Modelling and Construction of Pan-genome of *Plasmodium* Species: A Region Wise Analysis

By

Farhana Riaz

Fall-2018-MSBI-3-00000277747

Supervised by:

Dr Mehak Rafiq

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE in Bioinformatics

Research Centre for Modelling and Simulation (RCMS)

National University of Science & Technology (NUST)

Sept 2021

Dedication

This thesis is dedicated to my beloved family for their constant support and love.

Certificate of Originality

I, Farhana Riaz, hereby declare that the results presented in this research work titled "Modelling and Construction of Pan-genome of *Plasmodium* Species: A Region Wise Analysis" are generated by myself. Moreover, none of its contents is plagiarised nor set forth for any kind of evaluation or higher education purposes. I have acknowledged/referenced all the literary content used for support in this research work.

_____

Farhana Riaz

Fall-2018-MSBI-3-00000277747

Sept, 2021

Acknowledgement

Praise to the Almighty Allah, Lord of the worlds, for blessing humans with the ability to use their knowledge for the welfare of living beings and for giving me the strength and determination to complete my research work.

I want to record my sincere gratitude and heartfelt thanks to my mentor and supervisor, Assistant Professor Dr. Mehak Rafiq, for her continuous support, advice, and constructive criticism in my research. She had been able to establish the culture of the research group and a healthy lab environment where we worked as a family. It has been a life-changing journey with her.

I am also highly indebted to my guidance and examination committee members Dr. Rehan Zafar Paracha and Dr. Amjad Ali, who have counselled me in various aspects providing assistance and encouragement during my research. I admire Rector NUST, Principal RCMS, HOD RCMS and NUST administration for developing an admirable learning and grooming platform. I feel honoured to be part of this institution.

My special and affectionate gratitude is to my family, especially my father, M. Riaz Akhtar and my mother, Farhat Parveen Hashmi. They can never be thanked enough for their overwhelming love, care, financial and moral support. I would especially like to mention my brother M. Usman Riaz for his never-ending support and contribution to my research work.

I am much obliged to all my friends for being an amazing part of my Master's degree. I appreciate the support and continuous motivation of my Aqsa Khalid, Maleeha Ahmed, Tayyaba Alvi, Noor-Us-Subah, Mehar Masood and my entire research group. I wish them happiness and success in all their future endeavours.

Contents

List of Abbreviations

GO	Gene Ontology
PfEMP1	*P. falciparum* erythrocyte membrane protein 1
PPM	Parasite's Plasma Membrane
PVM	Parasitophorus Vacuolar Membrane
PTEX	*Plasmodium* Translocon of Exported Proteins
RBC	Red Blood Cells
rRNA	Ribosomal Ribonucleotide Acid
PF or *P. falciparum*	*Plasmodium Falciparum*
PV or *P. vivax*	*Plasmodium Vivax*
PF_Africa 1	*P. falciparum's* "GN01" strain from Guinea
PF_ Africa 2	*P. falciparum's* "MaliPS096_E11" strain from Mali
PF_Africa 3	*P. falciparum's* ‘‘2000708" strain from Tanzania
PF_Africa 4	*P. falciparum's* "Palo Alto/Uganda" strain from Uganda
PF_Asia 1	*P. falciparum's* "KH01" strain from Cambodia
PF_Asia 2	*P. falciparum's* "Dd2" strain from Indochina
PF_Asia 3	*P. falciparum's* "Vietnam Oak-Knoll" strain from Vietnam
PF_Asia 4	*P. falciparum's* "CAMP/Malaysia" strain from Malaysia
PF_Asia 5	*P. falciparum's* "FCH/4 "strain from Philippines
PF_Europe 1	*P. falciparum's* "3D7" strain from Netherlands
PF_Europe 2	*P. falciparum's* "NF54" strain from Netherlands
PF_Europe 3	*P. falciparum's* "NF54" strain from France
PF_North America 1	*P. falciparum's* "HB3" strain from Honduras
PF_North America 2	*P. falciparum's* "Santa Lucia" strain from Santa Lucia
PF_South America 1	*P. falciparum's* "7G8" strain from Brazil
PV_Africa 1	*P. vivax's* "Mauritania I" strain from Mauritania
PV_Asia 1	*P. vivax's* "India VII" strain from India
PV_Asia 2	*P. vivax's* "North Korean" strain from North Korea
Pv_Asia 3	*P. vivax's "PvT01"* strain from Thailand
Pv_Asia 4	*P. vivax's "PvP01"* strain from Indonesia

PV_Asia 5	*P. vivax's "PcC01"* strain from China
PV_North America 1	*P. vivax's "Salvador I"* strain from Salvador
PV_South America 1	*P. vivax's "Brazil I"* strain from Brazil
COG	Cluster of Orthologous Groups
PC	Principal Component
PCA	Principal Component Analysis
OG	Orthologous Clusters
MCL	Markov Clustering
DB	Database
AMA-1	Apical Membrane Antigen-1
PV-1	Parasitophorus Vacuolar protein

List of Figures

List of Tables

Abstract

Malaria, a mosquito-borne disease, continues to be a global health problem due to antibiotic resistance and mutations by host and environmental factors. It is caused by a single cell protozoan known as *Plasmodium,* of which *P. falciparum* and *P. vivax* are the only prevalent causative agents in Pakistan. Since the genomes are prone to variation from region to region, a single representative isolate is not sufficient to describe the entire heterogeneity of a species across different geographical regions. Modelling the pan-genome could provide us with a better insight into how the parasites develop the genetic variability and determine the minimum set of essential genes and common virulence factors. This study aims to model three pan-genomes region-wise (strains taken globally, only from Asian countries, from Asian countries excluding Pakistan's closest neighbour India. The Global pan-genome consisted of 2201 core genes (16.61%), 6716 accessory genes (70.7%) and 4329 unique genes (32.68%). The Asian pan-genome was comprised of 2587 core genes (23.06%), 4980 accessory genes (44.39%) and 3650 unique genes (32.53%). The Asian excluding India dataset was found out to be 2593 core genes (24.02%), 4719 accessory genes (43.72%) and 3480 genes (32.24%) as singletons. The study observed that when Pakistan's closest neighbour India was included in the Asian dataset, it increased the singletons by 170 genes. The pan-genome is found to be open where the gene pool kept on increasing by an average of 150 genes being added on the subsequent addition of a new genome. However, the core genome decreased with the addition of new strains until it became stable. The stable core genome contains three orthologous clusters (OG1.5_2955, OG1.5_ 2969 and OG1.5_3338) directly listed as being involved in the pathogenesis (GO:0009405). They encode Parasitophorus Vacuolar Protein 1(PV1), Apical Membrane Protein 1 (AMA-1) and 6-cysteine protein P47, respectively. These core viral proteins need to be evaluated in future for potential drug and vaccine candidates that can help combat the disease irrespective of its origin or type.

Introduction

1.1 Malaria

Malaria, a mosquito-borne infectious disease, is caused by the single-celled protozoan parasite known as *Plasmodium*. It continues to be a significant health problem globally [1]. The symptoms of malaria resemble those of common viral infections leading to delay in diagnosis. Some physical symptoms include high fever, chills, jaundice, myalgia, nausea, diarrhoea, and dry cough. Anaemia, splenomegaly and hepatomegaly often develop after some days caused by *Plasmodium* [2]–[4]. Metabolic complications include Hypoglycaemia and Acidosis. Severe malaria is often associated when the *P. falciparum* is detected by microscopy or having at least one criterion for severe malaria. It takes 48-72 hours to complete each cycle duration in *Plasmodium Falciparum* [2], [5].

1.1.1 Malaria in Pakistan

The World Malaria Report 2020 reports an increase in malaria cases from 228 million cases to 229 million cases in one year. Along with some other countries, Pakistan lies in the WHO's Eastern Mediterranean Region (EMRO), which is still considered a high-risk region. In the EMRO region. A 26% decrease in malaria cases from 7 million in 2000 to 5 million in 2019 was observed. Afghanistan and Pakistan were named the significant countries accounting for a quarter of the cases in the region (World Health Organisation, 2020).

Malaria is the second common clinically suspected disease in Pakistan [7]. According to Pakistan Malaria Annual Report 2019, 95 million, which makes roughly 60% of the population of Pakistan, are at high risk of malaria. The report also states that with 1 million estimated cases and 300,000 confirmed malaria cases, Pakistan is grouped with other countries to account for 95% of the total regional malaria burden [8]. A thorough review article indicates that despite Pakistan having a well-established Malaria control program, around 500,000 infections and 50,000 deaths in Pakistan can still be attributable to Malaria

[9]. Figure 1.1 shows the life cycle of *Plasmodium* species [10].

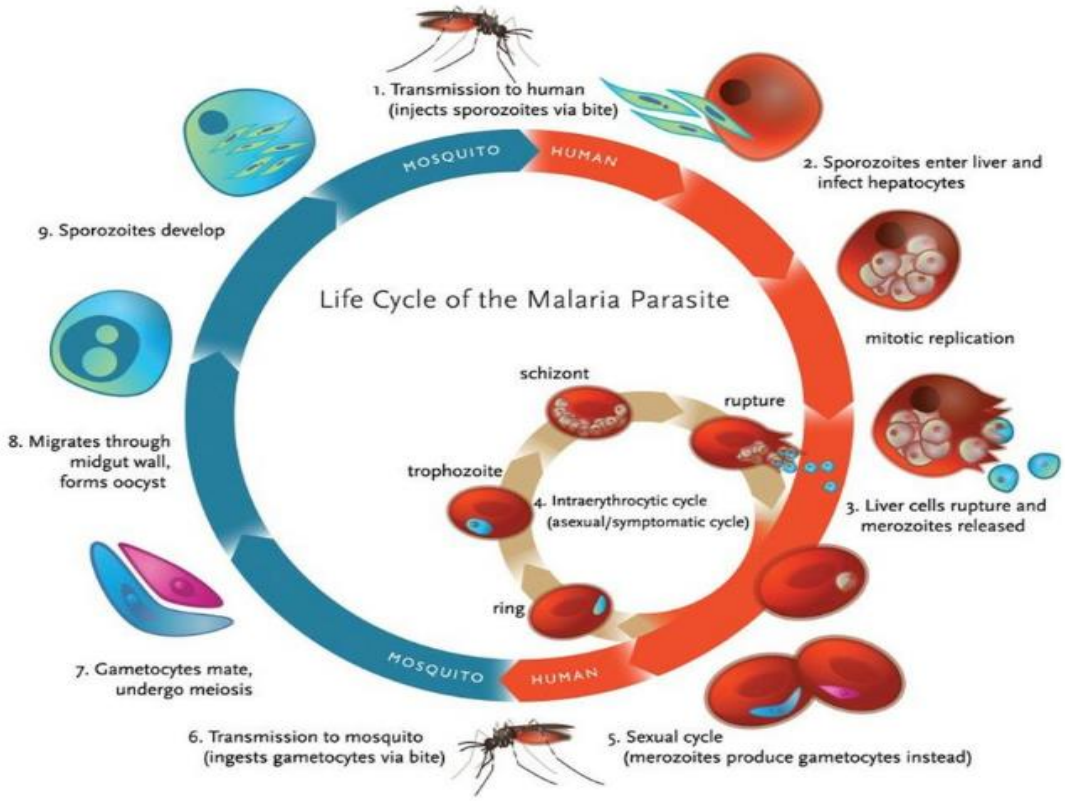

Figure 1.1: Life cycle of the malaria parasite Plasmodium. Plasmodium is transmitted to humans via the bite of a mosquito. Sporozoites are released that travel to the liver infecting hepatocytes. They undergo mitotic replication, after which liver cells rupture to release merozoites. In the intraerythrocytic cycle, the merozoites rupture to form schizonts and converts to trophozoite. It then enters the sexual cycle, where gametocytes are produced. Gametocytes are transmitted to another mosq+uito via a bite, where they undergo meiosis to mate. Finally, an oocyte is formed, and sporozoites develop, which ultimately are ingested into the blood of humans via a mosquito bite, causing malaria [10].

1.2 Types of *Plasmodium*

There are several different types of *Plasmodium*, of which five species are known as the causative agent of malaria in humans: *P. falciparum, P. vivax, P.ovale, P.malariae and P.knowlesi* [11]. Some other species of *Plasmodium* known to have mammalian hosts other than humans include *P. yoelli, P.bergheu, P.chabaudi, P.cynomolgi* [5], [12]. *P.knowlesi* primarily infects rodents and non-human primates, but it has led to some cases in humans where the mode of transmission is still not clear [13], whereas *P.malariae* additionally has hosts other than humans as African apes and South American monkeys

Each *Plasmodium* species and its interaction with the host determine the pathogenesis and severity of the disease [5], [12].

1.2.1 Prevalent Types of *Plasmodium* in Pakistan

P. falciparum and *P. vivax* are the only prevalent types of *Plasmodium* in Pakistan [8], [14]. *P. falciparum* is responsible for the most number of deaths globally due to its severe clinical forms. The second most prevalent species is the *P. vivax,* which also has the broadest geographical distribution. Refer to chapter 2, section 2.1, for details.

1.3 Need to Malaria Elimination

Elimination of malaria refers to zero incidences of locally transmitted cases of a specified parasite. In order to achieve complete elimination of malaria, the Global Technical Strategy (GTS) 2016-2030 was developed, which aims to decrease the rate of incidence and mortality at least 40% by 2020, 75% by 2025 and 90% by 2030. However, even though considerable efforts and progress had been made since 2000, the world was off track from achieving GTS 2020 milestone by 37% (56 incidence cases per 1000 population instead of 35 expected cases). If no actions are taken to reverse this trend, the world will be likely off track by 87% in 2030. Furthermore, an increased incidence in malaria cases from 2015 to 2020 was reported in 31 countries and 15 countries, increasing 40% more malaria cases incidence. Nine countries showed no change in their incidence rates between 2015 and 2020 (World Health Organization, 2020).

In endemic countries, efforts to achieve this are driven by ministries of health. Along with the strength of the national health system, biological determinants such as genomic diversity and demographics play a significant role in the rate of progress of elimination. Some of the additional factors to consider to achieve malaria elimination globally are discussed as follows and also in chapter 2, sections 2.3,2.4 and 2.5.

1.3.1 Antimalarial Drugs Resistance

Resistance to antimalarial drugs is a recurring issue that was first reported in the 1950s in *P. falciparum* and continues to be widespread. This issue shows an increasing concern for

researchers to treat and eliminate malaria. [15]. The earliest antimalarial treatment was Chloroquine which remained to be the first-line drug for quite many years. Over time, resistance was shown to this drug that resulted in a need for new drugs. The Chloroquine-resistance transporter (*pfcrt)* gene is known to be the mediator of chloroquine resistance (Fidock, D.A. *et al.*, 2000). Artemisinin-combination therapies (ACT) that combine Artemisinin with another antimalarial drug were recommended by WHO in 2005 to be used as first-line treatment. Even though the results were pretty impressive, with rapid clearance of parasites having a profound impact on malaria control, resistance to Artemisinin has now developed in *P. falciparum* (Bhatt, S. *et al.*, 2015). Multiple point mutations in the *Kelch13* gene affect the propeller region of the protein associated with Artemisinin resistance (Ashley, E.A. *et al.*, 2014). As per World Malaria Report 2020, the *PfKelch13* mutations are identified as molecular markers of Artemisinin resistance. As per a 2018 research paper, evidence suggests single nucleotide polymorphisms (SNPs) in genes were linked to clearance of the parasites being delayed upon treatment with ACT. The isolates on which this was tested were of *P.facliparum* from East and West Africa.

Similarly, SNPs were also detected in ACT resistance-associated genes of Southeast Asian strains (Hamilton, W.L. *et al.*, 2017). Also, most drugs do not effectively kill gametocytes, leading to a risk of transmission from treated patients. Some transmission-blocking drugs have reached pre-clinical and clinical phases that act against sexual stages. [20]. More on the biological markers for resistance can be read in Chapter 2, section 2.3.1.

1.3.2 Towards Vaccine

As described by World Health Organisation (WHO), a vaccine is the most cost-effective measure to eradicate malaria worldwide [21]. WHO's earlier focus used to be on malarial control rather than eradication, yet efforts to develop a vaccine for malaria began in the 1970s.

The World Malaria Report 2020 mentions malaria vaccine development as part of routine control efforts to be “The next major innovation” and “a new paradigm” in malaria control. (World Health Organisation, 2020). As per the position paper released by WHO in 2016, the European Medicines Agency (EMA) has given a positive recommendation to the

world's first malaria vaccine, RTS,S/AS01. About 500,000 people have already received their first dose of vaccination. Refer to Chapter 2, section 2.4, for a detailed quest of the malarial vaccine. The RTS,S/AS01 vaccine is a further ongoing evaluation to assess the vaccine's public health value, to be reviewed in late 2021 by WHO (World Health Organisation, 2020).

As limitations of the vaccine RTS, S/AS01 emerge as described in chapter 2, section 2.4, it consequently led to the loss of efficacy against divergent parasites [22], [23].

As per Penny *et al.*, there is still a need to identify new targets for malaria vaccines as clinical studies have been less than satisfying [24]. Finding out new biomarkers that can be easily measured can reduce the cost of clinical development. However, there is a considerable challenge pointing out that different populations show variation in immune response [25].

Hence, there is a need to understand the roles of population genetics in immune development.

Vaccines are generally designed by a single representative isolate, which is not sufficient to describe the entire genetic complexity of a species [11], [26]. The complete genome of *Plasmodium* has been sequenced and published worldwide in addition to publically available datasets of gene sequences available in various databases. This has emerged as a new hope for better potential antimalarial targets and vaccines.

1.3.3 Intra Species Variations are as Significant as Interspecies Diversity

At the beginning of the genomic era, a single representative isolate was thought to be sufficient to describe the genetic complexity of a species. However, microbial genomes are prone to variation from region to region. A plethora of studies on plants, animals and micro-organisms support the concept that intraspecies variation can be as significant as interspecies diversity [27], [28].

One such example is a study performed on rice varieties where the results showed that, on average, two genomes of the rice differed by 4000 genes (~10%). Compared with the reference genome, approximately 10,000 genes were missing in the reference strain [28].

1.4 Genome to Pan-genome – A Paradigm Shift

As *Plasmodium Falciparum* and *Plasmodium Vivax* strains from different geographical regions have evolved a successful parasite lifestyle, the common components in these strains may reveal critical adaptive features. The gene families belonging to specific lineages are also tied to the fundamental life cycle of *Plasmodium* species in that particular region. These genes of interest are involved in understanding pathogenesis and virulence mechanisms [11].

1.5 The Pan-genome

The pan-genome is the global gene repertoire or entire gene content pertaining to a species [29]. It is referred to as non-redundant homologous gene clusters found in a taxon [12], [30]. Phylogeny (post speciation gene content) and respective habitat(s) potentially shape a pan-genome of a species [30].

The Pan-genome concept was first introduced by Tettellin *et al.* In his study, he sequenced the complete genomes of different strains from major pathogenic serotypes of group B *Streptococcus agalactiae.* With thorough research containing six newly sequenced strains and two secondary sequences, it was concluded that eight genomes are not sufficient to study an entire gene repertoire. The mathematical models presented in the study found out that additional genes continue to be added in the gene pool no matter how many strains are added in the pan-genome. For every new sequence added, 33 new genes that are strain-specific on average were added to the gene pool, thus making the core genome only a fraction of the entire genome. Therefore, this theory nullified the concept presented in an earlier study on six strains of Streptococcus agalactiae by (Konstantinos T. Konstantinidis and James M. Tiedje, 2004) core genome carries almost 80% of any single genome, which showed variability within a bacterial species to be limited. Therefore, it was concluded that

its pan-genome could only study a bacterial genome as per its genetic variability [31]. Figure 1.2 depicts the pan-genome.

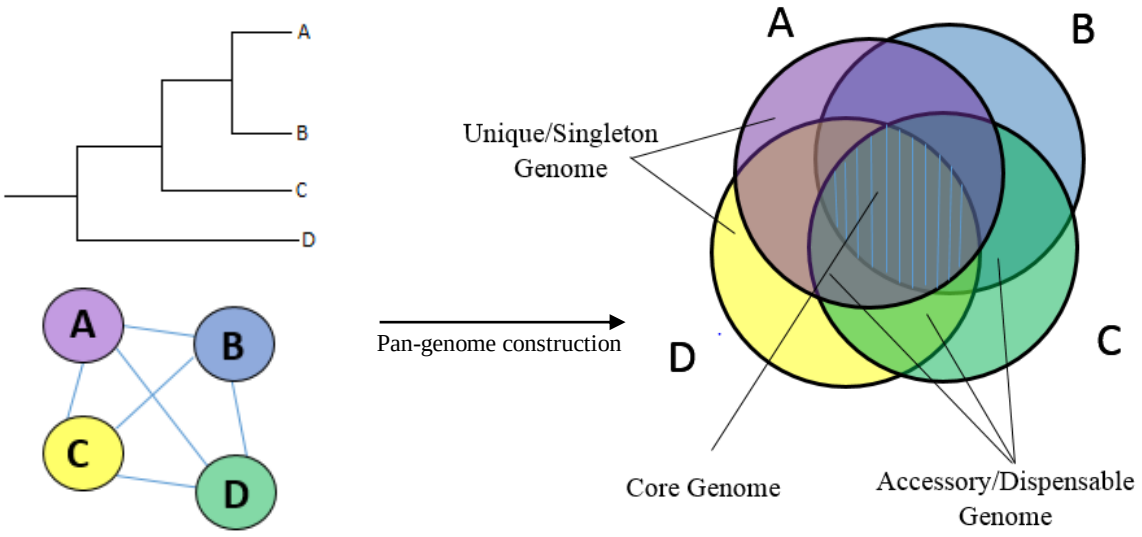

Figure 1.2: Four sequenced genomes A, B, C and D are related to each other, as shown in the phylogenetic clades from which the pan-genome is constructed. The pan-genome consists of a core, accessory and unique genomes. The core genome is a set of genes shared by all the genomes depicted in the centre lined region; the accessory/dispensable genome are the genes shared by some isolates and not by others, and the unique genome consists of strain-specific genes that are not shared by any genome under consideration[32]**.**

In general, a pan-genome can be divided into three parts, as described below.

1.5.1 Core Genome

The core genome consists of all the genes conserved and shared by all the isolates. The genes that lie in the core region of the genome are significant for the regulation of essential aspects of biology and generally relate to cell replication, protein translation and homeostasis balance of the cell [33]. Drastic changes in the core genome are inhibited because it undergoes substantial selective pressure concerning its function. Therefore, limited genetic diversity of the species of interest would lead to the number of core genes shared by all strains being higher, making a larger core genome. On the other hand, if the dataset considered is genetically more diverse, the number of shared genes would be less, and the core genome would become smaller [34]–[36].

1.5.2 Accessory or Dispensable Genome

The accessory genome is a set of genes present in some isolates but not in all of the isolates under consideration. The accessory genome consists of explicitly functional genes that may help survive the species in different niches. These genes are usually linked to antibiotic resistance and virulence. The accessory genes might be similar at the nucleotide level, but they are highly specific for their substrates. Horizontal gene transfer might be the source of these genes' emergence [33].

1.5.3 Unique or Singleton Genome

The unique genome consists of genes specific to strains, which lie in one single isolate. These are usually acquired by the transference of genes horizontally among species [33]. The strains that contain these genes show an adaptive benefit over those strains that do not possess them. These genes might be associated with virulence in pathogenic organisms, whereas they establish a connection with metabolism by being metabolic islands in non-pathogenic organisms. In contrast to the core genome, constant mutations occur in this set of genes as the mutational pressure is relaxed. When mutations have occurred successfully, they raise the adaptation of the organism to specific environments and conditions. The majority of the strain-specific genes are paralogous genes, as observed in a study (Jordan I.K. *et al.*, 2001). These genes encode surface-exposed proteins, which by their ability to confer on pathogenic bacteria and binding capability to cell hosts, they are considered virulence factors. [33].

1.5.4 Open and Closed Pan-genome

In order to know how many sequenced genomes may be required to acquire the complete gene repertoire of a given species, it is imperative to decide how many additional genes are to be added for each new genome that is sequenced. This leads to the concept of open and closed pan-genomes. Figure 1.3 is a pictorial representation of open and closed pan-genomes. The mathematical extrapolation of the *S.agalactidae* pan-genome data used by Tettelin and colleagues discovered that even after the addition of hundreds of sequenced genomes, the unique genes will always keep on being identified. Such a case is an open

pan-genome, where each new genome sequenced will provide novel genes [31]. This seemingly unbounded gene pool is created by mathematical extrapolation from existing strains. Thus, open pan-genomes depict that the genetic content of some species is very flexible. On the other hand, species living in an isolated *niche* tend to lack the ability to acquire foreign genes because they lack recombination and gene exchange mechanisms. This leads to species having a closed pan-genome [33]

1.5.5 Advantages of Pan-genome

The features of pan-genome that are desirable include 'completeness' by which enough functional elements are considered to serve as a reference to be used in analysing additional individuals; 'stability' by which discriminating and unique features can be explored again by any researcher at any time; 'comprehensibility' by which all structural complexities of the genome across entire species could be studied and 'efficiency' by which data can be organised so that the downstream analysis is more effortless and effective. Pan-genome can be used to characterise genes by individual gene sets by detecting virulence factors present in only one particular strain, develop targeted vaccines against pathogens, studying the horizontal gene transfer along with its impact on the evolution and exploring the impact of host and environment on variability in population genomics studies [38].

With the advent of rapid and cheap next-generation sequencing (NGS), the number of sequenced genomes is increasing steadily, making scientists reconsider the concept of reference genomes. The reference genome could take several forms, including considering either a single individual's genome selected on some criteria or a consensus drawn from an entire population as a reference. An alternative idea could be to consider and include all mutations in the form of a functional genome or a maximal genome that contains all sequences ever studied. However, many earlier references did not represent these points; instead, earlier reference genomes were sequence patches collected from different assays, often from unstructured individual biological sources.

To exploit the full advantages of recently determined complete genome sequences, an ideal 'reference genome' must-have capabilities afar from the aforementioned problems. The pan-genome can thus be used as a better reference standard representing the entire genomic content of the species [38].

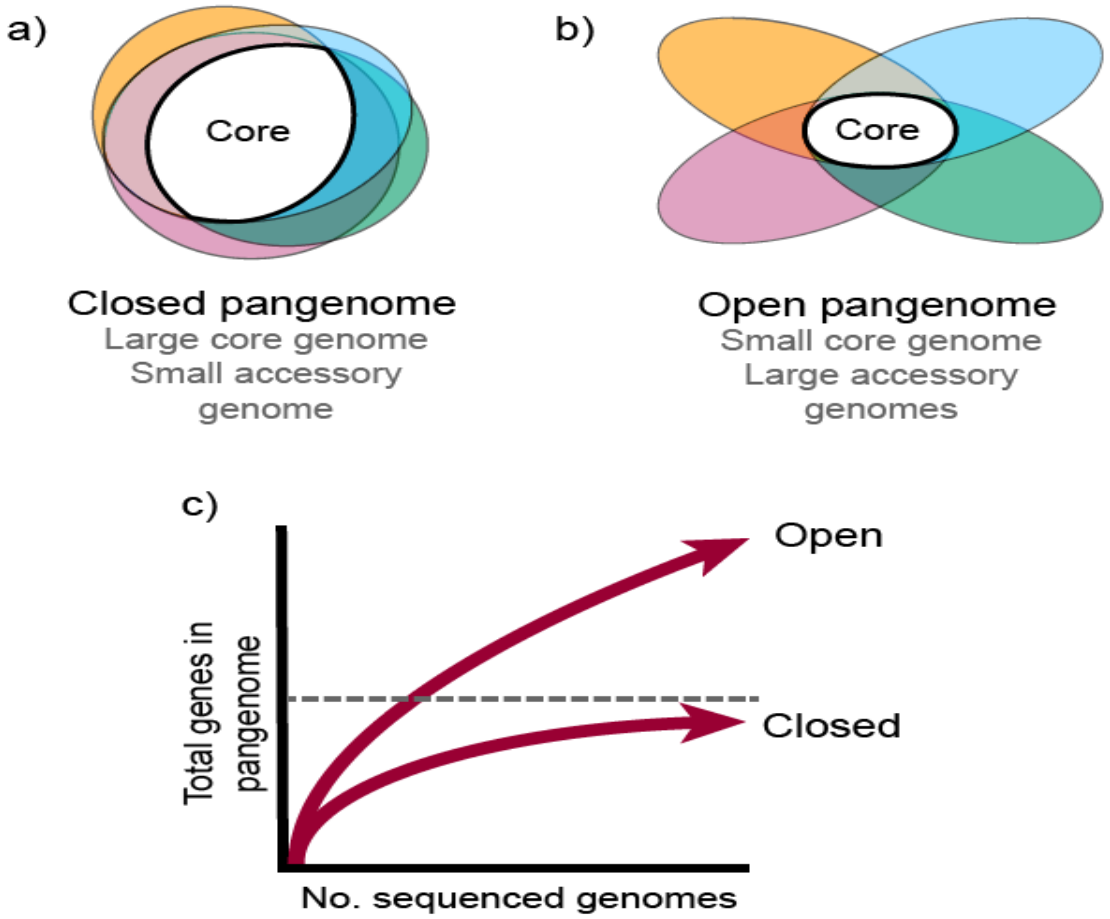

Figure 1.3: (a) Closed pan-genomes are characterised by a larger core and smaller accessory because, after some sequenced strains, additional strains do not provide new genes to the species pan-genome (b)Open pan-genomes are characterised by a small core genome because the number of genes of the pan-genome increases with additionally sequenced genomes. (c) The pan-plot of open pan-genome is infinite when extrapolated, whereas for closed pan-genome, the plot is finite, and pan-genome can be predicted (Creative Commons License Wikimedia)

In summary, computational pan-genomics has led the world to enter an era promising to resolve significant gaps in global maps of genomes and exploit their genetic variability. As a result, it is believed that the mid-term future will witness encompassing insights about nature, pace and extent of evolution.

This study is aimed to model and construct the pan-genome (or supra genome) of *Plasmodium* strains origin wise (i.e. among different continents around the world). The pan-genome analysis would enable the extraction of the core viral proteins common to all isolates, representing potential common virulence factors and potential therapeutic targets.

In an attempt to fight this endemic and identifying the viral proteins particular to the common core and Pakistani strains, this study is a subsequent contribution.

1.6 Problem Statement

A reference/single representative strain does not reflect the entire genetic heterogeneity of a species due to different geographical origins. A comprehensive analysis of the pan-genome of *Plasmodium* species has yet to be carried out.

A complete pan-genome study of *P. falciparum* and *P. vivax* is required by considering genomic strains distributed geographically worldwide to evaluate common core viral proteins and strain-specific novel genes that may account for the total genetic variability of the species.

1.7 Objectives

- To elucidate comparative genetic diversity of *Plasmodium* species region-wise.
- To model and analyse three complete pan-genomes of the species (Global, Asian, Asian excluding Pakistan's closest neighbour India).
- To investigate the common core viral proteins and unique strain-specific viral proteins present in the pan-genome.

Literature Review

This chapter discusses the concepts and earlier published studies that led to the formulation of the current research to answer the research questions and achieve objectives.

2.1 Prevalent *Plasmodium* Strains in Pakistan

The focus of this study was to perform pan-genome analysis on the strains of *Plasmodium* species that were prevalent within Pakistan. Each *Plasmodium* species and its interaction with the host is determinant of the pathogenesis and severity of the disease [5], [12]. *P. falciparum* and *P. vivax* are the only prevalent types of *Plasmodium* in Pakistan [8], [14]. *P. falciparum* is responsible for the most number of deaths globally due to its severe clinical forms. The second most prevalent species is the *P. vivax,* which also has the broadest geographical distribution.

In order to check the most prevalent types of *Plasmodium* causing malarial disease in Pakistan, several regional based and countrywide studies were carried out. *Aamer A et al.* undertook a comprehensive survey using blood samples from 801 febrile patients belonging to all age groups. Government and private facilities from 25 different cities of the four provinces and a hospital from the capital participated in the survey. The result showed that 18% of the malarial cases were of *P. falciparum,* whereas 76% were attributable to *P. vivax,* whereas the remaining 6% were mixed of both the species. None of the cases for *P.malariae* and *P.ovale* was detected in Pakistan [39].

Another retrospective study was carried out by *Bin et al.* containing 356 adults hospitalised with malaria in Pakistan. The only two types of *Plasmodium* infecting these patients were *P. vivax* (83% cases) and *P. falciparum* (13% cases). To compare the severity of the disease caused, it was concluded that 79.9% of the patients having a severe form of malaria in Pakistan were infected with *P. vivax* [14].

2.2 Genomics of *Plasmodium Falciparum* and *Plasmodium Vivax*

Among malarial parasites, the first whole genome to be sequenced was *Plasmodium Falciparum* using Sanger sequencing in seven years-long projects spanning 1995 to 2002

(Foster, J. *et al.*, 1995; Gardner, M.J. *et al.*, 2002; Hocking, 2020). Next, in 2008, the *Plasmodium Vivax* genome was sequenced using the Sanger method. (Gardner, M.J. *et al.*, 2002; Pain, A. *et al.*, 2008; Hocking, 2020). Finally, several years down the lane, approximately 3000 genomes of *Plasmodium Falciparum* have been sequenced under the *MalariaGEN Pf3k* project (*https://www.malariagen.net/projects/pf3k*). Some of the researches utilising large scale sequencing include population structure analysis (Assefa, S. *et al.*, 2015; Hocking, 2020), classification of loci that are drug-resistant (Miotto, O. *et al.*, 2013) and development of biological understanding of the species along with its evolution (Volkman, S.K. *et al.*, 2007; Manske *et al.*, 2012).

Genome sequencing of *P. falciparum* has revealed around 5400 novel open reading frames (ORFs) compared to only 20 proteins characterised earlier [11]. The reference genome of *P. falciparum 3D7* was completely sequenced in 2002, containing ~5300 protein-coding genes (Gardner, M.J. *et al.*, 2002). Notably, the canonical reference of *P. falciparum 3D7*, first published in 2002, was 22.9 Mb with 5268 genes on 14 chromosomes. It contained approximately 80 gaps (Gardner, M.J. *et al.*, 2002). However, this study used the latest improved version. The genome was recently improved to a current version (v3.0), having 5369 genes with no gaps. The new version has re-assembled some chromosomal regions of chromosomes 7,8 and 13 and has re-annotated it comprehensively (Bowman *et al.*, 1999; Gardner, M.J. *et al.*, 2002; N. Hall M. *et al.*, 2002).

P. vivax, being a significant pathogen in humans, is still understudied because it can only be transmitted in non-human primates rather than continuous *in vitro* culture. The reference strain of *P. vivax Salvador 1* was sequenced using whole-genome shotgun methods having a genome of 26.8 Mb. The contigs were large and assigned to 14 chromosomes having a size of approximately 22.6 Mb. About 4.3 Mb was unassigned because of repetition. *P. vivax* chromosomes have a unique feature of isochore boundaries where the sub-telomeric regions have a low G+C content as compared to the internal regions that exhibit a high content of G+C (Carlton, J.M., *et al.* 2008). Considering history, *P. vivax* is known to reduce the average lifespan from 58 to 33 years in the nineteenth century [51].

2.2.1 Comparison of *P. vivax* and *P. falciparum*

P. vivax has better tolerance of cooler climates as compared to other *Plasmodium* species and also has lengthy remission, due to which the risk of *P. vivax* infected malaria is estimated to target half of the world's population [51]. Even though the current mortality rate is higher for *P. falciparum,* evidence suggests *P. vivax* be the more virulent parasite. Furthermore, as per the recent studies, severe malaria syndromes are now being attributed to *P. vivax,* which were earlier known to be only for *P. falciparum* [51].

As per the study published by *Neafsey et al.* in 2012, newly sequenced *P. vivax* strains were compared to *P. falciparum* strains to know the genetic diversity. Table 2.1 compares the nuclear genome features of the two species. It is known that SNP rates, when compared to the reference, is a function of evolutionary or geographic distance; however, the SNP diversity rates in *P. vivax* were uniformly twice as much higher than *P. falciparum* across the genome. This reflects that *P. vivax* is stable in its demographic history and uniformly colonises globally. On the other hand, *P. falciparum* is shown to have undergone a significant population bottleneck where its size of the population is reduced drastically in recent history [51]. Furthermore, there is an additional complication of a dormant liver stage in *P. vivax,* which makes the reactivation of the species possible even in the absence of a mosquito bite, thus owing to its disproportionate demographic stability. [5], [12], [51].

Table 2.1 Comparison of nuclear genome features of the two Plasmodium species [52]

Feature		***P. falciparum***	***P. vivax***
Genome	Size (Mb)	23.3	26.8
	Number of chromosomes	14	14
	Coverage	5	10
	G+C content % AT content %	42.3% 80.6%	19.4% 57.7%
Genes	Average Number of Genes	5403	5433
	Mean Gene Length	2283	2164
	% coding	53.9%	52.1%
RNAs	Number of transfer RNA genes	43	44
	Number of 5S ribosomal RNA genes	3	3
	Number of 5S ribosomal RNA genes	7	7

2.3 Antimalarial Drug Resistance

As discussed in chapter 1, section 1.3.1, antimalarial drug resistance is a recurring issue that exhibits an increasing concern for researchers to treat and eliminate malaria [15]. From Chloroquine resistance to Artemisinin-combination therapies (ACT), the first-line antimalarial treatments improve, yet they develop resistance. Table 2.2 lists the first-line treatments region wise for *P. falciparum* and their treatment failure rates, and Table 2.3 lists species-wise treatments failures in Pakistan as per the WHO Malaria Threats Map.

Table 2.2 Species-wise treatments in Pakistan having failure rates in multiple studies as per WHO Malaria Threats Map(https://www.who.int/malaria/maps/threats/;retrieved 31st May 2021)

Type of *Plasmodium* infection in Pakistan	**Treatment Drugs**	**Follow Up Days**	**Study Years**	**No. of studies**
Plasmodium Falciparum	Artemether-lumefantrine (AL)	28	2012-2019	6
	Artesunate+sulfadoxine-pyrimethamine (AS+SP)	28	2011-2017	6
	Dihydroartemisinin-piperaquine	42	2015-2015	2
Plasmodium Vivax	Dihydroartemisinin-piperaquine	28	2013-2013	1
	Chloroquine	28	2013-2013	1

Table 2.3 First-line treatments currently available for malaria in WHO Regions with their failure rates [6]

WHO Region	First-Line Treatment	Treatment Failure Rate
WHO African Region	Artemetherlumefantrine (AL), Artesunate-amodiaquine (AS-AQ) and Dihydroatemisinin-piperaquine(DHAPPQ)	More than 10% observed in four studies
WHO Region of the Americas	Artemetherlumefantrine (AL), Artesunate-mefloquine (AS-MQ) and Chloroquine (CQ)	10.4% in one study of CQ from Bolivia
WHO South-East Asia Region	Artemetherlumefantrine (AL),Artesunate-sulfadoxine-pyrimethamine (AS+SP) and DHA-PPQ.	Exceeded 10% in three studies, one in Thailand and two in Bangladesh.
WHO Western Pacific Region	Artemetherlumefantrine (AL), Artesunate-amodiaquine (AS-AQ)	10%
WHO Eastern Mediterranean Region–	Artemetherlumefantrine (AL), Artesunate-sulfadoxine-pyrimethamine (AS+SP)	

2.3.1 Important Genetic Deletions and Mutations causing Treatment Delay

Rapid Diagnostic Tests (RDTs) have been widely used to detect *Plasmodium Falciparum* infections to diagnose malaria. When the infection is established, many parasite proteins are released into the blood of the host. Among these proteins are the antigens, namely *Plasmodium Lactate Dehydrogenase* (pLDH), *Plasmodium Aldolase* (aldolase), and *Plasmodium Falciparum-specific Histidine Rich Protein 2* (HRP2) that are targets for RDTs. The pLDH and aldolase can detect all human malaria, whereas the HRP2 is *Plasmodium Falciparum* specific. Sometimes the antibodies of HRP2 may react with HRP3 antigen since it has very similar amino acid sequences and epitopes. HRP2 is encoded by the *pfhrp2* gene located on chromosome 8, whereas the HRP3 antigen is encoded by the *pfhrp3* gene located on chromosome 13 [53]. Peru reported genetic deletions of *pfhrp2* or *pfhrp3* or sometimes both genes in naturally occurring populations of *P. falciparum* [54]. Over time, many of these deletions have been reported in other countries and negatively impact RDTs being used as diagnostic tools [55]. This leads to a lot of false negatives leading to increased morbidity, mortality and

transmission. Suppose the false-negative HRP2 based RDTs prevail for more than 5%. In that case, it is recommended to shift to alternative pLDH based RDTs or microscopy, which are less sensitive and have inferior performance for the detection [56]–[58].

Even though the results with ACT were pretty impressive, with rapid clearance of parasites having a profound impact on the control of malaria, resistance to Artemisinin has now developed in *P. falciparum*. (Bhatt, S. *et al.*, 2015). Multiple point mutations in the *Kelch13* gene affect the propeller region of the protein associated with Artemisinin resistance (Ashley, E.A. *et al.*, 2014). As per World Malaria Report 2020, the *PfKelch13* mutations are identified as molecular markers of Artemisinin resistance (World Health Organisation, 2020).

2.4 The Quest for Malarial Vaccine

As described by World Health Organisation (WHO), a vaccine is the most cost-effective measure to eradicate malaria worldwide [21]. The World Malaria Report 2020 mentions the development of malaria vaccine as part of routine control efforts to be “The next major innovation” and “a new paradigm” in malaria control (World Health Organisation, 2020). WHO's earlier focus used to be on malarial control rather than eradication, yet efforts to develop a vaccine for malaria began in the 1970s during the Vietnam war, hoping to get complete protection from malarial infection for the military. A pioneering study reported the vaccine candidates to be sporozoite antigens however the vaccine was not completely blocking the infection [59], [60].

In 1997, the leading candidate RTS, S, showed partial efficacy in humans. However, the RTS, S was based on the surface protein of *P. falciparum* known as circumsporozoite, which was pre-erythrocytic. It was then believed by *Saula et al.* that more successful vaccines could be made by targeting other antigens involved in the blood stages of parasites rather than being pre-erythrocytic [61]. Thus, by the 1990s, only the *SPf66* vaccine reached a phase III trial in children, among other candidates [62].

Some other vaccines were developed in the subsequent years that used a combination of sporozoite antigens as in the pioneering study along with the blood stages as the new studies had revealed. However, these vaccines had mixed results [63]–[65]. Some further studies on DNA vaccines having many antigens were also conducted, but the results were disappointing [66].

A practical approach to reducing malarial burden was to use RTS, S (most advanced candidate until now) as a paediatric vaccine. This led to partial protection in many controlled trials [67], [68].

In 2015, RTS, S combined with the AS01 adjuvant, was used in children's most extensive clinical study. Around 16,000 children in seven African countries were administered with RTS, S/AS01 vaccine, which showed the clinical efficacy after following up for 32 months to be moderate.

As per the position paper released by WHO in 2016, the European Medicines Agency (EMA) has given a positive recommendation to the world's first malaria vaccine, which is undergoing a phased pilot introduction in Kenya, Ghana, and Malawi starting from 2019. About 500,000 people have already received their first dose of vaccination. However, the RTS's/AS01 vaccine is a further ongoing evaluation to assess the vaccine's public health value, which will be reviewed in late 2021 by WHO (World Health Organisation, 2020).

There are limitations to the vaccine. Firstly, it can be observed that the vaccine's efficacy declines over time along with the rapid waning of antibody levels [23].

Secondly, Since The RTS, S/AS01 vaccine is based on genetic sequences from *Plasmodium Falciparum* strain *3D7*, it shows more efficacy against sequences that match the vaccine's protein allele. However, a study in Africa showed only <10% of infections matching the protein allele of the vaccine. This consequently led to the loss of efficacy against divergent parasites. Therefore, after the phase III trials, it was concluded that multiple factors affect the efficacy of RTS, S/AS01 such as: [22], [23], [67].

- Efficacy increases with an additional booster dose.
- Efficacy is more in older children as compared to infants.

- Efficacy is short-term and wanes over time.
- Efficacy is less in populations where transmission intensity is high.
- Efficacy is low with genetically diverse parasites [22], [23], [67].

Thus, there is a need to study the entire genetic complexities of the species based on regions to understand the core viral proteins that could help develop an effective vaccine target against all strains. For this, purpose the pan-genome concept comes to the rescue.

2.5 Natural Selection and Adaptation of Parasite due to Environmental Factors and Mutations

Strains belonging to the same species vary considerably in their gene repertoire due to host and environmental factors. In addition, adaptation to a new environment and natural selection occurs due to mutations such as insertions, deletions, single nucleotide polymorphisms, copy number variations, and microsatellites. Thus, mutations can be used to deduce genomic variability within and between species [11], [69], [70].

There is evidence from studies that provide insight into the impact on parasites at a molecular level due to a changing environment and geographical location, which are essential keys to eradicating malaria in the future.

For instance, a comparative analysis study was performed where Malaysian *P. vivax* strains in the pre-elimination stage were compared with 200 *P. vivax* songs taken from areas where transmission remains high. The objective was to check the reaction to natural selection prompted by near elimination. The analysis results showed multiple drug resistance loci in pre-elimination stage Malaysian samples, leading to their natural selection for elimination. Also, a large cluster of a particular strain could be seen rapidly involved in the clonal expansion. This indicates that mutations and SNPs as per geographical origin in different strains determines elimination [70].

For this reason, the in-depth comparative analysis of genomic data from multiple strains of a species must be studied to answer general questions about the fundamental processes such as pathogenesis, development of genetic variability in response to environmental

challenges, drug resistance as a mechanism of adaptation, and evolution to determine the minimum set of essential genes and common virulence factors. Pan-genome analysis can achieve those as mentioned above.

2.6 From Single Reference Genome to Pan-genome

As mentioned in Chapter 1 section 1.3.3, since diversity between the species can be significant enough compared to intra species variety, it is not sufficient to consider a single isolate as a reference genome to describe an organism's entire genetic complexity and diversity of species. Thus, a pan-genome being the whole genetic repertoire of the species must be used to define a phylogenetic clade [27], [40].

Analysis of the pan-genome (Refer to chapter 1 section 1.5) of species where its conserved core genome shared among nearly all members of the species, its accessory genome shared by some and absent in other strains of the species, and its unique genome shared with no other members of the species can be considered as a reference/representative of the entire genetic heterogeneity of the species and thus can answer a lot of previous biological queries.

The pan-genome can be open or closed, as discussed in Chapter 1 section 1.5.4. It depends upon the lifestyle of the species under consideration. For example, the allopatric species that live alone have a closed pan-genome because they are specialised as compared to the sympatric species that interact within a community and possess an open pan-genome along with high rates of horizontal gene transfer many ribosomal operons [69], [71].

2.6.1 Analogous Concepts: Pan-genomics vs Pan-metabolism vs Pan-regulon

Pan-metabolism refers to the metabolic reactions of all the strains in consideration, whereas the core metabolic reactions are those common to all the strains. Therefore, a study was carried out containing 29 species to determine the core and pan-metabolism of E.coli [72].

Results showed 1545 reactions in the pan-metabolism and out of which 885 metabolic reactions belonged to the core. The pan-metabolism distribution was independent of the proportion of core genes and pan-genome being open or closed. *E.coli* had a very low number of core genes as it had an infinite pan-genome but, on the other hand, had a large number of core metabolic reactions. Thus, Metabolic level diversity is higher than gene level. [69]

Another common analogy to pan-genome is pan-regulon. Pan-regulon contains all the genes that are controlled by a particular transcription factor in the genomes under consideration. For example, a study was carried out on *Sinorhizobium meliloti,* where its pan-genome and pan-regulon was considered. The core genome had 5124 genes, and the pan-genome had 7824 genes, whereas its pan-regulon was extremely small compared to its pan-genome [69].

2.6.2 Number of Strains

With a purpose to determine the required number of genomes to comprehensively analyse a pan-genome where it defines the totality of the genomes, a study was conducted where the ratio of core to the pan was compared among 27 bacterial species as a function of the number of genomes [69].
The results showed:

- For a very closed pan-genome, two strains may be sufficient.
- For a closed pan-genome, six strains may be sufficient.
- For a large open pan-genome, ten strains may be sufficient.
- For an infinite open pan-genome, it is not possible to close it by definition.

Another study by *Vernikos et al*. recommends at least five genomes to be the minimum number of genomes to be analysed for pan-genome analysis [73].

2.6.3 Types of Strains

It is crucial to select the types of strains carefully. If the species is a pathogen, some clinical aspects need to be studied since even different strains of the same species may cause different diseases, for example, in *E. coli*. [74].

The second important point is considering different geographical origins because these relate to the genotypes. The third criteria are to include strains having a phenotype such as antibiotic resistance and stress. By selecting strains based on these criteria, a larger panel describing the pan-genome diversity well could be selected [69].

2.7 Bioinformatics Tools for Pan-genomics

2.7.1Composition and Annotation

The composition of the pan-genome is estimated by orthologue search, which can be done by various methods such as BLAST or OrthoMCL. The annotation of the core, accessory and unique genome can be done by several tools such as Cluster of Orthologous Groups (COG), InterPro and KEGG (Kyoto Encyclopaedia of Genes and Genomes). In addition, metabolic pathways, transcription factors, regulation of protein expression are looked into for annotation. The Predicted Prokaryotic Regulatory Protein is another online tool for searching such proteins.

2.7.2 Alignment and Phylogeny

Plenty of tools are available for alignment purposes. For example, MAUVE might do a global alignment for comparison purposes, or multiple alignments for phylogeny construction might be done by CLUSTALW, MUSCLE or any other program. In addition, MEGA or MAFFT can do phylogeny reconstruction with an option of various algorithms to be used from neighbour joining, maximum parsimony or maximum likelihood.

2.7.3 Dedicated Pipelines

Many automated tools have now been developed to carry out the pan-genomic analysis. A tool Pan-genome analysis Pipeline (PGAP) is an automatic pipeline that runs through a

single line of command analysing gene functions, pan-genome profile construction, species evolution etc. For example, PGAP was used on *Streptococcus pyogenes.*

Another tool, PANSEQ (Pan-genome Sequence Analysis Program), is an online tool that helps to analyse unique and core regions of the pan-genome along with the SNPs in the core or accessory genome. One of the first comprehensive attempt pipelines to analyse eukaryotic pan-genomes is EUPAN that benchmarks the 453 rice genomes dataset. [75]. More on pan-genome pipelines and tools can be studied from [33].

2.8 Pan-genome Approaches

The reverse approach of pan-genome captures the genomic diversity of a group of interest. In contrast, the forward-thinking approach can be implemented to explore the minimum number of genomic sequences required to describe the totality of the group. The following three different approaches have been commonly used so far for pan-genome development.

2.8.1 De Novo Assembly

After the high depth sequencing and *de novo* assemblies of all the targeted accessions, the assemblies are compared one by one for conserved or variable regions without using a reference genome. The approach had been used in pan-genome construction in plants using SOAPdenovo, ALLPATHS-G and ABySS assemblers [76]. The De Novo assembly provided the physical position of genomic elements. However, assembly or annotation was often erroneous, thus confounding the actual differences between individuals [77]. Figure 2.1 depicts the *De Novo* approach.

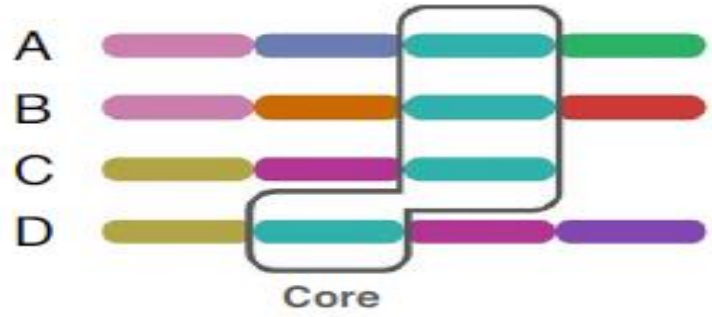

Figure 2.1: Each genome's original sequence is reconstructed by the sequencing reads. The variable segments and shared segments in the genome are shown in different and same colours, respectively. De novo assembly reconstructs each genome separately, allowing the genomes to be compared directly to find the core and variable regions. [77]

2.8.2 Reference Based Assembly and Iterative Mapping

This approach begins with a single reference genome and adds on non-redundant sequences in succession to build a pan-genome. The benefit is that the sequences can be from other individuals that might even have a lower sequencing coverage that is pooled up in large numbers. Presence Absence Variation analysis is then conducted by comparison and mapping to the constructed pan-genome. The limitation of this approach is that it causes issues while assembling highly repetitive sequences and considerable structural variations [77]. For example, mapping the resequencing data from strains of different species has used this approach [76]. Figure 2.2 shows the iterative mapping approach.

Figure 2.2: Figure shows the iterative approach in which a single reference is used along with all non-redundant sequences being added one by one. Those non-redundant sequences that cannot be added directly are first assembled by de novo assemblers and then added to make up the pan-genome. Choosing a different reference genome would place the genomic segments differently, as shown when Pan-genome A Vs Pan-genome E are used as references[77].

2.8.3 Graph and *k*-mer

A set of *k*-mer sequences resulting in a de Bruijn graph can be used to represent a pan-genome. The de Bruijn and string graphs assemble the genome via assemblers where graphs represent regions where chromosome varies. The pan-genome is thus created by a coloured graph that depicts multiple genomes with all of their variations and confiscates the contents that are non-redundant among the genomes under consideration. The tool using the de Bruijn graph approach is SplitMEM. The approach was used earlier in multiple studies of prokaryotic pan-genomes. However, since large sequences create a vast amount of vertices and a very large graph, using this approach for constructing complex eukaryotes' pan-genomes is limited and highly computationally intensive [76]. Figure 2.3 shows the constructed de Bruijn graph approach.

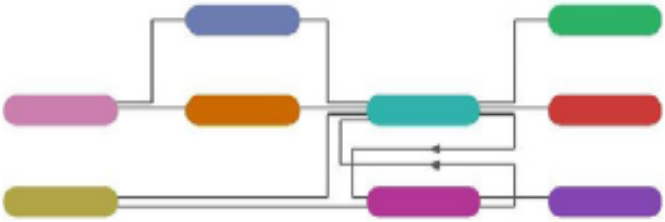

Figure 2.3: Figure shows the constructed pan-genome using De Bruijn graphs. By tracing back the paths on the De Bruijn graphs, the relationships within the genomic segments can be known. In recent years, hybrid approaches are being used to overcome the limitations of the respective pan-genome[77]**.**

2.9 Earlier Studies using Pan-genome Approach

The Pan-genome concept was first introduced by Tettellin *et al.* In his study, he sequenced the complete genomes of different strains from each of the significant pathogenic serotypes of group B *Streptococcus agalactiae* and found out the pan-genome of the species. Refer to chapter 1, section 1.5, for the pioneering study. Researchers have carried out pan-genome analysis at every phylogenetic resolution level, which exploits various frameworks of modelling. Table 2.4 shows pan-genome approaches being applied at different levels of phylogenetic resolution adapted from (Vernikos. 2020).

2.9.1 Pan-genome study on *Plasmodium* species

Cai carried out a study on core genome components of *Plasmodium*. H *et al*. The study considered six species of *Plasmodium,* out of which two infected humans, one infected monkey and three infected rodents. The core genome of the six species was extracted (3,351 genes) that made up 22% to 65% of the total gene repertoire. The functionality of the core components was further investigated. It was also found out in the results that 5% to 9% of the whole genome contained specific *Plasmodium* lineage radiations. The limitation of this study is that it did not investigate the entire pan-genome containing accessory and unique regions. The study also had only two genomes from humans that were not enough to represent the total genetic variability of the species [11]

Table 2.4:The Pan-genome approaches being applied at different levels of phylogenetic resolution adapted from [78]

Level	Organism	No. of Genomes	Core Size (No. of genes)
Species	*Streptococcus agalactiae*	8	1806
	Neisseria meningitidis	6 20	1337 1630
	Borrelia burgdorferi	21	1200
	Escherichia coli	17	2344
	Enterococcus faecium	7	2172
	Yersinia pestis	14	3668
	Streptococcus pyogenes	11	1376
	Clostridium difficile	15	1033
	Lactobacillus paracasei	34	1800
	Campylobacter jejuni	130	1042
	Campylobacter coli	62	947
	Haemophilus influenzae	13	1450
	Streptococcus pneumoniae	17 44	1400 1666
	Staphylococcus aureus	16	2245
	Moraxella catarrhalis	12	1755
	Lactobacillus casei	17	1715
	Gardnerella vaginalis	17	746
	Clostridium botulinum	13	2657
Group	*Bacillus cereus*	4	3000
	Bacillus subset of species	12	2009
Genus	*Streptococcus*	26 52	600 522
	Prochlorococcus	12	1273
	Bifidobacterium	14	967
	Listeria	13	2032
	Salmonella	35	2811
	Shewanella	24	1878
	Finegoldia	12	1202
Class	*Bacilli*	172	143
Phylum	*Chlamydiae*	19	560
Superkingdom	*Eubacteria*	573	250

Methodology

This study is focused on the pan-genome analysis of *Plasmodium Falciparum* and *Plasmodium Vivax* species from complete proteomes of the strains distributed geographically. With an aim to achieve previously formulated objectives, an integrated analysis was performed. Most of the work was done on a Linux (Ubuntu 18.08) environment on a workstation with 32 cores and 64 Gb RAM. Whereas phylogenetic analysis was carried out on a supercomputing cluster located at Research Centre for Modelling and Simulation, NUST. The complete framework of the study is described in this chapter. The generalised methodology of the study is shown in Figure 3.1.

3.1 Dataset Selection

The study required complete datasets of proteomes of *Plasmodium Falciparum* and *Plasmodium Vivax* strains from different origins. The publically available datasets of whole proteomes were collected from two resources, PlasmoDB and UniProt. Sequences were downloaded in FASTA format. The main dataset was further divided into two additional datasets to conduct the analysis on each separately. Since Pakistan does not have any completely sequenced strains of *Plasmodium* on public databases, this study thus created two additional pan-genomes with its closest neighbour India. The datasets for the three pan-genome analyses were as follows:

1. "Global Dataset" contains all the selected strains worldwide.
2. "Asian Dataset" contains all strains belonging only to Asian countries, including Pakistan's closest neighbour India.
3. "Asian excluding India Dataset" contains all strains belonging to Asian countries, excluding Pakistan's closest neighbour India.

Dataset for *P. falciparum* are shown in Table 3.1, and datasets for *P. vivax* are shown in Table 3.2.

Table 3.1:Datasets selected for P. falciparum.

Continent of Origin	Strain Name	Country of origin
Europe	3D7 NF54N NF54F	Netherlands Netherlands France
Asia	FCH/4 Dd2 CAMP/Malaysia Vietnam Oak-Knoll KH01	Philippines Indochina Malaysia Vietnam Cambodia
Africa	Palo Alto / Uganda MaliPS096_E11 Tanzania (2000708) GN01	Uganda Mali Tanzania Faranah, Guinea
North America	HB3 Santa Lucia	Honduras Salvadoran, Santa Lucia
South America	7G8	Manaus, Brazil

Table 3.2: Datasets selected for P. vivax.

Continent of Origin	**Strain Name**	**Country of origin**
Asia	North Korean India VII PcC01 PvT01 PvP01	North Korea India China Thailand Indonesia
Africa	Mauritania 1	Mauritania
North America	Salvador1	Salvador
South America	Brazil1	Brazil

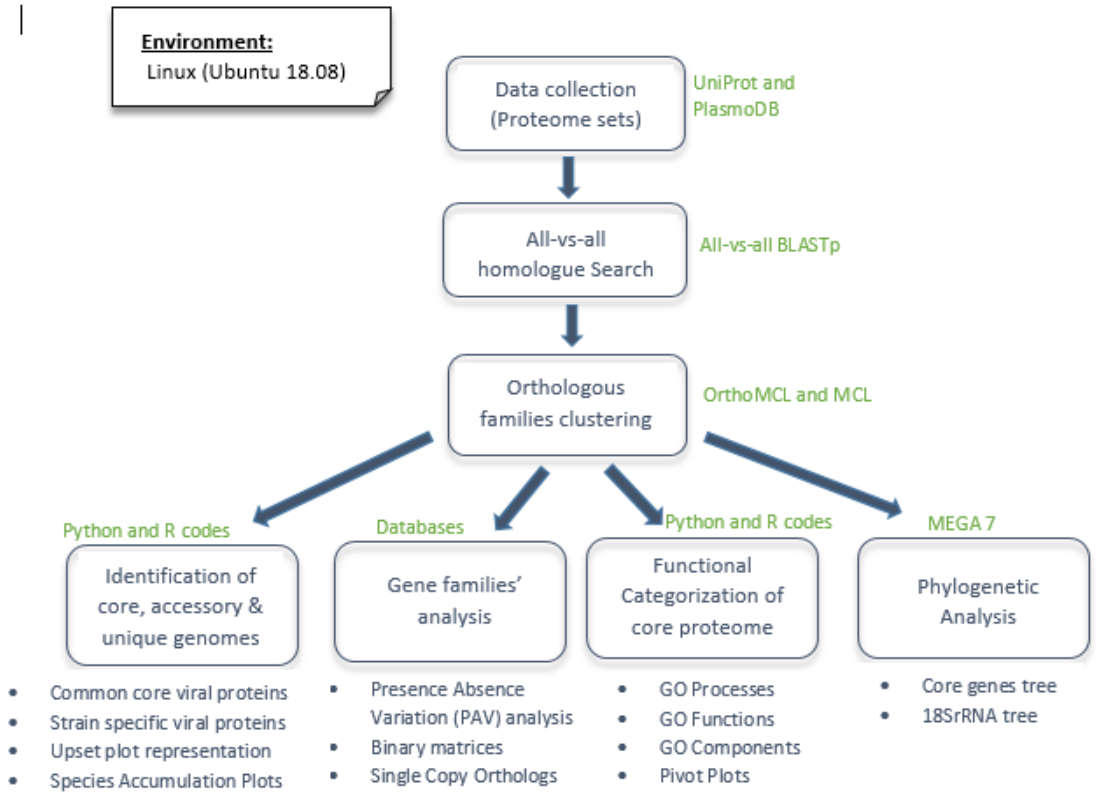

Figure 3.1: Complete Workflow of the study. Steps performed are displayed in blue, and their corresponding tools and software are mentioned in green alongside.

3.1.1 Criteria for Dataset Selection

The whole proteome sequenced datasets considered for this study were selected based on the following:

- Belonged to *P. falciparum* and *P. vivax* strains.
- Had origins from different regions of the world. An extensive literature study traced back the origin of some clones if not mentioned.
- Had the most recent updated version.
- Had Complete Proteome Detector (CPD) values as "Standard" or "Close to Standard."
- Had the highest relative BUSCO (Benchmarking Universal Single-Copy Orthologues) scores compared to other entries of the same isolate.

3.2 Homology Search by all-vs-all BLAST

This study used an extensive all-versus-all blastp approach using a custom made database of *Plasmodium* species of interest using a BLAST+ standalone application (version 2.10.0) [79], [80]. All-vs-all blast uses a large number of sequences to be compared to one another, i.e. having almost an equal size of query and database [81], [82].

All-versus-all BLAST is computationally demanding and almost intractable. Time taken to find all homologous sequences for a database of N sequences would be O(N2) [83]. Another constraint is the amount of data generated. It is estimated that the non-redundant subset of Genbank (NR database) has 6.5 million proteins and thus would produce approximately 1.8 Terabytes of tab-delimited one output containing 37 billion pairwise homology relationships [83].

For each dataset, the following steps were performed to run all-vs-all blastp locally on PC:

1. Configuration of BLAST+

A configuration file named *.ncbirc* specifying key configuration parameters was placed in the current working directory containing the following contents:

- BLASTDB: Specified the path where the custom database was saved.
- DATA_LOADERS: Specified “blastdb” as the data source to be used for the resolution of sequence identifiers automatically.
- BLASTDB_PROT_DATA_LOADER: Specified the “goodProteins.db” as the custom database to be used.

2. Adjustment of FASTA Proteins

For the purpose of this study, modified FASTA files were generated using the method *OrthoMCLAdjustFasta* of OrthoMCL (Version 2.0.9) [84], [85]. This allows the FASTA header to be in the form of >TaxonID|ProteinID where TaxonID is of three or four-letter taxon code and ProteinID is the sequence identifier unique within the taxon. Taxon IDs were manually set as file names to be used by command and saved in a directory “compliantFASTA” containing one file per proteome. The general command is:

```
OrthoMCLAdjustFasta filename file_path 1
```

where 1 indicates the column position of protein ID from original FASTA headers to be considered/kept in the final formatted header.

3. Filtration of Proteins

Poor quality sequences were defined as sequences having less than ten residues or more than 20% stop codons in the FASTA file. They were filtered by the *OrthoMCLFilterFasta* program, keeping default values. This produced a goodproteins.fasta file to run BLAST on. The command used was as follows:

```
OrthoMCLFilterFasta path_to_adjusted_proteins 10 20
```

where

- 10 refers to Min_length, i.e. sequences having at least ten residues
- 20 refers to Max_Percent_Stop, i.e. maximum per cent stop codons

4. Creation of Custom Database

Since the purpose of the study was to find common regions among only *Plasmodium* species using an exhaustive bi-directional blast, a locally created custom database containing only the said species was needed. This helped as lesser disk space was being used to do the thorough search, and shorter runs were being carried out in both directions. *makeblastdb* method is used to create a custom database from a multi-fasta sequence file containing headers in a specific format as modified above. The resultant good proteins from step 3 were combined into a FASTA file to act as the custom database of *Plasmodium* species and named "goodProteinsdb" using the *makeblastdb* method of BLAST+.

```
makeblastdb -in goodProteins.fasta -dbtype prot -
   parse_seqids -out goodProteinsdb.fasta
```

where

- –in specifies goodProteins multifasta file that is adjusted and filtered
- -dbtype mentions the type of database being used.
- -parse_seqids flag parses the IDs of sequences
- -out specifies the output filename

5. Splitting of Query Files

As mentioned in the constraints of Blast all-vs-all, the task is computationally exhaustive and thus time consuming. Splitting of query files reduces the search times significantly shorter allowing smaller data types in the lookup table. Each query sequence was split into files of 1000 proteins by using a FASTA splitter through an *awk* command.

```
awk 'BEGIN {n_seq=0;} /^>/
{if(n_seq%1000==0{file=sprintf("queryfile.fa",n_seq);}
print >> file; n_seq++; next;} { print >> file; }' <
queryfile
```

where

- n_seq%1000 defines 1000 sequences to be splitted in one file
- queryfile.fa defines the name of queryfile in FASTA format to be split.

6. Optimisation of the Command

Further optimisation of the BLAST run command was done to get faster and better results by using the following:

- -Num_threads 10: Parallel commands were run on ten threads of the CPU core for each split file.
- -outfmt 6: Tabular output format was chosen by selecting its key number ‘6’.
- -word_size 7: Initial search word size for seeding was increased to seven.
- -evalue 0.0001: Since e-value depends upon the size of the database being searched, it was chosen to be 0.0001

7. Merging of the Results

Since the query files had been split into chunks, the results of one strain were present in multiple notepad files. Also, the input format required to be loaded in the upcoming OrthoMCL database needed to be a single file of all blast results combined from all strains. Thus, all tab-delimited results from the split files of strains were concatenated into a *.tsv* file to act as input in upcoming steps by the bash script:

```
cat *final.tsv >> blastresults.tsv
```

The steps were repeated for each split query file against their goodProteinsdb to generate the three datasets mentioned earlier.

3.3 Orthologue Analysis and Protein Families' Clustering by OrthoMCL and MCL

In this study, standalone OrthoMCL v2.0.9 [84], [85] was used on a Linux machine to perform the analysis. Standalone OrthoMCL is used if the orthologue group analysis is to be carried out on at least two complete proteomes and up to hundreds. This software processes the proteins in four major phases. [84]

The following steps were carried out to complete clustering:

1. Configuration of OrthoMCL

A configuration file named *orthomcl.config* specifying key configuration parameters (percentMatchCutoff=50, evalueExponentCutoff=-5) alongwith enlisting table names to be formed in the database (similarSequencesTable=SimilarSequences, orthologTable=Ortholog, inParalogTable=InParalog, coOrthologTable=CoOrtholog) was placed in */etc* directory.The program *OrthoMCLInstallSchema* was then run to configure OrthoMCL and create required enlisted data tables in the database.

```
sudo orthomclInstallSchema /etc/orthomcl.config
install_schema.log
```

2. Formatting of Tab Delimited Format

The final merged FASTA output file generated by BLAST+ was in a tab-delimited format that required conversion into a format compatible for loading into the relational database. For this purpose, the *OrthoMCLBlastParser* program was used, which also computes Percent Identity and Percent Match. The command used was:

```
OrthoMCLBlastParser blastResults.tsv compliantFasta >>
similarSequences.txt
```

where

- blastResults.tsv is the output from BLAST+
- compliantFasta is the folder containing original strains files
- similarSequences.txt is the output file

3. Loading BLAST into Relational Database

The relational database used was MySQL. Due to the size of data to be loaded, the capacity of the relational database (MySQL) was increased during configuration using:

```
set global innodb_buffer_pool_size=53687091200;
set myisam_sort_buffer_size=34359738368;
```

OrthoMCLLoadBlast program was then used to load blast results into the relational database. For this command to work, the local-infile option has to be enabled, or it would give an error message "Used command is not allowed with this MySQL version". To prevent it, a client.cnf file was created in */etc/mysql/conf.d* with the following content:

```
[client]
loose-local-infile=1
```

Next, the SQL server was restarted for the change to take effect by:

```
sudo service mysql restart
```

Finally, the main command used to load blast results on the relational database was:

```
OrthoMCLLoadBlast /etc/OrthoMCL.config similarSequences.txt
```

4. Finding Orthologues, In-paralogs or Co-orthologue Pairs

OrthoMCLPairs program was used to find pairs of proteins that were potentially orthologues, in-paralogs or co-orthologues. This function is a series of 20 internal steps, each creating intermediary database tables and thus a computationally major step. The command used was:

```
OrthoMCLPairs /etc/OrthoMCL.config pairs.log cleanup=no
```

where

- cleanup=no keeps the intermediary tables in the database

5. Dumping Pairs Files into Outputs

OrthoMCLDumpPairsFiles program was used to create a set of result files from the results in the database made by *OrthoMCLPairs*. The output contained a pairs directory containing

orthologue.txt, co-orthologue.txt and Inpralogs.txt. They had a normalised similarity score to describe the pair relationships. A file called *mclInput* was also generated merged into a format expected by the MCL program. The command used was:

```
OrthoMCLDumpPairsFiles /etc/OrthoMCL.config
```

6. Markov Clustering by MCL software

Markov Clustering by MCL software was performed on candidate relationships of pairs given by OrthoMCL to group/cluster into orthologous groups of proteins and produced a *mcloutput* format. The inflation value for clusters was set as 1.5. Inflation values determine how tight the clusters should be. It can range from 1 to 6, but most publications use values between 1.2 -1.5 for detecting orthologous groups. Lower the inflation value, tighter the clusters. The command used was:

```
mcl mclInput --abc -I 1.5 -o mclOutput
```

7. Grouping of Proteins into Named Orthologous Groups

OrthoMCLMclToGroups program was used to form the clusters from the previous step into a groups.txt final file.

```
OrthoMCLMclToGroups my_prefix_ 1000 < mclOutput >
groups.txt
```

where

- my_prefix is an arbitrary custom string to use as a prefix of orthologous group IDs
- 1000 is an arbitrary number selected as a starting point for the group IDs.

8. Finding Singletons

OrthoMCLSingletons program was used to find singleton genes and proteins that were not clustered into any orthologous group. This required the goodProteins.fasta file prepared earlier as input alongside the groups.txt file. The command used was:

```
OrthoMCLSingletons goodProteins.fasta groups.txt >>
singletons.txt
```

3.4 Modelling of Pan-genome

Pan-genome is the entire global set of all genes pertaining to a species. In general, it can be divided into three parts:

- The core genome: Genes conserved and shared by all the strains.
- The dispensable/ accessory genome: Genes present in some but not all the strains.
- The strain-specific/unique: Genes unique to one strain [31].

The conserved core genome of *Plasmodium* parasites was estimated based on the established 50/50 rule. Significant BLAST hits with at least 50% alignment, with the length of alignment being 50% of the longest gene in comparison, were clustered together by OrthoMCL and MCL in gene families.

3.4.1 Extraction of Core, Accessory and Unique Genome

From the clustered gene families, the genes that did not belong in any family of genes were assigned to their own unique gene family, thus forming the unique genome. Gene families possessing at least one gene from each strain formed the core genome. The rest of the genes common to some gene families but not to all were termed accessory/dispensable genomes. A code written in python language was run in the IDE to extract the core, accessory and unique genome. The source of the published code was taken from https://github.com/microDM/Utility-codes/blob/master/parseOrthoMCLOutput.py and modified manually to fit the study's criteria.

```
- python3 parseOrthoMCLOutput.py -g groups.txt -f
goodProteins.fasta -n names.txt -s singletons.txt
```

where inputs and arguments specify:

- groups.txt file generated by MCL output was passed with –g flag.
- goodProteins.fasta containing all protein sequences of all strains combined generated via BLAST DB was passed with –f flag.
- Names.txt is the manually created list of names of genomes/organisms used passed with –n flag.
- Singletons.txt file generated by OrthoMCL Singletons was passed with –s flag.

3.4.2 Extraction of Sequences for the Core Genome

Since the protein families consist of multiple proteins clustered together based on their similarity. One complete family of proteins in the *groups.txt* file was coded by one gene. This implies considering anyone protein per family and reverse-translating could provide its corresponding gene's sequence. For this purpose, the following steps were performed:

1. The whole genome of the widely used *Plasmodium* reference strain 3D7 was downloaded from PlasmoDB. The genome file consisted of FASTA headers with the sequence of every gene in it.
2. Using Python code, a list of all proteins of the reference proteome 3D7 were extracted from the orthologous families file "groups.txt".
3. The protein IDs from the FASTA headers were scanned along with the gene IDs present in the FASTA headers of the 3D7 annotated file. If protein ID was the same as gene ID, the corresponding gene sequence was extracted and saved in another file.

The resultant file consisted of all core genes and sequences, providing sequences of the core genome shared by all species in consideration.

3.4.3 Extraction of Single-Copy Orthologue (1:1 Orthologue)

Single Copy Orthologue selects the orthologue groups that have precisely one gene per organism. The protein families having single linkage clustering were extracted so that each connection shared by proteins must be from two different genomes [86]. SCOs are required to deal with multiple proteins per gene, which is necessary for several reasons:

- To avoid pseudo-in-paralogs that are the alternative proteins within a family that look like in-paralogs.
- To avoid inaccuracy in gene duplication events.
- To reduce input dataset for downstream analysis requiring fewer resources.
- To build a phylogenetic relationship of the core single-copy genes that are not duplicated [86].

A python code was written to extract the SCOs from the core genome. The resultant list was termed as core Single Copy Orthologue.

3.4.4 Removal of Duplicate Proteins

Since multiple strains can have the same proteins expressed in them and the "goodProteins.txt" file contained all the proteins expressed in all the strains considered in this study; a python script was used to calculate and generate a file representing the number of duplicated proteins in one genome, two genomes and three genomes along with the strain names in which these were duplicated. The duplicate proteins count was thus determined to find out the exclusive initial number of proteins considered for this study. This was counter checked by considering the number of duplicates from the "groups.txt" file, which was formed after clustering proteins into orthologous groups.

Similarly, after the clustering of the proteins based on their similarity into orthologous groups, duplicates from the core, accessory and unique genomes were also filtered out into another output file mentioning the names of multiple strains in which the protein had appeared. This resulted in the exclusive list of proteins encoded by the core, accessory, and unique genomes, respectively, to be considered for functional analysis.

Furthermore, another analysis was performed on the proteins to determine if any proteins lay in both the core and accessory genomes.

3.4.5 Visualisation of Pan-genome by UpSet Plot

UpSet Plots were generated for each dataset to better visualise the intersecting sets of core, accessory and unique genes. The package UpSetR was imported for the construction of the plot. An input data frame was created with input file of presence-absence matrix (see chapter 3 section 3.5). The plot was generated with the following desired attributes:

```
upset(inputdf,order.by="freq", nsets=length(set_names),
  ….sets=set_names,keep.order=TRUE,mainbar.y.label="Gene
   Intersections Size",sets.x.label="Genes per Genome",
```

```
set_size.show=TRUE, set_size.scale_max=7000,
number.angles=0,  text.scale=c(1,1,1,1,1.1,1.8))
```

3.5 Presence Absence Variation Analysis

Pairwise comparisons of proteins were visualised in the form of matrices. The Presence Absence Variation (PAV) analysis was performed by creating the following two matrices:

3.5.1 Binary Matrix

A binary matrix of orthologous clusters vs genomes was created. Each corresponding box of the matrix had a value of either 0 or 1, indicating no proteins or at least one protein from a genome being present in the gene family, respectively.

3.5.2 Count Matrix

From the results of blastp, significant hits similar to query protein were grouped into families. One family may contain more than one protein from a genome. Thus, a count matrix indicating the number of shared proteins by each genomic strain falling together in a family was constructed. This gave us insights into in-paralogs, pseudo paralogs, and single-copy orthologue. The end of the matrix depicts significant hits of proteins within its proteome (internal paralogs) [86].

3.6 Functional Categorisation of the Core Genome

Extensive Gene ontologies of the core single copy genome were searched in the PlasmoDB database. To investigate how the species belonging to different origins preserve the common components essential to fundamental biology and find common GO components pertinent to parasite-specific lifestyles; the core GO components, GO processes and GO functions were retrieved from the PlasmoDB database (https://plasmodb.org). Pivot plots were made to calculate the number of genes and the percentage of the genome involved in respective GO functions in descending order.

3.7 Determination of Virulence Factors among Core Genome

Extensive literature search provided a list of virulence causing genes that were then looked into the extracted core, accessory and unique genomes to investigate if they exist in the common core region or the unique strain-specific regions of the pan-genome. The virulence factors were further filtered to check if they were well known to be potentially good vaccine candidates.

3.8 18S Ribosomal RNA (18SrRNA) Phylogenetic Analysis

In order to understand the evolutionary relationship of the strains and their virulence causing genes, the following phylogeny was constructed. Ribosomal RNA genes are termed standard phylogenetic markers by the pioneering studies on the tree of life [87]. In the 1980s, many studies concluded that phylogenetic relationships built using conserved regions of the genome were more stable to get phenotypic traits and other linked features (Woese and Fox, 1977; Francesca D. Ciccarelli et al., 2006; Staley, 2006). 18S ribosomal RNA (18SrRNA), the gene that encodes the RNA component of the smaller subunit of the eukaryotic ribosome, is termed as a chronometer in molecular evolution because they are:

- Universally distributed
- Functionally homologous
- Molecules of identical function
- Extremely conserved structures and sequences across broad phylogenetic distances.

For this phylogenetic tree, 18SrRNA sequences were downloaded from SILVA database SSU r138.1, released in December 2019 as per their official website (https://www.arb-silva.de/). SILVA Ribosomal RNA database project contains comprehensive quality checked rRNA small and large subunits datasets for bacteria, archaea and eukarya. The criteria to download was that the rRNA must belong to strains of interest for this study, having the highest available sequence quality and alignment quality relative to other entries in the database. The downloaded sequences were aligned using the MUSCLE (Version 3.8.31) tool to create phylogeny [90]. The cladogram was then constructed using five

hundred bootstraps and the maximum likelihood method in Molecular Evolutionary Genetics Analysis (MEGA-X) (Version 10.2.5) [91]. MEGA-X has been optimised to fully use 64-bit computing resources to analyse large datasets [91]. In addition, a representation of the tree colour coded based on the continents/ geographical area was made for better visualisation using the web interface iTol (Version 6.1.2).

3.9 Plots by Pagoo package Analysis

Multiple plots were generated to better visualise and understand the characteristics of the pan-genome using the R package 'Pagoo', which is used for comparative analysis of multiple datasets of the pan-genome.

3.9.1 Pan-Core Plots (Gene accumulation curves)

Accumulation curves are rare fraction curves to demonstrate gene accumulation when an increasing number of genomes are added one by one. The plots show the trends that the genomes follow to find novel genes with the addition of a new genome. This, in turn, determines whether the pan-genome and the core genome were open or closed. For this study, accumulation curves for core and pan genomes, also known as Pan-Core plots, were constructed for each of the three datasets in the R language using the package 'pagoo'. A data frame is constructed having gene name, organism name and COG names to which the gene belongs as an input file. Random permutations of the genomes were done to get a set of values that were plotted against each genome. An exponential decay function was applied to the core genome curve to better fit, and a power-law fit was applied to the pan-genome curve. The values of alpha were also calculated to determine open or closed pan-genome.

3.9.2 Frequency Plots

The frequency plots for each dataset were constructed to represent the pan-genome better. A frequency plot shows the pattern in a set of data by measuring how frequent particular values of genes occur in a set of genomes. This indicates the trend of the inclusion of genes with more genomes.

3.9.3 Heat Maps (Distribution Plots)

A heat map for each dataset was created to compute the distance between all pairs of organisms. The heat map function by pagoo used in this study returns a distance object containing all pairwise similarities and dissimilarities between genomes.

3.9.4 Core Level Plots

A core level was calculated for each dataset to determine the percentage of organisms a core cluster must have to be considered part of the core genome. Plots were generated for the core level to visualise it better with the help of Pagoo package.

3.9.5 Principal Component Analysis (PCA)

Principal Component Analysis was performed, and plots were generated for each dataset using Pagoo analysis. Principal components (PC) are the linear combinations of the initial variables (gene clusters) to construct new variables in order to reduce dimensionality and to make up clusters.

Results

To model the pan-genomes, Twenty-three genomic strains of two types of *Plasmodium* that were prevalent in Pakistan were considered, namely *Plasmodium Falciparum* and *Plasmodium Vivax*. The strains were taken based on their geographic origin. (chapter 3, section 3.1 for datasets). Since Pakistan does not have any completely sequenced strains of *Plasmodium* on public databases, this study thus created two additional pan-genomes with its closest neighbour India and analysed each separately. As a result, the following three pan-genomes were modelled and analysed:

1) "Global Pan-genome" contains all the selected strains worldwide.
2) "Asian Pan-genome" contains all strains belonging only to Asian countries, including Pakistan's closest neighbour India.
3) "Asian excluding India Pan-genome" contains all strains belonging to Asian countries, excluding Pakistan's closest neighbour India.

In addition to exploring pan-genomes, gene ontology analysis and phylogenetic analysis were also performed on the "Global Pan-genome." In this chapter, the results for each of the studies mentioned above are discussed separately.

4.1 Genome Organisation and Pathogen-omics

Twenty-three genomic strains of *Plasmodium* species were analysed in this study. The total number of proteins in all of the available complete genomes was 136,394, and an average protein count was found to be 5992 proteins with an average gene count of 6016 genes. The basic features of *Plasmodium Falciparum* vs *Plasmodium Vivax* are shown in Table 4.1.

Table 4.1:Comparison of nuclear genome features of the two Plasmodium species [52]

Feature		*P. falciparum*	*P. vivax*
Genome	Size (Mb)	23.3	26.8
	Number of chromosomes	14	14
	Coverage	5	10
	G+C content % AT content %	42.3% 80.6%	19.4% 57.7%
Genes	Average Number of Genes	5403	5433
	Mean Gene Length	2283	2164
	% coding	53.9%	52.1%
RNAs	Number of transfer RNA genes	43	44
	Number of 5S ribosomal RNA genes	3	3
	Number of 5S ribosomal RNA genes	7	7

4.2 Homology Search by all-vs-all BLAST+

Homologous genes were searched for each pan-genome since they share functional domains and are expected to maintain ancestral biological function. A bi-directional all-vs-all blastp was performed against their specific custom database to find potential homologs within the species based on sequence similarity. Refer to chapter 3, section 3.2 for methodology. The resultant file contained the complete collection of pairwise comparisons or Reciprocal Best Hits (RBH), which are the proteins encoded by two genes residing in two different genomes that found each other as best scoring match in the other genome. Results were ranked by similarity based on blast statistics and evolutionary statistics.

BLAST analysis of protein orthologue revealed a degree of amino acid conservation of each potential homologue. Figure 8.1 in the appendix shows a snapshot of a result file of the bidirectional blast for the proteins of *P. falciparum's* strain *3D7 (PF_Europe 1:Netherlands)* against the global pan-genome's database. Each tab-delimited output file contained query protein name, homologous hit name, percentage identity, length of the hit, mismatches, gaps, starting and ending positions of the match, e-value and bit score. The percentage identity describes how similar the query is to the aligned sequences. Since closely related species have a high percentage identity, it reflects relatedness. E-value is a measure of likeliness that sequence similarity is not by random chance. Hence, the lower the e-value, the better the hit. Finally, the bit score reflects the highest alignment score

between query and database segments. It is inversely proportional to the e-value; a large bit score is less likely to be obtained by random chance.

4.3 Identification of Pairs of Orthologue, In-paralogs and Co-orthologue by OrthoMCL

For each pan-genome, OrthoMCL was used to find pairs of proteins that were potentially orthologue, in-paralogs or co-orthologue. The function was a series of twenty internal steps, each creating intermediary database tables and thus was a computationally major step. Refer to chapter 3, section 3.3 for methodology. The output contained a pairs directory containing orthologue.txt, co-orthologue.txt and inparalogs.txt, which are explained one by one in the subsequent paragraphs.

The output of each pair file had a normalised similarity score to describe the pair relationships. This provided a measure of similarity between any two sequences. The similarity score is calculated by taking the average –log of the e-values obtained by blastp of gene A vs gene B and gene B vs gene A. The higher the similarity score, the more the alignment coverage and, in turn, the higher the chances of the pair to be a potentially strong pair of homologs.

The output pairs directory contained orthologue pairs found from the results of blastp, where the two genes were deemed potential orthologue if their proteins products had found each other as a reciprocal best hit in the opposite genome. Orthologues are homologous genes that diverge from a common ancestral gene in different species/genomes after a speciation event. However, they have shared functional domains and are expected to conserve ancestral biological function.

The output pairs directory also contained only the in-paralogs/recent paralogs. In-paralogs arise from a duplication that occurs after speciation. Ancient paralogs or out-paralogs were not identified in output. They were not going to be clustered with the true orthologue in later steps because they arose from duplication before speciation and thus may have diverged to acquire new functions.

The output pairs directory also contained co-orthologue pairs. Co-orthologues are two or more genes in one lineage that are collectively orthologous to one or more genes in another lineage due to duplication events in a specific lineage. Co-orthologues arise from gene duplication following speciation. Tables 4.2, 4.3 and 4.4 show the corresponding number of orthologue, in-paralogs and co-orthologue present in the Global dataset, whereas the tables for Asian and "Asian excluding India" datasets are attached in the appendix.

Table 4.2: Number of orthologue pairs present in corresponding strains of the Global dataset.

	Braz1	**FCH4**	**Dd2**	**7G8**	**Santa**	**Palo**	**CAMP**	**NF54F**	**Mauri**	**KH01**	**GN01**	**NF54N**	**MaliF**	**PcC01**	**PvP01**	**HB3**	**Sal1**	**Viet**	**India**	**PvT01**	**Tanz**	**Korea**	**3D7new**
Braz1		3616	2982	4068	4063	3927	4056	3929	6004	4134	4152	3786	3954	5785	5464	3925	5434	4043	6033	5638	4054	6007	4153
FCH4	3616		4460	5951	5758	6042	5669	5482	3624	6050	6378	5807	6054	3561	3647	5474	3589	5916	3620	3570	6638	3618	5937
Dd2	2982	4460		4923	4738	5097	4590	4988	2980	5526	5988	4903	5018	2935	2998	4945	2949	4956	2985	2944	5730	2968	5378
7G8	4068	5951	4923		6655	6768	6358	6291	4065	6893	7256	6502	6843	4038	4124	6142	4017	6741	4093	4050	7592	4086	6778
Santa	4063	5758	4738	6655		6639	6302	6071	4072	6557	6852	6369	6493	4032	4130	6058	4019	6504	4080	4054	7268	4074	6544
Palo	3927	6042	5097	6768	6639		6535	6503	3928	7038	7603	6843	6866	3882	3974	6338	3882	6671	3942	3895	7821	3938	7036
CAMP	4056	5669	4590	6358	6302	6535		5737	4052	6200	6533	6113	6438	4021	4106	5752	4011	6367	4075	4029	7135	4057	6135
NF54F	3929	5482	4988	6291	6071	6503	5737		3933	7180	7938	6460	6146	3919	3996	6297	3886	6061	3943	3920	7057	3928	7434
Mauri	6004	3624	2980	4065	4072	3928	4052	3933		4133	4148	3781	3952	5751	5452	3925	5423	4039	6015	5609	4052	5968	4153
KH01	4134	6050	5526	6893	6557	7038	6200	7180	4133		8635	6840	6785	4080	4194	6785	4056	6683	4118	4093	7828	4136	7766
GN01	4152	6378	5988	7256	6852	7603	6533	7938	4148	8635		7324	7395	4095	4207	7385	4073	7139	4133	4105	8596	4154	8573
NF54N	3786	5807	4903	6502	6369	6843	6113	6460	3781	6840	7324		6566	3740	3831	6060	3754	6400	3807	3756	7514	3798	7022
MaliF	3954	6054	5018	6843	6493	6866	6438	6146	3952	6785	7395	6566		3929	4004	6110	3915	6839	3962	3930	8061	3966	6584
PcC01	5785	3561	2935	4038	4032	3882	4021	3919	5751	4080	4095	3740	3929		5450	3860	5307	3990	5826	5790	4011	5786	4092
PvP01	5464	3647	2998	4124	4130	3974	4106	3996	5452	4194	4207	3831	4004	5450		3947	5199	4079	5472	5395	4095	5482	4213
HB3	3925	5474	4945	6142	6058	6338	5752	6297	3925	6785	7385	6060	6110	3860	3947		3876	6028	3927	3876	6836	3924	6760
Sal1	5434	3589	2949	4017	4019	3882	4011	3886	5423	4056	4073	3754	3915	5307	5199	3876		3987	5496	5242	4001	5446	4072
Viet	4043	5916	4956	6741	6504	6671	6367	6061	4039	6683	7139	6400	6839	3990	4079	6028	3987		4057	4005	7603	4047	6501
India	6033	3620	2985	4093	4080	3942	4075	3943	6015	4118	4133	3807	3962	5826	5472	3927	5496	4057		5714	4071	6028	4134
PvT01	5638	3570	2944	4050	4054	3895	4029	3920	5609	4093	4105	3756	3930	5790	5395	3876	5242	4005	5714		4024	5644	4102
Tanz	4054	6638	5730	7592	7268	7821	7135	7057	4052	7828	8596	7514	8061	4011	4095	6836	4001	7603	4071	4024		4055	7654
Korea	6007	3618	2968	4086	4074	3938	4057	3928	5968	4136	4154	3798	3966	5786	5482	3924	5446	4047	6028	5644	4055		4145
3D7new	4153	5937	5378	6778	6544	7036	6135	7434	4153	7766	8573	7022	6584	4092	4213	6760	4072	6501	4134	4102	7654	4145	

Table 4.3: Number of in-paralog pairs present in corresponding strains of the Global dataset.

	Korea	**PvT01**	**Santa**	**3D7 new**	**FCH4**	**CAMP**	**NF54N**	**Palo**	**Braz1**	**Mauri**	**PcC01**	**Sal1**	**7G8**	**Dd2**	**NF54 F**	**India**	**Tanz**	**Viet**	**GN01**	**HB3**	**KH01**	**PvP01**	**MaliF**
Korea	227																						
PvT01		162																					
Santa			485																				
3D7new				1339																			
FCH4					365																		
CAMP						380																	
NF54N							820																
Palo								924															
Braz1									182														
Mauri										179													
PcC01											364												
Sal1												166											
7G8													595										
Dd2														675									
NF54F															1073								
India																172							
Tanz																	1921						
Viet																		601					
GN01																			2345				
HB3																				828			
KH01																					1462		
PvP01																						82	
MaliF																							880

	3D7 new	India	Viet	Dd2	Korea	Braz1	CAMP	PcC01	GN01	FCH4	7G8	MaliF	NF54N	HB3	Palo	NF54F	Sal1	Tanz	Mauri	Mauri	PvP01	PvT01	KH01
3D7new		63	248	157	56	52	173	56	199	174	123	220	173	170	174	134	65	245	6033	49	37	37	140
India	63		67	35	125	99	49	230	41	58	76	53	55	64	45	59	58	67	3620	146	44	100	41
Viet	248	67		312	63	60	643	56	441	470	383	834	468	419	397	252	78	1342	2985	58	50	55	296
Dd2	157	35	312		35	32	220	41	229	187	188	301	191	239	197	166	47	393	4093	31	23	24	158
Korea	56	125	63	35		89	44	179	40	56	70	45	54	56	42	52	55	67	4080	113	42	103	41
Braz1	52	99	60	32	89		43	188	31	60	68	48	50	57	35	51	50	64	3942	112	49	94	31
CAMP	173	49	643	220	44	43		40	331	461	393	848	436	314	362	177	59	1322	4075	39	32	37	216
PcC01	56	230	56	41	179	188	40		43	48	69	40	48	80	41	54	104	58	3943	268	81	144	42
GN01	199	41	441	229	40	31	331	43		272	216	454	251	280	289	227	49	565	6015	33	23	25	240
FCH4	174	58	470	187	56	60	461	48	272		285	557	303	268	278	176	73	873	4118	55	45	49	151
7G8	123	76	383	188	70	68	393	69	216	285		487	265	237	265	119	87	679	4133	65	61	64	145
MaliF	220	53	834	301	45	48	848	40	454	557	487		470	420	469	240	63	1647	3807	42	39	45	319
NF54N	173	55	468	191	54	50	436	48	251	303	265	470		240	231	177	65	749	3962	47	39	42	173
HB3	170	64	419	239	56	57	314	80	280	268	237	420	240		245	180	76	495	5826	55	47	52	200
Palo	174	45	397	197	42	35	362	41	289	278	265	469	231	245		175	55	713	5472	37	29	32	193
NF54F	134	59	252	166	52	51	177	54	227	176	119	240	177	180	175		63	242	3927	44	35	37	158
Sal1	65	58	78	47	55	50	59	104	49	73	87	63	65	76	55	63		82	5496	58	26	55	49
Tanz	245	67	1342	393	67	64	1322	58	565	873	679	1647	749	495	713	242	82		4057	58	55	57	400
Mauri	49	146	58	31	113	112	39	268	33	55	65	42	47	55	37	44	58	58			58	118	32
PvP01	37	44	50	23	42	49	32	81	23	45	61	39	39	47	29	35	26	55	5714	58		31	23
PvT01	37	100	55	24	103	94	37	144	25	49	64	45	42	52	32	37	55	57	4071	118	31		25
KH01	140	41	296	158	41	31	216	42	240	151	145	319	173	200	193	158	49	400	6028	32	23	25	
Santa	142	69	414	210	65	65	359	54	263	300	236	500	268	251	244	144	79	739	4134	61	52	57	146

Table 4.4: Number of co-orthologue pairs present in corresponding strains of the Global dataset.

4.4 Identification of Clusters of Orthologous Groups (COGs) and Assignment of Protein Families by Markov Clustering (MCL)

The high similarity of recent paralogs relative to orthologue can bias the clustering process. Thus the score is normalised before applying the MCL algorithm to correct the systematic bias. To determine orthologous groups, high ranking alignments across multiple species were combined. As these groups do not provide elaborative insights and require meaningful units to be identified; Clusters of Orthologous Groups (COGs) are made. Markov Clustering by MCL software was performed on candidate relationships of pairs given by OrthoMCL to cluster into meaningful COGs by stochastic flow simulation. Refer to chapter 3, section 3.3 for methodology. The inflation value for clusters was set as 1.5; tight clusters were formed. Lower the inflation value, the tighter the clusters will be.

The final output of MCL was a groups.txt plain file. Figure 4.2 shows a snapshot of the groups.txt file. Each line is a Cluster of Orthologous groups (COG), i.e. proteins belonging to the same family and are homologues of each other. The term before the colon is simply a group identifier. The protein families consist of multiple proteins clustered together based on their similarity, and it is inferred that they have a similar function. Thus one complete family of proteins is coded by one gene. Those proteins that did not group into any functionally equivalent COGs were termed as singletons. The COGs made were interpreted as follows, and it can be seen that as the number of strains were reduced, the number of COGs being identified also decreased. Hence it can be concluded that each strain dis additional genes which were part of the core and also unique genes.

4.4.1 Global Dataset

8917 Clusters of Orthologous Groups (COGs) were formed, having 132063 proteins and 4329 proteins as singletons that did not lie in any group when all the Twenty-three strains were analysed. Refer to the next section for details on duplicate removal results.

4.4.2 Asian Dataset

Similarly, for just the Asian datasets, 7567 Clusters of Orthologous Groups were made (COGs) with 3650 singletons.

4.4.3 Asian excluding India Dataset

When the Indian strains were excluded, the number of clusters formed was reduced to 7312 Clusters of Orthologous Groups formed (COGs) with 3480 singletons.

Using these Groups, the next step was to model the pan-genome as depicted in chapter 3, section 3.4.

4.5 Modelling of the Pan-genome

As discussed in the introduction, the pan-genome is the entire global set of all genes pertaining to a species. In general, it is divided into three parts: the core, the accessory and the unique genome. In this study, three pan-genomes modelled on three different datasets were as follows:

4.5.1 Global Pan-genome

In the present study, comprehensive analysis of Twenty-three *Plasmodium Falciparum* and *Plasmodium Vivax* strains taken from different origins of the world. The global pan-genome revealed a total of 13246 genes, out of which the core genome was of 2201 genes (16.61%), the accessory or dispensable genome comprised of 6716 genes (70.7%) and the unique genome contained 4329 genes (32.68%), also known as singletons.

Similarly, the core proteome was also estimated in order to understand the proteomic conservation and to find out core viral proteins that might be utilised for potential vaccine candidates. The total count of proteins encoded by the core, accessory and unique genome were 50821, 74014 and 4329 proteins, respectively. Hence the total number of proteins encoded by the entire pan-genome (core + accessory + unique) was 129164.

Another analysis performed was to find common proteins found in both core and accessory regions. The number of proteins found common in both regions were 124835. Figure 4.1 shows the pan-genome of the global dataset.

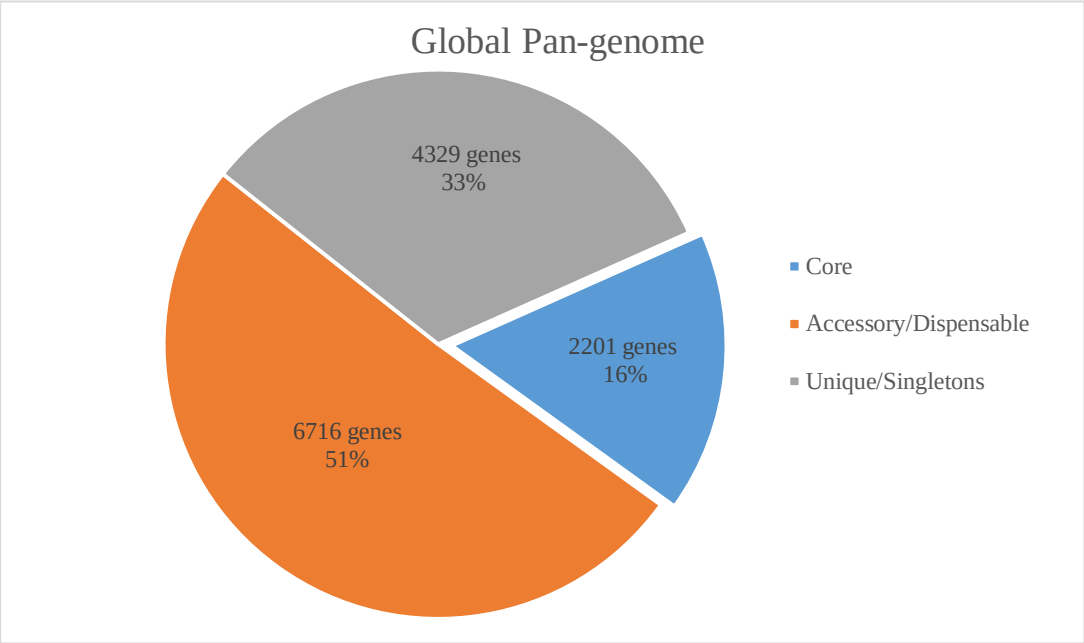

Figure 4.1: The Pan-genome modelled from the global dataset consists of 2201 core genes constituting 16%, 6716 accessory or dispensable genes constituting 51% and 4329 unique genes constituting 33% of the pan-genome.

4.5.2 Asian Pan-genome

The Asian pan-genome was constructed by the analysis of *Plasmodium Falciparum* and *Plasmodium Vivax* strains taken only from Asian countries. The pan-genome modelled from these contained 11217 total genes out of which the core genome was comprised of 2587 genes (23.06%), the accessory or dispensable genome was comprised of 4980 genes (44.39%), and the unique genome was comprised of 3650 genes (32.53%), also known as singletons.

As for the proteome analysis, the total count of core proteins encoded by the core genome was 24886, the accessory proteins were 26192 proteins, and the unique proteins were 3650. Hence, summing up the total number of proteins (core + accessory + unique) gave 48086.

The proteins found in both core and accessory regions were found out to be 51078. Figure 4.2 shows the pan-genome of the Asian dataset,

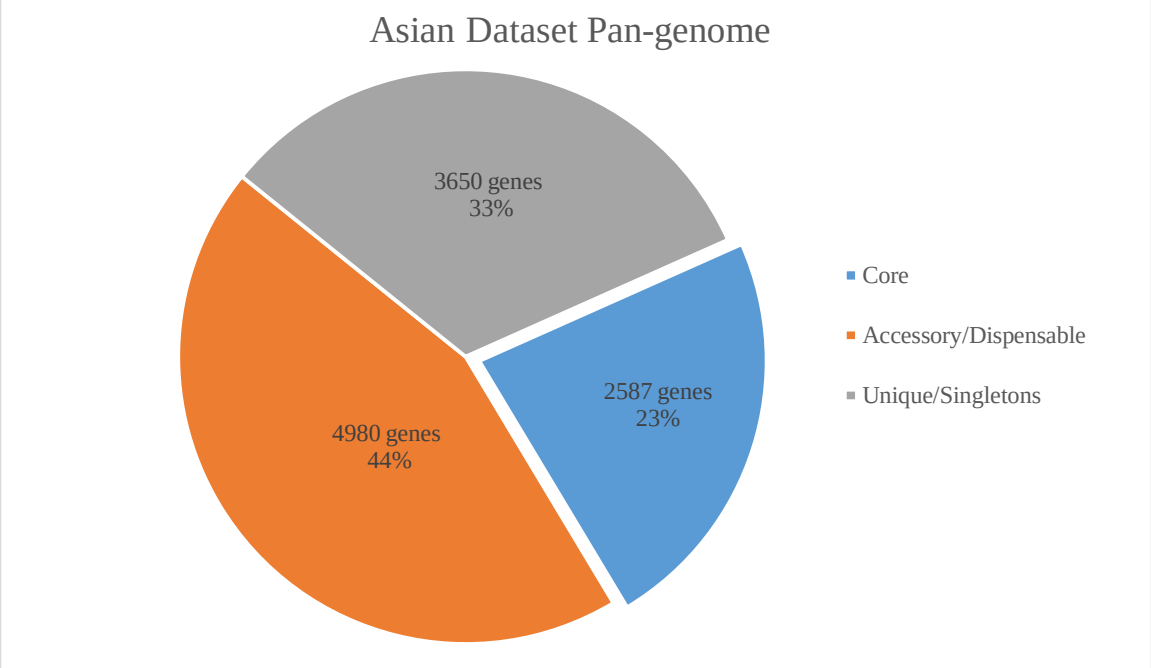

Figure 4.2: The Pan-genome modelled from the Asian dataset consists of 2587 core genes constituting 23%, 4980 accessory or dispensable genes constituting 44% and 3650 unique genes constituting 33% of the pan-genome.

4.5.3 Asian excluding India Pan-genome

The Asian excluding India pan-genome was constructed by the analysis of *Plasmodium Falciparum,* and *Plasmodium Vivax* strains take from Asian countries excluding Pakistan's closest neighbour India. This pan-genome revealed 10792 total genes having the core genome as 2593 genes (24.02%), the accessory or dispensable genome as 4719 genes (43.72%) and the unique genome as 3480 genes (32.24%).

The proteome analysis of this dataset revealed an equal number of core and accessory proteins as 22303 along with 3480 singletons. Hence, the total number of proteins encoded by the entire pan-genome (core +accessory +unique) was 48086.

The common proteins found in both core and accessory regions were 44606. Figure 4.3 shows the pan-genome of the Asian excluding India dataset.

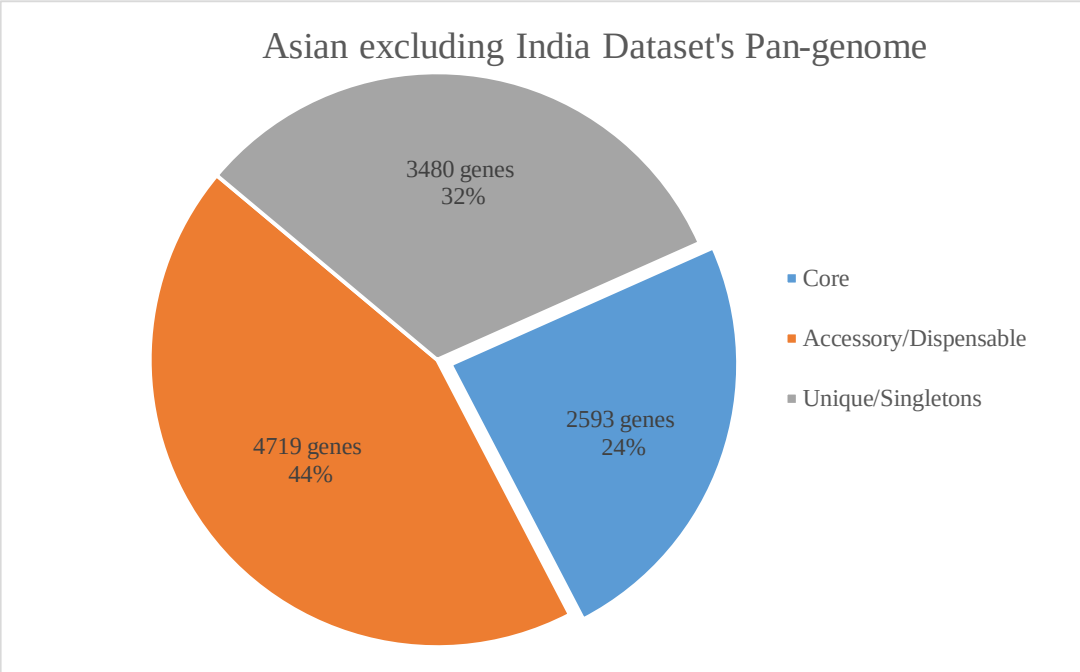

Figure 4.3: The Pan-genome modelled from Asian excluding India dataset consists of 2593 core genes constituting 24%, 4719 accessory or dispensable genes constituting 44% and 3480 unique genes constituting 32% of the pan-genome.

4.6 Filtration of Single-Copy Orthologues (1:1 True Orthologues)

Single Copy Orthologues select the cluster of orthologous groups that have exactly one gene per organism. The resultant file consisted of all core genes that were single-copy orthologue along with their sequences. A core level was calculated for each dataset to determine the percentage of organisms a core cluster must have to be considered part of the core genome. It shows the core level percentage between 85 and 100. If the core number is inversely proportional with the higher percentage of genomes (higher core level), no addition to the core has been made with the increment of core level percentage. Figure 4.4 shows the core level plot of the global dataset. The core level plots of the rest two datasets are attached in the appendix.

4.6.1 Global Pan-genome

The total number of clusters of orthologous groups (COGs) was equal to 8917, of which 2201 were core gene groups. Single copy orthologue extracted from the core having exactly one copy per genome, also known as 1:1 true orthologue, were found to be 1479 when the global dataset was considered. Figure 4.5 shows filtration of 1:1 orthologues of the Global dataset.

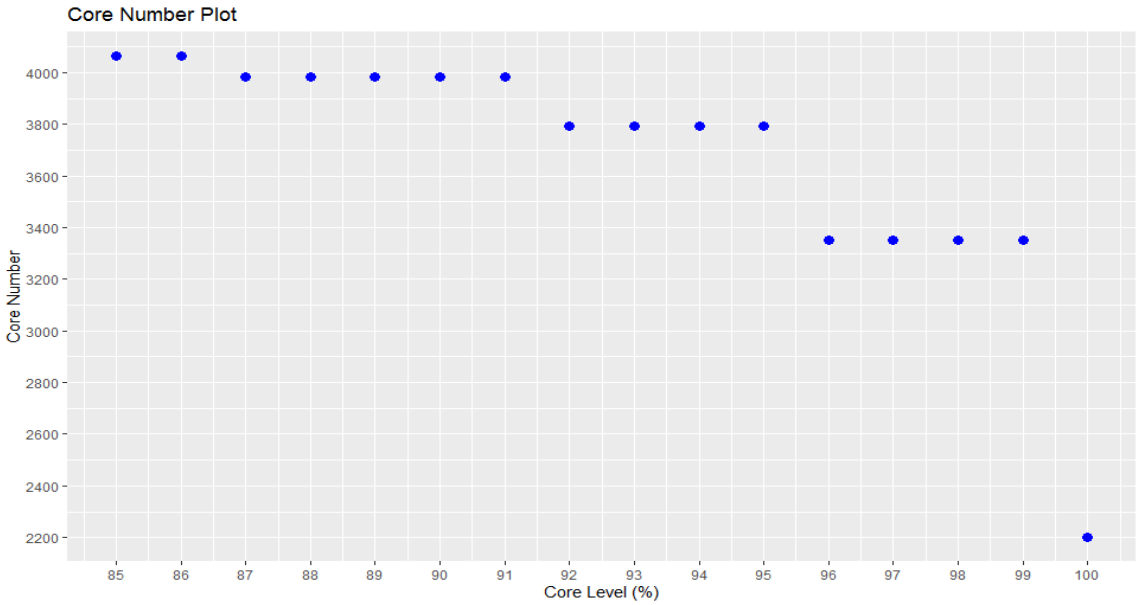

Figure 4.4: Core Number Plot of the Global dataset. It shows that when 100% genomes of the dataset were considered, the core number, i.e. total genes in the core region, were at 2201. However, the core number of genes decreased as the percentage of the core level kept increasing, i.e. with more genomes in the core. This indicates that new genes were not being added to the core when the core level percentage increased.

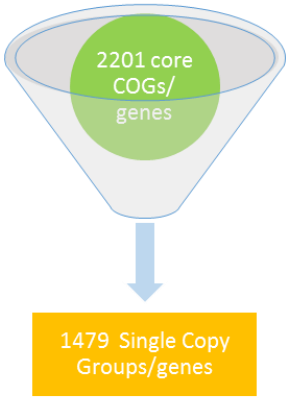

Figure 4.5: Filtration of single-copy genes, also known as true orthologue, from the core region of the global dataset.

4.6.2 Asian Pan-genome

When Asian strains were under consideration, the total number of cluster of orthologous groups (COGs) was equal to 7567, this was expected as the total number of genes are reduced when lesser strains are considered; in this case only the Asian strains. Similarly, as strains are added more and more genes are no longer part of the core gene groups hence the number of core gene increased from 2201 from the global to the 2587 for the Asian dataset. This can be depicted in This signifies that the pan-genome is open i.e. when more strains are added more diversity is introduced in the pan-genome. As the number of core genes increased this lead to an increase of single copy orthologue extracted (2060 genes). Figure 4.6 shows the filtration of Single copy orthologues from Asian dataset.

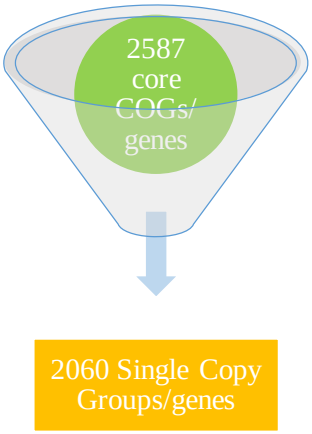

Figure 4.6: Filtration of single-copy genes from the core region of the Asian dataset.

4.6.3 Asian excluding India Pan-genome

As explained earlier that when strain number is reduced the COGs are reduced but the core genes are increases since this was the smallest dataset as the Indian strain was excluded from the Asian dataset, the number of COGs was reduced to 7312 but the core increases to 2593 genes. This results in the increase in single copy orthologue extracted from this dataset's core to 2072 genes. Figure 4.7 shows the filtration of single copy orthologues from Asian excluding India dataset.

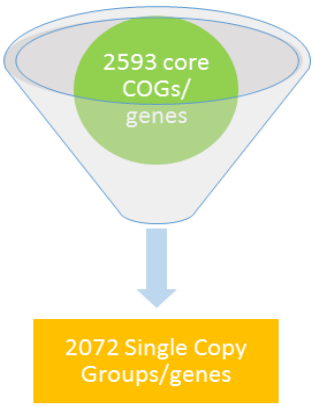

Figure 4.7: Filtration of single-copy genes also known as true orthologue from the core region of the Asian excluding India dataset.

4.7 Visualisation the Pan-genomes

UpSet plots are used to visualise and efficiently represent more than five intersecting sets and collections better than the traditional methods [92]. Earlier, methods such as Venn diagrams and Euler diagrams have been commonly used to represent biology sets, but the number of sets in this study i.e. 23 strains exceeded far greater than being represented efficiently. A famous example is the banana genome, where 64 intersections of overlapping genes are shared between multiple species where identifying participating sets required considerable effort [93]. Also, the Venn diagram only displayed the number of shared genes, making it hard to spot the largest or smallest overlaps. The solution is to construct UpSet plots that can scale up to 20 to 30 or even more sets depending on dataset properties [92].

For each of the three pan-genomes, UpSet plots were constructed to visualise the core and accessory intersections of the pan-genome. The image can be divided into the upper part having the bar chart and the lower part with a bar chart on the left, the names of genomic strains in the middle, and a dot matrix on the right. The horizontal bar chart on the bottom left shows the number of genes per genome as directly obtained from input files. The bar on the upper graph shows the number of elements for each intersection, i.e. the genes being shared by the corresponding strains having filled dots in the dot matrix. The dot matrix on the bottom right represents a filled dot for those intersecting strains that contain the

corresponding number of genes expressed on the top bar graph. For instance, if the bar indicates a number x while the strains 2,3 and 5 have filled dots, the x number of genes is shared by the mentioned three strains, i.e. 2,3 and 5. The UpSet plots for Global, Asian and Asian excluding India are represented in figure 4.8, 4.9 and 4.10, respectively.

4.8 Presence Absence Variation Analysis

Pairwise comparisons of proteins were visualised in the form of matrices. The Presence Absence Variation (PAV) analysis was performed by creating the following two matrices:

4.8.1 Binary Matrix

A binary matrix of orthologous clusters vs genomes was created. Each corresponding box of the matrix had a value of 0 or 1, indicating no proteins or at least one protein from a genome shared by the gene family, respectively. A table of the first few lines of the binary matrix for the Global dataset is shown in Table 4.5.

Table 4.5: Initial rows of the binary matrix for the Global dataset showing presence-absence variation in the genomes. The rows represent the orthologous groups clustered based on functional similarity. In contrast, the columns represent the genomic strain in which the presence or absence of proteins is indicated by either 1 or 0.

COGs/Genomes	CAMP	Dd2	FCH4	KH01	Korea	PcC01	PvP01	PvT01	Viet
OG1.5_1000	1	1	1	1	0	0	0	0	1
OG1.5_1001	1	1	1	1	1	1	1	1	1
OG1.5_1002	1	1	1	1	1	1	1	1	1
OG1.5_1003	1	1	1	1	1	1	1	1	1
OG1.5_1004	1	1	1	1	1	1	1	1	1

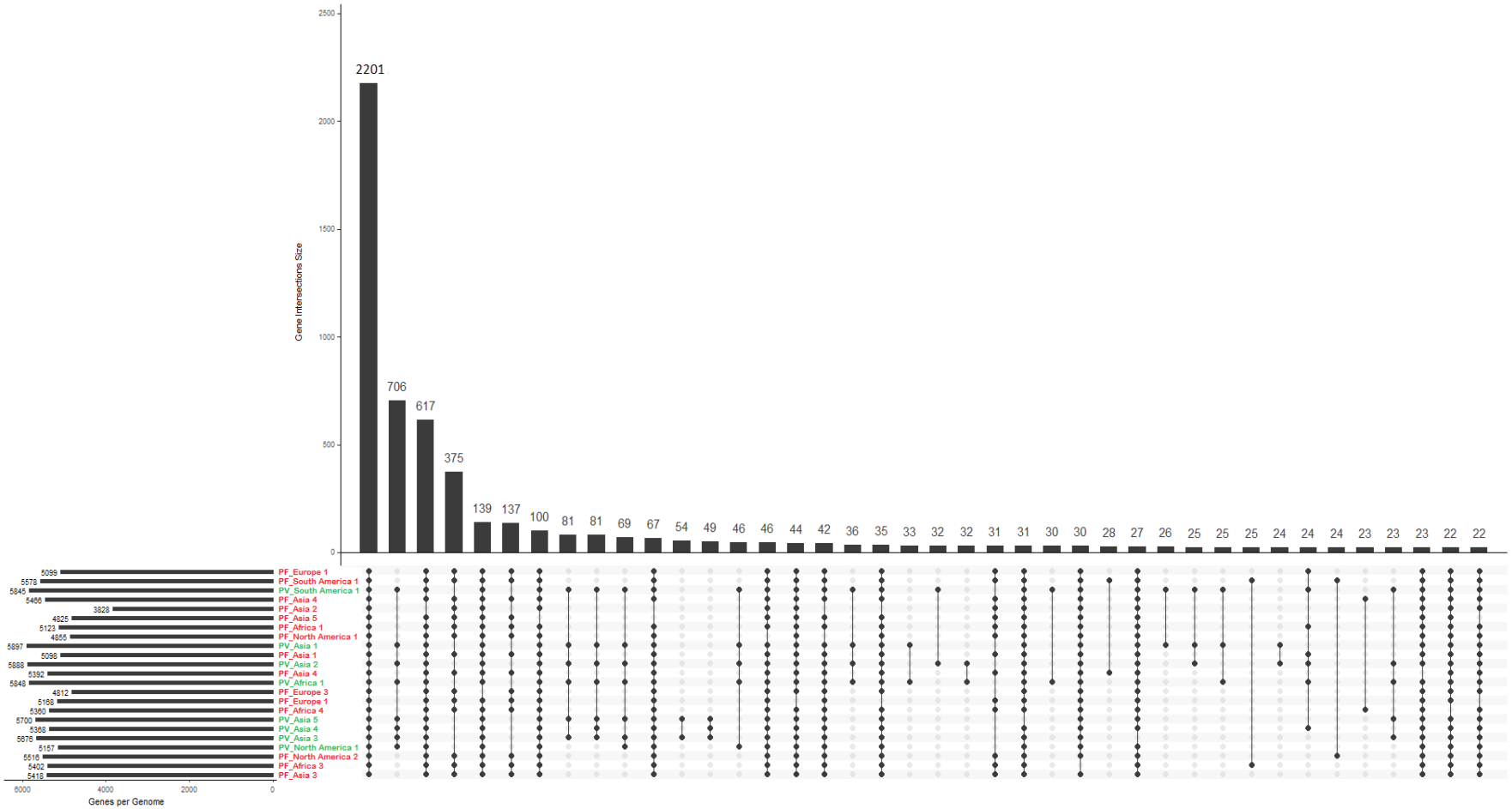

Figure 4.8: Upset Plot of the Global dataset. Plasmodium Falciparum strains and Plasmodium Vivax strains are represented by red and green colours, respectively. The number of genes per genome averages at 5000, as indicated by the horizontal bars graph. The vertical bars show the number of genes shared by the corresponding strains arranged in descending order of frequency. The core contains 2201 genes shared by all strains, as the first vertical bar shows. The filled dots in the dot matrix indicate the strains that share the common genes. The minimum number of accessory shared genes is 22.

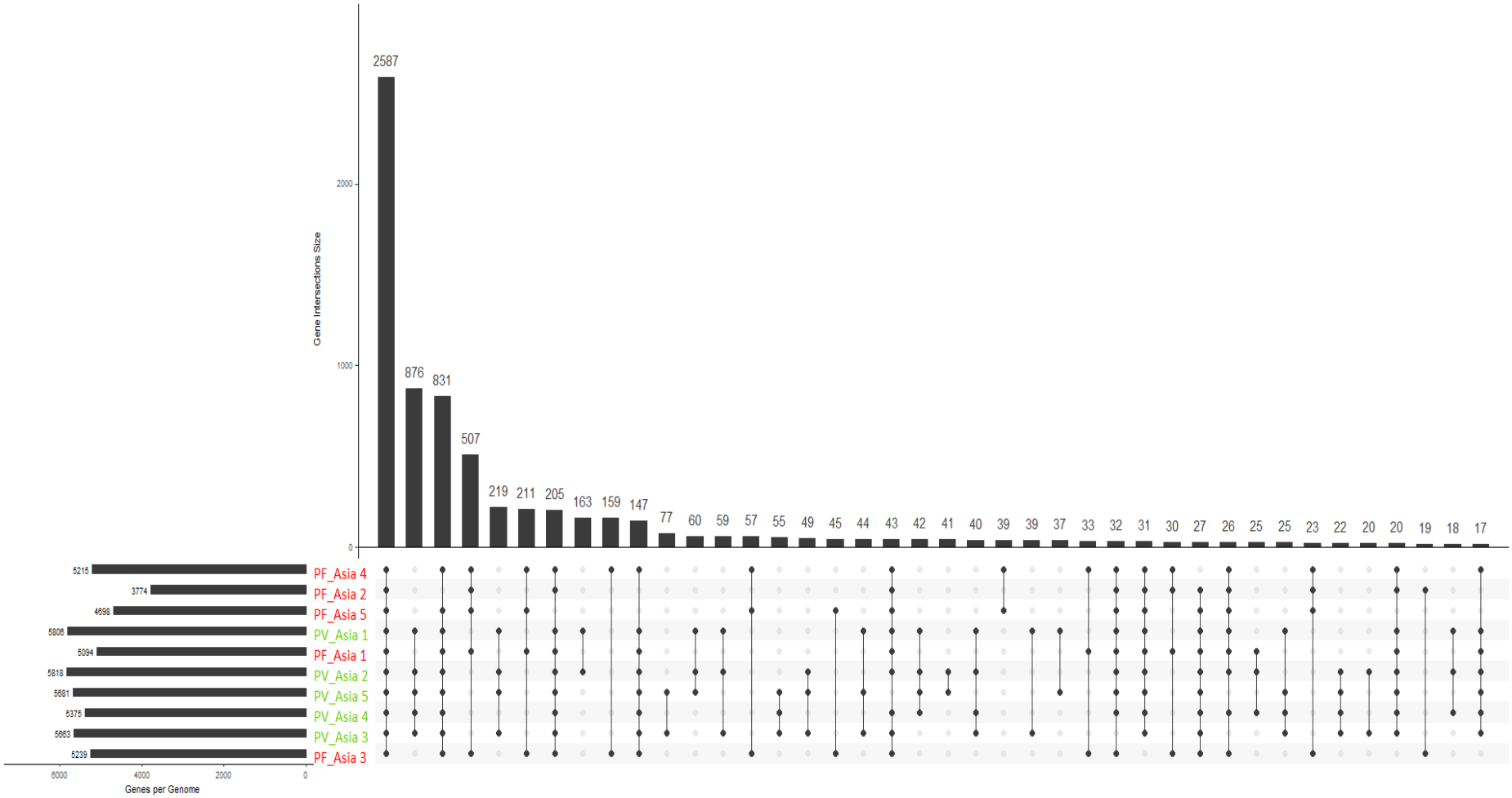

Figure 4.9: Upset Plot of the Asian dataset. Plasmodium Falciparum strains and Plasmodium Vivax strains are represented by red and green colours, respectively. The number of genes per genome averages at 5000, as indicated by the horizontal bars graph. The vertical bars show the number of genes shared by the corresponding strains arranged in descending order of frequency. The core contains 2587 genes shared by all strains, as the first vertical bar shows. The filled dots in the dot matrix indicate the strains that share the common genes. The minimum number of accessory genes are 17.

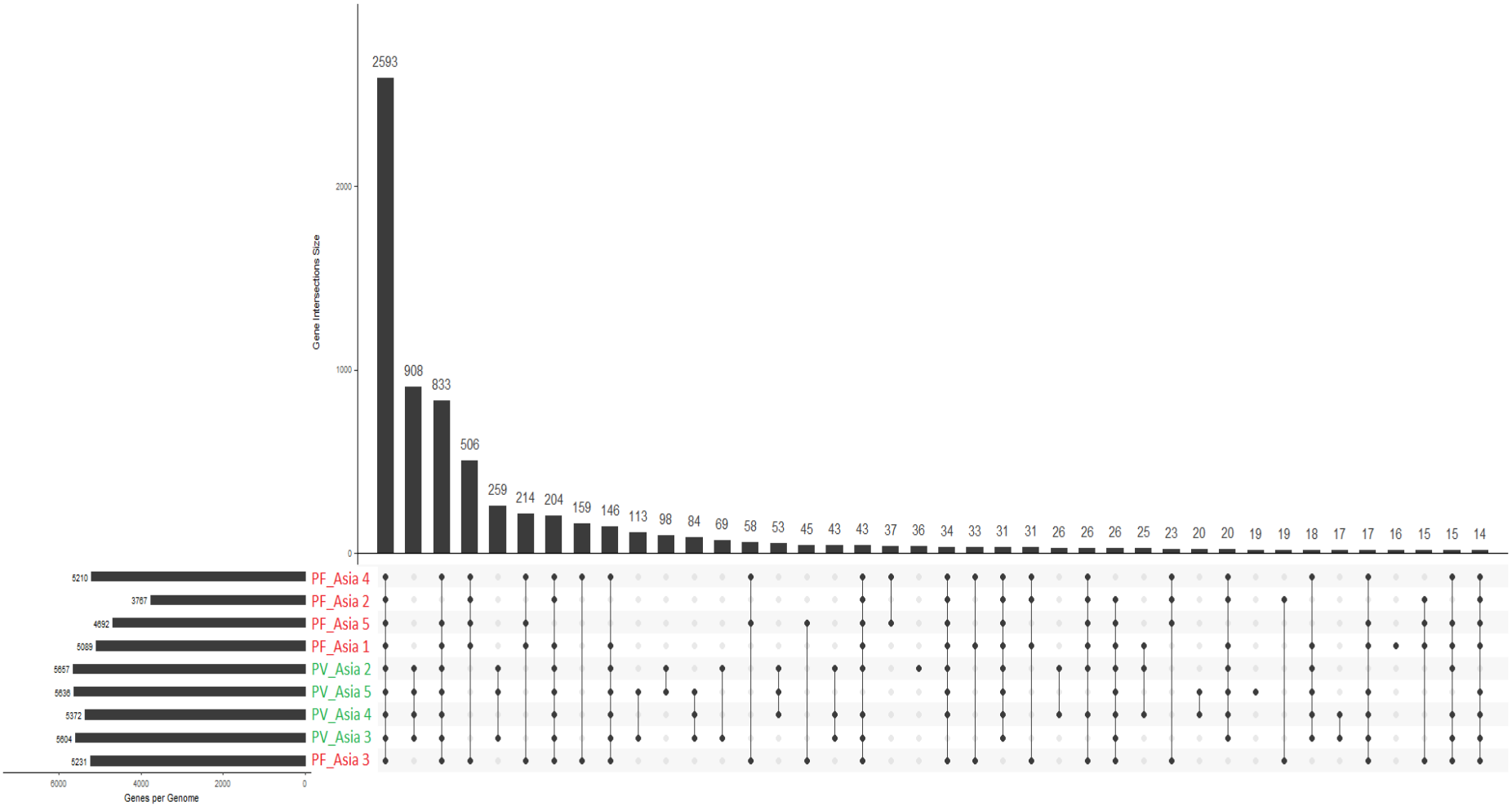

Figure 4.10: Upset Plot of the Asian excluding India dataset. Plasmodium Falciparum strains and Plasmodium Vivax strains are represented by red and green colours, respectively. The number of genes per genome averages at 5000, as indicated by the horizontal bars graph. The vertical bars show the number of genes shared by the corresponding strains arranged in descending order of frequency. The core contains 2593 genes shared by all strains, as the first vertical bar shows. The filled dots in the dot matrix indicate the strains that share the common genes. The minimum number of accessory genes are 14. The graph can also depict the strains that are more involved in sharing the genes.

4.8.2 Count Matrix

From the results of blastp, significant hits similar to query protein were grouped into families. One family may contain more than one protein from a genome. Thus, a count matrix was constructed indicating the number of shared proteins by each genomic strain falling together in a family. This gave us insights into in-paralogs, pseudo paralogs, and single-copy orthologue. The end of the matrix depicts significant hits of proteins within its proteome (internal paralogs) [86]. Table 4.6 shows the first few lines of the count matrix.

Table 4.6: The first few rows of the count matrix for the Asian excluding India Dataset. The rows represent the COGs, whereas the columns represent the genomes. Thus, the table shows the total number of proteins in a cluster of the orthologous group shared by the corresponding strain.

COGs/Genomes	CAMP	Dd2	FCH4	KH01	Korea	PcC01	PvP01	PvT01	Viet
OG1.5_1000	85	64	64	70	0	0	0	0	94
OG1.5_1001	9	8	6	8	9	11	10	11	8
OG1.5_1002	8	7	9	11	1	1	1	1	9
OG1.5_1003	5	5	6	5	3	3	3	3	7
OG1.5_1004	5	3	4	5	4	4	4	4	5
OG1.5_1005	6	8	3	6	0	0	0	0	8
OG1.5_1006	6	3	3	2	6	3	3	2	3
OG1.5_1007	3	4	3	3	3	3	3	3	3
OG1.5_1008	0	26	1	0	0	0	0	0	1

4.9 Frequency Plots

As with the concept of Moldovan and Gelfand, who described bacterial species using pan-genome data, the pan-genome representation can be done by a gene frequency spectrum G(k) that shows the correlation between the number of COGs/genes from k genomes. The set of genomes is said to be homogenous if the G(k) function chart presents a U-like shape. If a population of species is monophyletic, has a homogenous set of genomes and contains a maximal set of strains with these two conditions, the population should be classified in the same species [94]. Following the concept, the frequency plots for each dataset were constructed to represent the pan-genome better. A frequency plot shows the pattern in a set

of data by measuring how frequent particular values of genes occur in a set of genomes. This indicates the trend of the inclusion of genes with more genomes. Figure 4.11 shows the frequency plots of the global dataset.

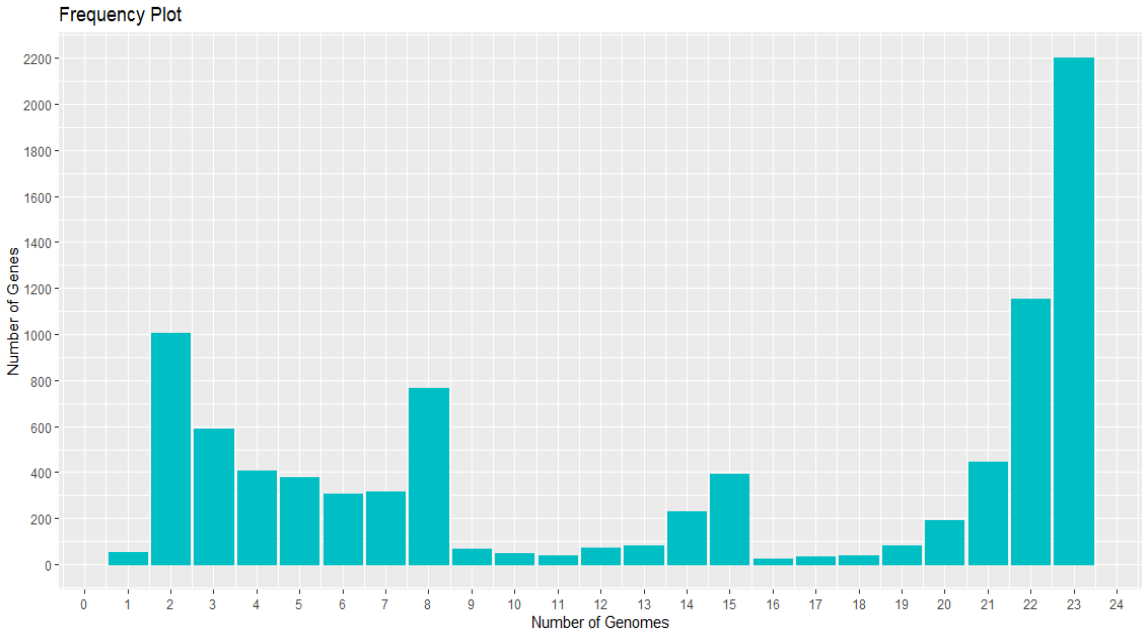

Figure 4.11: The frequency plot (No. genomes vs No. of genes) for the global dataset. . The internal peaks suggest that the genomes belong to different species. [94]***. The number of core genes for each dataset can also be determined by looking at the last peak when all strains are considered.***

4.10 Heat Maps (Distribution Plot)

Heat maps are used to visualise a phenomenon showing the intensity or magnitude by variation of the colours. It has fixed cell sizes into a matrix with varying shades of colours as per the corresponding intensity value. A heat map for each dataset was created to compute the distance between all pairs of organisms. The heat map function used in this study returns a distance object containing all pairwise similarities and dissimilarities between genomes. Figures 4.12 shows heat maps for the global dataset.

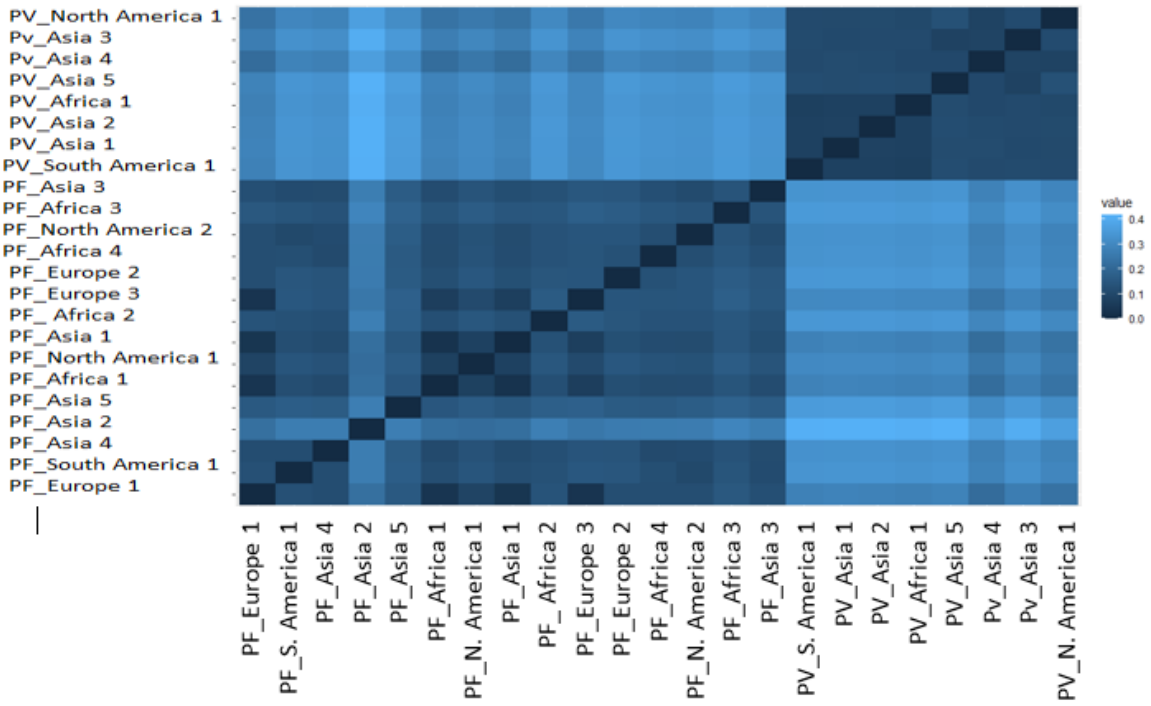

Figure 4.12: Heat maps of the global dataset. It shows the pairwise similarities between the two genomes. The darker the shade of the colours, the more similar the two corresponding genomes are. Two major clusters can be seen where the bottom left, and the top right regions are highly similar to their genomes than the rest indicating higher similarity within one type of species.

4.11 Genes Accumulation Plot/ Pan-Core Plot

Accumulation curves are rare fraction curves to demonstrate gene accumulation when an increasing number of genomes are added to the pan-genome. This concept is similar to generating a species accumulation curve that measures species' diversity in a sampling field. The number of shared genes is plotted as a function of the number *n* of strains sequentially added. For each *n*, circles are the values obtained for different strain combinations. The continuous curve represents the least square fit function to data. An exponential decaying function was used to extrapolate the size of the species core genome and for the pan-genome curve, a power-law fit was used. The results of all permutations for each of the added genomes for each dataset are shown in figures 4.13, 4.14 and 4.15, respectively.

4.11.1 Global Dataset

The comparative genomics analysis led to the pan-genome and core-genome estimation. As mentioned earlier the core genome helps elucidate the relative diversity of that species family. In this study 23 strains from diverse regions of the world were analysed for *Plasmodium* species. As can be seen from figure 4.15 the core genome of the *Plasmodium* in the start shows exponential decay i.e. adding new strains reduces the core. Nevertheless as can be seen in the plot after the inclusion of most regions the extrapolation curve reaches approximately 2000 genes and will remains relatively constant even if more genomes are added. A stable core genome also suggests that it can be explored further for the identification of conserved vaccine candidates. The red curve is made using equation 4.1, it represents the least-squares fit of the exponential decay to the medians

$$[A \times \exp(B \times x) + C]$$

***Equation 4.1** The values of the constants are A= 3799.02, B=- 0.19 and C=2201.00. The value of c represents the extrapolated core genome size which is 2201 in this case. The negative sign with the B indicates the curve to be decreasing*

Similarly, the curve of the pan-genome defines diversity and is in tandem with the core genome curve. Hence when the core genome decreases the pan-genome curve increases. The average number of new genes added by a novel sequence was 150 genes. In this study even though the core genome became relatively stable after the 16^{th} sequence the pan genome continued to rise and even the twenty-third genome added new genes. This finding suggests that the *Plasmodium* pan-genome is open and that its size grows with the number of independent strains sequenced. The blue curve is the least-squares fit of the power-law to medians.

$$(k \times x^{\Delta})$$

***Equation 4.2** The values of k=5652.72 and delta=0.16. Since delta >0, it indicates an open-pan-genome.*

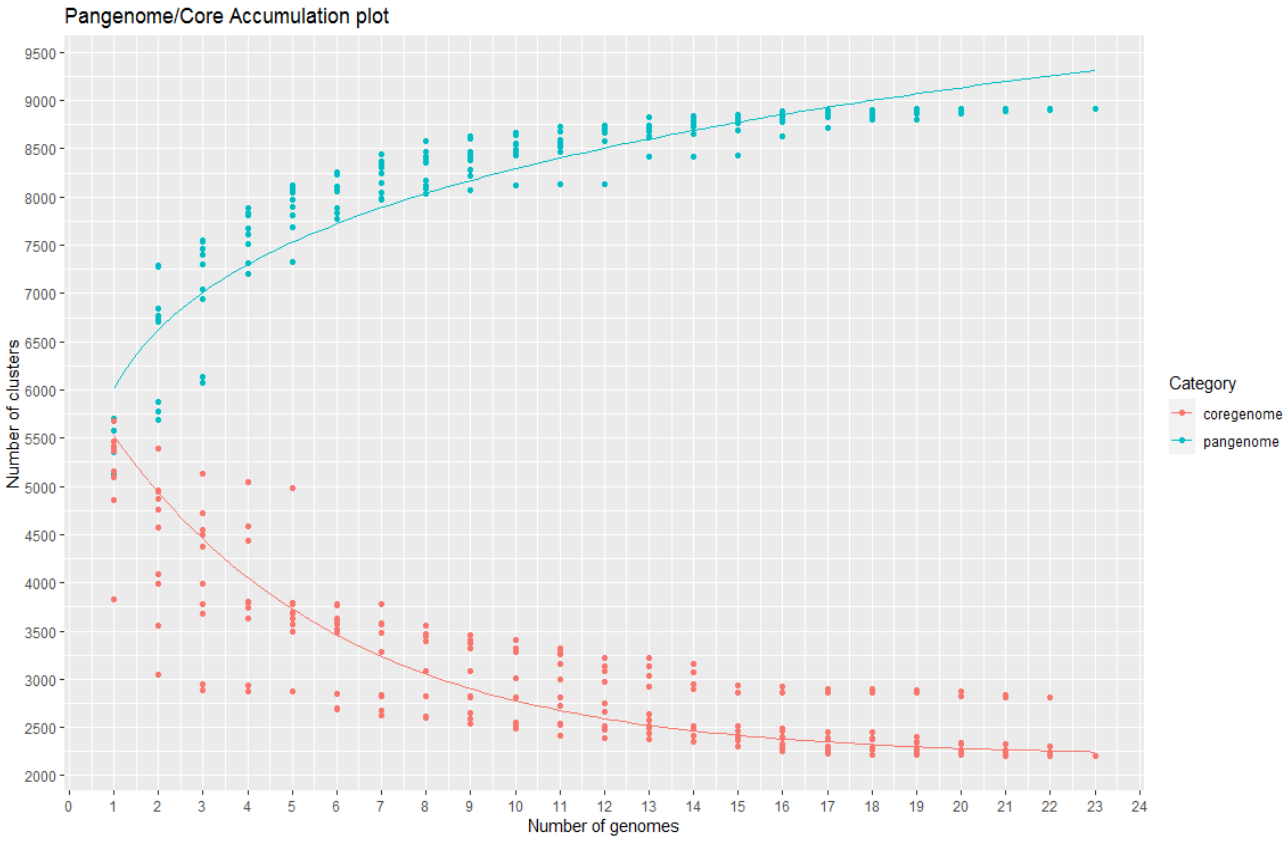

Figure 4.13: Gene accumulation curves for the pan-genome (blue) and core-genome (red) of twenty-three strains of the global dataset of Plasmodium. The x-axis and y-axis in the graph denots the cluster of orthologous groups (COGs) /genes expressed in each strain, and the number of genomes, respectively. The blue line shows the gradual pan-genome expansion, while the red line represents the core genome conservation with the addition of each genome. The number of shared genes is plotted as a function of the number n of strains sequentially added. For each n, circles are the values obtained for different strain combinations.

4.11.2 Asian Dataset

When strains from Asian countries were considered, the core genome undergoes exponential decay with the subsequent addition of new genomes i.e. new Asian strains reduce the core. As shown in the figure 4.14, the core drastically decreased with the addition of the first six genomes and gradually declined. The extrapolation could lead to zero genes being added to the core as more and more genomes from the same origin (Asia) are added. This signifies that the environmental factors of a geographical origin play a vital role in limiting the recombination events among the species and thus limiting genetic variability. The red curve is made using equation 4.3, it represents the least-squares fit of the exponential decay to the medians.

$$[A \times \exp(B \times x) + C]$$

***Equation 4.3:** The values of the constants are A= 5051.81, B= -0.49 and C= 2587.00. The value of c represents the extrapolated core genome size which is 2587 in this case. The negative sign with the B indicates the curve to be decreasing.*

Similarly, when the core genome decreases the pan-genome curve increases. The pan genome continued to rise with every new Asian strain being added. This finding suggests that the *Plasmodium* pan-genome still shows open behaviour when Asian countries are considered. The blue curve is the least-squares fit of the power-law to medians.

$$(k \times x^{\Delta})$$

Equation 4.4: The values of k= 5543.78 and delta=0.15. Since delta >0, it indicates an open-pan-genome.

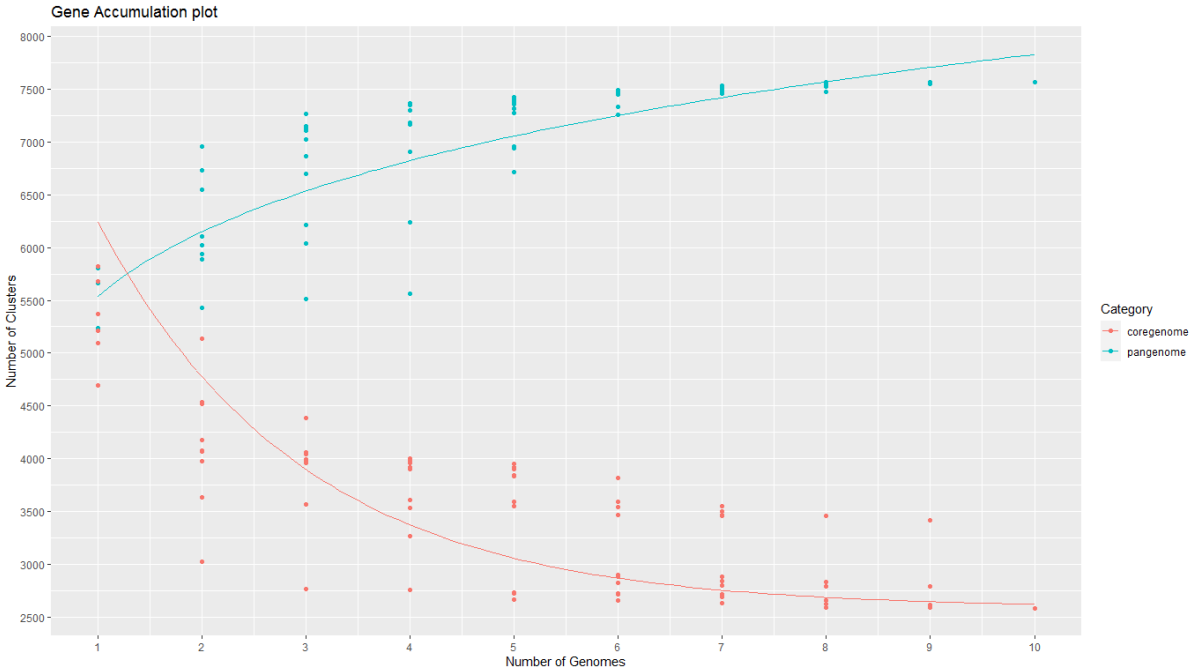

Figure 4.14: Gene accumulation curves for the pan-genome (blue) and core-genome (red) of Asian dataset of Plasmodium. The x-axis and y-axis in the graph denots the cluster of orthologous groups (COGs) /genes expressed in each strain, and the number of genomes, respectively. The blue line shows the gradual pan-genome expansion, while the red line represents the core genome conservation with the addition of each genome. The number of shared genes is plotted as a function of the number n of strains sequentially added. For each n, circles are the values obtained for different strain combinations

4.11.3 Asian excluding India Dataset

When the strain of Pakistan's closest neighbour India was excluded from the Asian dataset, the core genome reduces when new strains are added one by one. As shown in the figure 4.15, the exponential decay shows the core being drastically decreased till first six genomes and then becomes relatively stable but its extrapolation could lead the core to be zero i.e.

some genes are no longer part of the core. This signifies that the environmental factors of a geographical origin play a vital role in limiting the recombination events among the species and thus limiting genetic variability. The red curve is made using equation 4.5, it represents the least-squares fit of the exponential decay to the medians.

$$[A \times \exp(B \times x) + C]$$

***Equation 4.5:** The values of the constants are A= 3716.72, B= -0.50 and C= 2593.00. The value of c represents the extrapolated core genome size which is 2593 in this case. The negative sign with the B indicates the curve to be decreasing.*

Similarly, when the core genome decreases the pan-genome curve increases. The pan genome continued to rise with every new Asian strain being added. This finding suggests that the *Plasmodium* pan-genome still shows open behaviour when Asian countries are considered. The blue curve is the least-squares fit of the power-law to medians.

$$(k \times x^{\Delta})$$

Equation 4.6: The values of k= 5459.26 and delta=0.14. Since delta >0, it indicates an open-pan-genome.

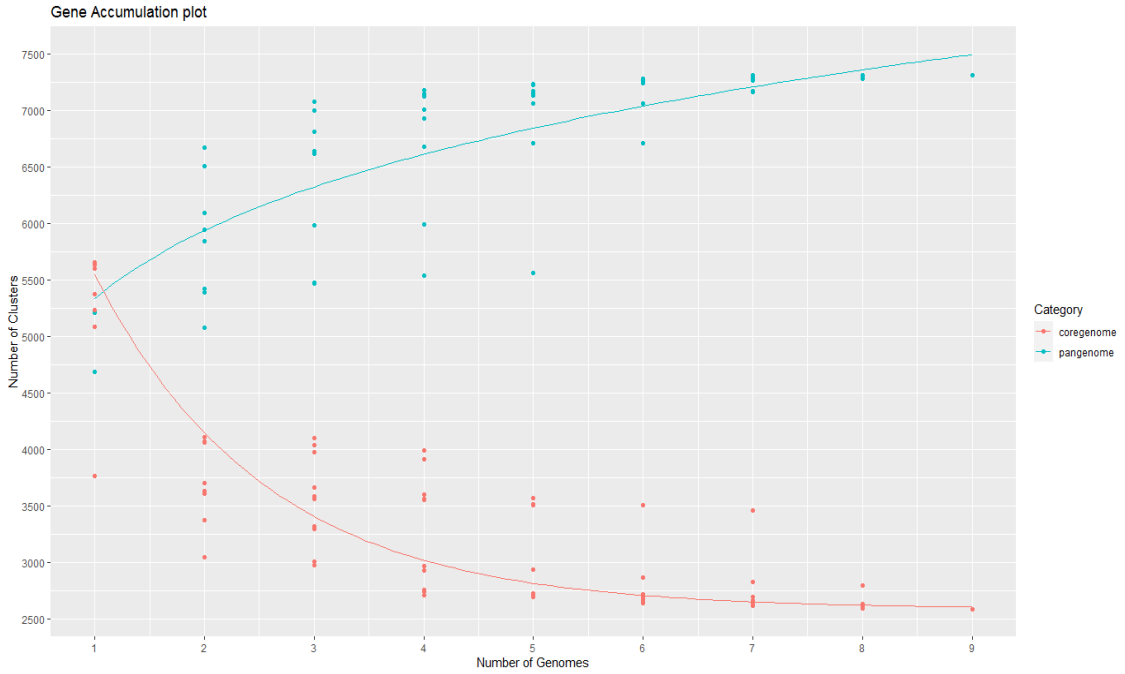

***Figure 4.15:** Gene accumulation curves for the pan-genome (blue) and core-genome (red) of Asian excluding India dataset of Plasmodium. The x-axis and y-axis in the graph denotes the cluster of orthologous groups (COGs) /genes expressed in each strain, and the number of genomes, respectively. The blue line shows the gradual pan-genome expansion, while the red line represents the core genome conservation with the addition of each genome. The number of shared genes is plotted as a function of the number n of strains sequentially added. For each n, circles are the values obtained for different strain combinations*

4.12 Open or Closed Pan-genome

The pan-core plot or gene accumulation curves returns the permutations of all genomes. The curve fitting of pan-genome is done either by applying the Power law or the Heap's law. In contrast, the expression regression decay is applied to the curve fitting of the core genome. The formula for Heap's law is shown in equation 4.7.

$$[n = kN\hat{} - \alpha]$$

Equation 4.7: where n=number of genes, N=number of genomes is the constant of proportionality and α is an exponent defined to fit the formula. [94]

The value of alpha (α) determines whether the pan-genome is open or close such that when α is lower than 1, the pan-genome is considered open, whereas a higher than one value of α determines a closed pan-genome [94]. The value of α was found for each of the three datasets using the package *Pagoo* by R. The value of α in the global dataset was 0.83, in the Asian dataset was 0.84. The Asian without India dataset was 0.85, which, as expected, were close to each other since all three datasets had the same species being extrapolated. Therefore, the value of alpha (α) is less than one indicates that the *Plasmodium* has an open pan-genome.

4.13 Principal Component Analysis (PCA)

Principal components (PC) are the linear combinations of the initial variables (gene clusters) to construct new variables. These are uncorrelated, and most of the information is placed in the first component. PCA reduces dimensionality and truncates the space without losing information by visualising the genomes in low dimensional space spanned by these directions instead of taking genomes as spots in a high dimensional space spanned by all gene clusters. As there are as many PCs as variables, the first PC accounts for the largest possible variance in the data set. The second PC accounts for the second-highest variance. This continues until a total of *n* PCs are calculated equal to the original number of variables. PCA on the pan-matrix of each dataset was computed. Each spot indicates a genome lying in the first two main components of the pan-matrix space. Each component's percentages show how much variation is seen in the total pan-matrix along with each principal

component. Summing up all the components, which is the relative explained variance, gives 1. The number of components shows that this number is required to capture the bulk of the total variation in the data. The first component always explains the maximum percentage of variation and is the most important one.

4.13.1 Global Dataset

PCA plot is also used to visualise the separation of data into distinct groups, i.e. two distinct species. Again, the principal components are differentiated well in two species, with PC1 showing 73.1% variability and PC2 shows 5.48%. Figure 4.16 shows the PCA plot for the global dataset. The Global dataset clustered the PF_Asia 2 (Indochina) strain with *Plasmodium Falciparum's* European strains. The PF_Asia 2 (Indochina) strain is genetically similar to the European region's strains. The strain PF_Africa 3 (Tanzania) was an outlier with no cluster formed with either species or origin in the global dataset. The Tanzanian strain shows high genetic diversity compared to the other strains from either type of species of *Plasmodium* found globally.

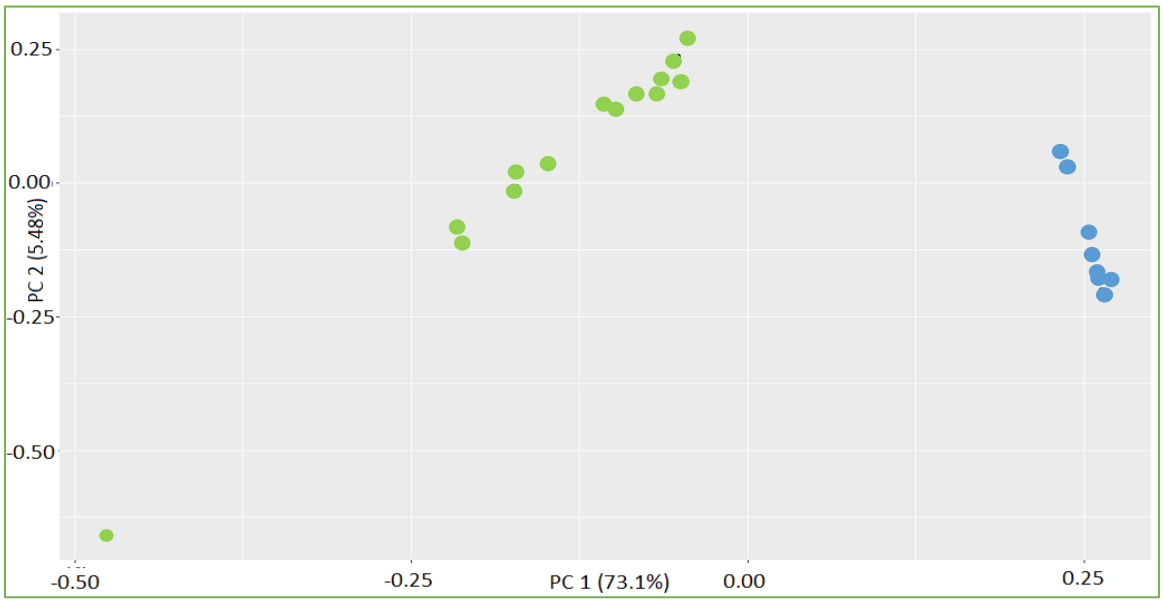

Figure 4.16: The principal components of the global dataset are differentiated well in two species, with PC1 showing 73.1% variability and PC2 shows 5.48%.

4.13.2 Asian Dataset

For the Asian dataset, the principal components are differentiated well in to two species, with PC1 showing 74.49% variability and PC2 shows 8.38%. Figure 4.17 shows the PCA plot for the Asian dataset. The PF_Asia 2 (Indochina) strain, namely “Dd2” was not grouped into any cluster. This signifies the higher genomic diversity of this strain as compared to other Asian strains.

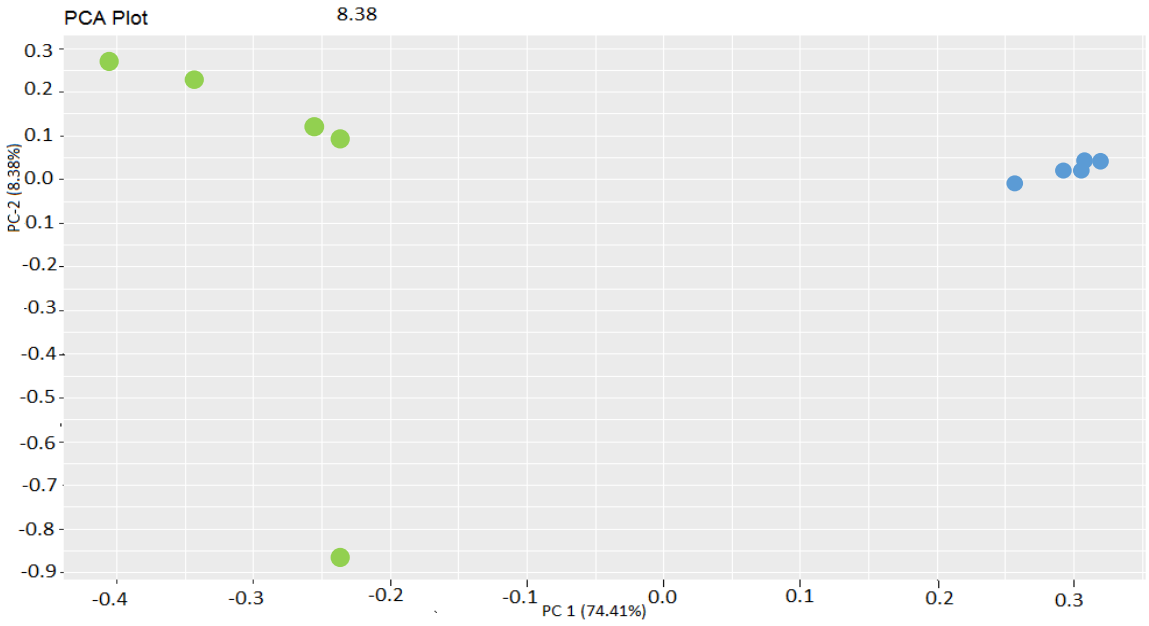

Figure 4.17: The principal components of the Asian dataset are differentiated well in two species, with PC1 showing 74.49% variability and PC2 shows 8.38%.

4.13.3 Asian excluding India Dataset

When the Indian strain was excluded from thew Asian dataset. the principal components were differentiated well in to two species, with PC1 showing 74.36% variability and PC2 shows 9.41%. Figure 4.18 shows the PCA plot for the Asian excluding India dataset. The PF_Asia 2 (Indochina) strain, namely “Dd2” was not grouped into any cluster as in the previous dataset. This signifies the higher genomic diversity of this strain as compared to other Asian strains.

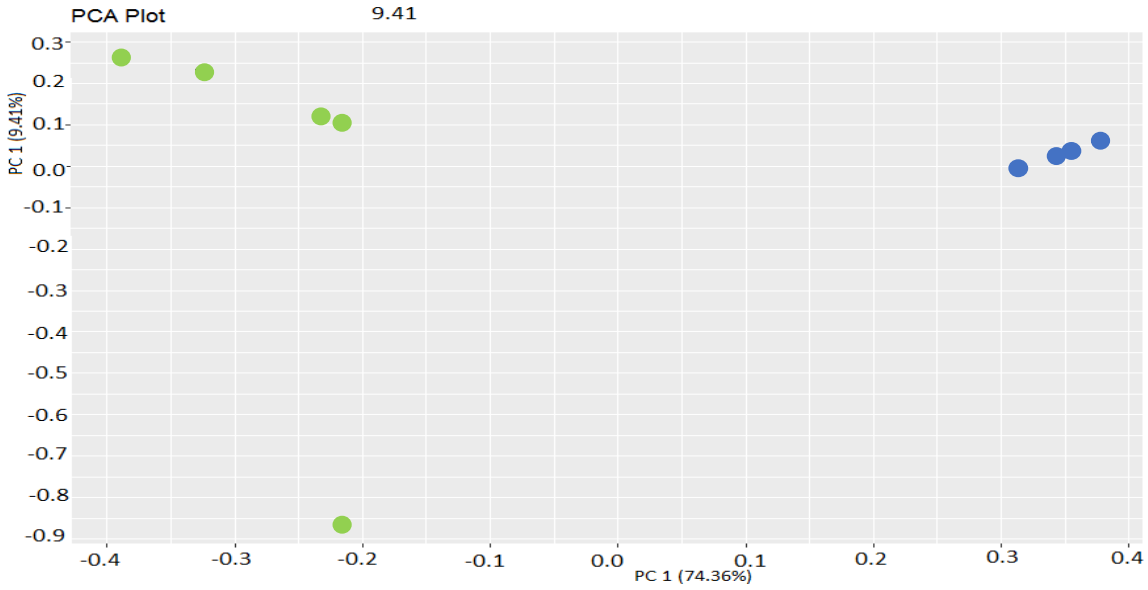

Figure 4.18: The principal components of the Asian excluding dataset are differentiated well in two species, with PC1 showing 74.36% variability and PC2 shows 9.41%.

4.14 Gene Ontologies

Gene ontologies introduce the concept of associating a list of genes with a biologically functional term in a systematic fashion. The GO database assigns GO terms to the genes based on the properties of their product. Extensive Gene ontologies of the core single copy genome were searched in the database. To investigate how the species belonging to different origins preserve the common components essential to fundamental biology and find common GO components pertinent to parasite-specific lifestyles, the core GO components, GO processes and GO functions were retrieved from the PlasmoDB database (https://plasmodb.org). Pivot plots were made to calculate the number of genes and the percentage of the genome involved in respective GO functions, GO processes and super-families involved in descending order. It is worth mentioning that 664 genes (44.8%) of the core genome were not having any information in the databases regarding GO functions. In comparison, 938 genes (63.29%) of the core genome did not have any GO processes, and 556 genes (37.53%) of the core genome did not have any superfamily mentioned in

the databases. Figures 4.19, 4.20 and 4.21 represent the top ten GO functions, GO processes, and super-families in descending order of many genes involved, respectively.

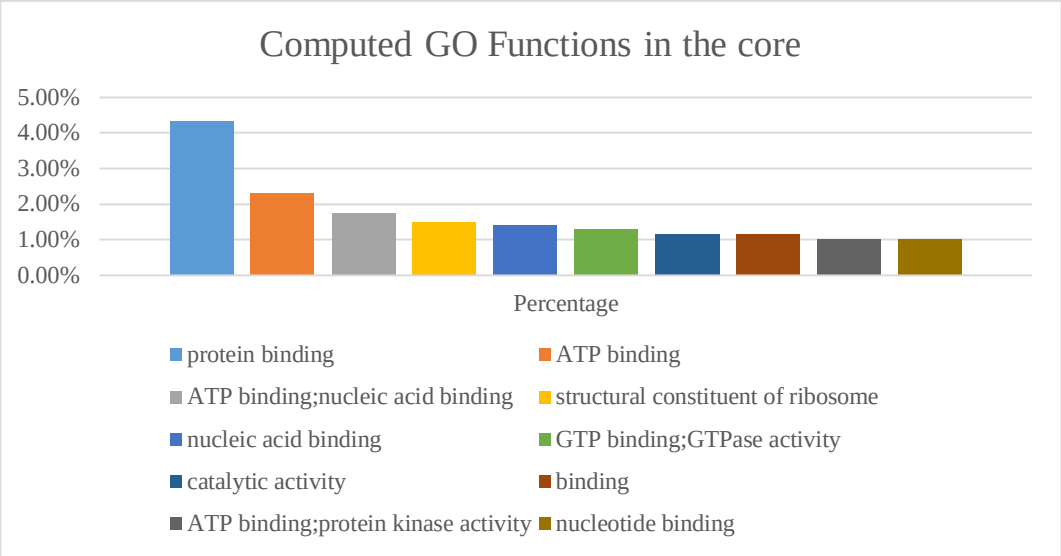

Figure 4.19: The ten major computed GO functions occurring in the Plasmodium genomes' core region and the percentage of genes involved in them are shown in descending order. Around 4% of the genes (64 genes in number) were engaged in protein-binding (GO:0005515), making it the most prevalent GO function in the core region.

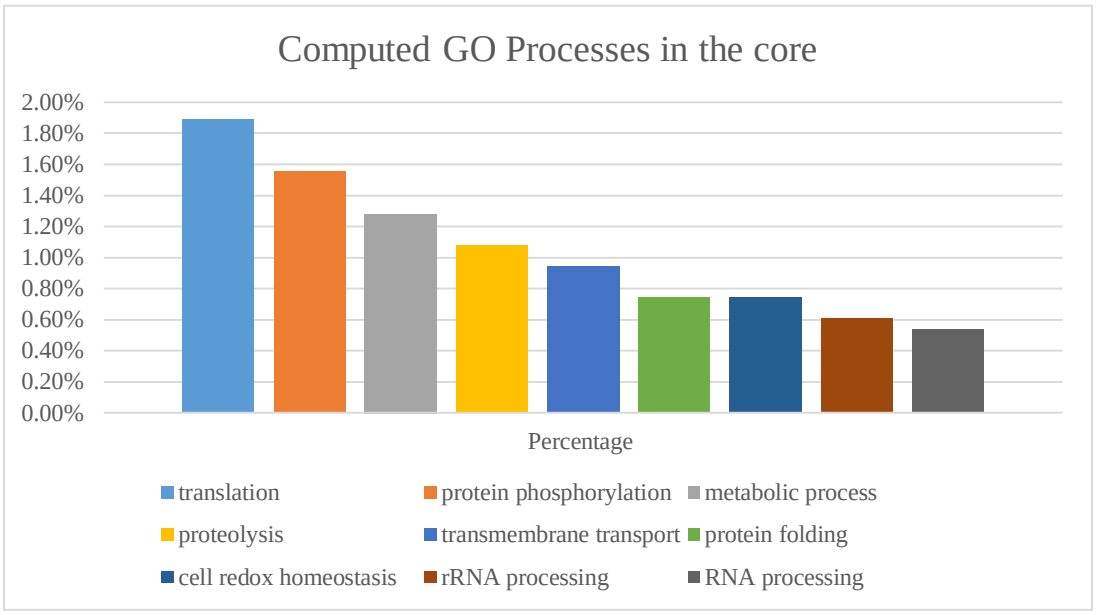

Figure 4.20: The ten major computed GO processes occurring in the Plasmodium genomes' core region and the percentage of genes involved in them are shown in descending order. Around 1.89% of the genes (28 genes in number) were involved in translation (GO:0006412), making it the most prevalent GO process in the core region.

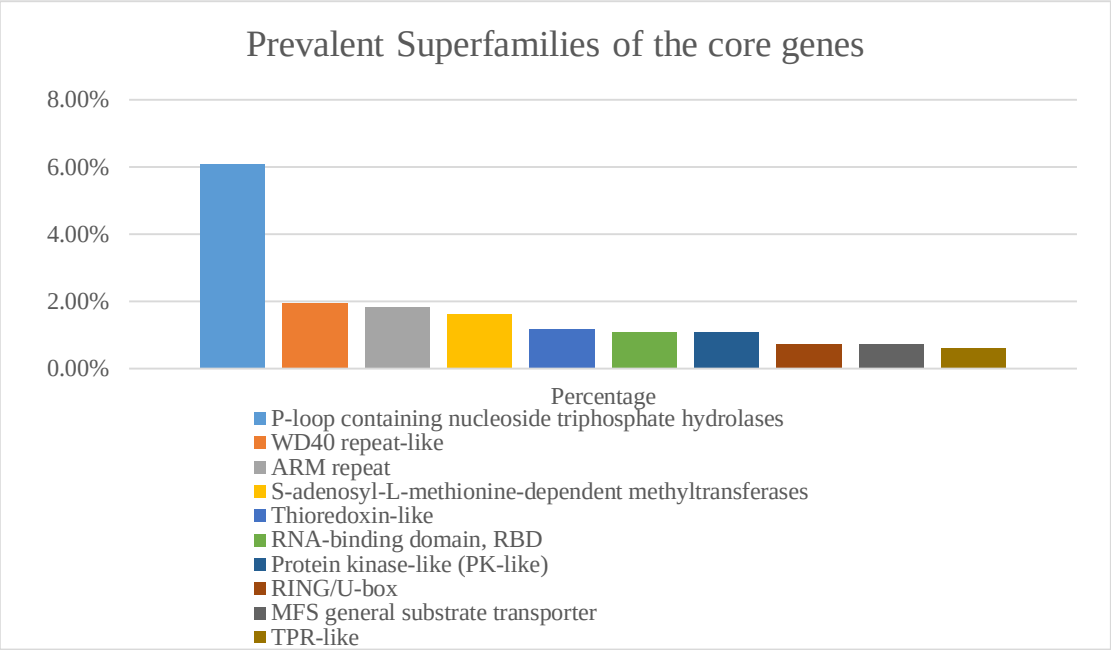

Figure 4.21: The ten major superfamilies to which the genes in the core region of the Plasmodium genomes. The percentage of superfamilies occurring is shown in descending order. Around 6% of the genes (90 genes in number) belong to the P-loop containing nucleoside triphosphate hydrolases superfamily making it the most prevalent in the core region.

4.15 Determination of Virulence Factors among Core and Unique Genomes

GO ontologies of the core genome revealed three genes involved in the process of the pathogenesis of malaria. These virulence factors were further checked if they encode surface-exposed proteins, which are well known as potentially good drug or vaccine candidates. Table 4.7 shows the details of the core SCO genes involved in the pathogenesis.

Table 4.7: COGs/Genes involved in pathogenesis of malaria belonging to the core region.

Gene ID and Gene Name/Symbol	Product Description	GO Function	GO Component	GO Process
Name: PV1 PF3D7_1129100	parasitophorous vacuolar protein 1	protein binding (GO:0005515)	Pathogenesis (GO:0009405)	Maurer's cleft (GO:0020036); food vacuole (GO:0020020); merozoite dense granule (GO:0020026); nucleus (GO:0005634); protein-containing complex (GO:0032991); symbiont-containing vacuole (GO:0020003);
Name: AMA1 PF3D7_1133400	apical membrane antigen 1	host cell surface binding(GO:0046812); protein binding(GO:0005515)	Pathogenesis (GO:0009405); entry into host (GO:0044409)	apical complex (GO:0020007); integral component of membrane (GO:0016021); microneme (GO:0020009); plasma membrane (GO:0005886)

Name: P47 PF3D7_1346800	6-cysteine protein P47	host cell surface receptor binding (GO:0046789)	modulation by symbiont of host cellular process (GO:0044068); negative regulation by symbiont of host innate immune response (GO:0052170)	anchored component of the plasma membrane (GO:0046658); cell surface (GO:0009986)

4.16 18S Ribosomal RNA (18SrRNA) Phylogenetic Analysis

Ribosomal RNA genes are termed standard phylogenetic markers by the pioneering studies on the tree of life [87]. In the 1980s, many studies concluded that phylogenetic relationships built using conserved regions of the genome were more stable to get phenotypic traits and other linked features (Woese and Fox, 1977; Francesca D. Ciccarelli et al., 2006; Staley, 2006). 18S ribosomal RNA (18SrRNA), the gene that encodes the RNA component of the smaller subunit of the eukaryotic ribosome, is termed as a chronometer in molecular evolution because they are:

- Universally distributed
- Functionally homologous
- Molecules of identical function
- Extremely conserved structures and sequences across broad phylogenetic distances.

For this phylogenetic tree, 18SrRNA sequences were downloaded from SILVA database SSU r138.1, and the cladogram was then constructed using five hundred bootstraps and the maximum likelihood method in Molecular Evolutionary Genetics Analysis (MEGA-X) (Version 10.2.5) [91]. Figure 4.24 represents the tree colour coded based on the continents/ geographical area made for better visualisation using the web interface iTol (Version 6.1.2).

The phylogenetic analysis of the 18SrRNA which are extremely conserved across broad phylogenetic distances reveals that the *Plasmodium falciparum* and *Plasmodium vivax* form clades with one another intermixing irrespective of their origin. The PF_Africa 3: Tanzania is the most diverse which shows the greatest evolutionary distance from the rest of the strains. This was in accordance to the Principal component analysis clusters formed. The PV_Asia 4, PV_Asia 3 formed one clade as expected. Similarly, strains from Philippines, Vietnam and Malaysia were found to be in closed clades. However, Ugandan strain from Africa was found to be sharing the same clade with Malaysian strain from Asia. Figure 4.22 shows the phylogeny of the 18SrRNA of *Plasmodium.*

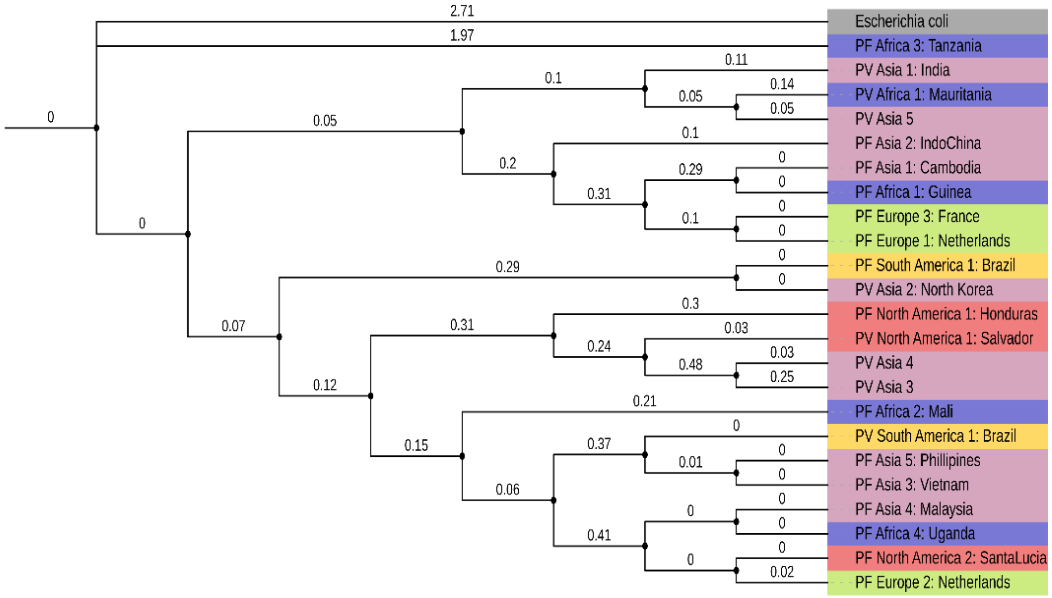

Figure 4.22: The cladogram made by 18SrRNA sequences of the entire dataset (23 strains) with an E. coli outgroup. Strains are colour coded based on their geographical origin (Blue: African, purple: Asian, green: European, pink: North American and yellow: South American). The branch lengths depict the evolutionary distances.

Discussion

Malaria, a mosquito-borne infectious disease, continues to be a significant health problem globally. (World Health Organisation, 2020). [1]. In Pakistan, 60% of the population is at high risk, and 1 million estimated cases are still attributable to malaria [8]. Even though considerable efforts had been made to eradicate malaria, the world is still off track by 37% in achieving the Global Technical Strategy 2020's milestone to eliminate the disease (World Health Organisation, 2020).

One of the factors was the continuous anti-malarial drugs resistance and the unavailability of an effective vaccine until recently when the world's first malarial vaccine passed the trials in late 2020 (World Health Organisation, 2020). However, there is a considerable challenge pointing out that different populations show variation in immune response [25]. In addition, vaccines are generally designed by a single representative isolate, which is insufficient to describe the entire genetic complexity of a species due to different geographical origins [11], [26].

Another major factor is that strains belonging to the same species vary considerably in their gene repertoire due to host and environmental factors. Adaptation to a new environment and natural selection occurs due to mutations which are essential keys to eradicating malaria in future. [70]. Since microbial genomes are prone to variation from region to region, a single representative reference genome is not enough. An ideal 'reference genome' must be a consensus drawn from a strain of different geographical origins.

Thus, there has been a paradigm shift from single reference genomics to pan-genomics. The pan-genome is the global gene repertoire pertaining to a species with core, accessory and unique regions regarding strains and their gene families.

To exploit the full advantages of computational pan-genomics and resolve significant gaps in global maps of genomes; the focus of this study is to model the pan-genome and perform the analysis on the only two prevalent types of *Plasmodium* in Pakistan, i.e. *Plasmodium falciparum* and *Plasmodium vivax o*rigin wise (i.e. among different continents around the

world). The pan-genome analysis would enable the evaluation of the core viral proteins common to all isolates, representing potential common virulence factors and potential therapeutic targets as well as strain-specific novel genes. Although pan-genome analysis has been performed on a number of species, a comprehensive analysis of *Plasmodium* has yet to be carried out.

For a pan-genome to be constructed comprehensively, at least five genomes must be analysed [73]. Furthermore, the types of strains must be from different geographical origins, and at least one must have an antibiotic resistance phenotype [69]. Considering the criteria, Twenty-three strains of *P. falciparum* and *P. vivax* were selected from publically available databases. The main dataset was further divided into two additional datasets to conduct the analysis on each separately. Since Pakistan does not have any completely sequenced strains of *Plasmodium* on public databases, this study thus created two additional pan-genomes with its closest neighbour India. The datasets for the three pan-genome analyses were as follows:

1. "Global Dataset" contains all the selected strains worldwide.
2. "Asian Dataset" contains all strains belonging only to Asian countries, including Pakistan's closest neighbour India.
3. "Asian excluding India Dataset" contains all strains belonging to Asian countries, excluding Pakistan's closest neighbour India.

Clusters of Orthologous Groups (COGs), which are functionally similar gene families, were formed for each dataset. As the number of strains was reduced, the number of COGs were also decreased, indicating that each strain added additional genes which were part of the core and unique regions. The pan-genomes consisting of core, accessory and unique regions were modelled for each dataset using these groups.

The core genome consists of the genes conserved and shared by all the isolates. The genes that lie in the core region of the pan-genome are significant for the regulation of essential aspects of biology and generally relate to cell replication, protein translation and homeostasis balance of the cell [33]. The core genome was of 2201 genes (16.61%), 2587 genes (23.06%), and 2593 genes (24.02%) for the Global pan-genome, Asian pan-genome

and Asian excluding India pan-genome, respectively. Drastic changes in the core genome are inhibited because it undergoes substantial selective pressure concerning its function [34]–[36]. Therefore, the larger core genome of both Asian datasets indicates that all these strains belonged to geographically similar regions and thus share more common genes. This shows the limited genetic diversity among the Asian strains as compared to those taken from different continents. As more strains are added to a previous dataset's core, more and more genes are no longer part of the core gene groups; hence the number of core genes decreased in the global dataset compared to both Asian datasets. This indicates the pan-genome being open, i.e. more diversity (smaller core region) in the pan-genome due to more strains.

The accessory genome is a set of genes present in some isolates but not in all isolates under consideration. The accessory or dispensable genome of the Global pan-genome comprised of 6716 genes (70.7%), the Asian pan-genome comprised of 4980 genes (44.39%) and the Asian excluding India pan-genome comprised of 4719 genes (43.72%). It can be seen that the global pan-genome contained 70% genes in the accessory region (more than both the Asian datasets) owing to the fact that the accessory genome consists of explicitly functional genes that may help survive the species in different niches. Hence, the wider the geographical distribution, the more the accessory genes. These genes are usually linked to antibiotic resistance and virulence. The accessory genes might be similar at the nucleotide level, but they are highly specific for their substrates. Horizontal gene transfer might be the source of these genes' emergence [33].

The unique genome consists of genes specific to strains, which lie in one single isolate. These are usually acquired by the transference of genes horizontally among species [33]. The unique genome of the global dataset contained 4329 genes (32.68%), while the Asian dataset contained 3650 genes (32.53%), and Asian, excluding India, contained 3480 genes (32.24%) as singletons. The strains that contain these genes show an adaptive benefit over those strains that do not possess them. These genes might be associated with virulence in pathogenic organisms, whereas they connect with metabolism by being metabolic islands in non-pathogenic organisms.

In contrast to the core genome, constant mutations occur as the mutational pressure is relaxed in this set of genes. When mutations have occurred successfully, they raise the adaptation of the organism to specific environments and conditions. The majority of the strain-specific genes are paralogous genes, as observed in a study (Jordan I.K. *et al.*, 2001). This study observed that when Pakistan's closest neighbour India was included in the Asian dataset, it increased the singletons by 170 genes. These genes can be explored further to look for virulence causing factors and have a unique strain-specific potential vaccine target specific to this origin.

Core genes can have experienced duplications which are often informational as opposed to metabolic genes. Within a group, several different types of orthology/paralogy relationships between genes are possible. Single copy genes are present in each species in exactly one copy (clear to one orthologue). They are more similar to each other than any other genes from the organisms being compared. These genes are essential as species cannot live without them. In this analysis, consideration of such genes that have not undergone any duplication events, i.e. no in-paralogs and out-paralogs, was important for phylogeny to handle the redundancy. By comparing the three datasets' results, it is evident that the number of single-copy orthologues, also known as 1:1 true orthologues increased with the increase in the number of core genes.

Heat maps for each dataset were constructed to show pairwise similarities between the genomes. It can be seen that all the *Plasmodium Vivax* strains shared very less similarity with the *Plasmodium Falciparum* strains irrespective of the geographical origin. On the other hand, within the *P. falciparum* strains, PF_Europe 3 (France), PF_Asia 1 (Cambodia), and PF_Africa 1(Guinea) strains were significantly similar to PF_Europe 1(Netherlands), which is also the reported reference genome 3D7. These strains also share high similarities. This indicates that these more similar genomes contribute to the core genome and are not genetically diverse. Hence, there is a high probability that if one vaccine candidate is found in any of these genomes, it can cater for the rest of the three genomes. As for the similarity of the strains in the Asian and Asian excluding India datasets, PF_Asia 3 (Vietnam) and PF_Asia 1 (Cambodia) showed the highest similarity. However, among the *P. vivax* strains, almost all strains were significantly similar to each other.

Accumulation curves performed an estimation of the core genome. In the global pan-genome, with the subsequent addition of new strains, the core keeps on decreasing, undergoing exponential decay. After the inclusion of most strains, the curve of the core genome becomes relatively constant even with the addition of new genomes. This curve, when extrapolated, reaches stability at approximately 2000 genes for the global dataset. A stable core genome also suggests that it can be explored further to identify conserved vaccine candidates. As for both the Asian dataset's core, the core drastically decreased with the addition of the first six to seven genomes and gradually declined. The extrapolation could lead to zero genes being added to the core as more and more genomes from the same origin (Asia) are added. This signifies that the environmental factors of a geographical origin play a vital role in limiting the recombination events among the species and thus limiting genetic variability.

Accumulation curves also performed an estimation of the pan-genome. In order to know how many sequenced genomes may be required to acquire the complete genetic repertoire of a given species, it is imperative to decide how many additional genes are to be added for each new genome that is sequenced. Analysis of the pan curve reflects that approximately 150 genes were added by a novel sequence over the global pan-genome's gradual expansion. The pan-genome kept on adding new genes even though the core was decreasing in all three datasets. The mathematical extrapolation of the pan-genome data discovered that even after the addition of hundreds of sequenced genomes, the unique genomes will always keep on being identified. This seemingly unbounded gene pool suggests that the *Plasmodium* pan-genome is open. Its size grows with the number of independent strains sequenced irrespective of their origin and the parasite type. Furthermore, open pan-genomes depict that the species is not living in an isolated *niche* and can acquire foreign genes due to recombination and gene exchange mechanisms [33]. Hence, the genetic content is very flexible for *Plasmodium.*

The curve fitting of pan-genome is done either by applying the Power law or Heap's law. The value of alpha α as determined by Heap's law was recorded as 0.838, 0.849 and 0.850 for Global, Asian and Asian excluding India datasets. When α is lower than 1, the pan-genome is considered open, whereas a higher than one value of α determines a closed pan-

genome [94]. Thus, it was concluded that the pan-genome of *Plasmodium* created from all three datasets is open irrespective of the geographical origin and type of the strains, as proven by accumulation curves and Heap's law. Furthermore, the values for alpha α of the Asian pan-genome and Asian excluding India pan-genome were closer since both datasets had the same strains being extrapolated except the Indian strain.

Principal Component Analysis (PCA) plots of the datasets indicate genomes lying in the first two main components. The bulk of the total variation of the data was distributed in multiple components. The first component, PC1, explains the maximum percentage of variation, showing 73.1% variability in the global dataset, 74.49% variability in the Asian dataset and 74.36% variability in the Asian excluding India dataset. The comparison of the three datasets shows that PC1, which explains maximum variation, almost showed the same variability percentage among all strains. Two very clear clusters were formed in the three datasets signifying *Plasmodium Falciparum* and *Plasmodium Vivax* as two different species. Each cluster had strains belonging to their specific type (*P. falciparum* or *P. vivax*) because they showed greater similarity within a species. There were outliers, in any case. The PF_Asia 2 (Indochina) strain, namely "Dd2" in both Asian datasets, was not grouped into any cluster. This signifies the higher genomic diversity of this strain concerning other Asian strains.

However, the Principal Component Analysis of the Global dataset clustered the PF_Asia 2 (Indochina) strain with *Plasmodium Falciparum's* European strains. The PF_Asia 2 (Indochina) strain is genetically similar to the European region's strains. The strain PF_Africa 3 (Tanzania) was an outlier with no cluster formed with either species or origin in the global dataset. The Tanzanian strain shows high genetic diversity compared to the other strains from either type of species of *Plasmodium* found globally.

To investigate how the species belonging to different origins preserve the common components essential to fundamental biology, the core genome was further evaluated by associating the genes with its biological functions and gene ontologies. Interestingly, it is worth mentioning that 664 genes (44.8%) of the core genome appear to have no identifiable ontology functions. In comparison, 938 genes (63.29%) of the core genome did not have

any identifiable GO processes, and 556 (37.53%) of the core genome did not have any superfamily mentioned in the databases. This is consistent with the fact that the best-annotated *Plasmodium* species *P. falciparum* had at least 60% of its 5460 ORFs annotated as "hypothetical protein", indicating the absence of reliable, functional characterisation(Cai, Gu and Wang, 2010).

Fundamental Biological Processes:

Despite their different specificities, the Twenty-three strains of *Plasmodium* species preserve the common components essential for their fundamental biology. There were plentiful orthologous families in the core region involved in processing genetic information such as replication, transcription, and translation. The most prevalent GO process in the core region was "translation" (28 orthologous families). Additionally, the translation machinery's common elements were present in the core genome, such as clusters regulating initiation, elongation, and termination of the processes, indicating basic survival elements. Similarly, the basic elements of the transcriptional machinery, such as general transcriptional factors and essential enzymes, were also part of the core region. It is well known that malarial parasites have to adapt to the developmental processes of their host, such as invasion of the parasite, sexual development and antigenic variation, which makes its transcriptional regulation very complex [95]. Associated with translation, RNA processing and splicesome is also conserved in the core clusters of these Twenty-three strains. Some other GO processes contained in the core region in the order of their prevalence after translation were protein phosphorylation, metabolic process, proteolysis, transmembrane transport, protein folding, cell redox homeostasis, protein catabolic process and DNA repair.

Parasite's Life cycle Processes:

The genes and gene products pertinent to parasite-specific lifestyles were also found in the core region and the fundamental processes. Multiple COGs for Cell cycle regulation (ten COGs), signal transduction and response to environmental challenges were found in the

core region. However, the cell cycle regulatory network remains largely unknown (Cai, Gu and Wang, 2010).

As for environmental challenges, families belonging to response to oxidative stress (GO:0006979), response to stress (GO:0006950), response to heat (GO_0009408) were found in abundance in the core.

Pathogenesis:

The core genome also contains orthologous clusters that may be relevant to pathogenesis or are virulence causing. Twenty-one orthologous clusters may be involved in the host cell entry process. For example, multiple COGs in the core encode multidrug resistance (MDR) proteins such as the core gene MDR1 (gene ID: PF3D7_0523000) encodes MDR1, and gene ID: PF3D7_1447900.1 encodes multidrug resistance protein 2 (MDR2), which are involved in ATP binding coupled to transmembrane movement of substances.

Regarding pathogenesis and virulence causing factors, there were three COGs, namely OG1.5_2955, OG1.5_ 2969 and OG1.5_3338 in the core region of the Twenty-three strains directly listed as being involved in the pathogenesis (GO:0009405).

The number of proteins that were clustered in OG1.5_2955 was 23. These proteins fall into the previously defined orthologous group OG6_533690 from OrthoMCL-DB. The OG6_533690 contains around 50 proteins, including strains from other types of *Plasmodium* not considered in this study. The gene is PV1 (ID PF3D7_1129100) which encodes the parasitophorous vacuolar protein 1(PV1). It is expressed in the GO components of Maurer's cleft, food vacuole, merozoite dense granule, nucleus, protein-containing complex and symbiont-containing vacuole. The PV1 is characterised as a novel merozoite dense granule protein which is crucial for proteins export role, specifically PfEMP1 protein. A common phenomenon observed in malarial pathogenesis is the cytoadhesion of the infected erythrocytes. The protein seen on the surface of the infected RBCs is the virulence causing *P. falciparum* erythrocyte membrane protein 1(PfEMP1). The surface-exposed antigen keeps changing by switching between protein variants and thus evade the host’s antibody immune response. Therefore, PfEMP1 is central to the pathology of the disease

and the acquisition of immunity. The PfEMP1 trafficks beyond the parasite's plasma membrane (PPM) and crosses the parasitophorus vacuolar membrane (PVM) to reach the host RBC's cytoplasm. PfEMP1 and other exported proteins remodel the erythrocytes that have been infected to make modifications crucial for the parasite's survival and virulence. A protein complex PTEX (*Plasmodium* translocon of exported proteins) is the only pathway mediating the passage of the protein across the parasitophorous vacuolar membrane (PVM). The parasitophorous vacuolar protein 1 (PF3D7_1129100) found in the core region of these Twenty-three strains co-precipitates with the PTEX complex, indicating it as an accessory molecule essential for the parasite's growth. A study also shows that the genetic knockdown of the PV1 compromises the remodelling of the host cell and decreases the cytoadhesion of infected erythrocytes. Knocking down PV1 protein has also rendered *Plasmodium berghei* less virulent. This indicates that limited physiological levels of PV1 are needed to adhere to endothelial ligands efficiently [96], [97]. Thus it is deduced that the Parasitophorus Vacuolar protein 1 could be a good drug target for both *P.falciparum* and *P.vivax* strains taken from different origins of the world.

Another core orthologous group OG1.5_2969 contained 23 proteins from the Twenty-three genomic strains that caused pathogenesis. The proteins from this group classify into the predefined OrthoMCL-DB group of OG6_130922. This protein family encodes the Apical Membrane Antigen 1. It belongs to the superfamily of Apical membrane antigens and is an integral component of the plasma membrane. Its primary GO function is host cell surface binding and protein binding, with the primary GO process being entry into the host. Many genes encode the AMA-1 protein in different populations indicating that the gene is under significant selection pressures[98].

Plasmodium parasite is asexually replicated in the host erythrocytes, which develops parasitemia. This ability of *Plasmodium* to recognise and invade the erythrocytes is a central phenomenon in the disease pathology process. Those molecules from the parasite that aid in invasion steps are agreed to be prophylactic immunisation targets. One such example is the AMA-1 antigen, a protein of the merozoite stage and a prime vaccine candidate. In the merozoite stage, a merozoite apical complex has micronemes organelles, which synthesises the AMA-1 molecule. There is evidence from animal models to consider AMA-1 as the

malarial vaccine. Antibodies can block its function by inhibiting asexual erythrocytes' multiplication of the parasite. Since recombination in the mosquito stage increases gene polymorphisms, hundreds of haplotypes can be observed in *Plasmodium* for some antigens. Polymorphisms in the AMA-1 gene have led to anti-AMA-1 vaccines because it has multiple antigens that can simultaneously overcome the species diversity by eliciting a broad protective response. The limitation is that antibody-based vaccine development targets polymorphic epitopes generally have limited efficacy [98], [99]. Thus, it is deduced that AMA-1 could be a good vaccine target as it is also surface exposed for both *P.falciparum* and *P.vivax* strains taken from different origins of the world.

The third core Cluster of orthologous group OG1.5_3338 involved in the pathogenesis is a 23 multi-proteins family encoded by the P47 gene (gene ID: PF3D7_1346800). The proteins clustered together fall into the category of a predefined cluster OG6_215372 on OrthoMCL-DB. The product is 6-cysteine protein P47. It is an anchored component of the plasma membrane and cell surface. It is involved in receptor binding to the host cell surface. Thus, it causes pathogenesis by modulation by symbiont of host cellular process and negative regulation by symbiont of innate immune response. It is observed in a study by U.N. *Ramphul et al.* in 2015 that a genetically selected strain taken from *Anopheles gambiae* mosquito, which is a vector in humans, eliminated most of the *P.falciparum's* strains taken from Asian and American countries. However, some strains from African region survived. A genetic cross between the African strains that survived and the Asian strains that were eliminated from the mosquito revealed that the gene P47 was present in the African strains that rendered the parasite invisible to the immune system and thus enhances the *P.falciparum* survival in *A.gambiae.* The 6-cysteine protein P47 gene basically disrupts the C-Jun N-terminal kinase (JNK) signalling pathway to evade the immune response. In the presence of P47 gene, JNK signaling pathway is disturbed leading to low caspase activity and no nitration. Consequently, the TEP1 molecules do not bind to the parasite which are usually detected by the immune system. This renders the parasite invisible to the immune system and the parasite survives. In P47 knockout studies, it has been observed the TEP1 molecules bind normally to the parasite and the parasite is lysed. Since 6-cysteine protein P47 is localised on the surface of female gametocytes, it is deduced that it can be a good

potential transmission blocking vaccine candidate for both *P.falciparum* and *P.vivax* strains taken from different origins of the world. The transmission blocking vaccines generate antibodies in immunised individuals that are transferred to mosquitoes during a blood meal to block the *Plasmodium* life cycle.

The potential virulence causing factors common to the core region of both types of *Plasmodium falciparum* and *Plasmodium vivax* from different geographical origins were found with the comprehensive modelling of the pan-genome. Future perspectives of the study require these candidates to be further evaluated to have an immune response to both the species of *Plasmodium* and tackle the problem of malaria mortality risk being on the rise.

.

Conclusion

The research is based on pan-genome analysis of *Plasmodium falciparum* and *Plasmodium vivax* to evaluate common core viral proteins and strain-specific genes from strains of different geographical origins. Three pan-genomes were modelled, namely Global pan-genome, Asian pan-genome and Asian excluding India pan-genome.

The Global pan-genome contains a core genome of 2201 genes (16.61%), while the Asian core genome contained 2587 genes (23.06%), and the Asian excluding India core genome contained 2593 genes (24.02%). The geographically similar strains shared more common genes and thus a larger core genome limiting the genetic diversity. In addition, two genes/protein families were added to the Asian excluding India's core when the Indian strain was added to the previous core.

The accessory or dispensable genome of the Global pan-genome comprised of 6716 genes (70.7%), the Asian pan-genome comprised of 4980 genes (44.39%) and the Asian excluding India pan-genome comprised of 4719 genes (43.72%). The accessory genes are specific to their substrates, and thus wider geographical distribution gives rise to more accessory genes. Single copy true orthologues also increased with the increase in the number of core genes for each dataset.

The unique genome of the global dataset contained 4329 genes (32.68%), while the Asian dataset contained 3650 genes (32.53%), and Asian, excluding India, contained 3480 genes (32.24%) as singletons. This study observed that when Pakistan's closest neighbour India was included in the Asian dataset, it increased the singletons by 170 genes.

The core genome decreases with the subsequent addition of new strains where it becomes relatively constant after including most strains. The core genome reaches stability at approximately 2000 genes in the global dataset. However, for the Asian dataset and Asian, excluding India's datasets, the core keeps declining without being stable.

On the other hand, the pan-genome gradually expanded with the addition of new genomes. Approximately 150 genes were added by a novel sequence in the global dataset. The genes

kept on adding, and the pan-genome's size grows with the number of independent strains sequenced irrespective of their origin and the parasite type. This seemingly unbounded gene pool suggests that the *Plasmodium* pan-genome is open. The values of alpha α from Heap's law used to fit the accumulation curves also indicate the three pan-genomes to be open.

The Principal Component Analysis for each dataset showed two very clear clusters being formed separating *Plasmodium falciparum* and *Plasmodium vivax* as two different species. Since strains from one type of species showed greater similarity, they were clustered together except for the outliers as the PF_Asia 2 (Indochina) strain in both Asian datasets and the Tanzanian strain in the global dataset that were not clustered with any type indicating higher genetic diversity from other strains.

The phylogenetic analysis of the 18SrRNA which are extremely conserved across broad phylogenetic distances reveals that the *Plasmodium falciparum* and *Plasmodium vivax* form clades with one another intermixing irrespective of their origin. The PF Africa 3: Tanzania is the most diverse which shows the greatest evolutionary distance from the rest of the strains. This was in accordance to the Principal component analysis clusters formed.

There were abundant orthologous families in the core region involved in processing genetic information such as replication, transcription, and translation. Some GO processes contained in the core region in the order of their prevalence were translation, protein phosphorylation, metabolic process, proteolysis, transmembrane transport, protein folding, cell redox homeostasis, protein catabolic process and DNA repair. In addition, multiple COGs for Cell cycle regulation (ten COGs), signal transduction and response to environmental challenges were also found in the core region. However, the cell cycle regulatory network remains largely unknown (Cai, Gu and Wang, 2010).

The core genome also contains orthologous clusters that may be relevant to pathogenesis or are virulence causing. For example, the group OG1.5_2955 contains 23 proteins that encode the Parasitophorous Vacuolar protein 1(PV1). The PV1 is a novel merozoite dense granule protein that is crucial for proteins export role, such as the export of virulence causing PfEMP1 protein to the host. It co-precipitates with the PTEX pathway. The genetic

knockdown of the PV1 hinders the alteration of the host cell and decreases the cytoadhesion of infected erythrocytes. This indicates that limited physiological levels of PV1 are needed to adhere to endothelial ligands efficiently [96], [97]. Thus it is deduced that the Parasitophorus Vacuolar protein 1 could be a good drug target for both *P.falciparum* and *P.vivax* strains taken from different origins of the world.

Another core orthologous group OG1.5_2969 also contained 23 proteins encoding the Apical Membrane Protein 1 (AMA-1). Its function is to multiply the infected erythrocytes asexually, which can be blocked by antibodies. There is evidence from animal models to consider AMA 1 as the malarial vaccine. However, since AMA-1 has highly polymorphic epitopes, the vaccine thus has limited efficacy [98], [99]. Thus, it is deduced that AMA-1 could be a good vaccine target as it is also surface exposed for both *P.falciparum* and *P.vivax* strains taken from different origins of the world.

Another core Cluster of orthologous group OG1.5_3338 involved in the pathogenesis is a 23 multi-proteins family that encodes 6-cysteine protein P47, which is involved in receptor binding to the host cell surface. It disrupts the JNK signalling pathway and evades the immune response rendering *Plasmodium* species invisible and surviving within the mosquito. Thus, it can be a good transmission blocking vaccine candidate. P47 is localised on the surface of female gametocytes, it is deduced that it can be a good potential transmission blocking vaccine candidate for both *P.falciparum* and *P.vivax* strains taken from different origins of the world.

The potential virulence causing factors common to the core region of both types of *Plasmodium falciparum* and *Plasmodium vivax* from different geographical origins were found with the comprehensive modelling of the pan-genome. The common core region of the strains contains known viral proteins that can be used as potential vaccine candidates once thorough evaluation and research are done on the candidates. Future perspectives of the study require these candidates to be further evaluated to have an immune response to both the species of *Plasmodium* and tackle the problem of malaria mortality risk being on the risk.

References

[1] N. Tangpukdee, C. Duangdee, P. Wilairatana, and S. Krudsood, “Malaria diagnosis: A brief review,” *Korean J. Parasitol.*, vol. 47, no. 2, pp. 93–102, 2009, doi: 10.3347/kjp.2009.47.2.93.

[2] A. Trampuz, M. Jereb, I. Muzlovic, and R. M. Prabhu, “Clinical review: Severe malaria,” *Crit. Care*, vol. 7, no. 4, pp. 315–323, 2003, doi: 10.1186/cc2183.

[3] “Malaria Fact Sheet N°94",” 2014.

[4] Caraballo H., “"Emergency department management of mosquito-borne illness: Malaria, dengue, and west Nile virus,” *Emerg. Med. Pract.*, vol. 16, no. 5, pp. 1–23, 2014.

[5] M. P. Girard, Z. H. Reed, M. Friede, and M. P. Kieny, “A review of human vaccine research and development: Malaria,” *Vaccine*, vol. 25, no. 9, pp. 1567–1580, 2007, doi: 10.1016/j.vaccine.2006.09.074.

[6] World Health Organization, *World Malaria Report 2020*, vol. 73, no. 1. 2020.

[7] S. Sinha, B. Medhi, and R. Sehgal, “Challenges of drug-resistant malaria,” *Parasite*, vol. 21, 2014, doi: 10.1051/parasite/2014059.

[8] Pakistan Directorate of Malaria Control, “Pakistan Malaria Annual Report 2019,” 2019, [Online]. Available: www.dmc.gov.pk.

[9] M. Mukhtar, “Killer number one: the fight against malaria: malaria strategy lags behind the global goals, Humanitarian news and analysis a service of the UN Office for the Coordination of Humanitarian Affairs. Nairobi: IRIN.,” *Nairobi: IRIN.*, 2006.

[10] P. Vale, Nuno & Aguiar, Luísa & Gomes, “Antimicrobial peptides: A new class of antimalarial drugs?,” *Front. Pharmacol.*, vol. 5, p. 275, 2014, doi: 10.3389/fphar.2014.00275.

[11] H. Cai, J. Gu, and Y. Wang, “Core genome components and lineage specific expansions in malaria parasites Plasmodium,” *BMC Genomics*, vol. 11, no. SUPPL. 3, pp. 1–10, 2010, doi: 10.1186/1471-2164-11-S3-S13.

[12] H. Kavunga-Membo *et al.*, “Molecular identification of Plasmodium species in symptomatic children of Democratic Republic of Congo,” *Malar. J.*, vol. 17, no. 1, pp. 1–7, 2018, doi: 10.1186/s12936-018-2480-5.

[13] J. D. F. M. Warrell, David A. Timothy M. Cox, “Malaria,” in *Oxford Textbook of Medicine*, 5th ed., 2010.

[14] A. Bin *et al.*, “Severe Plasmodium vivax Malaria in Pakistan,” vol. 19, no. 11, pp. 1851–1854, 2013.

[15] K. Mita, T. and Tanabe, “Evolution of Plasmodium falciparum drug resistance: Implications for the development and containment of artemisinin resistance,” *Japanese J. Infect. Dis. 65*, pp. 465–475, 2012, doi: 10.7883/yoken.65.465.

[16] M. T. Fidock, D.A., Nomura, T., Talley, A.K., Cooper, R.A., Dzekunov, S.M., Ferdig and et al. Ursos, L.M.B., Bir Singh Sidhu, A., “Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance,” *Mol. Cell 6*, pp. 861–871, 2000, doi: 10.1016/S1097-199 2765(05)00077-8.

[17] B. Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U. and et al. K., Moyes, C.L., “The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015.,” *Nat. 526*, 2015, doi: 10.1038/nature15535.

[18] S. Ashley, E.A., Dhorda, M., Fairhurst, R.M., Amaratunga, C., Lim, P., Suon, S., 194, and et al. S., Anderson, J.M., “Spread of Artemisinin Resistance in Plasmodium falciparum Malaria.,” pp. 441–423, 2014.

[19] J. C. and Hamilton, W.L., Claessens, A., Otto, T.D., Kekre, M., Fairhurst, R.M., Rayner and D. Kwiatkowski, “Extreme mutation bias and high AT content in Plasmodium falciparum,” *Nucleic Acids Res.*, vol. 45, pp. 1889–1901, 2017, doi: 10.1093/nar/gkw1259.

[20] A. P. Ashley, E.A. and Phyo, “Drugs in Development for Malaria,” *Drugs*, no. 9, pp. 861–879, 2018, doi: 10.1007/s40265-018-0911-9.

[21] WHO Africa, *World Malaria Report 2019*. 2019.

[22] N. Arora, L. C. Anbalagan, and A. K. Pannu, “Towards eradication of Malaria: Is the who’s RTS,S/AS01 vaccination effective enough?,” *Risk Manag. Healthc. Policy*, vol. 14, pp. 1033–1039, 2021, doi: 10.2147/RMHP.S219294.

[23] J. A. Stoute *et al.*, “ A Preliminary Evaluation of a Recombinant Circumsporozoite Protein Vaccine against Plasmodium falciparum Malaria ,” *N. Engl. J. Med.*, vol. 336, no. 2, pp. 86–91, 1997, doi: 10.1056/nejm199701093360202.

[24] M. A. Penny, P. Pemberton-Ross, and T. A. Smith, “The time-course of protection of the RTS,S vaccine against malaria infections and clinical disease,” *Malar. J.*, vol.

14, no. 1, pp. 1–13, 2015, doi: 10.1186/s12936-015-0969-8.

[25] L. Kurtovic *et al.*, “Induction and decay of functional complement-fixing antibodies by the RTS,S malaria vaccine in children, and a negative impact of malaria exposure,” *BMC Med.*, vol. 17, no. 1, pp. 1–14, 2019, doi: 10.1186/s12916-019-1277-x.

[26] M. E. Rauwane, U. V. Ogugua, C. M. Kalu, L. K. Ledwaba, A. A. Woldesemayat, and K. Ntushelo, “Pathogenicity and virulence factors of Fusarium graminearum including factors discovered using next generation sequencing technologies and proteomics,” *Microorganisms*, vol. 8, no. 2, 2020, doi: 10.3390/microorganisms8020305.

[27] N. Dorrell *et al.*, “Whole genome comparison of Campylobacter jejuni human isolates using a low-cost microarray reveals extensive genetic diversity,” *Genome Res.*, vol. 11, no. 10, pp. 1706–1715, 2001, doi: 10.1101/gr.185801.

[28] R. Ye *et al.*, “Genome-wide analysis of genetic diversity in plasmodium falciparum isolates from china–myanmar border,” *Front. Genet.*, vol. 10, no. OCT, pp. 1–8, 2019, doi: 10.3389/fgene.2019.01065.

[29] A. Caputo, P. E. Fournier, and D. Raoult, “Genome and pan-genome analysis to classify emerging bacteria,” *Biol. Direct*, vol. 14, no. 1, pp. 1–9, 2019, doi: 10.1186/s13062-019-0234-0.

[30] O. M. Maistrenko *et al.*, “Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity,” *ISME J.*, vol. 14, no. 5, pp. 1247–1259, 2020, doi: 10.1038/s41396-020-0600-z.

[31] H. Tettelin *et al.*, “Erratum: Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘pan-genome’ (Proceedings of the National Academy of Sciences of the United States of America (September 27, 2005) 102, 39 (13950-13955)),” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 45, p. 16530, 2005, doi: 10.1073/pnas.0508532102.

[32] A. A. Golicz, P. E. Bayer, P. L. Bhalla, J. Batley, and D. Edwards, “Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications,” *Trends Genet.*, vol. 36, no. 2, pp. 132–145, 2020, doi: 10.1016/j.tig.2019.11.006.

[33] L. C. Guimarães *et al.*, “Inside the pan-genome - methods and software overview,” *Curr. Genomics*, vol. 16, pp. 245–252, 2015, [Online]. Available: http://www.genomesonline.org.

[34] P. Lapierre and J. P. Gogarten, “Estimating the size of the bacterial pan-genome,”

2009, doi: https://doi.org/10.1016/j.tig.2008.12.004.

[35] A. J. van Tonder *et al.*, "Defining the Estimated Core Genome of Bacterial Populations Using a Bayesian Decision Model," *PLoS Comput. Biol.*, vol. 10, no. 8, 2014, doi: 10.1371/journal.pcbi.1003788.

[36] J. G. Lawrence and H. Hendrickson, "Genome evolution in bacteria: Order beneath chaos," *Curr. Opin. Microbiol.*, vol. 8, no. 5, pp. 572–578, 2005, doi: 10.1016/j.mib.2005.08.005.

[37] K. E. V. Jordan I.K., Makarova K.S., Spouge J.L., Wolf Y.I., "Lineage-specific gene expansions in bacterial and archaeal genomes," *Genome Res*, vol. 11, no. 4, pp. 555–565, 2001, doi: 10.1101/gr.GR-1660R.

[38] T. Marschall *et al.*, "Computational pan-genomics: Status, promises and challenges," *Brief. Bioinform.*, vol. 19, no. 1, pp. 118–135, 2018, doi: 10.1093/bib/bbw089.

[39] A. A. Khattak *et al.*, "Prevalence and distribution of human Plasmodium infection in Pakistan," *Malar. J.*, vol. 12, no. 1, p. 1, 2013, doi: 10.1186/1475-2875-12-297.

[40] S. E. Hocking, "Applying next generation sequencing of genomes and transcriptomes to investigate the population structure and biology of Plasmodium species," London School of Hygiene & Tropical Medicine, 2020.

[41] J. and W. T. M. G. C. Foster, J., Thompson, "The Plasmodium falciparum genome project: A resource for researchers," pp. 1–4, 1995, doi: 10.1016/0169-4758(95)80092-1.

[42] C. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W. and A. J.M., Pain, "Genome sequence of the human malaria parasite Plasmodium falciparum," *Nature*, pp. 498–511, 2002, doi: 10.1038/nature01097.

[43] T. Pain, A., Bohme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier and et al Mistry, J., "The genome of the simian and human malaria parasite Plasmodium knowlesi," *Nat. 455*, pp. 799–803, 2008, doi: 10.1038/nature07306.

[44] K. a Assefa, S., Lim, C., Preston, M.D., Duffy, C.W., Nair, M.B., Adroub, S. a, Kadir and et al. Goldberg, J.M., "Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi.," *Proc. Natl. Acad. Sci. United States Am. 112*, pp. 13027–13032, 2015, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26438871.

[45] A. Miotto, O., Amato, R., Ashley, E.A., Macinnis, B., Almagro-Garcia, J. and et al.

C., Lim, P., Mead, D., "Genetic architecture of artemisinin-resistant Plasmodium falciparum. Nature Genetics 47," pp. 226–234, 2013, doi: 10.1038/ng.3189.

[46] D. A. Volkman, S.K., Sabeti, P.C., Decaprio, D., Neafsey, D.E., Schaffner, S.F., Milner and et al. Daily, J.P., Sarr, O., "A genome-wide map of diversity in Plasmodium falciparum," *Nat. Genet. 39*, pp. 113–117, 2007, doi: 10.1038/ng1930.

[47] G. Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen and et al. O'Brien, J., Djimde, A., "Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing," *Nat. 487*, pp. 375–379, 2012, doi: 10.1038/nature11174.

[48] S. Bowman *et al.*, "The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum," *Nature*, vol. 400, no. 6744, pp. 532–538, 1999, doi: 10.1038/22964.

[49] A. P. N. Hall M. Berriman, C. Churcher, B. Harris, D. Harris, K. Mungall, S. Bowman, R. Atkin, S. Baker, A. Barron, K. Brooks, C. O. Buckee, C. Burrows, I. Cherevach, C. Chillingworth, *et al.*, "Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13," *Nature*, vol. 419, no. 527–531, pp. 527–531, 2002.

[50] C. Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E. and et al. J., Angiuoli, S. V., "Comparative genomics of the neglected human malaria parasite Plasmodium vivax.," *Nat. 445*, pp. 757–763, 2008, doi: 10.1038/nature07327.

[51] D. E. Neafsey *et al.*, "The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum," *Nat. Genet.*, vol. 44, no. 9, pp. 1046–1050, 2012, doi: 10.1038/ng.2373.

[52] J. M. Carlton *et al.*, "Comparative genomics of the neglected human malaria parasite Plasmodium vivax," *Nature*, vol. 455, no. 7214, pp. 757–763, 2008, doi: 10.1038/nature07327.

[53] T. E. Wellems and R. J. Howard, "Homologous genes encode two distinct histidine-rich proteins in a cloned isolate of Plasmodium falciparum," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 83, no. 16, pp. 6065–6069, 1986, doi: 10.1073/pnas.83.16.6065.

[54] D. Gamboa *et al.*, "A large proportion of P. falciparum isolates in the Amazon region of Peru lack pfhrp2 and pfhrp3: Implications for malaria rapid diagnostic tests," *PLoS One*, vol. 5, no. 1, 2010, doi: 10.1371/journal.pone.0008091.

[55] C. Bakari *et al.*, "Community-based surveys for Plasmodium falciparum pfhrp2 and pfhrp3 gene deletions in selected regions of mainland Tanzania," *Malar. J.*, vol. 19,

no. 1, 2020, doi: 10.1186/s12936-020-03459-3.

[56] WHO Global malaria programme, “Global response plan for pfhrp2/3 deletions,” 2017.

[57] WHO, *Surveillance template protocol for pfhrp2/pfhrp3 gene deletions*. 2020.

[58] A. B. Bosco *et al.*, “Molecular surveillance reveals the presence of pfhrp2 and pfhrp3 gene deletions in Plasmodium falciparum parasite populations in Uganda, 2017-2019,” *Malar. J.*, vol. 19, no. 1, pp. 1–14, 2020, doi: 10.1186/s12936-020-03362-x.

[59] W. R. Ballou *et al.*, “Safety and Efficacy of a recombinant DNA plasmodium falciparum sporozoite vaccine,” no. June, 1987, doi: 10.1016/S0140-6736(87)90540-X.

[60] R. S. N. and V Nussenzweig, “Development of sporozoite vaccines,” 1984, doi: doi.org/10.1098/rstb.1984.0113.

[61] A. Saula *et al.*, “Vaccine Human phase I vaccine trials of 3 recombinant asexual stage malaria antigens with Montanide ISA720 adjuvant,” *Vaccine*, vol. 17, no. 23–24, pp. 3145–3159, 1999, doi: https://doi.org/10.1016/S0264-410X(99)00175-9.

[62] B. Genton and Z. H. Reed, “Asexual blood-stage malaria vaccine development: Facing the challenges,” *Curr. Opin. Infect. Dis.*, vol. 20, no. 5, pp. 467–475, 2007, doi: 10.1097/QCO.0b013e3282dd7a29.

[63] A. Saul *et al.*, “Human phase I vaccine trials of 3 recombinant asexual stage malaria antigens with Montanide ISA720 adjuvant,” *Vaccine*, vol. 17, no. 23–24, pp. 3145–3159, 1999, doi: 10.1016/S0264-410X(99)00175-9.

[64] D. Sturchler *et al.*, “Safety, Immunogenicity, and Pilot Efficacy of Plasmodium falciparum Sporozoite and Asexual Blood-Stage Combination Vaccine in Swiss Adults,” vol. 53, no. 4, pp. 423–431, 1995, doi: https://doi.org/10.4269/ajtmh.1995.53.423.

[65] B. Genton *et al.*, “A recombinant blood-stage malaria vaccine reduces Plasmodium falciparum density and exerts selective pressure on parasite populations in a phase 1-2b trial in Papua New Guinea,” *J. Infect. Dis.*, vol. 185, no. 6, pp. 820–827, 2002, doi: 10.1086/339342.

[66] C. A. Long and S. L. Hoffman, “Malaria--from Infants to Genomics to Vaccines,” *Science (80-.).*, vol. 297, no. 5580, pp. 345–347, 2002, doi: 10.1126/science.1074484.

[67] Lancet, "Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: Final results of a phase 3, individually randomised, controlled trial," *Lancet*, vol. 386, no. 9988, pp. 31–45, 2015, doi: 10.1016/S0140-6736(15)60721-8.

[68] S. C. T. P. The RTS, "A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants," *N. Engl. J. Med.*, vol. 367, no. 24, pp. 2284–2295, 2012, doi: 10.1056/nejmoa1208394.

[69] L. Rouli, V. Merhej, P. E. Fournier, and D. Raoult, "The bacterial pangenome as a new tool for analysing pathogenic bacteria," *New Microbes New Infect.*, vol. 7, pp. 72–85, 2015, doi: 10.1016/j.nmni.2015.06.005.

[70] S. Auburn *et al.*, "Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics," *Nat. Commun.*, vol. 9, no. 1, pp. 1–12, 2018, doi: 10.1038/s41467-018-04965-4.

[71] K. Georgiades and D. Raoult, "Defining pathogenic bacterial species in the genomic era," *Front. Microbiol.*, vol. 1, no. JAN, pp. 1–13, 2011, doi: 10.3389/fmicb.2010.00151.

[72] G. Vieira *et al.*, "Core and panmetabolism in Escherichia coli," *J. Bacteriol.*, vol. 193, no. 6, pp. 1461–1472, 2011, doi: 10.1128/JB.01192-10.

[73] G. Vernikos, D. Medini, D. R. Riley, and H. Tettelin, "Ten years of pan-genome analyses," *Curr. Opin. Microbiol.*, vol. 23, pp. 148–154, 2015, doi: 10.1016/j.mib.2014.11.016.

[74] D. A. Rasko *et al.*, "The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates," *J. Bacteriol.*, vol. 190, no. 20, pp. 6881–6893, 2008, doi: 10.1128/JB.00619-08.

[75] Z. Hu *et al.*, "EUPAN enables pan-genome studies of a large number of eukaryotic genomes," *Bioinformatics*, vol. 33, no. 15, pp. 2408–2409, 2017, doi: 10.1093/bioinformatics/btx170.

[76] A. W. Khan, V. Garg, M. Roorkiwal, A. A. Golicz, D. Edwards, and R. K. Varshney, "Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement," *Trends Plant Sci.*, vol. 25, no. 2, pp. 148–158, 2020, doi: 10.1016/j.tplants.2019.10.012.

[77] M. F. Danilevicz, C. G. Tay Fernandez, J. I. Marsh, P. E. Bayer, and D. Edwards, "Plant pangenomics: approaches, applications and advancements," *Curr. Opin. Plant Biol.*, vol. 54, pp. 18–25, 2020, doi: 10.1016/j.pbi.2019.12.005.

[78] G. S. Vernikos., "A Review of Pangenome Tools and Recent Studies," in *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, T. H and M. D, Eds. Cham (CH): Springer, 2020.

[79] C. Camacho *et al.*, "BLAST+: Architecture and applications," *BMC Bioinformatics*, vol. 10, pp. 1–9, 2009, doi: 10.1186/1471-2105-10-421.

[80] C. Camacho, T. Madden, N. Ma, T. Tao, R. Agarwala, and A. Morgulis, "BLAST Command Line Applications User Manual, BLAST® Help [Internet]," *Natl. Cent. Biotechnol. Inf. (US), Bethesda, MD USA*, no. Md, pp. 1–14, 2008.

[81] G. Profiti, P. Fariselli, and R. Casadio, "AlignBucket: A tool to speed up 'all-against-all' protein sequence alignments optimizing length constraints," *Bioinformatics*, vol. 31, no. 23, pp. 3841–3843, 2015, doi: 10.1093/bioinformatics/btv451.

[82] I. Alam, S. A. Nadeem, and J. M. Brooke, "avaBLAST: A fast way of doing all versus all BLAST," *2008 Cairo Int. Biomed. Eng. Conf. CIBEC 2008*, no. January 2009, 2008, doi: 10.1109/CIBEC.2008.4786046.

[83] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastBLAST: Homology relationships for millions of proteins," *PLoS One*, vol. 3, no. 10, 2008, doi: 10.1371/journal.pone.0003589.

[84] S. Fischer *et al.*, "Using OrthoMCL to assign proteins to O," pp. 1–23, 2012, doi: 10.1002/0471250953.bi0612s35.Using.

[85] L. Li, C. J. J. Stoeckert, and D. S. Roos, "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research," *Genome Res.*, vol. 13, no. 9, pp. 2178–2189, 2003, doi: 10.1101/gr.1224503.candidates.

[86] A. Hassan *et al.*, "Pangenome and immuno-proteomics analysis of Acinetobacter baumannii strains revealed the core peptide vaccine targets," *BMC Genomics*, vol. 17, no. 1, 2016, doi: 10.1186/s12864-016-2951-4.

[87] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: The primary kingdoms," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 11, pp. 5088–5090, 1977, doi: 10.1073/pnas.74.11.5088.

[88] J. T. Staley, "The bacterial species dilemma and the genomic-phylogenetic species concept," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 361, no. 1475, pp. 1899–1909, 2006, doi: 10.1098/rstb.2006.1914.

[89] P. B. Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, "Toward Automatic Reconstruction of a Highly Resolved Tree of Life," *Science (80-.).*, vol. 311, no. 5765, pp. 1283–1287, 2006, doi: 10.1126/science.1123061.

[90] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004, doi: 10.1093/nar/gkh340.

[91] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets," *Mol. Biol. Evol.*, vol. 33, no. 7, pp. 1870–1874, 2016, doi: 10.1093/molbev/msw054.

[92] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, and A. Manuscript, "UpSet: Visualization of Intersecting Sets Europe PMC Funders Group," *IEEE Trans Vis Comput Graph*, vol. 20, no. 12, pp. 1983–1992, 2014, doi: 10.1109/TVCG.2014.2346248.UpSet.

[93] A. D'hont *et al.*, "The banana (Musa acuminata) genome and the evolution of monocotyledonous plants," *Nature*, vol. 488, no. 7410, pp. 213–217, 2012, doi: 10.1038/nature11241.

[94] S. S. Costa, L. C. Guimarães, A. Silva, S. C. Soares, and R. A. Baraúna, "First Steps in the Analysis of Prokaryotic Pan-Genomes," *Bioinform. Biol. Insights*, vol. 14, 2020, doi: 10.1177/1177932220938064.

[95] H. Cai, J. Gu, and Y. Wang, "Core genome components and lineage specific expansions in malaria parasites Plasmodium," *BMC Genomics*, vol. 11, no. SUPPL. 3, p. S13, 2010, doi: 10.1186/1471-2164-11-S3-S13.

[96] S. Batinovic *et al.*, "An exported protein-interacting complex involved in the trafficking of virulence determinants in Plasmodium-infected erythrocytes," *Nat. Commun.*, vol. 8, no. May, pp. 1–14, 2017, doi: 10.1038/ncomms16044.

[97] M. Morita *et al.*, "PV1, a novel Plasmodium falciparum merozoite dense granule protein, interacts with exported protein in infected erythrocytes," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1038/s41598-018-22026-0.

[98] E. J. Remarque, B. W. Faber, C. H. M. Kocken, and A. W. Thomas, "Apical membrane antigen 1: a malaria vaccine candidate in review," *Trends Parasitol.*, vol. 24, no. 2, pp. 74–84, 2008, doi: 10.1016/j.pt.2007.12.002.

[99] G. H. Mitchell, A. W. Thomas, G. Margos, A. R. Dluzewski, and L. H. Bannister, "Apical Membrane Antigen 1, a Major Malaria Vaccine Candidate, Mediates the

Close Attachment of Invasive Merozoites to Host Red Blood Cells," *Infect. Immun.*, vol. 72, no. 1, pp. 154–158, 2004, doi: 10.1128/IAI.72.1.154-158.2004.

Appendix

	Query name	Hit name	% identity	Length	Mis-matches	No.of gaps	Start query	End query	Start hit	End hit	e-value	bitscore
1	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_0100100.1:pep	100.000	2164	0	0	1	2164	1	2164	0.0	4487
2	3D7new|PF3D7_0100100.1:pep	NF54N|W7K1L5_PLAFO	99.885	1742	2	0	1	1742	1	1742	0.0	3611
3	3D7new|PF3D7_0100100.1:pep	7G8|W7FVH1_PLAF8	60.000	1885	572	46	17	1824	29	1808	0.0	2021
4	3D7new|PF3D7_0100100.1:pep	7G8|W7FVH1_PLAF8	74.359	234	53	5	1934	2163	1802	2032	4.44e-94	343
5	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_040027700-t41_1-p1	51.310	2329	834	71	8	2163	5	2206	0.0	1932
6	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_0115700.1:pep	50.236	2327	872	67	1	2164	1	2204	0.0	1865
7	3D7new|PF3D7_0100100.1:pep	NF54F|A0A2I0BXS9_PLAFO	50.215	2326	872	67	1	2163	1	2203	0.0	1865
8	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_100044800-t41_1-p1	50.151	2319	892	70	8	2163	5	2222	0.0	1819
9	3D7new|PF3D7_0100100.1:pep	NF54F|A0A2I0BYJ1_PLAFO	48.756	2332	932	63	16	2163	15	2267	0.0	1815
10	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_1255200.1:pep	48.757	2334	932	63	16	2164	15	2269	0.0	1814
11	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_010005100-t41_1-p1	48.878	2273	958	63	8	2163	4	2189	0.0	1791
12	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_1041300.1:pep	47.901	2263	976	64	19	2164	19	2195	0.0	1771
13	3D7new|PF3D7_0100100.1:pep	NF54F|A0A2I0BRG9_PLAFO	47.878	2262	976	64	19	2163	19	2194	0.0	1770
14	3D7new|PF3D7_0100100.1:pep	HB3|A0A0L7KFN8_PLAFX	50.463	2162	854	61	143	2163	1	2086	0.0	1769
15	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_070017200-t41_1-p1	48.202	2253	951	69	17	2163	19	2161	0.0	1748
16	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_070005300-t41_1-p1	46.898	2337	931	71	17	2163	19	2235	0.0	1745
17	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_050038400-t41_1-p1	47.161	2307	926	61	17	2163	14	2187	0.0	1739
18	3D7new|PF3D7_0100100.1:pep	HB3|A0A0L7KIS2_PLAFX	46.958	2334	947	71	16	2163	15	2243	0.0	1721
19	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_080005400-t41_1-p1	47.923	2335	917	76	1	2163	1	2208	0.0	1713
20	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_1200100.1:pep	46.399	2319	917	68	18	2164	15	2179	0.0	1711
21	3D7new|PF3D7_0100100.1:pep	NF54F|A0A2I0BZ51_PLAFO	46.376	2318	917	68	18	2163	15	2178	0.0	1710
22	3D7new|PF3D7_0100100.1:pep	NF54N|W7JQ66_PLAFO	46.964	2306	932	62	16	2163	15	2187	0.0	1710
23	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_030030500-t41_1-p1	47.834	2308	927	72	1	2163	1	2176	0.0	1708
24	3D7new|PF3D7_0100100.1:pep	Palo|W4IS60_PLAFP	47.147	2331	945	73	8	2160	5	2226	0.0	1698
25	3D7new|PF3D7_0100100.1:pep	Palo|W4J5Z5_PLAFP	47.559	2294	910	69	17	2163	12	2159	0.0	1691
26	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_070005000-t41_1-p1	46.921	2306	960	69	8	2163	2	2193	0.0	1685
27	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_0421300.1:pep	46.587	2329	944	70	1	2164	1	2194	0.0	1674
28	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_140005400-t41_1-p1	46.641	2307	966	70	8	2163	2	2194	0.0	1664
29	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_070017500-t41_1-p1	46.499	2271	920	68	17	2163	17	2116	0.0	1656
30	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_030030900-t41_1-p1	46.238	2286	948	73	9	2163	7	2142	0.0	1651
31	3D7new|PF3D7_0100100.1:pep	KH01|PfKH01_010021500-t41_1-p1	46.039	2348	970	77	6	2163	7	2247	0.0	1651
32	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_120046500-t41_1-p1	46.212	2257	950	72	9	2163	7	2101	0.0	1649
33	3D7new|PF3D7_0100100.1:pep	GN01|PfGN01_100005600-t41_1-p1	46.245	2290	992	63	19	2163	22	2217	0.0	1647
34	3D7new|PF3D7_0100100.1:pep	3D7new|PF3D7_0800100.1:pep	47.077	2275	938	70	2	2164	1	2121	0.0	1647
35	3D7new|PF3D7_0100100.1:pep	NF54F|A0A2I0BX63_PLAFO	47.054	2274	938	70	2	2163	1	2120	0.0	1646

Figure 8.1: A snapshot of result file of the bidirectional blast for the proteins of P. falciparum's strain 3D7 PF_Europe 1:Netherlands'against the global pan-genome's database. Each tab-delimited output file contained query protein name, homologous hit name, percentage identity, length of the hit, mismatches, gaps, starting and ending positions of the match, e-value and bit score.

```
OG1.5_1000: 3D7new|PF3D7_0100100.1:pep 3D7new|PF3D7_0115700.1:pep 3D7new|PF3D7_0200100.1:pep 3D7new|PF3D7_0223500.1:pep
OG1.5_1001: 3D7new|PF3D7_0207400.1:pep 3D7new|PF3D7_0207500.1:pep 3D7new|PF3D7_0207600.1:pep 3D7new|PF3D7_0207700.1:pep
OG1.5_1002: 3D7new|PF3D7_0215000.1:pep 3D7new|PF3D7_0215300.1:pep 3D7new|PF3D7_0731600.1:pep 3D7new|PF3D7_1200700.1:pep
OG1.5_1003: 3D7new|PF3D7_0220800.1:pep 3D7new|PF3D7_0302200.1:pep 3D7new|PF3D7_0302500.1:pep 3D7new|PF3D7_0831600.1:pep
OG1.5_1004: 3D7new|PF3D7_1335300.1:pep 3D7new|PF3D7_1335400.1:pep 7G8|W7ET30_PLAF8 7G8|W7F3F1_PLAF8 Brazl|A0A0J9SPH0_PLA
OG1.5_1005: 3D7new|PF3D7_0818900.1:pep 3D7new|PF3D7_0831700.1:pep 3D7new|PF3D7_0917900.1:pep 3D7new|PF3D7_1134000.1:pep
OG1.5_1006: 3D7new|PF3D7_0115000.1:pep 3D7new|PF3D7_0800700.1:pep 3D7new|PF3D7_0831100.1:pep 3D7new|PF3D7_1477600.1:pep
OG1.5_1007: 3D7new|PF3D7_0917600.1:pep 3D7new|PF3D7_1030100.1:pep 7G8|W7F1U0_PLAF8 7G8|W7FBE6_PLAF8 7G8|W7FFU9_PLAF8 Bra
OG1.5_1008: MaliF|A0A024WI94_PLAFA 3D7new|PF3D7_1407800.1:pep 3D7new|PF3D7_1407900.1:pep 3D7new|PF3D7_1408000.1:pep 3D7n
OG1.5_1009: 3D7new|PF3D7_1115300.1:pep 3D7new|PF3D7_1115400.1:pep 3D7new|PF3D7_1115700.1:pep 7G8|W7EYY3_PLAF8 7G8|W7F5J1
OG1.5_1010: Brazl|A0A0J9SJV3_PLAVI Mauri|A0A0J9T8Q5_PLAVI Mauri|A0A0J9TEW1_PLAVI Mauri|A0A0J9THI3_PLAVI Mauri|A0A0J9TJH2
OG1.5_1011: 3D7new|PF3D7_0208900.1:pep 7G8|W7FLD9_PLAF8 Brazl|A0A0J9SZ85_PLAVI CAMP|A0A024XFS2_PLAFC Dd2|A0A0L7LW99_PLAI
OG1.5_1012: 3D7new|PF3D7_0424400.1:pep 3D7new|PF3D7_0830800.1:pep 7G8|W7FKS6_PLAF8 7G8|W7FNL7_PLAF8 CAMP|A0A024XBD6_PLAI
OG1.5_1013: 3D7new|PF3D7_0112200.1:pep 3D7new|PF3D7_1229100.1:pep 7G8|W7FBD2_PLAF8 7G8|W7FK44_PLAF8 Brazl|A0A0J9SN43_PLA
OG1.5_1014: 3D7new|PF3D7_0413900.1:pep 7G8|W7F6M5_PLAF8 Brazl|A0A0J9VM79_PLAVI CAMP|A0A024XDX0_PLAFC Dd2|A0A0L7M4F5_PLAI
OG1.5_1015: 3D7new|PF3D7_0525100.1:pep 7G8|W7FHA8_PLAF8 Brazl|A0A0J9STM4_PLAVI CAMP|A0A024XC96_PLAFC Dd2|A0A0L7M2A7_PLAI
OG1.5_1016: 3D7new|PF3D7_1005600.1:pep 7G8|W7FCM3_PLAF8 Brazl|A0A0J9SV35_PLAVI CAMP|A0A024X869_PLAFC Dd2|A0A0L7LXA3_PLAI
OG1.5_1017: 3D7new|PF3D7_1018200.1:pep 7G8|W7F2R2_PLAF8 7G8|W7FCF1_PLAF8 CAMP|A0A024X7D5_PLAFC CAMP|A0A024X8E7_PLAFC Dd2
OG1.5_1018: CAMP|A0A024X6X6_PLAFC GN01|PfGN01_120007300-t41_1-p1 HB3|A0A0L7KJ84_PLAFX KH01|PfKH01_000031900-t41_1-p1 NF5
OG1.5_1019: 3D7new|PF3D7_0310300.1:pep 7G8|W7FEP9_PLAF8 Brazl|A0A0J9SV81_PLAVI CAMP|A0A024XF39_PLAFC Dd2|A0A0L7M5Z7_PLAI
```

Figure 8.2: A snapshot of the groups.txt output file representing each row as one COG containing a protein family.

Table 8.1: Number of orthologue pairs present in corresponding strains of the Asian dataset.

	KH01	PvT01	India	PvP01	PcC01	Korea	Viet	CAMP	FCH4	Dd2
KH01		4106	4133	4207	4096	4152	6892	6341	6137	5634
PvT01	4106		5715	5395	5791	5645	4011	4038	3580	2955
India	4133	5715		5474	5830	6030	4066	4085	3634	2992
PvP01	4207	5395	5474		5452	5482	4087	4117	3658	3008
PcC01	4096	5791	5830	5452		5789	3999	4031	3573	2948
Korea	4152	5645	6030	5482	5789		4058	4068	3630	2983
Viet	6892	4011	4066	4087	3999	4058		6374	5934	5058
CAMP	6341	4038	4085	4117	4031	4068	6374		5683	4676
FCH4	6137	3580	3634	3658	3573	3630	5934	5683		4511
Dd2	5634	2955	2992	3008	2948	2983	5058	4676	4511	

Table 8.2: Number of in-paralog pairs present in corresponding strains of the Asiandataset.

	PvP01	PvT01	Korea	Dd2	FCH4	Viet	CAMP	KH01	PcC01	India
PvP01	82									
PvT01		172								
Korea			243							
Dd2				992						
FCH4					370					
Viet						614				
CAMP							380			
KH01								1508		
PcC01									396	
India										197

Table 8.3: Number of co-orthologue pairs present in corresponding strains of the Asian dataset.

	Viet	FCH4	India	Korea	PcC01	CAMP	PvT01	Dd2	KH01	PvP01
Viet		2878	67	63	55	3503	55	3090	4268	50
FCH4	2878		58	56	48	2328	49	2083	2579	45
India	67	58		139	241	49	102	36	41	43
Korea	63	56	139		194	44	114	35	40	45
PcC01	55	48	241	194		40	160	42	42	84
CAMP	3503	2328	49	44	40		37	2513	3571	32
PvT01	55	49	102	114	160	37		25	25	32
Dd2	3090	2083	36	35	42	2513	25		2460	24
KH01	4268	2579	41	40	42	3571	25	2460		23
PvP01	50	45	43	45	84	32	32	24	23	

Table 8.4: Number of orthologue pairs present in corresponding strains of the Asian excluding India dataset

.

	Korae	KH01	PvP01	Viet	Dd2	PvT01	CAMP	FCH4	PcC01
Korea		4152	5482	4058	2983	5645	4068	3630	5789
KH01	4152		4207	6892	5634	4106	6342	6137	4096
PvP01	5482	4207		4087	3008	5395	4117	3658	5452
Viet	4058	6892	4087		5058	4011	6374	5935	3999
Dd2	2983	5634	3008	5058		2955	4677	4512	2948
PvT01	5645	4106	5395	4011	2955		4038	3581	5791
CAMP	4068	6342	4117	6374	4677	4038		5684	4031
FCH4	3630	6137	3658	5935	4512	3581	5684		3573
PcC01	5789	4096	5452	3999	2948	5791	4031	3573	

Table 8.5: Number of in-paralog pairs present in corresponding strains of the Asian excluding India dataset.

	PvP01	CAMP	Dd2	FCH4	PvT01	KH01	Viet	PcC01	Korea
PvP01	82								
CAMP		380							
Dd2			992						
FCH4				370					
PvT01					180				
KH01						1508			
Viet							614		
PcC01								439	
Korea									289

Table 8.6: Number of co-orthologue pairs present in corresponding strains of the Asian excluding India dataset
.

	FCH4	PvT01	PcC01	Viet	CAMP	Korea	Dd2	KH01	PvP01
FCH4		49	48	2877	2342	56	2082	2579	45
PvT01	49		179	55	37	129	25	25	32
PcC01	48	179		55	40	229	42	42	87
Viet	2877	55	55		3505	63	3091	4268	50
CAMP	2342	37	40	3505		44	2513	3570	32
Korea	56	129	229	63	44		35	40	49
Dd2	2082	25	42	3091	2513	35		2461	24
KH01	2579	25	42	4268	3570	40	2461		23
PvP01	45	32	87	50	32	49	24	23	

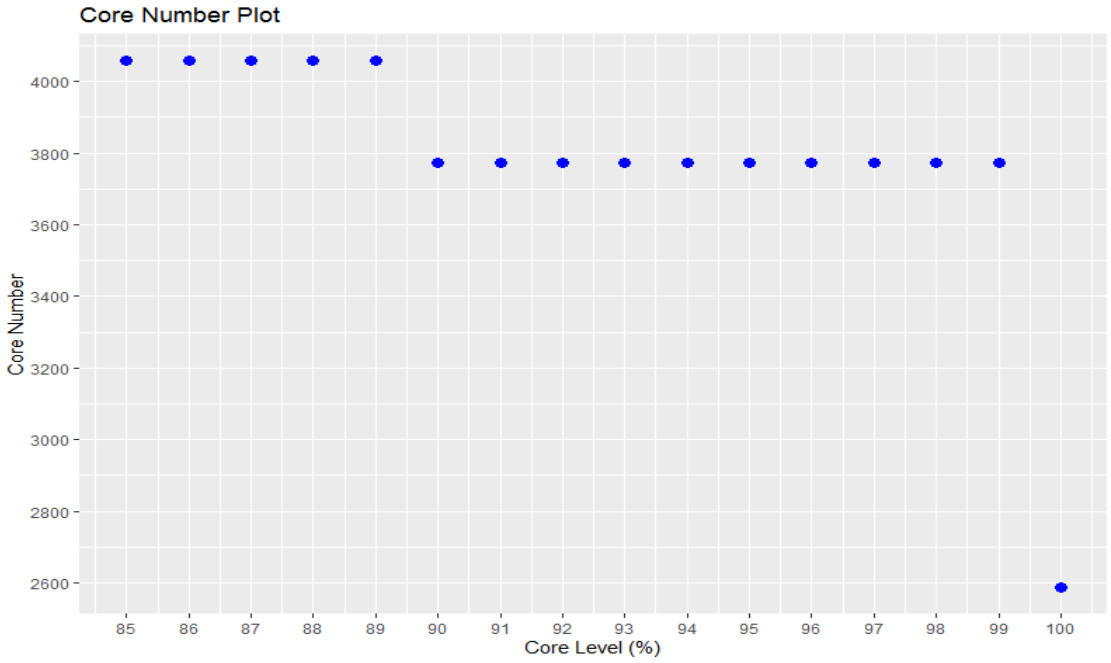

Figure 8.3: Core Number Plot of the Asian dataset.

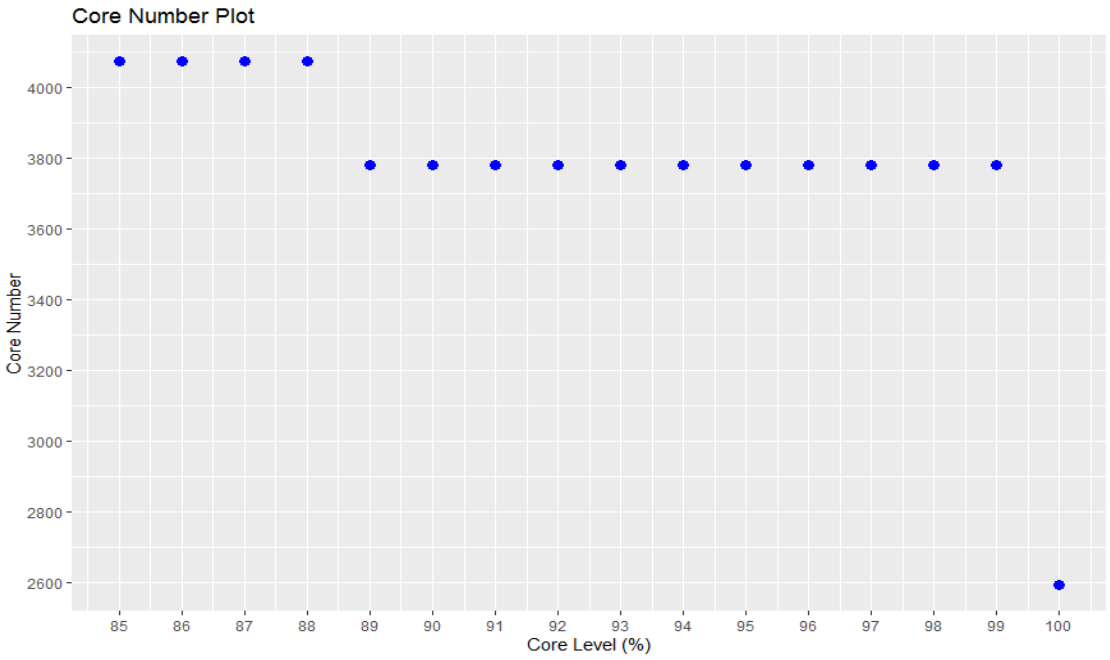

Figure 8.4: Core Number Plot of the Asian excluding India dataset.

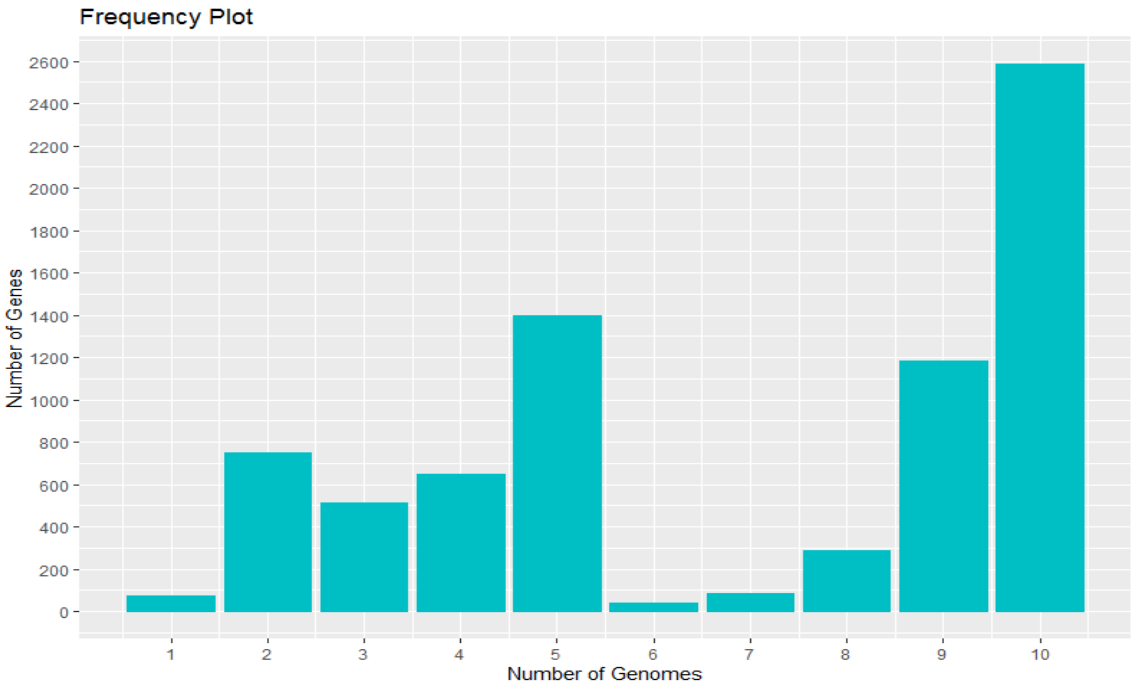

Figure 8.5: Frequency Plot of the Asian dataset.

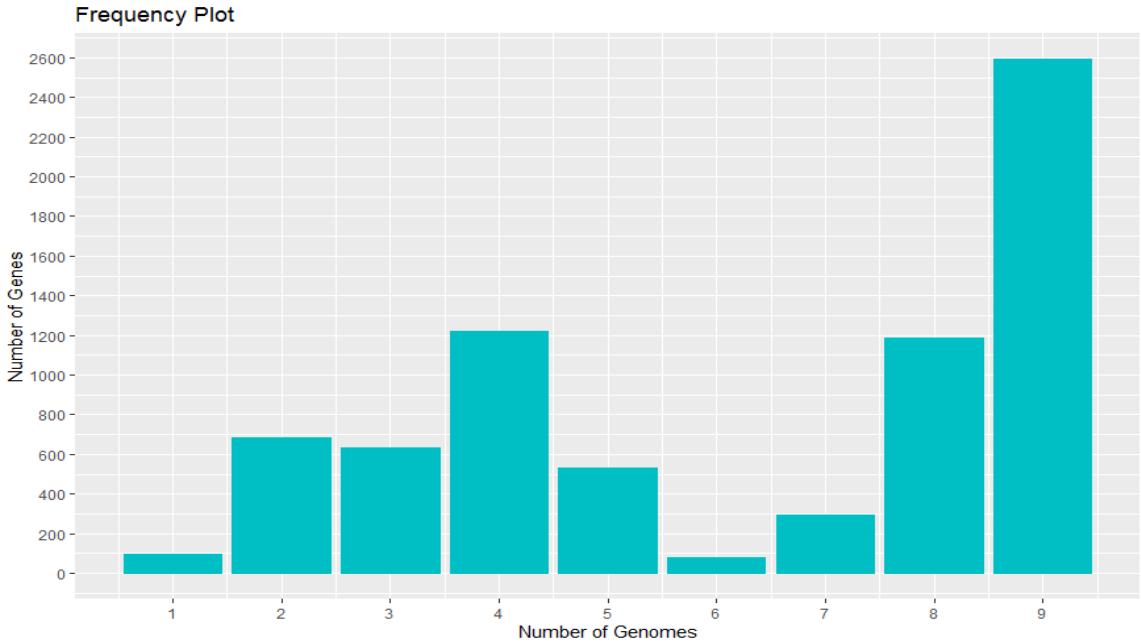

Figure 8.6: Frequency Plot of the Asian excluding India dataset.

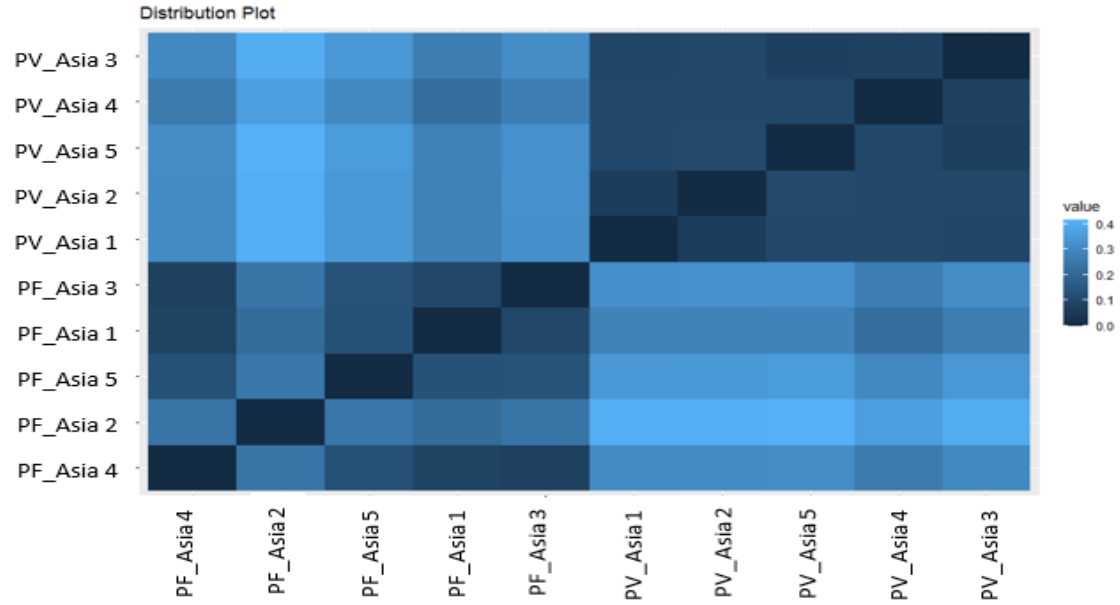

Figure 8.7: Heat Map of the Asian dataset.

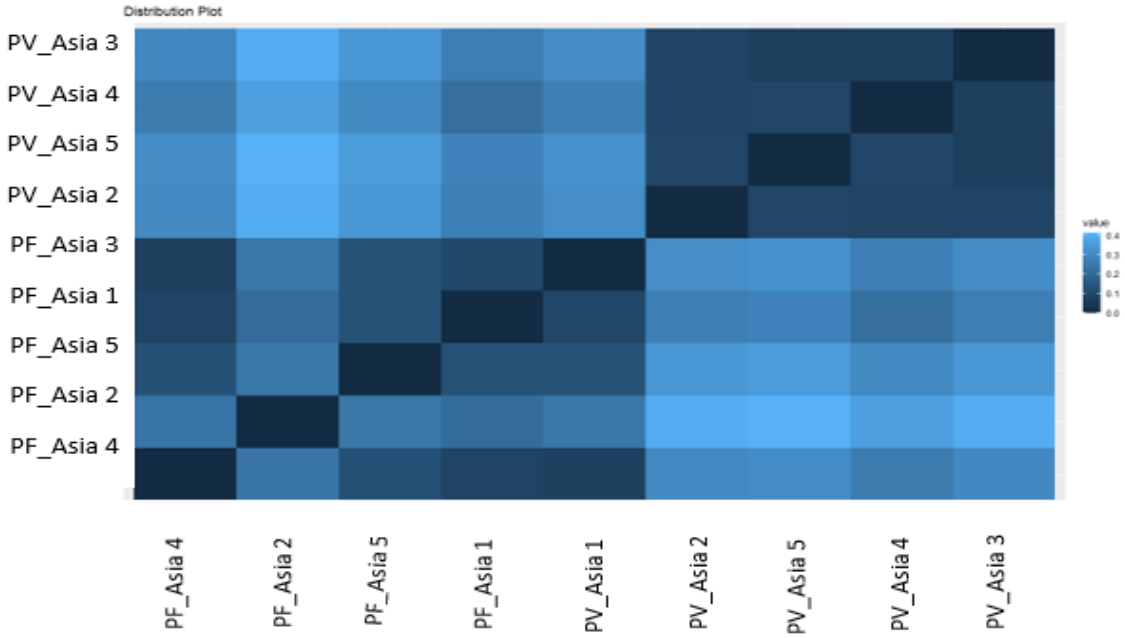

Figure 8.8: Heat Map of the Asian excluding India dataset.