Chapter I

Introduction

“By heaven, I'll make a ghost of him that lets me”
— William Shakespeare, Hamlet

Malaria is a deadly global disease. Every year the *Plasmodium* parasites are responsible for 219 million cases of malaria and over four hundred thousand deaths, mostly vulnerable young African children under 5 years old and pregnant mothers [1]. After decades of steady progress malaria incidence is no longer declining. According to the 2018 World Malaria Report, the number of malaria cases increased from 217 million cases in 2016 to 219 million cases in 2017, resulting in 435,000 deaths [1]. Furthermore, in 2017, 3.5 million more malaria cases were recorded in the 10 African countries with the highest disease burden. In the Americas, some countries are also seeing large increases in prevalence, namely Brazil, Nicaragua, and Venezuela [1].

Most malaria-related deaths (93%) are concentrated in Africa; particularly so (76%) in the 17 sub-Saharan countries with the highest incidence of disease. The stark death toll is due to the high prevalence of *Plasmodium falciparum* malaria – the most virulent form of the disease – in the African continent. Here, *P. falciparum* causes 99.7% of malaria cases, while outside of Africa *Plasmodium vivax* is the most common infection. Other *Plasmodium* species can also cause human disease, such as *P. malariae*, *P. knowlesi* and *P. ovale*, but their prevalence is lower [5]. In all cases, *Plasmodium* parasites are transmitted to humans through the bite of an infected mosquito of the *Anopheles* genus. *Anopheles gambiae* and *coluzzii* (former *A. gambiae* molecular “M form”[6]) are the main vectors of *P. falciparum* malaria in Africa. However, over 30 *Anopheles* species are major disease vectors in other geographic regions.

If left untreated, malaria is a chronic and often deadly infection, as the human immune system is unable to achieve sterile immunity. Only after years of exposure and repeated bouts of infection is the immune system able to contain the parasite, resulting in chronic

asymptomatic infections. During the first 5 years of life, malaria infections often cause symptoms such as fever and anemia, and a proportion of children (1-2%) develop life-threatening severe disease and can die[7]. Mild or asymptomatic disease is not observed until early adolescence, but febrile episodes can still occur. Only as they reach adulthood, do people transition to a chronic state where disease symptoms rarely occur [2, 8]. Because sterilizing immunity never develops, adults maintain asymptomatic blood-stage malaria infections throughout their lives. They also become asymptomatic gametocyte carriers, infecting mosquitoes in their communities and maintaining the transmission cycle [9]. Malaria immunity in humans is complex, involving early development of protection from severe disease, followed by asymptomatic uncomplicated disease, but rarely, if ever, involving complete resistance to infection. Similarly, complex innate immune responses to *Plasmodium* occur in *Anopheles* mosquitoes, the outcome of which determines disease transmission.

Recent calls for malaria eradication have led to considerable strides in controlling this deadly disease[10], but we are far from defeating it. Importantly, we might not even yet have the right tools for such a goal, as the first approved vaccine (RTS,S/AS01) only provides partial (32-41%) protection [11], and the current arsenal of anti-malarial drugs is becoming less effective as *Plasmodium* resistance spreads[12]. Two vector-control strategies, insecticide-treated nets (ITN) and indoor residual spraying (IRS), have been key for the successful reduction of the burden of malaria in the last ten years, but these gains are in peril as mosquitoes develop insecticide resistance[13].

The reproductive rate (R_0) is defined as the number of new infection one case can generate, on average, over the course of its infectious period. It has been recently calculated that the R_0 for malaria ranges from 1 to over 3,000, depending on location, parasite species, populations and vectors [14]. To put the number into perspective, the flu has an R_0 of 10. By definition, for malaria to be eradicated, R_0 has to drop below 1. From a public health perspective, that means no “one-size-fit-all” approach can work, while from a researcher stand-point, it indicates that vaccines alone might be insufficient to eradicate malaria in areas of high

transmission. New control strategies that reduce the rate of re-infection, such as transmission-blocking vaccines[15], or the use of *Metarhizium* fungi that rapidly kill insecticide-resistant mosquitoes[16], will be required to achieve eradication. Crucial to all such eradication efforts is a better understanding of the determinants of malaria transmission by mosquitoes. There is ample evidence that mosquitoes have the potential to mount effective anti-plasmodial immune responses[2]. The mosquito relies on epithelial, humoral and cellular innate immune responses, coordinated by the hemocytes, the equivalent of the human white blood cells in insects [17]. My thesis project involves the development of an atlas of the mosquito cellular immune system at single-cell resolution as it responds to *Plasmodium* infection.

Introduction

1 The malaria parasite

"There's more beauty in truth, even if it is dreadful beauty"
— John Steinbeck, East of Eden

The malaria parasite and mankind are old foes. A recent study has shown the parasite co-evolved with gorilla parasites and went through an evolutionary bottleneck when it gained the ability to infect human hosts between 40,000 and 60,000 years ago. Human-adapted *P. falciparum* thrived when the human population exploded 5,000 years ago thanks to advances in farming [18]. As such *P. falciparum* infection has exerted a strong selective pressure on human populations worldwide, perhaps more so than any other pathogen [19]. No example is more widely studied in medical schools all around the world than sickle cell anemia. Recent medical advances have transformed the life of sickle-cell disease patients, whose life expectancy keeps rising and was estimated at 57 years in the US (2014) and 60 years in the UK (2016) [20, 21]. However, the homozygous hemoglobin S (HbS) variant was historically uniformly lethal in children, and yet was still maintained in the population at a frequency of around 15%, thanks to partial protection against severe malaria in heterozygotes [22]. But the parasite also had to adapt to the mosquito vector. For example, the *Plasmodium* surface protein Pfs47 allows the parasite to evade the mosquito innate immune system. It is thought that for a parasite to be transmitted, it requires a Pfs47 haplotype compatible with the Pfs47 receptor of the mosquito. As such, Pfs47 functions as a molecular "key" that turns off mosquito immunity through interaction with a receptor ("the lock") specific for each vector species. Only the right "lock and key" combination allows parasites to survive in the mosquito and propagate [23]. The parasite's life cycle is exceedingly complex, in both its human and mosquito hosts. Hence, many more such host-parasite immune interactions surely remain to be discovered. One of the most widely used animal models to study host-parasite interactions, as well as *Plasmodium* life cycle and development in the mosquitoes is *P. berghei*, a malaria parasite that infects mice. I used this experimental model system extensively in my PhD thesis to investigate the transcriptional response of the mosquito hemocytes to *Plasmodium* infection.

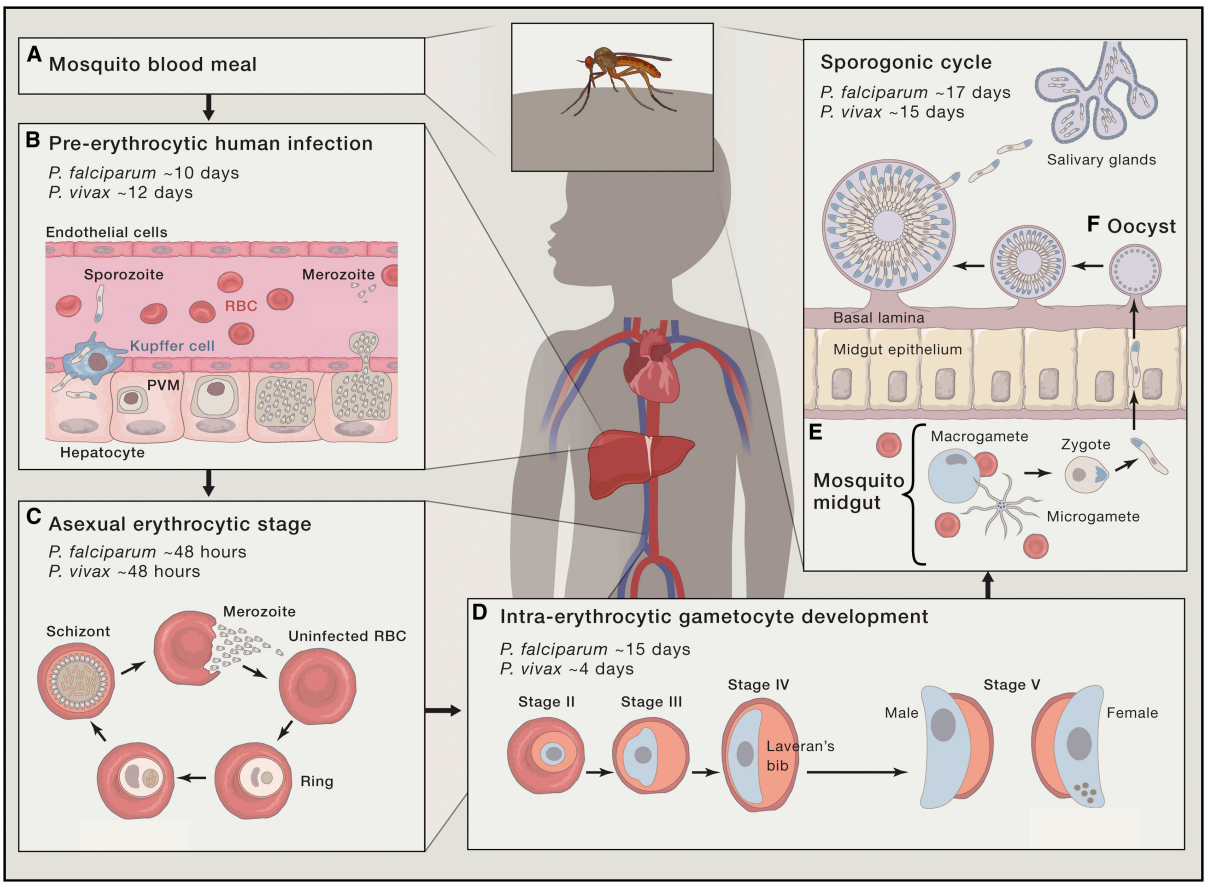


Figure I.1: ***Plasmodium falciparum*** **life cycle.** (A) Human malaria infection starts after an infected mosquito feeds and releases sporozoites from the salivary gland. (B) Some sporozoites escape the dermis, reach a blood vessel, travel to the liver through sinusoids or Kupffer cells and infect hepatocytes where they form a parasitophorous vacuole membrane (PVM) and undergo schizogony to released thousands of merozoites. (C) These merozoites travel in the blood and infect red blood cells, where multiple cycle of asexual reproduction (schizony) will occur before another burst and the repeat of the cycle. (D) Some merozoites are activated to differentiate to sexual gametocytes. (E) After bone marrow sequestration and maturation mosquitoes ingest gametocytes. In the midgut of mosquitoes, male and female gametocytes mate and form a zygote. In 24 hours the resulting motile ookinete penetrates the mosquito midgut epithelium and encysts. (F) In the oocyst asexual sporozoites replicate, are released in hemocoel, and colonize the salivary gland. Figure adapted from Cowman *et al*. [24]

Introduction

a. Malaria life cycle in humans

Before exploring in detail the mosquito phase of *Plasmodium* life cycle, it will be useful to review the human stages of parasite development. After a mosquito bite, sporozoites that had already colonized the salivary gland are injected into the human dermis. Some are able to survive local immune responses and move into blood vessels, travelling to the liver. Here they cross the hepatic sinusoidal barrier (fenestrated endothelial cells and Kupffer cells - resident macrophages), thanks to the action of SPECT, SPECT2, CelTOS, PL, and GEST proteins [24–28]. Sporozoites activate by binding higher sulfated forms of heparin sulfate proteoglycans (HSPGs), tetraspanin CD81, and scavenger receptor B1 (SR-B1) on hepatocytes [29, 30]. As circumsporozoite proteins (CSP) bind HSPGs, hepatocyte invasion commences [31]. Over the next 2-10 days sporozoites will morph into liver-stage (LS) schizonts, an exo-erythrocytic form (EEF) stage in which the parasite multiplies, eventually releasing over 40,000 merozoites per infected hepatocyte into the circulatory system [32].

Once released, merozoites infect circulating red blood cells in a three-step process [33]. The first and least understood is pre-invasion, in which low affinity binding between merozoites and erythrocytes orient the apical end of parasites towards red blood cells (RBCs) [34]. Then, specific binding mediated by erythrocyte binding-like proteins (EBA) and reticulocyte-binding protein homologs (PfRh) leads to actomyosin-driven host cell deformation and erythrocyte invasion [33, 35]. A PfRh5-PfRipr (Rh5-interacting protein) – CyRPA (cysteine-rich protective antigen) – basigin complex mediates the close interaction between erythrocyte and merozoite membranes, leading to microneme secretion and Ca^{2+} influx inside the red blood cells [36–39]. Merozoites are then irreversibly linked to erythrocytes through AMA1-RON tight junction complexes [40]. These are moving junctions, propelling the merozoites inside red blood cells just as rhoptry contents are released, which form the parasitophorous vacuole membrane (PVM) around the merozoites [41]. As the PVM seals, cytosolic water losses within host cells cause echinocytosis. Over the next 48 hours the

parasites take advantage of the established nutrient-rich cellular milieu to rapidly divide and produce 16 to 32 merozoites each, which then egress as they destroy the RBCs [24].

During these rounds of cellular replication (schizogony), a small proportion of *Plasmodium* parasites will differentiate into sexual forms, a required step for successful transmission to mosquito vectors. Male and female gametocyte differentiation is not fully understood, but is regulated by the master switch AP2-G [42] following sensing of environmental signals such as high parasitemia or presence of chloroquine in the blood stream. Gametocyte development lasts 11 days, during which time committed but not yet mature gametocytes hide sequestered in the bone marrow to avoid splenic clearance. Following development, mature stage V gametocytes are taken up by feeding mosquitoes to commence the mosquito life cycle [43] [Fig. I.2].

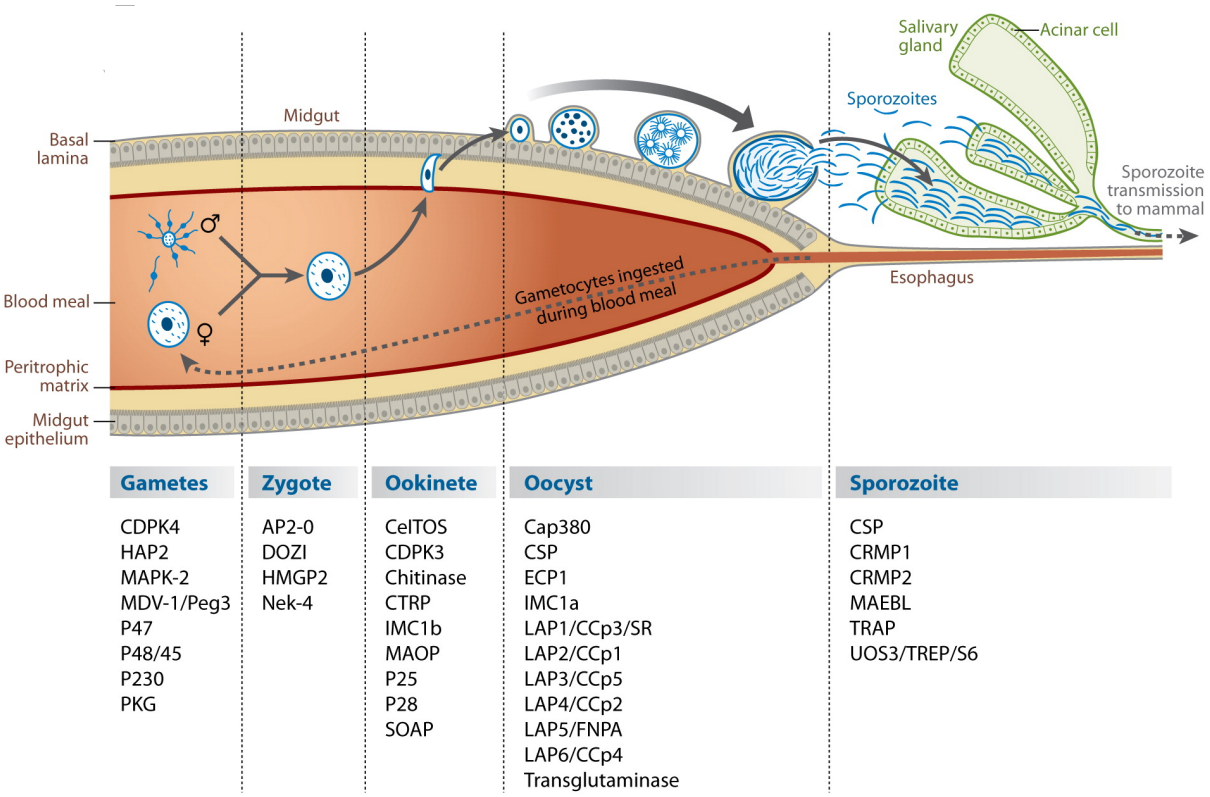


Figure I.2: Detailed *Plasmodium* life cycle in the mosquito and key parasite proteins. (A) *Plasmodium* gametocytes are first ingested when mosquitoes take an infected human blood meal. Gametes, both female (macrogametes) and male (microgametes) mature from gametocytes to form a zygote (B). After meiosis the zygote morphs into the motile, infective ookinete (C) which is able to penetrate through the mosquito midgut. After egressing from the midgut *Plasmodium* ookinetes encyst on the basal end, becoming sessile (D). After 10-14 days of growth and mitotic divisions, thousands of motile sporozoites are released into the mosquito circulation, travelling in the hemolymph until some reach the mosquito salivary glands (E). Here sporozoites attach to the basal side of salivary gland acinar cells, travel through them, and enter the ducts, where they await the next mosquito bite to continue the life cycle. Key proteins at each step are listed above. Figure adapted from Aly *et al.* [44]

b. *Plasmodium* life cycle in mosquitoes

Parasites ingested during a blood meal quickly undergo sexual reproduction in the mosquito midgut [44]. Gametogenesis starts the moment gametocytes are ingested by feeding mosquitos, leading to the formation of mature male and female gametes [45]. Gametogenesis is mediated by essential environmental signals such as a 5°C drop in temperature, the rise of extracellular pH (from 7.2 to 8), and xanthurenic acid (XA) sensing [46–50]. XA – a byproduct of mosquito metabolism – activates guanylyl cyclase, leading to increased second messenger cGMP production and protein kinase G (PKG) activation[51, 52]. In addition, XA increases inositol-(1,4,5)-trisphosphate (IP_3) production by activating phospholipase C, causing the opening of Ca^{2+} channels [53–55]. Heightened intracellular Ca^{2+} releases translational repression in male and female gametocytes by activating Ca^{2+}-dependent protein kinase 1 (CDPK1) [56]. Gametocyte activation is rapid, and within 15 minutes gametocytes egress from red blood cells by rupturing first the PVM and then the erythrocytic membrane (EM), steps respectively associated with osmiophilic bodies and egress vesicles [50]. The former is mediated by Pg377, MDV-1/Peg3 and GEST [57–59] while egress vesicles release perforin, which breaks the EM to release fertile gametes [60]. Activated microgametocytes undergo three rounds of replication, becoming octaploid, and producing eight flagellar mature microgametes by mitosis (exflagellation)[61]. Exflagellating microgametes adhere to nearby red blood cells, hiding within rosettes before detaching from the residual body, searching for macrogametes [45]. When a partner is found Pfs47, Pfs48/45, and Pfs230 proteins form complexes responsible for the binding of microgametes and macrogametes, commencing fertilization [62–64]. First, the plasma membranes of the two gametes fuse. The axoneme and male nucleus then enter the female cytoplasm, mediated by HAP2. Finally, nucleus fusion ensues, followed by meiosis and the production of a tetraploid zygote (as mediated by NIMA-related kinases Nek-2 and Nek-4)[65–67].

Next, the zygote morphs into a motile ookinete able to colonize mosquitoes. After fertilization ninety-one proteins were found to be specifically expressed, with silencing of

paternal alleles in both zygotes and ookinetes. These changes are orchestrated by the transcription factor AP-2O, the master regulator of ookinete development and motility. Furthermore, AP-2O also plays a key role in penetrating the mosquito midgut epithelium and protecting ookinetes from immune defenses [56, 68–70]. Among the upregulated genes are secreted proteins such as perforins (PPLP3, 4 and 5), PSOP1, 2, 6, 7 and 12, and SOAP, as well as WARP, POS1-10, and P25, P28 – all potential or actual targets of transmission-blocking vaccines [56, 70, 71]. Ookinete maturation completes between 19 and 36 hours after an infectious blood-meal, after which ookinetes exit the gut [72–74]. Ookinete motility is regulated by PKG and CDPK3 activity, as well as cGMP and Ca^{2+} levels [75–78]. In order to successfully infect a mosquito ookinetes must first penetrate the peritrophic membrane (PM) – a chitin structure that functions to protect mosquitoes from bacteria and gross food [79]. To break through the PM *Plasmodium* ookinetes produce a chitinase that is able to hydrolyse the chitin [80–82]. Traversal of midgut epithelial cells is mediated by CTRP, a protein secreted by the ookinete to form a bridge between the midgut epithelium and the actin/myosin motor of the ookinete [83–85]. Three perforins (PPLP3-5), SOAP, WARP, MAOP and CelTOS are all microneme proteins required to breach the epithelial membrane [70, 85–91]. Once ookinetes have crossed the midgut epithelium they are surrounded by the laminin and collagen of the basal lamina. The interaction of *Plasmodium* with laminin turns ookinetes sessile, which encysts on the basal side of the midgut epithelium, triggered by the proteins P25 and P28, with help from CTRP and SOAP [71, 92]. Cell transversal is a bottleneck, and only a few ookinetes are successful in invading the mosquito midgut [88, 93, 94].

Oocyst development and maturation lasts between 10 and 12 days, and is the only stage of the life cycle where the parasite is extracellular for an extended period of time. And yet, little is known of host-oocyst interactions. Nutrients flow through the oocyst capsule, formed by an outer layer of thick mosquito laminin, parasite transglutaminase, Cap380 (oocyst capsule protein)[95], and P25/P28, and an inner oocyst plasma membrane containing Cap93[96] and circumsporozoite protein (CSP), a GPI-anchored protein[97]. Oocysts grow to 50-60 μm in diameter, forming thousands of sporozoites after multiple rounds of mitotic divisions, mediated

by LAPs (LCCL/lectin adhesive-like proteins) expression [98]. CSP is essential in building syncytial lobes - called sporoblasts, coordinating the localization of microtubule organizing centers (MTOC) underneath sporoblast membranes to make mature sporozoites[99, 100].

Once sporozoites reach maturity, they egress from oocysts, in a process that involves digestion of the capsule mediated by a cysteine protease, ECP1 (egress cysteine protease 1)[101, 102]. CSPs and a hypothetical oocyst protein also have important roles in sporozoite release[103, 104]. Sporozoites in the hemocoel are then carried to all tissues in the mosquito body by circulatory flow. Some are deposited to the basal lamina of the salivary gland, where CSP again plays a role in attachment [105, 106]. Thrombospondin-related anonymous protein (TRAP) is essential for attachment and invasion[102, 107], binding to saglin, a mosquito receptor in the distal lobes of the salivary gland [108, 109]. Gliding motility and the actin-myosin motor are also involved in invasion, mediated by TRAP and TREP (TRAP-related protein), as well as cysteine repeat modular proteins (CRMP1 and 2) and MAEBL [110–112]. Cellular invasion mechanisms are largely conserved between human and mosquito life stages of the parasite. However, while MAEBL-deficient sporozoites can still invade human host cells, TRAP-deficient parasites cannot. In fact, acinar cells in the salivary gland are invaded by a slightly different mechanism than the parasitophorous vacuole (PV) involved in liver and blood-stage invasion. Rather, invasion happens through a specific vacuolar membrane produced by the host cell [113]. How the sporozoite is able to induce vacuole formation in the salivary gland epithelial cells is unknown.

2 Mosquito immune responses

"The world is, of course, nothing but our conception of it"
— Anton Chekhov

Throughout the life of *Plasmodium*, the mosquito is far from being a passive vector. There are active interactions between parasites and the immune system at each step of the life cycle, especially when *Plasmodium* is extra-cellular. In fact, insects are constantly exposed to a wide variety of micro-organisms and pathogens seeking to exploit the host mosquito for their own reproductive goals. Viruses, fungi, bacteria, protozoans, and nematodes all invade and infect mosquitoes [114–118]. Some pathogens are able to penetrate through the external exoskeleton of mosquitoes, formed by hydrophobic chitin, which also lines the foregut, hindgut, and tracheas. They accomplish invasion by degrading the cuticle [119, 120]. Other pathogens enter mosquitoes through the digestive tract, overcoming physical barriers such as pharyngeal armatures and the chitinous peritrophic matrix, as well as digestive enzymes, local microbiota, and a hostile pH. Some pathogens evolved mechanisms to penetrate through these defensive mechanisms to reach the hemocoel (blood) of the mosquitoes and replicate, while others remain within the gut itself. Mosquitoes have however developed sophisticated immune mechanisms to fight off and control these pathogens [117, 119, 121–123]. Mosquito hemocytes, the equivalent of human white blood cells, coordinate both cellular and humoral immune responses. Humoral immune responses are mediated by molecules that are secreted into the circulating mosquito hemolymph (equivalent to serum in vertebrates) by hemocytes, fat body cells or epithelial cells lining the haemocoel, such as midgut and salivary gland cells). For example, pattern recognition receptors (PPRs), phenoloxidase cascade components, antimicrobial peptides, and elements of the complement-like system are all key mediators of the mosquito humoral response [3, 124–127]. The different components of the immune system are all interconnected, crafting an exceedingly complex and well-coordinated immunological network able to kill pathogens by a variety of effector mechanisms [Fig. I.3].

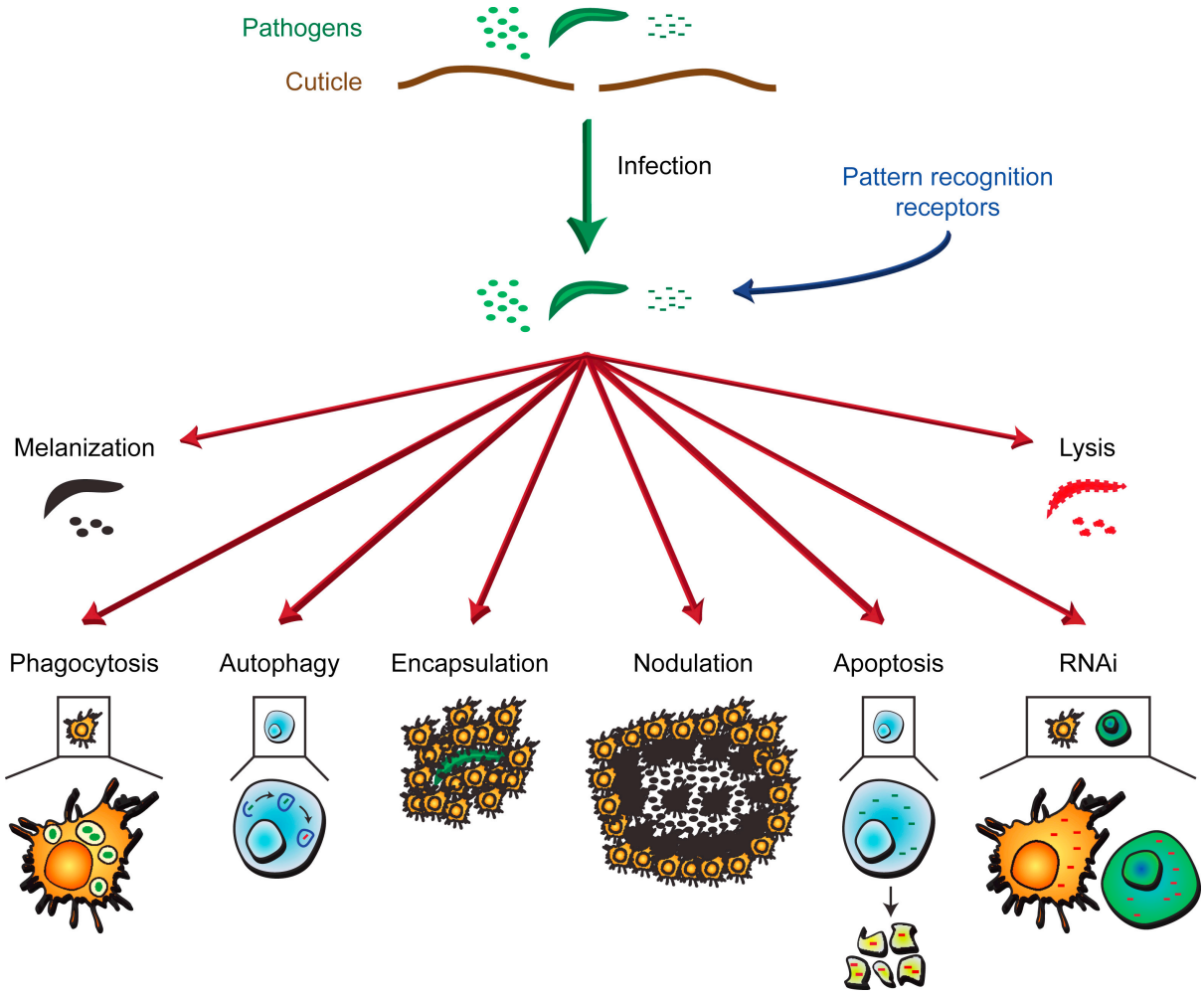


Figure I.3: Mechanisms of immune killing. Mosquitoes kill pathogens by melanisation, lysis, phagocytosis, autophagy, encapsulation, nodulation, apoptosis, and RNA interference. Adapted from Hillyer *et al.* [117]

a. Humoral immunity

Mosquitoes lack antibodies, but can activate highly effective humoral mechanisms to control infection. For instance, antimicrobial peptides (AMPs) – such as defensins, cecropins, gambicin, attacin and holotricin – are small charged molecules that are secreted into the hemolymph, with strong anti-bacterial or anti-fungal effects [128]. The composition and spatial expression of the "cocktail" of antimicrobial peptides secreted in response to an immune

challenge can differ widely among different mosquito species. For example, defensins can reach a concentration of 45 μM in *Aedes* [129], while in *Anopheles* they reach a maximum of 1-5 μM [130]. Furthermore, in *An. gambiae* antimicrobial peptides are produced at higher concentrations in the anterior midgut, and indeed *Plasmodium* oocysts colonise the posterior midgut[131]. Conversely, heightened AMP production can reduce parasite load. For instance, cecropin A (CecA) was shown to lower *P. berghei* oocyst counts in transgenic *An. gambiae* overexpressing CecA under the control of the *Aedes* carboxypeptidase promoter [132].

Melanisation is another powerful humoral mechanism to control infection. It is a biochemically conserved pathway to produce eumelanin from tyrosine and 5,6-dihydroxyindole (DHI) catalyzed by a cascade of prophenoloxidases (PPOs). The PPO enzymatic cascade leads to killing both by starving the invading pathogen of nutrients – walling it off from the rest of the body – as well as through the direct toxic effects of chemical byproducts. Melanisation is also involved in cuticle hardening and wound healing. Furthermore, it causes hemocytes to aggregate – an immune response akin to human granulomas, called in mosquitoes nodulation or encapsulation [Fig. I.4][128, 133–136]. The melanisation pathway begins with PRR sensing (C-type lectins, Gram-negative bacteria-binding proteins and beta-1,3 glucan recognition proteins), followed by a serine protease cascade leading to the activation of prophenoloxidases (mostly expressed by oenocytoids, a hemocyte subtype [136]). PPOs in turns activate melanin production by phenoloxidase. Melanisation is tightly regulated by serpins and C-type lectins. Similarly to human clotting, excessive activation would be deadly to the mosquito as widespread melanisation would damage the mosquito organs [133, 137, 138].

The complement-like pathway is one other crucial humoral effector mechanism, resulting in deposition of thioester-containing protein 1 (TEP1), a C3-like opsonin, on the surface of the microbe [139, 140]. Another four important proteins of the pathway are: two leucine-rich repeat proteins (APL1C and LRIM1), which stabilize TEP1 in circulation, and two clip domain serine protease homologs, (SPCLIP1 and CLIPA2), that modulate TEP1 activation. CLIPA2 is a negative regulator of TEP1, while SPCLIP1 promotes TEP1 activation.

It has been proposed that, following binding of TEP1 and SPCLIP1 recruitment and activation onto the pathogen surface, an endogenous TEP1 convertase is also deposited that further propagates local activation and binding of TEP1 [137, 141, 141–144].

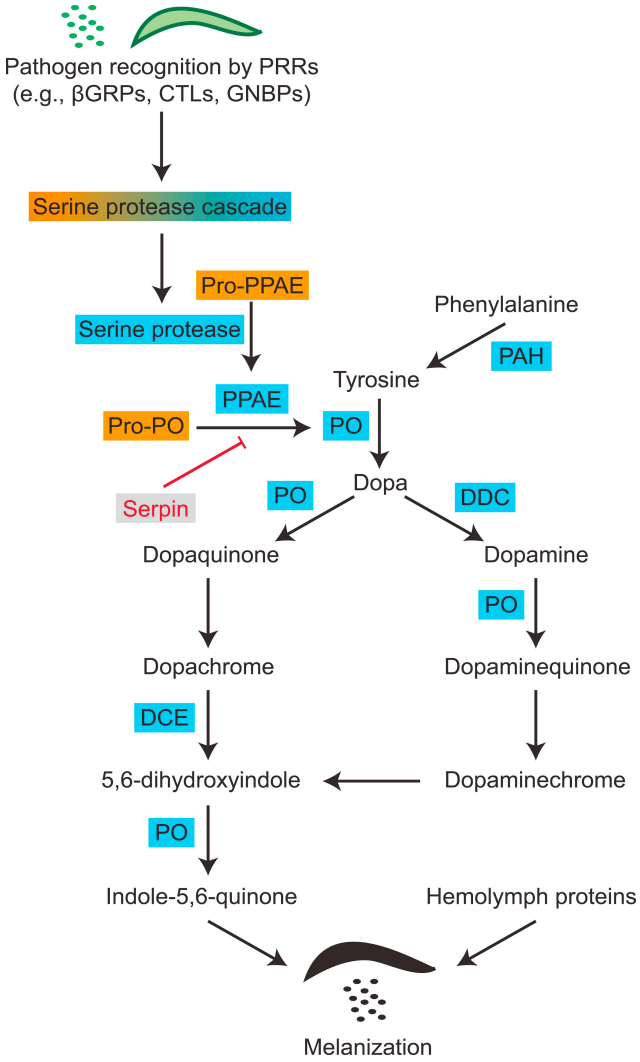


Figure I.4: Melanisation pathway. Abbreviations: PRR, pattern recognition receptor; βGRP, β-1,3 glucan recognition protein; CTL, C-type lectin; GNBP, Gram(−) binding protein; PPAE, phenoloxidase activating enzyme; PAH, phenylalanine hydroxylase; PO, phenoloxidase; DDC, dopa decarboxylase; DCE, dopachrome conversion enzyme. Adapted from Hillyer et al. [117].

b. Cellular immunity

Hemocytes are the primary immune cells circulating in the mosquito hemolymph. Mosquitoes have no antibodies or canonical adaptive immunity and entirely rely on innate immunity. Cellular responses are quick, with direct interactions between pathogens and the immune cells that include phagocytosis, encapsulation, and nodulation [117, 118, 145]. Experiments returned wildly discordant estimates of the total number of hemocytes in a mosquito, mostly due to technical differences in the collection methodology employed. The scientific community agrees that between 2,000 and 10,000 hemocytes patrol a mosquito, although only a fraction is motile (~ 500 - 2,000) and numbers vary considerably with blood-feeding and infection [4, 118, 146]. After morphological, enzymatic, and some functional characterization the consensus is that three main hemocyte subtypes exist: granulocytes, oenocytoids, and prohemocytes [Fig. I.5]. Of these, the vast majority are small prohemocytes (60-70%), followed by oenocytoids (20-30%) and granulocytes (1-10%), although estimates again vary considerably [4, 17, 17, 118]. Because the classification is largely morphological, subjective differences in interpretation and methodology are inevitable. Granulocytes are the main effector phagocytic cells in the mosquito, expressing AMPs, complement-pathway components, and low-level PO. Oenocytoids contain PPOs and POs at much higher levels. Prohemocytes are still a mystery. Originally thought to be progenitor cells, they have recently been shown to possess phagocytic capabilities, and are hypothesized to arise from asymmetric cell division of granulocytes [146].

While hemocytes remain in the hemocoel, and do not come in direct contact with the microbiome in the midgut lumen, transient bacteremia following blood feeding is thought to activate hemocyte replication after a blood meal [147–149]. A bacterium, yeast, fungus, or malaria parasite in the body cavity of a mosquito is usually quickly tagged, identified, and ingested by phagocytic hemocytes. Furthermore, hemocytes have been shown to aggregate around bacteria and form nodules [150]. Worms, fungi, or parasites become surrounded by melanocytic capsules [151–153]. In *Aedes aegypti* mosquitoes, hemocytes are also thought to play an important role in the systemic dissemination of arboviruses such as Sindbis or dengue virus [154, 155]. Hemocytes release microvesicles at sites of *Plasmodium* ookinete midgut

invasion that reduce parasite survival by promoting local activation of the complement-like system [156], thus coordinating epithelial and humoral antiplasmodial immune mechanisms to achieve an integrated and effective response.

Very little was known of hemocyte development in mosquitoes, except that blood-feeding increases their numbers[147]. In Drosophila, hematopoiesis is thought to occur in three waves: embryonic, larval, and lymph gland[157]. The first two are thought to be responsible for routine phagocytic and immunological functions of mosquitoes, whereas lymph gland hemocytes arise from synchronous differentiation of progenitors hemocytes within the gland following immune and environmental challenges. Crucially, after hemocytes differentiate these lymph glands disintegrate before adulthood[157]. Lymph glands have not been observed in mosquitoes.

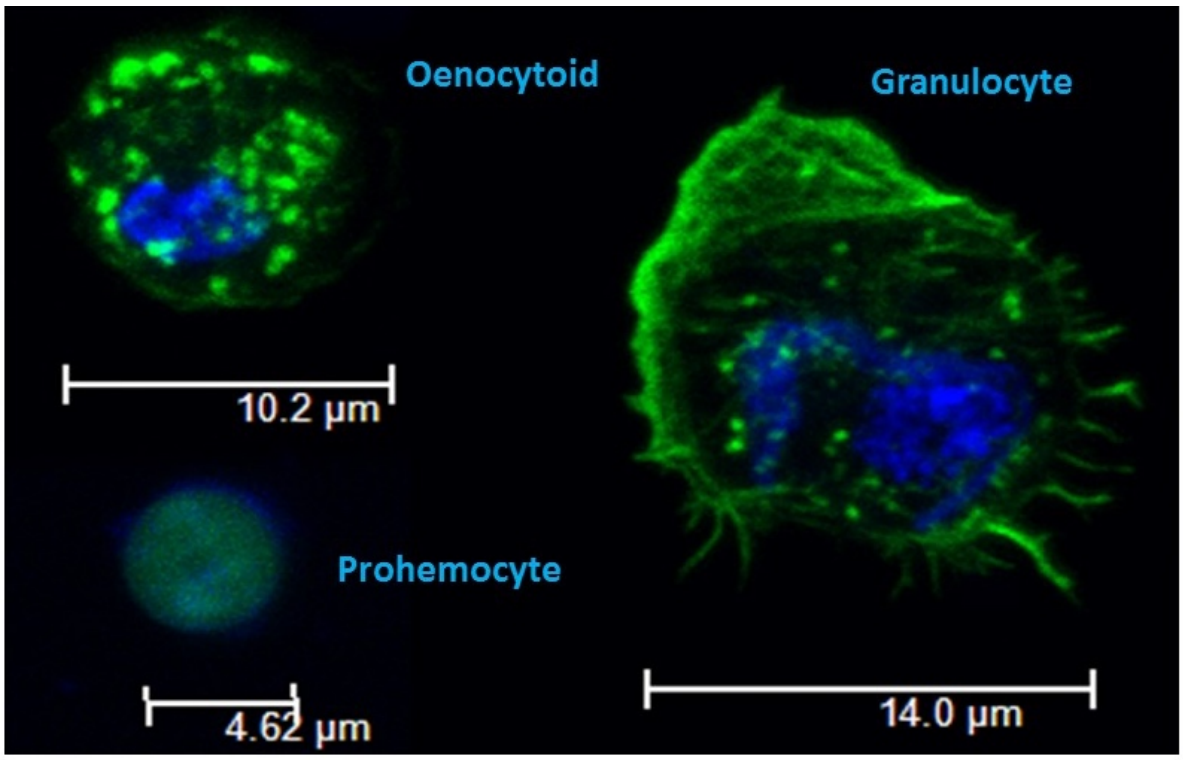


Figure I.5 Hemocyte subtypes. Average diameter and representative images of the three morphological subtypes of hemocytes in *A. gambiae*. Personal communication from Jose Luis Ramirez.

3 Specific *Anopheles* immune responses to *Plasmodium*

a. Midgut epithelial defenses

Mosquitoes ingest a very large, protein-rich meal during blood feeding which has to be digested to meet the nutritional needs of developing oocytes. At the same time, mosquitoes build a protective peritrophic matrix (PM), an acellular/semi-permeable layer of chitin polymers, that surrounds the blood meal and prevents direct contact between the microbiota in the lumen and the gut epithelium. The PM is also an important barrier against potential pathogenic organisms. At the end of the digestive process the PM matrix sack – containing remnants of the digested blood meal – is excreted from the mosquito. A new matrix will be formed when the mosquito ingests the next blood meal [158]. Mosquitoes also secrete mucin in the ectoperitrophic space between midgut epithelium and the PM, and actively modulate the permeability of this mucous layer through the activity of an immune-modulatory peroxidase (IMPer)/dual oxidase (Duox) system that catalyzes dityrosine cross-linking [2, 159, 160]. IMPer is secreted into the ectoperitrophic space, but is only active when immune elicitors activate Duox, which generates hydrogen peroxide a substrate required for IMPer to catalyze the formation of the dityrosine network. The end result is a dynamic modulation of the interaction between the immune elicitors released by gut flora and the gut epithelium following a blood meal, that allows the bacterial flora to survive by preventing constant activation of antibacterial immunity. This system also benefits *Plasmodium* parasites, because it allows them to develop within the midgut lumen without activating nitric oxide synthase (NOS) expression in epithelial cells. If this barrier is disrupted by silencing IMPer, mosquitoes mount a much stronger epithelial nitration response that eliminates *Plasmodium* [159, 161, 162].

b. Reactive oxygen/nitrogen species and complement-like defenses

Plasmodium ookinetes must traverse the midgut epithelium to complete their development in the mosquito. In doing so, they breach the peritrophic matrix, allowing the microbiota to come in direct contact with epithelial cells and cause irreversible damage as they invade midgut cells. Invaded mosquito epithelial cells express high levels of NOS, a response which is necessary,

but not sufficient to mount an effective response against *Plasmodium*. Specifically, increased NOS leads to nitric oxide production [161], which is unstable and is thought to rapidly convert to nitrite, a more stable molecule that accumulates in the cell [127, 161]. Similarly to what is observed in vertebrate macrophages, NOS activation is followed by activation of a peroxidase-mediated nitration reaction that uses nitrite as a substrate [161]. This is a highly efficient nitration reaction catalyzed by HPX2 (Heme peroxidase 2), that requires high local levels of hydrogen peroxide – provided by NOX5 (NADPH Oxidase 5) – and nitrite as substrates [159, 161, 162].

This NOX5/HPX2-mediated nitration is part of an apoptotic response in invaded/damaged midgut cell via JNK signaling that activates caspases expression, and is essential for mosquitoes to activate an effective immune response to *Plasmodium* by the mosquito complement-like system [161, 163, 164]. TEP1(C3-like factor), a key effector of the complement-like system, is produced by the hemocytes and circulates in the hemolymph as a stable complex with two proteins of leucine-rich (LRR) family, LRIM1 and APL1[140, 162, 165–167]. TEP1, APL1 and LRIM1 are form a MW complex responsible for TEP1 deposition on the surface of pathogens that promotes phagocytosis or leads to the formation of a complex that will lyse ookinetes [144, 166, 167]. The precise mechanism of killing and complement-activation is not completely understood, however work from our laboratory revealed that nitration of epithelial cells and the midgut basal lamina triggers the release of hemocyte-derived microvesicles (HdMv) into the basal lamina labyrinth, that is critical for activation of complement-mediated *Plasmodium* lysis [156].

c. Vector susceptibility and *Plasmodium* immune evasion

There are broad differences in compatibility, the extent to which the mosquito immune system limits infection, between different *Plasmodium*/mosquito combinations[168]. Intriguingly, while all ookinetes must come into contact with TEP1 in the mosquito hemolymph, only some are lysed [165]. That begs larger questions: how does *Plasmodium* evade the immune system

of mosquito vectors? And why are some *Anopheles* mosquitoes more susceptible to infection than others? Mosquito susceptibility to *Plasmodium* infection has a strong genetic component. For example, the mosquito *A. gambiae L3-5* strain was genetically selected to be highly resistant to *P. cynomolgi* (monkey malaria infection), but is also highly refractory to infection with *P. berhgei*, by expressing a TEP1 allele, with heightened anti-parasitic effects [169]. In addition, R strains have been shown to be in a state of chronic oxidative stress with increased basal levels of ROS and overactivation of JNK signaling, that is exacerbated by blood feeding [170].

While higher systemic ROS levels in the R strain result in loss of longevity and fecundity, the immune response to *Plasmodium* invasion is well localized both in time and space [171–173]. Invaded midgut epithelial cells activate ROS and nitration pathways, but these responses are localized and do not "spread" to healthy adjacent cells [162]. Furthermore, detox enzymes such as MnSOD (manganese-dependent superoxide dismutase), Gpx (hydrogen peroxide detox), and catalase are highly upregulated in healthy midgut cells and throughout the mosquito body (e.g. fat tissues), thus controlling any potential spillover and preventing damage. At the same time, catalase levels are downregulated in infected midgut cells to allow accumulation of ROS [162, 171, 172].

Parasite genetics are just as important in understanding *Plasmodium* transmission and infection. Our laboratory showed that epithelial nitration and microvesicle release are key for *P. berghei* destruction via TEP1-mediated lysis [156, 162]. Intriguingly, some reports had shown little [174, 175] to no [176] effects in disrupting the complement-like system when *A. gambiae* is infected with *P. falciparum*. Later studies demonstrated that susceptibility of *P. falciparum* killing by TEP1 is a *Plasmodium*-strain and mosquito-species specific response [162, 177]. For example, the *A. gambiae L3-5* refractory strain activates the complement-like system and kills the *P. falciparum* 7G8 strain from Brazil, while the African GB4 strain is able to evade the mosquito immune system and survive [178].

The *P. falciparum Pfs47* gene – a member of the 6-cystein protein family expressed on the surface of female gametocytes and ookinetes – allows the parasite to evade the immune responses mediated by TEP1: parasite killing with subsequent melanisation the *A. gambiae L3-5* refractory strain, as well as lysis without melanisation in the susceptible G3 strain [177]. *Pfs47* is polymorphic and exhibits a marked population structure and extreme fixation in non-African regions [179, 180]. The global populations structure of *Pfs47*, together with our laboratory experiments infecting anopheline mosquitoes vector species from different continents, provided strong evidence that distinct *P. falciparum* Pfs47 haplotypes were selected to be compatible with different mosquito vectors [177]. We then proposed the "lock-and-key theory", where *Pfs47* is a "key" that allows *P. falciparum* to evade the mosquito immune system by interacting with a mosquito receptor ("the lock"), different in each evolutionarily distant *anopheline* species [23, 181]. Only those parasites with a *Pfs47* haplotype compatible with a given mosquito species are able to evade the mosquito immune system, and this allows them to survive and become established in a given geographic area.

d. Signaling pathways of immune evasion / antiplasmodial immunity

Further work examined the mechanism through which *Pfs47* affects the response of the mosquito immune system to *Plasmodium* infection. JNK promotes TEP-1 lysis by inducing expression of HPX2 and NOX5 in midgut cells invaded by *P. berghei* ookinetes [182]. However, *Pfs47* disrupts JNK signaling, preventing caspases activation and downstream midgut nitration in response to *P. falciparum* invasion [163, 170]. A recent study has shown that in *P. berghei* Pbs47 is also required for ookinetes to avoid destruction by the complement-system [183]. *P. falciparum* ookinetes that do not express *Pfs47* activate JNK signaling, caspase activity and downstream epithelial nitration, triggering a strong activation of the mosquito complement system that is very effective killing the parasite [163, 184]. Other conserved immune-signaling cascades are important mediators of immune activation and killing of *Plasmodium*: *Toll***,** *Imd***,** and STAT. *Toll* and *Imd* activation promote TEP1-mediated lysis, but *Toll* appears to be more effective in limiting *P. berghei* (with silencing of repressor

protein *cactus*), while *Imd* is more effective against *P. falciparum*. These three pathways (*Toll*, *Imd*, JNK) all converge to TEP1 as the key effector of ookinete lysis [2].

If parasites evade and survive the early complement-mediated response, a different "late-phase" immune response is thought to further decreases parasite numbers by attacking the oocyst stage of *Plasmodium* [125, 185]. *Plasmodium* infection triggers a multi-pronged defense strategy by the mosquitoes, where an initial complement-mediated response that targets ookinetes is followed by activation of the STAT and LPS-induced TNFα transcription factor (LITAF)-like 3 (LL3) pathways that limits oocyst survival [185]. Interestingly, both STAT and LL3 seem to act independently. The STAT pathway is composed of STAT-B and STAT-A, with STAT-B regulating basal levels of STAT-A, which in turns regulated NOS, SOCS, and hemocytes differentiation. While STAT-dependent NOS expression reduces oocyst survival [125, 127, 185, 186], LL3-dependent midgut NOS induction has the opposite effect, increasing oocyst survival [186]. Other unknown factors are most likely at play, including the possibility of multiple isoforms of NOS [185].

In addition, while a double knock-down of SOCS (a suppressor of STAT) and NOS leads to higher oocyst survival than single SOCS silencing, single NOS knock-down unexpectedly leads to almost complete loss of infectivity due impaired epithelial cell invasion [125]. It appears that high levels of NOS are deleterious to oocysts, while a minimum level of NOS is required for ookinete midgut invasion to occur [125]. Finally, we still do not know what are the exact signals that lead to STAT and LL3 activation, although eicosanoids (see next section) or the wound-healing response might be implicated, as the AP-1/Fos-TGase2 axis has also been linked to increase TEP1 dependent *P. falciparum* killing [187].

e. Hemocytes are key coordinators of immunity in *Anopheles* and mediate mosquito immune memory

Our current understanding of both early and late cellular immune responses to *Plasmodium* is still limited, and much work is required to elucidate the precise molecular details of their immune effector functions. Furthermore, despite the ability of hemocytes to coordinate immune responses and respond to a variety of insults, including wound healing, and viral, bacterial, fungal, and parasitic infection, their exact molecular role in anti-*Plasmodium* immunity remains largely unknown [3, 185, 187–190]. We briefly discussed how three morphologically distinct subpopulations of hemocytes are believed to exist in *Anopheles*: the prohemocytes (putative undifferentiated precursors), granulocytes (phagocytic hemocytes), and oenocytoids (characterized by phenoloxidase activity) [4]. We have also discussed how hemocytes participate in the immune response against *Plasmodium* through cellular and humoral effector mechanisms. But what is the role of specific cell types in *Anopheles* defense mechanisms? And do only three cell types really exist? Already some recent studies suggest hemocytes could harbor greater complexity than originally thought, with three phagocytic subtypes found within $PPO6^{low}$ populations (equivalent to morphologic granulocytes)[191, 192].

Besides their conventional role as effectors of mosquito innate immunity, hemocytes have also been shown to mediate immunological memory. This phenomenon is called ‘immune priming’, and is defined as the ability of mosquitoes that have been infected with *Plasmodium* to develop a life-long, systemic state of enhanced immune surveillance, with an increased proportion of circulating granulocytes – the phagocytic cells that are more similar to vertebrate macrophages – which enhances their immune response to subsequent infections [193]. In addition, there are changes in the morphology and binding properties of granulocytes, with larger and more granular cytoplasm, pseudopodial extensions, and increased lectin-binding capabilities [193]. Interestingly, NK cells in vertebrates have recently been shown to also possess similar mechanisms [2]. The priming response in *A. gambiae* is activated when

Introduction

Plasmodium ookinetes breach the gut barrier and come into contact with the epithelial midgut cells, and damage to *Plasmodium* is due to a bystander effect [193]. Primed mosquitoes mount a stronger antiplasmodial response by greatly increasing the release of hemocyte-derived microvesicles [193].

Indeed, our laboratory has shown immune priming to be a 'two-step approach.' First, ookinete invasion induces expression of HXP7 and HPX8, two heme-peroxidases that catalyze prostaglandin E_2 (PGE_2) synthesis. Epithelial cells release PGE2 into the hemolymph and this attracts hemocytes to the basal surface of the gut. The chemotactic response is then followed by enhanced patrolling activity of the midgut basal lamina. If hemocytes detect a nitrated surface, they undergo apoptosis and release microvesicles into the basal labyrinth space, in close proximity to parasites that have traversed the midgut. The exact contents of these vesicles remains to be elucidated, but their release is essential for effective activation of TEP1-mediated anti-*Plasmodium* immunity [156, 194].

Following immune activation, primed mosquitoes constitutively release a hemocyte differentiation factor (HDF), and this factor persists in the hemolymph for the entire life of the mosquito [195]. HDF consists of a lipoxin/lipocalin complex [194]. Lipocalins are a family of proteins involved in lipid transport, while prostaglandins and lipoxins are all part of the eicosanoid lipid family [196]. Eicosanoids possess important signaling roles in homeostasis, inflammation and immunity not only in mammals, but also microbes and invertebrates like *Anopheles* [197]. Interestingly, suppression of host eicosanoid synthesis has been shown to be a mechanism of immune evasion by bacteria [197, 198].

Our laboratory has shown that immune priming involves an increase in lipoxin production (especially lipoxinA_4) from arachidonic acid, as well as increased expression of evokin, a lipid carrier protein of the lipocalin family. In addition, it appears as if LL3 is also necessary for HDF production, as silencing it stops HDF release. Priming can also be abolished by interfering with the function or movement of hemocytes by injecting water, PBS or Sephadex beads into the hemolymph [193]. Importantly, hemocyte differentiation factor

(HDF) is sufficient for effective priming, as transfer of both hemolymph, cell-free hemolymph, and HDF alone leads to hemocyte differentiation in the mosquitos and transference of enhanced antiplasmodial immune capabilities [2, 193]. Interestingly, although priming is elicited when ookinete invasion allows direct contact of the gut microbiota with midgut epithelial cells, *Plasmodium* species differ in their ability to establish a priming response, depending on their compatibility with the mosquito vector. For example, *A. gambiae* G3 mosquitoes mount a stronger immune response to *P. yoelii* ookinetes than to *P. berghei*, while *P. falciparum* NF54 fails to elicit an effective immune response. *P. yoelii*, the parasites that triggers the strongest immune response, leads to the strongest priming, while infection with the highly compatible *P. falciparum* NF54 strain results in weaker priming than *P. berghei* infection [163]. It is not clear whether strong epithelial nitration in midgut epithelial cells or the release of microvesicles also enhances the long-lasting priming response of hemocytes. Much remains to be discovered regarding the role of eicosanoids and hemocytes in insect immunity and immune memory.

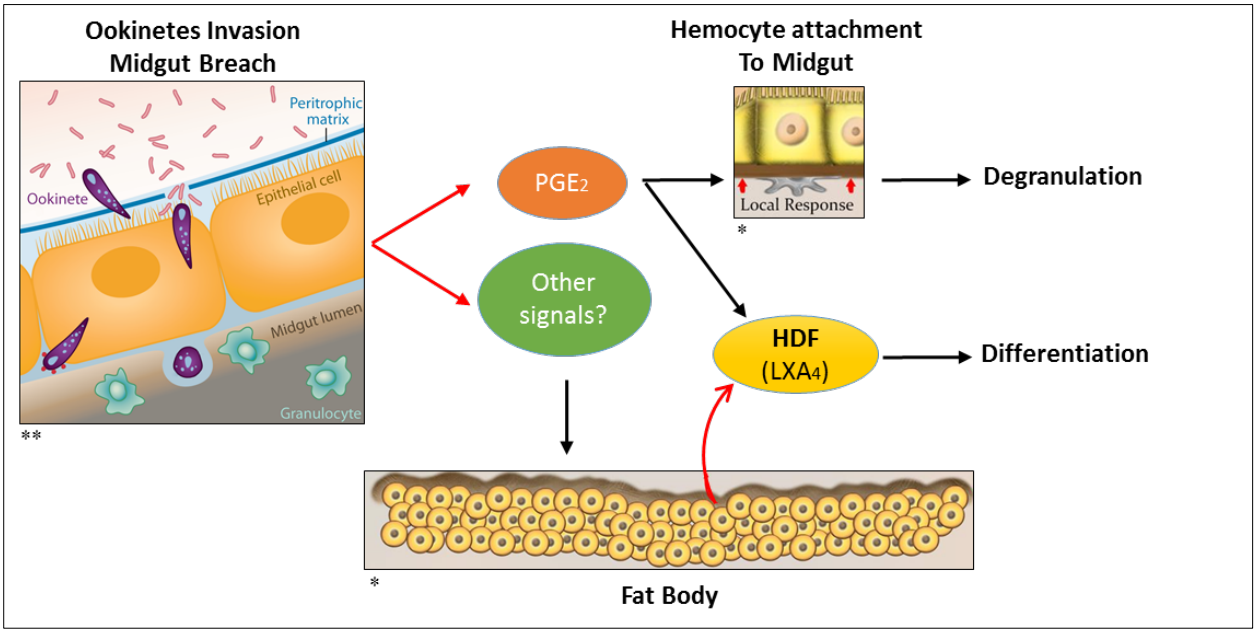


Figure I.6 Model of hemocytes activation and priming. Model follows *Plasmodium* ookinetes midgut invasion in *A. gambiae*. PGE_2 is released 24 hours post-invasion by the midgut, leading to chemotaxis and attachment of hemocytes to the basement membrane. Next, hemocytes release micro-vesicles in the basal surface of the midgut epithelial cells. PGE_2 and likely other signals lead to HDF release 48 hours post-invasion, which activates hemocytes for long-term priming and differentiation into granulocytes and oenocytoids. Priming can be abolished by interfering with function or movement of hemocytes[2, 193]. In addition, *LL3* is necessary for HDF production, as silencing it reduces HDF release [Dr. Barillas-Mury, personal communication]. None of the other immune pathways so far implicated in *Plasmodium* defense (*Toll*, *Imd*, *STAT*, *JNK*) are required for HDF release, although *Toll*, *STAT*, and *JNK* are all necessary for hemocyte differentiation in response to HDF[17]. Even for those pathways, we do not know the effector mechanisms, which receptors activate signaling cascades, or the sequence of kinases and transcription factor activation. Adapted from: * Crompton *et al.*[2] and ** Moreno-Garcia *et al.*[199]

4 Single-cell transcriptomics

"From out of all the many particulars comes oneness"
— Heraclitus

Past microarray studies in mosquitoes have uncovered *Plasmodium* and bacteria-mediated expression changes in genes regulating immunity [200–203]. However, few transcriptomic studies have been conducted to explore in depth how hemocytes respond to insults such as *Plasmodium* [192, 204]. And although terms such as "activation", "priming" and "innate memory" are used to describe immune phenomena in mosquitoes, their precise cellular basis is poorly understood [201]. As we have seen, many innate immune pathways are encoded in mosquito genomes, and have been linked to distinct immune responses by bulk transcriptomics and dissected through reverse genetics [17, 200, 202, 203, 205]. It is therefore highly likely that the number of relevant functional states in hemocytes is larger than currently known molecular markers suggest. In fact, two recent studies by the Levashina and Smith groups have started to explore the cellular heterogeneity of the mosquito immune system, but were limited by their chosen technology [191, 192]. They were largely unable to differentiate between hemocyte populations, since bulk approaches only look at the average expression level, and single-cell approaches conversely need large number of cells to make meaningful conclusions.

On the other hand, well designed single cell approaches such as single cell RNA-sequencing (scRNA-seq) enable researchers to thoroughly map whole immune systems, creating atlases of all immune cell type and states, describing their evolution in time and with infection. Critical biological questions can be explored, such as what transcript isoforms are variably expressed between different cell types [206–209], how cell types differentiate into one another[210, 211], and what is the precise lineage and cell cycle state of individual cells [207, 211]. In scRNA-seq we sample the transcriptome of each individual cell independently from one another, and the technique is quickly becoming the new state-of-the-art in cell biology.

Introduction

Seemingly homogeneous cell populations actually feature great transcriptional heterogeneity, both due to external factors such as individual microenvironments, but also internal stochastic processes [212]. Bulk approaches are unable to disentangle these differences, especially since the vast majority of transcripts in each cell is present in few copies, and most are not even messenger RNAs. The apparent randomness of transcriptional expression, even when cells are exposed to similar microenvironments – what we call transcriptional noise – is now recognised as crucial in determining cell fate decisions [213]. Single cell techniques are new, and both technical methodologies and analysis algorithms need to mature further, but a plethora of technologies have already emerged to make scRNA-seq the most sensitive, unbiased, and high-throughput technology to precisely capture these unique cell types, states, and transitions [214, 215].

The field has come a long way since its origins, having first been developed by Tang *et al*. in 2009 for hand-picked mouse blastomeres, which – thanks to their high RNA content (over 1 ng/cell) – could be more easily processed [216]. Now, tens of thousands of cells with only a few picograms of RNA can be successfully sequenced with highly automated pipelines [Fig. I.7]. However, all protocols share an initial reverse transcription to produce cDNA from RNA, which then needs to be amplified either by polymerase chain reaction (PCR) or in vitro transcription (IVT). As such, some of the original constraints of the technology remain [215, 217]. For example, it is still challenging to separate technical noise from biological variability [209, 210, 218]. In addition, any method only captures poly-adenylated RNA, and is severely limited by the suboptimal mRNA capture rate and reverse transcriptase efficiency [209]. The latter is the limiting step of scRNA-seq: it is estimated only 10-20% of all transcripts are reverse transcribed [219]. Direct RNA sequencing would represent a major step forward but it is still under active development [220].

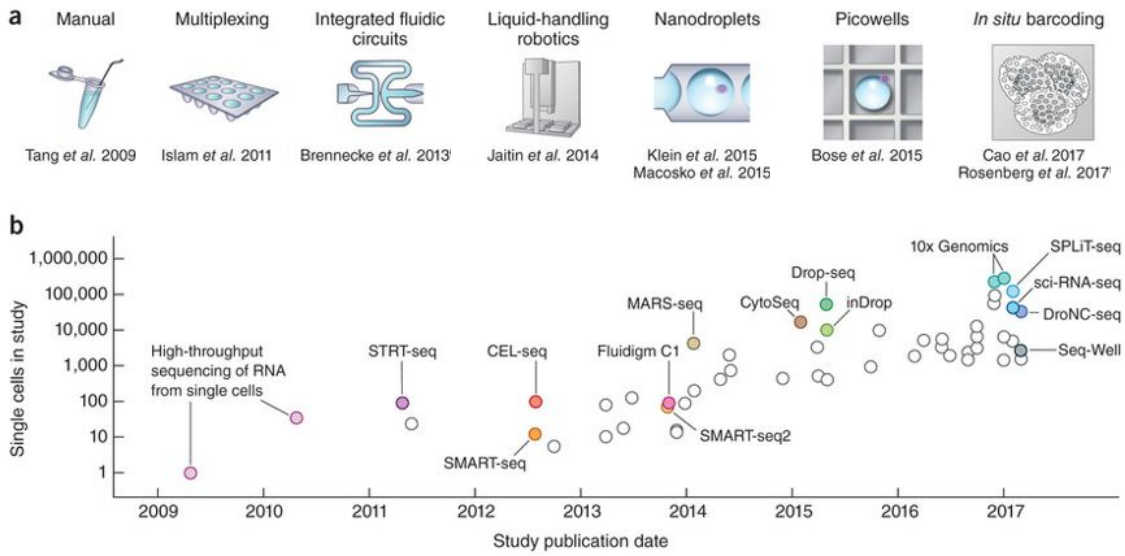


Figure I.7 Evolution of scRNA-seq technologies (A) Technical breakthroughs have increased the number of cells processed per run by orders of magnitude. Sample multiplexing was the first major innovation [221], followed robotics and fluidics[222, 223], which allowed researchers to study for the first-time thousands of cells in parallel. More recently, nanodroplets and picowells [224–226], and now *in-situ* barcoding, are pushing the field even further to its current scale [227, 228], as shown in panel (B). Key technologies are discussed below and summarized in Table I.1. Figure adapted from Svensson *et al*. [217].

a. Single-cell isolation and suspension

The first hurdle in a successful scRNA-seq experiment is creating a clean, pure, high-quality single-cell suspension of well-dissociated cells from the tissue of interest. Far from trivial, this initial step is crucial to the quality of downstream scRNA-seq data. The original Tang method – and one that is still in use when dealing with exceedingly fragile or rare cells – is low-throughput micromanipulation. As cells of interest are manually selected, the technique is precise, but it is also exceedingly labor-intensive. Alternatively, laser capture microdissection can likewise be used to isolate cells from solid samples. Fluorescence Activated Cell Sorting (FACS) on the other hand is able to quickly isolate of tens of thousands of cells. In addition, surface markers tagged with fluorescently-labelled antibodies can be used to purify cells of interest with high fidelity, and most scRNA-seq protocols are compatible with FACS. Nevertheless, FACS requires large amounts of starting material, can be rough on delicate cells,

and requires *a priori* knowledge of the system of interest. Alternatively, cells in suspension can be isolated and processed with microfluidics chips, which automate many of the required cell separation, selection, and collection steps, upstream of fully-automated scRNA-seq protocols. However, cellular stress can be high, capture-rate (number of cells sequenced per cells loaded) is low, and cell selection is highly dependent on chip-architecture. Recently, developments in microdroplet and microwell technologies have achieved significant reductions in hands-on time and reaction volumes (leading to lower costs), while increasing cellular throughput [209, 215, 229–231].

However, for tissues and cells rich in RNAses such as the pancreas or granular immune cells (e.g. neutrophils and macrophages), maintaining optimal cell and RNA integrity during sample preparation is an unsolved technical hurdle [232, 233]. Cells from tissues need to be dissociated and resuspended using enzymes such as collagenase and trypsin, which takes hours and inevitably affects both cell viability and transcriptome, further confounding biological differences. As a result, cells are stressed and their information altered by experimental manipulation, while RNA is lost due to the action of endogenous RNAses[234]. New protocols such as methanol and Lomant's fixation partially solved these issues, particularly for cells in suspension, however solid tissue dissociation remains a challenge [235, 236].

b. scRNA-seq technologies

Protocols can be roughly divided into two separate categories: full-length versus tag-based. The original Tang protocol was a full-length method, while the popular commercial 10X technology is a tag-counting protocol. Each has its strengths. Full-length scRNA-seq methods typically provide more genes per cell and allow a researcher to delve into the data deeper by exploring transcript isoform expression, allelic expression, and RNA editing thanks to the strand-specific information along the full length of each transcript [237]. Tag-based methods, on the other hand, feature higher throughput and lower costs, thus providing the necessary power to discover new and rare cellular subtypes or transcriptional states [231]. The most utilized methods are described and compared below (see also Table I.1).

Full-length protocols

Quartz-seq

An improvement over the original Tang method, with simplified workflow and improved performance, although quickly rendered obsolete by Smart-seq2 [238–240].

Smart-seq2

The classic full-length protocol –considered state-of-the-art in terms of genes per cell captured – begins with an RT reaction using the Moloney Murine Leukemia Virus (MMLV) RT enzyme and oligo-dT primers with template switching oligonucleotides (TSO) to synthesize cDNA. The cDNA is then amplified before library preparation. Though the protocol is time consuming, robotic handling can simplify the workflow [237, 239, 240]. Importantly, strand information is lost with standard Illumina sequencing, and technical errors due to unequal PCR amplification are not corrected by unique molecular identifiers (UMIs: unique molecular identifiers) as in tag-counting protocols, so that PCR amplification bias remains a concern [219]. When using UMIs, every transcript captured gets labeled with its unique barcode (e.g. 10-12 bp long with Chromium 10X) in addition to a cellular barcode. This allows to distinguish sequencing reads originating from unique mRNAs vis-à-vis PCR duplicates.

Tag-based protocols

CEL-seq, CEL-seq2, and MARS-seq

This tag-based protocol employs IVT rather than PCR amplification. CEL-seq (Cell Expression by Linear Amplification and Sequencing) starts with an RT reaction, before second strand cDNA synthesis, pooling, and IVT. Exonic reads are highly strand-specific (over 98% from sense strand), barcoding highly efficient, and no gene-length normalization is required. However, there is a strong 3' bias and spliced isoforms cannot be detected. CEL-seq shows poor sensitivity for lowly-expressed transcripts [241]. MARS-seq (Massively Parallel RNA single-cell sequencing) is a fully automated CEL-seq with UMIs, enabling the counting of individual RNAs [223, 242]. CEL-seq2 improved upon the original protocol by decreasing costs and hands-on time, while increasing sensitivity and implementing UMIs [243].

Introduction

STRT-seq

STRT-seq (Single-cell Tagged Reverse Transcription Sequencing), is a tag-based method that employs anchored oligo-dT primers and a MMLV-based enzyme as Smart-seq2, before PCR amplification. Later iterations of the protocol included UMIs and have been automated to allow multi-plexing and strand-specificity. Disadvantages are the same as for all tag-based protocols, including the inability to detect SNPs or splice variants [244].

DROP-seq, InDROP and Chromium 10X

All these three relatively newer technologies work in similar ways to increase throughput and lower reaction cost by carrying out all reactions in nanoliter emulsion droplets. These droplets contain the lysis buffer, RT, and barcoded microspheres with oligonucleotides to uniquely tag both the individual cells, as well as each transcript within those cells. inDrop and 10X are characterized by higher cell capture rate, 10X has the highest sensitivity and lowest technical noise. Drop-Seq on the other hand is the most cost-effective [224, 225, 245–247]. A detailed discussion of 10X follows in the materials and methods.

SeqWell

The latest addition to the arsenal and one of the most promising recent developments in scRNA-seq, SeqWell sports the same advantages of emulsion droplet methodologies, but employs microarrays and picowells to increase throughput even higher. Seq-Well utilizes PDMS arrays that each contain ~88,000 subnanoliter wells with uniquely barcoded poly(dT) mRNA beads. The uniquely barcoded mRNA capture beads and cells are both secluded in the wells, which are then sealed with semipermeable membranes, leading to a more efficient cell lysis and mRNA capture. Beads can then be pooled, thanks to double barcoding for cells and transcripts (UMIs). Seq-Well only requires a PDMS array, a polycarbonate membrane, a pipette, a clamp, an oven/heat source, and a tube rotator to produce stable cDNA product, making it functional in nearly every clinic and laboratory context. The protocol can also be adapted to use harsher lysis conditions, useful when dealing with fixed or otherwise challenging material [248, 249].

Combinatorial indexing

Recently, single-cell combinatorial indexing has emerged in different groups as another powerful high-throughput scRNA-seq methodology involving the split-pool barcoding of either cells or fixed nuclei. For RNAseq, the methods are similar and are alternatively called SPLiT-seq, sciRNA-seq, or sci-RNA-seq3 [227, 228, 250] . However, single-molecule combinatorial indexing can be used for many other omic techniques to explore chromatin accessibility (called sci-ATAC-seq)[251], genome sequence (sci-DNA-seq)[252], genome-wide chromosome conformation (sci-Hi-C)[253], and DNA methylation (sci-MET)[254].

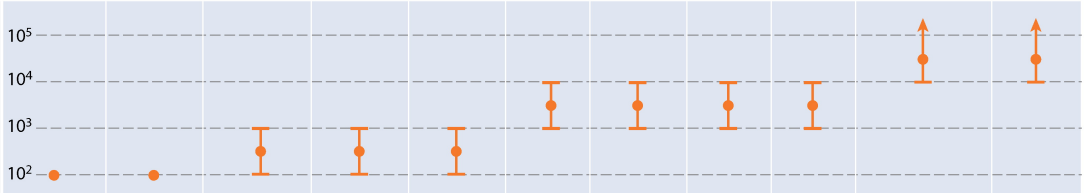
	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay											

Table I.1 Comparison of scRNA-seq methodologies. Abbreviations: cDNA, complementary DNA; DNA pol I, polymerase; RNase H, ribonuclease H. Adapted from Chen et al. [214]

c. scRNA-seq data analysis

After making a quality single-cell suspension, successfully loading the cells onto the scRNA-seq platform of choice, making a library of appropriate complexity, and then sequencing it to the desired depth (note that 50k reads per cell is thought sufficient to successfully cluster cells into subpopulations, including rare cell types) [255], data must be quality controlled before downstream biological analyses. Multiple packages have been created to integrate QC methodologies and analyses and simplify data exploration and interpretation. Seurat [256], Scanpy [257], Scater [258], Monocle [211] and Cellranger [246] are the most popular.

Read Alignment, Expression Quantification, and Quality Control

The reads to reference transcriptome mapping ratio is an early indicator of scRNA-seq data quality. Samples with low mapping percentages likely contain a high amount of damaged or degraded RNA and must be removed. Since sequencing output is the same as for bulk RNAseq the same software can be used for the first data analysis and QC steps. Alternatively, Kallisto and Salmon can both accurately estimate transcript abundance without relying on alignment to an existing transcriptome [259, 260]. Most users however use standard splice-aware alignment programs using reference assemblies. The most popular tools are TopHat2 [261], STAR [262], and HISAT2 [263], although 10X has implemented their version of STAR into a proprietary software suite called Cell Ranger [246]. Studies have compared these aligners highlighting trade-offs between speed, memory requirements, and alignment efficiency in all [264–266]. Which expression quantification method to use varies according to the scRNA-seq technique used. For whole-transcript protocols such as Smart-seq2 traditional bulk-RNAseq methods suffice. Tag methods such as Chromium 10X will either use the Cell Ranger pipeline or specifically-tailored algorithms such as SAVER (Single-cell analysis via expression recovery) to take advantage of UMIs and reduce technical noise [231, 267]. Data is then cleaned up to exclude reads originating from multiplets, broken cells, or dead cells (unless cell were fixed). Even the highest quality, healthiest cells will suffer from low mRNA capture efficiency, bias in transcript coverage, and dropout events (lack of transcripts that are known to be expressed

in the cells). Nevertheless, poor quality samples and cells will skew biological interpretation and must be removed. Some protocols use extrinsic spike-ins (e.g. ERCC) to estimate technical noise and cellular quality [268], as cells with high proportions of ERCC spike-ins likely feature broken, porous cellular membranes. Furthermore, while cytoplasmic RNA is usually lost when a cell ruptures, mitochondria remain within the cell. Thus, a high percentage of mitochondrial RNAs to total RNAs can indicate poor quality. Finally, low total gene counts or transcripts abundance within cells can also be an indication of low quality, although this can sometimes be due to technical limitations or the low total RNA content of the cell of interest [210, 269].

Normalization of scRNA-seq data, and removal of batch effects

Initial QC must be followed by careful data normalization in order to disentangle the biological signal of interest from the variability in capture efficiency, sequencing depth, dropouts, and all other technical effects in each individual sample. This intra-sample normalization is important, but as scRNA-seq datasets become larger, batch normalization is also becoming crucial. The latter takes into consideration all of the above, but also harmonizes samples often run on different days, platforms and laboratories. Normalization is an issue also in bulk-RNAseq, however it is far more complex in scRNA-seq. Bulk RNA-seq investigators standardize libraries by calculating quantities such as transcripts per million (TPM), fragments per kilobase of exon per million fragments mapped (FPKM, which takes into consideration both transcript length and library size), or size factors [210, 231, 270, 271].

That is not sufficient for single-cell RNA-seq, which features unique analytical challenges requiring specifically-tailored normalization algorithms. For example, scRNA-seq data matrices are characterized by abundant zeroes, but ‘zero inflation’ is due to both technical reasons (dropouts due to the low reverse transcription efficiency previously mentioned) as well as meaningful biological differences (e.g. quiescent or stem cells). Moreover, scRNA-seq is characterized by higher technical noise even for non-lowly expressed genes, further augmented by true biological heterogeneity. Any overcorrection by normalization algorithms will reduce such biological differences. Conversely, under-correction will lead to spurious biological conclusions. Traditionally, scRNA-seq normalization methods have employed off-the-shelf or

adapted bulk-RNAseq methods. Specifically, median normalization methods are used to identify genes whose expression does not change across cells. Non-differentially expressed (DE) genes are then used to calculate global scaling factors that are unique for each cell, but common for all transcripts within that cell. These methods all assume total RNA in each cell is more or less the same and that all variation is technical. However, that is not the case when dealing with single cells, especially in heterogenous samples. And further, RNA content in different libraries is influenced by both the cell-cycle and the dynamics of transcription (including transcriptional bursts) for each individual gene [14, 270–272]. A first attempt to circumvent this limitation was the use external spiked-in ERCC artificial RNAs at a set concentration in each individual cell library, thus adjusting for technical variation and improving the accuracy of global scaling factors [273]. However, issues with spike-ins can lead to inconsistent detection and few studies have used this technique [210]. Rather, researchers have been using UMIs to successfully remove or reduce cell-specific effects due to amplification and gene length, although UMIs can only be used with tagging scRNA-seq protocols [219, 224, 242, 270]. As the field matures, more and more sophisticated normalization methods are being specifically tailored for scRNA-seq datasets, such as SCnorm, SAMstrt, and SCTransorm (as integrated in Seurat V3.0) [272, 274, 275].

As larger scale scRNA-seq experiments become the norm, ‘batch effects’ – the aggregated technical variation of different cell dissociation methods, library preparation techniques, sequencing platforms, environments, handling, operating equipment, institutes and laboratories – are becoming ever more important confounding factors. These confounders are especially problematic for large consortium-scale projects such as the Human Cell Atlas, and must be minimized. The field is thronged with new computational methods that have begun addressing the issue, from the linear regression models of ComBat[276] to the nonlinear canonical correlation analysis (CCA) of Seurat [277] or the projection of mutual nearest neighbors method (MNNs)[278]. Separately, MAST, DESeq and limma can include batch effects as covariates in their DE testing model [279–281]. Principal component analysis (PCA) and visualization in low dimensional usually follows. Recently, a dedicated method (kBET)

has been developed to further explore batch effects (and their correction), in detail [282]. All these normalization, batch correction, and visualization methodologies are summarized below in Fig. I.8.

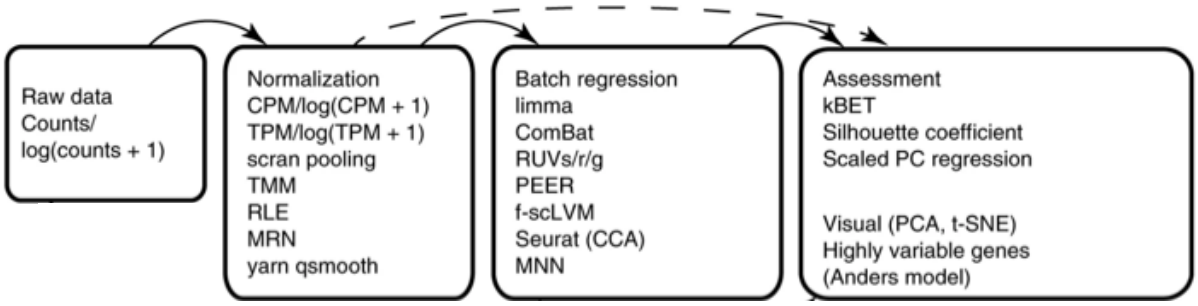


Figure I.8 Summary of normalisation, batch regression, and assessment techniques. Detailed information on each method and full figure in Büttner *et al*. [282]

Dimensionality reduction, feature selection, clustering, and differential expression analysis

Data matrices downstream of all QC and normalization processing feature many thousands of dimensions, with thousands of genes and tens of thousands of cells. Data must be simplified - dimensions reduced – to aid computations and interpretation while keeping intact key biological differences between cells and conditions. PCA is a linear dimensional reduction algorithm assuming normal distribution of data. It identifies new variables, called principal components (PCs), that are linear combinations of the variables from a dataset. Data is standardised so that each gene's mean expression across cells is zero and the PCs are then normalised eigenvectors of the genes' covariance matrix. Importantly, the PCS are ordered by how much dataset variation they describe. T-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction technique used for example by Seurat to visualize the scRNA-seq data [283]. Both are limited. PCA is unable to fully display data complexity, while t-SNE plots are inconsistent and do not preserve global information. Newer algorithms such as uniform manifold approximation and projection (UMAP) [284], and scvis [285] were designed specifically for scRNA-seq. UMAP is fast, reproducible, and cluster organization and display reflects inherent cellular similarity, unlike for t-SNE.

Feature selection is also used to reduce data dimensionality and free-up computational resources for downstream analysis such as clustering. Unsupervised algorithms for feature selection are divided into three main types. Highly variable genes (HVG) methods as in Seurat assume HVG are those that vary because of biology [222, 256]. Spike-in approaches (e.g. scLVM and BASiCS) identify genes that have higher variance than spike-ins with similar expression levels [270, 273]. Finally, dropout methods such as M3Drop use the dropout distribution characteristic of scRNA-seq data to efficiently select all important features in a dataset [286].

After dimensionality reduction we can finally interrogate our data and answer key biological questions. For example, what populations and subpopulations of cells do exist in our dataset? And what are their cell states? Cell clustering can be done either using known markers, or more commonly with unsupervised clustering methods. These are mainly divided into k-means, hierarchical clustering, density-based clustering, and graph-based clustering methodologies. K-means requires setting the number of clusters *a priori*, and assigns cells to nearest cluster center, while all others methods work in unsupervised fashion to establish the optimal number of clusters. Some methods such as single-cell consensus clustering (SC3) use a combination of methodologies, and the popular Seurat clusters mainly with a shared nearest neighbor algorithm (SNN) [287]. Differentially expressed genes (marker genes) are then found with differential expression analysis (DE) or analysis of variance (ANOVA). DE analysis is an active area of software development. Often, clustering algorithms return not only cell subpopulations, but also variable cell states for each of these populations. Software packages must be able to differentiate between the two, while dealing with the high noise of scRNA-seq data and the large sample size. While bulk RNAseq DE techniques are still used, in recent years specific tools such as MAST (linear model fitting and likelihood ratio testing), SCDE (Bayesian approach with low-magnitude Poisson), DEsingle (Zero-Inflated Negative Binomial), have been developed. Seurat uses the non-parametric Wilcoxon rank sum test as a default, but other methodologies such as MAST and DEseq2 can also be employed. Soneson et al tested over 36 methods in their recent review [288]. See Table I.2 below for a summary.

Clustering	References	DE analysis	Type	References
Seurat	Satija et al., 2015[256]	MAST	Single-cell	Finak et al., 2015[279]
SC3	Kiselev et al., 2017[289]	ROTS	Single-cell	Seyednasrollah, 2016[290]
Destiny	Angerer et al., 2016[291]	BCseq	Single-cell	Chen et al., 2018[292]
SNN-Cliq	Xu and Su, 2015[293]	SCDE	Single-cell	Kharchenko et al., 2014[294]
RaceID	Grun et al., 2015[295]	DEsingle	Single-cell	Miao et al., 2018[296]
SCUBA	Marco et al., 2014[297]	Cencus	Single-cell	Qiu et al., 2017[298]
BackSPIN	Zeisel et al., 2015[299]	D3E	Single-cell	Delmans et al., 2016[300]
PAGODA	Fan et al., 2016[301]	BPSC	Single-cell	Vu et al., 2016[302]
CIDR	Lin et al., 2017[303]	DESeq2	Bulk	Love et al., 2014[280]
pcaReduce	Zurauskiene, 2016[304]	edgeR	Bulk	Robinson et al., 2010[305]
TSCAN	Ji et al., 2016[306]	Limma	Bulk	Ritchie et al., 2015[281]
ZIFA	Pierson et al., 2015[307]	Ballgown	Bulk	Frazee et al., 2015[308]

Table I.2 Summary of clustering and DE analysis software packages. For more information please consult these excellent reviews: Andrews and Hemberg (clustering) [287] and Soneson *et al* (DE analysis) [288].

Cell lineage, pseudotime, alternative splicing and gene regulatory networks analysis

After probing the cellular complexity of tissues and cell populations, data can be used to explore the dynamics of cellular development and identify cell types lineages, for example by building a pseudotime ordering of cells which can showcase cellular differentiation. Pseudotime techniques order cells along a continuous trajectory, aligning cells based on transcriptional similarities rather than clustering them. These approaches not only allow investigators to probe the initial, transitional, and final cell states of a population, but also the genes that are involved in such transitions. Popular tools are Monocle (based on minimum spanning tree) [211], Monocle2 (reversed graph embedding) [298], Slingshot (cluster-based approach) [309], TSCAN [306], PAGA, and Cellrouter [310]. Saelens *et al.* recently evaluated most pseudotime and lineage approaches and found Monocle2, Slingshot, and PAGA to be superior, depending on the individual data structure of the dataset (e.g. linear, bifurcating, complex separate trees) [311].

Introduction

Gene regulatory network inference is a common feature of bulk RNAseq analyses, normally employing weighted gene co-expression network analysis (WGCNA), which assumes all genes that are highly correlated in expression to be co-regulated. By combining cells together to build a pseudo bulk-RNAseq dataset we can evaluate gene regulatory networks in the same way. However, the analysis needs to be run separately for each subpopulation. SCENIC is one such scRNA-seq method that can build gene regulatory networks from single cell data and predict transcription factors - target genes interactions [312]. PDIC is an alternative software suite to answer the same questions [313].

Finally, when data is generated with scRNA-seq protocols producing full-length transcripts (such as Smart-seq2) investigators can also analyze alternative splicing. Over 90% of human genes undergo alternative splicing, which plays important roles both in tissue homeostasis and disease [314]. Data on isoform usage could be crucial in understanding the expression dynamics of specific pathogenic isoforms for example, or to further characterize the importance of cellular subsets in immune process. However, bulk RNAseq methodologies are again unsuitable to the task. Recently new methods have emerged such as SingleSplice, Census, BRIE, and Expedition [298, 315–317].

5 Aims and outline of the thesis

This dissertation first focuses on dissecting the complexity of the *A. gambiae* M-form *(A. coluzzi)* immune system under baseline conditions. That knowledge is then leveraged to obtain an in-depth understanding of how mosquitoes responds to both blood-feeding and *Plasmodium* infection. In analogy with vertebrates, I posit the existence of different hemocyte subpopulations and states, each characterized by distinct gene expression profiles. I will further argue that hemocytes transition between distinct states along a range of predetermined routes, through which the diversity of functions associated with cellular immunity in invertebrates is generated. In addition, I will show that single-cell approaches, coupled with complementary bulk techniques and imaging validation, are an effective method to study the cellular arm of the immune system of mosquitoes.

In Chapter II, we explored different strategies to isolate hemocytes and create a clean, pure single cell suspension for downstream scRNA-seq. We evaluated different methods, enzymes, and fixatives to adapt single cell protocols to the unique challenges of mosquito immune cells, while maintaining high quality RNA and cellular integrity. As part of this work we developed a protocol to fix and sequence hemocytes at single cell resolution making use of the droplet-based Chromium 10X technology. We then validated our scRNA-seq results by adapting the commercial RNAscope RNA-FISH technology to mosquitoes.

In Chapter III, we used these methods to characterize the functional classes of *A. gambiae* mosquito hemocytes and build a comprehensive atlas of the cellular arm of the mosquito immune system to discover new hemocyte cellular subtypes. We then defined marker genes for each cell type, and identified surface markers for future functional studies. We uncovered different cell states within each hemocyte type, successfully building a lineage tree to explain how hemocytes differentiate into each cell type and cell state. Finally, we validated these scRNA-seq results with a combination of bulk-RNAseq and RNA FISH techniques and visualize each cell type and their spatial-temporal localization in the mosquito. Importantly, we not only recapitulated what previous knowledge existed, but also discovered novel effector

cellular subtypes, including a cellular subtype potentially akin to lamellocytes in *Drosophila*, likely involved in the 'late' cellular immunity against *Plasmodium*, as well as hemocytes responsible for secreting anti-microbial peptides in circulation, revealing a previously unknown complexity of this biological system.

In Chapter IV, we challenged *A. gambiae* mosquitoes first with blood-feeding and then with *P. berghei* and *P. falciparum* infection. We evaluated how mosquito hemocytes, guts, and carcasses respond to these challenges to explore how hemocytes differentiate into their distinct cellular states. We identified a trajectory of immune activation following the mosquito on a time-course after infection, reaching a peak of transcriptomic activity against the parasite at days 2-3 after infection, before returning to baseline at day 7. Finally, we describe how hemocytes dynamically respond to infection, going into circulation to respond to injury and replenish the immune cell pool. We identified rapidly dividing precursor cells, as well as the transcriptomic signatures of the response of hemocytes and fat body to *Plasmodium*, including what pathways are differentially activated in various cellular subtypes. Then, we explored how the upregulation of the Toll pathway affects hemocytes and their ability to mount an effective immune response to suggest how different hemocyte subtypes are the control of specific and distinct immune pathways.

In chapter V, I conclude by providing a summary of our findings and discussing what significance they hold in view of the emerging importance of vector-borne diseases for human health and disease, not only in the developing world, but increasingly also in the West.

6 List of publications

1. "Mosquito cellular immunity at single-cell resolution"
   Raddi G*, Barletta-Ferreira A*, Efremova M, Luis Ramirez J, Cantera R, Teichmann S, Barillas-Mury C*, Billker O* (*in review*)
   * These authors contributed equally to the work
2. "Seq-Well: A Sample-Efficient, Portable Picowell Platform for Massively Parallel Single-Cell RNA Sequencing"
   Aicher TP, Carroll S, **Raddi G**, Gierahn T, Wadsworth MH 2nd, Hughes TK, Love C, Shalek AK. *Methods Mol Biol*. (2019) 1979: 111-132
3. "Single-Cell RNA Sequencing with Drop-Seq"
   Bageritz J, **Raddi G**. *Methods Mol Biol*. (2019) 1979: 73-85