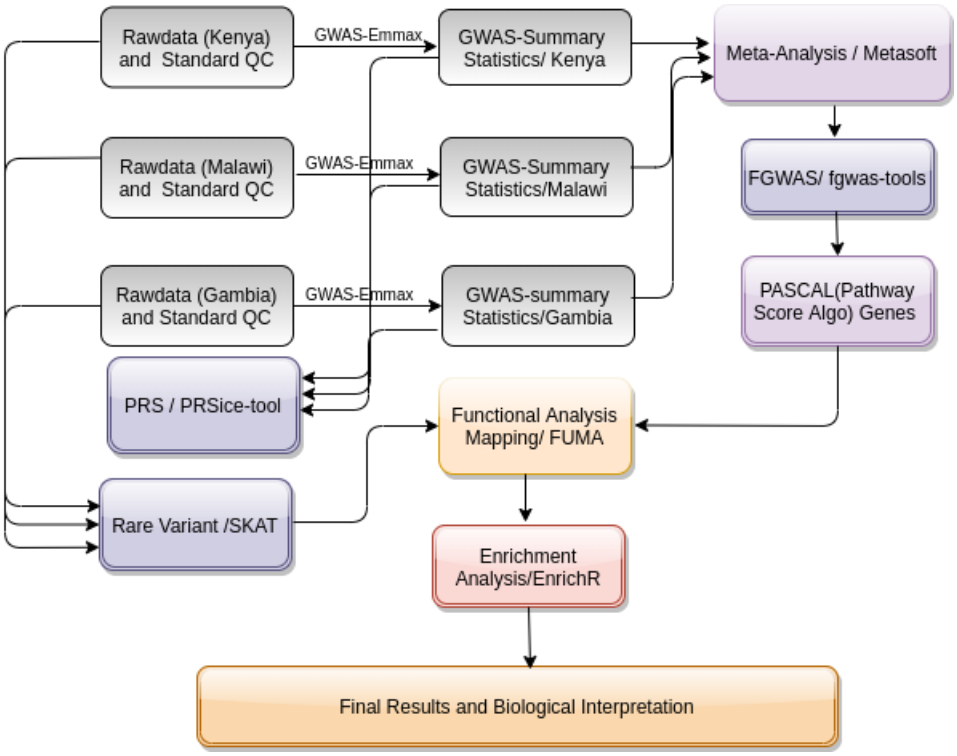


Figure 3.1: *Workflow of Analysis*

3.1 Genome Wide Association Study Analysis

3.1.1 Quality Control (QC)

We then performed basic quality controls (QCs) including SNP missingness, Hardy-Weinberg equilibrium, sample missingness, heterozygosity rate and minor allele frequencies(MAF) using Plink1.9 Software. These include

- We removed SNP with missing proportion greater than 0.05 (means < 95 % of no missing markers (geno 0.05)).
- The individuals having genotype missingness greater than 0.05% (mind 0.05) are excluded.
- SNPs with a low number of minor alleles are rare variants (with a frequency of minor alleles < 0.01), so there is a lack of ability to identify associations of SNP phenotypes. These SNPs are also more prone to genotype errors. We therefore filtered SNPs having MAF(Minor Allele Frequency) greater than 0.01.
- It is usually considered in GWAS that deviations from HWE (Hardy Weinberg Equilibrium) are the cause of errors in genotyping [95]. Thresholds are also less strict in cases than in controls, as in cases the breach of the HWE law may be representative of a real genetic association with the risk of disease. In our case, Markers that deviate from HWE(Hardy-Weinberg equilibrium) using p-value <1e-10 in cases and <1e-6 in controls would be omitted

We performed the Emmax method for the GWAS analysis. Currently, Emmax is one of the best tools used to perform GWAS, Many GWAS studies have demonstrated the success of Emmax, this can be attributed to the fact that only a small fraction of complex traits are described by each loci, which helps us to escape redundant variance component estimation process, resulting in a substantial increase in interaction mapping computational time using mixed model [96].
We therefore performed GWAS analysis then we found 8,133,118 variant SNPs for Gambia, 8,834,631 variant SNPs for Kenya and 7,761,423 variant SNPs for Malawi in our summary statistics. Based on the p-value threshold of 5e-7, we got 5 SNPs in Malawi populations, 10 SNPs in Kenya population and 0 SNPs in Gambia population (see result section 4.1). So, because of the small number of significant SNPs based on p-value threshold (< 5e-7), it becomes more relevant to improve the statistical power of variants to increase the number of significant SNPs associated with severe malaria by using Meta-Analysis approach. Then we have QC results filtered:

Table 3.2: *QC Result after filtering*

S/N	Study Name	Sample size	Size of SNP variant
1	Kenya Population data set for severe Malaria	1944 cases, 1708 controls, 180 parents & 33 other	15845917 variants have passed.
2	Gambia Population data set for severe Malaria	2807 cases, 2786 controls & 1 parents	8138576 variants have passed
3	Malawi Population data set for severe Malaria	1194 are cases and 1322 are controls	7786624 variants have passed

3.2 Cross Populations Meta-Analysis

We performed GWAS meta-analysis using tools including Metal and Metasoft. Briefly, using METAL the approach combines either test statistics and standard errors or p-values across studies (taking sample size and direction of effect into account).

MEtASOFT is considered to be a powerful method due to the integration of Binary effect method, usage of Random, Fixed effect method and Evaluation of Heterogeneity. To use it, we consider only SNPs that were common in all the data sets (Kenya, Gambia and Malawi) were retained for meta-analysis. We used METASOFT tool, which has the potential of carrying out meta-analysis using both fixed, random effect, binary effect and the Han and Eskin random-effects model. We used an in-house python script to obtain the set of SNPs that were common among the data sets in each case. To control any possible confounding from population stratification, we first run all the SNPs without -lambdamean and -lambdahetero parameters in the METASOFT. We then obtained the values as -lambdamean = 0.904629 and -lambdahetero = 0.62 from the log files of the meta-analysis run for the raw genotypes GWAS data. We then repeated the meta-analysis by supplying these values to the METASOFT as recommended. Variants with meta-analysis p-values less than < 5e-08 were considered significant.

From our three GWAS summary statistics, we performed GWAS Meta-Analysis using Metasoft. Then we got 5,942,350 SNP variants , the effects are assessed based on posterior probability of association called *"M-value"*. We will filter the variant SNP with *M-value*>0.9 across these 3 studies (Gambia, Malawi and Kenya), we will therefore consider variants that have passed threshold (see

Table 4.7). We will compare the results from these tools (Metal and Metasoft) and consider the replication variants, also we will observe the novel significant SNPs. .

3.3 Functional Genome Wide Association Study (FGWAS) Analysis

The detection of the genetic polymorphisms causing phenotypic variation in humans and identifying the target molecular mechanisms by which these variations exert their effects is a main goal of research in human genetics [97]. Although, GWAS has answered many questions related to trait and variant, to find causal variant it was always a challenge. Recent researches have improved these difficulties, It has been proven that from regulatory elements and Annotations of gene structures we can inform GWAS [57]. Unfortunately, despite the insurance and significant advances of this method, we always have challenges to find the relevant annotations able to interpret the association study related to the trait.
In the present study, we applied the specific model used in statistics using association computed across the genome to identify groups of genomic elements enriched or depleted for loci influencing the trait based on functional genomic information. This approach is called "FGWAS". To perform FGWAS in our study, we have used "FGWAS-tool" under the statistical model described by Pickrell JK (2014) [64].

To perform FGWAS approach using ("FGWAS-tool"), the tool requires at least 20 significant SNPs for the whole study. However, in our meta-analysis study we had 35 significant SNPs (see Table 4.7), then we will use this GWAS meta-analysis summary statistics to perform FGWAS.

3.3.1 Functional Genomics

A separate branch of genomics called functional genomics has developed to analyze genomic profiles of genes and their functions.Eventually, the elements of a genome express the characteristics to understand the function of genes or proteins. Scanning genomic regions to encoded proteins, SNPs may be detected by putative genes, based on features such as long open reading frames, transcriptional initiation sequences, and sites of polyadenylation. Further evidence would validate a sequence known as a putative gene, such as the similarity of the same organism to cDNA or EST sequences, from known proteins we can predict protein sequence, The association of sequences with promoters or the evidence that an identifiable phenotype is created by mutating the sequence.

The require database analysis contains 451 functional genomics annotated from genome, described the annotations (450 annotations for presence/absence and distance to the closest start point for transcription) which is presented as following:

- DNase-I hypersensitivity data (containing 116 blood cells samples)
- Chromatin State data
- Gene Models

These Annotations are downloaded from 1000 Genomes variants (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/).
There is a number of chromosomes, location (hg19 coordinates), and rs for each variant, accompanied by annotations. "1" indicates that the variant comes inside the annotation, and "0" implies that it does not. The "tssdist" annotation also exists, which is the gap in the "Ensembl gene databases" the closest to transcription start point [64].

3.3.2 Explanation of Analysis

The genome is segmented per block, such as each block contains 5,000 SNPs. from our meta-analysis, then extracted the significant SNP from each blocks. Thus,we reported the Z-score of each SNP from our Meta-Analysis summary statistics. So, the SNP with smallest P-value is called 'causal SNP', is tested into the annotations regions. We then test the annotations by putting all "non-causal" SNPs with annotation c_1, and "causal" SNPs with rate of the annotation c_2 . We considered different numbers of blocks included in the model.
For each simulation, the SNPs were assigned randomly to annotations based on determined rates. Therefore, the model was ran under the assumption that there is association, that is, all blocks include a causal SNP. We then measured power as the proportion of simulations in which there was no overlap of zero in the confidence intervals of the annotation effect.
In this Analysis, we have extracted the relevant functional genomics related to Severe Malaria (Brain and red blood), non-coding part, Binding site, transcription start site,etc ...
To proceed the analysis we have established a workflow that presents the different steps of the analysis.

1. GWAS and Genomic :
   Extraction of genotype information of our target data to the reference data. This new file will contain the following SNPID, CHR, BP, Z-Score, NCase, NControl and Functional Genomics information.

2. Test the annotation individually , it means we will test for each variant if it is depleted or enriched.
3. We will find the annotation able to increase the likelihood of the model, then test it in combination with all other significant annotations.
4 We will keep adding annotations in this manner until there are no annotations that improve the likelihood.
5 Because of over-fitting problems in the model sometimes, it's relevant to use cross-validation for the best fit model. We apply the cross-validation likelihood, and find the penalty with the best cross-validation likelihood.

STEP1

It's important to notice that, in Severe malaria there is 2 genomic annotations interested regions (Brain region and Red blood region), because as infection disease caused by parasite, most of pathogenesis is developing in red blood, also as severe malaria it has been shown that at the high stage this disease can cause and lead a cerebral Malaria, Which is the most significant neurological complication of *Plasmodium falciparum* infection, these attack the brain. We therefore decided to extract our genotype information using our target summary statistics to the reference annotation data integrated in the tool, only for Brain and Red blood, also to use binding site information (TSS, enhancer, promoter, etc ...). Then we have recorded 17 annotations, presented as following:

1. Non-Synonymous.
2. Enhancer (K562).
3. DNase (CMK Leukemia line)
4. TSS-dist (0-5 Kb, 5-10 Kb)
5. Repressed K562
6. Ens-exon coding
7. Ens-exon noncoding
8. CD3+ Cells
9. CD4+ Cells
10. CD56+ Cells
11. TH1 T Cells
12. Ens-utr3-exon
13. Ens-utr5-exon
14. TSS-distance

15. CD8+Cells
17. Brain-microvascular

We are going to test among these annotations, which one has a best likelihood. Find an annotation able to increase the likelihood of the model, then test it in combination with all other significant annotations.

The conclusion is made by 2 factors:

- If PPA (posterior probability of association) > 0.9 and p-value < 5e-8, then we have a true positive.
- Otherwise, we have true negatives.

3.4 Functional mapping and annotation of genetic associations (FUMA)

We will use the summary statistics of our GWAS meta-analysis across 3 populations to perform the functional annotations and mapping . The pipeline that recognizes genomic risk loci and prioritizes possible causal genes, has been introduced by integrating data from several sources (such as ENCODE, GTE, Roadmap Epigenomics Project and chromatin interaction information). From 1000 Genome version 3 of the African population, a pre-calculated LD structure has been used to test the risk loci and significant SNPs independent of GWAS summary statistics. These independent significant SNPs are classified as genome-wide significant SNPs (P-value < 5e-7) within (LD threshold $r^2 > 0.6$), near proximity SNPs (< 250 kb) are known to be a single locus, such that multiple independent significant SNPs and lead SNPs can be found in each genomic locus. We will identify the different genes mapped from our SNPs, and we will map the genes to cell-type tissue, and find their respective gene functions. Also will use GSEA to evaluate the statistical significant of different genes and to concorde differences between two biological states

3.5 GeneSet and Enrichment Analysis Using GENEMANIA

For Geneset analysis, we used GENEMANIA which has an aim to provide an affordable interface to query genomic [98], functional interpretation of gene data and proteomic. It can be used for several of single gene queries [99]. In our study GeneMANIA will help us to find the genes closely related to the networks and weighted these data based on gene co-annotation patterns from the biological function hierarchy of Gene Ontology . We will evaluate the interaction between genes,

their pathways.
For further enrichment analysis we used Enrichr, this database tool is highly worthy and flexible, it contains background libraries providing relevant information to pathways analysis,transcription regulation, protein proteins interactions, ontologies including GO(Gene Ontology) for human phenotype ontologies. Its cover signatures treat cells with drugs, gene expressions of different cells and tissues. This library covers over 200,000 sets of annotated genes from over 70 resources. The tool is accessible through the API and offers multiple ways to simulate the performance.
From our FUMA results, we used the gene-set result into Enrichr to identify the group of genes and proteins which are over-represented in a large set of genes and proteins database, verify the disease associated [100]. Transcriptomics tools and proteomics outcomes also classify thousands of genes that are used for study, using statistical methods to identify substantially enriched or deficient classes of genes.
Similarly as above the FUMA results, was analysis further analysis using Enrichr analysis to find the different pathways and disease related to our gene-list

3.6 Data Analysis for Rare and Common Variants

The hypothesis of GWAS is based on common variants and common disease. So,this hypothesis limitate the potential power for association of variants with lowest or smallest minor frequency (case of rare variants). To analyze rare variants, we applied the robust unified sequence kernel association test (SKAT-O), this test combine burden and variance component analysis to GWAS dataset.To harmonize genotype, we moved trash SNPs with strand and position different, we phased using SHAPEITv2. To improve the quality of data, we imputed our data using impute2, the result was $\sim$ 20 millions in each population. After quality control and imputation accuracy, we retained approximately 15,000 SNPs in each population. We filtered data based on SKAT-O procedure. To conduct the analysis of rare variants is important to filter the variants with MAF$<$ 1%, and from these filtered variants we will aggregate the effects by combining the different tests, using SKAT and Burden.

We have collected different samples from 3 study populations (Kenya, Malawi and Gambia) based on Malaria phenotype. We recorded the below information:

- 3142 Kenya samples and 8834631 variants
- 2516 Malawi samples and 7786624 variants.
- 4920 Gambia samples and 8133118 variants

From these above information, we merged the genotype data using plink, then we obtained a merge file containing the information of these 3 studies, in plink formats, we found merge files with 5658 samples and 9573062 variants. To perform SKAT analysis. We need 2 important files (SetID file and SSD file). To get a SetID file, we have to extract SNPs information from the bim file and we map these variants into the corresponding genes using dbSNP, this file will contain 2 columns (GeneSetid and SNPid). From the SetID file, you can generate an SSD file. then we got:

- 163 samples have either missing phenotype or missing covariates.
- 1276043 Number of SNPs.
- 5658 Number of Individuals.
- 1416 decoded.
- 40825 total number of sets.
- 30806 sets
- 22334 genes uniquely sorted.

We performed linear regression under the null hypothesis. As we are using binary traits, the phenotype information contains the binary values (1 and 0), then we run this model, by testing the model under "SKAT" and "SKAT-O". We will use the output to evaluate the MAPs (minimum achievement p-value) and map then at the gene-level.

3.7 Polygenic Risk score Analysis (PRS)

The present part is focused on analysis of PRS. This approach is based on summation of traits associated with alleles across many genetic loci, which we weighted the effect sizes estimated from GWAS. Using this method we are expecting to find out genetic signals of studies with low power, to infer a trait's genetic architecture, to screen for Malaria in clinical trials, and to act as a Malaria biomarker. Our PRS analysis will use PRSice ('precise') software. Its use to calculate, apply, evaluate and plot the results of PRS. It will also perform the best-fit PRS, as well as come out the results based on calculation at broad P-value thresholds. We use this study in each population. From GWAS, we have k SNPs, with i = 1, 2, ...,k. Each SNP is associated with the trait with P-value P_i, we define SNP genotype $W_{i,j} = \{0, 1, 2\}$ with j represent individual, where j = 1, 2, ... , N and the phenotype. Thus, based on the additive assumption made by GWAS, effect size for each variant is estimated by β_i , this is the effect of a unit able to increase in genotype, W_{ij} , on the phenotype.

We use SNPs for inclusion in polygenic risk score based on the degree of evidence, P-value, for

their association to base phenotype in a GWAS. For P_i smaller than a threshold we will consider SNP_i to be included in a PRS. We therefore calculate PRS by a different number of P-value thresholds P_T .
Thus, PRS for individual j based on threshold P_T is calculated as:

$$PRS_{P_T j} = \sum_{i=1}^{k} \beta_i W_{i,j} \tag{3.1}$$

Let's evaluate PRS on Malawi Data. To perform PRSice, we need to run the standard quality control using plink.
Generally, Polygenic risk scores (PRS) are calculated across a large number of individuals from the *"Target phenotype"* dataset uses n associated SNPs, the genotypes of which have either effect (or not) on the "Base phenotype". We therefore evaluate the distribution of genetic overlap phenotype between population using base and target dataset. This genotype results can be inferred from a univariate regression of the base phenotype on each SNP, such as from a genome-wide association study (GWAS).

3.7.1 Base Phenotype Dataset

Our base data set was downloaded from MalaGen, which is a Genome-wide study of resistance to severe malaria in eleven populations summary statistics, we recorded 18429068 variants from case-control added per population. From this data set, we extracted each GWAS summary statistics related to the concerned population. Since, the present Analysis have 3 studies populations (Kenya, Gambia and Malawi) then each population has its own base data set extracted from Genome wide study of Malaria resistance and severity. We excluded all variants with low quality of imputation (< 0.7).

3.7.2 Target Phenotype Data set

We used genotype data from the MalariaGen consortium for African population, presented as following:

Table 3.3: *Represent the raw data collected from MalariaGen, used as Target phenotype data set for each PRS analysis*

S/N	Study Name	Sample size
1	Genome wide study of resistance severe Malaria in eleven populations(Kenya)	1944 cases, 1708 controls, 180 parents
2	Genome wide study of resistance to severe Malaria in eleven populations(Gambia)	2807 cases, 2786 controls & 1 parents
3	Genome wide study of resistance to severe Malaria in eleven populations(Malawi)	1194 are cases and 1322 are controls

We did quality control on this dataset, by removing individuals with missingness > 1%, SNPs with MAF < 1% and SNPs not in HWE (P < 1e-05) were also removed, then our targets phenotypes data sets are presented as:

Table 3.4: *Represent the raw data collected from MalarianGen, used as Target phenotype data set for each PRS analysis*

S/N	Study Name	Sample size
1	Genome wide study of resistance severe Malaria in eleven populations(Kenya)	1505 cases, 1474 controls, 180 parents
2	Genome wide study of resistance to severe Malaria in eleven populations(Gambia)	2429 cases, 2491 controls & 1 parents
3	Genome wide study of resistance to severe Malaria in eleven populations(Malawi)	1194 are cases and 1322 are controls

The linkage disequilibrium (LD) was accounted by selecting the SNP in the base phenotype data set with the lowest discovery P-value in a sliding window of 250kb, only retaining variants with a pairwise LD with $r^2 < 0.1$, according to LD calculated in the target data set. We performed high-resolution scoring by testing every threshold between P_T = 0.0001 and P_T = 0.5 at increments

of 0.00005, when we apply PRSice, we exclude SNPs in LD (Linkage disequilibrium) and add PCA (principal components analysis) to control for population structure.

Chapter 4

Results

4.1 GWAS Results

4.1.1 Quality Control (QC)

From the Table 3.2, we have QC results filtered for the variants. We performed GWAS analysis for each population using Emmax. In our analyses, the biases from population structure was well-contained ($GC =\sim 1$) as shown in Figure 4.1. we identified a total of 10 ,0 and 5 GWAS significant SNPs (based on threshold < 5e-7) in Kenya, Gambia and Malawi respectively as shown in Figure 4.1 genomic control anlysis and Table 3.2

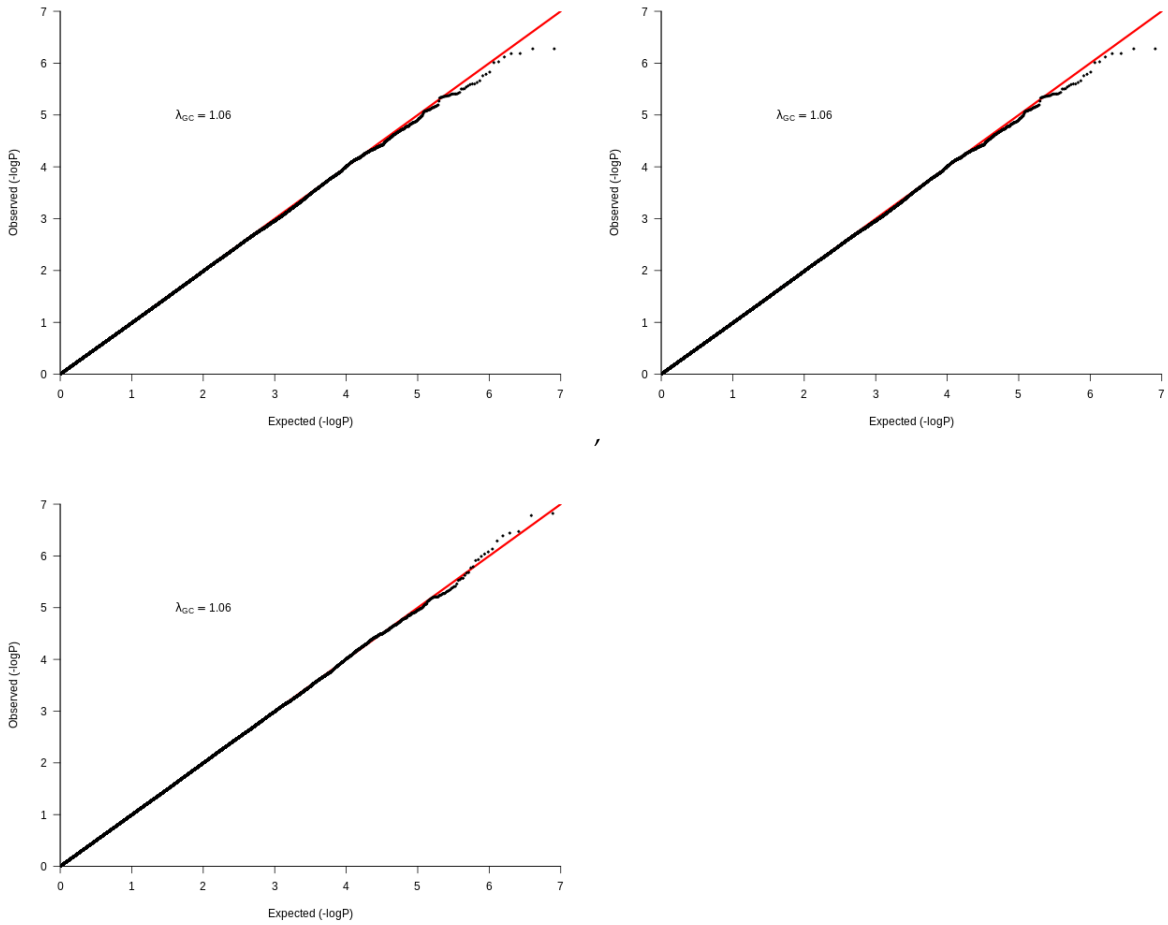


Figure 4.1: *From left to right, we have qqplot of Kenya, Gambia and Malawi based on population stratification*

a) Malawi Population

After performing GWAS using Emmax we got summary statistics result having 7,761,423 SNPs, and after filtering using based on threshold < 5e-7, we have the result presented on 4.1, given as :

Table 4.1: *Summary Statistics of significant SNPs*

CHR	BP	SNP	MAF	A1 / A2	BETA	S.E	P
2	21536042	rs11883957	0.1714	A/C	-0.09554687629	0.01868006014	3.375548093e-07
2	21536873	rs13427673	0.1721	A/G	-0.09493403663	0.01860868075	3.619283502e-07
15	75140207	kgp20017291	0.1898	A/G	-0.09533428839	0.01810268906	1.509921293e-07
15	75143525	rs56981059	0.1671	C /CAG	-0.1015045559	0.01933767982	1.656666137e-07
15	75146861	rs74831792	0.191	T/C	-0.09326829454	0.0183670664	4.093333867e-07

We have recorded 4 variants filtered based on threshold. Therefore, we can map them at the gene-level, then we have:

Table 4.2: *SNPs to Genes based on Biocard*

CHR	SNP	Band	gene	gene_biotype	P
2	rs11883957	NA	NA	NA	3.375548093e-07
2	rs13427673	NA	NA	NA	3.619283502e-07
15	rs56981059	q24.1	*SCAMP2*	protein_coding	1.656666137e-07
15	rs74831792	q24.1	*SCAMP2*	protein_coding	4.093333867e-07

Then we found only one gene "*SCAMP2*", This gene is part of the *SCAMP* family of proteins, it has as specifical function to secretory carrier of membrane proteins. He has a function of carrying the surface of cells in post-golgi recycling pathways. Therefore, we can visualize Malawi summary statistics in Manhattan plot:

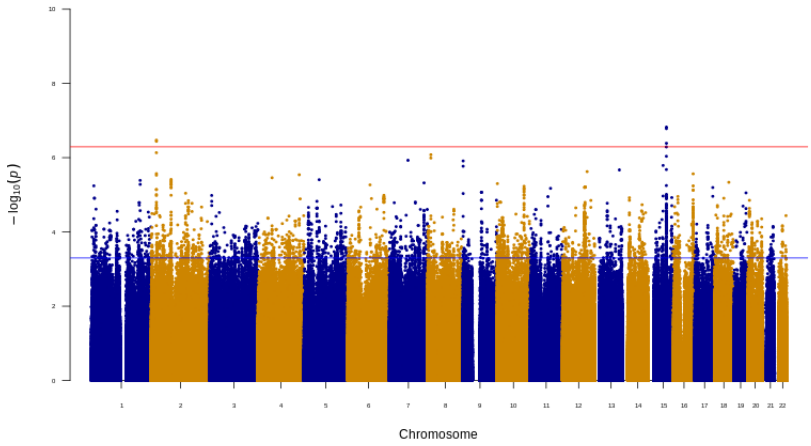


Figure 4.2: *Only 5 variants SNPs have passed the cut-off of p-value < 5e-07*

b) Gambia Population

In this study no variant snp has reached the threshold for p-value < 5e-07. We can visualize the result in Manhattan plot

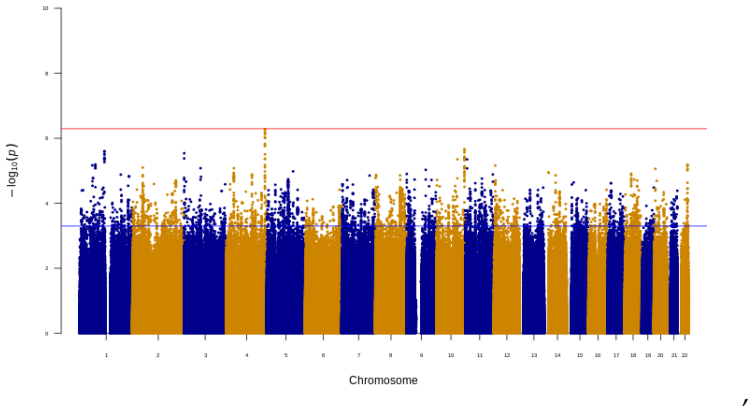


Figure 4.3: *Manhattan plot of Gambia GWAS based on his Summary Statistic*

c) Kenya Population

After performing GWAS Analysis based on Emmax applied in Kenya's Population and Filtering this summary statistics using threshold of p-value < 5e-07, we have results on Table 4.3, containing only 9 significant SNPs.

Table 4.3: *We find out 9 SNP variants passed the threshold*

CHR	BP	SNP	MAF	A1 / A2	BETA	S.E	P
1	238101255	rs570560	0.3322	C/A	-0.3289450132	0.06409746257	3.041926977e-07
4	37467839	rs1479764	0.1302	T/C	0.5691018221	0.0893084461	2.134275545e-10
4	37468127	rs148442339	0.1289	G/GTA	0.5686109224	0.08974262889	2.69457921e-10
4	37468270	rs77403150	0.09642	C/G	0.5968787134	0.1008968057	3.659750271e-09
4	37469435	rs16993572	0.1301	G/C	0.5694313557	0.0893034102	2.079329902e-10
4	37469735	rs78238616	0.1301	C/T	0.5694313557	0.0893034102	2.079329902e-10
4	37475006	rs375672236	0.0969	A/ATGAT	0.5876234886	0.1005603208	5.635308319e-09
5	102305398	rs6884577	0.1411	T/C	0.4371233347	0.08581364111	3.7154558e-07
7	83354218	rs6950669	0.3391	G/A	0.3244294498	0.06427159116	4.724085519e-07

We can map these significant SNP variants at the gene-level

Table 4.4: *Significant variants mapped onto the Gene-level*

CHR	SNP	Gene	Band	Gene_biotype	P
1	rs570560	NA	NA	NA	3.041926977e-07
4	rs1479764	*C4orf19*	p14	protein_coding	2.134275545e-10
4	rs148442339	*C4orf19*	p14	protein_coding	2.69457921e-10
4	rs77403150	*C4orf19*	p14	protein_coding	3.659750271e-09
4	rs16993572	*C4orf19*	p14	protein_coding	2.079329902e-10
4	rs78238616	*C4orf19*	p14	protein_coding	2.079329902e-10
4	rs375672236	NA	NA	NA	5.635308319e-09
5	rs6884577	*PAM*	q21.1	protein_coding	3.7154558e-07
7	rs6950669	NA	NA	NA	4.724085519e-07

Then we find only 2 Genes mapped at the gene-level (*C4orf19* and *PAM*). We can visualize GWAS Summary statistics with Manhatthan plot.

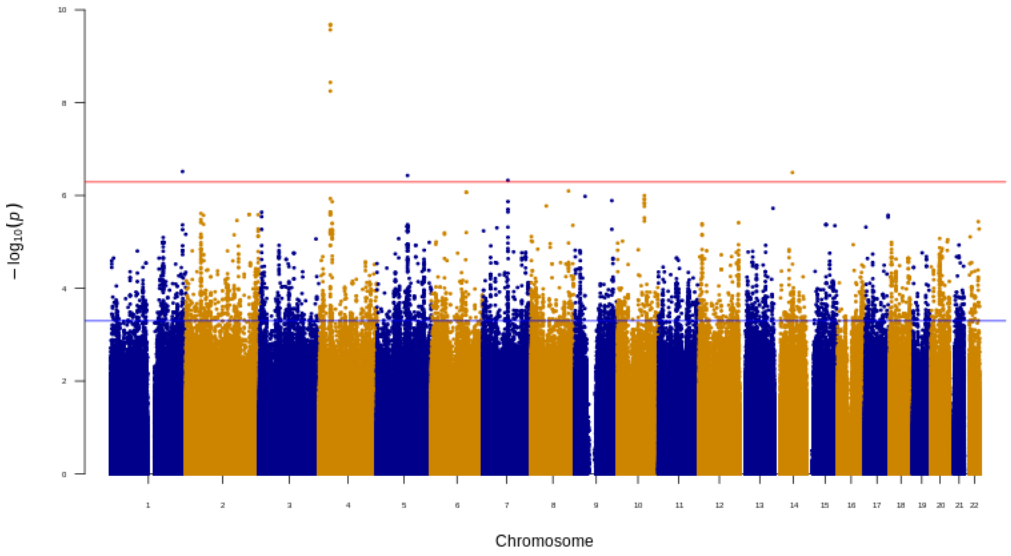


Figure 4.4: *Kenya Summary statistics Manhattan plot*

Our study is based on identification of causal variant snps. Despite many steps involved in this analysis, we still have a small number of SNPs and we discovered false positives due to the SNPs in LD. To challenge this issue, it becomes more relevant to perform Meta-Analysis.

4.1.2 Biological Functions of the identified variants

- From GWAS-Malawi summary statistics based-on Significant-SNPs we found only one gene *"SCAMP2"*, This gene is part of *SCAMP* family of proteins, it has as specifical function to secretory carrier of membrane proteins. He has a function of carrying the surface of cells in post-golgi recycling pathways.
- From GWAS-Kenya summary statistics based-on Significant-SNPs we found only 2 genes *"C4orf19"*,*"PAM"*.
    - *C4orf19* is the family of uncharacterized proteins, and is found in the tissue enhanced (pancreas, placenta), his protein can express membranous and cytoplasmic expression in most tissues.
    - *PAM means "Peptidylglycine Alpha-Amidating Monooxygenase")*, this gene is leading to code the protein, and is able to encode multiple functions of protein. It has proteolytically encoded protein, and is processed to provide a mature enzyme. So, these enzyme

has two domains having different catalytic activities such that :

- A peptidylglycine alpha-hydroxylating monooxygenase *(PHM)* domain.
- A peptidyl-alpha-hydroxyglycine alpha-amidating lyase *(PAL)* domain

So, the catalytic domains have a role of catalyzing the conversion neuroendocrine peptides to activate active alpha-amidated products sequentially. Multiple transcript variants arise from alternative splicing, proteolytic has processed at least one of them encodes an isoform. Diseases associated with PAM include Menkes Disease and Gestational Trophoblastic Neoplasm. The annotations of Gene Ontology (GO) have been associated with genes including *copper ion binding* and *L-ascorbic acid binding*.

We can summarize the tissue information and function for these two genes (*"PAM"* and *C4orf19*) into the Figures (4.5 and 4.6) :

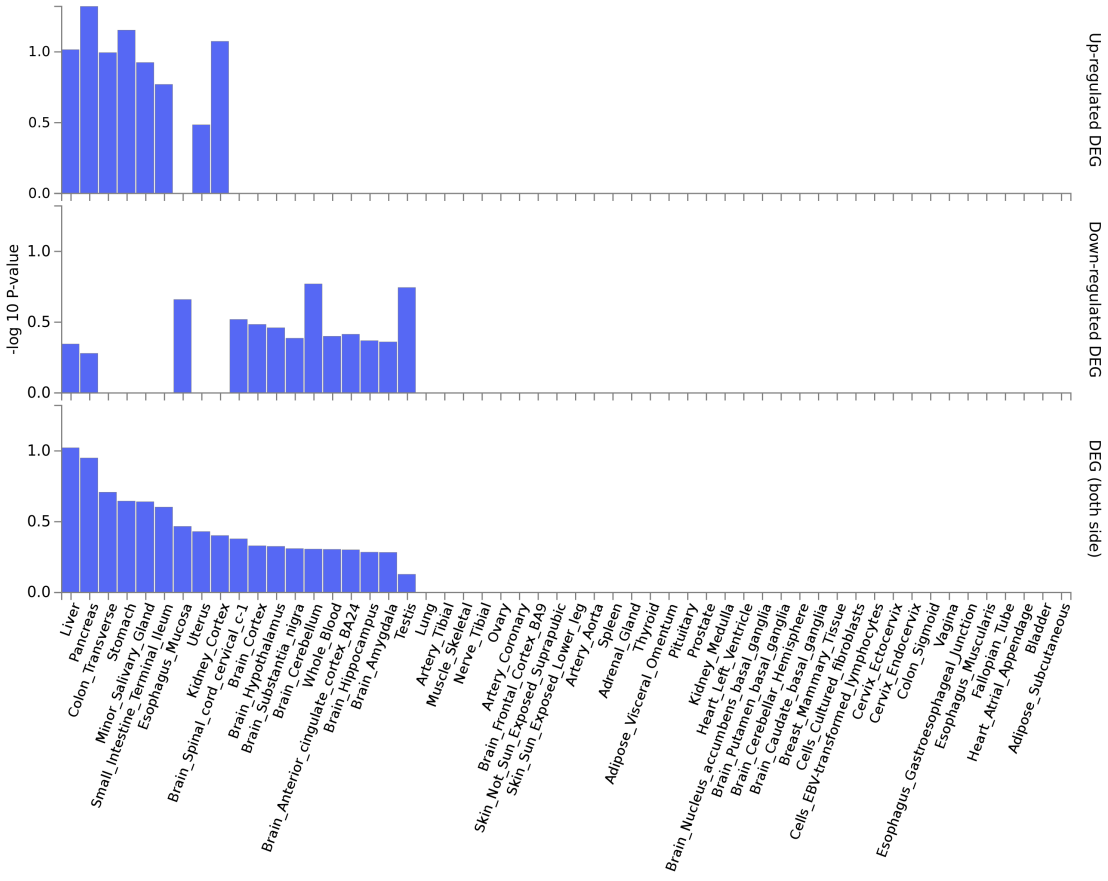


Figure 4.5: *Relevant information of "C4orf19" and "PAM" in cell type tissues*

,

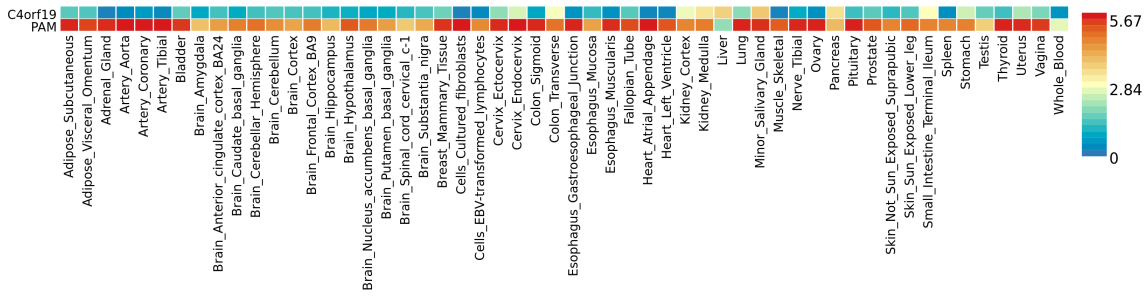


Figure 4.6: *Heatmap of expression for "C4orf19" and "PAM"*

4.2 Results from Cross Meta-Analysis

4.2.1 Metal

We performed Meta-Analysis 3 studies (Gambia , Malawi and Gambia) using Metal, then we found 8,402,059 variant SNPs. We sorted these SNPs in ascending order, then the lowest p-value is for **rs1479764** which is equal to 3.824e-07, shown on the Table4.5:

Table 4.5: *We got 3 SNPs variants have passed the threshold of 5e-07*

SNP	A1/A2	MAF	Z-score	P-value
rs1479764	t/c	0.8638	-5.078	3.824E-07
rs16993572	c/g	0.136	5.076	3.854E-07
rs78238616	t/c	0.136	5.076	3.854E-07

Before we visualize GWAS Meta-Analysis summary statistics, we have to convert that output to plink format,and then we can visualize it as:

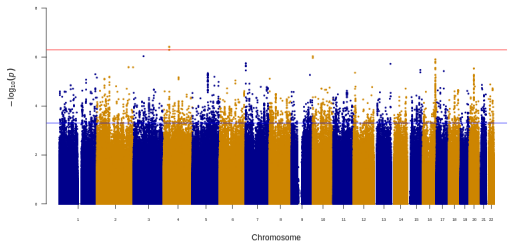


Figure 4.7: *Mahanatthan plot for Meta-Analysis find from Metal*

We can map these SNP variants at the gene-levels, then we got the Table 4.6:

Table 4.6: *We identify only one gene C4orf19 mapped by these SNP based on significant SNP threshold 5e-07*

CHR	SNP	Gene	Band	Gene_biotype	P	$OR \pm SE$
4	rs16993572	*C4orf19*	p14	protein_coding	3.854e-07	1.99439 ± 1.146
4	rs78238616	*C4orf19*	p14	protein_coding	3.854e-07	1.99439 ± 1.146
4	rs1479764	*C4orf19*	p14	protein_coding	3.854e-07	0.01245 ± 2.372

This result is temporary, we will validate it after using Metasoft.

We are uncertain which method has improved gene numbers with significant snp variants. Then we were 3 formats of output, which are *.mmap, *.meta.mcmc.out and *.meta, the summary information was containing in *.meta.mcmc.out.Then we got 5942350 SNP variants

The effects are checked based on posterior probability of association called *"M-value"*. We are going to filter the variant SNP with *M-value*>0.8, we got 35 variant SNPs that have passed this threshold, see in the Table Metasoft-tab.

We can visualize 3 top SNPs in forestplot, then we have the Figures 4.8.

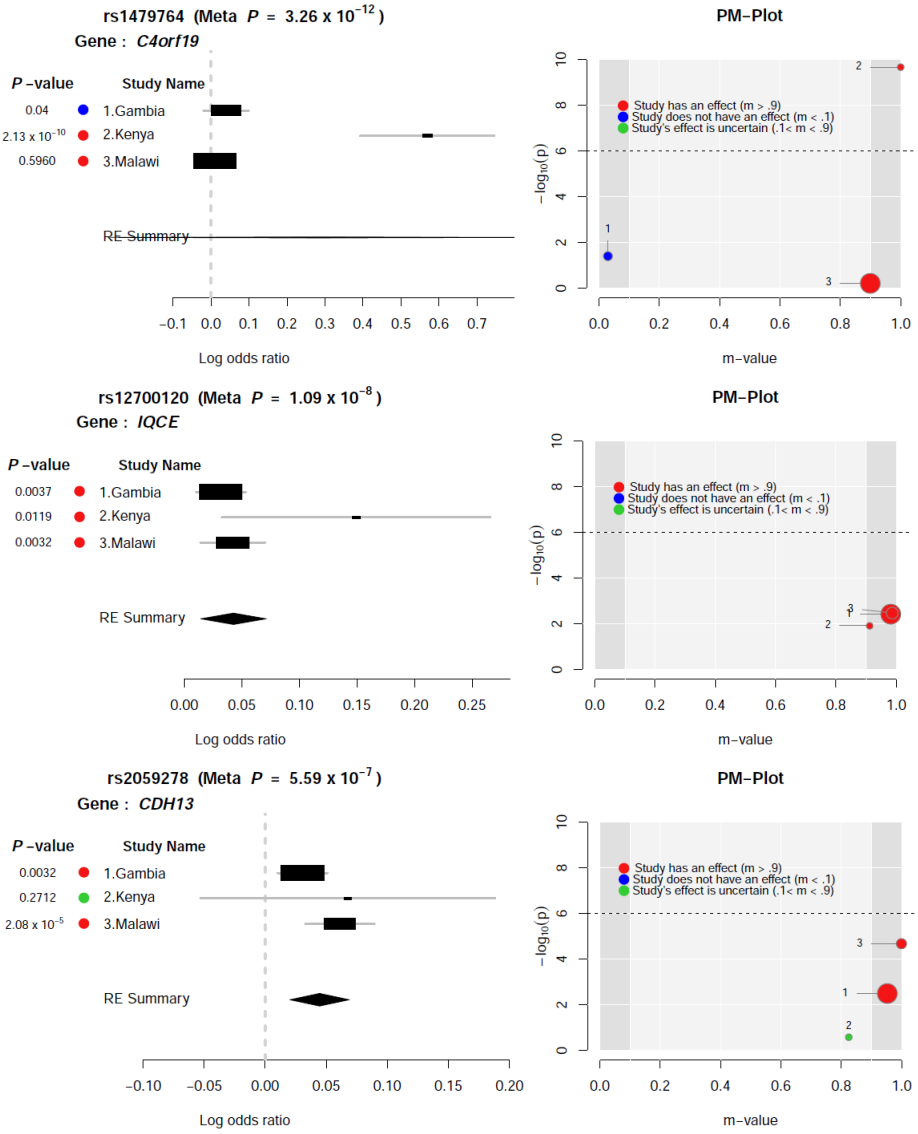


Figure 4.8: *The above figures represent the forestplot 3 top SNPs found in GWAS Meta-Analysis. We have:* ***rs1479764*** *mapped to the gene* ***C4orf19****,* ***rs12700120*** *mapped to the gene IQCE and rs2059278 mapped to the gene CDH13 describes*

Table 4.7: *We find out 20 SNP variants passed p-value threshold 5e-07 and m-value >0.8*

Snp-id	P Value-RE	Mvalue-GB	Mvalue-Kny	Mvalue-Mlw	Genes
rs1479764	3.27E-12	0.91	1	0.92	*C4orf19*
rs12700120	1.09061E-08	0.85	0.992	0.97	*IQCE*
rs2059278	5.56E-07	0.952	0.825	1	*CDH13*
rs1946283	4.27516E-07	0.99	0.872	0.998	*CDH13*
rs3214664	1.02375E-08	0.965	0.836	0.997	*IQCE*
rs28549904	1.05169E-07	0.97	0.828	0.997	*CDH13*
rs3922591	1.49E-07	0.992	0.897	0.978	NA
rs1364298	2.15E-07	0.971	0.868	0.998	*CDH13*
rs6563899	2.44383E-08	0.993	0.854	1	*CDH13*
rs1019569	4.06E-07	0.963	0.824	1	*CDH13*
rs7309982	5.2635E-07	0.998	0.827	0.983	*BORCS5*
rs10657376	8.39301E-08	0.977	0.818	0.996	*CDH13*
rs6665029	1.15328E-07	0.999	0.826	0.943	*LOC105378641*
rs146715151	1.3358E-07	0.987	0.825	0.993	NA
rs148009496	1.56514E-07	0.983	0.824	0.987	NA
rs113936393	1.76666E-07	0.995	0.83	0.966	*BORCS5*
rs691901	2.15472E-07	0.97	0.813	0.991	NA
rs4782771	3.19476E-07	0.981	0.853	0.998	*CDH13*
rs28419181	3.58101E-07	0.981	0.875	0.976	NA
rs7665590	3.9969E-07	0.983	0.83	0.982	*EIF4E*
rs6563897	5.50636E-07	0.972	0.836	0.998	*CDH13*
rs7191306	9.73333E-07	0.97	0.855	0.999	*CDH13*
rs8063612	0.000475776	0.957	0.858	1	*CDH13*
rs1019570	0.000512758	0.95	0.831	1	*CDH13*
rs1820255	0.000873478	0.924	0.815	0.999	*CDH13*
rs3837432	0.00107169	0.99	0.899	0.989	*BET1L*
rs1019568	0.0011107	0.923	0.825	0.999	*CDH13*
rs4721846	0.00201919	0.987	0.906	0.992	*IQCE*
rs12700112	0.00255554	0.978	0.913	0.99	*IQCE*
rs2833970	0.00294005	0.824	0.819	0.994	NA
rs12700120	0.00341397	0.981	0.912	0.987	*IQCE*
rs13242369	0.00423499	0.968	0.904	0.985	*IQCE*
rs12700094	0.00717821	0.96	0.908	0.975	*IQCE*

Comparing the results from Table 4.6 and Table 4.5, we have decided to use result from Table 4.6. Because the model incorporates the tool, it takes in account the random effect and it's based on a binary effects model.

4.3 Results from Gene and Pathway-based Association using PASCAL tool

Therefore, from our Meta-Analysis summary statistics, we conduct gene- and pathway-based association using PASCAL to generate the genescore of our analysis.
Our input summary statistics recorded 27 significant variants, then we performed the PASCAL analysis and we filtered significant genes (with p-value < 5e-08), we found 22 significant genes, see Table 4.8.

Table 4.8: *We find out 22 genes variants passed the threshold with p-value <5e-0.8*

chr	start	end	gene-id	gene-symbol	pvalue	Status
chr4	49189295	49381666	2492	C4orf19	7.07545134E-13	DAVIES-SUCCESS
chr16	116262692	116381921	2444	*CDH13*	5.39571388E-10	DAVIES-SUCCESS
chr6	116359893	116361107	728402	*TPI1P3*	5.39571388E-10	DAVIES-SUCCESS
chr5	74017030	74063042	84340	*GFM2*	5.79692505E-09	DAVIES-SUCCESS
chr20	29992647	30000641	245934	*DEFB121*	6.27940622E-09	DAVIES-SUCCESS
chr20	30009241	30016983	245935	*DEFB122*	6.27940622E-09	DAVIES-SUCCESS
chr20	30028410	30038060	245936	*DEFB123*	6.27940622E-09	DAVIES-SUCCESS
chr20	30053308	30060816	245937	*DEFB124*	6.27940622E-09	DAVIES-SUCCESS
chr20	30063104	30072708	28954	*REM1*	6.27940622E-09	DAVIES-SUCCESS
chr20	30073580	30075377	140875	*LINC00028*	6.27940622E-09	DAVIES-SUCCESS
chr5	74062815	74072737	10412	*NSA2*	6.32183805E-09	DAVIES-SUCCESS
chr5	129240522	129522327	337876	*CHSY3*	9.57275492E-09	DAVIES-SUCCESS
chr5	73935847	74017113	3074	*HEXB*	9.62195712E-09	DAVIES-SUCCESS
chr5	74073398	74191788	26049	*FAM169A*	1.36894532E-08	DAVIES-SUCCESS
chr20	29845466	29847435	245929	*DEFB115*	1.60585756E-08	DAVIES-SUCCESS
chr20	29891014	29896388	245930	*DEFB116*	1.60585756E-08	DAVIES-SUCCESS
chr3	114056946	114866127	26137	*ZBTB20*	1.81888481E-08	DAVIES-SUCCESS
chr15	64199234	64338521	23604	*DAPK2*	2.14797545E-08	DAVIES-SUCCESS
chr5	1050488	1112172	10723	*SLC12A7*	3.03389116E-08	DAVIES-SUCCESS
chr17	77085426	77512230	146713	*RBFOX3*	3.71772273E-08	DAVIES-SUCCESS
chr16	86365455	86379285	732275	*LINC00917*	3.74831993E-08	DAVIES-SUCCESS
chr2	58134785	58387055	7444	*VRK2*	4.68143136E-08	DAVIES-SUCCESS

Therefore we can analyse the function of these genes using FUMA. In doing so, we got 5 genes mapped and we have generated a heatmap for gene expression in Figure 4.9:

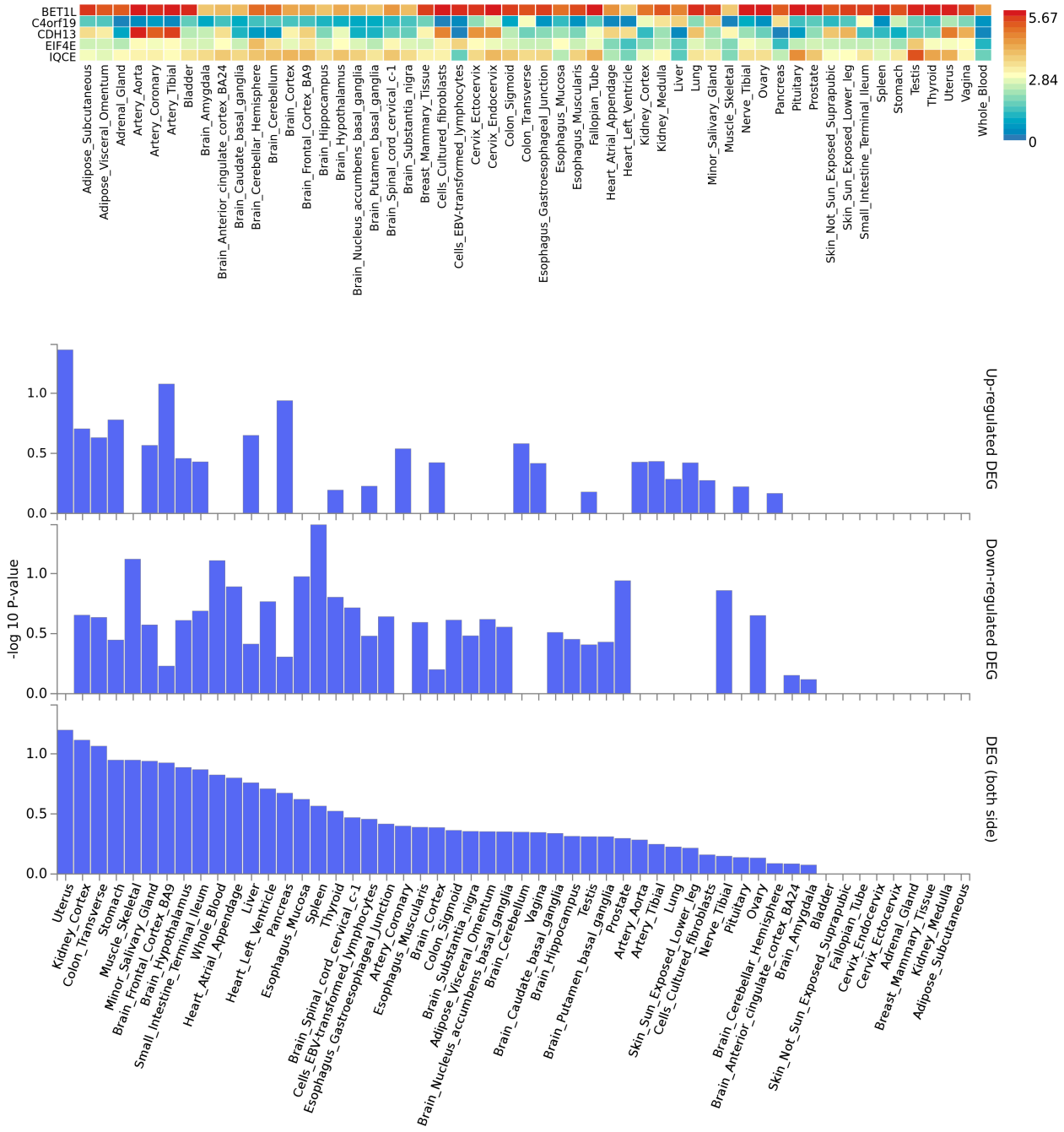


Figure 4.9: *Gene expression in different cell types and tissues type specificity of each gene.*

4.4 Result from FGWAS Analysis

FGWAS analysis is based on Metasoft output result, this is due to the large number of significant SNPs required by (the tool required at least 20 significant SNPs) the tool. We therefore got the result in the Table 4.9:

Table 4.9: *This table represents the likelihood of each annotation among 17 annotations selected (without tss_dist and Brain Microvascular)*

Annotation	ln(llk) [Likelihood]	AIC
nonsynonymous	0.708068	0.58
ens-coding-exons	4.21885E-14	2
CD3-DS17198	1.77636E-14	2
CD8-DS17203	1.55431E-14	2
hTH17-DS11039	1.17684E-14	2
HBMEC-DS13817	1.11022E-14	2
CD4-DS15947	1.06581E-14	2
DnaseTh1	9.76996E-15	2
k562.combined.R	7.9936E-15	2
k562.combined.E	0	2
CMK-DS12393	0	2
ens-utr3-exons	0	2
ens-utr5-exons	0	2
ens-noncoding-exons	0	2
CD56-DS16376	0	2

Also we have the maximum likelihood of each annotation parameters in the region-level and gene density, see the Table 4.10 and the Foresplot 4.10 .

Table 4.10: *Represent each annotation with the respect effect.*

Annotation	Marginal Effect [95 % CI]	log_2 (Effect) [95 % CI]
CD3-DS17198	27.0833 [-20,47.0833]	5.54872e-17 [2.51912e-21,0.00432037]
CD4-DS15947	27.0814 [-20,47.0814]	5.53697e-17 [2.51378e-21,0.00371303]
CD56-DS16376	30.625 [-20,50.625]	1.64746e-41 [7.47945e-46,0.00359436]
CD8-DS17203	27.0933 [-20,47.0933]	5.53206e-17 [2.51155e-21,0.00422478]
CMK-DS12393	5.82812 [-20,20]	9.09247e-20 [4.12798e-24,0.00356969]
DnaseTh1	27.0917 [-20,47.0917]	5.54028e-17 [2.51528e-21,0.00364998]
ens_coding_exons	27.8509 [20,47.8509]	5.54245e-17 [2.51627e-21,0.00653115]
ens-noncoding-exons	5.82812 [-20,20]	9.09247e-20 [4.12798e-24,0.00393848]
ens-utr3-exons	27.0975 [-20,47.0975]	5.52286e-17 [2.50738e-21,0.00465154]
ens-utr5-exons	5.82812 [-20,20]	9.09247e-20 [4.12798e-24,0.00380423]
HBMEC-DS13817	27.0917 [-20,47.0917]	5.54028e-17 [2.51528e-21,0.00405603]
hTH17-DS11039	28.7082 [-20,48.7082]	5.37557e-17 [2.4405e-21,0.00367723]
k562.combined.E	5.82812 [-20,20]	9.09247e-20 [4.12798e-24,0.00366803]
k562.combined.R	27.026 [-20,47.026]	5.46428e-17 [2.48078e-21,0.00397853]
Nonsynonymous	35.5014 [-20,55.5014]	0.00126895 [4.53979e-05,0.00691602]

Then we can represent it on the forestplot 4.10, we have:

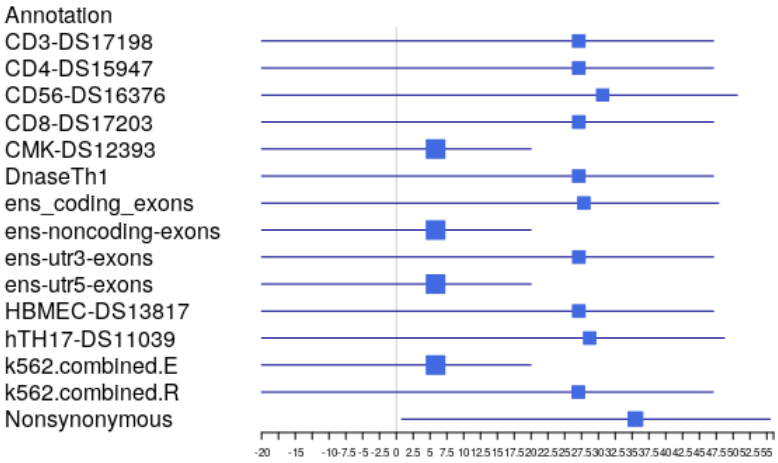


Figure 4.10: *Forestplot of each annotation*

STEP2

Among these annotations, we got nonsynonymous having the highest likelihood , then we can combine this latter to each annotation, therefore we observed the improvement annotation's likelihood.

Table 4.11: *Table represented the likelihood of each annotation among 17 annotations selected (without tss_dist and Brain Microvascular)*

Combine Annotations with NonSynonymous	ln(llk) [Likelihood]	AIC
ens-coding-exons + Nonsynonymous	0.797299	2.4054
CD3-DS17198 + Nonsynonymous	0	4
CD8-DS17203+Nonsynonymous	0	4
hTH17-DS11039+Nonsynonymous	0	4
HBMEC-DS13817 + Nonsynonymous	0	4
CD4-DS15947+Nonsynonymous	0	4
DnaseTh1 + Nonsynonymous	0	4
k562.combined.R+Nonsynonymous	0	4
k562.combined.E + nonsynonymous	0.784189	2.43162
CMK-DS12393 + nonsynonymous	0.703871	2.59226
ens-utr3-exons + nonsynonymous	0.792362	2.41528
ens-utr5-exons + nonsynonymous	0.711091	2.57782
ens-noncoding-exons + nonsynonymous	0	2
CD56-DS16376 + nonsynonymous	0	4

After adding 6 parameters of annotations progressively, we kept the best likelihood from combination of "K562.combined.E+ens-coding-exons+nonsynonymous+ens-utr3-exons+ens-utr5-exons+CMK-DS12393" called "C" which is given by llk = 0.924007. Based on this likelihood, we can switch to using the cross-validation likelihood, and find the penalty with the best cross- validation likelihood

Table 4.12: *In this table the values of cross validation likelihood penalized by 0.05,0.1,0.15,...,0.65 using C as annotation.*

Cross validation Likelihood	Ridge-parameter
-1.77369e-08	0.05
-3.29358e-08	0.1
1.06668e-08	0.15
-5.58249e-09	0.20
-2.54227e-09	0.25
-1.11384e-06	0.30
-2.52696e-09	0.35
2.5402e-09	0.40
-1.37317e-08	0.45
-6.60905e-09	0.50
-9.57009e-09	0.55
-6.61249e-09	0.60
-2.54214e-09	0.65

Therefore, p=0.40 has the maximum cross validation likelihood (xv = 2.5402e-09). Then, we can assess the posterior probability of association (PPA) for each variant belonging to these genomic regions of these 6 parameters, also we evaluate the p-value for each of them, see Table 4.13.

Table 4.13: *We have 29 significant SNPs with PPA > 0.9.*

SNP	CHR	Position	P-Value	PPA	Gene	Band	*Gene_biotype*
rs77403150	4	37468270	1.02E-15	0.96	*C4orf19*	p14	*protein_coding*
rs2918101	10	132984326	1.08E-14	0.97514	*TCERG1L*	q26.3	*protein_coding*
rs56392308	9	133255671	1.00E-13	0.9256	*ABO*	q34.2	*processed_transcript*
rs116478774	13	72979352	1.00E-13	0.987	NA	NA	NA
rs145143585	7	153331362	1.11E-13	0.9602	NA	NA	NA
rs4419390	3	175581823	1.12E-13	0.981	NA	NA	NA
rs7622028	3	13436336	1.13E-13	0.945	*NUP210*	p25.1	*protein_coding*
rs1863195	7	81480501	1.16E-13	0.9124	NA	NA	NA
rs111278730	1	8288486	1.20E-13	0.9125	NA	NA	NA
rs7068246	10	28464753	1.45E-13	0.9112	*MPP7*	p12.1	*protein_coding*
rs796073521	9	133261369	1.52E-13	0.9045	NA	NA	NA
rs37023	16	56968107	1.71E-13	0.9081	*HERPUD1*	q13	*protein_coding*
rs9839112	3	158981898	3.25E-13	0.9311	*IQCJ*	q25.32	*protein_coding*
rs1992521	3	105511196	4.67E-13	0.9012	*CBLB*	q13.11	*protein_coding*
rs9497849	6	148043646	6.57E-13	0.913	*SAMD5*	q24.3	*protein_coding*
rs147907945	18	20680489	2.11E-12	0.9411	NA	NA	NA
rs10920225	1	201615330	7.20E-12	0.9512	*NAV1*	q32.1	*protein_coding*
rs34012192	11	5234385	1.14E-11	0.975	*HBD*	p15.4	*protein_coding*
rs11047725	12	25104269	1.26E-11	0.9321	NA	NA	NA
rs2101086	18	8976498	4.10E-11	0.95862	NA	NA	NA
rs16852227	1	203727385	5.00E-11	0.922	*ATP2B4*	q32.1	*protein_coding*
rs75360548	1	203711045	8.00E-11	0.9103	*ATP2B4*	q32.1	*protein_coding*
rs56101813	22	47139583	1.40E-10	0.9188	NA	NA	NA
rs2728615	12	20159822	2.51E-10	0.9223	NA	NA	NA
rs33910209	11	5225648	4.00E-10	0.9166	*HBB*	p15.4	*protein_coding*
rs11158673	14	67965889	7.89E-10	0.92203	*TMEM229B*	q24.1	*protein_coding*
rs76621318	20	45742182	1.00E-09	0.925	*EYA2*	q13.12	*protein_coding*
rs7319528	13	74924732	1.48E-09	0.956	NA	NA	NA
rs11639047	15	67237369	1.26E-08	0.9781	NA	NA	NA

Therefore, we got 16 genes *(C4orf19, ABO, ATP2B4, CBLB, EYA2, HBB, HBD, HERPUD1, IQCJ, MPP7, NAV1, NUP210, SAMD5, TCERG1L, TMEM229B)*.

4.5 Result from Pathway Analysis, Enrichment Analysis and Functional Mapping and Annotation using FUMA

Genemania is one of the most popular web based application tools that has been applied in predicting pathway gene sets of the genes identified from a given study. The goal of this section is to put the results into biological context, we then sought to understand the biological significance of genes identified in the FGWAS result , and their relationship with severe malaria. For this analysis, we used the genes obtained from FGWAS result as the query genes and applied Genemania in predicting the genes they are related to in terms of co-expression, physical interaction, genetic interaction and shared pathways.

Generally, Genemania by default, is used to predict twenty other additional genes that are functionally similar to the query list of genes. However, for this analysis, we adjusted the number of genes, we obtained the prediction network with the best FDR (False Discovery Rate) values, we used genes identified from fGWAS as query list of genes Figure 4.11 shows the resulting network that was generated by the best FDR values and genes together with their top 7 scores see the Table 4.14.

Table 4.14: *Top 7 genes having high score.*

Index	Symbol Gene	Score
1	*ATP2B4*	0.966553
2	*C4orf19*	0.877459
3	*ABO*	0.864614
4	*HBB*	0.862778
5	*HBD*	0.859162
6	*TCERG1L*	0.7883008
7	*EYA2*	0.7501903

ATP2B4 was identified as the gene that had the strongest association with a score of 0.9665. The top biological functions which were associated with this network were blood microparticle at a P-value of 1.54e-08,gas transport at a P-value 2.31e-08, oxidoreductase activity, acting on peroxide

as acceptor at a P-value of 4.05e-08, peroxidase activity at a P-value of 5.2e-08, antioxidant activity at a P-value of 6.31e-07, hydrogen peroxide catabolic process at a P-value of 2.12e-06, endocytic vesicle lumen at a P-value of 1.25e-05 and protein heterooligomerization at P-value of 7.5e-04.

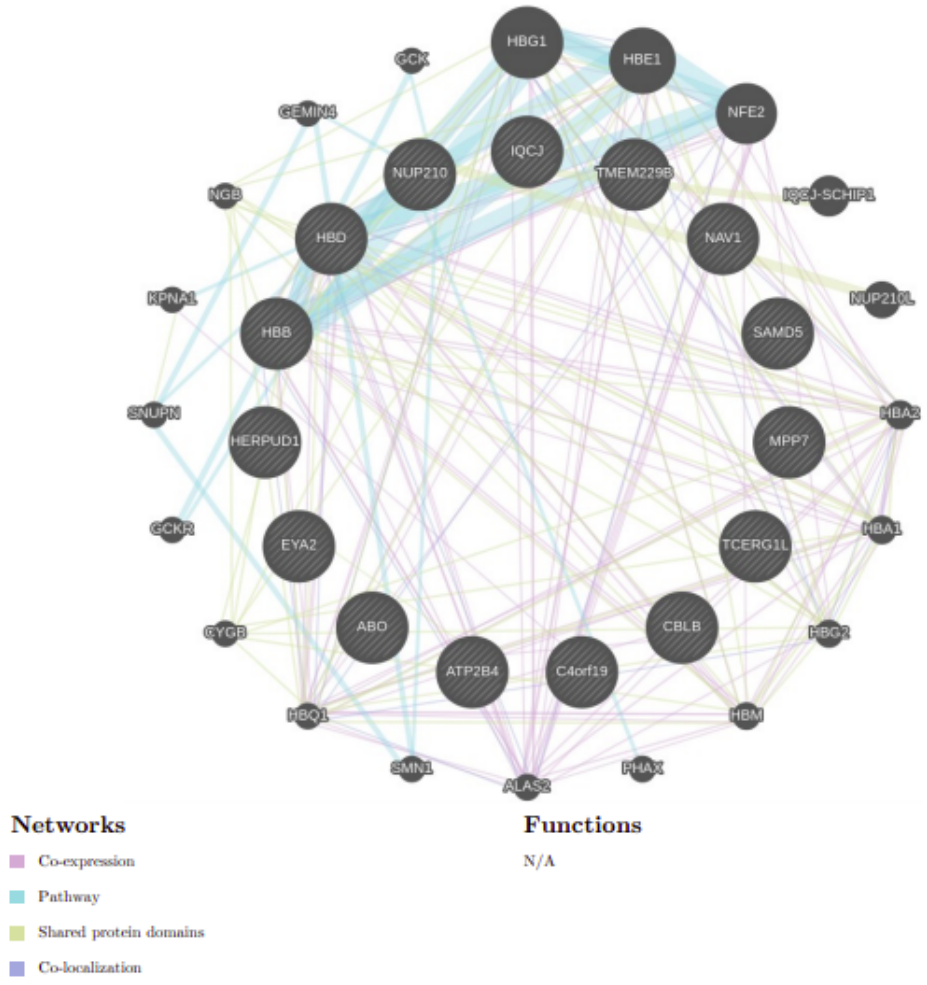


Figure 4.11: *Network analysis of genes from based on fGWAS result*

The platform is web-based using biological databases from several resources to facilitate GWAS results analysis and aggregate relevant functional annotation information, prioritize genes and their interaction. In this tool, they have integrated expression quantitative trait loci (eQTL) and chromatin interaction mappings information ,positional, provides gene-based, pathway and tissue enrichment results. from FUMA , the results contain relevant information which can characterize the functional experiment aimed to find the causal relation . We use genes resulted from FGWAS and perform deep functional analysis (to annotates genes in biological context, biological sources

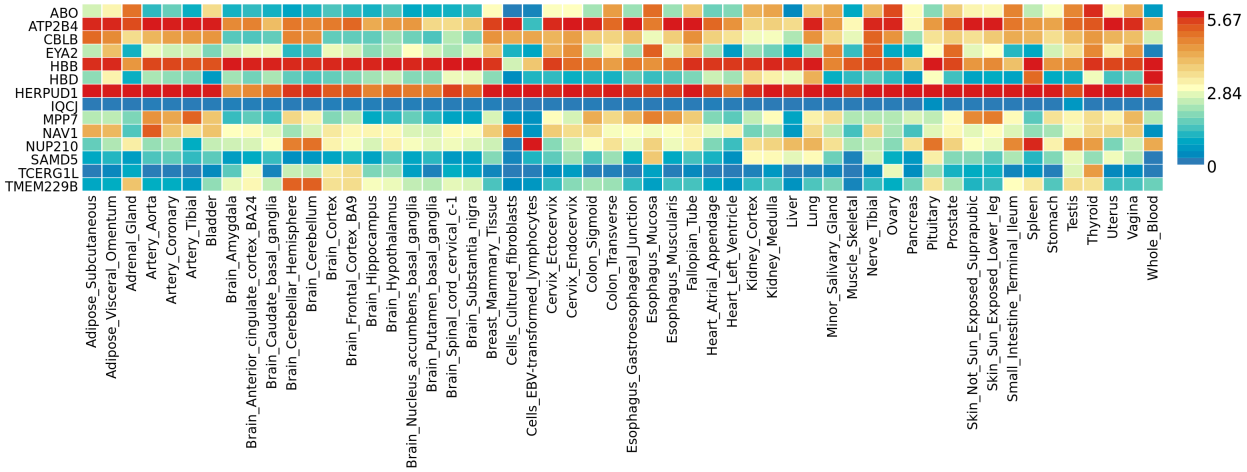


Figure 4.12: *Gene expression heatmap of our 14 genes.*

constitute an insight of association diseases from each gene find out, different specifical tissue expression pattern established based on GTEx v6 RNA-seq data from each gene we can visualize it on interactive heatmap Figure 4.12, over-representation in sets of differentially expressed genes Figure 4.13, GWAS catalogue Figure 4.16)

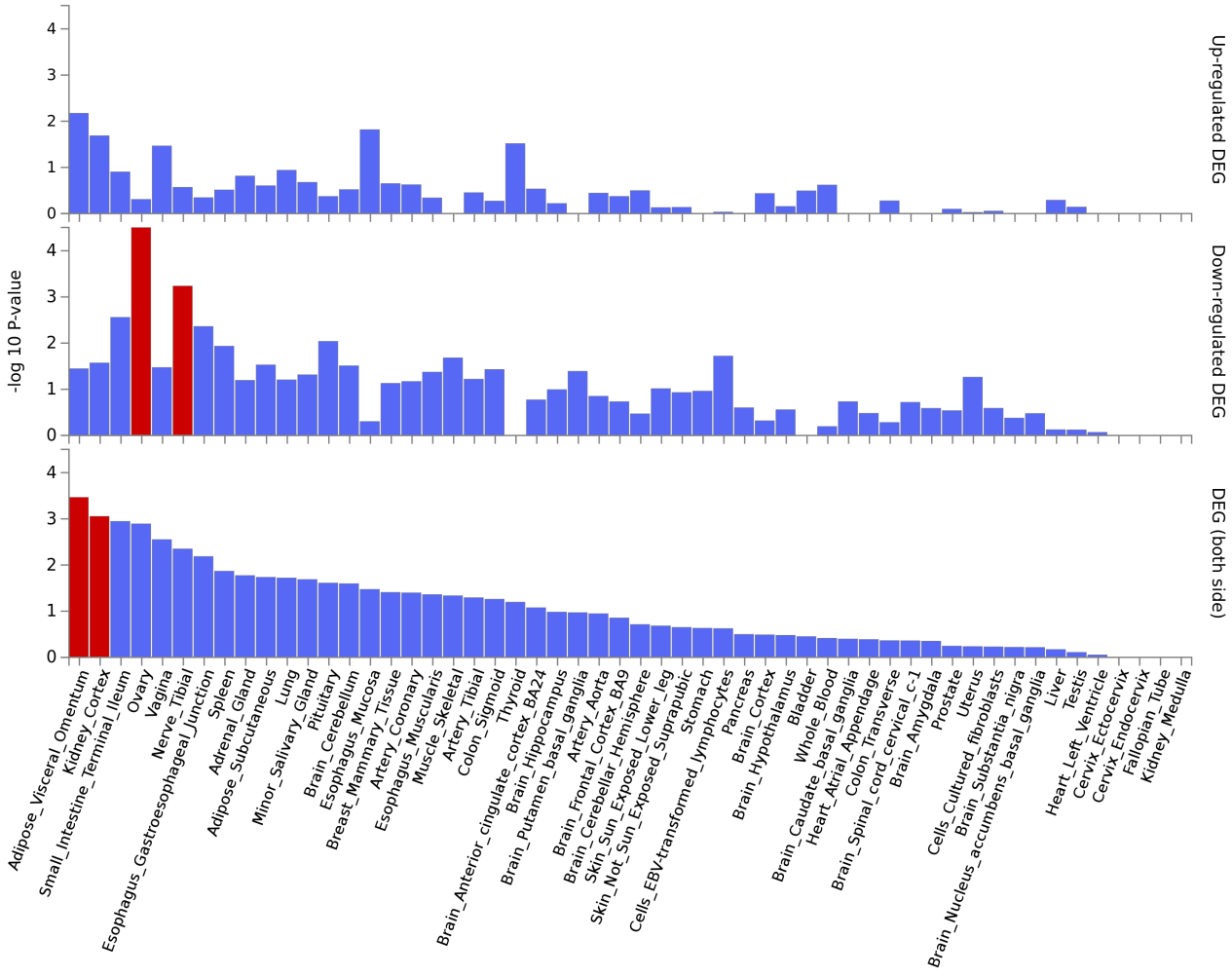


Figure 4.13: *Differentially expressed genes of tissues specificity, Pathway scores obtained from pascal analysis. Significant Pathways at Bonferroni corrected P-value ≤ 0.05 are coloured in red*

Therefore, we have enrichment of input genes in Gene Sets in Figures 4.15 and 4.16 below :

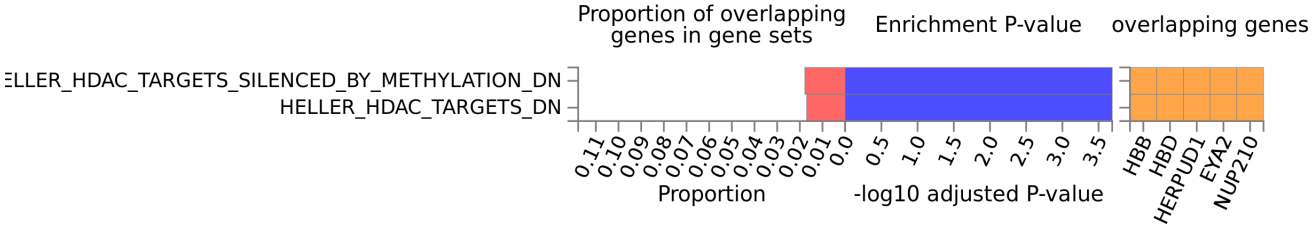


Figure 4.14: *Curate gene set*

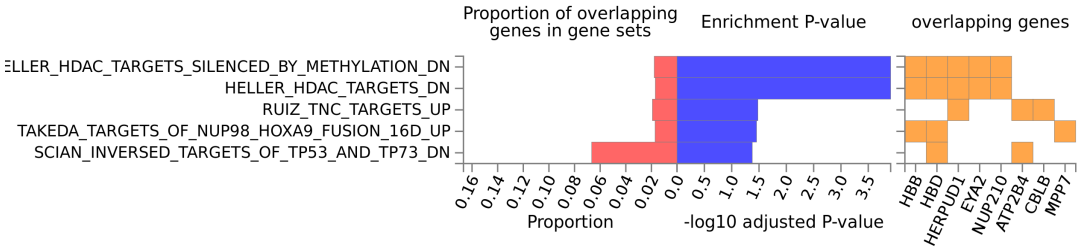


Figure 4.15: *Chemical and Genetic perturbation of our gene set*

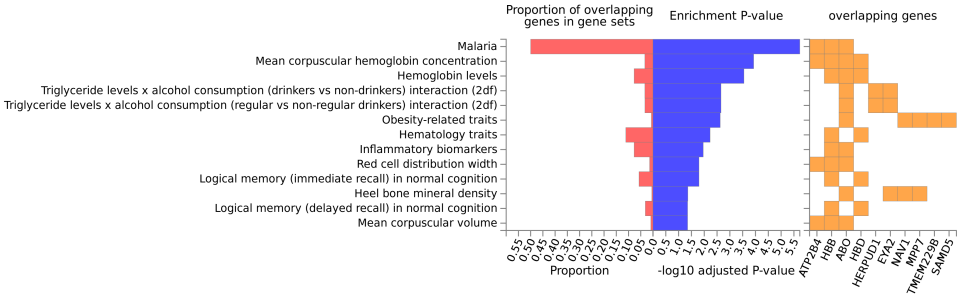


Figure 4.16: *GWAS Catalogue based on our genes*

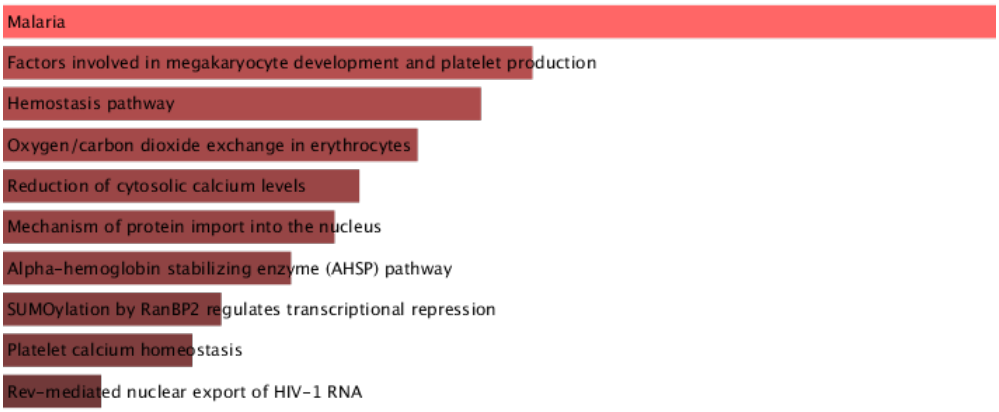
Enrichr is one of popular enrichment analysis tools, it contains a huge several ressources geneset database for genomic analysis. Currently, it contains 180184 annotated gene sets identified by 102 geneset libraries. This tool is a comprehensive engine resource with accuracy genesets and accumulates biological knowledge for deep analysis of biological discoveries. We have used our genes set result of FGWAS as input to perform Enrichment Analysis based on Enrichr to ascertain pathway relevance to the disease for network, then we got the result from Table 4.15 and 4.17:

Table 4.15: *Summary statistics of our genes and phenotype related.*

Index	Name	P-value Adjusted	p-value	Odds Ratio	Combined score
1	Malaria	6.816E-09	1.184E-05	666.67	12536.03
2	Hemoglobin levels	2.014E-06	0.001749	117.65	1542.99
3	Inflammatory biomarkers	0.0001559	0.06772	106.67	935.04
4	Hematology traits	0.0001823	0.06333	98.77	850.36
5	Insulin-related traits	0.003745	0.542	266.67	1489.98
6	Clinical laboratory measurements	0.005239	0.65	190.48	1000.31
7	Tumor biomarkers	0.005985	0.6116	166.67	853.08
7	Beta thalassemia/hemoglobin E disease	0.005985	0.5776	166.67	853.08
9	End-stage coagulation	0.006731	0.6154	148.15	740.89
10	Coagulation factor levels	0.007476	0.6493	133.33	652.8

The results show that according to the disease database incorporated into Enrichr GWAS catalog 2019, this pathway was highly associated with malaria with P-value of 6.816e-09 (Figures 4.17 and 4.18).

Figure 4.17: *BarGraph result from fGWAS results based on GWAS Catalog 2019*



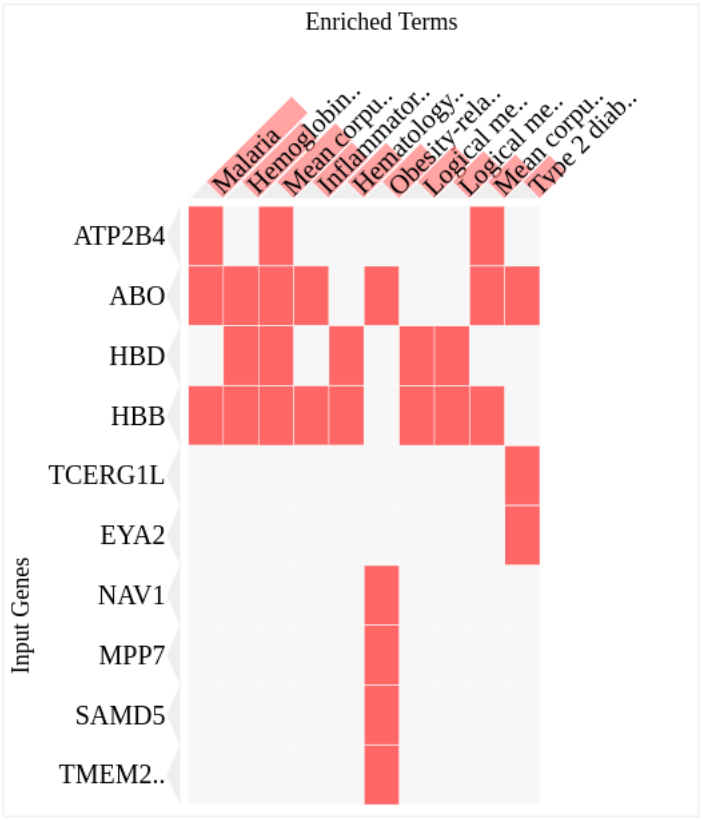


Figure 4.18: *Genes ordering in term of p-values related to trait*

We got 3 genes *(ATP2B4, ABO, HBB)* highly related to severe Malaria in terms of P-value.

4.6 Results from Rare Variant Association Analysis using SKAT

We got only 1276043 variants in our Setid file, this file contains 3806 genes based associated SNP using DBSNP (Data base).

We can run SKAT and SKAT-O also combine SKAT test and Burden test, then we filtered p-values different to "NA", we got 432 genes associated. Therefore, we selected the genes in terms of standard P-value threshold (< 0.05), then we have the Table 4.16:

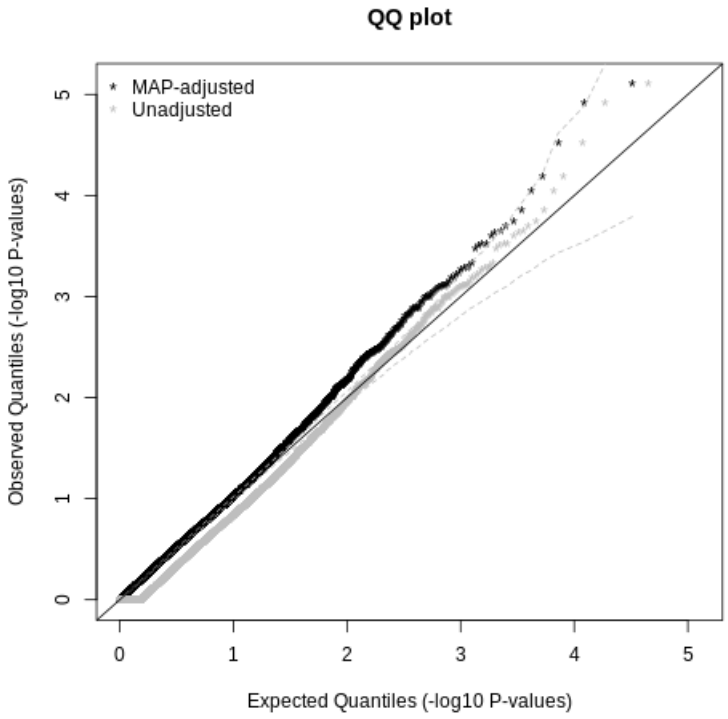
Table 4.16: *We have filtered SKAT output with p-values less then 0.05*

SetID	P.value	N.Marker.All	N.Marker.Test	MAC	m	Method.bin	MAP
LOC105376397	2.07493380545057E-06	22	1	1046	946	UA	-1
NDUFA4L2	0.000245621956271	3	1	3206	2595	UA	-1
MYO1H	0.002822185254877	16	1	5347	4065	UA	-1
LOC101927595	0.004415969202187	9	1	669	650	UA	-1
LOC100287010	0.006606340749515	7	1	843	807	UA	-1
LINC00656	0.007843379768197	7	1	1089	1029	UA	-1
C11orf53	0.008141825109537	24	1	775	751	UA	-1
PLPP7	0.008491857716472	11	2	1328	644	UA	-1
SAP30	0.009462094563975	1	1	626	609	UA	-1
MIR8485	0.010740681676086	4	1	901	867	UA	-1
LINC00184	0.012539988297798	13	1	1232	1167	UA	-1
LOC105373615	0.013139764642116	11	1	921	874	UA	-1
LOC105379343	0.014070335182194	2	2	1297	637	UA	-1
LOC105371688	0.016011662656197	1	1	2774	2398	UA	-1
LOC107984015	0.016736019748413	9	1	646	631	UA	-1
LOC101927752	0.018101212719709	1	1	4312	3441	UA	-1
KIRREL1-IT1	0.018972678934857	5	1	760	722	UA	-1
LOC107986630	0.020046134545665	3	1	810	780	UA	-1
LOC107985231	0.025307000869034	4	1	615	595	UA	-1
PAPOLB	0.026047472836158	2	1	4292	3215	UA	-1
RNF183	0.026817671945456	6	1	630	621	UA	-1
LOC105373937	0.027697882022767	14	2	2649	2184	UA	-1
FAM217B	0.028326633738291	14	2	8871	4738	UA	-1

GMPPA	0.030782567803414	6	1	2408	1980	UA	-1
LOC105376498	0.032353505849627	5	1	1345	1275	UA	-1
LOC105377364	0.037848943732829	24	1	649	627	UA	-1
CA9	0.039873242261575	8	1	4061	3293	UA	-1
TARP	0.040455590663857	20	1	876	836	UA	-1
LOC100506022	0.044020501929566	3	1	1825	1669	UA	-1
LOC105374475	0.045093453109974	10	1	640	620	UA	-1
LOC101929088	0.048646738374177	7	1	1059	987	UA	-1

The last column contains the minimum achievement p-values (MAP), if MAC is greater 20 than MAP is -1, this is due to large sample size M. We have recorded an effective number of tests greater than 30. Therefore, we can represent MAPs and P-values on QQplot (Figure 4.19), then we have:

Figure 4.19: *QQplot*



Our SKAT Analysis for rare variants came out with 32 genes. Therefore, we perform enrichment analysis using cell types and disease related to these genes.

4.6.1 Pathway Analysis and Enrichment Analysis

From **GENEMANIA**, we have performed the functional analysis of thesis pathways, then we have different functions associated with the false discovery rate (FDR), represented on Table 4.17.

Table 4.17: *Different functions and their FDR.*

No	Function	FDR (False Discover rate)
1	cellular response to oxygen levels	0.000007225018153732825
2	cellular response to decreased oxygen levels	0.000007225018153732825
3	cellular response to hypoxia	0.000007225018153732825
4	response to decreased oxygen levels	0.00006882197328502966
5	monosaccharide metabolic process	0.00006882197328502966
6	response to hypoxia	0.00006882197328502966
7	response to oxygen levels	0.00006882197328502966
8	glucose metabolic process	0.00034755787352911544
9	hexose metabolic process	0.0006375192027793193
10	glycolysis	0.0016277440094377258
11	glucose catabolic process	0.0031924260579812348
12	hexose catabolic process	0.006991228932961755

13	monosaccharide catabolic process	0.007202372463314842
14	regulation of transcription from RNA polymerase II promoter in response to hypoxia	0.014492573531483659
15	single-organism carbohydrate catabolic process	0.01834252995804484
16	carbohydrate catabolic process	0.020979000436203325
17	regulation of transcription from RNA polymerase II promoter in response to stress	0.02579799978496425
18	carbohydrate biosynthetic process	0.03201829501413785
19	regulation of DNA-templated transcription in response to stress	0.03428429338358941
20	gluconeogenesis	0.0593303951279183
21	hexose biosynthetic process	0.06358086247382655
22	monosaccharide biosynthetic process	0.07182329568828232

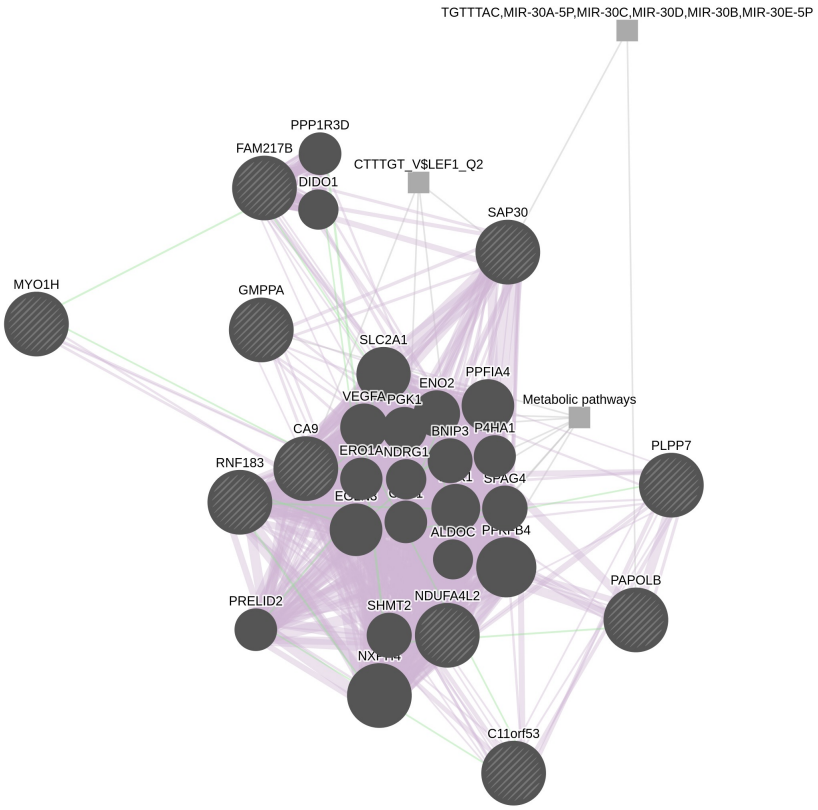
Using KEGG2019, we find 2 pathways significant based on P-value < 0.05, represented in the Table 4.18,The top biological functions which were associated with this network were Nitrogen metabolism at a P-value of 0.02604 and Fructose and mannose metabolism at a P-value 0.04994.

Table 4.18: *Significant Pathways for these genes.*

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Nitrogen metabolism	0.02604	1.000	37.95	138.45
2	Fructose and mannose metabolism	0.04994	1.000	19.55	58.59

Then we have a network of gene interactions represented in the Figure 4.20:

Figure 4.20: *Genes mapped by rare variants and their pathways.*



SAP30 was identified as the gene that had the strongest association with a score of 0.6414662253, followed by C11orf53 with score 0.62664044504, represented in the Figure 4.19.

Table 4.19: *Top score of 6 genes strongly associated.*

No	Symbol	Score
1	*SAP30*	0.6414662253602376
2	*C11orf53*	0.6266404450466443
3	*PAPOLB*	0.6133770303956222
4	*MYO1H*	0.6069617410969311
5	*RNF183*	0.5918503934112418
6	PLPP7	0.582797316665604

We can interpret the biological function of SAP30, this gene has known as **Sin3A-associated protein, 30kDa**, also known as **SAP30**, it participates in recrutement of Sin3-histone deacetylase complex (HDAC) functional to the part of N-CoR corepressor complexes. Able to make transcription repression through N-CoR. It has major function in eukaryotic gene expression regulation and its proteins have function to nucleolar localization signal.Also, capable to reaching *SIN3A* to the nucleolus [101]. His acide in the central region facilitates the interactions in nucleosomes to histone .

In malaria context, SAP30 participates in protective vaccination and DNA methylation of gene promoters localized in the liver of Balb which is modified by blood-stage malaria [102].

Based on **FUMA (Functional Mapping and Annotation)** We can see some specific tissue types related to these genes in Figure 4.21 and the Gene expression heatmap in Figure 4.22.

Figure 4.21: *The tissue specificity.*

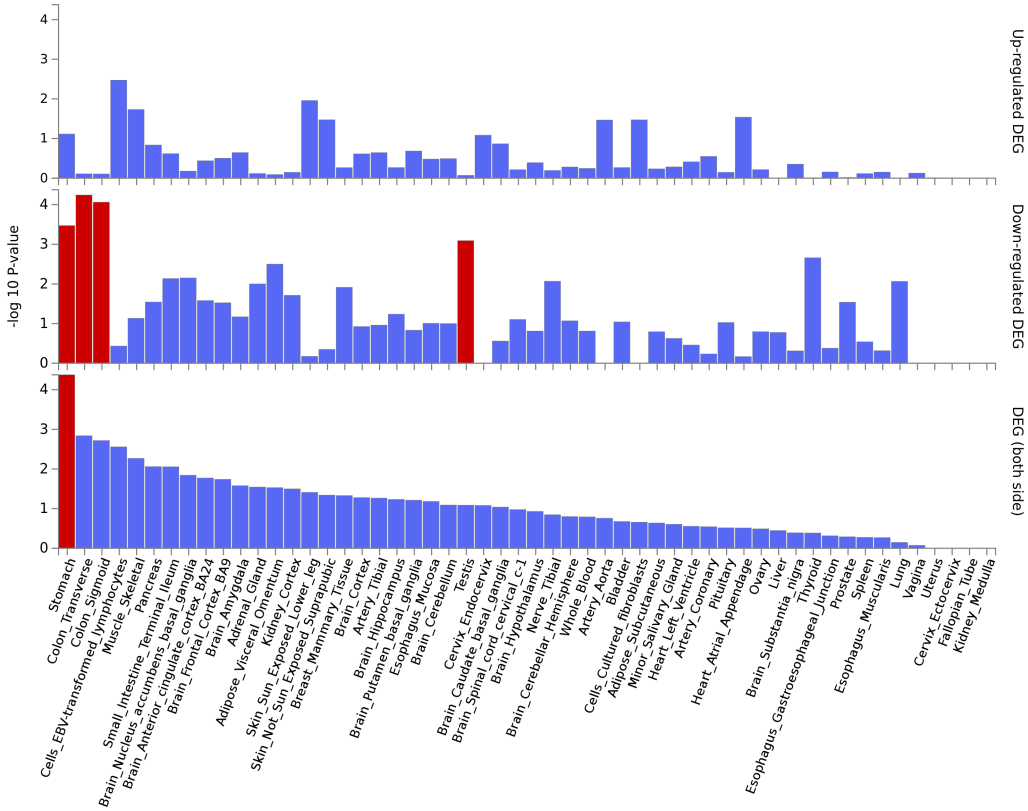
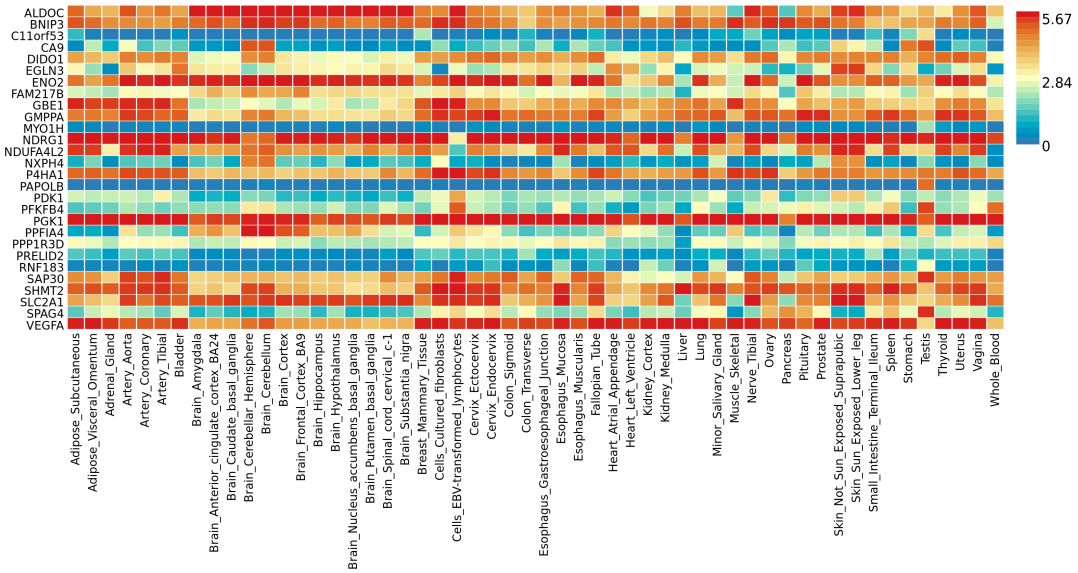


Figure 4.22: *Gene expression heatmap.*



From Figure 4.21, we have noticed that the stomach is highly localised than other tissues.

4.7 PRS Result

4.7.1 High-Resolution Polygenic Risk Scoring

We performed high-resolution polygenic risk scoring, with PRSice, we calculated PRS at a large number of evenly spaced P-value thresholds, between a minimum and maximum bound. For our analysis, we got:

Gambia Populations

We find significant evidence that genetic variants predict a severe condition and resistance of malaria. We have recorded 4921 under the approach of only testing PRS at several broad P-value thresholds find the most predictive threshold is at P_T = 0.00165005, with the P-value 2.93964e-06, in this region we recorded 2358 SNPs, with r^2 = 0.00443458. The lowest P-value threshold is 0.4, with p-value 2.93964e-06, then we can visualize in the Figures 4.23 and 4.24:

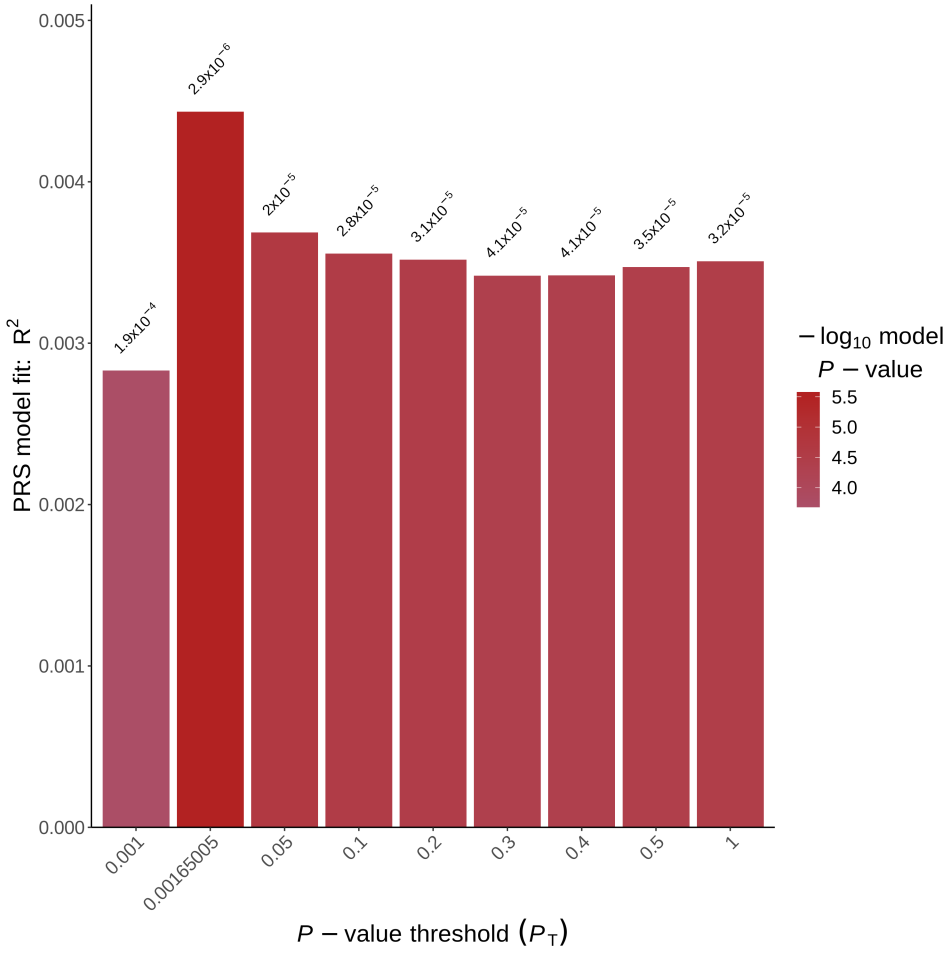


Figure 4.23: *Summary of Barplot, recorded High resolution of Risk predictive of Severe Malaria, the best PRS model fit has R^2=0.00443458*

We can observe it as:

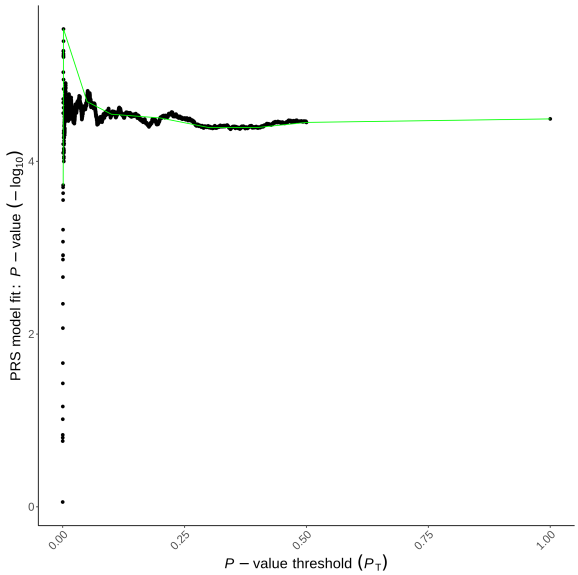


Figure 4.24: *This plot summarizes High-resolution PRS for 'Host having resistance to Malaria' predicting Severe status. The high-resolution best-fit PRS is 0.00443458 at P_T = 0.00165005.*

we can define the interval of 100 quantiles. This makes a link between odd ratio for score on the phenotype and strata for polygenic score.

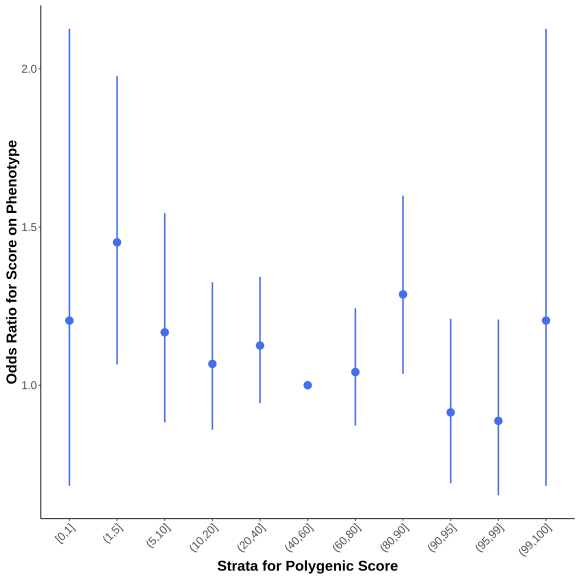


Figure 4.25: *This PRS Strata plot provide the information of each polygenic score interval and the odds ratio for score on Severe Malaria.*

Kenya Populations

At the same conditions we tried to evaluate the PRS for the individuals susceptible and resistant to severe malaria in the Kenya population. We have recorded 3142 (1637 controls and 1505 cases) based on approach, testing PRS of multiple P-value thresholds and find the best predictive threshold was at $P_T = 1$, with the P-value $8.16271e-158$, in this region we recorded 2358 SNPs, with $r^2 = 0.784666$. The lowest P-value threshold is 0.001, with p-value $1.44e-44$. then we can visualize in the Figures 4.26 and 4.27:

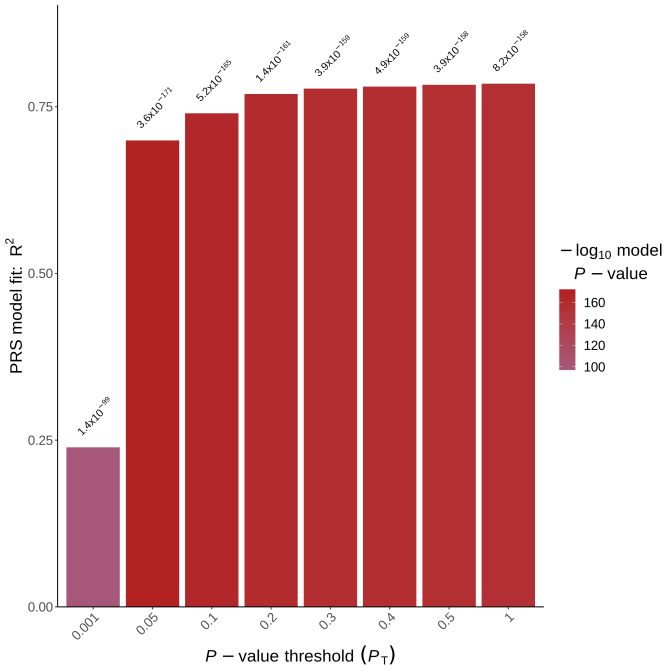


Figure 4.26: *Summary of Barplot, recorded High resolution of Risk predictive of Severe Malaria, the best PRS model fit has r^2=0.784666*

We can plot high resolution PRS score

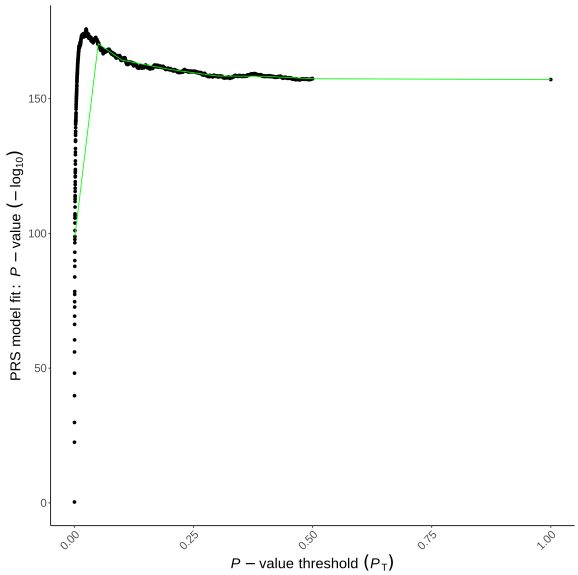


Figure 4.27: *This plot summarizes High-resolution PRS for 'Host having resistance to Malaria' predicting Severe status. The high-resolution best-fit PRS is 8.16271e-158 at P_T = 1*

we can define the interval of 100 quantiles. This make link between odd ratio for score on the phenotype and strata for polygenic score, we can visualize it in Figure 4.28:

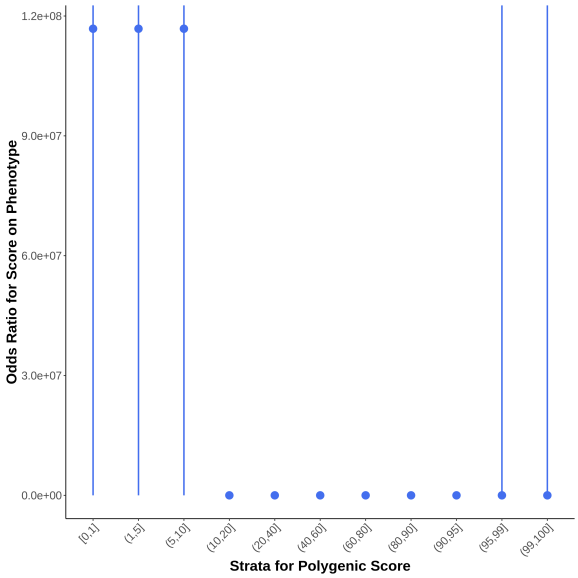


Figure 4.28: *This PRS Strata plot provide the information of each polygenic score interval and the odds ratio for score on Severe Malaria*

Malawi Populations

At the same conditions we tried to evaluate the PRS for the individuals susceptible and resistant to severe malaria in the Malawi population. We have recorded 2516 (1322 controls 1194 cases) based on approach, testing PRS of multiple P-value thresholds and find the best predictive threshold was at P_T = 1, with the P-value $1.51515e-55$, in this region we recorded 2358 SNPs, with $r^2 = 0.145282$. The lowest P-value threshold is 0.001, with p-value $5.8e-6$. Then we can visualize it in the Figures 4.29 and 4.30:

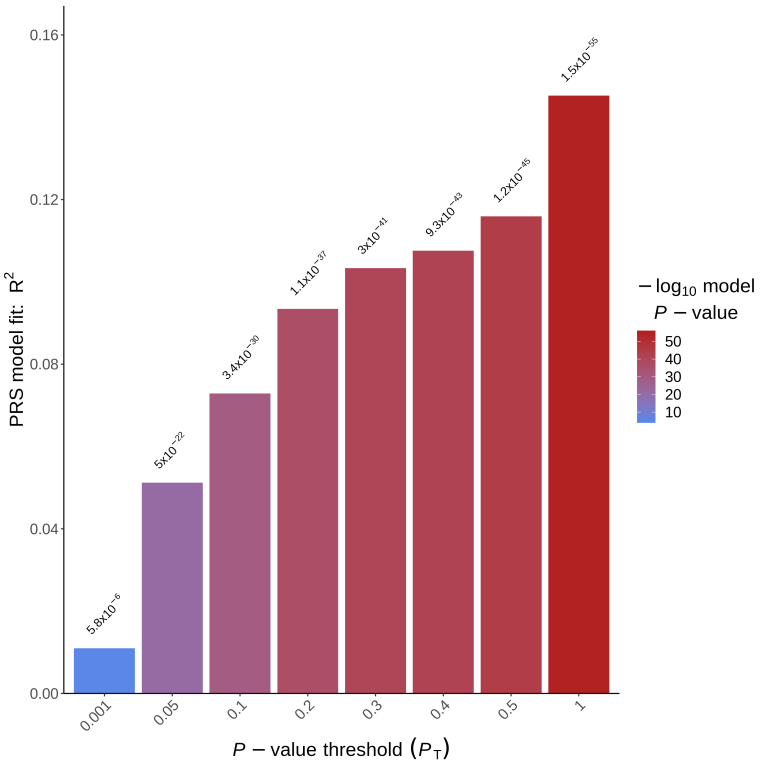


Figure 4.29: *Summary of Barplot, recorded High resolution of Risk predictive of Severe Malaria, the best PRS model fit has r^2=0.145282*

We can plot the high resolution PRS score for predictive risk as:

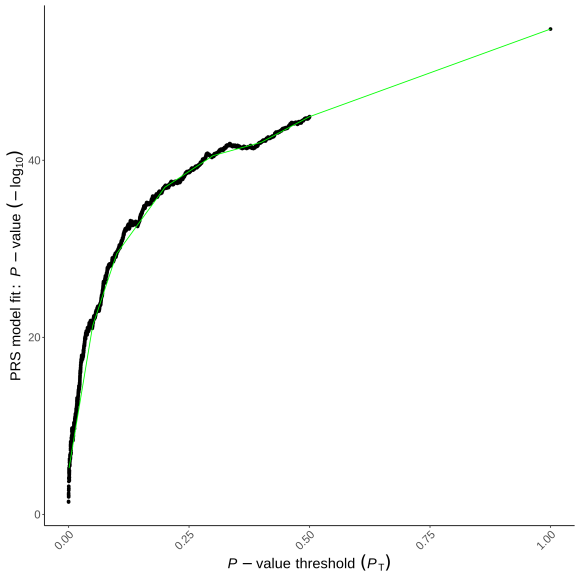


Figure 4.30: *This plot summarizes High-resolution PRS for 'Host having resistance to Malaria' predicting Severe status. The high-resolution best-fit PRS is 1.51515e-55 at P_T = 1.*

we can define the interval of 100 quantiles. This make link between odd ratio for score on the phenotype and strata for polygenic score, we can visualize it in Figure 4.31:

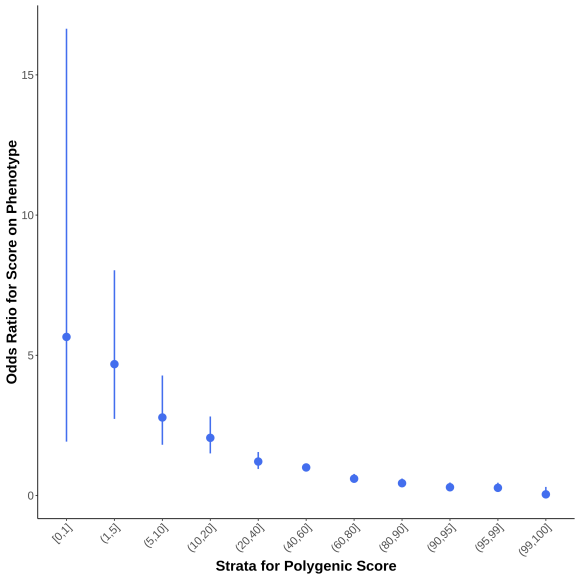


Figure 4.31: *This PRS Strata plot provide the information of each polygenic score interval and the odds ratio for score on Severe Malaria.*

Chapter 5

General Discussion and Conclusion

5.1 Discussion

The present section is based on discussion of our analysis. We noted that during the blood stage life cycle of the parasite, malaria resistance is due to multiple genes and pathways related to malaria pathogenesis, which include merozoite invasion, parasite development, cytoadherence, and signal transduction.

The genes identified by functional genome wide association study (fgwas) analysis might be relevant ; Through structural changes, these genes may function at the protein level, whereas the genes found by eQTL and chromatin interactions exert over quantitative variations at stages of gene expression their influences. In this study, we applied statistical methods for functional analytic to GWAS dataset of severe malaria susceptibility and identified the previous well-known loci associated with Malaria resistance and novel genes that were conducted to future functional experiments.

GWAS Meta-Analysis has been recognized as one of the robust post-gwas methods used to improve or boost the statistical power of genetic variant effects. Several tools for performing Meta-analysis have been proposed and used during the analysis, the evaluations were very important and informative. The studies of the best tool that suits a given population, however the challenge remains on the characteristic of the populations, which is either homogeneous (case of European populations) having a large linkage disequilibrium(LD), some are heterogeneous with short LD (Case of Africans, Americans) which can lead to underpower a GWAS analysis. In chapter4, we have analyzed GWAS study based on 3 different populations (Kenya, Malawi and Gambia), these studies have presented many improvement in terms of quality control (check

4.1) , but unfortunately many SNPs were underpowered (i.e based on P-value threshold 5e-07). For example, The identification of association at known malaria resistance loci in the Gambia population was highly attenuated owing to the low capacity of LD. This has been proved by the MRC Centre for Genomics and Global Health (CGGH) and the MRC Unit in the Gambia published on [94], none of SNPs in the Gambia population has reached the p-value threshold. In the Kenya population we recorded 10 significant SNPs (based on p-value threshold 5e-07) identified in 2 genes (*C4orf19* with 6 SNPs highly in LD and 1 one snp in *PAM*). Most of the identified SNPs were in strong LD , which means some were false positive. In terms of function, genetic variation in *C4orf19* has the impact into the concentration of glutamates in brains of patients with several sclerosis. Also, *PAM* PAM ("Peptidylglycine Alpha-Amidating Monooxygenase"), this gene is able to encode multiple function of protein. It has proteolytically as encoded protein, is processed to provide a mature enzyme. So, these enzyme has two functional domains having different catalytic activities such that peptidylglycine alpha-hydroxylating mono-oxygenase (PHM) domain, peptidyl alpha-amidating lyase (PAL) domain. So, the catalytic domains have a role of catalyzing the conversion neuro-endocrine peptides to activate active alpha-amidated products sequentially. Multiple transcript variants arise from alternative splicing, proteolytic has processed at least one of them encodes an isoform. Diseases associated with PAM include Menkes Disease and Gestational Trophoblastic Neoplasm. The annotations of Gene Ontology (GO) have been associated with genes including copper ion binding and L-ascorbic acid binding [103].

In parallel, Malawi GWAS analysis has presented the same challenge of GWAS underpower, we recorded 4 significant SNPs distributed as 2 SNPs in *SCAMP2*, and others identified into non-coding parts. The gene product *'SCAMP2'* is part of the *SCAMP* family of proteins that are proteins of the secretory carrier membrane. This works in post-golgi recycling pathways as carriers to the cell surface. So, most of these gene families are highly related and expressed together. These findings have shown that the SCAMPs may work at the same location during vesicular transport rather than in different pathways. Alternate splicing results in multiple transcript variants [104]. GWAS analysis of these 3 studies have shown that the signal across these studies was limited, this is due to many parameters including but not limited to heterogeneity of populations and environmental factors. The new strategy is to boost the statistical power by using GWAS Meta-Analysis. From our analysis, we have conducted GWAS Meta-analysis based on 2 tools (Metal and Metasoft). From Metal, we identified 3 significant SNPs (based on p-value threshold 5e-07), and identified only the gene *C4orf19*, thus revealing the weakness of the tool. We know that Metal is based on fixed effect. From this result we conclude that our study is not sufficiently adapted to fixed effects due to high degree of heterogeneity and random effects. To figure out this challenge, it was

relevant to consider the model which might capture the effects and minimize the bias like random effect, heterogeneity and aggregate the effects of variants across the studies. We therefore decided to use Metasoft tools. Metasoft is based-on Bayesian approach, this takes in account "Random effect model", "heterogeneity factors", "Binary effect" and evaluates the effect based on a posterior probability called "M-values". From Metasoft, we got 20 significant SNPs which mapped into 7 genes *(C4orf19, CDH13, IQCE, BET1L, EIF4E, BORCS5, LOC105378641)*.

We have C4orf19 repeatedly identified by both model (Metal and Metasoft) results, we perform further analysis of this gene; CDH13 encoding cadherin 13, was reported from [105] as Gene Associated with Protection against malaria. *Cadherin-13 (CDH13)*, a unique glycosylphosphatidylinositol anchored part of the cadherin family of cell adhesion molecules, this gene were identified as a risk gene for attention-deficit/hyperactivity disorder *(ADHD)*, various comorbid neurodevelopmental and psychiatric conditions, including depression, substance abuse, autism spectrum disorder and violent behavior [105].. The dysfunction mechanism of CDH13 can influence neuropsychiatric conditions to remain elusive. Also, *CDH13* plays a major role in brain activity by inhibiting modulation and identifying GABAergic interneurons synaptic function [106].

The purpose of this project was to analyze a statistical model which can conduct the identification of most relevant genomic annotations related to the biology of malaria Resistance and susceptibility that can improve gwas power, we were expecting the capacity of model of searching hundreds of genomic annotations having no prior knowledge of phenotype biology to group the collection of biologically interpretable annotations. Basically, fgwas-tool used to assess the relevance of from 450 different genomic annotations, enhancers,protein-coding genes and DNase-I hypersensitive sites in over 100 tissues and cell lines from *1000GenomesProjectannotatedVCFs, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/functional_annotation/annotated_vcfs/1000GenomesProjecthaplotypes, ftp://ftp-trace.ncbi.nih.go*. We have considered GWAS Meta-analysis summary statistics found from Metasoft (recorded 5.942.350 SNPs, with 20 significant SNPs based on threshold p-value 5e-07). We added the annotation that most significantly improves the likelihood to the model and we got the result shown in Table 4.9 describing 15 annotations which significantly improve the likelihood, Thus among these annotations, the best likelihood enriched is in non-synonymous region with likelihood (llk) 0.708068 and AIC 0.58. The idea is to assume that the probability of associate SNP in a specific genomic region to the phenotype depends on the density of SNP instead of physical size. This means that with a large number of SNPs, a concise genomic region It is as likely to have an association with few SNPs as a long genomic region a priori. We added "Non-synonymous " to others (14 remaining annotations), and we got the results shown in Table 4.11. The results showed that the more we are adding

the annotations, the more the model becomes accurate, because the previous probability added annotations will be considered as prior probability. After 6 iterations of adding annotations, the likelihood maintained was 0.924007, but the model has improved the challenge of overfitting, We have applied some machine learning technicalities like (using cross-validation likelihood) to find the best model penalized, because shrinkage estimators of the annotation parameters is used to maximize likelihood estimate of all parameters. Thus, we got the results from Table 4.12. To evaluate the best model we choose the penalty parameter of the highest cross-validation likelihood. We therefore considered the best model to perform the posterior probability of association (PPA). The significance is based on PPA > 0.9, from Table 4.13 we got 29 significant snps having PPA > 0.9 (with 16 snps identified to coding region and 13 into non-coding region).
From these variants identified into the coding regions we found 16 genes *(TCERG1L, C4orf19, ABO, ATP2B4, CBLB, EYA2, HBB, HBD, HERPUD1, IQCJ, MPP7, NAV1, NUP210, SAMD5, TCERG1L, TMEM229B)*. Among these genes, 5 of them were already identified for malaria, such as :

- *ABO* which is *ABO-blood-group* types, it plays the role of protection against severe Malaria and *Plasmodium falciparum* [107].
- *ATP2B4*, this has a polymorphism play the role of protection against malaria in case of pregnancy [108].
- *HBB* has a role of genetic control for resistance of human malaria [30].
- *HBD* the blood transcriptome of childhood malaria [109],serves in expression and regulation of the human β-defensins hBD-1 and hBD-2 in intestinal epithelium [110])
- *CBLB* plays a role in the analysis of Interactive transcriptome for malaria patients and infecting *Plasmodium falciparum*

From Genemania these 5 genes belong to the same pathway. These genes co-expressed the phenotype at 30.4%, pathway at 50.08%. From Enrichment analysis based on Enrichr the pathway has shown high significance of association to malaria with P-value at 6.816E-09, adjusted P-value 1.184E-05 , odd ratio at 666.67 and combined score 12536.03. Also, for Hemoglobin levels we found 2.014E-06 as associated p-value, 0.001749 adjusted P-value,odd ratio 117.65 and combined score 1542.99.
During the analysis, we have noticed that after performing GWAS and Post-GWAS, we always find some mutations into the gene *C4orf19*, this gene looks novel in this analysis. This may be addressed as a new question for the future works related to Malaria.

FGWAS has presented many advantages, it does not require raw data for analysis, the method is

based on summary statistics, he has several approaches used to decide the significance association, such as Bayes approaches, posterior probability of Association, etc ...
Among the limitations of FGWAS, we have noticed reference genome for annotations, most of the variant was not annotated in the reference incorporated in the tool. This makes the study limited and none accurate.
In addition, we also did the analysis of rare variants. We collected the variants with $MAF < 1\%$, we aggregate the effect by performing the "Skat and Burden" association test using from "SKAT-tool", than we mapped the significant variant based on standard p-value (0.05/M, with M the SNP quantities) onto gene-levels, then we found 31 genes, see Table 4.16. We performed the enrichment and pathway analysis for these genes. From Genemania we have established the scores of different genes, then we identified 6 genes with top scores, *SAP30, C11orf53, PAPOLB, MYO1H, RNF183, PLPP7* see Table 4.19. In this context, *SAP30* participate to protective vaccination and DNA gene promoters methylation of localized in liver of Balb which is modified by blood-stage malaria [102], In colorectal cancer mapping to chromosome 11q23.1, we have identified and characterized *C11orf53* which plays the role of functional risk variants [111], *PAPOLB (Poly(A) Polymerase Beta)* is a Protein Coding gene. Among its related pathways are mRNA surveillance pathways, RNA tissue specificity of this gene is enriched on Testis, *MYO1H (Myosin IH)* is a Protein Coding gene. Diseases associated with MYO1H include Central Hypoventilation Syndrome, Congenital and Spastic Paraplegia 36, Autosomal Dominant. Among its related pathways are Pathogenic Escherichia coli infection and Deltaf508-CFTR traffic / ER-to-Golgi in CF. Gene Ontology (GO) annotations related to this gene include actin binding and motor activity. An important paralog of this gene is *MYO1C.* [112].

5.2 Potential Impact of study

In this study we showed the advantage of using Meta-analysis as a post-gwas method and aggregate the effects based on functional genomics information incorporated in GWAS method. The approach was highly accurate and reliable due to some replicated genes found during analysis which have been reported from malaria by many international reviews, and the fixed novel discovered using several strategies.These finding will be potentially helpful to elucidate the clinical analysis and inform the relevant information . Malaria is reported to have significant measurable direct and indirect costs, thus, shown to be a major constraint to economic development by retarding productivity and growth. This is experienced mainly through diversion of resources to control malaria and loss of human lives which is very essential to the socio-economic development

of a nation. This study has therefore contributed to knowledge that could help in the translation of research findings in measurable outputs for malaria control.

5.3 Limitations and future work

Despite the impact of the study and robust analysis conducted, we have identified some limitations. However, fgwas method has incorporated functional annotations information characterizing each SNP identified during gwas analysis. Unfortunately, the reference genome of these functional annotations contains a limited number of annotations, some of SNPs are not annotated in the reference genome (it covers 450 functional annotations) this makes study limited in case of high number of SNPs unknown. Method requires frequently an updated reference genome to conduct a robust analysis.

This study was conducted for 3 regions (Kenya, Malawi and Gambia), during the analysis we have identified *C4orf19* repeatedly contains some significant SNPs from different tools and approaches, we hope a new future study to investigate the full spectrum of risk variants (rare variant , common variant and intermediate frequency variants; point (SNPs) mutation, small insertion deletion and copy number variations) potentially involved in malaria. This approach opens unique opportunities to interrogate the full spectrum genetic variations so far unexplored such as the low-frequency $(MAF < 5\%)$ and variants in the non-coding region of the genome, seat of the gene regulation.

References

[1] Karine G Le Roch, Yingyao Zhou, Peter L Blair, Muni Grainger, J Kathleen Moch, J David Haynes, Patricia De la Vega, Anthony A Holder, Serge Batalov, Daniel J Carucci, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639):1503–1508, 2003.

[2] Weihua Shou, Dazhi Wang, Kaiyue Zhang, Beilan Wang, Zhimin Wang, Jinxiu Shi, and Wei Huang. Gene-wide characterization of common quantitative trait loci for abcb1 mrna expression in normal liver tissues in the chinese population. *PloS one*, 7(9):e46295, 2012.

[3] Malaria Genomic Epidemiology Network. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from africa, asia and oceania. *Nature communications*, 10, 2019.

[4] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[5] Reedik Mägi and Andrew P Morris. Gwama: software for genome-wide association meta-analysis. *BMC bioinformatics*, 11(1):288, 2010.

[6] Charles M Poser, George W Bruyn, et al. *An illustrated history of malaria.* Parthenon Publishing Group, 1999.

[7] Francis EG Cox. History of the discovery of the malaria parasites and their vectors. *Parasites & vectors*, 3(1):1–9, 2010.

[8] Giovanni Battista Grassi. *Rapporti tra la malaria e peculiari insetti (zanzaroni e zanzare palustri).* Rome R. accad. d. Lincei., 1898.

[9] Henry E Shortt and Percy Cyril Claude Garnham. Pre-erythrocytic stage in mammalian malaria parasites. *Nature*, 161(4082):126–126, 1948.

[10] Wojciech A Krotoski, DM Krotoski, PC Garnham, RS Bray, R Killick-Kendrick, CC Draper, GA Targett, and MW Guy. Relapses in primate malaria: discovery of two populations of exoerythrocytic stages. preliminary note. *British medical journal*, 280(6208):153, 1980.

[11] Konstantinos Tzelepis, Hiroko Koike-Yusa, Etienne De Braekeleer, Yilong Li, Emmanouil Metzakopian, Oliver M Dovey, Annalisa Mupo, Vera Grinkevich, Meng Li, Milena Mazan, et al. A crispr dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell reports*, 17(4):1193–1205, 2016.

[12] Deepak Joshi, Jung-Yeon Kim, Wej Choochote, Mi-Hyun Park, and Gi-Sik Min. Preliminary vivax malaria vector competence for three members of the anopheles hyrcanus group in the republic of korea. *Journal of the American Mosquito Control Association*, 27(3):312–314, 2011.

[13] Karen Hayton and Xin-zhuan Su. Drug resistance and genetic mapping in plasmodium falciparum. *Current genetics*, 54(5):223–239, 2008.

[14] Xiao Yu, Baowei Cai, Mingjun Wang, Peng Tan, Xilai Ding, Jian Wu, Jian Li, Qingtian Li, Pinghua Liu, Changsheng Xing, et al. Cross-regulation of two type i interferon signaling pathways in plasmacytoid dendritic cells controls anti-malaria immunity and host mortality. *Immunity*, 45(5):1093–1107, 2016.

[15] Michael T Ferdig, Roland A Cooper, Jianbing Mu, Bingbing Deng, Deirdre A Joy, Xin-zhuan Su, and Thomas E Wellems. Dissecting the loci of low-level quinine resistance in malaria parasites. *Molecular microbiology*, 52(4):985–997, 2004.

[16] Fabrice Legros, Olivier Bouchaud, Thierry Ancelle, Amandine Arnaud, Sandrine Cojean, Jacques Le Bras, Martin Danis, Arnaud Fontanet, Rémy Durand, et al. Risk factors for imported fatal plasmodium falciparum malaria, france, 1996-2003. *Emerging infectious diseases*, 13(6):883, 2007.

[17] Juan Antonio Vizcaíno, Attila Csordas, Noemi Del-Toro, José A Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, Yasset Perez-Riverol, Florian Reisinger, Tobias Ternent, et al. 2016 update of the pride database and its related tools. *Nucleic acids research*, 44(D1):D447–D456, 2015.

[18] Authors/Task Force members, Stephan Windecker, Philippe Kolh, Fernando Alfonso, Jean-Philippe Collet, Jochen Cremer, Volkmar Falk, Gerasimos Filippatos, Christian Hamm, Stuart J Head, et al. 2014 esc/eacts guidelines on myocardial revascularization: the task force on myocardial revascularization of the european society of cardiology (esc) and the european

association for cardio-thoracic surgery (eacts) developed with the special contribution of the european association of percutaneous cardiovascular interventions (eapci). *European heart journal*, 35(37):2541–2619, 2014.

[19] Lisa C Ranford-Cartwright and Jonathan M Mwangi. Analysis of malaria parasite phenotypes using experimental genetic crosses of plasmodium falciparum. *International journal for parasitology*, 42(6):529–534, 2012.

[20] Patricia Schlagenhauf-Lawlor. *Travelers' malaria*. PMPH-USA, 2008.

[21] Alan F Cowman, Drew Berry, and Jake Baum. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *Journal of cell Biology*, 198(6):961–971, 2012.

[22] MM Stevenson, JJ Lyanga, and E Skamene. Murine malaria: genetic control of resistance to plasmodium chabaudi. *Infection and Immunity*, 38(1):80–88, 1982.

[23] John H Adams, Dlana E Hudson, Motomi Torii, Gary E Ward, Thomas E Wellems, Masamichi Aikawa, and Louis H Miller. The duffy receptor family of plasmodium knowlesi is located within the micronemes of invasive malaria merozoites. *Cell*, 63(1):141–153, 1990.

[24] DJ Weatherall. Host genetics and infectious disease. *Parasitology*, 112(S1):S23–S29, 1996.

[25] Stephen J Rogerson, Rushika S Wijesinghe, and Steven R Meshnick. Host immunity as a determinant of treatment outcome in plasmodium falciparum malaria. *The Lancet infectious diseases*, 10(1):51–59, 2010.

[26] AVS Hill. Genetic susceptibility to malaria and other infectious diseases: from the mhc to the whole genome. *Parasitology*, 112(S1):S75–S84, 1996.

[27] Margaret J Mackinnon, Tabitha W Mwangi, Robert W Snow, Kevin Marsh, and Thomas N Williams. Heritability of malaria in africa. *PLoS medicine*, 2(12):e340, 2005.

[28] F Verra, VD Mangano, and D Modiano. Genetics of susceptibility to plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies. *Parasite immunology*, 31(5):234–253, 2009.

[29] Dominic P Kwiatkowski. How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics*, 77(2):171–192, 2005.

[30] Malaria Genomic Epidemiology Network, Kirk A Rockett, Geraldine M Clarke, Kathryn Fitzpatrick, Christina Hubbart, Anna E Jeffreys, Kate Rowlands, Rachel Craik, Muminatou Jallow, David J Conway, et al. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature genetics*, 46(11):1197, 2014.

[31] Jason P Wendler, John Okombo, Roberto Amato, Olivo Miotto, Steven M Kiara, Leah Mwai, Lewa Pole, John O'Brien, Magnus Manske, Dan Alcock, et al. A genome wide association study of plasmodium falciparum susceptibility to 22 antimalarial drugs in kenya. *PLoS One*, 9(5):e96486, 2014.

[32] Adel Driss, Jacqueline M Hibbert, Nana O Wilson, Shareen A Iqbal, Thomas V Adamkiewicz, and Jonathan K Stiles. Genetic polymorphisms linked to susceptibility to malaria. *Malaria journal*, 10(1):271, 2011.

[33] DJ Weatherall and JB Clegg. Genetic variability in response to infection: malaria and after. *Genes & Immunity*, 3(6):331–337, 2002.

[34] Alfred Cortés, Ariadna Benet, Brian M Cooke, John W Barnwell, and John C Reeder. Ability of plasmodium falciparum to invade southeast asian ovalocytes varies between parasite lines. *Blood*, 104(9):2961–2966, 2004.

[35] CA Facer. Erythrocytes carrying mutations in spectrin and protein 4.1 show differing sensitivities to invasion by plasmodium falciparum. *Parasitology research*, 81(1):52–57, 1995.

[36] Hurng-Yi Wang, Hua Tang, C-K James Shen, and Chung-I Wu. Rapidly evolving genes in human. i. the glycophorins and their possible role in evading malaria parasites. *Molecular biology and evolution*, 20(11):1795–1804, 2003.

[37] SJ Allen, A O'donnell, NDE Alexander, MP Alpers, TEA Peto, JB Clegg, and DJ Weatherall. α+-thalassemia protects children against disease caused by other infections as well as malaria. *Proceedings of the National Academy of Sciences*, 94(26):14736–14741, 1997.

[38] G Pasvol and RJM Wilson. The interaction of malaria parasites with red blood cells. *British Medical Bulletin*, 38(2):133–140, 1982.

[39] Sarah A Tishkoff, Robert Varkonyi, Nelie Cahinhinan, Salem Abbes, George Argyropoulos, Giovanni Destro-Bisol, Anthi Drousiotou, Bruce Dangerfield, Gerard Lefranc, Jacques Loiselet, et al. Haplotype diversity and linkage disequilibrium at human g6pd: recent origin of alleles that confer malarial resistance. *Science*, 293(5529):455–462, 2001.

[40] Pierre M Durand and Theresa L Coetzer. Pyruvate kinase deficiency protects against malaria in humans. *Haematologica*, 93(6):939–940, 2008.

[41] Alfred Cortés, Mata Mellombo, Charles S Mgone, Hans-Peter Beck, John C Reeder, and Brian M Cooke. Adhesion of plasmodium falciparum-infected red blood cells to cd36 under flow is enhanced by the cerebral malaria-protective trait south–east asian ovalocytosis. *Molecular and biochemical parasitology*, 142(2):252–257, 2005.

[42] Susan Blumenfield. Reflections on effective leadership: strains and successes, strategies and styles. *Social Work in Health Care*, 20(4):21–37, 1995.

[43] J Alexandra Rowe, Ian G Handel, Mahamadou A Thera, Anne-Marie Deans, Kirsten E Lyke, Abdoulaye Koné, Dapa A Diallo, Ahmed Raza, Oscar Kai, Kevin Marsh, et al. Blood group o protects against severe plasmodium falciparum malaria through the mechanism of reduced rosetting. *Proceedings of the National Academy of Sciences*, 104(44):17471–17476, 2007.

[44] Atif A Elagib, Amel O Kider, Bo Åkerström, and Mustafa I Elbashir. Association of the haptoglobin phenotype (1—1) with falciparum malaria in sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 92(3):309–311, 1998.

[45] Maurine R Hobbs, Venkatachalam Udhayakumar, Marc C Levesque, Jennifer Booth, Jacquelin M Roberts, Ariana N Tkachuk, Ann Pole, Hilary Coon, Simon Kariuki, Bernard L Nahlen, et al. A new nos2 promoter polymorphism associated with increased nitric oxide production and protection from severe malaria in tanzanian and kenyan children. *The Lancet*, 360(9344):1468–1475, 2002.

[46] D Garcia-Santos and JAB Chies. Ho-1 polymorphism as a genetic determinant behind the malaria resistance afforded by haemolytic disorders. *Medical hypotheses*, 74(5):807–813, 2010.

[47] Masatoshi Nei. *Molecular evolutionary genetics*. Columbia university press, 1987.

[48] Carl E Rehfeld, JAMES W BACUS, Jeanne A Pagels, and Merlin H Dipert. Computer calculation of wright's inbreeding coefficient. *Journal of Heredity*, 58(2):81–84, 1967.

[49] E Azevedo, NE Morton, C Miki, and Shirley Yee. Distance and kinship in northeastern brazil. *American journal of human genetics*, 21(1):1, 1969.

[50] Masatoshi Nei. Genetic distance between populations. *The American Naturalist*, 106(949):283–292, 1972.

[51] Richard Durrett. *Probability models for DNA sequence evolution*. Springer Science & Business Media, 2008.

[52] WG Hill and Alan Robertson. Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, 38(6):226–231, 1968.

[53] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477, 2008.

[54] Andreas Ziegler, Inke R König, and Friedrich Pahlke. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons, 2010.

[55] Eran Halperin, Gad Kimmel, and Ron Shamir. Tag snp selection in genotype data for maximizing snp prediction accuracy. *Bioinformatics*, 21(suppl_1):i195–i203, 2005.

[56] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2013.

[57] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[58] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[59] Ben Hayes. Overview of statistical methods for genome-wide association studies (gwas). In *Genome-wide association studies and genomic prediction*, pages 149–169. Springer, 2013.

[60] Bridget Penman, Caroline Buckee, Sunetra Gupta, and Sean Nee. Genome-wide association studies in plasmodium species. *BMC biology*, 8(1):90, 2010.

[61] Matthew L Freedman, Alvaro NA Monteiro, Simon A Gayther, Gerhard A Coetzee, Angela Risch, Christoph Plass, Graham Casey, Mariella De Biasi, Chris Carlson, David Duggan, et al. Principles for the post-gwas functional characterization of cancer risk loci. *Nature genetics*, 43(6):513, 2011.

[62] Xiaoyang Zhang, Swneke D Bailey, and Mathieu Lupien. Laying a solid foundation for manhattan–'setting the functional basis for the post-gwas era'. *Trends in Genetics*, 30(4):140–149, 2014.

[63] Gleb Kichaev, Gaurav Bhatia, Po-Ru Loh, Steven Gazal, Kathryn Burch, Malika K Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L Price. Leveraging polygenic functional enrichment to improve gwas power. *The American Journal of Human Genetics*, 104(1):65–75, 2019.

[64] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014.

[65] George W Comstock. Tuberculosis in twins: a re-analysis of the prophit survey. *American Review of Respiratory Disease*, 117(4):621–624, 1978.

[66] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291, 2015.

[67] Chao Huang, Paul Thompson, Yalin Wang, Yang Yu, Jingwen Zhang, Dehan Kong, Rivka R Colen, Rebecca C Knickmeyer, Hongtu Zhu, Alzheimer's Disease Neuroimaging Initiative, et al. Fgwas: Functional genome wide association analysis. *NeuroImage*, 159:107–121, 2017.

[68] Xingbo Mo, Shufeng Lei, Yonghong Zhang, and Huan Zhang. Genome-wide enrichment of m 6 a-associated single-nucleotide polymorphisms in the lipid loci. *The Pharmacogenomics Journal*, 19(4):347–357, 2019.

[69] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[70] Charles E McCulloch and John M Neuhaus. Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online*, 2014.

[71] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–106, 2014.

[72] Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one*, 8(10):e75707, 2013.

[73] Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li, and Rongling Wu. The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523, 2011.

[74] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[75] Andrew F Jarosz and Jennifer Wiley. What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, 7(1):2, 2014.

[76] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1):79–86, 2009.

[77] William YS Wang, Bryan J Barratt, David G Clayton, and John A Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.

[78] Philippe Sanseau, Pankaj Agarwal, Michael R Barnes, Tomi Pastinen, J Brent Richards, Lon R Cardon, and Vincent Mooser. Use of genome-wide association studies for drug repositioning. *Nature biotechnology*, 30(4):317, 2012.

[79] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing gwas results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[80] Hae-Young Kim. Statistical notes for clinical researchers: effect size. *Restorative dentistry & endodontics*, 40(4):328–331, 2015.

[81] Ziyou Ren. *Development and Application of Innovative Algorithms for Transcriptome Profiling*. PhD thesis, Northwestern University, 2020.

[82] Elias Zintzaras and John PA Ioannidis. Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Computational biology and chemistry*, 32(1):39–47, 2008.

[83] Paul DP Pharoah, Ya-Yu Tsai, Susan J Ramus, Catherine M Phelan, Ellen L Goode, Kate Lawrenson, Melissa Buckley, Brooke L Fridley, Jonathan P Tyrer, Howard Shen, et al. Gwas meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature genetics*, 45(4):362–370, 2013.

[84] Ferdouse Begum, Debashis Ghosh, George C Tseng, and Eleanor Feingold. Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic acids research*, 40(9):3777–3784, 2012.

[85] Buhm Han and Eleazar Eskin. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet*, 8(3):e1002555, 2012.

[86] Cristen J Willer, Serena Sanna, Anne U Jackson, Angelo Scuteri, Lori L Bonnycastle, Robert Clarke, Simon C Heath, Nicholas J Timpson, Samer S Najjar, Heather M Stringham, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*, 40(2):161–169, 2008.

[87] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[88] Lennart C Karssen, Cornelia M van Duijn, and Yurii S Aulchenko. The genabel project for statistical genomics. *F1000Research*, 5, 2016.

[89] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D Buxbaum, and Xihong Lin. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, 2013.

[90] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 2009.

[91] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[92] T Tony Cai, X Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011.

[93] Malaria Genomic Epidemiology Network, Ea Achidi, T Agbenyega, S Allen, O Amodu, K Bojang, D Conway, P Corran, P Deloukas, A Djimde, et al. A global network for genomic epidemiology of malaria. 2008.

[94] Muminatou Jallow, Yik Ying Teo, Kerrin S Small, Kirk A Rockett, Panos Deloukas, Taane G Clark, Katja Kivinen, Kalifa A Bojang, David J Conway, Margaret Pinder, et al. Genome-wide and fine-resolution association analysis of malaria in west africa. *Nature genetics*, 41(6):657–665, 2009.

[95] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association

studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[96] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[97] Paul R Buckland. The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1762(1):17–28, 2006.

[98] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220, 2010.

[99] Max Franz, Harold Rodriguez, Christian Lopes, Khalid Zuberi, Jason Montojo, Gary D Bader, and Quaid Morris. Genemania update 2018. *Nucleic acids research*, 46(W1):W60–W64, 2018.

[100] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[101] Carol D Laherty, Andrew N Billin, Robert M Lavinsky, Gregory S Yochum, Angela C Bush, Jian-Min Sun, Tina-Marie Mullen, James R Davie, David W Rose, Christopher K Glass, et al. Sap30, a component of the msin3 corepressor complex involved in n-cor-mediated repression by specific transcription factors. *Molecular cell*, 2(1):33–42, 1998.

[102] Saleh Al-Quraishy, Mohamed A Dkhil, Abdel-Azeem S Abdel-Baki, Foued Ghanjati, Lars Erichsen, Simeon Santourlidis, Frank Wunderlich, and Marcos J Araúzo-Bravo. Protective vaccination and blood-stage malaria modify dna methylation of gene promoters in the liver of balb/c mice. *Parasitology research*, 116(5):1463–1477, 2017.

[103] Arvid Suls, Johanna A Jaehn, Angela Kecskés, Yvonne Weber, Sarah Weckhuysen, Dana C Craiu, Aleksandra Siekierska, Tania Djémié, Tatiana Afrikanova, Padhraig Gormley, et al. De novo loss-of-function mutations in chd2 cause a fever-sensitive myoclonic epileptic

encephalopathy sharing features with dravet syndrome. *The American Journal of Human Genetics*, 93(5):967–975, 2013.

[104] Lixia Liu, Haini Liao, Anna Castle, Jie Zhang, James Casanova, Gabor Szabo, and David Castle. Scamp2 interacts with arf6 and phospholipase d1 and links their function to exocytotic fusion pore formation in pc12 cells. *Molecular biology of the cell*, 16(10):4463–4472, 2005.

[105] Margaret J Mackinnon, Carolyne Ndila, Sophie Uyoga, Alex Macharia, Robert W Snow, Gavin Band, Anna Rautanen, Kirk A Rockett, Dominic P Kwiatkowski, and Thomas N Williams. Environmental correlation analysis for genes associated with protection against malaria. *Molecular biology and evolution*, 33(5):1188–1204, 2016.

[106] Olga Rivero, Martijn M Selten, S Sich, S Popp, L Bacmeister, E Amendola, Moritz Negwer, Dirk Schubert, F Proft, D Kiser, et al. Cadherin-13, a risk gene for adhd and comorbid disorders, impacts gabaergic function in hippocampus and cognition. *Translational psychiatry*, 5(10):e655–e655, 2015.

[107] SL Pathirana, HK Alles, S Bandara, M Phone-Kyaw, MK Perera, AR Wickremasinghe, KN Mendis, and SM Handunnetti. Abo-blood-group types and protection against severe, plasmodium falciparum malaria. *Annals of Tropical Medicine & Parasitology*, 99(2):119–124, 2005.

[108] George Bedu-Addo, Stefanie Meese, and Frank P Mockenhaupt. An atp2b4 polymorphism protects against malaria in pregnancy. *The Journal of infectious diseases*, 207(10):1600–1603, 2013.

[109] Angelica BW Boldt, Hoang Van Tong, Martin P Grobusch, Yvonne Kalmbach, Arnaud Dzeing Ella, Maryvonne Kombila, Christian G Meyer, Jürgen FJ Kun, Peter G Kremsner, and Thirumalaisamy P Velavan. The blood transcriptome of childhood malaria. *EBioMedicine*, 40:614–625, 2019.

[110] Deborah A O'Neil, Edith Martin Porter, Dirk Elewaut, G Mark Anderson, Lars Eckmann, Tomas Ganz, and Martin F Kagnoff. Expression and regulation of the human β-defensins hbd-1 and hbd-2 in intestinal epithelium. *The Journal of Immunology*, 163(12):6718–6724, 1999.

[111] Michela Biancolella, Barbara K Fortini, Stephanie Tring, Sarah J Plummer, Gustavo A Mendoza-Fandino, Jaana Hartiala, Michael J Hitchler, Chunli Yan, Fredrick R Schumacher, David V Conti, et al. Identification and characterization of functional risk variants for

colorectal cancer mapping to chromosome 11q23. 1. *Human molecular genetics*, 23(8):2198–2209, 2014.

[112] Jonathan S Berg, Bradford C Powell, and Richard E Cheney. A millennial myosin census. *Molecular biology of the cell*, 12(4):780–794, 2001.

Appendix

Rare Variant