Chapter 5

Additional gene mapping studies

Collaboration note *This chapter contains work performed in collaboration with Gemma Langridge and Dr. Keith Turner for bacterial transposon mutant mapping (Langridge et al., 2009), and Francisco Cubillos and Dr. Gianni Liti for yeast linkage analysis (Cubillos et al., 2011; Liti et al., 2009b).*
Gemma and Keith developed the transposon mutant library, performed the experiments, generated raw data, and did the high-level analysis; I contributed the statistical analyses of the data. Similarly, Francisco and Gianni designed and developed the yeast grid of crosses, performed the experiments, and high-level analysis; I contributed the statistical analyses and parts of interpretation of the data.

5.1 Gene mapping with one million bacterial transposon mutants

One trait mapping approach available in prokaryotic and simple eukaryotic organisms is generating a very large number of random mutants, and then examining which mutants survive selection (Chapter 1.1.3). This is related to the work in Chapter 4 on standing variation, but can access a wider variety of alleles. A version of this approach based on transposon insertions was recently developed for the bacterium *Salmonella enterica serovar Typhi* (*S. Typhi*, Langridge et al. (2009)).

5.1.1 Transposon insertion library for *Salmonella Typhi*

S. Typhi causes typhoid fever, and is responsible for hundreds of thousands of deaths in the developing world every year (Crump et al., 2004). One approach to fighting this disease agent is to map the genes essential for its survival in the permissive condition, restrictive conditions associated with its lifecycle in the human host, and under stress from therapeutic agents. To this end, our collaborators created a transposon insertion library with on the order of 1,000,000 mutants, each harbouring one transposon insertion. This large mutant library was then grown in a permissive condition and with added 10% ox bile to simulate gall bladder environment, followed by DNA extraction from the pool, amplification of DNA from the junction between transposon sequence and genomic DNA, and high-throughput sequencing. Mapping the sequencing reads to the genome results in a list of sites where some mutant had a transposon inserted.

Here, we focus on two mapping tasks. First, we look for essential genes that do not allow insertions, followed by study of genes essential for growth in bile, which is important for its persistence in the human host.

5.1.2 Mapping essential genes

To test whether a gene was essential, we quantified how unlikely it was to harbour a transposon insertion. Genes with no observed insertions are likely to be essential, while genes with many insertions are obviously not. For every gene g of length L_g, we calculated the insertion frequency $f_g = \frac{I_g}{L_g}$, where I_g is the observed number of insertions.

We noted that the distribution of f was bimodal with modes at 0 and roughly 0.05, and heavy-tailed (Figure 5.1). The mode at 0 corresponds to the essential genes that do not allow for any insertions, and the mode at 0.05 to all the other genes. Under the assumption of uniform incorporation of the transposon, we would expect the number of insertions in a gene to follow a Poisson distribution. However, the distribution is considerably more dispersed, indicating presence of unknown biases and potential sequence-specificity. Standard approaches to deal with overdispersion, such as using a negative binomial distribution, or a normal distribution with variance proportional to mean, did not give a substantially better fit, and were not straightforward to interpret.

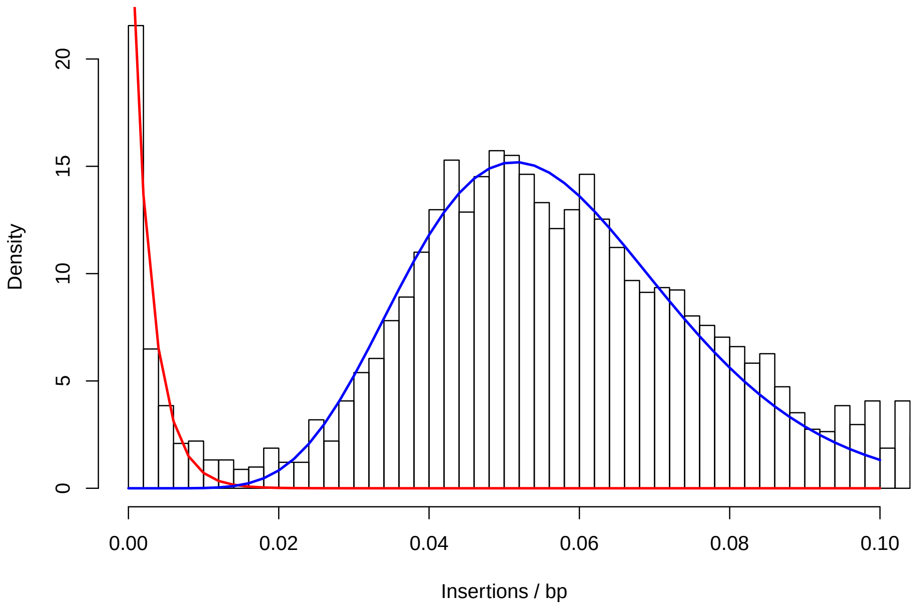

Figure 5.1: Histogram of per-base insertion frequency of individual genes. The blue line corresponds to the Gamma fit to the right mode (non-essential model), while the red line corresponds to the Gamma fit to the left mode (essential model).

Instead of modelling the generative process, we modelled the data directly. We fit Gamma distributions for the two modes of the distribution of the insertion site count in every condition using the R MASS library. For each f_g, we calculated the probabilities corresponding to the right tail of the essential model and left tail of the non-essential model. This represents our belief of observing an insertion index that is at least as extreme for both individual models. For every gene g, we calculated the base 2 logarithm of the likelihood ratio (L_g) between the two model fits, and classified the gene as essential ($L_g < -2$, essential model at least

4 times more likely), non-essential ($L_g > 2$, non-essential model at least 4 times more likely), or uncertain ($|L_g| < 2$).

We found 349 essential genes at false discovery rate less than 0.07. The analysis of these genes is presented in Langridge et al. (2009).

5.1.3 Mapping condition-specific essential genes

Next, we looked for genes essential for growth in bile. These genes should not be essential in general, but insertions in them should be observed less than in the permissive condition. There were three timepoints for growth in bile, we compared the data from each to the data from permissive condition. We analyse number of mapped reads instead of insertion events to avoid many more comparisons of very low frequency (1-2 insertions) events. From the raw data, it was clear that several genes had reduced insertion frequencies as assessed by the number of sequenced reads (Figure 5.2a).

For each pair of conditions (A, B), we calculated the $\log_2$ fold change ratio $S_{g,A,B}$ in the number of observed reads $R_{g,A}$, $R_{g,B}$ for every gene g as $S_{g,A,B} = \frac{R_{g,A}+100}{R_{g,B}+100}$. The correction of 100 reads in the numerator and denominator smooth out the high scores for genes with very low numbers of observed reads, and corresponds to a prior belief that if there is an insertion present, there should be an abundance of reads mapping to it.

Again, we modelled the data directly. We fit a normal model to the mode of distribution of $S_{A,B}$ over all genes, and calculated p-values for each gene according to the fit (Figure 5.2b). This procedure results in an ordered gene list. We chose an arbitrary cutoff, and considered a gene to be condition-specific if the fold-change between the conditions was greater than 4, which corresponds to p-value of 10^{-5}, and false discovery rate of 2.5×10^{-4}. These genes are analysed in depth in Langridge et al. (2009).

5.2 Gene mapping with grid of yeast crosses

Baker's yeast *Saccharomyces cerevisiae* has been successfully used in linkage studies over the last decade, focusing mainly on two F_1 crosses (Ehrenreich et al.

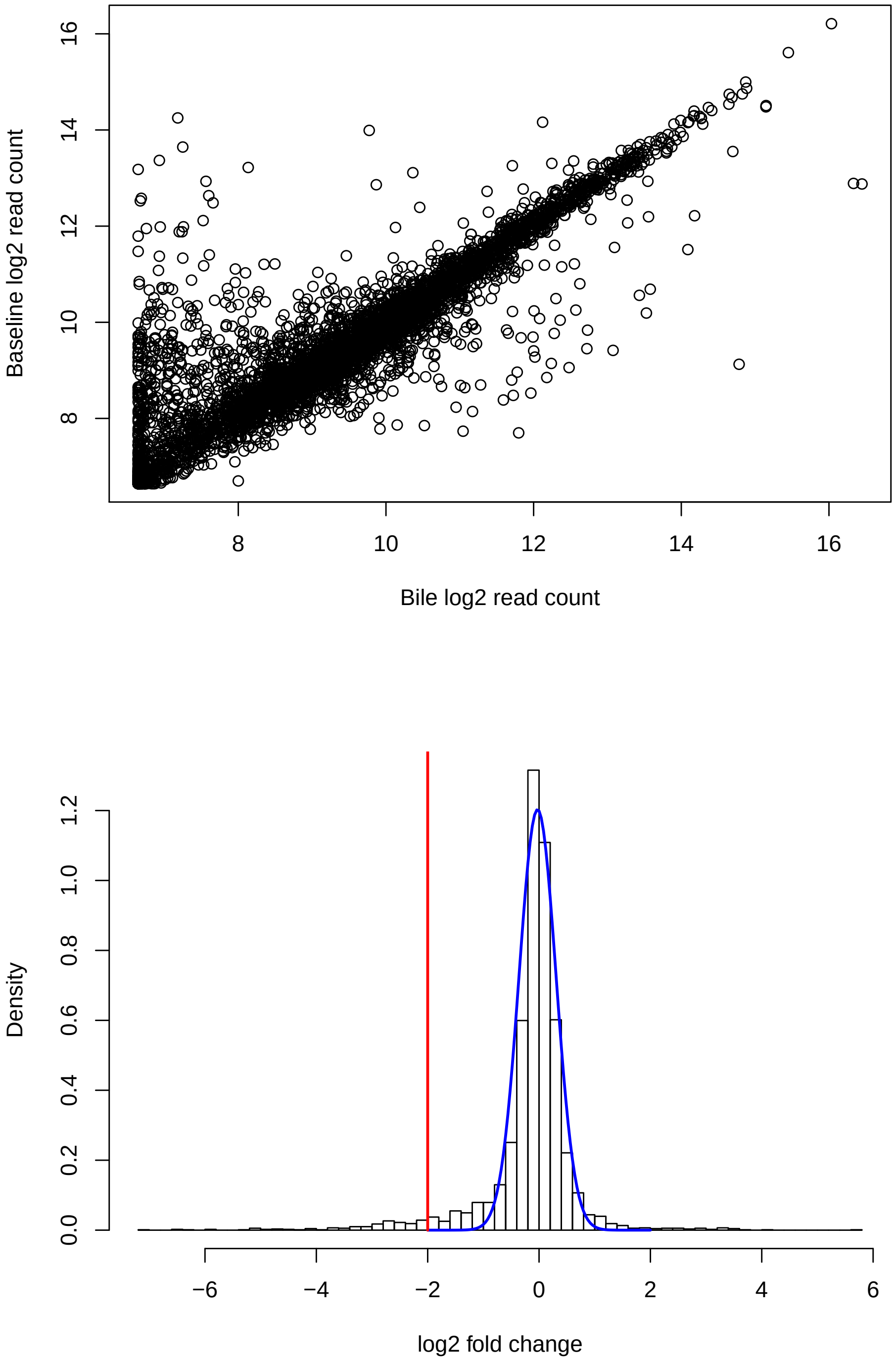

Figure 5.2: **(a)** Scatter plot of $\log_2$ read counts in two conditions. 100 is added to each gene's counts for smoothing. **(b)** Histogram of gene read count $\log_2$ fold change. The blue line corresponds to normal fit to the mode, red line is the cutoff used to determine condition-specificity.

(2009); Mancera et al. (2008); Steinmetz et al. (2002); Zheng et al. (2010), Chapter 1.1.1). However, a single cross gives restricted information about context-dependent allele effects, and is limited to variants present only in the two parental strains.

5.2.1 Grid of yeast crosses

To explore the effects of alleles in different genetic contexts, our collaborators generated a grid of crosses between all five clean (non-mosaic) lineages of *S. cerevisiae* sequenced as part of the *Saccharomyces* Genome Resequencing Project (Liti et al., 2009a). One of the strains (of Malaysian origin) was effectively reproductively isolated, and thus not included for further analysis. The remaining six crosses between the four strains captured 64% of the segregating sites identified by Liti et al. (2009a).

Ninety-six F1 segregants were isolated from 24 meiotic events for each cross. Every segregant was genotyped at 171 evenly spaced markers, followed by quantitative characterisation of growth curves in different conditions. Three growth environments were shared between all crosses, while the rest of the 32 tested environments were cross specific.

5.2.2 Recombination analysis

First, we characterised the global recombination landscape in the six crosses. We called a recombination event between two consecutive genotyped loci in one haploid segregant if the two observed alleles came from different parents.

We determined the average recombination rate ρ_k in each cross k as the number of observed recombination events divided by the genome size. We then used a Poisson model with mean ρ_k to assess the significance of hot- and coldspots in each cross k. A hotspot was deemed significant if the probability of observing as least as many recombination events under the model was less than $\alpha = 0.005$ (FDR$<$ 10%) in at least one cross. Similarly, coldspots were called significant, if the probability of observing up to that many recombination events was less than 0.005.

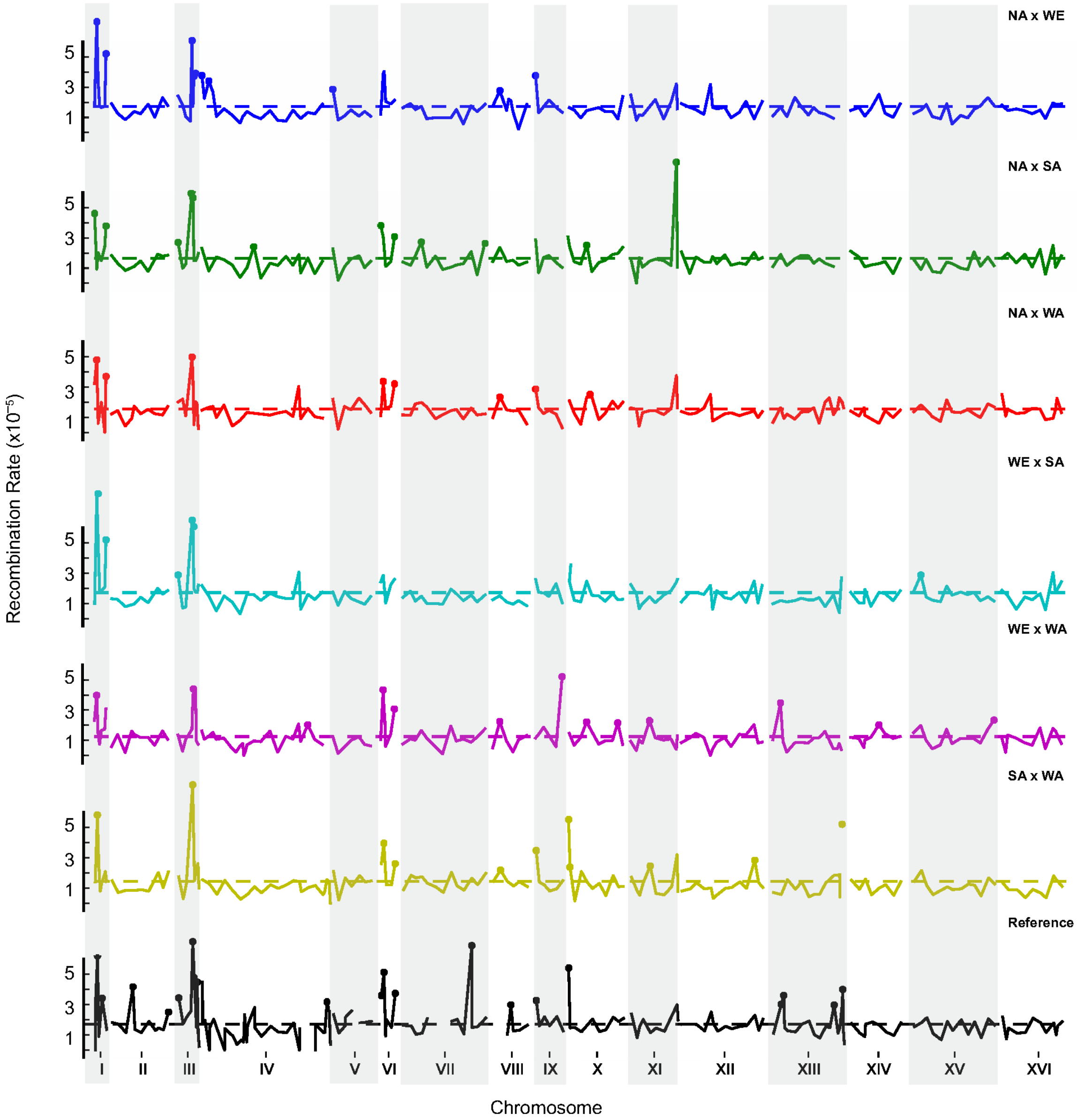

Figure 5.3: Observed recombination rate in each of six crosses, as well as a reference cross from previous work. Recombination hotspots are highlighted with a filled circular marker.

In addition, we used data from Mancera et al. (2008), who established a high-resolution crossover map in another cross using 56 meiotic events. For each marker we genotyped, we took the closest called genotype from their data for every segregant, and repeated our analysis on this dataset.

We found 32 hotspots (Figure 5.3) and 48 coldspots. Nine of the hotspots were not recovered with the high-resolution data form Mancera et al. (2008); seven of the nine were cross-specific, and the remaining two strain-specific, present in all crosses with one strain. We found ten of sixteen centromeric regions to be recombination coldspots in some cross, consistent with their reduced rate of meiotic recombination (Choo, 1998) while no other coldspots were shared between more than three crosses. These results suggest that hotspots, but not coldspots, are mostly conserved. In-depth analysis of these data is given in Cubillos et al. (2011).

5.2.3 Linkage mapping

We then mapped QTLs in all six crosses to determine the regions linked to growth phenotypes in different conditions. Linkage analysis was performed with the rQTL software (Broman et al., 2003) using the non- parametric (Kruskal-Wallis) test for QTLs and normal model for variance explained. LOD> 2.63 was used as cutoff (FDR< 5%) giving less than one QTL by chance per trait. We used the same approach to find strain-specific QTLs by performing one against all tests, pooling data from all crosses with a strain. We also searched for epistatic interactions using the normal model (Chapter 1.4.1), taking LOD> 5.8 (FDR< 5%) as a cutoff.

We found a plethora of QTLs. Two hundred and thirty-three marker-trait pairs were significant. Many of them were specific to a cross or strain, and other combinations were represented as well (Figure 5.4a). We found additional QTLs by pooling the genotypes across crosses (Figure 5.4b), and also detected putative epistatic interactions. Again, more analyses of the QTLs are provided in Cubillos et al. (2011).

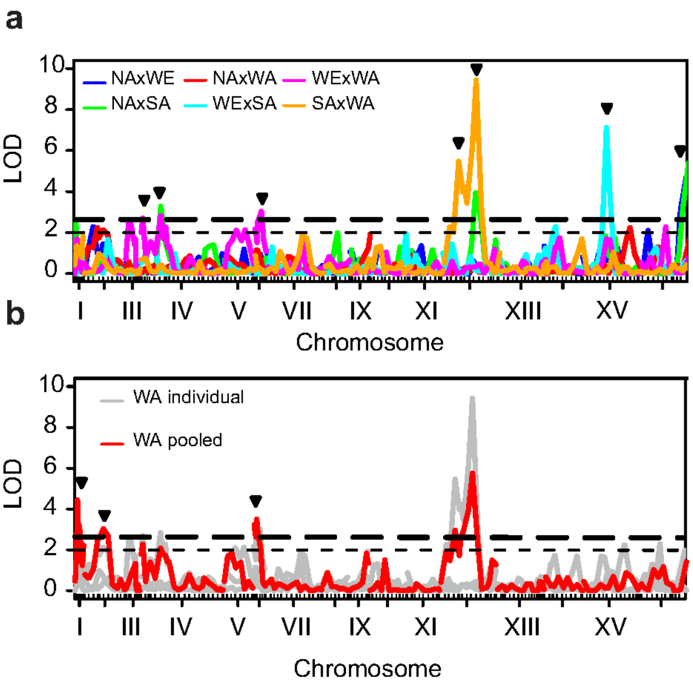

Figure 5.4: Paraquat growth rate QTLs found in each cross independently (a) and in a one-against-all test for the WA strain (b). The phenotyping approach and conditions used are described in Cubillos et al. (2011).