Chapter 4

STM analysis of protein kinase genes in *P. berghei*

4.1 Introduction

Following the development of an optimised method to count the relative abundance of barcodes of the *Plasmo*GEM vectors present in mixed pools, this chapter puts the new protocols to the test in a reverse genetics screen of eukaryotic protein kinases (ePK) genes of *P. berghei*. Protein kinases were chosen as test genes for these pilot experiments because they have recently been analysed systematically using conventional KO techniques [105]. According to the latter study, which served as a benchmark for the current analysis, out of a total of 66 ePKs, 23 are redundant for development of asexual stages, while 43 were identified as possibly essential as they could not be targeted.

The experimental setup of this pilot experiment included three stages (Fig. 4.1). Firstly, pools of *Plasmo*GEM targeting vectors were used to co-transfect triplicate pools of schizonts originally propagated for one passage in a Wistar rat[3].

Secondly, the resulting transfected parasites were injected into BALB/c mice as their inbred status reduced host variation to a minimum. Each of the three transfections gave rise to drug resistant parasites four days later, three days after drug selection was initiated. To monitor the composition of the parasite population as it developed in the host, small blood samples were collected from the tail vein, each day from day 4 to day 8 post-transfection, i.e. from the moment when parasitaemia became detectable on a Giemsa-stained thin blood film until the mice reached a critical health state. Additionally, our Home Office animal license did not permit daily blood sampling for longer than five consecutive days. Therefore, a total of five blood samples was collected from each mouse. Next, parasite gDNA was extracted using a phenol-chloroform method as maximal recovery was crucial.

Finally, these samples were used to amplify the vector specific barcodes by PCR, which were subsequently sequenced on a MiSeq instrument, using the optimised conditions described in Chapter 3.

An attempt to generate *Plasmo*GEM KO vectors for 63 ePK genes[4] by recombinase mediated engineering yielded a total of 41 barcoded vectors[5]. These included the CDPK group, MAP kinases, CRKs, RIO kinases, among others. Additionally, seven other KO vectors were prepared that served as additional references and controls. Four were vectors for

[3] The use of rats to boost transfection efficiency is documented in [159].
[4] Three ePK genes were not covered by the *Pb*G libraries.
[5] A list of these vectors can be found in appendix I

genes that were only expressed in sexual and mosquito stages: *p25*, *p28*, *soap* and *p230p*. The first three have well established functions in the ookinete: *p25* and *p28* are surface proteins, and *soap* is a secreted ookinete adhesive protein [164,165]. The last one is p230p gene, a redundant gene in *P. berghei* that is frequently used for transgene insertion [166]. A C-terminal tagging vector for this gene was used instead of a deletion vector because the latter did not integrate efficiently. For all these genes absence of a blood stage phenotype is well established and these mutants could therefore serve as references for normal growth against which fitness for all the other mutants could be calculated.

The other three constructs targeted metabolic enzymes, and were chosen as positive controls for detection of attenuated, i.e. slow, growth phenotypes. These were *plasmepsin IV*, an aspartic protease involved in haemoglobin degradation [167]; PBANKA_140160, a putative methyl transferase of unknown function[6] and PBANKA_110420, the E1β subunit of apicoplast branched chain α-ketoacid dehydrogenase [168].

Vectors were prepared as triplicate pools consisting of 100 ng of each of the 48 vectors.

[6] The selection of this target originated from personal observation in a parallel project.

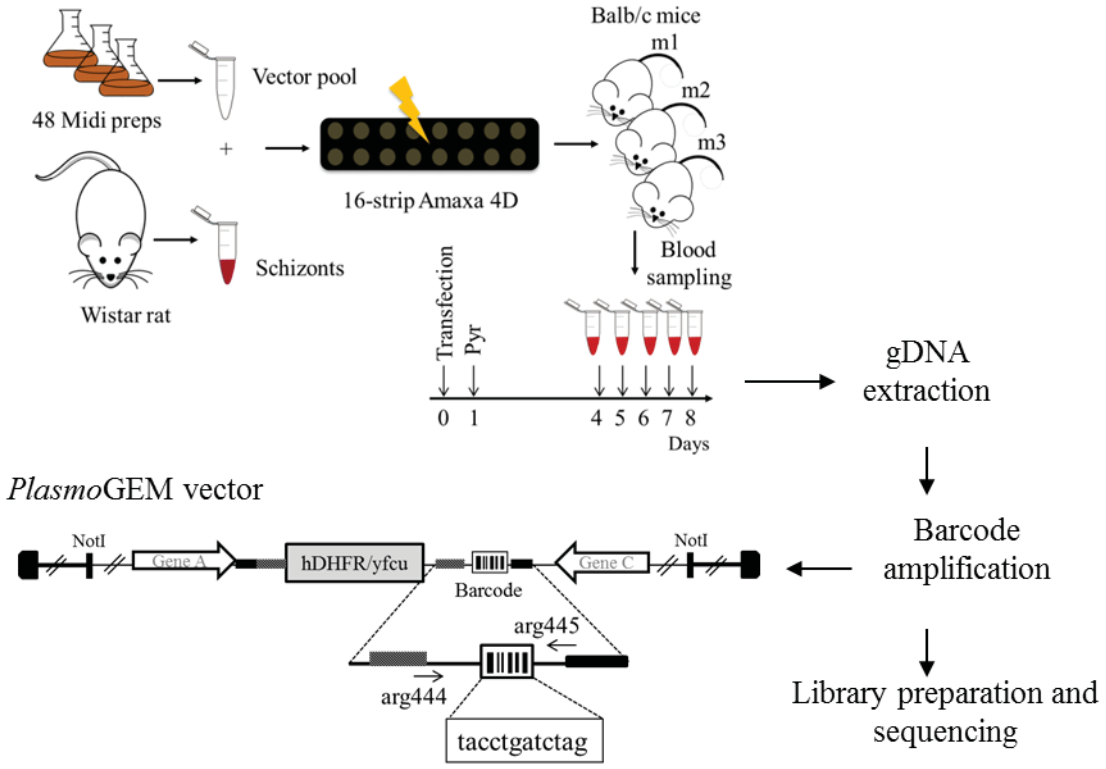


Fig. 4.1| Proposed experimental design for the *P. berghei* STM-Bar-seq experiments.
Midipreps for all vectors were prepared in order to ensure that enough material was available for all STM experiments presented in this thesis. Pools of vectors were transfected into schizonts using the 4D-Amaxa system and injected into three different mice. Drug selection with pyrimethamine started on day 1 post-transfection. Samples were collected on days 4-8 post-transfection and barcodes amplified and sequenced. Each time-point corresponded to an independent library. All libraries for each experiment were multiplexed in the same lane.

The main aims of this set of pilot experiments were: to determine if complex pools of mutants can be (1) generated through co-transfection of vectors and (2) followed over time; (3) to understand if the detection of barcodes corresponds to the detection of viable mutants and consequently if barcodes of essential genes are absent in the final pool; (4) how reproducibly is the generation of mutants from the same pool of vectors; (5) can the bar-seq strategy be used to measure growth rates during the infection, and (6) how can fitness costs be assessed.

4.2 Results

4.2.1 Barcode counting in *P. berghei* allows parallel phenotyping of mutants in a single mouse

The chosen workflow was successful, in that the overall parasitaemia of each mouse evolved as expected for a *P. berghei* infection, increasing rapidly during exponential phase at

ten-fold rate and then stabilised as it reached a growth plateau, with minimal variation across replicates (Fig. 4.2A). Five samples were collected from each of the three mice and a total of 16 were multiplexed and sequenced in the same MiSeq lane. The last sample, here termed the "input", was collected from the electroporation cuvette after the transfection. The data from this sample allowed detection and accurate measurement of the proportion of each vector at the moment of transfection.

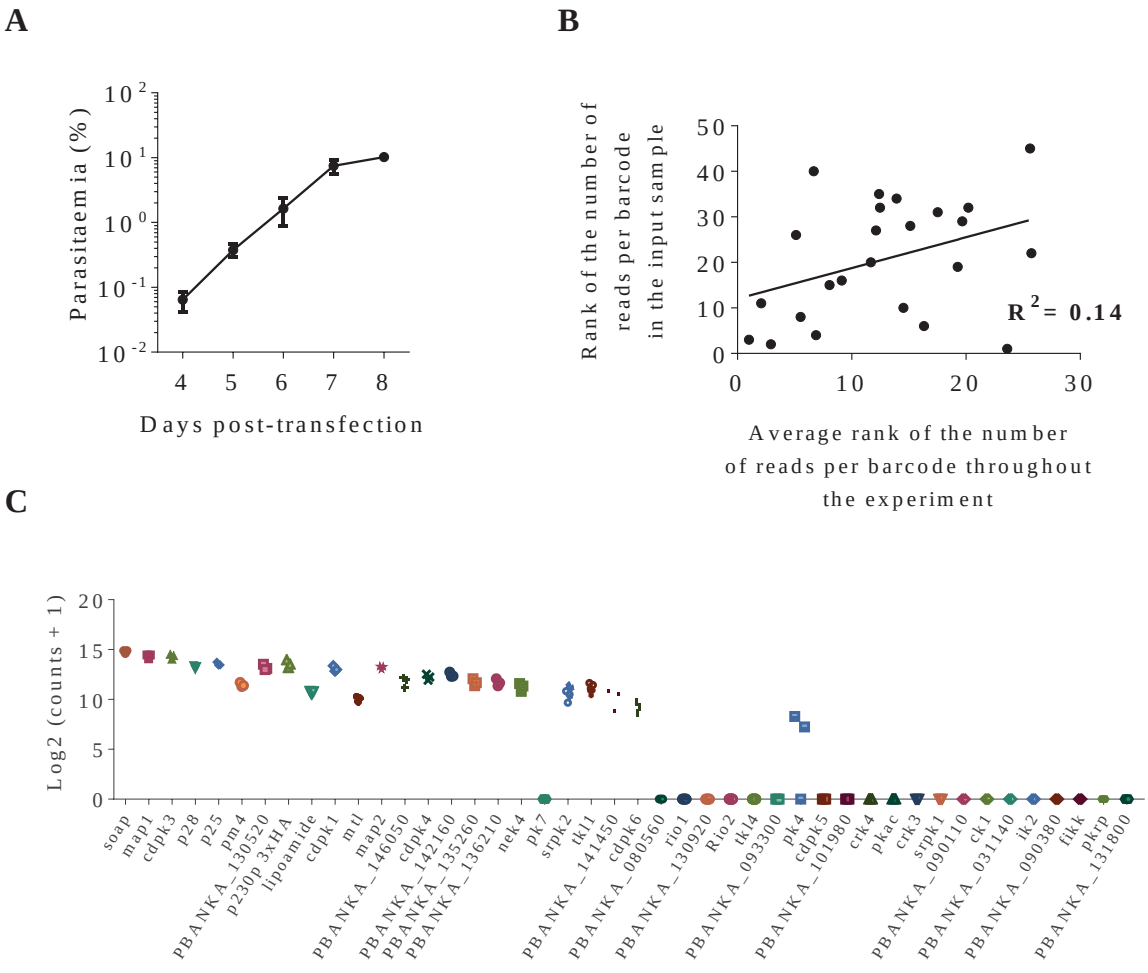


Fig. 4.2| Distribution of barcode counts for each gene and comparison with the input sample.
(A) Overall parasitaemia of the infections throughout the sample collection period showed very little variation between replicates. Error bars show standard deviations of the mean (n=3). (B) Pearson correlation analysis between the rank of the abundance of each barcode in the input sample and the average rank of the abundance of the corresponding barcode throughout the experiment. The low coefficient (R^2=0.14) indicated that the abundance of each vector in the transfection pool is not a major determinant of its subsequent abundance in the pool of mutants. (C) Distribution of the number of counts per barcode on day 7 post-transfection showed a range of abundances for the different barcodes within the pool.

Barcodes were extracted from the sequencing reads and counts per gene/barcode were calculated using a perl script developed by Frank Schwach. On average, three million reads

per lane (excluding PhiX reads) generated between 100 and 300 thousand barcode counts per library (time-point). The counting script was designed to be very conservative to reduce false positives to a minimum by counting only exact matches of a barcode flanked by at least five nucleotides of the barcode module.

The number of counts per barcode in the experimental libraries greatly varied from each other, unlike what was observed for the input sample (Fig. 4.2B, C). This suggested that a range of different abundances of each population of mutants (i.e. barcodes) had been generated despite the normalised quantity of DNA vector.

At this stage two different parameters were used to characterise each barcode: their abundance within the pool and their daily fold change. While the former reflects the integration efficiency and is also a consequence of the abundance of the respective vector in the DNA pool, the latter measures the growth rate, i.e. how the abundance of a given barcode changes over time.

Separate growth curves for all barcodes were calculated from their relative abundance and the overall parasitaemia over time (Fig. 4.3A).

As anticipated, the "normal growth" references (shown in green) were dominant throughout the experiment with highest growth rates (given by the slope of the curve). On the other hand, the attenuated references (shown in light orange) were less competitive and despite having started at similar numbers as the green references, their relative abundance decreased over time, also as expected.

Interestingly, the starting point of the growth curves was spread across at least two orders of magnitude despite the equivalent amount of DNA and parasites. No obvious correlation was found between the size of the homology arms of these vectors and the starting point of the growth curves (data not shown).

For the barcode counting to be used as a detection method of mutants it was crucial to investigate if detection of barcodes was equivalent to detection of integration events. Being linear, *Plasmo*GEM vectors are not expected to persist as episomes, which makes false positives very unlikely to start with. An analysis by Southern hybridisation of separated chromosomes (PFGE) showed genomic integration events throughout the genome. Interestingly, in the pool there were three vectors that targeted genes in chromosome eight[7] (PBANKA_080560, PBANKA_080800/*crk-4*, and PBANKA_083560/*pka*), but no band was present in the PFGE for this chromosome, which suggested that in the pool there were no

[7] The first two digits of the accession number indicate the chromosome number.

mutants with a gene disrupted in chromosome eight. Indeed, previously published data considered all three genes as possibly essential since integration of a KO vector in those loci was never obtained [105]. Further integration evidence was provided by PCR products showing integration of individual targeting vectors (Fig 4.3 C). Taken together, these data suggested that co-transfection of vectors can be used to generate pools of mutants.

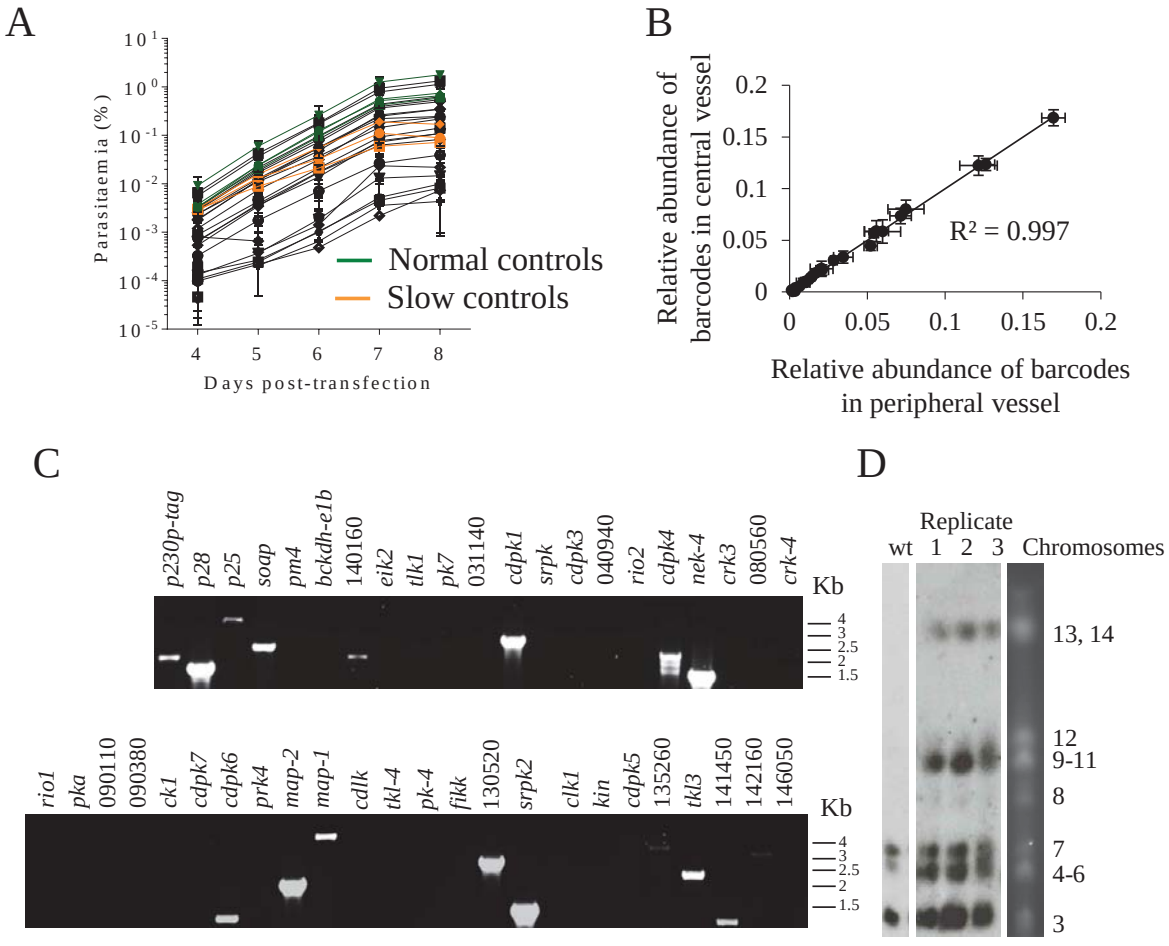


Fig. 4.3| Parallel transfection of pooled KO vectors generated pools of mutants

(A) Growth curves of the different mutants calculated from the overall parasitaemia and the relative abundance of each barcode. The "normal controls" were labelled in green and corresponded to the KOs of the following genes: *soap*, *p28*, *p25*, *p230p*-3xHA; while the "slow controls", labelled light orange, corresponded to KOs of *plasmepsinIV*, PBANKA_140160 and PBANKA_110420 (in descending order of abundance). The black lines represent all the other viable mutants present in the experiment. (B) Pearson correlation analysis between data generated from samples collected either from a peripheral or a central vessel. The high coefficient for these samples suggested that the route by which sampling was done was not a critical parameter. (C) PCR products across the predicted integration sites from the same experiment as in A and D, supporting genomic integration of many vectors whose barcodes were detected. The long homology arms of *Plasmo*GEM vectors can give rise to false negative results from PCR genotyping for some genes. (D) Southern hybridisation of chromosomes separated by PFGE with a ~500 bp probe for the 3' UTR of Pb*dhfr-ts*. This probe should target every integrated *Plasmo*GEM vector as it is present twice, flanking the drug resistance cassette and it should also detect the

endogenous *dhfr-ts*, located in chromosome 7 (positive control). Note that there is an exogenous site in on chromosome 3 in the background strain used due to the presence of a *gfp* gene that carried the same 3' UTR. Error bars show standard deviations from the mean (n=3).

Blood samples from days 4 – 7 were collected from tail blood (peripheral vessel), but on day 8 these were usually collected under terminal anaesthesia by cardiac puncture (central vessel) as larger samples were required for subsequent genotyping purposes. To ensure that the route by which sampling was performed would not impact on the distribution of parasites a correlation analysis was performed between the two types of samples collected on day 8 post-transfection. Very high correlation (R^2=0.997) between both samples was obtained suggesting that type of vessel used for sampling has no impact on the distribution of the parasites (Fig. 4.3B).

Growth curves were analysed using two parameters: the relative abundance of each barcode within the pool, and the relative fitness of each mutant, defined as the rate at which its abundance changed each day when compared to the average of the normal growth references. A second, independent experiment, also with triplicates, reproduced the data generated by experiment number one with remarkable accuracy. The parameters that were compared were the relative abundance of each population, the corresponding relative fitness, and the overall parasitaemia (Fig. 4.4A-C).

While both the relative abundance and the fitness parameters were highly reproducible between technical and biological replicates (Fig. 4.4A, B), only the latter, i.e. the shape of a growth curve, provided a quantitative measure for the fitness of a mutant. The relative abundance of a mutant within a pool was less informative since it was influenced by factors such as any local variation in recombination rates.

Fitness values of all four normal growth references were very similar and centred around the value one, i.e. their average, throughout time (Fig. 4.4D). Conversely, the average fitness of the attenuated references PBANKA_110420, PBANKA_140160, and *plasmepsin IV* was 0.65, 0.60, and 0.73, respectively, varying between 0.46 and 0.79 across all time points (Fig. 4.4 G-I). These significantly reduced fitness values (p-values < 0.05 for most time points, appendix V), confirmed my choice of attenuated reference mutants as adequate. As a result, these seven references were included in all STM experiments performed from this point onwards.

A closer look at the fitness of ePK mutants such as the CDPKs (Fig. 4.4E) revealed that the *cdpk3* mutants had an average fitness of 1.01 while *cdpk4* mutants were 10 % less fit. The

latter had not been detected by conventional methods where parasitaemias were counted manually [138]. Similarly, *cdpk6* mutants also had an average fitness of 0.88 but unlike *cdpk4* this decrease in fitness was significant on days 6 and 7 (appendix V). These findings reinforced the idea of a dynamic interplay between parasites and the host during infection that generates a multitude of phenotypes that can change daily, i.e. with every cycle. Despite being present in the input sample, no barcode counts were detected for either *cdpk5* or *cdpk7* in any of the replicates, an outcome that matched previously described data [105] where the disruption of these genes did not yield viable mutants. The same result, i.e. high number of counts in the input sample and complete absence in the experimental samples in this or other reverse genetics studies, was observed for other targets such as *fikk* and *crk4*, clearly strengthening the credibility of this approach. Quite surprising was the detection of the *cdpk1* KO mutant that, at the time when I performed these experiments, was still considered likely essential for asexual stage development [62]. In accordance with previously published data [129] the *map* kinase mutants (*map1* and *map2* KO) had growth curves very similar to the normal growth references and average fitness values of 1.02 and 1.00, respectively (Fig. 4.4F).

Importantly, barcode counting yielded reproducible fitness measurements that were independent of the relative abundance of a vector in the transfection pool as shown in Figure 4.4J. In this Figure it is also shown that the distribution of fitness values varies between 1.00 and around 0.50. Presumably lower fitness values could not be measured *in vivo* where fitter parasites will have outcompeted very slow growing mutants by the time the infection becomes patent.

For the analysis presented here, mutants were considered viable when counts for a given barcode were consistently present in every time-point for at least two out of three replicates.

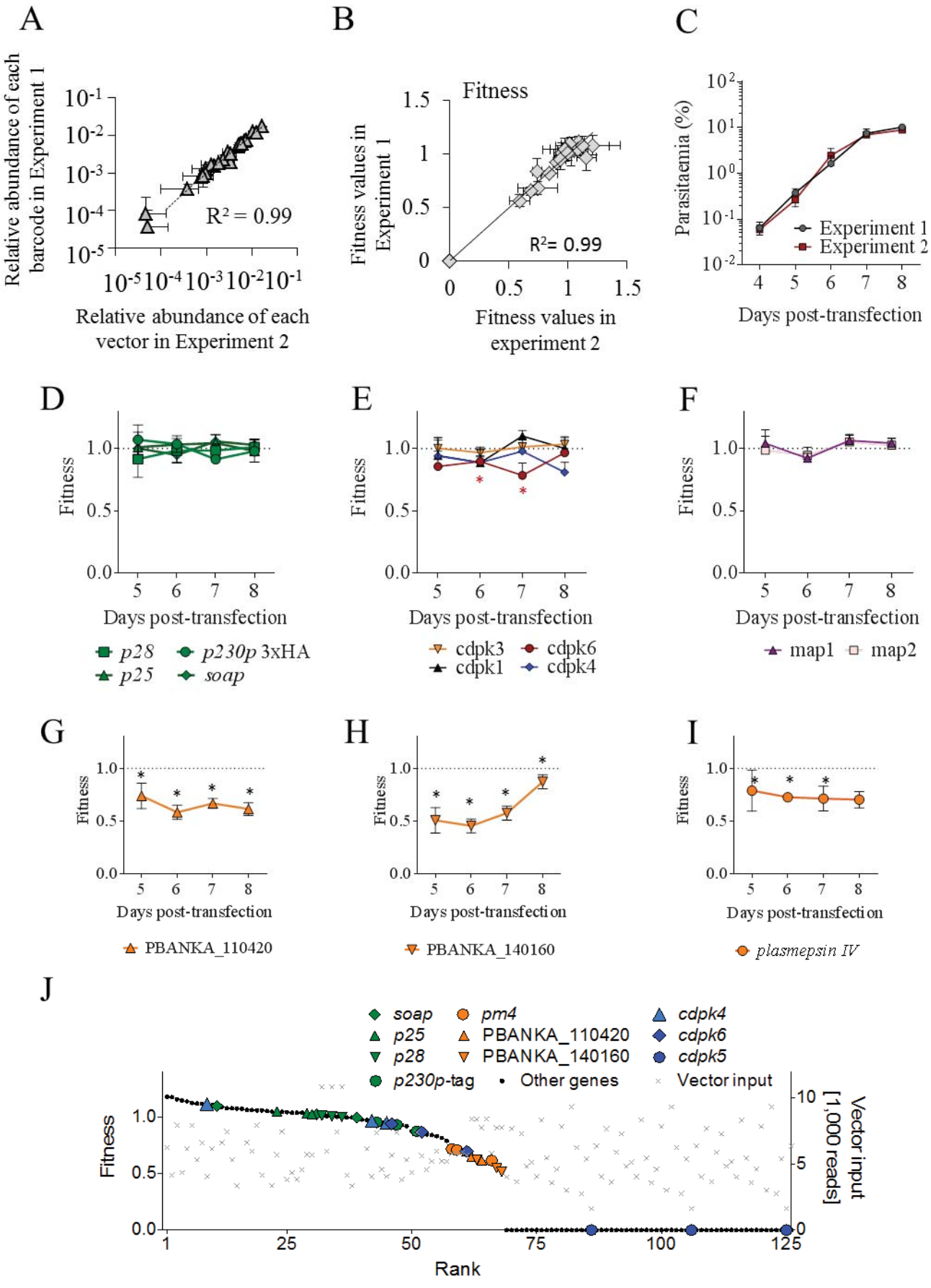

A
Relative abundance of each barcode in Experiment 1
R² = 0.99
Relative abundance of each vector in Experiment 2
B
Fitness
Fitness values in Experiment 1
R²= 0.99
Fitness values in experiment 2
C
Parasitaemia (%)
Experiment 1
Experiment 2
Days post-transfection
D
Fitness
Days post-transfection
p28
p230p 3xHA
p25
soap
E
Fitness
Days post-transfection
cdpk3
cdpk6
cdpk1
cdpk4
F
Fitness
Days post-transfection
map1
map2
G
Fitness
Days post-transfection
PBANKA_110420
H
Fitness
Days post-transfection
PBANKA_140160
I
Fitness
Days post-transfection
plasmepsin IV
J
soap
pm4
cdpk4
p25
PBANKA_110420
cdpk6
p28
PBANKA_140160
cdpk5
p230p-tag
Other genes
Vector input
Fitness
Vector input [1,000 reads]
Rank

Fig. 4.4| STM revealed a range of growth phenotypes.

(A) Pearson correlation analysis of average abundance values for each barcode, on day 7 post-transfection, of two independent STM experiments. (B) Pearson correlation analysis of average fitness values for each barcode, on day 7 post-transfection, of two independent STM experiments. Fitness was calculated from the replication rate of each barcode from day 6 to day 7 post-transfection relative to the average of the reference genes on the same period of time. (C) Overall average parasitaemia of the same two independent STM experiments. Note that the curves nearly overlap. (D) Fitness over time of the normal growth reference genes. Their consistency over time was striking, hence confirming that their choice as references was adequate. Fitness of reference mutants averages 1 by definition. (E, F) Fitness over time of selected mutants as labelled. The *cdpk1* gene had previously been considered essential for asexual development but it was not significantly less fit than the normal growth references; unlike the *cdpk6* KO parasites. (G-I) Fitness over time of the attenuated references. Statistical analysis indicated that these were significantly less fit than the normal growth references at nearly every time-point.
Error bars show standard deviations of the mean (n=3). * Different from reference mutants as determined by a two sided T-test corrected for multiple testing ($p<0.05$). (J) Distribution plot generated from a ranked list of day-6 fitness values measured for each gene in Experiment 1 (left axis). The relative abundance of a targeting vector in the electroporation cuvette at the moment of transfection, given by the input sample (grey crosses, right axis) did not predict whether a mutant could be obtained.

4.2.2 Comparison between barcode counting and a conventional deletion analysis

Different STM experiments yielded data for 46 ePKs for which there were published data available [105]. Barcode counting agreed with the previous study for 35 out of 47 ePK genes (76%), of which 15 were targetable and 20 were classified as likely essential ePKs genes. In the Tewari *et al* study [105], used as benchmark for this study, the authors targeted two ePKs that the STM approach failed to generate: *uis1* and *pk7*. It is uncertain whether such technical failures resulted from faults in individual vectors or from low recombinogenicity of the target loci. Excluding these two false negatives, the bar-seq strategy provided evidence for targetability for 25 genes, while the conventional approach targeted only 15 (plus *usi1* and *pk7*) [105]. The ten genes that had previously been considered possibly essential for erythrocytic development but were now targeted were *cdpk1*, *gsk-3*, *rio1*, *rio2*, PBANKA_082960, *tkl3*, PBANKA_141450, PBANKA_142160, *tkl1* and PBANKA_130520 [62,105]. As this corresponded to a rate of false positives of 34.5 %, I attempted to validate the first six by independently generating and carefully genotyping the mutants. This is further detailed in section 4.2.3. A table summarising this comparison between studies and including the daily fitness values for each mutant as well a description of the type of genotyping performed in each of the validation cases is presented in appendix V. The statistical analysis was also performed for each of the mutants comparing their growth rate to that of the normal growth references was also included in the above mentioned appendix.

4.2.3 Validation of false positives

If barcode counting detected non-integrated or incorrectly integrated targeting vectors, this would be a major disadvantage of the new method since it would result in false positives in STM screens. Therefore, it was important to validate mutants that were not previously considered viable. The generation and genotyping of KO clones was attempted for six of the ten new deletions detected by barcode sequencing. These were: *cdpk1*, *gsk-3*, *rio1*, *rio2*, *tkl3*, and PBANKA_082960 (Fig. 4.5). *cdpk1*, *gsk3*, PBANKA_082960 KO and *tkl3* parasites were easily generated and PFGE analysis showed clear integration of the vector in the expected chromosomes, i.e. 3, 4, 8 and 13, respectively (Fig. 4.5A). At this points *gsk3* and *cdpk1* KO parasites were cloned and further genotyping by PCR was performed (Fig 4.5 B,C) as these were two potential drug targets in *P. falciparum* [169,170]. Shortly after the generation of the *cdpk1* mutant, Jebiwott S. *et al*. reported a similar KO [146] and confirmed the phenotype at the ookinete stage, which previous work had described using a stage specific KO [62]. A short phenotypic analysis was carried out for the *gsk3* KO mutants (section 4.2.3.1). Cloned mutants for *tkl3* and PBANKA_082960 were generated as well but for genotyping purposes only. PCR data and gene maps are depicted in appendix VIII.

Kinases of the RIO family are atypical protein kinases involved in cell cycle progression and ribosome biogenesis [171–173] and have been considered essential in yeast and human cells [171–173]. These kinases have also been considered as possibly essential in both *P. berghei* [105] and *P. falciparum* [125]. However, barcode counts for both kinases, although sometimes at levels lower than the noise threshold, were consistently detected in different experiments.

The obtained RIO kinase mutants were harder to genotype and interpret. Unlike the other mutants, where harvesting took place on day 8 post-transfection, drug resistant parasites took until day 10 post-transfection to reach a parasitaemia of 2 % after transfection with a *rio1* KO vector. When targeting the *rio2* gene, parasites emerged even later, on day 14 post-transfection. PFGE results showed integration of the vector into chromosome 5 for *rio2*, as expected, but for the *rio1* KO mutant population, the intensity of the expected band on chromosome 14 was very weak (Fig. 4.5A). Since the 3'UTR of *Pbdhfr* was used as probe, the two copies present in the vector that flank the resistance cassette should have yielded twice the intensity of the endogenous sequence (internal control) on chromosome 7, which was clearly not the case, suggesting that not all parasites had integrated the vector. PCR

genotyping (Fig. 4.5D, E) showed the presence of the *rio* KO vectors (lanes 401/216, 403/218 and 396/218, 398/216)[8], but remained inconclusive since no integration PCR product could be obtained (lanes 405/216 and 400/218). In addition, both *rio* KO populations were still PCR positive for their respective WT sequence. Four attempts – two for each strain – at dilution cloning single parasites failed, and reducing the dilution factor to five parasites per mouse produced what appeared to be mixed populations of mutant and WT (not shown). Taken together these findings suggested that deletion of *rio1* or *rio2* produced mutants that were too unfit to produce infections as clones on their own. To characterise the target loci of both mutants further, non-clonal populations resulting from two different infections following dilution cloning for each mutant where subjected to whole genome sequencing (WGS).

Paired-end reads for the four libraries were mapped to the *P. berghei* ANKA reference using the Burrows-Wheeler Aligner (BWA) algorithm, manipulated with SAMtools and viewed with Artemis. The resulting alignments are depicted in Figure 4.6 for the *rio1* and Figure 4.8 for the *rio2* mutants. Targeting of *rio1* (lines blue and red) led to a substantial drop in coverage that corresponded exactly to the deletion region expected by the KO vector (Fig. 4.7A). The *rio2* populations served as controls. This indicated that the gene was deleted in all but a small number of WT parasites, thus confirming all other genotyping data.

The KO vector for *rio2* was expected to target the entire ORF of the gene (Fig.4.7B). However, a comparison of the coverage levels between all libraries at this locus (Fig. 4.8), revealed a ~29.7 kb duplication on either side of the *rio2* gene; coverage for the *rio2* itself was down to the reference levels. These results are consistent with a scenario in which a regional duplication including *rio2* had occurred in the parental parasite, and after transfection, the KO vector deleted only one of the two copies. These data are consistent with PFGE results, i.e. the vector integrated into the right chromosome, but also with the PCR data, where persistence of the WT locus was shown. Given these results, I speculated that, unlike *rio1*, *rio2* is likely to be essential and the integration of the KO vector in this locus was only possible due to the duplication. This also explained the delayed patency of these mutants as this duplication was, most likely, a very rare event.

[8] Gene maps illustrating binding sites of the different primers are available in appendix VI.

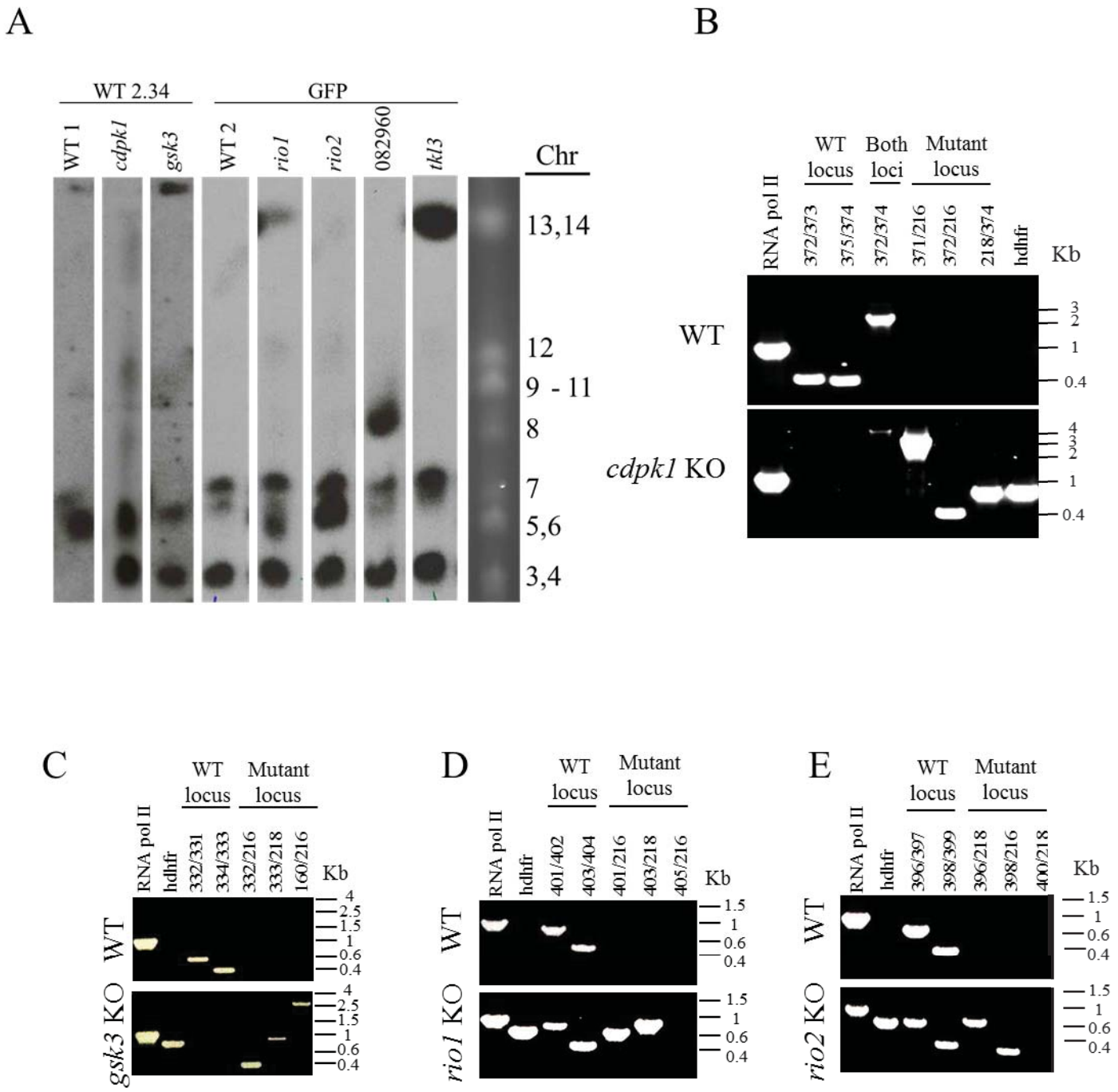


Fig. 4.5| Genotyping of the newly obtained mutants *cdpk1* KO, *gsk3* KO, PBANKA_08296 KO, *tkl3* KO, *rio1* KO and *rio2* KO.

(A) Southern hybridisation of chromosomes separated by PFGE with a ~ 500 bp probe for the 3' UTR of Pb*dhfr-ts*, as before. Note that there is an exogenous site in on chromosome 3 in the background strain labelled WT2 (GFP parasites). (B-E) PCR genotyping for *cdpk1* KO, *gsk3* KO, *rio1* KO and *rio2* KO to support the PFGE data shown in A. The first lane targeted *rna polymerase II* and was used as a positive control; the second lane targeted the resistance cassette and therefore should always be positive in the mutant but not in the WT lanes. The "WT locus" PCRs span either the 3' or 5' end of the genes. "Mutant locus" PCRs primed both inside the drug resistance cassette and outside the ORF. Maps showing the annealing sites of the genotyping primers are present in appendices VI and VII for the *rio* and *gsk3* KOs and in Figure 5.1A for *cdpk1* KO. PCR genotyping evidence as well as gene maps for PBANKA_082960 KO and *tkl3* KO can be found in appendix VIII.

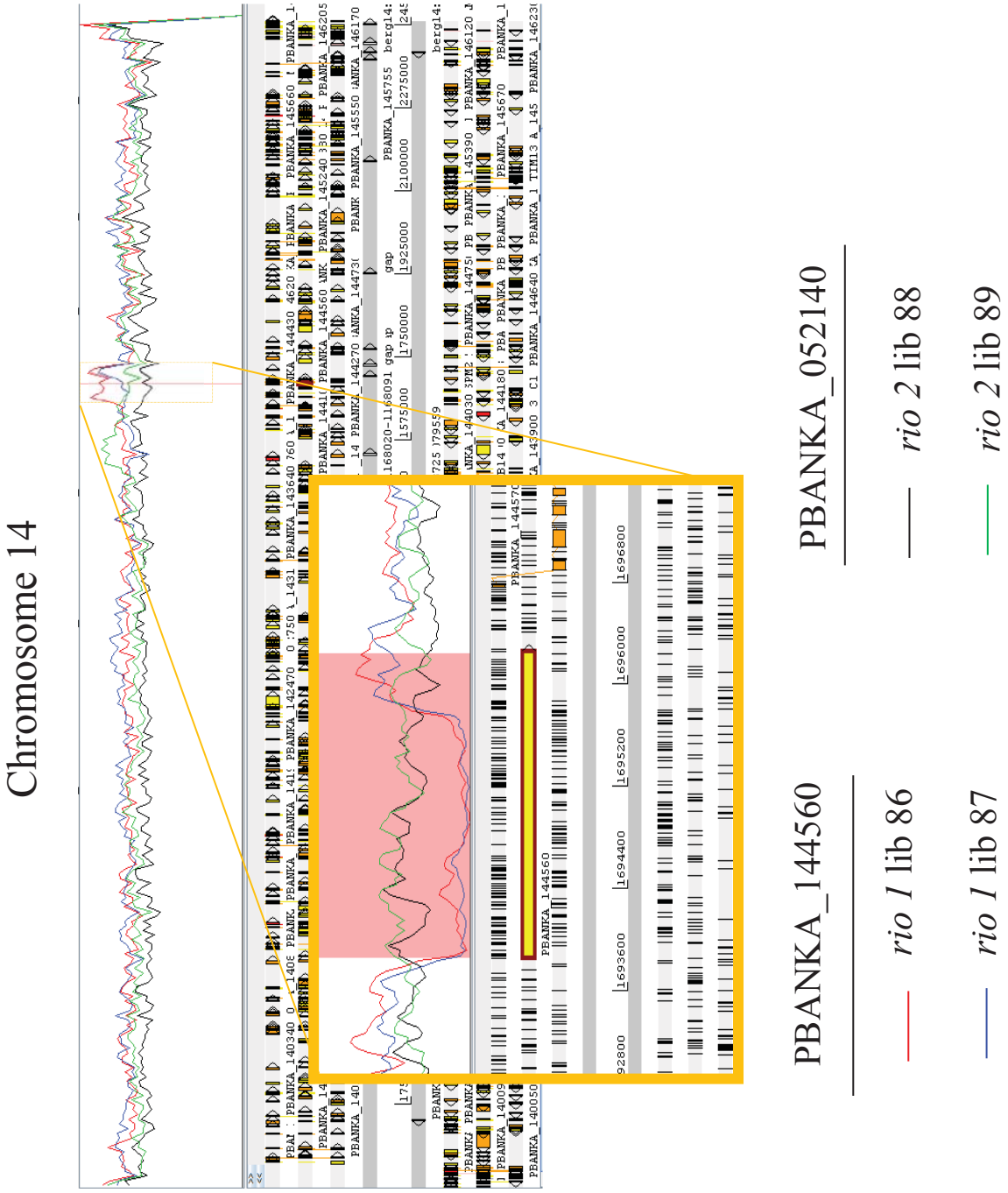
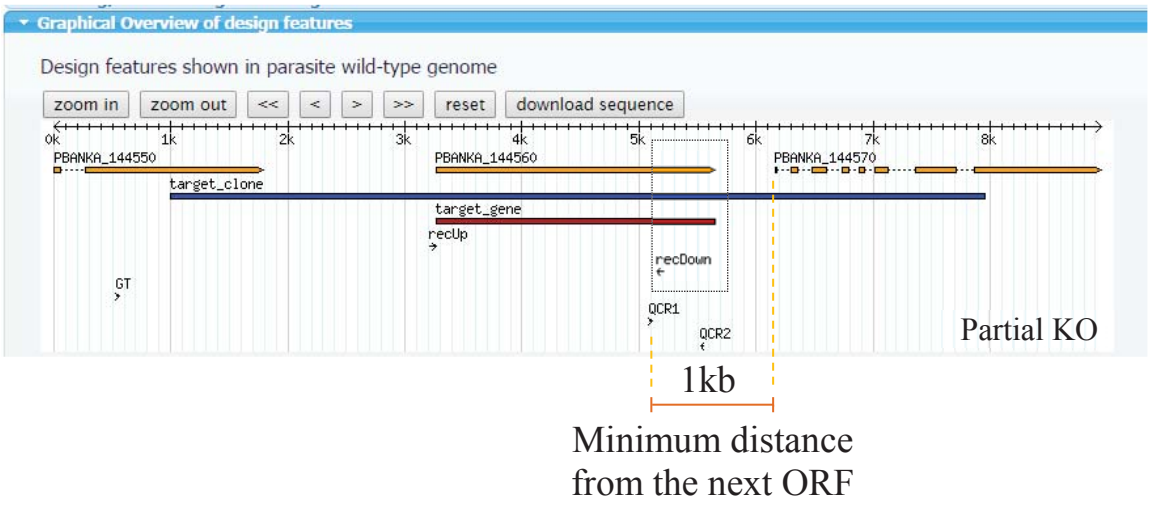


Fig. 4.6| WGS of RIO kinases - *rio1* locus is disrupted in the *rio1* mutants.
Mapped reads were visualised with the Artemis software. The Figure shows the coverage plots for the entire chromosome 14. Each line corresponds to each of the sequenced mutants (individual libraries – 86 to 89). Coverage across the chromosome was very even except for the KO region where the drop for libraries 86 and 87 indicated that *rio1* gene had been deleted. The rise in coverage at the last 500 bp of the gene is a result of the vector design, i.e. the vector did not target this region (please see Figure 4.7A). The pink highlighted region in the zoomed rectangle spans the *rio1* ORF (PBANKA_144560).
The numbers that are not preceded by "PBANKA_" are chromosome coordinates.

A

rio1 KO vector design



B

rio2 KO vector design

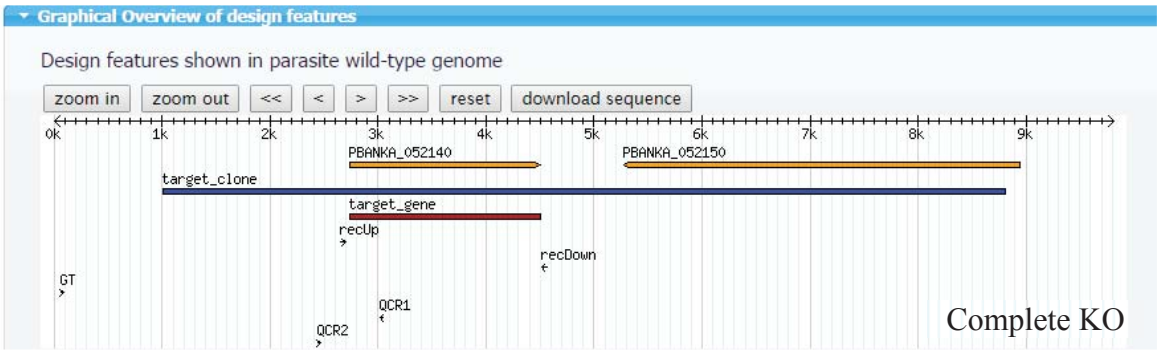


Fig. 4.7| RIO kinases KO vector designs.
Vector designs generated by the *Plasmo*GEM database for *rio1* and *rio2* genes (A and B, respectively). The standard *Plasmo*GEM KO designs aim to target the entire ORF. A few exceptions were implemented to include cases when: (1) There is not a library clone that covers the entire ORF; (2) One of the homology arms is less than 1 kb in length, in which case the design is adjusted to prioritise the homology arm; (3) There is not a minimum distance of 1 kb from the neighbour gene in which case the design is adjusted to avoid interfering with the other ORF's expression. This distance could be reduced to 0.8 kb if the downstream gene was in the reverse orientation, i.e. 1 kb from the next 5'UTR and 0.8 kb from the next 3'UTR. "recUp" and "recDown" represent the 50 bp homology regions between which the drug resistance cassette was inserted. (A) The design for *rio1* KO was included in the third category of exceptions and, as a result, the last ~500 bp were not included in the deletion region. (B) Design for *rio2* KO vector, showing the entire ORF was targeted for removal.
GT, QCR1 and QCR2 are quality control primers, automatically designed by the software.

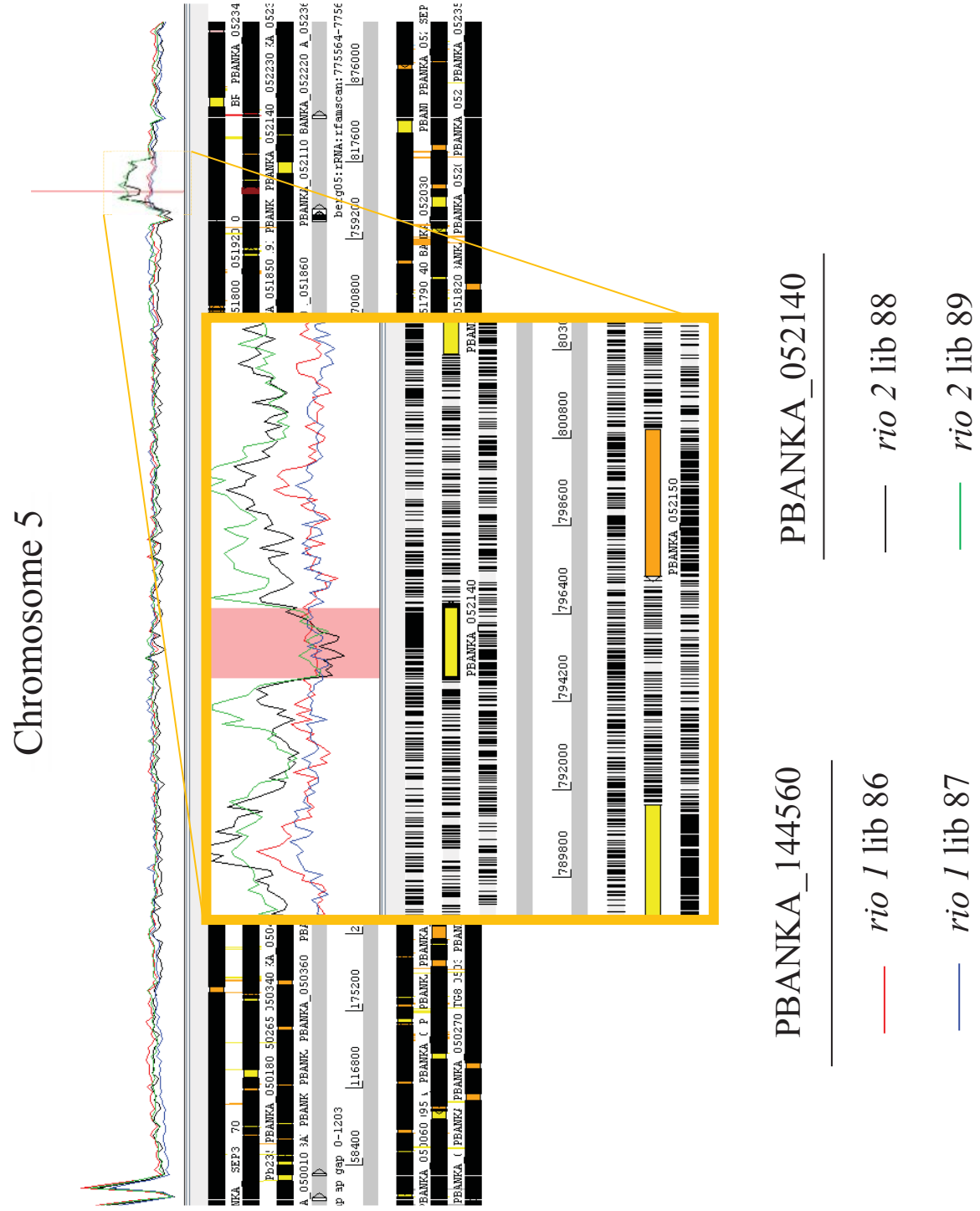


Fig. 4.8| WGS of RIO kinases - *rio2* locus.
Mapped reads for all mutant lines were visualised with the Artemis software. The Figure shows the coverage plots for the entire chromosome 5. Each line corresponds to each of the sequenced mutants (individual libraries – 86 to 89). Coverage across the chromosome was very even except for a ~29.7 kb region spanning the *rio2* gene, where a duplication was detected. This only affected the *rio2* libraries (lines green and black). This duplication did not include the *rio2* gene as the coverage for the ORF was down to the reference levels which meant that the targeting vector deleted one copy but not the other. The pink highlighted region in the zoomed rectangle spans the *rio2* ORF (PBANKA_052140).
The numbers that are not preceded by "PBANKA_" are chromosome coordinates.

4.2.3.1 *gsk3* KO phenotyping

As the *gsk3* gene had been considered essential for erythrocytic development by others [105], careful genotyping by PFGE (Fig. 4.5A), PCR (Fig. 4.5C) and Southern blot (Fig. 4.9A) confirmed that the *gsk3* gene had been deleted.

Barcode counting determined the average relative fitness of this mutant to be 1.01 for days 5-8 post-transfection (Fig. 4.9B). Interestingly, when comparing the growth pattern of a *gsk3* KO clone to the WT parasites for a longer period of time, a striking phenotype was revealed (Fig. 4.9C). On days 1-5, roughly the equivalent window of days 5-8 post-transfection in the screen in terms of overall parasitaemia, both curves seem to evolve evenly. However, beyond this time-point the growth pattern of the *gsk3* KO clone completely diverged from the WT. While the WT population increased over time until day 7 post-infection, the opposite was seen for the mutants. Only from day 8 onwards did their population increase.

Observation of Giemsa stained thin blood films indicated that the mutant parasites could only persist in reticulocytes. Figure 4.9D shows a snapshot of those smears, where blue arrows point at infected normocytes and black arrow at reticulocytes. Interestingly, the increase in parasitaemia registered from day 8 post-infection was accompanied by an increase in reticulocytaemia (~7 % on day 9 post-infection), presumably derived from the anaemia induced by the long-term infection.

The mutant parasites persisted in the mice for 16 days. At this point, severe hepatomegaly and splenomegaly, and severe anaemia became critical. Liver and spleen enlargement are some of the hallmarks of malaria pathology, but in these mice this was exacerbated, perhaps due to the unusual length of the infection (Fig 4.9E).

No significant differences were found between the *gsk* KO mutants and the WT for the number of exflagellation centres per 1000 RBCs (Fig. 4.9F), ookinete conversion rate (Fig. 4.9G), number of oocysts per midgut (Fig. 4.9H). However, the number of sporozoites per mosquito (Fig. 4.9I) was much reduced (p-value= 0.00022). Sufficient numbers of sporozoites reached the salivary glands and no delay in patency (day 4, for WT and *gsk3*, data not shown) was observed. These experiments were performed using a single clone and will therefore need to be reproduced with a second independent mutant.

Localisation experiments were performed by immunofluorescence of *gsk3*-3xHA tagged parasites but only a very diffuse and inconclusive signal was detected in the cytosol of schizonts, trophozoites, gametocytes and ookinetes (data not shown).

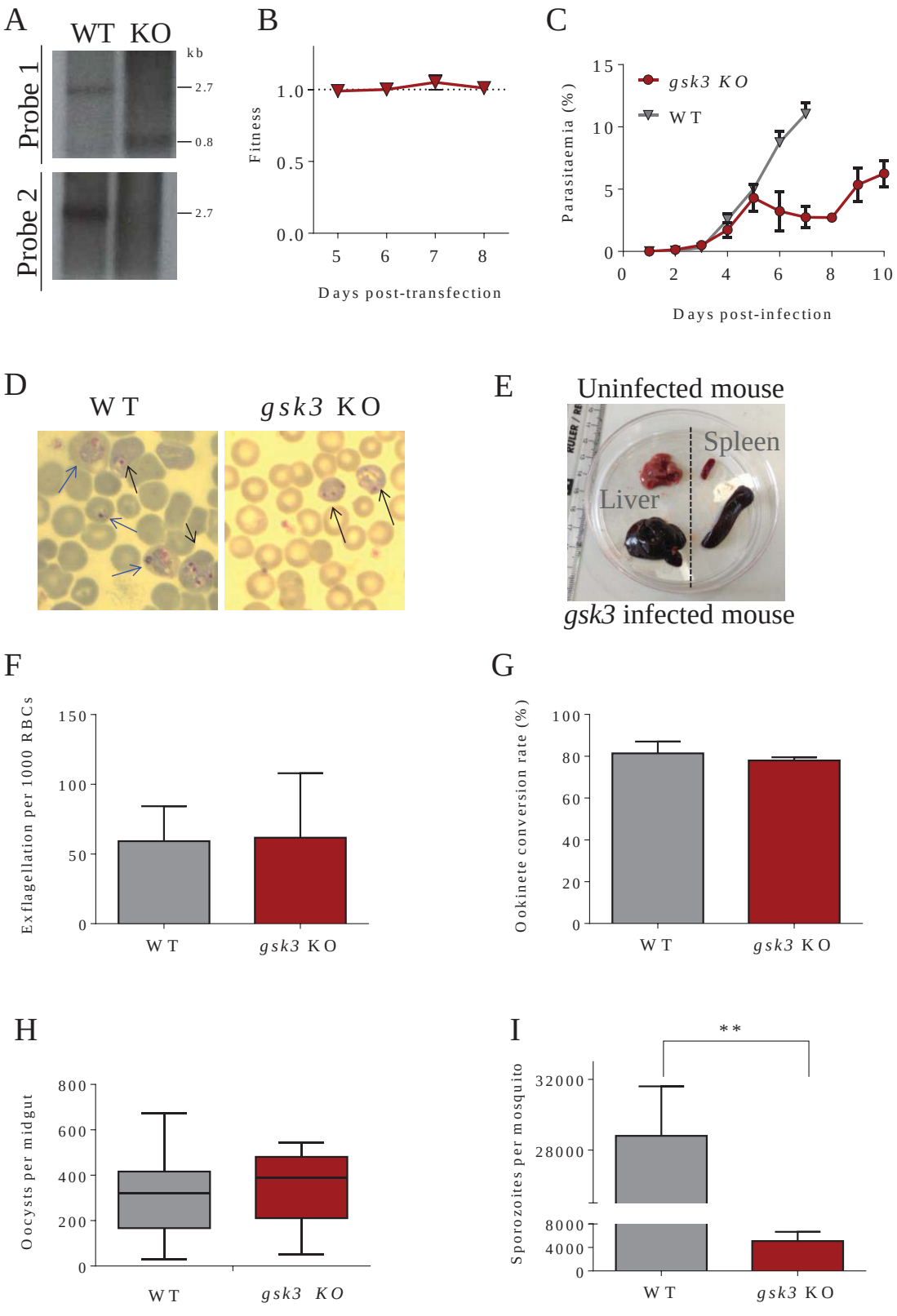


Fig. 4.9| *gsk3* KO genotyping and phenotypic analysis.
(A) Southern blot analysis of the *gsk3* locus to confirm gene disruption. Two probes were used: probe 1 to detect a difference in size between the WT (2.7 kb) and modified (0.8 kb) loci and probe 2 to hybridise only with WT-

specific sequences. A schematic representation of the locus with restriction sites used can be found in appendix VII. (B) Fitness over time for *gsk3* KO mutants obtained from an STM experiment showed that for early/low infections there was no fitness disadvantage. (C) Growth curves for the independently generated mutant, calculated from Giemsa stained thin blood smears prepared daily. Each mouse was injected with 10^6 infected RBCs. Error bars show standard deviations of the mean (n=3). During the first five days of this experiment where the parasitaemia was equivalent to the one in the STM experiments shown in B, the mutant and the WT growth curves were very similar. The attenuated phenotype was only detected beyond this time window. (D) Giemsa stained smears showing infected normocytes (blue arrows – the three left most arrows) and infected reticulocytes (black arrows). Reticulocytes are slightly larger than normocytes and stain purple due to a high content of nucleic acids. (E) Illustration of necrosis and enlargement of the liver and spleen of a *gsk3* KO infected mouse on day 16 post-infection. The other two replicates had a similar size and colouration. Imaging was performed *post-mortem*. (F-I) Phenotypic analysis of *gsk3* KO sexual stages: exflagellation centres per 1000 RBCs (F); ookinete conversion rate (n=3 cultures) (G); number of oocysts per mosquito midgut, dissected on day 12 post-feed (n=10 mosquitoes) (H); and number of sporozoites per mosquito, dissected on day 21 post-feed (n=10 mosquitoes) (I). * Different from the WT as determined by a two sided T-test (p-value= 0.00022).

4.3 Discussion

In this chapter I have shown that the *P. berghei* adapted STM strategy proposed in the previous chapter was fully validated. Parallel transfection of barcoded vectors proved to be a powerful strategy to screen large numbers of genes in *P. berghei*. Mutant barcoding enabled rapid and reliable measurement of the relative abundance of mutants in pools across three orders of magnitude and how it changed during infection. Fitness is a phenotype measurement central to many large-scale genetic studies and was originally measured in terms of population allele frequencies [101]. In this study, it was successfully calculated from the growth rates of each mutant line, as previously described [101], and used to assess the cost of each gene deletion.

The screen presented here, permitted the identification of nine kinases that are dispensable for blood stage development and had not been targeted before. This was likely due to a combination of factors that include (1) a higher integration efficiency of the *Plasmo*GEM vectors, (2) better transfection operating conditions and (3) higher sensitivity of the detection method. These discrepancies reinforce the idea that the absence of evidence for genomic integration is insufficient proof for the essentiality of a gene. Additional experiments such as gene tagging can supply supplementary information about the locus accessibility.

Five of the newly targeted genes were validated with generation of the independent mutants which included *cdpk1* and *gsk3* kinases, current targets of drug development research [143,169]. Data from these validation experiments confirmed the ability of barcode sequencing to identify new mutants and failed to detect evidence that STM screens will generate large numbers of false positives, as might be expected if *Plasmo*GEM vectors replicated as episomes or integrated by non-homologous recombination.

Eukaryotic ribosomes or 80 S ribosomes are so called due to their sedimentation coefficient (S). They are comprised of two unequal subunits, a small subunit (40 S) and a large subunit (60 S). Each of them contains multiple ribosomal proteins embedded on a scaffold of ribosomal RNA (rRNA). The small subunit is further assembled from 18 S rRNA and 32 ribosomal proteins [174]. The RIO (right open reading frame) family of atypical protein kinases plays an important role in ribosome biogenesis and is subdivided into three subfamilies Rio1, Rio2, and Rio3. Every organism from archaea to humans has both *rio1* and *rio2* kinases, while *rio3* genes are only found in multicellular eukaryotes [124,171]. In *S. cerevisiae* there are two *rio* kinases, *rio1* and *rio2*. Both were identified as necessary for processing of 20 S pre-rRNA into mature 18 S rRNA. Depletion of either of *rio* kinases leads to accumulation of 20 S pre-rRNA and cell cycle arrest [172]. In addition, *rio1* plays an important role in cell cycle, during G1 to S transition and is involved in the control of the onset of anaphase. On the other hand, *rio2* has been proposed to promote nuclear export of pre-40 S subunits [171,175]. The fact that deletion of either *rio1* or *rio2* is lethal for the cells, suggests that they perform distinct functions [103].

In malaria parasites these kinases have previously been considered essential for blood stage development [105,125]. However, the current study suggested that the *rio1* gene can be targeted. Despite the failure in obtaining a pure population of deletion mutants, WGS data clearly showed that the vast majority of the parasites had that gene deleted.

The *rio2* gene, on the other hand, is likely to be essential in *P. berghei* parasites as it is in other organisms. Interestingly, a random duplication of a segment of gDNA that included the *rio2* gene enabled the integration of the targeting vector into one copy of the duplicated loci. This allowed parasites to become resistant to pyrimethamine while still carrying a functional copy of the *rio2* gene. Random amplification of large regions of DNA characterises a remarkable phenomenon that has recently been shown to occur randomly amongst individual parasites. This is part of a two-step strategy, identified to be responsible for drug resistance development [176]. Step one involves random amplifications that include genes that confer resistance to a given antimalarial and therefore allow the parasite to survive. Then, to ensure tolerance of increasing drug pressure, additional copies of this amplification are generated. While the original copy of a given gene ensures the native role is still carried out, each of the other copies that generate a pseudo-ploidy state might carry mutations that enhance drug resistance or increase parasite fitness in a given context. Loss of these

amplification regions takes place when pressure is alleviated or when its fitness cost is no longer advantageous [176].

This is a very rare event that drives development of drug resistance [176] and, in this case, was the cause for the only confirmed false positive in the STM screen.

GSK3 was first described as a protein kinase that phosphorylates glycogen synthase [177]. Since then it has been implicated in a number of processes like metabolic control, embryonic development, cell proliferation and adhesion, Alzheimer's disease, circadian rhythm, and oncogenesis [178]. Due to its implication in several human diseases this multifunctional enzyme has been targeted for drug development research. Likewise, its likely essential role in *Plasmodium* development [105,145] has prompted the search for specific inhibitors [169]. In fact, treatment of *P. berghei* infections with LiCl led to a reduction in parasitaemia relatively comparable to treatment with chloroquine [179], a drug once widely used to treat *P. falciparum* infections but currently only recommended for the treatment of malaria caused by *P. vivax*, *P. malariae* and *P. ovale*.

Co-localisation experiments in *P. falciparum* parasites have shown the presence of GSK-3 in Maurer's clefts, within the erythrocyte cytoplasm, as vesicle-like structures [169]. Maurer's clefts are membranous structures that are present in *P. falciparum* infected RBCs [180]. These have a crucial role in virulence as they mediate protein sorting and export and have important implications on (1) the adherence properties of the infected RBC and consequently on splenic clearance, and (2) on the RBC permeability as a means to allow the acquisition of nutrients [181]. This observation regarding GSK3 localisation prompted the hypothesis of it being involved in protein trafficking. Localisation experiments with an epitope tagged allele of (Pb*gsk3*-3xHA) failed to reproduce such pattern (data not shown). At the time of these experiments this was not surprising since presence of Maurer's clefts-like structures had not been reported in *P. berghei*. However, recently biochemical and microscopic data revealed the presence of intra-erythrocytic membranous structures termed *P. berghei*-induced structures (IBIS). Although morphologically different from the Maurer's clefts it indicates that the rodent parasite also creates an intracellular network in infected RBCs [182,183]. In light of these recent developments, follow-up experiments on the localisation of PbGSK-3 will include a replacement of the HA epitope tag with a fluorophore such as mCherry to allow live imaging of this protein throughout the lifecycle.

Barcode counting experiments revealed that unlike what was previously suggested [105], *gsk3* can be disrupted in *P. berghei*. Generation of a thoroughly genotyped, cloned KO

mutant validated such results. Interestingly, a brief phenotypic analysis suggested that these mutants have a strong preference for reticulocytes.

Reticulocytes encompass a minor though heterogeneous population of RBCs precursors that can be grouped according to their levels of the transferrin receptor (CD71) expression. They mature into normocytes (normal RBCs) over a period of 72 hours during which these cells undergo profound changes such as loss of surface receptors and reticular matter. Some *Plasmodium* species have a preference for these young RBCs over normocytes. For instance, in the presence of equal numbers of erythrocytes and reticulocytes, *P. berghei* is ~150 times more likely to infect a reticulocyte [184]. Among the species infecting humans, *P. falciparum* is able to invade RBCs of all ages while *P. vivax* is restricted to reticulocytes. This tropism towards different sets of RBC fractions has a major impact on the course of infection and pathology which is why *P. falciparum* infections are life-threatening.

Preliminary phenotypic experiments suggested that the disruption of the *gsk3* gene exacerbated the reticulocyte preference of *P. berghei* parasites. Supporting this finding was the observation on Giemsa stained thin blood smears of (1) reticulocytes as the only infected RBCs and (2) a drop in parasitaemia between days 5 and 8 post-infection, i.e. moment when reticulocytes in circulation become limiting. It was unclear whether the *gsk3* mutants could not fully develop in normocytes or not invade them at all. The drop in parasitaemia, on the other hand, was attributed to the suppression of the erythropoietic response. The latter is a phenomenon that has been reported to occur during malaria infections, presumably as a host protective mechanism [184,185]. This response has not been seen during non-malaria anaemias and is only alleviated once haemoglobin levels drop below critical levels, which correlates with the increase in parasitaemia from day 8 post-infection onwards. As a consequence of this reticulocyte preference, *gsk3* mutants could persist for at least twice as long in the rodent bloodstream when compared to the WT *P. berghei* strains. Perhaps as a consequence of this prolonged infection, the liver and spleen of these mice showed extensive morphological signs of failure. A histological study of these organs would be required to best characterise this finding.

Furthermore, experiments that would help clarifying the nature of this phenotype would include complementation of the modified locus and validation of the reticulocyte preference phenotype through invasion assays with normal and reticulocyte enriched blood. The drop in parasitaemia from day 5 post-infection onwards is also consistent with adaptive immune clearance. It would therefore be interesting to access parasite growth in RAG mutant mice, where the disruption of the *rag* gene, involved in V(D)J recombination, prevents B and T cell

differentiation [186]. On the other hand, immunological profiling of cytokines during early and late stages of infection could also be quite informative.

In mammalian cells, GSK3 plays a central role in the control of proliferation through the repression of transcription factors such as CREB, NFκB and AP-1 [187]. Thus, transcriptome analysis of *gsk3* KO parasites could potentially reveal a similar function for this kinase in *Plasmodium* parasites.