Chapter 1

Introduction

The impact of malaria on human health is dramatic as financial constraints and wide-spread resistance to drugs hamper malaria control programmes. New, effective therapies are urgently required; their development however, relies on a better understanding of malaria parasite biology. The Evimalar network is one of the dedicated institutions created to promote malaria research as part of a global effort to eradicate this disease. As a PhD student of the EVIMalaR programme, I focused my efforts on the development of tools that enable large scale studies of the malaria parasite at the genetic level. To set the background for this work, this introduction outlines the disease burden, selected aspects of the malaria parasite biology, current high-throughput gene targeting strategies in model organisms and applications of such technologies to *Plasmodium* biology.

1.1 Malaria: A major global parasitic disease

Infectious diseases are still in the top 10 causes of death, having accounted for 18.4 % of total deaths worldwide in 2011 [1]. Although the widespread use of vaccines and drugs has dramatically decreased mortality from infectious diseases in developed countries, some them are beginning to emerge or re-emerge and are still prevalent in the developing countries [1].

Malaria is one of the oldest diseases known to mankind, the two appearing to have evolved together. Malaria literally means "bad air" (from the Italian words "*mal aria*"), named after the belief that this disease was caused by an unknown substance in the air arising from swamps [2]. Nowadays, it endangers half of the world's population (Fig.1.1). The WHO estimates that 219 million cases of malaria led to 660,000 deaths in 2012. Nearly 80 % of the cases and 90 % of the deaths are estimated to occur in sub-Saharan Africa, with children under the age of five and pregnant women being the most severely affected [3]. Malaria is a vector borne disease caused by parasites of the genus *Plasmodium*. The different species infect a wide variety of hosts including humans, monkeys, rodents, birds, and reptiles. The five human pathogens are: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi*. *P. falciparum* prevails in Africa and is the most deadly of these parasites, while *P. vivax* is less morbid but also widespread, and the other three species have a much lower incidence. The clinical symptoms of malaria include paroxysms (acute fever that is typically preceded by chills and rigor), vomiting, headache and anaemia. *P. falciparum*, in particular, can be responsible for cases of "severe malaria", a life-threatening condition that includes pronounced anaemia, disorders of the coagulation system and sequestration of the infected red

blood cells (RBCs) in the deep vasculature (e.g. brain, lungs and placenta in pregnant women) [4,5]. The neurological involvement may lead to a permanent coma and in some cases death. When a child recovers from a severe malaria episode, cognitive impairment is likely to occur thus reducing the child's lifelong potential [6]. The cyclic nature of the fever is a consequence of the strong inflammatory immune responses triggered by synchronous cycles of infection and release of parasites from circulating RBCs. This periodicity varies with the parasite; *P. malariae* causes fever every 72 hours, *P. knowlesi* every 24-28 hours and the remaining three species have 48-hour cycles. It is worth noting that *P. falciparum*, perhaps due to less synchronous growth, often causes an uninterrupted fever rather than periodic paroxysms [7].

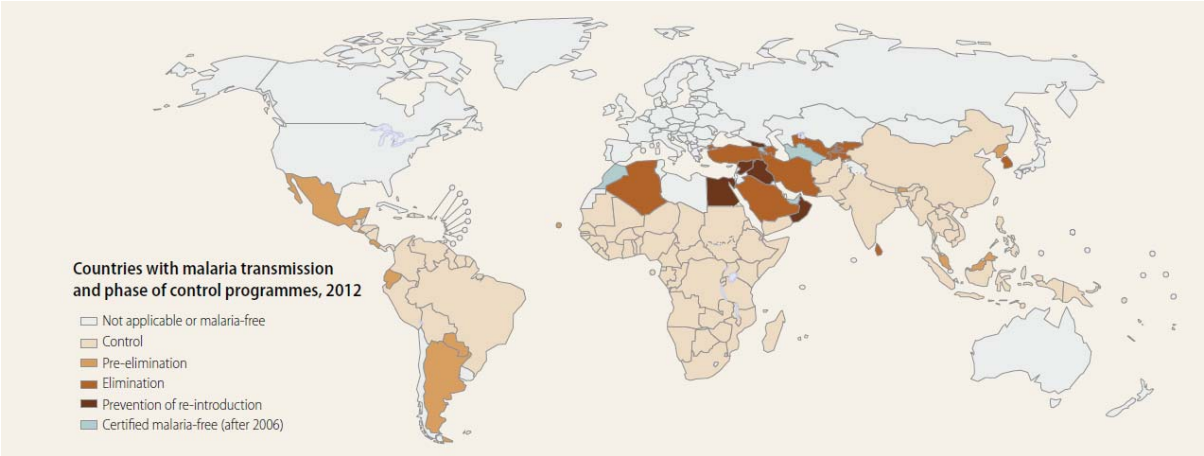

Fig. 1.1| Malaria endangers half of the world's population.
Malaria-free and malaria-endemic countries in phases of control, pre-elimination and elimination at the end of 2012. In "Pre-elimination" countries malaria test positivity rate is less than 5% during the malaria season (>5% "Control" phase); "Elimination" countries have zero incidence of locally transmitted infections; "Certified malaria free" countries have no locally transmitted infections for over a decade (Adapted from 2013 WHO report).

There is no natural acquisition of long lasting sterile immunity to malaria. In endemic areas, where exposure to infective mosquitoes is continuous, tolerance to infection may be seen in adults who, despite being asymptomatic, have a continuous low level of infection and thus act as reservoirs. Once these semi-immune individuals leave malaria endemic-areas, they lose their immunity within about six months [4]. This combined with genetic diversity, evolutionary plasticity and lifecycle complexity of *P. falciparum* complicate the development of malaria vaccines [8,9]. In the last two decades, over 40 vaccines designed to trigger an immune response against subunit components of liver or blood-stage parasites, or whole sporozoites (a mosquito stage form), have undergone clinical trials. Despite the promising results shown in pre-clinical and phase I-IIa trials, none accomplished full protection in the

field. Even the leading malaria vaccine candidate, RTS-S, demonstrated only modest protection against both clinical and severe malaria in young infants, in phase IIIb trials [10].

A range of malaria therapies exist. *P. falciparum* parasites have developed resistance against the majority of drugs introduced prior to the artemisinin combination therapies (ACT). For instance, chloroquine, once the drug of choice for prophylaxis and treatment, is no longer efficient in most areas where malaria is endemic [11]. Nowadays, it is only recommended for the treatment of malaria caused by *P. vivax* and *P. ovale* [12]. The ACT, the current first-line treatment for *P. falciparum* malaria combines the fast acting artemisinin-based compounds with a drug from a different class such as lumefantrine, mefloquine, piperaquine, among others, in order to reduce the chance of development and spread of resistance to either of the drugs. The second drug is chosen according to the local resistance patterns. Recent epidemiologic studies at the Thai-Cambodia border have, however, reported that the efficacy of ACT has decreased [8].

1.2 The life cycle of *Plasmodium* parasites

Malaria parasites belong to the phylum Apicomplexa, a large group of unicellular parasites that infect only animals. Other members of this phylum include *Cryptosporidium*, *Toxoplasma*, *Eimeria*, *Babesia* and *Theileria*. This phylum is defined by a specialised set of structures and secretory organelles at their apical tip, which is key to invasion and motility in –zoite stages [13]. This group of parasites features an unusual organelle named the apicoplast, which was acquired by secondary endosymbiosis between a free-living ancestor of these parasites and a red algae. The apicoplast is essential for parasite survival and contains biosynthetic pathways which have an equivalent in plants and bacteria, but not in animals such as type II pathway for *de novo* fatty acid synthesis [14,15].

Plasmodium species are obligate parasites with a complex life cycle that involves two different hosts: a vertebrate and a mosquito vector (Fig.1.2). Transmission to the vertebrate host is initiated by the bite of an infected female mosquito of the genus *Anopheles*, if the vertebrate is a mammal. During the blood meal *Plasmodium* sporozoites leave the mosquito salivary glands and enter the vertebrate's bloodstream (Fig. 1.2.1). The injected sporozoites migrate to the liver sinusoids where they traverse the vascular endothelium and invade hepatocytes (Fig. 1.2.2). There they multiply, giving rise to thousands of merozoites in 2-16 days, depending on the *Plasmodium* species (Fig. 1.2.3). Eventually merozoite-filled vesicles,

called merosomes, capable of infecting RBCs bud off from the hepatocyte into the liver sinusoids, thus starting the blood stage of the infection (Fig. 1.2.4) [16]. Inside the erythrocyte, a single merozoite replicates by schizogony and undergoes successive differentiations from ring through trophozoite stage eventually generating schizonts with 16-32 merozoites each. Finally the RBC ruptures and releases new merozoites, which in turn infect new erythrocytes (Fig. 1.2.5). While the initial stages of the infection (i.e. liver stages) are asymptomatic, repeated infection of erythrocytes by merozoites causes the symptoms and pathologies of malaria. During the asexual cycle in the blood a subset of parasites bypasses asexual multiplication and differentiates into sexually committed cells: the female and male gametocytes (Fig. 1.2.6). These forms are arrested in the G0 phase and only re-enter the cell cycle to produce gametes after being ingested by a mosquito (Fig. 1.2.7). In the mosquito midgut fertilisation of a female by a male gamete results in the formation of the zygote, the only diploid stage of an otherwise haploid parasite that develops into a motile and invasive ookinete. The ookinete crosses the midgut wall and forms an oocyst on the basolateral lamina (Fig. 1.2.8) where it will generate thousands of oocyst-derived sporozoites. When mature, these sporozoites migrate through the hemocoel to the mosquito's salivary glands (Fig. 1.2.9), making this mosquito infectious and hence completing the cycle (Fig 1.2.10) [17].

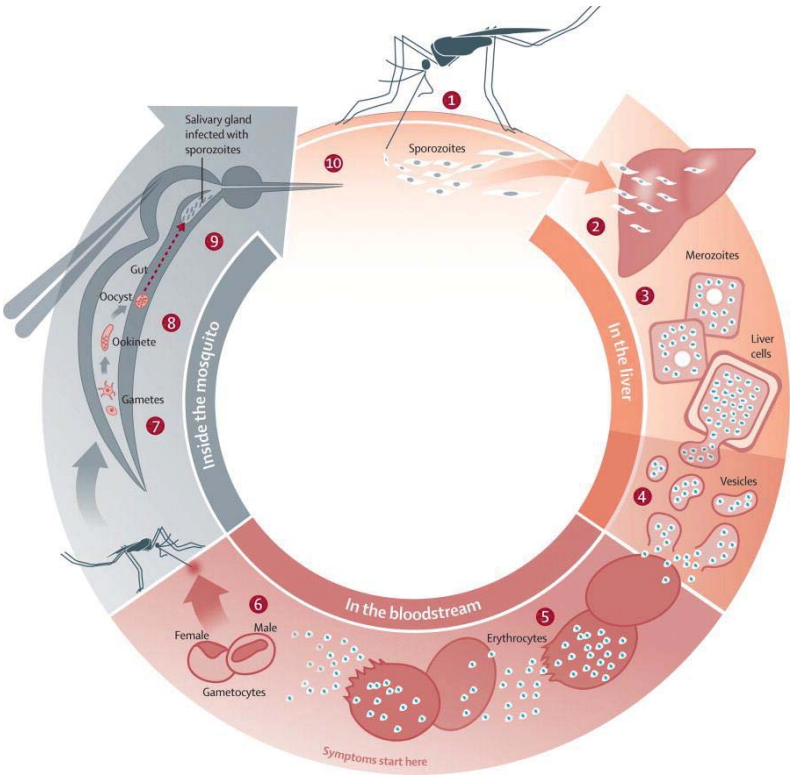

Fig. 1.2| Malaria life cycle.
The life cycle can be divided into three different stages: liver, blood and mosquito stages. The asexual cycle is initiated when the vertebrate host is bitten by an infected mosquito. The injected sporozoites migrate to the liver where they generate thousands of merozoites. After 2-16 days (63-72 hours in *P. berghei*, the rodent parasite in this study [18]) parasites leave the hepatocyte and invade erythrocytes, thus starting the symptomatic phase of the disease. Once gametocytes, the sexual precursors, are taken by another mosquito, fertilisation and mosquito colonisation occur. (Adapted from http://www.malariavaccine.org/malvac-lifecycle.php)

1.3 Next-generation sequencing technologies

DNA sequencing is the identification of the order of the four nucleotide bases adenine (A), guanine (G), cytosine (C), and thymine (T), in a molecule of DNA.

Until recently, DNA sequencing relied almost exclusively on Capillary/Sanger chemistry, a dideoxy chain termination method of sequencing [19,20]. Progress in technology across fields of microscopy, nucleotide chemistry, polymerase engineering, data storage and bioinformatics made next-generation sequencing (NGS) strategies possible.

Currently, the most commonly used NGS platforms are: 454 pyrosequencing (Roche/454 Life Sciences) [21], Illumina [22] and SOLiD (Applied Biosystems) [23]. These have produced an immense volume of accurate DNA sequence data at a fraction of the cost of the Sanger/Capillary method, which dramatically accelerated biological and biomedical research. Specifically, NGS have completely revolutionised several aspects of the field of genomics such as *de novo* genome sequencing, single-nucleotide polymorphism (SNP)

detection, chromatin immune-precipitation (ChIP-Seq) and transcriptome analysis (RNAseq). More recently, additional sequencing platforms such as Ion Torrent and Pacific Biosciences were introduced into the market [24].

NGS platforms differ in their sequencing biochemistry but their workflows are rather similar: DNA molecules are sheared into random short fragments which are then ligated *in vitro* to adaptor sequences at both ends to generate a so-called sequencing library.

A number of reviews have compared them to each other [25–28]. The advantages of NGS relative to Sanger sequencing are clear:

- *In vitro* construction and amplification of sequencing libraries;
- Higher degree of parallelism enabled by an array-based approach;
- Minimal volumes of reagents (picolitres or femtolitres) required since the array features are immobilized on a surface (flow cell).

Some disadvantages such as read-length and accuracy can nevertheless be listed; NGS reads are much shorter than Sanger sequencing and base-calls are, on average, at least tenfold less accurate. The latter is compensated by the huge number of reads generated, that together produce very accurate consensus sequences.

In this thesis I will focus solely on Illumina sequencing chemistry, as this was the only platform used for the present study.

1.3.1 Illumina sequencing

1.3.1.1 Illumina sequencing overview

The Illumina platform, sometimes still referred to as “the Solexa”, originated from work by Turcatti and colleagues [29,30]. It is optimised to generate large amounts of short DNA reads and is currently the cheapest sequencing technology per base of data. Read length has increased from 30 bp in 2008 to 300 bp in 2014, with the cost per bp also decreasing by several orders of magnitude in the same period. A recent study from the Sanger Institute estimated the error rate of Illumina reads to be below 0.4 % [27].

Table 1.1 summarises the main features of the different Illumina instruments currently on the market.

Table 1.1| Overview of current Illumina Instruments and their applications

	MiSeq	NextSeq 500	HiSeq 2500	Hi Seq X
Sequencing applications	Small genomes and amplicons	Genomes, exomes and transcriptomes	Production-scale genomes, exomes and transcriptomes	Population-scale human genomes
Output	0.3-15 Gb	20-39 Gb	10-180 Gb	1.6-1.8 Tb
Run time	5-55 hours	15-26 hours	7-40 hours	< 3 days
Reads per flow cell	25 Million	130 Million	300 Million	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

1.3.1.2 Overview of library preparation procedures

Most library preparation protocols share the following workflow: DNA fragmentation, end repair, A-tailing and adaptor ligation, as illustrated in Figure 1.3. Illumina technology performs the best with DNA fragments that are 200-600 bp. Therefore long molecules of DNA need to be sheared prior to library preparation (except for some RNAseq protocols where the mRNA is sheared before reverse transcription [31]). This can be achieved by one of four different methods: enzymatic digestion, sonication (e.g. Covaris), nebulisation or hydrodynamic shearing (Fig.1.3, step 1). Next, end repair is used to generate blunt-ended, 5'-phosphorylated DNA ends compatible with the adaptor ligation strategy (Fig. 1.3, step 2). The enzymes involved in this step are T4 polynucleotide kinase and T4 DNA polymerase, both originally isolated from a bacteriophage. After end repair, fragments are "dA-tailed" by the Klenow fragment, a process by which a dAMP nucleotide is added onto the 3' end of blunted DNA fragments (Fig. 1.3, step 3). This step maximises ligation efficiency of adaptors carrying complementary dT-overhangs. Finally, T4 DNA ligase is used to catalyse the adaptor ligation step (Fig. 1.3, step 4). These are partially single stranded, forming a Y-shape, which allows each strand to have two different sequences added, one at each end (5' or 3') – crucial for the Illumina sequencing chemistry. A final clean-up step ensures removal of free library adaptors and adaptor dimers. This is critical as adapter-dimers are co-amplified with the adapter-ligated library fragments and reduce the sequencing capacity of the platform. After this step, the final libraries (Fig. 1.3, step 5) are quantified, usually by qPCR. Depending on their concentration they can either be directly used for sequencing, or amplified by PCR so that the desired concentration can be achieved.

Different libraries can be pooled and run together in the same lane in a process called multiplexing. This enables efficient use of the sequencing capacity of the instrument and only requires the incorporation of differently barcoded adaptors at the library preparation stage.

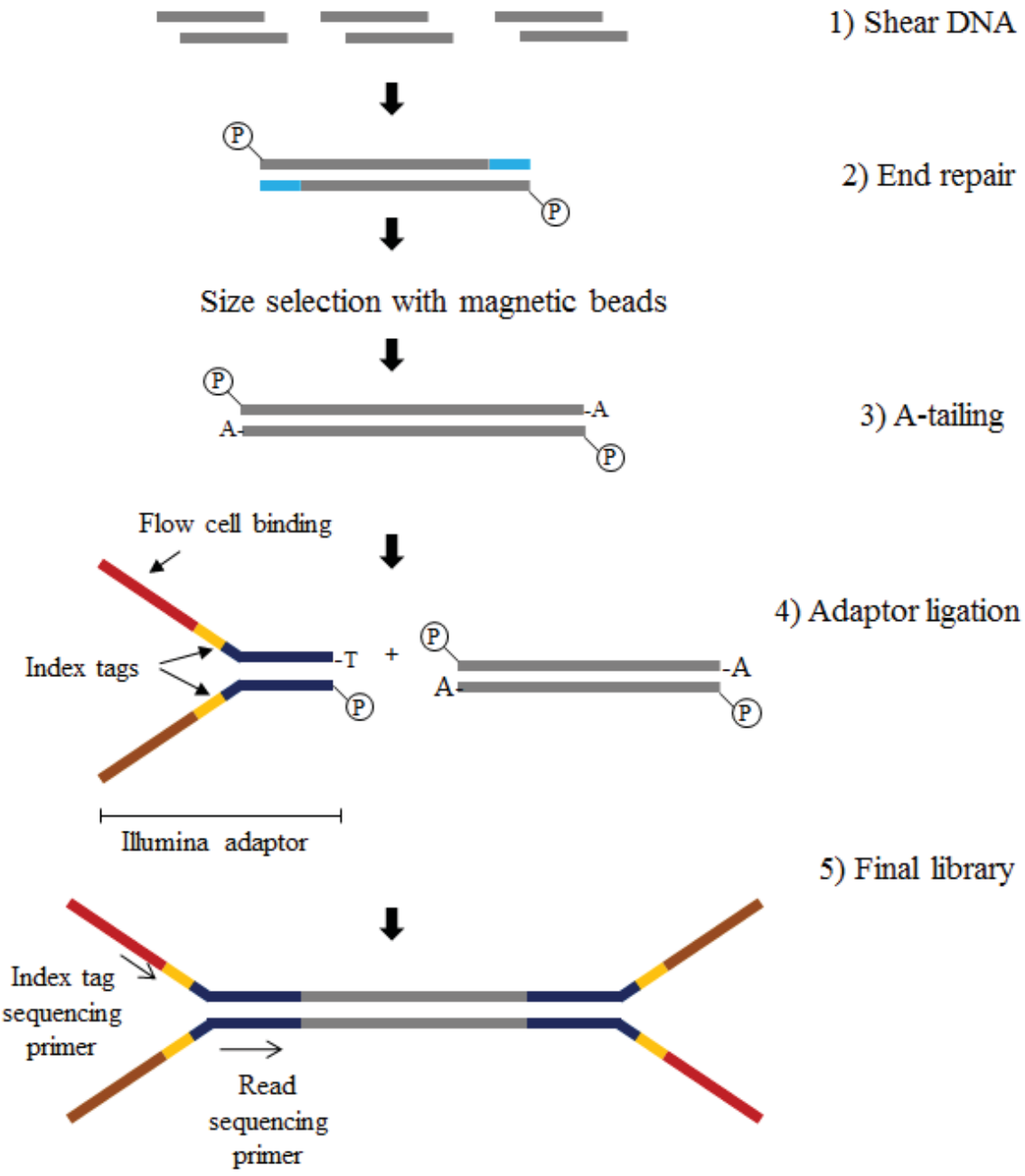

Fig. 1.3| Overview of a library preparation protocol.
Extracted gDNA is sheared in order to generate fragments of adequate size. Then end-repair and A-tailing follow. This enables the ligation of the Illumina adaptors. After this step the final libraries are ready to be quality controlled and sequenced.

1.3.1.3 Illumina sequencing chemistry

After careful quantification, a defined concentration (usually 1 to 4 nM) of each library is denatured and loaded into an Illumina flow cell. The latter is simply a glass surface coated with primers that are complementary to sequences present within the library adaptors. The Figure below illustrates, in nine steps, the Illumina sequencing workflow.

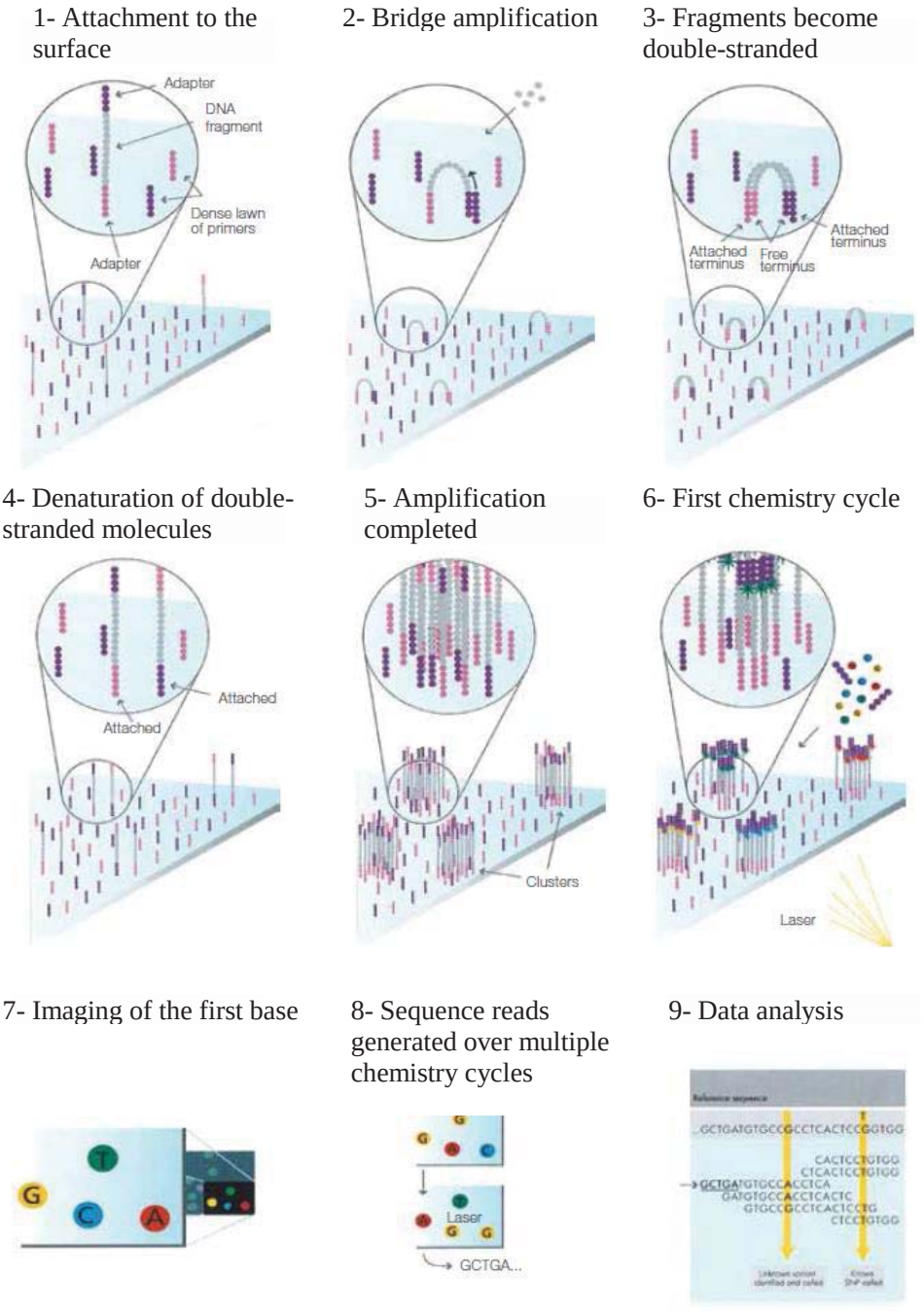

Fig. 1.4| Illumina sequencing chemistry.
1- Single stranded fragments randomly bind to complementary oligos on the surface of the flow cell. 2,3- unlabelled nucleotides and polymerase are added for solid-phase bridge amplification. At this stage the

fragments become double stranded. 4,5- After amplification is complete, a denaturation step results in millions of clusters of single-stranded molecules attached to the flow cell. 6- Determination of the first base; labelled reversible terminator nucleotides, primers and a polymerase are added to the flow cell thus initiating the first chemistry cycle. 7- The identity of the first base of each cluster is then recorded by imaging of the flow cell upon laser excitation. 8- After repeated cycles of chemistry the sequence of bases of a given library is determined. 9- Dedicated software converts the resulting image files into text files to allow data analysis by current bioinformatics tools. (Adapted from Illumina® Sequencing website)

After loading, the denatured, single-stranded library molecules are allowed to hybridise at one end with the primers on the flow cell (Fig. 1.4, step 1). This is followed by bridge amplification (Fig. 1.4, step 2), a process by which bound library molecules have their free adaptor end binding to a complementary primer sequence on the flow cell (forming a "bridge") and then act as template for the synthesis of the complementary strand of the duplex upon addition of the adequate reagents (Fig. 1.4, step 3). A subsequent denaturation step generates two single-stranded molecules that are both attached to the surface of the flow cell (Fig. 1.4, step 4). This "bridge amplification" is repeated until the desired cluster density is reached (Fig. 1.4, step 5).

After cluster formation sequencing is accomplished by SBS ("sequencing by synthesis") chemistry [22]. It is initiated with denaturation of the clusters and hybridisation of a sequencing primer complementary to the free end of the DNA molecules. Sequencing cycles follow, each consisting of the incorporation of a single modified nucleotide (Fig. 1.4, step 6) and subsequent high resolution imaging of the entire flow cell (Fig. 1.4, step 7). These nucleotides are modified in two ways: they are "reversible terminators" as they carry a chemically cleavable moiety at the 3' hydroxyl position that prevents incorporation of any other nucleotide in each cycle; and they have a chemically removable base-specific fluorophore that allows their identification by laser excitation in one of four different channels [30]. After imaging is complete, both the fluorescent dye and the terminator moiety are removed. Cycles of chemistry and imaging are repeated until the desired read length is reached (e.g. 150 bp). This is the end of a single-end read run. Paired-end sequencing requires subsequent removal of the first read products by denaturation followed by an equal number of cycles of SBS chemistry and imaging this time priming on the adaptor at the other end of the library. Positional information from the images of the flow cell links the two pieces of sequence data.

Computational analysis is then used to determine the base at each position (Fig. 1.4, step 8). In addition, a base-calling algorithm generates a quality value for each base call by quantifying the fluorescence signal from each cluster – any reads from mixed clusters are

filtered out. The percentage of reads that pass the "purity filter" constitute the purity filter (% PF) parameter that can be used for quality control purposes.

Mapping software can then be used to align the sequence data against a reference genome (Fig. 1.4, step 9). Reads produced by sequencing of genomic DNA are very useful to close gaps in genome annotations whereas reads from RNA-seq experiments provide information regarding gene expression and splicing.

1.4 Genetics of malaria parasites

The 24 Mb genome of *P. falciparum* was published in 2002, followed by rodent parasite species [32–35]. These data have shed light on the basic genome architecture and identified key structural elements, common metabolic and biosynthesis pathways and unique aspects that are shared among several *Plasmodium* parasites [34,36]. These genomes are organised in fourteen chromosomes and encode around 5400 genes, of which more than 60 % encode proteins with weak or no homology to other eukaryotes. Furthermore, their adenine and thymine (AT)-content is unusually high, i.e. close to 80 % in *P. falciparum*.

The human and rodent parasites share roughly 85 % of the genes at the level of content and order, and *P. falciparum*-specific genes that disrupt the conserved genome segments are predicted to play a role in host–parasite interactions [32–35]. This synteny strengthens the credibility of using the rodent model in many functional genomics studies [35].

1.5 Recombination in malaria parasites

Under ideal conditions, a DNA molecule could subsist for a maximum of 6.8 million years, although it would probably no longer be readable by polymerases after about 1.5 million years [37]. However, inside a living cell, DNA is constantly under the damaging effects of elements such as free radicals and other reactive species generated by metabolism, UV and ionizing radiation, including gamma rays and X-rays and its integrity relies on the constant action of repair systems. Additionally, normal progress of cell division and differentiation also generate DNA double-strand breaks (DSB) that might lead to transcriptional errors or chromosome rearrangements [38]. Two main repair mechanisms ensure repair of these breaks as faithfully as possible to prevent genome instability. These are non-homologous end joining (NHEJ) and homologous recombination (HR). The latter uses a

homologous template from which the lost DNA sequences are copied whereas in NHEJ the break is sealed in a template-independent fashion [39]. The choice of method varies with the organism – higher eukaryotes use predominantly NHEJ while prokaryotes prefer the high fidelity HR method [40,41]. A third method has recently been described named micro-homology-mediated end joining (MMEJ), where DNA breaks are repaired through re-connection of the broken ends at regions of micro-homology which are typically of 5 to 25 bp in length. This method always results in deletions and is highly associated with oncogenic chromosome rearrangements and genetic variation in humans [39].

In addition to the DSB caused by replication events *Plasmodium* DNA is continuously damaged by the immune response of the host [40]. Although haploid for most of their life cycle, malaria parasites are surprisingly thought to heavily rely on HR to repair DSBs. Accordingly, not only is the canonical machinery required for NHEJ, such as DNA ligase IV and KU 70/80 missing, but also *Plasmodium* parasites seem to be unable to close a linearized plasmid [32,40]. Given its role in closely related organisms like *Toxoplasma* parasites, it is thought that the loss of NHEJ pathways is a relatively recent event. This feature, i.e. predominance of HR repair mechanism, is particularly relevant for reverse genetic studies as it implies minimal off-target integration of targeting vectors.

In the absence of any template it has recently been shown that *Plasmodium* parasites are able to use a form of end-joining (EJ) repair method that resembles MMEJ and includes resection of single-stranded overhangs and insertion of templated short sequences at the site of the break [39]. This is consistent with the very limited set of products obtained in a study where DSBs were induced in unique areas of the genome [40]. This atypical EJ mechanism was however shown to be very infrequent and thus the prime mechanism of DSB repair in *Plasmodium* is resolutely HR [40].

1.6 The rodent model of malaria

Malaria research greatly relies on rodent parasite models, as maintaining the complete *P. falciparum* life cycle in the laboratory is still a difficult task. Rodents are the best non-primate models. Since their isolation from the wild between 1948 and 1974 in Central West Africa, rodent malaria parasites – *Plasmodium berghei* (used in this project), *P. chabaudi*, *P. vinckei* and *P. yoelii* – have allowed the study of several aspects of host-parasite-vector interactions, as we can easily access parasite stages from mosquitoes and from host's livers,

and evaluate potential interventions for malaria control [42,43]. Although none of the rodent *Plasmodium* species is the perfect model for *P. falciparum*, different species can be used to study different aspects of the infection. For instance, *P. chabaudi* is a model that best recapitulates the interplay between the parasite and the host immune response while *P. berghei* is a better model for studying the biology of transmission [42].

Apart from the fact that the whole parasite life cycle can be studied in laboratory conditions, rodent malaria parasites are more amenable to genetic modifications than *P. falciparum* [44]. In fact, it can take as little as two weeks to obtain a clonal population of *P. berghei* transgenic parasites, while for the human parasites this number is extended to at least two months. On the other hand, the availability of fluorescent rodent parasite lines and transgenic mice permits exploration of specific host parasite interactions that are not yet available for studies of human malaria [45].

1.7 Genetic engineering – new tools for reverse genetics

Our understanding of the molecular cues that drive *Plasmodium* development and differentiation has been hindered by the fact that only a small proportion of genes have been assigned functions experimentally. Gene function can be studied using forward or reverse genetics approaches. The former aims at the identification of gene(s) responsible for a particular phenotype. In malaria parasites such approaches are commonly mediated by transposon mutagenesis, a technology that has been used both in *P. berghei* [46] and in *P. falciparum* [47]. In contrast, reverse genetics involves alteration or disruption of a particular gene to enable the study of its biological function.

Reverse genetics is the favoured approach for analysing parasite gene functions. Usually, it comprises three main stages: assembly of the targeting vector, transfection of the parasites, and phenotypic analysis of the mutants. Several issues have delayed reverse genetics studies in malaria parasites. Although stable transfection technologies became available for *Plasmodium* parasites in the late 1990's [48–51] it was not until 2006 when efficient protocols were developed for *P. berghei*, the most genetically tractable of the malaria parasites [44]. The reason for this is the fact that *Plasmodium* parasites spend most of their life cycle intracellularly, inside a vacuole, with its genetic material enclosed within four membranes (i.e. erythrocyte membrane, parasitophorous vacuole, parasite membrane and nuclear envelope). Secondly, *Plasmodium* DNA is very AT-rich (~80 %) and unstable in *E. coli*, which makes preparation of targeting constructs difficult [32,33]. The reduced number

of selection markers has also been a limiting factor, especially for *P. berghei* studies where drug toxicity for the animals is an issue. Currently, the human dihydrofolate reductase (h*dhfr*) and the toxoplama *dhfr-ts* genes are used to select mutants with pyrimethamine. This drug inhibits DNA synthesis of the parasite by blocking the NADPH-dependent reduction of dihydrofolate to tetrahydrofolate, a crucial step for the biosynthesis of purines, pyrimidines, and certain amino acids [52]. The h*dhfr* also allows selection with the drug WR99210, but this requires subcutaneous injections unlike pyrimethamine which can be delivered orally. Until recently, the combination of both selection strategies was the only solution for performing a maximum of two sequential genetic modifications to *P. berghei* genomes, such as mutation of two non-consecutive genes or complementation of a KO. The recent development of a positive/negative recyclable selection cassette (*hdhfr/yfcu*) has made such studies more feasible**.** Negative selection is a process by which the loss of a marker is selected as its presence produces a toxic substance upon exposure to a suicide substrate (i.e. a prodrug). The system *hdhfr/yfcu* allows positive selection of mutant parasites with pyrimethamine and negative selection with 5-fluorocytosine (5-FC). After 5-FC treatment, this pharmacologically inactive compound is converted into the highly toxic form 5-fluorouracil (5-FU) by the enzyme coded by the *yfcu* gene – a bifunctional protein that combines yeast cytosine deaminase and uridyl phosphoribosyl transferase (UPRT)[53]. Exposure to this toxic metabolite forces mutant parasites to promote excision of the selection cassette through homologous recombination, facilitated by the presence of a directly repeated *Pbdhfr-ts* 3'UTR flanking this cassette [54].

Current technologies for the genetic manipulation of *Plasmodium* rely on the transfection of blood stages and subsequent drug selection, as described. Therefore, genes encoding products essential to blood stage development are not amenable to be disrupted in this manner. Conditional approaches are therefore required to study these genes.

RNA interference (RNAi) is widely used in model organisms [55,56] and parasites like trypanosomes [57], as a means to reversibly silence gene expression at the post-transcriptional level. However, *Plasmodium* lacks the enzymes required for RNAi-based approaches [58]. As a result, although less efficient, other conditional strategies have been developed such as the Tet-inducible system [59,60] and the destabilisation fusion domains strategy (for *P. falciparum* only) [61]. Conditional approaches for the study of essential genes in mosquito or liver stages are less complex and strategies like the Flp system and promoter swap can easily disrupt or silence genes beyond the blood stages [62–64].

So far, reverse genetics has only targeted 10 % of the *P. berghei* genes (http://www.pberghei.eu/). In order to increase the number of genes to which functions have been assigned experimentally it is necessary to move from a gene-by-gene approach towards high-throughput approaches.

1.7.1 The Gateway technology: DNA cloning using site-specific recombination

The bacteriophage lambda (λ) uses a site-specific recombination system when switching between the lytic and lysogenic pathways to promote its integration into the *E. coli* chromosome [65]. This recombination system has two major components: recombination sequences (*att* sites) and a set of proteins that mediate the recombination reaction (Integrase (Int), Integration host factor (IHF) and excisionase (Xis)). The lysogenic pathway is catalysed by Int and the *E. coli* IHF whereas the lytic pathway is mediated by the phage Int and Xis and the bacterial IHF. It is highly specific and conservative (i.e. no net gain or loss of nucleotides).

Hartley and colleagues exploited this system to devise a strategy by which these recombinases mediate the transfer of DNA fragments flanked by *att* sites into vectors also containing *att* sites [66]. This system has been commercialised by Invitrogen (currently Life technologies) since the late 1990s under the name of Gateway Technology ® (GW). It is used for the cloning and transferring of DNA fragments between different expression vectors in a high-throughput fashion while maintaining orientation and the reading frame of the fragment of interest [67].

This system carries out two reactions:

(1) *att*B × *att*P → *att*L + *att*R mediated by Int and IHF

(2) *att*L × *att*R → *att*B + *att*P mediated by Int, IHF and Xis.

The direction of the reactions is simply controlled by the enzyme cocktail and the available *att* sites. Each of these reaction mixes is commercially available as “BP clonase” (1) and “LR Clonase” (2). The original *att* sites have been mutated to ensure directionality and irreversibility of the *in vitro* *att*L x *att*R reaction. Specifically, *att*L1 sites react only with *att*R1 sites and *att*L2 sites react only with *att*R2 sites. GW technology was highly relevant for this project as it was part of a restriction enzyme-free cloning strategy to generate targeting vectors.

1.7.2 Recombineering, a homologous recombination based cloning strategy

Bacteriophages have been studied extensively as a means to understand the principles of homologous recombination. This process can be defined as a type of exchange of genetic material in which information is exchanged between two identical or nearly identical molecules of DNA, in a precise and accurate fashion. Recently, some of these phage recombination systems were explored as tools for genetic engineering of plasmids due to their precision and simplicity. In fact, their use at scale has greatly boosted functional genomics studies in model organisms [68].

The use of phage homologous recombination systems to carry out genetic engineering is termed recombineering [68]. This is a highly efficient recombination system that enables *in vivo* modification and subcloning of large fragments of DNA, without the need for restriction enzymes or error-prone PCR amplification [69]. Initially developed based on two proteins RecE/RecT from the Rac prophage, the system has also been developed for the analogous *red* operon of the λ phage (used in this project) [70]. The *red* operon includes three elements:

- Redα, a 5' to 3' exonuclease (functional equivalent of RecE);
- Redβ, an annealing protein (functionally equivalent of RecT);
- Redγ, an inhibitor of the major *E. coli* exonuclease and recombination complex (RecBCD, responsible for degradation of linear dsDNA).

The process is initiated after a double strand break, at which point Redα digests one of the DNA strands leaving the other strand as a 3' ended, single-stranded, DNA overhang. Then Redβ binds and coats the single strand and this complex aligns with the homologous DNA so that the 3' end can become a primer for DNA repair (Fig. 1.5). This is further assisted by Redγ, which inhibits the RecBCD exonuclease activity of *E. coli*. These enzymes are extremely efficient requiring only a minimal length of homology region of 42 bp to initiate recombination [71].

In practice, generating a targeting vector using recombineering involves two steps:

1) Flank the engineered DNA (e.g. resistance cassette, epitope tag) with short sequences homologous to the target where it is aimed to integrate. Due to their short size, the homology arms can be synthesised as primer overhangs and used to amplify the engineered DNA;

2) Transform the resulting PCR product into a bacterial host that carries both the Red system and the target DNA (either as an episome or within the chromosome). Provided that all elements have been provided, the recombinases promote the integration of the engineered

DNA into its target location and the final vector can easily be recovered by standard genomic DNA (gDNA) extraction.

In short, using recombineering we now can generate recombinant molecules without the need for unique or special sites, with greater precision and regardless of the size of the target molecules thus rendering traditional cloning methods obsolete for most applications.

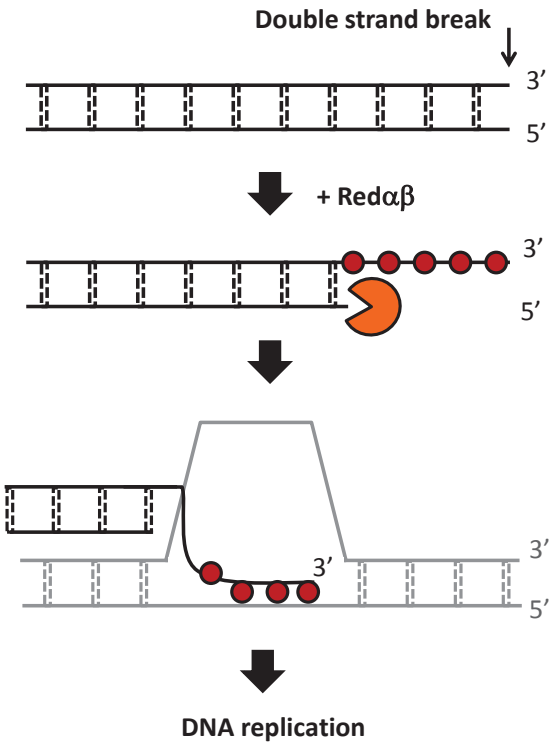

Fig. 1.5| Mechanism of Red recombination.
Upon a double strand break the 5'-3' exonuclease Redα (orange Pacman) digests one of the strands while Redβ (red circles) coats the resulting single strand. The homology regions from the engineered DNA (grey strand) are then used as templates for DNA repair.

1.7.3 Recombineering in *P. berghei* – the *Plasmo*GEM project

The recent generation of a *P. berghei* genomic DNA library using a low copy linear plasmid based on the bacteriophage N15 (pJAZZ, Lucigen) has now enabled the recombineering technology to be transferred to *Plasmodium* biology [72]. Currently, this library has 9113 clones with an average length of 9 kilobases (kb), and covers ~94 % of *P. berghei* genes. The high number of clones enables high coverage of most genes, for instance, 50 % of the genome is represented (at least partially) by at least five clones.

Together this library and the recombineering technology have set the basis for the development of the *Plasmodium* genetic modification project, *Plasmo*GEM. Launched in 2012 the *Plasmo*GEM project is based at the Wellcome Trust Sanger Institute (http://plasmogem.sanger.ac.uk/) and aims at producing a free community resource of genome-wide sets of genetic modification vectors for *P. berghei*. These vectors are generated through a highly efficient two-step recombineering pipeline, that has been scaled up to a 96-well plate format [73]. A brief description of each of the steps follows:

Step one: The gene of interest (GOI) present in a selected library clone is replaced *in vivo* by a positive/negative bacterial selection cassette *zeo-pheS* through recombineering (Fig. 1.6). This stage takes advantage of the short homology arms required to initiate recombination that are therefore synthetised as primer overhangs and used to amplify this cassette. At this point the Red/ET system is provided *in trans* as an inducible polycistronic unit in a temperature-sensitive, low copy number plasmid (pSC101-repA-BAD-gbaA). The pBAD promoter ensures that transcription of the recombinases (pSC101-repA-BAD-**gba**A) is activated only in the presence of arabinose, remaining otherwise inactive by action of the transcriptional repressor AraC. Two other layers of control have been added to the system: low copy number and thermo-sensitivity of the plasmid pSC101. The former is ensured by the origin of replication oriR101 and thermo-sensitivity is conferred by the protein RepA (pSC101-**repA**-BAD-gbaA), which is required for partitioning of plasmids to daughter cells at division [74]. This strategy therefore reduces the risk of undesired recombination events occurring as well as avoiding the recombineering plasmid to become a contaminant upon DNA extraction of the final product (*Plasmodium* vector) [68,75,76]. Finally, due to its role in general cellular integrity, recA (pSC101-repA-BAD-gba**A**) is also carried by the plasmid as the host cells are recA deficient in order to prevent recombination of the library clones [76]. After step one, the recombineered product is selected with the antibiotic zeocin, a broad-spectrum agent that induces double strand breaks of the DNA.

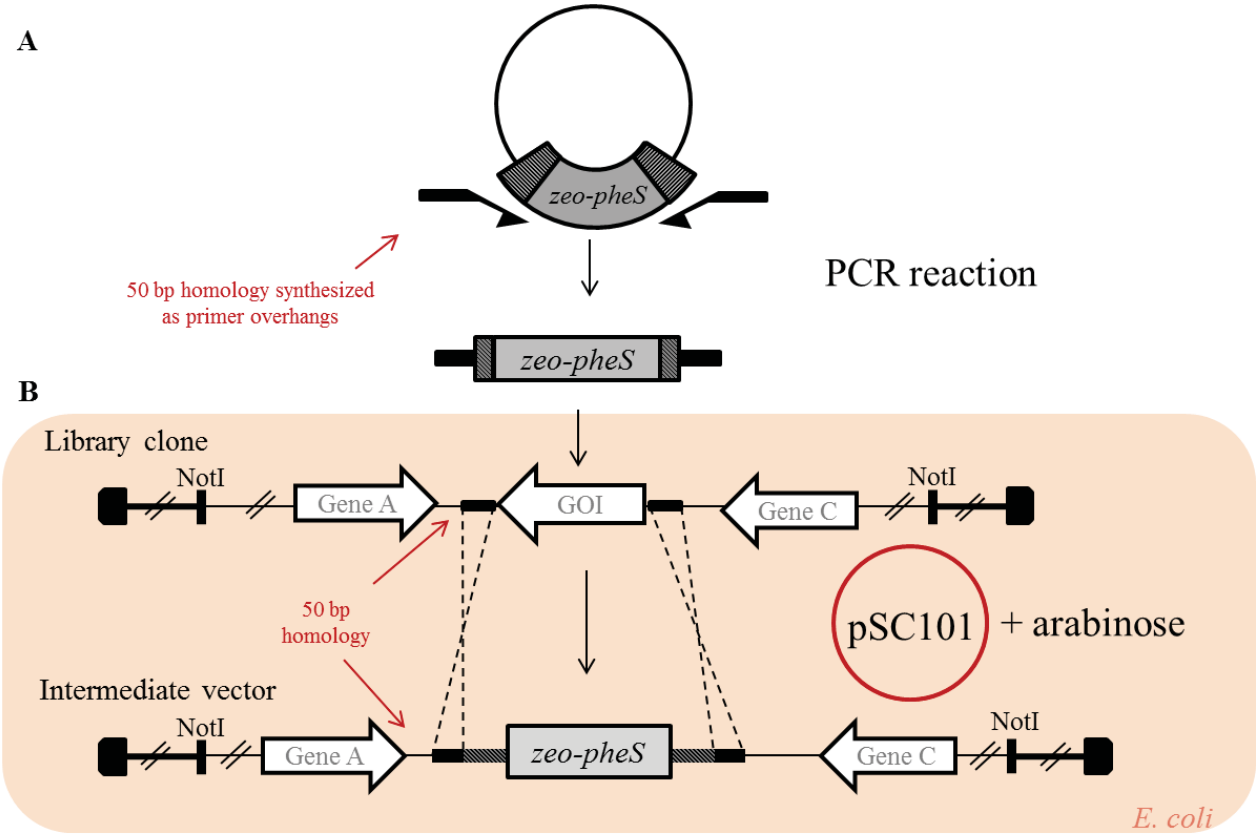

Fig. 1.6| First step – Recombineering reaction to generate a KO vector.
(A) A plasmid carrying the *zeo-pheS* cassette is used as template for a PCR reaction where the primers contain overhangs of 50 bp homologous to the flanking regions of the target gene. (B) This amplicon is electroporated into bacteria containing both the pSC101 plasmid and the library clone containing the target gene. Induction of the recombineering enzymes is accomplished with arabinose. Selection with zeocin is used to select the recombineered products. To generate a tagging vector, the approach is the same but the homology regions flank the stop codon in order to remove it.

Step two: the modified library clone is used as a substrate for an *in vitro att*L x *att*R GW reaction that replaces the *zeo-pheS* cassette by a parasite recyclable positive/negative cassette (*hdhfr-yfcu*) (Fig. 1.7) [54,73]. Negative selection against the *zeo-pheS* intermediate is then used to select the correct and final product – the *pheS* allele encodes a phenylalanyl-tRNA synthase *alpha* subunit that enables incorporation of the toxic phenylalanine analog p-chlorophenylalanine.

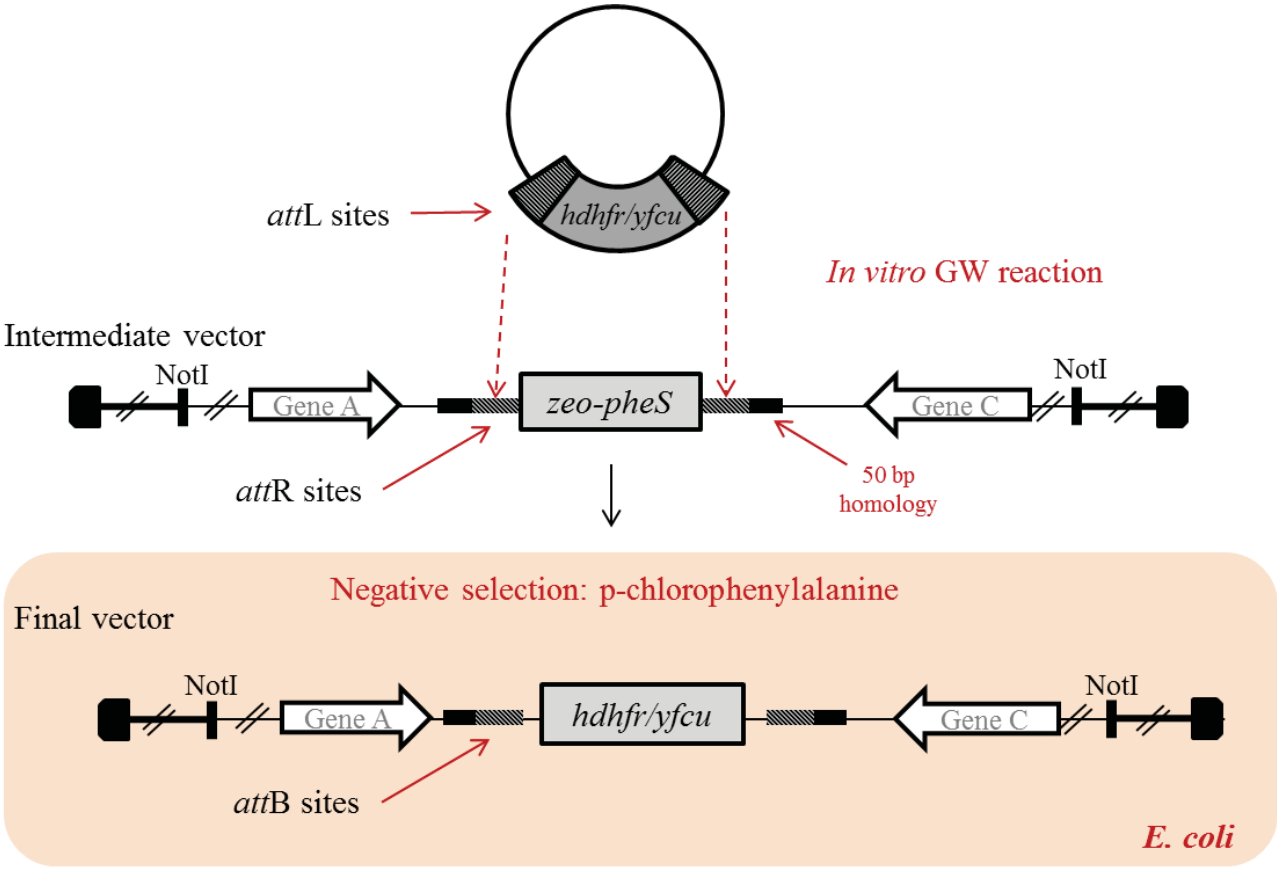

Fig. 1.7| Second step - Replacement of *zeo-pheS* cassette with *P. berghei* selection marker.
Gateway technology is used to replace the bacteria positive/negative selection cassette by the parasite selection marker. The reaction mix is then electroporated and the correct final product is isolated by negative selection.

This new approach has greatly reduced the time necessary for generation of large numbers of targeting vectors as its 96-well format has enabled the production of over 55 *Plasmo*GEM vectors in a period of less than two weeks, that only require *Not*I digestion to release the pJAZZ vector arms prior to transfection. Furthermore, as these vectors are generated from library clones they contain homology arms that are several kb long as opposed to the traditional 0.4-1 kb, which has greatly enhanced not only transfection efficiency but also transfection reproducibility [44,73,77]. Also, as these vectors are never circular, they are not expected to be kept in the cytosol as episomes.

1.8 High throughput reverse genetic screens

High throughput reverse genetics screenings have been vital for the understanding of model organisms at the molecular level. For instance, in *Saccharomyces cerevisiae* a library of a near-complete (96 %) collection of gene-deletion (KO) mutants has been generated [78], and around 9000 genes have already been targeted in highly germline-competent C57BL/6N mouse embryonic stem cells [79].

1.8.1 Signature tagged mutagenesis (STM)

Phenotypic analysis of mutants is the most laborious stage of any genetics screen, especially when *in vivo* infection models are required. Optimisation of plate-based bioassays has been accomplished, but only certain phenotypes can be analysed in this fashion, and although automated, each mutant is still screened individually. An efficient option to perform phenotypic analysis of large numbers of mutants is to analyse them in pools. Genetic footprinting was devised as a means to identify mutants within pools [80]. It involves the production of a large pool of mutants by transposon mediated mutagenesis, which will randomly include the disruption of the genes of interest, followed by inoculation of the pool under specific test conditions and the subsequent identification of the mutants that are present in the pool at the end of the experiment, by PCR. Some mutants, although present in the input inoculum will have been outgrown during the growth step due to the disruption of critical genes for the microbe's development. If, however, no detectable differences are observed between the pattern of PCR products before and after incubation it is likely that the disrupted gene is dispensable for survival under the tested conditions [80]. Genetic footprinting was first used to identify genes essential for the viability of *S. cerevisiae* under different growth conditions however, its low throughput encouraged the development of an alternative that would enable large-scale, parallel analysis of mutants [81]. This alternative was termed signature tagged mutagenesis (STM) and was presented by David Holden and co-workers in 1995 [82]. It is currently considered one of the most powerful and versatile large-scale genetic approaches to identify virulence determinants based on negative selection of attenuated mutants (i.e. mutants, which have lost the capacity to survive in a given host) [83]. Although very similar to genetic footprinting in its principle, the greatest advantage of STM is the ability to identify each mutant within a pool, not by individual PCRs, but by the presence of a

signature tag (i.e. unique short sequence of DNA) that is introduced into the microorganism upon integration of a transposon or targeting vector.

STM was first used to characterise virulence genes in *Salmonella typhimurium*. In this first report, a library of mutants was generated with a set of tagged transposons and then used to infect a rodent model of typhoid fever. Once infection was established, bacteria were recovered from the infected spleens and their tags were compared to the input pool using a dot-blot hybridisation approach. The tags included a 40 bp variable (and unique) central region that was flanked by constant annealing sites that also contained restriction sites. Such arrangement enabled not only the amplification of each tag by PCR and its subsequent cloning into the transposon, but also tag radio-labelling of the resulting pool prior to the hybridisation step [82].

STM has since been successfully applied to a wide range of microorganisms and yielded the identification of hundreds of new genes involved in virulence [84–89]. However, despite the preservation of the basic design (i.e. generation of tagged mutants followed by their propagation in pools and identification of the mutants present in the final pool through their tag), the variety of organisms to which this technique was applied led to the diversification of the mutagenesis method employed. For instance, *Neisseria meningitides* bacteria are not amenable to transposon mediated mutagenesis, therefore *in vitro* mutagenesis was used to generate mutants [90]. A summary of the mutagenesis methods is presented in Figure 1.8. These include *in vivo* and *in vitro* transposon mediated mutagenesis, shuttle mutagenesis, insertion–duplication mutagenesis by homologous recombination and gene replacement by homologous recombination [91].

The tag detection method was also optimised according to the available technologies and organisms. Hybridisation methods similar to the one used by Hensel *et al*. and variants that use digoxigenin or biotin labelled probes had as disadvantage the possibility of cross-hybridisation of probes that could lead to false positives. Although the construction of longer tags and pre-screen steps could help prevent such situations, alternative methods namely PCR-based approaches were developed. These included standard PCR reactions that relied on tag-specific primers and a flanking generic primer, less laborious multiplexed PCRs and also real-time PCR for a quantitative measurement of the abundance of each population of tags [91,92]. The yeast functional genetics field greatly benefited from STM. In 1996, shortly after Hensel's report, Shoemaker and co-workers implemented a new approach that enabled thousands of sequences to be analysed in parallel using high-density oligonucleotide arrays to detect the signature tags, termed barcodes, as they were gene specific [78]. This approach

allowed the measurement of the fitness cost of each gene deletion as differences in the intensities of the hybridisation signals reflected differences in the relative abundance of each population of mutants [93,94]. The libraries of barcoded mutants were generated by a gene-replacement strategy during which two 20-mer DNA barcodes – UPTAG and DOWNTAG, instead of one, were inserted for greater results confidence [94]. Despite being slower, the directed gene-replacement approach enabled prioritisation of targets and also made matching the presence of the barcode with the disrupted gene more straightforward [93]. Furthermore, the availability of libraries of single mutants simplified the validation of phenotypes seen in the pools as the original mutants were readily available.

Recently, NGS technologies were introduced as a means of barcode detection in STM experiments. This method, termed barcode analysis by sequencing or "Bar-seq" has been particularly relevant for the yeast functional genomics field and was shown to outperform the microarray detection methods (i.e. high-density oligonucleotide arrays) in terms of sensitivity, dynamic range, and limits of detection [95]. The Bar-seq strategy enables quantitative analysis of complex pools as the frequency at which a barcode is detected in the sequencing data is a measurement of the abundance of a given population of mutants within a pool [95,96].

In summary, STM is a very powerful and versatile technique that has revolutionised the functional genomics field as it can be applied to both *in vitro* and *in vivo* situations for a myriad of screens that include reverse and forward genetics and chemogenomic assays.

Mutagenesis method	
In vivo transposition	*Salmonella typhimurium, Mycobacterium tuberculosis, Vibrio cholerae, Yersinia enterocolitica, Legionella pneumophila, Brucella suis, Escherichia coli*
In vitro transposition	*Streptococcus pneumoniae, Neisseria meningitidis, Helicobacter pylori*
Shuttle mutagenesis	*Neisseria meningitidis*
Insertion-duplication mutagenesis	*Streptococcus pneumoniae*
Gene-replacement by homologous recombination	*Saccharomyces cerevisae*

In vivo transposition	1- Generation of a pool of tagged transposons. 2- Induce transposition into the organism of interest (OI). 3- Select clones with integrated barcoded transposons.
In vitro transposition	1- Transposition takes place *in vitro* (transposon activity independent of host factors) between DNA library of the OI and tagged transposons 2- Transformation of the OI with the mutagenized library. 3- Mutants are generated by homologous recombination.
Shuttle mutagenesis	1- Generation of a DNA library of the OI in a shuttle vector. 2- Transposition of barcoded transposons into this library takes place *in vivo* in a recipient organism. 3- Selected clones with the mutagenized library are introduced into the OI. 4- Mutants are generated by homologous recombination.
Insertion-duplication mutagenesis	1- Generation of targeting vectors by cloning DNA fragments from the OI. (synthetized by random or directed PCR) into barcoded vector. 2- Insertion of these vectors into the OI. 3- Mutants are generated by homologous recombination.
Gene-replacement by homologous recombination	1- Generation of a targeting vector that contains sequences that are homologous to the flanking regions of the gene of interest (GOI). 2- The vector should also contain a selection marker and a barcode. 3- Introduction of the vector into the target organism. 4- Homologous recombination generates a barcoded, resistant mutant.

Fig. 1.8| Summary of mutagenesis methods used for the generation of pools of mutants.
Top panel shows to which microorganisms the different methods were applied. Bottom set includes a brief description of each method. Adapted from ref [91].

1.8.2 Epistasis and genetic interactions

Genetic interactions can be described as biological phenomena that take place when the effect of a mutation depends on the genetic context in which it occurs. The same definition is given to the term epistasis as used in the population genetics field, although the classical definition by William Bateson restricts epistasis to a genetic interaction in which one mutation masks or suppresses the effects of another allele at another locus [97–99].

When a double mutant shows an expected, multiplicative phenotype compared to the corresponding single mutants it is very likely that these two genes do not interact either as parts of the same pathway or between pathways [100]. Genetic interactions can be classified as positive/alleviating or negative/aggravating (Fig. 1.9) [100,101]. The former refers to interactions where the simultaneous disruption of two genes yields a phenotype that is less severe than the phenotype expected from the sum of each independent mutation. Conversely, in a negative interaction the combined phenotype is more severe than expected, and in the most extreme cases (synthetic sick/lethal) the double mutation is lethal, unlike the single mutants [101]. A particular case of alleviating interaction is suppression. In suppression, the simultaneous perturbation (i.e. mutation or deletion) of two genes yields a wild type phenotype, despite the fact that each corresponding single mutant has an evident fitness loss. This is the case when one mutation counteracts the effects of another and is frequently associated with genes within the same pathway that also interact at the protein level [98]. Genetic interactions tend to occur among functionally related genes, although interactions of essential genes correspond to a broader functional range [100]. Although genetic interactions overlap with protein-protein interactions more often than expected by chance, such overlap is relatively rare, occurring at a frequency of less than 1% [102].

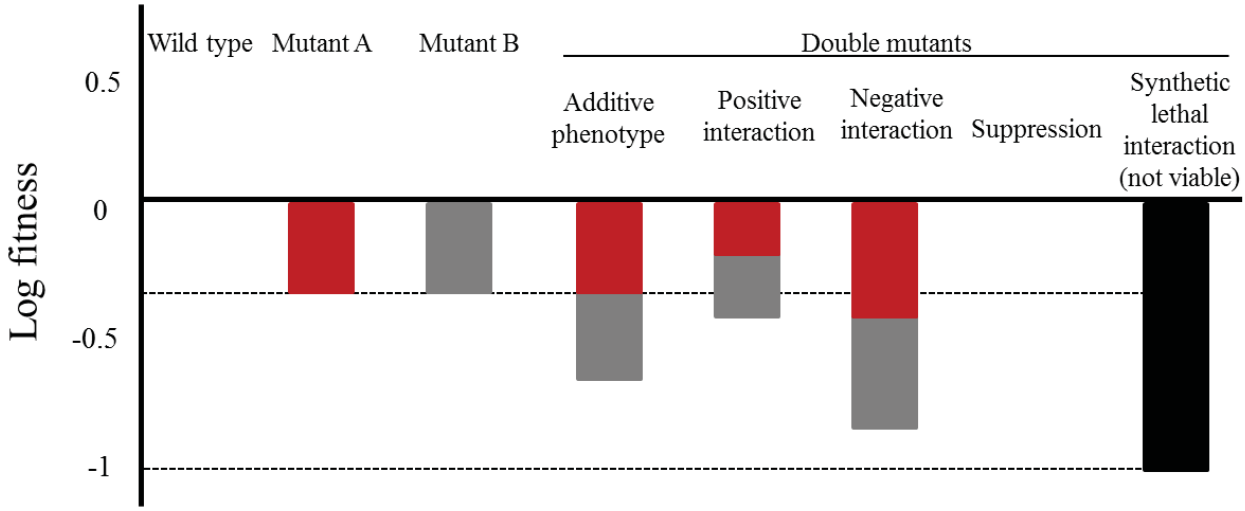

Fig. 1.9| Summary of genetic interactions.
The additive phenotype reflects the absence of genetic interactions. Positive or negative interactions take place when the phenotype of the double mutant is less or more severe than expected, respectively. Suppression happens when a second mutation counteracts the effect of a first mutation leading to a neutral phenotype. Synthetic lethal interaction is the most extreme case of a negative interaction and this double mutant is not viable.

Many recent insights into genetic interactions and networks have emerged from studies using *S. cerevisae.* Interestingly, under normal growth conditions, up to 80 % of *S. cerevisae* genes proved not to be essential for development, thus suggesting the presence of a high degree of interacting/compensatory pathways [94,103]. The availability of a collection of more than 6000 barcoded KO mutants allowed the test of 5.4 million gene-gene pairs for genetic interactions using a synthetic genetic array analysis (SGA) to screen the double mutants [104].

The highly complex life cycle of *Plasmodium* parasites suggests the existence of intricate signalling pathways. A systematic analysis of the *P. berghei* kinome suggested that only less than 35 % of the protein kinases were redundant for development unlike what was seen for the fission yeast (*Schizosaccharomyces pombe*) where 83 % of the protein kinases were amenable to deletion [105,106]. However, the scarcity of selectable markers available for *P. berghei* genetics has delayed genetic interaction studies as they require sequential modifications of the genome. In fact, only a very limited number of publications where sequential gene deletions were presented is available to date [107,108] and none involved signalling genes such as kinases, which due to their pivotal role in development and survival have been extensively targeted for genetic interaction studies in model organisms [109].

1.9 Protein kinases

A set of protein kinases was used throughout this dissertation to develop the screening method. For this reason a brief description of this family of proteins follows.

1.9.1 Eukaryotic protein kinases

Protein phosphorylation is the process by which kinases catalyse the transfer of phosphate groups from ATP to specific residues on their target proteins. It is a major regulatory mechanism that controls a myriad of cellular processes and is estimated to affect 30 % of the yeast proteome [110]. Protein kinases are one of the largest protein families, accounting for approximately 2 % of eukaryotic genomes [111]. For instance, the yeast genome encodes 127 protein kinases, while in humans the number increases to more than 500 [111,112]. The rapid and reversible nature of phosphorylation allows tight regulation of protein activity, localisation, stability, conformation and/or interaction with other proteins. Kinases themselves can be regulated in this fashion [113]. Kinase dysregulation is associated with a range of diseases, including vascular diseases, inflammatory disorders and cancers. For this reason kinases have been pursued as potential drug targets for the past three decades [114].

The catalytic domain of eukaryotic protein kinases (ePK) is characterized by highly conserved amino acids distributed in 11 subdomains (Fig. 1.10).

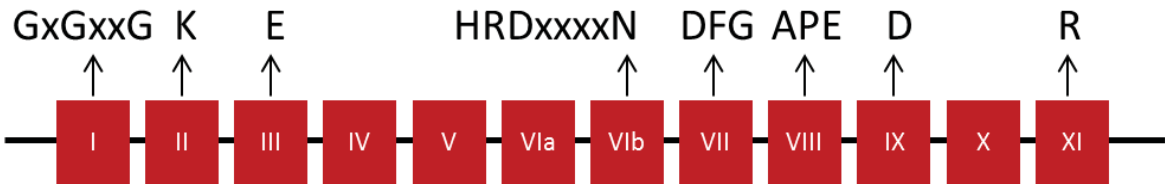

Fig. 1.10| Structure of ePKs catalytic domain.
The positions of amino-acid residues and motifs highly conserved throughout the ePK superfamily are indicated above the subdomains, using the single-letter amino-acid code with x being any amino-acid. The three glycine residues (GxGxxG) in subdomain I form a hairpin enclosing part of the ATP molecule; a lysine (K) in subdomain II, orientates the ATP molecules; a glutamate (E) in subdomain III forms a salt bridge with the former residue; in subdomain VIb aspartate (D) is thought to be the catalytic residue acting as a base acceptor; the aspartate in the DFG motif of subdomain VII, binds to the cation (Mg^{2+} or Mn^{2+}) associated with ATP; the glutamate (E) in subdomain VIII forms a salt bond with the arginine (R) in subdomain XI and provides structural stability of the C-terminal lobe; the aspartate in subdomain IX is involved in structural stability of the catalytic loop of subdomain VI through hydrogen bonding with the backbone. (Adapted from refs [115,116].)

According to their primary structure ePKs can be classified into seven major groups: ACG, CMGC, CaMK, CK1, STE, TKL and TyrK, that reflect broad functional categories [115,116]. Briefly:

1) **AGC**: includes the cyclic-nucleotide and calcium/phospholipid dependent kinases. E.g. PKA (cyclic-adenosine-monophosphate-dependent protein kinase), PKG (cyclic-guanosine-monophosphate-dependent protein kinase), PKC (protein kinase C) and related proteins;
2) **CMGC:** includes CDK (cyclin-dependent kinases), regulators of cell cycle progression, MAPK (mitogen-activated protein kinases), signal transducers that control effectors of cell cycle control and transcription, GSK3 (glycogen synthase kinase 3), also major regulators of cell proliferation and CLKs (CDK-like kinases) which play roles in RNA metabolism;
3) **CamK**: calcium/calmodulin-dependent kinases;
4) **CK1**: casein-kinase 1;
5) **STE**: includes PKs acting as regulators of MAPKs (STE stands for "sterile" and refers to the fact that these enzymes were first identified in a genetic screen of sterile yeast mutants);
6) **TKL:** tyrosine-kinase-like: related to tyrosine kinases but they are serine-threonine protein kinases;
7) **TyrK**: tyrosine kinases.

Atypical protein kinases (aPK) feature limited or no sequence similarity with ePKs however, they have demonstrated kinase catalytic activity experimentally. These can be divided in four groups: Alpha (e.g. myosin heavy chain kinase of *Dictyostelium*), PDHK (pyruvate dehydrogenase kinases), PIKK (phosphatidyl-inositol 3' kinase-related kinases) and RIO ('right open reading frame', as it was one of two adjacent genes that were found to be transcribed from a bidirectional promoter) [117,118].

A third group termed 'Other protein kinases' (OPK) allocates some ePKs that, although exhibiting some degree of sequence similarity to the main ePK groups, cannot be classified into any such group [117].

1.9.2 Protein kinases in *Plasmodium* parasites

Several independent groups have tried to assemble the kinomes (i.e. complete set of protein kinases) of *Plasmodium* parasites bioinformatically and according to the algorithm

used and the reference genomes available at the time, each study yielded slightly different results in terms of total numbers and classification [117–120]. The most recent analysis of *Plasmodium* was published by Miranda-Saavedra and co-workers and it compared twelve apicomplexan species using a validated computational tool, called Kinomer [117]. This kind of study relies on domain signature modelling of the conserved 11 subdomain structure of an ePK. Results showed that *Plasmodium* parasites have relatively small kinomes that represent between 1.2 % (*P. berghei* and *P. chabaudi*) and 1.6 % (*P. falciparum*) of their genomes [117].

The analysis of *P. falciparum* kinome yielded a total of 89 PKs with 65 ePKs, 19 FIKKs and 5 aPKs. Most of these enzymes were assigned to almost all major groups but some do not cluster with any group and some groups are not represented [117]. For instance, the CMGC group, which in other organisms gathers kinases involved in cell development and proliferation, contains the highest number of *Plasmodium* kinases (21 in *P. berghei* and 22 in *P. falciparum*). This is probably a requirement to meet the successive cycles of proliferation undergone as part of the parasitic life cycle. Examples include the CDKs, PfMAP-1/2, PfGSK3, PfCRK-1/3/4/5. The second most abundant group is the CamK (17 in *P. berghei* and 19 in *P. falciparum*) which evidences the importance of calcium signalling in *Plasmodium* development. Interestingly the CDPKs (calcium-dependent protein kinases) are part of this group which is only present in plants and other Apicomplexans – this topic is further developed in section 1.9.4. The CK1 group is the least represented group as in both *P. berghei* and *P. falciparum* only one *ck1* orthologue is found, unlike what has been found in other organisms where this group is expanded (e.g. 83/438 in *C. elegans* [121]). No *Plasmodium* PK clustered with either the STE or the TyrK groups. An important consequence of the former is the apparent absence of the canonical ERK1/2 pathways, despite the presence of MAP kinases homologues. In insects, plants, worms and mammals the TyrK group members are associated with hormone-response pathways. This suggests that this group of kinases arose as an adaptation to multicelullarity and therefore justifies its absence in both *Plasmodium* parasites and yeast [118]. Moreover, different mass spectrometry-based phosphoproteomic studies have confirmed that *P. falciparum* protein phosphorylation is involved in processes such as invasion and cytoadhesion, and also cell cycle control, DNA replication, transcription and translation [117,122].

Two main reasons for the seemingly low number of kinases found in *Plasmodium* genomes might be gene loss as an adaptation to the parasitic life cycle and the presence of *Plasmodium*-specific kinases that are too divergent to be found using the current methods that

are based on sequence similarity. Indeed, a set of *P. falciparum* PKs, named FIKK after their conserved four-residue motif in the kinase subdomain II (Phe-Ile-Lys-Lys) formed a tight cluster that despite containing most of the ePK conserved amino acids in the catalytic domain are not clearly related to any known ePK group [117,118]. So far, this family of kinases is restricted to Apicomplexa and is present as a single copy in *Plasmodium* parasites, except for *P. falciparum* and its closest relative *P. reichenowi* where it is expanded with 19 and six members, respectively [117,123]. All *P. falciparum* FIKK kinases have a variable extension N-terminal to their catalytic domain, which contain a PEXEL export signal motifs (required for exportation outside the parasitophorous vacuole), thus suggesting a role in host-parasite interaction for these enzymes [123].

Plasmodium parasites have also a small number of aPKs of the RIO and PIKK families. RIO proteins are widespread with at least two such enzymes present in organisms from Archaea to humans. In yeast these have been shown to be involved in rRNA processing and are essential for cell viability [117,124]. Similarly, both *P. falciparum* and *P. berghei* genomes have two orthologues and available data suggests that they are essential for the development of these parasites [105,125]. Enzymes belonging to the group PIKK (one in *P. berghei* and three in *P. falciparum*) are responsible for sensing DNA damage, nutrient-dependent signalling and nonsense-mediated RNA decay [117,126].

Altogether these studies showed that important divergences exist between malaria parasites and other eukaryotes such as the absence of the canonical ERK1/2 pathway components or the presence of Apicomplexa specific families of kinases. These differences mirror the obvious phylogenetic distance between Apicomplexa and Opisthokonta (lineage that includes *Homo sapiens*) and motivates the search for specific inhibitors as antimalarial drugs.

1.9.3 MAP kinases in *Plasmodium*

MAP kinases are serine/threonine PK responsible for regulation of cellular processes like mitosis, differentiation and cell survival that are activated through extracellular stimuli such as mitogens, osmotic stress or proinflammatory cytokines [127]. Two genes encoding atypical members of the MAP family, *map1* and *map2*, have been identified in both *P. falciparum* and *P. berghei*. However, while MAP1 seems to be related to the mammalian ERK7/8, MAP2, despite clearly belonging to the MAPK family, is somewhat divergent as it possesses a TSH motif in its activation loop motif instead of the canonical TxY (present in

MAP1) [119,128]. As previously mentioned, the STE group which includes a variety of kinases participating in MAPK signalling cascades is absent from the *Plasmodium* kinome. However, features of *Plasmodium* MAP kinases hint a different strategy for their activation. In the case of PfMAP-2, it has not only a TSH activation loop motif but also an insertion of about 26 amino acids at the N-terminal end. In addition, it seems to be phosphorylated and activated by the kinase PfNEK-1 which is not a member of the STE family but in this case acts as a MAPKK (MAPK kinase) equivalent. Moreover, the fact that all of these unusual features are present in other Apicomplexan parasites suggests a unique MAPK signalling mechanism [119].

Some functional differences have been observed between the MAP kinases of the rodent and the human parasites. While *Pfmap2* is likely to be essential for the erythrocytic cycle, *Pbmap2* can be deleted [129,130]. In the mutant, microgamete exflagellation is prevented, and hence parasite sexual reproduction and parasite transmission to the mosquito [130]. In contrast, the genetic inactivation of *map1* does not generate obvious phenotypic effects in either *P. berghei* or *P. falciparum* [129]. However, *Pfmap*2 is over-expressed in blood stages of the *Pfmap1* mutant [129] and *Pbmap2* transcripts are increased in liver stages of the *Pbmap1* mutant (Heussler, personal communication) suggesting that in both species the two MAP kinases may have partly overlapping functions and can partially complement each other's roles in development.

1.9.4 Calcium responsive kinases in *Plasmodium*

Signal transduction pathways allow cells to sense the environment and respond accordingly. This is a two-step process in which a first messenger (extracellular) binds a surface receptor followed by a second messenger that, within the cytoplasm, acts serving as chemical relay from the plasma membrane to the cytoplasm. Examples of second messengers are cyclic nucleotides (cGMP, cAMP) and calcium (Ca^{2+}). The latter is a major regulator of calcium-dependent protein kinases (CDPKs). These are abundant in plants (e.g. *Arabidopsis* has over 40 CDPKs) where they control a wide variety of processes including transcription and metabolism and also ion pumps and channels and the cytoskeleton [131]. Also present in apicomplexans (but not in animals), CDPKs seem to have a monophyletic origin as suggested by phylogenetic studies [132]. A CDPK contains three functional domains: a protein kinase catalytic domain, a carboxyl-terminal calmodulin-like domain with (usually) four EF-hands as

Ca^{2+} binding sites, and a junction domain between the kinase and the calmodulin-like domain [133].

In *Plasmodium* parasites calcium has been shown experimentally to regulate biological processes such as erythrocyte invasion by merozoites, motility and invasion by ookinetes and sporozoites, discharge of secretory organelles, and sexual differentiation in the mosquito vector [134–138]. In *P. falciparum* seven CDPKs have been identified (CDPK1-7) and found to be crucial for the development of the parasite. PfCDPK1 has been linked to motility and cell invasion mediated by microneme discharge and is thought to be essential for the intra-erythrocytic stages given the unsuccessful attempts to generate a KO mutant [139,140]. PfCDPK2 and PfCDPK3 are also likely to be essential during blood stage development although these have not been as explored [141]. Both PfCDPK3 and PfCDPK4 are thought to be involved in gametogenesis, while PfCDPK5 has been shown to have an essential role in the parasite egress from the erythrocytes [142–144]. Only PfCDPK4, PfCDPK6 and PfCDPK7 have been considered dispensable for parasite development at the blood stage [145].

The rodent parasites *P. berghei* have only six CDPKs, CDPK2 is missing, and unlike the human parasite four are not essential for completion of the intra-erythrocytic phase. The essential kinases include PbCDPK5 and PbCDPK7 and their function in the rodent parasites has not yet been elucidated [105]. Until recently, PbCDPK1 was considered to be essential during the asexual blood stages, therefore its function was best studied in the sexual stages where it was shown to activate translation of a subset of translationally repressed transcripts in the developing zygote stage and, its absence led to an arrest before the parasites could fully reach the ookinete stage [62,105]. Recent studies that include this project have, however, successfully deleted PbCDPK1 without an asexual growth phenotype despite its implication in invasion of *P. falciparum* parasites [146]. A function in ookinete's ability to invade the midgut epithelium has been proposed for PbCDPK3. Despite being able to glide through the blood meal in the mosquito gut the KO mutants are not capable of traversing the peritrophic membrane (i.e. protective layer formed from 12 h after a blood meal that protects the mosquito from pathogens) to reach the epithelial cells of the midgut [136]. Also not required for asexual stages, PbCDPK4 has a crucial role during male gametogenesis. Upon contact with xanthurenic acid (XA) in the mosquito gut, PbCDPK4 responds to a rapid increase in cytosolic calcium that occurs in gametocytes activating cell cycle progression, and promoting their differentiation into male gametes [138]. Finally, the absence of PbCDPK6 not only impairs sporozoite production but also renders the few produced sporozoites less infective to

hepatocytes likely due to a defect on the switch to the invasive phenotype, required for establishment of the liver infection [147].

Some CDPKs have their subcellular localisation determined by N-terminal acylation, which is thought to contribute to the specificity of such a small number of enzymes towards a ubiquitous secondary messenger [112,148]. Furthermore, the relatively small number of calcium effectors and the somewhat overlapping patterns of expression of the CDPKs at certain points suggest that not only do these enzymes interact with each other and perhaps other effectors but they might also have overlapping functions, thus conferring some redundancy to the system.

1.9.5 Defining the *Plasmodium* phospho-proteome

Fine-tuning of the intracellular machinery is achieved by several mechanisms that include transcriptional control, post-transcriptional control and post-translational modifications (PTMs) of proteins. PTMs are modifications of specific residues of proteins that often are reversible. Phosphorylation is one such alteration.

Advances in mass spectrometry based proteomic techniques have recently generated snap-shots of the phosphorylation status of organisms like *E.coli* [149], yeast [150] and even mice [151]. Since the apicomplexan parasites are quite divergent from most model organisms, dissecting *Plasmodium* signalling pathways is often a challenge as this cannot be fully achieved by bioinformatics predictions that use kinome and phospho-proteome data from other species [118]. For this reason in-depth phospho-proteome analyses of *Plasmodium* parasites have recently been performed [122,145,152,153]. Most phospho-proteome data has been produced through similar protocols and focused on the schizont stage as it is easily accessible in quantities that enable for mass-spectrometry based phospho-proteomics. Gene ontology (GO) analysis revealed that in schizonts the pospho-proteome is enriched for regulatory biological processes related to the basic transcription, translation and metabolic machinery which is in agreement with the schizogony process that these parasite undertake. Included in this dataset were at least 42 protein kinases [152].

The Treeck study [153] performed a comparison between the phospho-proteomes of *Toxoplasma gondii* and *P. falciparum* and showed that in these two related organisms the number of proteins implicated in secretory pathways that were phosphorylated suggests that these parasites might use phosphorylation as a means of regulating protein function outside their own boundaries. In *P. falciparum* parasites these included proteins located on the

parasitophorous vacuole and merozoite surface proteins such as MSP1 and MSP7. Secretion of the exonemes is required for merozoites egress in a process known to involve at least two kinases, CDPK5 and PKG. The latter was shown to be phosphorylated in its activation loop in a fashion consistent with autophosphorylation, thus, implying that PKG is not only regulated by changes in cGMP levels but also by phosphorylation. This phenomenon was shown to be present in at least 22 other protein kinases within the activation loop of the kinase, within one of the eleven kinase domains (e.g. CDPK1), or even outside the kinase domain (e.g. CDPK6)[145]. This suggests that the malaria protein kinases are organised in cascades, where the elements at the top are responsible for the phosphorylation, and therefore regulation, of their downstream partners.

Although we are still in the early stages of malaria phospho-proteomics research it is clear that phosphorylation is involved in most aspects of the parasite's biology and should thus be investigated to generate novel targets for pharmacological intervention.

1.10 Project aims

The major goal of this dissertation aims to overcome some of the limitations that currently prevent efficient reverse genetic screens in *Plasmodium* parasites. The specific aims of my work are:

(1) to explore the use of barcoded vectors with long homology arms as a means to carry out signature tagged mutagenesis in *P. berghei*;

(2) to develop barcode counting as a method for phenotyping the resulting complex mixtures of genetically modified parasites *in vivo*;

(3) to critically evaluate these new techniques by comparing their performance to conventional approaches using the parasite's protein kinases as a test case;

(4) to discover new kinase signalling pathways by conducting the first genetic interaction screen in *Plasmodium*.

The development of better therapies against malaria requires a deeper understanding of *Plasmodium* biology. Despite the establishment of methods for targeted genetic modification nearly two decades ago [51], only 10 % of the genes were assigned functions experimentally. As nearly half of the genome lacks any annotation, gene-by-gene strategies need to be scaled-up.

STM has been used extensively as a high-throughput method to identify microbial virulence genes by parallel phenotyping of pools of individually barcoded mutants [91]. These approaches encompass three different stages: (1) generation of pools of barcoded mutants, (2) propagation of the pool, and (3) detection of the mutants present in the final pool through their barcodes. The low transfection efficiency and high rate of false positives have prevented the development of such strategies for *Plasmodium* parasites. The recently developed *Plasmo*GEM vectors are linear and have improved integration efficiency due to long homology arms.

The first aim of this project was therefore to ask if STM technology could now be adapted to the rodent malaria parasite, *P. berghei*, and to use it to perform high throughput reverse genetic screens.

Chapter 3 describes the optimisation of the protocol. This involved maximisation of transfection efficiency, development of a parallel transfection strategy of pools of barcoded vectors and finally the development of a barcode sequencing approach, to read and count the barcodes of the pools of mutants generated. Chapter 4 presents the validation of the method

where a set of protein kinase genes was analysed by STM. In this chapter I also show how the barcode counting strategy enabled the analysis of the fitness of different mutants present in a pool.

As the STM technology enabled the accurate and reproducible measurement of fitness costs of single mutants growing within pools, another aim of this project was to detect genetic interactions through the analysis of the fitness of double mutants. Chapter 5 describes firstly how the presence of a recyclable selection cassette in each of the *Plasmo*GEM vectors enabled the generation of six selection marker free KO lines, and secondly, how these lines were used to perform genetic interaction screens. The latter involved measuring the fitness costs of the double mutants and comparing them to the fitness cost of the corresponding single mutants according to a multiplicative model for epistasis.