COMPARATIVE GENOMIC ANALYSIS OF PROTEOME REDUCTION IN THE APICOMPLEXANS

By

Mohammad Zillur Rahman

A dissertation submitted to the
DEPARTMENT OF BIOLOGY
FACULTY OF NATURAL SCIENCES
UNIVERSITY OF PUERTO RICO
RÍO PIEDRAS CAMPUS

In partial fulfillment of the requirements for the degree of
DOCTOR IN PHILOSOPHY

May 2020
San Juan, Puerto Rico

This dissertation has been accepted by faculty of the:

DEPARTMENT OF BIOLOGY
FACULTY OF NATURAL SCIENCES
UNIVERSITY OF PUERTO RICO
RÍO PIEDRAS CAMPUS

In partial fulfillment of the requirements for the degree of
DOCTOR IN PHILOSOPHY

Dissertation Committee:

_____
Steven E Massey, Ph.D., Advisor

_____
Tugrul Giray, Ph.D.

_____
Clifford Jaylen Louime, Ph.D.

_____
Jose Agosto Rivera, Ph.D.

_____
Adelfa E. Serrano, Ph.D.

Table of Contents

Abstract

Apicomplexans are alveolate parasites which include *Plasmodium falciparum*, the main cause of malaria, one of the world's biggest killers from infectious disease. Apicomplexans are characterized by a reduction in proteome size, which appears to result from metabolic and functional simplification, commensurate with their parasitic lifestyle. However, other factors may also help to explain gene loss such as population bottlenecks experienced during transmission, and the effect of reducing the overall genomic information content. The latter constitutes an 'informational constraint', which is proposed to exert a selective pressure to evolve and maintain genes involved in informational fidelity and error correction, proportional to the quantity of information in the genome (which approximates to proteome size).

In this dissertation, the dynamics of gene loss is examined in 41 Apicomplexan genomes using orthogroup analysis. This work shows that loss of genes involved in amino acid metabolism and steroid biosynthesis can be explained by metabolic redundancy with the host. There is a marked tendency to lose DNA repair genes as proteome size is reduced. This may be explained by a reduction in size of the informational constraint and can help to explain elevated mutation rates in pathogens with reduced genome size.

Effective population size (N_e) has a direct contribution to evolutionary changes. In these species, N_e is not well studied due to the morphological and genomic complexity. In order to measure N_e, model species *P. falciparum* is chosen whose mutation rate and generation time are already predicted. MSMC analysis indicates a recent bottleneck, consistent with predictions generated using allele-based population genetics approaches, implying that relaxed selection pressure due to reduced population size might have contributed to gene loss. However, the non-randomness of pathways that are lost challenges this scenario.

Malaria is an ancient disease and yet, there is no effective cure or prevention. This study looks for new antimalarial targets to identify unique orthogroups in malaria causing *Plasmodium* species that infect humans, with a high proportion of membrane associated proteins. Thus, orthogroup analysis appears useful for identifying novel candidate pathogenic factors in parasites, when there is a wide sample of genomes available.

In terms of biodiversity, Apicomplexa has many unclear taxonomic structures. In this study, a statistically robust phylogeny is reconstructed by concatenating 522 genes from the core Apicomplexan genome which account for 6068 amino acid sequences. Different biases and pitfalls among alignments and phylogeny inference methods are also discussed.

Lastly, this study provides a foundation for future experimental research, specific and comparative analysis of Apicomplexan proteomes.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Prof. Steven E Massey for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Dr. Tugrul Giray, Prof. Dr. Clifford Louime, Prof. Dr. Adelfa E. Serrano and Prof. Dr. Jose Agosto Rivera for their insightful comments, encouragement, and guidance which fostered me to widen my research from various perspectives. I also want to thank to Prof. Dr. Jose Carlos for giving me the opportunity to work with HPC. I express my gratitude to Dr. Julie Dutil for giving me the opportunity of practical training.

I would like to thank Natural Science faculties, office staffs, my fellow classmates and San Juan community who helped me to adapt here in this island.

Last but not the least, I would like to thank my family: my parents and to my brothers and sisters for supporting me spiritually throughout writing this thesis and my life in general.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Prof. Steven E Massey, Prof. Dr. Tugrul Giray, and Prof. Dr. Jose Agosto Rivera of the Department of Biology and Prof. Dr. Clifford Louime of the Department of Environmental Science and Prof. Dr. Adelfa E. Serrano of the Department of Microbiology, Medical Sciences Campus. The ID mapping method in chapter 2 was conducted by Dieunel Derilus of the Department of Environmental Science. All other work conducted for the dissertation was completed by the student independently.

This work was supported by the Biology Department, UPR – Rio Piedras, and DoD grant W911NF-11-1-0218, courtesy Dr Kai Gribenow (Chemistry Department, UPR – Rio Piedras). This work was also supported by an Institutional Development Award (IDeA) INBRE Grant Number P20GM103475 from the National Institute of General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH), and the Bioinformatics Research Core of INBRE.

List of Tables

List of Figures

Chapter 1

General Introduction of Apicomplexa and Malaria

1.1: Rationale for the Study

In 1996, the *Plasmodium falciparum* genome sequence project was launched and was subsequently completed and published in 2002, opening a new horizon in malaria research [1]. Six years is a long time to sequence this relatively small genome of ~23mb. In comparison human chromosome 1 is ~249 mb. The slow progress had three principal reasons. First, for this genome an *Escherichia coli* construct was not established for sequencing [2]. *E. coli* system is used in clone-by-clone sequencing in which the genome is split up into chunks of ~1.5-3 kb, mapped and then inserted into Bacterial Artificial Chromosome or BAC, facilitating genome sequencing and assembly. Second, this genome is highly AT rich. That is Adenine and Thymine represent more than 80% of the nuclear genome, and this ratio is even higher in introns and intergenic regions [3],[1]. Third, due to possible contamination with host genome. Chromosomes were separated by time consuming but stable and reproducible PFGE (Pulsed Field Gel Electrophoresis) technique, which was effective to reduce host contamination. Lastly, Gardner and colleagues sequenced this genome using whole chromosome shotgun sequencing method at a time when whole genome sequencing was not pragmatic or cost-effective.

Extremely AT or GC-rich regions in the genomes produce a reduced number of combinations of nucleotides with more overlaps which are difficult to clone for sequencers [4], [5]. AT-rich genome can produce biased Polymerase chain reaction (PCR) amplification, the assembly of primary reads and functional annotations can be problematic and the next-generation sequencing (NGS) methods can produce biased results [6], [7], [8],[9], [10]. Though some NGS technologies claim to be free from these types of biases [11],[12].

To get rid of host genome parts, scientists used different strategies. The principle of these strategies, is using a mixture of parasite and host genomes and treating them with some enzymes and reagents that specially affect or react or destabilize the host parts in the genome [13],[14],[15],[16],[17]. Whole genome capture method was

useful for *P. vivax* genome sequencing using cryopreservation. The limitation of this method is that the laboratory needs to be near the collection of field sample [18]. After sequencing, different bioinformatics tools can be used to remove host parts from the reads [19],[20],[21].

With the help of NGS technologies, within only one and half decade, there are now more than 78 genomes of the phylum Apicomplexa (*P. falciparum* belongs to Apicomplexa) in the latest release of EuPathDB (39th release, August 2018). This count does not include genomes of different strains within the 78 lineages. This achievement is a remarkable advance in exploring parasite genomes [22].

Gardner *et al.* predicted ~5300 protein-coding genes, and in the latest release of PlasmoDB (part of EuPathDB), there are 5548 open reading frames (ORFs) in *P. falciparum*. In the first published genome, there were 60.9 % hypothetical proteins due to lack of experimental evidence of in *vivo* expression. In the latest release, 35.6% proteins are of unknown function, 6 proteins are hypothetical, and 37.6% of proteins are putative. Putative proteins are those where function is not known but these have sequence similarity to the conserved regions of known proteins [22].

In other Apicomplexa species, the scenario is similar. This study focuses on 43 proteomes from the 33rd release of the EuPathDB. These belong to 41 Apicomplexa lineages and 2 related outgroups from Chromerids. In these taxa, 44.63% genes were annotated as hypothetical proteins, 9.36% of genes were unknown function and 25.09% of the genes correspond to putative proteins. In the same release, only six genes from *C. velia*, one from *E. mitis* and four from *V. brassicaformis* were annotated as pathogenesis-related proteins. These calculations were done using string search in the whole proteome fasta files. In Table 1.1, the unresolved part of 43 proteome data are summarized. In other words, the number of sequences increased yet there is a significant gap in understanding the biology of these parasites using genome sequences.

Comparative genomics analyses among closely related and distantly related species can help us understand the biology of the organisms. For instance, *Plasmodium* and *Cryptosporidium* comparative analysis divulged their adaptation through gene loss and gene innovation. This type of comparison is useful to find lineage-specific pathways [23]. Comparing six Plasmodium species, researchers found genes that are specific for different hosts and *P. falciparum*-specific genes that are involved in human virulence [24]. Relative comparison between model species from distantly related organisms can be helpful to elucidate complexity in the observed species. Many of the genes from *P. falciparum* are involved in host cell invasion and host immune evasion while fewer genes are involved in metabolism compared to another model eukaryote *Saccharomyces cerevisiae*. *P. falciparum* has less than half the number of metabolic genes compared to *S. cerevisiae* [25].

One problem with comparative genomics arises from the complex life cycle of Apicomplexa parasites. They have many stages in their life cycles. Different stages may have differentially expressed genes. DNA/RNA sequencing of different life cycle stages is useful to understand Apicomplexan biology. It is challenging to extract human *Plasmodium* in certain life stages to sequence. However, it is possible to purify rodent malaria parasites in many life stages, including liver stages. These can be studied in comparison to other species to illuminate similar stages of human parasites [26].

Comparing six *Plasmodium* species, Carlton *et al.* showed that coding regions of different organisms have more similarity than non-coding regions. This comparison was also useful in identifying the conservation level in the different chromosomal regions [27]. Several Apicomplexa species have an extensive range of cell types to infect. For example, *Toxoplasma* infects all cell types including RBC, WBC etc. But many of them infect only specific cell types such as *Plasmodia* infects only RBC whereas, *Cryptosporidium* infects intestinal cells. A comparison such as in the

current study that includes parasites with variable host range can elucidate mechanism and evolution of host specificity and virulence [28].

Another concern is the demonstrated divergence of the host-parasite proteomes. Understanding this divergence can be useful in reducing the number of candidate pathogenic genes. A large-scale comparison of the protein complements in 15 apicomplexan organisms revealed 9134 Apicomplexa specific protein families. This type of study was useful to understand genera-specific innovation of host cell invasion or host immune system evasion and diversity of housekeeping genes [29]. At first sight one can assume that incorporating more than the 15 genomes in a comparative study may have the potential to unearth more complexity. For instance, Templeton and colleagues (2004) reported at least 145 ‘apicomplexan' proteins including ~30 membrane and five secreted proteins [23]. These studies provide useful information in understanding parasitism. However, the authors compared only two species and orthologs (homologous gene sets) was inferred by one-to-one similarity comparison. As orthologous genes are descendants of a common ancestor, any method without reconciling evolutionary relationships can be error prone. In the current study, a tree based ortholog prediction methods is used which is discussed in the second chapter.

French and Chen (2011) showed only 16 genes that are specific for primate malaria parasites and not present in rodent malaria parasites [24]. They used a graph based ortholog prediction method (Inparanoid) to find orthologous genes among lineages of *Plasmodium*. As graph-based methods do not reconcile evolutionary trees, it is reasonable to believe that identifying orthologs using a tree-based method including more related species can reveal more species-specific as well as group-specific genes. In one of the most extensive data sets including 15 species, Wasmuth and colleagues (2009) found that major lineages of Apicomplexa share comparatively fewer parasite-specific genes [29]. These findings are discussed in the second chapter.

A comparative search of transcription factors across different lineages revealed that Apicomplexa are short of many known transcription factors along with DNA binding domains [30,31]. The shortage of transcription factors may cause significant differential expression of many genes in different life stages. Accurate orthologous gene identification may verify this result because these methods use only similarity search to find orthologous gene clusters. Protein-binding micro-array along with comparative genomics were useful in the exploration of transcription factors, DNA binding domains, their expression and involvement in the development and life stage changes [32]. Comparative evolutionary analysis of various Apicomplexa parasites demonstrated the divergence and conservation of many gene families namely cytochrome, oxidase and protease [33]. One of the primary responsibilities of membrane proteins is trafficking various materials across cells. As they are involved in transportation, membrane proteins have the potential to be pathogenic compared to cytosolic proteins. Surprisingly, several studies found fewer genes involved in membrane trafficking in Apicomplexa. The reason for the loss of the trafficking system in these organisms is not apparent [34]. Current work quantifies the number of genes in distinct functions as well as in membrane trafficking and measure the correlation with evolutionary constraint.

Apicomplexa have significantly smaller genomes compared to their algal ancestral clade [35]. Chromerids have significantly larger genome than all the Apicomplexa (Table 1.2). The reduction in genome size can affect diverse types of pathways in pathogens. For instance, metabolic pathways can be reduced in pathogens if they can use host resources [36,37]. Pathways can be lost randomly due to population bottlenecks [38,39]. DNA repair and recombination proteins also can be reduced as smaller genome has fewer nucleotides to repair [40]. This study discusses all these facets of genome reduction in Apicomplexa.

With the power of computation and new high throughput data, comparative genomics has revealed many issues that were not possible using traditional genetics. The expansion of genomics has resolved not only many fundamental

biological and evolutionary phenomenon but also provided new insight into thinking about the problems.

On the practical side, comparative genomics is also a useful tool to find antimalarial targets *in silico* [41, 27, 42, 43]. We need new antimalarial targets because there is a minimal number of antimalarial drugs and only one vaccine is currently available. The vaccine is not very effective. It is administered only to children under the age of 5, immunity is not long lasting, required multiple injections and protects only 30%. On the other hand, the malaria parasite has developed resistance to almost all antimalarial drugs commercially available. Malaria vaccine initiative and GlaxoSmithKline collaboratively developed the only available vaccine RTSS which is supported by Bill and Melinda Gates Foundation [44]. In a seven-year-long efficacy study of RTSS vaccine, it was found that this vaccine is initially protective, but efficacy declines in later years with the increased exposure of parasites [45].

Available drugs are quinine and its derivatives, artemisinin compounds, tetracycline, doxycycline, clindamycin, sulfadoxine, lumefantrine, and halofantrine. Some of these medications are limited in efficacy, safety, and dosage. In addition, most of these drugs are found to elicit resistance in parasites in most cases [46]. Understanding the molecular target or mechanism of the drug will help us to elucidate the evolution of drug resistance in parasites.

Parasites produce hemozoin or malaria pigment as a byproduct of hemoglobin digestion. The parasite consumes nutrition from erythrocytes, depositing the remains into the food vacuole in the trophic stage and release large amount of heme. The free heme causes oxidative stress and is toxic for the parasites. Hemozoin prevents the release of free heme which is important for the survival of parasite. Quinine compounds interfere with the conversion of free heme to hemozoin to kill the parasite. However, resistant parasites expel quinine before it reaches to the concentration which is sufficient to inhibit hemozoin formation [47]. Spontaneous mutations were often observed in the resistant parasites. The most

used and inexpensive drug is chloroquine (CQ), an aminoquinoline. Parasites developed resistance to CQ. Resistance occurs by mutations in membrane transporters pfCRT and MDR and other genes [48],[49]. Sulfa drugs and other antibiotic targets folic acid biosynthesis pathway in these parasites. Mutations were found in dihydrofolate reductase and dihydropteroate synthase in the resistant species. Some drugs such as atovaquone targets the electron transport chain of the parasites. Mutations (single-point) were reported in cytochrome-b genes in the resistance organism [50],[51],[52]. So, currently we have only 4-5 molecular targets against malaria parasites whereas, there are ~500 molecular targets against bacteria [53]. This research will search for possible new antimalarial targets for future experimental research.

It is reasonable that we have fewer drug targets against these organisms because they are eukaryotes, which rule out most of the prokaryotic molecular targets. Due to their lifestyle, parasites typically tend to be morphologically and genetically simpler, but they can be complex in terms of life cycle, evolution, host invasion and host immune system evasion. Understanding evolutionary relationships using phylogeny can help us explain the simplicity and complexity of organism's biology. In this thesis, Apicomplexa phylogeny is reconstructed to explain their relationships and complexity.

1.2: Thesis Aim and Objectives

There are three aims of this thesis. The first aim is to identify apicomplexan genes that are pathogenic for humans, to provide a foundation for new experimental studies of developing drug and vaccines against these organisms. The second aim is to measure the contributions of evolutionary constraints namely proteome size and effective population size on the genome reduction in these species. A third aim is to conduct a phylogenomic analysis to measure evolutionary relationships among these organisms.

To find genes involved in pathogenicity, I focus on genes that are membrane bound and specific to pathogenic organisms. The pathogenicity can be phylogenetically related or may arise independently. To reach this goal, first, a tree-based ortholog identification method is employed to find all orthologs from the whole proteome. Then the orthologs are annotated using blast and KEGG database. The dynamics of function loss and the effect of evolutionary constraints on the functions were also addressed. After the identification of species-specific orthogroups, the cellular components, molecular functions, and biological processes (aka gene ontology) of those orthogroups were assigned. To achieve the second aim, the proteome size in each of the observed species were calculated and its correlations with all the mapped functional orthologs was measured. After that, the effective population size of the model organism *Plasmodium falciparum* is measured using a whole genome coalescence method for a given period to explain the effect of proteome size and effective population on genome reduction. The correlation analysis between proteome size and abundance of functions is scrutinized using Phylogenetically Independent Contrasts. A phylogenomic analysis is also conducted concatenating all the genes that are found among all the observed species to validate orthologs and functional assignment and to provide an evolutionary relatedness of these species. Additionally, this work provides a robust foundation to analyze the parasite's functional dynamics.

1.3: Thesis structure:

Mathematically, we expect to see that, at least most of the functions will be positively correlated with genome size. These relations are explored in the second chapter. In a smaller population, many functions can be lost randomly as drift will be more effective. The effect of effective population size on genetic fidelity is discussed in the last chapter. In the third chapter, the evolutionary relationships in the observed organisms are discussed which also validates ortholog inference. The organization of these topics is summarized in Figure 1.1.

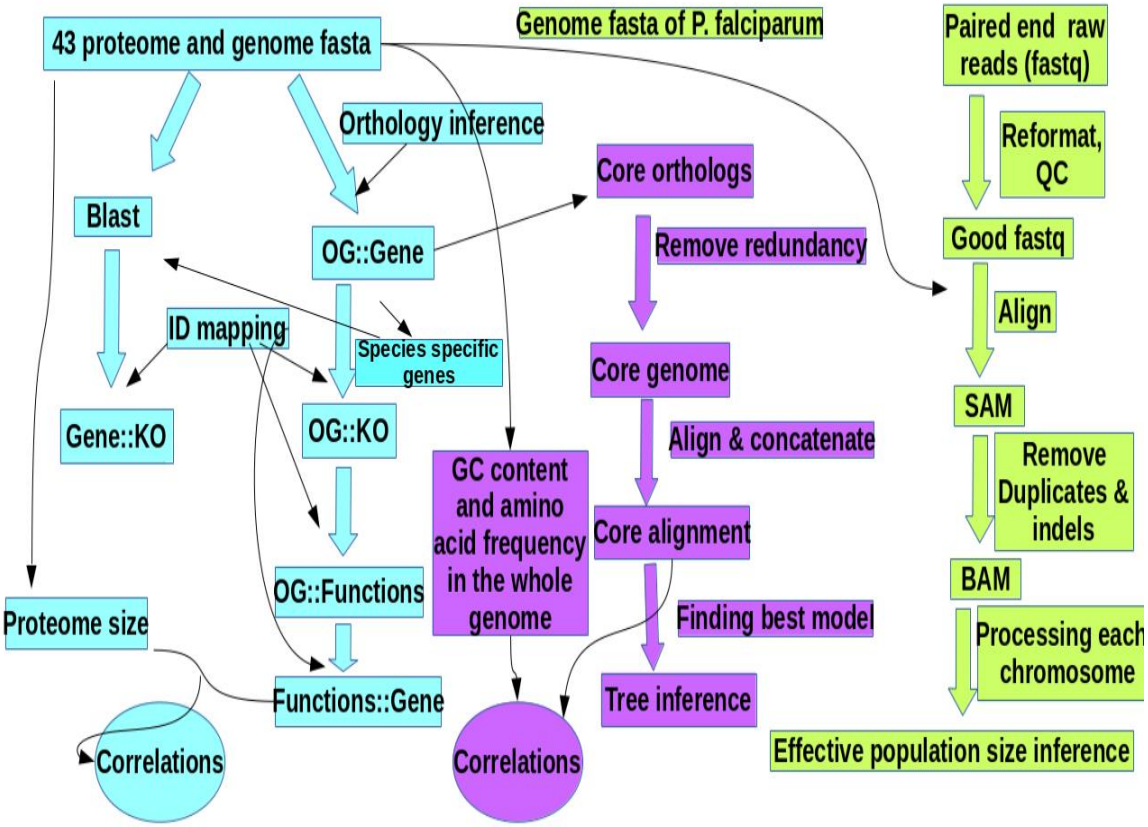

Figure 1.1: Schematic presentation of thesis workflow. KO=KEGG orthologs, OG=Orthogroups, QC=Quality control, SAM=Sequence alignment, BAM=Binary alignment. SAM and BAM are intermediate files in the pipeline. From 43 proteome, orthologs will be annotated to functions and correlations between proteome size and functional orthologs will be measured. Orthologs which have genes from all organism will be subjected for phylogeny. Finally, population size will be measured from validated SNP (Single Nucleotide Polymorphism) data.

1.4: General features of Apicomplexa

The Apicomplexa comprise one of the most abundant phyla of protozoa. As they invade one or multiple hosts, they evolved as one of the most complicated unicellular eukaryotes. Before starting the specific analysis, the complexity of their morphology, life cycle, and genomes are discussed briefly in this introductory chapter. This discussion will also articulate the motivations to study these organisms.

1.4.1: Biology and Disease

The Apicomplexa are widely distributed parasites. There are only seven genera that infect humans, and many others infect other animals like poultry, livestock, and pets. In table 1.2, pathogens with respective hosts are summarized. Organisms whose preferred environment is gastric or enteric can cause diarrhea and erythrocyte preferring pathogens cause malaria (Table 1.2).

Plasmodium causes malaria, *Cryptosporidium*, *Isospora* and *Cyclospora* cause watery diarrhea, *Toxoplasma* causes neurological problems, *Sarcocystis* causes sporadic infection, and *Babesia* causes rare zoonotic disease (fever in domestic and wild animals). *Toxoplasma* is mostly asymptomatic in immunocompetent persons. It is quite dangerous in pregnant women depending if the infection was acquired during the first trimester which may cause spontaneous abortion, stillbirth etc. *Cryptosporidium* and *Toxoplasma* are often associated with opportunistic infection in immunocompromised patients (AIDS). *Babesia, Theileria* and *Eimeria* infect cattle and poultry [54], [55], [56], [57],[58],[59]. These parasites infect pets and livestock that's why they are also economically important to study.

Apicomplexa are endoparasite alveolates. That means they have cortical alveoli which is a membrane-bound vesicle just underneath the plasma membrane. Cortical alveoli may involve in Ca^{++} storage, parasite motility and replications [60],[61]. As they are eukaryotes, these organisms have a nucleus**,** endoplasmic reticulum**,**

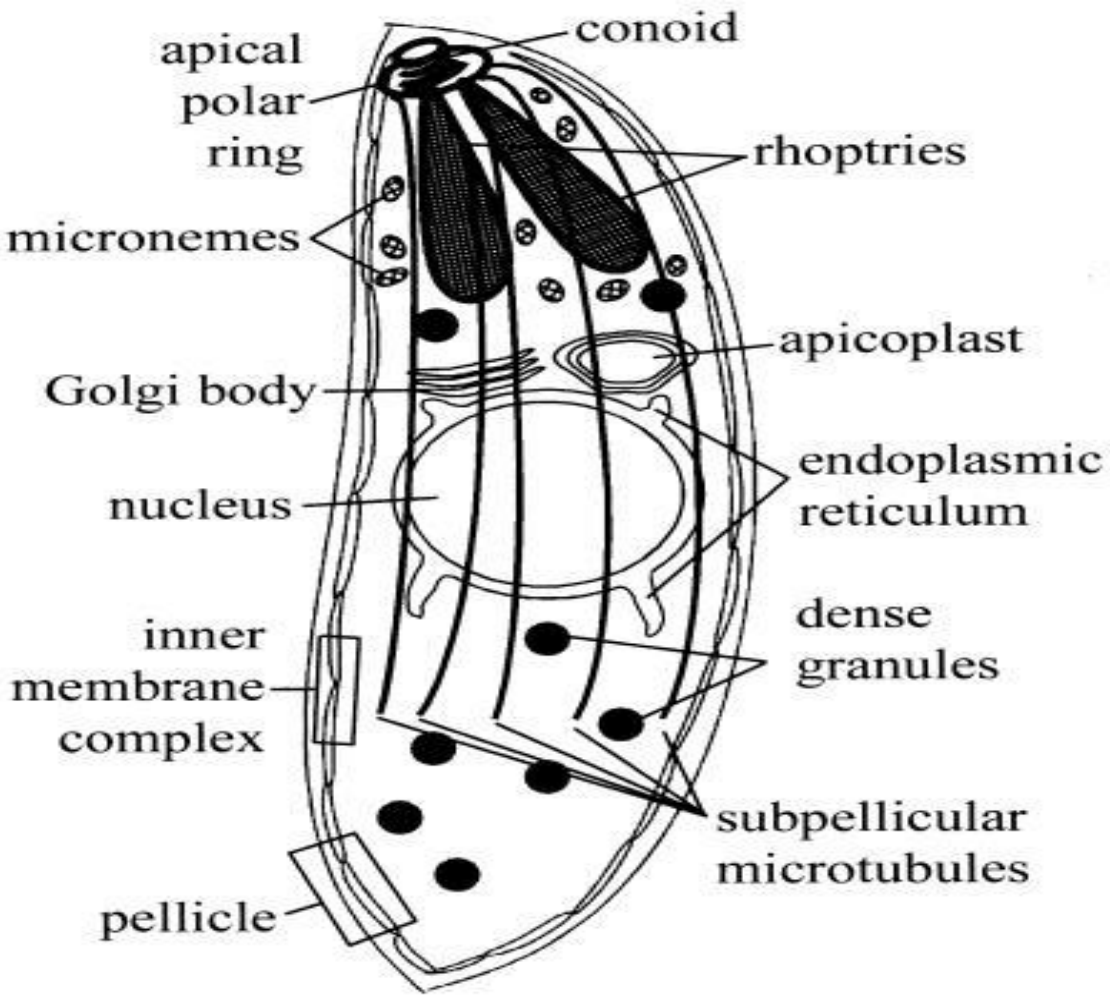

Figure 1.2: Morphology of a typical highly polarized Apicomplexa cell. This figure of Apicomplexa is adapted from [65]. A typical Apicomplexa can be a few micrometers in size. The shape of the cell can vary in different life stages.

Golgi apparatus and ribosomes like all other species of belonging to the eukaryotic domain. They also have endosymbiotic derived mitochondrion and an apicoplast. The apicoplast is a non-photosynthetic plastid, and always found in close contact with the mitochondrion among different life stages [62]. Both the apicoplast and the mitochondrion have their own genome. Predicted functions of the apicoplast include biosynthesis of fatty acids and isoprenoids, iron-sulfur clusters and heme production [62],[63], [64], [65].

Cryptosporidium has lost the apicoplast but no other Apicomplexa [66]. At the anterior part of the species, there is a group of organelles namely polar ring,

rhoptries, dense granules, and micronemes, collectively known as apical complex. At the most anterior part of the organism, there are contractile microtubule like organelles (micronemes) which are often associated with cone-shaped conoid (Figure 1.2). Recently, a novel apical compartment named monoeme, was identified in merozoite of *P. falciparum* [67].

The polar rings are involved in cell shape, and apical polarity and the conoids play a mechanical role in the invasion of host cells. Conoids are present in *Toxoplasma, Eimeria*, and *Sarcocystis* but not found in *Plasmodium* and *Theileria* [68,69]. Dense granules, micronemes, and rhoptries are secretory organelles. All the Apicomplexa has these organelles, but their number and shape vary in different genera. These three Apicomplexa specific organelles are involved in motility, adhesion to host cells, invasion of host cells, and the establishment of the parasitophorous vacuole [70, 71, 72, 73].

1.4.2: Life Cycle

Generally, members of the Apicomplexa have a complicated life cycle. Understanding their life cycle will help us to depict their adaptation and parasitism. Most of these organisms undergo a series of asexual and sexual reproduction in one or multiple hosts. The number of hosts involved, and the specific cell invaded in the life cycle varies. Some of them are monoxenous, such as, *Eimeria spp.*, *Cyclospora spp.* and *Gregarina spp. (*development occurs in a single host*)*.

They have several life stages, mainly sporozoite, merozoite, trophic stages, gametic stages, zygote and oocyst [74]. In general, sexual and asexual reproduction alternate, although some Apicomplexans lack one or the other. In *Plasmodium, Theileria, Babesia, Toxoplasma,* and *Cryptosporidium*, the sporozoite stages invades the host cell and undergoes asexual reproduction releasing merozoites which invade another host cell where a second round of asexual reproduction occurs. The specific cell type depends on the parasitic species. Sexual

differentiation follows forming gametocytes (male and female) in *Plasmodium*, piroplasms in *Theileria* and *Babesia*, bradyzoites in *Toxoplasma* and gamonts in *Cryptosporidium*. Sexual reproduction takes place in the insect vector in *Plasmodium spp.* (mosquitoes), while in ixodid ticks in *Theileria spp*. and *Babesia spp.*, followed by the formation of a zygote, and finally sporozoites which will initiate new infections.

Toxoplasma, a promiscuous parasite, invades all types of nucleated cells and sexual reproduction occurs in the intestine of the feline host. The male and female gametes fuse to produce diploid oocysts that are shed in the environment. The unsporulated oocysts of *Eimeria* are shed in the feces of a wide variety of vertebrate animals and sporulate in adequate conditions becoming infective. Within the vertebrate host, sporozoites are released in the gut, undergo asexual and sexual reproduction with formation of a zygote, which develops into unsporulated oocysts discharged in the feces. Similarly, *Cyclospora spp.*, (*C. cayetanensis* in human), transmission occurs by ingestion of oocysts. Gregarines are a diverse group of basal Apicomplexans parasitizing freshwater and terrestrial invertebrates. Their transmission occurs by oral ingestion of oocysts with cycles of asexual and sexual reproduction within the host [75,76]. The effect of life cycle complexity (mono vs heteroxenous) on pathway abundance is addressed in the 2nd chapter.

The size and shape of the cell and abundance of surface proteins in each stage differ significantly which helps them to evade the immune system. For example, *Plasmodium* parasites possess two gene families for antigenic variation (var genes and rifin genes) to escape immune responses. The *P. falciparum* life cycle and key differences with other human-infecting parasites will be discussed to get insight about virulence and their population. Throughout the life cycle, these organisms undergo massive population loss (bottle-necks) and revival [77, 78].
The sporozoite is the infective stage which is produced from a multinucleated oocyst (in *Plasmodium*) or zygote (*Toxoplasma*) through asexual reproduction

(sporogony). Sporogony produces single nucleated haploid invasive sporozoites from mature oocysts. Sporozoites travel to the saliva of a female A*nopheles* mosquito to invade human in the next blood meal.

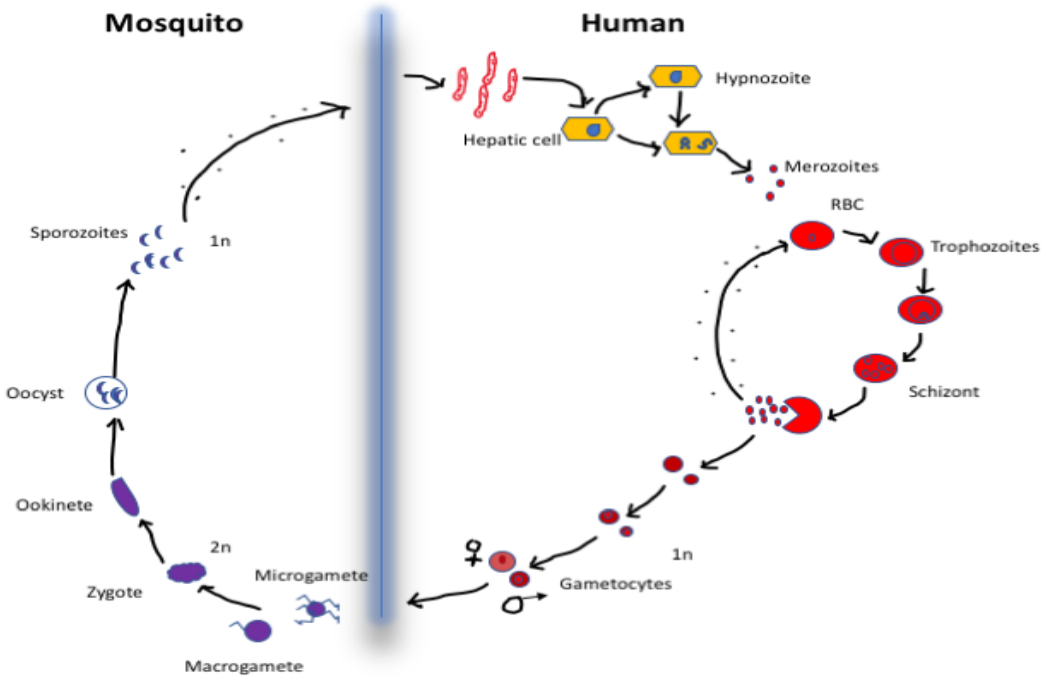

Figure 1.3: Alternating hosts and ploidy life cycle of *Plasmodium*. This figure was drawn using PowerPoint according to the description in https://www.malariasite.com/life-cycle/ and in literatures. Sporozoites, merozoites, and trophozoites are smaller compared to gametocytes, zygote, oocyst, and schizont. The diploid zygote develops into motile and elongated ookinete and secretes an elastic cyst wall to become round oocyst. Inside oocyst, it undergoes repeated mitotic division after nucleus reductions and produces many haploid sporozoites. These sporozoites travel to mosquito salivary gland to infect human. After infecting, they travel to the liver and develop into trophozoites or hypnozoite. Trophozoites grow in size and divide the nucleus several times to produce multinucleated schizont. These schizont releases numerous merozoites to infect red blood cells (RBC) or another liver cell. In the RBC, it gets sufficient nutrition to develop into trophozoites and undergoes erythrocytic schizogony to produce multiple merozoites. These merozoites attack other RBCs, and some of them undergo gametogenesis to make male and female gametocytes. These gametocytes travel to mosquito during blood meal to fertilize into the zygote and complete life cycle.

After the subcutaneous invasion, they travel to the liver within 30-60 minutes evading immune cells [79]. Sporozoites are small, ~15 µm in length and 1µm in width, spindle-shaped which have a distended middle part and pointed ends. They can

travel to both mosquito and vertebrate hosts [80]. Rhoptries secrete cytolytic enzymes which help in the invasion of the parasite into the liver cells. Many microtubule like organelles are found throughout the sporozoite stage [81], [82], [79].

After invading liver cells, sporozoites develop into trophozoites which feed on the infected cells and grow. Here, they undergo schizogony (asexual, multiple fission) to produce spherical schizont in 6-7 days. The schizont is ~40 µm in diameter and contains 2,000-40,000 merozoites. Number and size of different life stages and morphological structures varies in different species [83]. Some of the *P. vivax* and *P. ovale* sporozoites in the liver remains dormant for many days (days to years). They are known as hypnozoite and cause malaria relapse. Each schizont release thousands of merozoites which are ready to infect RBCs or liver cells again [84], [82], [85].

After invading RBCs, merozoites feed on nutrition and grow to become trophozoites. The early trophozoites create a central vacuole which pushes the cytoplasm and nucleus to the periphery and creates a ring-like structure. This stage is known as signet ring stage [86]. They consume food from RBC and become amoeboid which causes enlargement of the RBC. The RBC looks pale now because of lack of nutrition and Schuffner's dots (in *P. vivax*) are visible in the cytoplasm [87]. A developed trophozoite undergoes erythrocytic schizogony. The nucleus divides by multiple fission, and each of the divided nuclei gets surrounded by cytoplasm to produces more merozoites. These merozoites can infect an RBC again or can develop into gametocytes. In this stage, parasite breaks down hemoglobin into globin and hematin for nutrition. The digested hematin forms brown color hemozoin granules and manifests the symptoms of malaria. It takes ~10-14 days to produce symptoms after the entry of sporozoites into the body which is known as incubation period. The incubation period varies from species to species. For example, in *P. malariae*, incubation period is ~18-40 days, in *P. vivax* ~10-17, in *P. falciparum* ~8-11 and in *P. ovale* ~10-17 days [88], [89].

After completing a few erythrocytic schizogony, some merozoites develop into sexually differential gametocytes. Gametocyte production occurs in bone marrow and spleen. There are two types of gametocytes namely, macrogametocytes or female gametocytes and microgametocytes or male gametocytes. Female gametocytes are larger in size with a smaller nucleus, and male gametocytes are compared to small with a larger nucleus. There are more male gametocytes than female in number. These gametocytes wait for next mosquito bite to travel to mosquito for fertilization [90], [91].

In the mosquito crop (the expanded part of the alimentary tract), the microgametocytes divide their nucleus into eight nuclei and develop the flagellated body. Each of the flagella gets liberated from the flagellated body and produce microgamete or male gamete (exflagellation). Macrogametocytes develop into macrogamete or female gamete without any cell division. The male gametes contact with the reception cone of macrogametes and penetrate it and fuse to produce a zygote [92],[93]. This fertilization method is known as anisogamy where a fusion of two dissimilar gametes occurs.

Within a few hours of fertilization, the zygote develops into a motile and elongated ookinete. A ~316-fold loss in population occurs during the transformation from macrogametocyte to ookinete stage, and a ~100-fold loss occurs from the ookinete to oocyst stage. It penetrates the crop wall and develops an elastic cyst around it to create oocyst. Inside oocyst, nucleus undergoes reduction for repeated mitotic divisions to produce numerous nuclei (sporogony). Each of the newly divided nuclei gets surrounded by cytoplasm to produce a sporoblast. Sporoblasts are developed into sporozoites and break down the oocyst to travel to mosquito saliva [94], [95].

1.4.3: Comparative genomics of *P. falciparum*

Among all the Apicomplexa, *P. falciparum* is the most notorious and causes hundreds of thousands of deaths each year. A brief comparison of key genomic features between *P. falciparum* and *S. cerevisiae* will help us to understand the complexity of their genome. Because *S. cerevisiae* is a model unicellular eukaryote and has a closer number of genes as *P. falciparum*. Along with the nuclear genome, *P. falciparum* also have mitochondrial and apicoplast genome. They have 14 chromosomes which code for more than five thousand proteins. The smallest one is chromosome 1 (~0.64Mb), and the largest one is chromosome 14 (~3.3Mb). The gene density (Number of nucleotides/total number of protein-coding genes) is 4396.616, which is more than double compared to *S. cerevisiae* (2025.509).

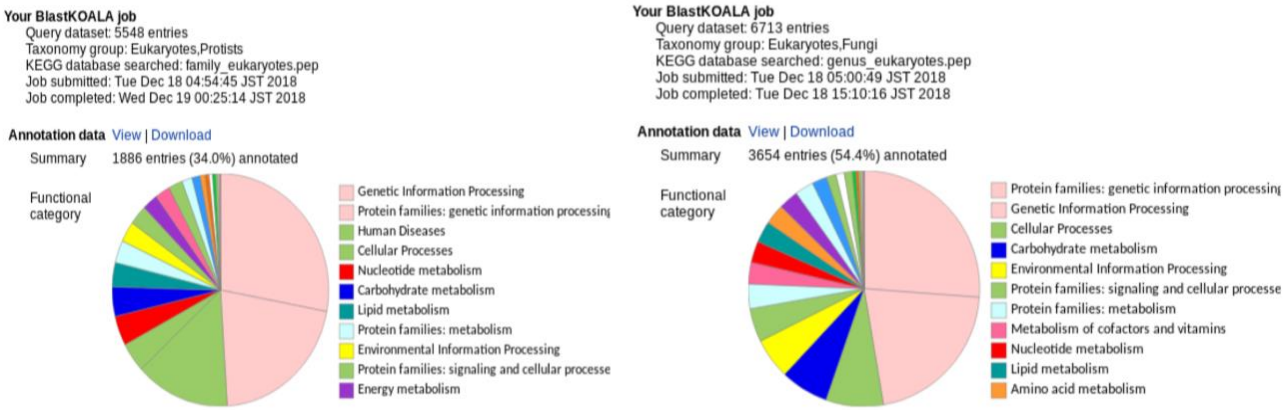

Figure 1.4: Comparison between the genomes of *S. cerevisiae* (right) and *P. falciparum* (left). This functional annotation was carried out using blastkoala. *P. falciparum* has a significantly fewer number of annotated genes. Both organisms have a substantial number of genes involved in genetic information processing. *S. cerevisiae* has a fewer number of genes for environmental information processing compared to *P. falciparum* which has more genes involved in cellular processes than *S. cerevisiae. P. falciparum* has some genes that are involved in human diseases.

However, the total number of protein-coding genes does not vary that much between these two model organisms. There are ~5307 and ~6002 proteins respectively in *P. falciparum* and *S. cerevisiae*. Though the total number of

nucleotides is almost double in *P. falciparum* (~23Mb) compared to *S. cerevisiae* (~12 Mb). *P. falciparum* has significantly fewer number of bases in its mitochondrial genome (5967), on the contrary *S. cerevisiae* has more than 14 times nucleotides (85779) in their mitochondrion. *P. falciparum* has half of the RNA genes (213) compared to *S. cerevisiae* (425). GC content also varies a lot in these two unicellular eukaryotes. Total GC is ~20% in *P. falciparum* and ~38% in *S. cerevisiae*. Extremely lower GC content is very much rare among eukaryotes. These unique features in the genome of *P. falciparum* bolster the survival and adaptation in different harsh environment [1], [25], [96]. In figure 1.4, some key genomic features of *P. falciparum* and *S. cerevisiae* are summarized.

1.4.4: Apicomplexan comparative genomics

Before starting the analysis of genomic data in detail, a concise introduction about the apicomplexan genome will give us some general clues about their genome. In table 1.1 and 1.2, some basic information about their genome, proteome and host ranges are summarized. In table 1.3, some other important features like gene density, number of the open reading frame and GC content are summarized. Apicomplexan genome structures are versatile in terms of size and compositions. The number of chromosomes is not always 14 in Apicomplexa. *Plasmodium, Toxoplasma,* and *Eimeria* all of them have 14 chromosomes while; *Cryptosporidium* has 8 and *Piroplasma* has only 4 chromosomes.

The smallest genome is *B. microti* (6.41Mb), and the largest genome is *S. neurona* (124.41Mb), excluding *V. brassicaformis* and *C. velia*. These organisms have a variable gene density range from 1795 (*B. microti)* to 8457 (*H. hammondi)*. *Piroplasma* is smaller, and they have a lower gene density compared to other Apicomplexa. A larger genome has a higher gene density (table 1.3). Apicomplexa have highly variable GC content in their genome. It ranges from ~19% (*P. gallinaceum* and *P. relictum*) to ~61% (*H. hammondi*). *Toxoplasma* and *Eimeria* have the highest GC content (~50%), *Cryptosporidium* and *Piroplasma* have an

average (~30-40%), and some of the *Plasmodium* have the lowest GC content (~20%). In the 2nd and 3rd chapter, these variables, their relationships, and potential biological impacts are explored.

Table 1.1: Number of hypothetical, unknown function and putative genes in 43 species. This information were collected using text string search from whole proteome fasta files (33rd release of EuPathDB). The strings were “hypot”, “unknow” and “putativ” for Hypothetical, Unknown function and putative respectively.

	Species	Hypothetical	Unknown function	Putative
1	*C. andersoni*	2166	0	0
2	*C. hominis*	1770	20	286
3	*C. muris*	2148	0	621
4	*C. parvum*	1523	3	254
5	*C. ubiquitum*	2223	0	4
6	*C. velia*	21480	0	10319
7	*G. niphandrodes*	3136	6375	1284
8	*V. brassicaformis*	14692	0	8720
9	*B. bigemina*	3153	4	1713
10	*B. bovis*	1846	4	876
11	*B. microti*	1006	487	219
12	*C. felis*	0	0	0
13	*T. annulata*	2128	0	1640
14	*T. orientalis*	1688	0	5
15	*T. parva*	2989	0	1068
16	*P. berghei*	2	1858	2486
17	*P. chabaudi*	6	1872	2787
18	*P. coatneyi*	44	0	17
19	*P. cynomolgi*	3143	0	27
20	*P. falciparum*	5	2054	2039

21	*P. fragile*	5178	0	0
22	*P. gaboni*	1841	112	2234
23	*P. gallinaceum*	107	1921	2885
24	*P. inui*	4763	0	0
25	*P. knowlesi*	95	1987	2779
26	*P. malariae*	64	2393	2917
27	*P. ovale*	117	2028	2900
28	*P. reichenowi*	8	2061	2675
29	*P. relictum*	128	1934	2919
30	*P. vinckei*	4086	0	0
31	*P. vivax*	2320	195	2230
32	*P. yoelii*	4	1859	2670
33	*C. cayetanensis*	4001	0	318
34	*E. acervulina*	4453	0	2096
35	*E. falciformis*	0	0	0
36	*E. maxima*	3929	0	1850
37	*E. mitis*	8078	0	1568
38	*E. necatrix*	5846	0	2223
39	*E. tenella*	5855	0	2198
40	*H. hammondi*	3966	1	790
41	*N. caninum*	4101	2	1469
42	*S. neurona*	2627	0	3
43	*T. gondii*	2833	1	1737

Table: 1.2: Genome and host range information of 43 parasites. Genome size and proteome size was calculated using bash and python from whole genome and proteome fasta files (33rd release of EuPathDB). Host and environment information was collected from different literature.

Species	Genome size (Mb)	Proteome size (Maa)	Definitive host	Intermediate host	Preferred environment
C. andersoni isolate 30847	9.09	2.25	None	Cattle	Gastric[2]
C. hominis isolate TU502_2012	9.1	2.30	None	Human	Intestine[1]
C. muris RN66	9.24	2.31	None	Mammals	Gastro-enteric[2]
C. parvum Iowa II	9.1	2.27	Mammal	Unknown	Intestine[1]
C. ubiquitum isolate 39726	8.97	2.31	None	Bovine	Intestine[3]
C. velia CCMP2878	193.89	17.14	Free living	Corals	Symbionts[4]
G. niphandrodes Unknown strain	14.01	2.92	Arthropods, nematodes annelids	Unknown	Intestine[1]
V. brassicaformis CCMP3155	72.7	12.17	Free living	Corals	Symbionts[5]
B. bigemina strain BOND	13.84	2.9	Tick	Cattle, Deer	Erythrocyte[2]
B. bovis T2Bo	8.18	1.86	Tick	Cattle, Deer	Erythrocyte[2]
B. microti strain RI	6.41	1.58	Tick	Rodents	Erythrocyte[2]
C. felis strain Winnie	9.11	2.01	Tick	Cat, Human	Enteric[2]
T. annulata strain Ankara	8.36	2.03	Tick	Bovine	Leukocyte[1]
T. orientalis strain Shintoku	9.01	2.05	Tick	Cattle	Erythocyte[6]
T. parva strain Muguga	8.35	1.9	Tick	Bovine	Leukocyte[1]
P. berghei ANKA	18.7	3.46	Mosquito	Rodent	Erythrocyte[1]
P. chabaudi chabaudi	18.9	3.49	Mosquito	Rodent	Erythrocyte[7]
P. coatneyi Hackeri	27.69	3.86	Mosquito	Monkey,	Erythrocyte[2]
P. cynomolgi strain B	26.18	3.29	Mosquito	Monkey,	Erythrocyte[2]
P. falciparum 3D7	23.33	4.19	Mosquito	Human	Erythrocyte[1]
P. fragile strain nilgiri	25.91	3.92	Mosquito	Monkey	Erythrocyte[8]
P. gaboni strain SY75	20.39	3.79	Mosquito	Chimpanzee	Erythrocyte[9]
P. gallinaceum 8A	25.03	3.73	Mosquito	Chicken	Erythrocyte[2]
P. inui San Antonio 1	27.41	3.76	Mosquito	Monkey	Erythrocyte[10]
P. knowlesi strain H	24.39	3.89	Mosquito	Monkey,	Erythrocyte[2]
P. malariae UG01	33.62	4.35	Mosquito	Monkey,	Erythrocyte[2]
P. ovale curtisi GH01	33.48	4.42	Mosquito	Human	Erythrocyte[2]

P. reichenowi CDC	24.06	4.31	Mosquito	Chimpanzee,	Erythrocyte[9]
P. relictum SGS1-like	22.61	3.65	Mosquito	Pigeon	Erythrocyte[2]
P. vinckei petteri strain CR	18.93	3.41	Mosquito	Rodent	Erythrocyte[11]
P. vivax Sal-1	27.01	3.92	Mosquito	Human	Erythrocyte[1]
P. yoelii yoelii 17X	23.08	3.81	Mosquito	Rat	Erythrocyte[1]
C. cayetanensis strain CHN_HEN01	44.03	4.26	Environment	Human	Intestine[12]
E. acervulina Houghton	45.83	4.59	Environment	Human	Intestine[2]
E. falciformis Bayer_Haberkorn_1970	43.67	4.85	Environment	Rodent	Intestine[13]
E. maxima Weybridge	45.98	4.03	Environment	Chicken	Intestine[2]
E. mitis Houghton	72.24	4.48	Environment	Chicken	Intestine[2]
E. necatrix Houghton	55.01	4.72	Environment	Chicken	Intestine[2]
E. tenella strain Houghton	51.86	4.34	Poultry	Unknown	Intestine[1]
H. hammondi strain H.H.34	67.7	6.47	Feline	avirulent	Wide range
N. caninum Liverpool	59.1	6.06	Dogs	Bovine/Equine/ Ovine	Wide range[1]
S. neurona SN3	124.41	6.97	Opossum	Equine	Leukocyte[1]
T. gondii VEG	64.52	6.63	Feline	Warm-blooded animals	Wide range[1]

Host and environment related information were collected from different sources listed below.

1. [29]
2. https://www.parasite.org.au/para-site/text/protozoa.pdf
3. [97]
4. [98]
5. [99]
6. [100]
7. [101]
8. [102]
9. [103]
10. [104]
11. [105]
12. https://www.cdc.gov/parasites/cyclosporiasis/index.html
13. [106]

Table 1.3: Comparative analysis in 43 genomes. The table was generated using text search in the whole proteome fasta file for total number of open reading frame (ORF), apicoplast and mitochondrial related proteins. Some of the organisms showed 0 mitochondria and apicoplast related proteins. It doesn't necessarily mean that they don't have any mitochondria or apicoplast related proteins, but they might poorly annotated. The search string was 'apicom' and 'mitochon' to find apicoplast and mitochondrion related annotations.

	Species	Chromosome	Total ORF	Gene density	GC content	Apicoplast	mitochondrial
1	*C. andersoni*	8	3904	2328	27	0	3
2	*C. hominis*	8	3745	2429	30	0	12
3	*C. muris*	8	3938	2346	33	0	3
4	*C. parvum*	8	3805	2391	30	0	17
5	*C. ubiquitum*	8	3766	2381	30	0	14
6	*C. velia*	4	31799	6097	48	2	170
7	*G. niphandrodes*	4	6375	2197	48	0	0
8	*V. brassicaformis*	4	23412	3105	56	1	148
9	*B. bigemina*	4	5090	2719	47	0	20
10	*B. bovis*	4	3706	2207	41	32	20
11	*B. microti*	4	3570	1795	35	8	61
12	*C. felis*	4	4323	2107	32	0	0
13	*T. annulata*	4	3796	2202	31	2	38
14	*T. orientalis*	4	4002	2251	35	6	16
15	*T. parva*	4	4082	2045	33	44	20
16	*P. berghei*	14	5076	3684	21	53	90
17	*P. chabaudi*	14	5217	3622	23	53	92
18	*P. coatneyi*	14	5516	5019	39	0	1
19	*P. cynomolgi*	14	5716	4580	26	0	40
20	*P. falciparum*	14	5548	4205	22	56	92
21	*P. fragile*	14	5672	4568	38	32	3
22	*P. gaboni*	14	5444	3745	18	56	87
23	*P. gallinaceum*	14	5307	4716	19	57	90
24	*P. inui*	14	5832	4699	34	1	11
25	*P. knowlesi*	14	5323	4582	40	44	70

26	*P. malariae*	14	6573	5114	20	48	88
27	*P. ovale*	14	7162	4674	22	46	82
28	*P. reichenowi*	14	5848	4109	21	56	92
29	*P. relictum*	14	5178	4366	19	55	91
30	*P. vinckei*	14	5160	3668	20	1	10
31	*P. vivax*	14	5552	4864	32	11	60
32	*P. yoelii*	14	6092	3788	21	53	90
33	*C. cayetanensis*	14	7455	5906	50	2	38
34	*E. acervulina*	14	6867	6673	46	0	29
35	*E. falciformis*	14	6588	6628	44	0	0
36	*E. maxima*	14	6057	7591	43	0	24
37	*E. mitis*	14	10077	7168	42	0	21
38	*E. necatrix*	14	8603	6394	49	0	27
39	*E. tenella*	14	8597	6032	52	1	32
40	*H. hammondi*	14	8005	8457	61	29	15
41	*N. caninum*	14	7125	8294	52	0	18
42	*S. neurona*	14	6965	17862	50	31	36
43	*T. gondii*	14	8410	7671	52	2	16

Chapter 2

Ortholog Inference, Pathogenic Genes, and Pathway Analysis

2.1: Abstract

Finding conserved and species-specific genes could lead us to find a potential vaccine and drug candidate in disease-causing organisms. 41 Apicomplexa parasites' proteomes are studied to find the homologous genes shared among these organisms. This study shows 2327 specific genes for malaria causing *Plasmodia* in human and chimpanzee. Most of these specific genes are in the membrane and potentially involved in host-parasite interactions. At least 98 genes known to be directly involved in pathogenicity are identified in *P. falciparum*, *P. reichenowi* and *P. gaboni*.

Additionally, the dynamics of gene loss and pathway abundance are discussed along with life cycle complexity. Many metabolic pathways that are redundant with the host are correlated with informational constraint or 'proteome size'. Genes involved in genetic fidelity such as DNA repair and recombination proteins are also found to be positively correlated with proteome size.

2.2: Introduction

Genes that have a common ancestor are known as homologous genes. If a homologous gene is found in another species that was derived from a speciation event, then it is termed an ortholog. When genes emerge from a duplication event within a genome, they are known as a paralog. Orthologs are a reliable source for gene function prediction[107]. Homologous genes can have similar functions in different lineages. Genes that were duplicated after a given speciation event are known as "in-paralogs", which were duplicated before speciation are known as out-paralogs. In-paralogs that are collectively orthologous to genes of another species are known as co-orthologs[108]. Ortholog inference can relate multiple lineages with their functional correspondence.

The orthology analysis can revolutionize the understanding of Apicomplexan biology, critical life stages, invasion, virulence, pathogenesis, and evasion from host immune response.

Even the comparison of mRNA abundance limited to *Plasmodium* species was insightful for understanding conservation, divergence, and evolution of their transcriptome. The conserved orthologs are thought to represent drug and vaccine targets [109]. Combined orthologous transcript analysis of human and rodent malaria parasites here can be used to illuminate liver stage maturation of these parasites [110].

Typically, AT-rich genomes can cause incomplete annotation for particular species. Orthologous groups identification is also very useful to annotate incompletely annotated parasite genomes [111]. Combined ortholog analysis of *Toxoplasma* and *Plasmodium* revealed conserved genes that are involved in parasite attachment and host cell invasion [112].

To date orthology analyses have provided important information. For example, analysis of orthologous genes illuminated the divergence and allelic dimorphism in merozoite surface protein1 in *Plasmodium* [113]. Orthologous gene identification using computational methods has been useful to find complex rhoptry proteins in different *Plasmodium* lineages [114]. Orthologous gene finding of variant surface proteins which are involved in evading the host immune response with the help of antigenic variation in various *Plasmodium* species has provided helpful insight into the evolution of virulence and pathogenesis [115].

As discussed in the previous chapter, Apicomplexa underwent genome reduction, and genome reduction causes a decrease in the number of genes in any given number of organisms. In that sense, a positive correlation of proteome size with most of the functions might be expected. Alternatively, the correlated functions are the effect of genome reduction. Overall, intracellular parasites are expected to lose metabolic functions redundant with host metabolism. This is reasonable; if they can use host resources to produce energy, they will cast away the metabolic genes to compensate genome reduction. In the Apicomplexans, a number of pathways have been lost which are redundant with the host, such as amino acid biosynthesis in *P. falciparum*, fatty acid biosynthesis and sterol biosynthesis (throughout the Apicomplexa) [116,36,117,37]. The inference is that the loss of diverse types of functions may have several explanations.

Orthologous and paralogous gene identification is not only helpful to elucidate biological processes but also useful to explain evolutionary and functional phenomena [118]. Divergence and evolution of sugar transporters were identified using orthology inference [119], for example.

From the above discussion, we can state that ortholog inference can unearth complex biological phenomena. The same function can independently arise in different lineages, just as bird's and bat's wings are not homologs but perform a similar function. The problem with AT-rich genome has been discussed in the introductory chapter. Another problem with AT-rich genomes is that gene prediction tools can produce biased results. Putative genes may miss microexons or noncoding sequences which affect the biological interpretations. Sequence similarity searches comparing the whole proteome can be an effective way to reduce this type of noise [120].

So, it is reasonable to believe that, ortholog identification with the whole proteome in much more closely related and distantly related species can reveal more complex biology and be helpful in finding drug and vaccine targets. Ortholog inference is useful for many purposes. The purpose of this study is to identify the number of genes in different pathways across genomes. After finding the number of genes in various functional orthologs, the variability and correlations with evolutionary constraint is measured. This chapter discusses the pattern of gene loss along with genome reductions. Moreover, this section shows species-specific and group-specific orthologous genes in order to find virulence genes.

There are two standard ways to infer an ortholog in a group of species for a given number of genes, namely graph-based methods and tree-based methods. Graph-based methods make a cluster of orthologous genes according to the similarities among them. Tree-based methods construct a phylogeny for genes and species to infer ortholog from tree topology and tree-reconciliation. Both methods depend on sequence similarity searching. Graph-based methods use the significance of similarity to infer ortholog. Tree-based

methods create alignment from significantly similar sequences to infer a phylogeny for ortholog identification.

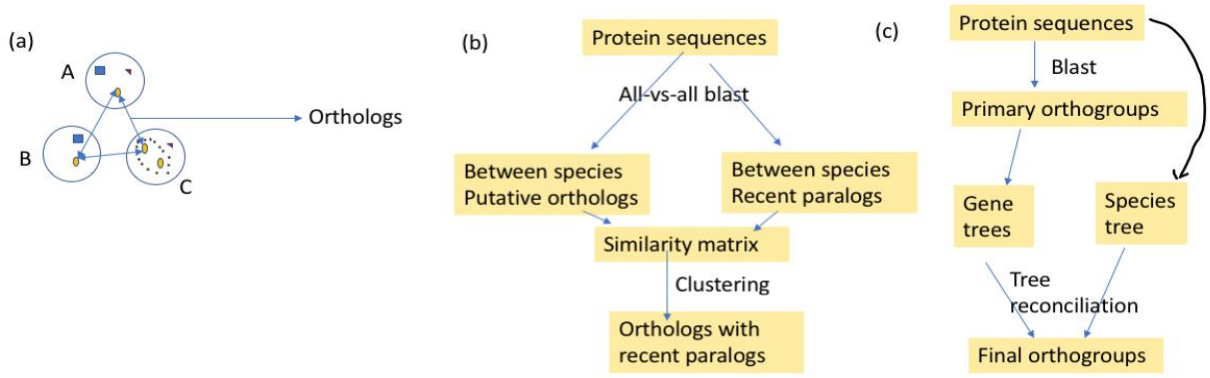

Figure 2.1.1: Basic workflow to identify orthologous genes. In (a) and (b) the graph-based method principle and workflow is summarized on the basis of description by Kristensen *et al.* (2011) and Li, L., Stoeckert, C. J., & Roos, D. S. (2003). (c) is the simplified workflow of orthology inference from gene tree and species tree reconciliation according to (Emms & Kelly, 2015) . This figure was drawn in powerpoint according to the description of the above-mentioned authors.

Graph-based methods infer ortholog based on best bidirectional hit (BBH) or best reciprocal hits (BRH) which can be pairwise or among all the observed species. Solely bidirectional hit is unable to detect paralogs, however using additional steps and evolutionary distances, paralogs also can be detected. There are several tools that use BBH, such as InParanoid and RoundUp, among others. These methods can produce biased results if some of the observed species have long evolutionary distances. This is an excellent way to identify orthologs, but gene duplications can cause ortholog relationships that are not one-to-one. This is because if gene A is an ortholog of gene B, and gene B is an ortholog of gene C, gene A and C may or may not be orthologs [107,121]. Multi-species graph-based methods infer orthologs, by clustering all the proteins among all the observed species.

There are different methods of making a cluster of orthologous genes after finding significant sequence similarity. One method to start and build a cluster is to select three-way BBHs in three separate species and then merge triangles that share a common side. Tools that use a cluster of orthologs are COG, eggNOG, OrthoDB, etc. There is another

way to build a cluster using a Markov clustering procedure which is known as the probabilistic method (OrthoMCL). To date, OrthoMCL is one of the best tools to infer orthologs. It was very much handy in identifying pathogenic genes in different lineages. Removing highly evolutionary distant hits and then using maximum weight cliques also can be a potential way to cluster orthologs (OMA).

Orthologs directly infer evolution and ancestry. So, any method to infer orthologs without using phylogenetic reconstruction cannot represent correct orthologous genes. Tree-based ortholog inference methods use homology searches and resolve orthology from the tree topology. A popular tool that uses tree topology to infer orthologs is LOFT. To resolve orthologs, reconciliation of gene trees against species tree is one of the most accurate methods. Tools that use tree-reconciliation are Ensembl-Compara, TreeFam, PhylomeDB, and OrthoFinder [111,122,123,124,125,126,127,128,129,130,131,132].

In Figure 2.1.1, the basic principle for all the current methods for ortholog inference are summarized. This work used OrthoFinder because it is a tree-based method which reconciles the gene tree with the species tree and infers orthologs from the topology of the reconciled tree. It is very useful to identify orthologs in different lineages [133,134,135]. This method is also extremely fast and produces summary statistics which are particularly useful for downstream analysis. OrthoFinder is also easy to use and best suited for massive data analysis. This study tests the accuracy of inferred orthologous groups and then, conducts species-specific analysis, pathway abundance, and correlations.

2.3: Methods

2.3.1: Genome data

Protein and nucleotide fasta files for 41 Apicomplexan species were downloaded from the 33rd release of EuPathDB (eupathdb.org, [22]). Data for *Chromera velia* and *Vitrella brassicaformis* (considered close relatives of the Apicomplexa) were downloaded from the same release.

2.3.2: Orthogroup identification

OrthoFinder v1.1.8 was used to find orthogroups. The following command was used, which utilizes all versus all Blast: orthofinder -M msa -A muscle -T raxml -f fasta_directory/. Here, -M is tree inference method for gene trees, -A the multiple sequence alignment methods, -T the tree program from and -f for fasta directory containing the genome files, which produces a list of orthogroups, gene trees, alignments, a rooted species tree and summary statistics.

2.3.3: Testing the inferred orthologs

Caveats and pitfalls of different methods of orthology prediction discussed above show that before starting pathway analysis, correlations, phylogeny, and other specific analysis, it needs to make sure that, the orthology inference is correct. This can be accomplished in two ways. First, the statistical outputs from the orthofinder analysis can be checked. Second, clustering can be used to validate the orthologous genes using different statistical methods and compared them. Orthology inference can be correct if the statistics and clustering reflect biological similarity and dissimilarity.

2.3.3.1: Verifying statistics

There are two statistical outputs namely, overall (Statistics_Overall.csv) and per-specific (Statistics_PerSpecies.csv). These CSV files are analyzed and visualized using Python and R. The Statistics_Overall.csv file produced by OrthoFinder contains information about

the number of orthogroups found in all species and the number and percentage of orthogroups and genes per species. Statistics_PerSpecies.csv contains the total number of genes, number and percentage of genes in orthogroups, the number and percentage of unassigned genes, the number and percentage of orthogroups across species and the number and percentage of species-specific orthogroups in all organisms (Data S2.1-S2.6). Statistics (overall and per-species) were imported to Python as pandas data frame and visualized using Matplotlib [136],[137],[138].

2.3.3.2: Cluster analysis

Along with other results and statistics, orthofinder also give us Orthogroups _SpeicesOverlaps.csv that contains the number of orthogroups shared between each species pair as a square matrix. This file was loaded in python as pandas data frame for hierarchical cluster analysis (HCA) to see how these organisms are related based on the number of pairwise shared orthogroups. An agglomerative approach was used to create the clusters, according to average distance using the Euclidean metric [136]. In this instance, agglomerative refers that each observation is different at the start, and then a cluster is built based on similarity. This can be done quickly using the seaborn library of python (example command: seaborn.clustermap(data frame)) which will give us the color coded clusters along with abundance. The same analysis is also carried out using different data matrix namely, phylogenetic trees and pathway abundance. Phylogenetic trees which were reconstructed using different methods were also incorporated in cluster analysis to compare the outputs from various methods. R (dendextend) and Python (Pandas, Matplotlib, seaborn) were used for the analysis [139],[140],[141].

2.3.4: Identification of species-specific orthogroups

Orthogroups that have genes from only one species but no genes from all other observed organisms are species-specific orthogroups, and the assigned genes are species-specific for that individual species. This is also true for a group of species, termed as group-

specific orthogroups and genes. The Orthogroups.csv file contains the names and corresponding gene symbols for each orthogroup among all species. The orthogroup table (Orthgroups.csv file) was imported into Python as a Pandas DataFrame.

	A	B	C	D	E
1	A1,A12	B1		D1	
2	A2	B2,B22	C2	D2	E2,E22,E23
3			C3,C33,C34		
4		B4			E4
5			C5	D5, D55	E5

Figure 2.2.1: A sample orthogroup table using imaginary data. There are five orthogroups 1, 2, 3, 4 and 5 including 21 genes which belong to 5 species A, B, C, D and E. From this Table, one can visually identify species-specific and group-specific orthogroups and genes. For example, species C has species-specific orthogroup which is 3. Orthogroups 4 is specific for species B and E and so on.

From Figure 2.2.1, one can identify species-specific and group-specific orthogroups and genes visually. However, finding those groups and genes in a table which holds 2,48,586 genes in 15380 orthogroups among 43 species is not visually possible. This can be done by sub-setting the data frame using certain conditions. For example, we can subset the orthogroups which have genes from all species but no genes from certain species. To do this, we have to specify how many blanks we can keep and in which position. In pandas data frame of python we can select species-specific orthogroups with the following command: data.iloc [:,np.r_[0: 43]].isna().sum(axis=1)==42. Here, we instruct python to subset the rows or orthogroups which has NaN values in any 42 columns or species.

In the same way, we can find overall and certain groups or species with the desired number of NaN. This strategy is also useful to find gene loss. Detailed commands and instructions are summarized in Appendix C2. Functional annotation was conducted using

OmicsBox (v1.2.4) for the fasta files of the orthgroups that were found specific for *P. falciparum, P. malariae, P. ovale, P. vivax, P. cynomolgi, P. knowlesi, P. gaboni* and *P. reichenowi* [142]. UpSetR (v1.4.0) was used to identify the intersections of sharing orthogorups between these 8 species [143].

2.3.5: Pathway analysis

A principal output from orthofinder is Orthogroups.csv which contains the name of orthogroups and species including gene symbols (like Figure 2.2.1 with real species and genes). The gene symbols look like, BBOV_III007500-t26_1-p1, BBOV_III007510-t26_1-p1 etc. We have to convert this information into biological or molecular functions to identify genes and orthogroups involved in different pathways. In the KEGG (Kyoto Encyclopedia of Genes and Genomes) database, there are assigned K numbers for each functional ortholog. So, if we can identify which gene symbol belongs to which K number, we can find the orthogroup that is related to a function. This process can be termed the functional annotation. In the case of well-annotated genomes, it can be found in the database as mentioned earlier. As most of our genomes are poorly annotated, the scenario is different. First, we have to do a blast (basic local alignment search tool) search to find a reference sequence which may be related with a K number to map. We can term this method as ID mapping.

2.3.5.1: ID Mapping

ID mapping is a process by which an identifier, such as a gene symbol, is converted into another identifier, such as a pathway name. Mapping genes to identifiers such as GO (Gene Ontology) or K number helps the acquisition of information regarding the biological and evolutionary process of each orthogroup. A Blastp search was conducted for the protein fasta file against the non-redundant protein NCBI database, with a maximum e-value cut-off of 1e−10. From the Blast output, the Genbank ID (GI) number was retrieved and converted to UniProt ID, and subsequently K number, using the UniProt

idmapping.dat mapping file obtained from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/.

Gene symbol	GI number	Uniprot ID	K number	Functions	Pathway
BBOV_III007500-t26_1-p1	568815597	P00750	00010	Glycolysis	Carbohydrate metabolism
BBOV_III007510-t26_1-p1	224589800	P00770	00620	Pyruvate metabolism	Carbohydrate metabolism

A

Gene symbol	K number	Functions	Pathway	Orthogroups
BBOV_III007500-t26_1-p1	00010	Glycolysis	Carbohydrate metabolism	OG00001
BBOV_III007510-t26_1-p1	00620	Pyruvate metabolism	Carbohydrate metabolism	OG00001

B

Figure 2.2.2: ID mapping workflow. The gene symbol is converted to GI number using blast search then mapped with Uniprot, K number, Functions, Pathways, and Orthogroups (A). Gene symbols were directly converted to K number using BlastKOALA and then mapped with Functions, Pathways, and Orthogroups. After blast search, the IDs can be converted using a python dictionary. For this table, six dictionaries were created with key and values are Gene symbol: GI number, GI number: UniProt ID, UniProt ID: K number, K number: Functions, Functions: Pathway, Pathway: Orthogroups.

The output of the analysis for each species is a two-column file with a K number assignment for the annotated genes. We can justify this method using BlastKOALA (KOALA (KEGG Orthology And Links Annotation). BlastKOALA is KEGG's internal annotation tool for K number assignment of KEGG genes using SSEARCH computation. BlastKOALA assigns the K number using Blast search of the protein sequences against a non-redundant dataset of pangenome sequences, identifying the K number content of each sequence. However, it is slow and time-consuming to process large datasets in BlaskKOALA, since it is a web and email-based service. Additionally, the maximum number of reads supported by BlastKOALA is 7500, which is limited.

Given that orthogroups are comprised of genes with orthologous function, we can infer the function of the entire orthogroup, from the K number assignment of a single gene. A python dictionary was created from the csv file to map the gene symbol to the corresponding GI number. This was repeated to get some other dictionaries of UniProt ID

and K (KEGG) numbers, K numbers and orthogroup names, K numbers and respective pathway names (Figure 2.2.2). After mapping, the data frame was grouped by K numbers and again mapped with orthogroups names of an abundance data set (Orthogroups.GeneCount.csv another output from OrthoFinder). Abundance data set was summed up and mapped with pathway names. K numbers were mapped to orthogroups for all the 43 proteomes individually, using a custom python script (Appendix C2). After K number to orthogroup assignment for each proteome, a list of K numbers for each orthogroup was generated, and the most common K number selected. This approach of mapping to each species significantly improves mapping, the efficiency of which is a common problem in comparative genomic studies [96],[144],[145],[146].

2.3.5.2: Correlation Analysis

The proteome size of each species was calculated using Biopython and then added to the final output from ID mapping. A correlation test between proteome size and pathway abundance was conducted and visualized using the Pandas, Scipy and Matplotlib libraries of Python. Given the non-independence of data points due to phylogenetic effects, a correction was applied using phylogenetically independent contrasts (PIC) using the R package ape, and the topology of the Apicomplexan phylogenomic species tree [147],[148], [149],[150].

2.4: Results and Discussion

2.4.1: Orthogroup testing and validation

A total of **2,90,278** genes were present in 43 species. Among these, **2,48,586** (85.6%) were assigned to 15380 different orthogroups with a mean and median orthogroup size of 16 and 6 genes, respectively. There were 335 species-specific orthogroups which includes 3274 genes. G50=39 (At least 50% of orthogroups have a minimum of 39 genes) and O50=1909 (a minimum of 1909 orthogroups have at least 50% of the genes). 522 orthogroups were found which have genes from all species, and this stands for the core genome. Detailed statistics of the orthogroup analysis is presented in Data S2.1-6.

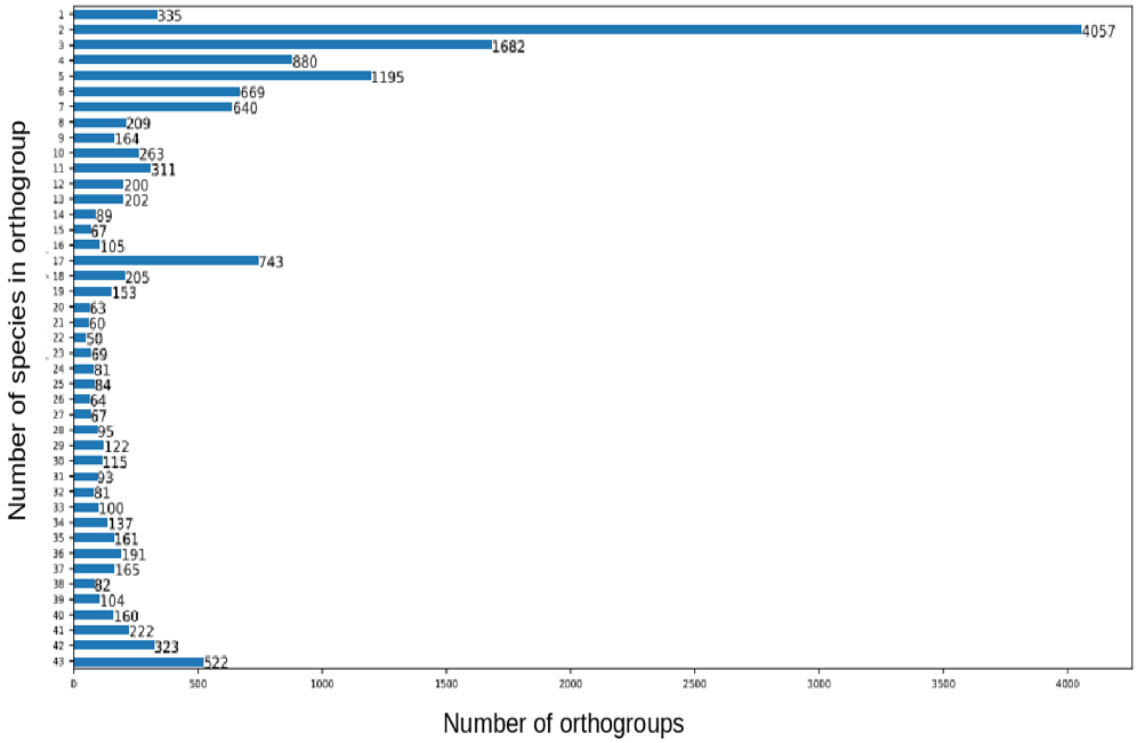

Figure 2.4.1.1: Distribution of orthogroups among 43 species. Pairwise-shared orthogroups is the largest group including 4057 orthogroups. All the species shares at least 522 orthogroups

26.4 % the orthogroups were shared between two species (Fig. 2.4.1.1). Most of the species (31 of 43) have species-specific orthogroups, ranging from 1 to 95 unique orthogroups (Fig. 2.4.1.2). 12 species do not possess any unique orthogroups; these 12 species have proteome size with mean ± std=3327452 ± 950669, whereas the remaining 31 species have a proteome size of 4517426 ± 3238842. The smaller proteome size of those species without unique orthogroups reflects a greater extent of genomic

simplification. As discussed in the introduction, adaptions to a parasitic lifestyle are expected, so the lack of unique orthogroups may merely be a statistical artifact given that there are fewer genes overall in a smaller genome.

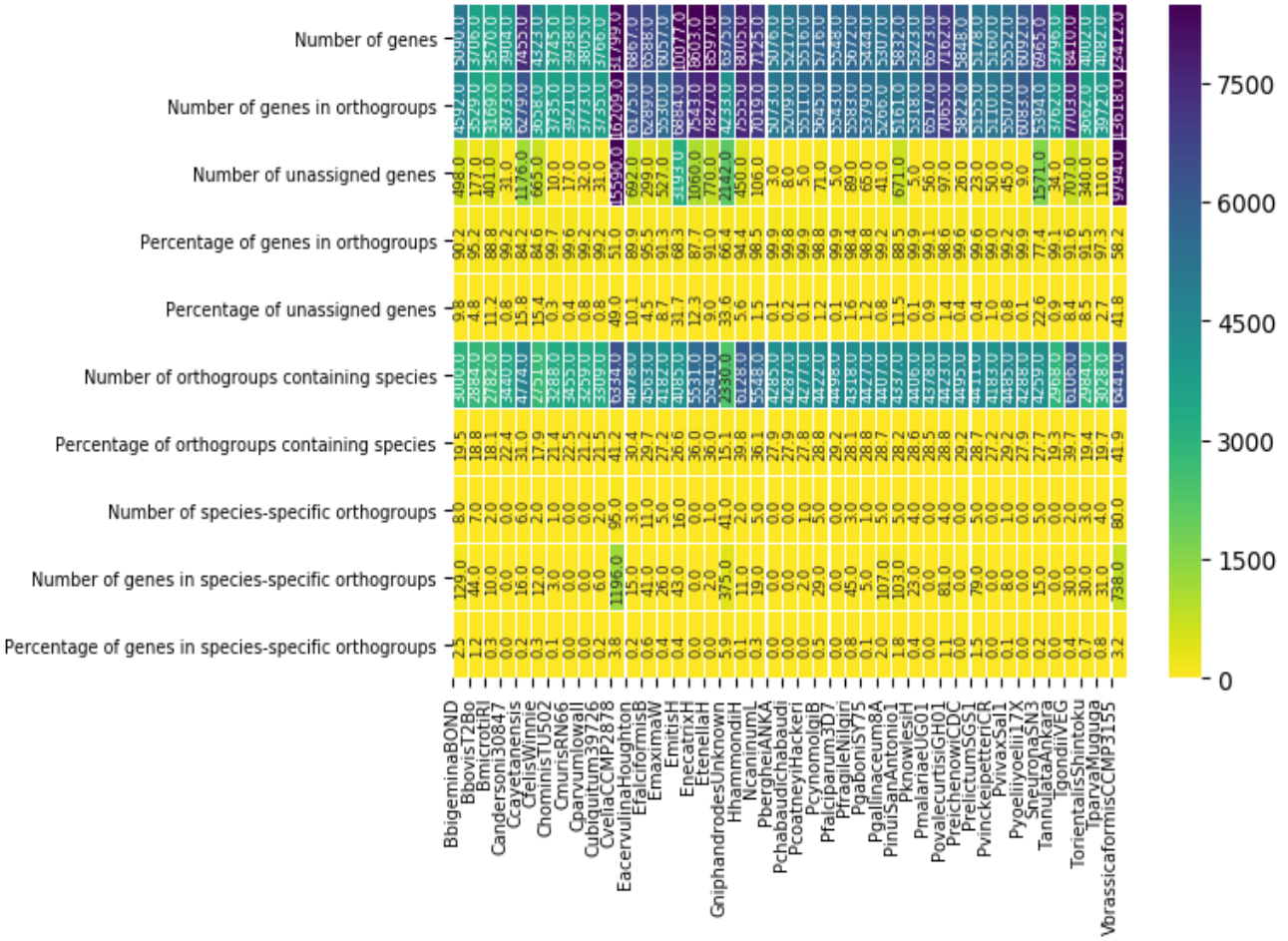

Figure 2.4.1.2: Per-species statistics of all orthogroups. The colors reflect observation values

Most of the genes (85.6% of total genes) were included in the orthogroups which indicates that there are sufficient genes in the next steps of the analysis (Data S2.5-6). In any given related group of genomes, the number of shared genes between any two genomes would be higher than the number of shared genes among 3 or 4 or more genomes. A distantly related species should have fewer genes in orthogroups compared to other closely related species. *C. velia* and *V. brassicaformis* showed compatibility with this argument according to the number of unassigned genes and the number of species-specific orthogroups (Fig. 2.4.1.2). *G. niphandrodes* is an early diverging apicomplexan compared to the others in the observed group [151]. A relatively smaller number of *G.*

niphandrodes's genes in orthogroups is present in comparison with other 40 observed Apicomplexans, in terms of the number of unassigned genes and percentage of genes in species-specific orthogroups (Fig. 2.4.1.2).

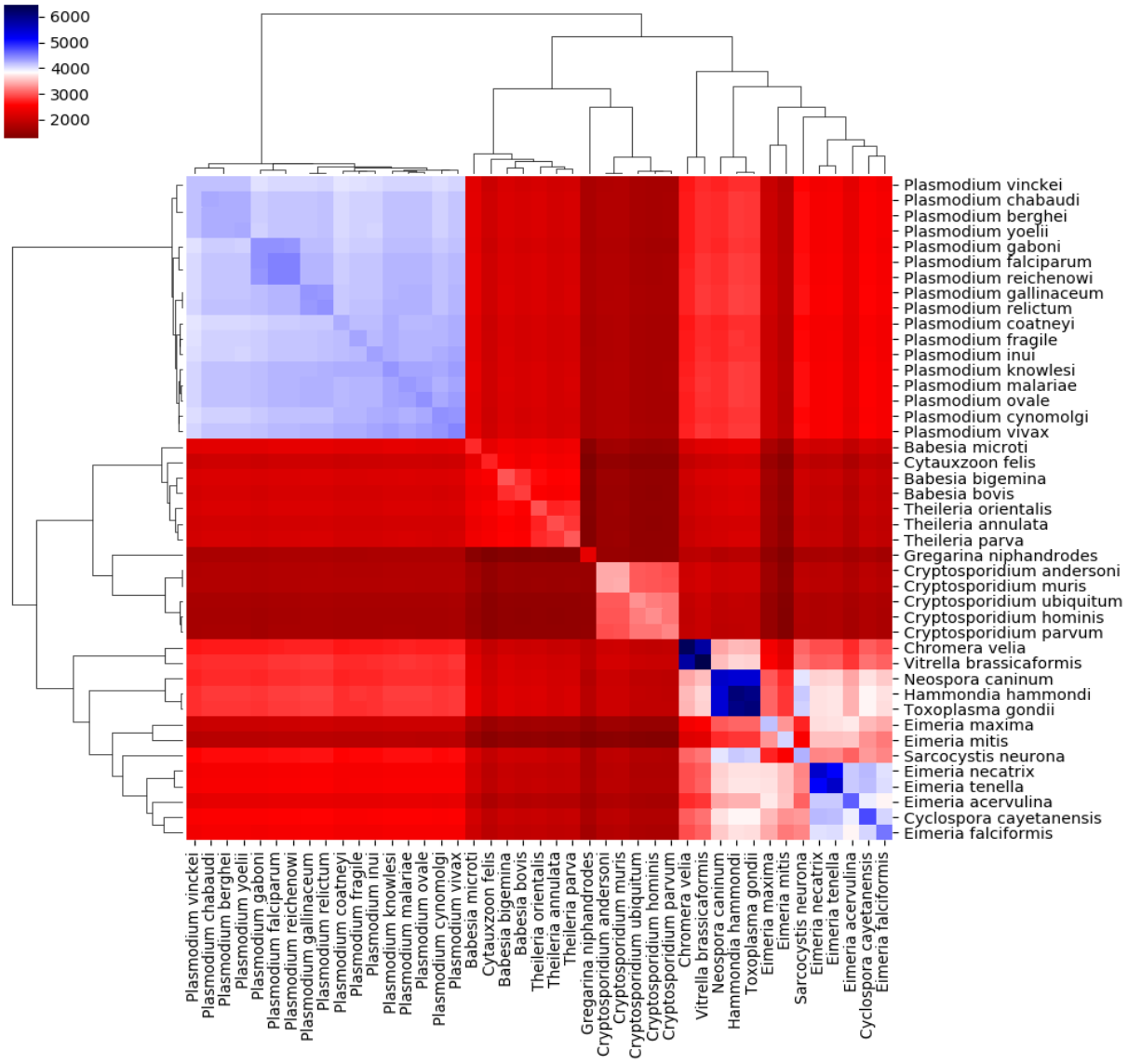

Figure 2.4.1.3: Hierarchical cluster according to pairwise shared orthogroups

Species of each genus *Piroplasma, Eimeria, Toxoplasma, Cryptosporidium*, and *Plasmodium* were clustered together in the hierarchical cluster built based on pairwise shared orthogroups, which indicates the robustness of the orthology prediction (Fig. 2.4.1.3). *C. velia* and *V. brassicaformis* shared most orthogroups among themselves and were clustered with *Toxoplasma*. These two have the most considerable number of genes

which supports why they share too many orthogroups. On average, *Toxoplasma* and *Eimeria* have more genes than other 3 groups (Tab. 1.3 and Fig. 2.4.1.3). This may explain why these two chromerids were clustered with *Toxoplasma*. So, from Tab. 1.3, Fig. 2.4.1.1 and 2.4.1.2, we can conclude that this orthology inference reflects biological relationships among the observed species. Clustering including all orthogroups, principal component analysis and descriptive boxplot with all genes and orthogroups also support the robustness of the ortholog prediction method (Fig. 2.4.1.2-6).

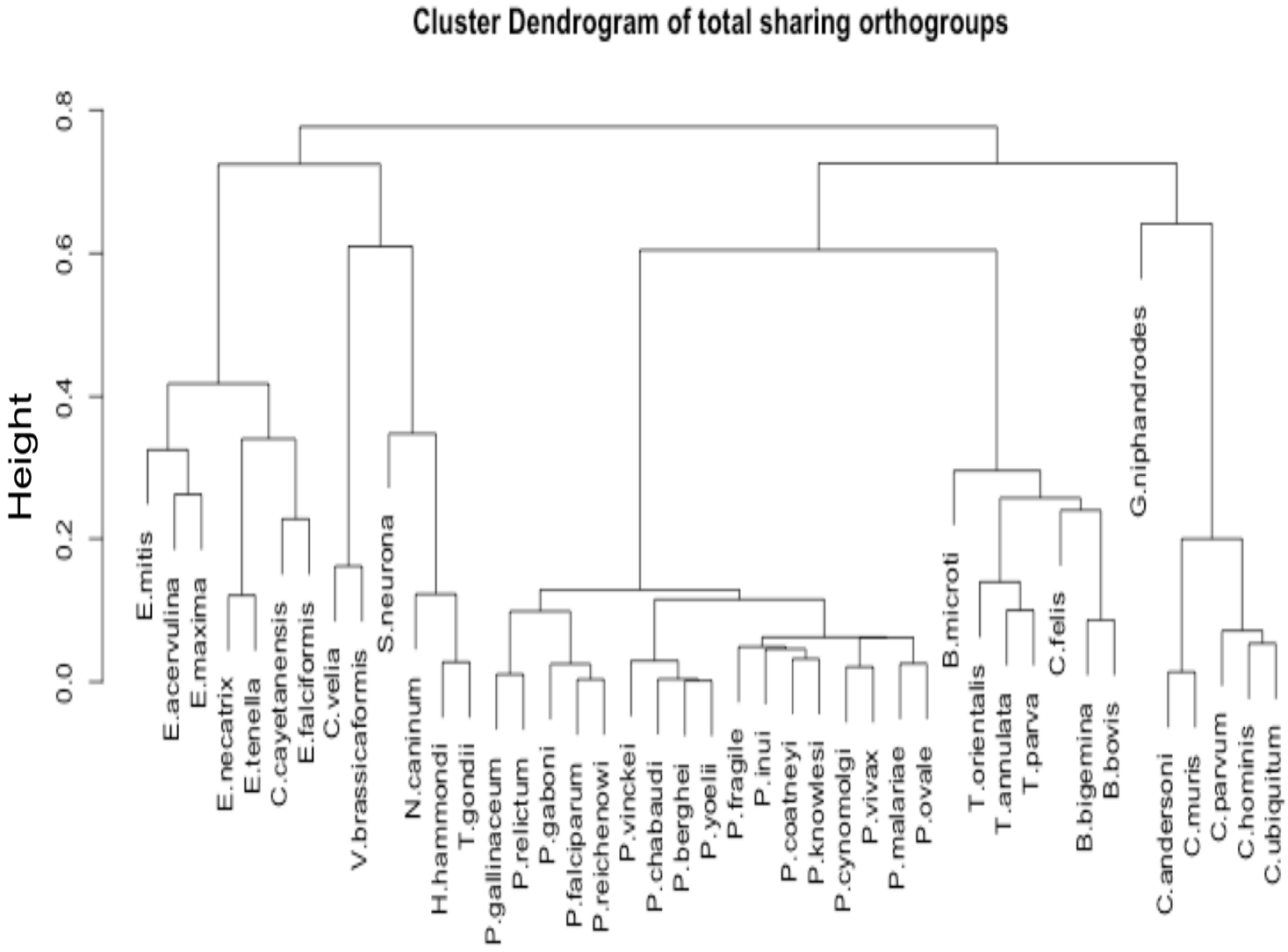

Figure 2.4.1.4: Cluster Dendrogram including all orthogroups. Major lineages clustered together here though not 100 % congruent with phylogeny. The Chromerids clustered together with Sarcocystidae indicates high number of genes in these two groups compared to other observed species

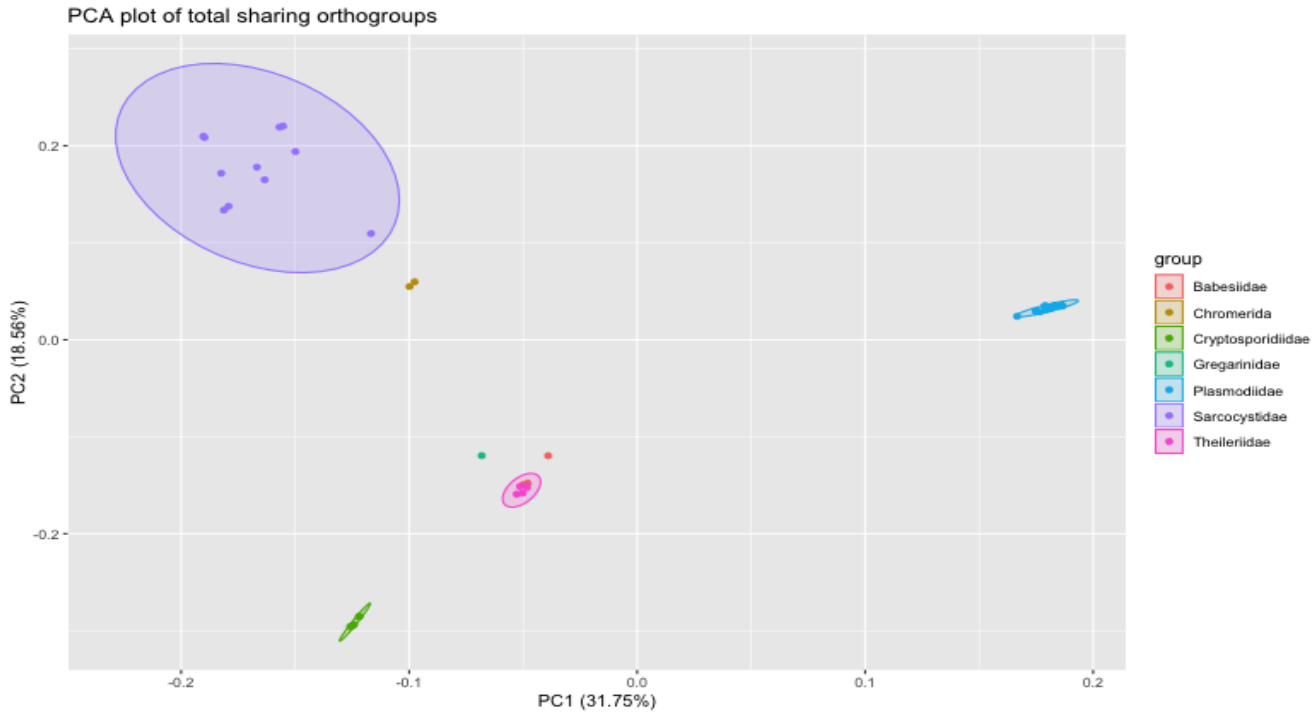

Figure 2.4.1.5: PCA plot with all sharing orthogroups. This is another and maybe clearer approach to find similar groups in a data set. Here also major lineages clustered together. This time *Gregarinidae* clustered with *Babesiidae* because of the very low number of orthogroups in these two lineages compared to others (exactly opposite as Fig. 2.4.1.3). These two phenomena can be explored in detail with boxplot (Fig. 2.4.1.6).

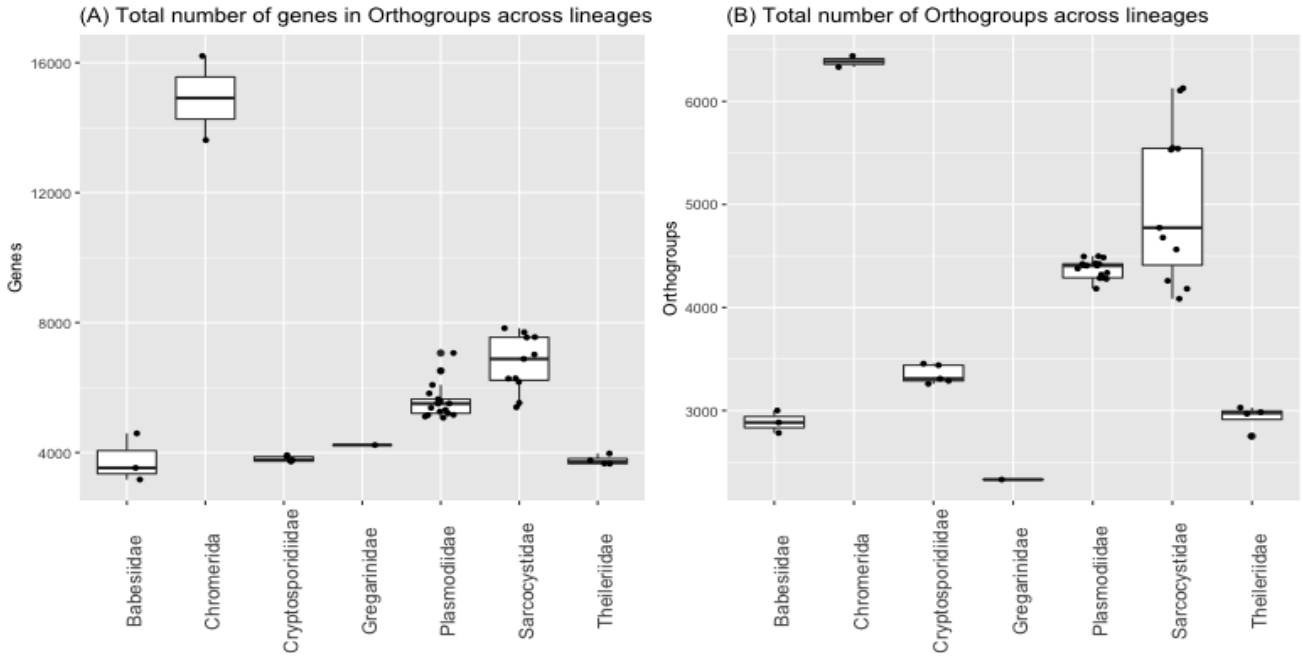

Figure 2.4.1.6: Boxplot of total number of genes (A) and orthogroups (B) across lineages. Here, B explain why Chromerids clustered with Sarcocystidae in Fig. 2.4.1.3 (high abundance) and clustering of Gregarinidae with Babesiidae (low abundance) in Fig. 2.4.1.5.

2.4.2: Unique orthogroups in malaria-causing *Plasmodium* species

The above-mentioned 8 Plasmodium species that cause malaria in humans and chimps were examined for unique orthogroups. These 8 species have 285 specific or pathogenic-specific orthogroups including 2327 genes. In this group, *P. knowlesi* has the lowest number of orthogroups (n=5) and genes (n=24). *P. ovale* has the largest number of genes (n=537) but *P. falciparum* has the largest number of orthogroups (n=189) (Fig. 2.4.2.1). This group of organisms shares genes among themselves in 19 diverse ways/sets. In these 19 sets, no set includes all the 8 species. 5 species (*P. malariae, P. reichenowi, P. falciparum, P. gaboni* and *P. ovale*) share 1 orthogroup with 1 gene from each. There are 3 sets with 4 species each; namely *P. cynomolgi, P. ovale, P. vivax, P. gaboni* (1 orthogroup with 6 genes), *P. cynomolgi, P. ovale, P. vivax, P. malariae* (2 orthogroups with 179 genes) and *P. malariae, P. reichenowi, P. falciparum, P. gaboni* (2 orthogroups with 29 genes). There are only 2 sets with 3 species each: namely, *P. falciparum, P. gaboni, P. reichenowi* (150 orthogroup with 944 genes, the largest group) and *P. cynomolgi, P. ovale, P. vivax* (18 orthogroups with 463 genes). There are total 8 sets of pairwise sharing orthogroups which constitute 96 orthogroups with 555 genes. 5 species have species-specific orthogroups totaling 15 orthogroups with 146 genes (Data S2.7).

These numbers (shared orthogroups and genes) are notably smaller compared to the sharing orthogroups and genes in all the species (522 orthogroups in all 43 species, Fig. 2.4.1.1) and even in the Plasmodia (673 orthogroups in only all of the Plasmodia, Fig. 2.4.2.2) which support [29], that, these organisms share fewer parasite specific genes compared to other gene families (Fig. 2.4.2.1).

In most cases, sequence similarities were not found for functional annotations. 1399 out of 2327 genes were unannotated. A GO annotation shows that membrane associated

proteins predominate in the orthogroups uniquely associated with malaria causing *Plasmodium* species (98.17% of annotated genes are in membrane (Fig. 2.4.2.1)).

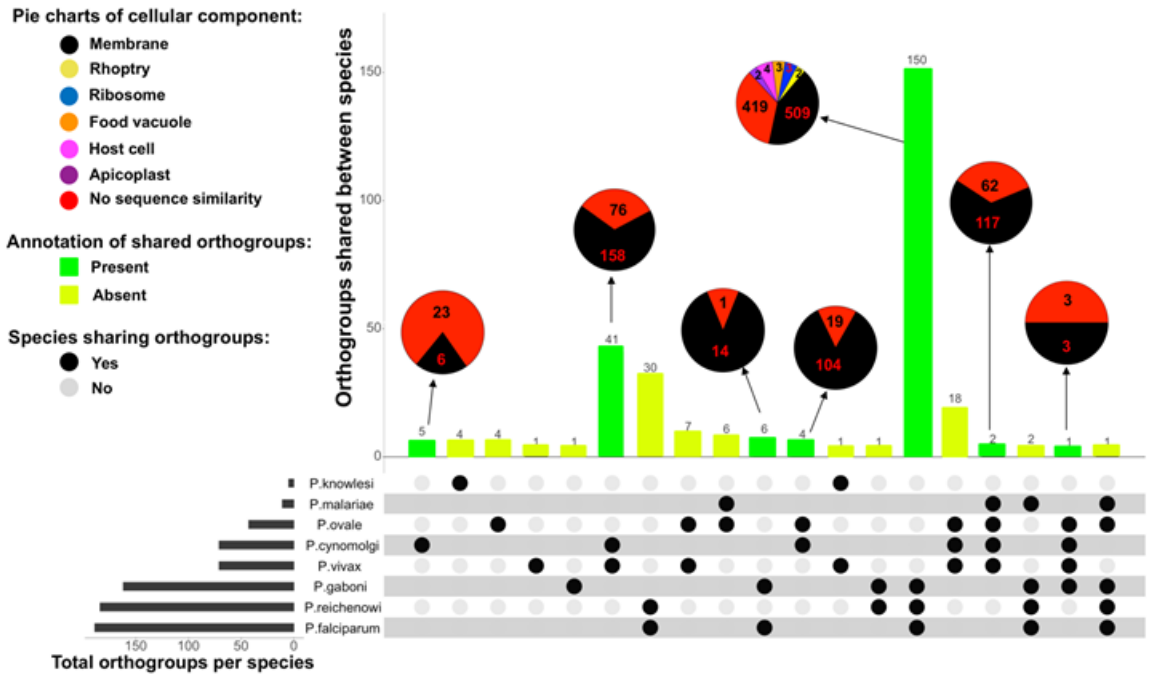

Figure 2.4.2.1: Unique orthogroups shared between 8 primate infecting *Plasmodium* species. The vertical bars represent orthogroups shared between species and the horizontal bars represent the total number of orthogroups per species. The pie charts show the GO terms (cellular components) of shared orthogroups (source annotation in Data S2.7). 285 orthogroups are shared in 19 different sets. Among these 19 sets, 12 have no known sequence similarity which comprises 76 orthogroups with 796 genes. In the remaining 7 sets (209 orthogroups with 1531 genes), 603 genes have no sequence similarity (species groups, number of genes and orthogroups are in the Data S2.7). Among 928 annotated genes, 911 are in Membrane, 2 in Rhoptry, 3 in Ribosome, 3 in Food vacuole, 4 in Host cell, 2 in Apicoplast, and 3 genes are annotated with biological process or molecular function whose cellular component were not found (details in Data S2.7).

P. falciparum shares some orthologous genes exclusively only with *P. reichenowi* [152] but no sequence similarity were found for those genes (This is maybe an example of database and tools version discrepancy; in our previous search (Tab. 2.4.1) and other author [152] also reported pathogenic homolog for this pair). 98 membrane associated genes were annotated as pathogenic which are shared only among *P. falciparum* (n=85), *P. gaboni* (n=6) and *P. reichenowi* (n=7). (Tab. S2.7). Of particular interest to malaria studies, the genes comprising the 150 unique orthogroups we have identified in *P. falciparum, P. gaboni* and *P. reichenowi* constitute a list of novel candidate pathogenic

factors, amenable to experimental analysis, such as genetic manipulation in *P. falciparum,* to silence or modify the genes in order to assess their function and potential role in pathogenicity; a better understanding of genes associated with pathogenicity could translate into potential interventions (such as vaccines, compounds blocking pathogenic genes, drugs) to help control or prevent disease.

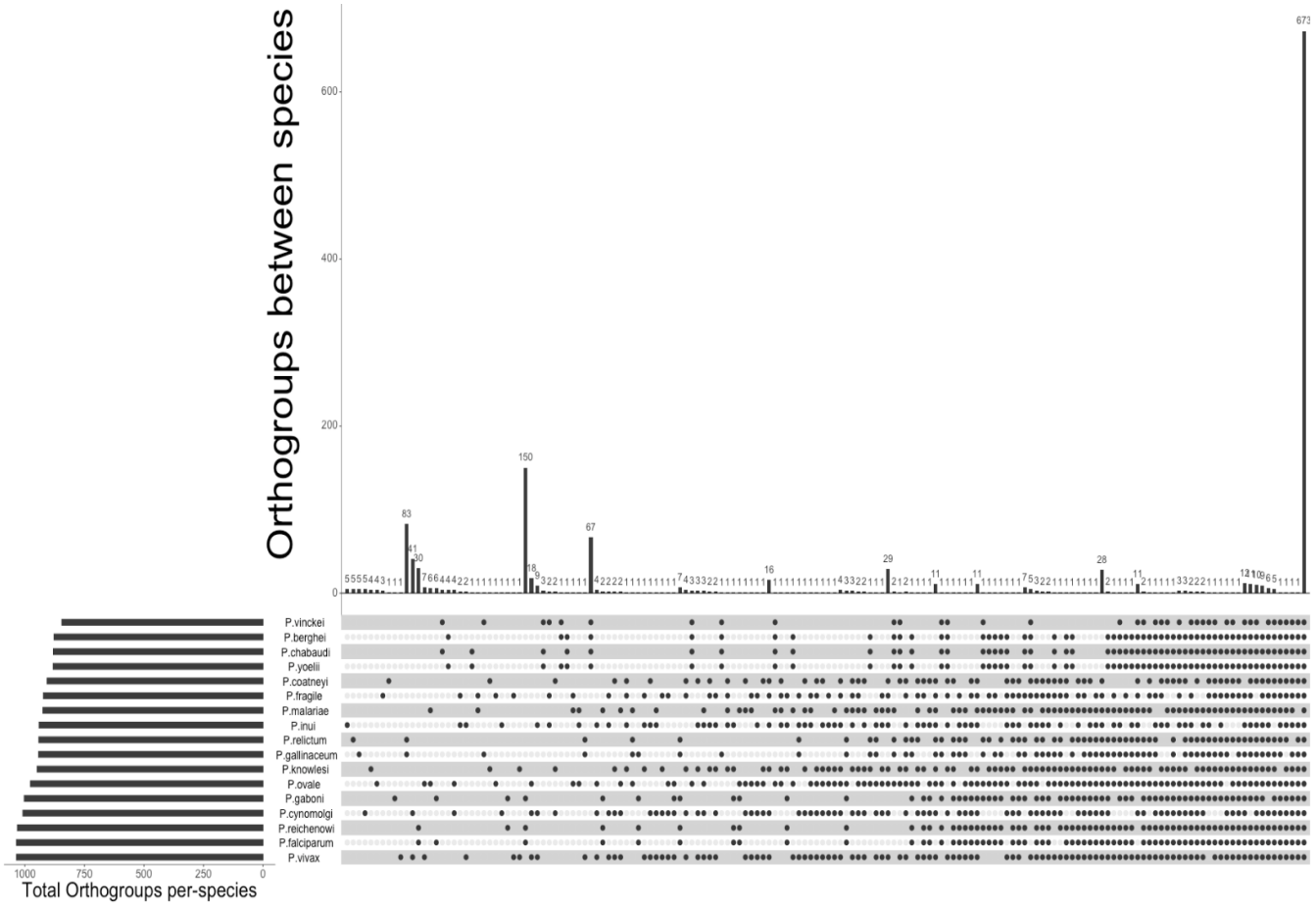

Figure 2.4.2.2: Unique orthogroups in all 17 *Plasmodium* species.

Previously, a comparison of 6 *Plasmodium* species that infect primates and rodents showed 16 conserved genes in 3 primate parasites, and 1118 common genes in all rodent and primate parasites (*P. falciparum, P. vivax, P. knowlesi, P. chabaudi, P. berghei* and *P. yoelii*) [24]. Our data show that 2327 genes conserved in 8 primate parasites that were not found in any rodent parasite (Fig. 2.4.2.1).

The thiamine biosynthesis pathway (KO00730) has been proposed as a potential

antimalarial target given that rodent parasites are dependent on thiamine uptake from hosts [153], [154], [24]. Our data show that *Plasmodium* species have more genes in this pathway than any other Apicomplexans, which may reflect the importance of this pathway in pathogenesis (Fig. 2.4.2.3).

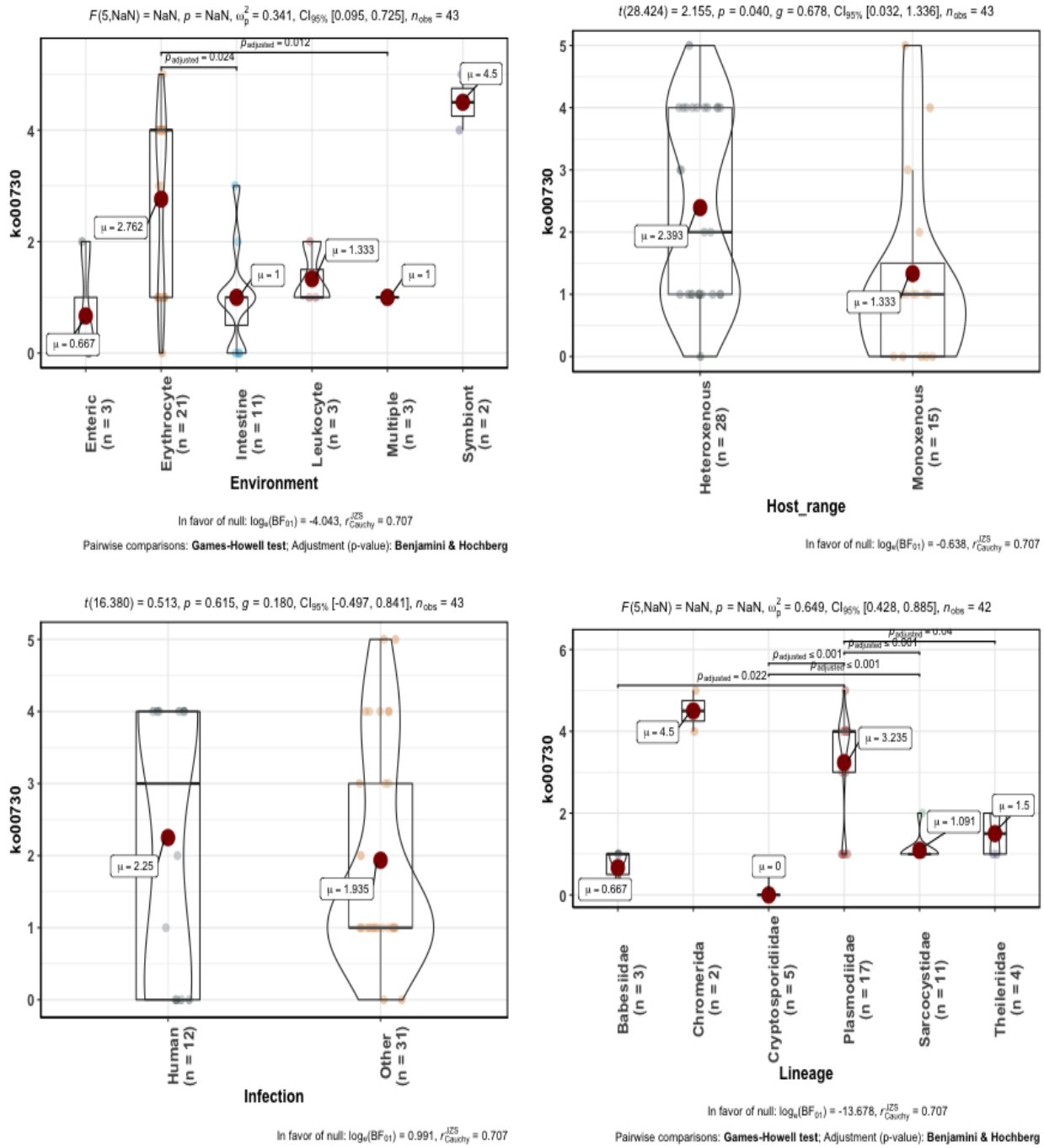

Figure 2.4.2.3: Abundance of Thiamine biosynthesis pathway in different group of species

2.4.3: Pathway Analysis

Among 43 proteomes (33rd release of EuPathDB, that were used in this analysis) 44.63% genes were annotated as hypothetical proteins, 9.36% genes as unknown function and 25.09% genes as putative (Table 1.1).

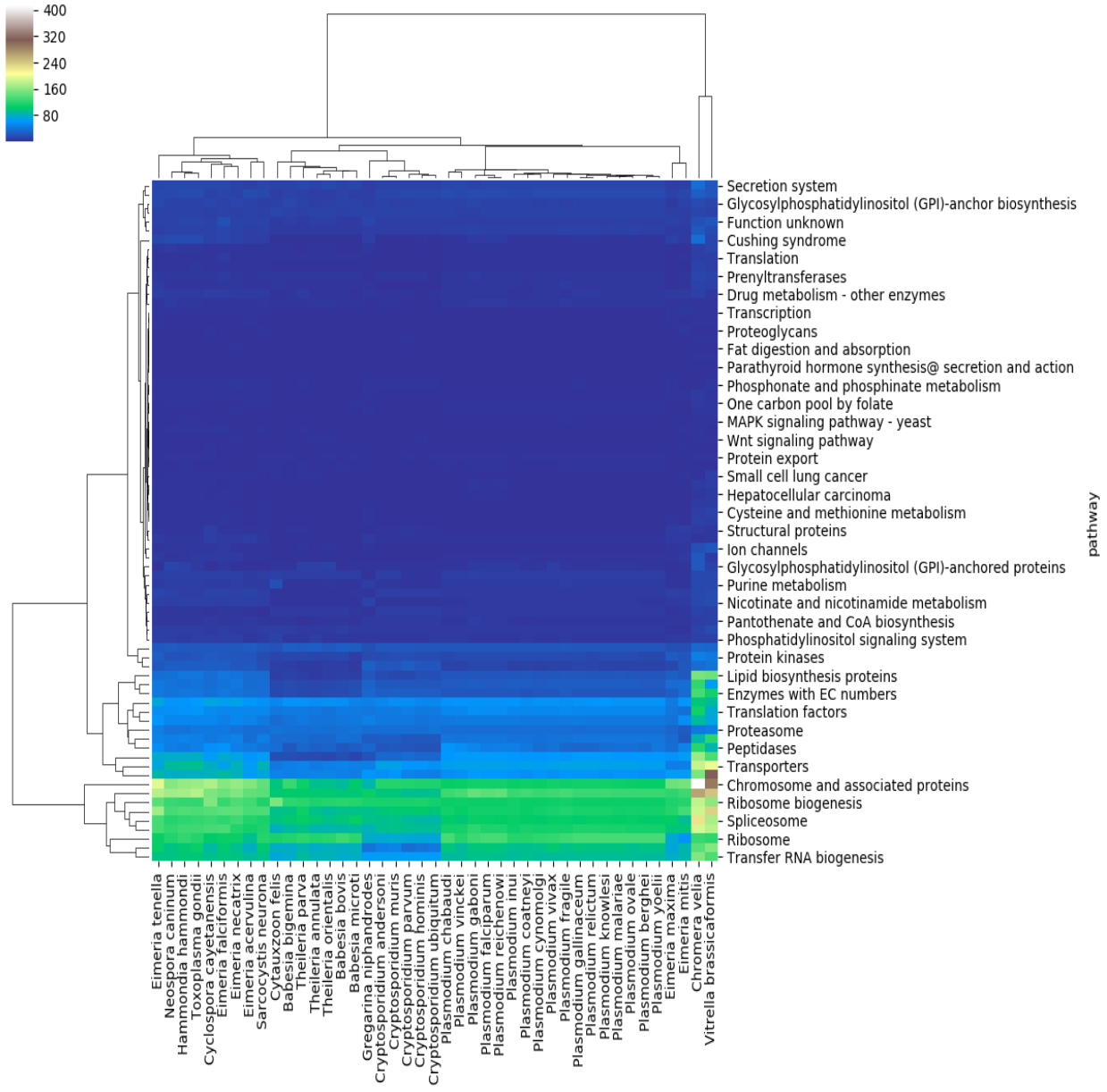

Figure 2.4.3.1: Hierarchical cluster analysis of pathways that are present in all Apicomplexan genomes. The colors represent the abundance of each pathway. The pathways that have fewer number of genes across genomes are less variable compared to pathways with abundant genes.

Among 15380 orthogroups which includes 248586 genes, 2865 orthogroups (95429 genes) were mapped into 2482 unique K numbers which are distributed in 195 functional orthologs (88063 genes). On average, each species has 2047 genes which were mapped to known functional orthologs (min: *C. ubiquitum*; 1510 and max: *C. velia*; 4901). Hence, 35.43% of assigned genes were included in 18.63% of total orthogroups which were successfully mapped with Kegg functional orthology.

The least abundant functions are Biosynthesis of secondary metabolites – unclassified, Carbohydrate metabolism, Dorso-ventral axis formation, Glycerolipid metabolism, Pentose and glucuronate interconversions, RIG-I-like receptor signaling pathway, Retinol metabolism, Sesquiterpenoid and triterpenoid biosynthesis, Steroid biosynthesis, Styrene degradation and Various types of N-glycan biosynthesis, each function has only 2 genes in 43 species. The number of genes in a few pathways are high and variable while, most of the pathways have very few genes and do not vary significantly species to species (Tab. 2.4.2 and Fig. 2.4.3.1).

Chromosome and associated proteins have the most significant number of genes (5751). Seventy-four functional orthologs were found in all the genomes with a total of 80430 genes (91.33% of annotated genes). Some functions vary a lot whereas many functions do not. The most abundant 10% of functions make up ~80% of genes (Fig. 2.4.3.2).

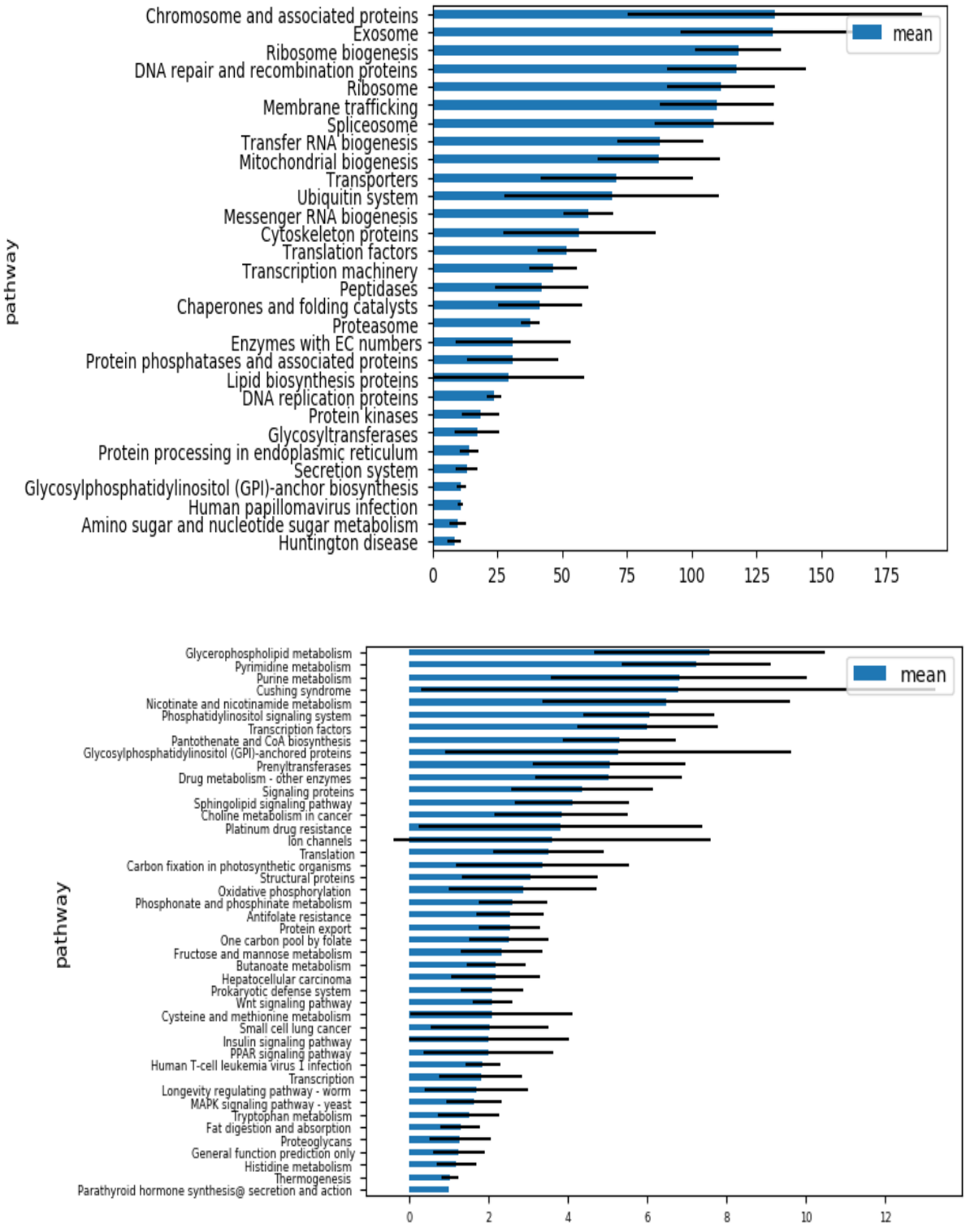

Figure 2.4.3.2: Abundance of mapped pathways with standard deviation. Most abundant 30 pathways are in the top, and the least abundant pathways are in the bottom.

2.4.4: Relationships between pathway loss and proteome size

C. velia and *V. brassicaformis* have an enormous proteome size compared to the Apicomplexan species, which affects the correlation analysis. These two organisms are considered as outliers, because, they are not Apicomplexa and have a larger genome than any other observed species, details in table 1.2, 1.3 and figure 2.4.1.2. The probability density function of the proteome size is presented in Figure 2.4.4.1. Among the most abundant pathways, proteome size was found significantly correlated with DNA repair & recombination proteins, exosome, membrane trafficking, and transfer RNA biogenesis (Pvalue < 0.05) (Fig. 2.4.4.2, Tab. 2.4.3).

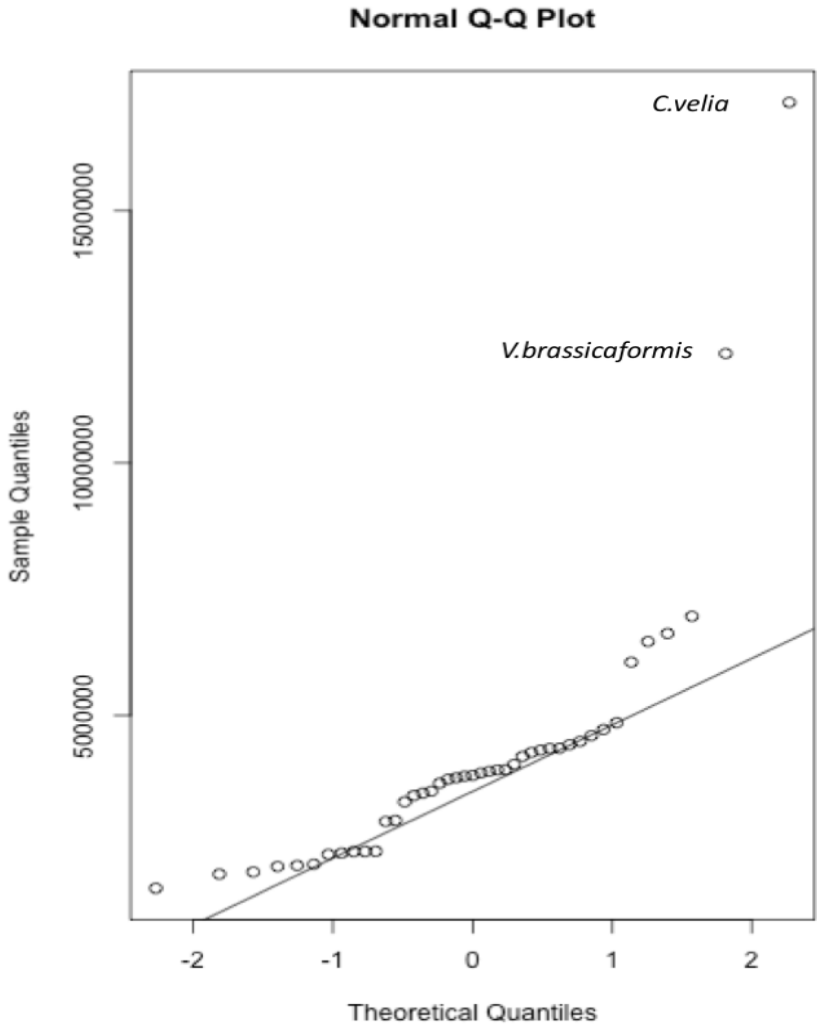

Figure 2.4.4.1: Probability plot of proteome size along with outliers. In the X-axis, a random variable with 0 mean and standard distribution and in the Y-axis, ordered values of proteome size was plotted.

All the functions that are correlated with proteome size are summarized in Tab. 2.4.3 with respective Pearson r value and p-value. Correlations between proteome size and

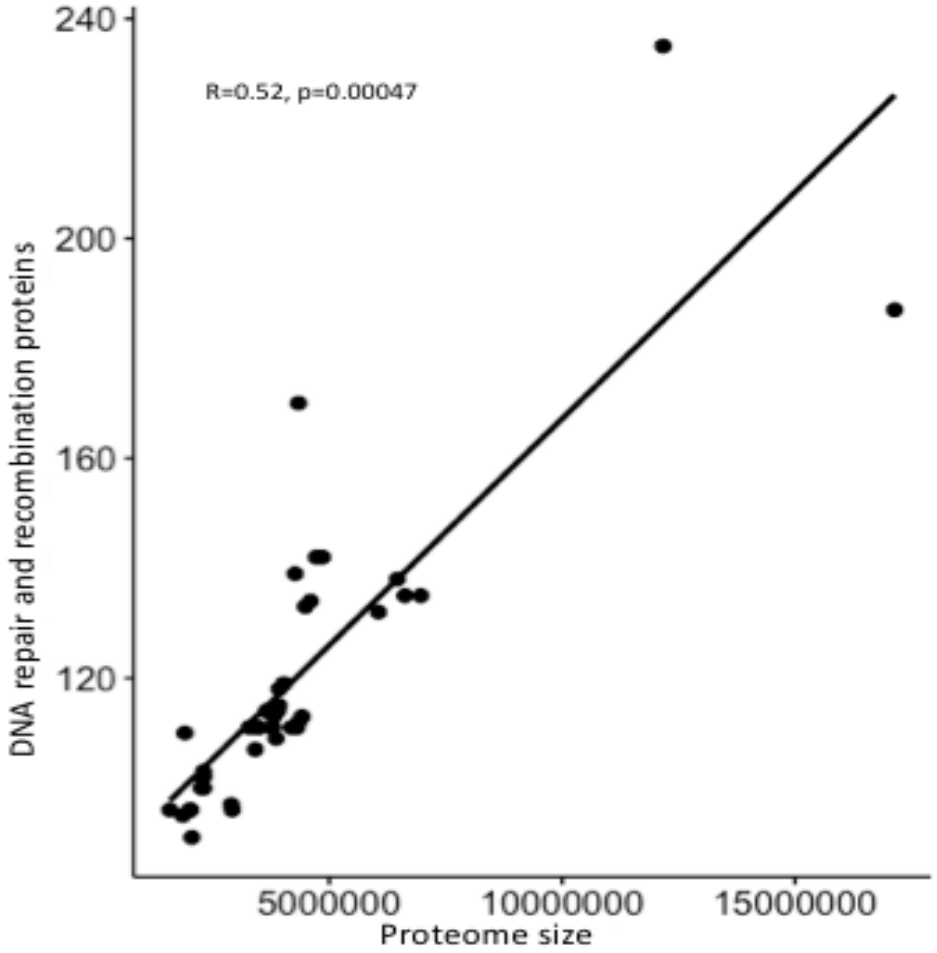

Figure 2.4.4.2: Correlation between proteome size and DNA repair system in all the observed species. This correlation is measured using phylogenetic independent contrast

pathways were corrected using phylogenetic independent contrasts method and summarized in table 2.4.3.

The Apicomplexa show differential gene loss during the process of genome reduction, with 14.95% (29 out of 195 annotated functional orthologs) of total orthogroups significantly correlated with proteome size (Tab. 2.4.3).

The following categories are completely lost from all Apicomplexa but exists in Chromerids: Biosynthesis of various secondary metabolites (KO00999), Carotenoid biosynthesis (KO00906), Cytochrome P450 (KO00199), Dorso-ventral axis formation (KO04320), Glycerolipid metabolism (KO00561), Pattern recognition receptors (KO04054), Pentose and glucuronate interconversions (KO00040), Retinol metabolism (KO00830), RIG-I-like receptor signaling pathway (KO04622), Sesquiterpenoid and triterpenoid biosynthesis (KO00909), Steroid biosynthesis (KO00100), Styrene degradation (KO00643) and Various types of N-glycan biosynthesis (KO00513) (Data. S2.8).

Those KEGG categories for amino acid metabolism that show a moderately strong correlation ($R \geq 0.6$) with proteome size are alanine, aspartate and glutamate metabolism (KO00250), cysteine and methionine metabolism (KO00270) and tyrosine metabolism (KO00350). Those that do not include arginine biosynthesis (KO00220) and tryptophan metabolism (KO00380).

Other metabolic categories that show a moderately strong correlation include lipid biosynthesis (KO00104), glycerolipid metabolism (KO00561), folate biosynthesis (KO00790), fructose and mannose metabolism (KO00051), carbohydrate metabolism (KO04973), steroid biosynthesis (KO00100), N-glycan biosynthesis (KO00510), carotenoid biosynthesis (KO00906), retinol metabolism (KO00830), ascorbate and aldarate metabolism (KO00053). Interestingly, orthogroups encoding transporter proteins (KO02000), which constitute potential drug targets [155], are correlated with proteome size, which seems counter-intuitive as the loss of biosynthetic capacity by the parasite might be expected to necessitate enhanced transport of metabolites provided by the host.

Additional categories that show a moderately strong positively correlation with proteome size include peroxisome proteins (KO04146), photosynthesis (ie. apicoplast) proteins (KO00194), cytoskeleton proteins (KO04812), secretion system proteins (KO02044), and lysosome proteins (KO04142). The loss of genes associated with these functions may reflect morphological simplification associated with adaptation to an intracellular lifestyle. Although adaptation to a parasitic lifestyle might be expected to be accompanied by the

evolution of new genes involved in host interaction, there were no KEGG categories that showed a strong negative correlation with proteome size.

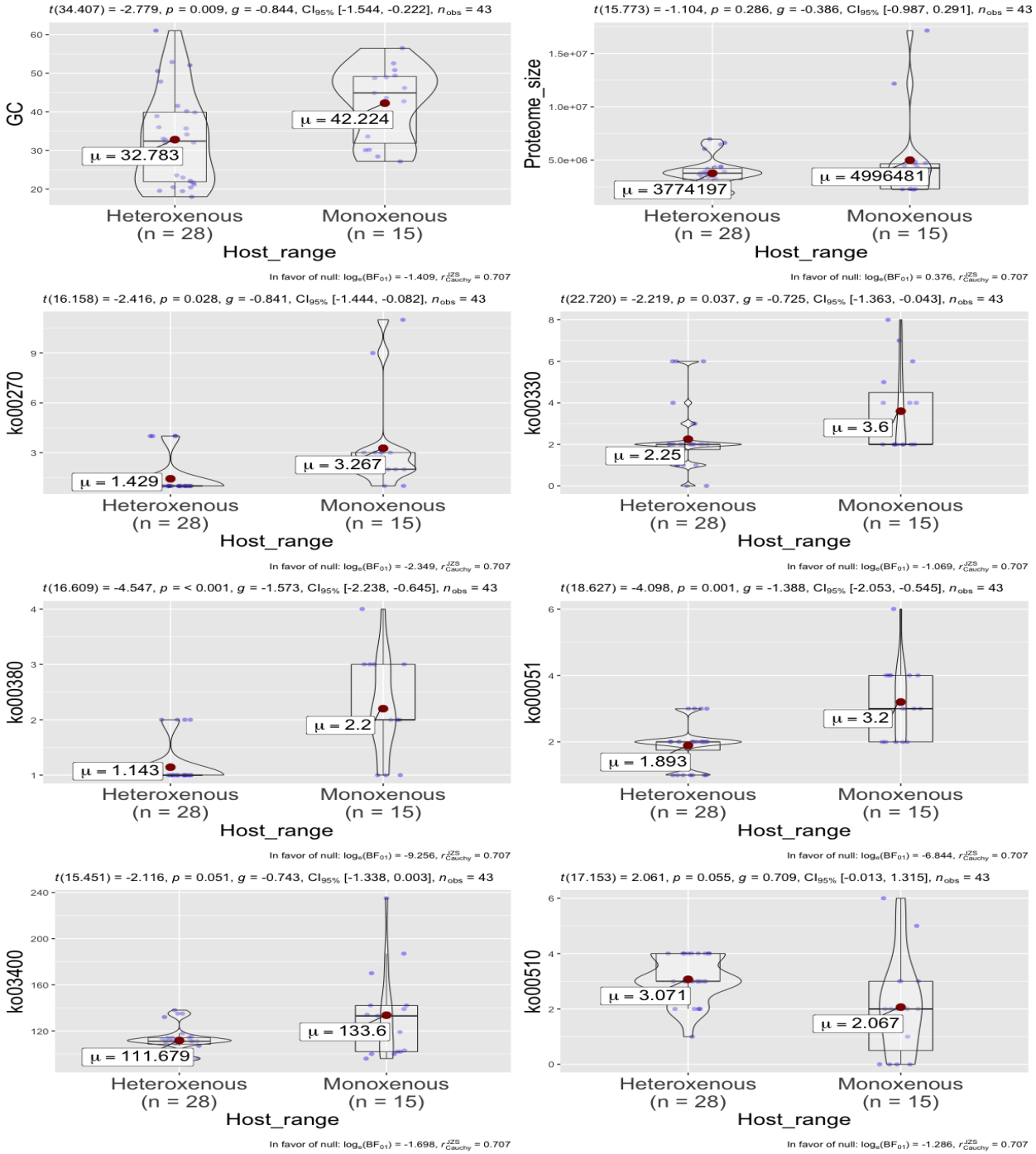

Figure 2.4.4.3: Comparison of different variables between monoxenous and heteroxenous species. Here, ko00270: Cysteine and methionine metabolism, ko00330: Arginine and proline metabolism, ko00380: Tryptophan metabolism, ko00051: Fructose and mannose metabolism, ko03400: DNA repair and recombination proteins, ko00510: N-Glycan biosynthesis

KEGG categories of heteroxenous species were compared to monoxenous species to assess if life cycle complexity has an influence on individual pathways. The mean value of proteome size in heteroxenous species is 3774197 and in heteroxenous species, it is 4996481 but these two means are not significantly different (parametric t-test result's Pvalue=0.286, Fig. 2.4.4.3). Heteroxenous species have fewer genes in Cysteine and

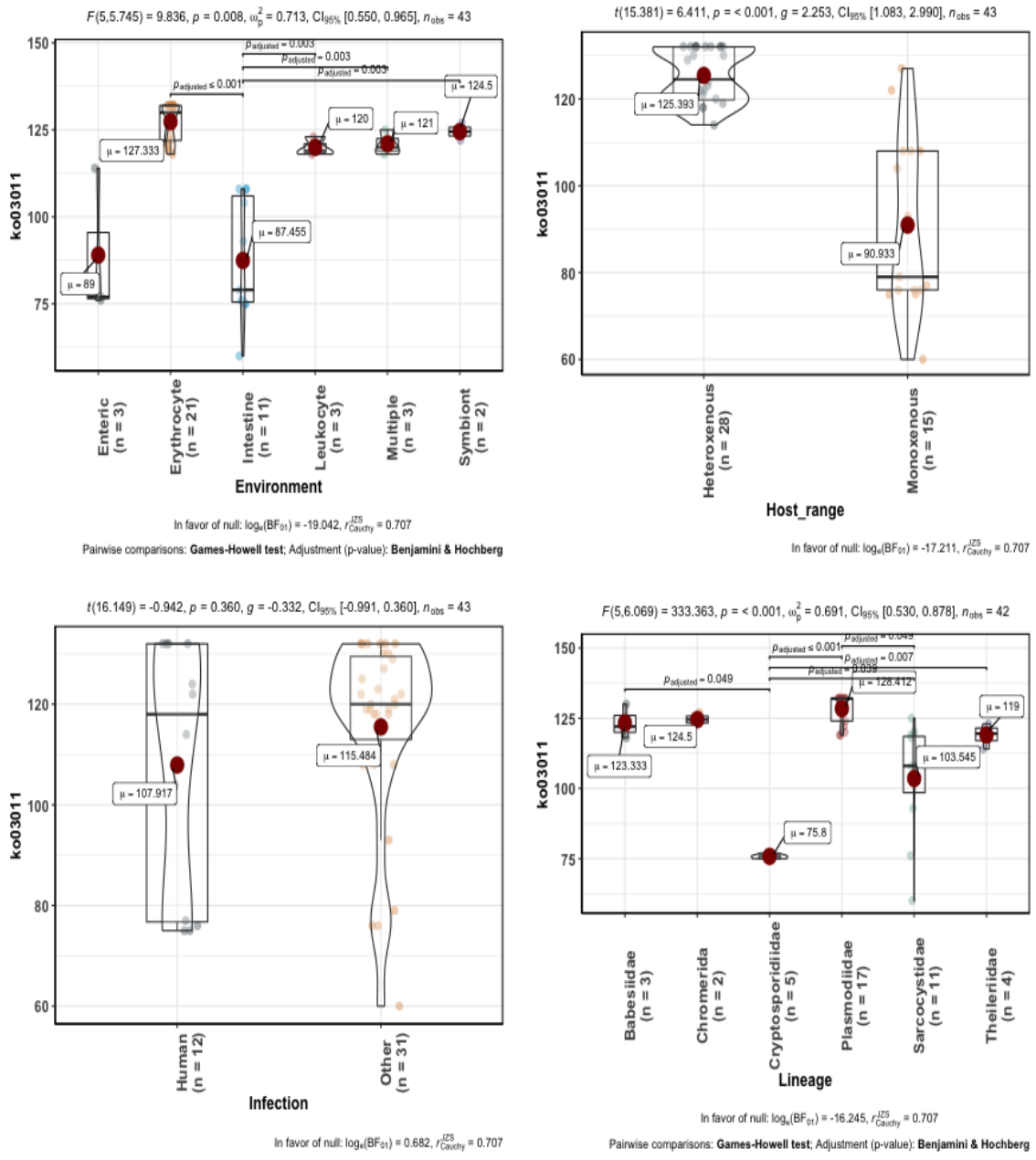

Figure 2.4.4.4: Abundance of Ribosome (KO03011) in different group of species

methionine metabolism (KO00270), Arginine and proline metabolism (KO00330), Tryptophan metabolism (KO00380), Fructose and mannose metabolism (KO00051) and DNA repair and recombination proteins (KO03400) compared to monoxenous species. (Fig. 2.4.4.3).

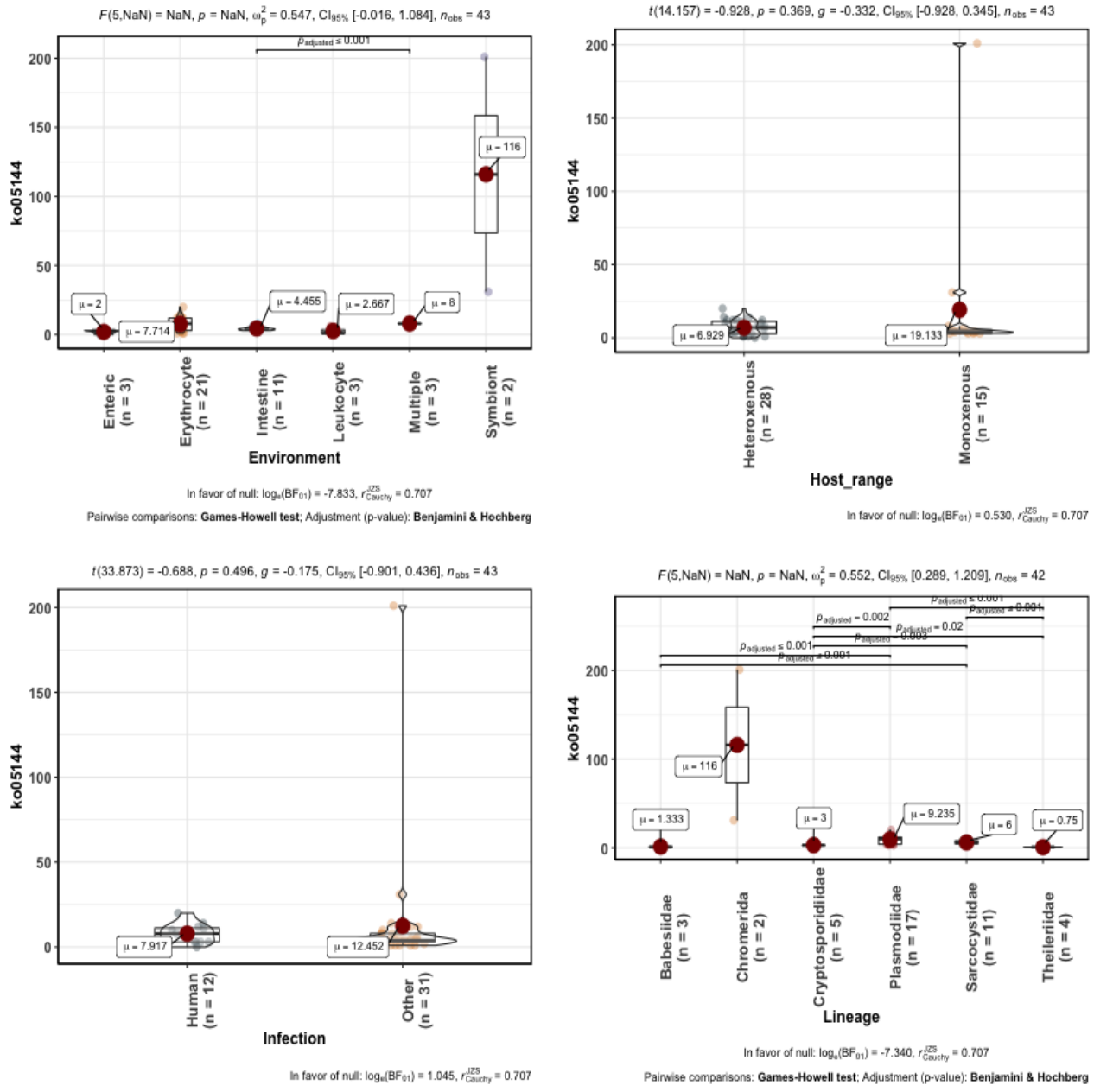

Figure 2.4.4.5: Abundance of genes in Malaria pathway in different group of species.

In contrast, heteroxenous species have more genes in the Thiamine biosynthesis (KO00730), Ribosome (KO03011) and N-Glycan biosynthesis pathway (KO00510) compared to monoxenous species (Fig. S2.4.4.3-4).

All the species except *C. felis*, have some genes in the Malaria pathway (KO05144) which is also significantly correlated with proteome size (PIC adjusted R=0.71; P value<0.0001). Chromerids have more genes in this pathway than any other lineage and Plasmodium has significantly more genes in this category than any other Apicomplexan, possibly due to annotation bias. The high abundance of genes in Chromerids implies an early origin of this pathway, which was then reduced in the Apicomplexans (Fig. 2.4.4.5).

The loss of orthogroups associated with the biosynthesis and metabolism of soluble metabolites implies that these are obtained from the host, presumably via trans-membrane transportation. These metabolites may be regarded as public goods within the host cell, and the blood stream, which are used for the benefit of the host organism as a whole. This perspective relies on the observation that cells of the body act in a cooperative fashion, with each other. Hence, game theory may be used to understand the production and distribution of metabolites (public goods) within the body.

In this scenario, the apicomplexan parasite represents an interlocutor, that intercepts the host's public goods for its own benefit, but to the detriment of the host. Thus, parasitism may be regarded as a form of freeloading behavior. However, the parasite would be expected to moderate its uptake of metabolites, given that excessive virulence may not be to the benefit of the parasite, as it may prematurely kill the host, which would curtail opportunities for dissemination [156]. Thus, parasite infection does not represent a zero sum game, but a situation where the players (parasite and host) have misaligned interests. The greater the degree of misalignment, the greater the expected virulence of the parasite. Thus, the mutualism-parasitism continuum, which represents the transition over evolutionary time from parasitism to mutualism, and a concomitant reduction in virulence [157], can be viewed as a gradual alignment of utilities between the infecting microbe and host.

Lastly, DNA repair and recombination proteins (KO03400) show a correlation with proteome size (R=0.52, p < 0.001). This observation is consistent with the Proteomic Constraint hypothesis, discussed next.

The amount of information in a genome (approximated to the proteome size) is expected to be related to the mutation rate (μ) as follows [158]:

$$\mu = k(2N_e\bar{s}\pi P)^{-1} \quad [1]$$

Where N_e is the effective population size,

$\bar{s}$ is the average selective pressure of a mutation (which will be deleterious on average), π is the genomic heterozygosity (per bp), P is the proteome size (in amino acids), and k is a proportionality constant.

The equation is a formal expression of the natural expectation that more genetic information in a genome should result in a higher selection pressure to reduce the mutation rate, given that the size of the mutational target is more substantial. This is expected to result in the evolution of enhanced DNA repair, reflected in the expansion of DNA repair pathways. The opposite is expected to occur when a genome reduces in size, which is expected to be accompanied by the loss and simplification of DNA repair pathways. Our results show that this is the case in the Apicomplexans, where a correlation is observed between the numbers of genes in orthogroups involved in DNA repair, and proteome size (Fig. 2.4.4.2)

Table 2.4.1: Blast results summary of *P. falciparum* and *P. reichenowi* specific orthogroups. Here only top hits were presented. gid=gene identification number from gene bank.

orthogroups	gene name	definition of top hit	Hit gid	GO:ID
OG0005027	PF3D7_0101400, PF3D7_0114000, PF3D7_0222000, PF3D7_0324000, PF3D7_0601300, PF3D7_0631300, PF3D7_0700900, PF3D7_0701700, PF3D7_0713200, PF3D7_1039600, PF3D7_1100900, PF3D7_1101800, PRCDC_0112200, PRCDC_0323300, PRCDC_0728800	exported protein family 1, DNAJ domain-containing protein, hypothetical protein PFHG_04793	86170384	0016021, 0020036
OG0011709	PF3D7_1039100, PF3D7_1102200, PRCDC_1100700	hypothetical protein PFFVO_03120	574749755	0016021,
OG0011710	PF3D7_1039200, PF3D7_1102100, PRCDC_1100600	hypothetical protein PFAG_03110	579333139	0016021
OG0015129	PF3D7_0100700, PRCDC_0035800	Plasmodium exported protein, unknown function,fragment	1371546166	0016021
OG0015130	PF3D7_0114400, PRCDC_0112600	hypothetical protein PFHG_05032	914550780	NA
OG0015131	PF3D7_0114500, PRCDC_0112700	Plasmodium exported protein (hyp10), unknown function	124505909	0016021, 0020011
OG0015132	PF3D7_0114600, PRCDC_0058900	stevor PIR protein, putative	1370985367	0016021, 0020036, 0020013
OG0015133	PF3D7_0114900, PRCDC_0800300	Plasmodium exported protein, unknown function	1370987008	0016021
OG0015134	PF3D7_0221100, PRCDC_0219800	hypothetical protein PFFVO_02284, Plasmodium	574750822	NA

		exported protein (PHISTa-like),		
OG0015135	PF3D7_0221800, PRCDC_0220100	hypothetical protein PFFVO_00367,	124801515	NA
OG0015136	PF3D7_0402900, PRCDC_0400500	probable protein, unknown function,	124505199	NA
OG0015137	PF3D7_0424200, PRCDC_0421400	reticulocyte binding protein homologue 4,	296004458	0016021, 0020008, 0008201, 0046789, 003026, 0044650
OG0015138	PF3D7_0500700, PRCDC_0007600	hypothetical protein PBILCG01_0400400	1370991177	NA
OG0015139	PF3D7_0532500, PRCDC_0531500	Plasmodium exported protein, unknown function	124506551	0016021, 0020011
OG0015140	PF3D7_0731300, PRCDC_0728100	hypothetical protein PFFVO_01899	574751304	0016021, 0020011
OG0015143	PF3D7_0917400, PRCDC_0915400	conserved Plasmodium protein, unknown function	124506895	NA
OG0015144	PF3D7_1001000, PRCDC_1000400	Plasmodium exported protein (hyp12), unknown function	124801890	0016021, 0020011
OG0015145	PF3D7_1008300, PRCDC_1007700	hypothetical protein PGO_080700	1194443552	NA
OG0015147	PF3D7_1102300, PRCDC_1100800	Plasmodium exported protein, unknown function	258597185	0016021
OG0015148	PF3D7_1129850, PRCDC_1128250	hypothetical protein PFNF135_03571	574980301	0016021
OG0015149	PF3D7_1219200, PRCDC_1218500	hypothetical protein PFMC_03765	575000005	0016021, 0030130, 0030132, 0071439, 000519, 0032051,0006886, 0016192, 0048268

OG0015150	PF3D7_1240200, PRCDC_1239400	erythrocyte membrane protein 1	910270594	0016021, 0046789, 0009405
OG0015151	PF3D7_1312450, PRCDC_1311450	apical ring associated protein 1, putative	1371546616	0016021, 0020011,
OG0015152	PF3D7_1334900, PRCDC_1333900	MSP7-like protein, partial	1304175241	NA
OG0015153	PF3D7_1335200, PRCDC_1334300	reticulocyte binding protein 2 homologue a, putative	1370979330	0016021
OG0015154	PF3D7_1409500, PRCDC_1408800	conserved Plasmodium protein, unknown function	124808239	NA
OG0015155	PF3D7_1412900, PRCDC_1412200	ubiquitin-conjugating enzyme, putative	1371546968	0005634, 0016874, 0061631, 0016567, 0043161
OG0015156	PF3D7_1413300, PRCDC_1412600	conserved Plasmodium protein, unknown function	258549184	0016021
OG0015157	PF3D7_1471800, PRCDC_1471000	conserved Plasmodium protein, unknown function	124810365	NA
OG0015158	PF3D7_1478000, PRCDC_1477000	Plasmodium exported protein (PHISTa), unknown function	124810562	0016021
OG0015159	PF3D7_1478700, PRCDC_1477700	hypothetical protein PFNF54_05896	583222326	0016021

Table 2.4.2: Descriptive statistics of pathway abundance in 41 Apicomplexa. *C. velia* and *V. brassicaformis* were excluded from this part of analysis.

Pathway	count	mean	std	Min	25%	50%	75%	max
Exosome	41	127.61	23.34	94	104	128	141	178
Chromosome and associated proteins	41	123.02	26.21	85	109	112	143	196
Ribosome biogenesis	41	117.76	11.95	108	108	114	121	157
DNA repair and recombination proteins	41	114.85	16.82	91	102	111	119	170
Ribosome	41	112.83	21.29	60	108	120	130	132
Membrane trafficking	41	108.15	14.91	81	94	112	116	134
Spliceosome	41	106.29	10.98	87	101	105	108	133
Transfer RNA biogenesis	41	86.68	10.94	64	79	90	94	101
Mitochondrial biogenesis	41	85.8	19.55	38	81	94	98	109

Transporters	41	66.07	12.59	44	58	69	70	93
Ubiquitin system	41	62	11.66	47	56	60	66	91
Messenger RNA biogenesis	41	59.51	5.48	46	57	60	61	74
Cytoskeleton proteins	41	52.73	21.15	13	34	63	64	80
Translation factors	41	50.73	5.67	39	45	53	54	62
Transcription machinery	41	45.68	5.06	36	43	44	46	61
Peptidases	41	39.54	11.38	21	32	39	48	61
Chaperones and folding catalysts	41	38.93	7.52	24	35	41	42	51
Proteasome	41	38.15	3.32	26	37	37	39	48
Protein phosphatases and associated proteins	41	28	8.22	14	24	27	35	44
Enzymes with EC numbers	41	26.8	8.78	15	22	23	33	44
DNA replication proteins	41	23.68	1.85	19	23	23	25	27
Lipid biosynthesis proteins	41	23.54	10.05	7	18	23	25	42
Protein kinases	41	17.46	3.98	8	15	16	18	27
Glycosyltransferases	41	16.37	7.34	6	11	13	23	29
Protein processing in endoplasmic reticulum	41	13.78	2.7	8	12	14	16	17
Secretion system	41	12.49	1.38	8	12	13	13	14
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	41	11	1.5	6	11	11	12	14
Human papillomavirus infection	41	10.9	0.86	8	11	11	11	13
Amino sugar and nucleotide sugar metabolism	41	9.37	1.87	6	9	10	11	13
Huntington disease	41	8.12	2.34	1	8	9	9	10
Glycerophospholipid metabolism	41	7.27	2.28	3	6	7	8	13
Pyrimidine metabolism	41	7.05	1.09	4	7	7	7	9
Purine metabolism	41	6.51	2.79	3	5	7	7	20
Nicotinate and nicotinamide metabolism	41	6.12	2.43	3	4	6	7	11
Phosphatidylinositol signaling system	41	5.95	1.45	5	5	5	6	11
Cushing syndrome	41	5.9	4.44	1	2	4	9	17
Transcription factors	41	5.88	1.52	2	5	6	6	9
Pantothenate and CoA biosynthesis	41	5.2	1.1	2	5	5	6	8
Drug metabolism - other enzymes	41	4.98	1.71	1	4	5	6	9
Prenyltransferases	41	4.76	1.02	1	5	5	5	7
Glycosylphosphatidylinositol (GPI)-anchored proteins	41	4.63	2.96	1	3	3	6	13
Sphingolipid signaling pathway	41	4.22	1.46	1	3	5	5	7
Signaling proteins	41	4.1	1.18	2	3	4	5	6
Choline metabolism in cancer	41	3.85	1.73	1	3	3	5	8
Translation	41	3.32	0.85	2	3	3	4	6
Platinum drug resistance	41	3.27	1.66	1	2	3	4	7
Carbon fixation in photosynthetic organisms	41	2.98	1.19	1	3	3	3	6

Structural proteins	41	2.95	1.66	1	2	2	4	9
Ion channels	41	2.76	1.2	2	2	2	3	6
Phosphonate and phosphinate metabolism	41	2.61	0.8	2	2	2	3	5
Oxidative phosphorylation	41	2.56	1.12	1	1	3	3	5
Protein export	41	2.54	0.78	1	2	3	3	4
Antifolate resistance	41	2.46	0.67	1	2	2	3	4
One carbon pool by folate	41	2.39	0.77	1	2	3	3	3
Butanoate metabolism	41	2.24	0.73	1	2	2	3	3
Fructose and mannose metabolism	41	2.22	0.85	1	2	2	3	4
Prokaryotic defense system	41	2.07	0.75	1	2	2	2	5
Wnt signaling pathway	41	2.05	0.38	1	2	2	2	3
Hepatocellular carcinoma	41	2.02	0.85	1	1	2	2	4
Human T-cell leukemia virus 1 infection	41	1.88	0.46	1	2	2	2	3
Small cell lung cancer	41	1.88	1.03	1	1	2	3	4
PPAR signaling pathway	41	1.76	1.2	1	1	1	2	5
Transcription	41	1.76	1.04	1	1	1	3	5
Cysteine and methionine metabolism	41	1.68	1.04	1	1	1	2	4
Insulin signaling pathway	41	1.63	1.13	1	1	1	2	5
MAPK signaling pathway - yeast	41	1.59	0.63	1	1	2	2	3
Longevity regulating pathway - worm	41	1.49	0.9	1	1	1	2	4
Tryptophan metabolism	41	1.44	0.71	1	1	1	2	4
Fat digestion and absorption	41	1.29	0.51	1	1	1	2	3
Proteoglycans	41	1.29	0.78	1	1	1	1	5
General function prediction only	41	1.15	0.42	1	1	1	1	3
Histidine metabolism	41	1.1	0.3	1	1	1	1	2
Thermogenesis	41	1.02	0.16	1	1	1	1	2
Parathyroid hormone synthesis@ secretion and action	41	1	0	1	1	1	1	1
GTP-binding proteins	40	6.78	1.31	4	5.75	7	8	8
Malaria	40	6.22	4.47	1	3	4	8.5	20
Toxoplasmosis	40	4.32	2.92	1	3	3	5.25	11
CD molecules	40	4.08	2.54	1	3	3	4.25	15
Ubiquinone and other terpenoid-quinone biosynthesis	40	2.05	0.39	1	2	2	2	3
Rheumatoid arthritis	40	1.52	0.72	1	1	1	2	3
Two-component system	40	1.3	0.46	1	1	1	2	2
Taste transduction	40	1.08	0.35	1	1	1	1	3
mRNA surveillance pathway	40	1	0	1	1	1	1	1
Arginine and proline metabolism	39	2.62	1.46	1	2	2	3	6
Endocytosis	39	1.72	0.65	1	1	2	2	4

Cofactor metabolism	39	1.67	0.58	1	1	2	2	3
Amino acid metabolism	39	1.23	0.43	1	1	1	1	2
Vibrio cholerae infection	39	1	0	1	1	1	1	1
Non-alcoholic fatty liver disease (NAFLD)	38	9.92	2.85	2	9.25	11	11	16
Peroxisome	38	2.68	1.97	1	2	2	2	8
Lipoic acid metabolism	38	2.47	0.98	1	1.25	3	3	4
Carbon fixation pathways in prokaryotes	38	2.26	0.69	1	2	2	2.75	5
Lysine degradation	38	1.66	1.1	1	1	1	2	7
Propanoate metabolism	38	1.42	0.89	1	1	1	1	4
Apoptosis - fly	38	1	0	1	1	1	1	1
Amino acid related enzymes	37	4.41	2.39	1	4	4	6	9
N-Glycan biosynthesis	37	2.86	0.86	1	2	3	3	4
Alanine@ aspartate and glutamate metabolism	36	4.11	0.85	3	4	4	5	7
Pyruvate metabolism	36	3.81	1.69	1	2.75	4.5	5	6
Riboflavin metabolism	36	1.94	0.33	1	2	2	2	3
Retrograde endocannabinoid signaling	36	1.75	0.5	1	1	2	2	3
Terpenoid backbone biosynthesis	35	7.97	1.4	7	7	8	8	15
Porphyrin and chlorophyll metabolism	35	5.14	2.17	1	5.5	6	6	8
Valine@ leucine and isoleucine degradation	35	4.6	1.06	3	4	5	5	7
Thiamine metabolism	35	2.23	1.42	1	1	1	4	5
Lipid metabolism	35	1.91	0.28	1	2	2	2	2
beta-Alanine metabolism	35	1.89	0.53	1	2	2	2	3
Glycine@ serine and threonine metabolism	35	1.66	1.16	1	1	1	2	5
Plant-pathogen interaction	35	1.6	0.5	1	1	2	2	2
Lectins	35	1.34	0.54	1	1	1	2	3
Renal cell carcinoma	35	1.06	0.24	1	1	1	1	2
Folate biosynthesis	34	4	1.63	1	4	4	4.75	7
Glyoxylate and dicarboxylate metabolism	34	2.29	1.71	1	1	3	3	10
Glutathione metabolism	34	2.26	0.9	1	1	3	3	3
Vitamin digestion and absorption	34	1.79	0.48	1	2	2	2	3
Phenylpropanoid biosynthesis	34	1.68	1.12	1	1	1	2	5
Pentose phosphate pathway	34	1.65	1.12	1	1	1	2	5
Glycosaminoglycan binding proteins	34	1.5	0.83	1	1	1	2	5
Glucagon signaling pathway	34	1.26	0.51	1	1	1	1	3
Taurine and hypotaurine metabolism	34	1.15	0.44	1	1	1	1	3
alpha-Linolenic acid metabolism	34	1.03	0.17	1	1	1	1	2
Steroid hormone biosynthesis	33	1.15	0.36	1	1	1	1	2
Aminobenzoate degradation	33	1	0	1	1	1	1	1
RNA transport	33	1	0	1	1	1	1	1
Streptomycin biosynthesis	32	1.25	0.44	1	1	1	1.25	2

Non-small cell lung cancer	32	1.16	0.37	1	1	1	1	2
Thyroid hormone synthesis	32	1.12	0.34	1	1	1	1	2
Photosynthesis proteins	32	1	0	1	1	1	1	1
Protein processing	32	1	0	1	1	1	1	1
Selenocompound metabolism	32	1	0	1	1	1	1	1
Arginine biosynthesis	31	1.58	0.5	1	1	2	2	2
Necroptosis	31	1.03	0.18	1	1	1	1	2
Others	31	1	0	1	1	1	1	1
Lysosome	30	1.93	0.37	1	2	2	2	3
Autophagy - animal	30	1.53	0.9	1	1	1	2.5	3
Phenylalanine@ tyrosine and tryptophan biosynthesis	30	1	0	1	1	1	1	1
Vitamin B6 metabolism	29	2.86	0.95	1	3	3	3	4
Nitrogen metabolism	29	1.38	0.56	1	1	1	2	3
Autophagy - yeast	29	1.28	0.53	1	1	1	1	3
Sphingolipid metabolism	29	1	0	1	1	1	1	1
Amoebiasis	28	1.5	1.2	1	1	1	1	6
Sulfur relay system	28	1.29	0.66	1	1	1	1	3
Biosynthesis of ansamycins	28	1.04	0.19	1	1	1	1	2
Central carbon metabolism in cancer	28	1	0	1	1	1	1	1
Viral carcinogenesis	27	1.04	0.19	1	1	1	1	2
Insulin resistance	26	2.15	0.83	1	2	2	3	4
Aminoacyl-tRNA biosynthesis	26	1.85	0.46	1	2	2	2	3
Primary immunodeficiency	26	1.15	0.37	1	1	1	1	2
Cell growth	26	1	0	1	1	1	1	1
Methane metabolism	25	1.04	0.2	1	1	1	1	2
MAPK signaling pathway - plant	22	1	0	1	1	1	1	1
Replication and repair	21	1	0	1	1	1	1	1
Chemical carcinogenesis	19	1.53	2.29	1	1	1	1	11
p53 signaling pathway	19	1	0	1	1	1	1	1
Starch and sucrose metabolism	17	3.35	1.77	2	2	3	4	9
Transport	16	1.25	0.58	1	1	1	1	3
Carbohydrate digestion and absorption	16	1.06	0.25	1	1	1	1	2
Carbapenem biosynthesis	15	1.93	0.26	1	2	2	2	2
mTOR signaling pathway	12	2.25	1.14	1	1	2	3	4
Pancreatic secretion	12	2.08	0.51	1	2	2	2	3
Monobactam biosynthesis	11	2.64	1.96	1	1	1	5	5
Ether lipid metabolism	11	1	0	1	1	1	1	1
Influenza A	11	1	0	1	1	1	1	1
Antigen processing and presentation	10	1.4	0.52	1	1	1	2	2

Amyotrophic lateral sclerosis (ALS)	10	1.2	0.63	1	1	1	1	3
D-Glutamine and D-glutamate metabolism	10	1.1	0.32	1	1	1	1	2
Quorum sensing	10	1.1	0.32	1	1	1	1	2
Toluene degradation	10	1	0	1	1	1	1	1
Cyanoamino acid metabolism	9	1	0	1	1	1	1	1
Benzoate degradation	7	1.14	0.38	1	1	1	1	2
Glycolysis / Gluconeogenesis	6	1.33	0.82	1	1	1	1	3
Energy metabolism	6	1	0	1	1	1	1	1
African trypanosomiasis	5	1	0	1	1	1	1	1
Morphine addiction	4	1.5	0.58	1	1	1.	2	2
Ascorbate and aldarate metabolism	4	1	0	1	1	1	1	1
Fatty acid elongation	4	1	0	1	1	1	1	1
Sulfur metabolism	4	1	0	1	1	1	1	1
Endocrine resistance	3	1	0	1	1	1	1	1
Tropane@ piperidine and pyridine alkaloid biosynthesis	3	1	0	1	1	1	1	1
Mineral absorption	1	1		1	1	1	1	1
Tyrosine metabolism	1	1		1	1	1	1	1

Table 2.4.3: Correlations between proteome size and different pathways. Pearson correlation, phylogenetic corrections with the Bayesian* and OrthoFinder** species tree

	R	p-value	R*	Pvalue*	R**	Pvalue**
Sulfur metabolism	0.94	< 2.2e-16	0.84	2.34E-12	0.92	< 2.2e-16
Lipid biosynthesis proteins	0.95	< 2.2e-16	0.82	2.41E-11	0.83	1.35E-11
Enzymes with EC numbers	0.93	< 2.2e-16	0.81	5.74E-11	0.89	2.92E-15
Cytoskeleton proteins	0.91	< 2.2e-16	0.79	3.27E-10	0.87	8.95E-14
Ascorbate and aldarate metabolism	0.89	7.68E-16	0.78	9.26E-10	0.81	8.69E-11
Alanine aspartate and glutamate metabolism	0.92	< 2.2e-16	0.77	2.70E-09	0.87	4.60E-14
Protein phosphatases and associated proteins	0.95	< 2.2e-16	0.76	3.75E-09	0.86	4.10E-13
Carbohydrate metabolism	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Retinol metabolism	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
RIG I like receptor signaling pathway	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Glycerolipid metabolism	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Porphyrin and chlorophyll metabolism	0.89	5.83E-16	0.76	4.57E-09	0.84	5.14E-12
Carotenoid biosynthesis	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11

Sesquiterpenoid and triterpenoid biosynthesis	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Pentose and glucuronate interconversions	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Cytochrome P450	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Biosynthesis of secondary metabolites unclassified	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Dorso ventral axis formation	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Steroid biosynthesis	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Styrene degradation	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Various types of N glycan biosynthesis	0.86	2.01E-13	0.76	3.73E-09	0.81	8.27E-11
Cysteine and methionine metabolism	0.88	1.02E-14	0.75	1.11E-08	0.64	4.17E-06
Photosynthesis proteins	0.86	1.61E-13	0.74	1.71E-08	0.77	2.38E-09
Chromosome and associated proteins	0.93	< 2.2e-16	0.74	1.65E-08	0.85	6.00E-13
Transfer RNA biogenesis	0.9	3.95E-16	0.73	5.08E-08	0.85	1.91E-12
Transporters	0.92	< 2.2e-16	0.72	1.02E-07	0.73	5.55E-08
Tyrosine metabolism	0.78	8.51E-10	0.71	1.53E-07	0.74	2.05E-08
Amino acid related enzymes	0.96	< 2.2e-16	0.71	1.32E-07	0.83	7.88E-12
Malaria	0.82	1.59E-11	0.71	1.41E-07	0.77	3.05E-09
Secretion system	0.86	1.63E-13	0.68	7.65E-07	0.76	4.63E-09
Ion channels	0.87	4.01E-14	0.68	6.76E-07	0.6	2.83E-05
Lysosome	0.85	3.54E-13	0.67	1.19E-06	0.78	9.78E-10
Nitrogen metabolism	0.87	3.40E-14	0.66	1.80E-06	0.61	1.82E-05
Mineral absorption	0.87	1.86E-14	0.65	3.00E-06	0.85	1.66E-12
Insulin signaling pathway	0.89	7.88E-16	0.65	3.44E-06	0.76	5.29E-09
Phenylpropanoid biosynthesis	0.89	1.03E-15	0.65	3.56E-06	0.78	1.26E-09
Spliceosome	0.93	< 2.2e-16	0.64	4.51E-06	0.82	1.86E-11
Pattern recognition receptors	0.74	1.12E-08	0.64	5.88E-06	0.71	1.53E-07
Thyroid hormone synthesis	0.91	< 2.2e-16	0.64	4.97E-06	0.69	4.24E-07
Fatty acid elongation	0.75	5.48E-09	0.64	5.78E-06	0.69	3.49E-07
Monobactam biosynthesis	0.8	1.23E-10	0.64	5.16E-06	0.52	0.0004148
Oxidative phosphorylation	0.92	< 2.2e-16	0.63	7.12E-06	0.77	2.00E-09
Fructose and mannose metabolism	0.78	9.32E-10	0.62	1.38E-05	0.72	1.01E-07
Exosome	0.93	< 2.2e-16	0.62	1.00E-05	0.77	3.55E-09
p53 signaling pathway	0.87	2.68E-14	0.61	1.58E-05	0.79	3.64E-10
Peroxisome	0.88	3.45E-15	0.61	2.11E-05	0.64	4.80E-06
Platinum drug resistance	0.86	7.92E-14	0.61	1.86E-05	0.75	1.23E-08
Translation factors	0.92	< 2.2e-16	0.61	1.96E-05	0.73	3.27E-08
alpha Linolenic acid metabolism	0.83	3.89E-12	0.61	1.63E-05	0.81	7.00E-11
Chaperones and folding catalysts	0.84	2.59E-12	0.61	1.51E-05	0.43	0.00477

Folate biosynthesis	0.78	9.26E-10	0.6	2.23E-05	0.63	7.85E-06
Mitochondrial biogenesis	0.71	8.38E-08	0.59	3.79E-05	0.54	0.0002215
Lectins	0.87	4.16E-14	0.58	6.21E-05	0.73	3.71E-08
Cushing syndrome	0.87	2.57E-14	0.57	9.59E-05	0.71	1.46E-07
Protein kinases	0.88	1.00E-14	0.57	7.92E-05	0.69	5.09E-07
Valine leucine and isoleucine degradation	0.67	1.09E-06	0.57	8.60E-05	0.64	6.11E-06
Energy metabolism	0.86	1.56E-13	0.56	0.000109	0.77	3.16E-09
Phenylalanine tyrosine and tryptophan biosynthesis	0.83	7.48E-12	0.56	0.000133	0.7	2.34E-07
Translation	0.84	1.43E-12	0.56	0.000123	0.74	2.02E-08
Membrane trafficking	0.9	< 2.2e-16	0.54	0.000247	0.69	3.39E-07
Nicotinate and nicotinamide metabolism	0.81	3.35E-11	0.54	0.000216	0.45	0.002632
Amino sugar and nucleotide sugar metabolism	0.89	2.93E-15	0.54	0.0002083	0.81	7.54E-11
beta Alanine metabolism	0.71	7.27E-08	0.53	0.000275	0.74	1.61E-08
Toluene degradation	0.82	1.42E-11	0.53	0.000323	0.68	8.82E-07
Central carbon metabolism in cancer	0.49	0.000900	0.52	0.000372	0.59	3.70E-05
DNA repair and recombination proteins	0.84	2.04E-12	0.52	0.000468	0.39	0.01086
One carbon pool by folate	0.74	1.25E-08	0.52	0.000458	0.66	1.85E-06
Ubiquitin system	0.77	1.43E-09	0.52	0.000412	0.21	0.189
Ubiquinone and other terpenoid quinone biosynthesis	0.79	1.95E-10	0.52	0.000366	0.61	1.87E-05
Carbon fixation in photosynthetic organisms	0.84	1.39E-12	0.51	0.000564	0.46	0.002114
Transcription machinery	0.91	< 2.2e-16	0.51	0.000597	0.62	1.38E-05
Others	0.75	5.56E-09	0.51	0.000531	0.82	4.57E-11
Longevity regulating pathway worm	0.83	4.18E-12	0.5	0.000852	0.52	0.0003898
Sulfur relay system	0.8	1.42E-10	0.5	0.000811	0.65	2.80E-06
Chemical carcinogenesis	0.56	1.00E-04	0.49	0.000928	0.3	0.05133
Riboflavin metabolism	0.75	8.40E-09	0.48	0.00128	0.48	0.001399
Histidine metabolism	0.85	8.90E-13	0.48	0.001144	0.56	0.0001246
Messenger RNA biogenesis	0.86	2.18E-13	0.47	0.001731	0.65	3.14E-06
Vitamin B6 metabolism	0.61	1.26E-05	0.47	0.001744	0.47	0.001732
Hepatocellular carcinoma	0.78	4.65E-10	0.46	0.002363	0.51	0.0005155
Morphine addiction	0.75	7.82E-09	0.46	0.002196	0.77	1.77E-09
Glycosyltransferases	0.64	4.15E-06	0.45	0.002559	0.21	0.1831
Function unknown	0.77	1.52E-09	0.45	0.002725	0.4	0.008783

Glycine serine and threonine metabolism	0.84	3.37E-12	0.44	0.00383	0.62	1.11E-05
Arginine and proline metabolism	0.76	3.79E-09	0.43	0.004303	0.37	0.01549
Pyrimidine metabolism	0.71	1.09E-07	0.43	0.004374	0.3	0.05595
Protein processing	0.57	6.63E-05	0.42	0.006041	0.53	0.0002658
Signaling proteins	0.74	1.54E-08	0.42	0.006001	0.55	0.0001448
Tropane piperidine and pyridine alkaloid biosynthesis	0.74	1.43E-08	0.42	0.005934	0.55	0.0001858
Antigen processing and presentation	0.8	1.21E-10	0.42	0.005798	0.57	9.36E-05
Propanoate metabolism	0.79	4.05E-10	0.41	0.006468	0.76	5.75E-09
Glycolysis Gluconeogenesis	0.43	0.004105	0.41	0.007256	0.13	0.4286
Protein processing in endoplasmic reticulum	0.8	1.46E-10	0.41	0.007252	0.62	1.00E-05
Peptidases	0.87	2.25E-14	0.4	0.008401	0.77	2.42E-09
Primary immunodeficiency	0.61	1.15E-05	0.4	0.007999	0.008	0.9605
DNA replication proteins	0.61	1.42E-05	0.4	0.008297	0.42	0.00551
Pyruvate metabolism	0.45	0.002407	0.4	0.008455	0.29	0.06518
Cell growth	0.39	0.009408	0.39	0.01022	0.59	3.98E-05
Autophagy yeast	0.77	1.16E-09	0.39	0.0116	0.48	0.001362
Pentose phosphate pathway	0.68	6.25E-07	0.39	0.01072	0.45	0.002777
Carbohydrate digestion and absorption	0.61	1.66E-05	0.39	0.01055	0.01	0.9318
N Glycan biosynthesis	0.43	0.004051	0.39	0.00985	0.34	0.02598
Methane metabolism	0.41	0.005903	0.39	0.01155	0.39	0.01053
Rheumatoid arthritis	0.42	0.005547	0.38	0.01235	0.58	6.77E-05
Non small cell lung cancer	0.48	0.001093	0.37	0.01703	0.45	0.002956
Prenyltransferases	0.77	1.27E-09	0.37	0.01498	0.69	5.04E-07
Ribosome biogenesis	0.66	1.66E-06	0.37	0.01573	0.52	0.0004359
Glycosaminoglycan binding proteins	0.79	2.37E-10	0.37	0.01544	0.52	0.0003691
Quorum sensing	0.7	1.31E-07	0.37	0.01603	0.54	0.0002075
Biosynthesis of ansamycins	0.66	1.47E-06	0.36	0.01948	0.47	0.001601
Glycerophospholipid metabolism	0.76	4.50E-09	0.36	0.02089	0.43	0.00441
General function prediction only	0.57	5.56E-05	0.36	0.02096	0.23	0.1501
Terpenoid backbone biosynthesis	0.44	0.00347	0.35	0.0228	0.5	0.000781
Carbapenem biosynthesis	0.5	0.000694	0.34	0.02605	0.53	0.0003085
Glucagon signaling pathway	0.76	4.90E-09	0.34	0.02733	0.64	5.27E-06
Lipid metabolism	0.5	0.000702	0.34	0.02983	0.08	0.6213
MAPK signaling pathway yeast	0.6	2.32E-05	0.33	0.03317	0.48	0.001407
Transport	0.64	3.26E-06	0.33	0.03283	0.08	0.5978
Ether lipid metabolism	0.6	1.84E-05	0.33	0.03204	0.43	0.004385

Sphingolipid metabolism	0.61	1.39E-05	0.32	0.03871	0.56	9.76E-05
CD molecules	0.66	1.34E-06	0.32	0.03945	0.34	0.02541
Purine metabolism	0.55	0.000124	0.31	0.04585	0.25	0.1103
Starch and sucrose metabolism	0.54	0.000175	0.31	0.04821	-0.07	0.6493
Carbon fixation pathways in prokaryotes	0.62	8.78E-06	0.31	0.04486	0.33	0.03338
Glutathione metabolism	0.43	0.004145	0.29	0.06509	0.12	0.4662
Ribosome	0.21	0.1864	0.29	0.05857	0.27	0.08879
Pancreatic secretion	0.55	0.000154	0.29	0.05834	0.32	0.04052
Huntington disease	0.55	0.000146	0.29	0.066	0.47	0.001635
Wnt signaling pathway	0.52	0.000355	0.29	0.06672	0.23	0.1391
MAPK signaling pathway plant	0.36	0.01855	0.29	0.05851	0.21	0.1847
Plant pathogen interaction	0.48	0.001261	0.28	0.07124	0.15	0.3357
Aminoacyl tRNA biosynthesis	0.44	0.003	0.27	0.08294	0.37	0.01673
Selenocompound metabolism	0.79	2.18E-10	0.27	0.08393	0.53	0.0002657
Transcription	0.47	0.001546	0.27	0.08962	0.26	0.09116
D Glutamine and D glutamate metabolism	0.56	0.0001068	0.27	0.08415	0.3	0.05348
Glycosylphosphatidylinositol GPI anchored proteins	0.68	6.59E-07	0.26	0.09001	0.55	0.0001677
Thiamine metabolism	0.36	0.01718	0.25	0.114	0.23	0.1455
Autophagy animal	0.32	4.00E-02	0.24	0.1223	0.16	0.3204
Human papillomavirus infection	0.29	0.06261	0.24	0.1256	0.13	0.4166
GTP binding proteins	0.49	0.000764	0.24	0.1315	0.44	0.003713
Benzoate degradation	0.64	3.70E-06	0.24	0.1186	0.22	0.1694
Endocytosis	0.4	0.007881	0.24	0.12	0.41	0.007317
Prokaryotic defense system	0.53	0.000242	0.24	0.1218	0.45	0.002748
Cyanoamino acid metabolism	0.49	0.000920	0.24	0.122	0.54	0.0002148
Drug metabolism other enzymes	0.62	8.10E-06	0.23	0.1368	0.36	0.01872
Phosphatidylinositol signaling system	0.56	9.49E-05	0.23	0.1417	0.26	0.09501
Two component system	0.33	0.02943	0.22	0.154	0.13	0.4269
mTOR signaling pathway	0.35	0.01988	0.22	0.1558	0.11	0.4948
Taurine and hypotaurine metabolism	0.38	0.01275	0.22	0.1674	0.39	0.01062
Insulin resistance	0.4	0.007209	0.22	0.156	0.06	0.69
Non alcoholic fatty liver disease NAFLD	0.29	0.06038	0.21	0.1728	0.24	0.1269
Transcription factors	0.28	0.07416	0.21	0.1832	0.33	0.03562
Cofactor metabolism	0.3	0.04778	0.2	0.1929	0.2	0.1962
Necroptosis	0.32	0.03793	0.2	0.211	0.17	0.284

Glycosylphosphatidylinositol GPI anchor biosynthesis	0.54	0.000199	0.2	0.1971	0.3	0.05202
PPAR signaling pathway	0.73	3.24E-08	0.2	0.2146	0.23	0.1462
RNA transport	0.34	0.02386	0.19	0.2297	0.23	0.1488
Vitamin digestion and absorption	0.42	0.005632	0.18	0.2571	0.28	0.06842
Viral carcinogenesis	0.42	0.004998	0.16	0.318	0.13	0.4273
Tryptophan metabolism	0.55	0.000116	0.16	0.3088	0.27	0.08441
Arginine biosynthesis	0.12	4.30E-01	0.16	0.2973	0.03	0.8537
Lipoic acid metabolism	0.44	0.002966	0.16	0.3133	0.18	0.2449
Glyoxylate and dicarboxylate metabolism	0.1	0.5421	0.15	0.3362	0.14	0.3763
Protein export	0.23	0.1424	0.15	0.3455	0.05	0.7476
Small cell lung cancer	0.34	0.02355	0.15	0.3527	-0.13	0.421
Proteasome	0.4	0.007208	0.14	0.3732	0.09	0.5841
Antifolate resistance	0.61	1.33E-05	0.13	0.427	0.37	0.01455
Amoebiasis	0.4	0.008	0.13	0.4183	0.29	0.06726
Vibrio cholerae infection	-.004	0.9772	0.12	0.4645	0.05	0.7536
Fat digestion and absorption	0.37	0.01401	0.12	0.4581	0.24	0.1276
Phosphonate and phosphinate metabolism	0.32	0.03423	0.11	0.4761	-0.08	0.6319
Lysine degradation	0.2	0.2061	0.11	0.4933	0.11	0.48
Choline metabolism in cancer	0.3	0.05464	0.09	0.5861	-0.09	0.5888
Streptomycin biosynthesis	0.34	0.02657	0.09	0.5839	0.18	0.266
Replication and repair	0.05	0.7288	0.09	0.5562	0.08	0.6364
Amino acid metabolism	0.38	0.013	0.06	0.7003	0.1	0.5333
Structural proteins	0.4	0.008102	0.05	0.7358	0.14	0.3663
Butanoate metabolism	-0.05	7.30E-01	0.04	0.8183	0.01	0.953
Proteoglycans	0.07	0.6728	0.03	0.8732	-0.006	0.9675
Endocrine resistance	0.22	0.1524	0.03	0.8638	-0.004	0.9795
Human T cell leukemia virus 1 infection	0.16	0.2913	0.03	0.8308	0.11	0.4838
Aminobenzoate degradation	-0.08	0.59	0.03	0.8293	0.07	0.6658
Thermogenesis	0.52	0.000313	0.03	0.8275	0.42	0.005411
Pantothenate and CoA biosynthesis	0.58	4.15E-05	0.03	0.8536	0.25	0.1035
Amyotrophic lateral sclerosis ALS	0.26	0.09	0.02	0.90s35	-0.11	0.4761
mRNA surveillance pathway	0.03	0.8644	0.01	0.9518	-0.003	0.9853
Influenza A	0.22	0.153	-0.0008	0.9959	-0.02	0.8915

Steroid hormone biosynthesis	0.33	0.03252	-0.004	0.9786	0.01	0.9289
Apoptosis fly	-0.03	0.83	-0.01	0.9456	0.05	0.774
Retrograde endocannabinoid signaling	0.17	0.287	-0.02	0.88	0.01	0.952
Renal cell carcinoma	0.22	0.1614	-0.02	0.8954	-0.05	0.7484
Sphingolipid signaling pathway	0.07	0.6472	-0.07	0.6373	0.1	0.5101
Toxoplasmosis	0.13	0.4056	-0.08	0.6066	-0.08	0.6084
Taste transduction	0.01	0.965	-0.14	0.3644	-0.06	0.6979
African trypanosomiasis	-0.26	0.095	-0.29	0.06079	-0.25	0.1164

Data S2.1: Orthogroups table (Gene symbol) (Data S2.1-S2. in github page; considering the size of the tables)
https://raw.githubusercontent.com/zillurbmb51/results/master/Orthogroups.csv

Data S2.2: Orthogroups table (Gene count)
https://raw.githubusercontent.com/zillurbmb51/results/master/Orthogroups.GeneCount.csv

Data S2.3: Species overlaps
https://github.com/zillurbmb51/results/blob/master/Orthogroups_SpeciesOverlaps.csv

Data S2.4: Unassigned genes
https://raw.githubusercontent.com/zillurbmb51/results/master/Orthogroups_UnassignedGenes.csv

Data S2.5: Over all statistics
https://github.com/zillurbmb51/results/blob/master/Statistics_Overall.csv

Data S2.6: Species-wise statistics
https://github.com/zillurbmb51/results/blob/master/Statistics_PerSpecies.csv

Data S2.7: Blast2go output of unique orthogroups found only in malaria causing *Plasmodia* in human and chimpanzee
https://github.com/zillurbmb51/results/blob/master/blast2go_combined2.csv

Data S2.8: Abundance of mapped functional orthologs in all 43 species
https://github.com/zillurbmb51/results/blob/master/functional_orthologs.csv

Chapter 3

Phylogenomics to Reconstruct the Species Tree

3.1: Abstract

In recent years, phylogenomics has appeared a promising method for reconstructions of evolutionary relationships among different lineages. In terms of biodiversity, Apicomplexa has many unclear taxonomic structures. In this study, a statistically robust phylogeny is reconstructed by concatenating 522 genes from the core Apicomplexan genome which accounts for 6068 amino acid sequences.

The amino acids frequencies in the whole genome are highly variable compared to the core genome and significantly correlated with GC (Guanine-Cytosine) content. The Bayesian and Maximum likelihood analysis of this data produced the same topology. The bipartition pattern of this Bayesian tree is better than the species tree inference from all genes (STAG; tree inference method used by OrthoFinder) in terms of evolutionary rate/branch lengths and at least one monophyletic clade. This study also discusses the effect of alignment filtering on tree topology.

3.2: Introduction

Phylogeny is the ultimate solution to resolve evolutionary relationships among observed organisms. It resolves both transformational and variational evolution. In transformational evolution, an entity changes over time. In variational evolution, groups of objects modify their comparative proportions through time which may lead to speciation [159,160]. The branch lengths of a phylogenetic tree represent transformational evolution. On the other hand, the relative branching order reflects variational evolution.

A homologous character matrix of morphological features, biochemical pathways, amino acids, and nucleotides sequences or any other characteristics can be used to infer phylogeny. Different methods can be employed to calculate evolutionary relationships from a given matrix, such as, least square, neighbor-joining, minimum evolution, Maximum parsimony, Maximum likelihood or Bayesian inference [161,162,163,164].

These methods use a substitution model to depict actual changes. In molecular phylogeny, substitution models describe the probability by which a set of characters changes into another set of homologous characters over time. So, to infer phylogeny, we

need a matrix, any of the above-mentioned methods and a substitution model. In some published phylogenies, different evolutionary relationships were inferred for same group of species that used different character matrix and evolutionary models [165].

Evolutionary relationship conjecture depends on types and numbers of characters in the input matrix, substitution model and method of inference. The character matrix should be carefully chosen for the given group of species. Phylogeny using molecular matrix aka DNA/RNA or amino acids sequence is usually considered as more correct and robust. The genes/proteins that are constitutive and mandatory to maintain essential cellular functions and expressed in all cell types in an organism (housekeeping genes) are suitable candidate to construct a phylogeny. Usually, ribosomal RNA genes/proteins are the most proper source to create a phylogenetic tree [166].

There are three nucleotide bases for each amino acid. This makes DNA/RNA sequences more informative than protein sequences. Amino acid sequences are more conserved than nucleotide sequences and relatively free from bias or noises such as GC content [167],[168],[169]. To explain this bias, the standard codon usage table of the model organism *P. falciparum* is presented in Table 3.1. From this table, we can assume that extremely AT-rich genome will produce fewer Arg (Arginine) compared to GC rich genome. There are six codons to translate Arg. Among these, two codons are completely GC, another three are GC rich, and only one is AT-rich. *P. falciparum* growth was found to be dependent on Arg concentrations in culture [170].

Table: 3.1: Codon usage table of *P. falciparum* (813541 codons). This table was excerpted from codon usage database for standard genetic code (https://www.kazusa.or.jp/codon/cgibin/showcodon.cgi?species=186763&aa=1&style=GCG)

Amino Acid	Codon	Number	Frequency/1000
Gly	GGG	2572	3.16
Gly	GGA	16033	19.71
Gly	GGT	14585	17.93
Gly	GGC	1785	2.19
Glu	GAG	8285	10.18
Glu	GAA	59540	73.19
Asp	GAT	44761	55.02
Asp	GAC	7412	9.11

Val	GTG	3278	4.03
Val	GTA	15727	19.33
Val	GTT	16063	19.74
Val	GTC	2201	2.71
Ala	GCG	1172	1.44
Ala	GCA	13502	16.6
Ala	GCT	11494	14.13
Ala	GCC	3276	4.03
Arg	AGG	2757	3.39
Arg	AGA	13606	16.72
Ser	AGT	17790	21.87
Ser	AGC	3080	3.79
Lys	AAG	14617	17.97
Lys	AAA	70309	86.42
Asn	AAT	76393	93.9
Asn	AAC	15261	18.76
Met	ATG	15015	18.46
Ile	ATA	32747	40.25
Ile	ATT	26226	32.24
Ile	ATC	4509	5.54
Thr	ACG	2943	3.62
Thr	ACA	18536	22.78
Thr	ACT	11589	14.25
Thr	ACC	4570	5.62
Trp	TGG	4647	5.71
End	TGA	197	0.24
Cys	TGT	13433	16.51
Cys	TGC	2069	2.54
End	TAG	236	0.29
End	TAA	1021	1.26
Tyr	TAT	31661	38.92
Tyr	TAC	4170	5.13
Leu	TTG	8213	10.1
Leu	TTA	41059	50.47
Phe	TTT	25698	31.59
Phe	TTC	6041	7.43

Ser	TCG	1987	2.44
Ser	TCA	15954	19.61
Ser	TCT	13361	16.42
Ser	TCC	4843	5.95
Arg	CGG	173	0.21
Arg	CGA	1750	2.15
Arg	CGT	2808	3.45
Arg	CGC	340	0.42
Gln	CAG	2705	3.32
Gln	CAA	21507	26.44
His	CAT	14172	17.42
His	CAC	3137	3.86
Leu	CTG	945	1.16
Leu	CTA	4537	5.58
Leu	CTT	7410	9.11
Leu	CTC	1443	1.77
Pro	CCG	844	1.04
Pro	CCA	14323	17.61
Pro	CCT	8702	10.7
Pro	CCC	2521	3.1

In table 3.1, it is observable that, there is more than one codon for each amino acid except met which is the start codon. So, a mutation in a codon may not change the actual amino acid in the protein (synonymous substitutions). If we use the amino acid matrix, there is a high chance of losing this information.

GC rich regions are inclined to have increased recombination rates than AT-rich regions, and thus high GC content can be negatively associated with phylogenetic accuracy [171]. It can be exemplified as that, two species which are distantly related to the evolutionary context can show significantly correlated if these two have the same GC content. Amino acids sequences are free from this type of bias [172],[173].

The least square method uses a linear distance to build a phylogenetic cluster. This method is accurate but limited in efficacy and speed. The neighbor-joining method uses an agglomerative approach to find pair of distinct taxa and then calculate distances to

add them as a node in an iterative algorithm. This method is faster but mostly dependent on the input matrix. In this method there are potentials of getting biased topology but still in most of the cases, it produced the right topology. The maximum parsimony method infers phylogeny from the optimal tree space which minimizes evolutionary events. This method is attractive in the field of evolution, but often it is difficult to find the most parsimonious tree and produce mistaken topology when to interpret distantly related lineages. Maximum likelihood uses probabilistic distributions to infer phylogeny. Till now, it is the most statistically standard process. Bayesian phylogeny uses maximum likelihood approach with a prior distribution and Markov chain Monte Carlo sampling algorithm [174],[175],[176],[177]. This study used both Maximum likelihood and Bayesian approach to construct phylogeny and compared the topology with other methods and with the cluster of orthologs.

Phylogenomics uses a concatenated core genome to resolve phylogenetic relationships of the observed group of species [178]. Phylogenomics is a relatively new concept in the field of evolution and genomics. It is a more reliable method to resolve evolutionary relationships compared to single gene phylogenies (more sequences, more information). Phylogenomic approaches are very much useful to improve functional annotations and evolutionary relationships even for uncharacterized genes [179].

The functional annotation is out of the scope of this study. Phylogenomic was handy to produce statistically robust and congruent results resolving previously controversial phylogenetic relationships in insect populations [180]. Previous studies using the whole genome as a homologous matrix showed significant improvements in resolving evolutionary events and unravel incongruity between gene trees and species trees [181].

C. velia and *V. brassicaformis* belong to the phylum *Chromerida* (a phylum of alveolates). They are photosynthetic and shows phylogenetically close relationships with the Apicomplexa. Recent multi-gene phylogeny revealed that these two species are more closely related to Apicomplexa than previously thought [182]. In the previous chapter, 522 orthogroups including more than 38,000 genes were found common all the observed species including these two. Here, these two species are used as outgroups.

Apicomplexan phylogeny was inferred from 30 apicoplast protein genes to determine phylogenetic relationships among the monophyletic rodent parasites [183]. A genome-scale phylogeny was also conducted to clarify the evolution of Apicomplexa and incongruence among gene trees [181]. SSU rDNA and ß-tubulin sequences were used to understand the evolution of parasitism in the species [184]. Genomic data from 15 Apicomplexan species was used to infer phylogeny and diversity of this group [29]. Phylogeny using apical membrane protein sequences divulged molecular diversity at the host-parasite interactions [185].

Conserved palmitoyltransferase sequence in Apicomplexan phylogeny provided evidence of the bacterial origin of this enzyme [186]. Molecular phylogeny revealed the independent origin of glucosamine-phosphate N-acetyltransferase in Apicomplexa which is essential for the survival and infectivity [187]. Phylogeny using sugar transporters proved their alveolate ancestry and showed their diversity in these organisms [188]. Phylogeny based on morphological and molecular data (SS rDNA) of marine gregarines explained the origin of *Cryptosporidium* and the initial stages of Apicomplexan evolution [189]. Interestingly, a comprehensive phylogeny using DNA sequence data of 34 *Haemosporidian* taxa showed a polyphyletic nature of *Plasmodium* [190].

Hence, a detailed phylogeny including all the common genes (core genome) will be helpful to define the exact position of the lineages. Though whole genome phylogeny is a promising method in the field of phylogeny and evolution, still it can be biased. In this study, the existence of bias in whole-genome phylogeny and how it can be avoided in the creation of homologous matrix is articulated.

3.3: Methods

Phylogenomics uses multiple genes to resolve phylogenetic relationships, in contrast with single gene phylogenies. In this work, a concatenated alignment from 522 alignments which have sequences from each species was used as a character matrix. Below is the description of the concatenation process, testing the robustness of the alignment and subsequently phylogeny inference

3.3.1: Concatenation

Concatenation means adding multiple genes, in this case, alignments together. There are two types of concatenation; vertical and horizontal. Vertical concatenation is easy; it adds files below one after another. In a computer, cat command can do this (cat file1 file2 file3 > all_file). Side by side concatenation is a little bit different. There is no direct command manual to do this. It can be done manually (Fig. 3.2.1).

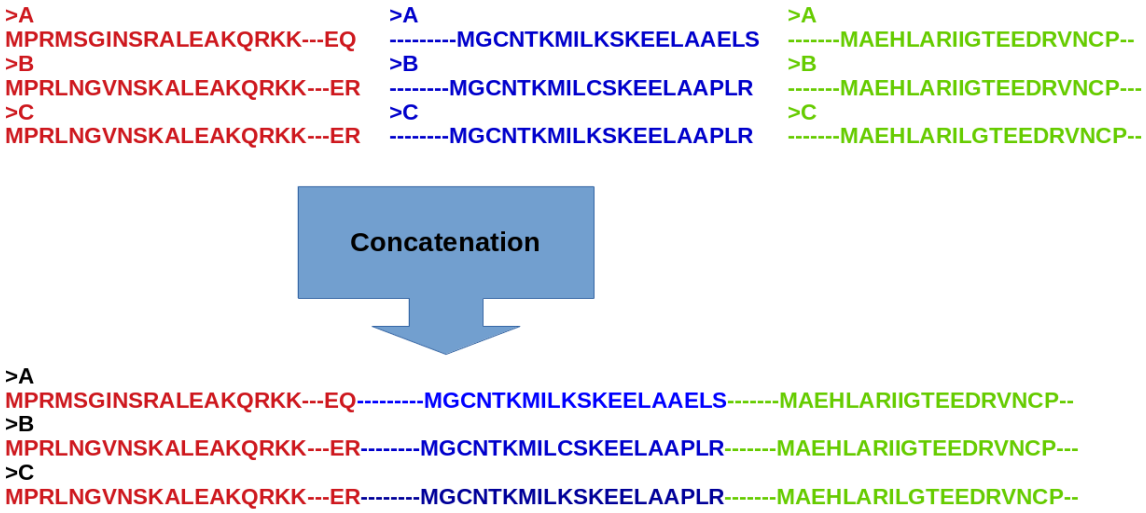

Figure 3.2.1: Side by side concatenation of multiple alignments. Three alignments were concatenated horizontally to make a single alignment.

In the Sequences directory (output of OrthoFinder), there is a sequence alignment fasta file for each orthogroup. The names of orthogroups which have genes from all organisms were extracted using a custom script, and then the individual fasta files were copied to remove duplicate sequences. The concatenating script was developed in collaboration with Dr. Kazutaka Katoh, Osaka University, Japan. After cleaning the header, the sequence was added side by side for each species. Poorly aligned positions and divergent regions were filtered using Gblocks (v0.91b) after trimming using the gapyout option of trimal (v1.4.15) [191,192].

3.3.2: Justifying the alignment

Before inferring phylogeny, we need to check the quality of the alignment. We can check it visually using different programs like Seaview and MegaX. Fewer undetermined

characters and missing sequence shows a better alignment. Amino acid frequency in the alignment also can determine the quality of the sequences.

As we used a core genome and filtered it carefully, a low variable amino acid frequency is expected. The amino acid frequency in the core genome was measured using MegaX. The amino acid frequency was also calculated for the whole proteome fasta files of 43 species. Two frequencies were imported to python to compare and visualize. The GC content of each genome was calculated from whole genome fasta files using a Bash script and added to the amino acid composition data to measure the correlations of GC content with both amino acid frequencies [193],[194]. ggplot2 was used to visualize the comparison of these two correlations [195].

3.3.3: Tree inference and testing

Iq-tree (v1.5.5) was used to find the best substitution model which enables free rate variation. RAxML was used to generate a maximum-likelihood-based phylogeny using the best model and 1000 bootstraps. A Bayesian phylogeny was generated using MrBayes (v3.2.6) in CIPRES science gateway [196],[197],[198],[199].

MrBayes uses MCMC (Markov Chain Monte Carlo) algorithm to generate tree samples from a list of probable trees. For each sample, it calculates the likelihood of the tree until convergence or reaches a stationary state. Tracer (v1.7.1) was used to test the convergence of MrBayes MCMC runs. MegaX was also used to generate neighbor-joining and maximum parsimony-based tree [200].

3.4: Results and Discussion

3.4.1: Core Alignment

In phylogenetics, the alignment is the most critical input for a true evolutionary relationship. The concatenated alignment was trimmed using different filtering criteria of Gblocks. The default trimming method did not produce factual phylogeny with reasonable branch lengths and branching patterns (Fig 3.4.1). That is why several trimming options were used to infer phylogeny (Table 3.2). All tree inference methodologies (Bayesian, Maximum likelihood) gave same topology for the default alignment.

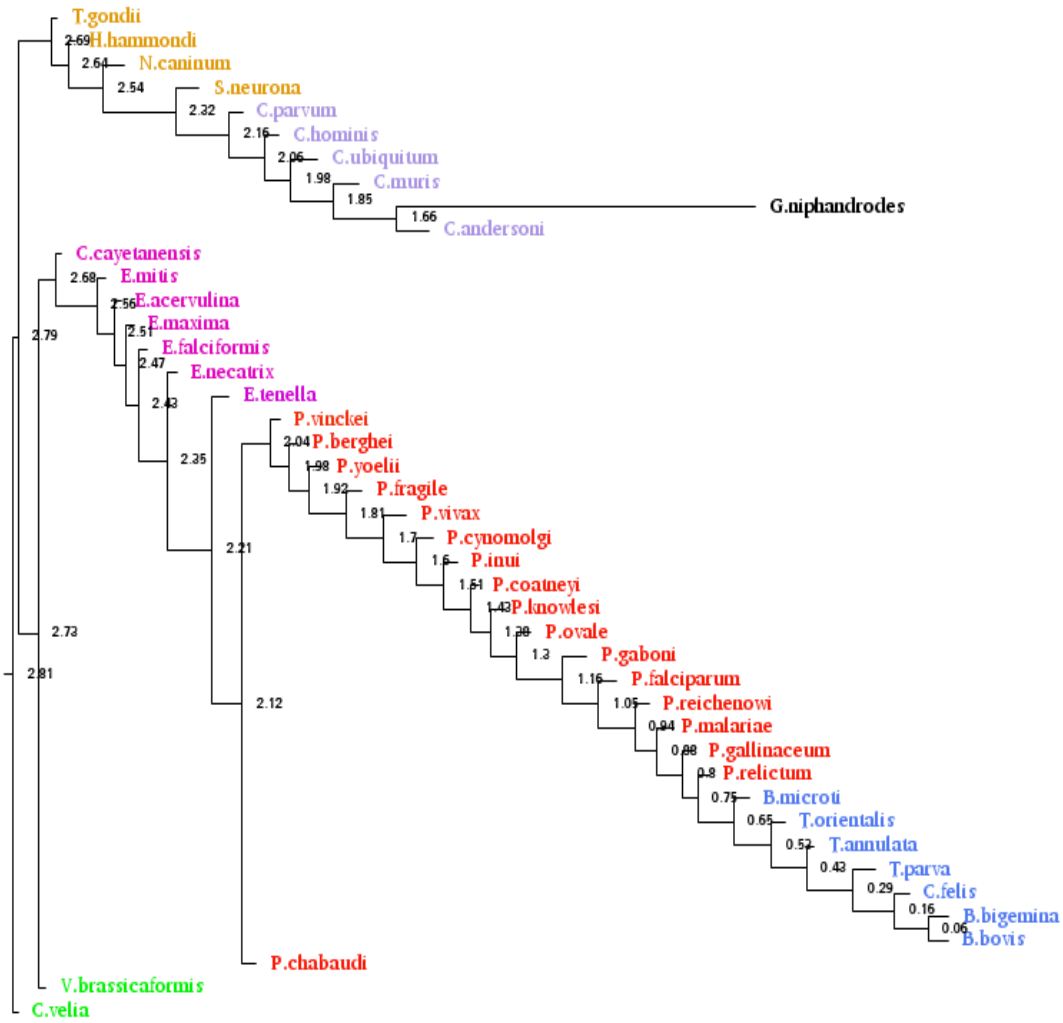

Figure 3.4.1: Bayesian phylogeny from the alignment that was trimmed using default filtering criteria. The numbers depict branch lengths.

Table 3.2: Sequence and topology for different filtering criteria.

No	Minimum number of sequences for a conserved position	Minimum number of sequences for a flanked position	Minimum number of contiguous nonconserved positions	Allowed gap position	Total number of sequence	Topology
1	22	36	0	All	1036	Bad
2	22	22	0	All	27078	Bad
3	36	43	0	All	1036	Bad
4	29	36	0	All	13218	Bad
5	29	36	0	None	3562	Bad
6	36	43	3	All	7582	Bad
7	36	43	2	All	5586	Bad
8	22	36	0	All	23822	Bad
9	36	43	2	All	6068	Good
10	36	43	0	All	1105	Bad
11	36	43	1	All	3415	Bad

3.4.2: Quality of alignment

In the final alignment, there were total 6068 selected sites including 5914 complete (no gaps, no X), 3383 variable (57.2% of complete) and 2620 informative (44.3% of complete)

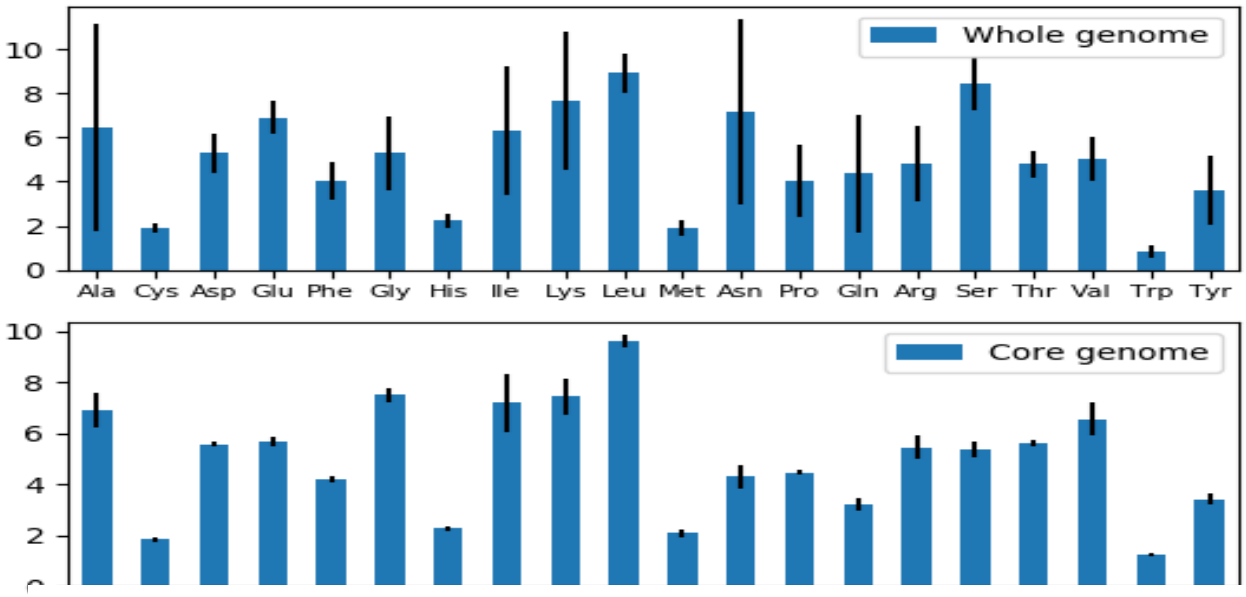

Figure 3.4.2.1: Mean and standard deviation of amino acid frequency in the core alignment and the whole genome. The bars indicate the mean values and the vertical black lines show sd.

amino acid sequence. Amino acid composition in the alignment are as follows: Group 1 : 34.0 %, E : 5.7 D : 5.6 Q : 3.2 N : 4.3 H : 2.3 R : 5.5 K : 7.4 Group 2 : 25.4% I : 7.2 L : 9.6 M : 2.1 V : 6.6 Group 3 : 29.8% A : 6.9 P : 4.5 S : 5.4 G : 7.5 T : 5.6 Group 4 : 7.6% F : 4.2 Y : 3.4 Group 5 : 3.1% W : 1.3 C : 1.9.

The amino acid frequency in the whole genome is highly variable compared to the core genome. From Figure 3.4.2.1, we can infer that the mean frequency of amino acids in the core genome and the whole genome are remarkably close, but the standard deviation of amino acid frequency in the whole genome is higher than core genome, in most cases. Many of these amino acid frequencies between these two groups are significantly different (Tab. 3.3 and 3.4).

Table 3.3: Independent ttest result of amino acid frequency in the core alignment and in the whole genome

	ttest_stat	Pvalue
Ala	0.59863941	0.5510243
Cys	-0.9245519	0.35784745
Asp	2.09422823	0.03925335
Glu	-10.644317	3.02E-17
Phe	1.4389828	0.15387132
Gly	8.4452414	7.76E-13
His	0.88253954	0.38000375
Ile	1.86453916	0.0657386
Lys	-0.4377743	0.66267353
Leu	4.77448543	7.54E-06
Met	2.95795694	0.00402122
Asn	-4.4813732	2.33E-05
Pro	1.75370492	0.08312866
Gln	-2.8082627	0.00619259
Arg	2.4159284	0.01786276
Ser	-16.112198	2.01E-27
Thr	8.75595297	1.84E-13
Val	8.5699205	4.35E-13
Trp	10.7582113	1.79E-17
Tyr	-0.8620088	0.39113688

Table 3.4: Descriptive statistics of amino acid frequency in the core alignment and whole genome

Core genome	count	mean	std	min	25%	50%	75%	Max
Ala	43	6.89	0.69	6.05	6.24	6.71	7.66	8.36
Cys	43	1.85	0.09	1.58	1.81	1.86	1.91	2.01
Asp	43	5.58	0.12	5.36	5.48	5.57	5.69	5.77
Glu	43	5.69	0.17	5.23	5.6	5.72	5.8	5.94
Phe	43	4.21	0.12	3.92	4.12	4.22	4.28	4.45
Gly	43	7.5	0.26	7.18	7.24	7.52	7.69	8.17
His	43	2.27	0.1	2.04	2.23	2.26	2.31	2.54
Ile	43	7.2	1.15	5.57	5.86	7.12	8.31	8.69
Lys	43	7.44	0.73	6.53	6.82	7.06	8.29	8.41
Leu	43	9.61	0.24	9.26	9.42	9.53	9.86	10.12
Met	43	2.08	0.17	1.9	1.93	2.03	2.18	2.47
Asn	43	4.3	0.45	3.6	3.89	4.19	4.76	4.93
Pro	43	4.45	0.09	4.33	4.38	4.46	4.52	4.7
Gln	43	3.22	0.25	2.9	2.96	3.24	3.44	3.63
Arg	43	5.46	0.48	4.83	4.93	5.57	5.91	6.15
Ser	43	5.38	0.33	4.85	5.18	5.31	5.48	6.2
Thr	43	5.63	0.12	5.36	5.55	5.64	5.71	5.92
Val	43	6.56	0.64	5.67	5.94	6.78	7.22	7.55
Trp	43	1.25	0.08	1.17	1.17	1.25	1.32	1.43
Tyr	43	3.42	0.23	3.03	3.2	3.48	3.64	3.73
Whole genome	count	mean	std	min	25%	50%	75%	Max
Ala	43	6.46	4.68	1.72	3.22	4.36	9.91	16.62
Cys	43	1.89	0.21	1.65	1.71	1.89	2.01	2.66
Asp	43	5.3	0.89	3.29	4.98	5.61	5.85	6.46
Glu	43	6.91	0.74	5.36	6.34	7.05	7.62	8.12
Phe	43	4.01	0.87	1.98	3.36	4.3	4.69	5.24
Gly	43	5.28	1.71	2.26	4.12	5.53	6.47	9.04
His	43	2.23	0.3	1.72	2	2.17	2.41	2.83
Ile	43	6.31	2.92	1.85	3.01	6.23	9.25	10.2
Lys	43	7.66	3.16	3.29	4.11	7.77	10.1	13.1
Leu	43	8.93	0.9	7.61	8.05	8.86	9.64	10.61
Met	43	1.9	0.36	1.05	1.7	1.97	2.18	2.62

Asn	43	7.19	4.2	2.05	2.56	7.71	8.84	14.76
Pro	43	4.02	1.63	1.66	2.73	3.61	5.72	6.78
Gln	43	4.38	2.67	2.17	2.84	3.47	3.99	14.29
Arg	43	4.8	1.71	2.48	3.9	4.48	5.85	8.26
Ser	43	8.47	1.21	6.38	7.7	8.1	9.41	11.08
Thr	43	4.81	0.6	3.62	4.35	4.75	5.2	6.01
Val	43	5.02	0.99	3.35	4.5	5.08	5.78	6.97
Trp	43	0.83	0.25	0.46	0.62	0.88	1	1.38
Tyr	43	3.63	1.57	1.29	1.69	4.08	4.97	5.91

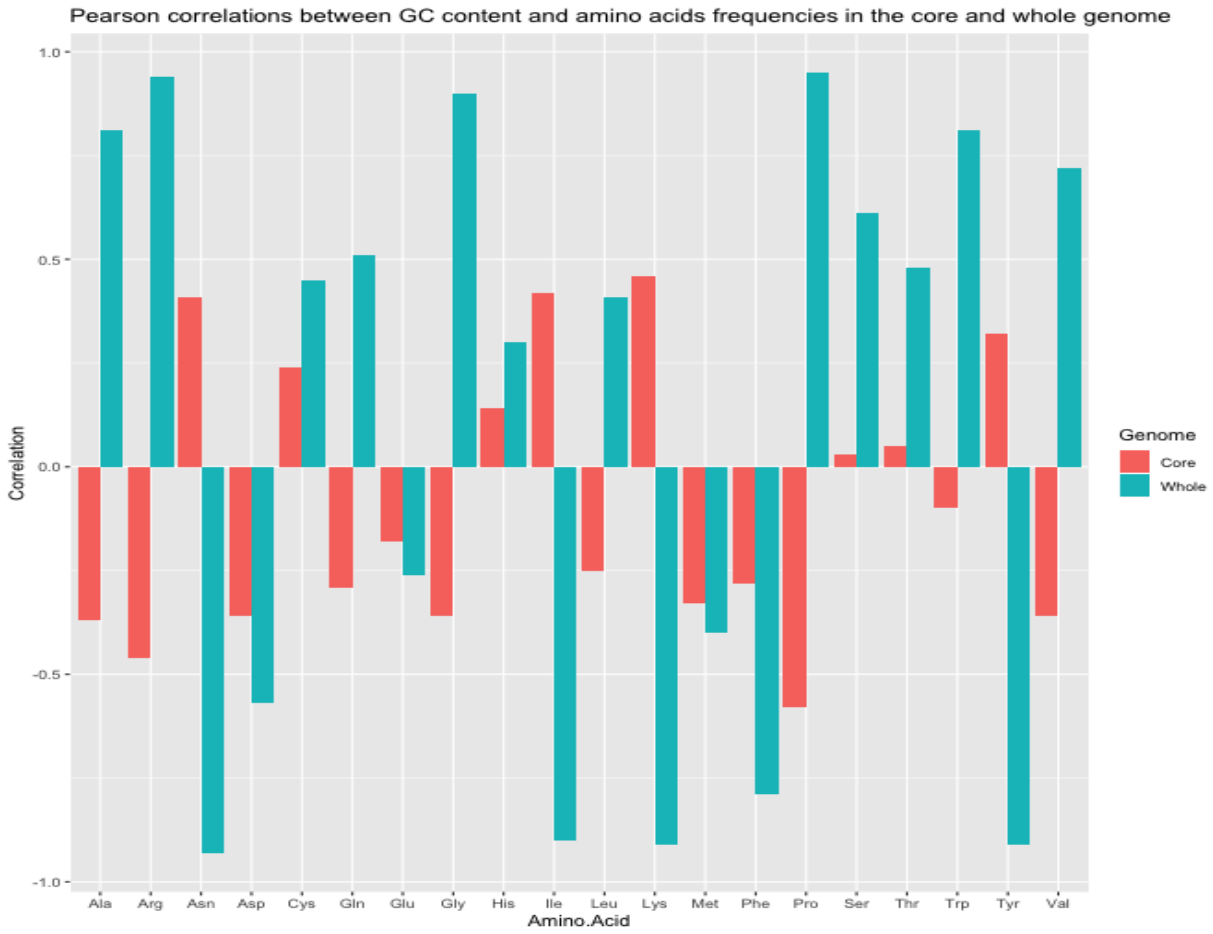

Figure 3.4.2.2: Correlations (R) of amino acid frequencies in the core and whole genome with GC content. In the core genome no amino acid is significantly correlated with GC content (-0.7<R<0.7), whereas in the whole genome at least 5 amino acids (Ala, Arg, Gly, Pro, Trp) are positively (R>0.7) and another 5 (Asn, Ile, Lys, Phe, Tyr) are negatively (R<-0.7) correlated.

Nearly half of the amino acids were found significantly correlated with GC content in the whole genome but not in the core genome (Figure 3.4.2.2). So, from Figure 3.4.2.1, 3.4.2.2 and Table 3.3 and 3.4, we can confer that, this alignment is highly conserved. Usually, conserved alignment is the best to infer true phylogeny [201]. Among 543 tested models, the best substitution model was LG+F+R4 (-LnL=89569.444, df=108, AIC=179354.889, AICc=179358.840, BIC=180079.653) [202],[203],[204] (Tab. 3.5 and 3.6).

Table 3.5: Model selection for phylogeny. Total 543 models were tested to find the best model. Considering the size of the table, it is placed in the accessible github page.

https://github.com/zillurbmb51/results/blob/master/concatenated.phy.log

Table 3.6: The nexus alignment file with parameter values. This describes all the parameters that were used to generate the Bayesian species tree.

https://github.com/zillurbmb51/results/blob/master/infile.nex

3.4.3: Tree Inference

More sequences do not necessarily ensure the real evolutionary relationships in the phylogeny. The phylogeny that is shown in Figure 3.4.1 is statistically robust (bootstrap support or probability value) but did not produce the right evolutionary relationship. The tree as mentioned above was generated using an alignment which consists of 14,216 amino acids sequences. After using one of the most stringent trimming processes (Table 3.2), the final alignment (6068 AA) produced a real bifurcated tree which depicted the genuine evolutionary relationship among observed species (Fig. 3.4.3.1).

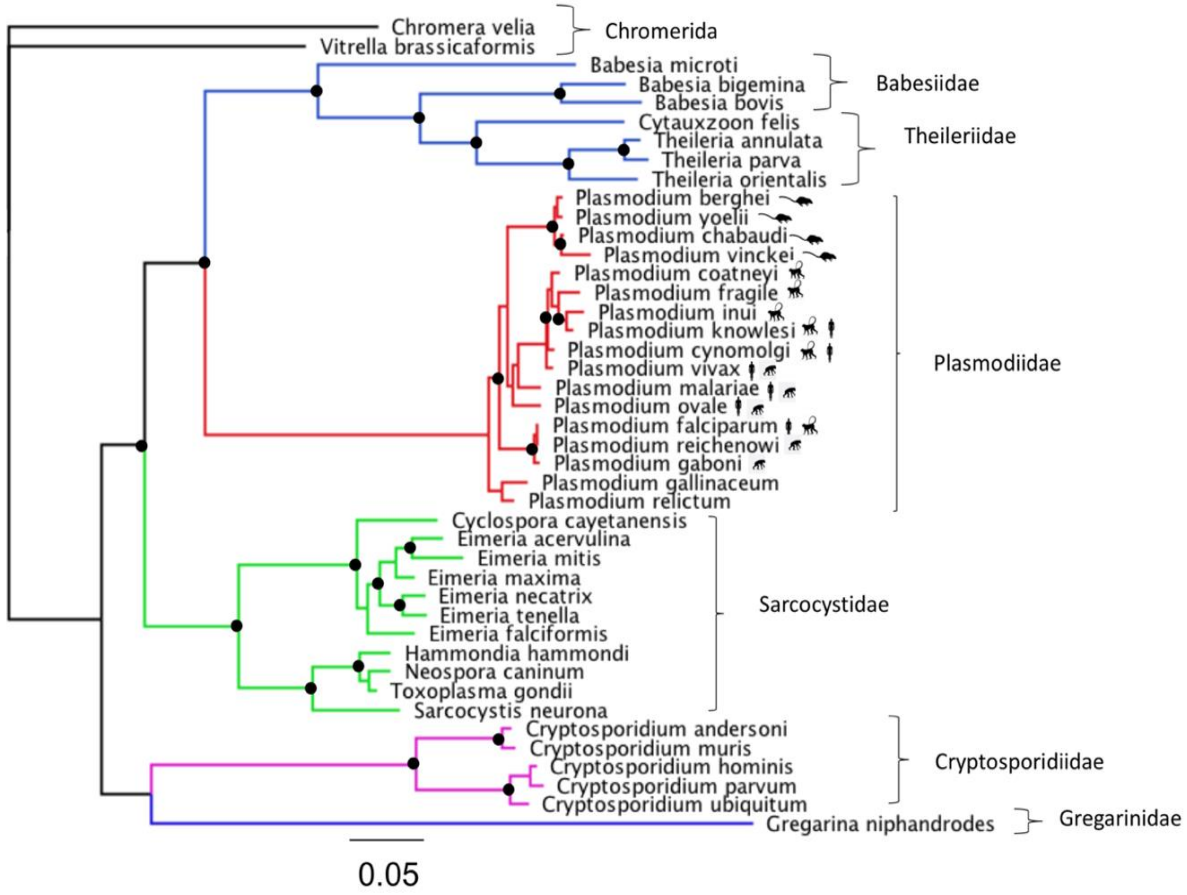

Figure 3.4.3.1: Phylogenomic analysis of 41 Apicomplexan species. The phylogenetic inference of 41 Apicomplexan species and 2 Chromerid species was generated from the core genome using maximum likelihood and Bayesian approaches. Branch lengths are in units of substitutions per site, derived from the Bayesian analysis. The black circles indicate that consensus support from MrBayes is 1 and RAxML bootstrap support is 100%. Here, = Rodent, = Monkey, =Human and =Chimpanzee. The tree was generated from a core alignment (522 genes with 6068 sequences after filtering) using best model (LG) in MrBayes (3.2.6) for 300000 MCMC runs with 8 chains for two runs.

This topology is not 100 % congruent with the topology of species tree inference from all genes (STAG; OrthoFinder species tree). The entanglement shows a subtle difference between these two trees. Here, only topological difference difference is considered not evolutionary rates (Fig. 3.4.3.2). The evolutionary rates aka branch lengths are also vague in STAG (Fig. 3.4.3.3). At least one monophyletic clade is totally different between these two trees. This clade consists of *T.gondii, H. hammondi* and *N. canium.*

From Figure 3.4.1, it can be inferred that, statistically robust tree does not always represent true evolutionary relationships. A statistically well-supported tree should give us genuine biological relationship. To understand the reason for getting the wrong topology, we need to know the source of errors. From Figure 3.4.1 and Table 3.2, we can infer that alignment trimming is one of the main factors that affect tree topology. Another factor that influences tree inference is the convergence of bootstrapping samples (in Maximum likelihood approach) or MCMC runs (in Bayesian method).

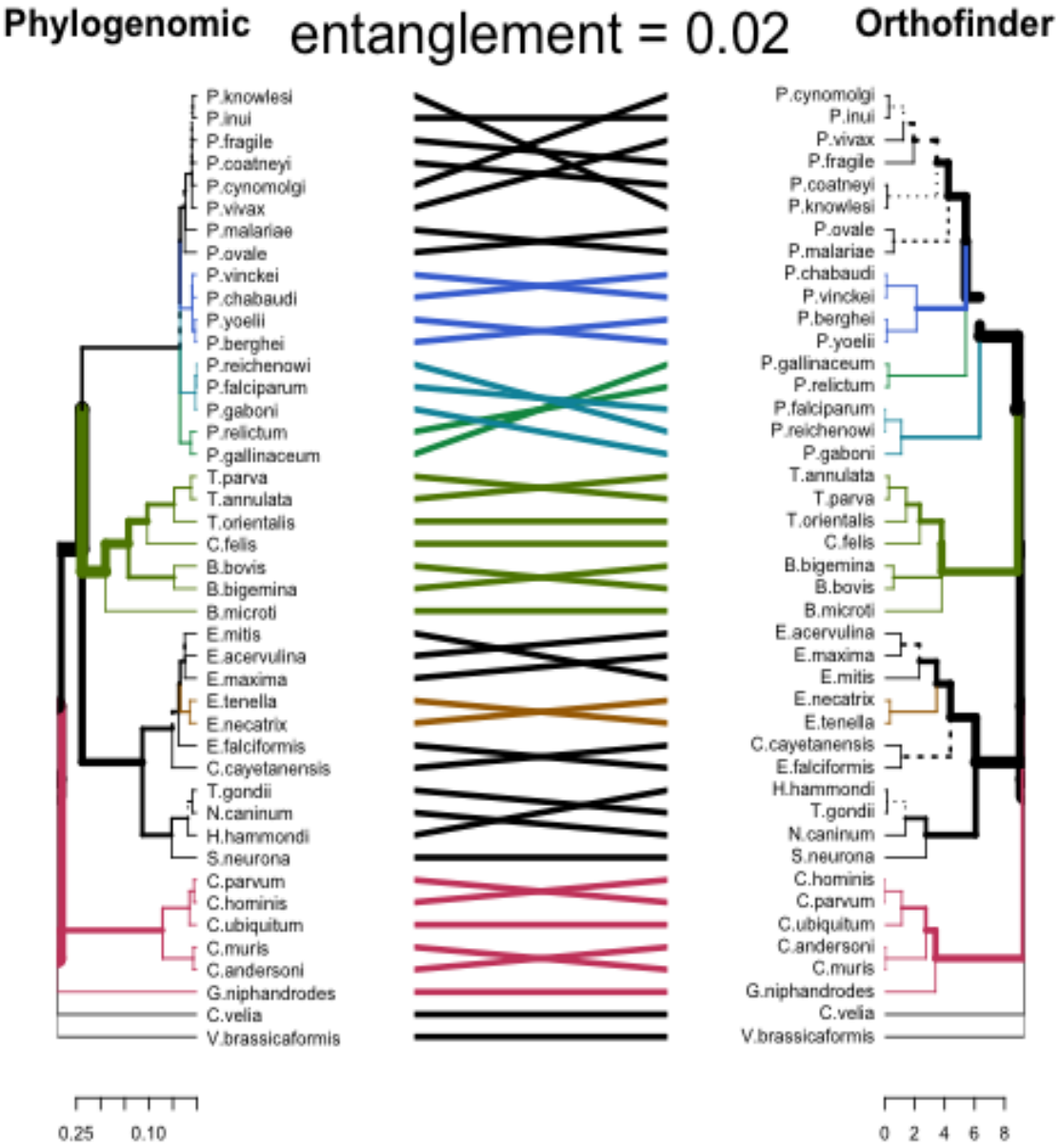

Figure 3.4.3.2: Congruence analysis between phylogenomic and OrthoFinder tree topologies. The dot lines indicate incongruence between two trees.

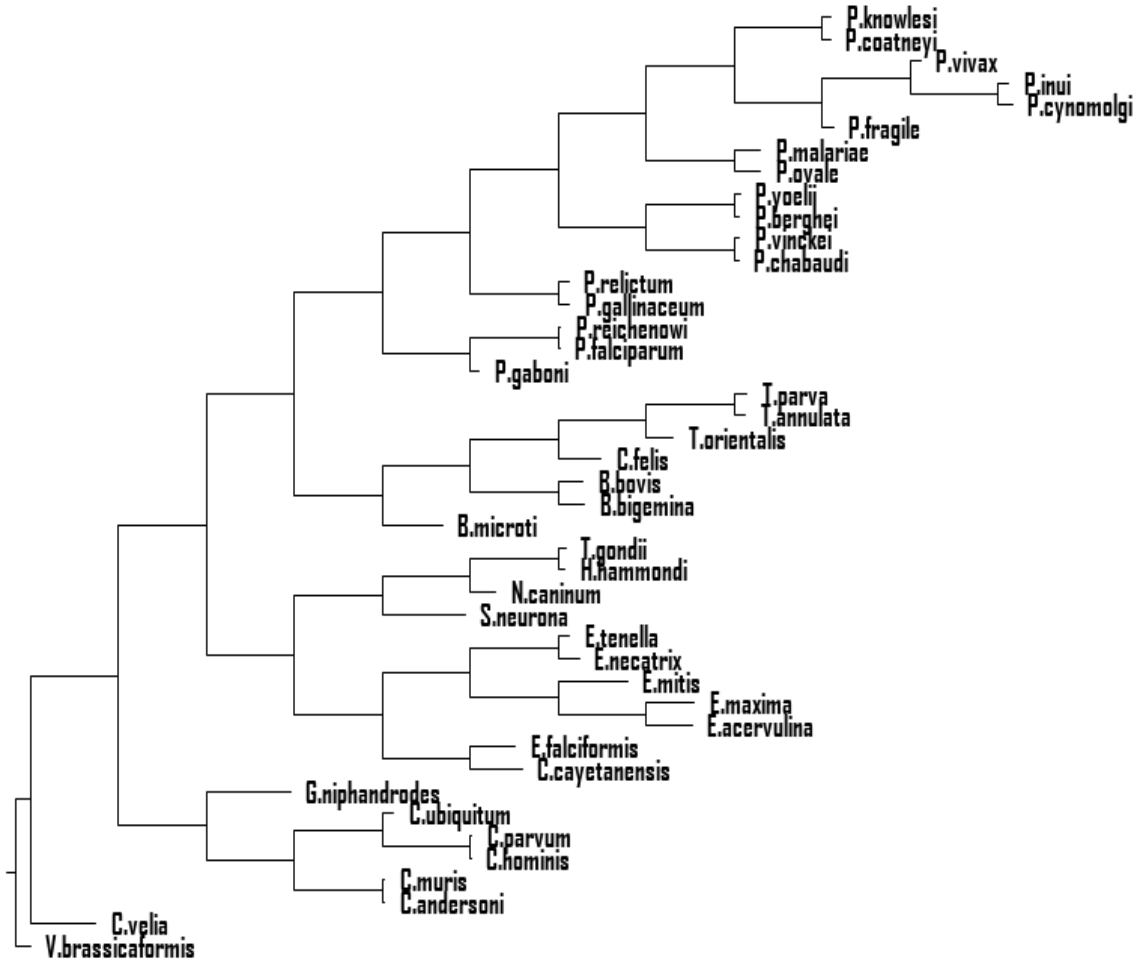

Figure 3.4.3.3: Species tree from all genes (OrthoFinder species tree). This tree is inferred using neighbor joining method.

3.4.4: Convergence test of MCMC run

In each MCMC run, MrBayes estimates the posterior distribution of model parameters namely, LnL, LnPr, TL, alpha along with a tree. To test the convergence, we can plot the sample number vs. these parameter values. Initially, an unstable curve will reach in a stationary phase as MCMC run reaches convergence. After completion of MCMC runs, we can calculate mean, variance, stdev, median, range and effective sample size for each parameter. Effective sample size should be more than 100 to be convergent. We also can

run two separate MCMC run and compare the parameter values. When an MCMC run reaches to stationary phase, there will be similar values in both runs.

The convergence also depends on the number of generations, the number of total samples and the number of total chains. An MCMC runs with 20000 generations is more likely to be convergent compare to an MCMC run with 2000 generations. If we increase the number of chains, the MCMC run will reach in the stationary phase at a faster rate. Another critical factor is the number of total samples and sample frequency. If we use a lower sample frequency, we will get more sample which is very much essential to reach stationary phase.

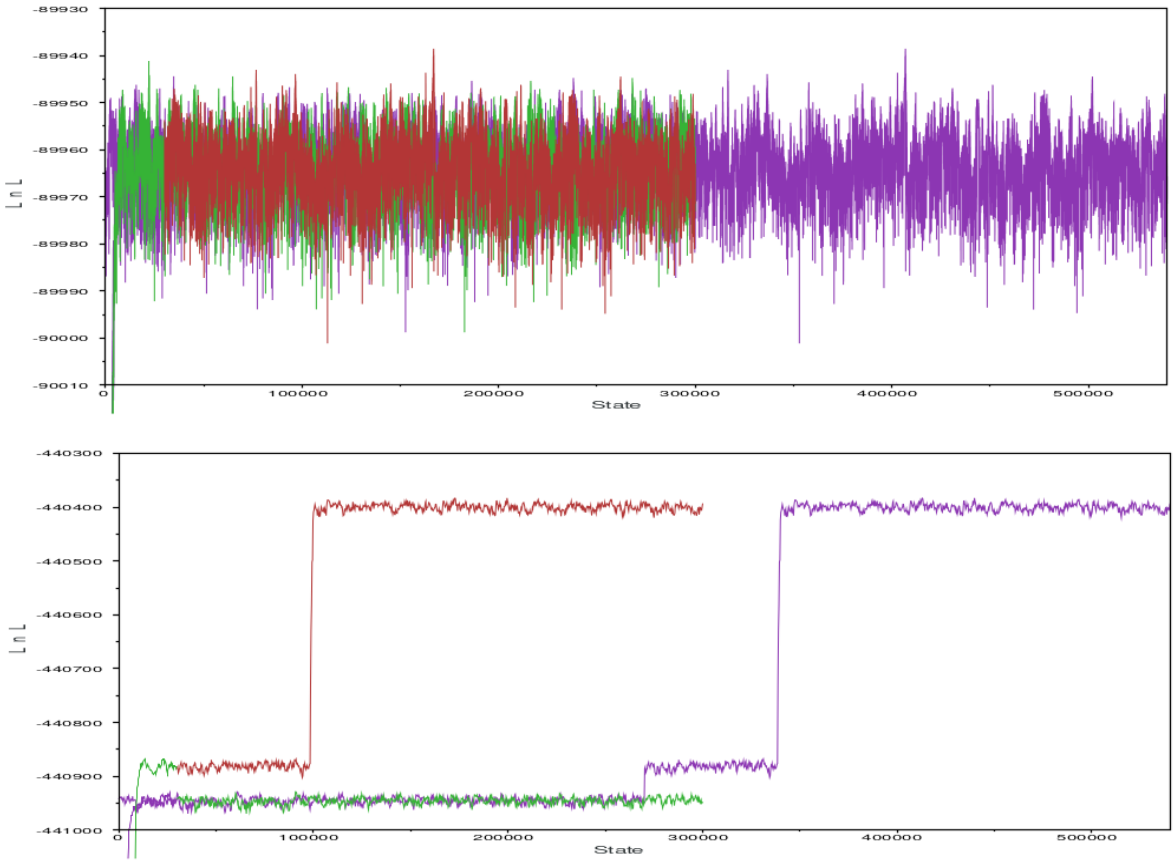

Figure 3.4.4.1: Trace values comparison between a convergent (top) and a nonconvergent (bottom) MCMC run for logLikelihood. The top one reached to stationary phase, and the bottom one did not. In both cases, MCMC runs were carried out for 300,000 generations with two chains and higher sample frequency.

The estimates of parameter values can also help us to find the convergence of MCMC runs. If an MCMC run reaches to the stationary phase, we can observe the trace of similar

parameter values for run1, run2 and combined (as Figure 3.4.4.1; top). For a nonconvergent MCMC run, we will get significantly different parameter values for different runs (as Figure 3.4.4.1 bottom).

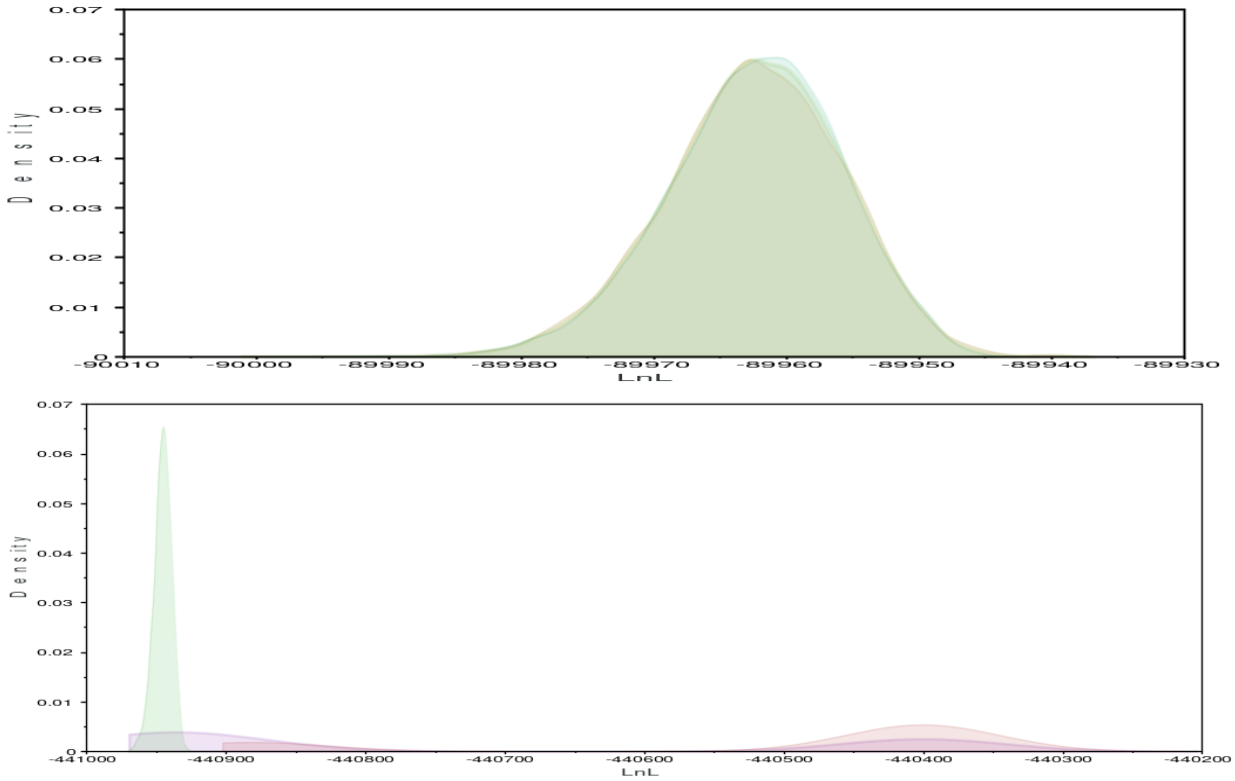

Figure 3.4.4.2: Probability density comparison between a convergent (top) and a nonconvergent (bottom) MCMC run for logLikelihood. The top one reached to stationary phase, and the bottom one did not. In both cases, MCMC runs were carried out for 300,000 generations

In the Table 3.7, the descriptive statistics of all parameters for the final MCMC run is presented. Probability density plots of logLieklihood of a convergent MCMC run will be similar for different runs (Figure 3.4.4.2; top) and will be different for a nonconvergent run (as Figure 3.4.4.2; bottom). The mean with variance for logLikelihood with both runs was

plotted in Figure 3.4.4.3 including convergent (bottom) and nonconvergent (top) MCMC runs to show the difference of parameter values between two types of runs.

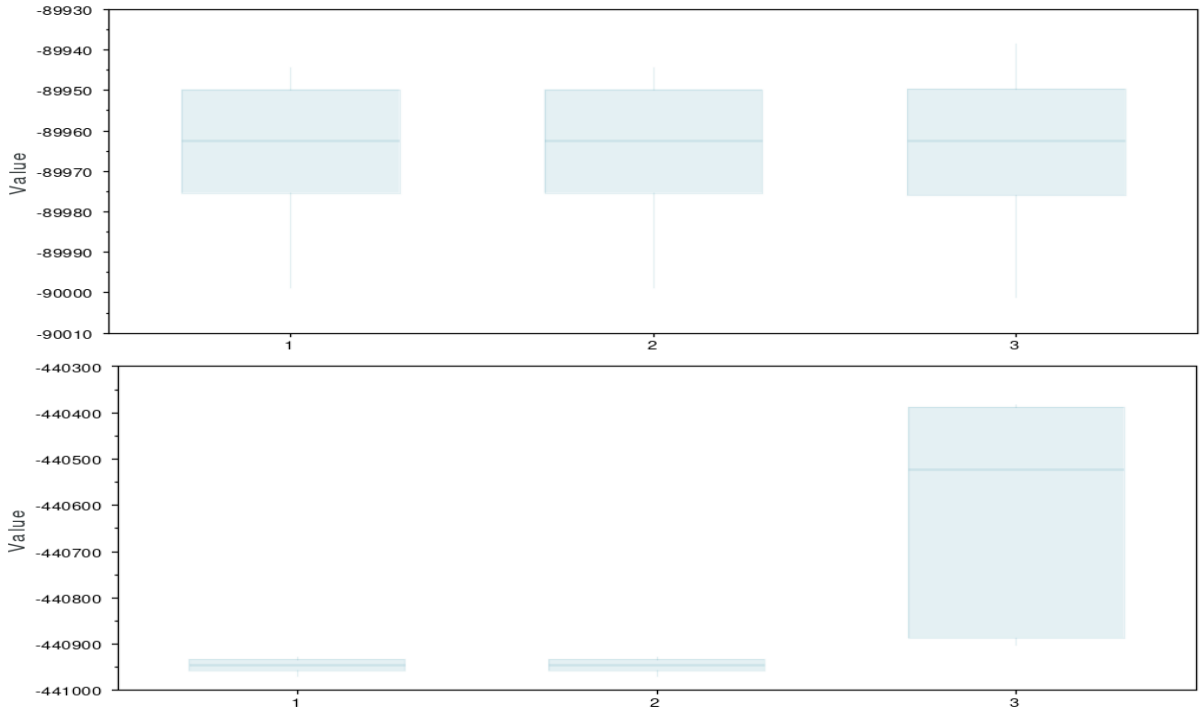

Figure 3.4.4.3: Estimates of logLikelihood values comparison between a convergent (top) and a nonconvergent (bottom) MCMC run. 1 and 2 indicate respective runs and 3 indicate the combined run. The bottom one reached to stationary phase and the top one did not. In both cases, MCMC runs were carried out for 300,000 generations

So, from Figure 3.4.4.1-3 and Table 3.7, it can be inferred that, the MCMC runs that were used to infer Bayesian phylogeny reached to a stationary state.

Table 3.7: Summary statistics of parameter values of MCMC run

Summary Statistic	LnL	LnPr	TL	alpha
mean	-89962.463	45.216	3.1531	0.4663
stderr of mean	0.255	0.0244	9.3417E-4	2.1637E-04
stdev	6.6709	1.514	0.058	0.0106
variance	44.5015	2.2922	3.3599E-3	1.1289E-04
median	-89962.19	45.2136	3.1526	0.4656
value range	-90001.11, -89938.56	39.0098, 51.3523	2.9257, 3.3986	0.4287, 0.5109
geometric mean	n/a	45.1907	3.1525	0.4662
95% HPD interval	-89975.41, -89949.49	42.1711, 48.1184	3.0379, 3.2657	0.4452, 0.4868
auto-correlation time	789.1518	140.2708	140.2626	223.9537
effective sample size	684.3	3849.8	3850.1	2411.3
number of samples	54002	54002	54002	54002

3.4.5: Comparing the species with other methods

The Bayesian tree and Maximum likelihood trees gave us the same topology. The node support values are also similar in these two trees. For example, where we got low posterior probability support in the Bayesian tree, maximum likelihood's bootstrap support was also low in the same position (Fig. 3.4.3.1).

The neighbor-joining method gave us a good tree with clear bifurcation, but the node support values are not as robust as Maximum likelihood or Bayesian methods (Fig. 3.4.5.1). Though it is the quickest way to reconstruct phylogeny from large scale molecular data [161].

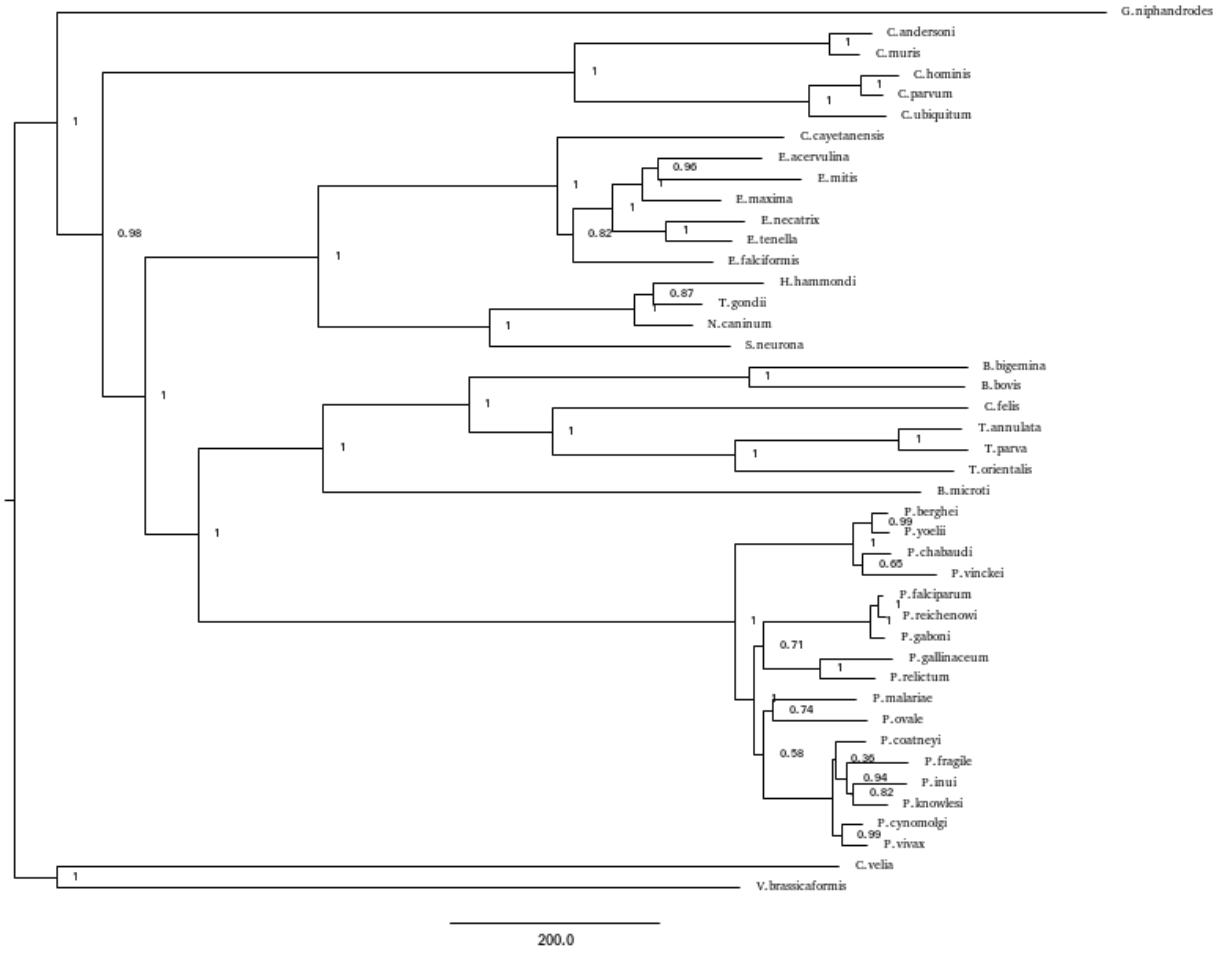

Figure 3.4.5.1: Species tree using neighbor joining method with the same alignment as input.

The minimum evolution method also produced a relevant topology (Fig. 3.4.5.2). In minimum evolution, the smallest sum of branch length estimates is most likely to be the true one. This method first introduced mathematical proof that the sum of branch length estimates for the true tree is the smallest among all possible trees [162]. The UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method also produced supportive topology as other methods (Fig. 3.4.5.3). This method also uses molecular distance to reconstruct phylogeny as neighbor-joining and minimum evolution [205].

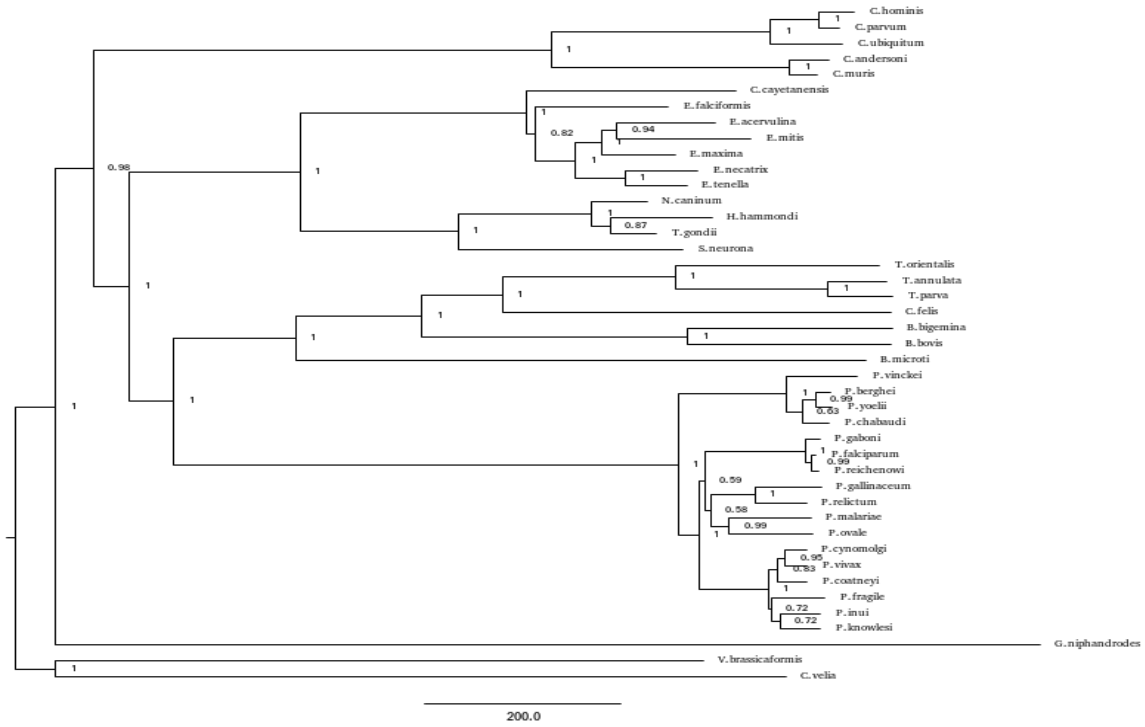

Figure 3.4.5.2: Minimum evolution-based tree with node label as bootstrap support values

The maximum parsimony method generated the worst tree among all the methods for this alignment. The clades were grouped as expected, but the evolutionary rates and history is ambiguous. The branch lengths seem to be the same for all the clades which cannot be true (Fig. 3.4.5.4). Maximum parsimony chooses the tree that minimizes the number of steps required to generate the observed variation in the sequences. In some cases (fewer taxa and character states, which did not happen in this case), this method may produce misleading topology [206],[176] .

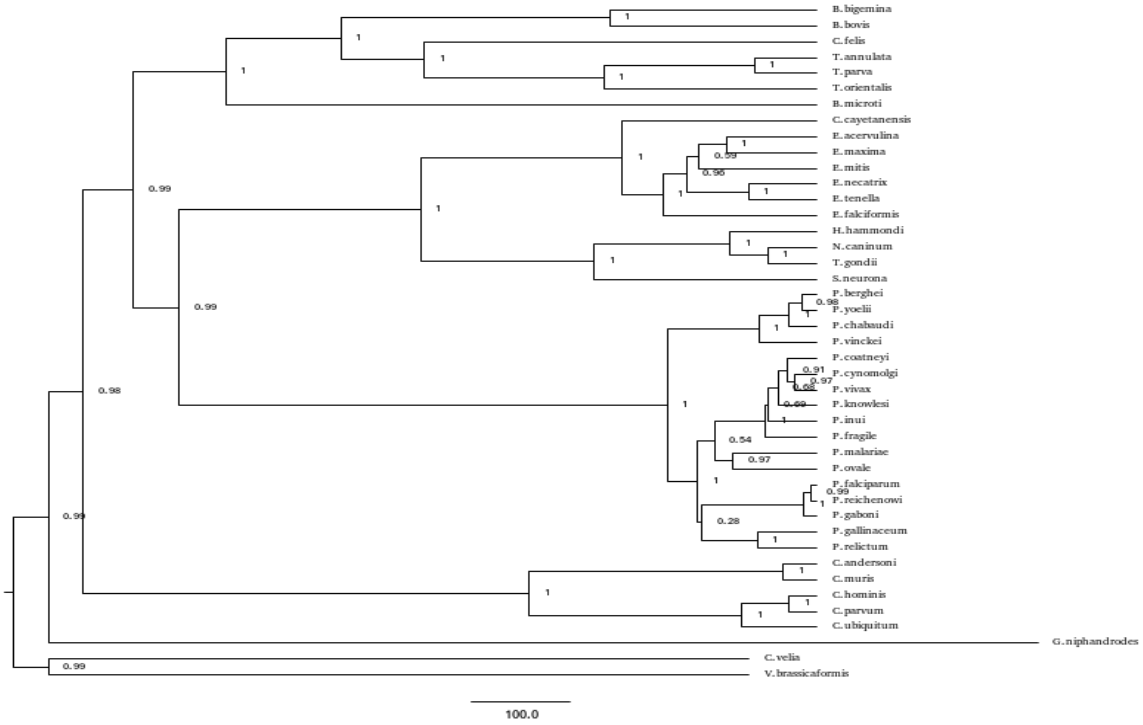

Figure 3.4.5.3: UPGMA based tree with node label as bootstrap support values

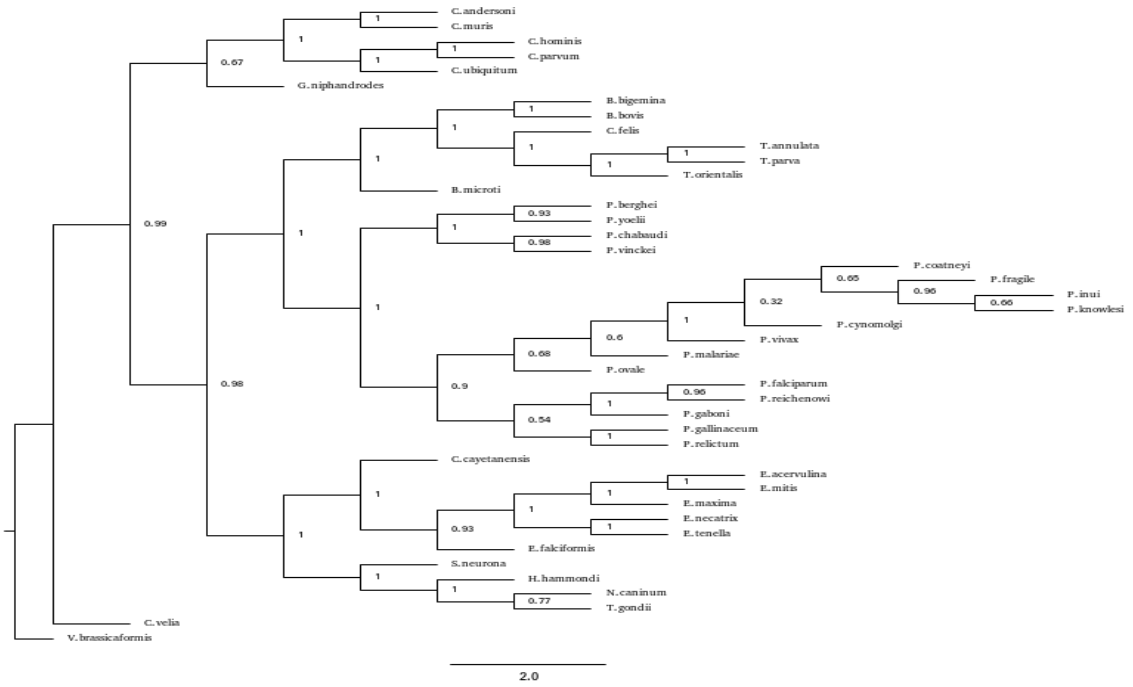

Figure 3.4.5.4: Maximum parsimony based tree. Nodes were labeled with the bootstrap support values

All but the maximum parsimony tree indicates that *G. niphandrodes* is the earliest taxon and still evolving at a faster rate than any other clade among the observed group of species.

3.4.6: Comparing the species with other sources

In the final tree (Fig. 3.4.3.1), the topology of the families *Babesiidae, Theileriidae, Plasmodiidae, Sarcocystidae* and *Cryptosporidiidae* generated in the phylogenomic analysis is consistent with that reported by [183]. The branching pattern of the piroplasmid taxa (*T. parva, B. bigemina* and *C. felis*), representing the difficulty to resolve genera *Theileria, Babesia* and *Cytauxzoon*, is consistent with [207], which used mitochondrial genome sequences. Taxa belonging to the *Cryptosporidiidae* show the same branching pattern as in [208]. The position of *T. gondii, N. caninum, H. hammondi* and *S. neurona* is consistent with the phylogeny reported by [209], while *C. cayetanensis* is basal to the *Eimeria* species, consistent with [210]. The branching pattern of the *Eimeria* species is consistent with [211]. Lastly, we found the unclassified *Gregarina* is most closely related to *Cryptosporidium* among all other Apicomplexa, consistent with [184].

The clade *Plasmodium* is largely congruent with a recent study by [190], which consists of a phylogeny of more than 30 *Plasmodium* species, constructed using 21 nuclear genes. The *Plasmodium* species from the phylogenomic analysis are congruent with theirs, with the exception the monophyletic clade that includes *P. knowlesi, P. inui, P. fragile, P. coatneyi, P. cynomolgi* and *P. vivax*, which is largely discordant with ours. Predominantly, four species of *Plasmodium* causes malaria in humans in our analysis: *P. falciparum, P. malariae, P. ovale* and *P. vivax* [212] (while human infections of *P. knowlesi* [212] and *P. cynomolgi* [213] have also been reported, indicated on Fig. 3.4.3.1). According to the

phylogenomic analysis, the four main human infecting species do not form a single clade but have evolved primate infectivity and pathogenicity independently (Fig. 3.4.3.1).

Rodent parasites *P. berghei, P. yoelii, P. chabaudi* and *P. vinckei* [214] created a monophyletic clade. *P. reichenowi, P. vivax. P. gaboni* and *P. ovale* [103] infect chimps and do not cluster together. *P. falciparum* was found distantly related to other human parasites, and clustered with chimpanzee parasites *P. gaboni* and *P. reichenowi*. *Plasmodium* species that infects monkeys include *P. knowlesi* [212], *P. inui* [215], *P. falciparum* [216], *P. malariae* [217], *P. coatneyi* [218], *P. cynomolgi* [219] and *P. fragile* [104]. These cluster together, with the exception of *P. falciparum*.

Lastly, it is important to remember that phylogenomic analysis is still subject to pitfalls [220]. This study shows that amino acid composition of the proteome is influenced by underlying GC bias (Fig. 3.4.2.2). Potentially, this could represent a conflating factor, although outside the scope of the present study.

Chapter 4

Effective Population Size Inference

4.1 Abstract

The effective population size (N_e) is the number of reproducing individuals among the total population. N_e acts as a deterministic factor in conserving the rate of evolutionary change due to genetic drift. Neutral evolution or drift is favored by low effective population size and selection is favored by larger effective population size. It was hypothesized that *Plasmodium falciparum* population suffered a recent population bottleneck. Alternative observations (larger N_e and N_e expansion) were also reported.

This study analyzed 35k variants across 14 chromosomes in the whole genome from an African (Mozambique) sample to calculate the effective population size changes in *P. falciparum*. Multiple Sequentially Markovian Coalescent (MSMC) method shows that *P. falciparum* has undergone serial population bottleneck. During the period ~26,000-5,000 years ago N_e reduced from ~$5.98*10^7$ *to* ~$4.5*10^3$. This observation also shows that there is a possibility of recent expansion of N_e of *P. falciparum*.

4.2 Introduction

N_e has direct effect on the evolutionary change [221]. In a Mendelian population (a group of inbreeding individuals who share all the available genetic information), when N_e is small drift is more effective whereas, for a large N_e, the selection is dominating [222]. This theory was tested by simulation using Populus (v6.0) [223]. In the simulation, a lower population size shows that randomly half of the allele fixation in fewer generations (Figure 4.1.1). This means that, in a smaller N_e, drift will exert larger deterministic effect on allelic fixation compared to natural selection.

In the stallion population, Calder (1927) found a very lower N_e compared to the actual population size [224]. From then, it was believed that N_e is always lower than the actual population size. In many cases, it may be right. However, in *African rhinoceros* populations, researchers found that N_e is as large as the actual population [225]. Using simulations of a population of annual plants and the temporal method, scientists showed that N_e could be same as actual population [226]. So, N_e can be smaller or the same as the actual census in different types of species.

In figure 4.1.1, one can observe that N_e has a significant contribution in determining the rate of evolutionary change. Now the question is, what is the effect of N_e in evolution? Is solely N_e determine the rate of evolutionary change or there are other factors? Using DNA

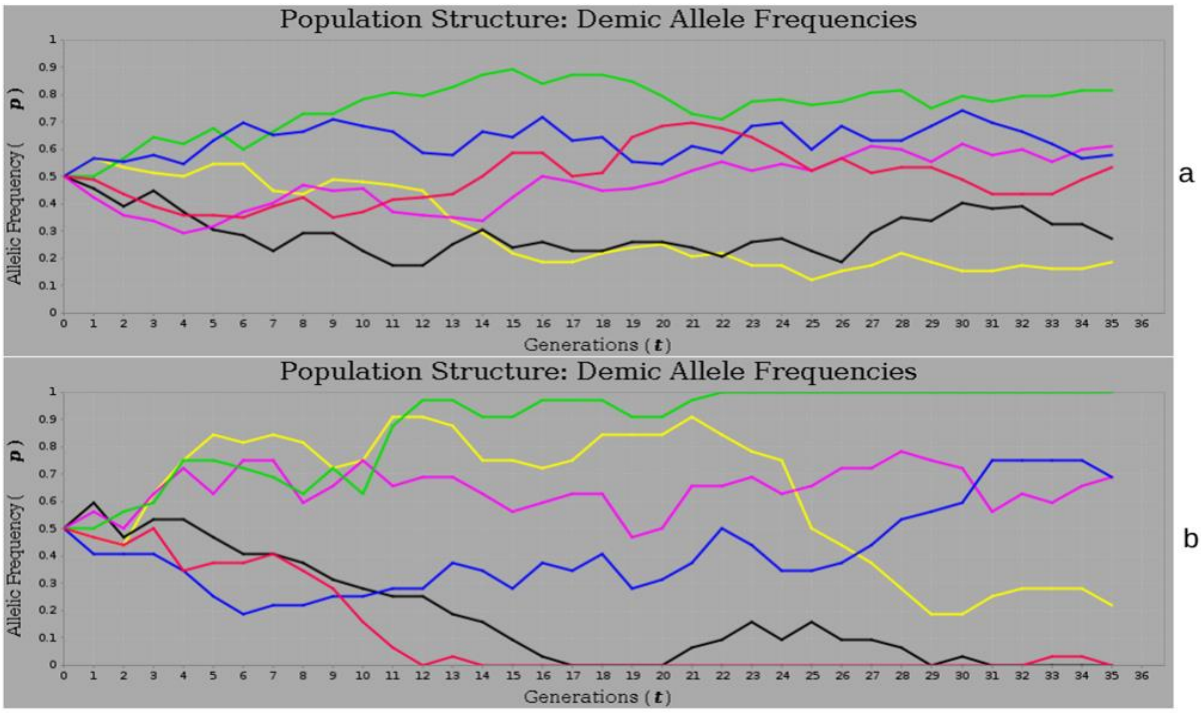

Figure 4.1.1: SImulation for the effect of effective population size in a mendelian population. The simulation was performed using Populus (v6.0). In (a) Ne=46 and in (b), Ne=16. In both cases migration rate =0.01, iteration=35, initial allele frequency=0.5 and number of alleles=6

sequences from seventy-five independent *Mesoplasma florum,* M. Lynch and colleagues showed that N_e determines 84% of mutations [227]. In 1998, John W. Drake showed that mutation rates are consistently higher in prokaryotes compared to eukaryotes [228]. This observation raised a question that, is there any effect of genome size on the mutation rate? Because, usually prokaryotes have smaller genome compared to eukaryotes. Eukaryotes have introns, which are not translated into proteins. Thus, if genome size affect mutation rate, coding regions will exert more selective pressure than the whole genome.

Then S. Massey proposed the Proteomic constraint theory, in which he elucidated that proteome size acts as a significant deterministic factor in the change of evolutionary aka mutation rates. Using several eubacterial, archaeal and DNA viral genome, Massey

showed that proteome size accounts for 94% of the variation in mutation rates [40],[229]. Later, in a letter to PNAS, Lynch showed that in eukaryotes, proteome size could explain ~41% of the variation in mutation rates [230]. The correlation between genome size and mutation rates are different in prokaryotes and eukaryotes [231]. In both studies, genome size is negatively correlated with mutation rates in prokaryotes.

The effective population size inference is important in these organisms because effective population size has a direct contribution to the genetic polymorphisms. A high level of polymorphisms was observed in microbial pathogen [232]. This observation says that pathogenicity might be correlated with genetic polymorphisms. Rich SM and colleagues (1998) found a lack of synonymous substitution in worldwide geographic strains of *P. falciparum.* From this observation, they hypothesized that this parasite is recently emerged from a single ancestor and spread all over the world through a demographic sweep which may account for its virulence (Malaria's Eve hypothesis) [233]. Later, extremely low synonymous nucleotide polymorphism in *P. falciparum* and radiation out of Africa was also observed in mitochondrial sequences from a different endemic area [234]. SNPs in introns of this parasite was found significantly lower compared to coding regions, which supports a recent origin of pathogenic *P. falciparum* [235].

The opposite scenario of Malaria's Eve hypothesis is a long time persistent, effective population size. Analyzing nucleotide substitution at 23 nuclear protein-coding loci, Hughes and Verra (2001) showed a large effective population size of *P. falciparum* (~100,000) for the past 300,000-400,000 years [236]. Variation in 100 worldwide mitochondrial DNA sequences suggests an early origin and recent expansion of this parasite population [237].

Synonymous allele-frequency spectrum analysis of 25 whole genomes also revealed a recent expansion of the population of this parasite [238]. Mutations in different regions of the genome also indicate a larger effective population size [239]. So, there are two types of evidence and hypothesis about the effective population size of *P. falciparum,* long term stable population and a recent bottleneck. These two hypotheses are summarized in Figure 4.1.2 from simulated data.

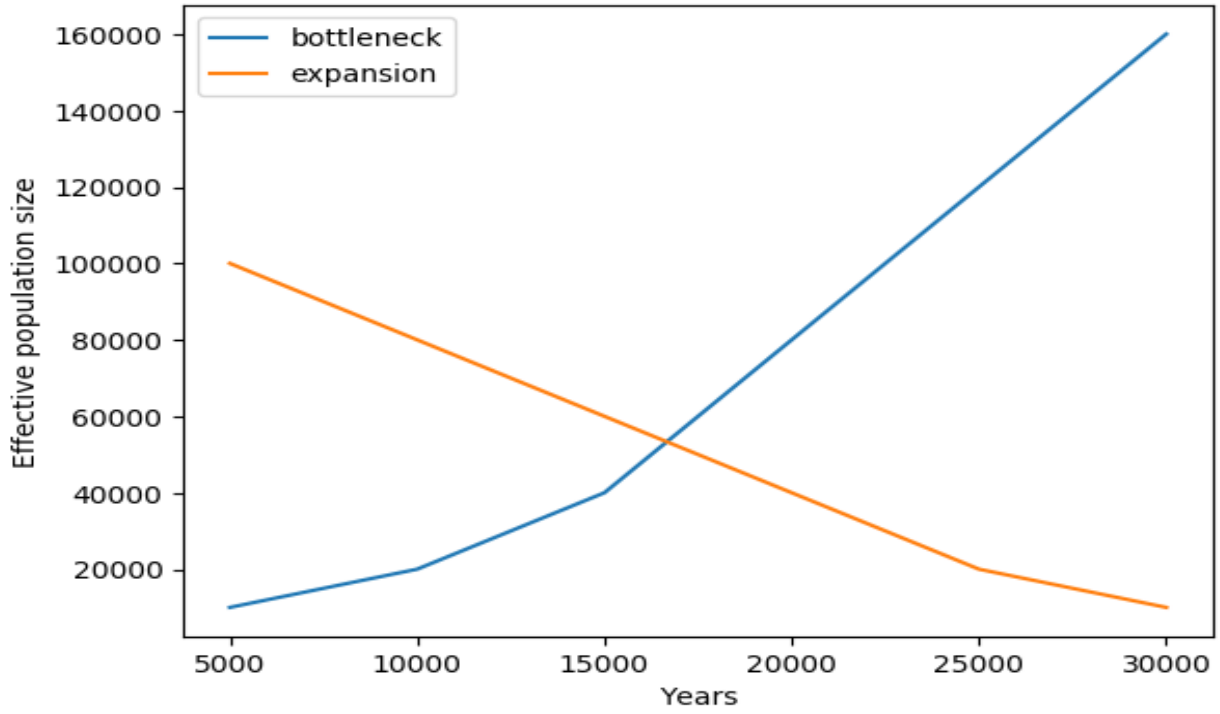

Figure 4.1.2: Recent expansion and bottleneck of population. This figure was created in matplotlib from simulated data inspired from the evidence of bottleneck and expansion of *P. falciparum* population in the past few thousand years.

As the breeding population has a direct contribution to genetic polymorphisms, a clear understanding of population structure will help us to identify variation in the genome which should be a critical strategic parameter in search of vaccine and drug targets. A strategy to control a highly polymorphic pathogen should be more complicated compared to a pathogen with a low level of genetic polymorphism.

There are several ways to calculate N_e for a given organism. In general, we can calculate the N_e by counting the number of adults who are directly contributing to the offspring. It is reasonable that this approach is not feasible for many types of species. N_e can be

calculated from genome data using different genetic markers. Measuring effective population size relies on the genetic markers (SNPs, microsatellites). Principally, there are four ways to infer effective population size namely, a) Current N_e estimated from heterozygote excess, (b) Short- or long-term N_e estimated from linkage disequilibrium, (c) Short-term N_e estimated from temporal samples and (d) Long-term N_e estimated from current genetic variation [240].

As mutation, selection and migration or the stochastic force of genetic drift affect the population size, a short-term N_e inference is more realistic (c). If we exclude all the forces, drift is solely responsible for changes in allele frequency. Markov Chain Monte Carlo (MCMC) evaluation is the best way to approximate the posterior distribution of N_e, especially for smaller genome [241],[242], [243].
In this thesis, Short-term N_e is estimated from whole genome SNPs data, using the Multiple Sequentially Markovian Coalescent (MSMC) method.

4.3 Methods

The Multiple Sequentially Markovian Coalescent (MSMC) method uses chromosome-wise biallelic SNPs to infer population size history from whole genome sequencing data [244]. One sample was selected for the final analysis from 12 whole genome sequence data. The samples were collected from uncomplicated malaria in Mozambican children [245]. The Study Accession is PRJNA315887, Sample accession: SAMN04573341 and Run accession: SRR3305687. Total of 12 samples from the same study was analyzed to get the best quality data (Table 4.1). The analysis is divided into three steps namely Mapping, SNP calling and MSMC2 inference.

4.3.1 Alignment

Paired end fastq files were downloaded from the European Nucleotide Archive (ENA). The quality of the reads was measured using FastQC [246]. Sequences were filtered and

reoriented using reformat.sh of BBTool before mapping with Novoalign (v2.08.02) against the *P. falciparum 3D7* reference genome obtained from 33rd release of EuPathDB [247],[22], [1], *(*Unpublished. Colin Hercus, 2008 Novocraft Technologies, www.novocraft.com).

4.3.2 Filtering and SNP calling

Next-generation sequencing data may contain errors as well as mapping output. The alignments were sorted and indexed using SAMtools (v1.7). Duplicate reads were removed. SNPs were called using samtools (mpileup), filtered (quality, ≥20; read depth, ≥5). The data set with the highest number of SNPs and reasonable Transition/Transversion (Ts/Tv) value was used for further analysis [248],[249], [250],[251].

4.3.3 Population size inference

MSMC2 suite has some scripts to process the SNP data for the input of the inference method. The generate_multihetsep.py script from msmc-tools was used to process variants with the following command: ./generate_multihetsep.py chr1.vcf.gz > chr1.msmc.input.txt (for all chromosomes). The output of this command is the input for MSMC2. The command was as follows: ./msmc2 -i 55 -p 1*2+15*1+1*2 -o test.msmc chr*.msmc.input.txt. (* indicates all chromosome input files). The parameter value of ‘i’ (Expectation-Maximization algorithm) was chosen from a range of values for which the Likelihoods reached to a stationary state. This produces a data frame which contains different parameter values from which we can calculate population size with a given mutation rate and generation time. An average mutation rate of 4.6E-09 was used for the analysis (generated from values of 3.8E-10 [252], 5.4E-09 [253], 6.8E-09 [238] and 5.9E-09 [237]). A generation time of two months was used approximating days in different life stages [82]. SNP calling method and MSMC inference was optimized for the smaller genome of *P. falciparum* according to [245] and [244].

MSMC2 gives us a data frame of 4 columns namely, time_index, left_time_boundary, right_time_boundary and lambda. Here, time_index means the simple index of time segments, left and right time boundary indicate the scaled start and end time for each time interval and lambda is the scaled coalescence rate estimate in that interval. From these parameters, one can calculate the population size in the past years. The equation to calculate population size is as follows:

Effective population size=(1/lambda)/(2*mu) ….. (Eq. 4.1)

And for timeline,

Years = left_time_boundary/mu*gen ……. (Eq. 4.2)

Here, mu=mutation rate and gen=generation time

4.4 Results and Discussion

NGS data are large and error prone. The sequencing error affects the alignment quality which causes misleading SNP calling and downstream analysis. In this study, data quality was measured in each stage and filtered for population size inference.

Table 4.1: Short summary of the data quality

Sample	Number of reads	Number of snps	Ts/Tv	Mean Read depth
SRR3305682	13,286,224	29037	0.89	73.99
SRR3305683	14,513,266	19973	0.92	64.61
SRR3305684	21,826,292	22611	0.90	98.01
SRR3305685	27,921,148	29987	0.89	183.74
SRR3305686	13,170,529	10909	0.87	64.35
SRR3305696	6,213,087	12933	0.86	64.35
SRR3305687	**26,319,513**	**35290**	**0.87**	**164.65**
SRR3305691	24,747,739	34608	0.89	181.52
SRR3305693	35,650,132	34425	0.89	288.96
SRR3305695	20,821,888	32248	0.86	98.52
SRR3305697	7,735,269	13250	0.80	40.46
SRR3305698	9,188,513	20000	0.90	36.94

4.4.1 Quality of the data

Among 26005830 sequences, 25963075 were passed in QC (99.835%) (Fig. 4.4.1.1-2). 90.71% of the reads were mapped to the reference genome with average mapping quality 52.21 and mean read depth 164.654 (Fig. 4.4.1.3). Q values of most of the reads are above 30 and read lengths are 100. This data produced total 35,290 SNPs with Ts/Tv 0.87 which is supported by [245] (Tab. 4.1).

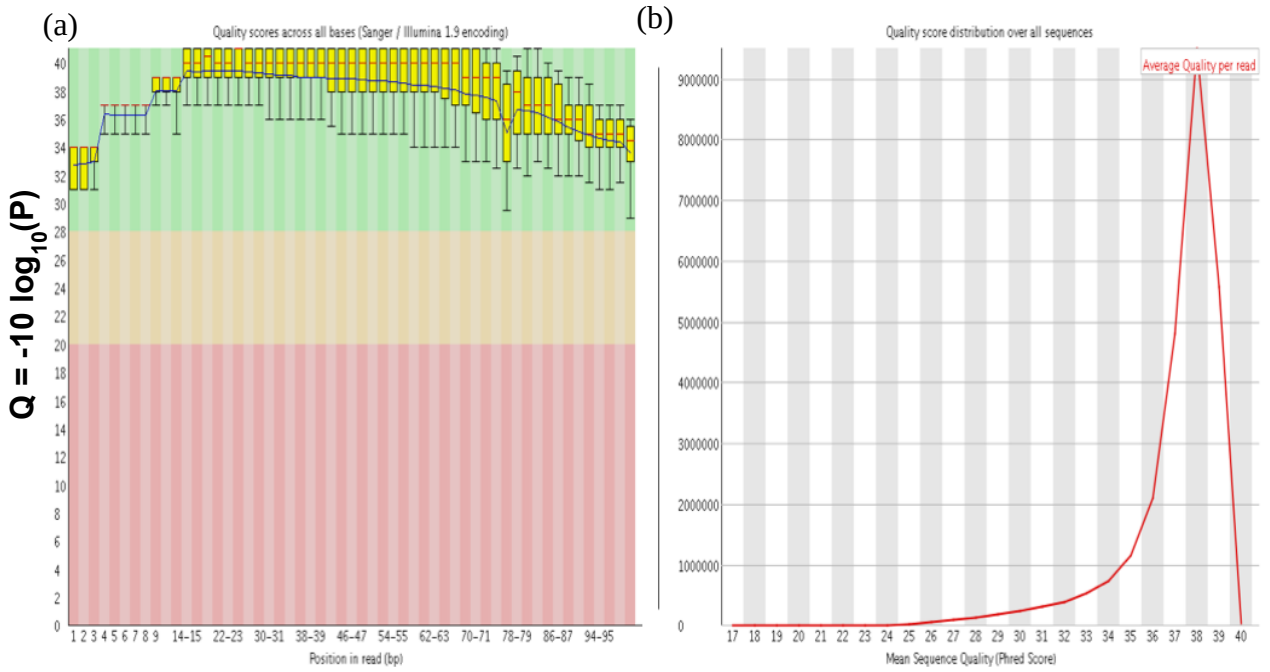

Figure 4.4.1.1: Quality score (Q value) and its distribution. (a) Q value of each base and (b) distribution of Q value in all sequences.

The quality of raw reads (fastq files) can be measured by Q value and the distribution of nucleotides. The fastq files contain the nucleotide sequence and corresponding quality score of each base. The quality score known as Phred or Q values. It is an integer value standing for the estimated probability of an error, i.e., that the base is incorrect. A larger Q score shows a lower probability of error, hence, a better-quality data. In this case, all bases had minimum Q=30 (Q=0 means P=1, Q=10 means P=0.1, Q=20 means P=0.01 and Q=30 mean P=0.001, here P=probability of error). According to Q-score, there are

minimum errors in the fastq file (Fig. 4.4.1.1). A steep normal distribution of Q score (after 30) says a lesser error in the data.

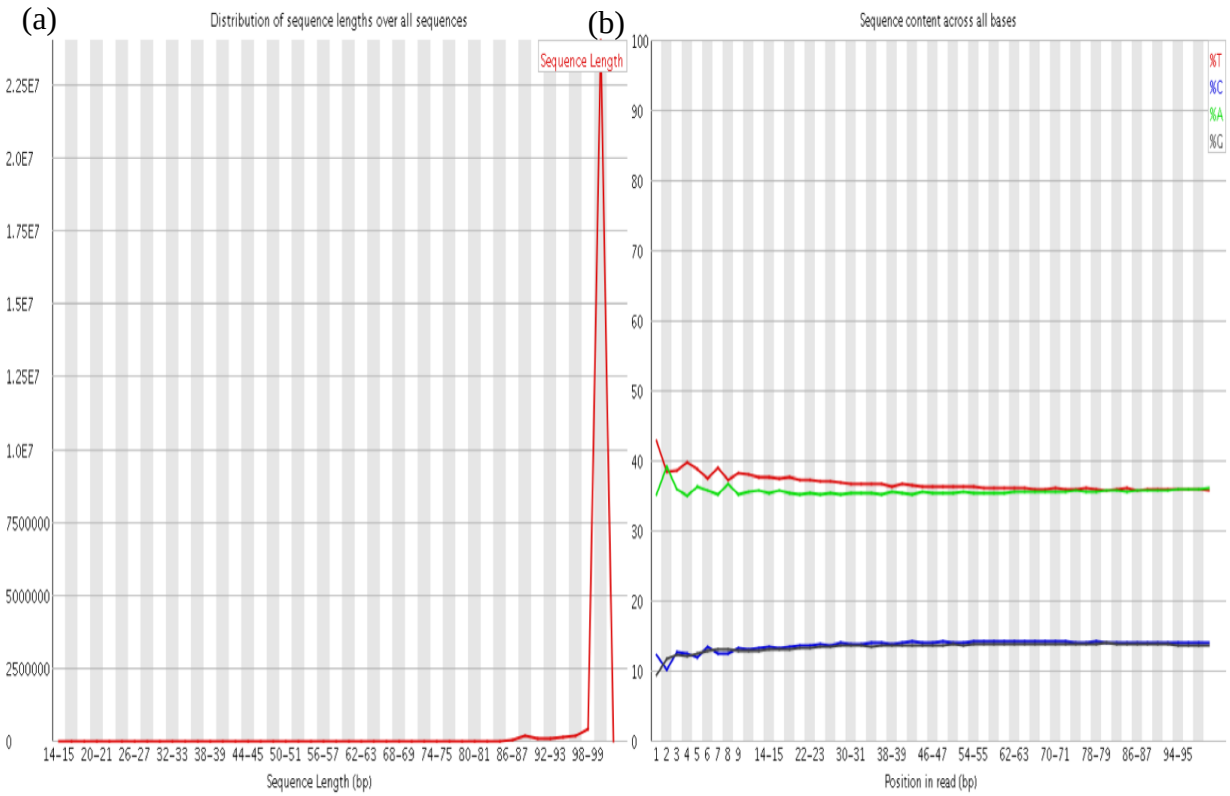

Figure 4.4.1.2: Distribution of sequence length and nucleotides. (a) Sequence lengths and its frequency, (b) percentage of nucleotides.

Per base sequence content is another important quality criterion. From a good fastq file, we can expect a uniform distribution of nucleotide in each position. We have found the number of A is almost equal to T as well as G equal to C (Fig. 4.4.1.2).

Read depth and mapping quality are important quality control parameters after mapping the reads to the genome. The depth can vary from region to region aka chromosome to chromosome. The lowest and the highest mean depth was observed in chromosome 7 (179.508) and chromosome 13 (293.812) respectively. The lowest number of SNPs were found in chromosome 5 (1166), and the highest number of SNPs were observed in chromosome 10 (3862) (Fig. 4.4.1.3).

According to the structure change there are two types of SNPs namely transition (Ts) and transversion (Tv). If a pyrimidine base (Cytosine; C and Thymine; T) is replaced by

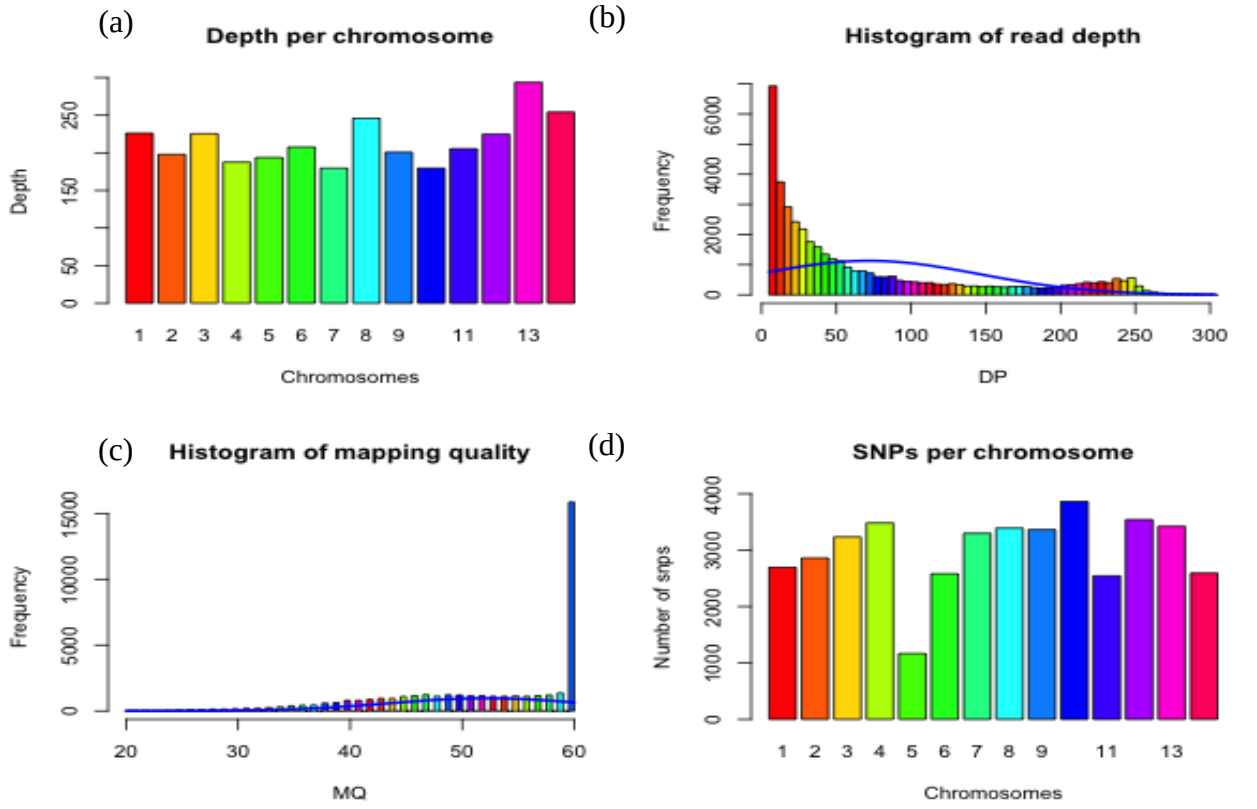

Figure 4.4.1.3: Quality of data after mapping. (a) Read depth per chromosome (b) Histogram with frequency of read depth (c) Histogram with frequency of mapping quality and (d) Number of SNPs in each chromosome.

another pyrimidine or a purine base (Adenine; A and Guanine; G) is substituted by another purine, then the SNP is called a transition. If a purine base is substituted by a pyrimidine or vice-versa, then the SNP is called a transversion (Fig. 4.4.1.4).

Another critical quality measurement is the ratio between Ts and Tv. Mathematically, for random sequence change the Ts/Tv ratio should be 0.5 (four possibilities for Ts, A to G and T to C or vice versa, whereas for Tv, there are eight possibilities, A to T, A to C, G to C and G to T or vice-versa (Fig. 4.4.1.4)). At the last step of our SNP calling pipeline, we will get all the information, statistics and Ts/Tv value, using samtools mpileup.

According to the structures of the bases, chemically and physically, it is easy to incorporate a purine instead of another purine or a pyrimidine in place of another pyrimidine; Ts (one amine group), whereas, incorporation of a purine instead of pyrimidine or vice-versa (Tv), is more complex (a carbon-nitrogen containing ring).

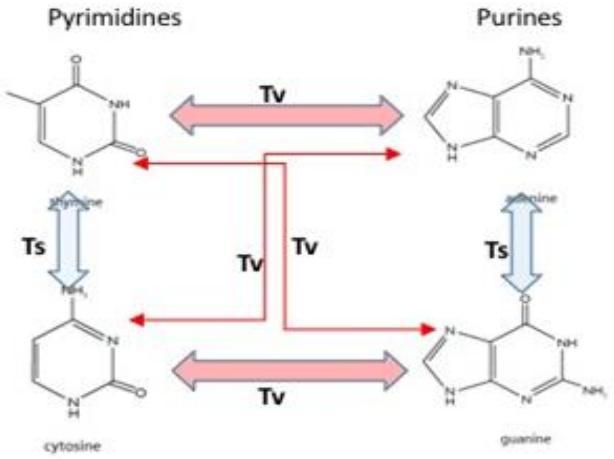

Figure 4.4.1.4: Transition and transversion processes

Random sequencing errors are expected to produce a Ts/Tv ratio of 0.5. So, the closer a Ts/Tv value is to 0.5 the more sequencing errors it may possess. For example, the overall Ts/Tv in the human genome is 2.1, and so values < 2.1 indicate significant levels of sequencing errors. Exonic regions have values > 2.1, due to the constraint exerted by the structure of the genetic code, which restricts the occurrence of transversions, which are more commonly associated with nonsynonymous changes [254],[255]. This means that in organisms with reduced genome sizes, and hence a higher proportion of coding DNA, the Ts/Tv ratio would be expected to be higher than the value in humans, 2.1. However, the genome wide Ts/Tv value in *P. falciparum* is 0.87 for this sample. The value is less than the expected because of its AT-rich genome but still acceptable according to [256] and [257] (Fig. 4.4.1.5). This value is also supported by data from different projects (Tab. 4.1-2). The rationale of being Ts/Tv ~2, can be clarified from Figure 4.4.1.4.

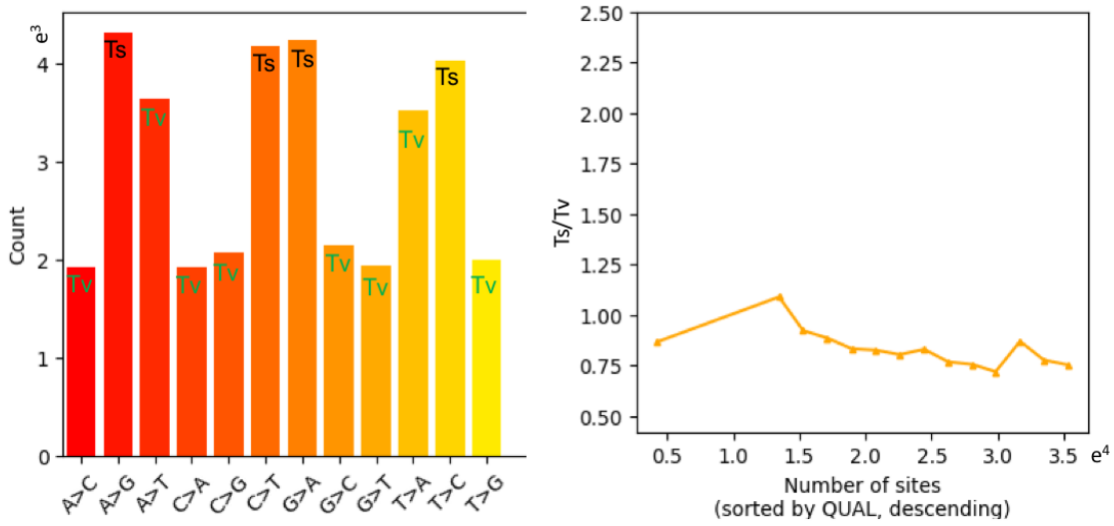

Figure 4.4.1.5: Number of SNPS and distribution of Ts/Tv. (a) Total number of each type of SNPS and (b) Distribution of Ts/Tv values

Table 4.2: Quality measurements of the reads from multiple samples from different projects.

No.	Accession No.	Instrument model	Library layout	GC content (in fastq)	GC content (in fasta)	Q value	Average read depth	Total SNPs	Ts/Tv
1	SRR034510	Illumina Genome Analyzer II	Paired End	29	19	31	188	3605	0.26
2	SRR081563	Illumina Genome Analyzer II	Paired End	40	19	22	182	1406	0.61
3	SRR1027715	Illumina HiSeq 2000	Paired End	22	19	38	109	20805	0.99
4	SRR1146611	Illumina HiSeq 2000	Paired End	22	19	38	102	23579	0.94
5	SRR1605090	Illumina HiSeq 2000	Paired End	22	19	39	41	28608	0.93
6	SRR3109233	Illumina HiSeq 2500	Paired End	25	19	38	81	6874	0.57
7	SRR349757	Illumina Genome Analyzer II	Paired End	32	19	15	45	1924	0.16
8	SRR3995443	Illumina HiSeq 2000	Paired End	42	19	37	6.5	12592	1.06
9	SRR3995833	Illumina HiSeq 2000	Paired End	43	19	10	1.4	0	0

10	SRR410819	Illumina HiSeq 2000	Paired End	35	19	34	22	51306	0.88
11	SRR520471	Illumina Genome Analyzer II	Paired End	19	19	39	48	542	0.31
12	SRR1335983	Illumina HiSeq 2000	Paired End	21	19	38	552	96946	0.79
13	SRR1619236	Illumina HiSeq 2500	Paired End	21	19	38	102	81076	0.83
14	SRR629016	Illumina HiSeq 2000	Paired End	25	19	38	313	86153	0.81
15	SRR609056	Illumina HiSeq 2000	Paired End	26	19	38	42	64788	0.84
16	SRR628956	Illumina HiSeq 2000	Paired End	21	19	38	29	59404	0.86
17	SRR646206	Illumina HiSeq 2000	Paired End	21	19	39	157	81327	0.81
18	SRR646211	Illumina HiSeq 2000	Paired End	21	19	38	141	80341	0.82
19	SRR650868	Illumina HiSeq 2000	Paired End	21	19	38	27	60715	0.85
20	SRR650892	Illumina HiSeq 2000	Paired End	21	19	38	21	57208	0.86
21	SRR767797	Illumina Genome Analyzer II	Paired End	36	19	37	1	11	1.75
22	SRR767799	Illumina Genome Analyzer II	Paired End	29	19	38	27	1061	0.49
23	SRR4005922	Illumina HiSeq 2500	Paired End	34	19	37	59	60519	0.86
24	SRR4005827	Illumina HiSeq 2500	Paired End	35	19	37	66	54059	0.87
25	SRR4006160	Illumina HiSeq 2000	Paired End	37	19	36	31	51928	0.86
26	SRR1573910	Illumina HiSeq 2000	Paired End	27	19	38	141	79687	0.87
27	SRR1605290	Illumina HiSeq 2000	Paired End	21	19	38	41	61313	0.86
28	SRR1612510	Illumina HiSeq 2000	Paired End	22	19	38	32	52729	0.88
29	SRR1605888	Illumina HiSeq 2000	Paired End	29	19	37	138	55978	0.95
30	SRR609055	Illumina HiSeq 2000	Paired End	26	19	37	43	62906	0.85
31	SRR628933	Illumina HiSeq 2000	Paired End	21	19	38	35	64959	0.85
32	SRR650858	Illumina HiSeq 2000	Paired End	22	19	38	18	48735	0.8665
33	SRR767807	Illumina Genome Analyzer II	Paired End	24	19	38	38	701	0.33

34	SRR071459	Illumina Genome Analyzer II	Paired End	25	19	38	6	31301	0.92
35	SRR071540	Illumina Genome Analyzer II	Paired End	26	19	37	8	35895	0.91
36	SRR3109616	Illumina HiSeq 2500	Paired End	22	19	38	65	8564	0.64
37	SRR1178929	Illumina HiSeq 2000	Paired End	22	19	38	100	76230	0.83
38	SRR2104404	Illumina HiSeq 2000	Paired End	24	19	37		59088	0.86
39	SRR2098662	Illumina HiSeq 2000	Paired End	33	19	37		39057	0.87

4.4.2 Effective Population Size History

Chromosome-wise positions, homozygous and ordered alleles are in table 4.3. The effective population size with history was calculated according to equations 4.1 and 4.2, from the MSMC2 output (Table 4.4). A population bottleneck (N_e from ~$5.98*10^7$ to ~$4.5*10^3$) was observed around ~26,000 up to ~5,000 years ago (Fig. 4.4.2.1-2). This observation supports part of Malaria's Eve hypothesis (a recent population bottleneck) but cannot confirm rapid population expansion, although a rise of effective population size was observed between last ~5000-~600 years ago (Table 4.4). Older history, like long term effective population size in the past 300,000-400,000 years is also out of the scope of this study [236].

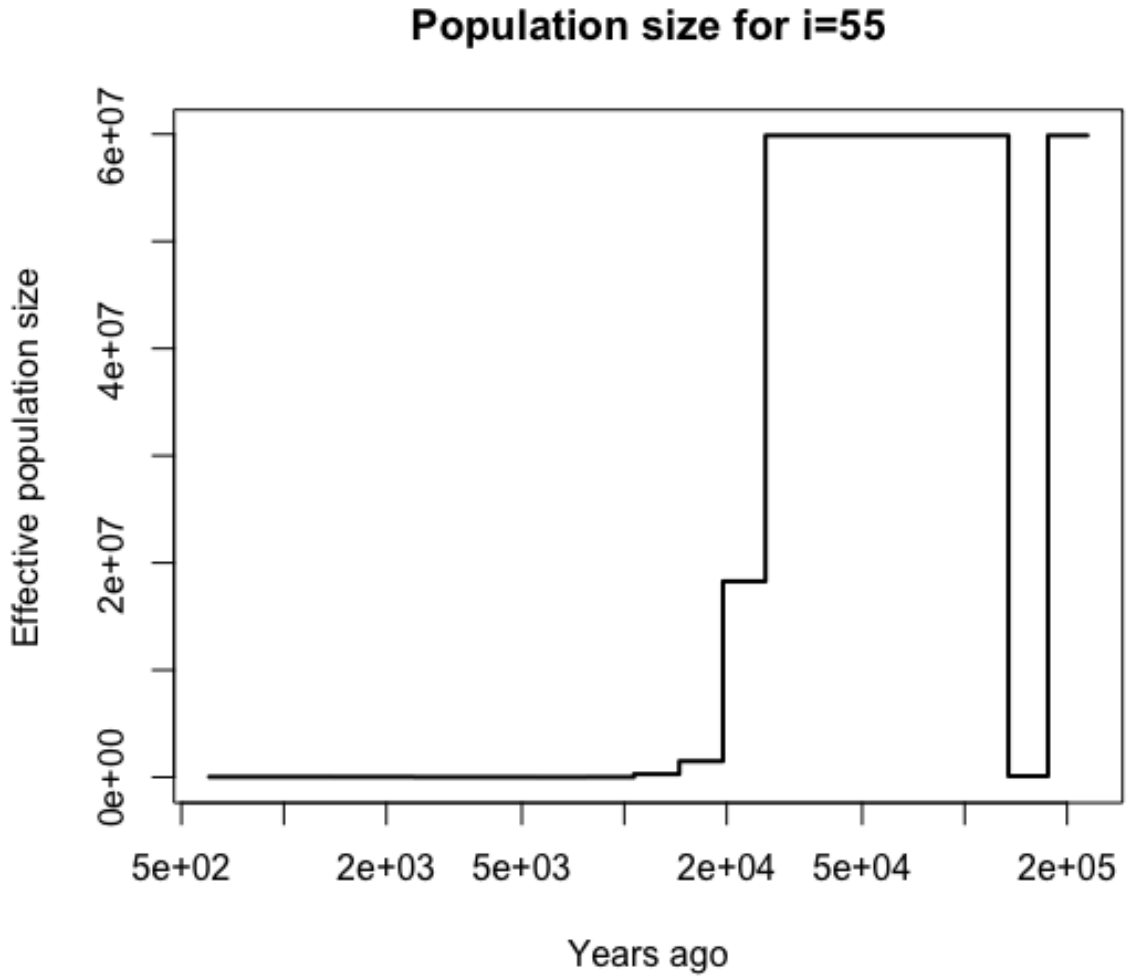

Figure 4.4.2.1: Effective population size history of *P. falciparum*.

The estimate of effective population size prior to the bottleneck is consistent with previous studies applying allele based population genetics approaches, which indicate a large N_e for *P. falciparum*, including estimates of 6.2E+05 [237], 6.9E+06 (an average, [238], and 2.6E+05 [239].

Table 4.3: Variants used in MSMC2 inference. Considering the size of the table, it is stored in the github page. This table can be retrieved using the following link:

https://github.com/zillurbmb51/results/blob/master/all.snps.txt

The description of the columns is as follows:

1. Chromosome name
2. Position of a segregating site
3. The number of called sites since the previous segregating site, including the current site
4. The ordered and phased alleles of the multiple haplotypes

Table 4.4: Population size with time history for average mutation rate. The first four columns are the output from MSMC2 program, and the last two columns were calculated using equations 4.1 and 4.2

Time Index	left_time boundary	right_time boundary	lambda	years_ago	effective_population_size
1	1.66E-05	3.83E-05	3649.38	603.21	29784.69
2	3.83E-05	6.66E-05	4043.05	1388.71	26884.57
3	6.66E-05	0.00010332	13115.3	2411.61	8287.70
4	0.00010332	0.0001512	14016.8	3743.62	7754.67
5	0.0001512	0.00021354	23702.5	5478.22	4585.83
6	0.00021354	0.00029473	8482.71	7737.03	12813.79
7	0.00029473	0.00040045	375.423	10678.51	289528.48
8	0.00040045	0.00053812	72.5763	14508.91	1497674.20
9	0.00053812	0.00071739	5.94791	19496.96	18274595.98
10	0.00071739	0.00095085	1.81528	25992.46	59878174.26
11	0.00095085	0.00125486	1.81528	34450.98	59878174.26
12	0.00125486	0.00165074	1.81528	45465.94	59878174.26
13	0.00165074	0.00216627	1.81528	59809.42	59878174.26
14	0.00216627	0.0028376	1.81528	78488.04	59878174.26
15	0.0028376	0.00371182	1.81528	102811.59	59878174.26
16	0.00371182	0.00485024	1116.4	134486.23	97362.64
17	0.00485024	0.00633271	1.81528	175733.33	59878174.26

The limitation of this study is that MSMC2 suit is optimized for human genome analysis. That is why the parameter value of "p" was changed from the default because of reduced proteome size to minimize the parameter space search area [244]. Another limitation of this thesis is that, the pathway annotations were totally dependent on the entries in different databases, which are not devoid of error [258].

This observation can be concluded as a relaxed selection pressure on the pathway loss in this species. But the non-randomness of gene loss contradicts with this scenario (Chapter 2; Highly varying pathways are either correlated with life cycle complexity or proteome size).

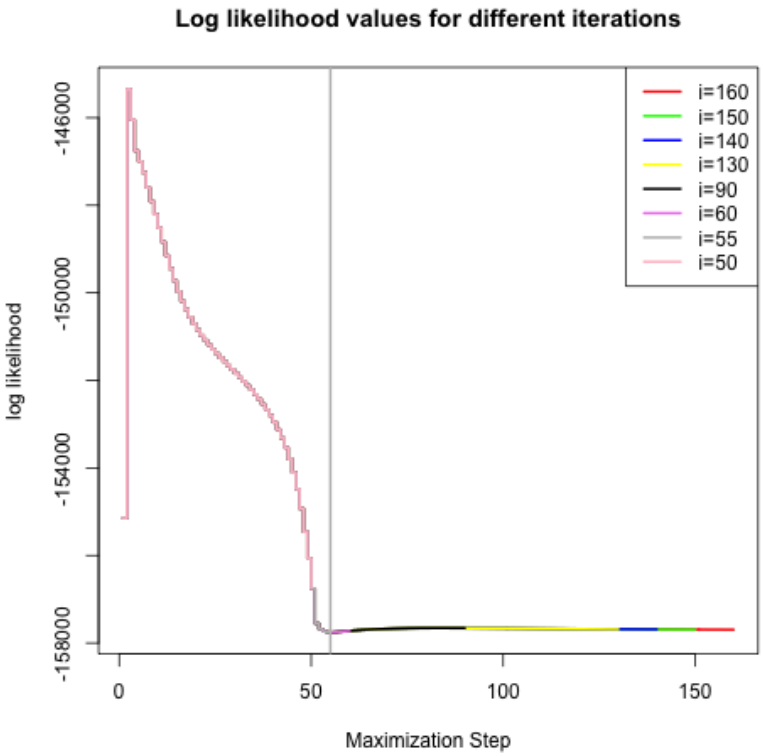

Figure 4.4.2.2: Optimum approximation parameter (i) measurement. Here, we ran msmc2 multiple times for different I to find at which stage log likelihood values becomes stationary.

In the earlier chapter, it has been shown that proteome size is significantly correlated with DNA repair and recombination proteins. DNA repair system directly contributes in mutation. The proteomic constraint theory can explain better than N_e, both the evolutionary change and pathway loss pattern in this group of organisms. The bottleneck could have an influence on pathogenicity rather than other pathways [259], though addressing this further is out of scope of this study. To reach a general theory about the effect of proteome size and effective population size on genome evolution, these two variables can be examined in other eukaryotic lineages.

Conclusion:

General Conclusion

This study explored the dynamics of pathway loss in Apicomplexans, and its relationships with evolutionary constraint, proteome size. A range of pathways associated with metabolites such as amino acids, carotenoids, folate, lipids and steroids have been lost. These metabolites may be provided by the host, and so appear to represent metabolic streamlining common in many endoparasites.

A number of pathways are differentially abundant in heteroxenous compared to monoxenous species which may reflect the adaptation with the lifestyle complexity. In particular, genes associated with N-glycan are more abundant in heteroxenous species, which may reflect an adaptation to multiple hosts.

The relation between DNA repair genes and proteome size is also observed in bacteria, archaea and DNA viruses. While in the eukaryote's *microsporidia* have lost DNA repair genes, this work is the first report of a formal correlation with proteome size in a eukaryotic lineage. This has special significance given that Apicomplexans are parasites, and elevated mutation rates act as a pathogenic factor.

A statistically robust phylogeny is reconstructed, and the biases in the correlations and the phylogeny inference are also pointed out.

Finally, using orthogroup analysis, this study showed candidate genes linked with pathogenicity in malaria-causing Plasmodium species. The high proportion of membrane and pathogenicity associated homologs implies the efficacy of this approach, which requires experimental confirmation. These thus constitute a list of candidate genes for future research.

The experimental design was focused on identifying pathogenic genes, but this study also creates a foundation to explore other lineage-specific genes and proteins.

References:

1. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* (2002) doi:10.1038/nature01097.

2. Triglia, T., Wellems, T. E. & Kemp, D. J. Towards a high-resolution map of the Plasmodium falciparum genome. *Parasitology Today* (1992) doi:10.1016/0169-4758(92)90118-L.

3. Pollack, Y., Katzen, A. L., Spira, D. T. & Golenser, J. The genome of plasmodium falciparum. I: DNA base composition. *Nucleic Acids Res.* (1982) doi:10.1093/nar/10.2.539.

4. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* (2012) doi:10.1093/nar/gks001.

5. Glockner, G. Large Scale Sequencing and Analysis of AT Rich Eukaryote Genomes. *Curr. Genomics* (2005) doi:10.2174/1389202003351472.

6. Baca, A. M. & Hol, W. G. J. Overcoming codon bias: A method for high-level overexpression of Plasmodium and other AT-rich parasite genes in Escherichia coli. *Int. J. Parasitol.* (2000) doi:10.1016/S0020-7519(00)00019-9.

7. Su, X. Z., Wu, Y., Sifri, C. D. & Wellems, T. E. Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. *Nucleic Acids Res.* (1996) doi:10.1093/nar/24.8.1574.

8. Hyman, R. W. *et al.* Sequence of Plasmodium falciparum chromosome 12. *Nature* (2002) doi:10.1038/nature01102.

9. Hall, N. *et al.* Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. *Nature* (2002) doi:10.1038/nature01095.

10. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* (2012) doi:10.1186/1471-2164-13-341.

11. No Title. *https://www.pacb.com/blog/malaria/*.

References

12. Imai, K. *et al.* A novel diagnostic method for malaria using loop-mediated isothermal amplification (LAMP) and MinION^{TM} nanopore sequencer. *BMC Infect. Dis.* (2017) doi:10.1186/s12879-017-2718-9.

13. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* (2009) doi:10.1038/nbt.1523.

14. Sandberg, R. *et al.* Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.* (2001) doi:10.1101/gr.186401.

15. Oyola, S. O. *et al.* Efficient depletion of host DNA contamination in malaria clinical sequencing. *J. Clin. Microbiol.* (2013) doi:10.1128/JCM.02507-12.

16. Auburn, S. *et al.* An effective method to purify plasmodium falciparum dna directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS One* (2011) doi:10.1371/journal.pone.0022213.

17. Melnikov, A. *et al.* Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* (2011) doi:10.1186/gb-2011-12-8-r73.

18. Bright, A. T. *et al.* Whole genome sequencing analysis of Plasmodium vivax using whole genome capture. *BMC Genomics* (2012) doi:10.1186/1471-2164-13-262.

19. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* (2011) doi:10.1371/journal.pone.0017288.

20. Langmead and Steven L Salzberg. Bowtie2. *Nat. Methods* (2013) doi:10.1038/nmeth.1923.Fast.

21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp324.

22. Aurrecoechea, C. *et al.* EuPathDB: The eukaryotic pathogen genomics database resource. *Nucleic Acids Res.* **45**, D581–D591 (2017).

23. Templeton, T. J. *et al.* Comparative analysis of apicomplexa and genomic diversity

in eukaryotes. *Genome Res.* (2004) doi:10.1101/gr.2615304.

24. Frech, C. & Chen, N. Genome comparison of human and non-human malaria parasites reveals species subset-specific genes potentially linked to human disease. *PLoS Comput. Biol.* (2011) doi:10.1371/journal.pcbi.1002320.

25. Goffeau, A. *et al.* Life with 6000 genes. *Science (80-.).* (1996) doi:10.1126/science.274.5287.546.

26. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* (2002) doi:10.1038/nature01099.

27. Carlton, J., Silva, J. & Hall, N. The genome of model malaria parasites, and comparative genomics. *Current Issues in Molecular Biology* (2005).

28. Reid, A. J. *et al.* Comparative genomics of the apicomplexan parasites Toxoplasma gondii and neospora caninum: Coccidia differing in host range and transmission strategy. *PLoS Pathog.* (2012) doi:10.1371/journal.ppat.1002567.

29. Wasmuth, J., Daub, J., Peregrín-Alvarez, J. M., Finney, C. A. M. & Parkinson, J. The origins of apicomplexan sequence innovation. *Genome Res.* (2009) doi:10.1101/gr.083386.108.

30. Balaji, S., Madan Babu, M., Iyer, L. M. & Aravind, L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.* (2005) doi:10.1093/nar/gki709.

31. Jeninga, M. D., Quinn, J. E. & Petter, M. Apiap2 transcription factors in apicomplexan parasites. *Pathogens* (2019) doi:10.3390/pathogens8020047.

32. De Silva, E. K. *et al.* Specific DNA-binding by Apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci.* (2008) doi:10.1073/pnas.0801993105.

33. Outlaw, D. C. & Ricklefs, R. E. Comparative Gene Evolution in Haemosporidian (Apicomplexa) Parasites of Birds and Mammals. *Mol. Biol. Evol.* (2010)

doi:10.1093/molbev/msp283.

34. Klinger, C. M., Nisbet, R. E., Ouologuem, D. T., Roos, D. S. & Dacks, J. B. Cryptic organelle homology in apicomplexan parasites: Insights from evolutionary cell biology. *Current Opinion in Microbiology* (2013) doi:10.1016/j.mib.2013.07.015.

35. Janouškovec, J., Horák, A., Oborník, M., Lukeš, J. & Keeling, P. J. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc. Natl. Acad. Sci. U. S. A.* (2010) doi:10.1073/pnas.1003335107.

36. Payne, S. H. & Loomis, W. F. Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Eukaryot. Cell* (2006) doi:10.1128/EC.5.2.272-276.2006.

37. Imlay, L. & Odom, A. R. Isoprenoid Metabolism in Apicomplexan Parasites. *Curr. Clin. Microbiol. Reports* (2014) doi:10.1007/s40588-014-0006-7.

38. Papkou, A., Gokhale, C. S., Traulsen, A. & Schulenburg, H. Host–parasite coevolution: why changing population size matters. *Zoology* (2016) doi:10.1016/j.zool.2016.02.001.

39. Moran, N. A. Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology* (2003) doi:10.1016/j.mib.2003.08.001.

40. Massey, S. E. The proteomic constraint and its role in molecular evolution. *Mol. Biol. Evol.* (2008) doi:10.1093/molbev/msn210.

41. Cai, H., Zhou, Z., Gu, J. & Wang, Y. Comparative Genomics and Systems Biology of Malaria Parasites Plasmodium. *Curr. Bioinform.* (2013) doi:10.2174/157489312803900965.

42. Cowell, A. N. *et al.* Mapping the malaria parasite druggable genome by using in vitro evolution and chemogenomics. *Science (80-.).* (2018) doi:10.1126/science.aan4472.

43. Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* (2008) doi:10.1038/nature07327.

References

44. https://www.malariavaccine.org/malaria-and-vaccines/first-generation-vaccine/rtss. *https://www.malariavaccine.org/malaria-and-vaccines/first-generation-vaccine/rtss*.

45. Olotu, A. *et al.* Seven-Year Efficacy of RTS,S/AS01 Malaria Vaccine among Young African Children. *N. Engl. J. Med.* (2016) doi:10.1056/NEJMoa1515257.

46. ter Kuile, F. O. *et al.* Halofantrine versus mefloquine in treatment of multidrug-resistant falciparum malaria. *Lancet* (1993) doi:10.1016/0140-6736(93)92409-M.

47. Foley, M. & Tilley, L. Quinoline antimalarials: Mechanisms of action and resistance and prospects for new agents. *Pharmacology and Therapeutics* (1998) doi:10.1016/S0163-7258(98)00012-6.

48. Ecker, A., Lehane, A. M., Clain, J. & Fidock, D. A. PfCRT and its role in antimalarial drug resistance. *Trends in Parasitology* (2012) doi:10.1016/j.pt.2012.08.002.

49. Price, R. N. *et al.* The pfmdr1 gene is associated with a multidrug-resistant phenotype in Plasmodium falciparum from the western border of Thailand. *Antimicrob. Agents Chemother.* (1999) doi:10.1128/aac.43.12.2943.

50. Krogstad, D. J. *et al.* Efflux of chloroquine from Plasmodium falciparum: Mechanism of chloroquine resistance. *Science (80-.).* (1987) doi:10.1126/science.3317830.

51. Singh Sidhu, A. B., Verdier-Pinard, D. & Fidock, D. A. Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. *Science (80-.).* (2002) doi:10.1126/science.1074045.

52. Plowe, C. V., Kublin, J. G. & Doumbo, O. K. P. falciparum dihydrofolate reductase and dihydropteroate synthase mutations: epidemiology and role in clinical resistance to antifolates. *Drug Resistance Updates* (1998) doi:10.1016/S1368-7646(98)80014-9.

53. https://www.sigmaaldrich.com/life-science/biochemicals/biochemical-products.html?TablePage=14837959.

54. Siński, E. *et al.* Apicomplexan parasites: Environmental contamination and

transmission. *Polish J. Microbiol.* (2004).

55. Kamani, J. *et al.* Molecular Detection and Characterization of Tick-borne Pathogens in Dogs and Ticks from Nigeria. *PLoS Negl. Trop. Dis.* (2013) doi:10.1371/journal.pntd.0002108.

56. Bowie, W. R. *et al.* Outbrak of toxoplasmosis associated with municipal drinking water. *Lancet* (1997) doi:10.1016/S0140-6736(96)11105-3.

57. O'Donoghue, P. J. Cryptosporidium and cryptosporidiosis in man and animals. *International Journal for Parasitology* (1995) doi:10.1016/0020-7519(94)E0059-V.

58. Di Gliullo, A. B., Cribark, M. S., Bava, A. J., Cicconetti, J. S. & Collazos, R. Cyclospora cayetanensis in sputum and stool samples. *Rev. Inst. Med. Trop. Sao Paulo* (2000) doi:10.1590/S0036-46652000000200009.

59. Dubey, J., Speer, C. & Fayer, R. Sarcocystosis of Animals and Man. *Parasitol. Today* (1989) doi:10.1017/S0031182000078902.

60. Stelly, N., Mauger, J. P., Claret, M. & Adoutte, A. Cortical alveoli of Paramecium: A vast submembranous calcium storage compartment. *J. Cell Biol.* (1991) doi:10.1083/jcb.113.1.103.

61. Klinger, C. M., Klute, M. J. & Dacks, J. B. Comparative Genomic Analysis of Multi-Subunit Tethering Complexes Demonstrates an Ancient Pan-Eukaryotic Complement and Sculpting in Apicomplexa. *PLoS One* (2013) doi:10.1371/journal.pone.0076278.

62. Van Dooren, G. G. *et al.* Development of the endoplasmic reticulum, mitochondrion and apicoplast during the asexual life cycle of Plasmodium falciparum. *Mol. Microbiol.* **57**, 405–419 (2005).

63. Stanway, R. R., Witt, T., Zobiak, B., Aepfelbacher, M. & Heussler, V. T. GFP-targeting allows visualization of the apicoplast throughout the life cycle of live malaria parasites. *Biol. Cell* (2009) doi:10.1042/BC20080202.

64. Ralph, S. A. *et al.* Tropical infectious diseases: metabolic maps and functions of the

Plasmodium falciparum apicoplast. *Nat. Rev. Microbiol.* (2004) doi:10.1038/nrmicro843.

65. Morrissette, N. S. & Sibley, L. D. Cytoskeleton of apicomplexan parasites. *Microbiol. Mol. Biol. Rev.* (2002).

66. Zhu, G., Marchewka, M. J. & Keithly, J. S. Cryptosporidium parvum appears to lack a plastid genome. *Microbiology* (2000) doi:10.1099/00221287-146-2-315.

67. Singh, S., Plassmeyer, M., Gaur, D. & Miller, L. H. Mononeme: A new secretory organelle in Plasmodium falciparum merozoites identified by localization of rhomboid-1 protease. *Proc. Natl. Acad. Sci. U. S. A.* (2007) doi:10.1073/pnas.0709999104.

68. Tran, J. Q. *et al.* RNG1 is a late marker of the apical polar ring in Toxoplasma gondii. *Cytoskeleton* (2010) doi:10.1002/cm.20469.

69. Morrissette, N. S. & Sibley, L. D. Cytoskeleton of Apicomplexan Parasites. *Microbiol. Mol. Biol. Rev.* (2002) doi:10.1128/MMBR.66.1.21-38.2002.

70. Blackman, M. J. & Bannister, L. H. Apical organelles of Apicomplexa: Biology and isolation by subcellular fractionation. *Molecular and Biochemical Parasitology* (2001) doi:10.1016/S0166-6851(01)00328-0.

71. Phillips-Howard, P. A. Malaria. Principles and Practice of Malariology. Vol 1 & 2. *J. R. Soc. Med.* (1989).

72. Carruthers, V. B., Giddings, O. K. & Sibley, L. D. Secretion of micronemal proteins is associated with toxoplasma invasion of host cells. *Cell. Microbiol.* (1999) doi:10.1046/j.1462-5822.1999.00023.x.

73. Scholtyseck, E. & Mehlhorn, H. Ultrastructural study of characteristic organelles (paired organelles, micronemes, micropores) of sporozoa and related organisms. *Zeitschrift für Parasitenkd.* (1970) doi:10.1007/BF00260383.

74. Striepen, B., Jordan, C. N., Reiff, S. & Van Dooren, G. G. Building the perfect parasite: Cell division in apicomplexa. *PLoS Pathogens* (2007)

doi:10.1371/journal.ppat.0030078.

75. Seeber, F. & Steinfelder, S. Recent advances in understanding apicomplexan parasites. *F1000Research* (2016) doi:10.12688/f1000research.7924.1.

76. Simdyanov, T. G. *et al.* A new view on the morphology and phylogeny of eugregarines suggested by the evidence from the gregarine Ancora sagittata (Leuckart, 1860) Labbé, 1899 (Apicomplexa: Eugregarinida). *PeerJ* (2017) doi:10.7717/peerj.3354.

77. Greenwood, B. M. *et al.* Malaria: Progress, perils, and prospects for eradication. *Journal of Clinical Investigation* (2008) doi:10.1172/JCI33996.

78. Florens, L. *et al.* A proteomic view of the Plasmodium falciparum life cycle. *Nature* (2002) doi:10.1038/nature01107.

79. Baum, J. *et al.* A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites. *J. Biol. Chem.* (2006) doi:10.1074/jbc.M509807200.

80. Kudryashev, M. *et al.* Structural basis for chirality and directional motility of Plasmodium sporozoites. *Cell. Microbiol.* (2012) doi:10.1111/j.1462-5822.2012.01836.x.

81. Touray, M. G. Developmentally regulated infectivity of malaria sporozoites for mosquito salivary glands and the vertebrate host. *J. Exp. Med.* (1992) doi:10.1084/jem.175.6.1607.

82. Soulard, V. *et al.* Plasmodium falciparum full life cycle and Plasmodium ovale liver stages in humanized mice. *Nat. Commun.* (2015) doi:10.1038/ncomms8690.

83. Antinori, S., Galimberti, L., Milazzo, L. & Corbellino, M. Biology of human malaria plasmodia including Plasmodium knowlesi. *Mediterranean Journal of Hematology and Infectious Diseases* (2012) doi:10.4084/MJHID.2012.013.

84. Markus, M. B. Malaria: Origin of the Term ‘Hypnozoite’. *J. Hist. Biol.* (2011) doi:10.1007/s10739-010-9239-3.

References

85. Prudêncio, M., Rodriguez, A. & Mota, M. M. The silent path to thousands of merozoites: The Plasmodium liver stage. *Nature Reviews Microbiology* (2006) doi:10.1038/nrmicro1529.

86. Cowman, A. F. & Crabb, B. S. Invasion of red blood cells by malaria parasites. *Cell* (2006) doi:10.1016/j.cell.2006.02.006.

87. Udagama, P. V *et al.* Immunoelectron microscopy of Schüffner's dots in Plasmodium vivax-infected human erythrocytes. *Am. J. Pathol.* (1988).

88. Weatherall, D. J. *et al.* Malaria and the red cell. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program* (2002) doi:10.1182/asheducation-2002.1.35.

89. Lew, V. L., Tiffert, T. & Ginsburg, H. Excess hemoglobin digestion and the osmotic stability of Plasmodium falciparum - Infected red blood cells. *Blood* (2003) doi:10.1182/blood-2002-08-2654.

90. Carter, R. & Beach, R. F. Gametogenesis in culture by gametocytes of Plasmodium falciparum. *Nature* (1977) doi:10.1038/270240a0.

91. Baker, D. A. Malaria gametocytogenesis. *Molecular and Biochemical Parasitology* (2010) doi:10.1016/j.molbiopara.2010.03.019.

92. Beier, J. C. MALARIA PARASITE DEVELOPMENT IN MOSQUITOES. *Annu. Rev. Entomol.* (1998) doi:10.1146/annurev.ento.43.1.519.

93. Walliker, D., Billingsley, P. & Currie, D. Random Mating in a Natural Population of the Malaria Parasite Plasmodium Falciparum. *Parasitology* (1994) doi:10.1017/S0031182000080665.

94. Vaughan, J., Noden, B. & Beier, J. Population dynamics of Plasmodium falciparum sporogony in laboratory-infected Anopheles gambiae. *J Parasitol.* (1992) doi:10.2307/3283550.

95. Vaughan, J. A., Noden, B. H. & Beier, J. C. Sporogonic development of cultured Plasmodium falciparum in six species of laboratory-reared Anopheles mosquitoes.

Am. J. Trop. Med. Hyg. (1994) doi:10.4269/ajtmh.1994.51.233.

96. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* vol. 27 29–34 (1999).

97. Wyatt, C. R., Riggs, M. W. & Fayer, R. Cryptosporidiosis in Neonatal Calves. *Veterinary Clinics of North America - Food Animal Practice* (2010) doi:10.1016/j.cvfa.2009.10.001.

98. Woehle, C., Dagan, T., Martin, W. F. & Gould, S. B. Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related Chromera velia. *Genome Biol. Evol.* (2011) doi:10.1093/gbe/evr100.

99. Oborník, M. *et al.* Morphology, Ultrastructure and Life Cycle of Vitrella brassicaformis n. sp., n. gen., a Novel Chromerid from the Great Barrier Reef. *Protist* (2012) doi:10.1016/j.protis.2011.09.001.

100. Watts, J. G., Playford, M. C. & Hickey, K. L. Theileria orientalis: a review. *New Zealand Veterinary Journal* (2016) doi:10.1080/00480169.2015.1064792.

101. Shutler, D., Reece, S. E., Mullie, A., Billingsley, P. F. & Read, A. F. Rodent malaria parasites Plasmodium chabaudi and P. vinckei do not increase their rates of gametocytogenesis in response to mosquito probing. *Proc. R. Soc. B Biol. Sci.* (2005) doi:10.1098/rspb.2005.3232.

102. Galland, G. G. Role of the squirrel monkey in parasitic disease research. *ILAR J.* (2000) doi:10.1093/ilar.41.1.37.

103. Kaiser, M. *et al.* Wild Chimpanzees Infected with 5 Plasmodium Species. *Emerg. Infect. Dis.* (2010) doi:10.3201/eid1612.100424.

104. Coatney, G. R., Chin, W., Contacos, P. G. & King, H. K. Plasmodium inui, a Quartan-Type Malaria Parasite of Old World Monkeys Transmissible to Man. *J. Parasitol.* (2006) doi:10.2307/3276423.

105. Ollomo, B. *et al.* A new malaria agent in African hominids. *PLoS Pathog.* (2009)

doi:10.1371/journal.ppat.1000446.

106. Dubey, J. P. & Ferguson, D. J. P. Life cycle of Hammondia hammondi (Apicomplexa: Sarcocystidae) in cats. *J. Eukaryot. Microbiol.* (2015) doi:10.1111/jeu.12188.

107. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science (80-.).* (1997) doi:10.1126/science.278.5338.631.

108. Kristensen, D. M., Wolf, Y. I., Mushegian, A. R. & Koonin, E. V. Computational methods for Gene Orthology inference. *Brief. Bioinform.* (2011) doi:10.1093/bib/bbr030.

109. Hoo, R. *et al.* Integrated analysis of the Plasmodium species transcriptome. *EBioMedicine* (2016) doi:10.1016/j.ebiom.2016.04.011.

110. Al-Nihmi, F. M. A. *et al.* A Novel and Conserved Plasmodium Sporozoite Membrane Protein SPELD is Required for Maturation of Exo-erythrocytic Forms. *Sci. Rep.* (2017) doi:10.1038/srep40407.

111. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* (2003) doi:10.1101/gr.1224503.

112. Huynh, M.-H. & Carruthers, V. B. A Toxoplasma gondii ortholog of Plasmodium GAMA contributes to parasite attachment and cell invasion. *mSphere* (2016) doi:10.1128/mSphere.00012-16.Editor.

113. Polley, S. D., Weedall, G. D., Thomas, A. W., Golightly, L. M. & Conway, D. J. Orthologous gene sequences of merozoite surface protein 1 (MSP1) from Plasmodium reichenowi and P. gallinaceum confirm an ancient divergence of P. falciparum alleles. *Mol. Biochem. Parasitol.* (2005) doi:10.1016/j.molbiopara.2005.02.012.

114. Wang, B. *et al.* Identification and characterization of the Plasmodium falciparum RhopH2 ortholog in Plasmodium vivax. *Parasitol. Res.* (2013) doi:10.1007/s00436-012-3170-9.

115. Prajapati, S. K. & Singh, O. P. Identification of a vir-orthologous immune evasion

gene family from primate malaria parasites. *Parasitology* (2014) doi:10.1017/S003118201300214X.

116. Poulin, R. & Randhawa, H. S. Evolution of parasitism along convergent lines: From ecology to genomics. *Parasitology* (2015) doi:10.1017/S0031182013001674.

117. Goodman, C. D. & McFadden, G. I. Fatty acid biosynthesis as a drug target in apicomplexan parasites. *Curr. Drug Targets* (2007) doi:10.2174/138945007779315579.

118. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* (2005) doi:10.1146/annurev.genet.39.073003.114725.

119. Mahmud, O. & Kissinger, J. C. Evolution of the Apicomplexan Sugar Transporter Gene Family Repertoire. *Int. J. Genomics* (2017) doi:10.1155/2017/1707231.

120. Vallender, E. J. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. *Methods* (2009) doi:10.1016/j.ymeth.2009.05.010.

121. Fitch, W. M. Homologya personal view on some of the problems. *Trends in Genetics* (2000) doi:10.1016/S0168-9525(00)02005-9.

122. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* (2015) doi:10.1093/nar/gku1203.

123. Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. in *Bioinformatics* (2006). doi:10.1093/bioinformatics/btl213.

124. Altenhoff, A. M., Schneider, A., Gonnet, G. H. & Dessimoz, C. OMA 2011: Orthology inference among 1000 complete genomes. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkq1238.

125. DeLuca, T. F. *et al.* Roundup: A multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* (2006) doi:10.1093/bioinformatics/btl286.

126. Trachana, K. *et al.* Orthology prediction methods: A quality assessment using curated protein families. *BioEssays* (2011) doi:10.1002/bies.201100062.

References

127. Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* (2000) doi:10.1093/nar/28.1.33.

128. Huerta-Cepas, J. *et al.* EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* (2016) doi:10.1093/nar/gkv1248.

129. Zdobnov, E. M. *et al.* OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkw1119.

130. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* (2009) doi:10.1101/gr.073585.107.

131. Huerta-Cepas, J. *et al.* PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkq1109.

132. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, (2015).

133. Fiebig, M., Kelly, S. & Gluenz, E. Comparative Life Cycle Transcriptomics Revises Leishmania mexicana Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. *PLoS Pathog.* (2015) doi:10.1371/journal.ppat.1005186.

134. Kao, D. *et al.* The genome of the crustacean parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife* (2016) doi:10.7554/eLife.20062.001.

135. Barker, M. S. *et al.* Most compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the calyceraceae. *Am. J. Bot.* (2016) doi:10.3732/ajb.1600113.

136. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Comput. Sci. Eng.* **9**,

10–20 (2007).

137. McKinney, W. & Team, P. D. Pandas - Powerful Python Data Analysis Toolkit. *Pandas - Powerful Python Data Anal. Toolkit* 1625 (2015).

138. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 99–104 (2007).

139. Bressert, E. *SciPy and NumPy*. *Journal of Chemical Information and Modeling* (2013). doi:10.1017/CBO9781107415324.004.

140. Seaborn. Seaborn. *https://seaborn.pydata.org/* (2012) doi:10.5281/zenodo.592845.

141. Galili, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv428.

142. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* (2005) doi:10.1093/bioinformatics/bti610.

143. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx364.

144. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* (2016) doi:10.1016/j.jmb.2015.11.006.

145. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkw1099.

146. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* (1990) doi:10.1016/S0022-2836(05)80360-2.

147. Chapman, B. A. & Chang, J. T. Biopython: Python tools for computational biology. *ACM SIGBIO Newsl.* **20**, 15–19 (2000).

148. Felsenstein, J. Comparative Methods with Sampling Error and Within-Species Variation: Contrasts Revisited and Revised. *Am. Nat.* (2008) doi:10.1086/587525.

149. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* (2004) doi:10.1093/bioinformatics/btg412.

150. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* (1985) doi:10.1086/284325.

151. Toso, M. A. & Omoto, C. K. Gregarina niphandrodes may lack both a plastid genome and organelle. *J. Eukaryot. Microbiol.* (2007) doi:10.1111/j.1550-7408.2006.00229.x.

152. Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat. Commun.* (2014) doi:10.1038/ncomms5754.

153. Müller, I. B., Hyde, J. E. & Wrenger, C. Vitamin B metabolism in Plasmodium falciparum as a source of drug targets. *Trends in Parasitology* (2010) doi:10.1016/j.pt.2009.10.006.

154. Du, Q., Wang, H. & Xie, J. Thiamin (vitamin B1) biosynthesis and regulation: A rich source of antimicrobial drug targets? *International Journal of Biological Sciences* (2011) doi:10.7150/ijbs.7.41.

155. Weiner, J. & Kooij, T. W. A. Phylogenetic profiles of all membrane transport proteins of the malaria parasite highlight new drug targets. *Microb. Cell* (2016) doi:10.15698/mic2016.10.534.

156. Mackinnon, M. J. & Read, A. F. Virulence in malaria: An evolutionary viewpoint. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2004) doi:10.1098/rstb.2003.1414.

157. EWALD, P. W. Transmission Modes and Evolution of the Parasitism-Mutualism Continuum. *Ann. N. Y. Acad. Sci.* (1987) doi:10.1111/j.1749-6632.1987.tb40616.x.

158. Acosta, S. *et al.* DNA Repair Is Associated with Information Content in Bacteria,

Archaea, and DNA Viruses. *J. Hered.* (2015) doi:10.1093/jhered/esv055.

159. Dobzhansky, T. Nothing in Biology Makes Sense except in the Light of Evolution. *Am. Biol. Teach.* (1973) doi:10.2307/4444260.

160. Darwin, C. *On the Origin of the Species*. *Darwin* (1859).

161. Saitou, N. & Nei, M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evo* **4**, 406–425 (1987).

162. Rzhetsky, A. & Nei, M. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.* **35**, 367–375 (1992).

163. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368–376 (1981).

164. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* vol. 294 2310–2314 (2001).

165. Philippe, H. *et al.* Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol.* (2011) doi:10.1371/journal.pbio.1000602.

166. Sogin, M. L. Early evolution and the origin of eukaryotes. *Curr. Opin. Genet. Dev.* (1991) doi:10.1016/S0959-437X(05)80192-3.

167. Gupta, R. S. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* **62**, 1435–1491 (1998).

168. Olsen, G. J. & Woese, C. R. Ribosomal RNA: a key to phylogeny. *FASEB J.* **7**, 113–123 (1993).

169. Hashimoto, T. *et al.* Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia. *Mol. Biol. Evol.* **11**, 65–71 (1994).

170. Awasthi, V., Chauhan, R., Chattopadhyay, D. & Das, J. Effect of l-arginine on the growth of Plasmodium falciparum and immune modulation of host cells. *J. Vector Borne Dis.* (2017).

171. Fullerton, S. M., Carvalho, A. B. & Clark, A. G. Local rates of recombination are positively correlated with GC content in the human genome [4]. *Molecular Biology and Evolution* (2001) doi:10.1093/oxfordjournals.molbev.a003886.

172. Romiguier, J. & Roux, C. Analytical biases associated with GC-content in molecular evolution. *Front. Genet.* (2017) doi:10.3389/fgene.2017.00016.

173. Bossert, S., Murray, E. A., Blaimer, B. B. & Danforth, B. N. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. *Mol. Phylogenet. Evol.* (2017) doi:10.1016/j.ympev.2017.03.022.

174. Fitch, W. M. & Margoliash, E. Construction of Phylogenetic Trees. *Science (80-.).* (1967) doi:10.1126/science.155.3760.279.

175. Mihaescu, R., Levy, D. & Pachter, L. Why neighbor-joining works. *Algorithmica (New York)* (2009) doi:10.1007/s00453-007-9116-4.

176. Felsenstein, J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst. Biol.* (1978) doi:10.1093/sysbio/27.4.401.

177. Penny, D. & Hendy, M. Estimating the reliability of evolutionary trees. *Mol Biol Evol* (1986).

178. Page, R. D. M. & Charleston, M. A. From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. *Mol. Phylogenet. Evol.* **7**, 231–240 (1997).

179. Eisen, J. A. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* (1998) doi:10.1101/gr.8.3.163.

180. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science (80-.).* (2014) doi:10.1126/science.1257570.

181. Kuo, C. H., Wares, J. P. & Kissinger, J. C. The apicomplexan whole-genome

phylogeny: An analysis of incongruence among gene trees. *Mol. Biol. Evol.* (2008) doi:10.1093/molbev/msn213.

182. Woo, Y. H. *et al.* Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* (2015) doi:10.7554/eLife.06974.

183. Arisue, N. & Hashimoto, T. Phylogeny and evolution of apicoplasts and apicomplexan parasites. *Parasitol. Int.* **64**, 254–259 (2015).

184. Leander, B. S., Clopton, R. E. & Keeling, P. J. Phylogeny of grenarines (Apicomplexa) as inferred from a small-subunit rDNA and β-tubulin. *Int. J. Syst. Evol. Microbiol.* (2003) doi:10.1099/ijs.0.02284-0.

185. Parker, M. L. *et al.* Dissecting the interface between apicomplexan parasite and host cell: Insights from a divergent AMA–RON2 pair. *Proc. Natl. Acad. Sci.* (2015) doi:10.1073/pnas.1515898113.

186. Mina, J. G. *et al.* Functional and phylogenetic evidence of a bacterial origin for the first enzyme in sphingolipid biosynthesis in a phylum of eukaryotic protozoan parasites. *J. Biol. Chem.* (2017) doi:10.1074/jbc.M117.792374.

187. Cova, M. *et al.* The Apicomplexa-specific glucosamine-6-phosphate N-acetyltransferase gene family encodes a key enzyme for glycoconjugate synthesis with potential as therapeutic target. *Sci. Rep.* (2018) doi:10.1038/s41598-018-22441-3.

188. Mahmud, O. & Kissinger, J. C. Evolution of the Apicomplexan Sugar Transporter Gene Family Repertoire. *Int. J. Genomics* (2017) doi:10.1155/2017/1707231.

189. Leander, B. S. Marine gregarines: evolutionary prelude to the apicomplexan radiation? *Trends Parasitol.* (2008) doi:10.1016/j.pt.2007.11.005.

190. Galen, S. C. *et al.* The polyphyly of Plasmodium: Comprehensive phylogenetic analyses of the malaria parasites (Order Haemosporida) reveal widespread taxonomic conflict. *R. Soc. Open Sci.* **5**, (2018).

References

191. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).

192. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp348.

193. Gouy, M., Guindon, S. & Gascuel, O. Sea view version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* (2010) doi:10.1093/molbev/msp259.

194. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* (2018) doi:10.1093/molbev/msy096.

195. Ginestet, C. ggplot2: Elegant Graphics for Data Analysis. *J. R. Stat. Soc. Ser. A (Statistics Soc.* (2011) doi:10.1111/j.1467-985x.2010.00676_9.x.

196. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

197. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

198. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).

199. Miller, M. A., Pfeiffer, W. & Schwartz, T. The CIPRES science gateway. in *Proceedings of the 2011 TeraGrid Conference on Extreme Digital Discovery - TG '11* (2011). doi:10.1145/2016741.2016785.

200. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* (2018) doi:10.1093/sysbio/syy032.

References

201. Phillips, N. E., Smith, C. M. & Morden, C. W. Testing systematic concepts of Sargassum (Fucales, Phaeophyceae) using portions of the rbcLS operon. *Phycol. Res.* (2005) doi:10.1111/j.1440-1835.2005.tb00353.x.

202. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* (2008) doi:10.1093/molbev/msn067.

203. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* (1974) doi:10.1109/TAC.1974.1100705.

204. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* (1978) doi:10.1214/aos/1176344136.

205. Michener, C. D. & Sokal, R. R. A Quantitative Approach to a Problem in Classification. *Evolution (N. Y).* (1957) doi:10.2307/2406046.

206. Mount, D. W. Maximum parsimony method for phylogenetic prediction. *Cold Spring Harb. Protoc.* (2008) doi:10.1101/pdb.top32.

207. Schreeg, M. E. *et al.* Mitochondrial genome sequences and structures aid in the resolution of Piroplasmida phylogeny. *PLoS One* (2016) doi:10.1371/journal.pone.0165702.

208. Kváč, M. *et al.* Cryptosporidium proliferans n. sp. (Apicomplexa: Cryptosporidiidae): Molecular and biological evidence of cryptic species within gastric cryptosporidium of mammals. *PLoS One* (2016) doi:10.1371/journal.pone.0147090.

209. OGEDENGBE, M. E. *et al.* Molecular phylogenetic analyses of tissue coccidia (sarcocystidae; apicomplexa) based on nuclear 18s RDNA and mitochondrial COI sequences confirms the paraphyly of the genus Hammondia. *Parasitol. Open* (2016) doi:10.1017/pao.2015.7.

210. Reiman, D. A. *et al.* Molecular Phylogenetic Analysis of Cyclospora, the Human Intestinal Pathogen, Suggests that It Is Closely Related to Eimeria Species. *J. Infect. Dis.* (1996) doi:10.1093/infdis/173.2.440.

211. Ogedengbe, J. D., Ogedengbe, M. E., Hafeez, M. A. & Barta, J. R. Molecular

phylogenetics of eimeriid coccidia (Eimeriidae, Eimeriorina, Apicomplexa, Alveolata): A preliminary multi-gene and multi-genome approach. *Parasitol. Res.* (2015) doi:10.1007/s00436-015-4646-1.

212. Singh, B. & Daneshvar, C. Human infections and detection of plasmodium knowlesi. *Clinical Microbiology Reviews* (2013) doi:10.1128/CMR.00079-12.

213. Singh, B. *et al.* Naturally acquired human infections with the simian malaria parasite, Plasmodium cynomolgi, in Sarawak, Malaysian Borneo. *Int. J. Infect. Dis.* (2018) doi:10.1016/j.ijid.2018.04.3581.

214. Perkins, S. L., Sarkar, I. N. & Carter, R. The phylogeny of rodent malaria parasites: Simultaneous analysis across three genomes. *Infect. Genet. Evol.* (2007) doi:10.1016/j.meegid.2006.04.005.

215. Collins, W. E., Chin, W. & Skinner, J. C. Plasmodium fragile and Macaca mulatta monkeys as a model system for the study of malaria vaccines. *Am. J. Trop. Med. Hyg.* (1979) doi:10.4269/ajtmh.1979.28.948.

216. Prugnolle, F. *et al.* African monkeys are infected by Plasmodium falciparum nonhuman primate-specific strains. *Proc. Natl. Acad. Sci. U. S. A.* (2011) doi:10.1073/pnas.1109368108.

217. Rayner, J. C. Plasmodium malariae Malaria: From Monkey to Man? *EBioMedicine* (2015) doi:10.1016/j.ebiom.2015.08.035.

218. Don E. Eyles, Yap Loy Fong, McWilson Warren, Elizabeth Guinn, A. A. S. and R. H. W. Plasmodium Coatneyi, a New Species of Primate Malaria from Malaya. *Am. J. Trop. Med. Hyg.* **11**, 597–604 (1962).

219. COATNEY, G. R. *et al.* Transmission of the M strain of Plasmodium cynomolgi to man. *Am. J. Trop. Med. Hyg.* (1961) doi:10.4269/ajtmh.1961.10.673.

220. Bleidorn, C. & Bleidorn, C. Sources of Error and Incongruence in Phylogenomic Analyses. in *Phylogenomics* (2017). doi:10.1007/978-3-319-54064-1_9.

221. Vahdati, A. R., Sprouffske, K. & Wagner, A. Effect of population size and mutation

rate on the evolution of RNA sequences on an adaptive landscape determined by RNA folding. *Int. J. Biol. Sci.* (2017) doi:10.7150/ijbs.19436.

222. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).

223. Alstad, D. *Basic Populus models of ecology*. (Prentice Hall, 2001).

224. Calder, A. The role of inbreeding in the development of the Clydesdale breed of horse. *Proc. R. Soc. Eninburgh* **37**, 118–140 (1927).

225. Braude, S. & Templeton, A. R. Understanding the multiple meanings of ‘inbreeding’ and ‘effective size’ for genetic management of African rhinoceros populations. *Afr. J. Ecol.* (2009) doi:10.1111/j.1365-2028.2008.00981.x.

226. Nunney, L. The effect of neighborhood size on effective population size in theory and in practice. *Heredity (Edinb).* **117**, 224–232 (2016).

227. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci.* (2012) doi:10.1073/pnas.1216223109.

228. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* (1998) doi:citeulike-article-id:610966.

229. Massey, S. Genetic Code Evolution Reveals the Neutral Emergence of Mutational Robustness, and Information as an Evolutionary Constraint. *Life* (2015) doi:10.3390/life5021301.

230. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Reply to Massey: Drift does influence mutation-rate evolution. *Proc. Natl. Acad. Sci.* (2013) doi:10.1073/pnas.1220650110.

231. Lynch, M. Evolution of the mutation rate. *Trends Genet.* (2010) doi:10.1016/j.tig.2010.05.003.

232. Matic, I. *et al.* Highly variable mutation rates in commensal and pathogenic Escherichia coli. *Science (80-.).* (1997) doi:10.1126/science.277.5333.1833.

References

233. Rich, S. M., Licht, M. C., Hudson, R. R. & Ayala, F. J. Malaria's Eve: Evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum (genetic polymorphism¯demographic sweep¯clonality¯selective sweep¯parasitic protozoa). *Evolution (N. Y).* (1998).

234. Conway, D. J. *et al.* Origin of Plasmodium falciparum malaria is traced by mitochondrial DNA. *Mol. Biochem. Parasitol.* (2000) doi:10.1016/S0166-6851(00)00313-3.

235. Volkman, S. K. *et al.* Recent origin of Plasmodium falciparum from a single progenitor. *Science (80-.).* (2001) doi:10.1126/science.1059878.

236. Hughes, A. L. & Verra, F. Very large long-term effective population size in the virulent human malaria parasite Plasmodium falciparum. *Proc. R. Soc. B Biol. Sci.* (2001) doi:10.1098/rspb.2001.1759.

237. Joy, D. A. *et al.* Early origin and recent expansion of Plasmodium falciparum. *Science (80-.).* (2003) doi:10.1126/science.1081449.

238. Chang, H. H. *et al.* Genomic sequencing of plasmodium falciparum malaria parasites from senegal reveals the demographic history of the population. *Mol. Biol. Evol.* (2012) doi:10.1093/molbev/mss161.

239. Anderson, T. J. C. *et al.* Population parameters underlying an ongoing soft sweep in southeast asian malaria parasites. *Mol. Biol. Evol.* (2017) doi:10.1093/molbev/msw228.

240. Wang, J. Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2005) doi:10.1098/rstb.2005.1682.

241. Kimura, M. & Crow, J. F. The Measurement of Effective Population Number. *Evolution (N. Y).* (1963) doi:10.2307/2406157.

242. Hill, W. G. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* (1981) doi:10.1017/S0016672300020553.

References

243. Anderson, E. C., Williamson, E. G. & Thompson, E. A. Monte Carlo evaluation of the likelihood for N(e) from temporally spaced samples. *Genetics* (2000).

244. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* (2014) doi:10.1038/ng.3015.

245. Guggisberg, A. M. *et al.* Whole-Genome Sequencing to Evaluate the Resistance Landscape Following Antimalarial Treatment Failure with Fosmidomycin-Clindamycin. *J. Infect. Dis.* (2016) doi:10.1093/infdis/jiw304.

246. Andrews, S. FastQC: A quality control tool for high throughput sequence data. Available at: www.bioinformatics.babraham.ac.uk/projects/fastqc/. *FastQC: A quality control tool for high throughput sequence data. Available at: www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2010).

247. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner*. *Joint Genome Instritute, department of energy* (2014) doi:10.1186/1471-2105-13-238.

248. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

249. Alkan, C., Coe, B. P. & Eichler, E. E. GATK toolkit. *Nat. Rev. Genet.* **12**, 363–76 (2011).

250. Broad Institute. Picard tools. *https://broadinstitute.github.io/picard/* https://broadinstitute.github.io/picard/%5Cnhttp://broadinstitute.github.io/picard/ (2016).

251. Danecek, P. & McCarthy, S. A. BCFtools/csq: Haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).

252. Claessens, A. *et al.* Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLoS Genet.* (2014) doi:10.1371/journal.pgen.1004812.

253. Bopp, S. E. R. *et al.* Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families. *PLoS Genet.* (2013)

doi:10.1371/journal.pgen.1003293.

254. Ding, J. *et al.* Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genet.* (2015) doi:10.1371/journal.pgen.1005306.

255. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btu668.

256. Hamilton, W. L. *et al.* Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Res.* **45**, 1889–1901 (2017).

257. Keller, I., Bensasson, D. & Nichols, R. A. Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes. *PLoS Genet.* (2007) doi:10.1371/journal.pgen.0030022.

258. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* (2009) doi:10.1371/journal.pcbi.1000605.

259. Geoghegan, J. L., Senior, A. M. & Holmes, E. C. Pathogen population bottlenecks and adaptive landscapes: Overcoming the barriers to disease emergence. *Proc. R. Soc. B Biol. Sci.* (2016) doi:10.1098/rspb.2016.0727.

Appendix:

Appendix C1: Publication from the thesis

Part of data from this thesis is submitted to the journal Infection, Genetics and Evolution with the following title:

Proteome size reduction in Apicomplexans is linked with loss of DNA repair, and host redundant pathways

Rahman, M.Z.[1], Derilus, D.[2], Serrano, A.E.[3], Massey, S.E.[1]

[1] Biology Department, University of Puerto Rico-Rio Piedras
[2] Environmental Sciences Department, University of Puerto Rico-Rio Piedras
[3] Department of Microbiology, University of Puerto Rico-School of Medicine, Medical Sciences

Appendix C2: Scripts developed to complete this work:

1. Proteome_pathway_cluster.py

```
#This script will create a hierarchical cluster map from abundance data
#To run: python Proteome_pathway_cluster.py ; in the same directory of the data set
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns #Load needed libraries
pt=pd.read_csv("pathway_count.csv",index_col=0) #Load the data
pt1=pt.drop(['proteome size']) #Drop unwanted columns
sns.clustermap(pt1,cmap="nipy_spectral_r") #Make the plot
plt.savefig("pathway_cluster.svg",bbox_inches='tight') #Save it aand then show and close
plt.show()
plt.close()
```

2. figure_2_3_s1_s2.py

This script will generate clustermap as number 1

```
import pandas as pd #Load the libraries
import matplotlib.pyplot as plt
```

Appendix:

```
import seaborn as sns

st=pd.read_csv("Statistics_PerSpecies.csv",index_col=0) #Load data
sns.heatmap(s4,annot=True,cmap='Set3',fmt="",linewidths=0.1,robust=True,annot_k
ws={"size":6,"rotation":'vertical'},alpha=1.0)
plt.savefig("all_stat_heat_fin_1.png",bbox_inches='tight') #make a heatmap and save
it

sp=pd.read_csv("Orthogroups_SpeciesOverlaps.csv",sep="\t",index_col=0)

a5=sns.clustermap(sp,cmap="spectral_r")
a5.savefig("species_overlap.png")

s3=sp.set_index('VbrassicaformisCCMP3155')
a7=sns.clustermap(s3,cmap="spectral_r")
a7.savefig("species_overlap_vbras.png")
s4=sp.set_index('CveliaCCMP2878')
a8=sns.clustermap(s4,cmap="spectral_r")
a8.savefig("species_overlap_cveli.png")
```

3. gc_avg.sh

```
#This script will calculate GC content of all the fasta file in a directory persequence and per file
#The 1st line will give GC content persequence
#The 2nd line will give mean GC for a file
#Run : bash gc_avg.sh

for f in *fasta; do bioawk -c fastx '{print $name, gc($seq)}' $f > $f.gc; done
for f in *gc; do awk '{ total +=$2 } END { print total/NR }' $f > $f.avg; done
```

Appendix:

4. get_ids.R

```
#This script will extract desired ids from database

source("http://bioconductor.org/biocLite.R") #Load bioconductor

biocLite("org.Pf.plasmo.db") #Install the p. falciparum databse

library(org.Pf.plasmo.db) #Load the database

ls("package:org.Pf.plasmo.db") #List available options

ids <- keys(org.Pf.plasmo.db, "SYMBOL") #Create a list with gene symbols

id_to_go=select(org.Pf.plasmo.db, ids, "GO", keytype="SYMBOL") #Create a datframe of corresponding symbol and GO

write.csv(id_to_go,"id_to_go.csv") #Write the mapping file to csv

gos=keys(org.Pf.plasmo.db,"GO") #Create a list with GO IDs

go_to_path=select(org.Pf.plasmo.db, gos,"PATH",keytype="GO") #Create dataframe of corresponding GO and K number

write.csv(go_to_path,"go_to_path.csv")
```

5. id_mapping.py

```
#This script will map different types of id from a mapping table

import pandas as pd #Import libraries

import numpy as np

import seaborn as sns

o11=pd.read_csv("Orthogroups.csv",sep='\t',index_col=0) #Load the data

idg=pd.read_csv("id_to_go.csv") #Mapping table generated from bioconductor database

o11['Pfalciparum3D7'].isnull().values.sum() #Count number of missing data in a column

idg2=idg[pd.notnull(idg['GO'])] #Create a new dataframe excluding missing values from a certain column

d11=dict(zip(idg2.SYMBOL,idg2.GO)) #Create dictionary to map

o11['go']=o11['Pfalciparum3D7'].str.extract('('+'|'.join(list(d11))+')').map(d11) #Create new column mapping with dictionary

o13=o11[pd.notnull(o11['go'])]
```

Appendix:

```
d13=dict(zip(o13.index,o13.go)) #Create dictionary with index and another column
gncnt=pd.read_csv("Orthogroups.GeneCount.csv",sep='\t',index_col=0)
gc=gncnt.drop(['Total'],axis=1) #Drop a certain column
gc2=gc.assign(orthogroups=gc.index) #Create a new column with orthogroups name
gc2['go']=gc2['orthogroups'].map(d13)
gc3=gc2[pd.notnull(gc2['go'])]
gtp=pd.read_csv("go_to_path.csv") #Load the corresponding GO IDs and K number dataframe
gtp2=gtp[pd.notnull(gtp["PATH"])]
d15=dict(zip(gtp2.GO,gtp2.PATH))
gc3['path']=gc3['go'].map(d15)
gc4=gc3[pd.notnull(gc3['path'])]
kegg=pd.read_csv("kegg_4.csv") #Load the K number and pathway name
kegg.columns=['path','pathway'] #Naming columns
kegg['pathway']=kegg['pathway'].map(str.strip) #remove extra white spaces from strings
kegg2=kegg[pd.notnull(kegg['path'])]
d17=dict(zip(kegg2.path,kegg2.pathway))
gc4['pathway']=gc4['path'].map(d17)
gc5=gc4.drop(['orthogroups','go','path'],axis=1)
gc6=gc5.groupby(['pathway']).sum() #Tidy data for next step of analysis
gc6.to_csv("pathway_count.csv")
sns.clustermap(gc6,cmap="nipy_spectral_r") #Hierarchical clustering
plt.savefig("pathway_cluster.svg",bbox_inches='tight')
```

Appendix:

6. proteome_size.py

```
#This is from biopython tutorial not modified
from Bio.SeqIO.FastaIO import SimpleFastaParser #Load fasta parser from
biopython
count=0 #Start counting as 0
total=0
with open(raw_input("Enter the fasta file:"),'rU') as fasta: #Load the fasta file
for t,s in SimpleFastaParser(fasta): #parse sequence and update within loop
count+=1
total+=len(s)
print("%i records with proteome size %i" %(count,total)) #Print the number of sequences
and Aas
```

7. tanglegrams.R

```
#This script will compare two dendrograms and find statistical differences

library(dendextend)

phyl1=read.csv("mega/d3.csv",row.names=1,header=F) #Phylogenomic distance from
mega

phyl1[is.na(phyl1)]<-0 #Replace NA with 0

dphylo=as.dendrogram(hclust(dist(phyl1))) #Create dendrogram of hierarchical cluster

ortho=read.csv("Orthogroups_SpeciesOverlaps.csv",sep='\t',row.names=1)

dortho=as.dendrogram(hclust(dist(ortho)))

path=read.csv("pathway_count.csv",row.names=1)

dpath=as.dendrogram(hclust(dist(t(path))))

dphylortho=dendlist(dphylo,dortho) #Create dendrogram list

dpathortho=dendlist(dpath,dortho)

dphylopath=dendlist(dphylo,dpath)

svg("tangle_phylo_ortho.svg") #Plot the tangle gram

dphylortho %>% untangle(method='step2side') %>%
tanglegram(common_subtrees_color_branche = T,margin_inner=10,main =
paste("entanglement =", round(entanglement(phylortho),
```

Appendix:

```
2)),main_left='Phylogenomic',main_right='Orthogroups',margin_outer=3,sort=F,highlight
_distinct_edges = T, highlight_branches_lwd = T)

dev.off()

svg("tangle_path_ortho.svg")

dpathortho %>% untangle(method='step2side') %>%
tanglegram(common_subtrees_color_branche = T,margin_inner=10,main =
paste("entanglement =", round(entanglement(pathortho),
2)),main_left='Pathway',main_right='Orthogroups',margin_outer=3,sort=F,highlight_disti
nct_edges = T, highlight_branches_lwd = T)

dev.off()

svg("tangle_phylo_path.svg")

dphylopath %>% untangle(method='step2side') %>%
tanglegram(common_subtrees_color_branche = T,margin_inner=10,main =
paste("entanglement =", round(entanglement(phylopath),
2)),main_left='Phylogenomic',main_right='Pathway',margin_outer=3,sort=F,highlight_dis
tinct_edges = T, highlight_branches_lwd = T)

dev.off()

#Statistical relationships among dendrograms

all.equal(dphylortho)

all.equal(dphylopath)

all.equal(dpathortho)

cor_bakers_gamma(dphylortho)

cor_bakers_gamma(dphylopath)

cor_bakers_gamma(dpathortho)
```

8. core-genome.R

```
# This script is to find the rownames which has values in all columns

df <- read.csv("Orthogroups.csv", header=T, sep = "\t", row.names = 1,
stringsAsFactors=F)

df[df==""]<-NA # Fill the blanks with NA

df_2<-df[complete.cases(df),] # Rows that have text in all columns
```

Appendix:

```
write(rownames(df_2),file="core_orthogroups.txt")
```

9. concatenate_alignment.sh

```
#sed -e 's/$/.fa/' -i core_orthogroups.txt #bash oneliner to add .fa in each line of the list

#cat ../../phylogenomics/test/core_orthogroups.txt | xargs mv -t ../../phylogenomics/test/ #move the listed files to desired directory

#sed -i '/^>/ s/_.*//' *.fa #Remove everything after 1st "_" to make all the names same

for f in *.fa; do awk '/^>/{f=!d[$1];d[$1]=1}f' $f > $f.nodup; done # remove duplicate before concatanation

#Line 6-14 will concatanate fasta files side by side, line 15-18 will fix the alignment, header and remove blank lines

#Line 6-14 is based on a csh script from Kazutaka Katoh <katoh@biken.osaka-u.ac.jp>

for i in *nodup;
    do awk '/>/{print "\n" $0} $1!~/>/{printf $0} END{print "\n"}' $i > $i.oneline
done


cat /dev/null > result

for i in *.oneline;
    do paste result $i | sed 's/    //' >> pasted
        mv pasted result
done

rm *.oneline

#Next three lines for tiding up the header lines

sed -e 's/\(>\).*\(>\)/\1\2/'g result > result_test_1.fa

sed -i 's/>>/>/' result_test_1.fa

sed -i '/^$/d' result_test_1.fa
```

Appendix:

10. MSMC2 analysis

#From fastq download to population size visualization

#Download

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR330/007/SRR3305687/SRR3305687_1.fastq.gz

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR330/007/SRR3305687/SRR3305687_2.fastq.gz

#unzip

gunzip SRR3305687_2.fastq.gz

gunzip SRR3305687_1.fastq.gz

#filter

fastp -i SRR3305687_1.fastq -I SRR3305687_2.fastq -o SRR3305687_1.fq -O SRR3305687_2.fq

#repair

repair.sh in=SRR3305687_1.fq in2=SRR3305687_2.fq out=SRR3305687_f1.fq out2=SRR3305687_f2.fq -Xmx22g t=16

#Indexing

novoindex pfal pfal.fa

#Mapping

novoalign -f SRR3305687_f1.fq SRR3305687_f2.fq -d pfal -o SAM | samtools view -@ 16 -bS > SRR3305687.bam

sort the bam file

samtools sort SRR3305687.bam -o SRR3305687.sorted.bam -@ 16

#Creating the fasta sequence dictionary file (this script should generate a "*.dict" file

java -jar /gondor/dieunel/picard.jar CreateSequenceDictionary R= pfal.fa O= pfal.dict

#Mark and remove duplicates

java -XX:ParallelGCThreads=16 -jar /gondor/dieunel/picard.jar MarkDuplicates INPUT=SRR3305687.sorted.bam OUTPUT=SRR3305687_Nodup.sorted.bam METRICS_FILE=metrics.tx

Appendix:

t REMOVE_DUPLICATES=true VALIDATION_STRINGENCY=LENIENT

Replace all read groups in the INPUT (sorted.bam)file with a single new read group and assign all reads to this read group in the OUTPUT BAM file.

java -jar /gondor/dieunel/picard.jar AddOrReplaceReadGroups I=SRR3305687_Nodup.sorted.bam O=SRR3305687_NoDup_addrplced.bam RGID=4 RGLB=lib2 RGPL=illumina RGPU=uni

t1 RGSM=20

inedxing the output bam file

samtools index SRR3305687_NoDup_addrplced.bam

variant calling

bcftools mpileup -f pfal.fa SRR3305687_NoDup_addrplced.bam | bcftools call -c > SRR3305687.vcf

remove snps with low coverage

vcfutils.pl varFilter -Q 20 -d 5 SRR3305687.vcf > SRR3305687-vf.vcf

#stats

bcftools stats SRR3305687-vf.adr.vcf > SRR3305687-vf_adr_FilteredVCFStats.txt

#For effective population size inference, we need snps in each chromosome

bcftools view -r Pf3D7_01_v3 SRR3305687-vf.adr.vcf > chr1.vcf

#Create input snps for msmc2

/gondor/zillur/thesis/psmc/msmc-tools/generate_multihetsep.py chr1.vcf.gz > chr1.msmc.input.txt #Repeated for all 14 chromosomes

#Effective population size history

/gondor/zillur/msmc2/build/release/msmc2 -p 1*2+15*1+2*1 -o final_11 *input.txt #All chromosomes that were generated in previous step

#Plotting the msmc2 output in R

f10=read.table('/Users/mzillur/thesis/msmc/final_11.final.txt', header = T)

plot(f10$left_time_boundary/mu*g, (1/f10$lambda)/(2*mu),log='x', type = 'n', xlab='Years ago',ylab = 'Effective population size')

lines(f10$left_time_boundary/mu*g,(1/f10$lambda)/(2*mu),type='s',col='red')