University of São Paulo
Interunit Bioinformatics Graduate Program

Lyang Higa Cano

Decoding Ubiquitination in the Fight Against Malaria: A Network-Based Exploration of E1-E2-E3 Enzyme Triples in *Plasmodium falciparum*

Master in Bioinformatics
Supervisor: Prof. Dr. Ronaldo Fumio Hashimoto

During the development of this work, the author received financial assistance from Coordination for the Improvement of Higher Education Personnel (CAPES), grant #88887.639065/2021-00.

São Paulo, 2023

À minha família, meus pais e irmãos, que foram fundamentais para que este trabalho fosse possível.

Expresso minha sincera gratidão ao meu orientador, Professor Ronaldo, e à Professora Wânia, cujas valiosas contribuições enriqueceram este estudo. Agradeço também pela paciência e gentileza demonstradas ao longo desta jornada.

Estendo meus agradecimentos aos professores e demais funcionários da USP, que sempre se mostraram prestativos e apoiaram meu percurso acadêmico.

Não posso deixar de expressar minha gratidão a todos os amigos e colegas do Laboratório E-Science e do PPG em Bioinformática. Em especial, gostaria de agradecer à Irina, Clara e Maria, que, em inúmeras tardes, compartilharam comigo conversas sobre grafos e malária.

Abstract

Plasmodium falciparum is the causative agent of malaria, a disease responsible for a significant number of global deaths. Decades of integrative research, encompassing genomics, transcriptomics, cell biology, and host interactions, have been dedicated to combating this parasite. As an eukaryotic intracellular pathogen, *P. falciparum* regulates its protein activity through the ubiquitin-proteasome system (UPS), orchestrating essential cellular processes.

The UPS pathway operates through a three-step enzymatic cascade involving three distinct groups: E1, E2, and E3 enzymes. An intricate puzzle lies in the identification of enzyme triples (E1, E2, E3) that collaborate within the same chain reaction during the intraerythrocytic developmental cycle (IDC) in *P. falciparum*. This quest is significant given the incomplete understanding of this phenomenon and its potential impact on malaria control.

To address this problem, we propose an innovative approach—a Gene Co-expression Network (GCN) model for the systematic ranking of enzyme triples (E1, E2, E3). This model, based on the concept that co-expressed genes are likely involved in the same biological processes, provides an avenue to identify triples operating in tandem.

The model's efficacy was tested across seven temporal RNA-Seq transcriptome datasets, each representing distinct experimental conditions and temporal stages during the IDC. Remarkably, our model revealed three triples (E1, E2, E3) that consistently collaborated across all seven datasets, demonstrating remarkable stability amidst varying experimental contexts.

This research not only enhances our comprehension of the UPS pathway in *P. falciparum* but also sheds light on potential targets for combating malaria. By deciphering the Ubiquitin Code, we aim to unravel the mechanisms underpinning critical biological processes, ultimately contributing to the global battle against malaria.

Keywords: Ubiquitination, Gene Co-expression Network, E1-E2-E3 matching, Malaria, *Plasmodium falciparum*.

Resumo

Plasmodium falciparum é o agente causador da malária, uma doença responsável por um grande número de mortes em todo o mundo. Décadas de pesquisas integradas, abrangendo genômica, transcriptômica, biologia celular e interações com o hospedeiro, têm sido dedicadas ao combate a esse parasita. Como um patógeno intracelular eucariótico, o *P. falciparum* regula sua atividade proteica por meio do sistema ubiquitina-proteassoma (UPS), orquestrando processos celulares essenciais.

A via do UPS opera por meio de uma cascata enzimática de três etapas envolvendo três grupos distintos de enzimas: E1, E2 e E3. Um quebra-cabeça reside na identificação de tríades de enzimas (E1, E2, E3) que colaboram na mesma reação em cadeia durante o ciclo de desenvolvimento intraeritrocítico (IDC) do *P. falciparum*. Essa busca é importante devido ao entendimento incompleto desse fenômeno e seu potencial impacto no controle da malária.

Para enfrentar esse problema, propomos uma abordagem inovadora — um modelo de rede de coexpressão gênica (GCN) para a classificação sistemática de tríades de enzimas (E1, E2, E3). Esse modelo, fundamentado na ideia de que genes coexpressos provavelmente estão envolvidos nos mesmos processos biológicos, oferece uma maneira de identificar tríades que operam em conjunto.

A eficácia do modelo foi testada em sete conjuntos de dados temporais de transcriptoma de RNA-Seq , cada um representando condições experimentais e estágios temporais distintos durante o IDC. Surpreendentemente, nosso modelo revelou três tríades (E1, E2, E3) que colaboram consistentemente em todos os sete conjuntos de dados, demonstrando uma notável estabilidade em contextos experimentais variados.

Essa pesquisa não apenas aprimora nossa compreensão da via do UPS no *P. falciparum*, mas também lança luz sobre possíveis alvos para o combate à malária. Ao tentarmos decifrar o Código da Ubiquitina, visamos desvendar os mecanismos subjacentes a processos biológicos críticos, contribuindo assim para a luta global contra a malária.

Palavras-chave: Ubiquitinação, Rede de Co-expressão Gênica, Correspondência E1-E2-E3, Malária, *Plasmodium falciparum*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter serves as the foundation for our exploration of *Plasmodium falciparum* and Malaria research, specifically focusing on the critical role of Ubiquitination in cellular processes. We begin with a comprehensive review of malaria and the biology of *P. falciparum*, highlighting its impact on human health. Our discussion encompasses various aspects of this disease, including potential threats, with a special emphasis on protein degradation systems as a promising strategy for combating malaria.

Shifting our focus to ubiquitination, a post-translational modification, and the ubiquitin-proteasome system (UPS), we explore them as alternatives to target the primary protein degradation system of *P. falciparum*, crucial for the parasite's life cycle. To provide context, we initiate with a brief historical overview, offering essential insights into the discovery of ubiquitination. We trace historical milestones in ubiquitination research, revealing pivotal moments that have shaped our current understanding of this fundamental biological process.

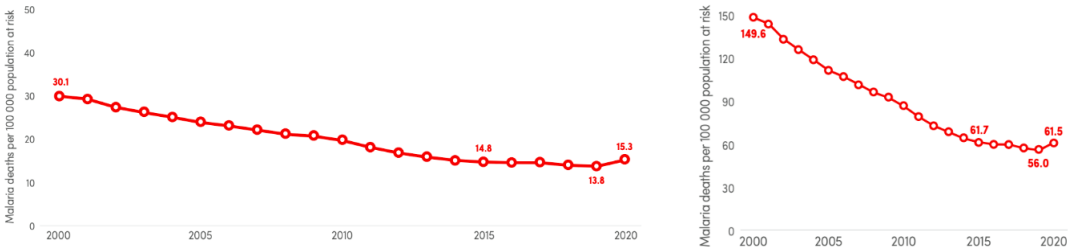
Moving from the historical context, we define the ubiquitin code and present the ubiquitin pathway. A central focus of our study emerges as we introduce the E2-E3 pairing problem, a key challenge in ubiquitination research. We offer a comprehensive discussion of this problem and provide a succinct overview of approaches employed by other researchers to tackle it.

In the final section of the Introduction (Section 1.3), we present our primary hypotheses and research objectives. We define the direction of our exploration and provide a clear vision of our goals for understanding the interplay between ubiquitination and Malaria.

1.1 Malaria

Despite dedicated efforts over the past few decades, malaria remains a global health challenge. Nearly half of the world's population resides in regions where the risk of malaria transmission persists across 85 countries [Pre21]. In the year 2020 alone, an estimated 241 million malaria cases and 627 thousand deaths were reported [Pre21]. This disease primarily afflicts impoverished tropical and subtropical areas, with Africa bearing the heaviest burden—accounting for 82% of all reported cases and 95% of the related fatalities [Wor21]. Among those most vulnerable are children under the age of five, constituting 77% of all malaria-related deaths, as well as pregnant women and individuals living with HIV/AIDS, whose compromised immune systems render them highly susceptible to this parasite [Pre21, Wor21, Tav19].

In 2020, an estimated US$3.3 billion was invested in malaria control and elimination efforts, according to the World Health Organization (WHO) [Wor21]. While funding has been



.

Figure 1.1: Malaria deaths per 100 000 population at risk. On the left: Death rates worldwide, on the right: Death rates on Africa Region. Adapted from [Wor21]

increasing annually, so has the gap between available resources and the actual requirements, reaching an estimated US$3.5 billion in 2020 [Wor21]. This substantial shortfall, which is over half of the required resources, poses a significant challenge to achieving WHO's goals for malaria control and elimination. The latest WHO Malaria Report highlights that critical milestones for the global malaria strategy in 2020 were missed, and without immediate and substantial action, the targets for 2030 may remain out of reach [Wor21].

Malaria is primarily caused by protozoans of the genus *Plasmodium*, which belongs to the larger group Apicomplexa. Among the hundreds of *Plasmodium* species, only five are known to infect humans: *P. falciparum*, *P. knowlesi*, *P. malariae*, *P. ovale*, and *P. vivax* [Sat21]. Among these, *P. vivax* and *P. falciparum* are the most prevalent, with *P. falciparum* being responsible for the majority of malaria-related deaths [CHMM16]. Our research primarily focuses on *P. falciparum*.

1.1.1 Parasite Life Cycle

These five species share a similar life cycle, which can be divided into two phases, as illustrated in Figure 1.2 for *P. falciparum*. The first phase begins when a female mosquito of the genus *Anopheles*, the vector, injects sporozoites into the human dermis during a blood meal. These sporozoites quickly migrate to hepatocytes through the bloodstream, where they undergo a process known as schizogony. This process results in the formation of thousands of new parasites called merozoites, which then re-enter the bloodstream. This initial phase is referred to as the Pre-erythrocytic phase and typically lasts for 10 days in the case of *P. falciparum*. Subsequently, these merozoites invade red blood cells (RBC), also known as erythrocytes, initiating the most critical phase of the life cycle. Inside the erythrocytes, the parasites undergo maturation, progressing through three stages: ring, trophozoite, and schizont, in chronological order. They then reproduce asexually through schizogony, resulting in the production of dozens of new merozoites and ultimately leading to the rupture of the infected red blood cells. The majority of these newly formed merozoites return to the bloodstream to initiate a new erythrocytic phase, which repeats every 48 hours. A smaller portion of merozoites undergo sexual differentiation to form gametocytes. In the case of *P. falciparum*, gametocyte development typically spans 15 days, after which the cycle restarts when a female *Anopheles* mosquito takes another blood meal [Sat21, Tav19, CHMM16].

The second phase of the parasite's life cycle begins when the mosquito feeds on blood-containing gametocytes. Once ingested, the gametocytes become activated upon reaching the

insect's midgut, leading to the formation of micro- and macrogametes. After fertilization, a diploid zygote is formed, which subsequently undergoes meiosis to produce the ookinete. The ookinete contains four haploid genomes and crosses the midgut before developing into an oocyst. Within the oocyst, numerous mitotic divisions occur, resulting in the production of a large number of sporozoites through a process known as sporogony. Eventually, the oocyst bursts, releasing sporozoites that migrate to the mosquito's salivary glands, where they become capable of infecting another human host when the mosquito takes its next blood meal [Sat21].

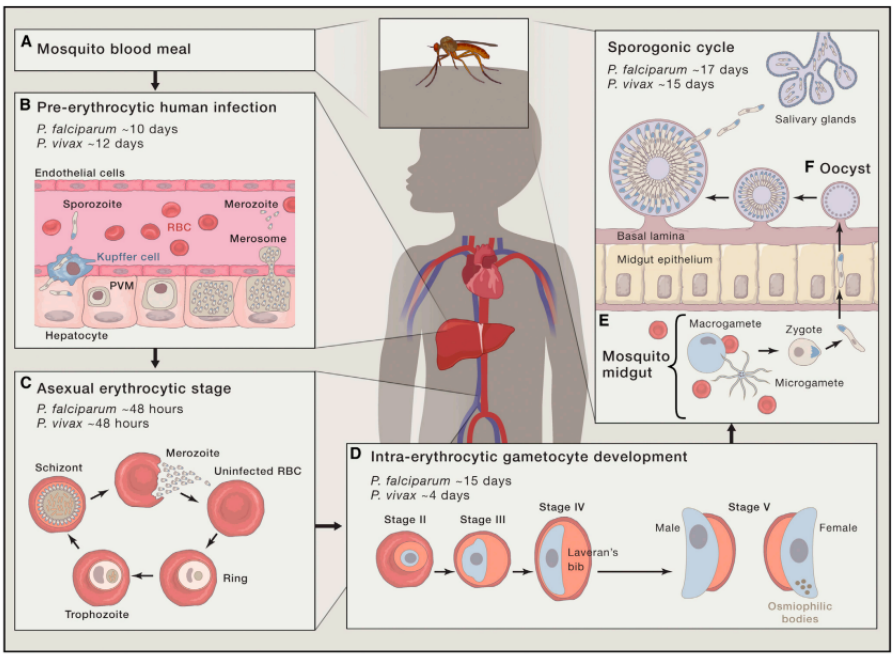


Figure 1.2: *Plasmodium falciparum* life cycle.[CHMM16]

1.1.2 Strategies to Combat Malaria

Vector control stands out as one of the most effective strategies for curbing malaria transmission and is a cornerstone of the World Health Organization's 'World Free of Malaria' initiative. The recommended methods primarily involve the deployment of insecticide-treated mosquito nets (ITNs) and indoor residual spraying (IRS). However, over the past decade, a concerning trend has emerged, with 78 out of 88 endemic countries reporting the detection of vector resistance to at least one class of insecticide [Wor21].

While the RTS,S/AS01 malaria vaccine, recently approved, showed promising results in children initially [RTS15], its efficacy waned over time, failing to meet the requirements of the Malaria Vaccine Technology Roadmap [GvS16, Mat17, vS19].

Artemisinin

In terms of treatment, malaria relies on five classes of drugs. Unfortunately, *P. falciparum* has developed at least partial resistance to all of them, with drug-resistant parasites now prevalent in endemic regions [Sat21]. Following the emergence of chloroquine resistance in the

1980s, artemisinin (ART) became the sole effective treatment for resistant parasites. Even today, the primary treatment approach heavily relies on artemisinin-based combination therapy (ACT). However, the independent emergence of partial resistance to artemisinin in various regions has raised significant global concerns [Wor21, Tav19].

Endoplasmic Reticulum Stress Response

To gain a deeper understanding of the mechanisms of ART, two fundamental concepts must be introduced: the Ubiquitin Proteasome System (UPS), which will be extensively discussed in Section 1.2.2, and the Unfolded Protein Response (UPR). These systems are crucial for responding to cellular stress induced by the accumulation of unfolded proteins [GNM+11, Tav19].

Misfolded proteins can either be refolded by chaperones or tagged with ubiquitin for subsequent degradation via the 26S proteasome. Both of these processes are regulated by the UPS [Pic01, KR12, Tav19]. On the other hand, the UPR initiates a pathway that activates three endoplasmic reticulum (ER) membrane proteins: Ire1, Atf6, and PERK. This leads to the up-regulation of transcriptional components involved in proteolytic activity, while general protein translation is reduced [GNM+11, Tav19]. The UPR in metazoans involves three ER membrane proteins, while in *Plasmodium*, it is restricted to a modified PERK pathway [GNM+11], also known as PK4, the homolog for the *Plasmodium* genus [BXC+18], as illustrated in Figure 1.3.

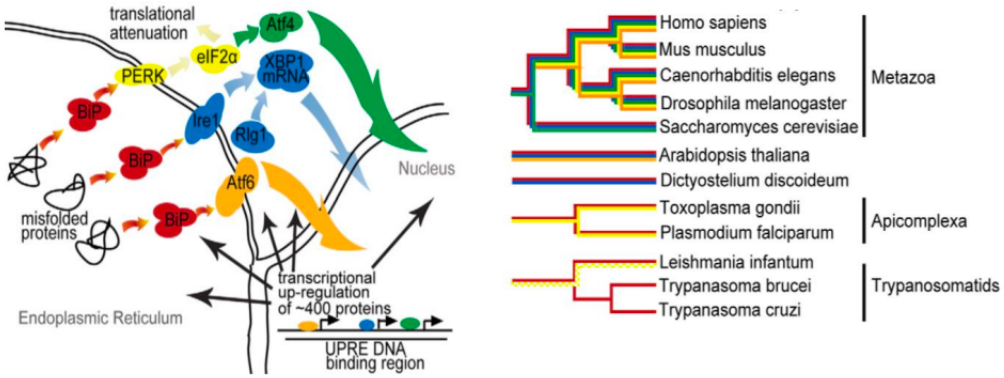


Figure 1.3: The Unfolded Protein Response (UPR) in different organisms. Proteins and branches with the same color indicate their presence in the respective organism. Adapted from [BXC+18] and [Tav19].

Upon ER stress in *Plasmodium* spp. due to the accumulation of misfolded proteins, the chaperone BiP dissociates, activating the PERK pathway. Subsequently, global protein translation is reduced as the α subunit of eIF2 is phosphorylated by the PRKR-like Endoplasmic Reticulum Kinase (PERK). This phosphorylation prevents the formation of the eukaryotic ribosome (80S complex) at the starting codon [GNM+11]. In summary, the response of *P. falciparum* to ER stress involves the refolding and degradation of misfolded proteins through the UPS and a decrease in global protein translation via the PERK pathway of the UPR, as depicted in Figure 1.4.

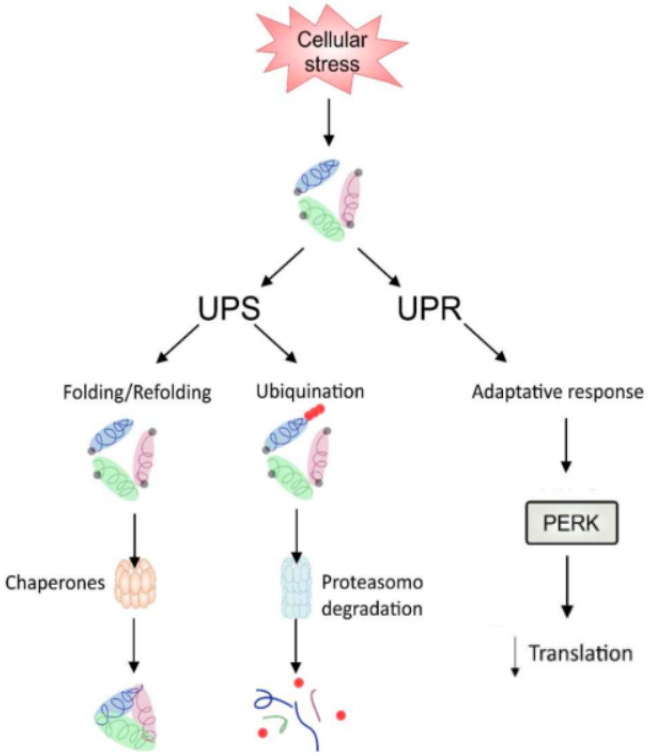


Figure 1.4: *Plasmodium* spp. response to cellular stress. The UPS refolds and degrades misfolded proteins, while the UPR attenuates global protein translation. Adapted from [Tav19].

Artemisinin Mechanism of Action and Drug Resistance

Notably, despite the critical importance of artemisinin, only in recent years have studies begun to elucidate its mechanism of action and the genes associated with partial resistance [BXC+18, XRT20]. In the study by Bridgford et al. [BXC+18], it was revealed that Dihydroartemisinin (DHA), the active metabolite of ART, not only inflicts damage on parasite proteins but also acts as a partial proteasome inhibitor. Consequently, it triggers the accumulation of polyubiquitinated and unfolded proteins, leading to ER stress. This stress, if left unresolved, can induce cell death, as shown in Figure 1.5. It is speculated that the low similarity between the UPS of *P. falciparum* and humans [PCG18] is the reason why DHA does not cause several effects at the host UPS. However, further studies on artemisinin are needed to completely understand this phenomenon.

The resistance to ART has been associated with various mutations in the β-propeller domain of the Kelch-13 protein (K13 - PF3D7_1343700) [XRT20], with over a thousand different mutations reported on the K13 gene [XRT20]. In the study by Straimer et al. [SGW+15], they demonstrated that the removal of mutations in the *P. falciparum* K13 gene through genetic modifications using zinc-finger nucleases significantly decreases the parasite survival rate when exposed to ART in vitro, from 13-49% to 0.3–2.4%, considering Cambodian isolates where drug resistance was first reported [SGW+15, BXC+18, XRT20]. Furthermore, when the researchers inserted the K13 mutations into the wild-type parasite, the survival rate increased from less than 0.6% to 2–29%.

The K13 protein is a 726-amino acid protein that shares sequence similarities with the

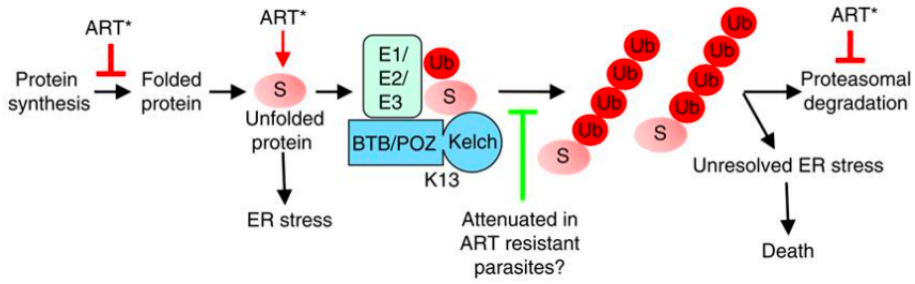


Figure 1.5: Activated Artemisinin (ART*) induces protein damage, leading to the accumulation of polyubiquitinated proteins that remain undegraded due to the partial inhibition of the proteasome by ART*. This accumulation triggers endoplasmic reticulum (ER) stress, which can ultimately result in the parasite's demise. Adapted from [BXC+18] and [Tav19].

Kelch/BTB/POZ family of E3 adaptors, which collaborates with the Cullin-3 E3 group to tag proteins with Ubiquitin [XRT20], as we will discuss in Section 1.2.2. This alignment is particularly significant given the crucial role of the UPS, as discussed earlier, where ART acts as a partial proteasome inhibitor of the 26S proteasome. Surprisingly, some unexpected interaction partners were found for the K13 protein, none of which belong to the E1-E2-E3 ubiquitin cascade [XRT20]. Therefore, further research is required to fully elucidate its precise function [XRT20].

In response to the imminent malaria crisis, Medicines for Malaria Venture (MMV), a leading authority in the battle against this disease, has placed significant emphasis on the development of innovative antimalarial therapies. This strategic focus is aimed at countering the escalating threat of drug resistance and advancing the global malaria eradication agenda.

1.1.3 Protein Degradation Systems of *P. falciparum*

Parasites belonging to the *Plasmodium* genus employ a diverse array of protein degradation systems, including the eukaryotic 26S proteasome, a prokaryotic proteasome caseinolytic protease Q (ClpQ) homolog, also known as heat shock locus V (HslV), in *P. falciparum* PfClpQ (PF3D7_1230400) is located in the parasite's mitochondria, and a ClpP protease homolog inherited from cyanobacteria, PfClpP (PF3D7_0307400) is localized in the apicoplast [NFB17, PCG18]. This extensive diversity underscores the vital role of protein degradation in the survival of these parasites, where proteolytic systems play a crucial role in the parasite's response to external stimuli, such as temperature variations and exposure to antimalarial drugs [NFB17]. Genes responsible for polyubiquitination are upregulated during these events, making the Ubiquitin pathway an attractive target for drug development [NFB17]. Inhibiting these protein degradation systems holds promise for the development of highly effective antimalarial drugs [NFB17].

In eukaryotes, about 80% of all cellular protein degradation is mediated by the 26S proteasome [CC17]. The 26S proteasome has a molecular mass of approximately 2.5 MDa [CC17], it can be found in both the cytoplasm and the nucleus of eukaryotic cells [CC17, PCG18], and is expressed throughout the lifecycle [NFB17, PCG18]. Comprising at least 32 subunits, the main component is the 20S core particle (CP), which has a mass of almost 700 kDa and is composed of 28 proteins arranged in four stacked rings. Each ring consists of seven proteins,

generating a cylindrical shape [CC17, PCG18]. The two inner rings are made up of β subunits (β1 to β7), while the two outer rings are composed of α subunits (α 1 to α 7) [CC17, PCG18], as illustrated in Figure 1.6A and B.

The proteolytic activity is guaranteed by the two inner rings, as the β1, β2, and β5 subunits exhibit caspase-like (CL), trypsin-like (TL), and chymotrypsin-like (ChTL) activities, respectively [CC17, NFB17, PCG18]. The two outer rings limit the ingress of polypeptides by controlling access to this proteolytic chamber [CC17, NFB17, PCG18]. Three different types of CPs have been identified: constitutive proteasome (cCP), immunoproteasome (iCP), and thymoproteasome (tCP). The cCP is present in all tissues, iCP in monocytes and lymphocytes, and tCP in cortical thymic epithelial cells [CC17]. The main difference between the human and plasmodial 26S proteasome lies in the CP complex, specifically in the unusual open β2 active site in *P. falciparum* CP [NFB17].

Access to CP is well-controlled to avoid the undesirable degradation of cellular proteins. Three different types of caps, which play this crucial role, have been identified: the bleomycin 10 cap (Blm10), the 11S cap, and the 19S regulatory particle (RP) [CC17], as shown in Figure 1.6 C. To allow the entry of polypeptides into CP, the cap requires a controlled opening of the α ring. Both the 11S cap and Blm10 are known to open the 26S proteasome in an ATP- and ubiquitin-independent manner [CC17]. Our interest in this work is in the 19S RP, as it is this cap that controls the degradation of polypeptides in a ubiquitin-dependent manner [CC17]. Different proteasome assemblies have been identified, as illustrated in Figure 1.6D.

The 19S RP is well-characterized, having about 900 kDa, and is responsible for recognizing and cleaving the ubiquitin chain from the polypeptide [CC17, PCG18]. The 19S RP is divided into two subcomplexes: the base and the lid, as shown in Figure 1.6 C. The base is composed of ten subunits, including six ATPases (Rpt 1 to 6), two organizing subunits (regulatory particle non-ATPase 1 and 2 - Rpn1 and Rpn2), and two ubiquitin receptors (Rpn10 and Rpn13) [CC17]. The lid is formed by nine subunits (Rpn3, 5 to 9, 11, 12, and 15), with Rpn11 being the only deubiquitination enzyme not only from the 19S RP but also from the entire 26S proteasome [CC17].

The ClpQ/HsIV and its chaperone ClpY/HsIU have been identified in *P. falciparum* (PfClpQ - PF3D7_1230400 and PfClpY - PF3D7_0907400) and are expressed in all phases of the IDC [NFB17]. Initially predicted to be located in the mitochondria, this localization was subsequently confirmed through immunofluorescence and immunoelectron microscopy using enhanced yellow fluorescent protein (EYFP) [NFB17]. PfClpQ, homologous to HsIV, is predicted to have a structure composed of two stacked rings [NFB17]. This proteasome exhibits protease-like, caspase-like, and chymotrypsin-like activities, with the latter two resembling the β1 and β5 subunits of the 26S proteasome, respectively. The AAA-ATPase chaperone PfClpY forms a complex with PfClpQ, significantly enhancing proteolytic activity [NFB17]. PfClpY plays a role in recognizing and unfolding polypeptides, similar to the function of the 19S RP in the eukaryotic proteasome [NFB17]. Since this type of protease is absent in humans, targeting this complex presents a promising opportunity for inhibition [NFB17].

The ClpP and its chaperone ClpC have also been identified in **P. falciparum** (PfClpP - PF3D7_0307400 and PfClpC - PF3D7_1406600) and are expressed during the late trophozoite and early schizont phases of IDC [NFB17]. Several studies confirmed their localization in the apicoplast [NFB17]. PfClpP primarily forms homoheptameric rings, and only a small fraction of PfClpP exists as an oligomeric complex, in both cases forming two stacked rings [NFB17]. This complex exhibits only chymotrypsin-like activity [NFB17], and again, it is the chaperone PfClpC that has the function of recognizing and unfolding polypep-

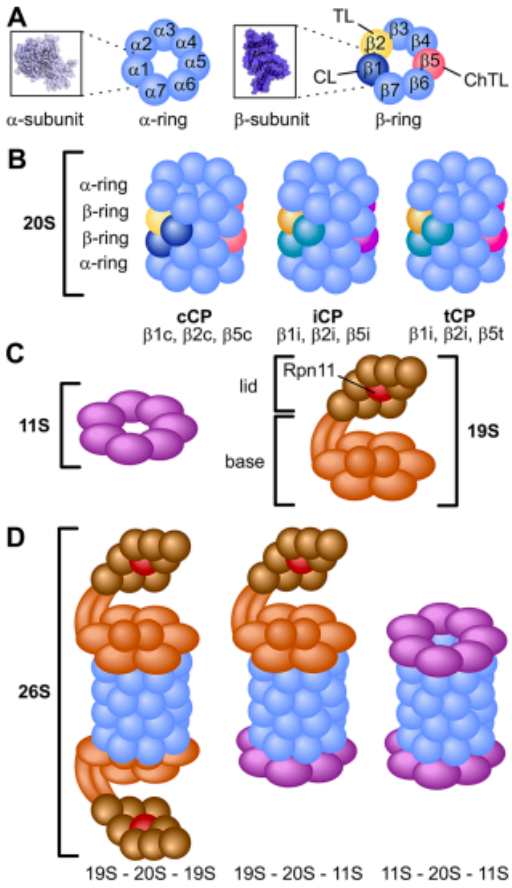


Figure 1.6: The 26S proteasome. A) α and β subunits from the outer and inner rings, respectively. The proteolytic activity is certified by $\beta 1$ (CL), $\beta 2$ (TL), and $\beta 5$ (ChTL). B) The three types of 20S: cCP, iCP, tCP, organized in four stacked rings, with the outer rings composed of seven α subunits and the inner rings of seven β subunits. C) Structure of two possible caps: 11S and 19S. D) Three different 26S proteasome assemblies have already been identified: 19S-20S-19S, 19S-20S-11S, and 11S-20S-11S. Figure adapted from [CC17].

tides [NFB17]. There is an antibiotic (acyldepsipeptide - ADEP4) that acts against *Staphylococcus aureus* ClpP; it blocks the assembly of ClpP and ClpC, preventing the chaperone from controlling access to the proteolytic chamber. Consequently, it allows the degradation of nonspecific polypeptides, resulting in the death of the bacteria [NFB17]. Further studies are needed to verify the potential use of this drug as an antimalarial.

As mentioned earlier, approximately 80% of all cellular protein degradation is mediated by the 26S proteasome [CC17]. Among the three proteasome systems, the eukaryotic proteasome is the most extensively studied for drug interventions. Various inhibitors, such as bortezomib, MLN-273, ZL3B, epoxomicin, and salinosporamides, have demonstrated remarkable efficacy in impeding the intraerythrocytic development cycle (IDC) of *P. falciparum* [LMA+13, JJWT17, PCG18]. However, the use of these compounds has also shown effects on the mammalian proteasome [NFB17, CC17], making direct application of these inhibitors less favorable. While the strategy of directly blocking the 26S proteasome is not yet effective, proteomic studies reveal that about 80% of all ubiquitin linkages in the parasite

are Lys48 [NFB17], indicating a predominant role of the ubiquitin-proteasome system (UPS), as further described in Section 1.2.2. Inhibiting ubiquitination could disrupt UPS homeostasis and impede parasite growth, given the UPS's active role throughout the parasite's life cycle [AAP12, KAB+14, NFB17, PCG18].

1.2 Ubiquitination, Ubiquitin Code and UPS

In the last section, we talked about the challenges posed by Malaria, acknowledging its continued impact as a major global health concern [Pre21, Wor21, Tav19]. We explored potential strategies to combat this disease, with a particular emphasis on degradation systems, such as the UPS. However, we faced the constraint that direct inhibition of the 26S proteasome, is not a viable approach, as it would also impact the human proteasome [NFB17, CC17].

In this section, our focus shifts to a deeper understanding of the UPS mechanism. Our objective is to unravel the intricacies of this system and identify a targeted approach to disrupt the parasite's UPS while sparing the human UPS from adverse effects. To better contextualize ubiquitination and the ubiquitin code, we will start by presenting a brief historical overview.

1.2.1 A brief history of Ubiquitination

The history of ubiquitination dates back to the 1970s when researchers began to observe the attachment of a small protein called ubiquitin to other proteins. This post-translational modification was found to mark proteins for degradation and influence their cellular fate. In the subsequent years, the work of Ciechanover, Hershko, and Rose unraveled the details of this process, elucidating the role of the proteasome—a large protein complex responsible for degrading ubiquitinated proteins [Wil05, Var06].

Their research revealed the cascade of events that lead to ubiquitination: a sequence of enzymatic reactions involving E1 (ubiquitin-activating enzyme), E2 (ubiquitin-conjugating enzyme), and E3 (ubiquitin ligase) enzymes. The E3 ligase determines the specificity of ubiquitin attachment to target proteins, allowing for precise regulation of various cellular functions. The discovery of the ubiquitin-proteasome system's pivotal role in protein turnover and cellular regulation revolutionized our understanding of cell biology and has far-reaching implications in fields such as cancer, neurodegenerative diseases, and immune responses. This recognition led to the 2004 Nobel Prize, which not only acknowledged the remarkable achievements of the laureates but also underscored the significance of the ubiquitin-proteasome system in shaping modern biology [Wil05, Var06].

The ubiquitin-proteasome system's discovery represents a remarkable convergence of intellectual curiosity, collaboration, and meticulous experimentation. Intriguing observations of energy requirements for intracellular proteolysis in mammalian cells laid the groundwork for understanding this phenomenon. The studies by Melvin Simpson in the 1950s initially highlighted this enigmatic process, while subsequent years saw limited progress in uncovering its mechanisms [Wil05, Var06].

In the early 1980s, two separate groups of researchers—Avram Hershko, Aaron Ciechanover, and Irwin A. Rose, as well as Alexander Varshavsky—concurrently made groundbreaking strides. Hershko's team embarked on a mission to elucidate intracellular proteolysis, leading to the discovery of ubiquitin as a small protein involved in targeting proteins for degradation. Varshavsky's laboratory, on the other hand, explored protein degradation and ubiquitin

conjugation. Both groups independently illuminated key aspects of the ubiquitin-proteasome system's functioning [Wil05, Var06].

The identification of E1, E2, and E3 enzymes and their roles in ubiquitin conjugation further enriched the understanding of this system. The concept of polyubiquitin chains, wherein ubiquitin molecules are covalently linked to form chains, emerged as a central theme. Additionally, the revelation of subunit selectivity in protein degradation highlighted the system's ability to dismantle specific protein subunits while sparing the rest of the complex. The implications of ubiquitination extended beyond protein degradation, with its involvement in DNA repair, cell cycle regulation, transcriptional control, stress responses, and more. This newfound knowledge expanded the paradigm of cellular regulation, demonstrating that regulated protein degradation was on par with transcription and translation in orchestrating cellular functions [Wil05].

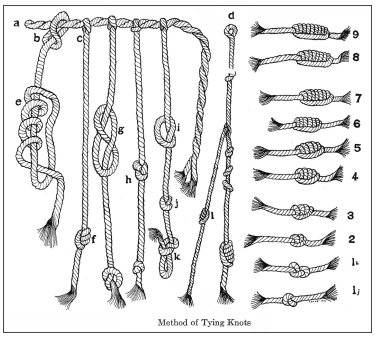
In conclusion, the discovery of ubiquitination and the ubiquitin-proteasome system represents a landmark achievement in the field of molecular biology. The collaborative efforts of scientists and their meticulous research endeavors revealed the machinery governing protein degradation and cellular regulation. This discovery's profound impact transcends traditional boundaries, influencing fields ranging from fundamental biology to medical research and drug development [Wil05, Var06].

1.2.2 Ubiquitination and The Ubiquitin Code

Ubiquitination is a post-translational modification that plays a crucial role in almost all eukaryotic cellular processes. It involves the attachment of ubiquitin, a 76-amino acid protein, to various substrates in diverse ways. These attachment methods range from single ubiquitin molecules to branched ubiquitin chains of varying topologies and sizes [Pic01, DJ09, KR12]. Each unique combination creates a distinct signaling pattern, leading to different biological outcomes. This process is orchestrated by a cascade reaction involving three enzymes [KR12].

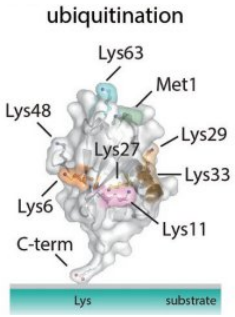
In their work, [KR12] aptly liken ubiquitination to the ancient Quipu language. The Quipu is a sophisticated system based on knots in a string, represented in Figure 1.7, where different types and combinations of knots generate distinct 'words' and 'phrases' with complex meanings. Remarkably, this language remains incompletely deciphered to this day. The case of ubiquitination, involves the attachment of ubiquitin, typically via its C-terminal glycine, preferably to a lysine residue on the substrate. Additional ubiquitins can be added, either to one of the seven lysines or the Met1 residue, as illustrated in Figure 1.8. These combinations form different chains, each corresponding to distinct biological processes, as shown in Figure 1.11. This complex ubiquitin code is present in almost all eukaryotes. Despite significant progress in recent years, the full understanding of how this code functions continues to elude us.

The ubiquitin pathway involves a series of enzymatic steps that result in the attachment of ubiquitin to target proteins. In the initial step of the ubiquitin pathway, the cysteine residue of E1 forms a thioester bond with the C-terminus of ubiquitin, an ATP-dependent process. Subsequently, ubiquitin is transferred to the E2 enzyme via another thioester bond. In the final stage, the Ub-E2 complex associates with an E3 ligase, facilitating the transfer of ubiquitin to the target protein. This results in the formation of an isopeptide bond between the C-terminal glycine of ubiquitin and, preferably, a lysine residue on the substrate or another ubiquitin molecule already attached to the chain [Pic01, DJ09]. E3 enzymes can be classified into three subgroups: Homologous to E6AP C-terminus (HECT), Really Interesting New



.

Figure 1.7: The Inca's Quipu language: examples of different types and combinations of knots. Figure from [dlC]



.

Figure 1.8: Ubiquitin structure and its seven Lysines and Met1 residues available for ubiquitination. Adapted from [SK16]

Gene (RING), and Ring-between-RING (RBR), each contributing to ubiquitin transfer in distinct ways [SS14], as illustrated in Figure 1.9. Additionally, the ubiquitin system allows for the removal of ubiquitin tags and the editing of ubiquitin chains, a role performed by deubiquitinating (Dub) enzymes [KR12].

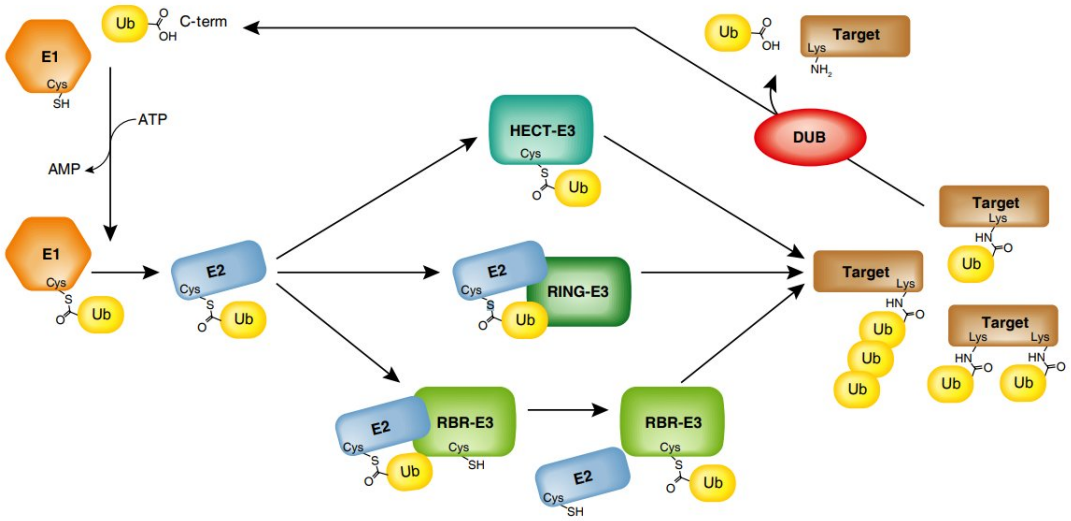


Figure 1.9: Ubiquitin Pathway: This schematic illustrates the diverse mechanisms of ubiquitin transfer to the substrate. E3 ligases are categorized into three main subgroups: HECT (Homologous to E6AP C-terminus), RING (Really Interesting New Gene), and RBR (Ring-between-RING). HECT E3 ligases form a thioester intermediate bond between their cysteine residue and ubiquitin. In contrast, RING E3 ligases, representing nearly 95% of all E3s in humans [DJ09], facilitate the direct transfer of ubiquitin from the E2 enzyme to the target protein. Some RING E3 ligases are components of protein complexes, such as CRLs (Cullin-RING Ligases), where substrate recognition is mediated by another subunit [DJ09]. RBR enzymes represent a hybrid of RING and HECT mechanisms. Figure adapted from [SS14].

Expanding on this language analogy, consider ubiquitin as the alphabet, with E1, E2, and E3 enzymes functioning as the scribes that assemble words and phrases. Deubiquitinating

(Dub) enzymes, on the other hand, take on the role of erasers, sometimes even acting as editors when working in conjunction with specific E3 ligases [KR12]. This system can be likened to grammar, with various components mixing to generate the ubiquitin code. While only a fraction of this code is currently understood, leveraging bioinformatics tools offers a promising approach to unraveling its complexities and comprehending the language of ubiquitination.

The ubiquitin protein, consisting of 76 amino acids [Pic01, DJ09, KR12], is essential in writing the ubiquitin code. Highly conserved across different species, ubiquitin exhibits only three conservative changes from yeast to humans [KR12]. In the case of *P. falciparum*, there is only one difference in ubiquitin when compared to human ubiquitin (E16 in *H. sapiens* and D16 in *P. falciparum*) [NFB17, PCG18].

Both ubiquitin and ubiquitin-like proteins (UBLs) share a common three-dimensional core structure known as the β-grasp fold [Hoc09, VDVP12, RBH16], as illustrated in Figure 1.10. Despite their common ancestry, UBLs, except for Nedd8, exhibit low similarity (less than 50%) [Hoc09, RBH16]. It is noteworthy that UBLs undergo the same E1-E2-E3 cascade reactions to modify target proteins [Hoc09].

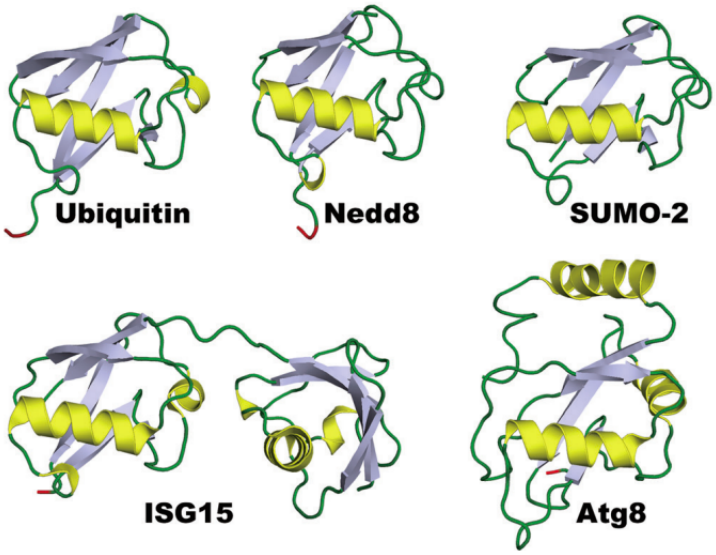


Figure 1.10: β-grasp fold structure conserved across ubiquitin and ubiquitin-like proteins.

Ubiquitin features two hydrophobic surfaces crucial for ubiquitin-binding domain (UBD) recognition. The Ile44 patch, formed by Ile44, Leu8, Val70, and His68, is bound by the 26S proteasome [KR12, RBH16]. Additionally, the Ile36 patch, consisting of Ile36, Leu71, and Leu73, mediates interactions between polyubiquitin chains, recognized by HECT E3s, Dubs, and various UBDs [KR12]. Interestingly, both patches are conserved in Nedd8 [RBH16]. An important distinction between ubiquitin and Nedd8 is the presence of Phe4, likely playing a pivotal role in protein trafficking [KR12]. This divergence enables several Dubs and UBDs to distinguish between them.

Another pivotal aspect in comprehending the ubiquitin code is the widely accepted division of labor between E3 and E2 enzymes. E3 enzymes primarily bear the responsibility for substrate selectivity, ensuring that the 'right' target protein is tagged. Conversely, E2 enzymes play a crucial role in determining the topology and size of the ubiquitin chain [DJ09, KR12, CZD22]. This orchestration, involving approximately 40 E2s and over 600 E3s in humans, transforms

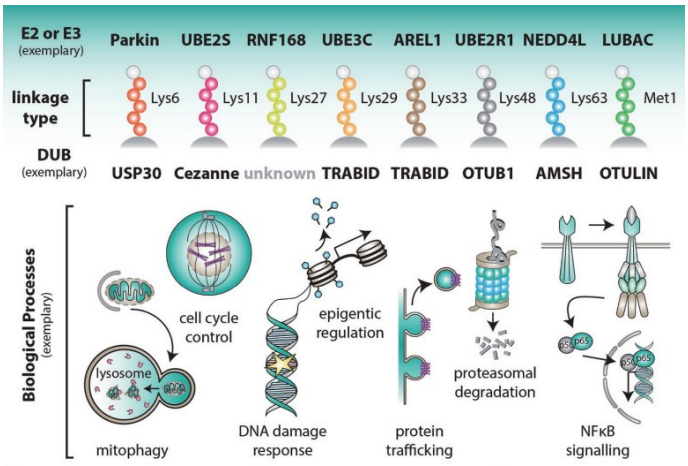


Figure 1.11: Dancing Partners of the Ubiquitin World: Insights into Ubiquitin Chain Types, Deubiquitinating Enzymes (DUBs), and Encoded Biological Processes. This figure showcases select E2 or E3 enzymes, the types of ubiquitin chains they create, the corresponding DUBs capable of disassembling these chains, and the biological processes regulated by these ubiquitin codes. Figure adapted from [SK16].

ubiquitin coding into a challenging combinatorial problem [DJ09, MKH+09, CZD22]. The complexity further escalates when considering that ubiquitin itself can undergo additional post-translational modifications, such as phosphorylation and acetylation, exponentially expanding the potential ubiquitin codes [SK16]. However, it's essential to note that this work will maintain its focus solely on the ubiquitination process.

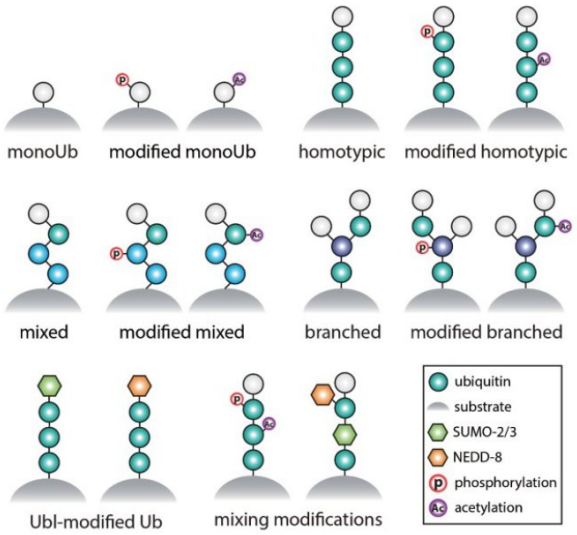


Figure 1.12: Ubiquitin Chains: Examples of possible topologies, sizes, and modifications in ubiquitin chains. Figure adapted from [SK16]

1.2.3 Importance of Ubiquitination

Much remains to be unveiled about the intricacies of the ubiquitination process, particularly concerning how diverse combinations of E1, E2, and E3 enzymes coordinate ubiquitin tagging. A central point of interest lies in the dynamic interplay between E2 and E3 enzymes, governing crucial aspects such as ubiquitin chain topology, size, and the specificity of target proteins [DJ09, KR12, CZD22]. Recent studies [MKH+09, VWDVK+09, KKNG12, DCJB+18, FP19, TZL+23] have introduced innovative approaches to predict these enzyme combinations and decipher the resulting biological outcomes. However, this remains a challenging and open problem. The implications of unraveling this complexity extend far beyond theoretical understanding. Advancements in this field hold great promise, offering potential insights and solutions for a range of diseases. For instance, in various types of cancer [CZD22], as illustrated in Figure 1.13, such insights may prove transformative. Additionally, research in this area holds relevance for combating malaria [NFB17, BXC+18], addressing neurodegenerative disorders [DB14], managing immune-related conditions [CC17], and potentially even novel drug development strategies [CC17].

Ubiquitin Proteasome System (UPS)

The Ubiquitin Proteasome System (UPS), one of the processes depicted in Figure 1.11, is a fundamental regulatory mechanism in cells, requiring a minimum Ub chain of four ubiquitins in Lys48 [DJ09]. This linkage type constitutes approximately half of all ubiquitin connections in humans [SK16] and about 80% in *P. falciparum* [NFB17]. Remarkably, the UPS is responsible for nearly 80% of all protein degradation processes in eukaryotes [CC17] and plays a pivotal role in regulating cancer progression or suppression [CZD22]. This regulation is achieved through the ubiquitination of proteins that function as either tumor promoters or suppressors, resulting in disruptions to cellular homeostasis [CZD22].

Role of E2 Enzymes in Chain Topology

As discussed earlier, E2 enzymes play a crucial role in determining the topology and size of ubiquitin chains, each corresponding to specific biological processes. Dysregulation of E2 enzymes can lead to erroneous processes, potentially contributing to tumor development. Several studies have already indicated that E2 enzymes malfunction in various types of cancer, emphasizing their potential utility as cancer biomarkers [CZD22].

A notable example illustrating how E2 enzymes determine chain size and topology is the monoubiquitination process catalyzed by the E2 enzymes Ube2W and Ube2T. When interacting with the E3 ligase FANCL, they mediate the monoubiquitination of the FANCD2 protein, coding to DNA repair [KR12]. Conversely, Ube2W can also collaborate with other E3 ligases such as Brca1-Bard1 and CHIP, generating monoubiquitin chains [KR12]. However, when these same E3 ligases interact with Ube2D instead of Ube2W, monoubiquitination is not observed [KR12]. This indicates that, at least in these cases, the E2 enzyme, specifically Ube2W, determines the chain size and topology.

HECT E3 Subfamily

In the last section, we introduced the fundamental idea—almost a rule—that E2 enzymes are responsible for determining the size and topology of the ubiquitin chain [DJ09, KR12,

E2-E3 pairing	Functional Outcomes	Disease
RING1B–BMI1 / UBE2D3	RING1B: tumor-promoter	Cervical cancer (HeLa cells)
		Stomach cancer (SGC-7901 cells)
	BMI1: tumor-promoter	
		Colon cancer (HCT116 cells)
	UBE2D3: tumor-promoter	
PUB22 / E2	PUB22 (U-box E3 ligase) mediate E2 docking	
CRL4^{CRBN}/ UBE2D3	CRL4^{CRBN}: tumor-suppressor	Multiple myeloma (HOVON-65/GMMG-HD4 trial in multiple myeloma patients)
CRL4^{CRBN}/ UBE2G1	UBE2D3: tumor-suppressor	Multiple myeloma (MM1.S cells)
	UBE2G1: tumor-suppressor	
ZNRF1 / UBE2N	ZNRF1: tumor-promoter	Non-small cell lung carcinoma (A549, H3255 cells)
		Cervical cancer (HeLa cells)
ZNRF1 / UBE2D2	UBE2N: tumor-promoter	Breast cancer
		Hepatocellular carcinoma
		Cervical carcinoma
		Melanoma
	UBE2D2: tumor-promoter	Breast cancer (MB231 cells)
E4B / UBE2D2	E4B: tumor-promoter	Breast cancer
	UBE2D2: tumor-promoter	Breast cancer (MB231 cells)
RNF8 / UBE2T	RNF8: tumor-promoter (oncogene)	Hepatocellular carcinoma (HCC) (MHCC-97H, Huh7 cells)
	UBE2T: tumor-promoter	
RNF167 / UBE2N	RNF167: tumor-promoter	Breast cancer (HCC1569, MCF7, T47D cells)
	UBE2N: tumor-promoter	Breast cancer (MB231 cells)

Figure 1.13: E2-E3 Pairs, Their Functional Roles, and Associated Cancer Types. Detailed references for each combination can be found in [CZD22].

CZD22], while E3 enzymes primarily dictate target protein specificity [DJ09, KR12, CZD22]. In most cases, this holds true. Now, we want to delve deeper into some special cases where this rule does not entirely apply.

The HECT E3 subfamily, illustrated in Figure 1.9, constitutes less than 5% of all E3 enzymes in humans [DJ09]. This subfamily is significant because HECT E3s feature a catalytic cysteine, as depicted in Figure 1.9. The E2 enzymes load this cysteine of HECT E3

before ubiquitination [DJ09, KR12, SS14]. Subsequently, the lysine from the target protein or the ubiquitin already in the polyubiquitin chain, attached to the target protein, attacks the thioester bond between the cysteine of HECT E3 and ubiquitin. Hence, the linkage specificity, which determines the topology of the ubiquitin chain, is dictated by the HECT E3 rather than the E2 [KR12].

Several examples highlight the role of HECT E3 in determining the linkage type. For instance, Rsp5 in yeast and Nedd4, a human HECT E3, are known to form Lys63 chains, while E6AP, another HECT E3, promotes Lys48 chains [KR12]. Studies have shown that Rsp5 and E6AP synthesize Lys63 and Lys48 chains, respectively, independently of their E2 partner [KR12]. This contrasts with the expectation that they would only influence the ubiquitin chain topology when interacting with nonspecific E2s like Ube2D [KR12]. The HECT E3 subfamily needs to orient and activate the acceptor Lys to determine linkage specificity [KR12]. Despite having insights into a few cases, our understanding of how HECT determines ubiquitin chain topologies remains limited [KR12].

RING E3 Subfamily

The RING subfamily, also known as U-Box, accounts for more than 95% of all E3 enzymes in humans [DJ09]. First described by Freemont et al., its canonical sequence is Cys-X_2-Cys-$X_{(9-39)}$-Cys-$X_{(1-3)}$-His-$X_{(2-3)}$-Cys-X_2-Cys-$X_{(4-48)}$-Cys-X_2-Cys [DJ09], where X represents any amino acid. While synthesizing ubiquitin chains requires modifying specific lysine residues of ubiquitin, the RING subfamily lacks a catalytic cysteine, unlike the HECT E3 subfamily [DJ09, KR12, SS14]. E3 enzymes in the RING subfamily facilitate the direct transfer of ubiquitin from the E2 enzyme to the target protein [DJ09, KR12, SS14]. Therefore, it is expected that the linkage specificity is established by the E2 enzyme, a notion supported by RING E3s capable of promoting different chain topologies depending on the E2 partner [KR12].

For instance, Brca1-Bard1 or Murf forms Lys63 linkages when paired with the Ube2N-Uev1A E2 partner but promotes Lys48 links with the Ube2K E2 enzyme [KR12]. Similarly, the CHIP E3 RING enzyme also assembles Lys63 links when interacting with Ube2N-Uev1A. However, when reacting with unspecific E2s like Ube2D, it produces unspecific linkages [KR12]. Another observation supporting this hypothesis is that RING E3s with a single E2 partner usually promote only the linkage type associated with that specific E2 [KR12].

RBR E3 Subfamily

Another important E3 subfamily is the RBR, which functions as a hybrid of HECT and RING, as illustrated in Figure 1.9. It employs the RING domain to bind with an E2, allowing the E2 to load the cysteine of the RBR E3, forming a HECT-like thioester before ubiquitination [KR12, SS14]. The RBR E3 group was initially described in 1999 by two separate groups [SS14]. It features a conserved catalytic unit composed of RING1, an in-between Ring (IBR), and RING2 domains [SS14]. In humans, only 14 RBR E3s are known [SS14].

The mechanism of ubiquitin chain formation by RBR is akin to HECT, where the E2 enzyme does not play a role in transferring ubiquitin to the target protein. Consequently, the topology of the ubiquitin chain is determined by the RBR E3 [KKNG12, SS14]. An illustrative example is the HOIP E3 RBR, which consistently promotes linear chains independent of its E2 partner [SS14], indicating that the E2 enzyme does not interfere in the topology of the ubiquitin chain in this case. In contrast, the Parkin RBR E3 interacts with multiple E2 partners, such as

Ube2A, Ube2D2, Ube2D3, Ube2L3, and Ube2L6 [SS14]. It facilitates the formation of various linkage types based on its E2 partner, underscoring the crucial role of the E2 enzyme in defining the ubiquitin chain topology in this scenario [SS14].

Ubiquitin Code Decoding: Ubiquitin-Binding Domains (UBDs)

After E2 and E3 enzymes write the ubiquitin code, ubiquitin-binding domains (UBDs) play a crucial role in decoding the information and converting it into biological processes [DWW09, KR12]. Despite the majority of ubiquitin chains being composed of Lys48 linkages, a crystal structure of ubiquitin chains with Lys48 linkages has not yet been reported. However, diubiquitin chains with Lys63 linkages have been observed in complexes with UBDs, providing insights into various ways ubiquitin chains are recognized [KR12].

One strategy for recognizing ubiquitin chains involves considering the distance between each ubiquitin molecule. This approach works because each chain type, which determines the topology of the chain, has its unique distance pattern [KR12]. Many proteins leverage this property by incorporating multiple UBDs at defined distances, each specialized to detect specific chain topologies. This is commonly observed in proteins with tandem repeats of ubiquitin-interacting motifs (UIMs) [KR12]. UIMs feature a hydrophobic α-helix that recognizes the Ile44 ubiquitin patch. An illustrative example is the Rap80 protein, which possesses two UIMs precisely separated to recognize Lys63 ubiquitin chains [KR12]. Compact Lys48 chains, on the other hand, are not recognized by this structure due to their more condensed structure [KR12]. In contrast, Ataxin-3, a deubiquitination enzyme, has two UIMs separated by a shorter distance, specializing in the recognition of compact Lys48 topologies [KR12].

Another property exploited by UBDs to recognize chain topology is the flexibility of the chain [DWW09, KR12]. Npl4-like zinc fingers (NZFs) can discriminate between Lys63 and Met1 chain topologies, which are structurally similar. Due to this property, NZFs can interact with the Lys63 topology via the Ile44 ubiquitin patch, while they are unable to interact with Met1-linkage chains due to their more rigid structure [KR12]. UBA domains present in proteasome shuttling factors likely employ this same property to recognize Lys48 topology. To interact with the Ile44 patch of ubiquitin, these UBA domains need to open the compact Lys48-linkage chains, making the recognition of the Lys48-linkage structure possible [KR12].

UBDs also employ a mechanism of combining binding sites to recognize ubiquitin chain topologies [KR12]. This is achieved by recognizing different surfaces of the ubiquitins available on the chain [KR12]. A notable example of this strategy is observed in A20, a well-known regulator of inflammatory processes, specialized in the recognition of Lys63-linkage chains. The zinc-finger domain of A20 interacts with three ubiquitin molecules, binding to the Ile44 patch of the first ubiquitin, the Tek-box of the second ubiquitin, and the surface around Asp58 of the third ubiquitin. This structure, observed only in Lys63 topologies, enables A20 to recognize this type of chain [KR12].

Theorical Models for UPS in Ubiquitin Code

In this work, our primary focus lies on the chain topology that encodes for the UPS. In recent years, the conventional view of a tetraubiquitin chain with Lys48-linkages as a consensus rule has been challenged [SK16]. New studies, particularly those with more details about the 19S RP structure, suggest that this rule may not fully explain the complexity of the UPS. For instance, certain receptors and Dubs of the proteasome do not necessarily require the

presence of four Lys48 ubiquitins to recognize a protein destined for degradation [SK16]. On the contrary, proteins like Cyclin B1, efficiently degraded by the UPS, may necessitate only short chains and not exclusively Lys48 linkages [SK16].

To address these evolving questions, a novel model has been proposed in recent years known as the 'ubiquitin threshold' model. This model posits that multiple linkage types, such as Lys48, Lys11, or even a combination of different linkage types, can contribute to the UPS code. In this model, the primary determinant of the coding mechanism is not the linkage type but rather the quantity of ubiquitin chains or branches. For each biological outcome, there exists an interval of ubiquitin chains or branches that encodes for that specific outcome [SK16]. While this model offers insights into previously unanswered questions, it simultaneously raises new inquiries, and its acceptance is not yet unanimous. In reality, our understanding of the ubiquitin code remains limited, and the complexity escalates further when considering other modifications that may occur concurrently with ubiquitination [SK16].

Remaining Gaps and Complexities

In this section, we discussed how the ubiquitin code is written and how it can be deciphered, yet certain gaps persist in our understanding, requiring further exploration of the underlying mechanisms. Regardless of the specific biochemical intricacies governing how a pair of E2-E3 adds ubiquitin molecules to a target protein or how precisely the 26S proteasome identifies tagged proteins for degradation, it is well-established that the E2-E3 pair plays a pivotal role in determining both the target protein and the ubiquitin code to be written [DJ09, KR12, CZD22]. Therefore, investigating E2-E3 pairs, elucidating their associated biological processes and discerning their protein targets is of paramount importance in addressing various diseases, including malaria [NFB17, BXC+18], understanding neurodegenerative disorders [DB14], managing immune-related conditions [CC17], unraveling the complexities of cancer evolution [CZD22], and exploring novel drug development strategies [CC17].

The landmark approval of Bortezomib in 2003 marked the advent of drugs targeting the UPS, gaining recognition for treating multiple myeloma. Subsequently, the FDA has approved additional proteasome inhibitors, signifying a significant expansion in this drug class [CC17]. Beyond their applications in cancer, proteasome inhibitors have been investigated for their potential in combating *P. falciparum* [CC17, NFB17]. Furthermore, innovative drugs designed to inhibit various UPS components in humans, such as E1s, E2s, E3s, and Dubs, have either entered clinical trials or are in pre-clinical development. Notably, KPG-818 (ClinicalTrials.gov Identifier: NCT04283097) is currently under investigation for the treatment of multiple blood cancers [CZD22]. The collective success of drugs targeting proteasomes, E1s, E2s, E3s, and Dubs underscores their immense potential for continued drug development, rendering this field of research both exciting and rapidly expanding [CC17, NFB17, CZD22].

1.2.4 Approaches to Address the E2-E3 Pairing Problem

In the preceding sections, we explored the significance of protein degradation systems for *P. falciparum* [GNM+11, Tav19], particularly highlighting the pivotal role of the 26S proteasome and, consequently, the UPS. Unfortunately, direct inhibition of the 26S proteasome has proven to be an impractical strategy, as it indiscriminately affects both the parasite and the human proteasome [NFB17, CC17]. Consequently, we turned our attention to the mechanism that governs UPS control — ubiquitination. This regulatory process involves a cascade of reactions

mediated by three key enzymes: E1, E2, and E3 [Pic01, KR12]. Recognizing the potential to disrupt this chain reaction, especially by targeting the E2-E3 pairing, emerges as a promising strategy to inhibit the parasite's UPS without causing collateral effects in the patient.

In this section, we will introduce two classical approaches to tackle the E2-E3 pairing problem. Additionally, we will initiate a discussion on our proposed methodology for addressing this challenge.

Yeast Two-Hybrid (Y2H)

The yeast two-hybrid (Y2H) technique stands as a pivotal method in the realm of protein interaction analysis. It enables the direct detection of protein-protein interactions within the confines of living yeast cells [BPL+09]. This methodology involves the examination of the interactions between two proteins, dubbed the "bait" and "prey," and hinges on the activation of reporter genes. These genes, when triggered, facilitate growth on specific media or induce a color reaction, effectively signifying successful interactions [BPL+09]. The beauty of Y2H lies in its adaptability and scalability, allowing for high-throughput investigations of protein interactions across diverse organisms, including bacteriophage T7, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and even humans [BPL+09]. This approach has played a seminal role in establishing comprehensive synthetic human interactomes and dissecting the underlying mechanisms of human diseases [BPL+09].

The origin of Y2H can be traced back to 1989 when Fields and Song introduced a groundbreaking genetic system designed to detect direct protein-protein interactions within the yeast *Saccharomyces cerevisiae* [FS89]. This innovation represents technological progress compared to prevailing practices, where protein interactions were predominantly studied via biochemical techniques [BPL+09]. The impetus behind this pioneering analytical tool stemmed from the molecular scrutiny of eukaryotic transcription factors, particularly the modular structure of Gal4, a yeast transcriptional activator. Gal4's modular nature was elucidated by the Ptashne Laboratory, which revealed that it comprised two distinct functional domains: an N-terminal DNA-binding domain (DBD) and a C-terminal transcriptional activation domain (AD) [BPL+09]. Remarkably, these domains exhibited autonomous functionality even in isolation, but when brought together in a non-covalent association, they could reconstitute a fully operational Gal4 transcription factor. Notably, when separated, these domains did not activate transcription in the presence of galactose [BPL+09]. Capitalizing on this modular property, Fields and Song devised a strategy where two proteins of interest, X and Y, were fused respectively to the DBD and AD of Gal4. This fusion allowed for X and Y to interact, ultimately reconstituting a functional transcription factor. The resulting functional transcription factor, Gal4, orchestrated the activation of reporter genes, initiating a colorimetric reaction via the lacZ reporter gene to label yeast cells [BPL+09], as illustrated in Figure 1.14.

Subsequently, the Y2H system underwent a series of refinements and expansions. The methodology was extended to encompass a range of DNA-binding proteins (e.g., LexA), transcriptional activators (e.g., Herpes simplex virus VP16), and a diverse array of reporter genes [BPL+09]. These reporter genes were chosen based on their capacity to offer straightforward readouts of interaction. Alongside the classic colorimetric reaction driven by the *lacZ* gene, researchers frequently employed auxotrophic markers (e.g., *LEU2*, *HIS3*, *ADE2*, *URA3*, *LYS2*) to facilitate growth on minimal media [BPL+09]. In contemporary Y2H setups, the simultaneous assay of multiple reporter genes has become commonplace, bolstering assay stringency and mitigating the risk of false positives stemming from indiscriminate in-

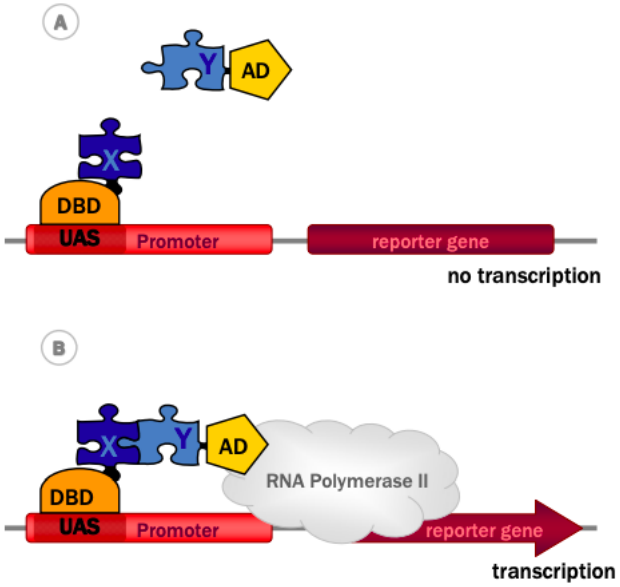


Figure 1.14: Illustration of the Yeast Two-Hybrid System. (A) The protein of interest X is fused to the DNA binding domain (DBD), creating the bait construct. The potential interacting protein Y is fused to the activation domain (AD) and referred to as the prey. (B) The bait, represented by the DBD-X fusion protein, binds to the upstream activator sequence (UAS) of the promoter. Interaction between the bait and the prey, in this case, the AD-Y fusion protein, results in the recruitment of the activation domain (AD), reconstituting a functional transcription factor. This leads to the subsequent recruitment of RNA polymerase II and initiation of transcription of a reporter gene [BPL+09]. Figure source: Bruckner et al. [BPL+09].

teractions [BPL+09]. While higher stringency has proven effective in minimizing false positives, it can also potentially obscure the detection of weak or fleeting interactions [BPL+09]. Researchers can adjust assay stringency by selectively inhibiting the enzymatic activity encoded by the reporter gene. For instance, the *HIS3* reporter's product, imidazole glycerol phosphate dehydratase, can be competitively inhibited by increasing concentrations of 3-aminotriazole [BPL+09].

Since its inception, the Y2H technique has evolved significantly, emerging as an invaluable tool for unraveling protein-protein interactions and shedding light on the intricacies of biological systems.

These studies [VWDVK+09, MKH+09] exemplify efforts to explore E2-E3 interactions using the Y2H technique. In these investigations, researchers aimed to uncover interactions between E2 enzymes (serving as the activation domain) and E3 ligases (serving as the DNA binding domain). The primary goal was to decipher the network of E2-E3 pairings, shedding light on the regulatory mechanisms governing protein ubiquitination.

The study [VWDVK+09] was interested in the E2-E3 pairing problem involving 35 human E2 enzymes and a set of 250 RING-type E3 ligases. Remarkably, this study unveiled over 300 high-quality interactions, significantly expanding our understanding of E2-E3 relationships. Moreover, [VWDVK+09] and [MKH+09] unraveled the existence of complex combinatorial interactions within this regulatory system. Notably, within the E2 enzyme family, UBE2U emerged as a pivotal player, forming interactions with multiple E3 ligases. This finding raised

intriguing questions about UBE2U's role in cellular processes and its potential implications in disease, particularly various types of cancer [VWDVK+09, CZD22].

Building on this discovery, subsequent studies (notably, [GAN+17]) delved deeper into the functional significance of UBE2U and its relevance to cancer. Utilizing RNA interference (RNAi)-based approaches, these investigations provided compelling evidence that underscored the importance of UBE2U in oncogenic processes. They validated the initial Y2H findings and established a strong link between UBE2U dysregulation and cancer pathogenesis [GAN+17, CZD22].

Co-immunoprecipitation (Co-IP)

Co-immunoprecipitation (Co-IP) is a powerful biochemical technique widely used to investigate protein-protein interactions within complex cellular environments. At its core, Co-IP relies on the selective binding of an antibody to a target protein of interest. The protein of interest is referred to as the "bait" while the interacting partner(s) are known as the "prey" proteins [NWSW+14].

In a typical Co-IP experiment, cells or tissues are lysed to release their protein content, generating a protein extract. This extract is then subjected to an immunoaffinity purification step, where an antibody specific to the bait protein is added. The antibody, now bound to the bait protein, forms an immune complex. This complex can be isolated using immobilized Protein A or Protein G, which serve as universal binding partners for antibodies. Importantly, this purification process allows for the capture of not only the bait protein but also any associated prey proteins [MBSU07].

Following the purification step, extensive washing removes non-specifically bound proteins, leaving behind only the bait-prey complexes. These complexes can then be eluted and subjected to further analysis, such as mass spectrometry, to identify the interacting proteins. Co-IP provides valuable insights into the composition of protein complexes, shedding light on the dynamic interactions that drive cellular processes, signaling pathways, and disease mechanisms. Its versatility and ability to study endogenous protein interactions make Co-IP a milestone technique in the field of molecular biology and proteomics [MBSU07].

The study [TZL+23] serves as an illustration of the utility of Co-IP in tackling the E2-E3 pairing problem. The researchers aimed to enhance our understanding of the molecular processes controlling grain yields, with the ultimate goal of improving agricultural productivity. Within this agricultural context, it is widely recognized that the UPS wields considerable influence [TZL+23]. To unravel these regulatory mechanisms, the scientists employed a multifaceted investigative approach. Initially, they harnessed genetic methodologies, leading to the identification of SGD1, a RING E3 enzyme, as a pivotal player in the regulation of grain yields [TZL+23].

However, delving further into the UPS, they recognized the critical role of E2 enzymes in orchestrating protein interactions within this system [TZL+23]. To explore these interactions comprehensively, the research team turned to the Y2H technique, which revealed SiUBC32, an E2 enzyme, as a crucial partner in this molecular regulation. To establish the biological significance of this newfound interaction, Co-IP experiments were meticulously conducted within living cells. These experiments revealed a significant finding: overexpressing these enzymes led to a 12.8% improvement in grain yields across various crops, including wheat, maize, and rice.

Gene Co-expression Network - Our Approach

In this study, our goal is to investigate potential interactions between E1, E2 and E3 enzymes of *P. falciparum* using computational methods exclusively. We plan to employ a Gene Co-expression Network (GCN) model. In simple terms, we select genes of interest for investigation. Subsequently, we apply a similarity score to their pairwise gene expression profiles for all possible gene pairs. We establish a minimum score as a threshold, and combinations exceeding this threshold are selected. The resulting network is constructed with the selected genes, where the genes represent the nodes or vertices, the pairwise combinations form the edges, and the similarity score serves as the edge weight, this is the basic idea of a GCN model [SNHL16].

This network-building approach has revealed several hypotheses to be true. Notably, genes that share the same function or participate in the same biological process tend to exhibit similar gene expression patterns, indicating co-expression [SNHL16, YMMS+21]. Thus, genes with known functions can help predict the functions of co-expressed genes with unknown functions [SNHL16, YMMS+21]. In our study, we leverage this concept, exploiting the notion that co-expressed genes are likely involved in the same biological process [WB04, SNHL16, YMMS+21], to predict potential triples of E1-E2-E3 enzymes working in the same chain reaction during the IDC of *P. falciparum*. Section 2.3 presents a formal definition of our GCN model.

A relevant precedent for this methodology can be found in the work of Williams et al. [WB04], where they employed GCN to analyze gene expression patterns in *Arabidopsis thaliana* and determine if neighboring genes are co-expressed. While exploring various characteristics of GCN, their study confirmed that, in the context of *Arabidopsis thaliana*, genes operating in the same biological process are indeed co-expressed, aligning with theoretical expectations [WB04].

1.3 Assumptions and Objectives

The Ubiquitin-Proteasome System (UPS) plays a pivotal role in regulating the Intraerythrocytic Development Cycle (IDC) of *Plasmodium falciparum*, likely involving an intricate network of regulators, with the UPS serving as a key component [AAP12, KAB+14, NFB17, PCG18]. Notably, the application of UPS inhibitors, such as bortezomib, MLN-273, ZL3B, epoxomicin, and salinosporamides, has demonstrated remarkable efficacy in impeding the IDC of *P. falciparum* [LMA+13, JJWT17, PCG18]. However, these compounds have been observed to exert effects on the human proteasome [NFB17, CC17].

Furthermore, the UPS is known for its high degree of conservation across different species [SF14, PCG18]. Notably, the UPS system in *Plasmodium* shows limited similarity to its human counterpart [PCG18]. This distinction creates an opportune avenue for developing specific, high-affinity molecules designed to inhibit UPS enzymatic activities. To achieve this, the use of precise and accurate bioinformatics tools becomes paramount, enabling the discovery of novel UPS networks and the identification of enzymes that may contribute to controlling the parasite's IDC.

Taking these factors into account, our central aim is to pinpoint groups of enzymes (E1, E2, E3) that are likely to cooperate within the same chain reaction during the IDC of *P. falciparum*, akin to the E2-E3 pairing problem discussed in previous sections. This task involves resolving an optimization problem, precisely locating triples of enzymes that maximize

a specified score, serving as an indicator of their synergistic functionality. By evaluating these trios based on their respective scores, we can readily identify the most promising candidates deserving of further investigation. Consequently, our primary focus is on establishing a reliable metric for identifying genes that collaborate. We operate under the assumption that genes exhibiting co-expression, indicating comparable expression patterns, are likely to engage in shared biological processes [WB04, SNHL16, YMMS+21]. Following this assumption, gene expression correlation emerges as a fitting metric for detecting genes with cooperative roles [SNHL16, YMMS+21].

In our study, we aim to uncover co-expressed genes among E1, E2, and E3 enzymes using an innovative Gene Co-expression Network (GCN) model. This model goes beyond our specific pathway and organism, allowing for the analysis of pathways across various organisms. By applying our GCN model, we conduct an exploration of the UPS within *Plasmodium falciparum*, promising insights into collaborative gene interactions underlying critical biological processes.

Chapter 2

Materials and Methods

In this chapter, we provide a general definition of our GCN model and outline the two essential requirements for building it—namely, the sets of genes of interest and RNA-seq datasets. We offer a brief overview of the specific datasets that will be used in this work. We also cover crucial implementation details, explaining how we calculate scores representing the probability of genes from different groups working together in the same chain reaction, along with the optional parameters used to construct the GCN model.

In this chapter, our goal is to present the most general idea of our model, using the specific problem of identifying collaborating genes from E1, E2, and E3 during the IDC of *P. falciparum* as an illustrative example. Consequently, we avoid introducing definitions that are specific to our problem. To assess the model's generalization capability, we include a computational complexity analysis in Section 2.6, which should be considered when applying this GCN model to another problem. More details about our specific problem will be provided in the next chapter (Chapter 3.1), where we apply the general ideas presented here.

2.1 Selecting Genes of Interest

Before applying the model proposed in this study (see Section 2.3), we needed to establish a set of target genes and compile pertinent information about them. In the context of our specific investigation into the UPS within *P. falciparum*, our focus was on all genes from E1, E2, and E3 gene groups.

On October 18, 2021, we accessed PlasmoDB [ABB+09], a comprehensive biological and bioinformatics database offering genomic, transcriptomic, and proteomic data for multiple Plasmodium species.

Our objective was to obtain RNAseq datasets during the IDC of *P. falciparum*. To achieve this, we conducted keyword searches related to UPS components on the PlasmoDB website, resulting in a transcriptome dataset that includes information about UPS genes expressed during the IDC. In Section 3.1, we present Table 3.1 with all the UPS-related genes we found for *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, and *Plasmodium malariae*.

It is important to note that, for this study, we will exclusively search for triples of E1, E2, and E3 enzymes; all other genes will not be used in our GCN model. However, this information can be useful for making comparisons between *Plasmodium* species and may be valuable for future research. PlasmoDB not only facilitated the identification of UPS genes but also provided access to crucial details, including their chromosomal location, cellular localization,

predicted function, and, notably, their expression patterns throughout the IDC.

Subsequently, we employed this acquired information to categorize genes into distinct groups, specifically E1, E2, and E3, aligning with our research objectives.

2.2 Datasets

To use our GCN model, users must provide a temporal RNA-seq dataset with multiple time points. Subsequently, our software can compute gene expression correlations among genes belonging to different groups, which also have to be provided by the user. A comprehensive explanation and visual demonstration of the entire pipeline will be presented in Section 3.2.

To exemplify how the dataset and the genes of interest are related and how the correlations are calculated we can use our problem as an example, therefore our genes of interest are organized in three groups: E1, E2 and E3, and we can consider the Broadbent et al., 2015 dataset which has nine time-points: [6h, 14h, 20h, 24h, 28h, 36h, 40, 44h, 48h], remembering that these time-points are the hours-post infection (hpi) of the IDC of *P. falciparum*. Therefore, we can calculate the correlations between any triple formed by one E1, one E2, and one E3, by taking the gene expression of each gene for these nine time-points of Broadbent et al., 2015 and calculating the correlation pair a pair, i.e. correlation considering these nine time-points for E1 and E2 enzymes, E1 and E3 enzymes and E2 and E3 enzymes.

For our specific case study, we employed seven distinct RNA-seq datasets Otto et al., 2010 [OWA+10], Broadbent et al., 2015 [BBR+15], Toenhake et al., 2018 [Toe], Wichers et al., 2019 [WSS+19], Subudhi et al., 2020 [SOR+20], Chappell et al., 2020 [CRO+20], Kucharski et al., 2020 [KTN+20], collected during the IDC of *P. falciparum*.

2.2.1 Otto et al., 2010

In Otto et al., 2010 [OWA+10], Illumina-based massively parallel sequencing was employed to delve into the transcriptome (RNA-Seq) of *P. falciparum* 3D7. Synchronized parasites were used to infect red blood cells, and RNA samples were collected at seven distinct time points [0h, 8h, 16h, 24h, 32h, 40h, and 48h], during the IDC from ring to mature schizonts. The RNA samples underwent Illumina sequencing, generating raw sequence data with read lengths of 37 bp and 54 bp, accessible at the European Nucleotide Archive (ENA) under study accession number ERP000069.

Gene expression data were normalized using Transcripts Per Million (TPM), and this normalized dataset for our genes of interest—comprising all 77 genes from E1, E2, and E3—was obtained from the PlasmoDB website. It's noteworthy that there are no missing values for these genes in the dataset.

2.2.2 Broadbent et al., 2015

In this study [BBR+15], two independent *P. falciparum* blood stage time courses were conducted, spanning 56 hours with a total of eleven samples. To enhance sample representation, equal ratios of samples harvested at 4 and 8 hours post-invasion were combined (referred to as T6), as were samples from 12 and 16 hours post-invasion (referred to as T14). Additionally, four samples representing late ring, early trophozoite, late trophozoite, and early schizont stages were collected around gross stage transitions. Strand-specific RNA sequencing

(RNA-seq) on an Illumina HiSeq 2000 platform generated approximately 614 million 101-bp paired-end reads.

The *P. falciparum* strain 3D7 clone was used for the time courses, cultured in human red blood cells under standard conditions. Synchronization was achieved through 5% sorbitol solution treatments, and RNA extraction involved RNeasy Midi and Mini columns, with optional on-column DNase I digestion. Strand-specific, non-polyA-selected library preparation was followed by sequencing on the Illumina HiSeq 2000 platform. The resulting gene expression data, normalized using the FPKM method, is available in [BBR+15]. The dataset covers nine time points [6h, 14h, 20h, 24h, 32h, 36h, 40h, 44h, 48h], and for these points, the dataset is complete with no missing values for our genes of interest.

2.2.3 Toenhake et al., 2018

In Otto et al., 2018 [Toe], directional RNA-Seq libraries were prepared from eight samples collected during the IDC of *P. falciparum* 3D7 parasites, the precisely time points are [5h, 10h, 15h, 20h, 25h, 30h, 35h, 40h]. Strand-specific RNA-Seq libraries were sequenced on the Illumina NextSeq 500 system, producing 75 bp single-end reads using the NextSeq500/550 HighOutput kit V2 (75 cycles) reagents (Illumina).

The identical parasite samples underwent ATAC-seq analysis to assess chromatin accessibility. The resulting data are available on PlasmoDB, and we obtained the gene expression data, already normalized using TPM. Importantly, there are no missing values for our genes of interest.

2.2.4 Wichers et al., 2019

The dataset from Wichers et al., 2019 [WSS+19], presents an RNA-seq dataset for Pf3D7 during the IDC. In a time-course experiment, three independent biological replicates of synchronized 3D7 cells were analyzed. RNA-seq samples were collected at eight developmental stages spanning from the young ring stage to the late schizont, precisely at [8h, 16h, 24h, 32h, 40h, 44h, 48h].

For the RNA-seq time course experiment, paired-end, unstranded sequencing was performed using Illumina HiSeq 4000. Synchronized 3D7 parasites served to establish a reference transcription profile. The experiment employed 100-bp paired-end RNA-seq, generating an average of 11.7 (1.9) million paired-end reads per sample in biological triplicates. This dataset offers enhanced coverage of the blood-stage transcriptome compared to previously published RNA-seq studies. Gene expression, normalized via TPM, is accessible on PlasmoDB, with no missing values for our genes of interest.

2.2.5 Subudhi et al., 2020

In Subudhi et al., 2020 [SOR+20] to investigate the 24-hour free-running transcriptome of *P. falciparum*, time-series RNA-seq experiments were conducted at a 2-hour resolution. Highly synchronized parasites were cultured at a constant temperature and in constant darkness, and samples were collected every 2 hours over 48 hours. Total RNA was isolated, and strand-specific mRNA libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit LS (Illumina). Sequencing took place on the Illumina HiSeq 4000 platform with paired-end 100/150 bp read chemistry. Subsequent steps included quality control, read trimming, and mapping to the *P. chabaudi* reference genome. Gene expression was estimated in raw

read counts using HTSeq-count. The normalized data, obtained through the TPM method, is publicly available on the PlasmoDB website, from where it was downloaded for our analysis. Importantly, there were no missing values for our genes of interest.

2.2.6 Chappell et al., 2020

In Chappell et al., 2020 [CRO+20], three *P. falciparum* strains (3D7, HB3, and IT) were cultured using standard methods, and RNA extraction used the TRIzol reagent. RNA quality and quantity were assessed with an Agilent Bioanalyzer 2100 Nano RNA chip.

For directional, amplification-free RNA-seq (DAFT-seq) library preparation, polyA+ RNA was selected with magnetic oligo-d(T) beads. The mRNA underwent directional encoding, shearing with a Covaris AFA sonicator, and subsequent library preparation steps in a "with-bead" approach. Barcoded sequencing adaptors and USER enzyme mix for second-strand digestion minimized amplification bias. Quantification with qPCR preceded sequencing on an Illumina HiSeq2000 (100bp paired-end).

Reads were mapped to version 3 of the 3D7 reference genome using TopHat2 with directional parameters, a maximum intron size of 5000nt. Gene expression data, obtained from PlasmoDB, was normalized using the TPM method. This dataset is complete, with no missing values for our genes of interest.

2.2.7 Kucharski et al., 2020

In the study by Kucharski et al., 2020 [KTN+20], a reference time course for the IDC of *P. falciparum* strain 3D7 was established. Parasites were double-synchronized using a 5% sorbitol solution, achieving approximately 6 hours of synchrony, and cultured under constant agitation. Sampling initiated at Time Point 1 (TP1), characterized by >95% early ring stage parasites (around 4 hours post-invasion). To ensure adequate mRNA for analysis, parasites were cultured in 25 individual flasks at 2% haematocrit and 8% parasitaemia for each of the 25 time points, sampled every 2 hours. Parasite development was monitored by Giemsa staining. A total of 24 time points were used for the asexual reference transcriptome, combining microarray and RNA-seq data, excluding the 9th time point due to dissimilarity, resulting in the following time points [4h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 22h, 24h, 26h, 28h, 30h, 32h, 34h, 36h, 38h, 40h, 42h, 44h, 46h, 48h, 50h, 52h].

RNA samples underwent integrity analysis with the Agilent Bioanalyzer 2100, with exclusion criteria based on the 18S-Pf/18S-Hs peak ratio and a minimum RNA Integrity Number (RIN) of 5. Complementary DNA (cDNA) samples were assessed using the DNA 12000 Kit. For RNA-sequencing, purified cDNA was used to generate sequencing libraries with the Illumina Nextera XT kit. Pooled libraries (20–24 samples per lane) were sequenced on the Illumina HiSeq4000 platform, producing 150 bp paired-end reads with 110 Gb data output per lane. The normalized gene expression data, obtained using the TPM method, is available on PlasmoDB, covering all 24 time points for all genes of interest.

2.3 Gene Co-expression Network (GCN) Model

Our proposed model is a weighted graph $G = (V, E)$, where V represents the vertices, each corresponding to a gene of interest. The set V consists of distinct, non-overlapping subsets, therefore for our problem, the groups of interest are E1, E2, and E3. Consequently, we establish

V as the union of these subsets, denoted as $V = V1 \cup V2 \cup V3$, where the subsets $V1, V2 and V3$ represents the genes of interest, i.e. E1, E2 and E3. It's essential to emphasize that these subsets do not overlap, which can be expressed as follows: $V1 \cap V2 = \emptyset$; $V1 \cap V3 = \emptyset$ and $V2 \cap V3 = \emptyset$.

E represents the set of edges within the graph. This set is composed by all possible pairs of vertices from the disjointed subsets, where $(v_i^x, v_j^y) \in E \, \forall \, x \neq y$. Here, x and y denote the specific subset ($V1, V2$, or $V3$) to which v belongs, and i and j are indices referring to the individual elements within these subsets, corresponding to genes from E1, E2, and E3.

Each edge within this graph is associated with a weight, represented by the function $w : E \to [-1, 1] \in \mathbb{R}$. In the context of our biological model, this weight serves as a measure of the likelihood that genes are involved in the same chain reaction. Our model is built upon the core hypothesis that these weights can be assessed using a correlation metric. Specifically, we adopt the Pearson correlation coefficient r for each pair of vertices, using the gene expression values from the RNA-seq datasets, from different subsets. By applying this metric, we can identify combinations with the highest degree of co-expression.

It's crucial to emphasize that we do not compute correlations among genes within the same group. Instead, our exclusive focus lies on genes originating from different groups that might collaboratively engage in the same UPS pathway. In essence, our model employs a Cartesian product approach to explore all potential combinations by selecting one gene from each group of interest. This process yields a tripartite graph, as illustrated in Figure 2.1.

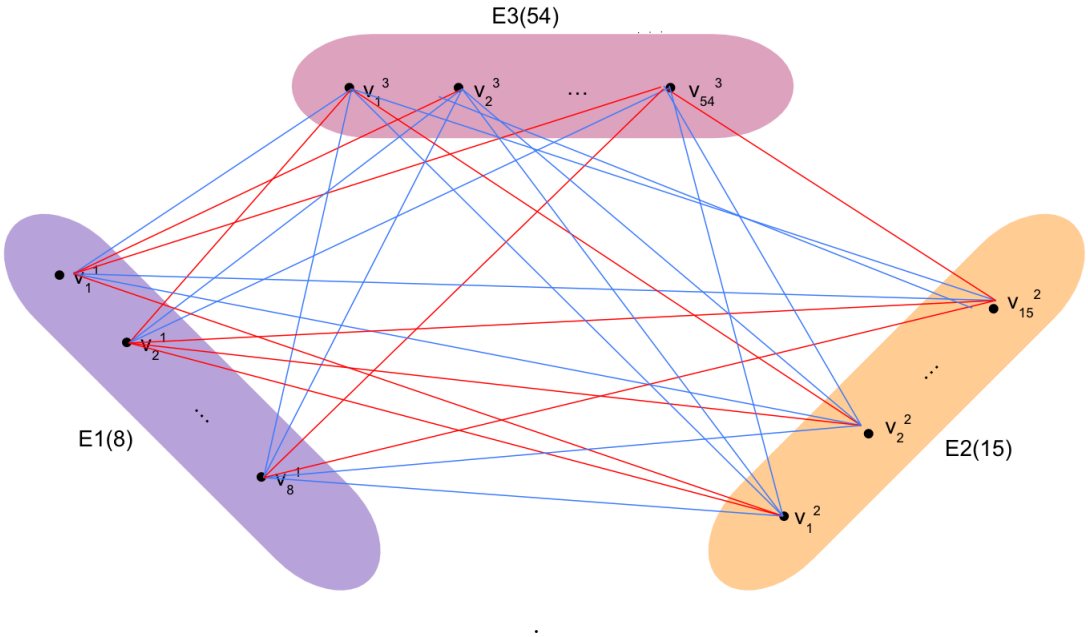


Figure 2.1: Tripartite graph generated by our model, representing E1, E2, and E3 genes, with red edges indicating negative correlations and blue edges indicating positive correlations. E1 comprises 8 genes, E2 consists of 15 genes, and E3 includes 54 genes, as detailed in Section 3.1

Using the Pearson correlation coefficient r relies on two key assumptions: firstly, that the gene expression profiles adhere to a normal distribution, and secondly, that a linear relationship exists between gene expressions [HP76]. For each computed correlation r, we derive an associated p-value, computed under the null hypothesis that the samples lack correlation and follow a normal distribution. This p-value is determined through the probability density func-

tion [VGO+20]. Subsequently, these individual p-values are combined to yield a composite p-value for the overall score. Conventional methods such as Fisher's or Pearson's, which assume independence of p-values, are not suitable for our purposes [HRD18], as the correlations $r_1(v_i^1, v_j^2)$, $r_2(v_i^1, v_j^3)$ and $r_3(v_i^2, v_j^3)$ are not independent variables. Consequently, we employ an adaptation of Brown's method [PGS+16], specifically designed to handle dependent p-values.

2.4 Defining Gene Collaboration Scores

We've established gene correlations as a metric for assessing potential collaborative interactions. However, to identify the most promising candidates likely to participate in the same chain reaction, we need a ranking system. For this purpose, we've devised two distinct scoring methods.

2.4.1 Sum Score

The Sum Score, our first scoring method, simply sums the weights of the edges. In essence, it quantifies the strength of collaboration among genes. In our tripartite graph G, each maximal clique, a cyclic path containing precisely one vertex from each group, represents a potential reaction. Cliques with the highest sum of correlations are considered the most likely reaction chains.

Mathematically, this sum is expressed as:

$$SumScore = r(v_i^1, v_j^2) + r(v_j^1, v_k^3) + r(v_k^2, v_i^3) \tag{2.4.1}$$

Where r represents the Pearson correlation function, and v_i^1, v_j^2, v_k^3 are the genes indexed by i, j and k from E1, E2 and E3 groups, respectively.

2.4.2 Geometric Mean Score

The Geometric Mean Score provides a measure of central tendency for a set of correlation values. It is defined as the nth root of the product of n correlation values, in our specific case, n = 3 since we have three correlations:

$$GeometricMean = (\prod_{z=1}^{3} r_z)^{1/3} = \sqrt[3]{r(v_i^1, v_j^2) \times r(v_j^1, v_k^3) \times r(v_k^2, v_i^3}} \tag{2.4.2}$$

Where r represents the Pearson correlation function, and v_i^1, v_j^2, v_k^3 are the genes indexed by i, j and k from E1, E2 and E3 groups, respectively.

In this case, it is important to handle negative correlations properly. Negative values are maintained as positive until the nth root is taken, and then the result is multiplied by -1 to ensure a negative final score. This penalizes negative correlations, indicating that two genes are less likely to work in the same reaction. The Geometric Mean Score also penalizes low positive correlations because even one low correlation value results in a small final score. This aligns with the hypothesis that genes working in the same reaction should be highly co-expressed.

2.4.3 Score Comparison

To illustrate the impact of these scoring methods, consider two candidate combinations for E1, E2, and E3:

1. Correlation values: 0.9, 0.9, and 0.1
2. Correlation values: 0.6, 0.6, and 0.6

For each scoring method, the rankings are as follows:

- Sum Score:
  - Combination 1: $0.9 + 0.9 + 0.1 = 1.9$
  - Combination 2: $0.6 + 0.6 + 0.6 = 1.8$
- Geometric Mean Score:
  - Combination 1: $(0.9 * 0.9 * 0.1)^{1/3} \approx 0.4327$
  - Combination 2: $(0.6 * 0.6 * 0.6)^{1/3} = 0.6$

These varied ranking outcomes illustrate the significant impact of the scoring method on the selection of candidate gene combinations. Consequently, users should carefully consider which method best suits their specific context and research goals.

2.5 Optional Parameters

Our model offers the flexibility of incorporating two optional parameters: t and β. The t parameter represents a threshold corresponding to the minimum Pearson correlation value required to establish an edge in the graph. For example, setting t to 0 will include only edges that have positive correlations as weights, disregarding all edges with negative weights. The default value is $t = -1$, which connects all possible gene pairs.

On the other hand, the β parameter operates similarly to its use in the WGCNA package [LH08]. It raises all edge values to the power of β as in $r(v_1, v_2)^\beta$, where r denotes the Pearson Correlation function, and v_1 and v_2 represent the expressions of two given genes. This parameter is valuable because, for $\beta > 1$, it introduces a penalty for smaller values. When $\beta > 1$, a small value in the range $[0, 1]$ will decay more significantly when raised to β than a larger value within the same range. This mechanism helps fine-tune the model's sensitivity to correlations.

2.6 Generalization to Other Biological Pathways

As previously discussed, this model is not limited to the UPS pathway, the *P. falciparum* organism, or any specific biological pathway. Instead, it can be easily adapted for the analysis of any biological pathway that adheres to a cascade structure, forming n-partite graphs. To create a new model, users only need to provide lists of genes for each desired group within the pathway they wish to investigate, along with an RNA-seq dataset. From this dataset,

gene expression correlations are computed. This flexibility empowers researchers to study the specific gene groups that align with their research interests.

However, it is essential to acknowledge the combinatorial nature of this approach. Constructing a model entails taking the Cartesian product of the provided gene groups to generate all feasible combinations. For each of these combinations, we calculate the Pearson correlation for every pair of genes. To enhance computational efficiency, it is advisable to employ a threshold parameter, denoted as t. This parameter can be set to a value such as $t = 0$ or any other threshold within the interval $[-1, 1]$. When t is used, any correlation value smaller than t leads to the immediate exclusion of that combination. This optimization conserves computational resources by bypassing the calculation of correlations for the remaining pairs within that combination. Moreover, the threshold aids in filtering out weak correlations, thus refining the analysis to focus on stronger gene relationships.

To measure the computational complexity, we utilize Big-O notation. In the worst-case scenario, with $t = -1$, the number of operations can be described as:

$$O(m \cdot \binom{n}{2} \cdot \prod_{i=1}^{n} |V_i|) \tag{2.6.1}$$

In this equation, m denotes the number of time points in the dataset, as the number of operations required for calculating Pearson correlations grows linearly with the number of time points. V represents a gene group, indexed by i, and $|V|$ represents the cardinality of that particular set. Finally, n corresponds to the number of gene groups included in the analysis.

In the subsequent sections 4.1.2 and 5, we will provide two illustrative examples of models extending beyond the (E1, E2, E3) scenario.

Chapter 3

Results

This chapter applies all methods introduced in the previous chapter, presenting the main results following the pipeline outlined in Figure 3.1.

We start by classifying genes related to the UPS of various *Plasmodium* species, with a focus on the E1, E2, and E3 groups from *P. falciparum*, which are our genes of interest.

Next, we demonstrate the construction of the GCN model using the list of genes of interest (presented in Section 3.1) and the datasets (previously introduced in Section 2.2).

Finally, we demonstrate the filtering of triples to extract the top results (0.7%), representing gene triples with the highest likelihood of collaborating in the same chain reaction for each dataset. Subsequently, we present the robust triples, which are the intersections of the best results from each dataset.

3.1 Classification of Ubiquitin Proteasome System (UPS) Components in *Plasmodium* Genomes

We search for the UPS genes in PlasmoDB [ABB+09] within *Plasmodium* genomes, including *Plasmodium falciparum* (Pf), *Plasmodium vivax* (Pv), *Plasmodium ovale* (Po), and *Plasmodium malariae* (Pm). Based on predicted domains and functions, UPS genes were systematically categorized into various groups, encompassing Ubiquitin (Ub)/Ubiquitin-like proteins (Ubl), Ub activating enzymes (E1), Ub conjugating enzymes (E2), Ub ligases (E3), Ub binding proteins (UBP), Deubiquitinating enzymes (DUB), and Proteasome components. Table 3.1 provides an overview of the predicted gene numbers for UPS components across the aforementioned four genomes.

P. malariae exhibits the lowest number of UPS genes among the analyzed *Plasmodium* species. This variation in UPS gene counts across different *Plasmodium* species may arise from variations in respective genome projects, leading to occasional truncation of UPS components. Moreover, these differences could reflect species-specific variations in host specificity, life cycle, or virulence.

For *P. falciparum*, we identified 8 genes from the E1 group, 15 genes from E2, and 54 genes from E3. These are the genes we will use to build our GCN model. Detailed information about these genes is available in the file `ups_net.xlsx` as described in Section 6.3.

UPS Components	Pf	Pv	Po	Pm
Ubiquitin and Ubiquitin-Like Proteins				
Ubiquitin	10	10	10	10
Ubiquitin-like	11	10	10	10
Ubiquitin Activating Enzyme (E1)				
THIF	8	8	8	8
Ubiquitin-Conjugating Enzyme (E2)				
UQ_con	15	15	15	14
Ubiquitin Ligases (E3)				
	54	54	53	51
Deubiquitinases (DUB)				
	21	21	19	19
Ubiquitin Binding Proteins (UBP)				
	2	2	2	2
Proteasome				
	43	43	43	43
Total	164	163	160	158

Table 3.1: Predicted gene numbers of UPS system components for the four analyzed genomes (Pf, Pv, Po, Pm). Pf: *Plasmodium falciparum*, Pv: *Plasmodium vivax*, Po: *Plasmodium ovale*, Pm: *Plasmodium malariae*

3.2 Utilizing the Proposed Model to Rank Genes of Interest

To implement our proposed model, we developed a comprehensive pipeline, visualized in Figure 3.1. In pursuit of robustness, we conducted a thorough evaluation, using diverse RNA-seq datasets from multiple sources. These datasets, detailed in the last Section 2 and available in the supplementary materials, in Section 6.3, all of them contain the normalized gene expression of all 77 genes of interest categorized as E1, E2, and E3 in Section 3.1. Our dataset selection encompassed studies by Otto et al., 2010 [OWA+10], Broadbent et al., 2015 [BBR+15], Toenhake et al., 2018 [Toe], Wichers et al., 2019 [WSS+19], Subudhi et al., 2020 [SOR+20], Chappell et al., 2020 [CRO+20], and Kucharski et al., 2020 [KTN+20], all focusing on the IDC of *P. falciparum*, as described in the last Section 2.

Although all datasets were already normalized, Broadbent et al., 2015 used FPKM method while all the other six used TPM, it is important to emphasize that the gene expression values in these datasets exhibit a wide range of scales. To enable meaningful comparative analysis and model construction, we conducted a series of preprocessing steps. Specifically, we performed a logarithmic base-2 (log_2) transformation and standardized all the datasets. The standardization process entailed subtracting each gene expression value by the mean expression of the corresponding gene and subsequently dividing it by the standard deviation. This approach standardizes the expression signals to a consistent scale, which is a critical step in the development of our model. This standardization is particularly valuable since our model relies on calculating correlations between pairs of gene expression values, with an emphasis on capturing similarities in expression patterns rather than absolute intensities.

Another crucial observation concerns the temporal scale represented in hours across all datasets. This temporal dimension contrasts with the rapid reactions typically witnessed

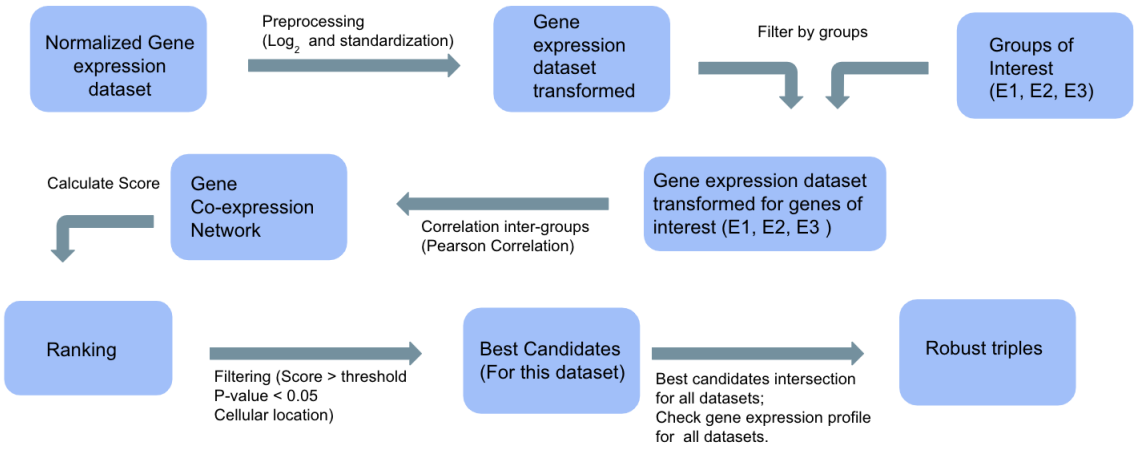


Figure 3.1: Overview of the Pipeline: The values in parentheses correspond to the parameters employed in this specific study. The pipeline begins with the selection of an RNA-seq dataset from the seven experiments chosen for this investigation. This pipeline is applied to all datasets. Data transformation occurs through a series of preprocessing steps. Subsequently, we construct the proposed Gene Co-expression Network (GCN) model, using the genes of interest (E1, E2, E3). Utilizing the resulting network, we compute Pearson correlations for each edge, which are then assigned as edge weights, forming the basis for ranking gene triples, to calculate the score we use the Sum Score, described in the Equation 2.4.1. To identify the most promising triples, we employ three filtering criteria. Subsequently, we investigate analogous triples among the top candidates from all seven datasets, assessing whether they exhibit comparable gene expression profiles. These consistently identified triples, referred to as "robust," demonstrate high scores and consistent gene expression patterns, regardless of the specific RNA-seq experiment chosen.

in the UPS pathway, which often last within seconds [DJ09]. Given this temporal discrepancy, it is reasonable to consider that these reactions occur simultaneously. Consequently, it becomes plausible to anticipate that genes from E1 and E3, participating in the same chain reaction, should demonstrate elevated co-expression at corresponding time points. With $|E1| \times |E2| \times |E3| = 8 \times 15 \times 54 = 6480$ possible combinations, our ranking encompasses 6480 positions, representing the potential reactions. To establish all conceivable edges, we can employ a threshold value of $t = -1$, as illustrated in Figure 2.1. However, our primary focus is exclusively on capturing positive correlations. Thus, we adopt $t = 0$ and $\beta = 1$, permitting only positive edges within the network.

3.3 Triple Filtering

The resulting triples were subjected to three filtering criteria: a minimum score value for each dataset, to obtain approximately the top 45 triples (0.7%), a p-value less than 0.05, and a requirement that the proteins encoded by these co-expressed genes are localized in the same subcellular compartment. The complete process is summarized in Figure 3.1. To determine the minimum sum score threshold, two criteria were employed. The first criterion was derived from the analysis depicted in Figure 3.2, illustrating scatterplots of Score against p-value for triples from Broadbent et al., 2015. As expected, an inverse relationship between the Score and p-value is evident, supported by a Spearman correlation of -0.9. Notably, all triples with a sum score ≥ 1.8 also exhibit a p-value < 0.04 (as observed in the right panel of Figure 3.2),

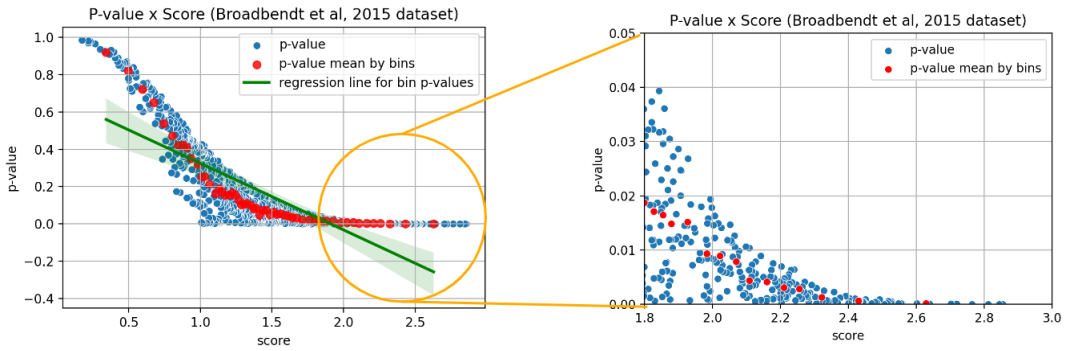


Figure 3.2: Scatterplots of P-value versus Sum Score for Broadbent et al., 2015 [BBR+15], using $t = 0$ (only positive correlations). The right panel shows data for triples with Sum Score ≥ 1.8. The red dots represent the average p-values calculated in 55 bins, while the green line in the left panel represents the best-fit linear regression on the 55 average p-values.

indicating their significance when considering a p-value threshold of 0.05. Applying the filters of sum score ≥ 1.8 and p-value ≤ 0.05, while restricting the analysis to triples of genes expressed in the same cellular location, resulted in 25 triples. This quantity of triples is amenable to manual inspection. Similar plots for all datasets are provided in Section 6.1 in Figures 6.1, 6.2, 6.3, 6.5, 6.4, and 6.6.

In addition to identifying significant triples, we aim to give priority to the top-ranked ones. To accomplish this, we set minimum sum score thresholds that yield approximately 45 triples, or 0.7% of all possible 6480 triples, a number amenable to manual scrutiny.

Applying these three filters to all datasets resulted in the identification of 25 triples for the Broadbent et al., 2015 dataset, 33 triples for the Otto et al., 2010 dataset, 35 triples for the Toenhake et al., 2018 dataset, 32 triples for the Wichers et al., 2019 dataset, 42 triples for the Subudhi et al., 2020 dataset, 34 triples for the Chappell et al., 2020 dataset, and 29 triples for the Kucharski et al., 2020 dataset. A comprehensive list of these triples, including details such as the minimum sum score threshold used, ranking, sum scores, p-values, and other relevant information, can be found in the Supplementary Material, in Section 6.3. From these finalized results for each dataset, our objective was to identify overlapping triples by seeking intersections among the outcomes. This comparative analysis revealed just three triples (E1, E2, E3) that consistently appeared across all datasets. They are:

1. (PF3D7_1333200, PF3D7_1345500, PF3D7_0319100);
2. (PF3D7_1333200, PF3D7_1345500, PF3D7_1210900);
3. (PF3D7_1333200, PF3D7_1345500, PF3D7_0303800)

Detailed results of these shared triples for each dataset are presented in Tables 3.2, 3.3, and 3.4. For visual representation, Figures 3.3, 3.4, and 3.5 display the gene expression profiles of these three triples across all seven datasets.

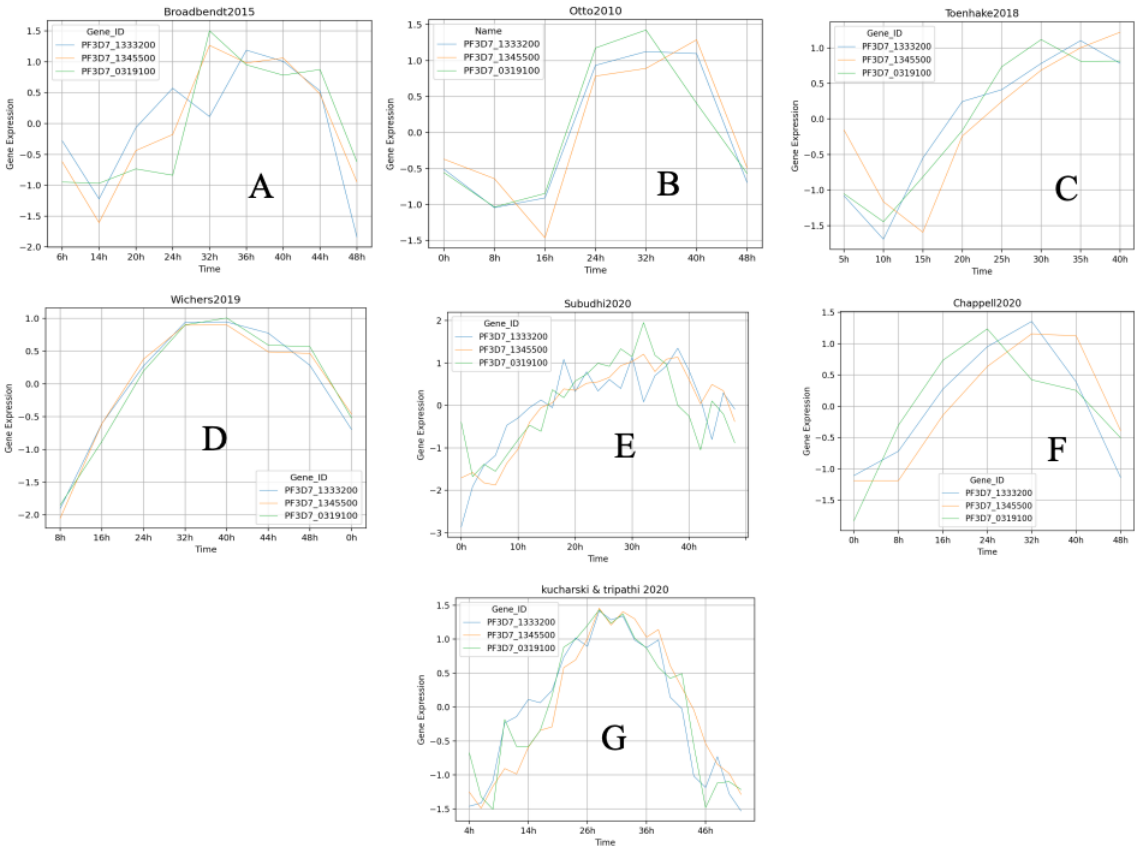


Figure 3.3: Gene expression profiles of the first robust triple for all datasets. A: Brodbendt et al., 2015 [BBR+15]; B: Otto et al., 2010 [OWA+10]; C: Toenhake et al., 2018 [Toe]; D: Wichers el al.,2019 [WSS+19]; E: Subudhi et al., 2020 [SOR+20]; F: Chappell et al., 2020 [CRO+20]; G: Kurcharski et al., 2020 [KTN+20].

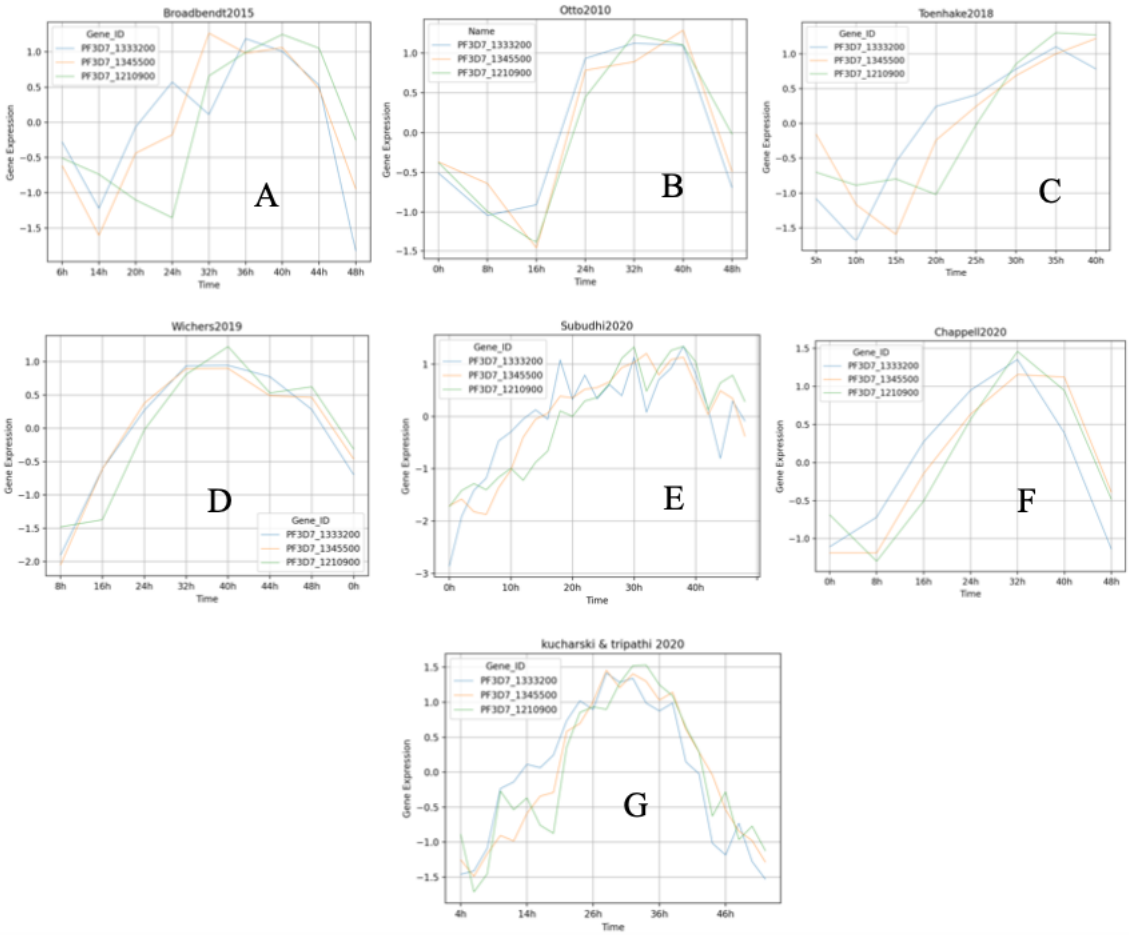


Figure 3.4: Gene expression profiles of the second robust triple for all datasets. A: Brodbendt et al., 2015 [BBR+15]; B: Otto et al., 2010 [OWA+10]; C: Toenhake et al., 2018 [Toe]; D: Wichers el al.,2019 [WSS+19]; E: Subudhi et al., 2020 [SOR+20]; F: Chappell et al., 2020 [CRO+20]; G: Kurcharski et al., 2020 [KTN+20].

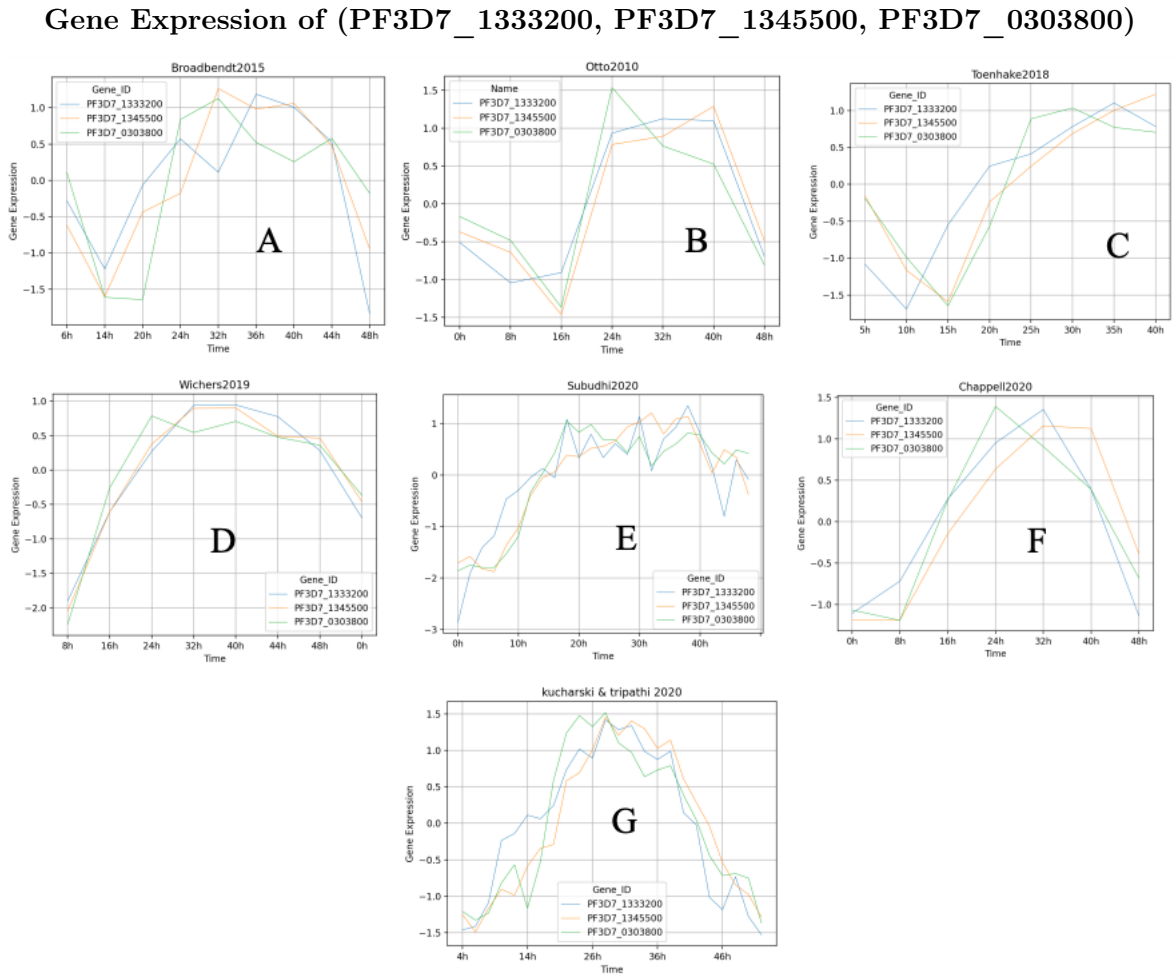


Figure 3.5: Gene expression profiles of the third robust triple for all datasets. A: Brodbendt et al., 2015 [BBR+15]; B: Otto et al., 2010 [OWA+10]; C: Toenhake et al., 2018 [Toe]; D: Wichers el al.,2019 [WSS+19]; E: Subudhi et al., 2020 [SOR+20]; F: Chappell et al., 2020 [CRO+20]; G: Kurcharski et al., 2020 [KTN+20].

Results of (PF3D7_1333200, PF3D7_1345500, PF3D7_0319100) for all datasets:

Dataset	Sum Score(Range[0;3])	Normalized Score	P-value
Broadbent et al., 2015	2.302	0.767	0.001375
Otto et al., 2010	2.743	0.914	0.000329
Toenhake et al., 2018	2.613	0.871	0.000270
Wichers et al., 2019	2.956	0.985	8.18e-08
Subudhi et al., 2020	2.263	0.754	5.75e-08
Chappell et al., 2020	2.371	0.790	0.008552
Kucharski et al., 2020	2.772	0.924	2.13e-14

Table 3.2: Results of the triple (PF3D7_1333200, PF3D7_1345500, PF3D7_0319100), the first robust triple, for all datasets.

Results of (PF3D7_1333200, PF3D7_1345500, PF3D7_1210900) for all datasets:

Dataset	Sum Score(Range[0;3])	Normalized Score	P-value
Broadbendt et al., 2015	2.050	0.683	0.006651
Otto et al., 2010	2.815	0.938	0.000137
Toenhake et al., 2018	2.469	0.823	0.001801
Wichers et al., 2019	2.807	0.936	0.000016
Subudhi et al., 2020	2.541	0.847	3.0e-10
Chappell et al., 2020	2.651	0.884	0.008552
Kucharski et al., 2020	2.732	0.910	3.527e-13

Table 3.3: Results of the triple (PF3D7_1333200, PF3D7_1345500, PF3D7_1210900), the second robust triple, for all datasets.

Results of (PF3D7_1333200, PF3D7_1345500, PF3D7_0303800) for all datasets:

Dataset	Sum Score(Range[0;3])	Normalized Score	P-value
Broadbendt et al., 2015	2.044	0.681	0.011828
Otto et al., 2010	2.712	0.904	0.000527
Toenhake et al., 2018	2.482	0.827	0.001213
Wichers et al., 2019	2.884	0.961	0.000002
Subudhi et al., 2020	2.629	0.876	1.0e-12
Chappell et al., 2020	2.674	0.891	0.000732
Kucharski et al., 2020	2.725	0.908	2.38e-12

Table 3.4: Results of the triple (PF3D7_1333200, PF3D7_1345500, PF3D7_0303800), the second robust triple, for all datasets.

Chapter 4

Discussion

4.1 Implications of Variability in Gene Expression Profiles

Upon analyzing the final results from each dataset, a prominent trend emerges: a substantial proportion of genes within subsets E1, E2, and E3 exhibit divergent gene expression profiles across the various RNA-seq experiments. Notably, in a similar vein, researchers in [WSS+19] made a pertinent observation during their exploration of the stevor gene, underscoring the significant variability in gene expression patterns across diverse experimental contexts. This variability can be attributed to a multitude of influential experimental factors, including discrepancies in experimental design parameters such as replicates, read counts, sequencing lengths, and the timing of sample collection. Furthermore, distinct culture conditions, such as variations in serum versus AlbuMAX supplementation, gas mixtures and synchronization methods, contribute significantly to this observed divergence. Additionally, the accumulation of genetic variations across different cell lines also plays a role in this variability [WSS+19].

The noteworthy variance in gene expression patterns among distinct experiments significantly impacts our study. This divergence arises from the heterogeneous nature of the input data, which forms the cornerstone of our computational model. Given the substantial disparities in these inputs, the resulting outputs naturally diverge as well. In essence, gene triples identified as potential components of the same biological reaction in one RNA-seq experiment may receive a score indicating non-cooperativity in another experiment. To explore this variability, we analyzed how the top-scoring triples from the Broadbent et al., 2015 dataset performed when the other six datasets were used as input instead.

4.1.1 Analysis of the Top Candidates from Broadbent et al., 2015

To verify how varying gene expression patterns across different experimental contexts influenced our study's outcomes, we conducted a rigorous analysis. We selected the best-ranked triple from the Broadbent et al., 2015 dataset, which comprises the following genes: (PF3D7_1225800 (UBA1), PF3D7_1033900 (2ONU), PF3D7_0826500 (UFD2)). Subsequently, we assessed how this same triple performed in the remaining six datasets, yielding the results outlined in Table 4.1.

Upon scrutinizing these results, it becomes apparent that the top-performing candidate from the Broadbent et al., 2015 dataset exhibited notably inferior performance in all other six datasets under consideration. This discrepancy strongly suggests that this particular triple

Dataset	Sum Score (Range[0;3])	Normalized Score	P-value
Broadbent et al., 2015	2.514	0.838	0.0001
Otto et al., 2010	0.574	0.191	0.22
Toenhake et al., 2018	1.086	0.362	0.01
Wichers et al., 2019	1.461	0.487	0.018
Subudhi et al., 2020	0.766	0.255	0.000017
Chappell et al., 2020	1.279	0.426	0.052
Kucharski et al., 2020	1.535	0.512	1.584e-07

Table 4.1: Results of the triple (PF3D7_1225800, PF3D7_1033900, PF3D7_0826500), which represents the top-ranked triple in the Broadbent et al., 2015 dataset, across all datasets.

likely does not participate in the same cooperative interactions in these alternative experimental contexts. we also examined the gene expression profiles of this triple across each dataset, revealing significant variations, as depicted in Figures 4.1 and 4.2.

Figure 4.1: Gene expression profile of (PF3D7_1225800 (UBA1), PF3D7_1033900 (2ONU), PF3D7_0826500 (UFD2)) for the Broadbent et al., 2015 dataset.

These results, along with the gene expression profiles, clearly demonstrate that variations in gene expression data, which serve as inputs to our model, yield distinct outcomes for each dataset. To assess the influence of experimental context on our model's predictions, we compared the results of the top 25 candidate triples from the Broadbent et al., 2015 dataset, when the other six datasets were used as input.

To gain a comprehensive view of the results obtained from the top 25 candidates in the Broadbent et al., 2015 dataset, we constructed a graph as described in Section 2.3. In this graph, vertices represent the genes from these 25 triples, while edges signify predicted gene interactions, with edge weights indicating the correlation (r). The resulting graph is illustrated in Figure 4.3.

Upon visualizing the graph depicted in Figure 4.3, a distinct pattern emerges: the graph is disjointed, with one cluster of genes, named Cluster 1, exhibiting low expression profiles during the trophozoite stage, and another cluster, Cluster 2, showing high expression profiles in the mature parasite stage (after 32 hours post-infection). This pattern is further confirmed in Figure 4.4.

These significant findings align with our hypothesis of the UPS playing a pivotal role in regulating the IDC of *P. falciparum*, as the expression profiles of these enzymes correlate with

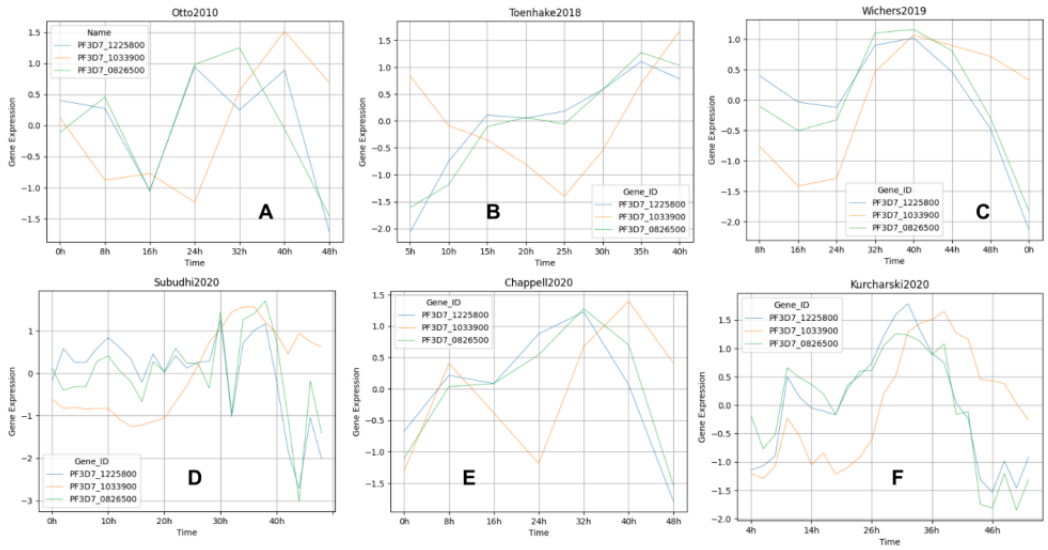


Figure 4.2: Gene expression profiles of the best scoring triple from Broadbent et al., 2015 across various datasets. A: Otto et al., 2010; B: Toenhake et al., 2018; C: Wichers et al., 2019; D: Subudhi et al., 2020; E: Chappell et al., 2020; F: Kucharski et al., 2020.

the progression of the IDC. However, when we examine the graphs generated by the same 25 triples and their corresponding gene expression profiles (available in Section 6.2), using the other six datasets as input, an entirely different outcome emerges. This once again emphasizes that our software effectively predicts potential gene interactions, but these predictions are linked to the specific context of individual RNA-seq experiments. This constraint arises from the inherent variability in gene expression profiles, leading to distinct outcomes across experiments.

4.1.2 Exploring Genes within the Tubulin Complex

To assess the predictive capabilities of the GCN model and gain a more profound understanding of the observed variability in gene expression across diverse experimental contexts, we turned our attention to PF3D7_0903700 (Alpha tubulin 1) and PF3D7_1008700 (Tubulin beta chain). These two genes are widely recognized for their indispensable roles within the Tubulin complex, an integral component of microtubules with crucial functions in the *P. falciparum* life cycle [SFK+19, HFK+22]. Unlike our previous focus on predicting potential gene interactions, these two genes are well-established for their cooperative action in the same biological reaction, making them ideal candidates for use as a positive control.

Unexpected Insights

Despite consistently achieving notably high scores and reporting minuscule p-values across six of the seven datasets (Table 4.2), these widely acknowledged gene pairs exhibit diverse gene expression profiles across the various datasets. For instance, around the 20-hour post-infection (20hpi) time frame, while certain experiments depict elevated expression levels for

both genes, others display a pattern of reduced expression, as visualized in Figure 4.5. This unexpected finding challenges conventional assumptions about the uniformity of gene expression and underscores the predictive prowess of our GCN model in identifying potentially cooperating genes. Simultaneously, it emphasizes the critical importance of cautious interpretation when comparing outcomes across different experimental contexts, highlighting that while predictions hold within a specific experimental setting, establishing universal rules remains a complex undertaking.

Dataset	Sum Score (Range[0;1])	P-value
Broadbent et al., 2015	0.996	1.6e-5
Otto et al., 2010	0.960	0.012
Toenhake et al., 2018	0.996	6e-5
Wichers et al., 2019	0.997	5.4e-5
Subudhi et al., 2020	0.992	1.3e-13
Chappell et al., 2020	0.729	0.246
Kucharski et al., 2020	0.987	1.3e-11

Table 4.2: Results for the gene pair PF3D7_0903700 (Alpha tubulin 1) and PF3D7_1008700 (Tubulin beta chain) across all datasets.

4.1.3 The Robust Triples

In light of our recognition of the significant influence of experimental inputs, our investigation was geared towards identifying triples that consistently ranked at the top across all seven datasets. Remarkably, only three triples emerged as consistent performers in the outcomes of all seven datasets, and their detailed results are provided in Tables 3.2, 3.3, and 3.4. Upon scrutinizing their gene expression profiles across a wide array of experiments, as exemplified in Figures 3.3, 3.4, and 3.5, we found a new pattern. Beyond merely achieving high scores across all datasets, these triples showcased similar gene expression trends consistently, although not precisely identical at every moment, demonstrating unwavering behavior within each phase of the parasite life cycle (ring, trophozoite, and schizont). This consistency implies that these genes possess robustness, consistently yielding congruent results and gene expression patterns, regardless of the specific RNA-seq experiment chosen.

Our scrutiny unveiled that E1 and E2 enzymes remained constant across these triples, while E3 enzymes exhibited variations among them. It is widely acknowledged that, in eukaryotes, E2 enzymes, despite their abundance, exhibit less diversity in comparison to the more diversified E3 enzymes [RDB+03]. Remarkably, this pattern is mirrored in *P. falciparum*. Within the three consistently robust triples, the E1 enzyme (PF3D7_1333200/PF13_0182) and the E2 conjugating enzyme (PF3D7_1345500) remained unchanged. However, the E3 enzyme exhibited variability: in the first triple, we identified the E3 ubiquitin-protein ligase RBX1 (PF3D7_0319100); in the second triple, the GPI mannosyltransferase 1 (PF3D7_1210900), a Cullin-RING E3 ubiquitin ligase; and in the third triple, the IBR domain protein, putative (PF3D7_0303800).

Since comprehensive molecular and functional characterizations of E1, E2, and E3 ubiquitin enzymes in *P. falciparum* remains elusive, our knowledge of their activities and interactions with partner molecules remains limited. Consequently, our primary focus revolves around un-

covering the potential functional activities within the highest-ranking robust triple identified in this study. Notably, we place particular emphasis on analyzing the Broadbent et al., 2015 dataset, a fundamental cornerstone of our research. Within this context, we concentrate our investigation on the first robust triple, distinguished as the top performer in the Broadbent et al., 2015 dataset.

The Pfsuba1 protein, encoded by the PF3D7_1225800 gene, plays a pivotal role as a ubiquitin-activating enzyme within the ubiquitin pathway of *P. falciparum*, and it stands out as one of the most extensively studied members of the E1 group in this species [SHM+09]. In contrast, the uba1 protein, encoded by the PF3D7_1333200 gene and prominently featured in the first robust triple, achieves the highest score but remains an area of limited exploration. An alignment analysis, available in Section 6.3, revealed that the amino acid sequences of the two uba1 proteins, XP_001350655.1 and XP_001350063.1, encoded respectively by the PF3D7_1225800 and PF3D7_1333200 transcripts, exhibit low similarity. While variations in specific E1 enzymes may exist among different species, the fundamental function and structure of E1 enzymes remain conserved. These enzymes share common features and mechanisms of action, including the adenylation of ubiquitin and the formation of a thioester bond with the activated ubiquitin molecule [ZIC14].

In our research, we identified two conserved ThiF domains in uba1/PF3D7_1333200, which were also detected in Pfsuba1. This confirmation was made through a Pfam analysis accessible via the Swiss Browser, as depicted in Figure 4.6. The ThiF domain is a NAD/FAD-binding fold commonly found in ubiquitin-activating E1 family members and bacterial ThiF/MoeB/HesA family proteins [LWRS01, LL05, LS08]. Notably, Pfsuba1, encoded with features like a ubiquitin-activating enzyme active site, two ubiquitin-activating enzyme catalytic domains, two ThiF repeats, and a catalytic cysteine at the N-terminus, align with the characteristics of this enzyme [SHM+09, ACP+13].

The E2 ubiquitin-carrying enzyme identified in our analysis is PF3D7_1345500, also known as UBC or 2H2Y. This UBC protein serves as the E2 conjugation enzyme and is associated with a family of specific ERAD-like proteins within the parasite, much like the E1 enzyme [SHM+09, ACP+13]. These ERAD-like proteins not only feature essential ubiquitination domains but also possess a signal peptide responsible for directing them to the apicoplast. The apicoplast, arising from a secondary endosymbiosis event involving a red alga, is a non-photosynthetic plastid characterized by its attachment to four membranes. It plays a crucial role in fatty acid metabolism and isoprenoid biosynthesis in Apicomplexa parasites, such as *Plasmodium* and *Toxoplasma* [FSH+05, ACP+13, FCAS17].

Beyond its role in targeting the apicoplast, the E2 enzyme, once ubiquitin-conjugated, interacts with the E3 ubiquitin ligase and contributes to the UPS pathway. During High-Throughput Mass Spectrometry (HMM) research on Apicomplexan parasites, proteins associated with a parasite-specific ERAD-like system encompass all the essential ubiquitination enzymes required for activation (PF3D7_1333200 and PF3D7_1365400), conjugation (PF3D7_1345500), binding (PF3D7_0316900 and PF3D7_0312100), and deconjugation (PF3D7_1031400) [ACP+13, FCAS17].

E3 ubiquitin ligases play a crucial role in the UPS, ensuring a high level of specificity and selectivity when targeting substrates within cells [DJ09, KR12, CZD22]. In the *P. falciparum* 3D7 strain, the PF3D7_0319100 transcript encodes a putative E3 ubiquitin-protein ligase known as RBX1. It is characterized by its catalytic RING domain, a vital component in the ubiquitination process. In higher eukaryotes, the E3 ubiquitin ligase RBX1 is an integral part of the tetrameric E3 ubiquitin ligase complex referred to as SCF (Skp1-Cullin1-F-box protein).

This complex comprises four essential components: RBX1 (or RBX2), SKP1, an F-box protein, and a cullin [DJ09].

Recent research has illuminated the functional diversity of E3 ubiquitin ligases in *P. falciparum*. Two distinct types have been identified: the conventional SCF-like E3 ubiquitin ligase, named PfSCF, and a human CRL4-like E3 ubiquitin ligase, known as PfCRL4 [RRG+23]. PfSCF is thought to consist of several key components based on *in vitro* experiments, including PfCullin-1, PfSkp1, PfRbx1, PfFBXO1, and PfCacyBP. In contrast, PfCRL4 comprises PfCullin-2, PfCPSF_A, two WD40 repeat proteins, and PfRbx1. PfCRL4 is known to play significant roles in cell division, maintaining membrane integrity, and is believed to be essential for the proper function of cellular organelles like mitochondria and the endoplasmic reticulum (ER) [RRG+23].

Nonetheless, the precise function of the E3 ubiquitin-protein ligase RBX1, specifically in the context of *P. falciparum*, remains a topic of ongoing investigation. Recent studies, such as one by Rizvi and colleagues [RRG+23], provide evidence of interactions between the cullin protein and PfRbx1 (PF3D7_0319100) during the trophozoite stage in the D10 strain of *P. falciparum*. These findings suggest that PfRbx1 is expressed in various cellular compartments, including the cytoplasm, nucleus, and chromatin, further emphasizing its potential significance. As depicted in Figure 3.3, the predicted results indicate an increase in the expression of E1, E2, and E3 transcripts approximately around the 16-hour mark of the intraerythrocytic cycle, with elevated expression persisting until approximately 46 hours. However, it's worth noting that some variability exists based on the specific transcriptome dataset under consideration. This evidence of PfRbx1 expression in trophozoites leads us to speculate that transcript expression is particularly pronounced during the early and late trophozoite stages, as well as during the schizont stage, suggesting a potential role for PfRbx1 in regulating critical cellular processes during these phases.

In mammals, it is well-established that this enzyme exhibits a wide tissue distribution and plays a pivotal role in cell survival and division. Remarkably, RBX1 and RBX2 have undergone extensive scrutiny in the context of anti-cancer therapies. Inhibiting these proteins has demonstrated the capacity to induce apoptosis and cellular senescence, while their overexpression directly correlates with the proliferation of tumor cells [JSS09, SFJ+22]. Although the analysis in question identified only the RBX1 transcript, it is plausible to surmise that enzymes such as E1, E2, and notably E3 ligases also assume crucial functions in the parasite's survival by modulating protein activity through the UPS pathway.

Given the potential of the E3 ligase enzyme RBX1 as a therapeutic target, there arises a compelling need to characterize this protein in *P. falciparum*. Furthermore, the development of inhibitors presents a promising avenue for exploring its role during the parasite's intraerythrocytic cycle.

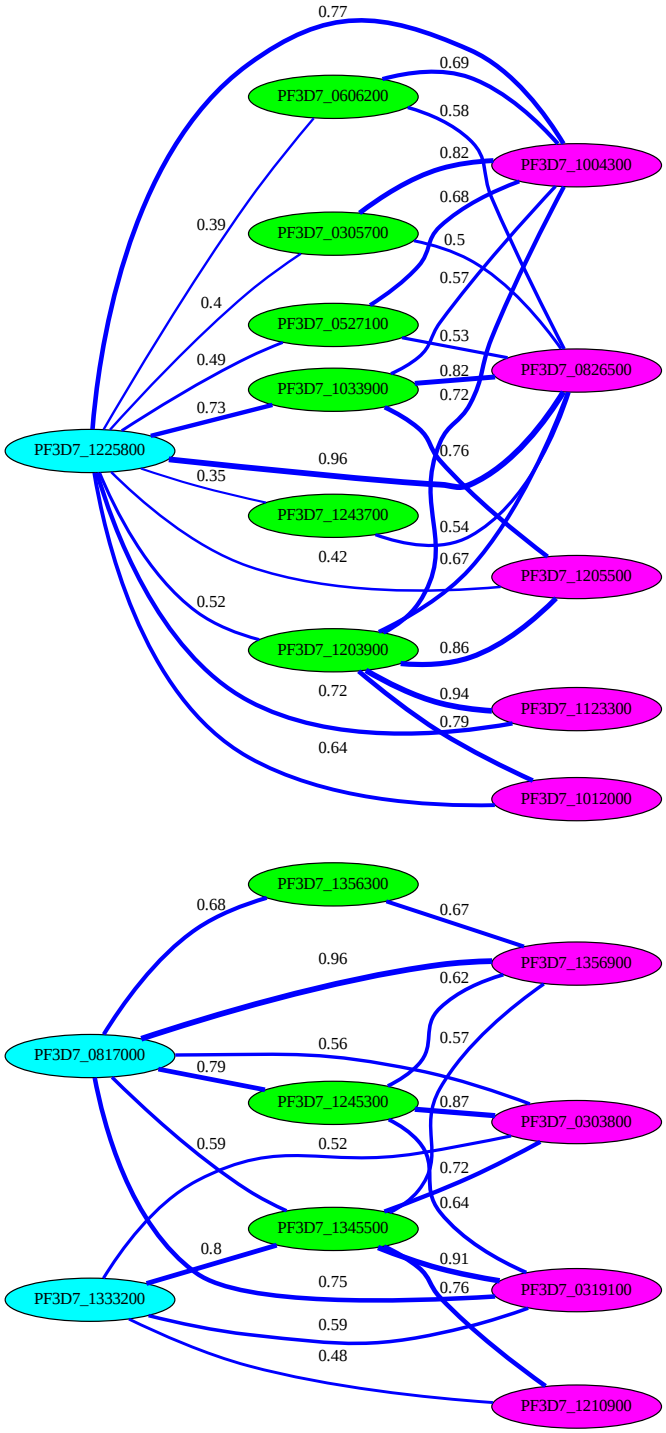


Figure 4.3: Graph generated using the top 25 candidates from the Broadbent et al., 2015 dataset with the same RNA-seq data as input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

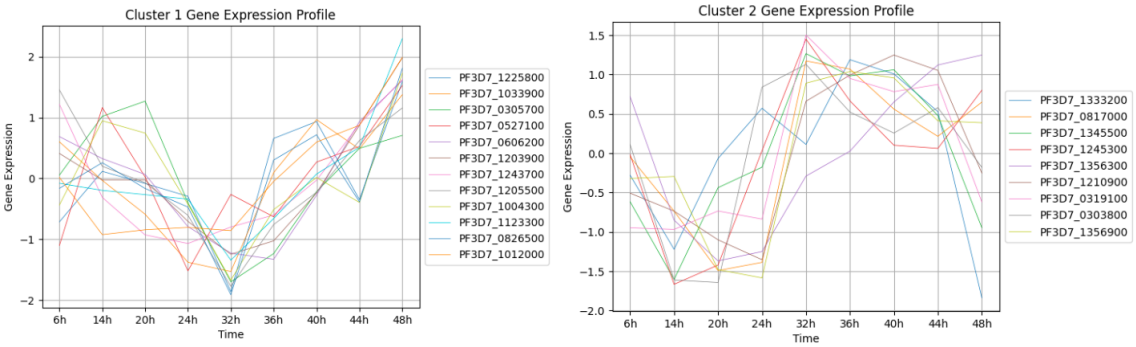


Figure 4.4: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

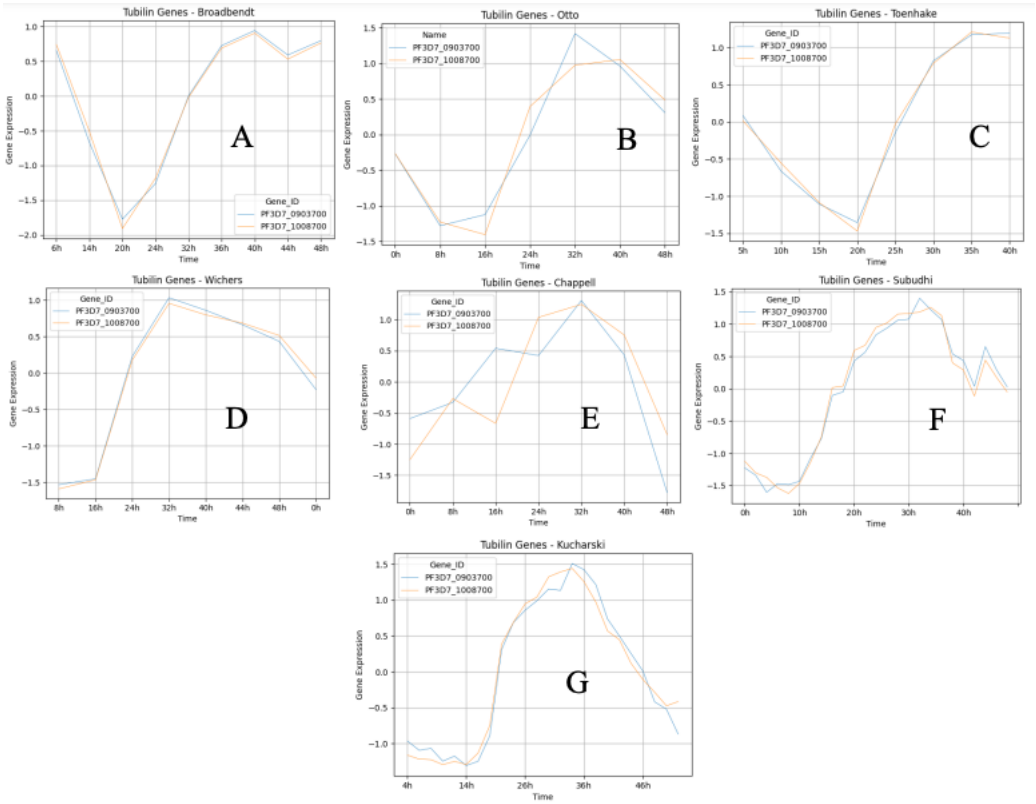


Figure 4.5: Gene expression profiles for the gene pair PF3D7_0903700 (Alpha tubulin 1) and PF3D7_1008700 (Tubulin beta chain) across all datasets. A: Broadbent et al., 2015; B: Otto et al., 2010; C: Toenhake et al., 2018; D: Wichers et al., 2019; E: Subudhi et al., 2020; F: Chappell et al., 2020; G: Kucharski et al., 2020.

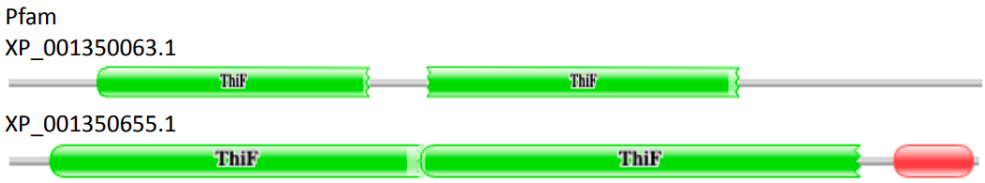


Figure 4.6: Pfam analysis for the two uba1 proteins, XP_001350655.1 and XP_001350063.1, encoded respectively by the PF3D7_1225800 and PF3D7_1333200 transcripts.

Chapter 5

Conclusion

Our model, designed as a predictive instrument for unraveling potential gene interactions, functions within the context of individual RNA-seq experiments, since as we saw they show a high divergence. Within this context, it's intriguing to observe the emergence of genes that exhibit remarkable robustness, transcending the constraints of experimental diversity.

Our GCN model has, without a doubt, displayed its prowess in forecasting and guiding further investigations. Section 4.1.2 serves as an example of the model's capabilities, where we rigorously tested its predictive prowess using the well-known Alpha tubulin 1 gene and tubulin beta chain gene. These genes, integral components of the Tubulin complex, have been long-established collaborators [SFK+19, HFK+22]. The model adeptly illuminated their collaborative potential, correctly predicting their likely cooperation, underscoring the value of our predictive tool.

Yet, we have to deal with the inherent challenges stemming from the gene expression variability across datasets. This inherent diversity prevents us from forging universal rules or patterns. Even as we scrutinize the expression dynamics of the tubulin genes through the stages of the IDC across these disparate datasets, it becomes evident that we remain constrained by the divergent RNA-seq data.

In our quest, we did not merely predict gene interactions. We took an additional step after filtering and selecting the best candidates, leading to the formation of networks, exemplified in Figures 4.3 and 4.4 for the Broadbent et al., 2015 dataset [BBR+15]. These networks and gene expression profiles align with our hypothesis, suggesting the key role of the UPS within the regulatory network governing the parasite's life cycle. Yet, it's crucial to acknowledge the dependency of these results on the experimental context. While our findings offer valuable insights, they also underscore the need for further investigations to unveil the full extent of the UPS's regulatory influence throughout the *P. falciparum* IDC.

Our search for intersections of the best candidates across datasets revealed the presence of three robust triples, meticulously detailed in Tables 3.2, 3.3, and 3.4. These triples consistently demonstrated their robustness across all seven RNA-seq experiments, as illustrated in Figures 3.3, 3.4, and 3.5. Our research indicates a fundamental role of these E1, E2 and E3 enzymes in the parasite development and replication during the trophozoite and schizont stages of the *P. falciparum* IDC. This work has culminated in a paper submitted to the Heliyon journal (https://www.cell.com/heliyon/home), which is currently under review.

For future work, a natural generalization of our problem is investigating E3 substrate interactions (ESIs) and deubiquitinase (DUB) substrate interactions (DSIs). ESIs and DSIs

represent the challenge of identifying the target proteins of E3s and DUBs, as highlighted in recent literature [HLPY12, PKW20, SHCW23]. However, this problem has a high complexity—virtually any protein can undergo ubiquitination or deubiquitination, as expounded in Section 1.2.2. Our current Gene Co-expression Network model cannot address this combinatorial challenge encompassing all E1s, E2s, E3s, and proteins within *P. falciparum*.

It is important to note that the scientific community is actively exploring ESIs and DSIs. Recent studies [HLPY12, PKW20, LCJ+21, WLH+22, HLW+22, SHCW23] have significantly contributed to enhancing our understanding of these interactions. Their findings shed light on the complex network of ubiquitin-mediated regulation in various organisms, providing valuable insights that can complement our understanding of *P. falciparum* IDC regulation.

To align our approach with the research objective of comprehensively exploring the UPS's regulatory impact on the *P. falciparum* IDC, we have strategically shifted our focus to 19 widely recognized cell cycle regulators specific to *P. falciparum* during the IDC, as documented in Matthews et al., 2018 [MDM18]. This strategic adjustment allows our model to address a more expansive challenge that includes ESIs/DSIs. By incorporating this new group of 19 genes into our E1, E2, and E3 model, we can ensure that our combinatorial software can handle the complex landscape of UPS regulation.

Chapter 6

Appendices

6.1 Triple Filtering Appendix

The following Figures 6.1, 6.2, 6.3, 6.4, 6.5, and 6.6 are the plots used to select the minimum score value used as a threshold, as described in Section 3.3.

6.2 Graphs Generated from the Best Candidates of Broadbent et al.,2015

In Section 4.1.1, we detailed our analysis of the top 25 results from Broadbent et al., 2015. Now, in this section, we examine how these same 25 triples perform across the remaining six datasets.

Cluster 1 (C1) experiences a significant loss of coherence across all datasets, often displaying negative correlations. These findings suggest that the genes within C1 might not be operating collaboratively, as evident from their random gene expression profiles.

Cluster 2 (C2) shows a relatively milder impact, although its performance remains below expectations when compared to the results derived from the Broadbent et al., 2015 dataset. While there are discernible patterns in C2's gene expression profiles, they deviate substantially from what was anticipated based on the Broadbent dataset. These observations underscore the profound influence of the experimental context on our model's performance, as even the

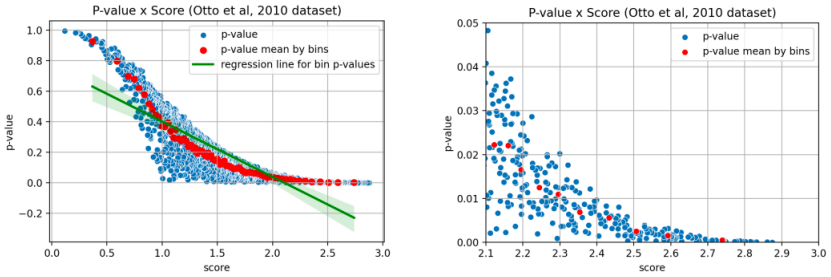


Figure 6.1: Scatterplots of P-value versus Score for Otto et al., 2010 [OWA+10], using $t = 0$ (only positive correlations). The right panel shows data for triples with Score ≥ 2.1. The red dots represent the average p-values calculated in 68 bins, while the green line in the left panel represents the best-fit linear regression on the 68 average p-values.

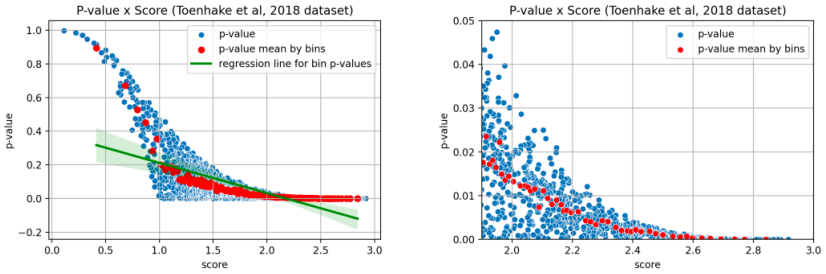


Figure 6.2: Scatterplots of P-value versus Score for Toenhake et al., 2018 [Toe], using $t = 0$ (only positive correlations). The right panel shows data for triples with Score ≥ 1.9. The red dots represent the average p-values calculated in 120 bins, while the green line in the left panel represents the best-fit linear regression on the 120 average p-values.

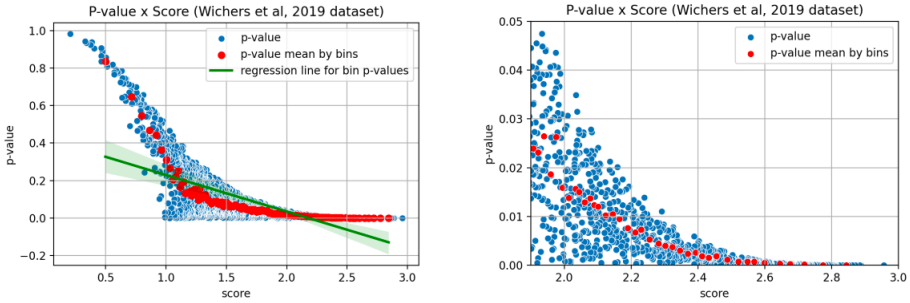


Figure 6.3: Scatterplots of P-value versus Score for Wichers et al., 2019 [WSS+19], using $t = 0$ (only positive correlations). The right panel shows data for triples with Score ≥ 1.9. The red dots represent the average p-values calculated in 109 bins, while the green line in the left panel represents the best-fit linear regression on the 109 average p-values.

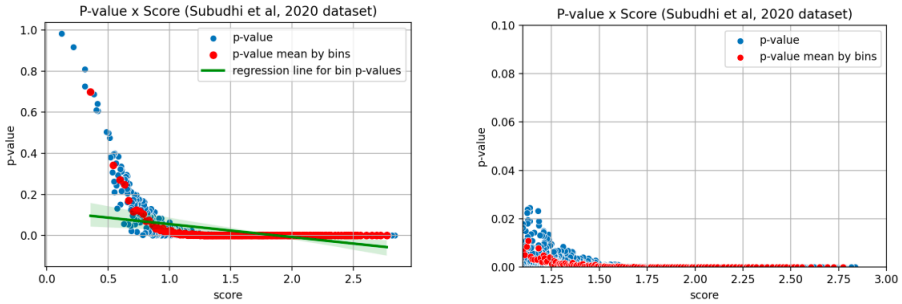


Figure 6.4: Scatterplots of P-value versus Score for Subudhi et al., 2020 [SOR+20], using $t = 0$ (only positive correlations). The right panel shows data for triples with Score ≥ 1.1. The red dots represent the average p-values calculated in 150 bins. It's noteworthy that this dataset requires a lower minimum Score threshold due to its extensive time point sampling from 0h to 48h, enhancing correlation and Score confidence.

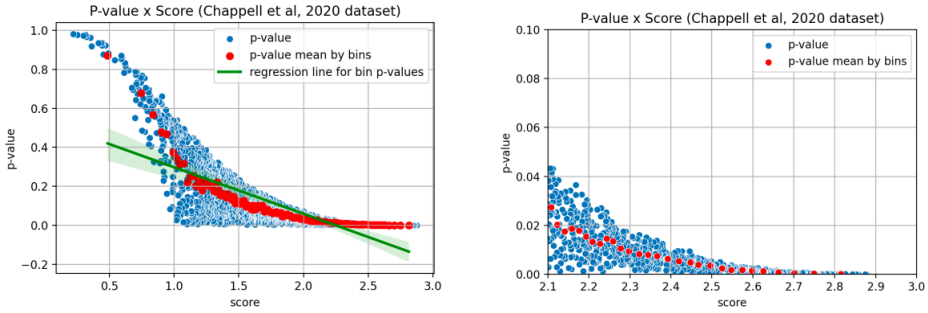


Figure 6.5: Scatterplots of P-value versus Score for Chappell et al., 2020 [CRO+20], using $t = 0$ (only positive correlations). The right panel shows data for triples with Score ≥ 2.1. The red dots represent the average p-values calculated in 104 bins, while the green line in the left panel represents the best-fit linear regression on the 104 average p-values.

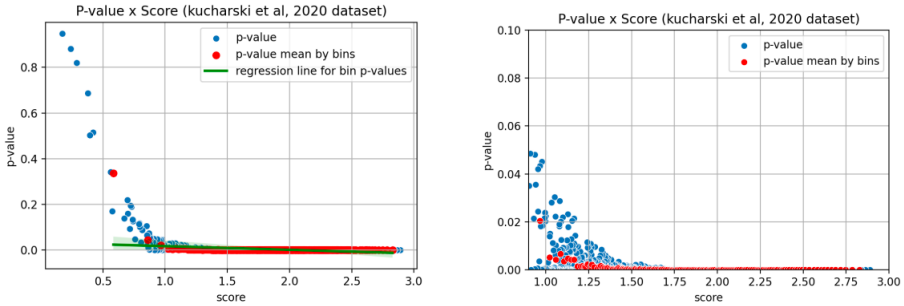


Figure 6.6: Scatterplots of P-value versus Score for Kucharski et al., 2020 [KTN+20], using $t = 0$ (only positive correlations). The right panel shows data for triples with Score ≥ 0.9. The red dots represent the average p-values calculated in 151 bins. This analysis highlights the impact of time point density on Score confidence. Kucharski et al. (2020) benefits from denser time point sampling, resulting in improved correlation and Score confidence.

most promising candidates from one dataset exhibit subpar performance across the other six

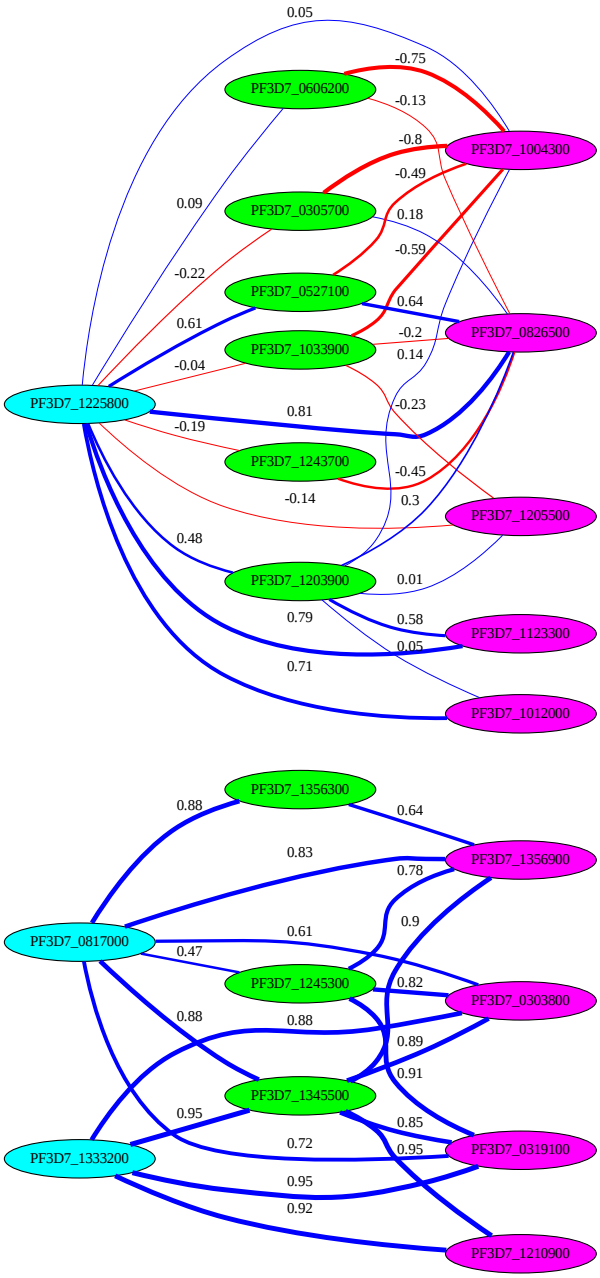


Figure 6.7: Graph generated from the best candidates of Broadbent et al., 2015 dataset using Otto et al.,2010 as RNA-seq input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. Blue edges represent positive correlation and red edges negative correlation The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

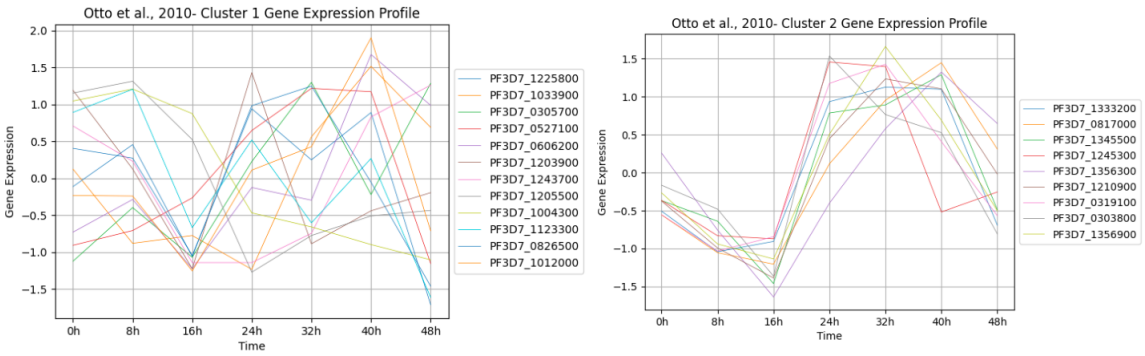


Figure 6.8: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015 using Otto et al., 2010 as RNA-seq input. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

6.3 Supplementary Materials

The supplementary materials include the following files and resources:

- **RNA-seq Datasets (`datasets.xlsx`):** This file contains the RNA-seq datasets used in this study.
- **Genes of Interest Classification and Relevant Information (`ups_net.xlsx`):** This file provides classification and relevant data about the genes of interest.
- **Best Candidates for All Datasets (`best_candidates.xlsx`):** Here, you can find a list of the best candidate genes selected for each dataset, along with supporting data.
- **Alignment of Amino Acid Sequences:** This resource includes the alignment of amino acid sequences of two Uba1 proteins, XP_001350655.1 and XP_001350063.1, encoded by the PF3D7_1225800 and PF3D7_1333200 genes. You can access this alignment in the PDF file named `XP_001350655.1_XP_001350063.1_alignment.pdf`.

Furthermore, all the code, including the GCN model, auxiliary libraries, and the pipeline used to obtain these results, is available in this GitHub repository:
https://github.com/LyangHiga/gcn_p_falciparum and in this Google Drive:
https://drive.google.com/drive/folders/1_lGCRsT1v06SG-4BUvJQ0yyRpvXBycQ6?usp=sharing

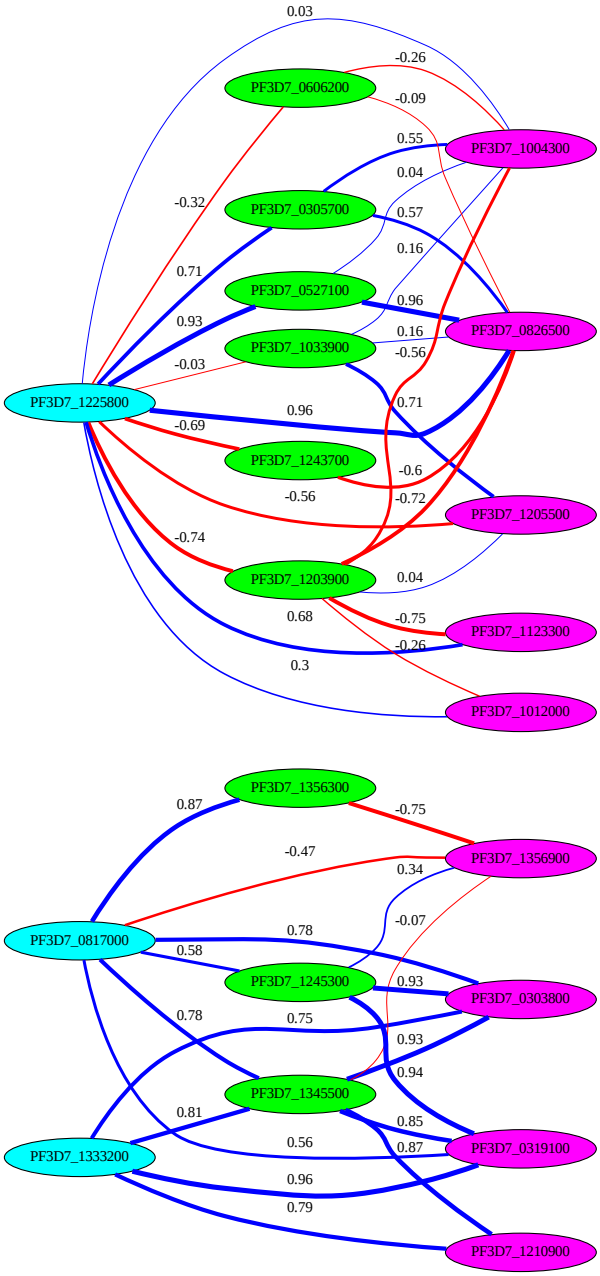


Figure 6.9: Graph generated from the best candidates of Broadbent et al., 2015 dataset using Toenhake et al.,2018 as RNA-seq input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. Blue edges represent positive correlation and red edges negative correlation. The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

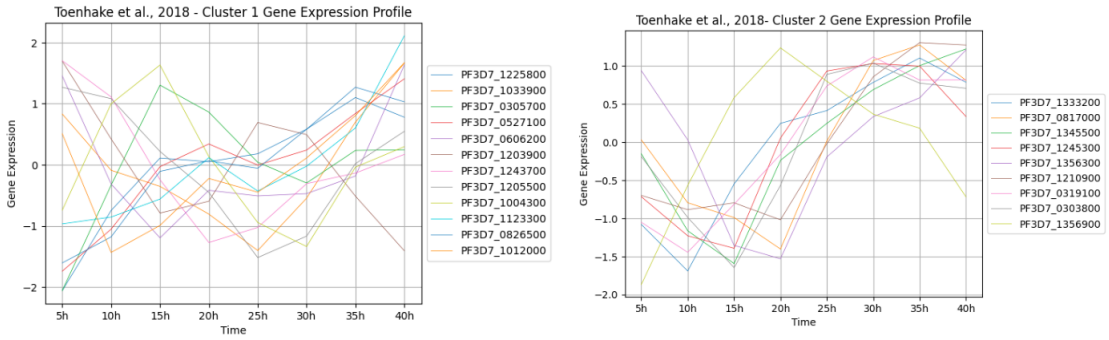


Figure 6.10: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015 using Toenhake et al., 2018 as RNA-seq input. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

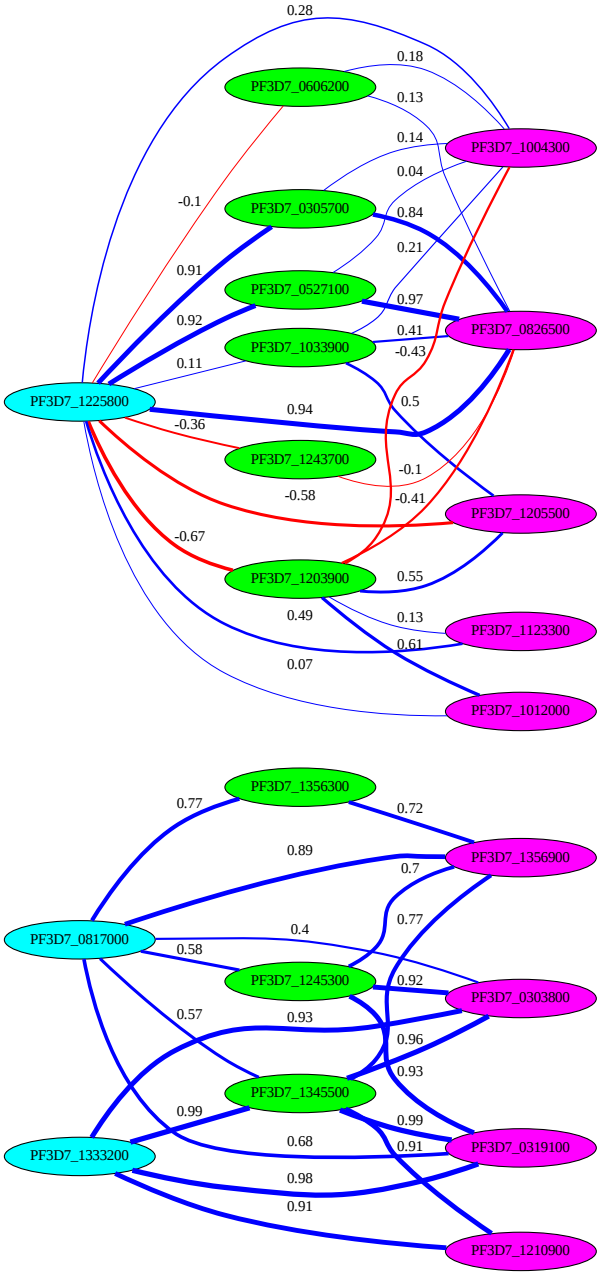


Figure 6.11: Graph generated from the best candidates of Broadbent et al., 2015 dataset using Wichers et al.,2019 as RNA-seq input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. Blue edges represent positive correlation and red edges negative correlation. The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

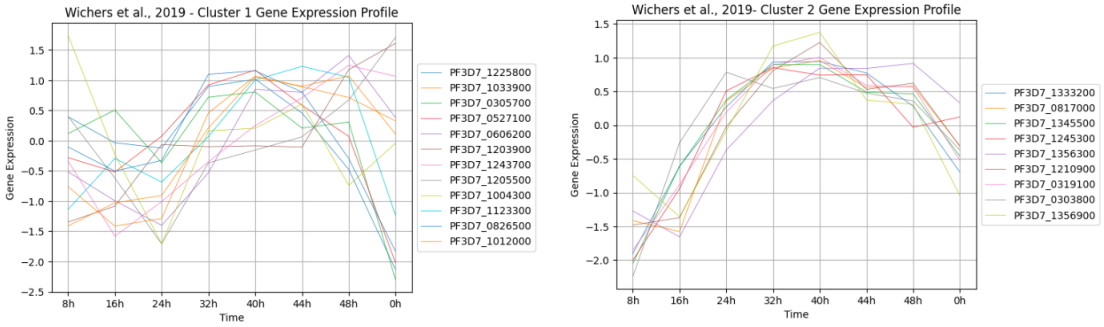


Figure 6.12: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015 using Wichers et al., 2019 as RNA-seq input. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

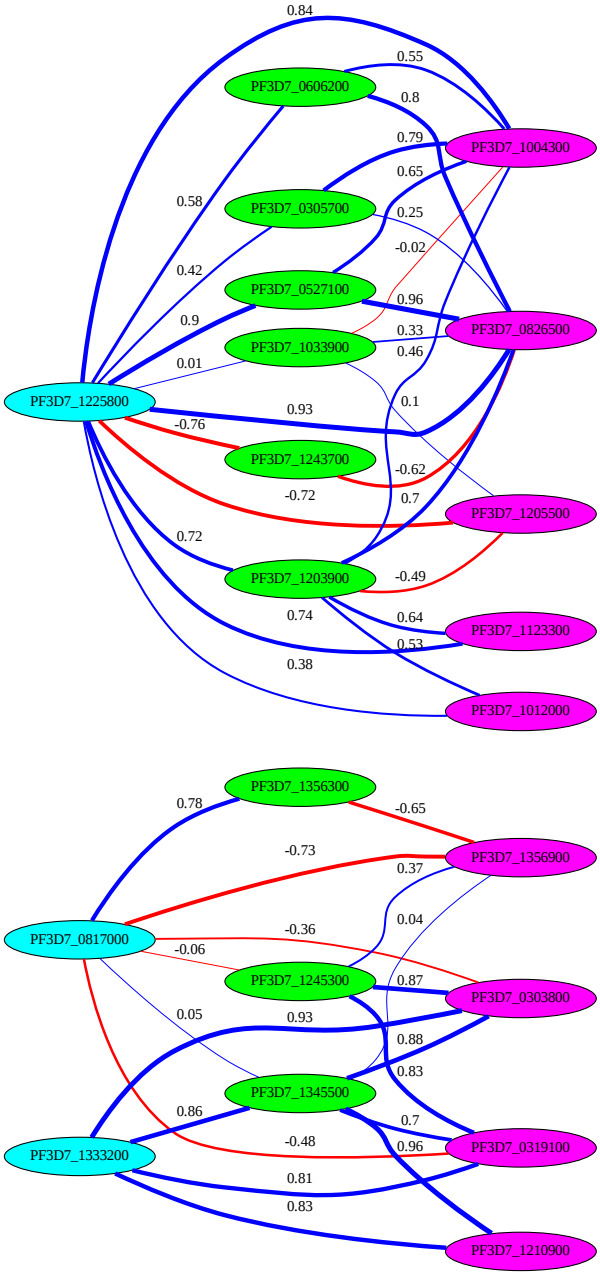


Figure 6.13: Graph generated from the best candidates of Broadbent et al., 2015 dataset using Chappell et al.,2020 as RNA-seq input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. Blue edges represent positive correlation and red edges negative correlation. The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

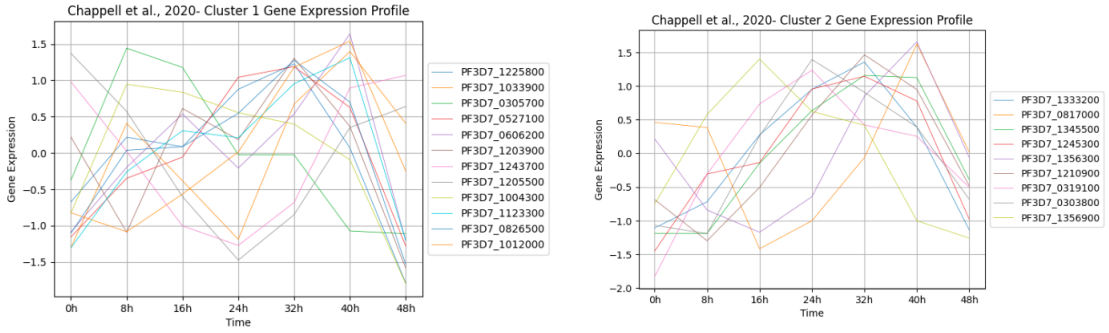


Figure 6.14: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015 using Chappell et al., 2020 as RNA-seq input. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

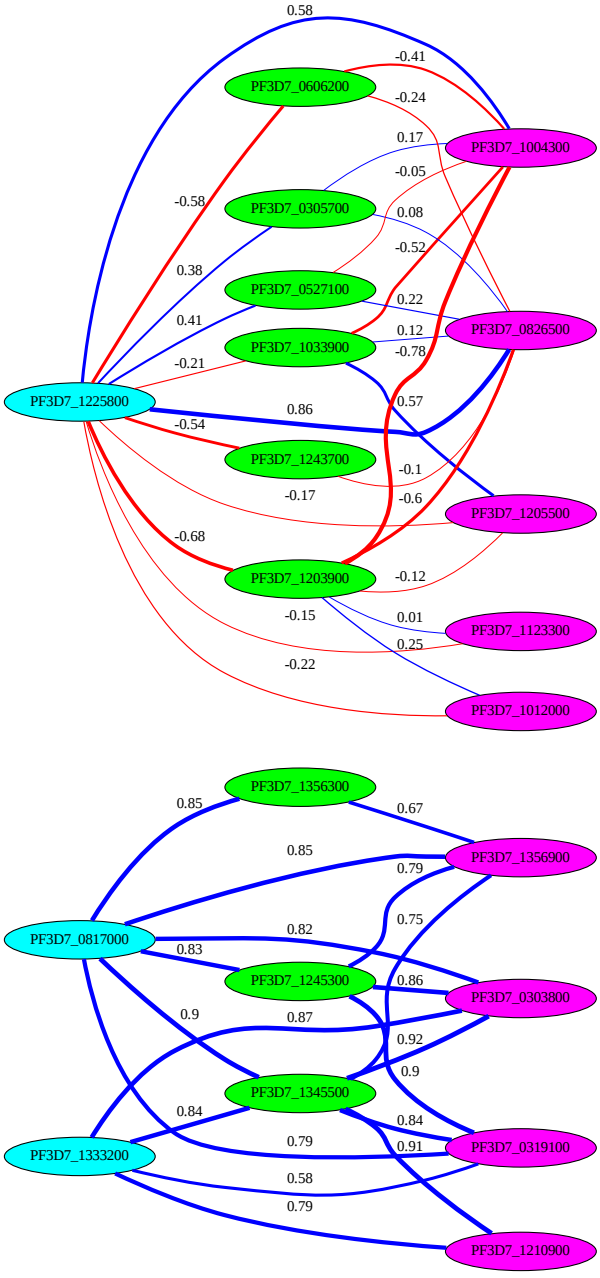


Figure 6.15: Graph generated from the best candidates of Broadbent et al., 2015 dataset using Subudhi et al.,2020 as RNA-seq input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. Blue edges represent positive correlation and red edges negative correlation. The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

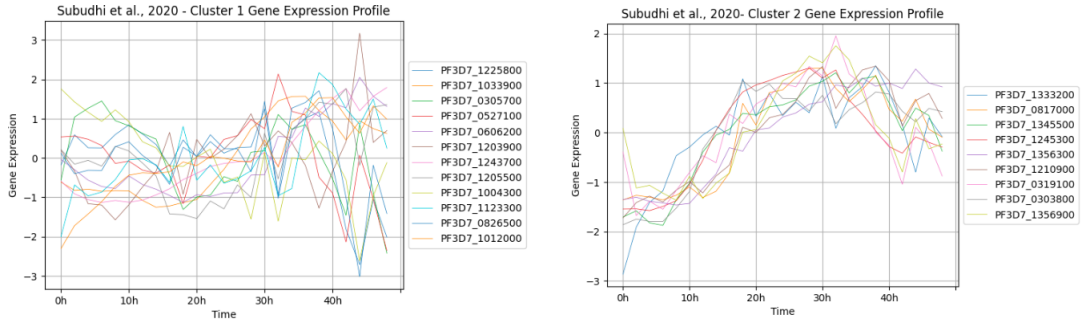


Figure 6.16: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015 using Subudhi et al., 2020 as RNA-seq input. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

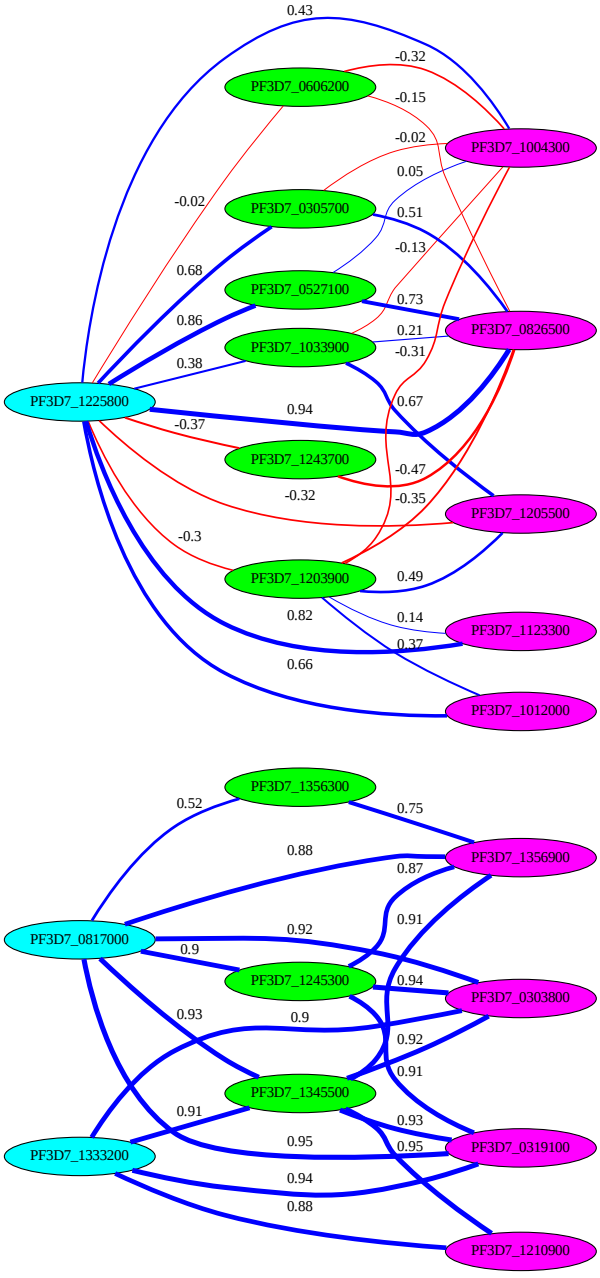


Figure 6.17: Graph generated from the best candidates of Broadbent et al., 2015 dataset using Kucharski et al.,2020 as RNA-seq input. Genes from E1 are shown in cyan, E2 in green, and E3 in magenta. Blue edges represent positive correlation and red edges negative correlation. The upper disjointed component is labeled as "Cluster 1" (C1), while the lower component is labeled as "Cluster 2" (C2).

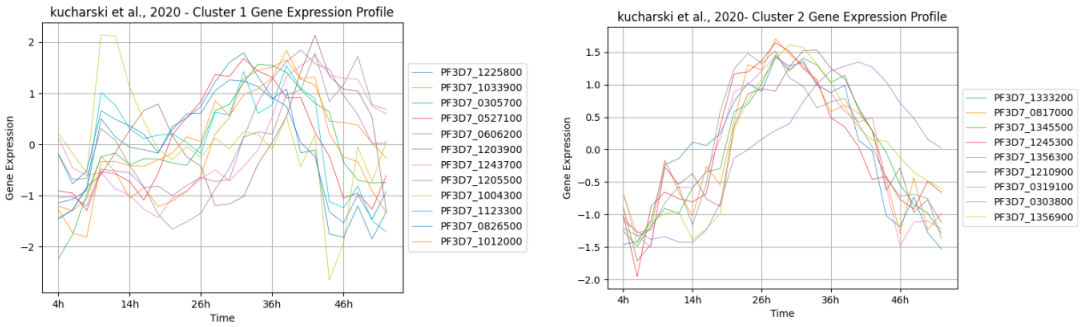


Figure 6.18: Gene expression profiles for the 25 best candidates from Broadbent et al., 2015 using Kucharski et al., 2020 as RNA-seq input. The left panel displays the gene expression profile for Cluster 1, while the right panel shows the gene expression profile for Cluster 2.

Bibliography

[AAP12] Makoah Nigel Aminake, Hans-Dieter Arndt, and Gabriele Pradel. The proteasome of malaria parasites: A multi-stage drug target for chemotherapeutic intervention? *International Journal for Parasitology: Drugs and Drug Resistance*, 2:1–10, December 2012. URL: https://www.sciencedirect.com/science/article/pii/S2211320711000145, doi:10.1016/j.ijpddr.2011.12.001. 9, 22

[ABB+09] Cristina Aurrecoechea, John Brestelli, Brian P. Brunk, Jennifer Dommer, Steve Fischer, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Omar S. Harb, Mark Heiges, Frank Innamorato, John Iodice, Jessica C. Kissinger, Eileen Kraemer, Wei Li, John A. Miller, Vishal Nayak, Cary Pennington, Deborah F. Pinney, David S. Roos, Chris Ross, Christian J. Stoeckert, Jr., Charles Treatman, and Haiming Wang. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Research*, 37(suppl_1):D539–D543, January 2009. doi:10.1093/nar/gkn814. 24, 32

[ACP+13] Swati Agrawal, Duk-Won D. Chung, Nadia Ponts, Giel G. van Dooren, Jacques Prudhomme, Carrie F. Brooks, Elisadra M. Rodrigues, John C. Tan, Michael T. Ferdig, Boris Striepen, and Karine G. Le Roch. An Apicoplast Localized Ubiquitylation System Is Required for the Import of Nuclear-encoded Plastid Proteins. *PLOS Pathogens*, 9(6):e1003426, June 2013. Publisher: Public Library of Science. URL: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003426, doi:10.1371/journal.ppat.1003426. 44

[BBR+15] Kate M Broadbent, Jill C Broadbent, Ulf Ribacke, Dyann Wirth, John L Rinn, and Pardis C Sabeti. Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long noncoding RNA and circular RNA. *BMC Genomics*, 16(1):454, December 2015. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1603-4, doi:10.1186/s12864-015-1603-4. ix, 25, 26, 33, 35, 36, 37, 38, 49

[BPL+09] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *International Journal of Molecular Sciences*, 10(6):2763–2788, June 2009. URL: http://www.mdpi.com/1422-0067/10/6/2763, doi:10.3390/ijms10062763. viii, 19, 20

[BXC+18] Jessica L. Bridgford, Stanley C. Xie, Simon A. Cobbold, Charisse Flerida A. Pasaje, Susann Herrmann, Tuo Yang, David L. Gillett, Lawrence R. Dick, Stuart A. Ralph, Con Dogovski, Natalie J. Spillman, and Leann Tilley. Artemisinin kills malaria parasites by damaging proteins and inhibiting the proteasome. *Nature Communications*, 9(1):3801, December 2018. URL: http://www.nature.com/articles/s41467-018-06221-1, doi:10.1038/s41467-018-06221-1. vii, 4, 5, 6, 14, 18

[CC17] Philipp M. Cromm and Craig M. Crews. The Proteasome in Modern Drug Discovery: Second Life of a Highly Valuable Drug Target. *ACS Central Science*, 3(8):830–838, August 2017. URL: https://pubs.acs.org/doi/10.1021/acscentsci.7b00252, doi:10.1021/acscentsci.7b00252. vii, 6, 7, 8, 9, 14, 18, 22

[CHMM16] Alan F. Cowman, Julie Healer, Danushka Marapana, and Kevin Marsh. Malaria: Biology and Disease. *Cell*, 167(3):610–624, October 2016. URL: https://linkinghub.elsevier.com/retrieve/pii/S009286741631008X, doi:10.1016/j.cell.2016.07.055. vii, 2, 3

[CRO+20] Lia Chappell, Philipp Ross, Lindsey Orchard, Timothy J. Russell, Thomas D. Otto, Matthew Berriman, Julian C. Rayner, and Manuel Llinás. Refining the transcriptome of the human malaria parasite Plasmodium falciparum using amplification-free RNA-seq. *BMC genomics*, 21(1):395, June 2020. doi:10.1186/s12864-020-06787-5. ix, xi, 25, 27, 33, 36, 37, 38, 53

[CZD22] Shu-Chun Chang, Bo-Xiang Zhang, and Jeak Ling Ding. E2-E3 ubiquitin enzyme pairing - partnership in provoking or mitigating cancers. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1877(2):188679, March 2022. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304419X2200004X, doi:10.1016/j.bbcan.2022.188679. viii, 12, 13, 14, 15, 18, 21, 44

[DB14] Nico P. Dantuma and Laura C. Bott. The ubiquitin-proteasome system in neurodegenerative diseases: precipitating factor, yet part of the solution. *Frontiers in Molecular Neuroscience*, 7, July 2014. URL: http://journal.frontiersin.org/article/10.3389/fnmol.2014.00070/abstract, doi:10.3389/fnmol.2014.00070. 14, 18

[DCJB+18] Virginia De Cesare, Clare Johnson, Victoria Barlow, James Hastie, Axel Knebel, and Matthias Trost. The MALDI-TOF E2/E3 Ligase Assay as Universal Tool for Drug Discovery in the Ubiquitin Pathway. *Cell Chemical Biology*, 25(9):1117–1127.e4, September 2018. URL: https://linkinghub.elsevier.com/retrieve/pii/S2451945618301946, doi:10.1016/j.chembiol.2018.06.004. 14

[DJ09] Raymond J. Deshaies and Claudio A.P. Joazeiro. RING Domain E3 Ubiquitin Ligases. *Annual Review of Biochemistry*, 78(1):399–434, June 2009. URL: https://www.annualreviews.org/doi/10.1146/annurev.biochem.78.101807.093809, doi:10.1146/annurev.biochem.78.101807.093809. viii, 10, 11, 12, 13, 14, 15, 16, 18, 34, 44, 45

[dlC] Antonio Rafael de la Cova. Latin american studies: Incas quipu. URL: http://www.latinamericanstudies.org/quipu.htm. vii, 11

[DWW09] Ivan Dikic, Soichi Wakatsuki, and Kylie J. Walters. Ubiquitin-binding domains — from structures to functions. *Nature Reviews Molecular Cell Biology*, 10(10):659–671, October 2009. URL: https://www.nature.com/articles/nrm2767, doi:10.1038/nrm2767. 17

[FCAS17] Justin D. Fellows, Michael J. Cipriano, Swati Agrawal, and Boris Striepen. A Plastid Protein That Evolved from Ubiquitin and Is Required for Apicoplast Protein Import in Toxoplasma gondii. *mBio*, 8(3):e00950–17, June 2017. doi:10.1128/mBio.00950-17. 44

[FP19] Tyler G. Franklin and Jonathan N. Pruneda. A High-Throughput Assay for Monitoring Ubiquitination in Real Time. *Frontiers in Chemistry*, 7:816, December 2019. URL: https://www.frontiersin.org/article/10.3389/fchem.2019.00816/full, doi:10.3389/fchem.2019.00816. 14

[FS89] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, July 1989. URL: https://www.nature.com/articles/340245a0, doi:10.1038/340245a0. 19

[FSH+05] Bernardo J. Foth, Luciana M. Stimmler, Emanuela Handman, Brendan S. Crabb, Anthony N. Hodder, and Geoffrey I. McFadden. The malaria parasite Plasmodium falciparum has only one pyruvate dehydrogenase complex, which is located in the apicoplast. *Molecular Microbiology*, 55(1):39–53, January 2005. doi:10.1111/j.1365-2958.2004.04407.x. 44

[GAN+17] Yingying Guo, Liwei An, Hoi-Man Ng, Shirley M. H. Sy, and Michael S. Y. Huen. An E2-guided E3 Screen Identifies the RNF17-UBE2U Pair as Regulator of the Radiosensitivity, Immunodeficiency, Dysmorphic Features, and Learning Difficulties (RIDDLE) Syndrome Protein RNF168 *. *Journal of Biological Chemistry*, 292(3):967–978, January 2017. Publisher: Elsevier. URL: https://www.jbc.org/article/S0021-9258(20)40275-3/abstract, doi:10.1074/jbc.M116.758854. 21

[GNM+11] Sara J. C. Gosline, Mirna Nascimento, Laura-Isobel McCall, Dan Zilberstein, David Y. Thomas, Greg Matlashewski, and Michael Hallett. Intracellular Eukaryotic Parasites Have a Distinct Unfolded Protein Response. *PLoS ONE*, 6(4):e19118, April 2011. URL: https://dx.plos.org/10.1371/journal.pone.0019118, doi:10.1371/journal.pone.0019118. 4, 18

[GvS16] Roly Gosling and Lorenz von Seidlein. The Future of the RTS,S/AS01 Malaria Vaccine: An Alternative Development Plan. *PLoS Medicine*, 13(4):e1001994, April 2016. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4829262/, doi:10.1371/journal.pmed.1001994. 3

[HFK+22] William G. Hirst, Dominik Fachet, Benno Kuropka, Christoph Weise, Kevin J. Saliba, and Simone Reber. Purification of functional Plasmodium falciparum tubulin allows for the identification of parasite-specific microtubule

inhibitors. *Current Biology*, 32(4):919–926.e6, February 2022. Publisher: Elsevier. URL: https://www.cell.com/current-biology/abstract/S0960-9822(21)01736-X, doi:10.1016/j.cub.2021.12.049. 42, 49

[HLPY12] Youngwoong Han, Hodong Lee, Jong C. Park, and Gwan-Su Yi. E3Net: A System for Exploring E3-mediated Regulatory Networks of Cellular Functions *. *Molecular & Cellular Proteomics*, 11(4), April 2012. Publisher: Elsevier. URL: https://www.mcponline.org/article/S1535-9476(20)30487-4/abstract, doi:10.1074/mcp.O111.014076. 50

[HLW+22] Chao Hou, Yuxuan Li, Mengyao Wang, Hong Wu, and Tingting Li. Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning. *BMC Biology*, 20(1):162, December 2022. URL: https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-022-01364-6, doi:10.1186/s12915-022-01364-6. 50

[Hoc09] Mark Hochstrasser. Origin and function of ubiquitin-like proteins. *Nature*, 458(7237):422–429, March 2009. URL: https://www.nature.com/articles/nature07958, doi:10.1038/nature07958. 12

[HP76] Larry L. Havlicek and Nancy L. Peterson. Robustness of the Pearson Correlation against Violations of Assumptions. *Perceptual and Motor Skills*, 43(3_suppl):1319–1334, December 1976. URL: http://journals.sagepub.com/doi/10.2466/pms.1976.43.3f.1319, doi:10.2466/pms.1976.43.3f.1319. 28

[HRD18] Nicholas Heard and Patrick Rubin-Delanchy. Choosing Between Methods of Combining p-values. *Biometrika*, 105(1):239–246, March 2018. arXiv:1707.06897 [stat]. URL: http://arxiv.org/abs/1707.06897, doi:10.1093/biomet/asx076. 29

[JJWT17] Jagrati Jain, Surendra K. Jain, Larry A. Walker, and Babu L. Tekwani. Inhibitors of ubiquitin E3 ligase as potential new antimalarial drug leads. *BMC Pharmacology and Toxicology*, 18(1):40, June 2017. doi:10.1186/s40360-017-0147-4. 8, 22

[JSS09] Lijun Jia, Maria S. Soengas, and Yi Sun. ROC1/RBX1 E3 ubiquitin ligase silencing suppresses tumor cell growth via sequential induction of G2/M arrest, apoptosis, and senescence. *Cancer research*, 69(12):4974–4982, June 2009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2744327/, doi:10.1158/0008-5472.CAN-08-4671. 45

[KAB+14] Fernanda C. Koyama, Mauro F. Azevedo, Alexandre Budu, Debopam Chakrabarti, and Célia R. S. Garcia. Melatonin-induced temporal up-regulation of gene expression related to ubiquitin/proteasome system (UPS) in the human malaria parasite Plasmodium falciparum. *International Journal of Molecular Sciences*, 15(12):22320–22330, December 2014. doi:10.3390/ijms151222320. 9, 22

[KKNG12] Gozde Kar, Ozlem Keskin, Ruth Nussinov, and Attila Gursoy. Human Proteome-scale Structural Modeling of E2–E3 Interactions Exploiting Interface Motifs. *Journal of Proteome Research*, 11(2):1196–1207, February 2012. URL: https://pubs.acs.org/doi/10.1021/pr2009143, doi:10.1021/pr2009143. 14, 16

[KR12] David Komander and Michael Rape. The Ubiquitin Code. *Annual Review of Biochemistry*, 81(1):203–229, July 2012. URL: https://www.annualreviews.org/doi/10.1146/annurev-biochem-060310-170328, doi:10.1146/annurev-biochem-060310-170328. 4, 10, 11, 12, 14, 15, 16, 17, 18, 19, 44

[KTN+20] Michal Kucharski, Jaishree Tripathi, Sourav Nayak, Lei Zhu, Grennady Wirjanata, Rob W. van der Pluijm, Mehul Dhorda, Arjen Dondorp, and Zbynek Bozdech. A comprehensive RNA handling and transcriptomics guide for high-throughput processing of Plasmodium blood-stage samples. *Malaria Journal*, 19(1):363, October 2020. doi:10.1186/s12936-020-03436-w. ix, xi, 25, 27, 33, 36, 37, 38, 53

[LCJ+21] Zhongyan Li, Siyu Chen, Jhih-Hua Jhong, Yuxuan Pang, Kai-Yao Huang, Shangfu Li, and Tzong-Yi Lee. UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase–substrate interactions. *Database*, 2021:baab010, September 2021. doi:10.1093/database/baab010. 50

[LH08] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559, doi:10.1186/1471-2105-9-559. 30

[LL05] Luisa Maria Lois and Christopher D Lima. Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1. *The EMBO Journal*, 24(3):439–451, February 2005. URL: http://emboj.embopress.org/cgi/doi/10.1038/sj.emboj.7600552, doi:10.1038/sj.emboj.7600552. 44

[LMA+13] Wânia R. Lima, Miriam Moraes, Eduardo Alves, Mauro F. Azevedo, Dario O. Passos, and Célia R. S. Garcia. The PfNF-YB transcription factor is a downstream target of melatonin and cAMP signalling in the human malaria parasite Plasmodium falciparum. *Journal of Pineal Research*, 54(2):145–153, March 2013. doi:10.1111/j.1600-079X.2012.01021.x. 8, 22

[LS08] Imsang Lee and Hermann Schindelin. Structural Insights into E1-Catalyzed Ubiquitin Activation and Transfer to Conjugating Enzymes. *Cell*, 134(2):268–278, July 2008. Publisher: Elsevier. URL: https://www.cell.com/cell/abstract/S0092-8674(08)00709-5, doi:10.1016/j.cell.2008.05.046. 44

[LWRS01] Michael W. Lake, Margot M. Wuebbens, K. V. Rajagopalan, and Hermann Schindelin. Mechanism of ubiquitin activation revealed by the structure

of a bacterial MoeB–MoaD complex. *Nature*, 414(6861):325–329, November 2001. Number: 6861 Publisher: Nature Publishing Group. URL: https://www.nature.com/articles/35104586, doi:10.1038/35104586. 44

[Mat17] Kai Matuschewski. Vaccines against malaria-still a long way to go. *The FEBS journal*, 284(16):2560–2568, August 2017. doi:10.1111/febs.14107. 3

[MBSU07] Kelly Markham, Yu Bai, and Gerold Schmitt-Ulms. Co-immunoprecipitations revisited: an update on experimental concepts and their implementation for sensitive interactome investigations of endogenous proteins. *Analytical and Bioanalytical Chemistry*, 389(2):461–473, September 2007. URL: https://link.springer.com/10.1007/s00216-007-1385-x, doi:10.1007/s00216-007-1385-x. 21

[MDM18] Holly Matthews, Craig W. Duffy, and Catherine J. Merrick. Checks and balances? DNA replication and the cell cycle in Plasmodium. *Parasites & Vectors*, 11(1):216, December 2018. URL: https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-018-2800-1, doi:10.1186/s13071-018-2800-1. 50

[MKH+09] Gabriel Markson, Christina Kiel, Russell Hyde, Stephanie Brown, Panagoula Charalabous, Anja Bremm, Jennifer Semple, Jonathan Woodsmith, Simon Duley, Kourosh Salehi-Ashtiani, Marc Vidal, David Komander, Luis Serrano, Paul Lehner, and Christopher M. Sanderson. Analysis of the human E2 ubiquitin conjugating enzyme protein interaction network. *Genome Research*, 19(10):1905–1911, October 2009. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.093963.109, doi:10.1101/gr.093963.109. 13, 14, 20

[NFB17] Caroline L. Ng, David A. Fidock, and Matthew Bogyo. Protein Degradation Systems as Antimalarial Therapeutic Targets. *Trends in Parasitology*, 33(9):731–743, September 2017. URL: https://linkinghub.elsevier.com/retrieve/pii/S1471492217301320, doi:10.1016/j.pt.2017.05.009. 6, 7, 8, 9, 12, 14, 18, 22

[NWSW+14] Armand G. Ngounou Wetie, Izabela Sokolowska, Alisa G. Woods, Urmi Roy, Katrin Deinhardt, and Costel C. Darie. Protein–protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cellular and Molecular Life Sciences*, 71(2):205–228, January 2014. URL: http://link.springer.com/10.1007/s00018-013-1333-1, doi:10.1007/s00018-013-1333-1. 21

[OWA+10] Thomas D. Otto, Daniel Wilinski, Sammy Assefa, Thomas M. Keane, Louis R. Sarry, Ulrike Böhme, Jacob Lemieux, Bart Barrell, Arnab Pain, Matthew Berriman, Chris Newbold, and Manuel Llinás. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular Microbiology*, 76(1):12–24, April 2010. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2009.07026.x, doi:10.1111/j.1365-2958.2009.07026.x. ix, x, 25, 33, 36, 37, 38, 51

[PCG18] Pedro H. Scarpelli Pereira, Chiara Curra, and Celia R. S. Garcia. Ubiquitin Proteasome System as a Potential Drug Target for Malaria. *Current Topics in Medicinal Chemistry*, 18(5):315–320, 2018. doi:10.2174/1568026618666180427145308. 5, 6, 7, 8, 9, 12, 22

[PGS+16] William Poole, David L. Gibbs, Ilya Shmulevich, Brady Bernard, and Theo A. Knijnenburg. Combining dependent *P*- values with an empirical adaptation of Brown's method. *Bioinformatics*, 32(17):i430–i436, September 2016. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw438, doi:10.1093/bioinformatics/btw438. 29

[Pic01] Cecile M. Pickart. Mechanisms Underlying Ubiquitination. *Annual Review of Biochemistry*, 70(1):503–533, June 2001. URL: https://www.annualreviews.org/doi/10.1146/annurev.biochem.70.1.503, doi:10.1146/annurev.biochem.70.1.503. 4, 10, 12, 19

[PKW20] Seongyong Park, Shujaat Khan, and Abdul Wahab. E3-targetPred: Prediction of E3-Target Proteins Using Deep Latent Space Encoding, June 2020. arXiv:2007.12073 [cs, q-bio]. URL: http://arxiv.org/abs/2007.12073. 50

[Pre21] CDC-Centers for Disease Control and Prevention. CDC - Malaria - Malaria Worldwide - Impact of Malaria, December 2021. URL: https://www.cdc.gov/malaria/malaria_worldwide/impact.html. 1, 9

[RBH16] Judith A Ronau, John F Beckmann, and Mark Hochstrasser. Substrate specificity of the ubiquitin and Ubl proteases. *Cell Research*, 26(4):441–456, April 2016. URL: https://www.nature.com/articles/cr201638, doi:10.1038/cr.2016.38. 12

[RDB+03] Eddy P. Risseeuw, Timothy E. Daskalchuk, Travis W. Banks, Enwu Liu, Julian Cotelesage, Hanjo Hellmann, Mark Estelle, David E. Somers, and William L. Crosby. Protein interaction analysis of SCF ubiquitin E3 ligase subunits from Arabidopsis. *The Plant Journal*, 34(6):753–767, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-313X.2003.01768.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-313X.2003.01768.x, doi:10.1046/j.1365-313X.2003.01768.x. 43

[RRG+23] Zeba Rizvi, G. Srinivas Reddy, Somesh M. Gorde, Priyanka Pundir, Divya Das, and Puran Singh Sijwali. *Plasmodium falciparum* contains functional SCF and CRL4 ubiquitin E3 ligases, and CRL4 is critical for cell division and membrane integrity. preprint, Microbiology, April 2023. URL: http://biorxiv.org/lookup/doi/10.1101/2023.04.18.537323, doi:10.1101/2023.04.18.537323. 45

[RTS15] RTS,S Clinical Trials Partnership. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet (London, England)*, 386(9988):31–45, July 2015. doi:10.1016/S0140-6736(15)60721-8. 3

[Sat21] Shigeharu Sato. Plasmodium—a brief introduction to the parasites causing human malaria and their basic biology. *Journal of Physiological Anthropology*, 40(1):1, December 2021. URL: https://jphysiolanthropol.biomedcentral.com/articles/10.1186/s40101-020-00251-9, doi:10.1186/s40101-020-00251-9. 2, 3

[SF14] Marion Schmidt and Daniel Finley. Regulation of proteasome activity in health and disease. *Biochimica Et Biophysica Acta*, 1843(1):13–25, January 2014. doi:10.1016/j.bbamcr.2013.08.012. 22

[SFJ+22] Jun Shao, Qian Feng, Weifan Jiang, Yuting Yang, Zhiqiang Liu, Liang Li, Wenlong Yang, and Yufeng Zou. E3 ubiquitin ligase RBX1 drives the metastasis of triple negative breast cancer through a FBXO45-TWIST1-dependent degradation mechanism. *Aging (Albany NY)*, 14(13):5493–5510, July 2022. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9320552/, doi:10.18632/aging.204163. 45

[SFK+19] Benjamin Spreng, Hannah Fleckenstein, Patrick Kübler, Claudia Di Biagio, Madlen Benz, Pintu Patra, Ulrich S Schwarz, Marek Cyrklaff, and Friedrich Frischknecht. Microtubule number and length determine cellular shape and function in *Plasmodium*. *The EMBO Journal*, 38(15):e100984, August 2019. URL: https://www.embopress.org/doi/10.15252/embj.2018100984, doi:10.15252/embj.2018100984. 42, 49

[SGW+15] Judith Straimer, Nina F. Gnädig, Benoit Witkowski, Chanaki Amaratunga, Valentine Duru, Arba Pramundita Ramadani, Mélanie Dacheux, Nimol Khim, Lei Zhang, Stephen Lam, Philip D. Gregory, Fyodor D. Urnov, Odile Mercereau-Puijalon, Françoise Benoit-Vical, Rick M. Fairhurst, Didier Ménard, and David A. Fidock. K13-propeller mutations confer artemisinin resistance in Plasmodium falciparum clinical isolates. *Science*, 347(6220):428–431, January 2015. Publisher: American Association for the Advancement of Science. URL: https://www.science.org/doi/abs/10.1126/science.1260867, doi:10.1126/science.1260867. 5

[SHCW23] Yixuan Shu, Yanru Hai, Lihua Cao, and Jianmin Wu. Deep-learning based approach to identify substrates of human E3 ubiquitin ligases and deubiquitinases. *Computational and Structural Biotechnology Journal*, 21:1014–1021, 2023. URL: https://linkinghub.elsevier.com/retrieve/pii/S2001037023000211, doi:10.1016/j.csbj.2023.01.021. 50

[SHM+09] Simone Spork, Jan A. Hiss, Katharina Mandel, Maik Sommer, Taco W. A. Kooij, Trang Chu, Gisbert Schneider, Uwe G. Maier, and Jude M. Przyborski. An Unusual ERAD-Like Complex Is Targeted to the Apicoplast of Plasmodium falciparum. *Eukaryotic Cell*, 8(8):1134–1145, August 2009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2725561/, doi:10.1128/EC.00083-09. 44

[SK16] Kirby N Swatek and David Komander. Ubiquitin modifications. *Cell Research*, 26(4):399–422, April 2016. URL: http://www.nature.com/articles/cr201639, doi:10.1038/cr.2016.39. vii, viii, 11, 13, 14, 17, 18

[SNHL16] Elise A. R. Serin, Harm Nijveen, Henk W. M. Hilhorst, and Wilco Ligterink. Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in Plant Science*, 7, April 2016. URL: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00444/abstract, doi:10.3389/fpls.2016.00444. 22, 23

[SOR+20] Amit K. Subudhi, Aidan J. O'Donnell, Abhinay Ramaprasad, Hussein M. Abkallo, Abhinav Kaushik, Hifzur R. Ansari, Alyaa M. Abdel-Haleem, Fathia Ben Rached, Osamu Kaneko, Richard Culleton, Sarah E. Reece, and Arnab Pain. Malaria parasites regulate intra-erythrocytic development duration via serpentine receptor 10 to coordinate with host rhythms. *Nature Communications*, 11(1):2763, June 2020. doi:10.1038/s41467-020-16593-y. ix, x, 25, 26, 33, 36, 37, 38, 52

[SS14] Judith J Smit and Titia K Sixma. RBR E3-ligases at work. *EMBO reports*, 15(2):142–154, February 2014. URL: https://onlinelibrary.wiley.com/doi/10.1002/embr.201338166, doi:10.1002/embr.201338166. viii, 11, 16, 17

[Tav19] Tatyana Almeida Tavella. UNIVERSIDADE ESTADUAL DE CAMPINAS INSTITUTO DE BIOLOGIA. page 130, 2019. vii, 1, 2, 4, 5, 6, 9, 18

[Toe] Christa Geeke Toenhake. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. page 23. ix, x, 25, 26, 33, 36, 37, 38, 52

[TZL+23] Sha Tang, Zhiying Zhao, Xiaotong Liu, Yi Sui, Dandan Zhang, Hui Zhi, Yuanzhu Gao, Hui Zhang, Linlin Zhang, Yannan Wang, Meicheng Zhao, Dongdong Li, Ke Wang, Qiang He, Renliang Zhang, Wei Zhang, Guanqing Jia, Wenqiang Tang, Xingguo Ye, Chuanyin Wu, and Xianmin Diao. An E2-E3 pair contributes to seed size control in grain crops. *Nature Communications*, 14(1):3091, May 2023. URL: https://www.nature.com/articles/s41467-023-38812-y, doi:10.1038/s41467-023-38812-y. 14, 21

[Var06] A. Varshavsky. The early history of the ubiquitin field. *Protein Science*, 15(3):647–654, February 2006. URL: http://doi.wiley.com/10.1110/ps.052012306, doi:10.1110/ps.052012306. 9, 10

[VDVP12] Annemarthe G. Van Der Veen and Hidde L. Ploegh. Ubiquitin-Like Proteins. *Annual Review of Biochemistry*, 81(1):323–357, July 2012. URL: https://www.annualreviews.org/doi/10.1146/annurev-biochem-093010-153308, doi:10.1146/annurev-biochem-093010-153308. 12

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors.

SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi:10.1038/s41592-019-0686-2. 29

[vS19] Lorenz von Seidlein. The Advanced Development Pathway of the RTS,S/AS01 Vaccine. *Methods in molecular biology (Clifton, N.J.)*, 2013:177–187, January 2019. doi:10.1007/978-1-4939-9550-9_13. 3

[VWDVK+09] Sjoerd J L Van Wijk, Sjoerd J De Vries, Patrick Kemmeren, Anding Huang, Rolf Boelens, Alexandre M J J Bonvin, and H Th Marc Timmers. A comprehensive framework of E2–RING E3 interactions of the human ubiquitin–proteasome system. *Molecular Systems Biology*, 5(1):295, January 2009. URL: https://www.embopress.org/doi/10.1038/msb.2009.55, doi:10.1038/msb.2009.55. 14, 20, 21

[WB04] Elizabeth J.B. Williams and Dianna J. Bowles. Coexpression of Neighboring Genes in the Genome of *Arabidopsis thaliana*. *Genome Research*, 14(6):1060–1067, June 2004. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.2131104, doi:10.1101/gr.2131104. 22, 23

[Wil05] Keith D. Wilkinson. The discovery of ubiquitin-dependent proteolysis. *Proceedings of the National Academy of Sciences*, 102(43):15280–15282, October 2005. URL: https://pnas.org/doi/full/10.1073/pnas.0504842102, doi:10.1073/pnas.0504842102. 9, 10

[WLH+22] Xun Wang, Yang Li, Mengqi He, Xiangren Kong, Peng Jiang, Xi Liu, Lihong Diao, Xinlei Zhang, Honglei Li, Xinping Ling, Simin Xia, Zhongyang Liu, Yuan Liu, Chun-Ping Cui, Yan Wang, Liujun Tang, Lingqiang Zhang, Fuchu He, and Dong Li. UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase–substrate interactions in eukaryotic species. *Nucleic Acids Research*, 50(D1):D719–D728, January 2022. URL: https://academic.oup.com/nar/article/50/D1/D719/6406468, doi:10.1093/nar/gkab962. 50

[Wor21] World Health Organization. *World malaria report 2021*. World Health Organization, Geneva, 2021. Section: liv, 263 p. URL: https://apps.who.int/iris/handle/10665/350147. vii, 1, 2, 3, 4, 9

[WSS+19] J. Stephan Wichers, Judith A. M. Scholz, Jan Strauss, Susanne Witt, Andrés Lill, Laura-Isabell Ehnold, Niklas Neupert, Benjamin Liffner, Renke Lühken, Michaela Petter, Stephan Lorenzen, Danny W. Wilson, Christian Löw, Catherine Lavazec, Iris Bruchhaus, Egbert Tannich, Tim W. Gilberger, and Anna Bachmann. Dissecting the Gene Expression, Localization, Membrane Topology, and Function of the Plasmodium falciparum STEVOR Protein Family. *mBio*, 10(4):e01500–19, July 2019. doi:10.1128/mBio.01500-19. ix, x, 25, 26, 33, 36, 37, 38, 40, 52

[XRT20] Stanley C. Xie, Stuart A. Ralph, and Leann Tilley. K13, the Cytostome, and Artemisinin Resistance. *Trends in Parasitology*, 36(6):533–544, June 2020. URL: https://linkinghub.elsevier.com/retrieve/pii/S1471492220300714, doi:10.1016/j.pt.2020.03.006. 5, 6

[YMMS+21] Wencheng Yin, Luis Mendoza, Jimena Monzon-Sandoval, Araxi O. Urrutia, and Humberto Gutierrez. Emergence of co-expression in gene regulatory networks. *PLOS ONE*, 16(4):e0247671, April 2021. URL: https://dx.plos.org/10.1371/journal.pone.0247671, doi:10.1371/journal.pone.0247671. 22, 23

[ZIC14] Alice Zuin, Marta Isasa, and Bernat Crosas. Ubiquitin Signaling: Extreme Conservation as a Source of Diversity. *Cells*, 3(3):690–701, September 2014. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. URL: https://www.mdpi.com/2073-4409/3/3/690, doi:10.3390/cells3030690. 44