Chapter 3

Establishment of Signature tagged mutagenesis in *P. berghei*

Setting the scene

3.1 Introduction

The first aim of this project was to adapt the STM strategy to *P. berghei* parasites in order to enable large scale genetic screening.

STM screens have been designed in many ways, reflecting the diversity of genetic systems of different taxa. Common to all is a workflow that starts with (1) mutagenesis, i.e. generation of barcoded mutants, which is followed by (2) propagation of these mutants in pools and finally (3) identification of the mutants present after propagation through their barcode.

One main approach used to generate barcoded mutants in bacteria is *in vivo* transposition (Fig. 1.6). Fonager and colleagues have applied a *piggyBac* transposition system to *P. berghei* parasites, but fine tuning is yet to be achieved [46]. In yeast, directed gene-replacement has been the most used method to generate libraries of thousands of mutants that are then pooled and used in STM approaches [94]. Modifications that rely on homologous recombination are probably the most reliable method for genetic modification of *Plasmodium* parasites, although at different frequencies according to the species. However, the approach taken by the yeast field would be of very little use for *P. berghei* parasites as at least 12 mice are needed to generate a single clonal line.

Recently, a new generation of *P. berghei* targeting vectors was developed – the *Plasmo*GEM vectors [73]. These are linear vectors in which the length of homology arms is increased from 0.3 – 1.0 kb to several kb, to improve homologous integration frequencies. Additionally, they have not been reported to persist as episomes after transfection, which decreases the rate of false positives, as drug selection ensures elimination of the parasites where integration did not take place. A combination of improved integration with reduction of false positives made these vectors promising tools for a *P. berghei* adapted STM. To allow identification of mutants generated by these vectors within a pool, gene-specific barcodes were introduced into their basic design that labels mutants upon genome integration. These barcodes consisted of a 10-11 mer DNA sequence that was inserted into a ~100 bp-long module, located next to the B2 gateway site in all vectors. The length of the barcode permitted

that enough sequences with a hamming distance of four[1] (i.e. single error correction plus double error detection) [158], could be generated to cover the entire *Plasmodium* genome.

The barcode module was flanked by constant annealing sites, which enabled a bias-free amplification of all barcodes from a pool through a single PCR reaction.

Taken together these tools offered the opportunity to perform STM-like experiments in *P. berghei*. We hypothesised that the properties of the *Plasmo*GEM vectors would enable the generation of complex pools of mutants, thus circumventing the need to generate each mutant independently prior to parallel phenotyping. In other words, transfection of pools of barcoded *Plasmo*GEM vectors (mutagenesis step) would generate pools of barcoded mutants that after being expanded in a single mouse (propagation step) could be identified through their barcodes (identification step).

To test this hypothesis, various parameters needed to be optimised. These included transfection conditions, barcode detection, sequencing library preparation and sequencing run conditions.

3.2 Results

3.2.1 Optimisation of operating conditions for transfection

The ability to develop STM based screening in *Plasmodium* depends on the complexity of mutant pools that can be easily generated by co-transfecting multiple vectors. This in turn depends on the transfection efficiency that can be achieved. My first objective was therefore to identify the most suitable electroporator and the most adequate DNA concentration of each *Plasmo*GEM vector to use.

3.2.1.1 Choice of electroporator

One aspect that is critical for transfection efficiency is the type of the electroporation system used. The traditional Bio-Rad instruments have been surpassed by the Lonza electroporators, which are now the most efficient devices used to generate *P. berghei* transgenics.

[1] A Hamming distance of four enables single error correction plus double error detection, i.e. it takes four mutations, or sequencing/synthesis errors for one barcode to become another; one mutation can be corrected and two can be detected.

Two different Lonza electroporator systems were tested for their efficiency – Nucleofector II and 4D Nucleofector X unit (Fig. 3.1A). Using the same pool of schizonts cultured from two different mice, four different transfections were performed with each electroporator.

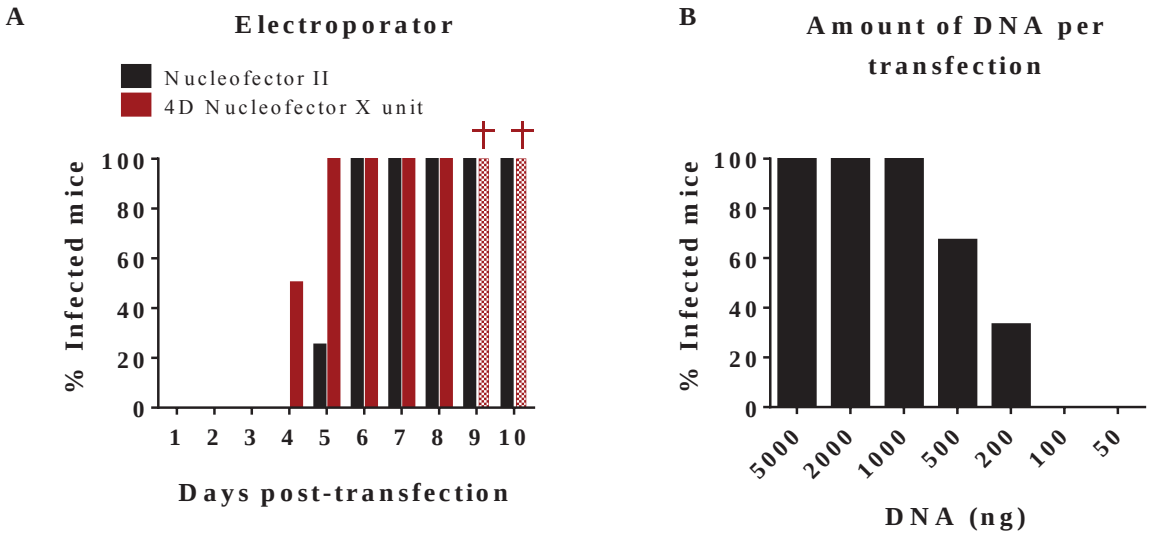


Fig. 3.1| The choice of electroporation system and DNA concentration are critical for maximum transfection efficiency.
(A) Impact of the choice of electroporator on patency. Four different transfections were performed using either the Nucleofector II or the 4D Nucleofector X unit electroporator to transfect 5 µg of a KO vector targeting *map1* gene. The graph shows the duration after transfection required for mice to develop parasitaemias visible on a Giemsa stained smear, which was used to infer transfection efficiency. † culled mice due to high infection. (B) Assessment of how little DNA is required for a successful transfection using the 4D Nucleofector X unit system (n=3 per concentration). All mice were injected intra-venously (i.v.) in the tail vein.

Transgenic parasites were obtained for all replicates. However, the 4D Nucleofector system proved to be more efficient since parasites were visible in the blood of two out four mice by day 4 post-transfection, whereas with the Nucleofector II system only one mouse had visible parasitaemia by day 5 post-transfection. All eight mice were diagnosed as infected on day 6.

Transfection efficiency was determined according to Janse *et al* [159] by comparing the number of surviving parasites in mice before and after pyrimethamine selection, based on the daily 10x multiplication rate of *P. berghei* in mice [160]. Mathematically this is defined as: $(n2/n1) \times 1/10^{d}$. To this end, the parasitaemia after injection of the transfected parasites (*n1*) was determined by counting Giemsa stained thin blood films at ~24h post-transfection, just before the start of the selection with pyrimethamine. Later, usually four to seven days (*d*) after

transfection, the number of drug-resistant parasites (*n2*) was determined from the parasitaemia counted on day *d*, when infection was patent.

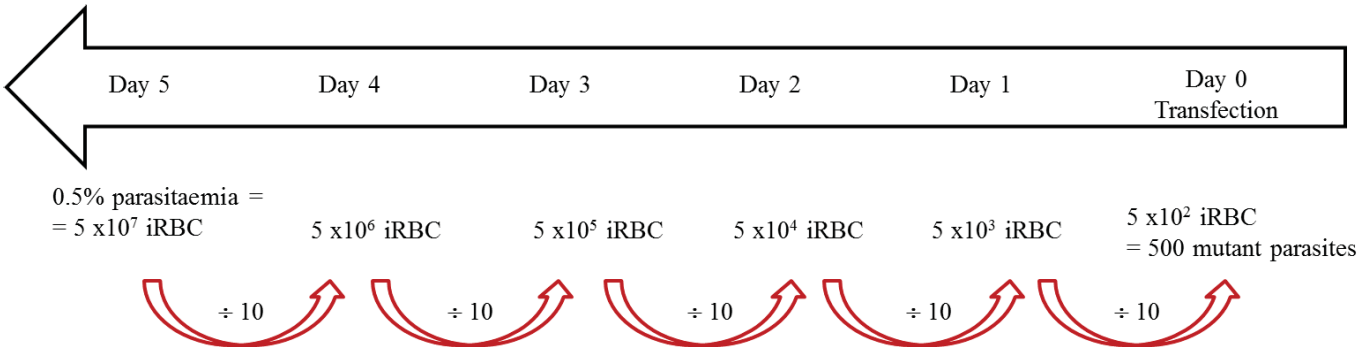
These results translated into a transfection efficiency of ~ $2x10^{-5}$ and ~$5x10^{-4}$ for the Nucleofector II and 4D, respectively.

3.2.1.2 Optimal DNA concentration

As transfection of pools of vectors was envisaged I explored what the minimal DNA concentration of each vector was to generate transgenic parasites, and whether very high concentrations would be more efficient.

For this, a dilution series of a KO vector for the *map1* gene (PbGEM-036210) was prepared to generate the following range: 5000 ng, 2000 ng, 1000 ng, 500 ng, 200 ng, 100 ng and 50 ng. Each of them was transfected in triplicates using the 4D Nucleofector X unit and the percentage of mice that became infected by day 10 post-transfection was registered. As expected, 2000 ng (the standard concentration for a *P. berghei* transfection) generated transgenic parasites for all replicates as did 5000 ng and 1000 ng, whereas concentrations below 200 ng were not successful. The transfection of 200 ng and 500 ng generated one and two infections, respectively (Fig. 3.1B). In addition, there was no difference in patency day for the highest concentrations, hence suggesting that above 2000 ng the system is probably close to saturation.

On day 5 post-transfection of the experiments where 1 µg of DNA was transfected, the observed parasitaemia was around 0.5 %. Based on the daily 10x multiplication rate of *P. berghei* parasites [160] during the exponential phase of growth this translates into approximately 500 independent integration events as the number of circulating RBCs in a mouse[2] is on average $1x10^{10}$, as detailed below:



[2] This RBC concentration applies to an average 6-8week-old, 30 g mouse.

Since transfecting 1 µg of DNA generates around 500 independent integration events while 0.1 µg produced none, I concluded that efficient transfection might be a threshold phenomenon, as is electroporation [161,162], and not linearly related to the amount of vector DNA used. I therefore predicted that even smaller quantities of vector DNA should reliably give rise to mutants when delivered as part of a vector pool as long as the total amount of DNA was at least 1 µg. To this end, I transfected two different pools (n=3) of equimolar amounts of 10 vectors with a total DNA concentration of 1000 ng (100 ng/vector) and 2000 ng (200 ng/vector). These pools included the KO vector used in previous experiments, as a positive control (*map1* KO). The choice of number of vectors had the objective of assessing the limit suggested by the previous experiment (i.e. no transgenic parasites < 200 ng/vector). All transfections were positive by day 6 post-transfection and integration for the control vector was detectable in all experiments by PCR.

The detection of *map1* KO vector in both conditions strongly indicated that the overall DNA concentration is a more important variable than the concentration of each vector for transfections of pools in the context of STM experiments. However, the same phenomenon that enables the transfection of small amounts of each vector and therefore simplifies the preparation of each pool prior to transfection (i.e. one miniprep yields enough DNA for several experiments) also raised the concern of integration of multiple vectors in the same genome. Mathematically, assuming a transfection efficiency of $5x10^{-4}$, as determined previously for the experiments shown in Figure 3.1A, and assuming that the integration of two different vectors are independent events, then the likelihood of a double integration event should be the product of the individual likelihoods, i.e. close to $2.5x10^{-7}$.

However, in reality integration events are almost certainly not independent since the efficiency with which DNA is delivered will vary between individual schizonts. I addressed this question experimentally by asking whether vectors lacking selection markers (i.e. intermediate vectors) could become "passengers" of vectors with selection marker.

KO vectors for known targetable genes were transfected with a 20-fold excess of intermediate vectors for different other targetable genes (Fig. 3.2A) that lacked a selection marker for *P. berghei* and could therefore only replicate when integrated into genomes that carried a second insertion of a final KO vector. PCR analysis of resistant parasites on day 9 post-transfection failed to detect such events (Fig. 3.2B), suggesting that double integration events are rare and that the large majority of the transgenic parasites are single mutants.

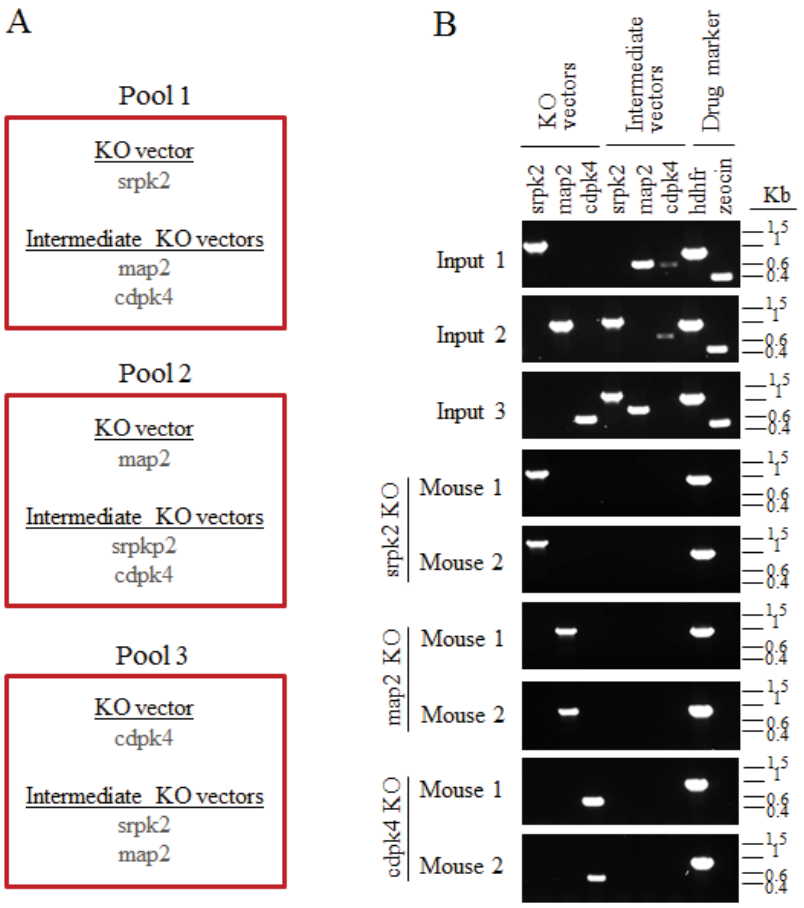


Fig. 3.2| Absence of passenger vectors lacking a selection cassette.
(A) Three different pools consisting of one final KO vector in the presence of a 20-fold excess of intermediate vectors (10 µg total DNA per transfection), which have the same homology arms but a zeocin resistance cassette that cannot be selected in *P. berghei* parasites were transfected in duplicates. (B) PCR genotyping performed on gDNA samples, from day 9 post-transfection from each of the six infected mice, failed to detect the presence of intermediate vectors that could only be selected if integrating into a genome where a KO vector was present.

3.2.2 Optimisation of barcode detection using Illumina sequencing

Before any attempt to perform STM experiments with *P. berghei* parasites, I investigated whether a "bar-seq" strategy (section 1.7.1) using Illumina sequencing would be feasible to read the *Plasmo*GEM barcodes. To this end, several pools of barcoded vectors were prepared and sequenced.

Pools 1 and 2 (Fig. 3.3A) were devised to look at the sequencing accuracy across two orders of magnitude of different ratios of each of 19 vectors that varied in abundance, while trying to reproduce realistic concentrations of vectors in real experiments. In some cases, as little as 5 ng (0.05 ng/µL) of vector was added (vectors 4, 15 and 17) to verify the sensitivity of the PCR reaction for low abundant barcodes in an attempt to mimic a pool of parasites containing a less fit (and therefore less abundant) population of mutants. Conversely, some

vectors (8 and 14) were more abundant (5.00 ng/μL) than the rest so that the impact of having a dominant barcode in the pool could be assessed. Some vectors were used as negative controls and were therefore not added to the pool but their barcode was searched for in the sequencing data (vectors 1 and 11) and in other cases a vector would be present in one of the pools but not the other (vectors 5 and 17).

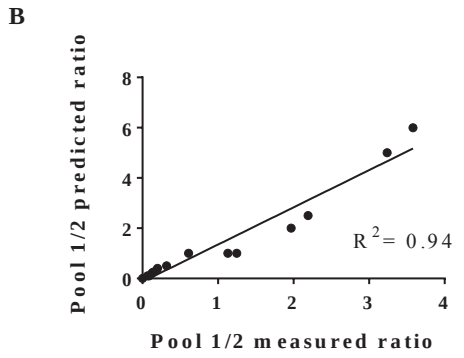
A correlation analysis between the predicted and the measured ratios yielded a high correlation coefficient (R^2) of 0.94 and is depicted in Fig. 3.3B. For instance, for vectors 2, 4, and 6 the predicted ratio was 2.00, 0.10, and 1.00 and the corresponding measured ratios were 1.97, 0.09, and 1.13, respectively. The negative controls were not detectable in either of the pools (1 and 11) and the same was seen for vectors 5 and 17 in pools 2 and 1, respectively.

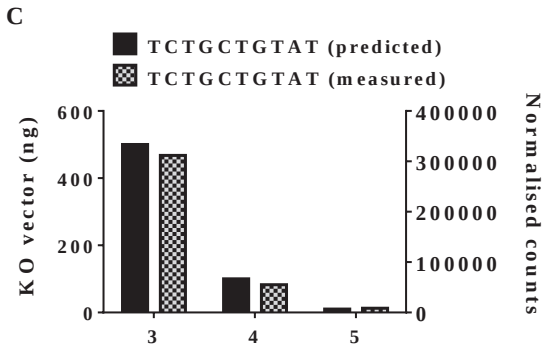
In parallel, three other less complex pools were prepared where the concentration of each of three vectors was adjusted to mimic three different hypothetical outcomes of real experiments: decline (Fig. 3.3C), steady maintenance (Fig. 3.3D) and increase (Fig. 3.3E) in abundance over time. The obtained patterns greatly resembled parasite growth and again very similar numbers between measured read counts and their corresponding prediction were obtained, clearly showing that less abundant vectors yield fewer read counts and vice-versa.

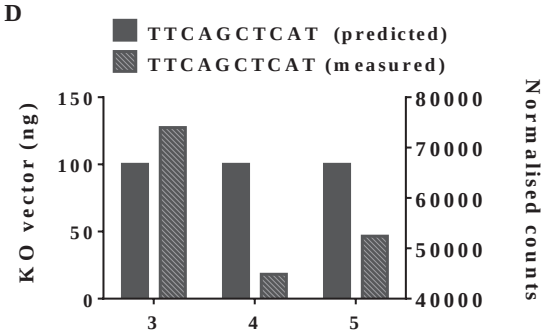
Together these data showed that barcode counting provides quantitative measurements that are sufficiently accurate to measure differences in barcode abundance within pools.

A

Barcode	pool 1 (ng)	pool 2 (ng)	Ratio pool 1/2
1)TGACCACTAT	0	0	-
2)TCGAAGTGAT	200	100	2.00
3)TGTACCTTAT	50	200	0.25
4)ACAGTGGTAT	5	50	0.10
5)TCTCGGTTAT	50	0	-
6)TGGCTCAGAT	300	300	1.00
7)TAGCTTGTAT	50	100	0.50
8)TGATACGTAT	100	500	0.20
9)TCCAGTCGAT	300	50	6.00
10)TCTGCTGTAT	100	300	0.33
11)TTAGGCATAT	0	0	-
12)TCCTCAATAT	100	100	1.00
13)TCGAGCGTAT	100	100	1.00
14)TAAGTTCGAT	200	500	0.40
15)GCCAATGTAT	5	50	0.10
16)TTCAGCTCAT	500	10	50.00
17)CGATGTTTAT	0	5	0.00
18)TTCTGTGTAT	50	10	5.00
19)TCAGATTCAT	500	200	2.50







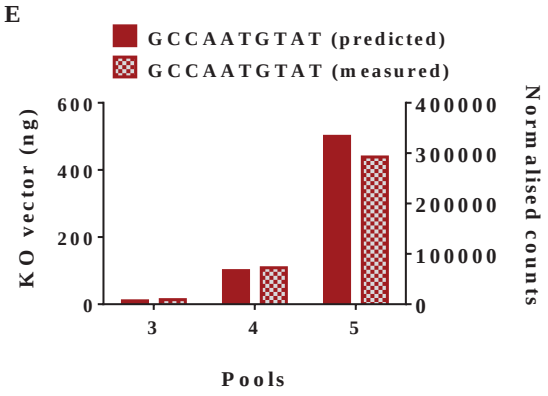


Fig. 3.3| The barcodes within the *Plasmo*GEM vectors are compatible with a bar-seq strategy.
(A) Known concentrations of the same set of vectors were used to prepare two different pools. The final volume of each pool was 100 µL. (B) Both pools were sequenced and the ratio pool 1/2 was calculated from the sequencing reads generated. A Pearson correlation analysis was performed to compare the measured with the predicted ratios of vectors. This analysis excluded all vectors for which a ratio could not be calculated. (C-E) Three other smaller pools (3-5) were prepared to look at feasibility of measuring growth patterns (decline, steady maintenance and increase) through time. The normalised counts patterns were very similar to the predicted ones.

3.2.3 Optimisation of Illumina library preparation

The *Plasmo*GEM barcodes are flanked by constant annealing sites. This ensures that barcodes can blindly be amplified in a single PCR reaction and eliminates any multi-template PCR bias [163]. This is particularly relevant for an STM approach as the pool of barcodes is expected to change during infection.

Amplification of barcodes from purified *Plasmo*GEM vectors was routinely done in 25 cycles. However, the same number of cycles was not enough to reach saturation in samples originated from infected blood. In order to reach the minimal concentration required for Illumina library preparation, either ten different reactions needed to be pooled for each sample or the number of cycles needed to be increased to 35. To exclude any data bias that this increase might induce, the impact on the quality of the data generated after either 25 or 35 cycles was compared (Fig. 3.4).

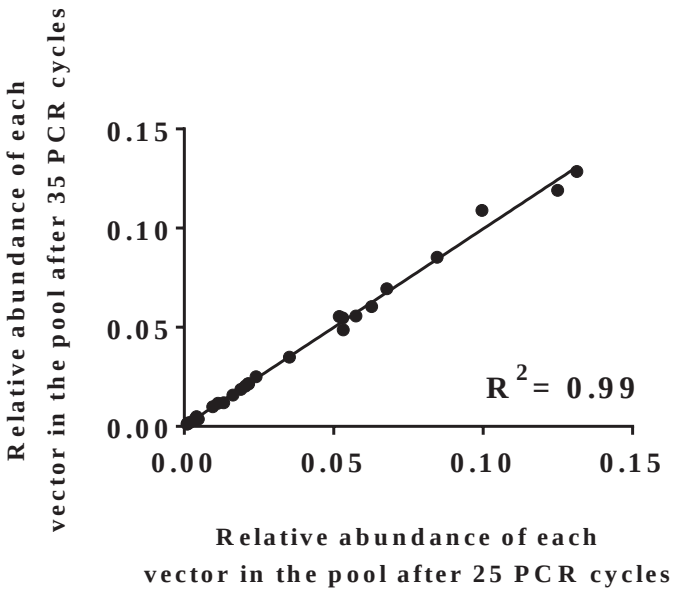


Fig. 3.4| Increasing the number of PCR cycles does not have a high impact on data quality.
The impact on the relative abundance of each vector within the pool after 25 or 35 PCR cycles at the library preparation stage was assessed. A Spearman correlation analysis between the two datasets yielded a high correlation coefficient (R^2= 0.994). Each library had 28 different vectors.

The number of cycles had very little impact on the data as shown by the high correlation coefficient (R^2=0.99) between samples where 28 different vectors were present at different abundances. As a result, 35 cycles were used to amplify all samples in this project since these yielded a more suitable concentration of amplicon.

Two different strategies can be used to prepare sequencing libraries from PCR amplicons: adaptor ligation (AL) or direct amplification (DA).

The AL method is by far the most common library preparation method used to generate next-generation sequencing libraries. Essentially, it requires the ligation of specific adaptor oligos to the pre-processed fragments of the DNA to be sequenced as illustrated in Figure 3.5 (left panel). As the STM amplicons are only 100 bp in length, the original protocol was simplified as shown (Fig. 3.5, right panel). Briefly, 500-1000 ng of the purified barcode PCR product were dA-tailed and then used to ligate Illumina adaptors. When needed, a PCR amplification step priming on the adaptor sequences could be used to boost the library yield.

The AL method, although robust, required high amounts of DNA and was very time-consuming for large numbers of samples. The second method, DA, relied exclusively on two rounds of PCR followed by one purification step (Fig. 3.6A). The first PCR amplified the barcode as previously and introduced priming sites for the second reaction at which stage Illumina adaptors were incorporated. A final clean-up step removed primer/adaptor dimers ensuring that the libraries were ready for loading and sequencing.

In order to choose the best approach, both methods were analysed in terms of preparation time, cost, input material needed and data generated.

On average, preparing a set of 32 samples using AL method took five days, cost £42 per sample and required at least 500 ng of purified PCR amplicon. Conversely, the DA method proved to be substantially faster as only two days were required to process the same set, more affordable (£18/sample) and used less sample (<10 ng of unpurified primary PCR product). The most important parameter, data generated, was not influenced by the library preparation method. High correlation coefficients were obtained for the analysis of six libraries that were prepared by both methods (Fig. 3.6 B, C).

Given that the DA method proved to be a faster and more cost-effective alternative to AL, it was chosen as the preferred method to generate samples throughout this project.

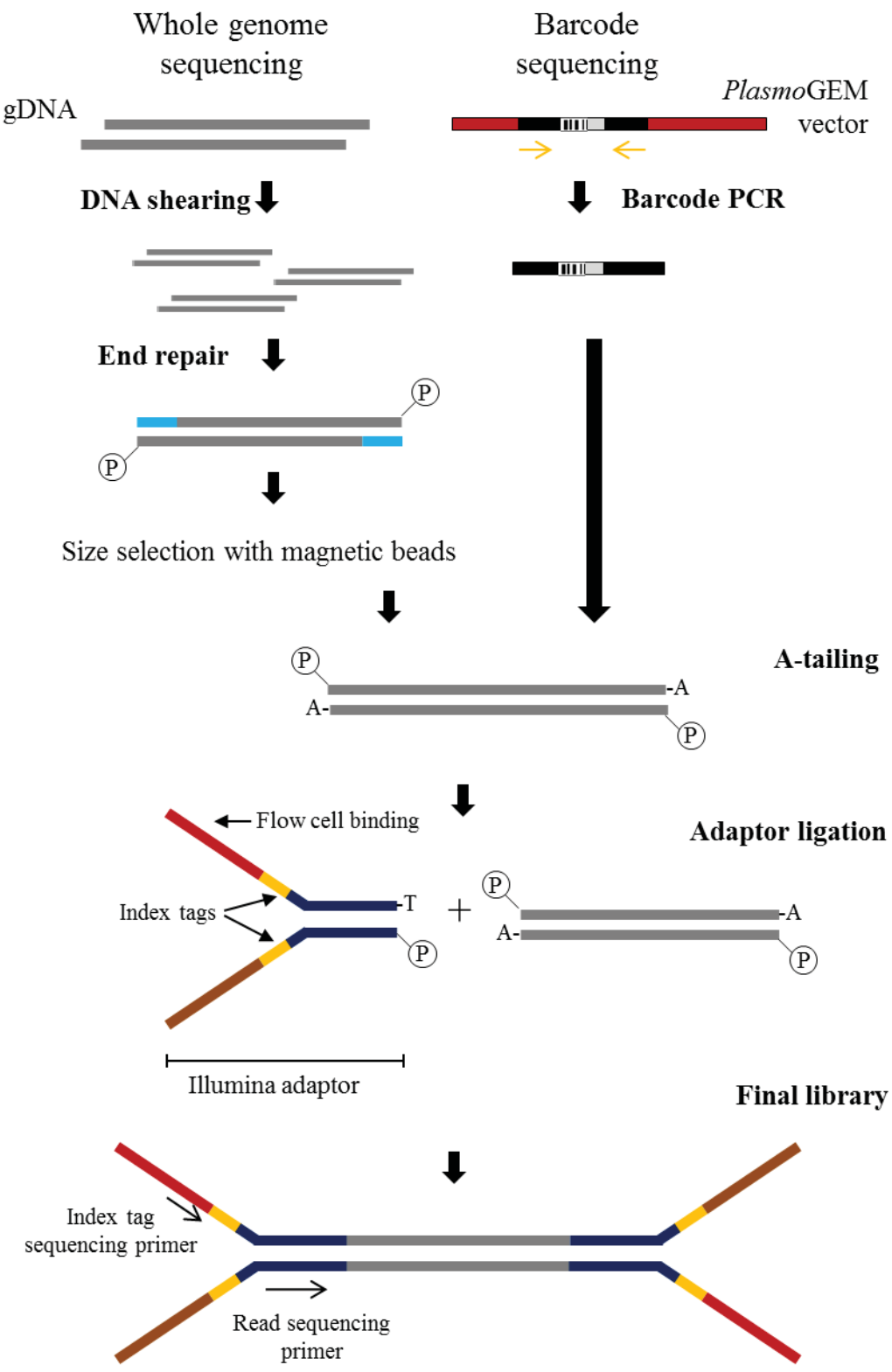
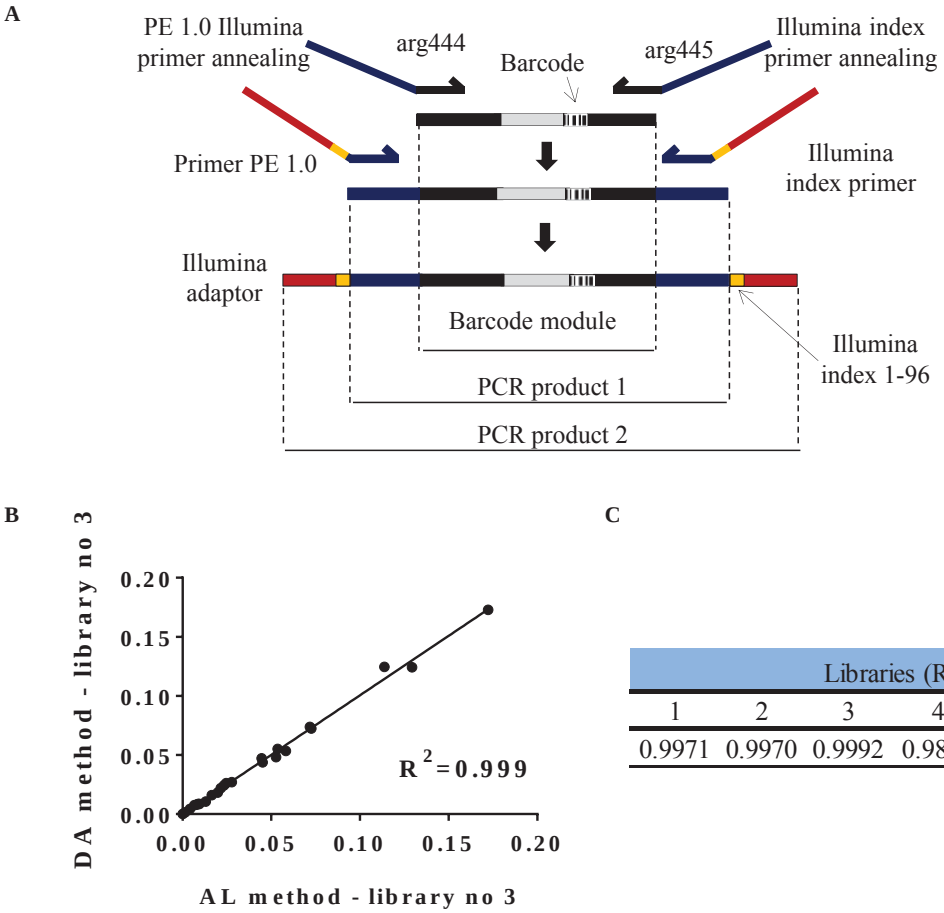


Fig. 3.5| Adaptor ligation method overview.
Typically, an adaptor ligation protocol for WGS samples involves the following steps: DNA shearing, end repair, a-tailing and adaptor ligation. In the case of the STM samples, no shearing or end repair steps are required. Instead, barcodes are amplified by PCR using the flanking annealing sites that are common to all *Plasmo*GEM vectors. Next a-tailing and subsequent adaptor ligation follow. The structure of the final library is depicted at the bottom.



Libraries (R^2)					
1	2	3	4	5	6
0.9971	0.9970	0.9992	0.9898	0.9979	0.9979

Fig. 3.6| Comparison between AL and DA library preparation methods.
(A) DA method overview. A nested-PCR approach amplified the barcodes (PCR1) and generated Illumina compatible libraries (PCR2). The priming sites that enabled the barcode amplification (black regions in primers arg444 and arg445) were unchanged from the AL method but they carry, as overhangs, priming sites for the next round of PCR. The second PCR reaction introduces the Illumina adaptors shown in red and yellow. (B, C) Spearman correlation analysis for six different libraries that were prepared by both AL and DA methods. Very high coefficients were obtained for all of them discarding the possibility of bias introduced by the DA method. (B) Regression analysis plot for library number three.

3.2.4 Optimisation of Illumina MiSeq run conditions

Whole genome sequencing libraries undergo a shearing step that generates random fragments of DNA. This creates a high degree of diversity at the sequencing level as the distribution of bases throughout the flow cell in each sequencing cycle becomes random. STM PCR amplicon libraries are of very low complexity as the order of each base is the same for every cluster except at the barcode region. Also, shearing is not recommended as they already are very short. This makes differentiation of the individual clusters at the imaging level a very difficult task. As very uniform clusters tend to be highly error-prone due to imaging limitations of the platform, it is essential that enough diversity is present to ensure accurate base-calling.

The PhiX spike-in is a base-balanced DNA library derived from the genome of a PhiX 174 bacteriophage. It is commonly used as a spike-in to generate diversity in the flow cell or as a control lane to validate the quality of each run.

The density of clusters on the flow cell is of vital importance to the throughput of the instrument as overload will impede accurate imaging of individual clusters. Normal MiSeq run conditions for high complexity libraries are ~ 8 x10^5 clusters/mm^2 and up to 5 % of PhiX. In order to optimise the run conditions for the STM libraries I tested different conditions of PhiX spike-in concentration and cluster density and determined their impact on the purity filter (%PF). Anticipating problems with cluster density due to the low complexity, the first tests were run at low density (< 4 x10^5 clusters/mm^2). In these conditions, a high correlation (R^2= 0.97) was observed between the spike-in abundance and the %PF of six different runs that were run at similar cluster densities (3 – 3.8 x10^5 clusters/mm^2) (Fig. 3.7 A). The absence of PhiX was highly detrimental to the run quality at both normal and low cluster densities (Fig. 3.7 B and C - Runs 8396 and 10499, respectively); neither of these runs achieved the minimal QC threshold. The addition of 5 % of PhiX increased the %PF of runs 10666 and 10600, 56.7 % and 56.5 %, respectively, at both low and normal density (2.9 and 6.5 x10^5 clusters/mm^2, respectively). However, only when the concentration of PhiX was increased to > 50 % did the %PF reach acceptable levels (at both densities) as shown in Figures 3.7 B and C, run IDs 10343 and 10537.

From these results I established that the ideal run conditions for STM libraries to achieve a %PF of at least 85 % were: a spike-in concentration of 40-50 % (as given by the linear regression shown in Fig. 3.7A) and cluster density of 4-6 x10^5 clusters/mm^2.

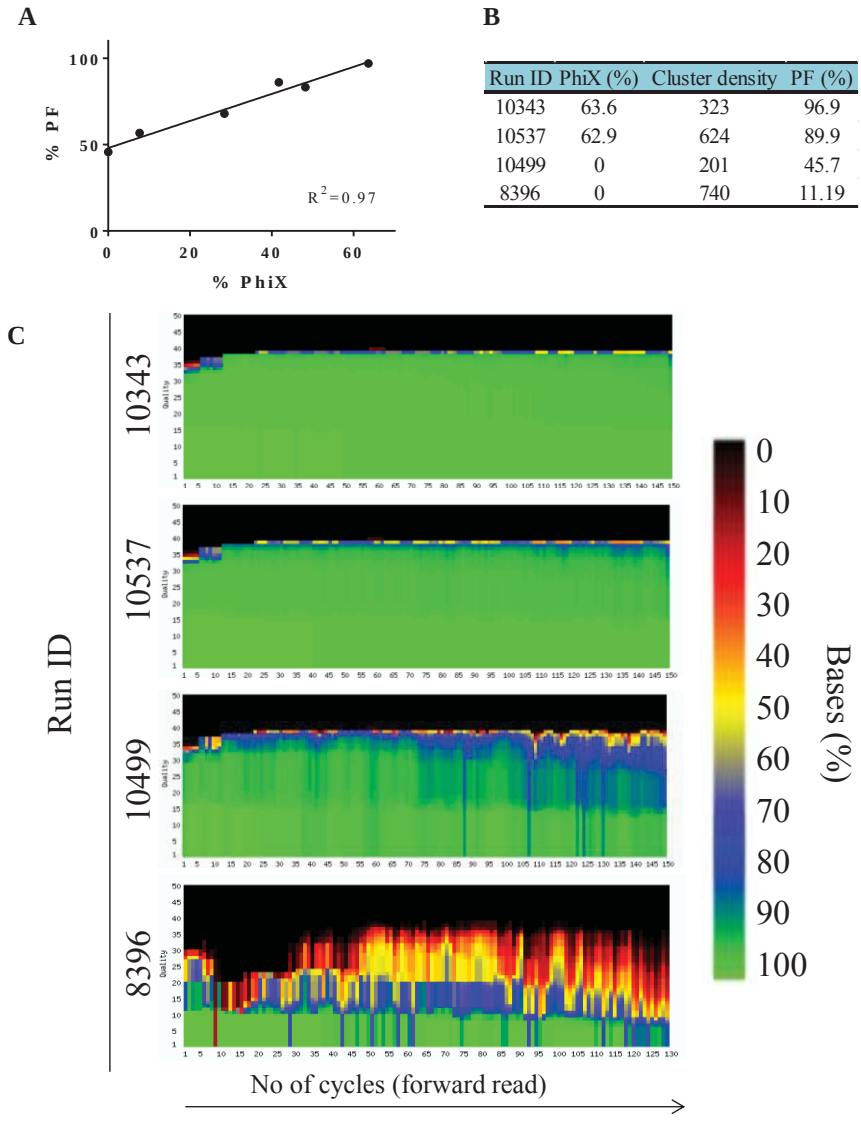


Fig. 3.7| Miseq run quality analysis.
(A) Comparison of the %PF of six different libraries run at similar cluster densities with different proportions of PhiX. The quality of the run highly correlated with the levels of PhiX. (B) Impact on %PF of different cluster densities in the presence and absence of PhiX. (C) Quality histograms of the runs shown in B. The x axis shows the number of sequencing cycles of each run while the y axis shows the quality of base calling: the colour gradient represents the percentage of bases that on a given cycle reached a given quality. Quality values higher than 30 (Q30) are optimal.

Finally, data reproducibility within and between runs for the same sample was verified. For this, two libraries were prepared from the same STM sample with different index adaptors. These were multiplexed and run in the same conditions twice. High correlation was obtained for data generated within runs ($R^2_{(10572)} = 0.9987$ and $R^2_{(10537)} = 0.9999$) and between runs ($R^2 = 0.9995$) thus suggesting that any variation between biological replicates should be seen as real variation (Fig. 3.8).

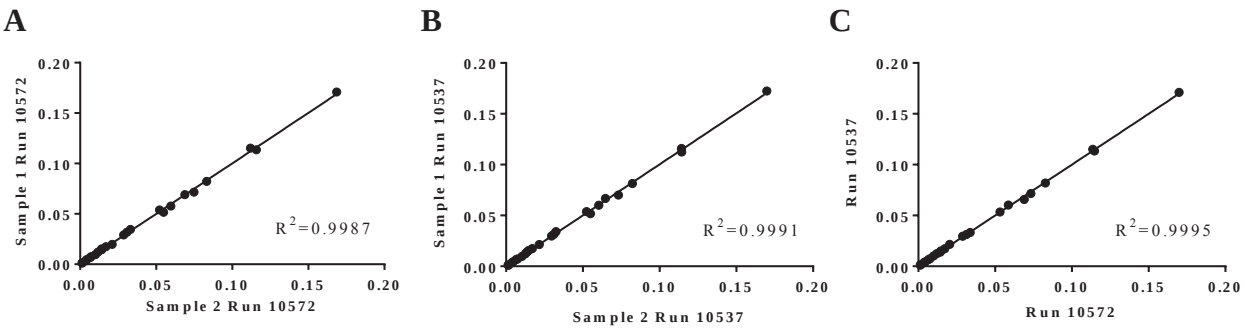


Fig. 3.8| Sequencing reproducibility within and between runs.
(A and B) Pearson correlation analyses of the relative abundances of each barcode within the pool between replicates run in the same lane. (C) Pearson correlation analysis between the averages of each run.
Run IDS: 10537 and 10572, both run at ~6x10^{5}clusters/mm^2; PhiX was spiked in at ~ 60 % and PF values were 89.9 % and 95.9 %, respectively.

3.3 Discussion

In this chapter I have shown that adequate choice of electroporation system can boost transfection efficiency by more than one order of magnitude, which is crucial for the development of an STM approach. In addition, as low as 100 ng of a single *Plasmo*GEM vector is enough to generate transgenic parasites, provided that they are part of a larger pool of at least 1 µg of DNA. This will, in theory, enable the generation of highly complex pools of transgenics in a single transfection.

The library preparation method (DA) optimised here will enable rapid processing of high numbers of samples for multiplex sequencing. However, one consequence of its simplicity is the generation of low complexity libraries. This was overcome by the optimisation of the run conditions, which included low cluster density (4-6 x10^5 clusters/mm^2) and the presence of a base-balanced spike-in (PhiX).

As the *Plasmo*GEM vectors have not been reported to be maintained as episomes and double integration events in the same transfection were shown to be too rare to be detected, if at all existent, I expect that each mutant will carry only one barcode, integrated in the genome.

Barcode sequencing proved to be a reliable method to count barcodes since the abundance of sequencing reads reflected the abundance of the barcodes in the samples. Applied to real STM samples this means that it will be possible to analyse the relative abundance of each barcode, i.e. mutant, and how it changes during infection.

Taken together these data indicated that from the technical point of view *P. berghei* STM approaches are feasible.