3 Fundamental factors for plasmid stability

3.1 Introduction

In this genomic era, vast amounts of DNA sequence data are being generated. However, the rate of processing these data into information about biological functions is lagging. This situation calls for the development of high throughput methods for the simultaneous functional analysis of multiple genes within genomes. Defining the genes that are essential under specific conditions is of importance for defining the basic materials of synthesis biology and identifying potential targets for new antimicrobial agents. A gene can be defined as essential under given conditions when it is impossible to obtain the knockout of that gene in the condition investigated – one such condition is life itself, i.e. the gene is essential for the survival of the bacterium. Several experimental approaches have been used to define essential gene lists for bacterial isolates including single-gene deletion (Baba *et al.* 2006, de Berardinis *et al.* 2008, Kobayashi *et al.* 2003), ordered or random global transposon mutagenesis (Akerley *et al.* 2002, Hutchison *et al.* 1999, Salama, Shepherd & Falkow 2004, Sassetti, Boyd & Rubin 2003), antisense RNA inhibition (Ji *et al.* 2001) and trapping lethal insertions (Knuth *et al.* 2004).

Random transposon mutagenesis has been the method of choice for many studies because of its speed and cost effectiveness. However, the major drawback of these methods is the possibility of missing essential genes due to (a) sub-saturation knock-out of the whole genome by the transposon and (b) inaccuracy in identifying transposon insertion sites. Problems can also be encountered if the given transposon is too specific in terms of target sequence selection. The majority of transposon mutant libraries contain only a few thousands mutants per genome (Hutchison *et al.* 1999, Salama,

Shepherd & Falkow 2004, Sassetti, Boyd & Rubin 2003), which account for only a fraction of the genes in a given genome, inevitably lead to the missing of essential genes by chance. Signature–tagged mutagenesis (STM) (Hensel *et al.* 1995), transposon-site hybridisation (TraSH) (Sassetti, Boyd & Rubin 2001) and transposon-mediated differential hybridisation (TMDH) (Chaudhuri *et al.* 2009) are transposon-based mutagenesis methods that make use of PCR and hybridisation on microarray respectively for identifying of transposon insertion sites. Although these methods allow simultaneous investigation of genome-wide transposon insertion sites, they are all sub-optimal due to the numbers of transposons that can be located and the inherent inaccuracy of microarray to identify transposon insertion sites. Transposon insertion can also be used to investigate the role of single genes in the stability of single copy plasmids but new methods are needed to address current technological disadvantages and improve the reliability of identifying essential gene function.

Plasmids are extra-chromosomal DNA molecules capable of autonomous replication within their host cells. The genes on plasmids are therefore normally believed to be non-essential to the host. However, genes that contribute to the stable maintenance of a plasmid within a bacterial cell are of great interest not only to the understanding of plasmid biology but also to the discovery of novel drug targets to limit the transmission of antibiotic resistant plasmids.

IncHI1 plasmids have become strongly associated with *S.* Typhi after the introduction of chemotherapy for typhoid fever (WOODWARD, SMADEL 1948, Wain *et al.* 2003, Wain, Kidgell 2004, Wain *et al.* 2003). The *sfh* gene on IncHI1 plasmids has been shown to play a role in their stability in *S.* Typhimurium by reducing the regulatory disruption caused by the presence of the large plasmid (Doyle *et al.* 2007, Doyle, Dorman 2006). However, our knowledge on many IncHI1 plasmid encoded genes is

otherwise very limited; 43% of the genes are still annotated as encoding a "hypothetical protein".

In this chapter, we used a novel random global mutagenesis method called transposon-directed insertion-site sequencing (TraDIS) to investigate the genes important for IncHI1 plasmid stability inside *S.* Typhi during growth in rich media. By using TraDIS, we combined the use of a large transposon insertion mutant library of over one million mutants, with Illumina (formerly known as Solexa) sequencing technology to identify insertion sites with the accuracy of a single base-pair. This technique allows the precise identification of essential genes on the chromosome of *S.* Typhi and, with knowledge of those essential genes, those which are of important for plasmid stability. Here, the work is aimed at identifying which genes are likely to have an important role in plasmid stability.

3.2 Results

3.2.1 The generation of a one million mutant library

The mutant library was generated using a Tn*5*-derived transposon carrying a kanamycin resistant gene (see 2.3.8.2). PCR amplicons of the transposon were coupled with commercial transposases before being electroporated into an attenuated strain of *S.* Typhi Ty2, WT26 harbours the pHCM1 plasmid (see Table 2-1) (Figure 3-1). Transposon inserted mutants were grown on selective media with kanamycin before being collected into pools. The final transposon mutant library contained an estimated 1.1×10^9 individual mutants. The optimisation of the protocol for transposon insertion was undertaken with the assistance of Keith Turner and the library was generated as part of this PhD.

One aliquot of the mutant library was used to investigate plasmid stability (Figure 3-2).

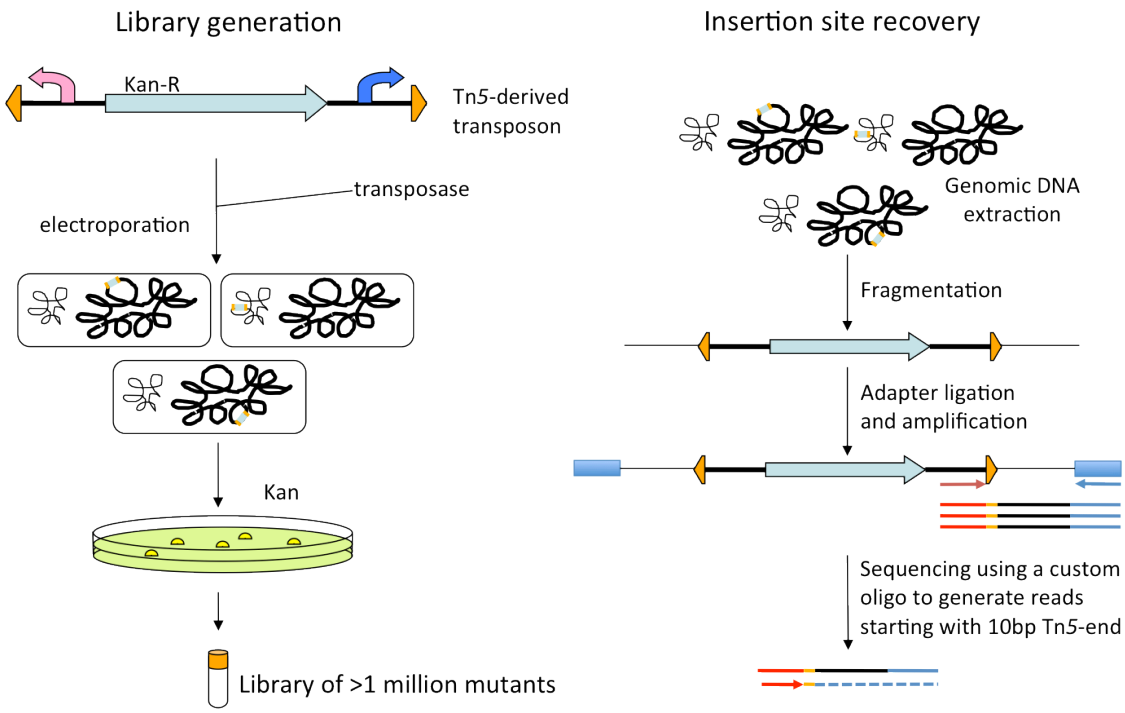


Figure 3-1 Schematic illustration of TraDIS steps

3.2.2 Experimental design

Plasmid pHCM1 encodes resistance to chloramphenicol. To investigate genes involved in stability of IncHI1 plasmids in S. Typhi, the TraDIS transposon mutant pool harbouring pHCM1 was grown in LB broth either supplemented with chloramphenicol (CmP) or without chloramphenicol (non-CmP) for six overnight culture passages (equal to approximately 60 cell generations, Figure 3-2).

Mutation by transposon insertion into any chromosomal gene that is required for stable plasmid inheritance will result in plasmid loss following passage, rendering the bacterial cell chloramphenicol sensitive. In the cultures supplemented with chloramphenicol this will result in loss of those mutants from the mutant pool, but no loss of such mutants will occur from the unsupplemented cultures. Thus, such genes may be identified by having transposon insertions when grown without chloramphenicol, but significantly fewer insertions when grown in its presence.

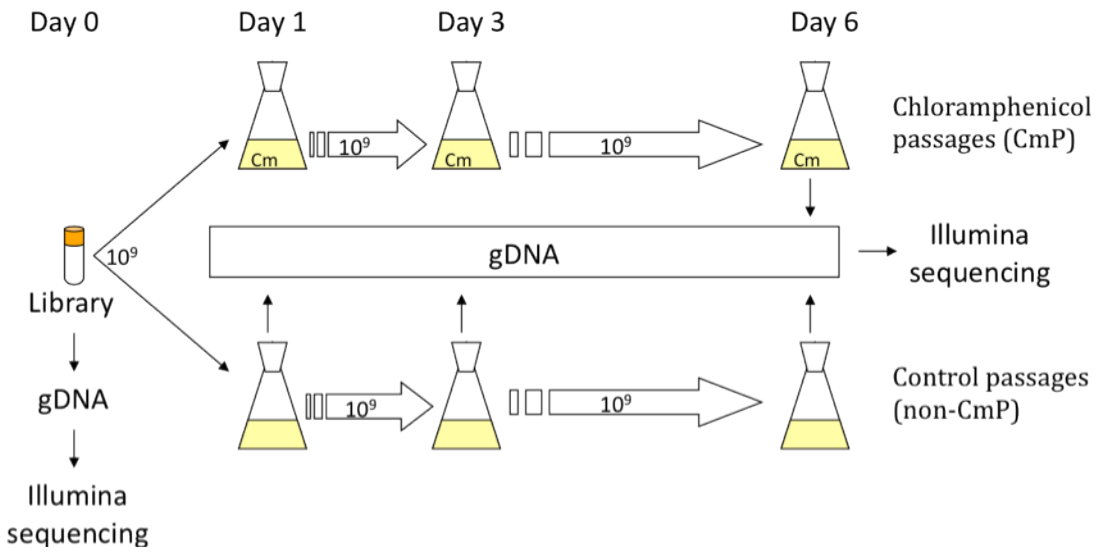


Figure 3-2 Selection assay using the mutant library to investigate plasmid stability

Disruption of plasmid-encoded genes required for stable plasmid inheritance will result in loss of the plasmid following passage. Such genes would therefore be expected to show few insertions regardless of the presence or absence of chloramphenicol. Mutations in genes from post-segregational killing systems would result in the killing of plasmid-free cells instead of just the loss of plasmid. This, however, would also lead to decrease in number of insertions into those genes in both passage conditions.

3.2.3 Identification of insertion sites from the library by Illumina sequencing

DNA samples were extracted from the mutant libraries on day 0 (the initial library), day 1, day 3 and day 6 of the control passages (non-CmP). One DNA sample was obtained from the CmP passaged cells on day 6. Illumina sequencing of DNA prepared from these samples were performed by the sequencing group (Daniel Turner) at WTSI. Briefly, fragmented DNA was sequenced using paired end adaptors and transposon

specific primers. This gave 10bp of transposon sequence to serve as a tag for reads that were transposon-directed.

Each Illumina sequencing lane produced between 1.8 to 6.5 million reads, almost 90% of which contained the 10bp sequence tag. The plasmid specific sequence from each tagged read was then mapped to the reference sequences (NC_004631 for Ty2 and NC_003384 for pHCM1) to identify up to 294,588 insertion sites. The Perl scripts for sequence data manipulation (Minh-Duy Phan and Gemma Langridge) are included in Appendix 8.4.

To identify maximum unique insertion sites from the mutant library, samples from day 0 and day 6 were sequenced on 4 and 5 Illumina lanes respectively, producing up to 12 million reads (day 6, combining of 5 lanes). Figure 3-3 shows the linear increase in reads when combining lanes from the same samples whilst the number of insertion sites reaches saturation after 3 lanes. To balance the maximum unique insertion sites identified and the sequencing cost, it was decided to sequence two lanes for each sample. Thus, the insertion sites identified from this point onwards are from two sequencing lanes (with the exception of the data from day 0 which used 4 lanes).

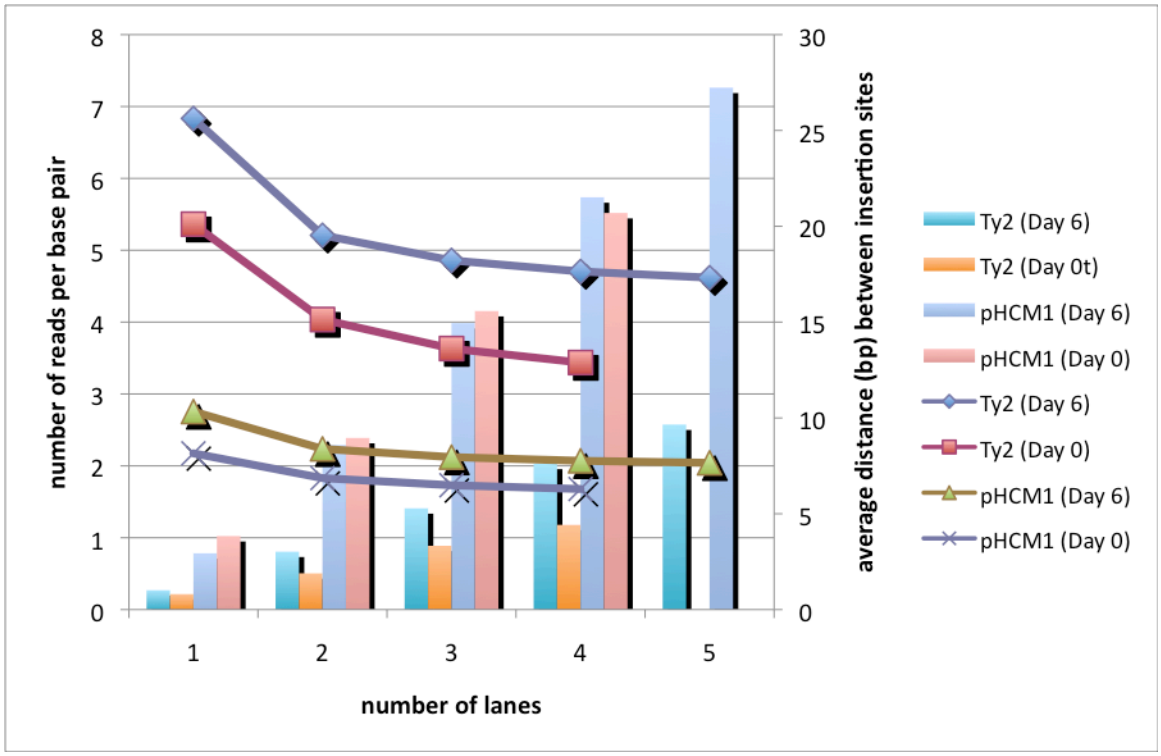
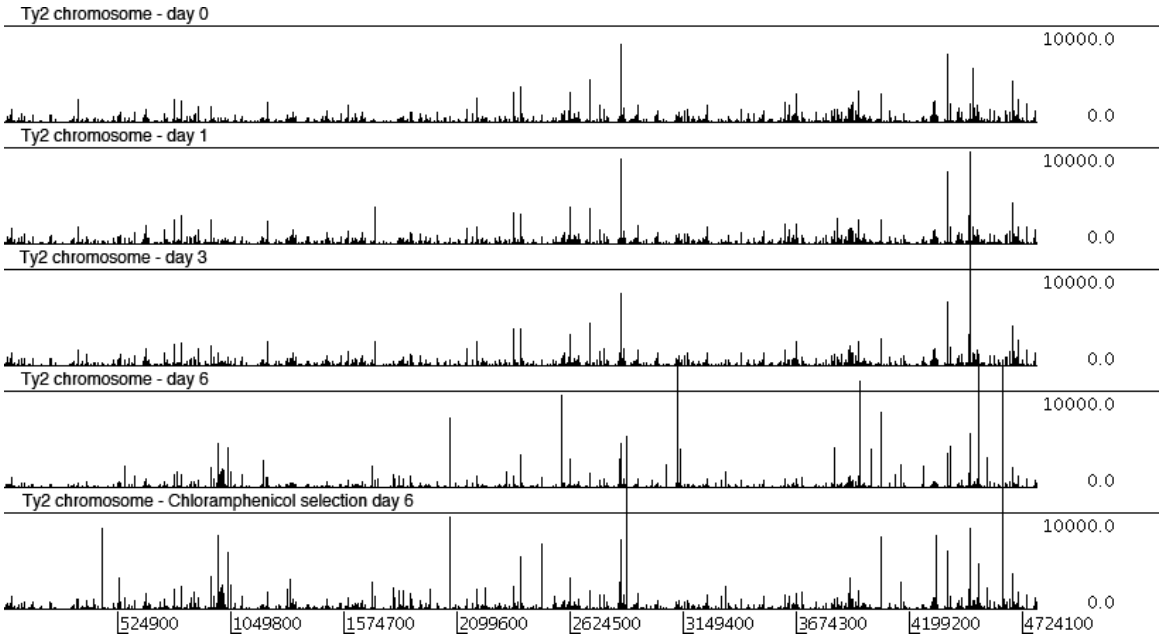


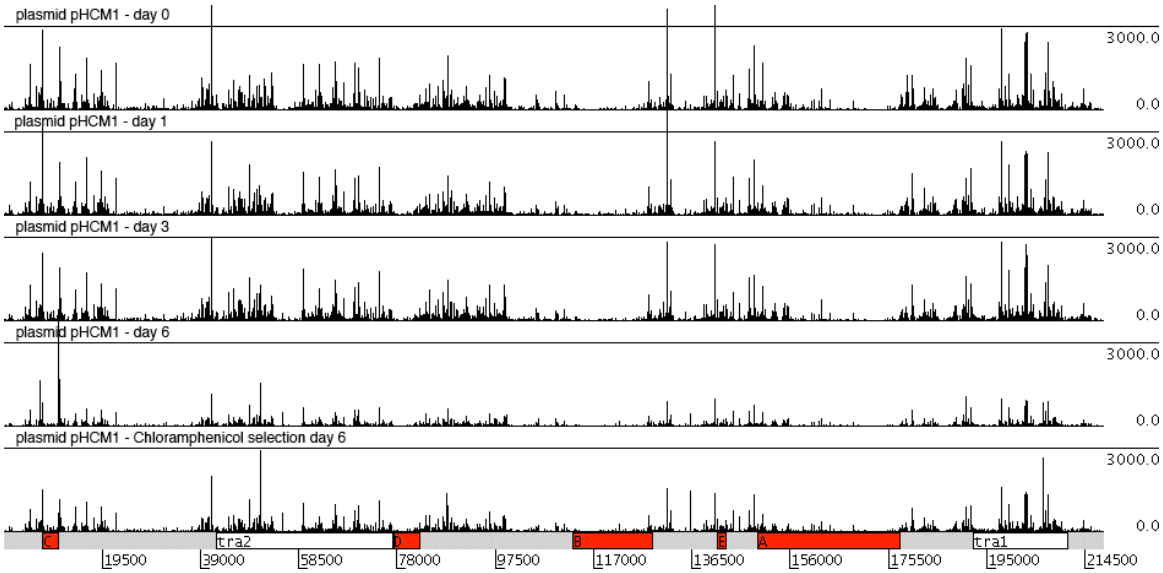
Figure 3-3 Saturation of the transposon insertion sites.
The increase in number of sequencing reads used to map to reference genome leads to near saturation of transposon insertion sites after three lanes
The columns show number of sequence reads per base pair (left axis); the lines show average distance between insertion sites (right axis).

The general distribution of insertion sites across the genome is shown in Figure 3-4. Overall, the insertion sites were well-spread across the genome with some occasional hotspots. Day 0, 1 and 3 showed a similar distribution whilst day 6 and CmP6 were quite different with the reduction in number of reads at the majority of insertion sites. Some insertion sites in day 6 however showed a marked increase in the number of reads. The plasmid plots showed clear cold spots, some of which correspond to the position of mobile elements on the plasmids. Closer inspection of these cold spots revealed some insertion sites, the number of which is still high enough to conclude that very few genes on the plasmid are completely protected (no insertions).

The probability of any gene being missed by transposon insertion was also calculated: The probability of mis-identifying the shortest gene in Ty2 genome, *hisL* (23bp), as essential was 0.186.



(a)



(b)

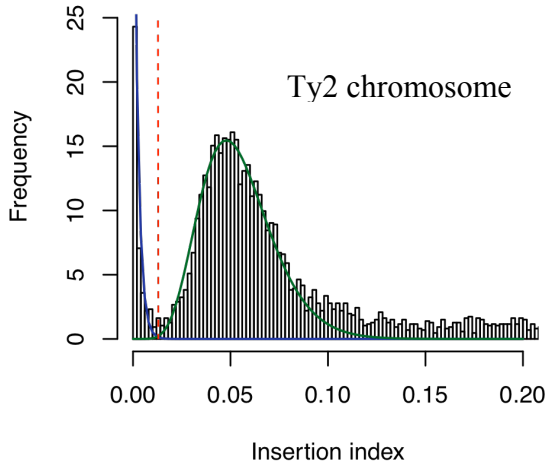
Figure 3-4 Frequency and distribution of transposon directed insert-site sequence reads across the genome of Ty2 (a) and pHCM1 plasmid (b) over time.
The red regions in (b) are those in pHCM1 but not in R27 plasmid, an earlier incHI1 plasmid. The x-axis shows nucleotide position within the genome, the y-axis shows number of reads mapped to each insertion sites. The maximum number of reads shown for Ty2 is 10,000 and for pHCM1 is 3,000.

3.2.4 Essential genes and genes require for long-term survival

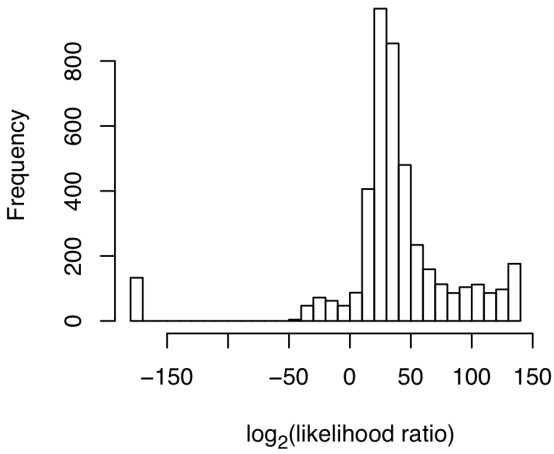
In order to compare the level of insertions across genes of different length, the data were normalised by dividing the number of unique insertion sites within any gene by the gene length to give an insertion index. A frequency distribution of insertion index for all the annotated genes on the Ty2 chromosome gives a clear bimodal distribution (Figure 3-5a). The leftmost peak includes genes with 0 or very low number of insertions. Transposon insertions into these genes were probably lethal to the cells or are required for cellular growth hence their corresponding mutants did not survive or were greatly diminished in the library pool. The rightmost peak represents genes with tolerance to transposon insertions. The gene knock out is either neutral or even advantageous to cellular growth which allows the mutants to survive or even thrive within the pool. This bimodal distribution allowed us to calculate the likelihood ratio of any gene to be on the leftmost peak i.e. essential to the bacteria. The histogram of $\log_2$ likelihood ratio ($\log_2$LR) for all the genes on Ty2 chromosome from day 0 shows a clear cluster at -175 which represents essential genes with no tolerance to transposon insertions (Figure 3-5b). A $\log_2$LR of -2 corresponds to the lowest point between the peaks of the bimodal distribution for all samples investigated. We therefore chose $\log_2$LR of -2 as the global cut-off for essentiality (at which point a gene is four times more likely to belong to the "essential" peak). A $\log_2$LR of 2 was chosen as the cut-off for non-essentiality (four times more likely to be non-essential). Genes with a $\log_2$LR between -2 and 2 could not be assigned as either essential or non-essential.

The insertion index for all the annotated genes on pHCM1 however does not fall into a bimodal distribution (Figure 3-5c) This may be due to a number of factors including the small number of genes, the non-essential nature of plasmid genes and perhaps the redundancy of genes caused by a copy number effect. It is therefore not possible to

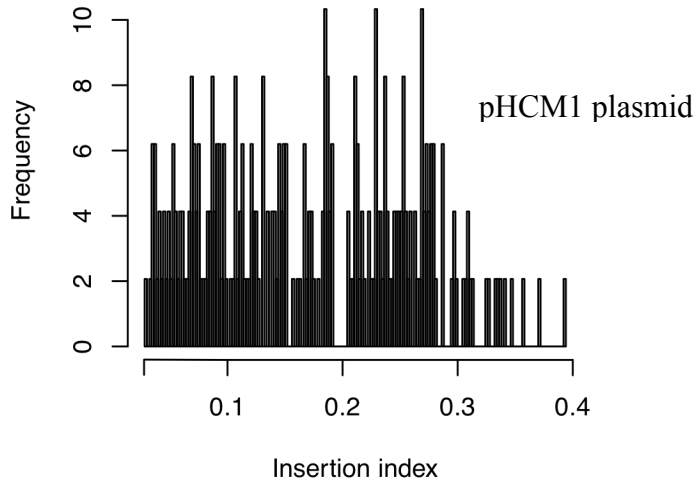
calculate a cut-off for essentiality of plasmid genes. The data however still gives us a ranking list of genes that show low tolerance to transposon insertions suggesting their contribution cellular growth and/or plasmid stability.



(a)



(b)



(c)

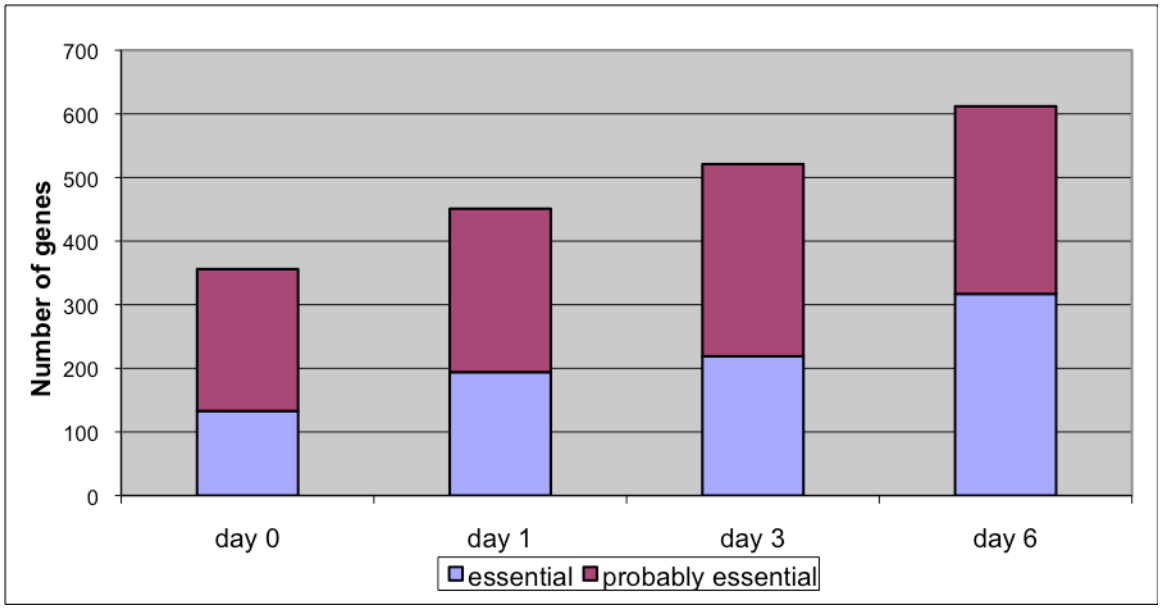
Figure 3-5 Identification of essential genes
(a) Frequency distribution of insertion index (the red dotted vertical line represents the cut-off selected to distinguish between essential and non-essential genes) and (b) Frequency distribution of $\log_2$ likelihood ratio of genes on Ty2 chromosome from day 0 sample; (b) Frequency distribution of insertion index of genes on pHCM1 plasmid on day 0.

From the day 0 sample, 4301 out of 4323 genes on Ty2 chromosome (99.49%) could be assigned to a specific group. Of 356 genes on the chromosome that were protected from transposon insertion 133 had no insertions (essential) and a further 223 genes with

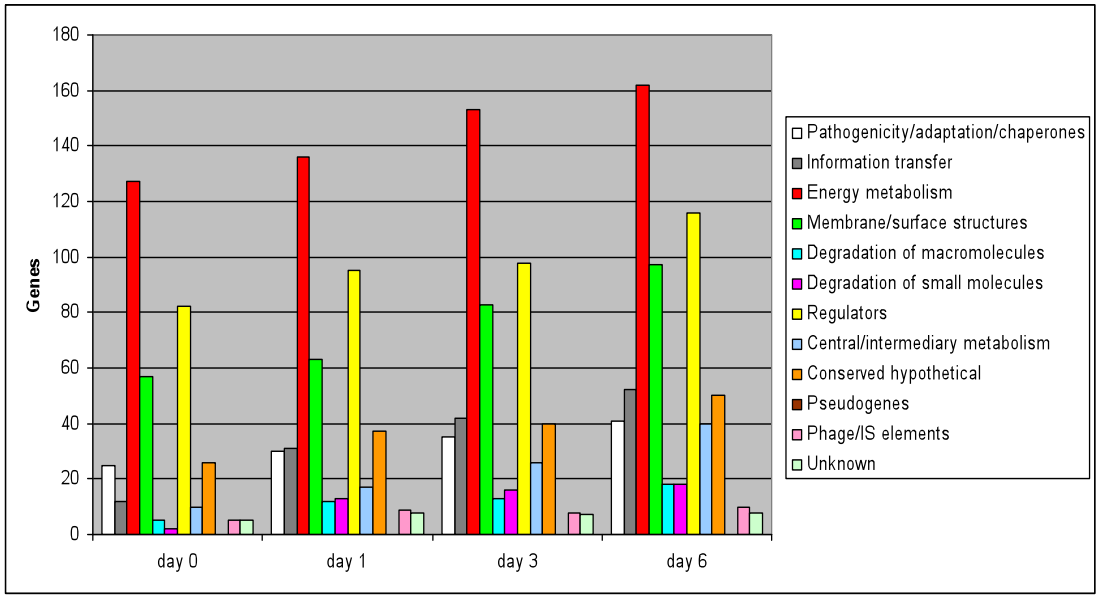
log_2LR lower than -2 (probably essential) (Figure 3-6a). In addition 22 genes could not be confidently assigned as essential or non-essential due to their log_2LR of between -2 and 2, leaving 3945 (91%) non-essential genes at day 0.

In rich, non-selective media the number of essential and probably essential genes increased over time (Figure 3-6a) from 356 on day 0 to 612 on day 6 (see appendix 8.1 for full list of essential genes from day 0). This gradual dropping out of mutants from the library pool likely highlights gene disruptions that do not have an immediately lethal effect on the cell. Such mutants, however, adversely affect cellular growth so that eventually cell death occurs after several generations possibly because the cell is unfit to compete with others and is thus driven out of the pool. Together these are the genes required for long-term survival of the bacterial cells and, as such, cannot be considered as directly involved in plasmid stability. This is important as the plasmid stability experiment was carried out over 6 days and thus 3711 (85.8%) of chromosomal genes could be tested for their effect on plasmid stability.

The long-term survival genes were investigated to assess if they share the same or similar function (Figure 3-6b). The functional classes were assigned based on the CT18 genome annotation of corresponding Ty2 homologues. Genes from the energy metabolism class account for the majority of essential genes. The number of essential genes in all functional classes increases overtime. However, no particular functional classes are significantly associated with genes required for long-term survival.



(a)



(b)

Figure 3-6 The number of essential and probably essential genes for survival in rich, non-selective media at different time points during 6 day passages.
(a) The number of Ty2 essential genes (genes without insertions) and probably essential genes (genes with insertions but have $\log_2$LR lower than -2) at different time points during 6 day passages (b) The number of essential genes within functional classes.

3.2.5 Chloramphenicol resistant and plasmid stability genes on Ty2 chromosome

Plasmid stability is defined as a measure of the likelihood with which a plasmid is inherited by daughter cells at cell division (Nordstrom, Austin 1989). However the term can be used more loosely as the collective results of different mechanisms to ensure the stable maintenance of a plasmid in a bacterial population. Plasmid encoded machineries have predominantly been the focus of researchers who study plasmid stability. However, there have been suggestions of chromosome-plasmid interaction and co-evolution that lead to enhanced fitness of the host cells and hence the stable maintenance of plasmids in the population (Dionisio *et al.* 2005, Lenski, Simpson & Nguyen 1994). However, it is rare to find reports about chromosome-encoded genes that affect plasmid stability. By using a saturated insertion mutant library in combination with long-term passages in selective and non-selective media, we attempted to look for candidates on the Ty2 chromosome that might contribute to the stable inheritance of IncHI1 plasmid in an *S.* Typhi population.

Insertion mutations in plasmid stability gene(s) on the chromosome would be gradually decreased in the library pool overtime under the presence of chloramphenicol as a selective agent for plasmid positive cells. The comparison between non-Cm passage day 6 and CmP day 6 highlighted the mutants that disappear in CmP day 6. Mutants in essential genes defined previously were not considered as candidates for plasmid stability. Care should be taken though to interpret the data because the CmP also selects for genes on the chromosome that contribute to survival in the presence of Cm independent of plasmid mediated resistance.

Three measurements were calculated for each gene: $\log_2$ read ratio ($\log_2$RR) (the ratio of reads in day 6 of non-Cm and Cm passage), the probability of a gene having more reads

in non-Cm than in Cm passage, and the real difference in number of insertion sites within a gene. A list of genes showing high insertion site difference and high probablitity (>0.98) of difference in the two passages is shown in Table 3-1 and the visual comparisons of them are in Figure 3-7.

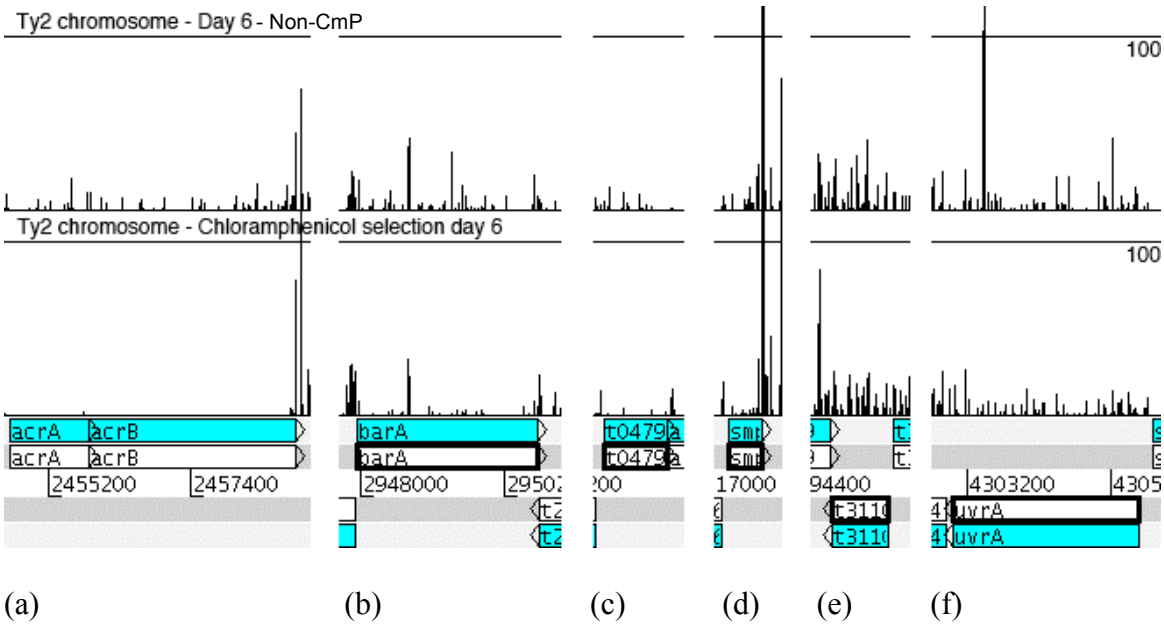


Figure 3-7 Genes on chromosome potentially contributing to survival in chloramphenicol
Artemis plots comparing the insertion sites and their frequency of genes in non-CmP day 6 and CmP day 6; (a) *acrAB,* efflux pump operon; (b) *barA,* encoded a sensor protein; (c) t0479, predicted N5-glutamine S-adenosyl-L-methionine-dependent methyltransferase; (d) *smpB,* tmRNA-binding protein; (e) t3110, predicted extradiol ring-cleavage dioxygenase, class III enzyme, subunit B; (f) *uvrA,* nucleotide excision repair protein subunit A.

The gene with the lowest read difference, t3110 - a conserved hypothetical protein, has 2.7 times more reads in CmP day 6 as in non-CmP day 6 ($Log_2LR = 1.93$). This gene is conserved (with lowest nucleotide identity of 97%) across many serovars within subspecies *enterica* (Typhi, Paratyphi A, B, C, Newport, Choleraesuis etc.). The latest prediction for the function of this gene is extradiol ring-cleavage dioxygenase, class III enzyme, subunit B, which involves in oxidation reduction activity and plays a key role in degradation of aromatic compounds (Interpro entry IPR004183). This is interesting

because chloramphenicol, the selective agent in this experiment, does have an aromatic ring in its structure.

Two genes in a single operon encode an efflux pump (*acrAB)* and show the highest log_2LR (first, 3.04 and third, 1.64 in the list). This pump has been shown to significantly contribute to multiple-antibiotic resistance phenotype in *E. coli* (Okusu, Ma & Nikaido 1996) and *S.* Typhimurium (Piddock *et al.* 2000). This suggests a clear link between *acrAB* mediated resistance and chloramphenicol in the media. The role of *acrAB* in Cm resistance in a Cm acetyltransferase expressing strain has been previously reported in *E. coli* (Potrykus, Baranska & Wegrzyn 2002) and so verifies the assay.

BarA (Log_2LR = 2.37) is a sensor protein that plays a global response regulatory role in cell division, carbon metabolism, iron metabolism and pili formation (Sahu *et al.* 2003). This gene is also well conserved across many serovars of *S. enterica* (nucleotide identity of 98% or more) and also across many *Enterobacteriacae*. BarA belongs to a two-component signal-transduction system: BarA-SirA (Altier *et al.* 2000). In *S.* Typhimurium, BarA-SirA activates SPI1 genes, including the type III secretion system and its effector proteins (Sips), in response to high salt concentration (300mM NaCl) (Mizusaki *et al.* 2008).

The updated annotation of t0479 predicts the gene to encode a N5-glutamine S-adenosyl-L-methionine-dependent methyltransferase, involves in methylation of ribosomal protein L3 (STY2617, NP_456926.1).

SmpB is a tmRNA-binding protein which binds to SsrA RNA to mediate the addition of a short peptide tag to the C-terminus of the partially synthesized polypeptide chain for degradation. SmpB knockout results in phage development defects and failure to tag protein translated from defective mRNA (Karzai, Susskind & Sauer 1999). A SmpB-SsrA mutant in *Yersinia pseudotuberculosis* suffered severe deficiencies in expression

and secretion of *Yersinia* virulence effector proteins, resulting in avirulent phenotype and inability to proliferate in macrophages (Okan, Bliska & Karzai 2006). An SmpB deletion mutant in *S.* Typhimurium affects the expression of 189 proteins in the bacterium proteome, rendering the mutant avirulent and defective in intramacrophage proliferation (Ansong *et al.* 2009). Deletion of *ssrA* in cyanobacterium *Synechocystis* sp. strain PCC6803 results in mutants that are not viable in the presence of the protein synthesis inhibitors chloramphenicol, lincomycin, spiramycin, tylosin, erythromycin, and spectinomycin at low doses that do not significantly affect the growth of wild-type cells (de la Cruz, Vioque 2001). This hyper-sensitivity phenotype of *ssrA* suggests that a SmpB knock-out would also show sensitivity to chloramphenicol.

UvrA is a nucleotide excision repair protein subunit A (Selby, Sancar 1990) involved in DNA damage repair, such as UV radiation damage. This subunit recognises damage DNA and delivers subunit UvrB to the damage site. UvrC then recognises UvrB-damage DNA complex for the excision and repair of the damage DNA.

Table 3-1 Top genes on the chromosome contributing to the survival in chloramphenicol passages

Non-CmP6 Total inserts	Non-CmP6 Total reads	Gene length	Sys ID	Name	CmP6 Total inserts	CmP6 Total reads	$\log_2$(RR)	probability	Insert site diff	Function
67	804	3131	t2385	acrB	3	10	3.04	1	64.00	acriflavin resistance protein B
71	1200	2738	t2867	barA	37	151	2.37	0.9999	34.00	sensor protein
21	220	1175	t2384	acrA	1	3	1.64	0.9981	20.00	acriflavin resistance protein A precursor
23	259	914	t0479	-	8	17	1.62	0.9978	15.00	conserved hypothetical protein
33	562	464	t2642	smpB	23	114	1.63	0.9980	10.00	SsrA (tmRNA)-binding protein
49	1091	812	t3110	-	39	340	1.44	0.9891	10.00	conserved hypothetical

Non-CmP6 Total inserts	Non-CmP6 Total reads	Gene length	Sys ID	Name	CmP6 Total inserts	CmP6 Total reads	log_2(RR)	probability	Insert site diff	Function
										protein
93	1976	2807	t4160	uvrA	83	446	1.93	0.9999	10.00	excision nuclease subunit A

Note: see appendix 8.2 for the full list of genes.

3.2.6 Plasmid mediated cell death and plasmid stability genelist

In order to identify plasmid borne candidate genes for plasmid stability the changes in the insertion index (the number of reads per base pair) of each gene from day 0 to day 6 were compared. Non-CmP and CmP conditions should give a similar gene list because disruptions in stability genes would cause plasmid loss in both conditions. Table 3-2 shows the most significant candidates identified from non-Cm and Cm passages. The level of significance was measured by log_2 of read ratio (log_2RR) and the probability of reads in a gene from day 6 being higher than those in day 0. We used *hok*, a member of hok/sok toxin/antitoxin system involved in the post segregational killing of plasmid free cells to ensure the stable inheritance of plasmid in the population, as a known marker and considered genes with a log_2RR higher than that of *hok* to be candidates for plasmid stability.

Table 3-2 Top plasmid gene candidates for plasmid stability recovered from control and Cm passages after 6 days

Non-CmP day 6 against day 0

Total inserts	Total reads	Gene length	Systematic ID	Name	Day 0 reads	log_2RR	Probability	Gene function
23	**848**	**608**	**HCM1.243**	***tetR***	**5851**	**-2.65017**	**1**	tetracycline repressor protein
61	**686**	**989**	**HCM1.87**	***parA***	**3724**	**-2.28248**	**1**	putative plasmid partition protein
63	**1034**	**1235**	**HCM1.86**	***parB***	**3265**	**-1.56919**	**1**	putative plasmid partition protein
14	**369**	**125**	**HCM1. 141ac**	**-**	**619**	**-0.6164**	**0.999919**	hypothetical protein
79	**7911**	**386**	**HCM1. 178ac**	***sfh***	**10039**	**-0.33986**	**0.996986**	putative DNA-binding protein

Total inserts	Total reads	Gene length	Systematic ID	Name	Day 0 reads	log_2RR	Probability	Gene function
4	69	38	HCM1.166c	-	102	-0.25733	0.992663	putative aminoglycoside acetyltransferase
23	**322**	**437**	**HCM1.128**	**-**	**401**	**-0.24757**	**0.991894**	putative membrane protein
10	274	122	HCM1.53	-	329	-0.19794	0.986795	hypothetical protein
18	231	374	HCM1.145	-	261	-0.12517	0.974434	hypothetical protein
58	**2688**	**368**	**HCM1.245c**	**-**	**2793**	**-0.05334**	**0.953917**	hypothetical protein
18	**669**	**260**	**HCM1.130**	**-**	**697**	**-0.0516**	**0.95329**	hypothetical protein
62	**1558**	**572**	**HCM1.125**	**-**	**1618**	**-0.05129**	**0.953178**	putative membrane protein
23	**4312**	**341**	**HCM1.45**	**-**	**4403**	**-0.02945**	**0.944704**	hypothetical protein
61	**1706**	**695**	**HCM1.277**	**-**	**1743**	**-0.02925**	**0.944623**	putative periplasmic protein
11	**275**	**140**	**HCM1.290c**	***hok***	**281**	**-0.02290**	**0.941937**	putative stable plasmid inheritance protein

CmP day 6 against day 0

Total inserts	Total reads	Gene length	Systematic ID	Name	Day 0 reads	log_2RR	Probablitity	Gene function
11	187	641	HCM1.206	*cat*	4781	-4.08805	1	chloramphenicol acetyltransferase
27	**417**	**608**	**HCM1.243**	***tetR***	**5851**	**-3.5249**	**1**	tetracycline repressor protein
55	**258**	**989**	**HCM1.87**	***parA***	**3724**	**-3.41705**	**1**	putative plasmid partition protein
74	**481**	**1235**	**HCM1.86**	***parB***	**3265**	**-2.534**	**1**	putative plasmid partition protein
81	**4009**	**386**	**HCM1. 178ac**	***sfh***	**10039**	**-1.30306**	**0.997981**	putative DNA-binding protein
48	1548	251	HCM1.124	-	3373	-1.07547	0.979964	hypothetical protein
230	7431	1016	HCM1.92	-	15677	-1.06691	0.978417	putative plasmid stability/partition protein
97	1522	659	HCM1.93	-	3164	-1.00887	0.965039	hypothetical protein
30	**2144**	**341**	**HCM1.45**	**-**	**4403**	**-1.00481**	**0.963893**	hypothetical protein
65	**783**	**572**	**HCM1.125**	**-**	**1618**	**-0.96024**	**0.949159**	putative membrane protein
18	113	332	HCM1.182	-	314	-0.95877	0.948603	hypothetical protein
12	**272**	**125**	**HCM1. 141ac**	**-**	**619**	**-0.95068**	**0.945448**	hypothetical protein
152	3217	854	HCM1.106c	-	6256	-0.93824	0.940296	putative lipoprotein
29	234	302	HCM1.246c	-	536	-0.92917	0.936311	hypothetical protein
20	**326**	**260**	**HCM1.130**	**-**	**697**	**-0.90372**	**0.924011**	hypothetical protein
131	4487	761	HCM1.190	-	8470	-0.90174	0.922982	hypothetical protein
58	831	1202	HCM1.183	-	1630	-0.89391	0.918816	hypothetical protein
102	1609	467	HCM1.269	-	3047	-0.88082	0.911467	hypothetical protein
70	**910**	**695**	**HCM1.277**	**-**	**1743**	**-0.86770**	**0.903616**	putative periplasmic protein
64	**1488**	**368**	**HCM1.245c**	**-**	**2793**	**-0.86535**	**0.902161**	hypothetical protein
60	4772	302	HCM1.24c	-	8700	-0.85298	0.894220	hypothetical protein
115	2581	989	HCM1.100	*trhU*	4730	-0.84925	0.891731	plasmid transfer protein
62	2182	416	HCM1.95	*htdF*	3985	-0.84003	0.885415	putative periplasmic

Total inserts	Total reads	Gene length	Systematic ID	Name	Day 0 reads	log_2RR	Probablitity	Gene function
								protein
104	1104	986	HCM1.209c	-	2047	-0.83448	0.881486	putative transposase
24	**181**	**437**	**HCM1.128**	**-**	**401**	**-0.83424**	**0.881309**	putative membrane protein
18	294	323	HCM1.44	-	600	-0.82915	0.877626	hypothetical protein
107	4160	341	HCM1.199c	-	7416	-0.81911	0.870108	hypothetical protein
116	2881	689	HCM1.61c	-	5126	-0.80991	0.862949	hypothetical protein
38	440	251	HCM1.112	-	846	-0.80888	0.862131	hypothetical protein
566	17887	2663	HCM1.77	*trhC*	31095	-0.79436	0.850245	plasmid transfer protein
10	**121**	**140**	**HCM1.290c**	***hok***	**281**	**-0.78574**	**0.842879**	putative stable plasmid inheritance protein

Genes in bold appear in both non-CmP and CmP gene list.

Chloramphenicol acetyltransferase gene *cat* was highly protected in CmP as oppose to non-CmP (Table 3-2 and Figure 3-8). This enzyme inactivates Cm by covalently binding one or two acetyl groups to the hydroxyl groups on the Cm molecule. The known mechanism of the product of *cat* to confer resistance to Cm is our positive confirmation that the passage worked in selection against mutations within the *cat* gene.

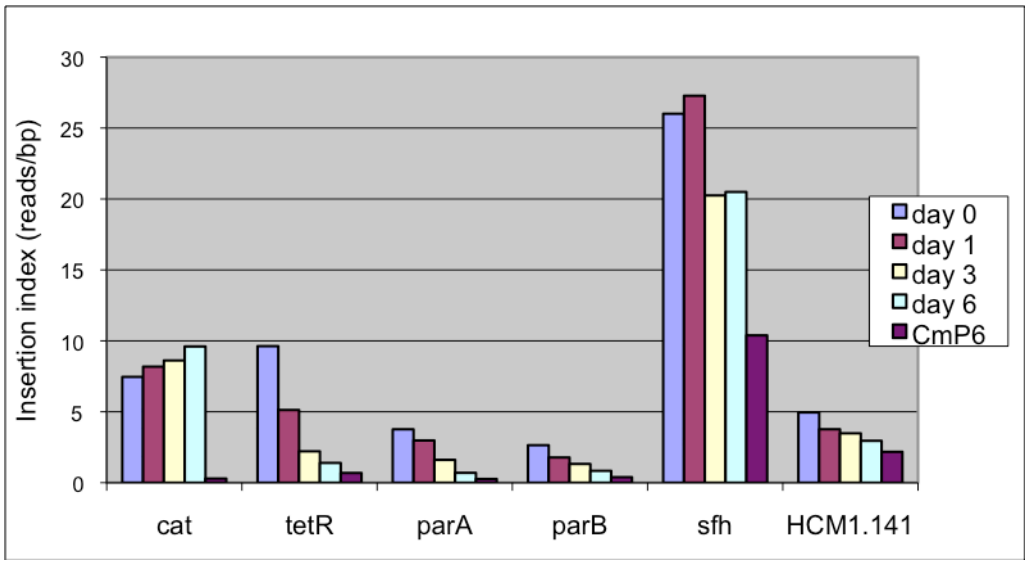
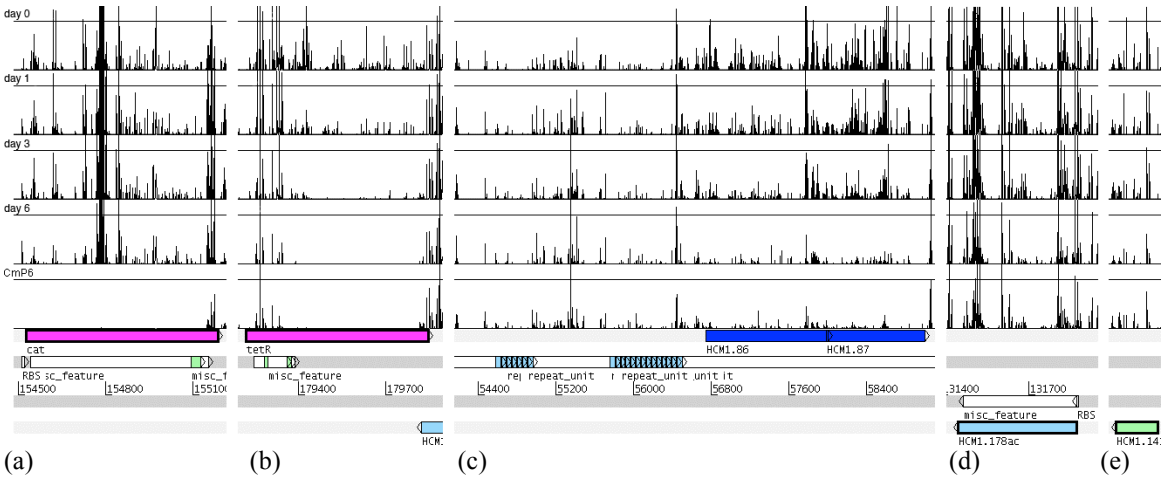


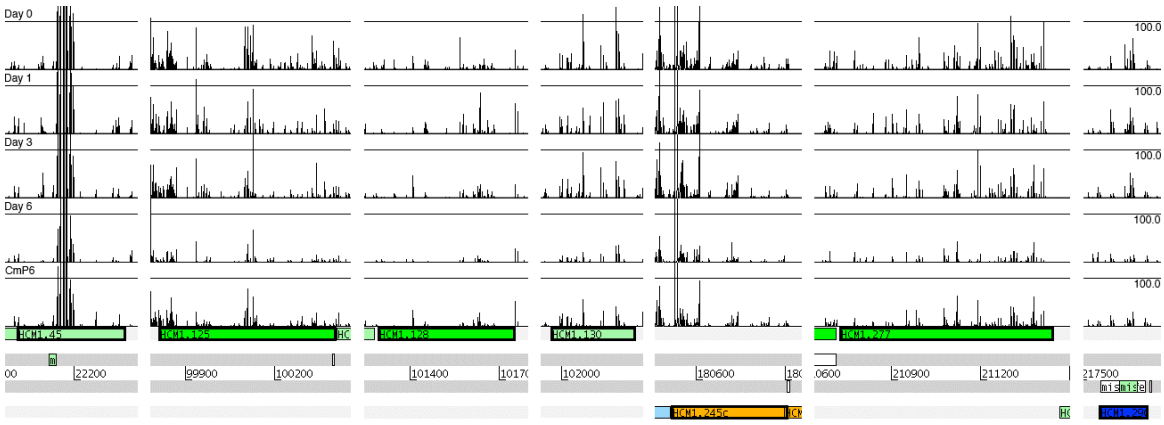
Figure 3-8 Changes in insertion index (number of reads per base pair) of candidate genes across passages and in the chloramphenicol selection passage at day 6 (CmP6)

The changes in insertion index (i.e. number of reads per bp) for several top genes are shown in Figure 3-8. With the exception of *cat*, mutants in other genes gradually

decrease overtime and were lowest in CmP. Care should be taken in interpretation of these data because although a pattern of a gradual decrease in mutants of a gene indicates good candidates for plasmid stability, it may also indicate survival mutants; insertions in plasmid genes that adversely affect cellular growth would also slowly disappear from the passages. One example of this plasmid mediated killing phenomenon is perhaps the *tetR* repressor. The Tet operon on Tn*10* is well characterised: TetR is a tetracycline dependent regulator and *tetA* encodes a metal-tetracycline/H^+ antiporter (Yamaguchi, Someya & Sawai 1992). The expression of *tetA* in the absence of tetracycline causes partial collapse of the membrane potential, arrest of growth and killing of the cells (Eckert, Beck 1989). The repressor *tetR* is also protected in experiments to select spontaneous mutations on the Tet operon within Tn*10* encoded on a multicopy plasmid (Moyed, Bertrand 1983). The data presented here confirms that a pattern of a gradual decrease in mutants of a gene indicates good candidates for plasmid stability but may also indicate insertions in plasmid genes that adversely affect cellular growth. Our plasmid stability gene list in fact likely contains a mixture of real plasmid stability genes and genes, which when knocked out, cause plasmid associated killing of the host cell.



(a) (b) (c) (d) (e)



(f)

Figure 3-9 Changes in number of insertion in top candidates for plasmid stability genes
(a) chloramphenicol resistant gene (*cat*); (b) repressor of tetracycline resistant operon (*tetR*); (c) partition region including *parAB*; (d) *hns*-like gene HCM1.178ac; (e) hypothetical gene HCM1.141ac. (f) seven other candidates for plasmid stability. The graphs were drawn with window size 1, maximum read of 100.

The involvement of *parA* and *parB* in IncHI1 plasmid stability has been shown previously (Lawley, Taylor 2003). The double deletion of *parA* and *parM,* the minor partition genes, resulted in the integration of R27 plasmid into the chromosome (Lawley, Taylor 2003).

```
>lcl|35551 Sfh_shigella
Length=134

 Score =  257 bits (657),  Expect = 4e-74, Method: Compositional matrix adjust.
 Identities = 132/134 (98%), Positives = 132/134 (98%), Gaps = 0/134 (0%)

Query  1    MSEALKSLNNIRTLRAQGRELPLEILEELLEKLSVVVEERRQEESSKEAELKARLEKIES  60
            MS ALKSLNNIRTLRAQGRELPLEILEELLEKLSVVVEERRQEESSKEAELKARLEKIES
Sbjct  1    MSGALKSLNNIRTLRAQGRELPLEILEELLEKLSVVVEERRQEESSKEAELKARLEKIES  60

Query  61   LRQLMLEDGIDPEELLSSFSAKSGAPKKVREPRPAKYKYTDVNGETKTWTGQGRTPKALA  120
            LRQLMLEDGIDPEELLS FSAKSGAPKKVREPRPAKYKYTDVNGETKTWTGQGRTPKALA
Sbjct  61   LRQLMLEDGIDPEELLSPFSAKSGAPKKVREPRPAKYKYTDVNGETKTWTGQGRTPKALA  120

Query  121  EQLEAGKTLDDFLI  134
            EQLEAGKTLDDFLI
Sbjct  121  EQLEAGKTLDDFLI  134

(a)
```

```
>lcl|54641 Hns_CT18
Length=137

 Score =  141 bits (355),  Expect = 4e-39, Method: Compositional matrix adjust.
 Identities = 82/135 (60%), Positives = 99/135 (73%), Gaps = 1/135 (0%)

Query  1    MSEALKSLNNIRTLRAQGRELPLEILEELLEKLSVVVEERRQEESSKEAELKARLEKIES  60
            MSEALK LNNIRTLRAQ RE  LE LEE+LEKL VVV ERR+EES+  AE++ R  K++
Sbjct  1    MSEALKILNNIRTLRAQARECTLETLEEMLEKLEVVVNERREEESAAAAEVEERTRKLQQ  60

Query  61   LRQLMLEDGIDPEELLSSFSAKSGAPKKVREPRPAKYKYTDVNGETKTWTGQGRTPKALA  120
             R++++ DGIDP ELL+S +A     K  R  RPAKY Y D NGETKTWTGQGRTP  +
Sbjct  61   YREMLIADGIDPNELLNSMAAAKSGTKAKRAARPAKYSYVDENGETKTWTGQGRTPAVIK  120

Query  121  EQL-EAGKTLDDFLI  134
            + + E GK L+DFLI
Sbjct  121  KAMEEQGKQLEDFLI  135

(b)



>lcl|35767 StpA_CT18
Length=133

 Score =  140 bits (354),  Expect = 5e-39, Method: Compositional matrix adjust.
 Identities = 77/134 (57%), Positives = 101/134 (75%), Gaps = 1/134 (0%)

Query  1    MSEALKSLNNIRTLRAQGRELPLEILEELLEKLSVVVEERRQEESSKEAELKARLEKIES  60
            M+  L++LNNIRTLRA  RE  +++LEE+LEK  VV +ERR+EE  ++ +L  + EKI +
Sbjct  1    MNLMLQNLNNIRTLRAMAREFSIDVLEEMLEKFRVVTKERREEEELQQRQLAEKQEKINA  60

Query  61   LRQLMLEDGIDPEELLSSFSAKSGAPKKVREPRPAKYKYTDVNGETKTWTGQGRTPKALA  120
              +LM  DGI+PEEL +  SA   + KK R+PRPAKY++TD NGE KTWTGQGRTPK +A
Sbjct  61   FLELMKADGINPEELFAMDSAMPRSAKK-RQPRPAKYRFTDFNGEEKTWTGQGRTPKPIA  119

Query  121  EQLEAGKTLDDFLI  134
            + L AGK+LDDFLI
Sbjct  120  QALAAGKSLDDFLI  133

(c)
```

Figure 3-10 Pair-wise comparisons of Sfh protein on pHCM1 with its homologues
Comparison of pHCM1 Sfh against (a) Sfh protein from *Shigella flexneri* 2a 2457T, (b) Hns protein from *S.* Typhi CT18 and (c) StpA protein from *S.* Typhi CT18. The comparison was run using BLAST for protein (blastp).

The *sfh* gene also contributes to plasmid stability by silencing plasmid genes to minimise the interference to chromosomal gene regulation (Doyle *et al.* 2007, Banos *et al.* 2009). The Sfh protein, the third member of H-NS-like protein family, was first reported in *Shigella flexneri* 2a 2457T to be encoded on an R27-like plasmid, (Beloin *et al.* 2003). Two other H-NS-like proteins are H-NS and StpA (Dorman, Hinton & Free 1999). All three proteins were also found in *S.* Typhi CT18 harbouring pHCM1 plasmid. The homology of these proteins is shown in Figure 3-10. The DNA binding profile of Sfh to promoters of virulence genes and to DNA curvature (similar to the

binding of H-NS and StpA) suggests its role in regulating virulence genes and the interaction of these paralogues in a complex regulatory network within the cell (Beloin *et al.* 2003). It was also shown that the Sfh in R27 interacts with Hha to thermo-regulate the conjugation of IncHI1 plasmid (Alonso *et al.* 2005, Forns *et al.* 2005). The Hha protein in pHCM1 (HCM1.135), however, was not identified in our experiment as contributing factor to the stable plasmid inheritance. This might be due to the fact that our passages were performed at non-permissive temperature (37°C) for plasmid conjugation.

The remaining genes in our list are good candidates for further investigation, especially those genes highlighted from both passage conditions (Table 3-2 and Figure 3-9). These include one putative periplasmic protein (HCM1.277), two putative membrane proteins (HCM1.125 and HCM1.128) and four hypothetical proteins (HCM1.45, 130, 141ac and 245c). HCM1.277 belongs to the nuclease-related domain (NERD) superfamily. HCM1.125 encodes a potential ribonucleotide-diphosphate reductase subunit alpha domain (PRK07632). With the exception of HCM1.245c, which shares close similarity to proteins in other plasmids, all these genes are unique for IncHI1 plasmids.

3.2.7 Growth curves of sfh knock-out

Previous evidence for the involvement of *sfh* in the silencing of plasmid genes to avoid the disruption of chromosomal regulation is based on experiments in *S.* Typhimurim (Doyle *et al.* 2007, Banos *et al.* 2009). Although similar IncHI1 plasmids have evolved in *S.* Typhi, it is possible that there are unique interactions between pHCM1 and *S.* Typhi. We generated *sfh* deletion mutants in pHCM1 to further investigate the plasmid-chromosome interaction in *S.* Typhi. The *sfh* deletion was generated by an allelic exchange strategy (Turner, Nair & Wain 2006). The strategy was designed to use homologous recombination to swap the *sfh* gene with a kanamycin resistant marker.

The mutant genotype was then confirmed by sequencing. Three Δ*sfh* mutants were generated independently as biological replicates.

This section presents the growth curves of Δ*sfh* mutants in comparison with the wild type (Figure 3-11). There is no significant difference in the observed growth rates between the mutants and wildtype strains. The effect of Δ*sfh* on cellular growth is perhaps too subtle to be detected in the growth curves. Other methods such as competitive growth, long-term plasmid stability assay or gene expression analysis are needed to characterise these mutants. Data from Doyle et al (2007) suggests that the knock out of *sfh* in *S.* Typhi might also reduce the relative fitness of the mutant whilst enhance the level of survival in macrophage.

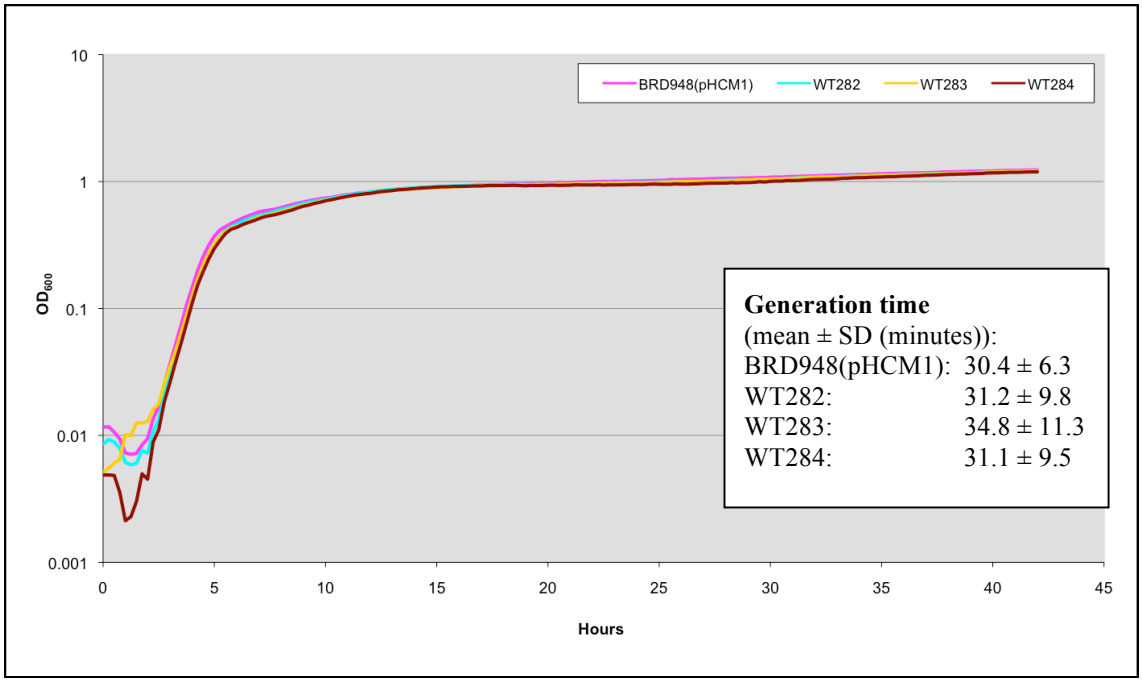


Figure 3-11 Growth curves of Δ*sfh* mutants (WT282, WT283 and WT284) and wildtype strains
(See Methods section 2.2.4.3 for the calculation of generation time)

3.3 Discussion

Transposon insertion mutagenesis is the method of choice for genome wide high throughput investigation of genetic essentiality due to its speed and cost effectiveness. However, major drawbacks of this method involve the capability to generate a saturated

mutant library and the ability to accurately identify each insertion site. In this chapter, we presented a novel method to simultaneously and precisely identify a very large number of transposon insertion sites using Illumina sequencing. The actual sequence of every insertion site provides unprecedented clear signals for each transposon insertion in a semi-quantitative manner. Our mutant library consisted of more than 1 million mutants, which proved essentially to be close to saturation for the method employed. Our transposon-directed insertion-site sequencing (TraDIS) method has therefore addressed some of the disadvantages of previous transposon mutagenesis methods.

TraDIS should be easily applicable to other transposon libraries without the need to re-generate them. One simply needs to re-design PCR primers and sequencing primers for Illumina sequencing for the new transposon. One bottleneck of this method, however, is the generation of a large mutant library. Optimisation of a suitable transposon for a particular organism is of importance to achieve the high frequency of randomised mutagenesis.

The semi-quantitative data enabled us to compare not only the location of insertion events but also the frequency of each event. This is particularly useful in the discovery of long-term survival genes over serial passages. The gradual disappearance of certain mutants over time identifies any genes contributing to survival or fitness. This is of particular importance in assessing plasmid stability because the copy number of plasmids in some cells may be more than one and so knocking out only one copy of a multi-copy gene may not affect the plasmids stability. The measurement of the comparative “success” of each mutant identifies genes that are costly to cellular growth (in the conditions tested) and the deletions of such genes are therefore beneficial (data not shown). The ability to follow the dynamic changes of almost every mutant overtime could also be powerful in multi-stage experimental designs such as cell adhesion and

invasion assays. One mutant library can therefore be used to investigate a range of the biological features associated with bacterial cells.

We used TraDIS to identify plasmid stability genes on both the pHCM1 plasmid and the *S.* Typhi Ty2 chromosome. Two parallel serial passages of up to 6 days in rich media, one with Cm and one without, were used as negative selection for plasmid stability mutants. For chromosome-encoded plasmid stability genes, non-Cm passages were also used to set a baseline for essential genes and long-term survival genes. The genes identified were subtracted in the data from the Cm passages. For plasmid-encoded plasmid stability genes the two passages were effectively duplicates and generated similar gene lists ranked by the difference in insertion index for each gene between day 0 and day 6. The chromosome-encoded genes identified by our assays are a mixture of Cm resistance associated genes and stability genes. The mutants highlighted in Cm passages are those that rendered the cells unfit to remain after 6 days of serial culture in rich media (LB) broth supplemented with Cm. Based on literature information, it is most likely that *acrA* and *acrB* are involved in Cm resistance (Okusu, Ma & Nikaido 1996, Piddock *et al.* 2000). It is known that the action of chloramphenicol acetyl transferase (Cat) alone leads to the depletion of intracellular acetyl coenzyme A, hence the *acrAB* efflux pump system may compensate for Cat activity (Potrykus, Baranska & Wegrzyn 2002). This phenomenon however appears to be in a background specific for *E. coli* strain CM2555 which has a dysfunctional *acrA* gene (Potrykus, Baranska & Wegrzyn 2002), suggesting that in other *E. coli* there might be another pump (*acrEF*) contributing to Cm resistance. In our *S.* Typhi background however, it appears that *acrAB* is the sole efflux pump system for Cm resistance. SmpB is a tmRNA that plays an important role in the degradation of partially synthesized polypeptide chain (Karzai, Susskind & Sauer 1999). There is also evidence suggesting that SmpB deletion might

cause hypersensitivity to Cm (de la Cruz, Vioque 2001). We also discovered a hypothetical protein t3110 that has never been reported as involved in Cm resistance. This gene is predicted to encode an extradiol ring-cleavage dioxygenase class III enzyme, which potentially has a role in degradation of aromatic compounds, of which Cm is one. Experimental evidence is still needed but our preliminary conclusion for this gene is that it is involved in antibiotic resistance.

Two genes *barA* and *uvrA* have known functions (Sahu *et al.* 2003, Selby, Sancar 1990) but none of these suggested a role in either Cm resistance or plasmid stability. The t0479 gene is also not previously predicted to be involved in plasmid stability. We can conclude based on our assay that *barA, uvrA* and t0479 are candidates for plasmid stability genes encoded on Ty2 chromosomes.

We were able to rank the plasmid stability candidate genes based on the decrease of mutants between day 0 and day 6 ($\log_2$(read ratio)). It is however difficult to define a cut-off for plasmid stability because we have no prior knowledge of how such mutant differences should be accounted for. Stability and copy number may both be contributing factors. Our statistical analysis can only provide a measure of how significant a difference is but not how likely a gene is to be responsible for stability. The *hok* gene on pHCM1 is similar to the host-killing gene on plasmid R1, which has been shown to contribute to the maintenance of plasmids (Gerdes, Rasmussen & Molin 1986). It is therefore very likely that pHCM1 *hok* gene contributes to the stable maintenance of this plasmid. We thus used this gene as a phenotypic cut-off. Genes that show higher level of difference than *hok* are more likely to contribute to plasmid stability.

Because any mutants that de-stabilise the plasmid would cause plasmid loss in both non-Cm and Cm passages, the plasmid stability genes were identified as the genes

showed significant decreased in mutants between day 0 and day 6 in both conditions. Non-Cm and Cm passages were in this case considered as two replicates. It is worth noting that the difference is bigger in CmP than in non-CmP for the same gene. For example, $\log_2$RR of *sfh* in non-CmP is -0.33986 whilst in CmP is -1.30306. This means the number of *sfh* mutants in CmP decreases more rapidly than in non-CmP. The presence of chloramphenicol is likely to have attributed to this because of competition during growth; with no plasmid free cells (killed by chloramphenicol) plasmid positive cells could grow to higher densities, or during the sequencing reactions, chromosomal DNA in plasmid free cells may have diluted plasmid DNA.

Any plasmid mutants that are lethal to the cells were also selected by this analysis. TetR mutants are potentially an example of plasmid mediated cell death. The disruptions of *tetR*, the *tetA* repressor, would result in constitutive over expression of *tetA*. The presence of TetA, a proton antiporter, in the absence of tetracycline causes loss of membrane potential resulting in cell death (Eckert, Beck 1989).

A literature search of other top genes on the list did not provide alternative evidence to support their role in plasmid mediated bacterial cell death. We therefore believed that they are candidates for plasmid stability genes. Apart from *tetR*, the two partition genes *parA* and *parB* showed most significant decrease in their mutants after 6 days in both conditions. The *parA* gene encodes a Walker-type ATPase similar to those in P1/F plasmids and *parB* encodes a DNA-binding protein that binds to the centromere region. This partitioning module has been shown to contribute significantly to IncHI1 plasmid stability, especially in condition causing slow growth (Lawley, Taylor 2003). The identification of *parAB* partitioning module proves that our method is picking up plasmid stability genes.

Other genes that were highlighted include HCM1.178ac (*sfh*), HCM1.45, HCM1.125, HCM1.141ac, HCM1.130, HCM1.277, HCM1.245c and HCM1.128. The *sfh* is an *hns*-like gene on IncHI1 plasmids that has been shown to play an important role in minimising the bacterial fitness cost by minimising the regulatory disruption caused by the presence of a large plasmid (Doyle *et al.* 2007, Doyle, Dorman 2006, Banos *et al.* 2009). HCM1.125 and HCM1.128 encodes two putative membrane proteins with the predicted signal peptides and transmembrane domains. HCM1.277 carries a nuclease-related domain (NERD) superfamily, which suggests a role in DNA processing and this may have nuclease function (IPR011528). The remaining genes encode hypothetical proteins. These genes are all conserved within IncHI1 plasmids. Thus, our assay has successfully identified candidate genes for plasmid stability along with genes known to be involved in chloramphenicol resistance and plasmid mediated cell death.

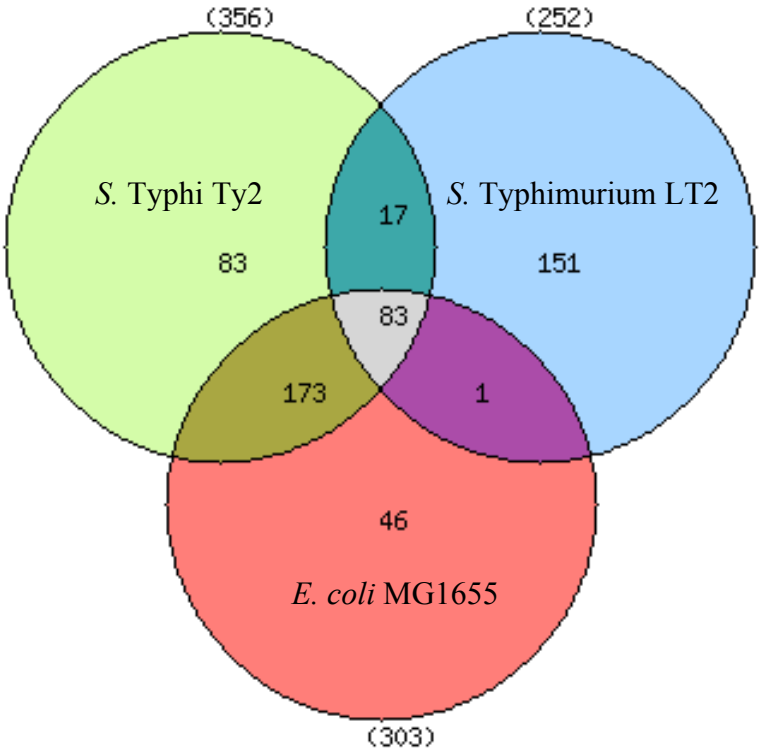


Figure 3-12 Comparing essential genes of *S.* Typhi, *S.* Typhimurium and *E. coli*
The essential genes of *S.* Typhi are taken from this study, *S.* Typhimurium from Knuth *et al.* (2004) and *E. coli* from Baba *et al.* (2006)

It is encouraging that our method provides high enough resolution to precisely discern and trace the dynamic changes of any mutant of interest among million of other mutants. Our list of 356 essential genes is compatible with other studies (Baba *et al.* 2006, Knuth *et al.* 2004, Zhang, Lin 2009, Zhang, Zhang 2008) (Figure 3-12). There are 256 genes shared between our S. Typhi essential gene candidates (70%) and *E. coli* (84%) (Baba *et al.* 2006). Surprisingly, only 40% of the essential genes in *S.* Typhimurium LT2 (100 genes) (Knuth *et al.* 2004) are also in our list. It is worth noting that the methods to identify essential genes are different in these three studies. Our passage data also suggests that a minimal bacterium with full fitness competency and long-term survival might need a larger set of genes, one of approximately 600 genes. Our plasmid stability gene candidates have provided insightful information to prioritise future research on this field.