Transposon and Exponential Mutagenesis Approaches for Determining Essential Genomic Functions

Claire Hill

100001541

University of East Anglia (UEA)

Quadram Institute Bioscience (QIB)

In collaboration with Nanna Therapeutics

This thesis is submitted for the degree of Doctor of Philosophy

2022

Abstract

Transposon mutagenesis is an increasingly used technique in molecular microbiology, antimicrobial development, and bioproduction of compounds. However limited experimental evidence is available for the reproducibility and intrinsic biases of the technique. The reliance on annotated genomes for determination of essentiality is problematic as annotations can change rendering a gene list inaccurate. Genomic features form networks and functional redundancies can be overlooked using the technique, so methods producing double mutations in a similar fashion would provide better resolution of essential processes within a cell.

In this thesis, the reproducibility of transposon library construction was assessed by generating mutant libraries. One library is demonstrated to be reproducible when using the BioTradis pipeline, as long as a sufficient density of mutants with insertions inside coding regions is achieved. Transposon insertion frequency was found to be dependent on factors other than gene essentiality, and these should be taken into account during insertion frequency analysis. Some of these insertion biases can be measured and accounted for and insertion counts normalised to reflect this.

Also presented is an annotation-independent approach for determining essentiality using change point analysis that can accept normalised insertion data. This enables identification of essential regions that may or may not code for a known product. Finally, the development of Exponential Mutagenesis, the production of mutants containing multiple transposon insertions is demonstrated, producing preliminary data identifying known genetic interactions in metabolic pathways for folate biosynthesis; these are targeted by sulphonamides and trimethoprim, two therapeutic antimicrobials.

A high throughput paired mutant approach will lead to a better understanding of the mechanisms and redundancy involved in bacterial survival and antimicrobial resistance. Investigating the full complement of double mutants should also allow us to understand gene interactions in both metabolic and biosynthetic pathways and highlight essential genomic functions rather than gene products.

Acknowledgements

I would like to thank the MRC and the Nanna Therapeutics Ltd. for funding my PhD project. I am extremely grateful to my supervisors, Professor John Wain and Dr. Gemma Langridge for their continued support and guidance (especially Gemma's patience editing tenses). I want to thank Dr. Emma Manners for having the crazy idea for this project. I would also like to thank everyone in the groups, formerly MMRL, who have helped, listened, offered advice and who told me I should do a PhD, you win!

I am grateful to Nanna Therapeutics agreeing to collaborate and make this project happen. Everyone in the company for being so welcoming, and particularly Neil for the crash course in molecular microbiology; essential for me to be able to do this project.

I would like to thank my family for always being supportive and at the end of the phone whenever I needed them. I would also like to thank my friends, who this would have been much more difficult to write without, especially Leanne and Matt for dealing with me being stressed and dragging me away to the pub when needed.

And lastly to the fluff balls, Miso and Maggi who made sure I knew it was time to take a break and pay attention to them, the most important 'people'.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A Adenine

AMR Antimicrobial Resistance

ANOVA Analysis of Variance

bp Base Pairs

C Cytosine

CDS Coding Sequence

CFU Colony Forming Units

DDR DNA Damage Response

DEG Database of Essential Genes

DNA Deoxyribonucleic Acid

dsDNA Double Stranded DNA

EM Exponential Mutagenesis

G Guanine

GSK Glaxo Smith Klein

HMM Hidden Markov Model

IPTG Isopropyl β-D-1-thiogalactopyranoside

IR Inverted Repeat

IS Insertion Sequence

LB Lysogeny Broth

MCS Multiple Cloning Site

MDR Multi Drug Resistant

ME Mosaic End

MIC Minimum Inhibitory Concentration

MOA Mechanism of Action

MRSA Methicillin Resistant *S. aureus*

NHEJ Non-homologous End Joining

OD_{600} Optical Density at 600 nm

ORF Open Reading Frame

PBS Phosphate Buffered Saline

PCR Polymerase Chain Reaction

PEC Profiling *E. coli* Chromosome

PNK Polynucleotide Kinase

RNA Ribonucleic Acid

SME Small to Medium Enterprise

SPRI Solid Phase Reversible Immobilisation

ssDNA Single Stranded DNA

STM Signature Tagged Mutagenesis

T Thymine

TIS Transposon Insertion Sequencing

Tm Melting Temperature

Tn-Chr Transposon-Chromosome

WGS Whole Genome Sequence

WHO World Health Organisation

1 Introduction

1.1 Antimicrobials and Resistance

Antimicrobials have become an integral part of life whether that be for the treatment or prevention of infection, in humans, animals or crops. The dependence placed upon these molecules has led to widespread use, which in turn has led to a major selective pressure on bacteria, which need to adapt to tolerate the effects of these ubiquitous compounds in order to survive (Abdulmahdi et al., 2021) . Antimicrobial resistance (AMR) is in fact a current medical crisis with a conservatively estimated 700,000 deaths per year attributable to antimicrobial resistant organisms and that figure is projected to increase to 10 million by 2050 (O'Neill, 2014). Bacterial evolution occurs on a relatively short timescale and within a single cell there could be any number of inherent mechanisms to subvert antimicrobials (Aslam et al., 2018), a brief overview of clearance mechanisms in *Figure 1-1.*

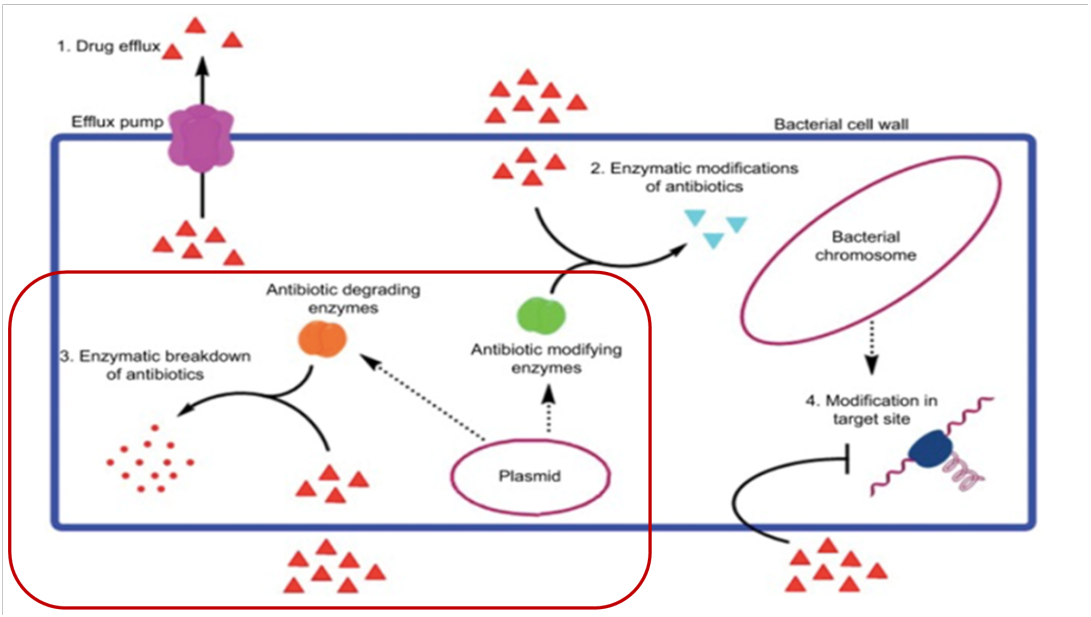

Figure 1-1 *Overview of drug clearance mechanisms in a bacterial cell.*

The processes within the red box can be either intrinsic or acquired. Image modified from (Aslam et al., 2018).

Intrinsic resistance can be attributable to any number of tactics that a cell is able to employ to mitigate the effect of an antibiotic (*Figure 1-1).* Naturally resistant phenotypes include features such as modifications in cell membranes or changes to permeability brought about by porins or efflux pumps (Aslam et al., 2018). Mutations in cellular processes that reduce the potency of an antimicrobial, therefore offering a fitness benefit, cause a selection drift towards resistance to the antimicrobial. The presence of subinhibitory concentrations of antimicrobials provides a selective pressure, increasing the evolutionary rate (Lee Ventola, 2015).

Acquired resistance mechanisms are more commonly found between bacterial species and are spread by horizonal gene transfer, so can provide multispecies reservoirs of resistance in the environment (Lerminiaux & Cameron, 2019). Acquired resistance mechanisms are more likely to be enzymatic degradation of the antimicrobial. When this degradation is of an external compound, such as β-lactamase degradation, a few organisms are able to provide resistance and lower the concentration of the drug (Saebelfeld et al., 2021).

Another communal evasion mechanism used by bacterial cells is to form extracellular matrices, or biofilms (Sharma et al., 2019). Biofilms are microbial communities that are spatially associated, the cells are bound in matrices composed of deoxyribonucleic acid (DNA), polysaccharides, and proteins. These rigidly adhere to surfaces and are ubiquitous within the environment but in the case of an infection, the ability to form a biofilm is considered a virulence factor. This can be due to the increased difficulty for drug penetration due to physical characteristics such as the barrier effect of the matrix (Singh et al., 2016), and the increased proximity of cell communities, which generates a pool of resistance genes. Additionally, the low metabolic rate of a stable biofilm is counterproductive for many antimicrobial actions.

1.1.1 The AMR Problem in Relation to Human Health

The O'Neill report is a review of the global problem of AMR, it was conducted by economist Jim O'Neill supported by the UK Government and the Wellcome Trust (O'Neill, 2014). The aim of the review was to approach the topic from a social and economic perspective and the impact of drug resistance at the global scale. The review was conducted over two years and O'Neill published his final analysis of the growing problem and proposed actions in 2016. The report states ten areas to focus on for tackling antimicrobial resistance; only one of these is the development of novel antimicrobials, the rest focus on reducing use of existing and or novel antimicrobials, improving sanitation, making antimicrobial drug discovery financially viable and alternative treatment strategies.

Pan-drug-resistance describes bacteria that are resistant to all current classes of antimicrobials currently in therapeutic use and while there are relatively few reports of such organisms, the evolution of single-, multi-, extensive- and now pan- resistant organisms is of concern and is an indication that resistance is increasing (Bhagirath et al., 2019).

1.1.1.1 Combination Therapies

In a clinical setting, acute infections, especially sepsis, require rapid antibiotic treatment and in the case of a severe infection where the causative agent is unknown it is often advised to begin treatment with two antibiotics with differing modes of action (Díaz-Martín et al., 2012). The notion of this approach is that an infective agent is less likely to develop resistance to two modes of antibiotic attack, so this offers the best defence before knowing the causative agent and its susceptibilities. However, there is limited data about the effectiveness of these approaches, what combinations are safe to use in terms of toxicity and some studies even suggest a detrimental effect (Kumar et al., 2010; Pletz et al., 2017; Safdar et al., 2004). Some report that taking this approach is an example of poor antimicrobial stewardship and is encouraging multiple resistances in pathogenic bacteria, thus decreasing the scope of treatment currently available (Pletz et al., 2017). If the best treatment plan for severe bacterial infection is controversial and enhancing resistance evolution, then this highlights the urgent need for the development of novel antimicrobials and the technology to predict potential resistance mechanisms.

1.1.2 Approaches to Antimicrobial Discovery

The solutions to the growing AMR problem are to find either compounds with novel antimicrobial properties or to refine the targets of currently used drugs. With the 'low hanging fruit' of antimicrobials having already been discovered, finding effective drugs has become more challenging and conventional discovery methods are no longer producing viable antimicrobial compounds. A range of creative approaches have been used in the search for novel antimicrobials but all fall into either searching for novel molecular structures or understanding the processes within an organism that lead to proliferation or resistance. Ultimately, the goal would be to target only those processes that are deemed essential for the specific niches inhabited during infection, but many of these are difficult to recreate under laboratory conditions (Payne et al., 2007).

1.1.2.1 Bioprospecting

Most of the antibiotics in use had previously been discovered as natural products from bacteria (Jackson et al., 2018). It has been suggested that these compounds are used by a saprophyte to outcompete in an ecological niche, however, their natural occurrence at sub lethal doses (Aslam et al., 2018) implies either an alternative function or highlights environmental adaptation and is

possibly indicative of emerging resistance. Current approaches have included attempts to culture increasingly diverse bacteria under challenging growth conditions, co-culturing bacterial species or communities and manipulation of quorum sensing signalling pathways (Peraman et al., 2021). This is to coerce the production of diverse molecules.

1.1.2.2 Molecule Optimisation

Compounds currently in clinical trials are mostly derived from chemical structures already in use as antimicrobials, meaning that there is already evidence for underlying resistance mechanisms to these compounds before they are even approved for clinical use (Aslam et al., 2018). Variations in structures may improve efficacy, albeit temporarily. But with the high risk, expense and long time required to get a compound to market as an antimicrobial and no promise of return. There is the addition of reservoirs of resistance awaiting (Payne et al., 2007), therefore antimicrobials are becoming less financially appealing for development.

For example, there is a polymyxin derivative currently in phase I clinical trials (MRX-8) that has shown activity against *Escherichia coli*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii;* three of the ESKAPE pathogens. The structure is undisclosed but *in vitro* testing in mouse models demonstrated a 2- to 4-fold reduction in the minimum inhibitory concentration (MIC) of MRX-8 comparted to polymyxin B (Lepak et al., 2020).

1.1.2.3 Functional Genomics

With the increasing availability of whole genome sequence data from numerous organisms, comparative genomics has been employed to uncover novel targets for a broad-spectrum antimicrobial. Glaxo Smith Klein (GSK) used their genomic pipeline to identify a putative set of highly conserved 'essential genes' based on *Streptococcus pneumoniae* and five other gram positive pathogens. However, due to the low return of investment over six years at GSK, it was decided to move to screening synthetic chemical libraries for novel molecular structures (Payne et al., 2007).

A successful mining approach has produced a candidate drug, which is in the pre-clinical pipeline (World Health Organisation (WHO), 2021). The gene *ftsZ* is conserved across both gram positive and gram negative organisms, this gene is required for growth and as such is a potential target with a broad spectrum of activity. There is a compound (TXA709) ready for stage I trial, targeting

ftsZ, that has shown efficacy against methicillin resistant *Staphylococcus aureus* (MRSA) (Lepak et al., 2015).

1.1.2.4 Unconventional Approaches

The approaches discussed above describe conventional approaches to identify a molecule that inactivates bacterial cells is identified. However, in the search for novel antimicrobials some investigations have turned to drugs that are already in use for other health conditions (Konreddy et al., 2019). A major advantage is that these have already been approved for use in humans and therefore can bypass much of the cost and lengthy timescale of the clinical development pipeline.

For bacterial species with known and common resistance mechanisms, prodrugs or adjuvants are being investigated. These small molecules target and inactivate resistance mechanisms for therapeutic antimicrobials that have been identified, thus making the clinically administered antimicrobial effective (Evans et al., 2019).

1.1.3 The Antimicrobial Development Pipeline

The World Health Organisation (WHO) published a report in November 2021 stating that there were currently 217 agents in clinical development and that in the time since their last report. 2017, only 12 antimicrobials had been approved. Extrapolation of this would suggest that by 2025 there could be another 12 antimicrobial therapies approved. Within this, around 80 per cent (10) belonged to classes of antibiotics with known resistance mechanisms (World Health Organisation (WHO), 2021).

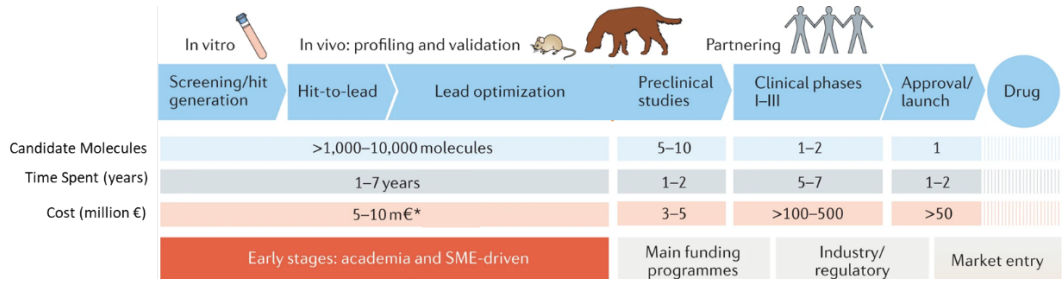

Figure 1-2 *The pipeline from hit discovery to clinical approval.*

Modified from (Miethke et al., 2021).

The WHO have suggested four attributes that make a molecule innovative; an innovative molecule indicates the potential to overcome known resistance mechanisms. The four criteria are:

1. Belong to a novel chemical class
2. Have a new bacterial target
3. Have a new mechanism of action (MOA)
4. Absence of known cross-resistance mechanisms

Only 2 of the 12 antimicrobials approved for clinical use since 2017 were considered innovative and both only fulfilled one of the criteria (World Health Organisation (WHO), 2021).

1.1.3.1 The market problem in drug development

Currently, the success rate of a compound completing clinical trials and entering the market is not a model that is profitable (Renwick et al., 2015; Silver, 2016). Even with incentives, partnership groups and funding strategies for the development of antimicrobials (GARDP, CARB-X, BARDA, NIAIS, BEAM Alliance), (Balasegaram and Piddock, 2020; Outterson et al., 2016; Sciarretta et al., 2016), the process from compound to product is costly, estimated in the hundreds of millions per compound, assuming no failures, and takes from 10-15 years (Miethke et al., 2021). Furthermore, once a product reaches the market, its use is controlled, and treatment plans are only for a brief period of time.

For the investment to be worthwhile, there must be a marketable and profitable product at the end of the pipeline, with antibiotics this is not the case. Broad spectrum treatments are more profitable but more prone to resistance evolution (Sciarretta et al., 2016). Therefore, drugs targeting lifelong medical conditions are far more lucrative for investment. As such Sanofi and AstraZeneca, along with other large drug development companies have withdrawn from the antimicrobial market (Blaskovich et al., 2017; Mahase, 2020; Tillotson & Blondeau, 2014). The majority of antimicrobials in the pre-clinical development pipeline have come from privately owned companies with fewer than 50 employees (Miethke et al., 2021; World Health Organisation (WHO), 2021). If a candidate drug fulfils more than one of the innovative criteria or there are improved methods for identifying the routes towards development of resistance or minimising evolution towards resistance, then this would lead to fewer failures during pre-clinical trials. An increased success rate during pre-clinical trials could potentially lead to a reduction in development cost and therefore a better return of investment.

1.2 Transposons in Nature

Transposable elements, transposons, were discovered as a genetic regulation mechanism in yeast in the 1940s (Comfort, 2001; McClintock, 1950) and have since been identified across prokaryotes and eukaryotes (Hayes, 2003). Transposons have been identified as molecules of evolution and participate in gene expression, inactivation, or mobilisation. Transposons are often considered selfish; for their proliferation, they must offer a fitness benefit to the target cell (Muñoz-López & García-Pérez, 2010). Therefore, it is common to see evidence of ancestral transposition events scattered throughout stable genomes that are no longer mobile (Bourque et al., 2018).

1.2.1 Insertion Elements

There is some cross over between the naming of mobile genetic elements covered in this umbrella term. A term widely used interchangeably with "transposon" is insertion sequence (IS). Generally, an IS refers to a transposase gene (*tnp)* between two inverted repeats (IRs) and the class of the IS depends on the transposase encompassed, *Figure 1-3.* ISs are found scattered throughout prokaryotic and eukaryotic genomes and are now mostly defunct. A transposon consists of two IS elements and a larger section of DNA between them(Muñoz-López & García-Pérez, 2010) . A mini transposon consists of two IRs and the section of DNA between them. This thesis describes the use of two mini-transposon systems, these will be referred to as transposons.

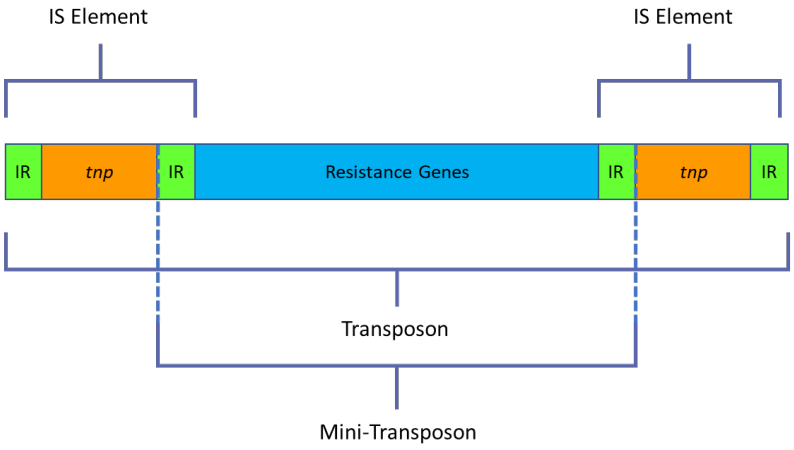

Figure 1-3 *The composition of insertion sequence elements and transposons.*

1.2.2 Transposition methods

There are two observed methods of transposition. Replicative transposition occurs via a ribonucleic acid (RNA) intermediate (Bourque et al., 2018). More commonly referred to as "copy and paste" transposons, these generate a replica of the DNA sequence to be transferred, the transposon, and then it randomly inserts into the genome, generating two copies. The alternative mechanism of transposition is conservative, or "cut and paste". Integrity of the transposon is maintained as no copies are made (no opportunity for replication errors). The two transposons used in this work (Tn*5* and the mariner family *himar*1C9), are of the latter type, and are conservatively transposed (Muñoz-López & García-Pérez, 2010).

1.2.3 The Role of Transposons in AMR

Transposons can be considered to be reservoirs of resistance genes. In the same way as plasmids, these mobile genetic elements can horizontally transfer resistance genes between organisms, and across species (Lerminiaux & Cameron, 2019). In fact, transposons carrying resistance genes to carbapenems, β-lactams, aminoglycosides and trimethoprim have been documented (Partridge et al., 2018).

The mobility of a transposon within a genome is an example of genome plasticity and therefore offers a selective advantage to a cell. Beyond the introduction of resistance genes, transposons can alter gene expression (Lipszyc et al., 2022). It has been observed that transposition rates increase amongst stressed cells, such as under antimicrobial treatment, and increases genetic variation amongst a population (Wu et al., 2015). A mutant that offers a selective advantage, such as antimicrobial tolerance or resistance, proliferates and generates a population of resistant cells via vertical transmission. This is an example of rapid evolution.

1.3 Transposons as Synthetic Genetic Tools

The insertion of a transposon within a genome causes disruption to the genetic element at the insertion site. As such, the random nature of transposon insertion has been used to generate untargeted insertion, therefore loss of function, mutants for functional genomic screens.

1.3.1 Transposon Mutagenesis

Targeted gene deletions can be used to determine the function of a gene but are dependent on an accurate annotation of the genome sequence and an underlying functional hypothesis. An alternative approach is to generate random gene knockouts using transposons that insert into the genome aberrating the function at the site of insertion. Once transposition has occurred a pool of mutants can be phenotypically screened in a negative selection process for loss of function. Generally, most studies utilising transposon mutagenesis use either the bacterially derived Tn*5* system (Barquist et al., 2013) or the eukaryotic derived *himar1* mariner family *system*.

Both Tn5 and mariner systems are cut and paste type. With one difference to note, while both are reported to have no real insertion site bias, mariner has an absolute requirement for TA nucleotides at the insertion site (Barquist et al., 2013) and this should be considered when considering an organism with a GC rich genome.

1.3.2 Signature Tagged Mutagenesis

Due to the random nature of transposon insertion, the location of the transposon within the chromosome needs to be identified. Signature tagged mutagenesis (STM) was the first iteration of combining transposon mutagenesis and DNA hybridisation (Hensel et al., 1995) to identify individual mutants surviving the negative selective process. Each transposon is barcoded with a known DNA sequence, a pool of mutants is generated, and the selective pressure is applied. Barcodes from both the input and output pool are amplified and compared to identify the barcodes that have been lost from the pool. These barcodes represented a transposon insertion in a region required for infection. However, the barcodes of interest need to be translated back to a genomic region or gene using DNA hybridisation methods. Hensel et al. developed the protocol while using transposon mutagenesis to identify virulence genes in *Salmonella typhimurium;* genes absolutely required for infection but not necessarily survival make ideal candidates for antimicrobials as there is less of a selection pressure outside of the infection niche. As transposon mutagenesis experiments have become more commonplace, STM has been used as an umbrella term when describing hybridisation assays detecting transposon insertion sites.

A more sophisticated transposon site hybridisation (TraSH) approach has been used to identify conditionally essential genes in *Mycobacteria bovis* (Sassetti et al., 2001). Briefly, an array was constructed using complimentary DNA for each one of the annotated open reading frames

(ORFs) in the organism under investigation. The RNA is then radio-labelled and hybridises to the corresponding ORF, determining the location of the chromosome. Using this method, more than one condition can be assayed simultaneously by labelling each with a fluorophore emitting at a different wavelength, so the difference in the fitness contribution of the genomic region across conditions can be achieved by comparing ratios of emittance. The advantage of TraSH over STM is that the DNA is hybridised to an ORF, so the genome location of the transposon is immediately known without needing further analysis.

1.3.3 Transposon Insertion Sequencing

Transposon Mediated Differential Hybridisation (TMDH) was used by Chaudhuri et al. to further refine STM methods, they used transposon specific primers to amplify the genomic region flanking the transposon insertion site. The PCR products were sequenced with capillary DNA sequencing, therefore only the genomic regions that did not have a transposon insertion, determined by the hybridisation assay were sequenced (Chaudhuri et al., 2009). This study identified 351 essential genes in *Staphylococcus aureus*.

The throughput of transposon mutant experiments has developed alongside sequencing capabilities. As high throughput sequencing became accessible, transposon mutagenesis experiments were combined with the sequencing technology rather than hybridisation assays (Chao et al., 2016). Historically there have been protocols for transposon insertion localisation for the most commonly available instruments: 454 pyrosequencing (Bronner et al., 2016), Ion torrent (Perry et al., 2016), and Solexa (Z. Xu et al., 2017). Today, most experiments use Illumina sequencing. However, as nanopore sequencing is becoming ubiquitous, there have been approaches developed to use this technology for high throughput transposon insertion sequencing (Baltrus et al., 2019; Lott et al., 2022; Yasir et al., 2022), the advantage of long read sequencing is enhanced mapping, especially around large repeat regions in the genome.

Four groups simultaneously developed protocols for combining transposon mutagenesis experiments with high throughput sequencing technologies. (Gawronski et al., 2009; Goodman et al., 2009; Langridge et al., 2009; Van Opijnen et al., 2009). All following a similar methodology and will collectively be referred to as Transposon Insertion Sequencing (TIS). Briefly, the general workflow is to:

1. Perform a large-scale mutant generation and pool the mutants as a library of single gene knockouts, *Figure 1-4A.*
2. Allow permissive and selective growth*, Figure 1-4B.*
3. Extract genomic DNA from the input pool and the surviving mutants.
4. Enrich the Transposon – Chromosome (Tn-Chr) junction.
5. Sequence the amplicons.
6. Use the chromosomal portion of the sequenced amplicons to map to an organism specific reference genome, to locate the transposon insertion sites.
7. Genes lacking mapped insertions are not represented in the pool, meaning that the organism is unable to proliferate under the tested condition; this suggests that the gene is required for survival. This provides candidate essential genes, *Figure 1-4C.*
8. Optionally, compare the relative transposon insertion rate between permissive and selective growth to identify genes of interest. For example, genes required for infection, these genes would be considered conditionally essential, *Figure 1-4C.*

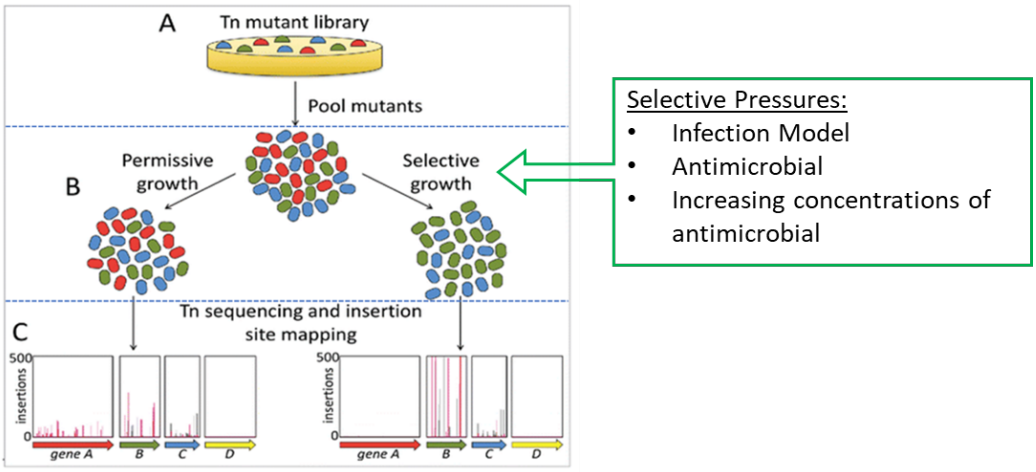

Figure 1-4 *Overview of the transposon insertion sequencing process.*

Figure modified from (Paulsen et al., 2017)

The addition of massively parallel sequencing means that with one large transposon mutagenesis experiment there is a pool of mutants generated. This pool can then be applied to any range of selective pressures to identify the genes absolutely required for survival under a range of conditions, such as antimicrobial treatment or in infection models. (Barquist et al., 2016; van Opijnen & Camilli, 2013). The simplest of applications is to determine which genes are essential for growth under standard laboratory conditions. The four groups that pioneered the approach were not only investigating genes in terms of essentiality for growth but wanted to gain an insight into a specific virulence factor of the organism under investigation.

1.3.4 TIS in Drug Discovery

TIS is an invaluable tool that can be used to identify potential drug targets, the MOA, and any potential resistance mechanisms (Coward et al., 2020; Holden et al., 2021; Sargison & Fitzgerald, 2021). The assumption is that the relative number of transposon insertion events is directly proportional to the contribution of a gene to the fitness of the organism. This enables genes to be sorted into the following categories:

1. Neutral – disruption of these genes offers neither an advantage or disadvantage to survival and the number of insertions is consistent across the test conditions, *Figure 1-4C*, gene C.
2. Essential – disruption of these genes results in non-viability under the test conditions, *Figure 1-4C*, gene D.
3. Detrimental – disruption of these genes results in a reduced fitness under the test conditions, *Figure 1-4C*, gene A.
4. Advantageous – disruption of these genes results in increased fitness under the test conditions, *Figure 1-4,* gene B.

In the context of drug discovery, *Figure 1-5,* essential genes are used to identify candidate drug targets for novel compounds (Meredith et al., 2012). If screening for a bacterial response to a candidate compound, the advantagous category can highlight the genes that a mutant pool uses to subvert antimicrobial action. Therefore, any novel compounds identified can be screened for the likelihood of resistance development and the mechanisms of this resistance.

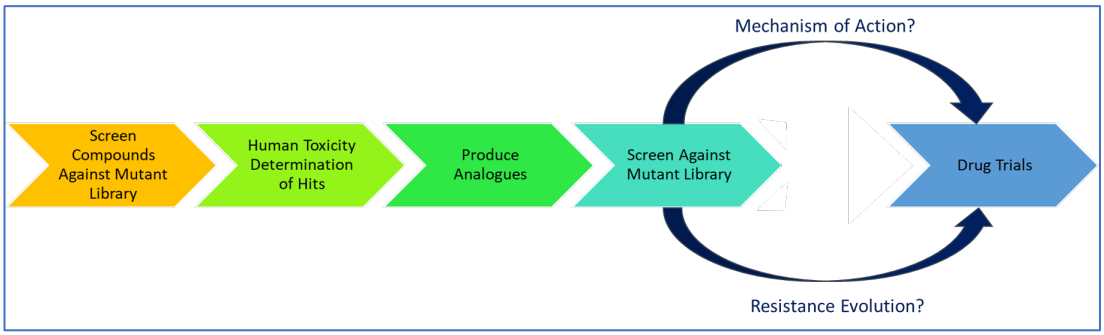

Figure 1-5 *An example drug discovery pipeline utilising transposon insertion sequencing.*

An example of a drug discovery pipeline utilising TIS is given in *Figure 1-5.* This patented approach (US 2015/0307873 A1 and US 2017 / 0342460 A1) has been successfully used to identify compounds with a novel mechanism of action against MDR *Neisseria gonorrhoeae* (Jacobsson et al., 2020). A promising novel antimicrobial may be incredibly effective against its target but if it is expected to rapidly evolve resistance or there are known cross resistance mechanisms, then the time and expense of clinical trials will not yield a clinically useful drug. Hence the WHO innovative criteria, *section 1.1.3.* A Leucyl-tRNA Synthetase Inhibitor GSK2251052 for gram negative infections was suspended in Phase II trials due to more than 20 percent of the urinary tract infection causative agents (*E. coli*) developing resistance to the novel compound (O'Dwyer et al., 2015). However, this does not mean that a potent antimicrobial is not worth pursuing, just that it may have restricted use or be reserved for use in extreme life-threatening infections. As discussed in *section 1.1.3.1*, a drug with a restricted market is not financially appealing to develop. Therefore, early MOA and resistance mechanism screens can eliminate hits prior to significant investment. Once the platform has been established, TIS screens are rapid and inexpensive.

1.4 Gene Essentiality

1.4.1 Absolutely Essential Genes

For any organism there is a minimum number of cell processes that are required for a cell to survive. An essential gene is defined as a gene that is absolutely required for an organism to proliferate; it can therefore be assumed that these genes carryout essential functions within the organism(Christen et al., 2011; Davis et al., 2008). If these genes or gene functions were to be removed, there would be a fitness cost, potentially leading to cell death. Such genes are difficult to validate experimentally as the cell cannot be cultivated, due to this, often genes essential for growth of an organism are also deemed essential for survival. In functional genetic screens, if a target gene is disrupted and the mutant is non-cultivatable under standard laboratory growth conditions, then it is classified as an essential gene.

1.4.2 Conditionally Essential Genes

While a gene product may be dispensable when a cell is grown under standard laboratory conditions, once a stress has been applied, a non-essential gene, or function of may be required for survival. The stresses applied for the context of drug development could be the selective

pressure of an antimicrobial or the nutrient limitation of an infection niche. Genes that are required under stress conditions but not standard-permissive growth, are considered conditionally essential (Joyce et al., 2006; Zhao et al., 2017).

Uncovering conditionally essential genes is particularly important for the development of novel antimicrobials. Conditionally essential genes identified during infection or host colonisation would be ideal novel antimicrobial targets (Peraman et al., 2021; Silver, 2016). The general notion is that any essential function would make a good antimicrobial target, however, targeting an absolutely essential gene would target the entire population of cells, adding a constant selective pressure and encouraging resistance (Aslam et al., 2018). Therefore, targeting only the genes involved in pathogenesis would reduce the selection pressure applied and lead to reduced rate of evolution towards resistance. Ultimately, the goal would be to target only those processes that are deemed essential during infection, many of these conditions are difficult to recreate under laboratory conditions.

1.4.3 Essentiality as a Metabolic Function

Genes and gene products do not operate as discrete units and most contribute to a metabolic pathways or reaction networks leading to measurable functions. Therefore, rather than looking at the individual essential genes constituting one step in a pathway, it is important to consider that a cell requires a set of essential metabolic processes rather than the specific genes involved. Understanding the metabolic pathways where a gene is implicated in can add depth to MOA identification or resistance evolution during a TIS screen.

1.4.3.1 Metabolic Models

Genetic maps of metabolism are being replaced by network maps. Reactions and substrates can be considered as absolutely required rather than gene products (Costanzo et al., 2006). One upcoming area of research, genome scale metabolic reconstruction, is trying to generate accurate *in silico* models of the metabolic reactions that an organism is capable of. By creating these models, different stresses can be placed on the model to investigate how it will respond, a useful tool referred as constraint based metabolic modelling. Models are constructed based on genomic sequence data and a group have developed a toolkit, Tn-Core, that links metabolic models to TIS data (Dicenzo et al., 2017). Highlighting indispensable genes for growth under a certain condition and linking them to metabolism, then utilising this information to guide novel

antimicrobial development would provide the best representative screen of conditions that are not achievable in a laboratory. This would enable replication of infection niches *in silico*. This is the first incidence, to our knowledge, of a pipeline to directly import TIS data into a metabolic model.

1.5 Available Resources for Determining Essentiality in *E. coli*

1.5.1 Profiling *E. coli* Chromosome Database

Genes identified from essentiality studies in *E. coli* have been collated as a database, the Profiling *E. coli* Chromosome (PEC) database (Yamazaki et al., 2008). Importantly, it contains information on the essentiality of genes, and each gene is classified as essential, non-essential (dispensable) and unknown. While the designation of essential is unambiguous, it is important to note that the definition for the database is in relation to cell growth and not necessarily survival, if a mutant were able to survive but replication was hindered, then it would be outcompeted and then considered an essential gene.

Within the database, conditionally essential genes are listed as essential. As per any collated databases, there are exceptions to the rules and particularly when there is a lack of experimental data available to reference. For example, the genes encoding ribosomal proteins are classed as essential unless there is evidence supporting that they are dispensable (Yamazaki et al., 2008).

1.5.1.1 Database of Essential Genes

The Database of essential genes (DEG) list studies that have investigated gene essentiality by organism. Notably, there has been a recent update (Luo et al., 2021; R. Zhang et al., 2004) which now includes some online analysis tools to give information on whether genes are on the lagging or leading strand, with essential genes being more often on the leading strand (Luo et al., 2021). There are also tools to identify processes, or functions that the gene products are involved in, and the metabolic pathways implicated. Due to the increased use of transposon insertion mutagenesis approaches, there are twice as many prokaryotic essential genomes listed than the previous iteration in 2014 (Luo et al., 2014). There are two differing studies for essential genes in *E. coli* MG1655. One reports 609 essential genes highlighted by transposon insertion and *in situ* hybridisation (Gerdes et al., 2003). The second reports 296 essential genes that were identified by knockout mutants (Baba et al., 2006). There have not been any additions since high

throughput transposon insertion sequencing has been widely accepted, so the discrepancy in the number of reported essential genes for this organism remains when using DEG.

1.5.2 A Defined Plasmid Library

Another major molecular tool available for use in studying *E. coli* is The ASKA library of clones (Kitagawa et al., 2006). This is a collection of *E. coli* K12 ORFs (excluding start and stop codons) cloned into a multi copy plasmid. This resource enables the functional analysis of any ORF to help allude to a function for unknown genes and to be used as to assay the function once known. To do this, the ORF is Histidine tagged at the N terminal and green fluorescent protein tagged at the C terminus to generate a fusion protein. The fluorescence is used as an indicator of successful cloning and transcription. These clones have been used in DNA microarray construction (Oshima et al., 2002), protein production and functional analysis of ORFs (Awano et al., 2005). These clones have been invaluable for studies investigating genetic interactions and have been used in combination with the Keio collection of knockout mutants (Baba et al., 2006; Butland et al., 2008; Typas et al., 2008).

1.5.3 Targeted Knockout Mutant Libraries

In 2006, prior to TIS, a group generated and catalogued a library of single gene knockouts in *E. coli* K12 strain BW25113. This collection is provided as an ordered knockout array with every non-essential gene systematically replaced by a kanamycin resistance cassette and comprises of 3985 deletions, duplicated as separate insertions. Flanking recognition sites enable the cassette to be removed using flippase, to generate in frame knockouts (Baba et al., 2006). This collection of mutants is referred to as the Keio collection and is used to validate TIS functional screens (Goodall et al., 2018; Holden et al., 2021; Yasir et al., 2020)

The whole genome sequence of *E. coli* BW25113 became available in 2014 (Salama et al., 2004) but this collection was generated in 2006 so construction of the Keio collection was based on an annotation and similarities to two other K12 species – MG1655 (4.5Mb) and W3110 (2.6Mb) (Riley et al., 2006). Despite the two very different genome sizes, they were combined to generate a composite K12 genome containing 4453 genes with 4296 ORFs (Baba et al., 2006); which was regarded as the most accurate genome of any organism at the time.

One issue addressed with targeted mutagenesis, rather than with random insertion, is that there can be an overlap with genes in ORFs, so knockouts can impact more than one gene, leading to

the classification of a gene as essential when in fact the lethality is caused by the other interrupted gene. At the time of construction of the Keio collection, there were 742 overlapping genes of up to 260 nucleotides. For example, *folC* has an 11 nucleotide overlap with the conserved *dedD* protein. As such, *dedD* was mistakenly classified as essential due to alteration of the C-terminus of *folC,* which is an essential gene (Baba et al., 2006).

1.6 Variability in Classifying Gene Essentiality

In TIS, the essentiality of genes is determined depending on the number of transposon insertions. In a mutant library pool, repeated insertions provide biological replicates of gene knock out events, therefore, computational determination algorithms use statistical approaches to designate essential genes (Barquist et al., 2013; Chao et al., 2016). Using software to analyse gene essentiality provides methodological consistency for repeated experiments. However, there is not one standardised statistical approach used for this, and as such, a number of approaches have been developed, some of the accessible tools are described in *Table 1-1.*

Genome annotation is particularly important when considering essentiality of a gene or region of the genome, and some of the available tools take annotation into account. Automated analysis can lead to an underestimation in the number of essential genes (Larivière et al., 2021). If only part of the protein coding sequence is essential, then it would be feasible for there to be a number of transposon insertions detected in any other region of the gene. This is most often seen in terminal regions of genes and some analyses exclude a portion of the termini; others do take the terminal regions into account so may determine that such genes would not be essential as they contain sufficient transposon insertions within the coding sequence (CDS) (Goodman et al., 2009; Van Opijnen et al., 2009).

Some of the available tools have been developed for use with mariner transposon sequencing data, these tools use the number of insertions in TA sites to determine essentiality rather than using gene length. Therefore, their use is limited by the transposon family used (DeJesus et al., 2015; Mccoy et al., 2017; Van Opijnen et al., 2009). Due to the nature of DNA condensation, there may be areas where it is structurally unachievable for a transposon to insert, this region may serve no purpose or be essential in any way but statistical analyses within this region would determine that the number of insertions is below the cut off and be deemed essential (Goodall et al., 2018). This phenomenon would occur for all of the tools listed in *Table 1-1.*

The range of tools available can be overwhelming (Van Opijnen et al., 2015) for the analysis of TIS and with different essential genes reported between studies lead to confusion and wasted investments during drug discovery and optimisation. Further descriptions of TIS analysis and variability are provided in *chapters 3 and 5.*

Table 1-1 *A summary of publically available transposon insertion sequencing gene essentiality determination tools.*

TOOL	STATISTICAL APPROACH	ANNOATION	TRANSPOSON FAMILY	REFERENCE
BIOTRADIS	Negative Binomial Distribution	Dependent	Any	(Barquist et al., 2016; Langridge et al., 2009)
ESSENTIALS	Negative Binomial Distribution	Dependent	Any	(Zomer et al., 2012)
TRANSIT	Gumbel and Hidden Markov Model	Independent	Mariner	(DeJesus et al., 2015)
ARTIST	Hidden Markov Model	Independent	Any	(Pritchard et al., 2014)
TN-SEQ EXPLORER	Fit to Exponential function within Sliding windows	Independent	Any	(Solaimanpour et al., 2015)
MAGENTA	Comparison of Standard Deviations	Dependent	Mariner	(Mccoy et al., 2017)
ALBA TRADIS	Negative Binomial Distribution (BioTraDIS) Log fold change	Dependent	Any	(Page et al., 2020)
LORTIS	Negative Binomial Distribution (BioTraDIS) - Long read input data	Dependent	Any	(Lott et al., 2022; Yasir et al., 2022)

1.6.1 Classifying conditionally Essential genes

While conditional essentiality is not explored within this work, a number of bioinformatic pipelines can be used to identify these. The BioTradis pipeline (Barquist et al., 2016; Langridge et al., 2009) identifies essential genes but can also compare transposon insertion profiles of genes under different experimental conditions. A Log_2 fold change between conditions is used to determine conditionally essential genes. This has been developed into the standalone tool AlbaTRaDIS for assessing multiple conditions (Page et al., 2020). Other available tools also use the log fold change approach to determine conditional essentiality, TnSeq-Diff and MAGenTA (Mccoy et al., 2017; Zhao et al., 2017).

1.7 Genomic Redundancy

1.7.1 Redundancy as a Genetic Concept

High throughput functional genetic screens using TIS are successfully used to identify essential and conditionally essential genes. However, there is plasticity within genomes, and essential cellular functions may be rescued by an alternative gene product or reaction pathway, so while essential genes can be targets for antibiotics (Juhas et al., 2012), understanding the interaction of genes may offer more robust targets. Redundancy can be defined as the ability of an organism to compensate for the loss of a genetic function in order to achieve the same essential outcome, even at a fitness cost.

If survival of a cell is considered as a series of interconnected essential functions determined by reaction pathways; it is conceivable that within such a network of reactions, there would be more than one possible pathway (Mahadevan & Lovley, 2008). This is the concept of redundant genetic interactions, that if one pathway it obliterated then there is an alternative that may be able to compensate for the loss(Láruson et al., 2020).

While it has become almost routine to use a saturated transposon insertion library to determine the essentiality of a gene for survival under a range of conditions, it is well documented that genes do not operate in isolation (Mani et al., 2008; Phillips, 2008; Van Opijnen et al., 2009) and that within a genome there are countless interactions. While there has been some research, most studies have been looking at the interactions of eukaryotic genomes, namely *Saccharomyces cerevisiae.* A genetic interaction can be described as any case where the effect of a double mutant is not equal to the compounding effects of the two single mutants, and it is observed as a measurable phenotype, for

example fitness (Mani et al., 2008). One area lacking in exploration is analysis of genetic interactions within prokaryotic genomes, with a small number of groups trying to uncover such networks. However, due to the massive number of mutants required, it is difficult for a genome wide interaction study to be undertaken. Three low throughput approaches that have been taken are summarised below.

1.7.2 GIANT coli Approach

Using *E. coli* as a model organism, Typas et al. report that in 2008 (1 year before transposon insertion sequencing studies published) there were 200 gene interactions reported for *E. coli*, 10,000 times fewer than were for *Saccharomyces cerevisiae* (Typas et al., 2008) . They developed a medium throughput method for generating double mutants deemed GIANT-coli, using two well defined libraries of single mutants of around 4,000 single mutants. Double knockouts were obtained by conjugation And the Keio collection as the donor strains and the ASKA clones as the recipient, or vice versa, to screen for genetic interaction (Baba et al., 2006; Kitagawa et al., 2006).

1.7.3 *E. coli* Synthetic Genetic Array

At the same time, a second group that were considering genetic interactions in *E. coli*, published in the same issue as the study described above. The second method developed was *E. coli* synthetic Genetic Array (eSGA) screening (Butland et al., 2008). The aim was to automate as much of the screening process as possible, in order to investigate functional crosstalk and genetic interactions among a gene of interest and any other non-essential gene, using *E. coli* as a model organism. Using bacterial conjugations and recombination, a donor strain with the queried gene knocked out was recombined with each of the Keio collection mutants and screened for fitness change compared to wild type. This was one of the first methods developed to undertake genome wide gene interaction studies, however, only one query gene could be screened at a time. This method focusses on one specific gene and while useful for pinpointing the function of that particular gene, this method does not provide the potential depth of information that can be generated with the proposed Exponential Mutagenesis methods.

1.7.4 Transposon Insertion

A decade before either of the above approaches were used to investigate genetic interactions, Elena & Lenski used a mini-Tn10 (transposon) system to generate mutant pools of one, two or three random gene knockout mutants (Elena & Lenski, 1997). They compared the relative fitness of the mutants against the fitness of the wild-type and found there to be no significant interactions. To refine their methods, the next experiment was to generate targeted knockout mutants, three genes of varying fitness detriments across 3 genomes were combined to give a total of 27 double knockout variants. The advantage of this method was that the genes could be evaluated for fitness detriment as a single knockout before being combined. When comparing this set of mutants, they found 7 gene combinations to have a synergistic effect and 7 to have an antagonistic effect. This rationalised the earlier negative result, and they concluded that the previous experiment failed to give any significant results not because interactions are rare but because when analysed as a pool, the synergistic and antagonistic interactions negate each other to give no measurable difference.

1.7.5 Stress Networks

Any potential antimicrobial targets are going to be more lucrative when the organism is stressed or constrained within the niche of infection. One of the major targets of current antibiotics is the DNA damage response (DDR). One example of this is ciprofloxacin use, where DNA gyrase and topoisomerase IV are the targets (Kumar et al., 2016), an example of functional redundancy. Despite being an effective target, there is still evidence for emerging resistance, and much of the functionality and interaction of the genes involved are still unknown (Fasugba et al., 2015). An extension to the eSGA method described above was developed to investigate genetic interactions among 398 genes whose annotation suggest involvement in DDR, and a further 151 that are predicted to be involved but not assigned to a specific pathway (Kumar et al., 2016). A knockout mutant was generated in the query genes and recombined with 526 of the Keio mutants. Fitness was determined by colony size, digitised, and assigned a score indicating the relative fitness compared to wild type. Generally, a positive score implied parallel pathways and a negative score was given when the two genes were in a linear pathway. A parallel pathway indicates redundancy whereas a linear pathway indicates a directional process.

Using methyl methanesulfonate, a DNA alkylating agent provided stress, as a treatment when the single knockout mutants were of genes involved in DNA repair, recombination, cell division or a combination, resulted in the mutant being hyper-sensitive to antibiotics inhibiting protein synthesis such as trimethoprim. Exposing the double mutants to stress, Kumar et al found that some interactions were only observed under stress (Kumar et al., 2010). This is a particularly important finding because understanding redundancies within a network will better enable appropriate antibiotic choices. Furthermore, observing the physiological changes under stress provides a more accurate representation of the capabilities of an organism and the accessible resistance mechanisms. Under the stressed condition there were 373 pairs of genes where the fitness of the cell was significantly decreased or increased as a result of lost genetic interactions, demonstrating that there are compensatory changes within the DNA damage response networks (Kumar et al., 2016) and it can be concluded that to combat resistance, antibiotics targeting the genes involved in DDR may be used as adjuvants or helper drug in combination with other drugs.

1.7.6 Secondary Resistome

The secondary resistome is the term used to describe genes essential for growth in the presence of an antimicrobial, particularly where the organism shows a level of tolerance rather than enzymatic degradation of the antimicrobial. Using a dense TIS library averaging an insert every 19 bp in a strain of *Klebsiella pneumoniae*, an opportunistic pathogen often resistance to multiple drug classes(O'Neill, 2014). There were 35 novel genes identified as conditionally essential for growth in colistin, one for ciprofloxacin and one for imipenem, in addition to genes already identified as determinants of resistance. While it seems that Colistin resistance is a long way off, only small mutation would be required before this multi-drug resistant (MDR) strain is resistant to ciprofloxacin or imipenem (Jana et al., 2017).

Known genes conferring resistance to the antimicrobial under investigation having no insertions validated that the method worked (*gyrA* and *nrdA* were used as controls*)* and a previously unreported gene, *dedA,* was implicated. Thus, resistance identified this gene as essential for survival in colistin, indicated by a 512-fold decrease in insertions when the therapeutic concentration of colistin is used.

A targeted deletion of *dedA* rendered *K. pneumoniae* susceptible to colistin as evidenced by a 16-fold reduction of the MIC from 8 to 0.5 µg/mL, with resistance being restored by

complementation. This demonstrates the use of TIS to identify novel targets and presents *dedA* as a potential target for a helper drug to restore colistin sensitivity. This positive result was not mirrored when the imipenem pathway was investigated. The authors note that this could be due to the fact that the therapeutic dose used was much lower than the MIC and therefore did not simulate the stressed condition where the secondary resistance genes would be required. This is supported by the MIC remaining at 16 ug/mL even after *nhaA* inactivation (Jana et al., 2017). Reportedly, *nha*A encodes a pH dependent Na+/H+ antiporter that is active under stress (Berlyn, 1998), confirmation that the MIC dose used was too low or indicative of a secondary resistance mechanism.

1.7.7 Synthetic Lethality

Synthetic lethality describes two non-essential gene products where a knockout of both would result in lethality for the organism. These pairs are most often determined *in silico* using genome scale metabolic models and altering fluxes through non-conventional pathways to simulate stressed environments (Aziz et al., 2015; Ellis et al., 2009). There is a degree of plasticity within the genome and disruptions to metabolic pathways are overcome by redirecting fluxes. If a second disruption occurs in the rerouted pathway, then this may lead to synthetic lethality.

It is important to note any lethal pairs require confirmation by creating the respective mutants. This may not always be achievable and the pairs may have been identified using non-culturable conditions (Aziz et al., 2015). Additionally, this method may fail to identify redundancies that cause a significant detriment to growth, which if done experimentally would be classified as lethal. The transposon mutagenesis approach can also identify genomic regions that contribute to a phenotype, for example a protein domain that is essential for function rather than assessing genes as discrete genomic units (Benstead-Hume et al., 2019; B. Li et al., 2011). Due to the reliance of this method on a curated GSM, the number of organisms that this approach can be used on is a small. This is expanding, but currently it is limited to only those with extensive prior genome research, as a model requires genes and their functions to be known. Identifying pathways that are relevant for antibiotic resistance or tolerance depends on known mechanisms and accurate genome annotation.

Redundancy is costly to an organism, however, in an evolution experiment it appears that single gene deletions increase network resilience across the whole organism when

randomly removed *in silico*. Resilience is determined as how the network of cellular processes adapts to being fragmented and nodes removed (Maddamsetti, 2021). This suggests that as an organism evolves under selection, the redundancy within its genome is increased to ensure survival despite the fitness costs.

1.8 Project Aims

1.8.1 What is Exponential Mutagenesis?

Exponential Mutagenesis (EM) is a proposed new method for generating paired double knockout mutants using the high throughput nature of TraDIS (Langridge et al., 2009).

While the studies described above (Butland et al., 2008; Elena & Lenski, 1997; Typas et al., 2008) were able to assay a number of paired knockouts, there was always the constraint that in order to make the experiment feasible, there had to be a query gene to focus on, otherwise the generation of a double knockout library gets beyond the limit of practicality. An overview of the EM workflow is shown in *Figure 1-6*.

The intention is that with a high throughput method for paired mutant generation there will be a significant advance in the construction of genetic maps uncovering novel targets for antimicrobials or highlighting pathways that can be manipulated to produce novel bioactive compounds.

While assessing gene essentiality under a range of experimental conditions remains relevant, particularly in respect to antimicrobial resistance, my project aims to investigate the effect of removing more than one gene per cell in an organism pool to investigate functional redundancy, EM. Essential cellular functions may be rescued by an alternative gene product or reaction pathway, so while essential genes can be targets for antibiotics, understanding the interaction of genes may offer more robust targets.

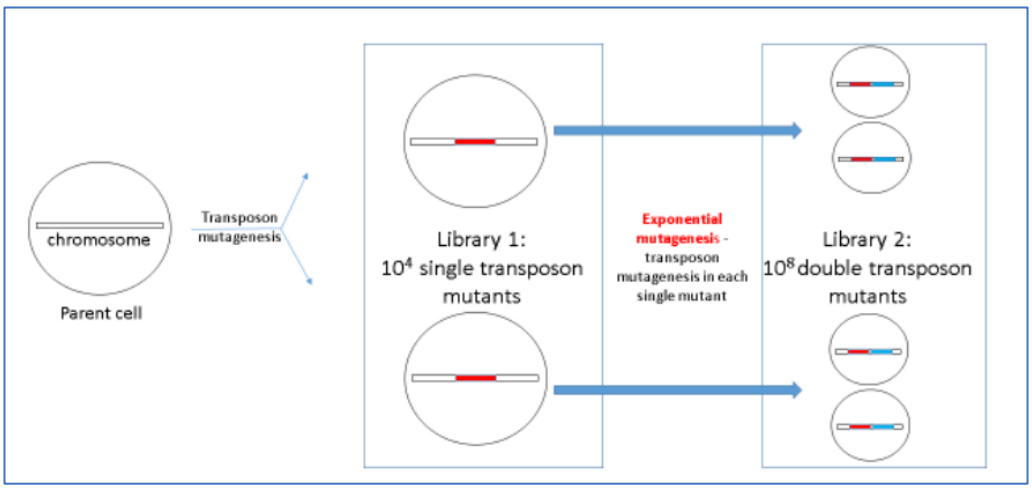

Figure 1-6 *A summary of the workflow of exponential mutagenesis.*

Using the *E. coli* genome and the estimated 4200 genes (Baba et al., 2006), it would require 8.8 million individual mutants to represent each gene inactivated as a pair. This minimum number of mutants is a conservative target, in reality, a library would need to be denser than one insertion in every gene and the initial plan is to generate 10^4 insertions of Transposon 1 and follow with 10^4 insertions of Transposon 2 within the same chromosome. This would generate pool of 10^8 mutants with unique insertion sites, with two transposons in one chromosome (Manners et al., 2018).

1.8.2 *E. coli* as the experimental organism

E. coli was chosen as the target organism due to the ease of propagation and genetic manipulation. *E. coli* has been extensively used as a model organism and is much more receptive to manipulation than gram positive bacteria and many other gram negative bacteria. Due to its extensive use over decades, there is a wealth of protocols for molecular manipulation, cloning and growth.

The target organism for the development of EM was chosen as *E. coli* BW25113, the parent strain for the Keio collection (Baba et al., 2006), *section 1.5.3.* Due to the knockout collection being a widely distributed resource, there have been numerous studies into the functions and biological relevance of the individually knocked out genes (Goodall et al., 2018; Grenier et al., 2014; Holden et al., 2021; Marzan et al., 2017; Yamamoto et al., 2009; Yasir et al., 2020). This means that there is already a vast amount of information relating to this strain published, there is the complete genome sequence in NCBI (Grenier et al.,

2014a), there have been numerous whole genome annotations and due to the relevance of the collection (Baba et al., 2006; Keseler et al., 2017; Riley et al., 2006; Yamamoto et al., 2009), the published resources are maintained and updated as more research is conducted. This makes BW25113 an ideal strain to use for the development of a novel technique to explore genotype-phenotype relationships where findings can be validated with the extensive resources available.

1.8.3 Gaps in drug discovery that this project addresses

This project aims to develop a molecular method that can be used to identify areas of the genome that contribute to bacterial pathogenesis. It is hoped that the method will be beneficial in the context of drug development by identifying mechanisms involved in antimicrobial resistance, particularly when screening new molecules with potentially novel targets. This would filter out drug candidates that are likely to fail pre-clinical trials before there has been a significant investment both financially and with time. Currently, there is not a high throughput genome wide genotype - phenotype assay that is used to identify redundant pathways in the development of antimicrobial resistance. Being based on existing technologies already in use, it will be transferrable across a range of bacterial species.

2 Materials and Methods

2.1 Bacterial Strains Used

Table 2-1 *A description of the E. coli strains used in this study.*

Strain	Referred to as	Used For	Reference
E. coli BW25113 [Δ(araD-araB)*567* Δ(rhaD-rhaB)*568 Δ*lacZ*4787* (:*rrnB*-3) hsdR*514 rph-1*]	BW25113	The background strain used throughout this work for generating transposon mutant libraries and exponential mutant libraries in this work.	(Baba et al., 2006; Datsenko & Wanner, 2000; Grenier et al., 2014a)
E. coli ST18 S17 λpirΔhemA	ST18	A plasmid donor strain used to develop conjugation protocols.	(S. A. Jackson et al., 2020; Thoma & Schobert, 2009)
E. coli MFD*pir+* MG1655 RP4-2-Tc: [Δ*Mu1*:*aac* (3) IV-Δ*aphA*-Δ*nic*35-Δ*Mu2*: *zeo*] *ΔdapA*: :(*erm-pir*) Δ*recA*	MFDpir+	The plasmid donor strain used to generate the transposon mutagenesis libraries and exponential mutagenesis libraries in this work.	(Ferrières et al., 2010; S. A. Jackson et al., 2020)

All strains of *Escherichia coli* ,*Table 2-1,* were grown in LB Miller broth (Tryptone 10 g/L; Yeast Extract 5 g/L; Sodium Chloride 10 g/L; Formedium Ltd., Swaffham, UK) at 37 °C, shaking at 200 rpm or on LB Miller Agar (Tryptone 10 g/L; Yeast Extract 5 g/L; Sodium Chloride 10 g/L; Agar 11 g/L; Formedium Ltd., Swaffham, UK) at 37 °C, unless stated otherwise. Depending on the plasmid used or stage of library production, antibiotics were added to select for donor strains or mutants, see *Table 2-2***.**

Table 2-2 *Selective compounds and additives used in growth media.*

Compound	Stock Concentration	Working Concentration	Purpose of Use
5-aminolevulinic acid hydrochloride[1] (ALA)	50 mg/mL	50 µg/mL	Required for cultivation of *E. coli* ST18.
2,6-Diaminopimelic acid[2] (DAP)	300 mM	0.3 mM	Required for cultivation of *E. coli* MFD*pir+.*
D-Glucose[1]	30 % (w/v)	0.3 % (w/v)	Added to the growth medium to repress the second transposition during exponential mutagenesis.
Ampicillin Sodium[2]	100 mg/mL	100 µg/mL	Used to select for the backbone of all plasmids used during this work.
Carbenicillin Disodium[2]	100 mg/mL	100 µg/mL	Used for the same applications as ampicillin, but more stable during prolonged incubation.
Gentamicin Sulphate[2]	10 mg/mL	8 µg/mL	Used to select for Tn*5* transposition when using pBAMD1-6. Also used to select for the second transposition event during EM.
Kanamycin Monosulphate[2]	50 mg/mL	50 µg/mL	Used to select for transposition when using pSAM_Ec or pBAMD1-2. Also used to select for the first transposition event during EM.

[1]Thermo Scientific, Waltham; USA [2]Formedium Ltd., Swaffham, UK

2.2 Plasmids Used

The plasmids used in this work are briefly described below. These were either: already present in the laboratory (pIMAY), a gift from researchers that have deposited their work in the plasmid repository Addgene or constructed for this project.

2.2.1 pIMAY

pIMAY (plasmid # 68939, Addgene, gifted by Tim Foster), *Figure 2-1:Left*, was used previously in the group and had been modified for transposon mutagenesis prior to this project. Dr Emma Manners cloned a Tn*5* transposon system, *Figure 2-1:Right*, and a mariner transposon system into this vector. It is a temperature sensitive plasmid for allelic exchange from *E. coli* into staphylococci. The backbone of the plasmid has a chloramphenicol resistance cassette (*cat*) and a tetracycline inducible counterselection system, for tetracycline-induced expression of Anti-*secY* leading to cell death when expressed (Monk et al., 2012).

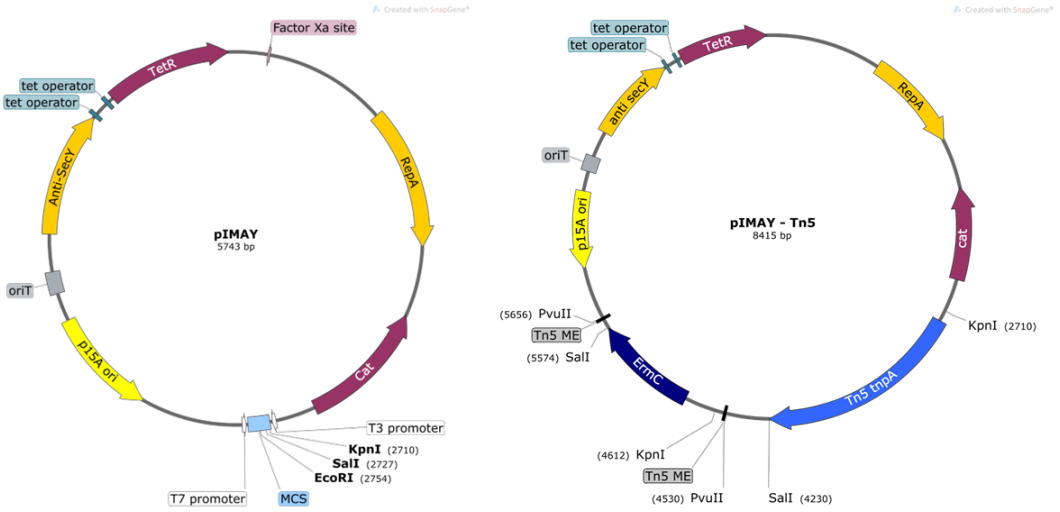

Figure 2-1 *Maps of the pIMAY and pIMAY-Tn5 plasmids.*

Left: pIMAY Right: pIMAY-Tn5

2.2.2 pSAM_Ec

The pSAM_Ec plasmid (plasmid # 102939, Addgene, gifted by Matthew Mulvey), *Figure 2-2*, has been used by other groups to generate transposon insertion mutant libraries (Goodman et al., 2009; Sivakumar et al., 2019; Zhao et al., 2017). It contains a mariner

transposon flanked by the MmeI modified inverted repeats (IRs) described in (Goodman et al., 2009) . The Himar1C9 transposase is under an inducible *lac* promoter, and the transposon consists of an aminoglycoside phosphotransferase, conferring kanamycin resistance, flanked by two termination sequences to prevent downstream transcriptional changes upon insertion into the chromosome (Wiles et al., 2013).

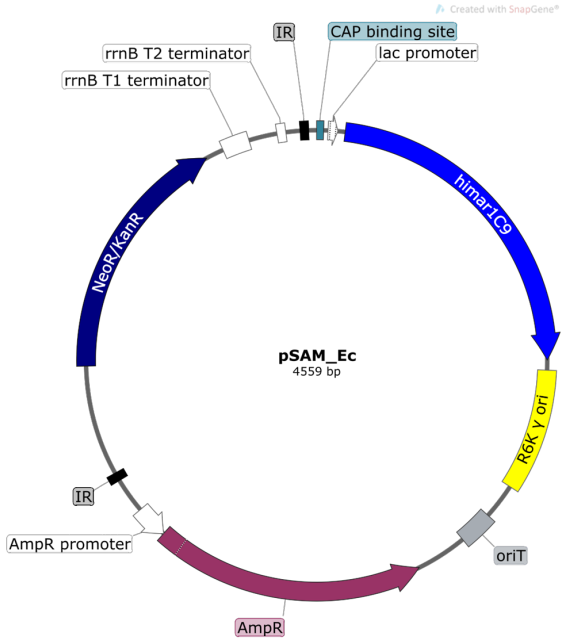

Figure 2-2 *Map of the pSAM_Ec plasmid.*

2.2.3 pBAMD1-6

A derivative of pBAM1 (Martínez-García et al., 2011), pBAMD1-6 (plasmid # 61566, Addgene, gifted by Víctor de Lorenzo) is a modular mini-Tn*5* delivery vector that has been previously used to generate transposon insertion libraries (Martínez-García, Aparicio, Lorenzo, et al., 2014). It uses the same mosaic ends (MEs) as the EZ: Tn*5* kits (LGC Biosearch Technologies, Petaluma, USA, formerly Lucigen). The transposase is the hyperactive mutant of Tn*5 tnpA,* containing three well described amino acid substitutions (D. Liu & Chalmers, 2014; Martínez-García et al., 2011). The Tn*5* transposon consists of the two MEs and a gentamycin acetyltransferase constituently expressed by a Pc promoter, *Figure 2-3:Left*.

2.2.4 pBAMD1-2

pBAMD1-2 (plasmid # 61564, Addgene, gifted by Víctor de Lorenzo) is the same vector as pBAMD1-6 (Martínez-García, Aparicio, Lorenzo, et al., 2014), except that the gentamycin acetyltransferase on the transposon is replaced by an aminoglycoside phosphotransferase conferring resistance to kanamycin, *Figure 2-3:Right.* The two plasmids were constructed for the same study and function in the same way.

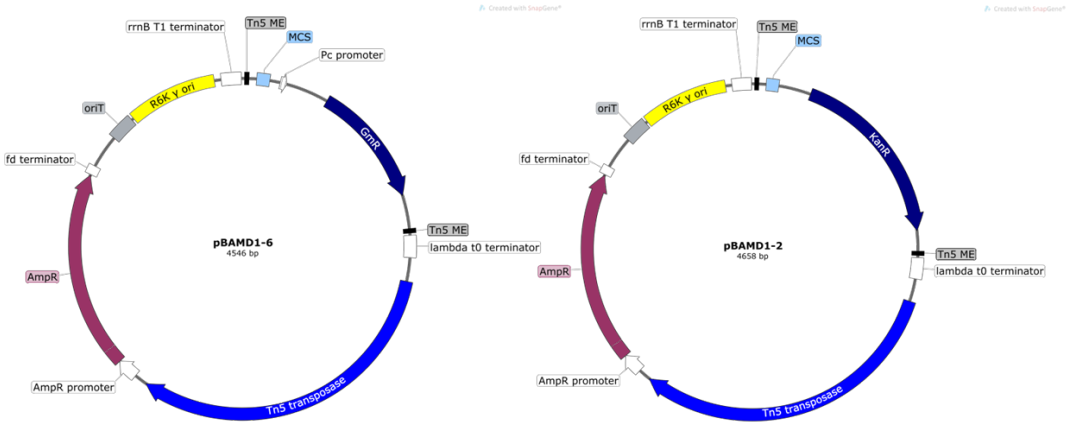

Figure 2-3 *Maps of the pBAMD1-6 and pBAMD1-2 plasmids.*

Left: pBAMD1-6 Right: pBAMD1-2

2.2.5 pExM

pExM is the exponential mutagenesis plasmid constructed during this work for generating exponential transposon mutants. The Himar1C9 transposase, the preceding catabolic activator protein (CAP) binding site and lac promoter were amplified from the plasmid pSAM_Ec (Wiles et al., 2013), and ligated into the Tn5 transposon between the 5' Tn5 ME and the acetyl transferase gene for gentamicin resistance. The mariner transposon was cloned between the acetyl transferase gene conferring gentamicin resistance and its constitutive promoter in the Tn5 transposon on the pBAMD1-6 plasmid (Martínez-García, Aparicio, Lorenzo, et al., 2014). A full description on the generation of the pExM plasmid can be found in *section 6.2.1.*

2.3 Molecular Cloning Techniques

2.3.1 Polymerase Chain Reaction

High fidelity (low error rate) polymerase chain reaction amplification (PCR) was performed with primers designed using SnapGene® Viewer, version 6.0.4.The appropriate annealing temperature was determined for each primer pair using the New England Biolabs (NEB, Ipswich, USA) online annealing temperature calculation tool (https://tmcalculator.neb.com/#!/main). Primers were designed to be the same as the sequence (5'-3') flanking the region to be amplified. Each pair was designed to be a minimum of 15 bp long and to have the same melting temperature (Tm), or within 4 ˚C where equal was not possible. A 2x master mix of NEBNext Ultra II Q5 or NEB Q5 polymerase (NEB) was used per manufacturer's instructions.

Lower fidelity colony PCR was performed using NEB 2x Taq master mix (NEB) following manufacturer's instructions. This was used to confirm the presence of transposons or plasmid backbone and successful cloning attempts. Primers were designed with suitable annealing temperatures as described above.

2.3.2 Restriction Endonuclease Digestion

Most of the endonucleases used were from NEB (high-fidelity option when available) and compatible with the 10x CutSmart™ buffer (NEB). Reactions were typically performed in 30 µL volumes, with 1/10 of reaction volume of 10x CutSmart™ buffer, up to 1 µg of DNA to be digested, 1 µL (20 units) of the appropriate restriction endonuclease and molecular water up to the final volume. All reactions were incubated at the optimal temperature of 37 °C for 15 minutes per microgram of DNA to be digested.

The exception to this was the endonuclease Bsp1407I (Thermo Scientific);). The digestion rection contained 1 µg of plasmid DNA, 1 µL (10 U) of Bsp1407I, 3 µL of supplied 10 x Tango buffer, 1 µL (1 U) of FastAP and was adjusted to 30 µL with nuclease free water, *section 6.2.1.3.2*. The digested products were either cleaned with SPRI beads, *section 2.3.3* or size selected with agarose gel, *section 2.3.4.*

2.3.3 Solid Phase Reverse Immobilisation (SPRI) Bead Purification

DNA preparations were purified after PCR, restriction endonuclease digestions, or modifications where either buffer incompatibilities or smaller DNA fragment removal was required. AMPure XP beads (Beckman Coulter, Pasadena, US) were equilibrated to room temperature prior to use, then typically 1.5x the initial volume of starting material was mixed with the nucleic acid to be purified. This ratio of beads to DNA was used for buffer exchange, purifying or concentrating, but was decreased to as low as 0.7x the starting volume for applications where size selection of large DNA fragments was required, one such application is described in *section 2.8.8*. For size selection, the ratios used were in accordance with manufacturer's instructions. The DNA and bead combinations were thoroughly pipette-mixed in 1.5 mL tubes and left at room temperature for five minutes. Tubes were then placed on a magnetic rack for 3 minutes or until the supernatant was clear, then the supernatant was removed. The beads were washed twice with 200 μL 80% ethanol and allowed to air dry for 5 minutes or until it was judged by eye that there was no longer ethanol present. Tubes were then removed from the magnetic rack and the beads were resuspended in 25 μL of molecular water as standard (this was reduced to 10 μL if more concentrated DNA was required) and incubated at room temperature for a further 5 minutes. Tubes were then placed back on the magnetic stand for 5 minutes until the supernatant was clear and the supernatant (containing DNA) was retained.

2.3.4 Gel Electrophoresis

Gel electrophoresis was performed using a 50x concentrated TAE buffer and diluting in water to a final concentration of 40 mM Tris-acetate and 1mM EDTA (Ethylene Diamine Tetra-acetic Acid; Thermo Scientific, Waltham, USA). Typically, 1% w/v agarose (Sigma Aldrich, St. Louis, US) gels were made and run in the TAE buffer at 120 V, 400 mA for 30 minutes. DNA was combined with 6x Purple DNA Loading Dye (NEB) and was stained with 10,000x SYBR Safe DNA Gel Stain (Invitrogen, Waltham, USA), which was added to the agarose prior to pouring. DNA was run alongside the NEB 1kb Plus ladder (NEB) for size markings and visualised using a Syngene™ Blue LED transilluminator.

2.3.5 Agarose Gel Size Selection

Once the gel was run, the band corresponding to the size of the digested plasmid backbone (*section 6.2.1.2.3*) was excised from the gel using a scalpel, and the DNA was purified using

the Macherey Nagel PCR and Gel clean up kit (Macherey Nagel, Düren, Germany). The gel was melted in the supplied DNA binding buffer NTI, at a ratio of 200 µL per 100 mg of gel and heated to 50 ˚C for 10 minutes. Then the sample was run through the provided column at 11,000 x *g* for 1 minute. The column was washed twice with 700 µL of the supplied wash buffer NT3, centrifuging at 11,000 x *g* between washes. Finally, the DNA fragment was eluted from the column by adding 30 µL of the supplied elution buffer and incubated at room temperature for 2 minutes, then centrifuged at 11,000 x *g* for 1 minute. The eluate was used for downstream ligations.

2.3.6 DNA Phosphorylation

Phosphorylation of PCR products or primers was performed with polynucleotide kinase (PNK) from NEB, as per manufacturer's protocol but using the 10x T4 Ligase buffer (NEB) in place of the buffer provided. Typically, 2 µL of buffer was added to 17 µL of DNA and 1 µL of PNK. The reaction was incubated at 37 ˚C for 30 minutes, then 65 ˚C for 20 minutes to inactivate the enzyme. When used to phosphorylate oligos prior to PCR, to generate the custom Tn*5* transposon (*sections 2.5.2 and 6*), the reactions were cleaned with SPRI beads (*section 2.3.3*). For ligation reactions, there was no clean up and DNA ligase was added directly to this reaction (*section 2.3.8*).

2.3.7 DNA Dephosphorylation

Dephosphorylation of endonuclease digested products was performed with FastAP thermosensitive alkaline phosphatase (Thermo Scientific) as per the manufacturer's protocol. Up to 1 µg of phosphorylated DNA suspended in a maximum of 17µL of water was added to 2 µL of 10x reaction buffer and 1 µL of alkaline phosphatase. The reaction was incubated at 37 ˚C for 10 minutes and then 65 ˚C for 15 minutes. Alternatively, for dephosphorylation of a linearised vector, 1 µL of alkaline phosphatase was added directly to the restriction endonuclease reaction (*section 2.3.2*), as the enzyme was effective in 10x CutSmart™ buffer (NEB) supplied with the restriction endonuclease.

2.3.8 DNA Ligation

Ligation of PCR amplified, or restriction endonuclease digested products was performed by adding 1 µL of T4 ligase (NEB) to the heat inactivated PNK reaction (*section 2.3.6*) and

incubated as per manufacturer's instruction: 10 minutes at 22˚ C for cohesive ends (*section 6.2.1.3.3*) or 1 hour at 22˚ C for blunt ended ligation (*section 6.2.1.2.4*), then heat inactivated at 65 ˚C for 10 minutes. Alternatively, reactions were left at 16 ˚C overnight and then heat inactivated at 65 ˚C for 10 minutes.

2.4 Transformation of Cells

2.4.1 Making Chemically Competent Cells

Chemically competent cells were made using a modified protocol (Chang et al, 2017), based on that published by (Chung et al., 1989). Cells were grown to stationary phase overnight and diluted 1:100 into fresh LB Miller broth (Formedium Ltd.), permissible for organism growth (*Table 2-*). The subcultures were incubated at 37 °C and under shaking (200 rpm), until they reached an optical density at 600 nm (OD_{600}) of 0.4 – 0.5 (determined using an Eppendorf BioPhotometer® spectrometer, Eppendorf, Hamburg, Germany). The cells were then centrifuged at 4000 x *g* for 5 minutes and resuspended without vortexing or pipetting in 1/10 of the initial volume in ice-cold 30mM $CaCl_2$ (Sigma Aldrich, St. Louis, USA), with 15% glycerol (Thermo Scientific). This wash was repeated twice, and the cells were finally divided into 50 µL aliquots and stored at -80 °C.

2.4.2 Chemical Transformation

This was performed again following the protocol published by (Chang et al., 2017; Chung et al., 1989). Aliquots of competent cells (*section 2.4.1*) were thawed on ice, and up to 100 ng of plasmid DNA in a maximum of 5 µL was added to the cells. The tubes were carefully flicked to mix without pipetting or vortexing and were then incubated on ice for 30 minutes. The cells were heated to 42 °C for 45 seconds using a dry block then immediately placed on ice for 5 minutes to recover. Then, 950 µL of 37 °C prewarmed SOC broth (Formedium Ltd.) supplemented with any of the additives required to support auxotrophic growth (*Table 2-),* was added to the cells. The cells were then incubated at 37 °C, 200 rpm for 60 minutes, before being centrifuged at 4500 x g for 1 minute. Next, 900 µL of the supernatant was discarded, and the cell pellet was resuspended in the remaining 100 µL. The cell suspension was then spread onto prewarmed selective agar plates *(Table 2-) and* incubated at 37 °C overnight.

2.4.3 Making Electrocompetent Cells

Electrocompetent cells were made using a modification of the method published by Dower et al. (Dower et al., 1988). A stationary culture of cells was diluted 1:1000 into prewarmed LB broth and grown at 37°C with shaking (200 rpm). Once the subculture had reached OD_{600} of 0.3-0.4 (Eppendorf BioPhotometer®), it was cooled on ice, with all the following steps being performed at 4 °C (centrifugation) or on ice. The cells were then pelleted by centrifugation at 4000 x g or 10 minutes, and the supernatant discarded. Next, the cells were gently resuspended in half of the volume of 10% (v/v) glycerol and incubated on ice for an hour. The centrifugation and resuspension in half volume 10% (v/v) glycerol was repeated three times more, to wash and concentrate the cells. Finally, the cells were aliquoted as 60 µL in a 1.5 mL centrifuge tube and stored at –80 °C.

2.5 Transposon Library Generation

2.5.1 Conjugation

Conjugation was performed using a protocol with elements from previously published work (Freed, 2017; Naorem et al., 2018). The donor and recipient cells were grown overnight to stationary phase in LB broth, with addition of the appropriate supplements or antibiotics (*Table 2-).* The cells were washed three times in the starting volume of Phosphate Buffered Saline (PBS) to remove any antibiotics and resuspended in 1/20th of the initial volume in PBS. The concentrated donor and recipient cells were then mixed in a ratio of 1:1 volume. The mix was spotted as 10 µL independent matings on an LB agar plate for standard library preparation. When the conjugations were for the production of Exponential Mutagenesis libraries (*section 6.2.2.1*), the cells were spotted onto LB agar containing 0.3% (w/v) glucose (Thermo Scientific), to supress expression of the second transposase during conjugation. The plate was incubated at 37 °C for five hours to allow conjugation and transposition to occur.

2.5.2 Constructing Transposomes

Custom Transposomes were constructed using the EZ: Tn5 kit (LGC Biosearch Technologies, Hoddesdon, UK). The composite transposon was amplified from the plasmid pExM (*section 2.2.5*) using the 5’ phosphorylated Tn5 ME primer (Pho-CTGTCTCTTATACACATCT) and the NEBNext Ultra II Q5 2x Master mix as per the manufacturer's protocol (*section 2.3.1*). The

amplified transposon was concentrated to 500 ng/µL using SPRI beads (*section 2.3.3*), to give 400 fmol in 2 µL, and mixed with the purified transposase and glycerol as per the manufacturer's instructions. The reactions were incubated at room temperature for 30 minutes and then stored at -20 ˚C.

2.5.3 Electrotransformation of Transposomes

Electroporation parameters were obtained from the protocol published by Dower et al. (Dower et al., 1988). One aliquot of electrocompetent cells (*section2.4.3)* per transformation was thawed on ice, and 0.4 µL of prepared transposomes (*section 2.5.2)* was added and mixed by stirring with a pipette tip. The transformation mix was incubated on ice for 30 minutes and then transferred to an ice cold 2 mm electroporation cuvette (Bio-Rad, Hercules, USA). Transformations were performed using the Bio-Rad GenePulser Xcell™ exponential decay wave electroporator set to 2500 V, 25 µF and 200 Ω. SOC, prewarmed to 37 ˚C, was immediately added to the cells following the pulse, and the cells were incubated for 90 minutes at 37 ˚C, 200 rpm. Then the cells were pelleted by centrifugation, 900 µL of the supernatant removed and the recovered cells were resuspended in the remaining 100 µL. The recovered cells were spread onto LB agar selecting for the transposon, *Table 2-.*

2.5.4 Outgrowth from Conjugation to a Transposon Library

Following conjugation, the mating spots were harvested from the LB agar conjugation plates, using a sterile 10 mm loop, into sterile PBS. The cells were washed twice with PBS and resuspended in 3 mL of PBS. Then, 500 µL of the cell suspension was spread onto a large square plate (245 mm x 245 mm) of LB agar containing the antibiotics needed to select for the transposon in use *(Table 2-*). Six plates were spread per library, aiming for a colony density of around 80,000 mutants per plate. The plates were incubated at 37 ˚C overnight. Following incubation, the plates were harvested by resuspending the cells on the plates into PBS using a sterile spreader, the mutants from all six plates were pooled to generate one library. An equal volume of 40% (v/v) glycerol was added to give a final glycerol concentration of around 20%, before being stored at -80˚ C.

2.5.5 Estimating Library Purity

For this work, library purity was considered as a determination of the true transposition events compared to cells that were harbouring the conjugated plasmid. Following conjugation, the cells were serially diluted to 10^{-8} in PBS, with each dilution plated in duplicate as 5 µL spots onto LB agar selecting for the transposon and a separate plate with LB agar selecting for both the transposon and the plasmid backbone *(Table 2-).* The plates were incubated at 37 ˚C overnight. Once the colonies had grown, the colony forming units (CFU) per mL were calculated and the ratio used to estimate the library purity.

$$Purity = \left(\frac{CFU\ Transposon\ Selection}{CFU\ Backbone\ Selection}\right) \times 100$$

2.5.6 Estimating Unique Insertions

The number of unique insertions was estimated by serially diluting the recipient cells to a 10^{-8} final dilution, prior to and following the five-hour conjugation. Thus, the number of recipient cells used per conjugation and those obtained following five hours of incubation at 37 °C were calculated. The difference was used to calculate the number of doublings of the recipient cells during incubation at 37°C.

$$Doublings = \log_2\left(\frac{Input\ CFU}{Output\ CFU}\right)$$

Once the doublings were calculated, the number of unique insertions could be estimated from the number of mutants obtained from the purity calculation (*section 2.5.5).*

$$Unique\ Insertions = \frac{CFU\ Transposon\ Selection - CFU\ Backbone\ Selection}{2^{Doublings}}$$

2.6 Nucleic Acid Extraction, Quantification and Quality Determination

2.6.1 Genomic DNA Extraction

Genomic DNA was extracted using the GeneJET Genomic DNA Purification Kit (Thermo Scientific) and following the manufacturer's protocol for Gram-Negative bacteria. Briefly, up to 10^9 cells from either a stationary culture or from harvested transposon library stocks, were lysed with the supplied lysis and Proteinase K solutions; then treated with RNase A (provided in the extraction kit). Then the DNA was bound to the silica column and washed twice with the supplied wash buffer containing ethanol. After removal of the ethanol from the column, the genomic DNA was eluted in molecular grade water. The protocol was slightly modified to increase the concentration of DNA obtained hence, the DNA was eluted in 40 µL and then a further 40 µL to give a final volume of 80 µL rather than the recommended 200 µL.

Alternatively, DNA extraction was automated using the Maxwell® RSC 48 instrument (Promega, Madison, USA), following the manufacturer's protocol for gram negative bacteria. Extraction was from a maximum of 1 x 10^9 cells using the Maxwell® RSC Cultured Cells DNA extraction Kit (Promega) and using the AS1620 program for the instrument.

2.6.2 High Molecular Weight Genomic DNA Extraction

High Molecular weight genomic DNA was extracted from cells using the FireMonkey and FireFlower combined extraction kit (Revolugen Ltd., Glossop, UK) and following the manufacturer's updated protocol (August 2021) for extraction from gram negative bacteria. No modifications were made to the protocol and extra care was taken at any pipetting steps to maintain DNA integrity.

2.6.3 Plasmid DNA Extraction

Plasmid DNA was extracted using the NucleoSpin Plasmid Mini kit (Macherey-Nagel) and following the manufacturer's protocol for Gram-Negative bacteria. The kit used alkaline lysis for extraction; the protocol was modified to elute in 30 µL of molecular water to achieve a higher concentration of final product.

2.6.4 Spectrophotometry for Purity Determination

Spectrophotometry was performed using the NanoDrop® ND-1000 (Thermo Scientific). The spectrum was visualised using the ND-1000 V3.8.1 software (Thermo Scientific). Nucleic acid was deemed to be adequately pure for downstream applications if the 260/280 nm ratio was above 1.60 and the 260/280 nm ratio was above 1.80, based on the optimums stated in the Nanodrop Technical Note 52646 (Matlock, 2015).

2.6.5 Fluorescence Quantification

The Qubit™ (Invitrogen) assay was used to determine nucleic acid concentrations. Both the dsDNA High Sensitivity (0.1 – 120 ng) and dsDNA Broad Range (4 – 2000 ng) assays were used throughout this work. The first step in the protocol was to dilute 1 µL of dye in 199 µL of buffer per sample. Then, 2 µL of the nucleic acid to be quantified was added to 198 µL of freshly diluted dye in a Qubit™ Assay tube, vortexed to mix and incubated at room temperature for two minutes. The fluorescence was measured and converted to a double stranded DNA (dsDNA) concentration using the Qubit™ 4 fluorometer.

2.6.6 Automated Electrophoresis

Automated electrophoresis was performed with the TapeStation 2200 instrument (Agilent, Santa Clara, USA) to determine the integrity, size distribution and molarity of DNA samples for short read sequencing. As fragments were expected to be under 3kb, the D5000 Screen tape, buffer, and ladder, (Agilent), were used according to manufacturer's instructions. The DNA ladder sample was made by adding 1 µL of concentrated ladder to 10 µL of sample buffer and reused for multiple measurements. For each sample to be measured, 0.5 µL of sample was added to 5 µL of sample buffer and mixed by stirring with a pipette tip, then loaded into the instrument as per the manufacturer's protocol. Analysis was performed through the TapeStation 2200 software where concentration, molarity and integrity were reported alongside a curve representing nucleic acid size distribution.

2.7 Whole Genome Sequencing

Whole genome sequencing (WGS) was performed using a minimised version of the Illumina DNA Prep protocol, (Illumina, 2022), developed by David Baker and provided as an institute service. Genomic DNA was extracted as detailed in *section 2.6.1* diluted to 5 ng/ µL as

determined by Qubit®, *section 2.6.5,* and submitted for whole genome sequencing on a 150-cycle paired end flow cell of the NextSeq 500 instrument (Illumina, Cambridge, UK).

2.8 Transposon Directed Sequencing

Transposon directed sequencing was performed using elements from Illumina's standard DNA whole genome sequencing protocol (Illumina, 2022), and adaptations made by Dr. Keith Turner and Dr. Muhammed Yasir to generate a pool of Tn-Chr junctions for sequencing (Yasir et al., 2020) the adaptations to the standard Illumina protocol include custom amplification primers, biotin affinity capture and increased PCR cycle numbers.

2.8.1 Sequencing Amplification Primer Design

The custom sequencing primers used in this work were designed with the same rationale as those used by Dr. Keith Turner and Dr. Muhammad Yasir (Yasir et al., 2020) including the following 5 distinct regions, detailed in *Table 2-2,* which are required to be compatible with the Illumina sequencing platform.

Table 2-2 *The elements required for designing Illumina compatible transposon insertion sequencing custom sequencing primers.*

Element	Length (bp)	Sequence (compatible Illumina barcode)
Illumina i5 Adapter	29	AATGATACGGCGACCACCGAGATCTACAC
Illumina Nextera i5 Index	8 8	TATCCTCT – Tn*5* (i503) AGAGTAGA – mariner (i504)
Read 1 Sequencing Primer Site	33	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Diversity Spacer	8 8 8 8	CTGACCAG TGACATCA GACTGAGC ACTGTGTT
Transposon Binding Site	29 20 22	TTTACAAGCATAAAATCTCTGAAGATGTG – Tn*5*_Forward GCCTGAGACACAAAGATGTG – Tn*5*_Reverse TACGAAGACCGGGGACTTATCA – mariner

Designed to be used with the Illumina Nextseq platform. The i5 adapter and Transposon binding site are variable regions, with the binding site being specific to the transposon used for the study. The compatible Illumina barcode enabled multiplexing.

2.8.1.1 The Illumina i5 Adapter

This region annealed the single stranded DNA (ssDNA) to be sequenced to the flow cell and anchored it for sequencing to occur. This results in cluster generation on the instrument so the sequence in the custom primer was designed to be the same as the one that is the standard Illumina Nextseq oligo (Illumina, 2019).

2.8.1.2 The Illumina i5 Index

This was the first variable region in the designed primers; intended to include any of the standard Illumina Nextera eight base pair barcodes to enable sample multiplexing. This was included in the design to maintain dual barcode multiplexing opportunities. The primers for this study were designed with Tn*5* transposon libraries having the i503 barcode and mariner transposon libraries having the i504 barcode (Illumina, 2019).

2.8.1.3 The Illumina Read 1 Sequencing Primer Site

This sequence is complimentary to the Read 1 standard sequencing primer that is used for Illumina sequencing on the flow cells. This is where sequence reads originate. Illumina does have the option for custom primers to be used in this place, however, keeping the standard option made TIS sequencing compatible with any other routine sequencing runs.

2.8.1.4 The Diversity Spacer

The diversity spacer was included in the primer design to mitigate the fluorescence problems encountered when sequencing low diversity pools. Transposon mutant libraries were an example of a low diversity pool as all of the reads will start with the 3' end of the transposon. Adding the eight-base pair diversity spacer ensures that there are fluorescent readings for all channels and removes the need for altering the sequencing reaction protocols to include dark cycles where no imaging occurs (Barquist et al., 2016; Cain et al., 2020).

2.8.1.5 The Transposon Binding Site

This section of the primer was designed to specifically bind to the transposon and be outward facing, to ensure amplification of the Tn-Chr junction. The binding site spanned the 3' end of the transposon and into the IR for mariner, ending 10 bp before the end of the IR. This 10 bp allowed each read to be verified as a true transposon read. Similarly, the Tn*5* custom primers were designed to anneal within the 3' end of the transposon, through the ME and leave 12 bp for transposon validation. The binding sites were designed to have an annealing temperature of 64 ˚C using NEB Tm Calculator (NEB, https://tmcalculator.neb.com/#!/main).

2.8.2 Blocking Primer Design

Blocking primers were introduced into the reaction to minimise amplification of the transposon-plasmid junctions that may be present, particularly in the limited growth transposon libraries (*section 4.2.2.2*). The rationale was based on previously published studies that have used a C3 spacer on an oligonucleotide to prevent polymerase progression; with the intent to enrich rare mutation events amongst known sequences (Lee

et al., 2011; Vestheim & Jarman, 2008). The primers were designed to anneal to the relevant plasmid sequences immediately after the transposon ME or IR; this would prevent polymerase progression and prevent amplification of transposon-plasmid junctions. The oligos were designed to have an annealing temperature of 64 ˚C determined by NEB Tm Calculator tool (NEB, https://tmcalculator.neb.com/#!/main). The oligo sequences are listed in *Table 2-3.*

Table 2-3 *Sequences of the blocking primers designed and used in this study.*

Primer Name	Sequence
Tn*5*_Block_F	ACTAGTCTTGGACTCCTGTTG – c3
Tn*5*_Block_R	AAAGGCATCAAATAAAACGAAAGG – c3
Mar_Block_F	TAGGATCCAGTGAGCGCAACG – c3
Mar_Block_R	TAGGTACCGAGGACGCGTG – c3

2.8.3 Biotinylated Primer Design

A biotinylated primer was designed for use during amplification of the Tn-Chr junctions so that the junctions could be purified from the pool of DNA, to get a deeper sequence depth of the target. The primer was designed to be used with the Illumina DNA Prep kit, (Illumina). As such, the sequence was the same as the annealing sequence of the Illumina i7 barcoding primer (Illumina, 2019) but with a 5' biotin added. Sequence: [BTN] – GTCTCGTGGGCTCGG.

2.8.4 Tagmentation

Tagmentation of the genomic DNA extracted from harvested libraries *(section 2.6.1)* was normalised to 50 ng in 15 µL of molecular water based on Qubit® concentration (*section 2.6.5).* All reagents were from the Illumina DNA prep kit (Illumina) unless otherwise stated. Following the Illumina DNA Prep protocol (Illumina, 2022), but as half reactions, 5 µL of bead linked transposase (BLT) and 5 µL of transposase buffer 1 (TB1) were added to the normalised DNA and mixed well. The reaction was incubated in a thermocycler at 55 ˚C for 15 minutes with the heated lid set to 100 ˚C, then cooled to 10 ˚C. The Tagmentation reaction was stopped with the addition of 5 µL of tagment stop buffer (TSB), mixed well, and incubated for 15 minutes at 37 ˚C in a thermocycler with the lid set to 100 ˚C, before being cooled again to 10 ˚C. The reactions were incubated on a magnetic plate for five

minutes to separate the beads from supernatant, and the supernatant was removed. The beads were then washed twice in 100 µL of Tagmentation wash buffer (TWB) by removing the tube from the magnet and fully resuspending the beads, then returning to the magnet before removing the supernatant.

2.8.5 Amplification

Amplification of the Tn-Chr junctions was performed using the bead linked tagmented DNA as the PCR template. For each sample, the beads with the tagmented DNA *(section 2.8.4)* were resuspended in a reaction mixture comprising of 20 µL of NEBNext Ultra II 2x Q5 master mix (NEB), 5 µL of Custom transposon primer (10 mM) *(section 2.8.1)*, 5 µL of Biotinylated i7 primer (10 mM) *(section 2.8.3)*, 5 µL of transposon specific forward blocking primer (100 mM) and 5 µL of transposon specific reverse blocking primer (100 mM) *(section 2.8.2)*. The PCR program was 72 ˚C for 3 minutes followed by 16 cycles of 98 ˚C for 10 seconds, 64 ˚C for 60 seconds and 72 ˚C for 20 seconds and finally held at 4 ˚C until ready to proceed. The PCR products were purified with a 1:1 ratio of Ampure XP beads *(section 2.3.3).* The final product was eluted in 40 µL of elution buffer (EB) (Qiagen, Hilden, Germany).

2.8.6 Biotin Affinity Capture

Affinity capture was performed following the protocol developed by Dr. Muhammed Yasir. The biotinylated PCR products were bound to Streptavidin coated magnetic beads for purification. The beads used were Dynabeads™ in the KilobaseBINDER™ kit from (Invitrogen). The beads were prepared by transferring 10 µL (100 µg) of beads to a microcentrifuge tube for each sample, placing the tube on a magnetic stand and removing the supernatant once clear. The beads were washed with 40 µL of the supplied binding solution by resuspending the beads and then placing on the magnetic stand and removing the supernatant once clear. The beads were then resuspended in 40 µL of the supplied binding solution and the 40 µL of clean PCR product *(section2.8.5)*, was added and mixed well by pipetting. The tubes were incubated on a rolling shaker for 4 hours at room temperature. Following incubation, the tubes were placed on the magnetic stand and the clear supernatant removed. The beads were washed by resuspending in 40 µL of the provided washing solution, removing the tube from the magnetic stand each time and then repeated twice with 40 µL molecular grade water. The beads were placed back onto the

magnetic stand, the supernatant from washes was removed and then the beads were resuspended in 15 µL EB (Qiagen, Hilden).

2.8.7 Second Amplification

A further amplification of the DNA fragments bound to the streptavidin beads *(section 2.8.6)* was performed. The 15 µL resuspended beads were combined with 25 µL of NEBNext Ultra II 2x Q5 master mix (NEB), 5 µL of Custom transposon primer (10 mM) *(section 2.8.1),* 5 µL of unique barcoded Illumina Nextera i7 primer (10 mM) and mixed well by pipetting. The PCR program used was 72 ˚C for 3 minutes followed by 12 cycles of 98 ˚C for 10 seconds, 64 ˚C for 60 seconds and 72 ˚C for 20 seconds and finally held at 4 ˚C until ready to proceed. The streptavidin beads were separated from the PCR products by placing the tubes on the magnetic stand and retaining the supernatant containing the PCR products.

2.8.8 Pooling and Loading

Each sample was run on the TapeStation and quantified by the Qubit® fluorometer *(section 2.6.6 and 2.6.5)*. Samples for a single run were pooled in equimolar quantities determined by five times the lowest molarity. This pool was purified using a 0.7x SPRI wash *(section 2.3.3)* and run on the TapeStation *(section 2.6.6)* to get a final pool molarity. The pool was diluted to a 4 nM concentration and loaded onto the NextSeq 500 for a 75-cycle single ended read run, following the Illumina loading guide (Illumina, 2018). The DNA was denatured using sodium hydroxide, diluted to 20 pM, combined with 35% of PhiX (v/v) at 20 pM (Illumina) then further diluted to 1.8 pM and loaded onto the instrument for sequencing.

2.9 Nanopore Sequencing

2.9.1 Library Preparation

Nanopore sequencing libraries were prepared using the Oxford Nanopore SQK-LSK109 ligation sequencing kit (Oxford Nanopore Technologies, Oxford, UK) and a hybrid method, taking elements from a protocol developed by Dr. Emma Waters (Library Preparation for Native 48/96-Plex Barcoding Sequencing with Ligation kit for Minion DNA Sequencer,

unpublished) and the manufacturer's instructions for adapter ligation and loading the MinION (Oxford Nanopore Technologies).

2.9.1.1 Concentrating DNA

The high molecular weight DNA, *(section 2.6.2)* was concentrated to 1320 ng in 7.5 µL of EB (Qiagen) using 1.2x SPRI beads *(section 2.3.3).*

2.9.1.2 End Preparation

The NEBNext Ultra II End repair/dA-tailing Module (NEB) was used to repair the sheared ends of DNA and add a poly-A tail for ligation in the next step. For each sample, 6.25 µL of concentrated DNA *(section 2.9.1.1)* was combined in a PCR tube with 0.375 µL of enzyme mix and 0.875 µL of reaction buffer. This was incubated at 20 ˚C for five minutes and 65 ˚C for five minutes in a thermalcycler.

2.9.1.3 Native Barcode Ligation

Following End Preparation, 2.5 µL of Oxford Nanopore native barcode, (Oxford Nanopore Technologies) was added to each sample, in order to multiplex the run. Then, 5 µL of NEB Blunt/TA ligase master mix was added to each sample and heated in a thermal cycler at 20 °C for 20 mins and 65 °C for 10 mins.

2.9.1.4 Purification and Quantification

Each prepared sample was quantified using Qubit® *(section 2.6.5)* and pooled to be equimolar. Then the DNA was purified with 1.2x SPRI beads *(section 2.3.3)*, resuspending in 65 µL of EB (Qiagen).

2.9.2 Adapter Ligation

In a separate tube, 60 µL of the purified DNA was combined with 25 µL of Ligation Buffer (Oxford Nanopore Technologies), 10 µL of NEB Next Quick T4 DNA Ligase (NEB) and 5 µL of Adapter mix (Oxford Nanopore Technologies). This was thoroughly mixed by gentle pipetting and incubated at room temperature for 10 minutes. This mix was cleaned with 40

µL SPRI beads *(section 2.3.3),* with Long Fragment Buffer (Oxford Nanopore Technologies) being used instead of 80 % (v/v) ethanol in the wash steps. The DNA was eluted in 15 µL EB (Qiagen).

2.9.3 Loading the MinION

To prepare the flow cell priming mix, 30 µL of flush tether (Oxford Nanopore Technologies) was mixed into a tube of flush buffer and mixed by vortexing to make the priming mix (Oxford Nanopore Technologies). The flow cell was prepared by using a P1000 pipette to draw 20-30 µL of storage buffer from the priming port. Without the introduction of air bubbles, 800 µL of priming mix was added to the flow cell and left to incubate while preparing the DNA sample, or for five minutes. The DNA loading mix was prepared by combining 12 µL of the prepared library *(section 2.9.2)* with 37.5 µL Sequencing Buffer and 25.5 µL of mixed Loading Beads (LB) (Oxford Nanopore Technologies). Then, a further 200 µL of priming mix was added to the flow cell and 75 µL of prepared DNA Library was added to the SpotON port in a dropwise fashion. The run was set up on MINKnow and run to depletion of the flow cell.

3 Reproducibility of Essential Gene Determination using Tn*5* and mariner Transposons and the BioTradis Pipeline

3.1 Introduction

3.1.1 Differences in the Essential Genes Determined in *E. coli* K12

Transposon insertion libraries are commonly used to investigate gene function through essentiality (or conditional essentiality) in an ever-increasing diversity of bacterial species. The gene lists provided by such studies can contribute to publicly available databases of conditionally essential genes that are regularly updated, for example the Database of Essential Genes (DEG) (Zhang, Ou, and Zhang, 2004; Luo et al., 2021). For *E. coli* specifically there is the "Profiling *E. coli* Chromosome" (Yamazaki, Niki, and Kato, 2008) (PEC) and EcoCyc (Keseler et al., 2017) though the reproducibility of these datasets is not clear.

Many **T**ransposon **I**nsertion **S**tudy (TIS) libraries are generated in different pathogenic bacterial species to investigate virulence and drug tolerance, but these are seldom repeated. There are three studies reporting essential genes in the model organism *E. coli* K12 BW25112 and a further five in the parent strain MG1655 and related strains, therefore this is an ideal organism to use for assessing reproducibility. One study in MG1655 (Gerdes et al., 2003) reports 609 essential genes, the highest number of reported essential genes by transposon insertion. Contrasting with this is a report of 299 essential genes that were identified by knockout mutants (Baba et al., 2006).

In the 2020 update of DEG, two TIS experiments in *E. coli* were added to DEG reporting: 315 and 722 essential genes using Tn*5* (Phan et al., 2013) and mariner (Warr et al., 2019) transposon libraries, respectively. Warr et al. further report an additional 275 genes that have essential regions but are not wholly essential, making the total 1086. Studies not listed in this database report 358 (Goodall et al., 2018) and 357 (Mccarthy et al., 2018) essential genes for differing *E. coli* strains, BW25113 and K1 A192PP respectively, both using Tn*5*. The lack of alignment in these data sets, summarised in*, Table 3-1,* suggest that the current methods used to determine gene essentiality are not entirely reproducible even when the same strain is used.

Table 3-1 *A summary of works with published essential gene lists for varying strains of E. coli.*

Study	Strain	Data Generation	Essential Genes Reported
Baba et al., 2006	BW25113	Targeted Knockout	299
Goodall et al., 2018	BW25113	Tn*5*	358
Ghomi et al., 2022	BW25113	Tn*5*	438
Warr et al., 2019	EHEC EDL933	mariner	880[1]
Mccarthy et al., 2018	K1 A192PP	Tn*5*	357
Gerdes et al., 2003	MG1655	Tn*5* (TraSH)	609
Kato and Hashimoto, 2007	MG1655	Targeted Knockout	505
Yamazaki et al., 2008	MG1655	Curated Database	302
Byrne et al., 2014	MG1655	Tn*5*	651
Warr et al., 2019	MG1655	mariner	786[1]
Phan et al., 2013	ST131	Tn*5*	315
Ghomi et al., 2022	ST131 EC958	Tn*5*	334
Ghomi et al., 2022	ST131 NCTC13441	Tn*5*	340
Yamazaki et al., 2008	W3110	Curated Database	300

[1]*This work does not distinguish between essential and non-essential, the value stated here have been determined to be underrepresented in the transposon pool.*

Generating concise but accurate essential gene lists under any condition, particularly stress, is vital for antibiotic drug development. In the case of target site identification, any hits could indicate a potential target requiring further investigation. Any screens that are intended to profile potential resistance mechanisms to a candidate drug may miss an essential gene in the pathways and lead to failure or identify too many genes to be comprehensible. The question driving this chapter is: What is causing the variation seen in published results for TIS results for E. coli K12 essential gene sets?

Fortunately, there is a gold standard for assessing gene essentiality using targeted ORF deletions (Ghomi et al., 2022). The Keio knockout mutant collection (Baba et al., 2006) is a set of E. coli K12 isolates each containing one specific coding sequence knockout and PEC (Yamazaki, Niki, and Kato, 2008) is a curated database of gene knockout mutants from the literature that is updated regularly. The Keio collection of knockout mutants has been used previously to validate gene function and TIS data (Goodall et al., 2018; Guzmán et al., 2018;

Holden et al., 2021; Koo et al., 2017; Yasir et al., 2020). The collection consists of 3985 non-essential deletion mutants. A CDS was considered essential if the deletion mutant was not viable.

One issue with determining essential genes is genome annotation status. There should be an element of caution applied when using the Keio collection today as the coding sequences targeted for the collection were in accordance with a 2005 annotation of E. coli MG1655 (Riley et al., 2006). Baba et al. themselves note that there were 12 ORFs where the coding region had changed (Riley et al., 2006) between construction of the collection and publishing their results. A review in 2009 (Yamamoto et al., 2009) identified 27 genes that were not targeted in construction of the mutant collection. The 2005 annotation (Riley et al., 2006) identifies 4292 coding sequences including 74 pseudogenes. The 2020 annotation of MG1655 comprises of 4364 coding sequences including 166 pseudogenes (GenBank accession GCA_904425475.1). TIS studies are less likely to be subject to errors arising from updated genome annotations and may supersede knockout studies as the gold standard if robust gene determination is achieved. The impact of annotation on determining gene essentiality is important and will be discussed further in this chapter.

Goodall et al. published a comparison of the genes listed as essential from their highly saturated Tn*5* library compared to the Keio collection and PEC (Goodall et al., 2018). The authors state that there are 81 genes identified unique to their TIS experiment compared to only 25 and 18 for Keio and PEC respectively. From these values, it appears that TIS data are overestimating the number of essential genes. This comparison was repeated, *Figure 3-1* where the number of genes unique to the TIS was much higher, this is most likely due to differences in annotation and gene names used and highlights one of the challenges encountered when comparing essential gene sets.

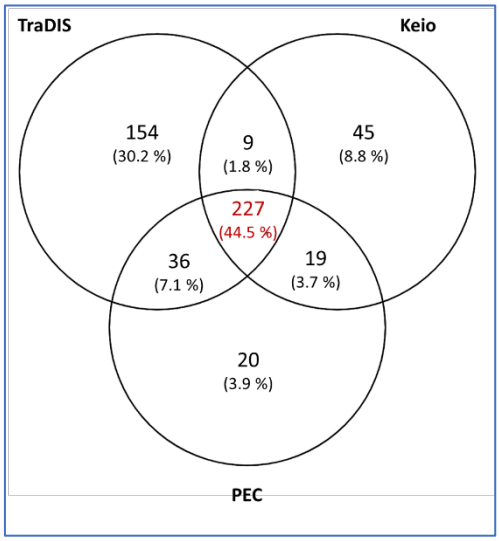

Figure 3-1 *Comparison of the number of essential genes identified using different approaches.*

Summary of the Keio Collection (Baba et al., 2006) *, the Profiling E. coli Database (PEC)*(Yamazaki, Niki, and Kato, 2008) *and a saturated TraDIS library* (Goodall et al., 2018)*. The gene lists were taken from each study, gene names not corrected between studies.*

3.1.2 How is Gene Essentiality Determined in TIS?

A gene is classified as essential through analysis and interpretation of the number of insertions across the chromosome (Cain et al., 2020; Chao et al., 2016). For specific gene knockouts, such as the Keio collection or low throughput transposon insertion protocols such as Signature Tagged Mutagenesis (STM) or Transposon Site Hybridisation (TraSH) (Sassetti et al., 2001), essentiality is determined in a binary fashion: if a mutant is viable then the gene is considered non-essential. Similarly for low insertion density transposon pool data, if a gene can tolerate a transposon insertion, then it is considered non-essential.

For high density TIS libraries, disruption of any non-essential gene is represented repeatedly in the mutant pool providing replicate data for those genes. Having more replicates of a knockout effect can provide increased confidence in the designation of a gene as being non-essential. A gene is considered essential by comparing the expected insertion frequency vs. observed (Barquist et al., 2016; Langridge et al., 2009); if insertions in a coding sequence are underrepresented then the gene is considered detrimental to fitness or essential if below a threshold.

3.1.3 Factors Affecting Essential Gene Determination

In this chapter the following factors affecting the calling of essential genes were investigated.

1. Genome annotation. Genome annotation is not perfect (start and stop codons can vary, and gene names can appear arbitrary) and some genes have essential regions which do not extend over the whole annotated CDS. For example, a TIS study in Caulobacter crescentus, another model gram negative, found that more than six per cent of essential genes appeared shorter than the annotation and around 12 per cent could tolerate insertions in the C-terminus (Christen et al., 2011). Such genes may be missed by automated analyses that only account for the presence or absence of insertions within the entire CDS of the gene. An example in E. coli is ftsK, essential at the N-terminal domain (Draper et al., 1998; Yu et al., 1998); this gene would be essential in a knockout collection but may not be in a TIS study. To this end, some analysis pipelines employ annotation independent approaches; one has been developed by Dr. George Savva in conjunction with this work and will be discussed further in Chapter 5. For this study, the genome of E. coli BW25113 (a K12 derivative) was reannotated using Prokka (Seemann, 2014) and the EcoCyc database (Keseler et al., 2017) (section 3.2.1).

2. The transposon used. Tn*5* and mariner family transposons were investigated in this project. In the case of Tn*5*, the mutant used was the hyperactive form of the native enzyme (Reznikoff, 2003). For the mariner system, the Himar1 transposon was chosen and contained inverted repeats with a single nucleotide change from native to introduce a MmeI restriction site to simplify the sequencing protocols (Van Opijnen et al., 2009). Mariner has a requirement for a TA dinucleotide at the site of insertion but is otherwise promiscuous (Lampe et al., 1999). Whereas Tn*5* has no required insertion site but a preference for GC rich regions (Chao et al., 2016).

3. Transposon mutant library saturation. It is generally accepted, from a statistical standpoint, that a saturated library will generate more robust results for essential genes (genes protected from insertion) (Chao et al., 2016) which raises the question, what defines a saturated library? Theoretical saturation is when a

transposon has inserted into every accessible site within the genome (Nlebedim et al., 2021). In practice this is not the case due to regions of the genome being required for the organism to survive and limitations in library generation. Therefore, the practical definition of saturation could be a transposon insertion in a sufficient proportion of the accessible insertion sites to accurately determine the essential genes for the conditions of library generation. In practice this will be transposon, organism, and condition dependent. For example, Chao et al. comment that saturation is easier to achieve using a mariner transposon due to the requirement for a TA insertion site, thus limiting the number of possible insertion sites within a genome (Chao et al., 2016) .

Some analysis tools developed are exclusively for analysing the data from mariner transposon libraries and therefore only account for TA dinucleotide sites (Miravet-Verde et al., 2020; Solaimanpour et al., 2015). These tools are unsuitable for use with TIS using non-mariner transposons, and may not be suitable for all organisms, for example genomes with high GC contents. While there have been a vast number of tools developed to deal with TIS insertion data, there are similarities in the statistical approaches underpinning each. The differences between tools arise from the processing of sequence read data and the calculation of a cut-off to designate a gene as essential or not. In this work the BioTradis pipeline (Barquist et al., 2016; Langridge et al., 2009) has been used to determine gene essentiality in libraries generated with Tn*5* and mariner transposons.

3.2 Methods for this Chapter

Some of the general methods used for this section of work have been described previously in *Chapter 2.* This section describes the specific protocols that were used to generate the data discussed in this chapter.

3.2.1 Generating a Reference Genome

A reference genome for the *E. coli* BW25113 used for this work was generated as a hybrid assembly using both long read (Oxford Nanopore) and short read (Illumina) sequencing platforms, *sections 2.7 and 2.9.* Firstly, an assembly was produced from the long reads using Flye (with a maximum coverage set to 50x), Galaxy version 2.9 (Kolmogorov et al.,

2019) and two rounds of Racon polishing, Galaxy Version 1.3.1.1 (Kolmogorov et al., 2019). A consensus assembly was generated from the raw reads and the Flye draft assembly using Medaka Galaxy Version 1.4.3 (Oxford Nanopore Technologies, Oxford). Then, short read sequence data was applied, and a final assembly produced using two rounds of minimap2, Galaxy Version 2.17 (H. Li, 2017) and Pilon, Galaxy Version 1.20.1 (Walker et al., 2014) before annotating with Prokka, Galaxy Version 1.14.5 (Seemann, 2014). The Prokka annotation was corrected manually using the EcoCyc (Keseler et al., 2017) database and Uniprot BLAST (Pundir, Martin, and O'Donovan, 2016).

3.2.2 Transposon Library Generation, Sequencing and Data Processing

Two transposons were selected for use in this work: mariner, reported to have a TA insertion site bias and Tn*5*, reported to be promiscuous.

3.2.2.1 Mariner

The mariner libraries used in this work were generated as described in *section 2.5.1*. One spot of 10 µL mating mix between the donor strain *E. coli* MFD*pir+:* pSAM_Ec (Wiles et al., 2013) and the recipient strain *E. coli* BW25113 was sequenced to generate an estimate of insertion density per mating. For this trial mating, there were around 80,000 unique insert sites in the reference genome, data not shown. To make the full libraries, the process was multiplied by five; this was calculated as an overestimate in an attempt to reach saturation. The mating spots were harvested, spread onto selective media, and incubated overnight. The mutants were harvested from the selective plates using LB and pooled. The pools were mixed with a 1:2 ratio of 40 per cent glycerol to mutant pool by volume, final of 20 per cent glycerol, and stored at -80 ˚C.

3.2.2.2 Tn*5*

The protocol for Tn*5* transposon library generation is described in *section 2.5.1*. Similarly, to mariner, there was a trial mating with one spot of 10 µL mating mix between the donor strain *E. coli* MFD*pir+:* pBAMD1-2 (Martínez-García, Aparicio, de Lorenzo, et al., 2014) and the recipient strain *E. coli* BW25113. Sequence analysis indicated that one mating gave around 64,000 unique insertion sites. The aim was to generate 500,000 unique insertion mutants for Tn*5*, and therefore each full library was made with 10 mating spots.

3.2.2.3 Transposon Library Sequencing

For both mariner and Tn*5* transposon directed sequencing was performed with 50 ng extracted genomic DNA from 1 x 10^9 mutants. Genomic DNA extraction using the Maxwell instrument is described in *section 2.6.1*. Sequencing library preparation was prepared following a modified version of the Illumina standard DNA whole genome sequencing protocol (Illumina, 2022). Full details of each step are described in *section 2.8.* The prepared sequencing libraries were run on an Illumina Nextseq 75-cycle single-end run. The base called sequence data was downloaded from the in-house Quadram IRIDA platform (Matthews et al., 2018) and processed as described below.

3.2.2.4 Concatenating Fastq Files to Batches

The transposon libraries were generated as three technical replicates within three biological replicates on separate days. The raw `.fastq` files from the three technical replicates were concatenated, prior to any processing, in Bash using the `-cat` command to form "Batches" where a batch represents a biological replicate. Per transposon, all nine biological and technical replicates were concatenated in the same way to generate "All" libraries. The Batch and All `.fastq` files underwent the same analysis pipelines as the raw `.fastq` files generated.

3.2.2.5 Cutadapt Trimming

The raw, Batch and All `.fastq` files were then trimmed using Cutadapt version 4.1 (Martin, 2011). Trimming was necessary to remove the diversity introduced with the custom sequencing primers and ensure that the reads began with the 10-12 bp transposon tag (*section 2.8.1*). Cutadapt was run to remove bases up to and including the primer annealing site 10 bp upstream of the transposon ME or IR specific for each transposon, detailed in *Table 1-1*. The bases were trimmed from the 5' end of the sequence with the default error rate of 10 per cent, i.e., 1 base error tolerated. The Tn*5* libraries were processed twice, the forward and reverse custom transposon primers have different annealing sites so there were two adapters to trim. The output was a `.fastq` file with any adapter sequences removed, such that the start of the sequence was the transposon tag, *Table 3-2*.

Table 3-2 *The ten base pair sequences provided to Cutadapt.*

Transposon	Sequence used for Trimming
Mariner	`GGACTTATCA`
Tn*5*	F – `TGAAGATGTG` R – `AAAGATGTGT`

To trim as the first stage of TIS sequence read processing, the sequence is unique to each transposon used and is specific to the custom sequencing primer designed for each.

3.2.2.6 The BioTradis Analysis Pipeline

The BioTradis analysis pipeline was developed for use with Tn*5* insertion data (Langridge et al., 2009); it has since been optimised further (Barquist et al., 2016) and is updated regularly. BioTradis is free and licensed under GPLv3. This software was chosen due to experienced users working within the Quadram Institute.

Following adapter trimming, the `.fastq` files were processed through the BioTradis mapping pipeline, Galaxy Version 1.4.5 (Barquist et al., 2016). In the first step of this pipeline the reads were filtered on whether they start with the 'transposon tag'; this was the 12 or 10 base pairs at the 3' end of the ME or IR for either the Tn*5* or mariner transposons respectively, the tags used are listed in *Table 3-3.*

Table 3-3 *The sequences provided to the BioTradis mapping pipeline.*

Transposon	Sequence used for Trimming
Mariner	`TCCAACCTGT`
Tn*5*	`TATAAGAGACAG`

To ensure that a read has originated from a true transposon-chromosome junction. The sequence is specific to the transposon used.

Once a tag was identified, allowing for a one base error, it was trimmed from the 5' end of the read which was then mapped to the hybrid reference genome described in *section 3.3.1.* The reads were mapped using the smalt software (Ponsting & Ning, 2012) with custom parameters described in *Table 3-4* and the minimum mapping quality set to zero to account for repeat sequences. One of the outputs from the BioTradis mapping pipeline was a table containing the number of filtered and mapped reads, these have been summarised and will be presented further in the chapter. The main output of the BioTradis mapping

pipeline (`insert_site_plot.gz` file) was a table with 4,631,419 rows representing each base in the length of the reference genome. The first column contained the number of reads mapping (transposon insertions) in the forward direction; the second column contained the number of reads mapping to the reverse direction. These files underwent further processing as described below.

Table 3-4 *the smalt parameters used to map the trimmed and filtered fastq files to the hybrid reference sequence used in this work.*

Command	Parameter	Default	Custom
smalt_k	Word Length	20	13
smalt_s	Step Size	5	1
smalt_y	Minimum	0.96	0.8
smalt_r	Seed	1	0

3.2.2.7 Insert Plot File Processing

To process the insert plot files output from BioTraDIS I wrote a series of scripts in Windows Visual Studio Code as a Jupyter notebook executed in Python version 3.10.7. The following sections, 3.2.2.8-3.2.2.10, describe the purpose of the scripts which were wrapped into a function and executed with `insert_file_process()`. The inputs required to run this function were the `insert_site_plot.gz` files and the `.stats` file from BioTraDIS.

3.2.2.8 Generating Insert Count Read Numbers

First the `insert_site_plot.gz` files were unzipped and saved as `.txt` files. The plot files contained two columns, and the number of rows equalled the length of the reference genome, one row for each base pair. The two columns represented the forward and reverse orientation of transposon reads. Due to the design of the custom sequencing preparation primers, the reads could be in either direction. To enable further analysis, the two columns were combined to give one insertion count for each base pair and saved as a list in a separate `.txt` file.

3.2.2.9 Filtering Low Insertion Counts

The files were amended to filter out any sites that had low insertion counts. These could represent spurious reads from sequencing or errors from mapping. These reads would interfere with essentiality prediction statistics so needed to be removed. The number of mapped reads was divided by the number of unique insertions to give an average number of expected counts per insertion site. Any count lower than 10 per cent of this average was replaced by zero to disregard these reads. These new counts were saved as both a list in a `.txt` file and as the original two column format to maintain the directionality observation and for visualisation in Artemis if required.

3.2.2.10 Counting Instances of the TA Dinucleotide

A Python script, *appendix 9.4,* was written (in a Jupyter notebook and executed in Python version 3.10.7) to determine the number of instances of the dinucleotide TA. This script was wrapped in the `mariner_analysis()` function and is discussed further in *chapter 4*. Briefly, the same reference genome provided to the BioTradis pipeline was supplied in `.fasta` format and used to search for and count the TA dinucleotides within the sequence.

3.2.2.10.1 Extracting the number of insertions at TA sites

A further script was written to list the base pair location of the TA dinucleotides calculated above *appendix 9.4*. The list was exported as a `.txt` file. Then, a further extension extracted the number of insertions reported at the base pair from the insert plot file for each of these TA dinucleotides, exported as another `.txt` file.

3.2.3 Determining Gene Essentiality

To determine essentiality, the R package within the BioTradis software was used, version 1.4.5 (Barquist et al., 2016). The insert plot files generated from the pipeline were combined with the annotated reference and the `tradis_gene_insert_sites` command aligned the genes with the number of insert site. Then, the `tradis_essentiality.R` command determined gene essentiality and output three .csv files per sample: one for essential genes, one for ambiguous genes, one for non-essential genes. The R package normalised the number of insertions per gene by dividing by

gene length to give each gene an insertion index (Langridge et al., 2009). When plotted as a histogram, the insertion indexes were bimodal, one zero and one non-zero. The R package then fitted a gamma curve to both populations, representing the essential (zero) and non-essential (non-zero) insertion indexes within the library. An example is given in *Figure 3-2*. Ambiguous genes were those where the essentiality algorithms used could not statistically determine whether the gene in question belonged to the essential or non-essential population for whatever the condition being tested was, such genes had an insertion index between the two red lines in *Figure 3-2.*

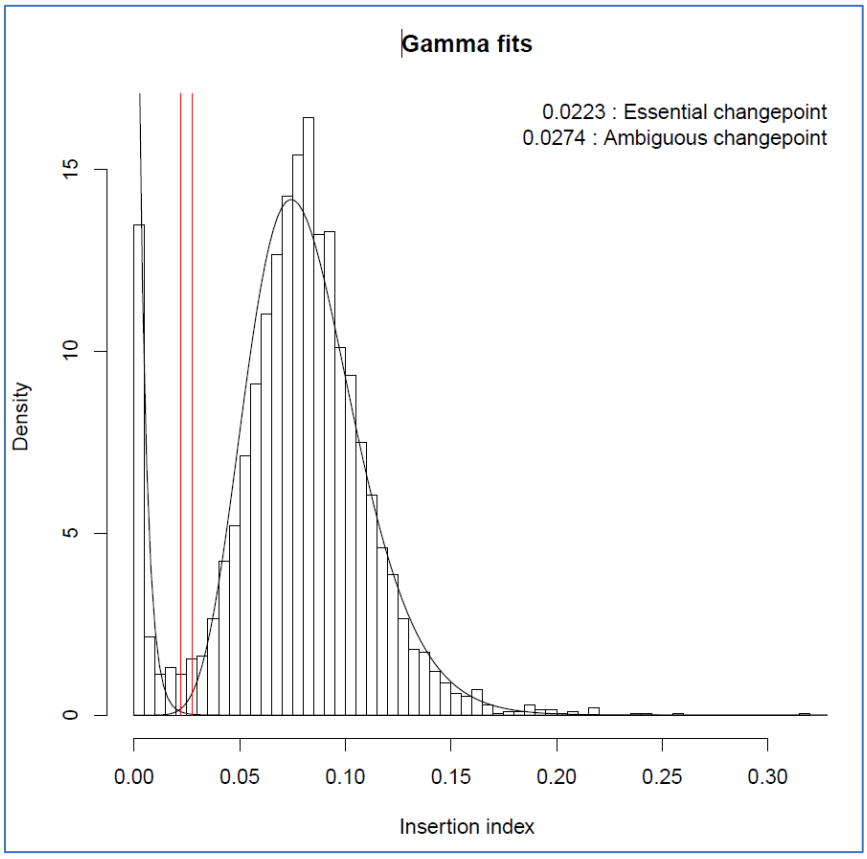

Figure 3-2 *An example of gamma fit curves applied to a histogram of insertion indexes.*

These plots were used to determine gene essentiality, one of the outputs from the BioTradis analysis pipeline. The essential and ambiguous gene changepoints are listed and denoted with red lines; indexes to the left are associated with essential genes, indexes to the right are associated with non-essential genes and indexes falling between the lines associated with ambiguous or undetermined genes.

A Log_2 likelihood ratio was then applied to determine a cut off value for gene essentiality. If the likelihood ratio fell below -2 then a gene was considered essential, if it was above 2 then the gene was deemed non-essential. There was a third category where the likelihood

ratio fell between these two values which was reported as ambiguous. The R package converted the likelihood ratios to an insertion index for each library to demonstrate these cut offs and can be seen as the red lines in *Figure 3-2.*

3.3 Results

3.3.1 Generating a Reference Genome

The hybrid assembly of the BW25113 genome produced one contig of 4,631,919 base pairs in length. This was 450 base pairs longer than the published genome length which is 4,631,469 base pairs. The annotation of the hybrid assembly contained 83 tRNAs and 22 rRNAs; these were not included in essentiality analyses. There was a total of 4354 coding sequences including 102 pseudogenes annotated and all were included in essential gene determination. The FASTA sequence and EMBL annotation files are included as *appendix files 9.1 and 9.2.*

This new reference genome for this strain was generated, rather than using the published sequence deposited in NCBI in 2014 (Grenier et al., 2014) primarily to ensure that the most accurate template was used to map the transposon directed reads and account for any single nucleotide polymorphisms, duplications, inversions, or deletions that may have occurred during multiple passaged growths prior to obtaining the strain.

Furthermore, the published genome sequence was generated using 300 bp paired-end reads, then assembled and manually inspected. The assembly was polished using sanger sequencing reads from PCR products; the length of these was not stated. In comparison my reference was generated using long reads from nanopore sequencing with an N50 of 21.5 kb, meaning that half of the data obtained from this run was in reads greater than 21.4 kb in length. Longer reads are easier to assemble, especially across repetitive regions of genome that are longer than the short read lengths (Yasir et al., 2022; T. Zhang et al., 2022); the addition of long read data could explain the difference of 450 bp seen between the published sequence and my reference.

3.3.2 Transposon Library Generation, Sequencing, and Data Processing

The following section describes optimisation and generation of transposon insertion libraries. Aside from *section 3.3.2.1* these results are organised by transposon for ease of reference.

3.3.2.1 Donor Selection

The plasmid donor strain for both transposon families used in this work was originally *E. coli* ST18 (Thoma & Schobert, 2009). The transposon plasmids contained an *R6K* origin of replication so that they would not replicate within the recipient cells and gave confidence that kanamycin resistance was due to a successful transposition and not conjugation. However, initial experiments indicated that there was plasmid persistence in the recipient or alternatively, homologous recombination had occurred at a higher rate than transposition.

From reading the literature it became apparent that this was a poor choice of donor strain due to Mu phage or *hfr* being transferred with the vector during conjugation (Ferrières et al., 2010) which enabled persistence of the plasmid despite the *R6k* origin of replication. From the sequence data presented later in this chapter, the plasmid retention reduced to 15-36 per cent for the mariner libraries, *Table 3-9,* and 11-29 per cent for the Tn5 libraries, *Table 3-11*, when the donor was changed to *E. coli* MFD*pir+*.

Table 3-5 demonstrates that the donor strain MFD*pir+* is an improvement on ST18 in terms of conjugation efficiency, and therefore transposition efficiency for the mariner transposon system. The most important improvement is that the plasmid retention dropped from roughly 28 per cent to 2 per cent when changing the donor strain; this would result in purer final transposon mutant libraries.

Table 3-5 *Comparison of the estimated efficacy of the two donor strains ST18 and MFDpir+.*

		CFU/spt	Conjugation (%)	Transposition (%)	Plasmid Retention (%)
ST18 : BW25113	**LB**	7.27E+08	0.07	0.05	27.78
	Ampicillin	1.33E+05			
	Kanamycin	4.80E+05			
MFDpir+ : BW25113	**LB**	8.47E+08	0.20	0.20	1.95
	Ampicillin	3.33E+04			
	Kanamycin	1.71E+06			

Estimates were made based on colony counts on the specified media, using serial dilutions from the conjugation pools following 5hr incubation. This experiment was performed with the pSAM_Ec plasmid (mariner).

Plasmid persistence was confirmed by sequencing, where most reads contained transposon-plasmid junctions rather than transposon-chromosome junctions, *Table 3-6.* Changing the transposon plasmid donor strain resulted in fewer reads coming from the plasmid, indicating reduced plasmid retention for both the Tn5 and mariner systems used. Based on the data presented in *Table 3-5* and *Table 3-6* the donor strain was changed to E. coli MFD*pir+* for both transposon families.

Table 3-6 *Plasmid retention of ST18 versus MFDpir+ determined by sequence data.*

		Total Reads	Chromosome Mapped (%)	Plasmid Mapped (%)
ST18 : BW25113	**Mariner**	1.54E+07	15.60	44.90
	Tn5	2.22E+06	8.90	70.10
MFDpir+ : BW25113	**Mariner**	4.17E+06	39.74	13.88
	Tn5	1.92E+06	36.50	56.69

Plasmid retention was determined by running the BioTraDIS pipeline with the relevant plasmid sequence as the reference genome.

3.3.2.2 Mariner Transposon Library Generation

Mariner transposons are frequently used to generate TIS libraries in prokaryotic organisms. For this work a conjugation-based protocol with an established TIS plasmid was used to generate a library of *in vivo* transposition mutants.

3.3.2.2.1 Mariner Transposase Optimisation

In pSAM_Ec the *himar1C9* transposase gene was under a *lac* inducible promoter (Goodman et al., 2009) however, expression appeared to be constitutive. Adding IPTG as an inducer did not enhance transposition rates ; conversely adding glucose did not repress transposition as the literature indicated should happen (Griffiths et al., 2000). After comparing the promoter layout to other *lac* inducible promoters, I identified that an operator element was frequently present between the promoter and the start codon which would prevent transcription (Wiles et al., 2013). To enable control over transposase expression levels and therefore when transposition occurred, I cloned this operator sequence into the region before the start codon, *Figure 3-3.*

Figure 3-3 *The original layout of the promoter region of the Himar1C9 transposase in pSAM_Ec.*

A: An overview of the features as they are arranged in the plasmid. B: the original arrangement of the promoter region, zoomed. C: The modified promoter region of pSAM_Op showing the added operator element.

The optimised plasmid was named pSAM_Op, and trial experiments showed that the addition of 0.03 percent glucose (w/v) to the growth medium suppressed transposition, *Table 3-7.* Glucose prevented the constitutive expression previously seen with LB only and provided a transposition control mechanism. However, the addition of IPTG did not enhance successful transposition rates, *Table 3-7*, so further experiments were performed using the original pSAM_Ec plasmid.

Table 3-7 *Transposition and plasmid retention following conjugation using plasmid pSAM_OP.*

	Transposition (%)	Plasmid Retention (%)
LB	0.759	0.07
IPTG	0.744	0.09
Glucose	0.002	0.02

3.3.2.2.2 Estimated Number of Mutants in the mariner Libraries

The number of unique insertions was estimated as described in *section 2.5.6* and provided an estimate of unique insertions for an individual library within the batch, these estimations were recorded in *Table 3-8*. There was an overestimation in the number of unique insertions for library Batch 2 relative to the sequencing results where the individual libraries had around one third of the estimated unique insertions. But otherwise, the estimates were in the range that would be expected.

The number of mating spots was scaled based on the number of unique insertions in one mating spot, determined by preliminary sequencing before full library production, data not shown. Batch 3 had the fewest donor cells, but this did not lead to the fewest number of estimated unique insertions. As would be expected, Batch 1 demonstrated that recovering fewer cells from mating led to fewer expected unique insertions. Interestingly, despite fewer cells being recovered from the conjugation in Batch 1, the number of transconjugants was not the lowest but the transposon efficiency was predicted to be the lowest. The calculations indicated that there was growth of the recipient and transconjugants during the conjugation, this translated as one to four representations of every transposon insertion site, when plated across multiple outgrowth plates this gave detrimental mutants a chance to not be outcompeted and maintain representation in the final pool.

Table 3-8 *Estimations of transposon insertion efficiencies of the mariner library biological replicates.*

	Input CFU Per Mating		Recovered CFU per Mating			Doublings	Transconjugants	Transposon mutants	Unique Insertions	Plasmid Retention (%)	Transposition Efficiency
	Donor	Recipient	Donor	Recipient	Mating						
MARINER BATCH 1	2.50E+08	3.30E+08	7.20E+06	9.40E+08	5.00E+08	1.51	2.90E+06	2.87E+05	9.13E+04	91.03	5.53E-05
MARINER BATCH 2	2.40E+08	2.20E+08	1.08E+07	8.40E+08	8.60E+08	1.93	1.10E+06	9.2E+05	2.41E+05	16.36	2.19E-04
MARINER BATCH 3	8.30E+07	3.20E+08	9.40E+06	8.80E+08	7.60E+08	1.46	5.00E+05	3.50E+05	1.27E+05	30.00	7.95E-05

Estimations generated by the CFU recovered across LB agar supplemented with combinations of 0.3mM DAP, kanamycin sulphate 50 µg/mL and ampicillin sodium 100 µg/mL.

3.3.2.2.3 BioTradis Mapped mariner Insertions

The sequence data from each of the mariner libraries and the Batches were run through the BioTradis pipeline as described, the output is summarised in *Table 3-9*. The number of reads per library were similar within biological replicates except for library 8 which had half of the reads compared to library 7 and a quarter of the reads compared to library 9. The number of reads for each replicate was not indicative of the quality or density of a library and would have been due to inexact normalisation of molarities when pooling the samples prior to sequencing. Batch 1 had the lowest number of reads and Batch 3 had the highest, this could be attributed to longer storage at -80 ˚C and possibly indicated a reduction of DNA accessibility or integrity in the course of storage. Regardless, each library had sufficient reads originating from the transposon.

The percentage of reads mapped to the chromosome was lower than would be expected, in other work this has been as high as 70 per cent, (Yasir and Turner, personal communication, 2022). The level of plasmid persistence was lower than the estimations, *Table 3-8,* the aim was around ten per cent, the estimations in *Table 3-8* were not representative of plasmid retention. Closer to zero would have been ideal but was not achievable when using conjugation methods without multiple passages of the library. While this is an option for future investigations, the design of this experimental work was to minimise growth of the mutants. These results indicated that using the methods described in *section 2.5.1*, it was possible to repeatedly generate mariner transposon mutant libraries with a density of up to an insertion every 15 base pairs, this reduced to every nine base pairs for biological replicate pools. This further reduced to six base pairs when all mariner mutants were combined bioinformatically, giving a total of 7.42×10^5 unique insertions in the *E. coli* BW25113 chromosome.

Table 3-9 *A summary of the BioTradis mapping pipeline output for each of the mariner libraries and replicates and batch concatenations.*

LIBRARY	Total Reads	Transposon Reads		Chromosome Mapped Reads		Plasmid Mapped Reads		Unique Insertion Sites	Insertion Distance (bp.)	Average reads / Insertion
		Count	%	Count	%	Count	%			
MARINER_1	8.38E+06	7.74E+06	92.41	2.25E+06	28.99	1.13E+06	13.43	1.30E+05	35.75	17.33
MARINER_2	5.14E+06	4.77E+06	92.81	1.76E+06	36.77	6.12E+05	11.90	1.37E+05	33.81	12.81
MARINER_3	5.13E+06	4.74E+06	92.37	1.83E+06	38.68	5.79E+05	11.29	1.63E+05	28.49	11.27
MARINER_BATCH_1	1.87E+07	1.73E+07	92.51	5.83E+06	33.80	2.32E+06	12.42	3.07E+05	15.07	18.98
MARINER_4	9.40E+06	8.76E+06	93.12	4.01E+06	45.81	9.07E+05	9.65	2.27E+05	20.38	17.65
MARINER_5	1.35E+07	1.26E+07	93.28	6.64E+06	52.82	1.07E+06	7.94	2.83E+05	16.36	23.45
MARINER_6	1.08E+07	9.98E+06	92.54	5.41E+06	54.24	7.69E+05	7.13	2.97E+05	15.57	18.20
MARINER_BATCH_2	3.37E+07	3.13E+07	93.00	1.61E+07	51.31	2.75E+06	8.16	5.08E+05	9.11	31.61
MARINER_7	1.05E+07	9.74E+06	92.50	4.72E+06	48.46	9.56E+05	9.08	2.78E+05	16.65	16.97
MARINER_8	4.11E+06	3.81E+06	92.59	1.71E+06	45.04	4.00E+05	9.72	1.85E+05	24.99	9.25
MARINER_9	1.76E+07	1.62E+07	92.20	6.76E+06	41.63	1.75E+06	9.95	2.85E+05	16.27	23.75
MARINER_BATCH_3	3.22E+07	2.98E+07	92.35	1.32E+07	44.30	3.11E+06	9.63	4.77E+05	9.72	27.69
MARINER_ALL	8.46E+07	7.83E+07	92.64	3.51E+07	44.79	8.17E+06	9.66	7.42E+05	6.24	47.30

3.3.2.2.4 Potential mariner Insertion Sites in *E. coli* BW25113

It has been repeatedly reported throughout the literature that mariner family transposons have an absolute requirement for a TA dinucleotide at the site of insertion and that there is little sequence bias in the flanking regions (Cain et al., 2020.; Lampe et al., 1999; Miravet-Verde et al., 2020; Morris et al., 2016). In the reference genome used, *E. coli* BW25513 has theoretically 211,712 potential insertion sites for a mariner family transposon. This was calculated according to the requirement that the transposon can only insert into TA dinucleotide sites. Calculation of this is described in *section 3.2.2.10.* In practice, a saturated library is expected have fewer insertions than this number as some sites may be protected from insertion or be in an essential region of the genome and thus unable to tolerate insertions.

In my data, I did find that there was a strong bias towards insertion at TA dinucleotides, but that insertion was not exclusive to a TA dinucleotide as shown in *Figure 3-4.* Also, *Figure 3-4* demonstrates that whilst not as predominant as the insert site bias, the nucleotide immediately before and after the insertion site suggested a preference for either a Thymine or an Adenine.

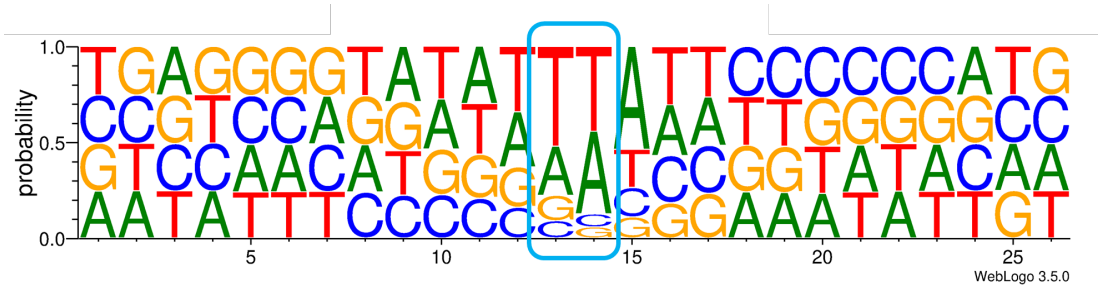

Figure 3-4 *A sequence logo plot of the insertion site using all of the mariner libraries.*

+/- 12 base pairs from the insertion dinucleotide, which is circled in blue. Constructed from every insertion present in the combined mariner library, mariner_All. Nucleotides are coloured for identification and their height denotes the probability of the base occurring at each location with the most probable at the top of each column and the least at the bottom. This was generated using WebLogo version 3.0.5 in Galaxy.

3.3.2.3 Transposon Library Generation – Tn*5*

For TISs that use Tn*5*, it is common for researchers to use the commercially available EZ:Tn*5* transposome kit (LGC Biosearch Technologies), where a purified Tn*5* transposase is incubated with a DNA transposon that has been generated by PCR or cloned into the supplied vector pMOD2. This transposome is then electroporated into the organism of choice (Goodall et al., 2018; Langridge et al., 2009).

As there was not an equivalent purified mariner transposase kit available, a conjugative approach was taken in this work. A plasmid containing both the transposase and relevant transposon was conjugated into *E. coli* BW25113 and transposition occurred *in vivo* upon transposase expression. While less commonly used, this is an approach that has previously been applied to gram negative organisms (Freed, 2017; Martínez-García, Aparicio, de Lorenzo, et al., 2014; Wiles et al., 2013), there are multiple vectors available for such applications. The Tn*5* mutant generation methodology was the same as discussed for the mariner transposon system and was chosen to generate comparable data between the two transposon families.

3.3.2.3.1 Tn*5* Transposase Optimisation

It was assumed that the conjugation efficiency was equivalent for the Tn*5* pBAMD1-2 and mariner pSAM_Ec plasmids (they are a comparable size with the same origin of transfer), Tn*5* transposition appeared to be less efficient than mariner based on the estimation calculations. This was concordant with the literature and reflected in *Table 3-8* and *Table 3-10.*

One factor potentially limiting Tn*5* transposition efficiency was the presence of a putative *lexA* binding site between the -35 and -10 elements of the T1 promoter region. This region is a weak binding site for the repressor protein and reports suggest that a mutation here can improve transposition efficiency by 2.6-fold. *LexA* is a global SOS repressor that blocks transcription of genes until a cell is stressed (Kuan & Tessman, 1991; Ross et al., 2014). If this was the case, then transposition would not occur unless the transconjugants were stressed and LB is not typically a stressful environment for a cell.

To see if this putative weak binding region had an effect on Tn*5* transposition rate, I replaced the sequence between the -35 and -10 boxes of the T1 promoter with the *lac* promoter from pSAM_Ec. *Figure 3-5* shows the changes to the promoter region that were made. The new plasmid was named pBAMDL and was evaluated for transposition efficiency compared to the native plasmid but there was no measurable increase in transposition efficiency so pBAMDL was not used for further experiments.

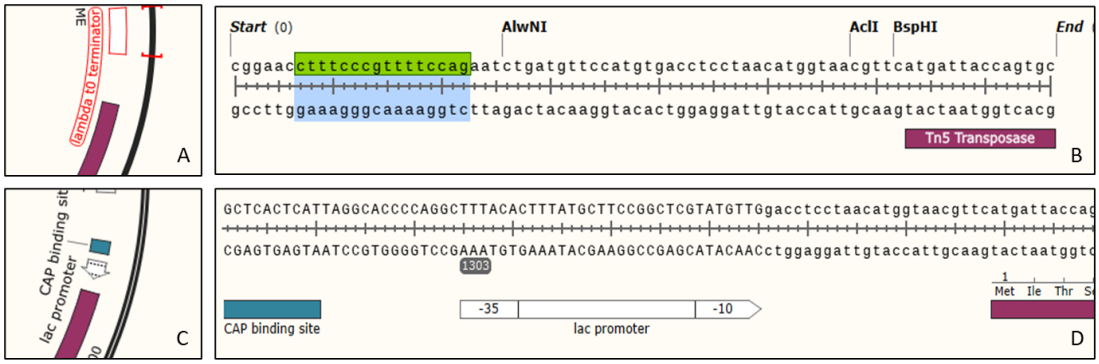

Figure 3-5 *The original promoter region of tnpA the Tn5 transposase in pBAMD1-2.*

A: showing the promoter region arrangement in pBAMD1-2 B: highlighting the putative lexA binding site in pBAMD1-2 C: The modified promoter region D: showing the replacement sequence in pBAMDL.

3.3.2.3.2 Estimation of Tn*5* Transposon Libraries Generated

The Tn*5* libraries were generated in a similar fashion to the mariner libraries and as described in *section 2.5.1*, and the number of expected unique insertions was estimated as described in *section 2.5.6*. The number of estimated insertions is recorded in *Table 3-10.* The colony counts for Batches 1 and 2 indicated that the libraries were scaled correctly to generate 5 x 10^5 unique insertion mutants. Library 2 showed a higher amount of plasmid retention, 27 per cent. Library Batch 3 estimations suggested that Tn*5* was 10-fold more efficient than the literature reports, in the region of 10^{-5} (Krebs & Reznikoff, 1988). As the estimated transposition rate was 10-fold higher than expected, this indicated that there may have been an error introduced during serial (10-fold) dilutions. Hence the calculation of the estimated number of unique insertions may have also been miscalculated by a factor of 10. In fact, from *Table 3-10, column 9* it can be seen that the estimated number of unique insertions was higher (around double) than for Batches 1 and 2; it would have been expected that the number of unique insertions for the libraries in Batch 3 would be around half of the estimation. The doublings or growth on the conjugation plate were comparable to the mariner libraries.

Table 3-10 *Estimations of transposon insertion efficiencies of the Tn5 library biological replicates.*

	Input CFU Per Mating		Recovered CFU per Mating			Doublings	Transconjugants	Transposon mutants	Unique Insertions	Plasmid Retention (%)	Transposition Efficiency
	Donor	Recipient	Donor	Recipient	Mating						
TN5 BATCH 1	1.20E+08	2.6E+08	4.00E+06	8.40E+08	6.00E+08	1.69	1.80E+06	1.66E+06	5.14E+05	7.78	3.95E-05
TN5 BATCH 2	4.20E+08	5.10E+08	3.80E+07	1.25E+09	3.80E+08	1.29	2.20E+06	1.60E+06	6.53E+05	27.27	2.56E-05
TN5 BATCH 3	4.60E+07	1.80E+08	2.4E+07	5.20E+08	5.00E+08	1.53	3.80E+06	3.44E+06	1.19E+06	9.47	6.62E-04

Estimations generated by the CFU recovered across LB agar supplemented with combinations of 0.3mM DAP, kanamycin sulphate 50 µg/mL and ampicillin sodium 100 µg/mL.

3.3.2.3.3 BioTradis Mapped Tn*5* Insertions

The output of the BioTradis insert mapping pipeline for the Tn*5* libraries is summarised in *Table 3-11*. The number of reads per library were similar within and between biological replicates with the exceptions of libraries 3 and 7, this is most likely to have been due to pipetting inaccuracies or estimated molarity calculations on a broad range of DNA sizes when pooling sequence preparations. There were a consistent number of reads across all three Batches. Around 84-87 per cent of the reads originated from a transposon-chromosome junction, this was lower than for mariner but likely due to differences in the custom transposon amplification sequencing primers used. Generally, more reads had mapped to the reference genome for Tn*5* than mariner, 55-59 per cent for Tn*5* and 29-54 per cent for mariner.

As mentioned previously, mapping percentages can be 70 per cent or above (Yasir and Turner, personal communication, 2022) but the Tn*5* libraries demonstrated a high rate of plasmid persistence, around 28 per cent except for library 9 at 11 per cent. When the proportion of reads mapping to the chromosome and plasmid were combined, for each library, the total mapping percentages were within the range expected. The plasmid persistence, *Table 3-11,* in library 9 (11.08 per cent) was in the region of estimations for Batches 1 and 3 (7.78 – 9.47 per cent) whereas all the others were in the range estimated for Batch 2 (27.27 per cent), *Table 3-10.* While there was generally more plasmid retention than would be acceptable for most TIS studies, there was evidence, library 9, that it was possible to generate libraries with low levels of plasmid contamination using these methods.

Table 3-11 *A summary of the BioTradis mapping pipeline output for each of the Tn5 libraries and replicates and batch concatenations.*

LIBRARY	Total Reads	Transposon Reads		Mapped Reads		Plasmid Mapped Reads		Unique Insertion Sites	Insertion Distance (bp.)	Average reads / Insertion
		Count	**%**	**Count**	**%**	**Count**	**%**			
TN*5*_1	3.62E+06	3.13E+06	86.32	1.74E+06	55.59	1.02E+06	28.23	6.26E+04	73.93	27.74
TN*5*_2	5.91E+06	5.02E+06	84.94	2.78E+06	55.45	1.59E+06	26.90	7.00E+04	66.15	39.73
TN*5*_3	8.79E+06	7.52E+06	85.53	4.16E+06	55.36	2.38E+06	27.04	1.00E+05	46.28	41.60
TN*5*_BATCH_1	1.83E+07	1.57E+07	85.50	8.68E+06	55.44	4.99E+06	27.23	1.98E+05	23.38	43.82
TN*5*_4	3.41E+06	2.90E+06	85.13	1.59E+06	54.80	9.81E+05	28.80	9.00E+04	51.47	17.66
TN*5*_5	4.26E+06	3.65E+06	85.85	2.16E+06	59.20	1.07E+06	25.22	9.56E+04	48.47	22.64
TN*5*_6	3.26E+06	2.82E+06	86.54	1.59E+06	56.54	8.92E+05	27.40	7.82E+04	59.23	20.38
TN*5*_BATCH_2	1.09E+07	9.37E+06	85.83	5.35E+06	57.04	2.95E+06	26.99	2.13E+05	21.73	25.08
TN*5*_7	7.45E+06	6.22E+06	83.51	3.68E+06	59.15	1.73E+06	23.23	9.75E+04	47.52	37.75
TN*5*_8	3.49E+06	2.99E+06	85.67	1.70E+06	56.90	9.39E+05	26.92	6.65E+04	69.70	25.58
TN*5*_9	3.35E+06	2.88E+06	86.10	1.64E+06	56.80	3.71E+05	11.08	6.53E+04	70.93	25.08
TN*5*_BATCH_3	1.43E+07	1.21E+07	84.64	7.02E+06	58.03	3.56E+06	24.91	1.75E+05	26.50	40.15
TN*5*_ALL	4.35E+07	3.71E+07	85.30	2.10E+07	56.69	1.15E+07	26.41	4.57E+05	10.13	46.04

3.3.2.3.4 BW25113 Potential Insert Sites for Tn*5* Transposition

Libraries made with a mariner transposon are expected to have insertion site requirements and these have been discussed in the literature (Cain et al., 2020; Lampe et al., 1999). However, libraries made with a Tn*5* transposon are generally assumed to be random and the transposon distribution to be uniform. My data suggested a slight bias towards Guanidine (G) and Cysteine (C) at the immediate insert site, seen in *Figure 3-6,* but not to the extent that has been suggested by other researchers (Goryshin et al., 1998; Green et al., 2012). The data presented here confirmed that Tn*5* was essentially random and did not confer any particular sequence site bias for insertion.

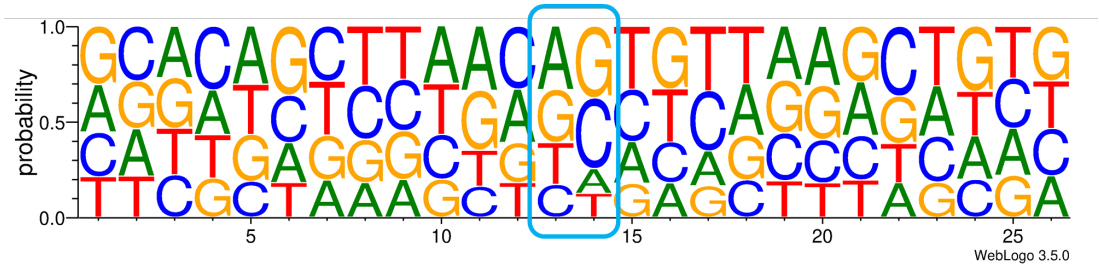

Figure 3-6 *A sequence logo plot of the insertion site of the Tn5 libraries.*

+/- 12 base pairs from the insertion dinucleotide, which is circled in blue. Constructed from every insertion present in the combined Tn5 library, Tn5_All. Nucleotides are coloured for identification and their height denotes the probability of the base occurring at each location with the most probable at the top of each column and the least at the bottom. This was generated using WebLogo version 3.0.5 in Galaxy.

3.3.3 Determining Gene Essentiality

3.3.3.1 Essential Genes Determined Using a mariner Transposon

All nine mariner libraries produced an essential gene list. The number of genes classed as essential ranged from 442 to 535 for the individual libraries, Table 3-12. Once batched into biological replicates, the number of essential genes reduced, and the library “All” had the fewest number of essential genes reported. This showed that increasing the number of unique insertion points increased the average insertion index and reduced the number of essential genes, as would be expected as transposon insertion saturation increases. Apart from mariner Batch 1, a similar trend could be seen with ambiguous genes.

Table 3-12 *The number of essential, ambiguous and non-essential genes for each of the mariner libraries.*

LIBRARY	CDS Unique Insertion Sites after Filtering	Average Insertion Index	Essential Changepoint	Ambiguous Changepoint	Essential Genes	Ambiguous Genes	Non-Essential Genes
MARINER_1	7.04E+04	0.0184	0.0026	0.0036	572	54	3728
MARINER_2	1.18E+05	0.0295	0.0036	0.0048	469	38	3847
MARINER_3	1.39E+05	0.0350	0.0081	0.0111	533	84	3737
MARINER_BATCH_1	1.80E+05	0.0450	0.0115	0.0151	468	64	3822
MARINER_4	1.32E+05	0.0332	0.0067	0.0090	460	61	3833
MARINER_5	1.66E+05	0.0415	0.0100	0.0134	478	70	3806
MARINER_6	1.74E+05	0.0437	0.1080	0.1420	465	81	3808
MARINER_BATCH_2	2.52E+05	0.0633	0.0152	0.0191	408	37	3909
MARINER_7	1.65E+05	0.0415	0.0105	0.0138	467	78	3809
MARINER_8	1.59E+05	0.0399	0.0082	0.0110	457	66	3831
MARINER_9	1.65E+05	0.0415	0.0103	0.0137	516	61	3777
MARINER_BATCH_3	2.36E+05	0.0593	0.0154	0.0196	428	39	3887
MARINER_ALL	3.05E+05	0.0767	0.0223	0.0274	403	36	3915

The changepoints recorded for each library determine the classification of each gene based on its insertion index, average insertion index for each library is also shown.

Figure 3-7A shows that there was a negative linear relationship between the average insertion index and the number of essential genes for libraries at the density of those reported here; this relationship was expected and suggested by *Table 3-12*. Analysis of variance (ANOVA) demonstrated that there was no real batch effect meaning that there were no differences between the biological replicates for mariner transposon libraries provided that there is sufficient transposon insertion saturation. This can be seen in the overlapping variances for the batches in *Figure 3-7.*

This saturation was achieved when the technical replicates for each biological replicate were combined, achieved with at least 1.80×10^5 unique insertions *Table 3-12.* The ANOVA was applied to the number of essential genes, p-value = 0.199; ambiguous genes, p-value = 0.627; non-essential genes, p-value = 0.060; and insertion index, p-value=0.4419. Box plots shown in *Figure 3-7B-E* demonstrate these variances. The variation between the technical replicates could not be assessed as for this work, each library was only sequenced once and repeat mappings would have been required for the analyses. However, when the mutant libraries were generated, each conjugation spot could have been considered a technical replicate, so pooling the mutants from multiple mating events to the insertion density required overcame variation in technical replicates.

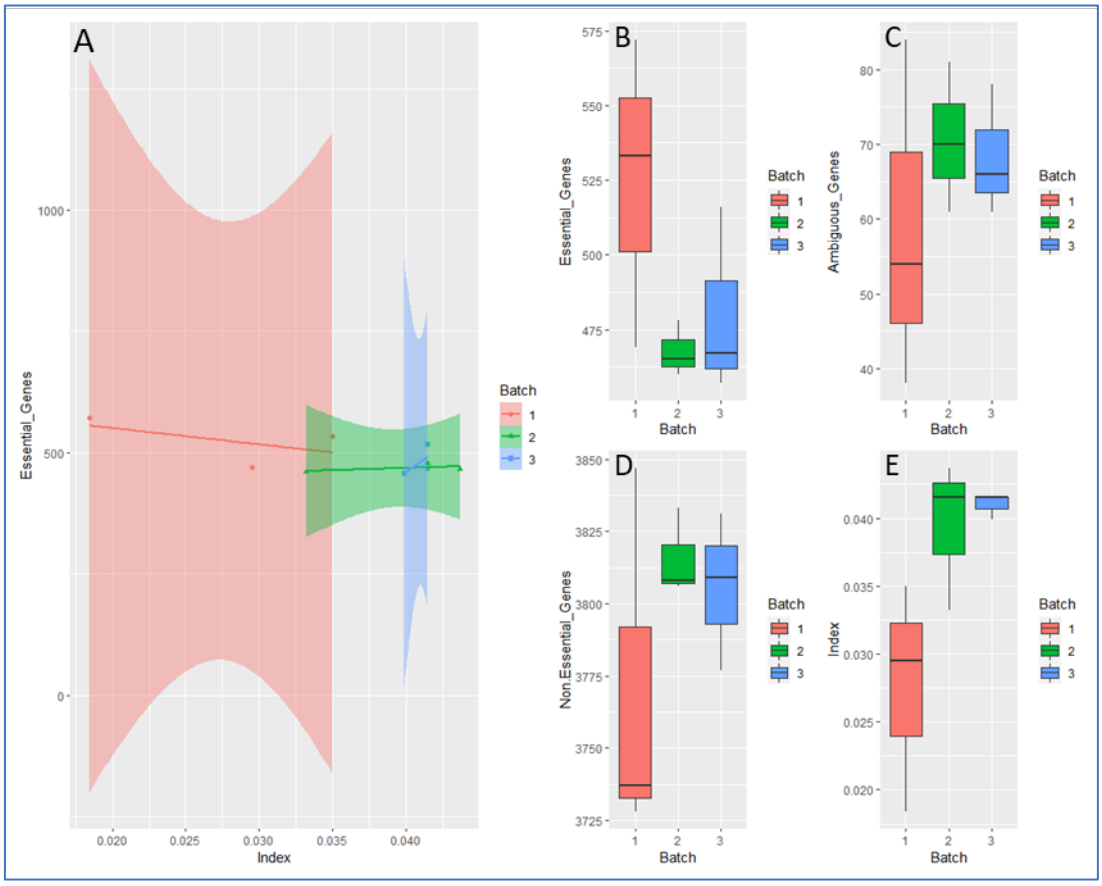

Figure 3-7 *Scatter plot of the number of essential genes determined by the average insertion index for each of the mariner libraries.*

A: The linear regression trendline is included with the area around each representing the variance amongst biological replicates, the colours represent biological replicate groups; red is Batch 1, green is Batch 2 and blue is Batch 3. B-E: Boxplots showing the average and variance in the essential genes, ambiguous genes, non-essential genes, and average insertion index, respectively. Coloured by biological replicate.

When the gene lists produced for each batch of libraries were compared, the majority, 400 (out of 532, 445 and 467 for Batches 1, 2, 3 respectively *Figure 3-8*), were found as a consensus in all three batches, demonstrated in *Figure 3-8*. Mariner Batch 1 appeared to have the most genes reported unique to that biological replicate, this fitted the trend mentioned above as this library had the lowest number of unique insertions and consequently the lowest average insertion index. Interestingly, some individual libraries and all of the biological replicate batches exceeded the predicted maximum number of possible insertions based on TA sites generated in *section 3.2.2.10*; this will be discussed further in *chapter 4*.

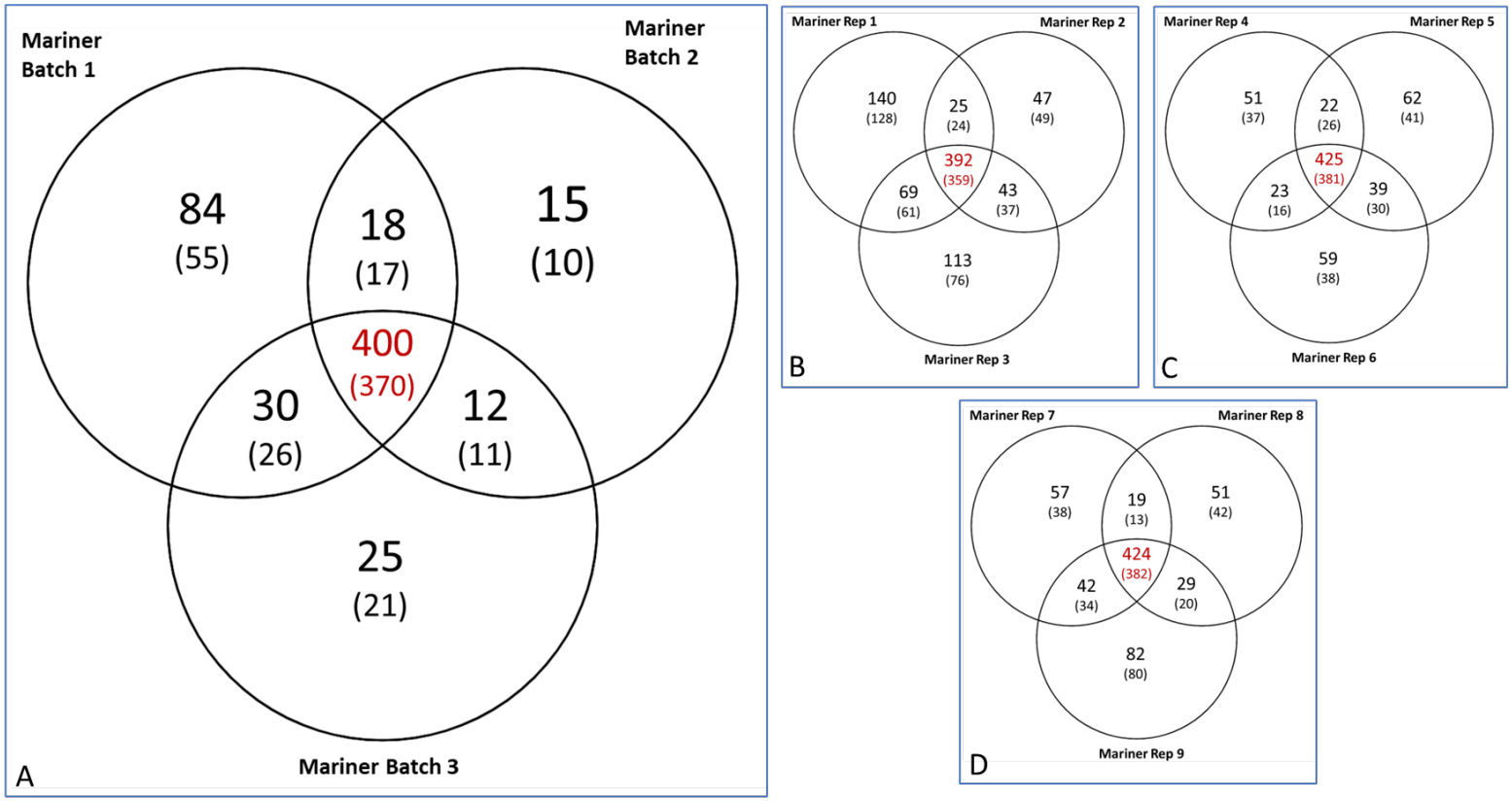

Figure 3-8 *Venn diagrams showing the overlap in the essential and ambiguous genes from replicates of mariner libraries.*

A: The overlapping essential plus ambiguous genes determined for each biological replicate (batch), the numbers in brackets represent the number of essential genes only. B-D Venn diagrams to show the overlap of genes reported as essential in the technical replicates within the biological replicates for Batches 1-3 respectively. Created in Venny 2.1.0.

3.3.3.2 Essential Genes Determined Using a Tn*5* Transposon

From *Table 3-13*, it is apparent that essential gene determination using a Tn*5* transposon was more variable than using the mariner family transposon, with the number of essential genes ranging from 448 to 2165 essential genes reported for individual libraries. Library 4 produced the lowest number of essential genes and had the second most insertion sites, whereas library 9 produced the most essential genes and did have the fewest insertion sites. This supported the generalisation that increasing the number of unique insertion sites gives a more refined list of essential genes. Batch 1 had around 30 per cent more unique insertions than library 4 yet produced more essential genes and ambiguous genes, this demonstrated that while the number of insertion sites in a library was important, the distribution of insertions across the genome was also contributed to the determination of essential genes.

Whilst the variation amongst the number of essential genes determined across the individual libraries was much larger than the mariner libraries, within biological replicates the consensus number of essential or ambiguous genes was similar. *Figure 3-9A* shows the overlap of essential genes predicted for all three biological replicates, there were 396 essential, or ambiguous, genes identified in all three biological replicates (out of 697, 706 and 1119 for Batches 1, 2, 3 respectively, *Table 3-13 columns 5 and 6*), this was in the same range as the number of essential genes from previous studies (Baba et al., 2006; Ghomi et al., 2022; Goodall et al., 2018; Yamazaki et al., 2008). *Figure 3-9A* shows Batch 3 was the most different from the other Batches, with the most genes unique to that biological replicate, most likely because the number of essential, or ambiguous, genes determined for this group was nearly double that of the other two replicate groups.

Table 3-13 *The number of essential, ambiguous and non-essential genes for each of the Tn5 libraries.*

LIBRARY	CDS Unique Insertion Sites after Filtering	Average Insertion Index	Essential Changepoint	Ambiguous Changepoint	Essential Genes	Ambiguous Genes	Non-Essential Genes
TN*5*_1	2.15E+04	0.0057	0.0001	0.0002	1157	0	3197
TN*5*_2	2.35E+04	0.0063	0.0001	0.0003	1013	2	3339
TN*5*_3	3.15E+04	0.0082	0.0010	0.0017	896	232	3226
TN*5*_BATCH_1	6.30E+04	0.0165	0.0015	0.0023	521	178	3655
TN*5*_4	4.07E+04	0.0107	0.0001	0.0003	448	1	3905
TN*5*_5	4.46E+04	0.0118	0.0011	0.0020	528	198	3628
TN*5*_6	3.47E+04	0.0091	0.0002	0.0004	588	3	3763
TN*5*_BATCH_2	6.91E+04	0.0182	0.0021	0.0037	493	213	3648
TN*5*_7	2.89E+04	0.0075	0.0027	0.0046	2041	545	1768
TN*5*_8	2.25E+04	0.0060	0.0019	0.0033	2091	529	1734
TN*5*_9	2.09E+04	0.0055	0.0017	0.0030	2165	489	1700
TN*5*_BATCH_3	4.69E+04	0.0124	0.0009	0.0018	816	305	3233
TN*5*_ALL	1.14E+05	0.0299	0.0016	0.0028	351	103	3900

The changepoints recorded for each library determine the classification of each gene based on its insertion index, average insertion index for each library is also shown.

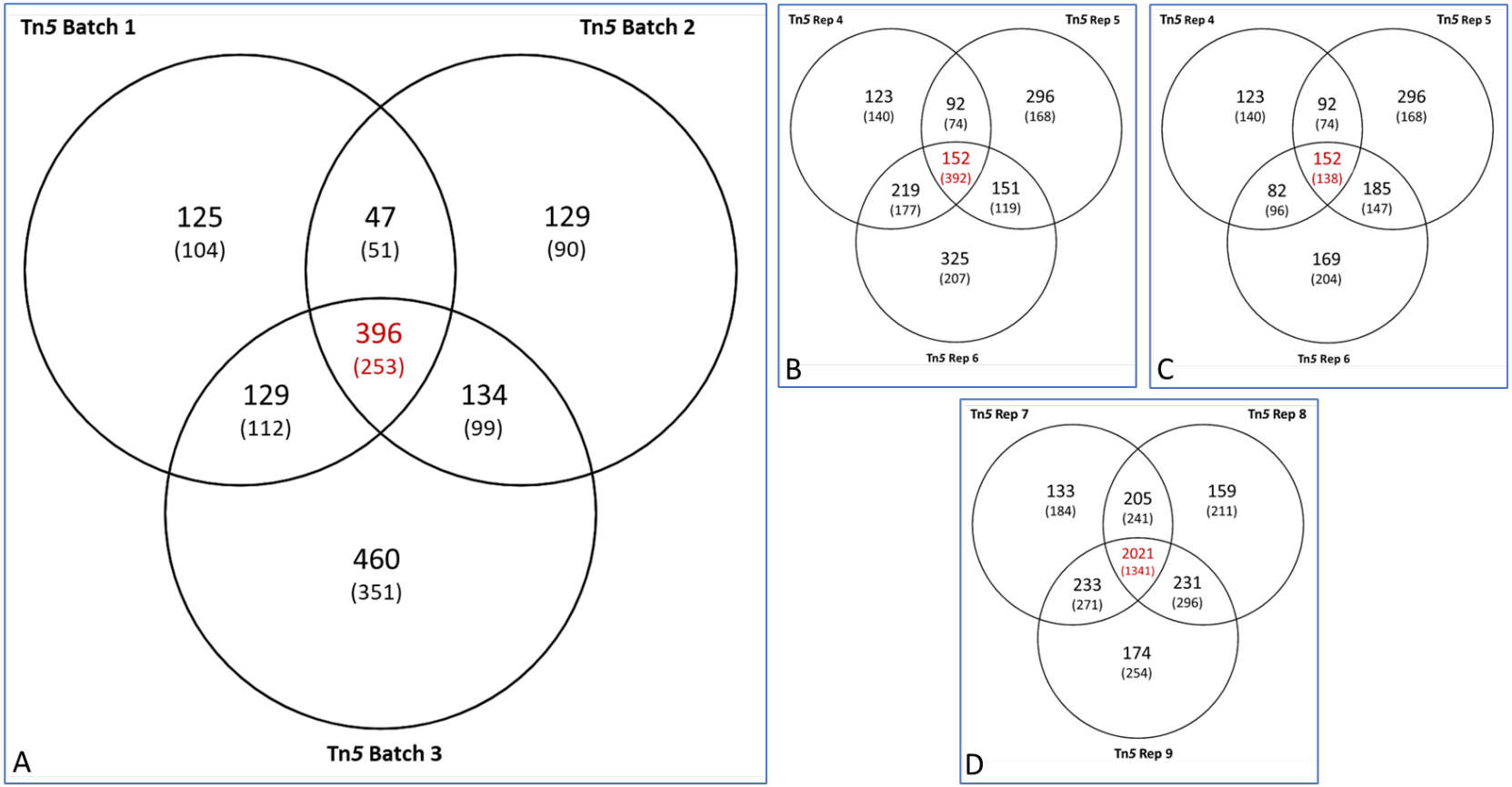

Figure 3-9 *Venn diagrams showing the overlap in the essential and ambiguous genes from replicates of Tn5 libraries.*

A: The overlapping essential plus ambiguous genes determined for each biological replicate (batch), the numbers in brackets represent the number of essential genes only. B-D Venn diagrams to show the overlap of genes reported as essential in the technical replicates within the biological replicates for Batches 1-3 respectively. Created in Venny 2.1.0.

Figure 3-9D suggests that the technical replicates within Batch 3 were the most similar in terms of the percentage of genes called essential or ambiguous, with 64 per cent of genes identified being in all three. This value fell to 23 per cent and 14 per cent for biological replicates Batch 1 and Batch 2 respectively, Figure 3-9B and C. This could indicate that the technical replicates within biological replicate Batch 3 were more similar; alternatively, this could have been due to the vast number of genes reported as essential or ambiguous in those libraries. As with the mariner technical replicates, there were not enough sequencing replicates of the technical replicates to evaluate the relatedness.

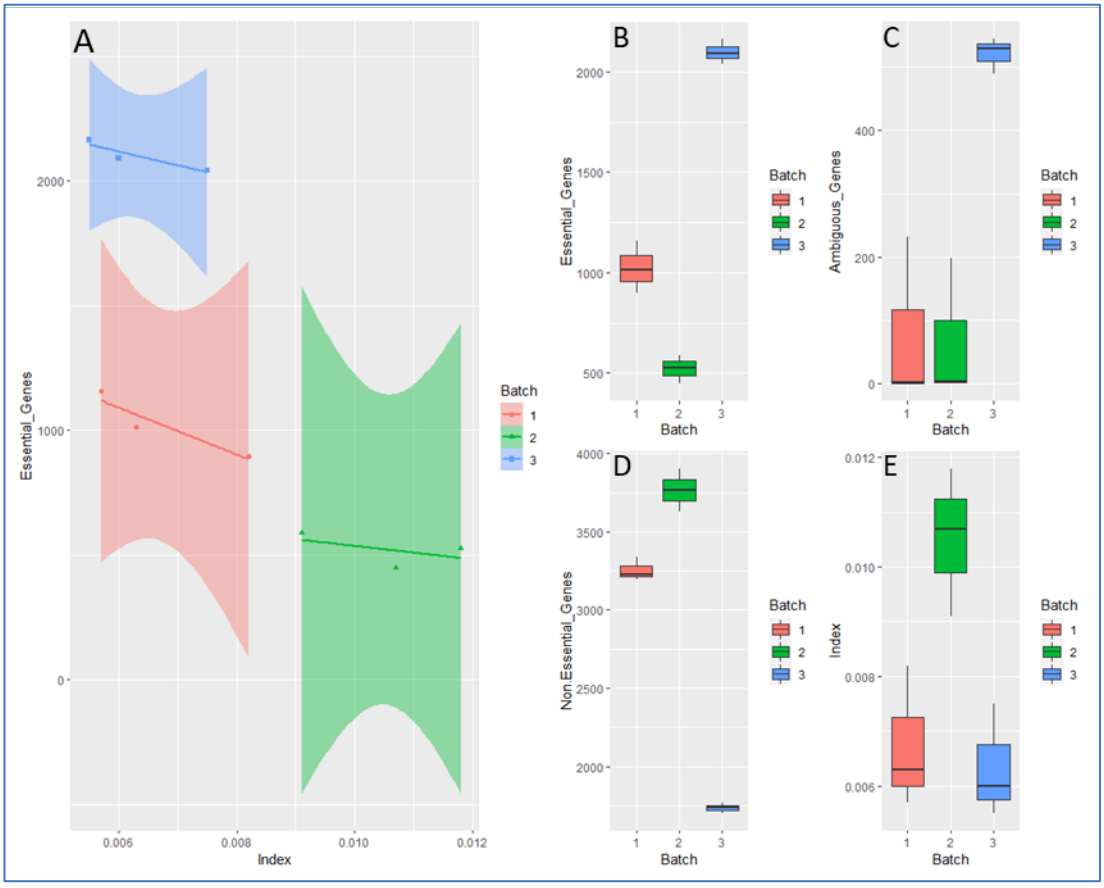

Figure 3-10 *Scatter plot of the number of essential genes determined by the average insertion index for each of the Tn5 libraries.*

A: The linear regression trendline is included with the area around each representing the variance amongst biological replicates, the colours represent biological replicate groups; red is Batch 1, green is Batch 2 and blue is Batch 3. B-E: Boxplots showing the average and variance in the essential genes, ambiguous genes, non-essential genes, and average insertion index, respectively. Coloured by biological replicate.

Unlike the mariner libraries, there was a batch effect observed with the biological replicates of the Tn*5* libraries, this is clearly seen in *Figure 3-10A* where the technical replicates and their associated variance form distinct populations as replicates. This batch effect is further demonstrated in *Figure 3-10B-E* showing a statistical difference between the biological replicates for the essential genes reported p-value = 2.26×10^{-6}; the ambiguous genes reported, p-value = 0.00247; the non-essential genes reported, p-value = 4.56×10^{-7}; and the average insertion index, p-value = 0.111.

3.3.3.3 Differences in Essential Genes Determined by either a Tn*5* or a mariner Transposon

When all of the libraries were pooled per transposon, giving the maximum library density available in this work, there were a similar number of essential genes reported, 403 and 351 for mariner and Tn*5,* respectively. When the ambiguous genes were added, the two transposons reported 439 and 454 essential or ambiguous genes. This would suggest that TIS was reliable for determining the majority of essential genes regardless of transposon used. However, when the lists of genes were compared there was an overlap of only 325 genes, *Figure 3-11*. Also, 26 per cent of the genes designated essential by the mariner transposon libraries were not for Tn*5* and 29 per cent of genes determined by the Tn*5* transposon were not for mariner. This confirmed that at this insertion density, the two transposons used did not provide a consistent gene list between them despite initial similarities in the number of essential genes reported.

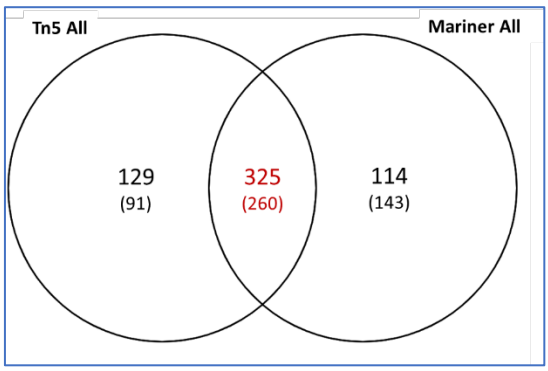

Figure 3-11 *Venn diagram showing the overlap in the essential and ambiguous genes from all of the mariner and Tn5 combined libraries.*

The numbers represent the total number of genes that have not been designated as non-essential, the numbers in brackets are for essential genes only. Created in Venny 2.1.0.

The libraries produced for each transposon differed and there was much more variation in the Tn*5* libraries in terms of insertion index and essential gene determination. *Figure 3-12A* shows that Tn*5* and mariner libraries formed two distinct populations, demonstrating that insertion densities and essential gene lists produced for the two transposons were statistically different. A key factor for the difference between the two transposons could be that there were almost three times as many insertions in coding sequences (CDS) for the combined mariner libraries (3.05×10^5) than the combined Tn*5* libraries (1.14×10^5); this could also be a reason for increased variance observed in the Tn*5* libraries.

The outputs of each library were different between the transposons and can be seen in *Figure 3-12B-D*, where the variability amongst the Tn*5* libraries was greater than for mariner and there was a difference in the essential genes reported, p-value = 0.00714, and non-essential genes reported, p value = 2.64×10^{-8}. However, this was not the case for the ambiguous genes reported, p value = 0.0702. *Figure 3-12E* shows that the average insertion index for the mariner libraries was greater than for the Tn*5* libraries, p value = 0.0111. As the insertion index is determined by the number of insertions per gene it is linked to library saturation.

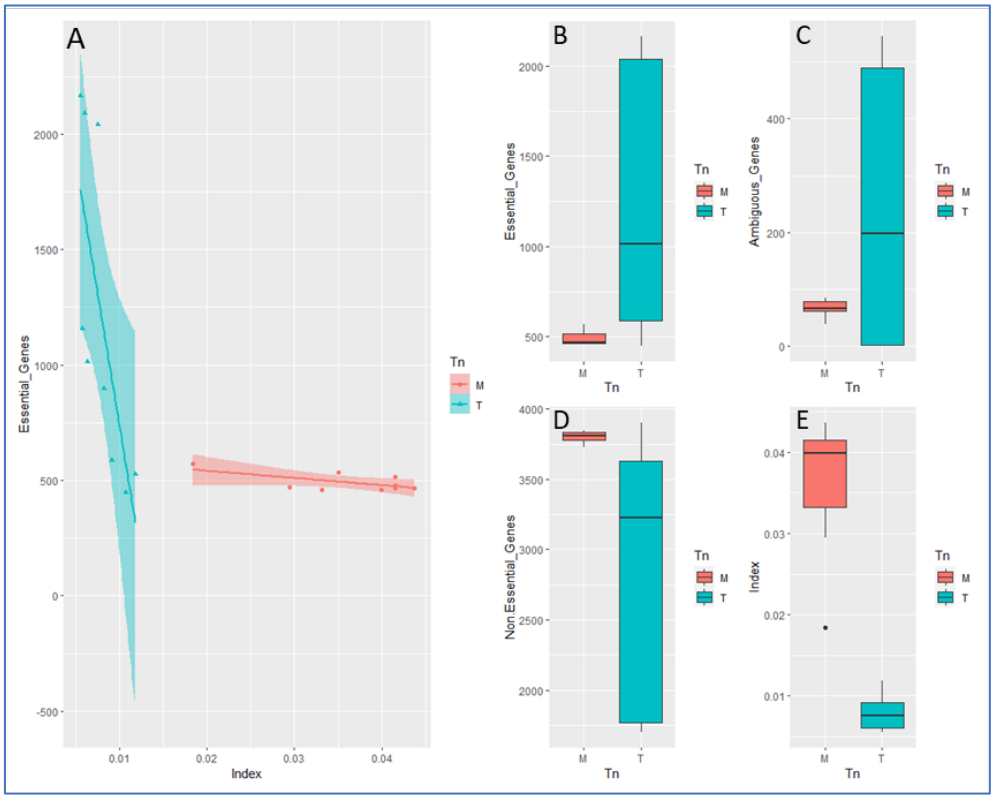

Figure 3-12 *A scatter plot of the number of essential genes determined and the average insertion index for all the mariner libraries and all the Tn5 libraries.*

A: The linear regression trendline is included with the area around each representing the variance amongst biological replicates, the colours represent the transposon family; red (M) is mariner and, blue (T) Tn5. B-E: Boxplots showing the average and variance in the essential genes, ambiguous genes, non-essential genes, and average insertion index, respectively.

For both of the transposons used, library saturation was investigated by sequentially concatenating the sequence data from the individual transposon libraries until all nine were incorporated. These files were run through the BioTradis pipeline and essentiality determination as described in *sections 3.2.2.4-3.2.2.9* and *3.2.3*. The unfiltered unique insertions, unique insertions within CDS, essential genes and ambiguous genes were plotted and are shown in *Figure 3-13.*

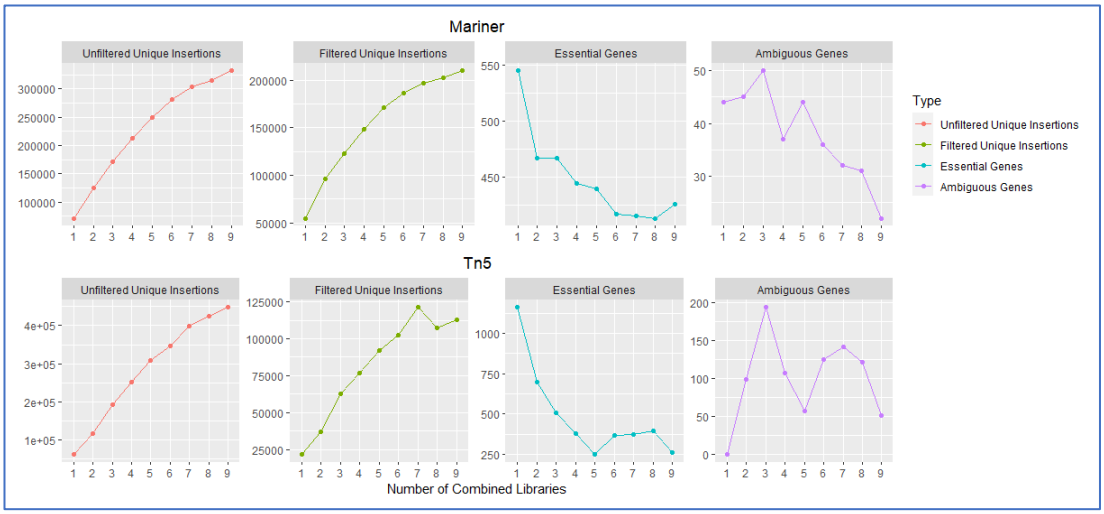

Figure 3-13 *Line plots to investigate the effects of increasing insertion number towards saturation.*

Top: Mariner, Bottom: Tn5. Split to show the changes in the number of unfiltered unique insertions, filtered insertions[1], essential genes and ambiguous genes designated using the BioTradis pipeline. Consecutive libraries were concatenated to observe whether library saturation had been achieved. [1] Insertion counts were filtered to remove noise, insertions with fewer than 1/10th of the average insertion expected per site based on sequence read depth was replaced with zero.

Assuming that saturation is achieved when all permissive insertion sites contain a transposon, the curves in *Figure 3-13,* especially for unfiltered insertion sites and filtered insertion sites show that saturation had not been achieved for libraries generated in this work and not even if all of the replicates for each transposon were pooled. A plateau in the number of insertions would have demonstrated saturation. Another interpretation of saturation could be that there are sufficient insertions that the number of essential genes determined is unchanged, again demonstrated by a plateau. The data here was inconclusive, for the mariner libraries the number of essential genes began to level out until the ninth library was added. The essential genes determined by the concatenated Tn*5* libraries decreased, suggesting increasing saturation until library addition six where the number of essential genes reported increased.

3.3.3.4 Essential Genes Compared to a Knockout Collection

Although individual Tn*5* and mariner libraries are different and some of this could be attributed to insertion density differences, the Keio collection (Baba et al., 2006) does not have any insertion density dependencies. The essential, or ambiguous, gene lists produced, from the mariner and Tn5 transposons in this work were compared to the Keio knockout

collection (Baba et al., 2006). There were 235 genes that were identified across all and a further 38 identified as essential in the Keio collection and by only one of the two transposons. The overlaps in the lists of essential genes are visualised in *Figure 3-14*; there were 90 genes identified by TIS only and 27 identified by the Keio collection only.

Interestingly, in the list of genes exclusive to TIS compared to Keio, the genes atpABCDEGHI, not atpF were identified as essential. These form the ATP synthase F complex for H+ transport across the membrane (Senior, 1990). Both the Keio collection and the Goodall library state these genes to not be essential for growth in LB (Baba et al., 2006; Goodall et al., 2018). Baba et al. also grew their library on MOPS supplemented with 0.4% glucose, they found that all nine genes were not essential. Feist et al. used a metabolic reconstruction of MG1655 and identified only three of these genes as essential (Feist et al., 2007). Joyce et al. grew the Keio mutants on M9 supplemented with 1% glycerol to find metabolic differences and found six of the nine genes to be essential (Joyce et al., 2006). The libraries used in this work were cultivated on LB yet when looking at this operon the essential gene profile was more similar to experiments conducted when mutants were grown on glycerol or with limited glucose. Both transposons had determined eight of the nine genes to be essential. This may suggest that the protocols used here were limiting access to glucose or growing cells in a culture that was too dense. Alternatively, the libraries from this study were stored at -80˚C in 20% glycerol prior to DNA extraction and sequencing, glycerol is common across the studies.

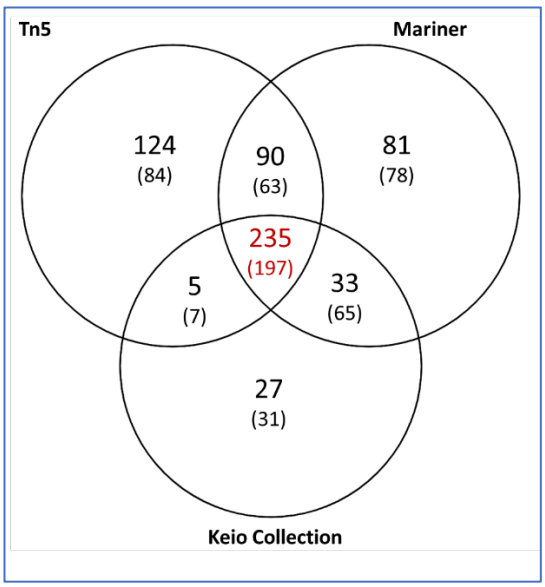

Figure 3-14 *Venn diagram showing the overlap in the essential and ambiguous genes from all of the mariner, Tn5 libraries and the Keio collection.*

The numbers represent the total number of genes that have not been designated as non-essential, the numbers in brackets are for essential gene only. Ambiguous genes were only reported for mariner and Tn5 as Keio does not have that category. Created in Venny 2.1.0.

If the Keio collection is assumed the gold standard (Ghomi et al., 2022), then the data here suggest that the TIS libraries generated in this work fail to identify up to ten per cent of essential genes and over call 30 per cent as essential. However, TIS should not be dismissed, the Keio collection was constructed based on an annotation identifying 4296 CDS with 74 pseudogenes and did not target seven already disrupted genes or *ldrABCD*.

3.3.3.5 Essential Genes Compared to a Published Tn*5* Library

Goodall et al. published the results from their high density Tn*5* transposon mutant library, with around 900,000 unique insertion mutants, and an average insertion index of 0.195 (Goodall et al., 2018). When the essential, plus ambiguous, genes reported for the two transposons used in this work were compared with the same output from the Goodall library, there was an overlap of 297 genes, seen in *Figure 3-15*. However, there were still around one third of genes unique to both of the Tn*5* transposon libraries in this comparison. In this comparison, the mariner library appeared to be the least different from the published data and had more genes in common with the Goodall library than was observed between the two Tn*5* datasets.

The list of essential genes reported by Goodall et al. did not include the ambiguous genes and these are reported in the supplemental data as undetermined. When these undetermined or ambiguous genes are removed from the comparison between transposons the Goodall data remained unexpectedly more closely related to the mariner library than the Tn*5* library. There were fewer overlapping essential genes listed for all the libraries.

One gene identified exclusively in the Goodall data was *ybfQ,* an inactive transposase (Keseler et al., 2017); if the gene is inactive then it should not be essential, so this suggested either there was an active gene product, the genome was unfavourable for transposon insertion or that the statistical approach used was not appropriate.

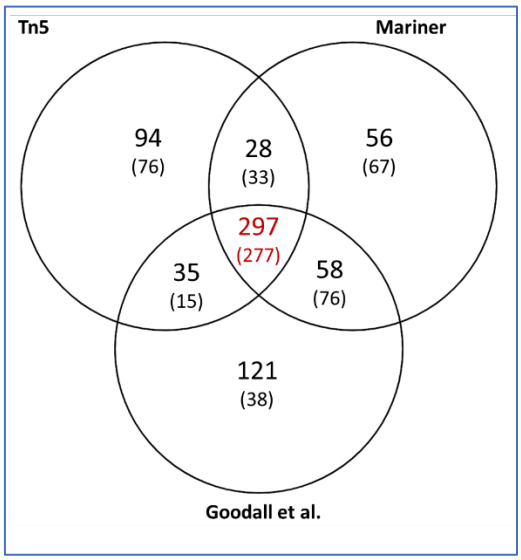

Figure 3-15 *Venn diagram showing the overlap in the essential and ambiguous genes from all of the mariner, Tn5 libraries and the Goodall el al. library.*

The numbers represent the total number of genes that have not been designated as non-essential, the numbers in brackets are for essential gene only. Ambiguous genes were reported for mariner and Tn, for the data from Goodall et al the ambiguous genes are referred to as undetermined. Created in Venny 2.1.0.

It was expected that the two Tn*5* libraries would be the most similar and previous data reported in this chapter suggested that as the density of a library increases the more refined the essential gene list became. The Goodall mutant set did not follow this trend (Goodall et al., 2018).

3.4 4. Discussion

3.4.1 Generating Transposon Libraries

The results presented here demonstrate that the methods described can be used to generate multiple and scalable TIS libraries for both mariner and Tn*5* family transposon systems. Conjugation methods using the relevant transposase plasmids generate relatively unbiased transposon mutant libraries with transposition efficiencies at similar levels to those stated in the literature (Krebs & Reznikoff, 1988; Lampe et al., 1999). The modified Illlumina DNA sequencing protocol described is effective in enriching and sequencing Tn-Chr junctions for multiple transposon families and is compatible with any normal Illumina WGS prep without the need for dark cycles so can be run with routine WGS sequencing. The chromosomal mapping percentages for the mariner and Tn*5* libraries were lower than previously seen however, Miravet-Verde et al. suggest that longer reads generate better mappings. The sequencing protocol for this work was a 75-cycle single-end run as previously described in *chapter 2,* in future this could be increased to a paired-end 150 cycle run. Preliminary data to guide library production showed increased mapping percentages and was run as 150 bp paired end reads.

The mariner transposon used here primarily inserts at TA dinucleotides as reported, but not exclusively. A study using mariner insertion sites in *Mycobacterium tuberculosis* libraries (Choudhery et al., 2021) report the probability of a Guanidine (G) or Cysteine (C) to be the first nucleotide to follow the insertion TA, to be around 70 per cent, and possibly indicating an insertion site preference beyond the TA dinucleotide. My data suggest that in *E. coli* K12, these occur at around 65 per cent. Choudhery et al. observe that this increases to around 90 per cent for low frequency insertion sites so conclude that a G or C flanking the TA dinucleotide is preventative to insertion. The frequency of insertion at different sites was not investigated in this chapter, a more detailed investigation into specific insertion site patterns and probabilities is described in *chapter 4*.

The Tn*5* transposon used in this work shows no real preference for any obvious insertion site, though it could be argued that there is a slight preference towards GC as is reported in the literature (Green et al., 2012) and others. However, the authors did not normalise the Tn*5* data (as they had done with other transposons), as such the GC insertion site bias may be less evident in this case.

3.4.2 Are Replicates Required for Essentiality Determination?

When comparing biological and technical replicates of transposon library generation, the technical replicates appeared more similar than the biological, but this was not statistically proven. For the mariner libraries there was no statistical difference in either the insertion index or the essential genes determined, yet there was for the Tn*5* libraries. Therefore, when constructing a transposon mutant library, it is prudent to pool mutants generated at several timepoints rather than to increase the repeated number of mating spots on any one day to achieve the library density that the study requires.

The work presented here has demonstrated that this is more important when using a Tn*5* based preparation method. It is plausible that these larger differences could be due to minor but unavoidable changes on any one day. For example, the exact concentration of antibiotic on a batch of plates, the slight temperature difference in incubators (repeated opening, room temperature), differences in media preparations. Here the batch effect has only been investigated with conjugation as a method for transposon library generation and it should be noted that mutant libraries are frequently constructed with purified transposases (Goodall et al., 2018; Holden et al., 2021; Langridge et al., 2009).

Goodall et al. used the purified Tn*5* transposase approach and suggest an alternative approach to determining gene essentiality, using the insertion index to calculate a length of the genome that would be expected to be insertion free and then use that metric to determine whether a gene has an under or over representation of insertions (Goodall et al., 2018). This is different to the BioTradis essentiality algorithm used in this work which uses probability distributions to estimate the probability of a gene belonging to an essential or non-essential population based on the insertion index (number of unique insertions divided by the gene length). Analysis of the transposon mutant libraries in this work has indicated that there is an association between the number of unique transposon insertions and the number of essential genes reported by the BioTraDIS pipeline. However, it has been demonstrated that the location of a transposon insertion is a contributing factor. This is included in the Goodall et al. analysis where the expected distance between insertion sites is included in essentiality determination.

3.4.3 The Impacts of Library Saturation

In this work, the Tn*5* libraries were the most variable at reporting essential genes. Interestingly the total number of unique insertions for Tn*5* was 4.57 x 10^5 yet, only around one quarter occurred within an annotated CDS (1.14 x 10^5). For mariner, the total number of unique insertions was 7.42 x 10^5 and around half occurred in an annotated CDS (3.05 x 10^5). This would suggest that Tn*5* is more likely to insert into intergenic regions of the chromosome, intergenic insertions are disregarded when using the BioTradis pipeline to determine gene essentiality. Therefore, the Tn*5* library is considered to be at lower saturation than the mariner library when looking at the proportion of permissive insertion sites containing a transposon (Chao et al., 2016).

Hypothetically, there is a minimum saturation point of a transposon library required to give the most comprehensive essential gene list using the BioTradis pipeline. The work here has been unable to define this for when a transposon mutant library will produce a robust essential gene list. It has shown that these metrics (insertion index and number of unique insertions) alone are insufficient, demonstrated mostly by the Tn*5* libraries where in Batch 3, the fewest unique insertion sites report around 30% more genes than Batch 1 and Batch 2. For both transposons there is an inverse trend where fewer genes are reported as essential with the greater number of unique insertion points. For the mariner library this inverse relationship is mirrored with the number of ambiguous genes. For the Tn*5* libraries it is not and is most likely due to the more promiscuous transposon. Attempts to define whether the libraries generated had reached saturation were inconclusive; when plotting the increase in unique insertion sites as more libraries are combined the curve did not plateau so suggests that genome saturation had not been achieved. When using the number of essential genes to determine saturation, the curve did flatten so would indicate that saturation was achieved but then increased again. Whereas, using the number of ambiguous genes did not show any trends.

The steady reduction in essential genes as more insertions are used to determine essentiality and then an increase is most likely to be a limitation of the statistics underpinning determination. The number of ambiguous genes determined is more variable in the pattern and this is also most likely due to the essentiality determination process. Ambiguous genes have an insertion index that falls between the essential and non-essential distributions so cannot be confidently assigned to either population (Barquist et al., 2016).

Some of the differences between the essential genes listed for the two transposons could be due to the design of the transposons within the plasmid. They are identical except that the 3' end of the mariner transposon has transcription terminators to prevent activation of downstream genes. When genes are organised as operons, one promoter can initiate transcription of all of the genes, if the mariner transposon inserts after this, it will inactivate gene expression of the remaining operon. If any of the genes downstream are essential for growth, then the mutant will die; this is considered a polar effect and can lead to whole operons being identified as essential.

Conversely, Tn*5* does not have transcriptional terminators so there is the potential for induction of downstream genes adjacent to the transposon insertion. This can cause expression or over expression under conditions that would not otherwise occur. This can lead to misclassification of essential genes, for example expression of a toxin (Hutchison et al., 2019).

Both the BioTradis essentiality algorithm (Barquist et al., 2016) and the approach taken by Goodall et al. (Goodall et al., 2018) led to a list of ambiguously essential or undetermined essential genes. These should not be confused with conditionally essential genes whose essentiality is dependent on the conditions being tested and under different stresses this list would change (Khatiwara et al., 2012). Some of the differences between the essential gene lists produced by the Goodall et al. data and data generated from this work can be explained due to the differences in mutant library cultivation, the Goodall et al. mutants were selected with chloramphenicol compared to Kanamycin in this work. Any genes that are essential to growth on either medium (or the resistance mechanism introduced during transposition) would be determined as essential in the relevant data.

Alternative analysis pipelines may find that the Tn*5* and mariner transposon systems used in this work are not statistically different based on the libraries and insertion density presented here, this is an issue that needs to be addressed as there is no standardised approach to determining gene essentiality for one of the easiest to manipulate and extensively studied model organisms.

TIS can be used to detect potential resistance mechanisms to antimicrobials. A robust essential gene list prior to treatment with the compound is needed to ensure that a pre-treatment essentiality screen correctly identifies essential genes. Under stress, or treatment conditions, a second screen highlights genes that are required for survival under the stress and the two lists are compared to find genes that are conditionally required

under the stress, highlighting potential drug targets. If the essential gene lists between treated and untreated or increasing concentration of antimicrobials show differencing numbers of essential genes, this can indicate evolution towards tolerance and in some cases resistance mechanisms if multiple passages are analysed. If the original list of essential genes is not robust then development of antimicrobials is more likely to result in a failed investment either through the wrong target (not an essential gene) or due to rapid evolution towards resistance.

3.5 Conclusions

Gene essentiality determination is closely linked to the accuracy of genome annotation, therefore as annotation becomes more extensive, the essential genes listed will become more accurate. The difference in essential genes reported by the BioTradis essentiality pipeline was reproducible for the mariner transposon libraries but not the Tn*5* transposon libraries, and the variability was associated with the number of unique insertions within CDS rather than intragenic regions.

Rather than biological and technical replicates, it is more important to replicate independent mating procedures to generate a diverse final pool of mutants and perform multiple sequencing reactions to obtain the maximum depth of coverage. The essential gene lists produced with either the mariner or Tn5 libraries were comparable but not congruent indicating that there were differences in transposon insertion between the two families. These differences, along with other factors affecting transposon library generation insertion will be explored further in *chapter 4.*

4 Defining Bias in the Use of Tn*5* and mariner Transposons for Determining Protected Regions of the *E. coli* Genome

4.1 Introduction

Chapter 3 demonstrated that the insertion frequency and distribution of transposon insertions can affect the reliability of essential gene prediction for TIS. This chapter will explore potential factors affecting variability of transposon insertion within a genome. Biases can be considered as either methodological or biological. Methodological biases are considered as arising from the protocols used to prepare genomic DNA for sequencing and the selection for specific insertion sites during growth of the library prior to exposure to experimental conditions. Biological bias encompasses the transposon family chosen, the preferred nucleotide sequence for transposon insertion and the genomic composition of the organism under investigation.

The protocols used to identify the transposon insertion sites can introduce methodological biases. The TIS methods used throughout this work could be a source of bias leading to under or overrepresented sites of transposon insertion. In this chapter, I focus on accounting for biases that arise from the procedures used to prepare DNA for sequencing, the sequencing itself and then mapping of the data to the reference genome.

4.1.1 Methodological Bias Introduced During Sequencing Preparations

Generally, TIS experiments are designed such that minimal bias is introduced in the sequencing preparation reactions or the sequencing itself. Two main areas that introduce bias are PCR amplification and the sequencing (PCR like) reactions. PCR free sequencing protocols have been used but provide other sources of variation, (Chao et al., 2016). It has been shown that there is a loss of specificity when determining genes as essential with increasing PCR cycles leading to an increased number of genes determined essential (Alkam et al., 2021) . Gaio et al. observed common sequence motifs amongst areas of low sequence coverage in three different organisms while minimising Nextera Flex protocols. They suggest that lllumina's sequencing chemistry does not amplify and sequence all sequence motifs equally (Gaio et al., 2022). The work in this chapter describes a whole genome sequencing (WGS) approach to highlight not only PCR amplification biases but any methodological bias introduced during the sequencing and data mapping protocols.

4.1.2 Variability Introduced by Mutant Growth

During generation of a transposon mutant library, the mutants are selected for with an appropriate antibiotic. During this growth period, a mutant will typically undergo 30 generations to produce a colony (Mashimo et al., 2004). Under permissive growth conditions, mutants lacking the genes or genomic features required for growth will be lost from the library. Additionally, mutants that exhibit a slower growth phenotype will be outcompeted and underrepresented in the final library (Mahmutovic et al., 2020).

4.1.3 Transposon Biases

This work focusses on two transposons with a DDE type cut and paste mechanism(Bourque et al., 2018). However, there are differences in the mechanism of action and target site requirements for each. Each transposon will not necessarily produce the same insertion pattern in the same genomic background; to test this we have used a model *E. coli* K12 with no GC bias in genome composition.

4.1.4 Organism Dependent Bias

The organism under investigation is a variable factor in TIS studies. Chromosome structure and configuration may leave areas impenetrable to transposon insertion therefore providing physical protection (Chao et al., 2016; Green et al., 2012; Morris et al., 2016). It may be that there is in fact a biological reason to provide physical protection to some areas of the genome and essential genes may appear more frequently in these regions due to the inaccessibility of transposon insertion or recombination with foreign DNA. Another prevention to transposon insertion is the blocking of DNA by binding proteins and transcriptional regulators, such as the global gene silencer H-NS (Verma et al., 2019).

4.1.5 Aims of this Chapter

Transposon insertion variation should be considered in terms of the methods used to prepare libraries, the transposon system chosen and the target organism. The aim of this chapter is to investigate the factors affecting the reproducibility of transposon insertion and how that can be used to refine the determination of protected regions of the genome that are designated as essential in analysis pipelines.

4.2 Methods for this Chapter

There is an overlap in the methods used throughout this work; generalised protocols are fully described in *chapter 2*. The specific conditions used to generate data presented in this chapter are described below.

4.2.1 Whole Genome Sequencing of *E. coli* BW25113

The genome of *E. coli* BW25113 was sequenced on the same Illumina NextSeq run as the TIS libraries described in this work. This increased the read diversity for the run and would highlight any biases in sequence library preparation.

4.2.1.1 Sequence Library Preparation

The whole genome sequencing (WGS) of *E. coli* BW25113 was performed using a combination of the minimised Tagmentation reaction from the TIS sequencing protocol detailed in *section 2.8.4* and the Illumina DNA prep standard protocol for WGS (Illumina, 2022). Briefly, 50 ng of genomic DNA was tagmented and amplified with the standard Illlumia i5 primer and the custom biotinylated i7 primer, sequences provided in *section 2.8.3*. This amplification differed from the TIS protocol; the annealing temperature was lowered to 62 ˚C for WGS. Then biotin affinity capture was performed using the Dynabeads™ kilobaseBINDER™ Kit (Invitrogen, Massachusetts, US) following manufacturer's instructions. This was followed by a further PCR amplification step; again, the annealing temperature was reduced to 62 ˚C for the WGS sample. The PCR products were cleaned and normalised to 4 nM with nuclease free water, then pooled with the TIS samples as described in *section 2.8.8.* The pool was denatured and further diluted according to the Illumina protocol, then sequenced on a 75-cycle, single end run.

4.2.1.2 Mapping the *E. coli* BW25113 Reads to the Reference Genome

The genomic sequence reads were mapped to the hybrid reference, *section 3.3.1, appendix 9.1*. The BioTradis pipeline, Galaxy Version 1.4.5 (Barquist et al., 2016), was employed in tag-less mode and the reads were mapped with smalt (Ponsting & Ning, 2012) using the custom parameters described in *Table 4-1* and the minimum mapping quality set to zero (to keep reads with multi-mappings, such as insertion elements that occur frequently in the *E. coli* genome). These parameters were the same for mapping all the TIS sequencing data.

Table 4-1 *The smalt parameters used to map the trimmed and filtered fastq files to the hybrid reference sequence.*

Command	Parameter	Default	Custom
smalt_k	Word Length	20	13
smalt_s	Step Size	5	1
smalt_y	Minimum	0.96	0.8
smalt_r	Seed	1	0

4.2.2 Generation of Limited Growth Libraries

Limited growth libraries were made to identify the distribution of transposon insertions regardless of their impact upon growth. Donor and recipient matings were incubated for five hours to allow sufficient time for conjugation but to restrict growth to an estimated maximum of two doublings.

4.2.2.1 Conjugation to Generate Mutants

The limited growth libraries were generated in the same way as the standard libraries described in *section 2.5.1*. Spots of 10 µL mating mix containing equal volumes of the donor strain *E. coli* MFD*pir+:* pSAM_Ec or MFD*pir+:* pBAMD1-2, mariner and Tn*5* respectively, and the recipient strain *E. coli* BW25113 were incubated on LB agar for five hours. For the mariner limited growth library there were five spots and for the Tn*5* library there were ten. The spots for each library were then harvested, pooled per transposon, and stored at -80˚C in 20% glycerol.

4.2.2.2 Sequencing Limited Growth Libraries

The limited growth libraries were sequenced and mapped to the reference genome using BioTradis and the parameters listed in *Table 4-1.* Sequencing was performed as previously described in *sections 2.8 and 3.2.2.3.*

4.2.3 Dinucleotide Distribution Throughout the Genome

The dinucleotide occurrence for each of the 16 combinations was determined using a custom python script to identify the pattern, dinucleotide, in the hybrid reference sequence and return the number of occurrences of each. The script that I wrote for this is included in *appendix 9.4.*

4.2.4 Insertion Point Visualisation

Two tools were used to visualise the transposon insertion sites across the genome, described below.

4.2.4.1 Artemis Genome Browser

Artemis Genome Browser version 17.0.1 (Carver et al., 2012) was used to visualise the transposon insertion sites across the genome. The annotated genome was provided and then transposon mutant libraries were loaded in as the insertion plot `.txt` files. Artemis was primarily used for detailed inspection of insertion patterns in specific areas of the genome and for annotating the genome to provide context to the insertion patterns.

4.2.4.2 TraDIS Viewer

TradisViewer is a tool scripted by Dr. George Savva and implemented in R version 4.2.2 (R Core Team, 2022). It used the insertion plot `.txt` files. This tool offered customisable visualisation of transposon insertion data with user defined plotting parameters using common R plotting language. The main benefit to using TradisViewer was that it allowed for normalisation of input datasets. Additionally, TradisViewer allowed more insertion profiles to be viewed at the same time than Artemis; all nine replicates could be visualised simultaneously.

4.3 Results

4.3.1 Methodological Biases identified by WGS

The WGS reads were processed through the BioTradis mapping pipeline (Barquist et al., 2016; Langridge et al., 2009), the output is summarised in *Table 4-2.* Unlike the transposon libraries, there were no reads mapped to either of the plasmids. All the reads by default started with the transposon tag. Approximately 89 per cent of the reads mapped to the genome; this is on a par with the Tn*5* libraries where up to 88 per cent of the reads mapped to either the chromosome or the plasmid. These were both higher than for mariner where up to 67 per cent of reads mapped to either the chromosome or plasmid.

Table 4-2 *Summarised BioTradis output from mapping WGS.*

Library	Total Reads	Transposon Reads		Mapped Reads		Unique Insertion Sites	Insertion Distance (bp.)
		Count	%	Count	%		
BW25113 WGS	1.92E+07	1.92E+07	100	1.71E+07	89.14	3.12E+06	1.40

The reads were mapped to the E. coli BW25113 hybrid reference genome.

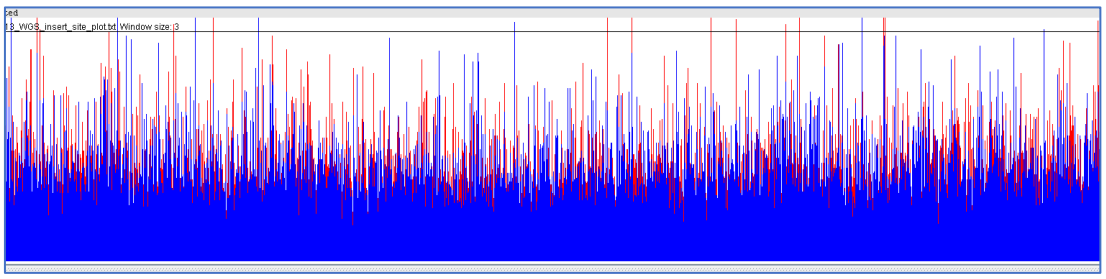

Figure 4-1 *Artemis whole genome sequence plot.*

Artemis visualisation of the whole genome sequence mapping across the E. coli BW25113 hybrid genome assembly from 1,050,000 bp to 4,375,00 bp. Reads mapping to the forward strand are coloured blue and reads mapping to the reverse strand are coloured red. A read was mapped, on average, approximately every 1.5 bases.

When the whole genome sequence library was visualised with Artemis *Figure 4-1,* it showed that amplification and sequencing occurred evenly across the chromosome, consistently at around 30 reads per site, with no areas lacking coverage. There were some sites that had an increased number of insertions, indicated by spikes in *Figure 4-1*, but these appeared random.

Visualisation in Artemis was limited by the window of observation and scaling is often required. The spikes, or 'jackpot amplification events'(Chao et al., 2016) demonstrated chance extreme overrepresentations that were not reflective of insertion frequency (*Figure 4-2*). These overrepresented jackpot events became less apparent when viewed at $\log_{10}$ scale, *Figure 4-2B*.

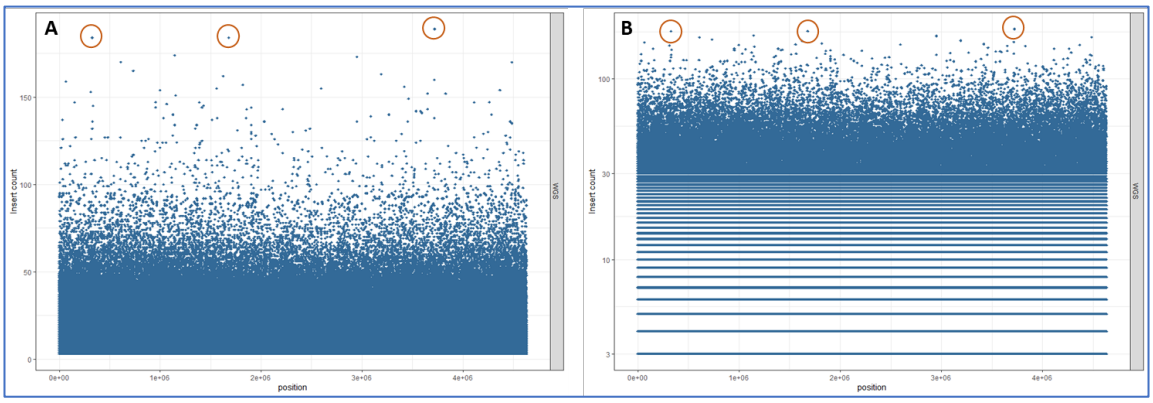

Figure 4-2 *TradisViewer whole genome sequence plot.*

Visualisation of the whole genome sequence mapping across the entire E. coli BW25113 hybrid genome assembly, A: raw counts, B: on a $\log_{10}$ scale. The orange circles highlight three examples of random amplification jackpots.

When the WGS mapping was observed alongside the mariner and Tn*5* libraries, there was no evidence that any areas or specific base pairs with an increase in read mapping had transmitted through to the transposon libraries; this can be seen in *Figure 4-3*. There was variation in the peak height indicating the number of reads mapped or insertion frequency, observed in the libraries themselves. Each library was individually scaled to demonstrate the extreme values of the overrepresented insertion locations. Increased insertion counts at any site could be due to random amplification or a genuine increase in the likelihood of transposon at that genome location; one example is indicated by the solid arrow in *Figure*

4-3. Conversely, an area where there appeared to be a lack of insertions, or reads mapped, across all three is indicated by the dashed arrow. In downstream essentiality analysis, this could be an example of a misidentified essential region when in fact it is simply underrepresented due to sequencing and mapping.

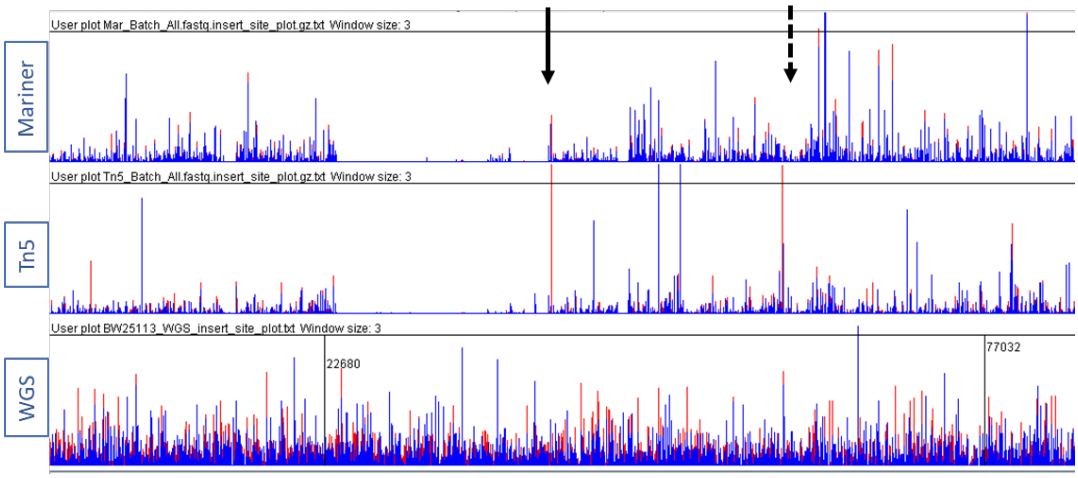

Figure 4-3 *Artemis visualisation of mariner, Tn5 and WGS insertion (mapping) sites from 0-84,500 bp.*

Mapping to the E. coli BW25113 hybrid assembled genome. Mariner scaled to a maximum of 1000 reads, Tn5 to a maximum of 200 reads and WGS to a maximum of 50 reads. The solid arrow marks a site where insertion is likely to be beneficial to growth. The dashed arrow marks an area where a lack of insertion density could be due to methodological biases unrelated to growth of the mutants. Blue lines represent reads mapped to the forward stand, red lined represent reads mapped to the reverse strand.

TradisViewer showed that the whole genome sequence mapping did not appear to correlate with any peaks or troughs in the transposon insertions mapped for either the Tn*5* libraries or the mariner libraries, this was visualised at the genome scale and on a Log_{10} scale, (*Figure 4-4*). This showed that there was no obvious bias in insertion site counts caused by the PCR during library preparation for *E. coli*. When the library insertion data was normalised by the WGS mapping, the insertion data did not appear changed and maintained the same profile across the genome with the same peaks and troughs (*Figure 4-4B)*. This showed that any variation provided by the methods used to generate the sequence reads and mapping insertions had little or no observable effect on the variation in insertion density of either the Tn*5* or the mariner transposons.

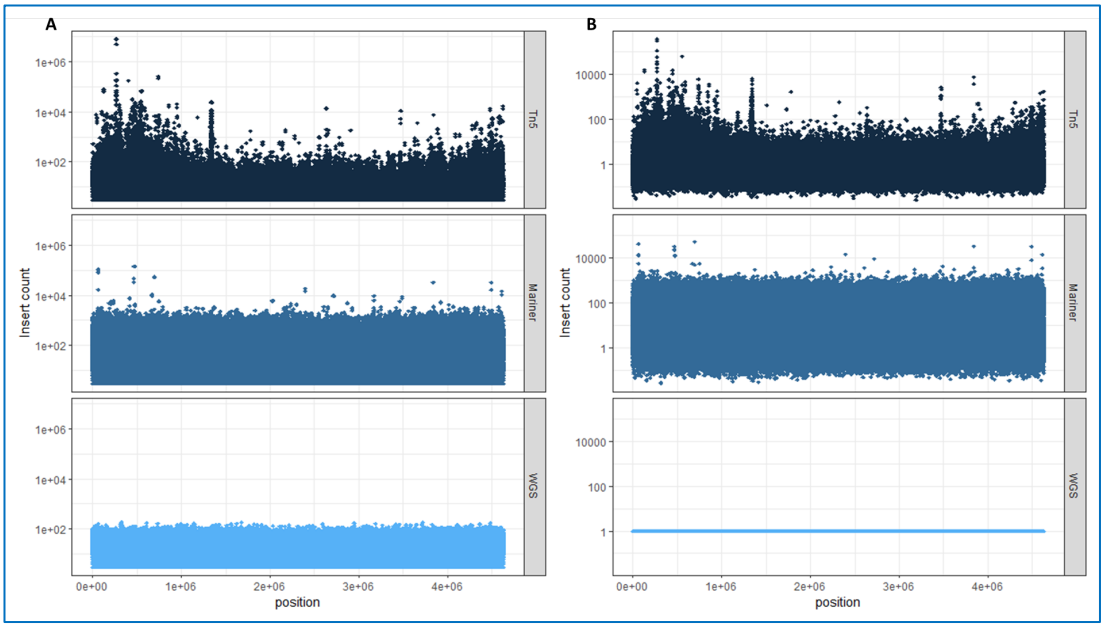

Figure 4-4 *Whole genome comparison of WGS and transposon insertion points.*

TradisViewer visualisation to compare the mariner and Tn5 insertion points to the whole genome sequence mapping of the E. coli BW25113 hybrid assembled genome, log_{10} scale. A: Raw counts, B: Tn5 and mariner insert counts normalised to the whole genome mapping.

An analysis of variance between the Tn*5* insertion libraries and the WGS mapping confirmed that the methodological variance had minimal impact on the variance observed in the insertion density of the Tn*5* transposon library across the chromosome, with R^2 being 1.91×10^{-8}. For the same analysis of variance between the mariner libraries and the WGS mapping, the R^2 value was 4.07×10^{-4}. This indicated that the amplification steps during sequence library preparation, the sequencing process, and the sequence data processing had a negligible contribution to the variance observed in transposon insertion density for both the Tn*5* and mariner transposon libraries. This was the case in *E. coli* K12 and for the sequencing run used to obtain the data here; alternative sequencing protocols or TIS in less straightforward organisms may have more of an effect on TIS library variability and so it would be prudent to check for any bias in PCR before analysing TIS data.

4.3.2 Variability Provided by Growth of Mutants

Growth of mutants to stationary phase on transposon selective media was another factor to consider as a source of bias (Mahmutovic et al., 2020). To investigate the variation in transposon insertion density provided by growth, a limited growth library was constructed for both the Tn*5* and the mariner transposons. The DNA was extracted immediately following conjugation meaning that the mutants could only undergo growth during conjugation, and this was estimated to be an average of 1.5 doublings.

A summary of the BioTradis mapping output from the two limited growth libraries is given in *Table 4-3*. For both transposons used, far fewer reads mapped to the genome than any of the previous libraries constructed. These libraries did not contain as many unique insertions as the standard transposon libraries produced; for Tn*5* the number of unique insertions for the limited growth library was between one fifth and one third as dense as the fully grown libraries. For mariner, this fell lower to between one twentieth and one tenth.

Table 4-3 *A summary of the BioTradis mapping pipeline output for the mariner and Tn5 limited growth libraries.*

Library	Total Reads	Transposon Reads		Mapped Reads		Plasmid Mapped Reads		Unique Insertion Sites	Insertion Distance (bp.)	Average reads / Insertion
		Count	%	Count	%	Count	%			
Mariner Limited Growth	8.10E+06	6.99E+06	86.29	9.03E+05	12.92	2.14E+06	26.37	9022	513	100.10
Tn*5* Limited Growth	4.37E+06	3.81E+06	86.98	8.35E+04	2.19	3.00E+06	68.60	21608	214	3.86

Although these limited growth libraries were not at an insertion density across the genome sufficient to determine “essentiality” they were able to highlight specific areas of the genome that were more or less susceptible to transposon insertion *(Figure 4-5).* The dispersion of identified insertion sites across the genome appeared to differ between the two transposons used and suggests a differing propensity for insertion of each of the transposons at any given genomic site. However, there were some sites, denoted by solid arrows in *Figure 4-5* that had an increased insertion rate for both transposons.

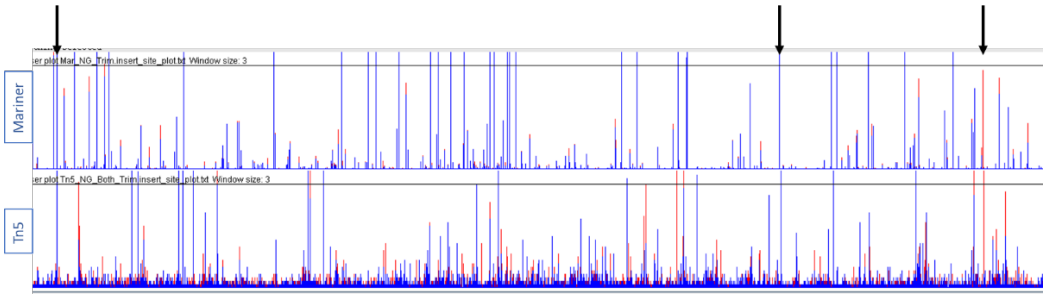

Figure 4-5 *Artemis visualisation of common elements in the mariner and Tn5 insertions for the limited growth libraries.*

Common “jackpot” insertion events between the mariner insertion data and the Tn5 insertion data from 875,000 bp to 4,025,000 bp. The solid arrows highlight examples of sites that have an increased insertion density for both transposons used.

4.3.2.1 Mariner Limited Growth Library

The mariner limited growth library contained 9022 unique insertions, a density of around 1 insertion every 500 bp. This was too sparse to indicate specific regions of low transposon insertion density and therefore regions of the genome that could be protected from transposon insertion. However, it was possible to see areas lacking insertions in both the limited growth library and the growth libraries when they were aligned, these were indicated by the dashed arrows in *Figure 4-6*.

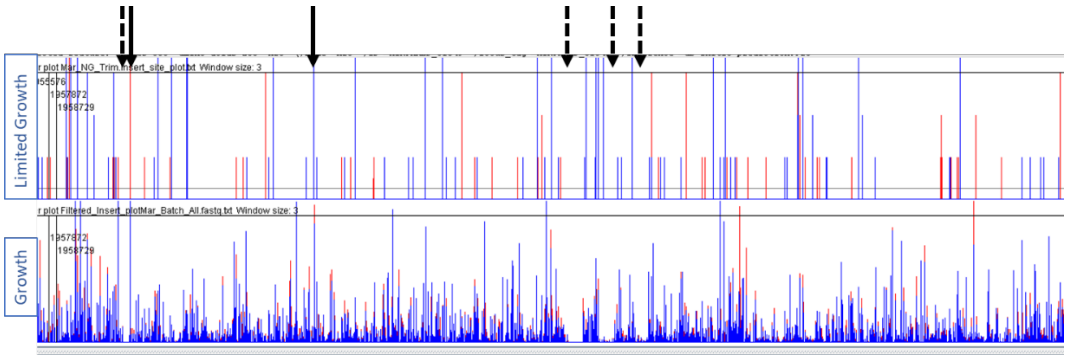

Figure 4-6 *Artemis visualisation of the limited growth and mariner library insertions.*

From 1,956,500 bp to 2,067,000 bp; (top) limited growth library; (bottom) the mariner library. The solid arrows highlight sites that may be favourable to mariner transposon insertion. The dashed arrows indicate a lack of insertions common to both data.

The reduced insertion density could suggest that while transposon insertion is possible, it is less favourable at these locations. Had the limited growth library been at a higher level of saturation (work in progress) this would have been demonstrated more clearly. Alternatively, the lower insertion density could indicate that these regions were beneficial to the growth of *E. coli* K12 and that disruptions led to slower growing mutants, hence a lower representation in the pool of mutants. The solid arrows in *Figure 4-6* highlight sites where there was an increase in transposon insertion in both the limited growth and growth libraries, this suggested a preference for insertion at these locations.

When TradisViewer was used to look in more detail at areas of the genome, *Figure 4-7*, there were evidently patterns in the limited growth library that had impacted the nine growth libraries. The right side showed the raw insertion count data, the left was normalised to the limited growth library.

Figure 4-7 (drawn from 200,000 bp to 250,000 bp) showed that, at around 215,000 bp to 218,000 bp (inside the box), there was a region of high insertion density for the limited growth library that had not translated through to the growth libraires; insertion density for the grown libraries was consistent with the rest of the chromosome. The next genome section, 218,000 bp to 220,000 bp, showed a higher insertion density in the limited growth library but not as extreme as the previous segment. The increased in insertion within this region was absent from any of the grown libraries, this suggested that this region of the genome was required for growth on LB with kanamycin at 37˚C.

Figure 4-8 demonstrated four examples of areas that lacked transposon insertion in all of the growth and the limited growth libraries. Additionally, there was a peak at around 3,845,000 bp (indicated by an arrow) that was present in all of the growth and the limited growth libraries. When normalised, the absence of insertions remained the same, but the peaks were removed. The region from 3,770,000 bp to 3,780,000 bp (inside the box) lacked insertions in the limited growth library across parts of the highlighted region but the growth libraries had areas of normal insertion density, this demonstrated that the limited growth library required more unique insertions to give a clearer representation of growth independent transposon insertion bias.

Figure 4-9 is a zoomed visualisation of *Figure 4-8,* from 3,840,000 bp to 3,845,000. There was a peak in the limited growth library (within the box) that transmitted through to all of the growth libraries to some extent. This peak in the context of an essentiality study, would be interpreted as an area where mutation would be beneficial to the organism. When the growth libraries were normalised to the limited growth libraries this peak became less prominent, therefore interrupting this gene was unlikely to be beneficial to growth under the test conditions. Generally, these figures show that areas of low insertion density were likely to transfer to grown libraries. Single overrepresented insertions in the limited growth library did not appear to transmit through growth but regions of increased insertion did.

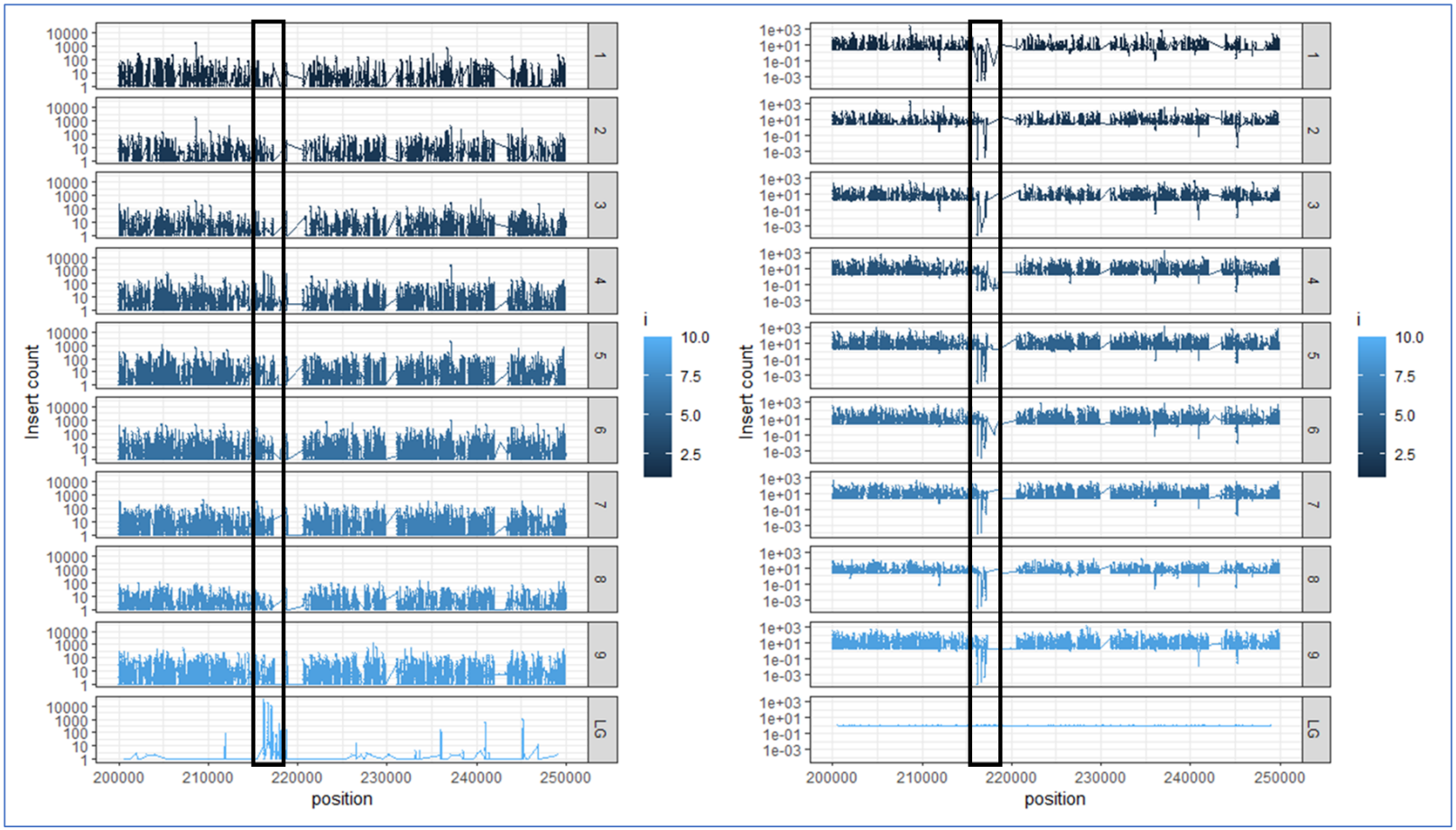

Figure 4-7 *TradisViewer visualisation of the nine mariner libraries and the mariner limited growth library from 200,000 bp to 250,000bp.*

The black box surrounds an area of high transposon insertion that had not translated to growth libraries. Left: Raw insert counts. Right: Normalised to Limited growth.

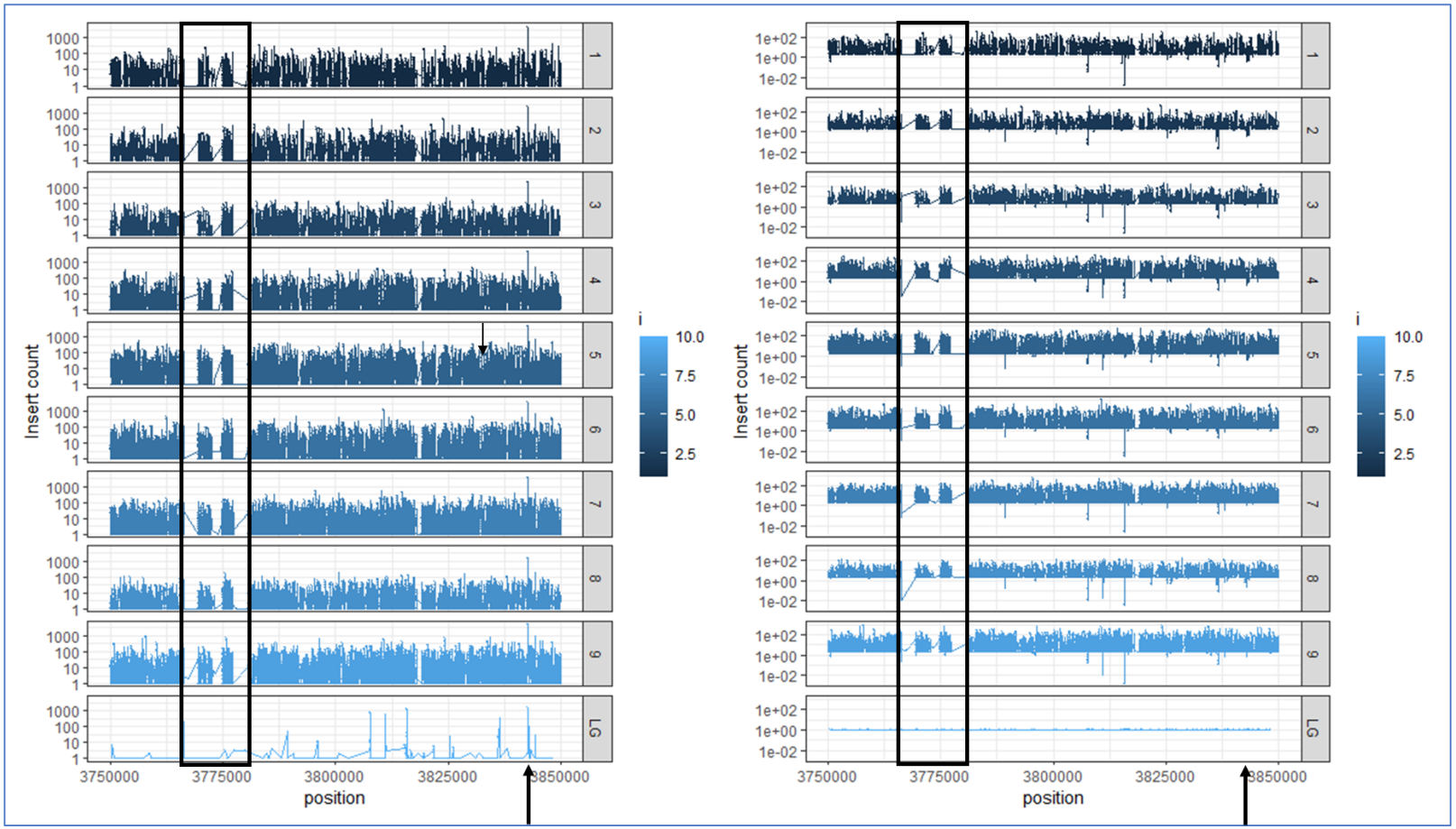

Figure 4-8 *TradisViewer visualisation of the nine mariner libraries and the mariner limited growth library from 3,750,000 bp to 3,850,000 bp.*

The black box shows a lack of insertions, independent of growth. The arrow highlights an insertion hotspot. Left: Raw insert counts. Right: Normalised to Limited growth.

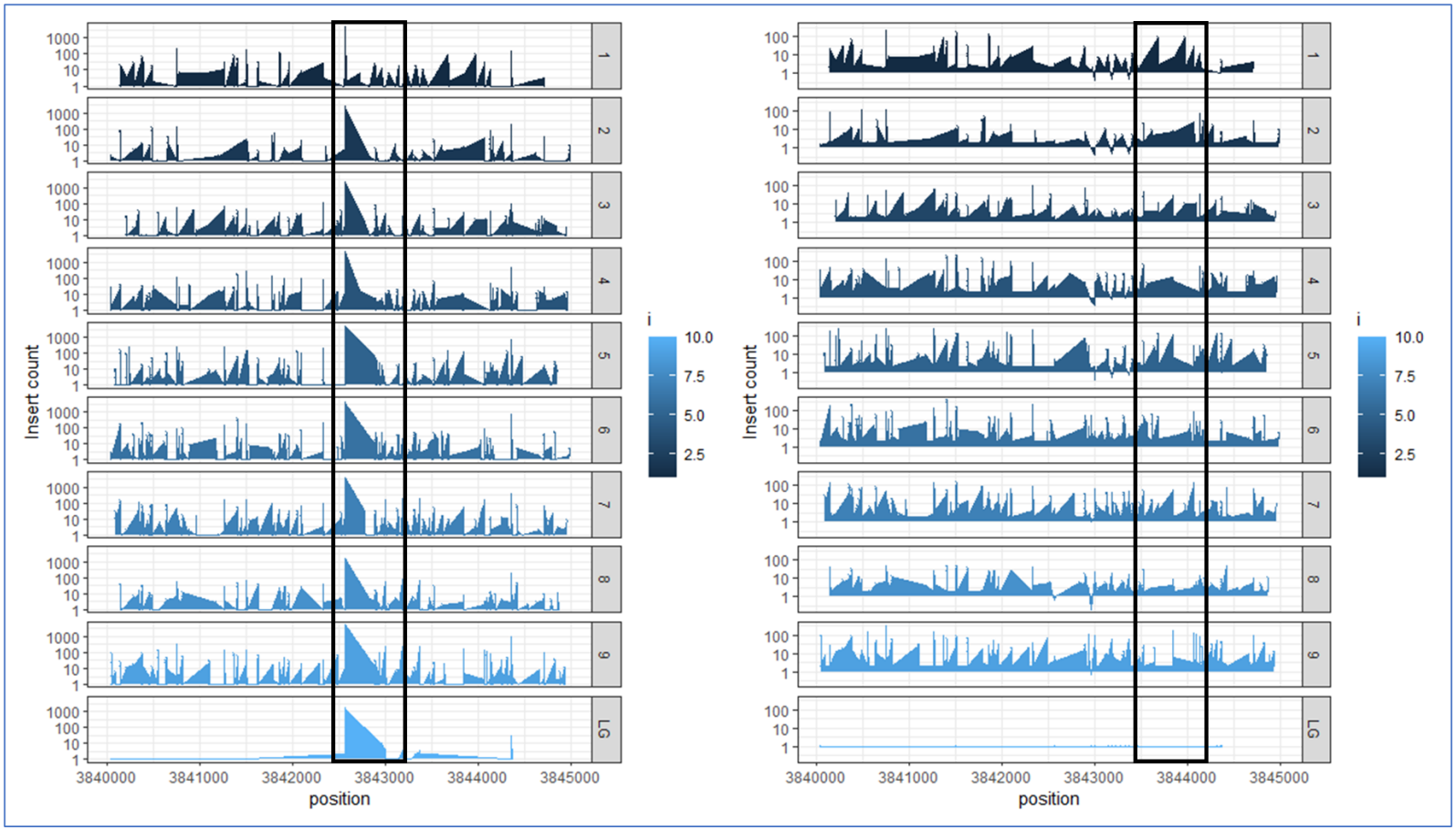

Figure 4-9 *TradisViewer visualisation of the nine mariner libraries and the mariner limited growth library from 3,840,000 bp to 3,845,000 bp.*

The black box shows where the insertion profile changed when normalised to the limiited growth library. Left: Raw insert counts. Right: Normalised to Limited growth.

4.3.2.2 Tn*5* Limited Growth Library

Similarly to the mariner libraries, patterns between the insertion profiles of the fully grown and limited growth Tn*5* libraries emerged, *Figure 4-10.* The solid arrows highlight insertion sites that were overrepresented in both; previous results from this chapter show that amplification and sequencing protocols did not account for this. Therefore, there must have been some sites that the Tn*5* transposon is more likely to insert into and these had transmitted through to the growth libraries.

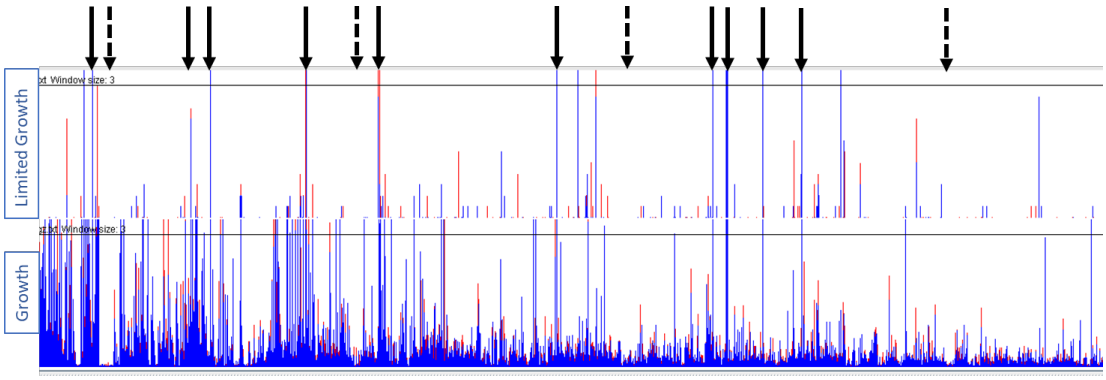

Figure 4-10 *Artemis visualisation of the limited growth and Tn5 library insertions.*

From 1,956,500 bp to 2,067,000 bp; (top) limited growth library; (bottom) the Tn5 library. The dashed arrows indicate a lack of insertions common to both data. The solid arrows highlight sites that may be favourable to mariner transposon insertion.

Importantly, the dashed arrows in *Figure 4-10* show regions that lacked transposon insertions or where Tn*5* transposon insertion was reduced compared to the surrounding areas. In the grown library these areas would be considered essential for growth, yet they were present in a library without growth. This showed that there were regions of the chromosome where Tn*5* transposition was less likely to occur and lacked insertion regardless of growth or, in other applications, stress applied. These areas were protected from transposon insertion and not determined functionally essential in this context.

When using TradisViewer to visualise the insertion counts on a $\log_{10}$ scale, *Figure 4-11*, clear patterns between the growth and limited growth libraries emerged. The insertion profile of the limited growth library matched the growth libraries before normalisation, seen in the left side of *Figure 4-11.* This would suggest that for Tn5, the variance in transposon insertion was determined by more than the fitness cost of the mutation.

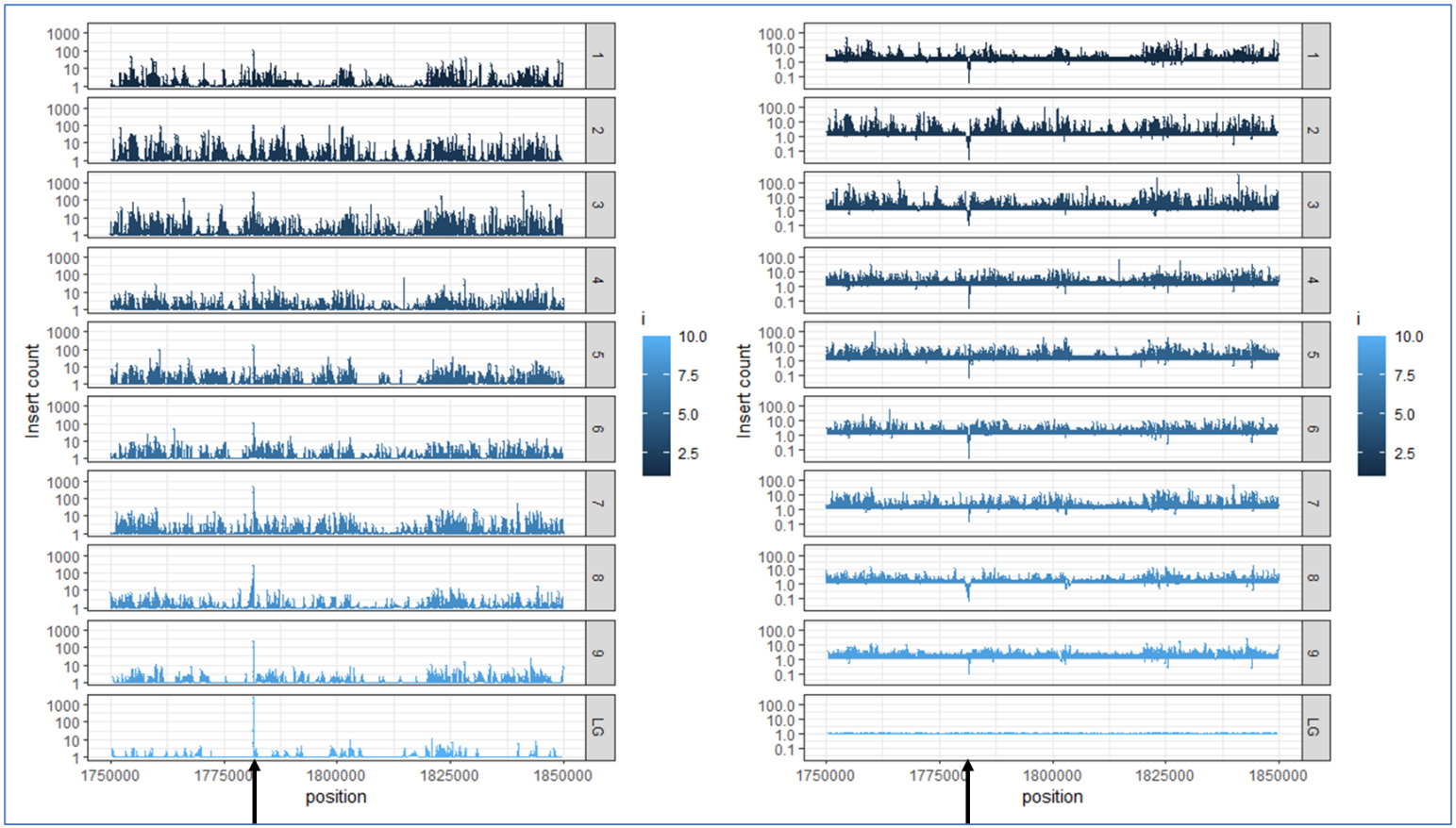

Figure 4-11 *TradisViewer visualisation of the nine Tn5 libraries and the Tn5 limited growth library from 1,750,000 bp to 1,850,000 bp.*

The arrow highlights an insertion hotspot. Left: Raw insert counts. Right: Normalised to Limited growth.

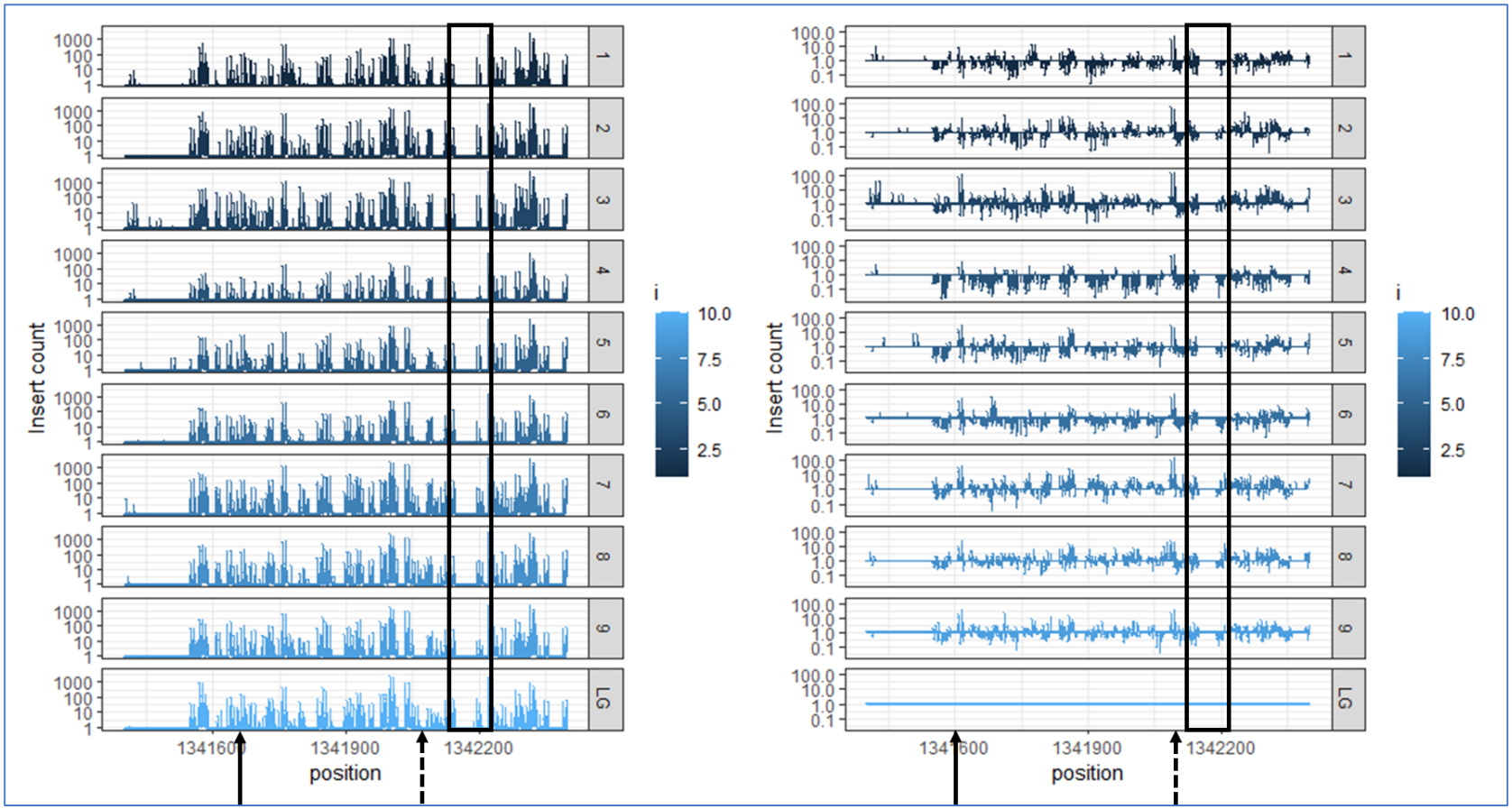

Figure 4-12 *TradisViewer visualisation of the nine Tn5 libraries and the Tn5 limited growth library from 1,341,600 bp to 1,342,200,000 bp.*

The arrows show areas where a transpson insertion was benficial to growth. The box highlights an area that was protected from transposon insertion. Left: Raw insert counts. Right: Normalised to Limited growth.

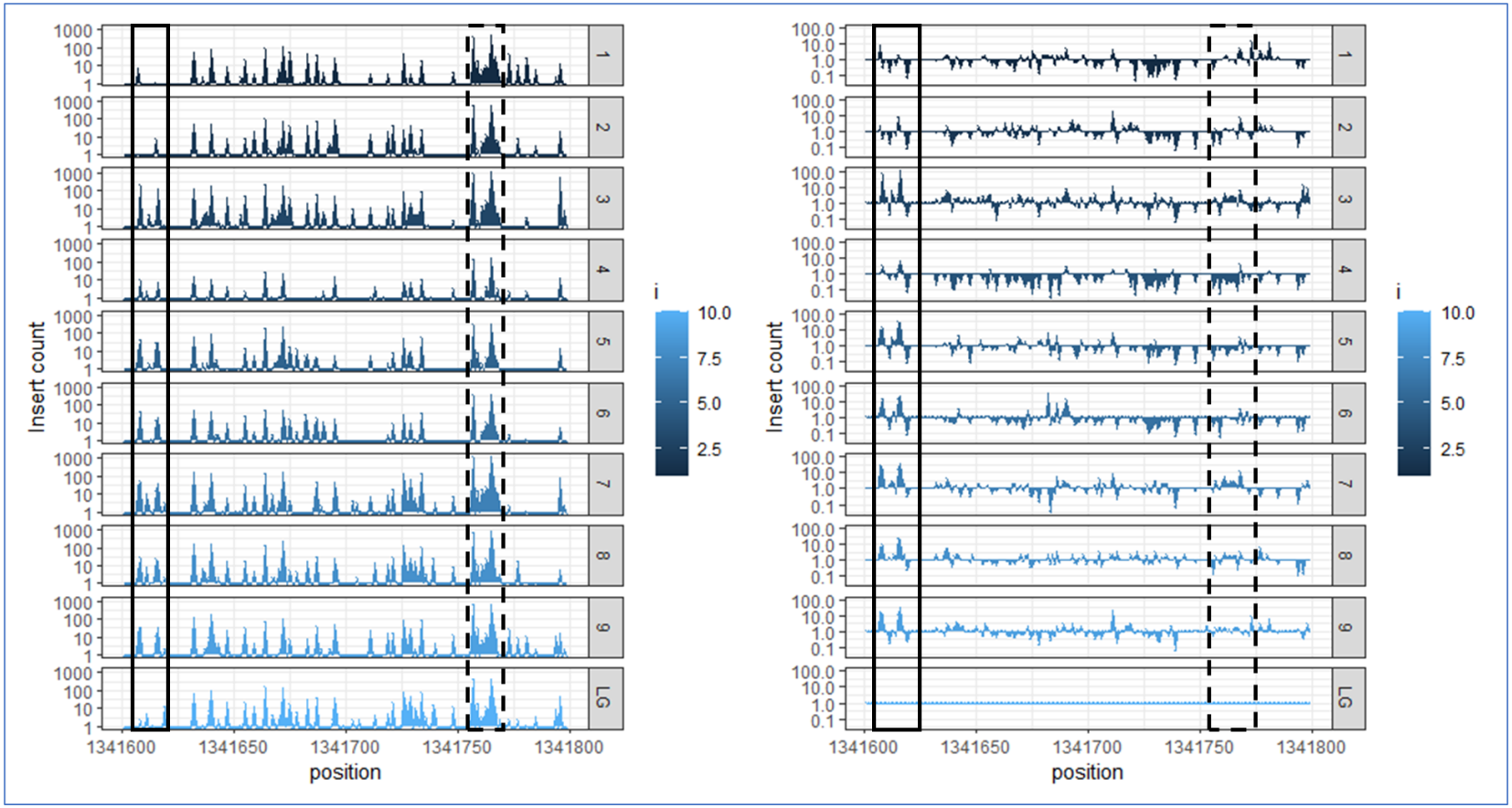

Figure 4-13 *TradisViewer visualisation of the nine Tn5 libraries and the Tn5 limited growth library from 1,341,600 bp to 1,341,800,000 bp.*

The boxes highlight areas that changed insertion profile when normalised to the limited growth library. The solid vox highlights an areas of increased transposon insertion leading to fitness advantages. The dashed box represents an area where transpoosn insertion appeared to offer a fitness advantage but this not the case upon normalisation to a limited growth library. Left: Raw insert counts. Right: Normalised to Limited growth

In *Figure 4-11A* there was a site, around 1,780,000 bp (denoted with an arrow), with an increased number of insertions in all libraries that was not present once normalised to the limited growth library. In general, once the insertion counts had been normalised to the limited growth library, there were fewer defined peaks, and the insertion density appeared more consistent across the region.

Figure 4-12 shows areas that were protected in all nine of the growth libraries and the limited growth library, for example from 1,342,150 bp to 1,342,200 bp (within the boxed regions). Across this 1 kb window of the genome, there were similar insertion profiles between the limited growth library and the growth libraries despite differences in insertion density. When normalised to the limited growth library, the profile of the grown libraries changed. There were areas that had an increased number of insertions following growth, so interruption was advantageous to growth, these are at around 1,341,600 bp and 1,342,100 bp and denoted with a solid and dashed arrow, respectively.

Figure 4-13 is further zoomed to show 200 bp of the genome. There were some differences in insertion pattern between the replicates, but the general patterns were the same. When normalised, the 2-3 peaks from around 1,341,610 bp to 1,341,625 bp (within the solid box) became two clear peaks, excluding libraries one and two. Additionally, when normalised, the distinctive shape seen at around 1,341,750bp to 1,341,770 bp (within the dashed box) was removed from the profile in the grown libraries, suggesting that growth was not impacted by insertions in this region and that the variability in insertion was due to another factor.

4.3.3 Comparing the Insertion Profiles of Both Transposons

The two transposons used in this work do not insert into the same nucleotide pattern across the genome. However, when the insert plot files were viewed in Artemis, across the genome there appeared to be very little difference in the distribution of transposon insertion sites as seen in *Figure 4-14*. Across the first million base pairs there were clear regions lacking insertion for both transposons and the insertion density across this region was consistent between the two; there were instances where the mariner libraries showed a decreased insertion frequency but the Tn*5* libraries did not, one prominent example was indicated by an arrow in *Figure 4-14.* This indicated that there was a difference in the insertion distribution between the Tn5 and mariner libraries therefore, the choice of

transposon used for TIS will impact the number of essential genes reported and could explain some of the discrepancies discussed in *chapter 3*, particularly *Figure 3-11*.

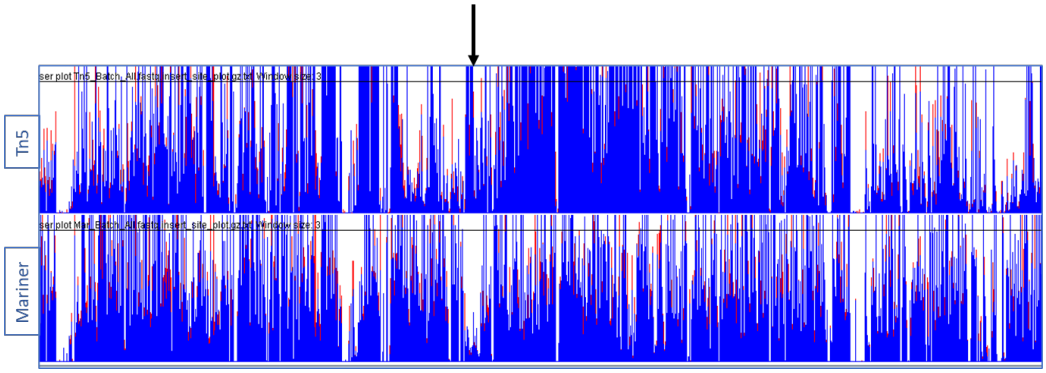

Figure 4-14 *Artemis visualisation of all the Tn5 and mariner insertions.*

From 0 to 934,000 bp. The arrow indicates a region in the genome where there is a difference in the number of insertions for the Tn5 and mariner libraries so demonstrates that Tn5 and mariner transposon insertion is not consistent across the genome. The Tn5 was scaled to a maximum of 75 reads mapping and mariner to a maximum of 300.

At the genomic scale, the insertion profiles of the two transposons appeared similar, however, there were instances where the profiles differed and in a way that would lead to differences in the classification of genes. *Figure 4-15A-D* shows genes where this was observed. Based on the mariner library insertions, these genes would have been considered essential. However, all except for *ubiJ* have, using Tn*5*, been determined non-essential by Goodall et al. (Goodall et al., 2018) and the Keio collection (Baba et al., 2006). The gene *ubiJ, Figure 4-15B,* was determined non-essential by Goodall et al. but essential by the Keio collection so is accessible for Tn*5* transposition but a knockout mutant was not cultivatable. Interestingly, in the Tn*5* insertion profile presented in this work *Figure 4-15B* there was a lack of insertions at the C-terminus relative to the rest of the gene. This suggested that this region of the gene was essential for growth and that the remainder of the gene was dispensable.

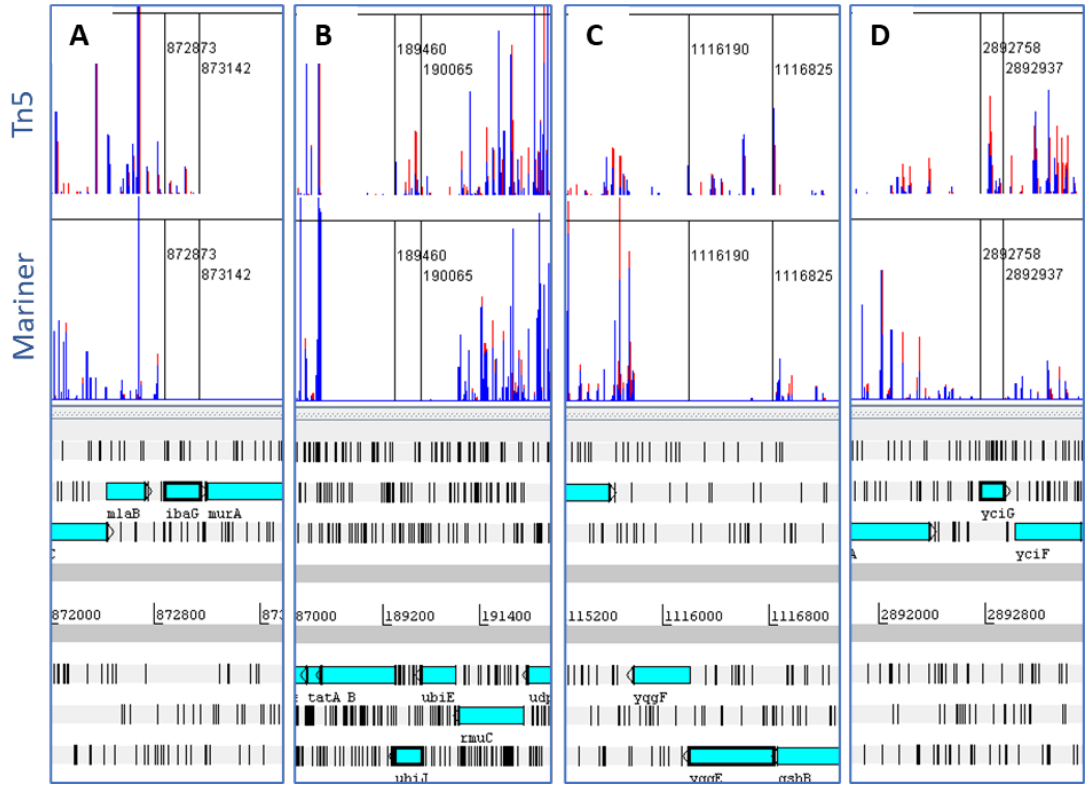

Figure 4-15 *Artemis visualisation of genes that have a different insertion site profile between the Tn5 and mariner transposons.*

These genes would be classified as non- essential using the (top) Tn5 insertion data but essential using the (bottom) mariner insertion data. A: ibaG, B: ubiJ, C: yggE, D: yciG.The black lines in the plot areas mark the boundaries of the selected gene.

Similarly to above, there are genes that would have been classified as essential by the Tn*5* libraries but not the mariner libraries, these are shown in *Figure 4-16A-D*. The gene *azuC*, *Figure 4-16B* was not given a classification in the Keio collection (Baba et al., 2006) or by Goodall et al. (Goodall et al., 2018), *punR* was determined non-essential by both so it is only in this work that it was deemed essential based on the Tn*5* libraires, *Figure 4-16D*. The gene *hda, Figure 4-16A*, was classified as essential using Tn*5* TIS (Goodall et al., 2018) but not the Keio collection (Baba et al., 2006); when considered with these data, *hda* was a candidate that was potentially inaccessible to Tn*5* transposition. *FtsK* was another example where part of a gene was essential, *Figure 4-16C* shows that the N-terminus was lacking insertions in the mariner library but there were sufficient insertions in the rest of the gene to be classified as non-essential. Goodall et al. have determined this gene non-essential, yet Baba et al. were unable to cultivate mutants lacking this gene.

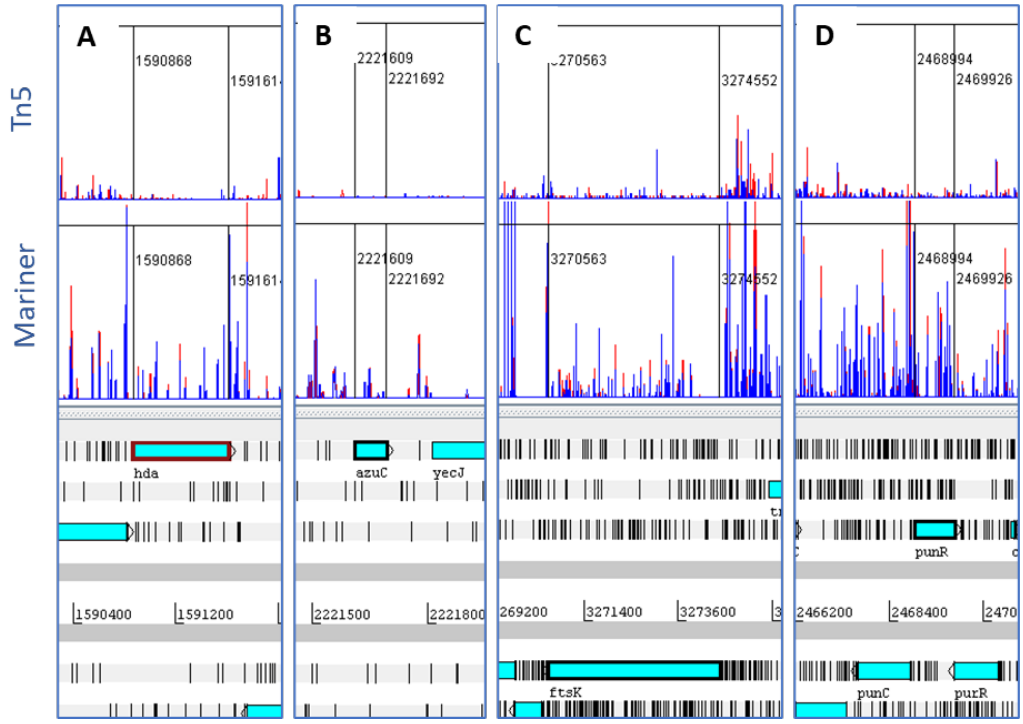

Figure 4-16 *Artemis visualisation of genes that have a different insertion site profile between the Tn5 and mariner transposons.*

These genes would be classified as essential using the (top) Tn5 insertion data but non-essential using the (bottom) mariner insertion data. A: hda, B: azuC, C: ftsK, D: punR. The black lines in the plot areas mark the boundaries of the selected gene.

When the two transposon libraries were observed in TradisViewer, the distribution profile of insertion sites appeared to differ between the two transposons, *Figure 4-17*. The mariner transposon inserted at a constant level across the entire genome, with a few jackpot locations having a much higher insertion count (from around 3,000 up to around 10,000) than the base rate of around 1,300. The Tn*5* insertions were not as consistent across the genome and demonstrated a phenomenon where insertion density was increased around the origin of replication due to gene dosage (Chao et al., 2016; Larivière et al., 2021; Zomer et al., 2012), where dividing cells have more copies of genes on replication forks near the origin and terminus available for transposon insertion. For this assembly the origin of replication started at 286,056 bp, marked by the red dashed line.

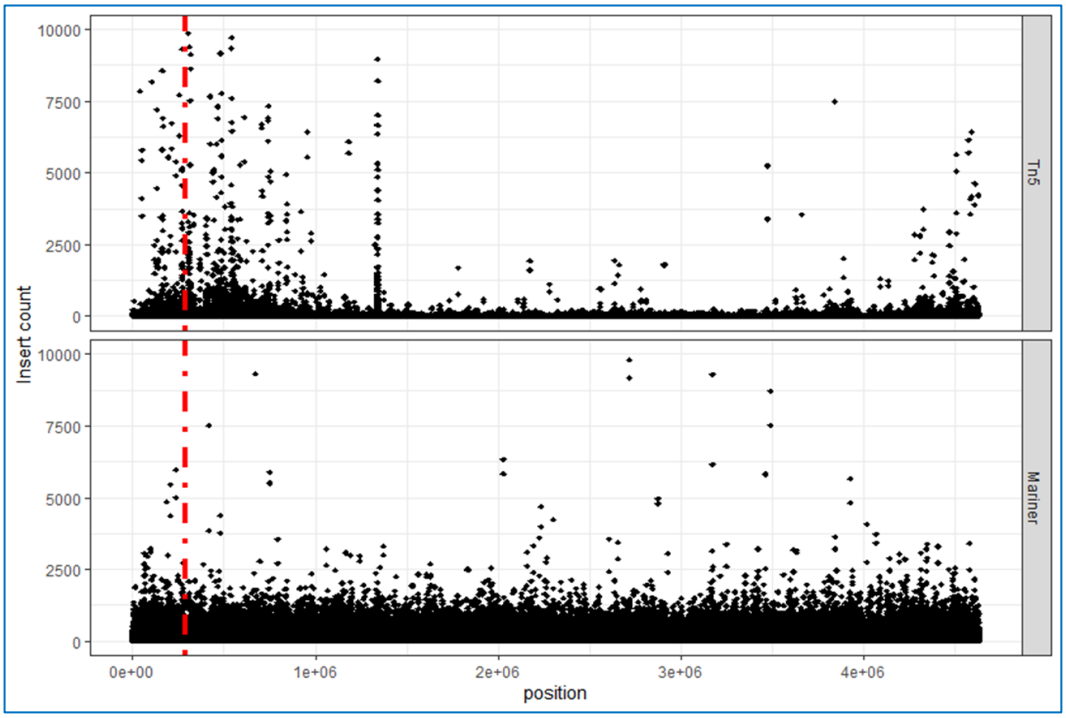

Figure 4-17 *TradisViewer visauliation of all of the Tn5 insertions and all of the mariner insertions.*

The insertion data for the growth libraries across the whole genome. The red dashed line indicates the origin of repliation.

There did not appear to be the same gene dosage effect with the mariner libraries. An explanation for this could be that the recipient cells of *E. coli* BW25113 were undergoing growth at the time of Tn*5* transposition, whereas for mariner, the cells were in a stationary state. The protocol for conjugation and transposition was the same for both transposon experiments; however, the mariner transposon was estimated to have a greater transposition efficiency than Tn*5*. Differences in transposition rates between the two transposons could mean that the recipient cells were in different growth states when the transposition event occurred.

The other feature present in the Tn*5* insertion data but not in the mariner insertion data was a dramatic increase in insertions at around 1,350,000 bp. This peak could have been overlooked in Artemis unless the maximum peak size was increased; at this resolution, the majority of the peaks in the library were not visible. This can be seen in *Figure 4-18* where the area with these high numbers of insertion have been located to the genes *nlpD* and *rpoS.* The insertion density in this region was increased for the Tn*5* data but not the mariner or the WGS, this suggests an increased insertion propensity for the Tn*5* transposon and was evidence of insertion site or chromosome access bias.

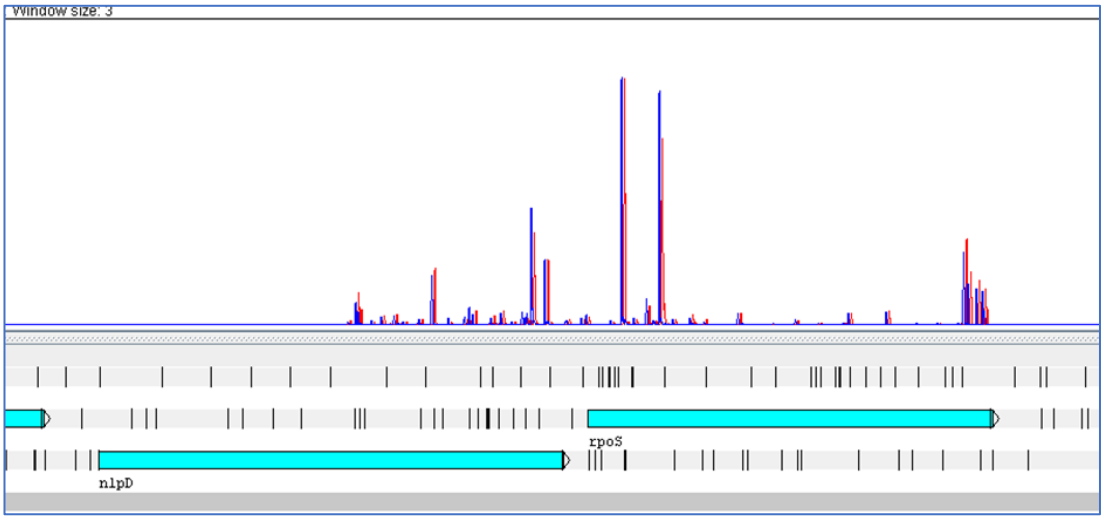

Figure 4-18 *Zoomed in view of the area of high Tn5 transposon insertion.*

Visualised in Artemis from 1,340,700 bp to 1,343,400 bp. Maximum peak size set to 10,000.

4.3.3.1 Mariner Insertion Site Bias – Dinucleotides

Mariner is widely reported to insert exclusively at TA dinucleotides (Bourque et al., 2018; Chao et al., 2016; DeJesus et al., 2015; Larivière et al., 2021; Morris et al., 2016; Zhao et al., 2017) and to leave a TA duplication at the insert site. Yet when mapping the mariner libraries from this work to the reference sequence, the number of unique insertions exceeded the number of TA sites within the genome. This indicated that the mariner transposon must have inserted into sites that were not TA dinucleotides. The number of insertions at each dinucleotide across the entire genome is listed in *Table 4-4* and shown in *Figure 4-19.* TA was the most common dinucleotide for insertion, with 38 per cent of insertions occurring at a TA dinucleotide, but mariner insertion was not exclusive to TA sites.

Table 4-4 *The number of sites with insertions or no insertions at each dinucleotide in the mariner libraries and the proportion of the total insertions at each.*

	Insertion		Non-Insertion			
	Count	% Of dinucleotides in genome	Count	% Of dinucleotides in genome	Count in genome	% Of Total Insertions
AA	46943	13.85	292069	86.15	339012	12.89
AC	38984	15.28	216132	84.72	255116	10.70
AG	21930	9.30	213818	90.70	235748	6.02
AT	49012	15.84	260335	84.16	309347	13.46
CA	7513	2.34	314194	97.66	321707	2.06
CC	5054	1.87	264601	98.13	269655	1.39
CG	5792	1.67	340169	98.33	345961	1.59
CT	6429	2.71	230998	97.29	237427	1.77
GA	4375	1.64	262417	98.36	266792	1.20
GC	5308	1.39	377770	98.61	383078	1.46
GG	3331	1.23	267811	98.77	271142	0.91
GT	8037	3.14	248269	96.86	256306	2.21
TA	136866	64.65	74846	35.35	211712	37.58
TC	6734	2.52	260167	97.48	266901	1.85
TG	9164	2.82	315303	97.18	324467	2.52
TT	8772	2.60	328775	97.40	337547	2.41
All	364244	7.86	4267675	92.14	4631919	100.00

After TA sites, insertions occurred most frequently at, AA, AC and AG. All of these dinucleotides have A as the first nucleotide and so this would be expected if the nucleotide preceding insertion is a T thus giving a TA site. It is undetermined whether the insertion point is considered to be at the start of the TA, the middle, or the end. When combined as a hypothetical TAN site these dinucleotides accounted for 43 per cent of the insertion sites, more than TA alone (38 per cent). When combining the proportion of TA and TAN sites, still only 81 per cent of insertions were accounted for, suggesting that there were a significant number (19 per cent) of insertions at other dinucleotides. These data suggest that if only TA dinucleotides were considered as potential insertion sites, then around 20 per cent of possible insertion sites would be disregarded. These values are for *E. coli* K12

and are likely to differ based on the organism used but may be important when calculating genome saturation levels for mariner transposon insertion.

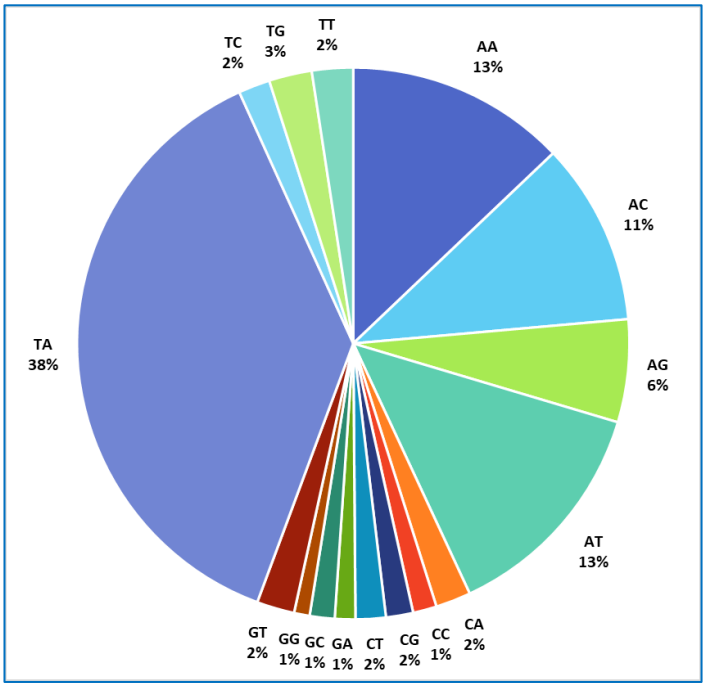

Figure 4-19 *The proportion of mariner transposon insertions occuring at each dinucleotide across the E. coli BW25113 genome.*

Figure 4-20 shows the proportion of sites with an insertion versus without an insertion for each of the 16 dinucleotide possibilities. For TA dinucleotides, it was more likely that the site contained an insertion than not, 65 per cent of the TA sites had an insertion. If the hypothesis that TAN can also be considered as a TA insertion then it would be expected that 16.3 per cent of each of the TAN sites would have insertions versus no insertions; this was the case for AA, AC and AT with all having close to the predicted average (14-16 per cent). However, AG was lower than the other dinucleotides that start with A and, as shown in *Table 4-4,* only accounted for six percent of the total number of insertions, whereas the other dinucleotides starting with A represented 10-12 per cent of the total insertions each; this suggested a role for guanidine in protection from transposon insertion by mariner.

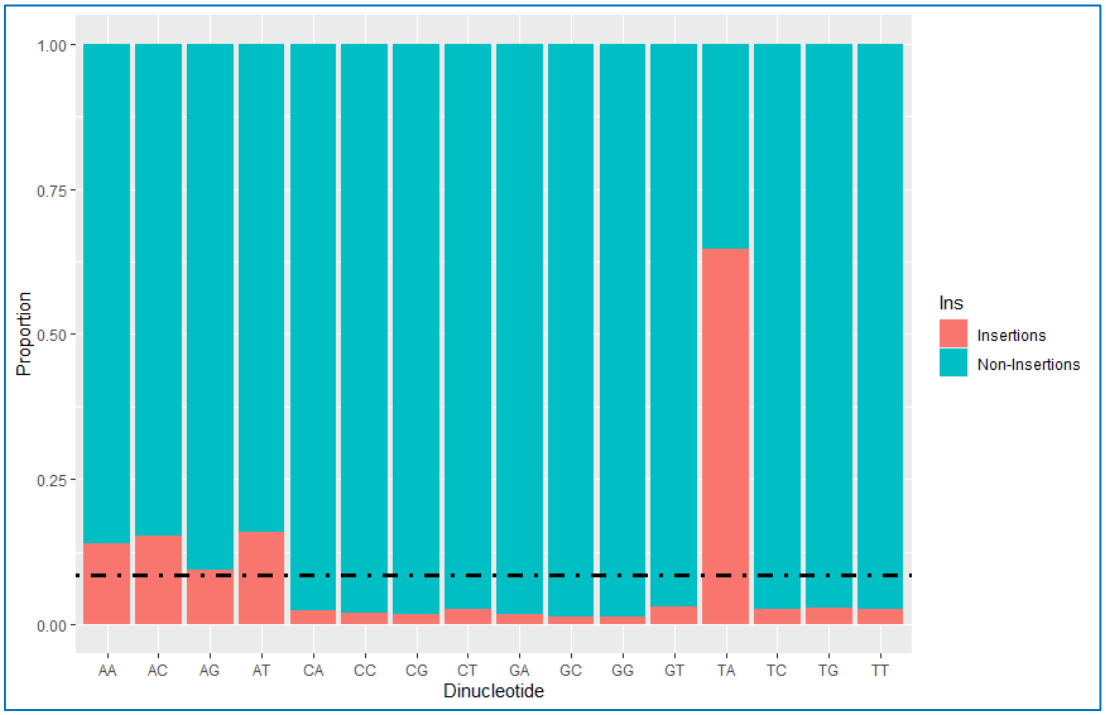

Figure 4-20 *The proportion of insertions or non-insertions at each of the sixteen dinucleotides.*

Data shown for all of the mariner growth libraries, the dashed line represents the expected insertion proportion (average) for libraries at this density assuming no insertion bias.

To investgate the impact of the sequence surrounding the insertion dinucleotide, the occurrence of each nucleotide at the positions +/- 12 base pairs from the non-insertion sites for each dinucleotide were plotted, *Figure 4-21*. There were few observable patterns other than a marginal increase in the occurrence of G, or C in the 2-3 base pairs preceeding or following the dinucleotide of interest. This was particularly apparent when the dinucelotide under observation is TA, the most permissive site for insertion. This would imply that G/C around the dinucleotide of insertion was preventative as suggested by *Figure 4-19* and *Table 4-4,* this pattern was palindromic around the TA dinucleotide. This led to a suggested non-permissive motif of CGN(TA)NCG around the insertion dinucleotide.

Figure 4-22 shows the occirence of each nucleotide at the positions +/- 12 base pairs from the insertion sites for each dinucleotide. The suggestion that G/C dinucleotides were protective is further demonstrated, where there was generally a decrease in the occurrence of G or C in the regions surrounding the dinucleotide of insertion. However, this could be a byproduct of insertions occuring more frequently in regions of genomic DNA with increased TA content. *Figure 4-22* demonstrates that when a mariner transposon

insertion occured in a site that is not TA, there was an increased proportion of T and A nucleotides preceeding the insertion dinucleotide. The nucleotide preceeding any dinucleotide beginning with A is predominently T supporting the assumption that insertions occur at TA or after the A in a TAN motif.

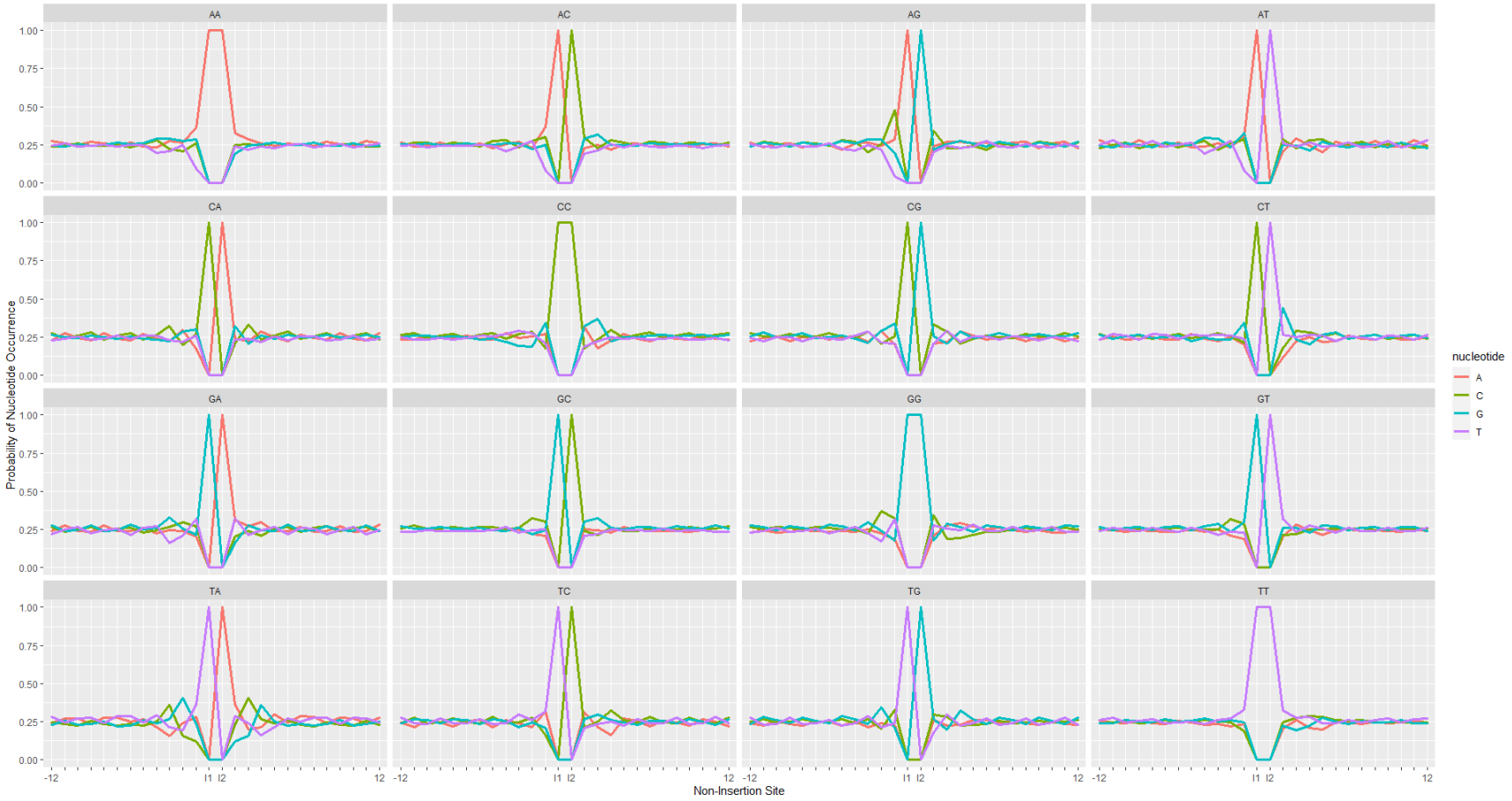

Figure 4-21 Non-Insertion: *Profiles of the probabilities of each nucleotide.*

+/- 12 base pairs from the dinucleotide of non-insertion for the mariner libraries, displayed by dinucleotide.

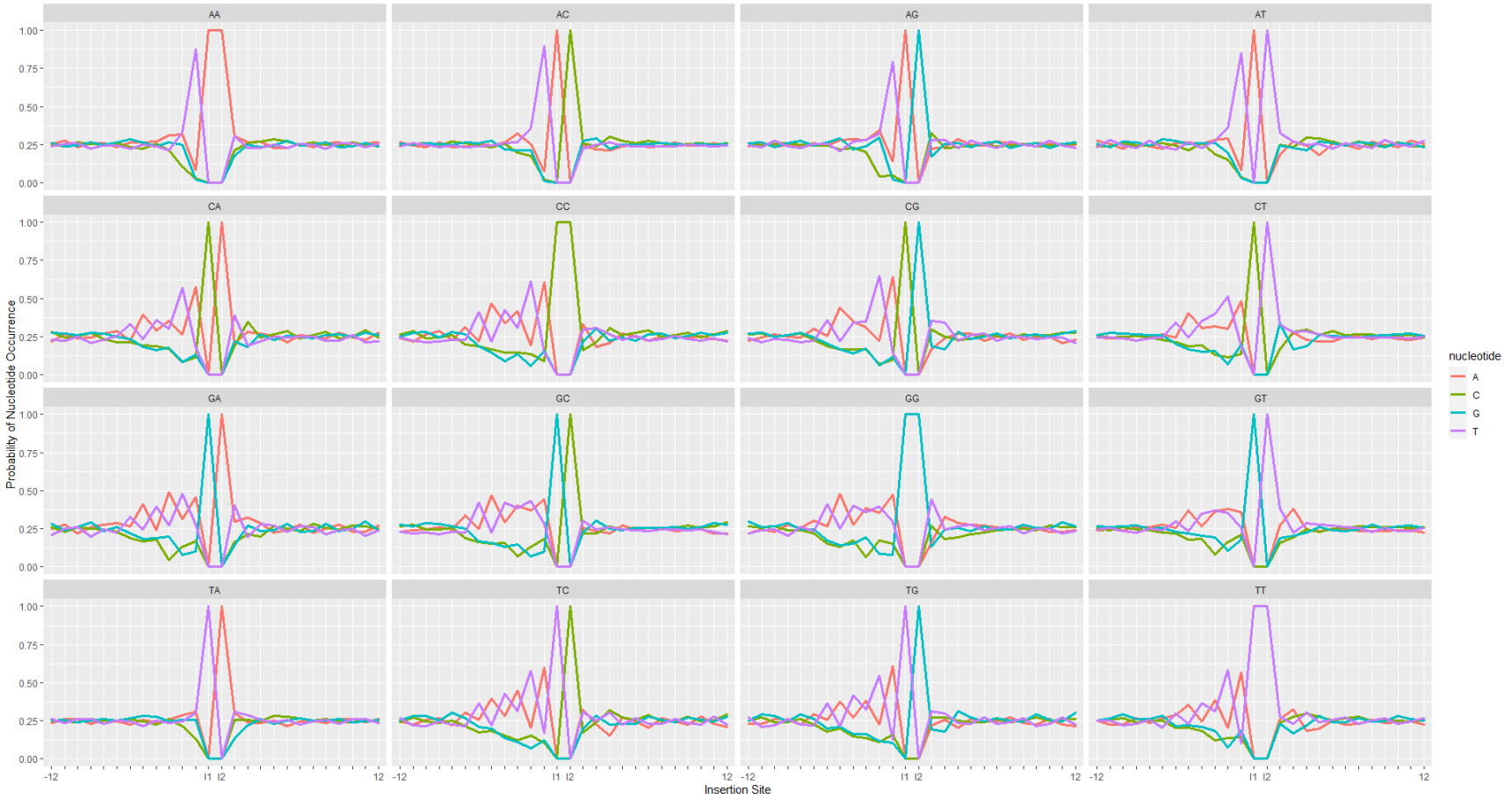

Figure 4-22 Insertion: *Profiles of the probabilities of each nucleotide.*

+/- 12 base pairs from the dinucleotide of insertion for the mariner libraires, displayed by the dinucleotide at the insertion point.

When the nucleotide occurrence surrounding all dinucloetides was plotted by insertion or non-insertion, *Figure 4-23*, the permissive and preventative insertion site preferences became apparent. The increase in TA preceding insertion in non-TA dincucleotides was a non-symmetrical phenomenon, *Figure 4-23:Left*. In other insertion site motif observations, the motifs have been symmetrical or palindromic around the site of insertion. Similarly *Figure 4-23:Right* highlights a non-symmetric pattern where G/C was predominant around dinucleotides with no insertion. However, this was to be expected because a TA dinucleotide is more likely to have an insertion than not have an insertion, *Figure 4-20.*

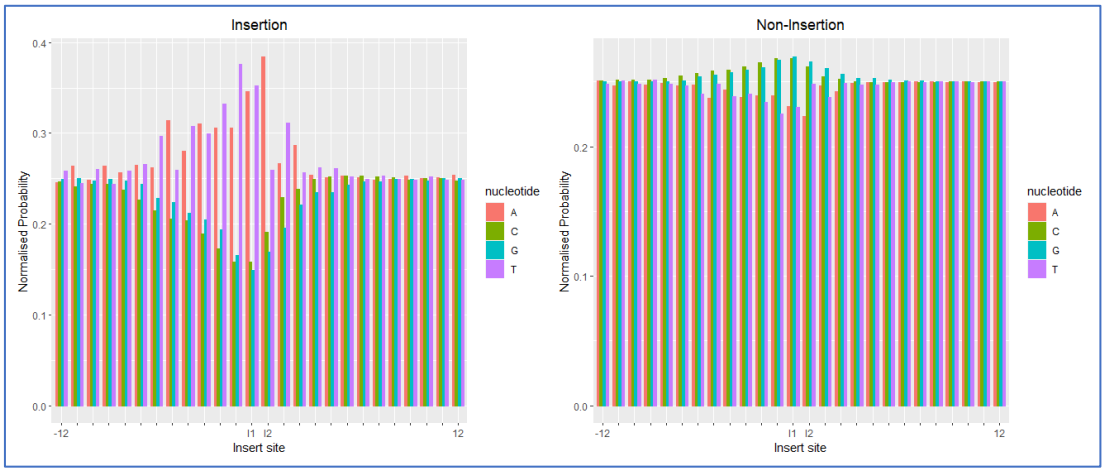

Figure 4-23 *The probability of nucleotide occurrence insertion +/- bp at any dinucleotide for the mariner growth libraries.*

Left: Dinucleotides where insertion had occurred. Right: Dinucleotides where there was no insertion.

The lack of insertion site symmetry suggested that mariner transposon insertion was promiscuous in areas with repeating T/A nucleotides. The insertion did not occur within the repeats (demonstrated by the TA repeats leading up to the TA dinucleotide insertion) but rather at the end of the repeats. This is important to consider when looking at mariner transposon insertion in genomes with high TA content, when looking at promoter regions or in genomic features with repeating T/A nucleotides.

4.3.3.2 Tn*5* Insertion Bias – Dinucleotides

The hyperactive Tn*5* transposon used for this work is documented to have little insertion bias (Morris et al., 2016; Reznikoff, 2003) with reports in the literature of preferential

insertion sites (Goryshin et al., 2003; Green et al., 2012; Morris et al., 2016). Looking at the 26 base pairs spanning the insertion site indicated a marginal preference for G or C nucleotides at the immediate insertion site and +/- one base pair in the data presented here. As with the mariner libraries, insertion site was investigated in regard to the dinucleotide at the site of insertion. For Tn*5*, the dinucleotide distribution throughout the genome, the number of instances of an insertion and non-insertion are listed in *Table 4-5* and the proportion of insertions occurring at each dinucleotide is shown in *Figure 4-24*.

Table 4-5 *The number of insertions or non insertions at each dinucleotide in the Tn5 growth libraries and the proportion of the total insertions at each.*

	Insertion		Non-Insertion			
	Count	% Of dinucleotides in genome	Count	% Of dinucleotides in genome	Occurrence in Genome	% Of Total Insertions
AA	3397	1.00	335615	99.00	339012	2.57
AC	5038	1.97	250078	98.03	255116	3.82
AG	3028	1.28	232720	98.72	235748	2.29
AT	5570	1.80	303777	98.20	309347	4.22
CA	13104	4.07	308603	95.93	321707	9.93
CC	13376	4.96	256279	95.04	269655	10.13
CG	6517	1.88	339444	98.12	345961	4.94
CT	15817	6.66	221610	93.34	237427	11.98
GA	8505	3.19	258287	96.81	266792	6.44
GC	13457	3.51	369621	96.49	383078	10.19
GG	10389	3.83	260753	96.17	271142	7.87
GT	17283	6.74	239023	93.26	256306	13.09
TA	3715	1.75	207997	98.25	211712	2.81
TC	4462	1.67	262439	98.33	266901	3.38
TG	3509	1.08	320958	98.92	324467	2.66
TT	4850	1.44	332697	98.56	337547	3.67
All	132017	2.85	4499902	97.15	4631919	100

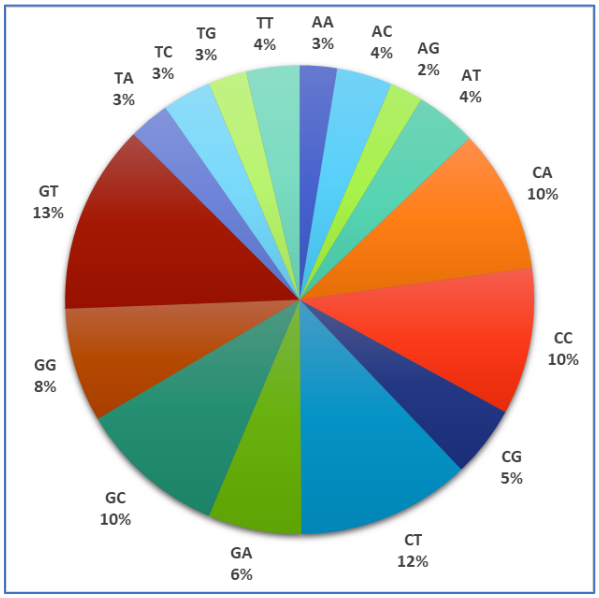

Figure 4-24 *The proportion of Tn5 transposon insertions occuring at each dinucleotide across the E. coli BW25113 genome.*

The dinucleotides were not represented equally across the *E. coli* BW25113 genome, however the proportion of the dinucleotide sites containing a transposon insertion varied from 1 per cent to 6.74 per cent depending on the dinucleotide. At this saturation level, and assuming no dinucleotide bias, it was expected that transposon insertion would occur at 2.85 per cent for any dinucleotide. The proportions of dinucleotide sites that contained an insertion and those that did not are shown in *Figure 4-25* with the average for a library at this insertion density, denoted by the dashed line.

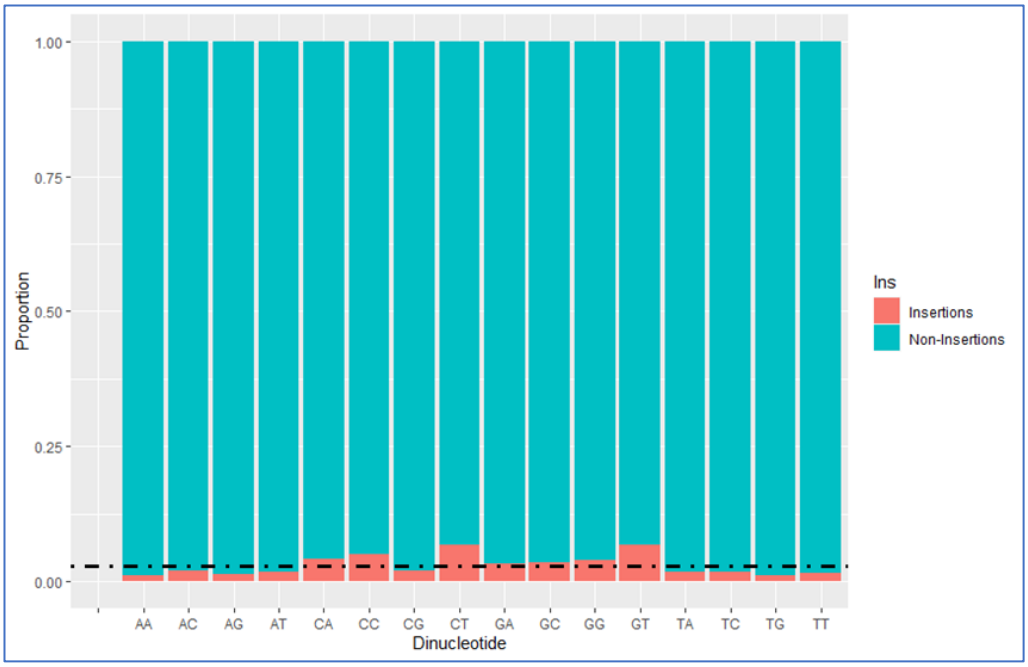

Figure 4-25 *The proportion of insertions or non-insertions at each of the sixteen dinucleotides.*

Data shown for all of the Tn5 growth libraries, the dashed line represents the expected insertion proportion for libraries at this density assuming no insertion bias.

There was an increase in insertion when either G or C was the first nucleotide in the dinucleotide under investigation, but insertion occurred at a similar rate whether it was G or C, 37.6 and 37 per cent respectively. Insertions at dinucleotides starting with A or T occurred less frequently but relatively evenly at 12.9 and 12.5 per cent respectively. These results are summarised in *Table 4-6* and suggest that at the dinucleotide level, insertion was determined by the first nucleotide, which was more likely to be a C or a G, this was consistent with the literature reports that Tn*5* insertion is more frequent in high GC content areas. After that, insertion was more likely to occur if the second nucleotide was a T or a C based on data presented here.

Table 4-6 *The proportion of nucleotide occurrence as the first or second member in the dinucloetide.*

	1st Nucleotide (%)	2nd Nucleotide (%)	Occurrence In Genome (%)
A	12.90	21.76	24.60
C	36.98	27.52	25.36
G	37.60	17.76	25.42
T	12.53	32.97	24.62

For all of the Tn5 libraries and how that compares to the genomic proportion for each nucleotide.

If the structure of DNA at the insertion site was either permissive or preventative for transposon insertion, this would have been dependent on the sequence motifs beyond the dinucleotide at insertion. The twelve bases upstream and downstream of the insertion dinucleotide were taken and the probabilities of nucleotide occurrence was plotted for the non-insertions, *Figure 4-26.* Unlike with the mariner insertions, there did not appear to be any pattern across all dinucleotide combinations that suggested a non-permissive pattern.

The nucleotide occurrence surrounding the dinucleotide of insertions was also plotted, *Figure 4-27* . The insertion data from this work suggested an increased probability of nucleotides around the insertion dinucleotide and gave a motif of nn**CAG**nnn**G**nn(II)n**C**nnn**CTG**nnn; this motif is palindromic around the first nucleotide of the dinucleotide insertion site observed. The palindromic nature of this sequence may have been attributable to the bidirectionality of Tn*5* transposon insertion and amplification during sequence preparation. The markedness of this pattern was varied amongst the dinucleotides observed but elements were common throughout all 16 nucleotide combinations.

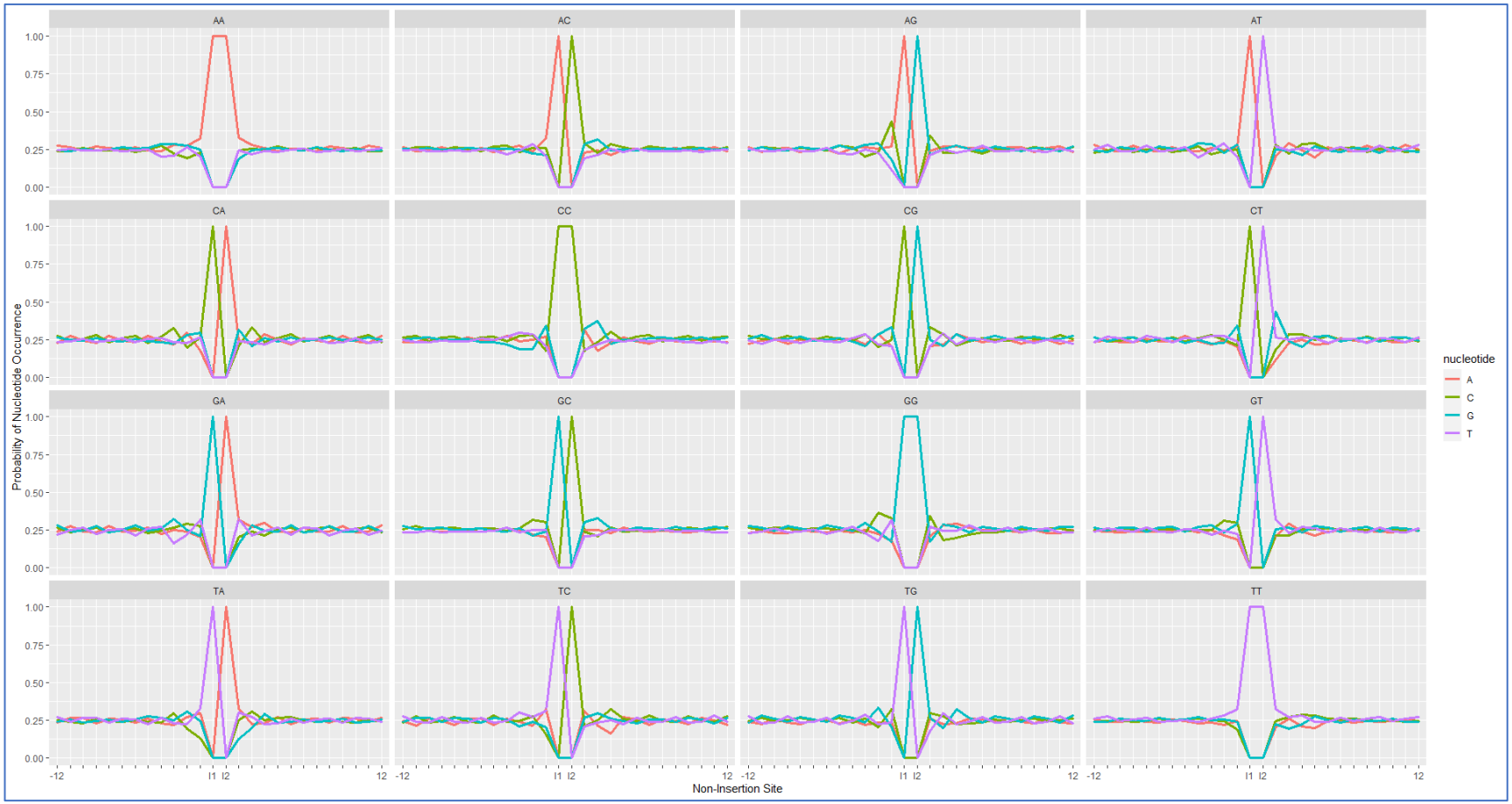

Figure 4-26 Non-Insertion: *Profiles of the probabilities of each nucleotide.*

+/- 12 base pairs from the dinucleotide of insertion for the Tn5 libraries, displayed by the dinucleotide at the insertion point.

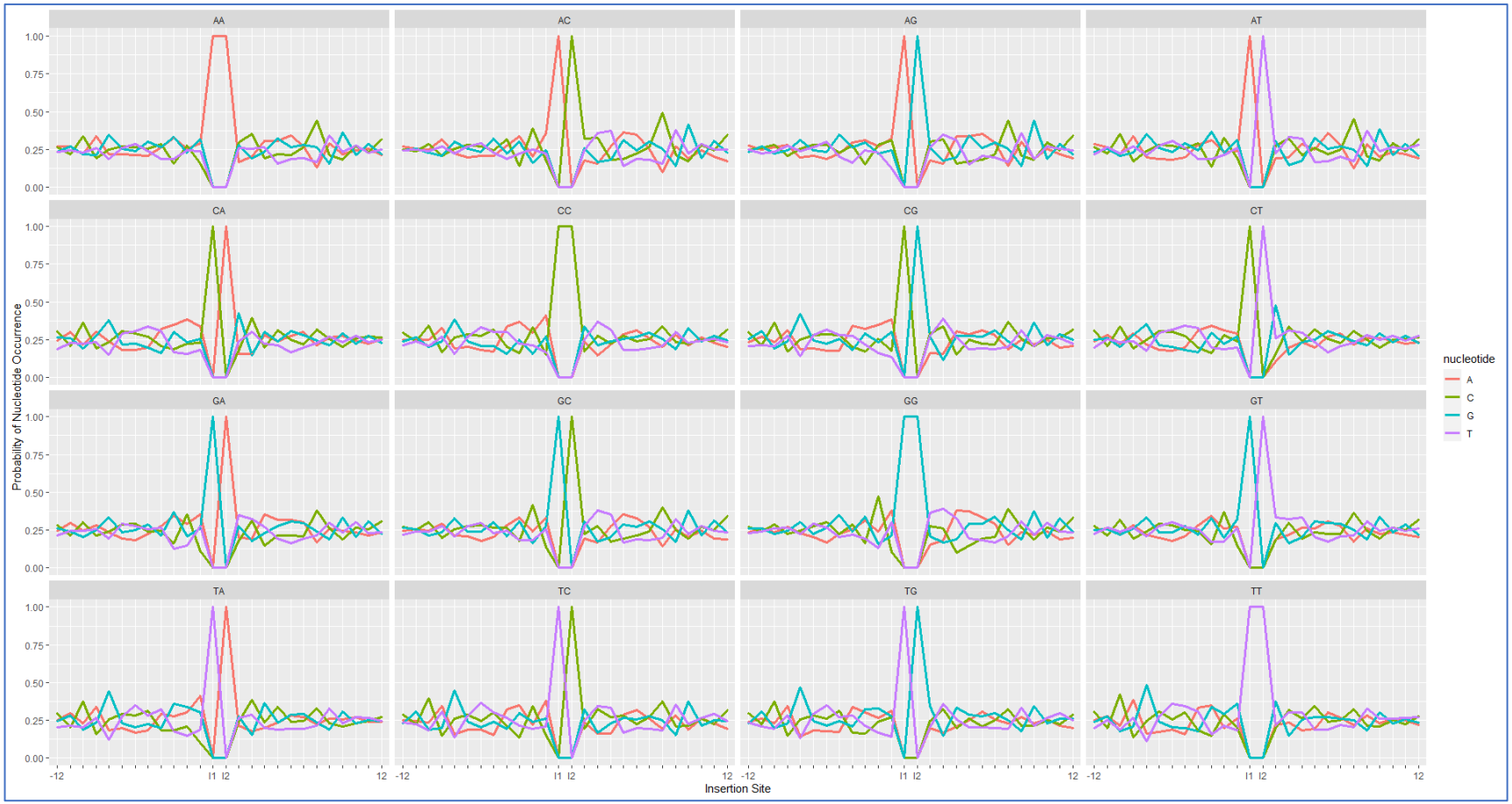

Figure 4-27 Insertion: *Profiles of the probabilities of each nucleotide.*

+/- 12 base pairs from the dinucleotide of insertion for the mariner libraries, displayed by the dinucleotide at the insertion point.

This was not the same consensus sequence that Green et al. found in their work (Green et al., 2012). However, their investigation was using *Candida glabrata* genomic DNA which has a median GC content of 38.6 per cent (NCBI genome ID:192). Additionally, they have reported insertion site preferences for Tn*7* and Mu transposons in the same work but have only normalised the Mu insertion site preferences to the genome content of each nucleotide. Therefore, the insertion site sequence for Tn*5* may be different once normalised. The general consensus sequence identified is mostly an increased abundance of G and C around the insert site.

Figure 4-28 shows the probability of each nucleotide that was in the 24 base pairs around the insertion dinucleotide for the Tn*5* transposon libraries made in the work presented here. For Tn*5* it appeared that the insertion occurred at the first nucleotide within the centred dinucleotide, so this was considered be the reference point and the insertion site. *Figure 4-28:Left* shows that there was much more variation in the probability of a nucleotide occurring around sites of insertion than where there was not an insertion, *Figure 4-28:Right,* this indicated that there was an insertion site bias for Tn*5*.

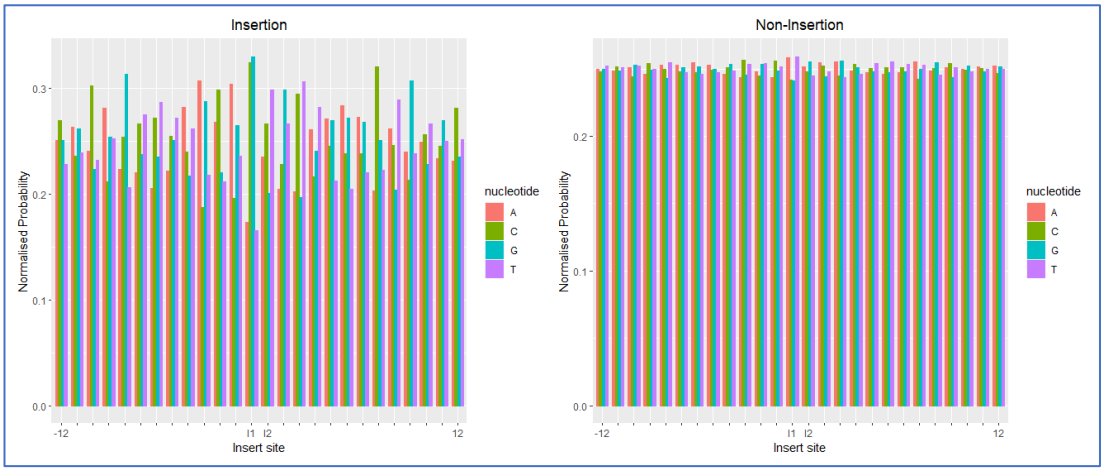

Figure 4-28 *A: Figure 4-29 The probability of nucleotide occurrence insertion +/- bp at any dinucleotide for the Tn5 growth libraries.*

Left: Dinucleotides where insertion had occurred. Right: Dinucleotides where there was no insertion.

Goryshin et al. have studied the insertion site of Tn*5* with 384 *in vitro* insertions into the pRZTL1 plasmid, specifically the Cam^R gene originally from the pACYC184 plasmid(Goryshin et al., 1998). Transposed DNA was transformed into *E. coli* K12 to select for successful transposition events. These insertions identified a preference for a G at the site of insertion and a C nine bases from insertion. The data from this work was consistent with that finding, Green et al. report the same. The resulting consensus insertion sequence reported from Goryshin et al is A-GNTYWRANC-T (N = any nucleotide, Y = T or C, W = A or T, R = A or G); with the first G being the site of insertion and that the pattern is symmetrical around the G of insertion. This pattern could be identified in *Figure 4-28:Left*.

Using *Figure 4-28*, the observable pattern appeared to be (G/C)TGYWNRRCAGNNN, there did appear to be symmetry around the insertion G/C, but palindromic; as previously mentioned, this may have been be due to bidirectionality of the transposon insertion and sequencing. However, Goryshin et al. also comment on symmetry in their identified motif and state that their methods only enable one direction to be considered (Goryshin et al., 1998); so, the insertion site was symmetric.

It is hypothesised but not yet determined that as Tn*5* transposition generates a nine-base pair duplication, the target site is nine base pairs long. Shevchenko et al. suggest that the target site is 11-13 base pairs long with a core of nine base pairs (+/- up to 4 base pairs from insertion G/C) (Shevchenko et al., 2002). Using data generated here and *Figure 4-28* there was evidence to suggest that the target site is at least 11 base pairs long (+/- 10 base pairs from insertion G/C).

The motif identified was not absolutely required for insertion; *Figure 4-27* and *Figure 4-28* demonstrated that other nucleotides can occur within the insertion site. There was no evidence of a sequence motif that is preventative to transposon insertion. The same two figures show that there was an equal chance of any nucleotide occurring in the 26 base pair sequence across an area of no insertion. The data for this figure was normalised to the genomic average occurrence of each nucleotide, which for *E. coli* K12 rounds to 25 per cent for each. In an organism where the nucleotides are not present in equal amounts and GC content is not around 50 per cent, a non-permissive or less permissive pattern may emerge.

4.3.4 Genome Composition

It is generally accepted that mariner transposons insert at TA dinucleotides (Lampe et al., 1999); the data presented in this chapter showed that this was the case most of the time. Hence it would be expected that in areas of low GC content of the genome there would be an increase in mariner insertion and a reduction in areas of high GC content. However, this did not appear to be the case and can be seen in *Figure 4-30* where the opposite occurred. Regions within the dashed boxes show that in regions of above average GC content, left, there was no change to the insertion density compared to the surrounding regions. When there were regions of below average GC content, right, there appeared to be a reduction in the insertion density compared to the surrounding areas.

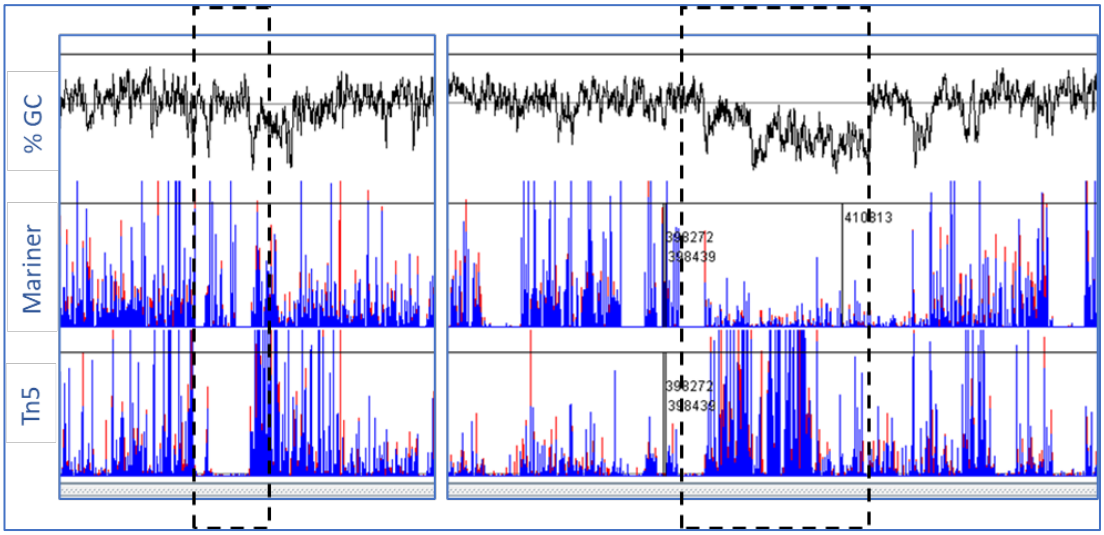

Figure 4-30 *Artemis visualisation to compare the insertion profiles to GC content.*

Showing insertions from all of the grown mariner library insertions and all of the grown Tn5 library insertions with the average GC content plotted in the top panel. The dashed boxes higlight an are of above average GC content (left) and below average GC content (right) relative to the rest of the genome.

The opposite was true for Tn*5*, despite having an insertion site preference containing mostly G or C, areas with above average GC content had decreased insertion density and areas of below average GC content had an increased insertion density. The apparent Tn*5* preference for AT rich regions was also observed with Tn*4001,* a member of the Tn*5* family (Miravet-Verde et al., 2020). Both of the highlighted regions in *Figure 4-30* contain sites

that were essential or protected from transposon insertion for both mariner and Tn*5* transposons irrespective of the regional GC content. These observations were in *E. coli* K12 BW25113 which has an average GC content of around 50 per cent across the genome. Organisms where there is more fluctuation in GC content of the genome may have differing transposon insertion patterns relating to regional variations in GC content.

4.4 Discussion

The work in this chapter aimed to identify factors that can influence transposon insertion variabilities that are not commonly considered when using transposon insertion profiles to classify genes as essential or non-essential. Not all of the factors can be quantified, for example it is assumed that all mutants will be equally represented in the pool and with an assumed 5 x 10^5 unique mutants, there should be 2,000 cells of each mutant type in the extraction pool of 1 x 10^9 cells; this may not be the case. This is a bottleneck (Mahmutovic et al., 2020) and could potentially be overcome with replicate extractions and sequence preparations of a single transposon mutant library, as discussed in *Chapter 3*. Additionally, should transposon insertion disrupt cell wall biosynthesis then there will be an over or under representation. This sampling bias could be the cause of the large peaks seen in the Tn*5* library in the genes *nluD* and *rpoS*.

In this work, as demonstrated by WGS mapping to the reference genome, there was no real methodological bias introduced when identifying Tn-Chr junctions in TIS libraries despite it being documented that Illumina short read sequencing is susceptible to variations in coverage based on genome content (Gaio et al., 2022), but to a lesser extent than previous sequence library preparation protocols (Bruinsma et al., 2018) . It is important to consider that this may not be the case with every organism encountered, *E. coli* K12 is a model organism and easy to manipulate yet shows a high level of laboratory adaptation (Browning et al., 2022). While the Illumina sequencing protocols are developed to be robust across a broad range of genomes, the efficacy across organisms varies and the unbiased WGS mapped data presented here for *E. coli* K12 may not be representative of pathogenic organisms of interest in using TIS methods for antimicrobial development (Subashchandrabose et al., 2013; Mobegi et al., 2014). The WGS mapping approach shows that preparation and sequencing of the genomic DNA does not account for variability in transposon insertion, however, it cannot represent any bias introduced from the selection of mutants or procedure for DNA extraction.

The impact of growth on the variability in transposon insertion has been shown in data presented here. Despite the low insertion density, libraries that had undergone limited growth showed that transposon insertion across the genome was not consistent. Data from this chapter highlights that there was a non-consistent profile of transposon insertion across the genome and that patterns could be observed in libraries of permissive growth and where growth had been restricted to fewer than two generations. The insertion profiles showed regional patterns rather than increased insertions at specific base pairs suggesting factors other than DNA sequence influence insertion probabilities. Analysis of a limited growth library also highlighted areas of the genome that are absent of transposon insertion for both a Tn*5* and a mariner transposon. This may not present a problem when comparing conditionally essential genes, but it could falsely highlight a candidate drug target to be perused.

Normalisation to the limited growth library can change the insertion profile of a library that has undergone permissive growth. The impact of normalisation on essential gene determination was not explored in this work. However, Miravet-Verde et al. investigated the effect of passaging mutant libraries on the determination of essential genes, showing that one passage introduces a sampling bottleneck and cell division. They found that as more passages, or cell divisions occur, the accuracy of identifying essential genes decreases and more genes are reported to be essential; they recommend allowing fewer than 30 cell divisions in generating a library (Miravet-Verde et al., 2020).

The two transposons used in this work are commonly used for TIS studies (Holden et al., 2021; Langridge et al., 2009; Nazareno et al., 2021; Nlebedim et al., 2021; Sivakumar et al., 2019; Thibault et al., 2018; van Opijnen et al., 2009; Weerdenburg et al., 2015) and at the genome level there are few observable differences in the insertion patterns, assuming that the insertion densities are comparable. However, some differences in insertion became apparent at the gene level. The benefit of high throughput mutagenesis combined with in depth sequencing is resolution and at increased resolution the differences between mariner and Tn*5* transposon insertion increase, and this is likely due to the insertion site preferences of the two transposases. The data presented here demonstrates that while amplification jackpots are present in TIS sequence mappings, these occur randomly and are not sequence dependent whereas transposon insertion is sequence dependent.

Mariner does not only insert at TA dinucleotides as commonly stated (Chao et al., 2016; Morris et al., 2016; Larivière et al., 2021); this work indicates that there is a strong

preference but not an exclusive requirement for this. The generally accepted notion that insertion is TA exclusive has made mariner a preferable family to use as it is easier to achieve chromosome saturation (Chao et al., 2016). Hence, it has been used in *Mycobacterium tuberculosis* to generate saturated mutant libraries, with a genomic GC content of 65.5 percent (Dejesus et al., 2017). The authors achieved insertions in 84 per cent of the available TA sites but their analysis was limited to only TA dinucleotides.

The authors themselves state that 13 per cent of ORFs were discounted for not containing sufficient TA sites. Using data from the work presented in this chapter, DeJesus et al. may have overlooked up to 20 per cent of insertion sites or up to around 60 per cent depending on what is considered as the site of insertion. Additionally, many of the non-TA insertions occur at the end of TA repeats, these could be determined essential if exclusively looking at TA sites.

Tn*5* shows less overall insertion site requirement but still has a preferred motif that is G/C rich. This has previously been reported (Goryshin et al., 1998; Green et al., 2012; Miravet-Verde et al., 2020) and can be identified in this work. However, the published motif had not been normalised to the genomic average for each nucleotide and may account for the identified motif being less apparent in this work. The motif for Tn*5* is predicted to be 11 base pairs long. There does not appear to be a non-permissive motif so each 11 base pair combination will offer a different chance of transposon insertion across the genome.

The two transposons used have differing insertion site preferences, however, these regions are incredibly small when compared to the size of a genome. In terms of generalised sequence as a predictor for insertion propensity, GC content is not an accurate predictor of transposon insertion density (Miravet-Verde et al., 2020), therefore there are other factors affecting the likelihood of a transposon to insert into the DNA. More importantly, understanding the areas of the genome that are inaccessible for transposon insertion are important, whether that be determined by sequence or architecture. In some cases, DNA sequence determines architecture (Murlidharan Nair, 2010) with TA contributing to DNA curvature particularly in area of high GC rich content (Dlakic & Harrington, 1995).

The growth state of the organism at the time of transposition will also have an influence on the distribution of insertion sites. Rapidly growing cells have more insertions located at the origin and terminus of chromosome replication due to the increased copy number of DNA on replication forks (Chao et al., 2016); this was seen with the Tn*5* data. Goryshin et al. state that Tn*5* transposition occurs more frequently in actively transcribing or super-coiled

DNA (Goryshin et al., 1998) and could explain why this effect was only observed for Tn*5* and not mariner. If cells are in stationary phase, then chromosome replication is not occurring, so the distribution of transposon insertion would be expected to be more consistent, as seen with the mariner data. Another consideration is that if cells are in exponential phase during transposition, there will be increased replication machinery and other DNA binding proteins blocking access for the transposon.

One of the predicted factors affecting transposon insertion are DNA blocking enzymes. HNS blocking has been observed with a mariner transposon(Kimura et al., 2016). This has only been reported for HNS but is likely to be common for all DNA binding proteins. In *E. coli* K12 there are 256 transcription factors with 3980 binding sites (Keseler et al., 2017). The limited growth libraries used here did not have sufficient density to recognise regions at this resolution. Future work would be to increase the density of these libraries. Alternatively, CHIP-seq could be used to locate DNA binding proteins (Bailey & MacHanick, 2012).

Chromosome condensation and architecture will affect the access of a transposon, this effect has not been studied here. Other than limited or no growth transposon libraries, Hi-C could be used to investigate the conformation of the chromosome (Eagen, 2018) and if this is related to transposon insertion. Another factor not investigated in this work is whether methylation of the DNA influences transposon insertion. It may be that a larger proportion of essential genomic regions are physically protected from transposon insertion as an evolutionary tactic.

4.5 Conclusions

The use of TIS for essential gene calling is now well established but the results presented here suggest that with more careful experimental and analytic design a much higher resolution could be achieved if the following are more carefully controlled. The likelihood of transposon insertion is dependent on the transposon used, access to the DNA and any growth advantage or disadvantage provided by the mutation. Understanding these factors will enable development of more accurate models for determining essential genomic regions. The BioTradis pipeline uses the number of unique insertions within each gene for insertion index prediction. The model may be more refined if the number of total insertions within a gene is used. In this case, normalisation to the number of sequence reads would be required. Work in this chapter highlights that additional normalisation to a limited

growth library may further improve the accuracy of determining gene essentiality. To further explore the advantage of a limited growth library, more sequence data would be required, these libraries require a greater sequencing depth than standard due to donor and recipient cells being in the final pool.

A full analysis model would have to be able to account for insertion site sequence and normalise against a limited growth library. Additional layers of refinement could be included, such as modelling transposon insertion site preferences, identifying DNA binding sites for proteins and chromosome structure. Once these factors are accounted for, the insertion count data would be viewed in the context of a genome annotation. This approach would require a lot of organism specific testing and library production which may not be accessible to every laboratory. However, limiting the growth of a mutant library during generation can be included in protocols. The next chapter will discuss approaches towards generating an analysis model that can incorporate some of the biases highlighted in this chapter.

5 Developments Towards an Annotation Independent Essentiality Prediction Model

5.1 Introduction

5.1.1 Approaches to Analysing TIS Data

Currently, there is no consistent method for analysing TIS data (Larivière et al., 2021) as there are multiple tools available that have been used to determine gene essentiality. There are however three main statistical approaches used to classify genes as essential or non-essential. Firstly, there is the Gumbel or Extreme Value distribution approach (DeJesus et al., 2015; Goodall et al., 2018) which provides an estimate of the maximum distance that would be expected between insertion sites. If a distance greater than this occurs, then the gene can be considered essential. A Gumbel or Extreme Value distribution is limited by the transposon used and any insertion site preferences, any region lacking suitable insertion sites will be designated essential as discussed in *chapter 4*.

The second approach is to use regression and has been developed from studies in RNA-seq data (Barquist et al., 2016). The chosen distribution model determines the number of insertions anticipated in a specified length, for example a gene (DeJesus et al., 2015; Sivakumar et al., 2019; Zomer et al., 2012). This is then regressed to provide limit of insertion to designate the region, gene, as non-essential.

The third commonly used approach is to use a Hidden Markov Model (HMM) in which the gene, or region of interest, is assigned a state (DeJesus & Ioerger, 2013). In TIS the states would be considered essential for growth or non-essential for growth.

5.1.1.1 Genome Annotation

Most of the tools developed for TIS approach the data from the perspective of generating a list of genes, the products of which are essential for growth under the conditions that are being investigated. Genome annotation has been incorporated in the algorithms for some tools, for example BioTradis (Barquist et al., 2016; Langridge et al., 2009) used gene length to determine the number of expected transposon insertions in that region or Tnseq-Explorer (Solaimanpour et al., 2015) uses the number of TA sites within a gene. These approaches have been successfully used to generate prospective essential genes but as previously discussed, genome annotations change as more research is undertaken; this makes essential gene lists somewhat variable over time.

Some tools, namely TRANSIT (DeJesus et al., 2015) offer analysis in an annotation independent fashion; they employ a sliding window approach to assess whether a

proportion of the genome has the number of insertions expected, this approach is similar to those used for annotation dependent models except the size of the region of interest is set. These windows can be overlapping, or discrete, discrete windows offer faster processing at a lower resolution.

5.1.2 Proposed New Modelling Approach

The aim of this chapter is to explore the use of an HMM or a changepoint detection algorithm to determine essential or protected regions of the genome in an annotation independent manner. Following protected region determination, annotation overlay should give protected areas of the genome, independent of CDS.

5.2 Methods for this Chapter

The procedures used to generate insert count numbers from sequence data were fully described in *sections 3.2.2.4– 3.2.2.8*. The methods detailed below describe the use of R scripts written by Dr. George Savva that were used for importing insert count files, *appendix 9.5.1,* and executed in R 4.2.1 (R Core Team, 2022). The TIS libraries that were modelled have been described in *chapters 3 and 4*.

5.2.1 Hidden Markov Model

A Hidden Markov Model (HMM) was developed by Dr. George Savva using mariner TIS insertion data from nine replicate libraries (using preliminary data not shown). Each base pair was assigned to one of three states: the first designated state represents areas of the genome that can tolerate insertions and was assigned to most of the genome. The second and third states both represent regions that are unable to sustain insertions. State two and three had a different threshold for the transition between each and the first state. Hence state three represented large gene scale protected areas. State two represented shorter areas. The model was estimated using the Baum-Welch algorithm with the Viterbi algorithm (Y. Zhang et al., 2014) used to predict states, using the HiddenMarkov package version 1.8-13 (Harte, 2021) for R version 4.1.1 (R Core Team, 2022).

The genome annotation overlaid with the HMM was manually added to the image. The annotation was visualised in Geneious version 2021.2.1 (www.geneious.com), and the

genes coloured as essential based on the Keio (Baba et al., 2006) collection the Goodall et al. TraDIS data (Goodall et al., 2018)and listed in the Ecocyc database (Keseler et al., 2017).

5.2.2 Segmentor3IsBack Model

Segmentation of the genome was performed using the R package Segmentor3ISBack (Cleynen et al., 2014) version 2.0. The package was downloaded from GitHub as the package was unavailable from the CRAN repository. The package was run in R version 4.2.1. Again, Dr. George Savva wrote the original R scripts to model and visualise the data using the ggplot2 R package version 3.4.0. I modified some elements of the original scripts to enable annotation to be added to the visualisations. The final script used is included in *appendix 9.5.3*. The model was constrained to a maximum of 300 segments per 100,000 bp and the maximum number of insertions at any base point was restricted to 100 to avoid tailing negative binomial distributions. For tuning, the model was set to the oracle penalty. The output was visualised using the ggplot2 (Wickham, 2016) R package version 3.4.0. The script used is included in appendix G. For the segmentation models, the R package gggenes (Wilkins, 2020)version 0.4.1 was used with the ggplot2 package (Wickham, 2016) to draw annotated features as arrows and plot on a ggplot object. This allowed annotation overlay onto the modelled data in the same plot area.

5.2.3 OnlineBcp Model

The other segmentation package used to identify break points in the insertion data was OnlineBcp (H. Xu et al., 2022). Version 0.1.8 was run in R version 4.2.1 with the default parameters and the insertion count at each base. The output was visualised using the ggplot2 package version 3.4.0. The script used is included in *appendix 9.5.4*. The gggenes package was used to provide annotated features to the visualisations of the model.

5.3 Results

5.3.1 Hidden Markov Model

The HMM predicted protected regions in the first 125,000 bp of a published *E. coli* BW25113 genome, whether that be physical or functional. *Figure 5-1A* shows the number of replicates where there was a transposon insertion at that genome location. The HMM

effectively predicted areas of the genome that were protected from transposon insertion or were required for growth, *Figure 5-1B*. Around 96 per cent of the predicted protected regions aligned with essential genes reported in the literature for the first 125,000 bp of the reference BW25113 genome, *Figure 5-1C*. However, in around 9 per cent of cases, a gene determined as non-essential in the literature was determined essential by the HMM. The gene *dnaK* has been circled, it was reported as essential by TIS using a dense Tn*5* library (Goodall et al., 2018) but has been knocked out with the Keio collection (Baba et al., 2006). The HMM model presented here determined that it was an essential or protected gene, this may have represented an area of physical protection rather than essential function. Work with the HMM was discontinued in favour of a break point detection model due to the break point detection approach being easier to manipulate with normalisation and the segment boundaries were easier to extract.

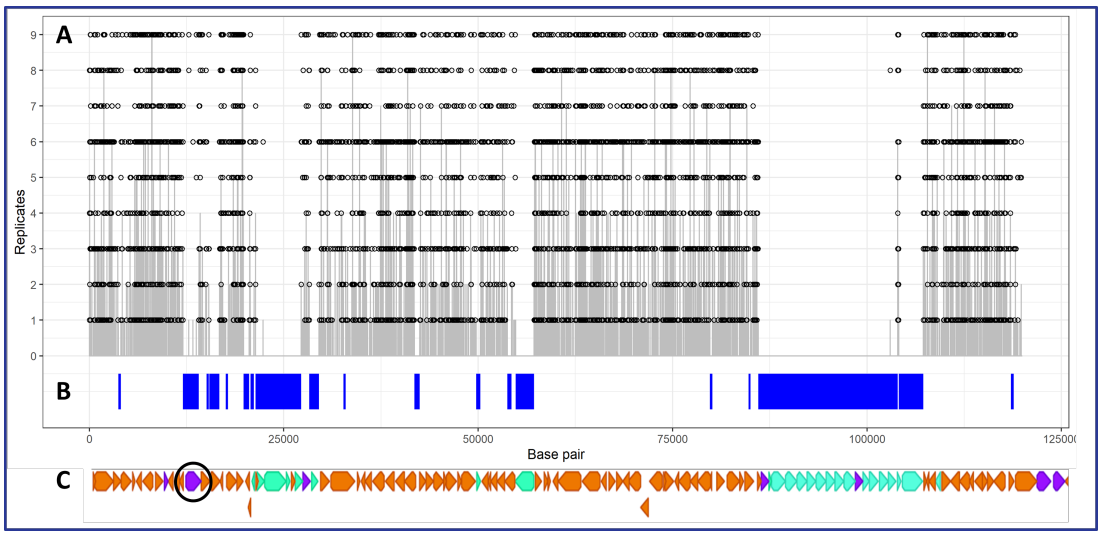

Figure 5-1 *Visualisation of the first 125,000 base pairs of E. coli BW25113 and insertions across all nine replicates.*

A: Circles denote a reported insertion at that position in the genome, the height of the lines represents the total number of insertions. B: regions in the genome, annotation independent, predicted to be essential by the HMM model used. C: Annotated genes, orange – non-essential, turquoise – essential, purple – classified as essential by TraDIS but not knockout studies.

5.3.2 Segmentor3IsBack Model

Segmentor3 (Cleynen et al., 2014) was successful at segmenting the chromosome. The package was developed to detect copy number variation in RNA to determine transcription profiles. The authors used the assumption that the copy number of a region was relative to the level of gene expression. A similar assumption could also be made of TIS sequencing data, where the number of transposon insertions within a genomic region was relative to a fitness cost or benefit when it was interrupted.

5.3.2.1 Model Output Tn*5*

The Segmentor3 model effectively segmented the first 100,000 base pairs of the *E. coli* BW25113 genome based on the Tn*5* insertion data as seen in *Figure 5-2.* This portion of the genome was divided into 128 segments of differing means. The segment sizes were reasonable to correspond with genes, smaller segments were observed, for example the orange section at around 90,000 bp in *Figure 5-2* below.

When the genome annotation was overlaid with the segmentation image, *Figure 5-3,* the model was shown to be accurate. For example, the genomic regions containing the CDS for the essential genes: *hemE*, *murB* and *murI* (Keseler et al., 2017) had a lower mean number of insertions than the adjacent segments and therefore were successfully identified as protected by the model. Additionally, the model identified the essential N-terminus of *ftsN* (Goodall et al., 2018) at around 90,000 bp as described in *chapter 3* and demonstrated that an annotation free approach could provide a better resolution of protected regions of the genome.

Another feature that the Segmentor3 model was able to provide, was to highlight regions of the genome where there was a fitness increase upon transposon insertion; this information is not available using the BioTradis pipeline (Barquist et al., 2016) where it is determined by eye using Artemis plots (Carver et al., 2012). Interestingly, the model identified a fitness increase when interrupting only the N-terminus of some genes; *ppc, ptsA* and *katG* at around 60,000, 71,000 and 79,000 bp in *Figure 5-3* respectively. This phenomenon would have been unnoticed when using the BioTradis essentiality pipeline (Barquist et al., 2016); this would be important in the case of a bifunctional protein where only one function is required. In the event that an antagonist is selected as a potential drug candidate, the non-essential function may be targeted.

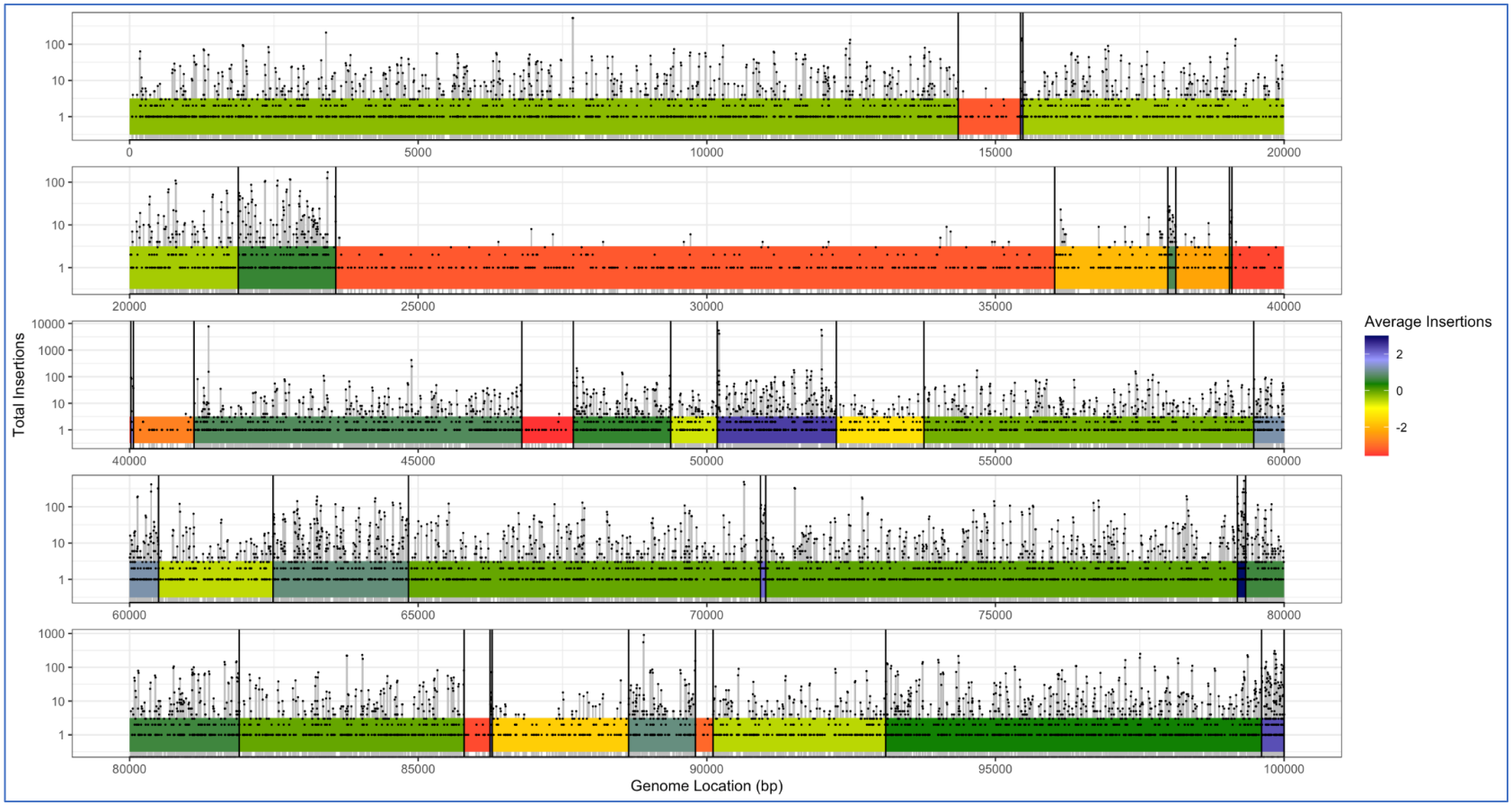

Figure 5-2 *Visualisation of the Tn5 segmentor model.*

The breakpoints identified in the first 100,000 bp of the E. coli BW25113 hybrid reference genome when the Tn5 insertion data was used. Breakpoints are identified by a vertical line, grey lines indicate the number of insertions at each point and the colour indicates the Log_{10} *of the mean number of insertions in that segment.*

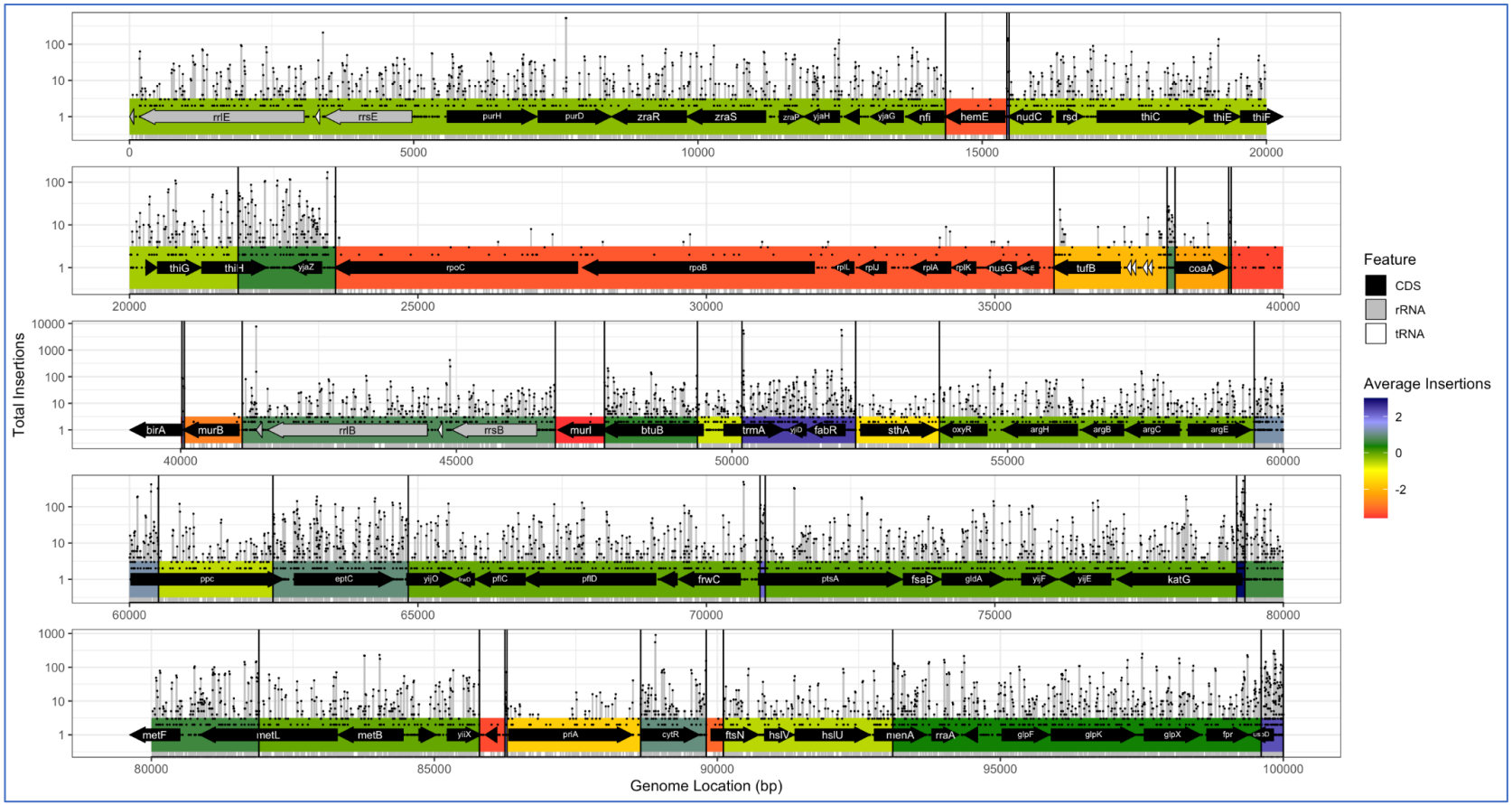

Figure 5-3 *Annotation of the Tn5 segmentor model.*

Annotation overlaid onto the Tn5 model breakpoints , as above, the breakpoints are identified by a vertical line, grey lines indicate the number of insertions at each point and the colour indicates the Log_{10} *of the mean number of insertions in that segment.*

5.3.2.2 Model Output Mariner

The mariner libraries were used to model changes in mean insertion counts within the first 100,000 bp of the *E. coli* BW25113 reference genome and similarly to Tn*5* this region of the genome was effectively segmented. With the mariner insertion data, the genomic region was split into 104 segments and can be seen in *Figure 5-4.*

Addition of the annotation overlay*, Figure 5-5,* showed that again most of the CDSs fell within a segment. The model predicted that *hemE, murB* and *murI* are protected, the same as for Tn*5*. This model also showed the essential portion of *ftsN.* Again, this model showed a fitness advantage when interrupting the N-termini of a handful of genes, this could have been a genuine fitness advantage or could have indicated errors in annotation.

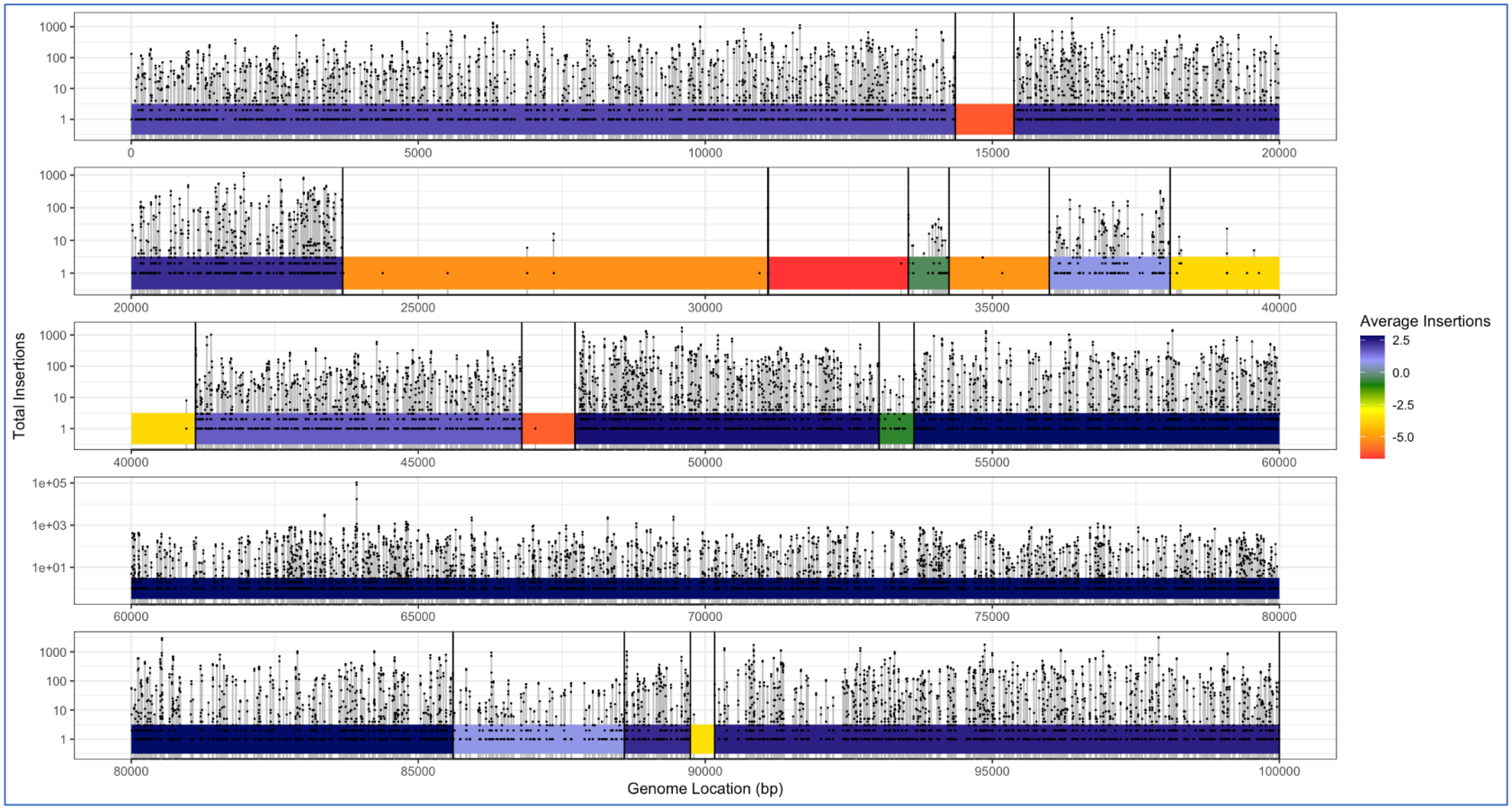

Figure 5-4 *Visualisation of the mariner segmentor model.*

The breakpoints identified in the first 100,000 bp of the E. coli BW25113 hybrid reference genome when the mariner insertion data was used. Breakpoints are identified by a vertical line, grey lines indicate the number of insertions at each point and the colour indicates the Log_{10} *of the mean number of insertions within that segment.*

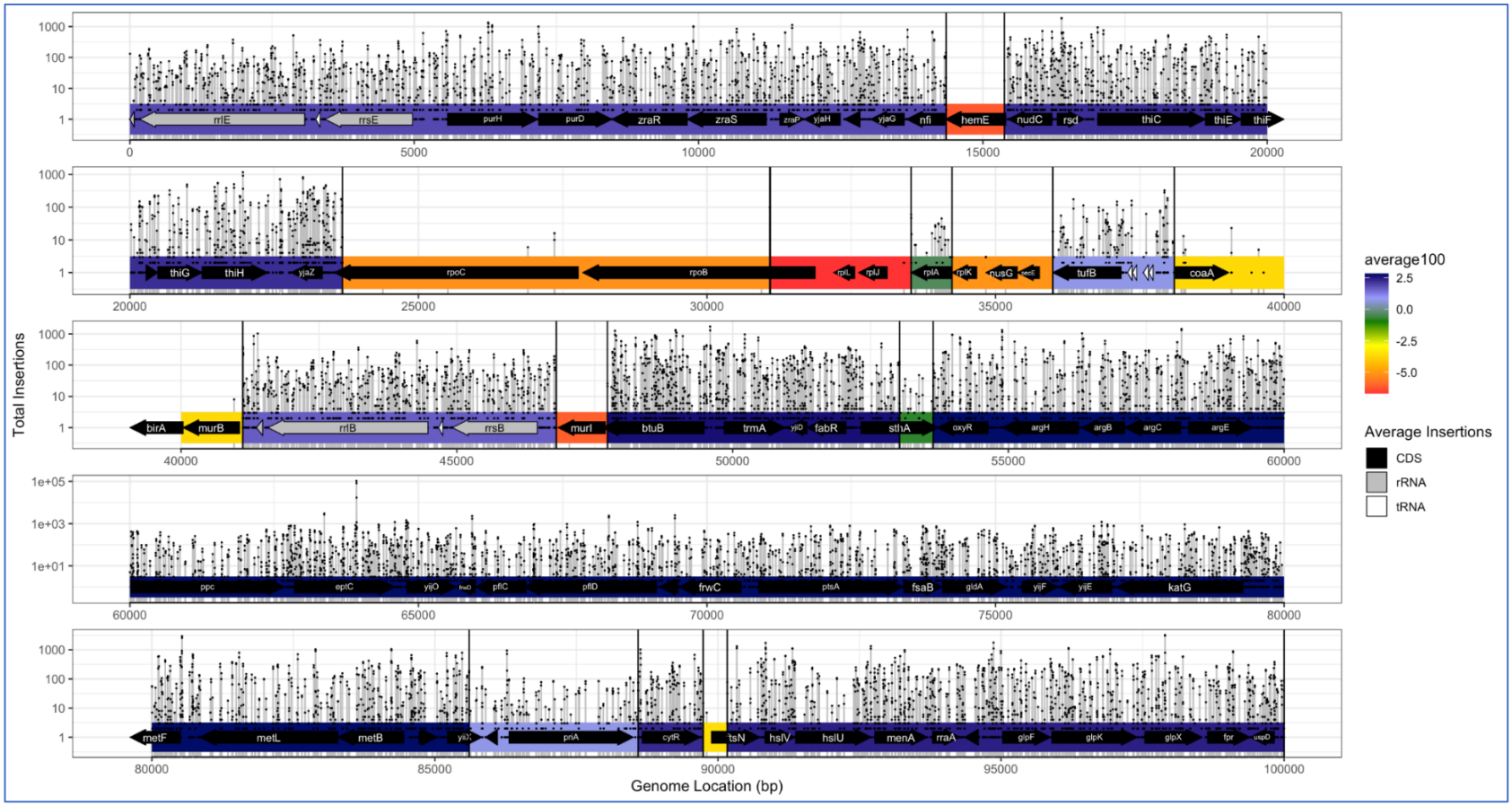

Figure 5-5 *Annotation of the mariner segmentor model.*

Annotation overlaid onto the mariner model breakpoints , as above, the breakpoints are identified by a vertical line, grey lines indicate the number of insertions at each point and the colour indicates the Log_{10} *of the mean number of insertions in that segment*

5.3.2.3 Comparison of the Tn*5* and mariner outputs

Overall, modelling both the Tn*5* and mariner data effectively identified known protected areas of the genome. The Tn*5* model, *Figure 5-3* , indicated that there were 128 segments compared to the 104 segments for the mariner model, *Figure 5-5,* in the 100,000 bp region of genome discussed here. The difference in segmentation is likely due to differences in transposon insertion counts, these differences were explored in *chapter 4.*

The mariner model provided improved resolution from around 23,000 bp to 36,000bp; in the mariner model *rplA* was identified as not protected compared to the surrounding region. Another disparity between the two models was that the Tn*5* model highlighted that *rpmE,* at around 86,000 bp in *Figure 5-3* is protected; the mariner model *Figure 5-5* indicates that this region had a relatively high mean number of insertions compared to other segments. This could have been an example of inaccessibility to the Tn*5* transposon, this gene was not designated essential by the Keio collection (Baba et al., 2006) but not stated for the Goodall et al. Tn*5* transposon library (Goodall et al., 2018).

A striking difference between the two models is the scales and corresponding colours used to visualise the insertion count data, the mariner libraries had fewer insertion counts across the segments and as such the colour scale goes from -5.0 to 2.5 compared to Tn*5* where the scale ranged from -2.0 to 2.0; this has affected the colours that each segment is assigned. A simple modification would be to colour by deviation from the genomic mean number of insertions (z-score) to have more comparable visualisations.

5.3.3 OnlineBcp Model

The OnlineBcp (H. Xu et al., 2022) package was chosen as an alternative to Segmentor3ISBack (Cleynen et al., 2014) as it was recently updated. The package was faster to run than Segmentor3 and able to detect break points in the sequence of insertion counts across the whole genome. However, the algorithm was too sensitive and applied breakpoints excessively, as seen in *Figure 5-6.*

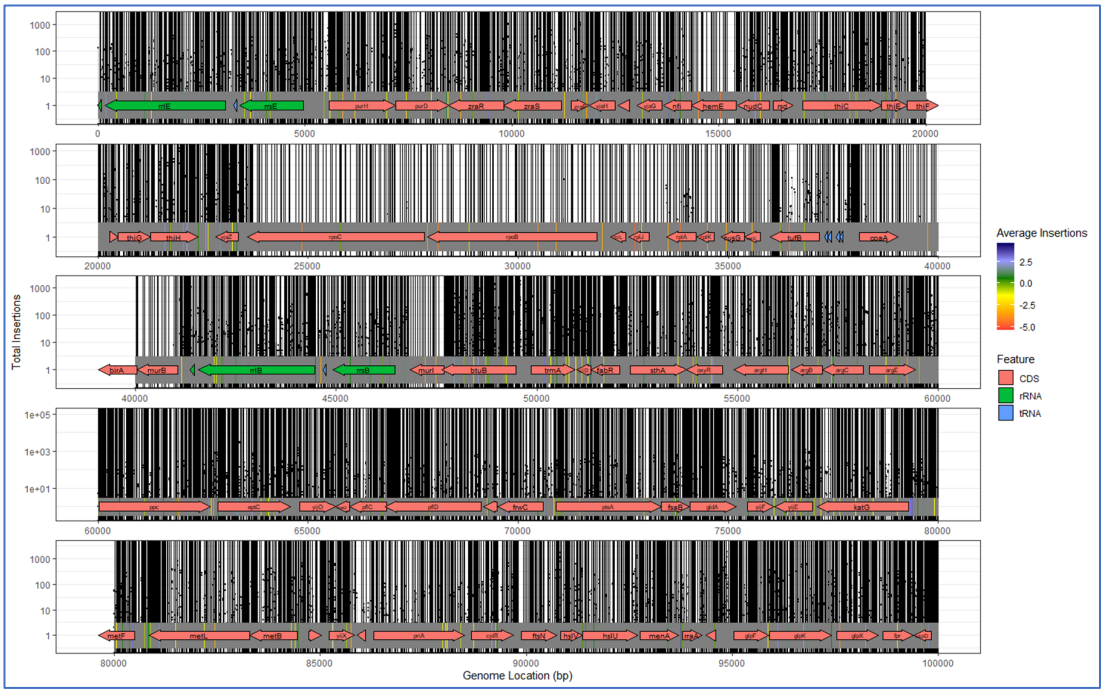

Figure 5-6 *Visualisation of the onlineBcp model with overlaid annotation.*

The breakpoints identified in the first 100,000 bp of the E. coli BW25113 hybrid reference genome when the Tn5 insertion data was used. Breakpoints are identified by a vertical line.

From *Figure 5-6* it can be seen that the breakpoints were assigned at very short intervals, at this spacing it was difficult to determine whether any region of the genome was protected. However, when visualised, there appeared to be more breakpoints in the non-protected regions (e.g., *hemE, murB, murI*), this was due to the low or zero insertion counts having similar means, therefore fewer changes in the mean were detected. There was no option to restrict the number of break points detected (as with the Segmentor3 model), at this sensitivity the package proved not to be useful in using transposon insertion patterns to highlight protected areas of the genome. Further development of this model was not pursued.

5.4 Discussion

TIS data analysis using HMMs is an approach that others have taken, (DeJesus & Ioerger, 2013; Pritchard et al., 2014) implemented an HMM but their approach used and only looked at insertions in TA sites. In an HMM, the state of the previous site is used to partially infer the state of the current site so groups essential or non-essential sites into regions and as such HMMs require calibration to keep the transition probabilities and gene lengths within reasonable sizes. The HMM model developed for this work did determine protected regions of the genome that are concordant with the literature (Baba et al., 2006; Goodall et al., 2018; Keseler et al., 2017). However, there was a false essentiality detection rate of around eight percent when compared to the gold standard knockout collection (Baba et al., 2006; Ghomi et al., 2022). Therefore, we also looked at a segmentation approach and this was favoured due to the ease of manipulation of the data after being modelled.

The segmentation approach using Segmentor3IsBack was the most effective model explored in this work and provided candidate protected genes, determined by visualisation of the data, when the model was run with both the Tn*5* insertion data and the mariner insertion data. The required input is a sequence of count data representing the number of insertions at each base in the genome, with a simple input there can be pre-processing steps. These steps could include normalisation to any of the biases discussed in *chapter 4.*

The work in this thesis has been focussed on essential or protected regions versus non-essential. The focus of many studies is to look at changes in the state of genes under different growth conditions to determine conditionally essential genes (Holden et al., 2021; Mccarthy et al., 2018; Salama et al., 2004; Shields & Jensen, 2019; Warr et al., 2019; Yasir et al., 2020). The proposed segmentation model offers two approaches to identifying these. Approach one is that breakpoints are determined under a control condition, then the breakpoints are left unchanged between test conditions and the mean number of insertions within breakpoints compared. The other approach is to observe how the estimated break points change between a control condition and any tests.

Miravet-Verde et al. developed a tool to standardise TIS data, ANUBIS (Miravet-Verde et al., 2020) which was published in 2020 and offers normalisation of GC content and chromosome location, but is annotation driven. The tool can implement a sliding window approach, overlapping, or not overlapping. However, to combine all of the approaches currently in use with a lot of user defined parameters makes study comparisons difficult

because all state use of the same tool but not the same parameters. Moreover, the parameters set for one TIS library may not be appropriate for analysis of a different library.

The Segmentor3 package is outdated and has been removed from the CRAN repository, while the code remains available from GitHub. However, it is no longer being updated (most recently in 2016) and its use within R (R Core Team, 2022) is limited to older versions of R as the software develops. Using R enabled visualisation using the familiar plotting package ggplot2 (Wickham, 2016). However, R is not the best program for a full analysis tool and was not built for such applications; it is relatively slow compared to more sophisticated programming languages such as Java or C/C++ (Fourment & Gillings, 2008; Johnson & Chandran, 2021). The changepoint detection package used in this work was not developed to analyse data as large as a whole genome sequence. The maximum size that Segmentor3 could reliably analyse was 100,000 bp which would take 47 runs to complete the whole genome; the program is slow and demanding on the computer. An up-to-date application of the Segmentor3IsBack algorithm would be more appropriate for a TIS modelling package. The R package OnlineBcp was able to model the whole genome but appears to be too sensitive and does not offer user defined parameters such as a maximum number of breakpoints to make it more applicable to TIS.

The use of machine learning could be considered an extension to this work to further improve protected region determination, and machine learning approaches have been used for this purpose (Gale et al., 2020; Zhu et al., 2018). An algorithm could be trained with datasets and this approach may be useful for applications involving analysis of many libraries in similar organisms. Relatively few people want to look at TIS at that scale so it is more useful to have something that is accessible and can be tailored to individual use.

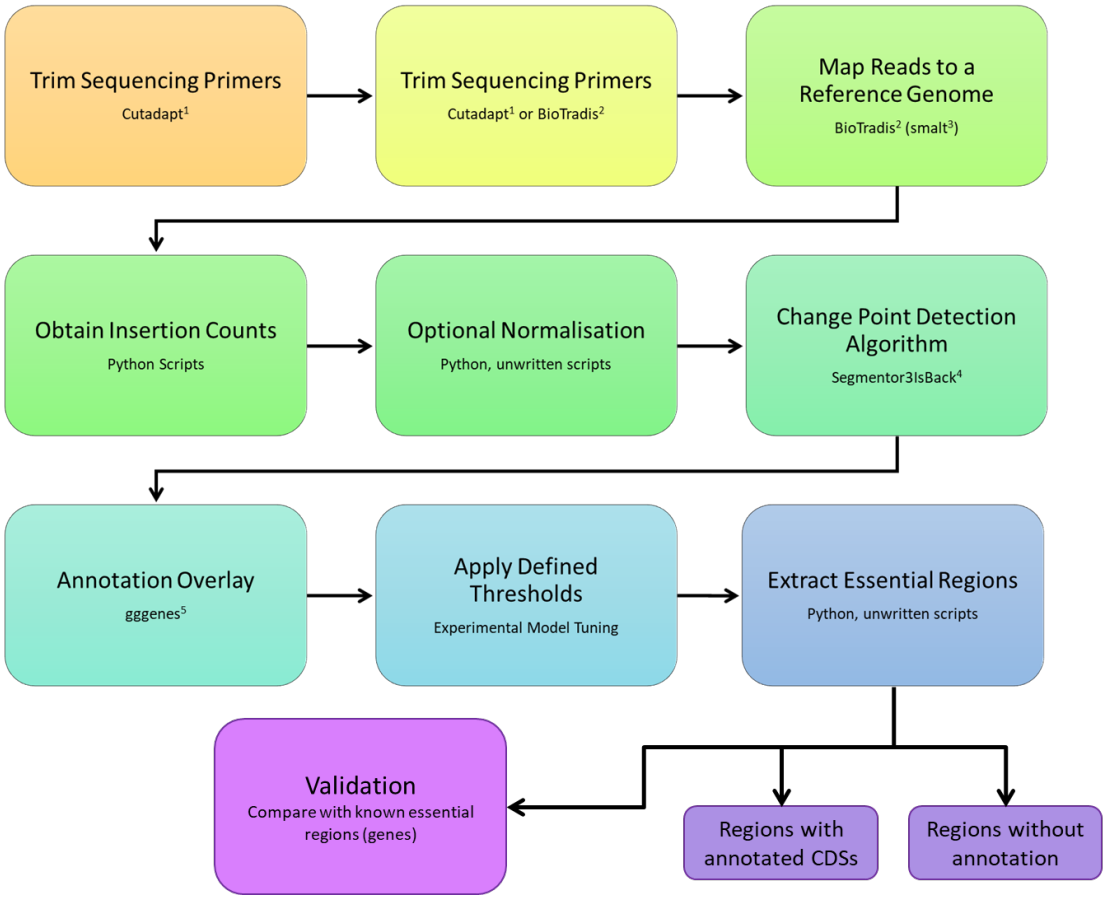

Figure 5-7 *Proposed TIS analysis pipeline*.

The pipeline of analysis for TIS using the software and tools discussed in this chapter. [1](Martin, 2011), [2](Barquist et al., 2016), [3]*(Ponsting & Ning, 2012),* [4](Cleynen et al., 2014), [5](Wilkins, 2020).

5.5 Developments Required for an Analysis Tool

The segmentation or changepoint detection approach presented here identified protected regions of the genome that corresponded with known essential genes. The models could detect changes in mean at a high enough resolution to predict protected regions within genes. At the current state, *Figure 5-7*, the segmentation model described in this work would need a number of developments to generate a reliable, accessible tool that could analyse TIS data. These are as follows:

- Use of the underlying statistics and implementation in a more sophisticated programming language to improve the speed of analysis.
- Have validated thresholds to call a region of the genome beneficial or detrimental to fitness and have these listed as outputs.
- Have annotation independent analysis but with the option to overlay an annotation, this enables the same model to be observed as improvements are made in annotation.
- Be combined with some of the python scripting from this work to enable a full workflow from sequence data to model.
- Have optional normalisation functions to enable normalisation to sequencing biases, transposon insertion biases and biases introduced by growth, as described in *chapter 4.*
- Maintain visualisation of the model as an output.

5.6 Conclusions

Annotation independent analysis is more informative than only focussing on the essentiality of an annotated gene. Non-coding genomic regions can be functional and contribute to the survival and growth of an organism and may provide alternative, unconventional drug targets. Transposon insertion patterns can be used to determine the physical configuration of the chromosome, this information is lost when only considering CDS. Changepoint detection is consistent even if annotation changes, and annotation can be overlaid provided that the reference genome sequence is unchanged.

6 Development of Exponential Mutagenesis

6.1 Introduction

One of the limitations of TIS is that there is only one transposon per cell. This makes analysis more straightforward as there is a direct association between insertion event and phenotype but means that any redundancy in the system is not seen by disrupting only one part of the genomic network. Generating double knockouts in all combinations would mean creating a library of mutants around each single mutant. This increases the number of insertion events exponentially and requires improved analysis methods.

6.1.1 Metabolic Networks

In reality, genes rarely operate in isolation and gene products interact in networks. The interactions of these genes can be mapped after extensive research into networks of biochemical activity (Gillis & Pavlidis, 2011). The notion is that the fewer the reactions that a gene product is involved in, the more dispensable the gene is. However, within a cell there are linked and unlinked processes (van Opijnen & Camilli, 2013). Eliminating genes in this network can often have no phenotypic consequences due to the network compensating for the lost pathway. This is where the concept of genomic redundancy arises.

6.1.2 Definitions of Redundancy

While there is evidence of redundant or interchangeable genes, there are degrees to which redundancy is observed. Within a genome or network, types of redundancy can be classified as:

1) Functional Redundancy – two or more gene products can perform the same or very similar functional roles within an organism, for example in UPEC strains, *iucB* and *entD* are both involved in iron acquisition but act via different siderophores (Garcia et al., 2011).

2) Target Redundancy – two or more genes are responsible for producing the same molecule, for example *ileS* and *glyS* are both tRNA synthases, only one is required for survival (Baba et al., 2006).

3) Alternative route – if a biochemical pathway is interrupted then the metabolic flux is rerouted through alternative reactions to achieve the same biochemical outcome, for

example the ability of electron transport to occur through differing coenzymes (Goldford et al., 2022).

Using TIS to identify functional and target redundancies can identify potential new drug targets, whereas understanding alternative biochemical pathways can help to determine resistance evolution to a novel compound.

6.1.3 Synthetic Lethality and computational approaches

A major development in synthetic biology is the generation of genome scale computational metabolic models of an organism. Models use defined chemical reactions encoded within a genome and allow parameters to be changed to mimic stresses; allowing the flux of energy through the cell to be modelled (Grimbs et al., 2019). Within this, reactions and pathways can be removed or altered to mimic genetic mutations or knockouts; multiple genes can be removed and the impact on cell growth can be observed *in silico* (Aziz et al., 2015). There are two levels for this phenomenon, one for detailing when there is significant detriment to growth, synthetic sick and one where removal of both genes leads to cell death, synthetic lethality. Though less studied, there are examples of interactions where a second deletion can cause overcompensation of the first and lead to increased mutant fitness (Côté et al., 2016).

Once a model has been curated, endless hypothetical experiments can be run relatively quickly and without much additional cost. Therefore, a lot of gene interaction or redundancy data is available from these metabolic models. However, these models rely on accurate genome sequencing and annotation of gene function. Furthermore, assumptions and predictions used to curate the models are validated by laboratory data. So, while in silico models can provide predictive interaction or redundant pairs of genes, any hits would require further validation.

6.1.4 Low throughput approaches into EM

There has been work undertaken to identify gene interactions and redundancies, but these have been approached in a low or medium throughput manner where specific genes are targeted (Garcia et al., 2011) or where a handful of genes is tested against a pool of mutants (van Opijnen & Camilli, 2013).

When *E. coli* is grown in nutrient-limited media, more than 100 genes become essential - those required for the synthesis of amino acids and vitamins (Côté et al., 2016; Joyce et al., 2006). If more genes become essential during nutrient stress, such as infection, these could provide antimicrobial targets. Côté et al. performed a medium throughput screen of double mutants by mating each mutant from the Keio collection with their own targeted gene deletions. Across 315,400 double deletion mutants, 1,881 synthetic sick or lethal interactions were identified under nutrient stress conditions. All the tryptophan biosynthesis genes formed a beneficial interaction with the uncharacterized gene *yhdU* (Côté et al., 2016).

The aim of this chapter is to describe the work undertaken to investigate the possibility of producing a TIS library with every combination of two genes are knocked out in the mutant pool. Using a "piggyback" approach, with nested transposons, each initial insertion would result in a library of second insertions generating in the region of 20,000,000 double mutants in the *E. coli* BW25113 genome.

6.2 Methods for this Chapter

6.2.1 Building the pExM Plasmid

The Tn*5* mutagenesis plasmid pBAMD1-6 (Martínez-García, Aparicio, Lorenzo, et al., 2014) was chosen as the vector to build the EM construct. The approach was to clone the *himar1C9* transposase and the mariner transposon from pSAM_Ec (Wiles et al., 2013) into the pBAMD1-6 vector to create pExM.

6.2.1.1 Designing the EM Transposon

The EM transposon was designed to be composite with a mariner transposase and a mariner transposon nested within a Tn*5* transposon. This design assumes chromosomal repair of the Tn*5* transposon following the mariner transposition.

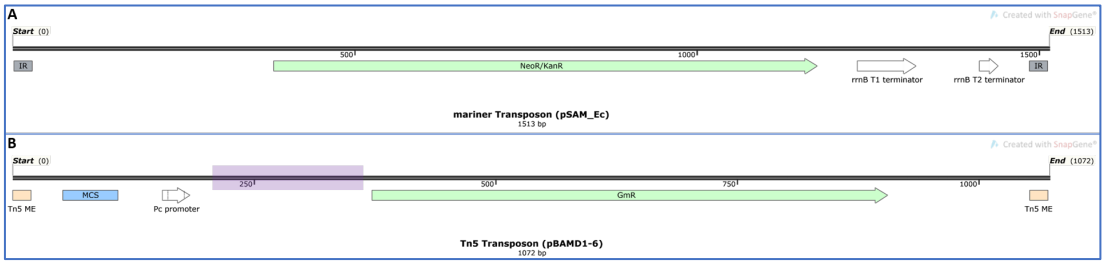

Figure 6-1 *The Design of the Exponential Mutagenesis Composite Transposon.*

A: the mariner transposon from pSAM_Ec. B: the proposed insertion site for the mariner transposase within the MCS (blue) and the himar1C9 mariner transpososase in the region within the purple box.

The transposase was cloned into the multiple cloning site (MCS), blue box in *Figure 6-1,* upstream of the gentamicin resistance cassette. The transposon was cloned into the exogenous region between the gentamicin resistance gene and its constitutive Pc promoter. The intended composite transposon is shown in *Figure 6-2.*

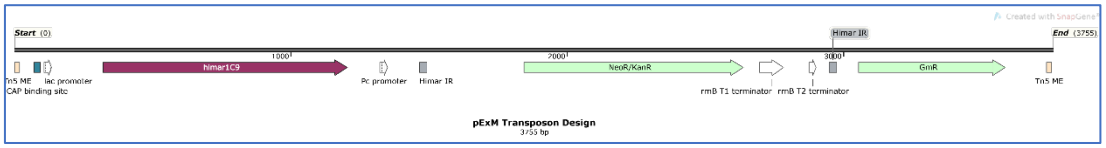

Figure 6-2 *The Intended Composite Transposn Arrangement.*

A genetic map of the designed pExM transposible element containing the Himar1C9 transposase, mariner transposible element and Tn5 MEs.

The Tn*5* transposon contained the himar1 inverted repeats around the kanamycin resistance gene and transcriptional terminators; the same mariner transposon was used as for the work in previous chapters. Outside of the *himar1* inverted repeats there was a second resistance gene, an acetyl transferase conferring resistance to gentamicin and its native Pc promoter as in the plasmid pBAMD1-6. The aim was to clone the mariner transposon into the Bsp1407I restriction endonuclease site between the gentamicin resistance cassette and its Pc promoter, the rationale (and assumption made) was that the distance between the promoter and acetyltransferase gene (1.5kb) prevented transcription; readthrough transcription was prevented by the transcriptional terminators

at the 3' terminus of the mariner transposon. Following the second transposition and repair of the excision site, the Pc promoter was within range to express the acetyl transferase gene to confer resistance to gentamicin. This enabled selection of successful exponential (double) mutants using gentamicin. An intergenic region was chosen, rather than to disrupt the gene itself, to mitigate any effect of imperfect repair followed excision.

Outside of the mariner transposon, the Himar1C9 mariner transposase, with an inducible *lac* promoter, was cloned using the KpnI and SalI restriction endonuclease sites in MCS of pBAMD1-6 upstream from the gentamicin acetyltransferase cassette. The plasmid was digested in a double reaction following the manufacturer's (NEB, Ipswich) double digest protocol using 1 µL of KpnI exonuclease, 1 µL of SalI exonuclease, 5 µL of Cutsmart 10x buffer and 43 µL of plasmid DNA at 25 ng/ µL. The reaction was incubated at 37 ˚C for 30 minutes. The reaction was then cleaned by 1.5 x SPRI bead purification *section 2.3.3.*

6.2.1.2 Cloning – Mariner Transposase

The mariner transposase (insert) was PCR amplified; pBAMD1-6 (vector) was digested and then the mariner transposase was ligated into the vector.

6.2.1.2.1 Transposase Amplification

The insert (transposase) was amplified from the plasmid pSAM_Ec (Wiles et al., 2013) using PCR and the primers detailed in *Table 6-1.* The reaction was set up in triplicate as follows: 25 µL of NEB Q5 2x Mastermix (NEB), 2.5 µL of 10 mM CAP_F primer (Eurofins Genomics, Ebersberg), 2.5 µL of 10 mM Mar_Frag_R primer (Eurofins Genomics), 1ul of 1 ng/µL pSAM_Ec template and 19 µL of nuclease free water (Invitrogen, Waltham). The reactions were mixed well with a pipette and run in a thermocycler with the heated lid option, the cycle conditions are listed in *Table 6-2.*

***Table 6-1** Sequences of the PCR primers used to amplify the mariner transposase from pSAM_Ec.*

Name	Sequence	Manufacturer
CAP_F	TAATGTGAGTTAGCTCACTCAT	Eurofins Genomics, Ebersberg
Mar_Frag_R	TGGGAATTCCCCTCCACCGCG	Eurofins Genomics, Ebersberg

Table 6-2 *The PCR cycling conditions used to amplify the mariner transposase from pSAM_Ec.*

Number of Cycles	Temperature	Time
1	98 ˚C	30 s
30	98 ˚C	10 s
	61 ˚C	30 s
	72 ˚C	45 s
1	72 ˚C	2 min
1	10 ˚C	Hold

The PCR products were pooled and cleaned using a 1:1 ratio of SPRI beads, this protocol is fully described in *section 2.3.3.*

6.2.1.2.2 Insert Phosphorylation

Phosphorylation of the amplified transposase was performed as described in *section 2.3.6*. Briefly polynucleotide kinase (NEB) was used to add a phosphate group to the 5' ends of the DNA fragment. No clean-up was performed.

6.2.1.2.3 pBAMD1-6 Digestion and Dephosphorylation

The vector pBAMD1-6 was digested using the restriction endonuclease SmaI and its supplied buffer Cutsmart (NEB). Dephosphorylation was performed at the same time using FastAP (Themo Scientific, Waltham). The reaction was set up with 1 µg of purified plasmid, 3 µL of 10x Cutsmart buffer, 1 µL (20 U) of SmaI, 1 µL (1 U) of FastAP and the volume adjusted to 30 µL using nuclease free water (Invitrogen). The reaction was incubated at 37 ˚C for 15 minutes and then cleaned using a 1:1 ratio of SPRI beads as detailed in *section 2.3.3.*

6.2.1.2.4 Ligation and Transformation

The insert and vector were quantified using the Qubit® fluorometer, *section 2.6.5*, then combined at a 3:1 molar ratio of transposase to vector. The ratio was calculated using NEBioCalculator version 1.15.1 using the DNA Ligation option and a vector mass of 50 ng.

Once combined, 1 µL (400 U) of T4 Ligase (NEB), 2 µL of 10 T4 ligase buffer and nuclease free water to 20 µL were added and mixed well with a pipette. The reaction was incubated at room temperature for 2 hours and then 65 ˚C for 10 minutes. Then 5 µL of the ligation was transformed, as per *section 2.4.2*, into chemically competent *E. coli* MFD*pir+* cells*, section 2.4.1.* Successful transformants were recovered on LB agar supplemented with 10 µg/mL gentamicin sulphate (Formedium, Swaffham).

6.2.1.3 Cloning – Mariner Transposon

The mariner transposon (insert) was synthesised by Genewiz, Leipzig. The pBAMD1-6-Tnp vector was extracted from an overnight culture of a successful transformant described above. Extraction was performed as per *section 2.6.3.*

6.2.1.3.1 Transposon Preparation

The synthesised DNA was designed to have a restriction site for the Bsp1407I enzyme flanking the IRs, to allow for cohesive ligation into the vector. The synthesised gene is shown in *Figure 6-1.*

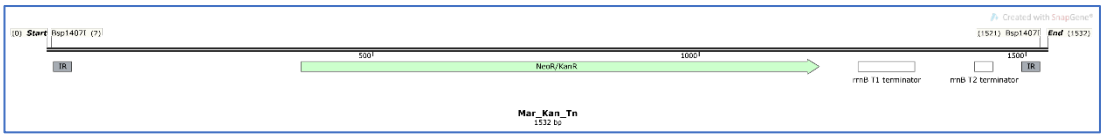

Figure 6-3 *Map of the synthesised mariner transposon.*

Genetic map showing the organisation of the mariner transpososn synthesised by GeneWiz.

The synthesised transposon was isolated from the supplied vector by digestion with Bsp1407I (Thermo Scientific). The digestion rection contained 1 µg of plasmid DNA, 1 µL (10 U) of Bsp1407I, 3 µL of supplied 10 x Tango buffer and was adjusted to 30 µL with nuclease free water. The reaction was incubated at 37 ˚C for one hour. Then the ~1.5 Kb product was selected for using agarose gel size selection, *section 2.3.5*.

6.2.1.3.2 pBAMD1-6-Tnp Digestion and Dephosphorylation

The vector pBAMD1-6-Tnp was digested using Bsp1407I (Thermo Fisher Scientific). The digestion rection contained 1 µg of plasmid DNA, 1 µL (10 U) of Bsp1407I, 3 µL of supplied 10 x Tango buffer, 1 µL (1 U) of FastAP and was adjusted to 30 µL with nuclease free water. Dephosphorylation was performed at the same time using FastAP (Themo Fisher Scientific). The reaction was incubated at 37 ˚C for 15 minutes and then cleaned using a 1:1 ratio of SPRI beads as detailed in *section 2.3.3.*

6.2.1.3.3 Ligation and Transformation

The insert and vector were quantified using the Qubit® fluorometer, *section 2.6.5,* then combined at a 3:1 molar ratio of transposase to vector. The ratio was calculated using NEBioCalculator version 1.15.1 using the DNA Ligation option and a vector mass of 50 ng. Once combined, 1 µL (400 U) of T4 Ligase (NEB), 2 µL of 10x T4 ligase buffer and nuclease free water to 20 µL were added and mixed well with a pipette. The reaction was incubated at room temperature for 2 hours and then 65 ˚C for 10 minutes. Then 5 µL of the ligation was transformed, as per *section 2.4.2*, into chemically competent *E. coli* MFD*pir+* cells*, section 2.4.1.* Successful transformants were recovered on LB agar supplemented with 50 µg/mL kanamycin sulphate.

6.2.1.4 Plasmid Sequencing

The pExM plasmid was extracted from *E. coli* MFD*pir+:* pExM donor cells using the Macherey Nagel NucleoSpin kit (Macherey Nagel, Düren) and following manufacturer's instructions. The plasmid DNA was diluted to 5 ng/µL with nuclease free water (Invitrogen) and submitted to the Quadram sequencing service. Plasmid DNA was prepared for sequencing using a minimised version of the Illumina DNA Sequencing protocol. The prepared library was run as a low percentage spike into a paired-end 150 cycle Illumina NextSeq run (Illumina, San Diego).

6.2.1.5 Plasmid Sequence Assembly

The plasmid sequence was assembled using Unicycler version 0.4.8.0 in Galaxy. The assemblies were visualised in Snapgene Viewer version 6.0.4.

6.2.2 Exponential Mutant generation

Two methods of transformation were explored for generating mutants; conjugation (as had been successfully used in previous chapters) and electroporation of *in vitro* prepared transposomes.

6.2.2.1 Conjugation

Conjugation was performed as described in *section 2.5.1.* Briefly, overnight, stationary cultures of both donor *E. coli MFDpir+:* pExM and recipient *E. coli* BW25113 cells were concentrated 20x and mixed in equal volumes. Then 10 µL alliquotes of this mixture were repeatedly spotted onto LB agar plates containing 0.3% (w/v) glucose. The plates were incubated at 37˚C for five hours to allow conjugation and transposition but to restrict growth.

6.2.2.2 Transposome Generation

Transposomes were generated using the EZ: Tn*5* transposome (LGC Biosearch Technologies, Petaluma, USA, formerly Lucigen) as described in *section 2.5.2.* The concentration of purified PCR product was increased from 100 ng/µL to 400 ng/µL to account for the increased transposon size and maintain molar ratios of enzyme to DNA ends. The transposome reaction contained 2 µL of purified and concentrated PCR product from the same day, 2 µL of pure glycerol and 4 µL of purified transposase from the kit. The reaction was prepared on ice and mixed well, then incubated at room temperature for 45 minutes. The reaction was then stored at -20 ˚C until use.

6.2.2.3 Electroporation of Transposomes

Electroporation was performed as described in *section 2.5.3*. Electrocompetent *E. coli* BW25113 were prepared from exponentially growing cells by washing with ice cold glycerol and concentrating. Each 60 µL aliquot was combined with 0.4 µL of prepared transposome in a 2 mL electroporation cuvette and incubated on ice for 30 minutes. The cells were transformed using a single pulse of 2500 V, 25 µF and 200 Ω. The cells were recovered with prewarmed SOC for 90 minutes and then plated onto LB agar supplemented with 50 µg/mL kanamycin sulphate and 0.3% w/v glucose and grown at 37 ˚C overnight.

6.2.2.4 Second Transposition

The second transposition, mariner, occurred without induction despite the transposase being under control of a *lac* promoter. As discussed in *chapter 3,* the promoter is leaky, so constitutive expression occurs. Rather than inducing the second transposition event by activating the promoter, the promoter was repressed using glucose catabolite suppression (Griffiths et al., 2000)during conjugation and outgrowth of the first transposition mutants. Successful second transposon mutants were selected for by spreading onto LB agar, no glucose, and supplemented with 8 µg/mL gentamicin sulphate and 50 µg/mL kanamycin sulphate.

6.2.3 Sequencing EM Mutants

6.2.4.1 Nanopore Sequencing

Individual mutants were isolated from the selective LB agar plate (section 6.1.1.4); each mutant was grown overnight in LB broth supplemented with 8 µg/mL gentamicin sulphate and 50 µg/mL kanamycin sulphate at 37 ˚C. Then HMW DNA extraction was performed using the FireMonkey kit (Revolugen, Glossop, UK) following the manufacturer's instructions, as described in *section 2.6.2.*

The HMW DNA was prepared for nanopore sequencing using a minimised method developed by Dr. Emma Waters and then followed the Manufacturer's instructions for adapter ligation and flow cell loading. The details of the protocol are described in section 2.2.10.1. The prepared DNA was sequenced on a MinION flow cell version R9.4.1 and controlled using MinKNOW. The raw data was base called using Guppy. The instrument, flow cell and software are all produced by Oxford Nanopore Technologies, Oxford.

6.2.4.2 Illumina Sequencing

For each mutant pool, DNA was extracted using the RSC Maxwell instrument as per *section 2.6.1*. Both mariner and Tn*5* transposon directed sequencing was performed on the extracted DNA from each pool using 50 ng for each preparation. Sequencing library preparation was prepared following a modified version of the Illumina standard DNA whole genome sequencing protocol (Illumina, 2022) full details of each step is described in *section 2.8*.

6.2.3.1.1 Concatenating Fastq Files

The raw sequence `.fastq` files from multiple sequencing runs using the Illumina NexSeq were concatenated per transposon; fewer EM mutants were generated so this provided one sequence pool to enable analysis of gene essentiality.

6.2.3.1.2 Adapter Trimming, Genome Mapping and Essentiality Determination

The rest of sequence file processing was performed in the same way as described in *sections 3.2.2.4-3.2.2.8.* The BioTraDIS (Barquist et al., 2016; Langridge et al., 2009) pipeline was used to map transposon insertions to the genome and to determine gene essentiality.

6.3 Results

6.3.1 Building the pExM Plasmid

Sequencing and assembly of the pExM plasmid generated 2 contigs, one contained the sequence of the pBAMD1-6-Tnp plasmid (6061 bp). The SmaI recognition site within the MCS had been interrupted, this interruption showed that an insert had been cloned into this site. The second contig of size 1168 bp was the PCR amplified *himar*1C9 with a *lac* promoter and CAP binding site. This showed that the cloning to generate the composite transposon was successful. The organisation of the sequenced composite transposon is shown in *Figure 6-4.*

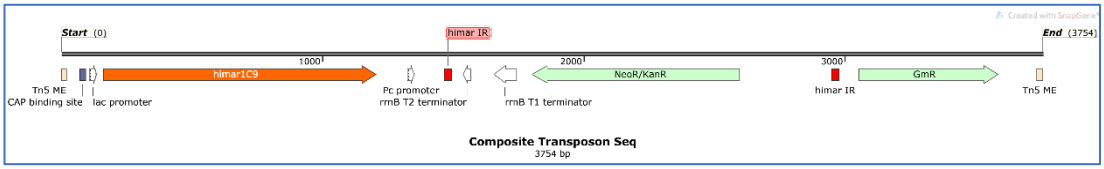

Figure 6-4 *Sequence assembly of the pExM composite transposon.*

Assembly of 150 bp paired end Illumina sequences, assembled using Unicycler and visualised in SnapGene Viewer. The himar1C9 transposase is shown as an orange aroow and the himar IRs are red boxes.

The sequence showed that as a result of cloning the mariner transposon as a blunt fragment, it had ligated in the opposite direction to the mariner transposase. There is no evidence that this affects transposition but is a point to note. The fully assembled pExM plasmid can be seen in *Figure 6-5B*. The intermediate pBAMD1-6-Tnp plasmid was also sequenced before constructing the pExM vector; the assembly and annotation of pBAMD1-6-Tnp can be seen in *Figure 6-5A.*

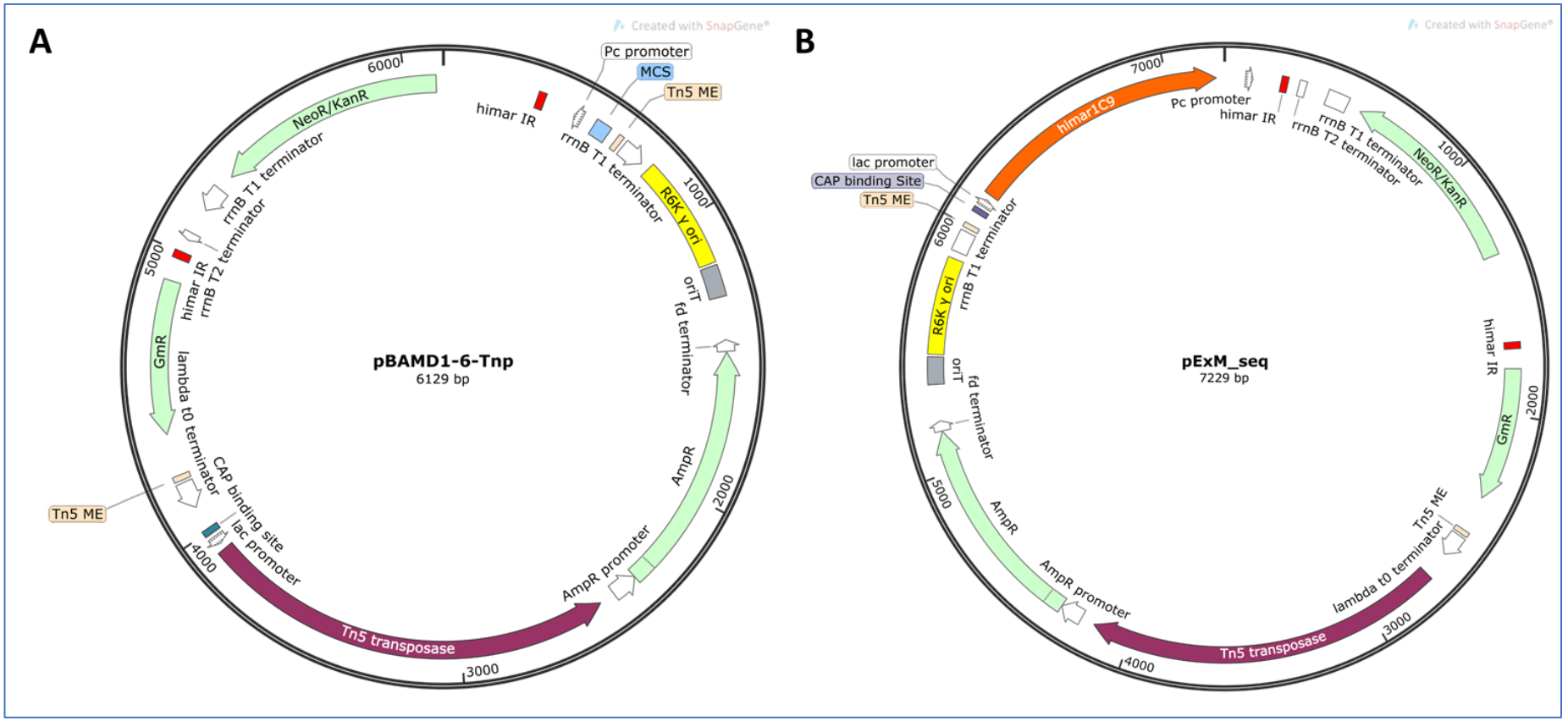

Figure 6-5 *Plasmid maps showing the assembled sequences of the plasmids cloned in this work.*

A: pBAMD1-6 showing the mariner transpoosn cloned into the pBAMD1-6 vector with the himar IRs shown as red boxes. B: pExM showing the himar1C9 mariner transposase cloned into the pBAMD1-6-Tnp vector (A). Both were sequenced using Illumina DNA Prep protocols, 150 bp paired end sequencing. The sequences were assembled using Unicycler and visualised in SnapGene Viewer

6.3.2 Exponential Mutant Generation

6.3.2.1 Conjugative Approach

Conjugation of pExM was successful using the protocols developed for TIS and described in *chapter 3*, but at a lower efficiency. This was to be expected due to the increased size of the conjugative plasmid. The transposition efficiencies are summarised in *Table 6-3*.

***Table 6-3** Transposon efficiencies calculated for EM*

	Tn5 Mutants	***Tn5 Rate***	***Mariner Mutants***	***Mariner Rate***	***Plasmid Retention (%)***	***Doublings***
LB	3.94E+05	1.19E-04	0.00E+00	0	1.50	1.18
Glucose	1.80E+05	7.80E-05	0.00E+00	0	70.00	0.66

The efficiencies for each transposon were calculated using colony counts on selective agar. Two conjugation media were used, LB agar and LB agar with 0.03% (w/v) glucose. Rates were calculated for both media.

6.3.2.2 Transposome

Electroporation of the EM transposome (*section 6*) produced an estimated 3400 kanamycin resistant colonies. This indicated 3400 successful Tn*5* transformants to be screened for mariner transposase induction methods. Since the aim was to generate 10,000 Tn*5* mutants for EM, this method was not pursued, and conjugation was chosen as the preferred method of mutant generation.

6.3.2.3 Induction of the Second Transposition

Control of the second transposition was key to ensure there was enough representation of each individual single Tn*5* knockout mutant to generate a comprehensive double mutant library. Using glucose as a catabolite repressor (Santillán & Mackey, 2004) was effective at supressing the second transposition, demonstrated by a reduced second transposition rate for glucose seen in *Figure 6-6A*. Induction of the second transposition event was less reliable. Using IPTG did not offer any substantial benefit. *Figure 6-6A:D* shows the different media components that were tested for conjugation (*Figure 6-6A*), selection (*Figure 6-6B*) and induction (*Figure 6-6C*) when generating the EM mutants. The aim was to achieve the highest second transposition rate while minimising plasmid retention (*Figure 6-6D*)

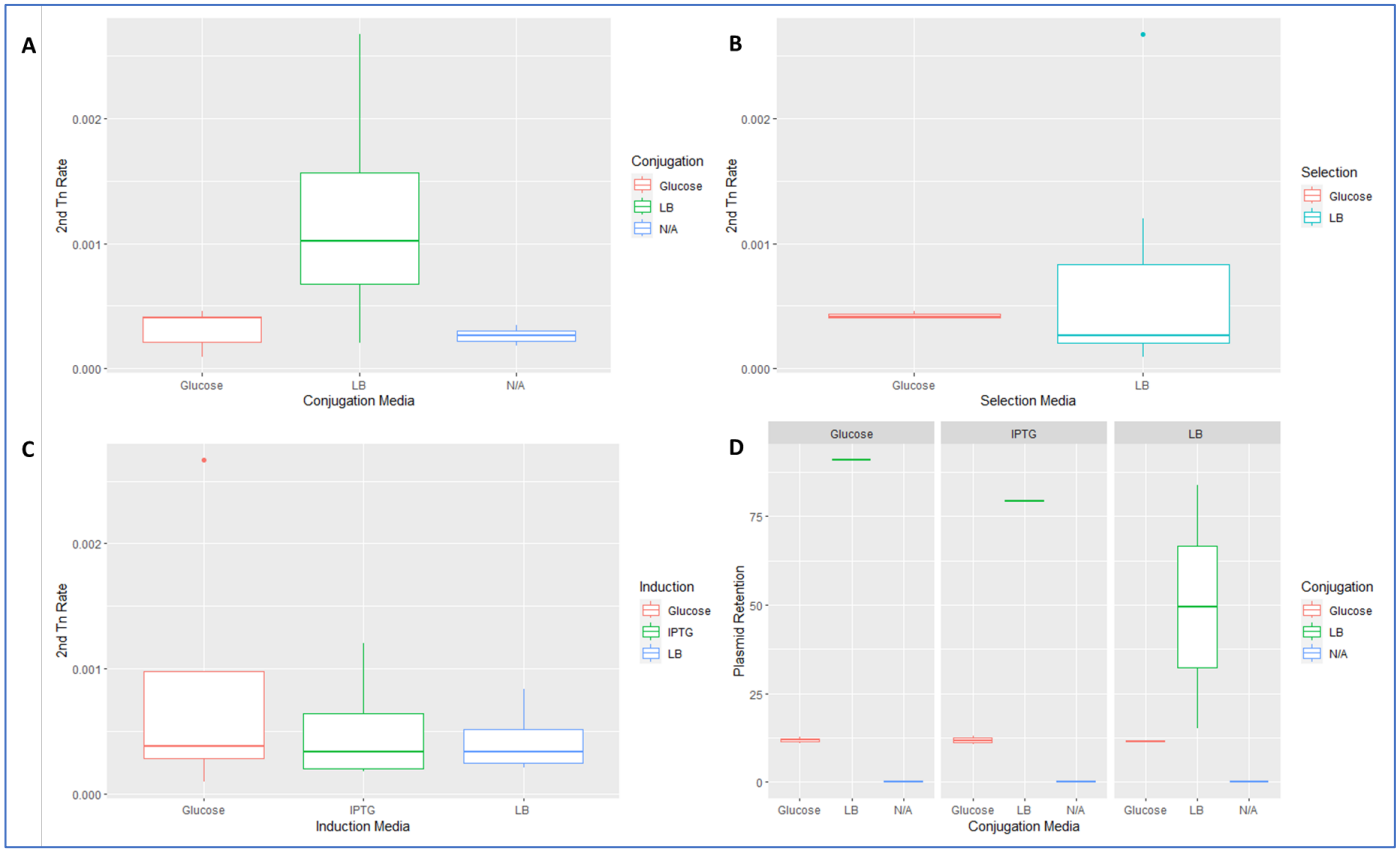

Figure 6-6 *The effect of media choice on mariner transposition rate throughout mutant generation.*

A: The difference between conjugation on LB or LB with 0.03% (w/v) glucose on second transposition rate, N/A refers to mutants generated using transposomes. B: The difference between using the same two media with 50 µg/mL kanamycin sulphate for first mutant selection. C: The difference between using the same media or Lb with 1mM IPTG during the induction of the second transposition event. D: The effect of media choice on the percent of plasmid retention, split by Induction medium.

The boxplots in *Figure 6-6* suggested that for the maximum second transposition rate and the least plasmid retention, LB should be used as the medium throughout the procedure. Using LB gave the most variable results, removing glucose during conjugation enabled the second transposition to occur at any time, potentially independent from the first transposition event. Therefore, all further EM mutant generation was performed by adding glucose to the conjugation medium and LB for the remainder of the process. Induction took place over five hours at 37 ˚C to keep growth to a minimum but allow sufficient time for transposition.

6.3.3 Sequencing EM Mutants

The EM mutants were analysed as individual TIS libraries because identifying two transposons inserted into the same chromosome was not possible at this time. To check that there really were two transposons in each chromosome long read sequencing was used.

6.3.3.1 Nanopore Assemblies

The long read sequence data assemblies were comprised of one or two contigs for each mutant chosen. A summary of the assemblies; checkm (Parks et al., 2015) completeness, contamination, and the number of contigs is summarised in *Table 6-1.*

Table 6-4 *Genome assembly quality metrics for single EM mutant colonies.*

Mutant	***Assembly Completeness (%)***	***Sequence Contamination (%)***	***Contigs***
BW25113	98.20	0.04	1
2	98.52	0.04	1
3	98.13	0.04	2
4	98.54	0.06	2
5	98.59	0.08	1
6	98.00	0.07	1
7	98.13	0.04	1
8	98.72	0.07	1
9	98.48	0.06	1

The assembly for *E. coli* BW25113 was comparable with all the mutants, this indicated that the transposon insertions were not detrimental to long read genome assembly. *Figure 6-7A* shows the consensus sequence for BW25113 was 4,631,773 bp long so 146 bp shorter than the hybrid reference genome (*section 3.3.1*). Importantly, there were no instances of the Tn*5* ME or the mariner IR sequences within the WGS assembly, therefore these could be used to identify the composite and independent mariner transposon within a mutant genome assembly.

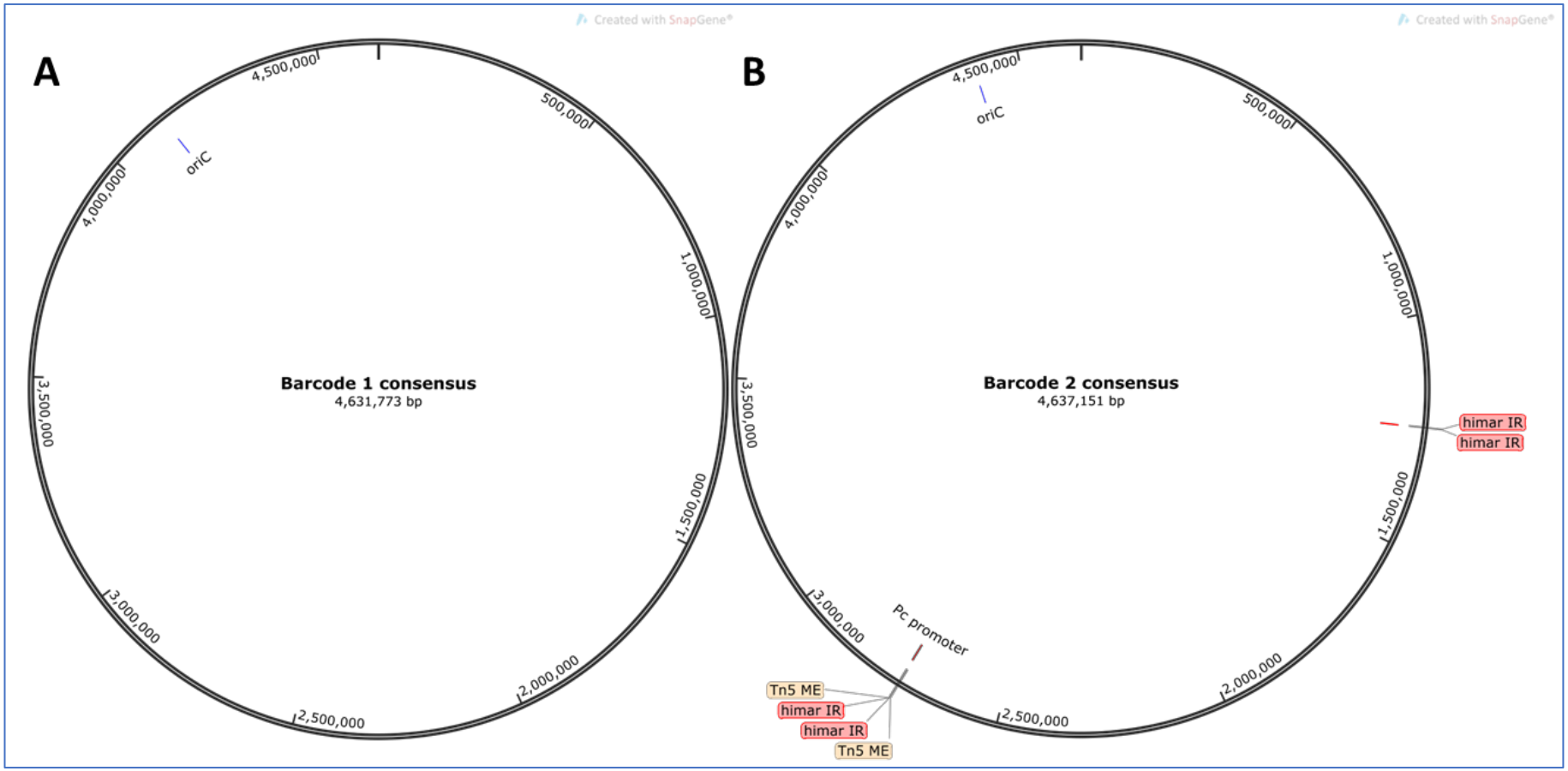

Figure 6-7 *Consensus long read assembled sequences.*

A: for BW25113 and B: mutant number 2

The consensus sequence from mutant number two, *Figure 6-7B* showed that both Tn*5* and mariner transposition had occurred and indicated that the method had generated double transposon mutants. However, the composite transposon was still intact meaning that the mariner transposon had not been excised and moved. Mariner family transposons move by a cut and paste mechanism (Lampe et al., 1999) so this should not have been the case. It is possible that within the mutant colony and subsequent growth for DNA extraction, the mariner transposon had been excised in come cells but not others, and that the consensus sequence showed both states.

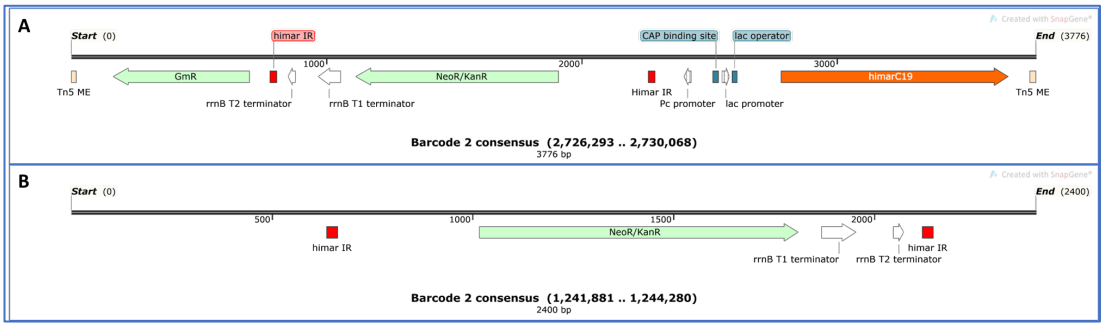

Figure 6-8 *The consensus sequences for mutant number two showing the two separate transposon insertion events identified.*

A: From 2,726,293 – 2,730, 068 bp showing the full composite EM transposon intact. B: From 1,241,881 – 1,244,280 bp showing only the mariner transposon and the flanking regions to show an independent mariner insertion.

Mutants were selected using gentamicin, it is conceivable that mutants where the second transposition had occurred were able to neutralise the antimicrobial, enabling survival of all cells. *Figure 6-8A* shows the Tn*5* transposon from the consensus sequence of barcode two, it was the intact composite transposon. *Figure 6-8B* shows the consensus sequence of the independent mariner transposon and the region upstream, absence of the Pc promoter demonstrated that the transposon was independent from Tn*5*.

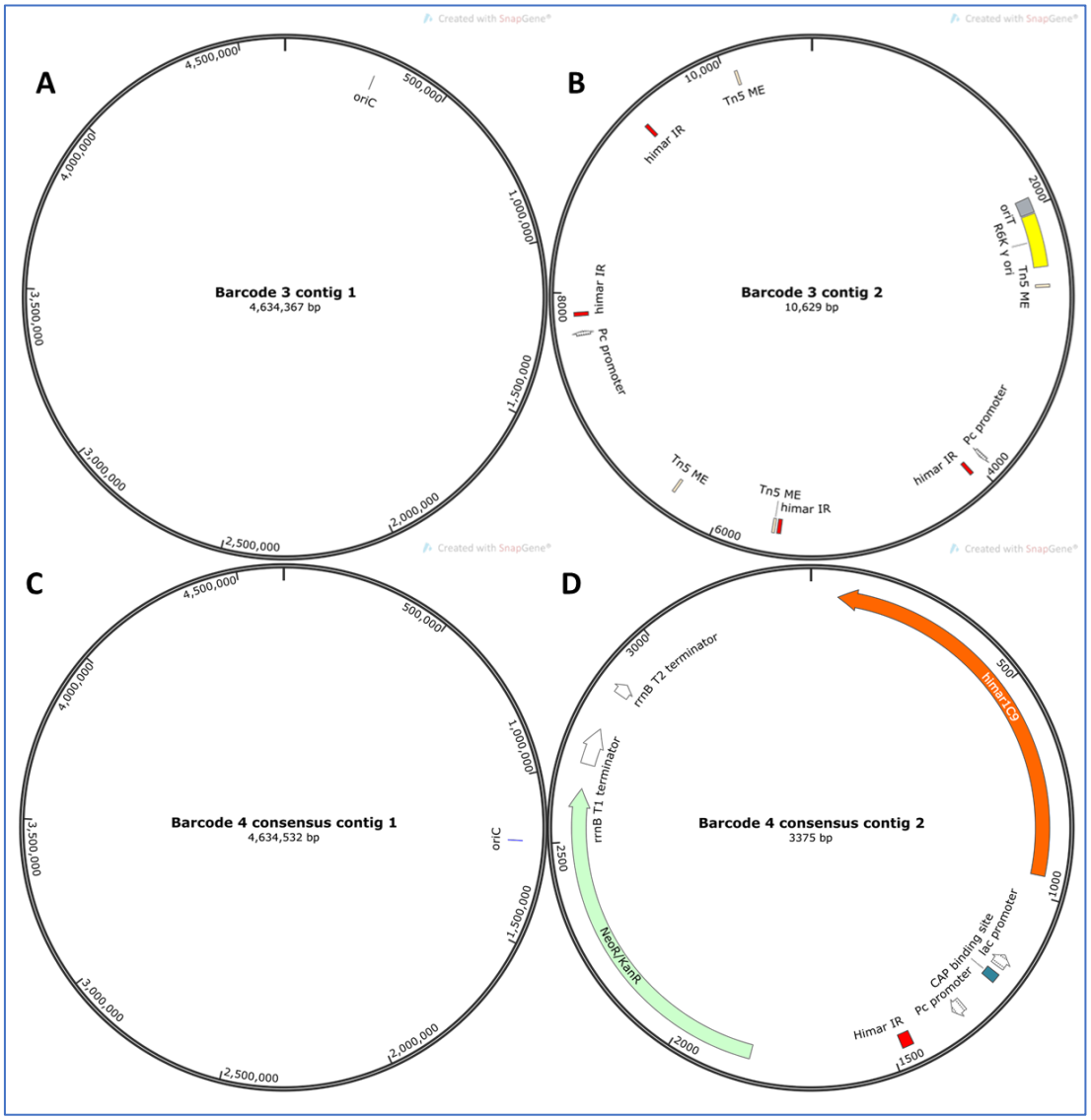

Figure 6-9 *Consensus long read assembled sequences.*

A-B: Mutant number 3 as two contigs showing elements from the pExM plasmid so not an EM mutant. C-D Mutant number 4 as two contigs showing the composite transposon in contig 2(D) but no evidence of a second transposition event.

Mutant numbers three and four did not show the same, for both, contig 1 *Figure 6-9A and C*, had no evidence of transposons being present. For mutant number three contig 2 *Figure 6-9B,* had the Tn5 MEs and mariner IRs but also the R6k origin of replication, this was from the plasmid and so mutant three was unlikely to be either a single or a double transposon mutant. Contig 2 for mutant number four, *Figure 6-9D,* showed the mariner transposon and elements of the Tn5 transposon but not the gentamicin acetyltransferase gene or a Tn5 ME. Again, this mutant is unlikely to be a double transposon mutant.

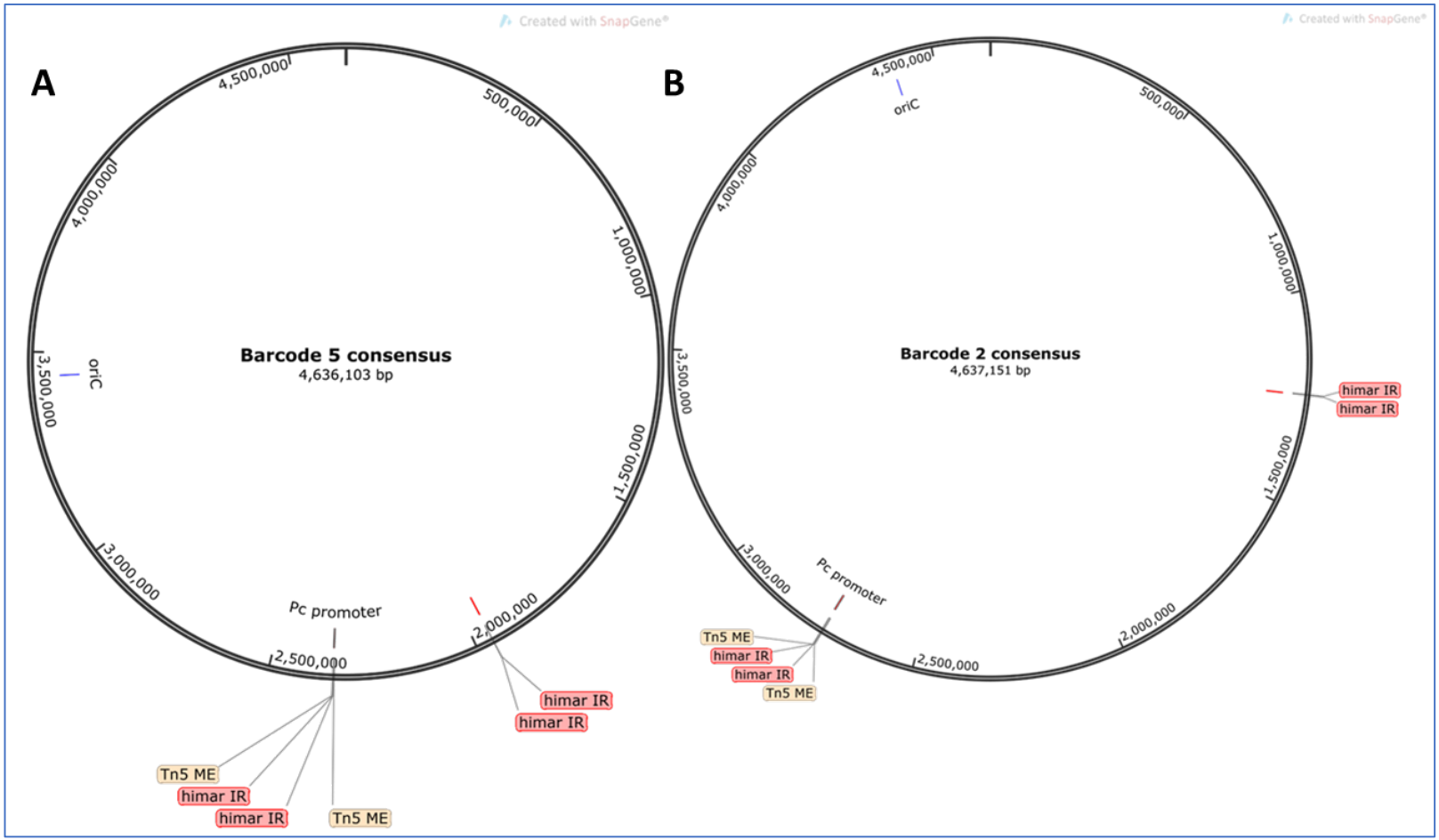

Figure 6-10 *Consensus long read assembled sequences.*

A: for mutant number 5 and B: mutant number 2

As with mutant number two, mutant number five was a potential double transposon mutant. The consensus assembly, *Figure 6-10A,* showed an independent mariner transposon. Again, the Tn5 transposon contained the composite mariner elements, and was likely due to a mixed colony of mutants as described previously. Comparison to mutant two, *Figure 6-10B,* showed that this was not a duplicate colony, the Tn5 transposons were both in different genomic locations relative to oriC and the independent mariner transposons were also spaced differently around the chromosome so provided two examples of EM occurring. However, the double mutants could have been produced by separate independent transposition events rather than linked.

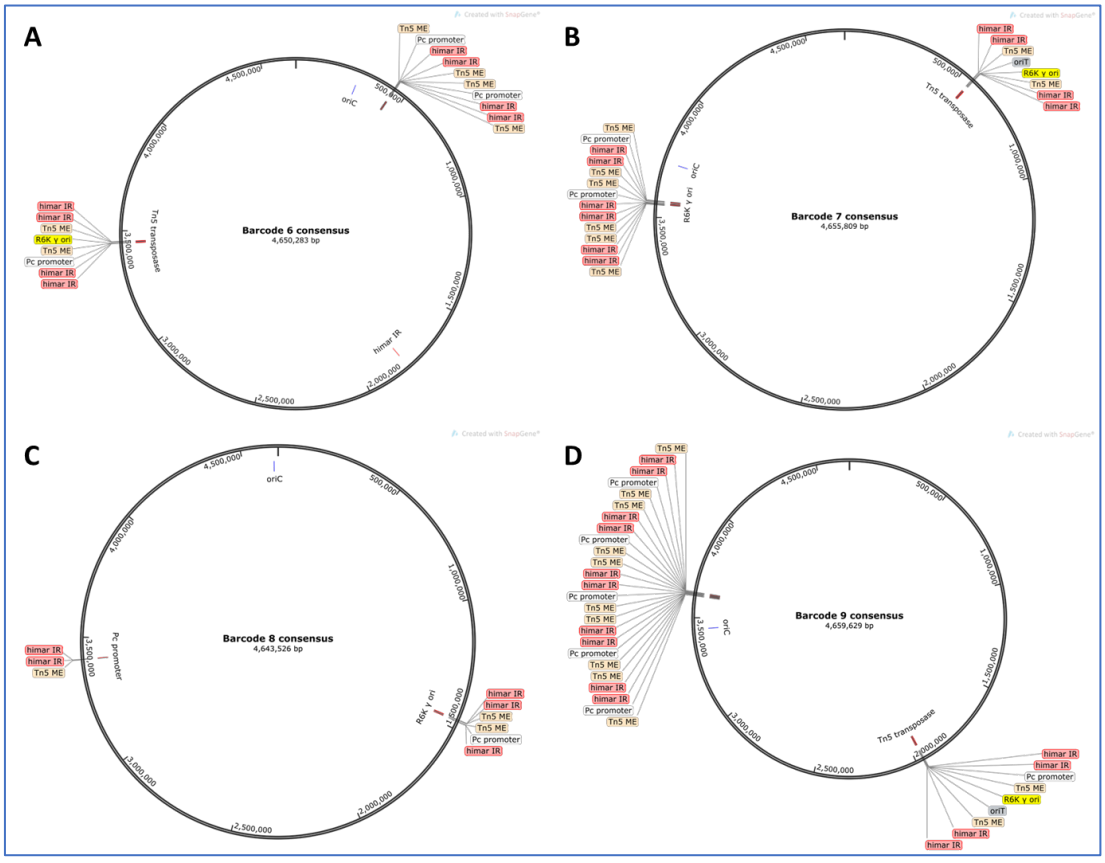

Figure 6-11 *Consensus long read assembled sequences.*

A: Mutant 6 showing the plasmid origin of replication (yellow). B: Mutant 7 showing the plasmid origin of replication (yellow) C: Mutant 8 showing two transposn events and potentially a succesful EM mutant. D: Mutant showing the plasmid origin of replication (yellow).

Mutants six, seven, eight and nine were unsuccessful, the sequence assemblies are shown in *Figure 6-11.* Generally, it seemed that there were errors introduced in the sequence assembly, there was multiple instances of MEs or IRs at potential insertion sites and in all cases the mariner IRs were identified outside of the Tn*5* ME. The repetitive sequence regions would have aligned several times in the consensus sequence, generating the issues seen. Mutants six, seven and nine all had the *R6k* origin of replication from the pExM plasmid in the consensus sequence so indicated that they were not real mutants or that there was an amount of plasmid remaining in the mutants. Overall, the long read sequence assemblies suggested that two of the chosen eight mutants were successful EM mutants.

6.3.3.2 Illumina Sequence Mapping

A summary of the BioTradis pipeline output for the concatenated sequence files is shown in *Table 6-5.* The mariner TIS had far fewer mapped transposon sites than the Tn*5* TIS. Overall, the number of reads mapping to the chromosome was lower than had been observed with the standard TIS libraries described in *chapter 3*. The mariner sequence library preparation was more efficient at enriching Tn-Chr junctions, 42 per cent of sequence reads started with a transposon compared to 21 per cent for Tn*5* but both were considered low compared to data presented in chapters three and four.

Table 6-5 *Summary of the BioTradis mapping output for EM sequencing, split by transposon.*

TIS	Total Reads	Transposon Reads		Chromosome Mapped Reads		Unique Insertion Sites	Insertion Distance (bp.)	Average reads / Insertion
		Count	%	Count	%			
Tn*5*	1.18E+07	2.49E+06	21.08	5.71E+05	22.90	1.03E+05	44.83	5.53
Mariner	4.40E+07	1.84E+07	41.84	3.23E+06	17.54	7.56E+03	612.93	427.15

The mariner libraries were not sufficiently saturated to get an essential gene list using the BioTradis pipeline (Barquist et al., 2016; Langridge et al., 2009), at this density and assuming equal distribution, each gene would have been expected to contain between one and two insertions and the average distance between transposons would have been 613 bp, this would exclude shorter genes from analysis. A genome scale overview of the mapped transposon insertions is shown in *Figure 6-12.*

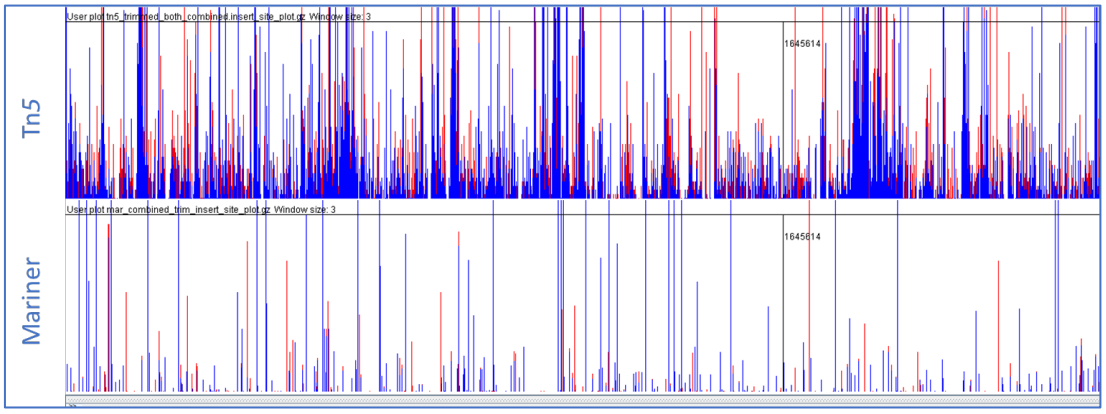

Figure 6-12 *Artemis visualisation of EM sequence mapping*

View of the mapped transposon insertions for the Tn5 Tn-Chr junctions enriched (top) and the mariner Tn-Chr junctions enriched (bottom) across the majority of the E. coli BW25113 genome.

The number of Tn*5* insertions mapped was sufficient to generate an essential gene list. This list highlighted genes that were essential and that were required if another, unknown, gene had been disrupted. Only 219 genes were determined to be essential (or ambiguous) with the EM Tn*5* insertion data compared to the 454 and 439 for Tn*5* and mariner libraries respectively from *chapter 3.* When the three lists of essential, plus ambiguous, genes were compared, there were 38 designated essential in the EM TIS data only, the number of overlapping genes for all 3 libraries was 131 *Figure 6-13*

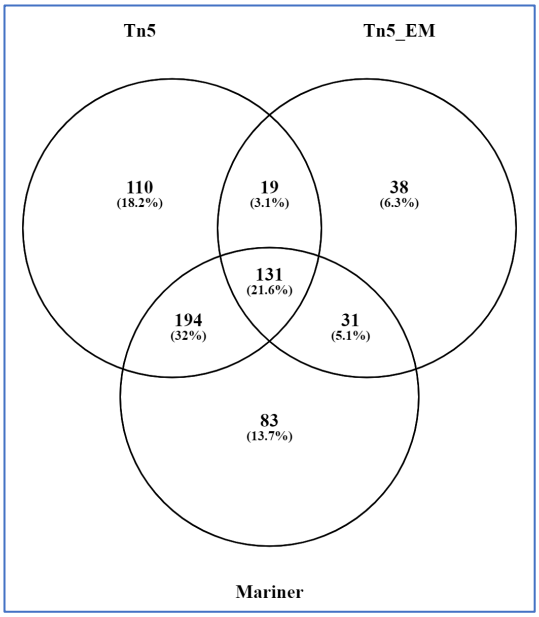

Figure 6-13 *the number of overlapping genes designated essential for mutant libraries in this work.*

The 38 genes exclusive to the Tn*5* EM data provided a list of potential candidate genes for redundancy, the list of these genes is provided in *appendix 9.3, sheet 5*. Amongst these genes was *ibsC,* this produces a toxic peptide that is repressed by an RNA antitoxin encoded on the opposite strand but overlapping *ibsC* (Mok et al., 2010). Another interesting gene in this list was *ruvC,* this participates in DNA repair at Holliday junctions and junctions formed with inverted repeats (Iwasaki et al., 1991); this gene would be required for DNA repair during transposition.

Visualisation of the insertion data for mariner in Artemis (Carver et al., 2012) showed that insertions in the essential gene *folC* (Baba et al., 2006; Goodall et al., 2018)*, Figure 6-14B* could be tolerated. The essential gene *folE* was free of insertions and remained essential for both EM and TIS*, Figure 6-14.*

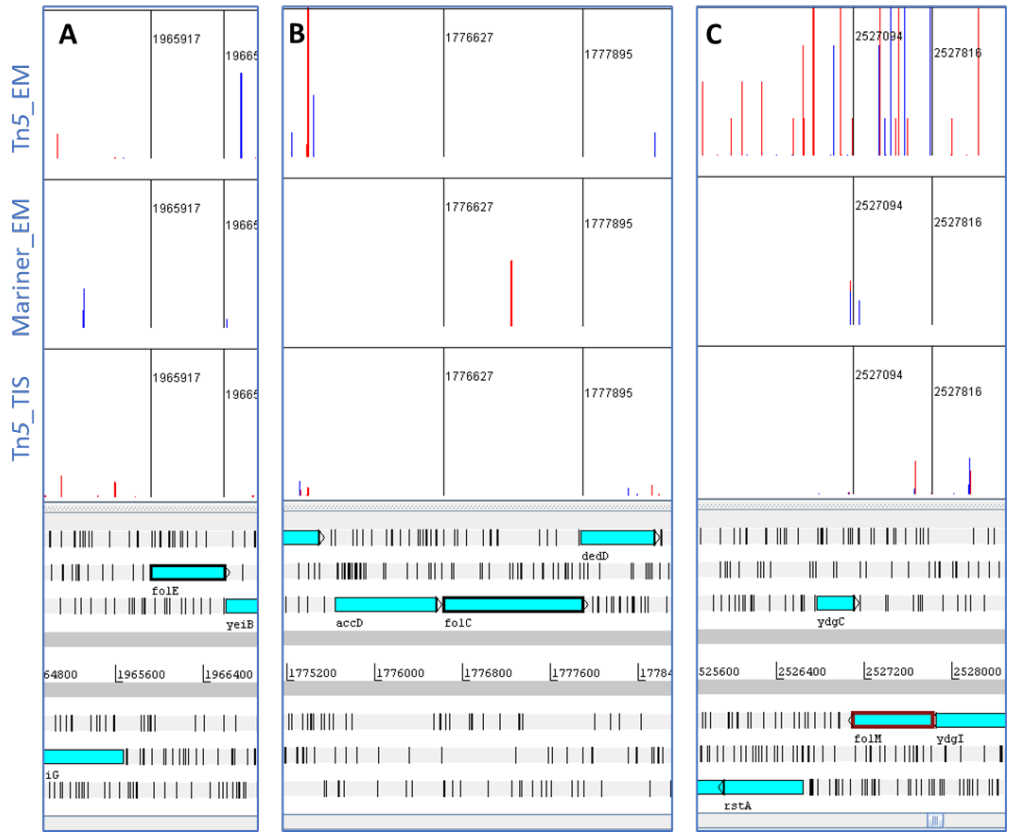

Figure 6-14 *Visaulisation of protected genes in a TIS library compared to EM libraries.*

A: folE, protected in both library types. B: folC, protected in the TIS library (bottom) but could tolerate an insertion in the EM library (middle). C: folM, protected in the TIS library (bottom) but beneficial to be inactivated in EM (top).

These genes are both involved in folate biosynthesis pathways, *folC* is bifunctional and is a dihydrofolate synthase (Keshavjee, Pyne and Bognars, 1991). The gene product of *folE* catalyses tetrahydrofolate synthesis (Kellermann et al., 1991). *Figure 6-14C* shows a different phenomenon that occurred in the EM data, it appeared that disruption to the gene *folM* was beneficial when another, unknown, gene or genes have been disrupted. Again, *folM* is involved in folate biosynthesis (Pribat et al., 2010) and the unexpected insertion patterns observed for these genes suggested that this is an example of a biosynthetic pathway with levels of redundancy and interconnected metabolic pathways.

6.4 Discussion

Data presented in this chapter demonstrate that using a composite transposon system to generate paired knockout mutants in a high throughput manner is possible, but the experimental methods require optimisation and data analysis needs to be improved.

A key element to successful double mutant generation was a controlled second transposition event, for this work the approach of catabolite repression using glucose (Griffiths et al., 2000; Santillán & Mackey, 2004) was used to restrict transposition. Transposition was induced by removing the suppression. Using IPTG (Naorem et al., 2018) to induce transposition was not successful in increasing the rate beyond that without induction. It is likely that overexpression of the transposase had a negative effect, as previously described (Tellier & Chalmers, 2020; Weinreich et al., 1994).

A *lac* promoter system may not be the most suitable, a scalable expression promoter such as arabinose (Khlebnikov et al., 2000) or rhamnose (Kelly et al., 2018) where the promoter activity is directly linked to the concentration of inducer may be of greater use. Another benefit would be that the metabolism of the organism does not need to change between transposon insertions.

EM should be as accessible as TIS, but further developments are required to ensure that the data acquired accurately represent a double knockout mutant. In the current state, the two transposons have not been linked and any comparison was made on the difference in genes listed as essential for one transposon insertion versus two transposons. This can provide a candidate list of genes to be further investigated but as demonstrated in *chapter 3*, there are variations in the number of genes determined to be essential between dense transposon libraries. Generating a saturated EM library will be challenging but necessary to achieve reproducible results. For development of antimicrobials, the candidate gene should reflect the contribution of a gene or gene pair to mutant fitness under antibiotic stress and not represent variation within the libraries.

Insertion data generated here identified a biosynthetic pathway, folate synthesis, where there is evidence to suggest redundancy and genetic interactions. This has been shown to be the case in the literature; *folM* forms a synthetic lethal pair with *folA* (Giladi et al., 2003). Mutants lacking *folM* show no observable growth defect and overexpression of *folM* can compensate for a *folA* mutant. This is an important association as *folA* is the target of the antimicrobial trimethoprim, as is *folC* due to the build-up of inhibitory compounds (Kwon et al., 2008). Disruption of *folM* leads to reduced susceptibility of trimethoprim and

sulfanomethoxine. The disruption in *folC* observed in data presented here suggests that in the absence of another gene, *folC* becomes dispensable. This may be via metabolic flux rerouting and may highlight alternative resistance mechanisms to known or novel compounds. Understanding these mechanisms will further narrow potential antimicrobial candidates to those where mutations leading to resistance evolution are not pursued.

In order to fully analyse pools of double mutants there needs to be an effective method for linking the two transposon insertion sites. One approach would be to use barcoding to identify mariner transposons that have originated from a specific Tn*5* transposon e.g., using a pool of random oligos to make a pool of barcoded Tn*5* transposons in a similar approach to (H. Liu et al., 2018). Alternatively, the sequencing could be performed in two stages and incorporating RBTn-seq (Wetmore et al., 2015), where one round of sequencing would identify the first transposon location and a further sequencing run to locate the second. This could be performed after the first transposition but before the second, each Tn*5* barcode would then be associated with an insertion site.

A limitation of the methods described here is that the genes encoding proteins required in conjugation or in DNA repair mechanisms will be deemed essential to fitness when they are essential to the methods used for library production. DNA repair is particularly important for repairing the chromosome post mariner excision from the composite transposon. For example, in their conjugative approach (Côté et al., 2016) found that *recA* had synthetic lethal interactions with every other gene, this was to be expected as *recA* is required for homologous recombination following conjugation (Kuzminov, 2011). Côte et al. utilised conjugation and homologous recombination to generate their mutants, and therefore *recA* was essential to the methodology and not to mutant fitness. This work showed that *ruvC*, a DNA repair enzyme is required for survival during EM.

An issue arising from the use of *E. coli* BW25113 is that *E. coli* is reported to have inefficient Non-Homologous End Joining (NHEJ) repair mechanisms (Chayot et al., 2010), DNA repair was one of the fundamental assumptions made during the composite transposon design and a damaged chromosome following transposon excision would likely be detrimental to a mutant. Accounting for the NHEJ rate (Chayot et al., 2010), the estimated double mutant efficiency rate is estimated at around at around 5^{-10}. As a future development, the first (Tn*5*) transposon could be modified to include the DNA repair genes *LigD* and *Ku* (Amare et al., 2021) from *Pseudomonas* or *Mycobacterium* (Malyarchuk et al., 2007; Paris et al.,

2015), as these enzymes have been used to improve DNA repair rates in *E. coli* previously (Malyarchuk et al., 2007).

Adding these two genes should allow the transposon to encode its own repair machinery and so increase the efficiency of EM double mutant production. Furthermore, if the transposon carried its own repair machinery, then it is possible that EM would be applicable to any bacterial species regardless of repair capabilities; there would be a reduction in conjugation and transposition efficiencies, but it would still provide an overall increase in double mutant generation. There is no reason that functional redundancy is limited to two genes, but EM is limited by transposon size, (this would be adding a minimum of 2kb for each iteration) and there is generally an inverse relationship between transposon size and transposition efficiency. Additionally, there is unlikely to be multiple layers of redundancy as it would increase the genome size of an organism and a larger genome is more prone to mutation; evolution has scaled towards minimised genomes.

6.5 Conclusions

This chapter has presented an approach to generate thousands of double knockout mutants using TIS methodologies. There are still caveats in terms of linking transposons, transposition efficiency, and data analysis. The preliminary data presented here has identified genes known to be involved in genetic interactions in conditions of limited stress. At the current state of development, EM would likely be able to provide a more comprehensive screen than TIS alone, particularly if the second transposition is performed after sequencing to confirm the location of the composite transposon in the genome.

7 Conclusions and Future Work

Firstly, before any targeted or random genome disruption occurs, accurate and consistent genome annotation is primarily important when looking at gene essentiality. Mis-annotated genomes or using inconsistent gene names generates confusion when comparing gene lists from different research groups.

Gene essentiality determination pipelines are not consistent in the classification of genes, even in the same organism. Most antimicrobial and drug discovery now occurs in research environments followed by small to medium enterprises (SMEs) that have niche applications of technology rather than the large corporations where the focus is on guaranteed profitable drugs . However, a SME is less likely to tolerate a failed investment so there is the potential loss of innovative drug discovery techniques. Genes that are misidentified as essential could be detrimental to a SME drug discovery as this could lead to wasted progress along the discovery pipeline. Conversely, should there be a gene that is required for growth but not identified by TIS screens then a missed opportunity for a number of compounds.

Essential cellular functions may be rescued by an alternative gene product or reaction pathway so while essential genes can be targets for antibiotics, understanding the interaction of genes may offer more robust target.

The aim of this project was to develop a methodology for high throughput production of paired double transposon knockout mutants. The intention was to use such libraries to advance the construction of genetic interaction network maps to uncover novel targets for antimicrobials or highlighting pathways that can be manipulated to produce novel bioactive compounds.

This work has identified the sources of some of the variability in transposon insertion and has suggested approaches to accommodating these in analysis pipelines so that the data better represents biological fitness, the measured outcome. This work has proposed a novel approach to defining essential (or protected) regions of the genome that can accommodate a normalised input. The model is annotation independent but can be contextualised with annotation following the main analysis. This analysis pipeline can offer a further dimension to analysis, one that clearly shows a region (or a specific genomic location) where an interruption is beneficial to the fitness of an organism. This has previously been observed in TIS data but had been determined by eye, the proposed analysis model would be standardised and measurable. In the context of drug discovery,

this is an important insight into the potential emergence of resistance or for indications of the mechanism of action of a candidate antimicrobial.

Finally, work towards a high throughput method of paired double knockouts has proven that the methods described can be used to generate a pool of thousands of mutants. While it still requires development, at the current state analysis of the mutant pool identified genes known to be synergistic and involved in essential metabolic pathways targeted by antimicrobials in clinical use.

7.1 Limitations of this work

One major limitation in this work was the saturation of the individual transposon mutant libraries; this was overcome by concatenating sequence files prior to analysis. This was prominent for the analysis of EM where one single sequence run did not have enough data to draw any conclusions. On the same note, the limited growth libraries provided sufficient transposon insertions to draw conclusions and observe differences in transposon insertion patterns across the genome, but there were not sufficient transposon insertions to automate normalisation and analysis.

When comparing gene essentiality using different transposon libraries (both amongst and between the Tn*5* and the mariner transposons), only one analysis pipeline was utilised for the analysis. There are other tools with alternative statistical approaches that may have produced more similar or different essential gene lists that may have led to alternative conclusions being drawn in *chapter 3.*

For the development of EM and the composite transposon, the methodology requires a bacterium that is manipulatable, to give statistical power to the screens then there needs to be sufficient mutants to an undetermined library saturation point. Not all organisms are as easy to manipulate as *E. coli,* methods were developed in a laboratory adapted strain that is known to have emerged from typical bacteria of the same species. Some organisms, such as Staphylococci, are much more difficult to manipulate with molecular genetics, at the current efficiency, EM may not be accessible for use.

7.2 Future Work

To further develop the TIS analysis methods described and EM mutant generation I would undertake a number of approaches both in the lab and computationally. These have been split into three timeframes based on the amount and the complexity of work that would be required to achieve them.

7.2.1 Short term:

Firstly, increasing the amount of DNA from the limited growth mutants to be sequenced for the limited growth libraries made with both the Tn*5* and mariner transposons would be recommended. The limited growth libraries require more depth of sequencing than a standard TIS mutant library so multiple preparation reactions are required to enrich sufficient Tn-Chr junctions. More mapped insertions would provide a better representation of the propensity for transposon insertion across the genome before growth (or mutant fitness) is assessed.

Also, with the increased saturation of a limited growth mutant library, normalise the TIS mutant library counts and run through BioTradis essentiality pipeline. To observe whether normalisation to transposon insertion patterns impacts the genes that are reported as essential for the same libraries and using the same reference genome, would be interesting. Running the normalised data through the Segmentor3IsBack model to see how the two and segmentor3 model to compare the essential genes that are identified by the two analysis pipelines, would be an interesting short-term approach.

The final short term development would be to generate more EM mutants to give a more comprehensive double knockout mutant library. The two disrupted genes would not be linked but in data presented in this work, linking the two transposons is not necessary to identify genomic interactions and genes of interest.

7.2.2 Medium Term:

These developments would take longer to achieve, and again are aimed towards improving EM mutant library saturation and to continuing to account for transposon insertion variability in an insertion site analysis model. As discussed in *chapter 5*, there are two genes (*ku* and *ligD*) that have previously been used to increase DNA repair efficiency (Malyarchuk et al., 2007). Cloning these onto the composite transposon, not within the second

transposon, would be of interest. The rationale behind this would be that if NHEJ repair is enhanced, the EM mutant generation efficiency should increase making it easier to achieve library saturation. This modification should also increase the range of organisms that the composite transposon will be active in.

Another modification that would be interesting to test into the composite transposon would be to exchange the constitutive *lac* promoter upstream of the mariner transposase to a fully inducible promoter, which would not be active unless the inducer is provided. Ideally this would be a *tac* promoter so that auto induction can be utilised, making the EM process higher throughput. This would allow single mutants to proliferate to a density required to provide sufficient mutant numbers for the second transposition, e.g., 10^9 of each single mutant to generate one insertion in each gene at an efficiency of 10^{-5}. This would also require the target organism to be able to metabolise glucose and lactose, *E. coli* BW25113 cannot utilise lactose so this approach was not explored in this work.

Another further development would be to refine the TIS analysis model by using the increased saturation limited growth library to model the insertion site preferences of a transposon. This could be used assign a probability matrix that is incorporated into normalisation based on the likelihood of the transposon of choice inserting at any site in the genome to further refine the essentiality determination. Additional developments would be to incorporate chromosome profiling data such as using ATAC-seq and Hi-C sequencing to infer transposon accessibility and the configuration of the chromosome (Buenrostro et al., 2015; Eagen, 2018). Using nanopore long read sequencing to investigate the epigenetic patterns of the chromosome would be ideal to determine if some of the observed variation in transposon insertion can be attributed to methylation.

7.2.3 Longer term:

1) A possible longer term development, originating from this work could be to construct a fully functional analysis model that is annotation independent and uses change point detection algorithms that is easy to use. This tool would be accessible and have optional normalisations to any of the factors listed in this work, that way the user would be able to normalise to data that they have available for their organism of interest. Annotation overlay would also be an optional and essential (or protected) gene lists along with those that are beneficial to remove would be exported.

2) Developments to EM would be to incorporate barcodes into the composite transposon, that way the two transposition events could be linked. If the location of the two transposons was known, any mutant of interest could be recreated, or alternatively, if there was the infrastructure then mutants could be arrayed individually for further analysis. Knowing the linked transposon insertion sites would make EM a valuable tool for helping to refine metabolic models, especially for organisms where much of the functional genome is unknown.

Just because a gene is within a genome, it is not transcribed in every situation. TraDIS Xpress (Yasir et al., 2020) is a modification on TIS protocols where the transposon contains an outward facing inducible promoter. Where the gene at the insertion site would normally be aberrated, this system instead allows control of expression. This enables conditionally essential genes to be assayed, in this instance their involvement in triclosan sensitivity. When the organism is put into the stressed state, the promoter can be induced and transcription occurs, in this instance a transposon knocking out a gene essential for resistance to triclosan would not be detected but if the transposon inserts upstream of the same gene, then the mutant would show resistance. This study identified one gene known to be involved in triclosan resistance and a further three novel genes, the novel genes were verified by using the specific knockouts from the Keio collection (Baba *et al.*, 2006). Incorporating inducible promoters, initially to one of the transposons would help to elucidate how the expression of synergistic genes can recover essential or conditionally essential functions within a cell.

Adding promoters to both transposons for EM would most likely overcomplicate analysis at this point in time but could be done with two scalable promoter types such as rhamnose and arabinose so that the expression of both can be tightly controlled. However, it may be preferable to investigate how expression is incorporated with inducible promoters once EM has identified a synergistic pair and the mutant genotype has been recreated.

7.3 Final Conclusions

TIS is an incredibly powerful tool for surveying the effects of the genome on bacterial fitness. This approach has been successfully utilised to identify targets for antimicrobials and to determine both the mechanism of action and resistance mechanisms to novel antimicrobials. A drug discovery pipeline would typically use multiple rounds of TIS screening to refine a candidate molecule. Using EM in the place of TIS will provide more depth to an essentiality screen and highlight gene interactions that have hidden key physiological processes that are enacted to overcome treatment.

Beyond investigating the mechanisms involved in tolerance to a novel compound, EM could help to identify new molecular scaffolds for antimicrobials. Interruptions to key metabolic pathways can lead to the accumulation of toxic compounds. If the method of degradation of these toxins also be interrupted (efflux or enzymatic), then the toxin will cause cell death. These compounds, and analogues of, can be screened against known pathogens for antimicrobial activity. In many organisms, intermediate compounds are not known or form part of cryptic metabolic pathways that are only active under specific environmental conditions that would be challenging to recreate in a laboratory condition.

8 Bibliography

Abdulmahdi, R., Jasim, F., & Jasim, R. A. F. (2021). Strategies for Challenging Development in Antimicrobial Resistance. *J Babylon*, *18*, 172–179. https://doi.org/10.4103/MJBL.MJBL_35_21

Alkam, D., Wongsurawat, T., Nookaew, I., Richardson, A. R., Ussery, D., Smeltzer, M. S., & Jenjaroenpun, P. (2021). Is amplification bias consequential in transposon sequencing (TnSeq) assays? A case study with a Staphylococcus aureus TnSeq library subjected to PCR-based and amplification-free enrichment methods. *Microbial Genomics*, *7*, 655. https://doi.org/10.1099/mgen.0.000655

Amare, B., Mo, A., Khan, N., Sowa, D. J., Warner, M. M., Tetenych, A., & Andres, S. N. (2021). LigD: A Structural Guide to the Multi-Tool of Bacterial Non-Homologous End Joining. *Frontiers in Molecular Biosciences*, *8*, 1161. https://doi.org/10.3389/FMOLB.2021.787709/BIBTEX

Aslam, B., Wang, W., Arshad, M. I., Khurshid, M., Muzammil, S., Rasool, M. H., Nisar, M. A., Alvi, R. F., Aslam, M. A., Qamar, M. U., Salamat, M. K. F., & Baloch, Z. (2018). Antibiotic resistance: a rundown of a global crisis. *Infection and Drug Resistance*, *11*, 1645–1658. https://doi.org/10.2147/IDR.S173867

Awano, N., Wada, M., Mori, H., Nakamori, S., & Takagi, H. (2005). Identification and functional analysis of Escherichia coli cysteine desulfhydrases. *Applied and Environmental Microbiology*, *71*(7), 4149–4152. https://doi.org/10.1128/AEM.71.7.4149-4152.2005

Aziz, R. K., Monk, J. M., Lewis, R. M., In Loh, S., Mishra, A., Abhay Nagle, A., Satyanarayana, C., Dhakshinamoorthy, S., Luche, M., Kitchen, D. B., Andrews, K. A., Fong, N. L., Li, H. J., Palsson, B. O., & Charusanti, P. (2015). Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. *Scientific Reports 2015 5:1*, *5*(1), 1–12. https://doi.org/10.1038/srep16025

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., Mori, H., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*. https://doi.org/10.1038/msb4100050

Bailey, T. L., & MacHanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, *40*(17), e128–e128. https://doi.org/10.1093/NAR/GKS433

Balasegaram, M., & Piddock, L. J. V. (2020). The Global Antibiotic Research and Development Partnership (GARDP) Not-for-Profit Model of Antibiotic Development. *ACS Infectious Diseases*, *6*(6), 1295–1298. https://doi.org/10.1021/ACSINFECDIS.0C00101/ASSET/IMAGES/LARGE/ID0C00101_0002.JPEG

Baltrus, D. A., Medlen, J., & Clark, M. (2019). Identifying transposon insertions in bacterial genomes through nanopore sequencing. *BioRxiv*, 765545. https://doi.org/10.1101/765545

Barquist, L., Boinett, C. J., & Cain, A. K. (2013). Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biology*, *10*(7), 1161–1169. https://doi.org/10.4161/rna.24765

Barquist, L., Mayho, M., Cummins, C., Cain, A. K., Boinett, C. J., Page, A. J., Langridge, G. C., Quail, M. A., Keane, J. A., & Parkhill, J. (2016). The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics (Oxford, England)*, *32*(7), 1109–1111. https://doi.org/10.1093/bioinformatics/btw022

Benstead-Hume, G., Chen, X., Hopkins, S. R., Lane, K. A., Downs, J. A., & Pearl, F. M. G. (2019). Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. *PLOS Computational Biology*, *15*(4), e1006888. https://doi.org/10.1371/JOURNAL.PCBI.1006888

Berlyn, M. K. (1998). Linkage map of Escherichia coli K-12, edition 10: the traditional map. *Microbiology and Molecular Biology Reviews : MMBR*, *62*(3), 814–984. http://susi.bio.unigiessen.de/ecdc.html,

Bhagirath, A. Y., Li, Y., Patidar, R., Yerex, K., Ma, X., Kumar, A., & Duan, K. (2019). Two Component Regulatory Systems and Antibiotic Resistance in Gram-Negative Pathogens. *International Journal of Molecular Sciences*, *20*(7). https://doi.org/10.3390/IJMS20071781

Blaskovich, M. A. T., Butler, M. S., & Cooper, M. A. (2017). Polishing the tarnished silver bullet: the quest for new antibiotics. *Essays in Biochemistry*, *61*(1), 103–114. https://doi.org/10.1042/EBC20160077

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements 06 Biological Sciences 0604 Genetics. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1577-z

Bronner, I. F., Otto, T. D., Zhang, M., Udenze, K., Wang, C., Quail, M. A., Jiang, R. H. Y., Adams, J. H., & Rayner, J. C. (2016). Quantitative insertion-site sequencing (QIseq) for high throughput phenotyping of transposon mutants. *Genome Research*, *26*(7), 980–989. https://doi.org/10.1101/GR.200279.115/-/DC1

Browning, D. F., Hobman, J. L., & Busby, S. J. W. (2022). *Laboratory strains of Escherichia coli K-12: not such perfect role models*. https://doi.org/10.1101/2022.06.29.497745

Bruinsma, S., Burgess, J., Schlingman, D., Czyz, A., Morrell, N., Ballenger, C., Meinholz, H., Brady, L., Khanna, A., Freeberg, L., Jackson, R. G., Mathonet, P., Verity, S. C., Slatter, A. F., Golshani, R., Grunenwald, H., Schroth, G. P., & Gormley, N. A. (2018). Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics*, *19*(1). https://doi.org/10.1186/S12864-018-5096-9

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *109*, 21.29.1. https://doi.org/10.1002/0471142727.MB2129S109

Butland, G., Babu, M., Díaz-Mejía, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O., Mori, H., Wanner, B. L., Lo, H., Wasniewski, J., Christopoulos, C., Ali, M., Venn, P., Safavi-Naini, A., Sourour, N., … Emili, A. (2008). eSGA: E. coli synthetic genetic array analysis. *Nature Methods*, *5*(9), 789–795. https://doi.org/10.1038/nmeth.1239

Byrne, R. T., Chen, S. H., Wood, E. A., Cabot, E. L., & Cox, M. M. (2014). Escherichia coli genes and pathways involved in surviving extreme exposure to ionizing radiation. *Journal of Bacteriology*, *196*(20), 3534–3545. https://doi.org/10.1128/JB.01589-14

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. In *Nature Reviews Genetics* (Vol. 21, Issue 9, pp. 526–540). Nature Publishing Group. https://doi.org/10.1038/s41576-020-0244-x

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, *28*(4), 464–469. https://doi.org/10.1093/bioinformatics/btr703

Chang, A. Y., Chau, V. W., Landas, J. A., & Yvonne. (2017). *Preparation of calcium competent Escherichia coli and heat-shock transformation*. *1*.

Chao, M. C., Abel, S., Davis, B. M., & Waldor, M. K. (2016a). The Design and Analysis of Transposon-Insertion Sequencing Experiments. *Nat Rev Microbiol*, *14*(2), 119–128. https://doi.org/10.1038/nrmicro.2015.7

Chao, M. C., Abel, S., Davis, B. M., & Waldor, M. K. (2016b). The Design and Analysis of Transposon-Insertion Sequencing Experiments. *Nature Reviews. Microbiology*, *14*(2), 119. https://doi.org/10.1038/NRMICRO.2015.7

Chaudhuri, R. R., Allen, A. G., Owen, P. J., Shalom, G., Stone, K., Harrison, M., Burgis, T. A., Lockyer, M., Garcia-Lara, J., Foster, S. J., Pleasance, S. J., Peters, S. E., Maskell, D. J., & Charles, I. G. (2009). Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics*, *10*(1), 1–18. https://doi.org/10.1186/1471-2164-10-291/FIGURES/6

Chayot, R., Montagne, B., Mazel, D., & Ricchetti, M. (2010). An end-joining repair mechanism in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(5), 2141–2146. https://doi.org/10.1073/PNAS.0906355107

Choudhery, S., Brown, A. J., Akusobi, C., Rubin, E. J., Sassetti, C. M., & Ioerger, T. R. (2021). *Modeling Site-Specific Nucleotide Biases Affecting Himar1 Transposon Insertion Frequencies in TnSeq Data Sets*. https://doi.org/10.1128/mSystems

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., Fero, M. J., McAdams, H. H., & Shapiro, L. (2011). The essential genome of a bacterium. *Molecular Systems Biology*, *7*(1), 528. https://doi.org/10.1038/MSB.2011.58

Chung, C. T., Niemela, S. L., & Miller, R. H. (1989). One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(7), 2172. https://doi.org/10.1073/PNAS.86.7.2172

Cleynen, A., Koskas, M., Lebarbier, E., Rigaill, G., & Robin, S. (2014). Segmentor3IsBack: An R package for the fast and exact segmentation of Seq-data. *Algorithms for Molecular Biology*, *9*(1), 1–11. https://doi.org/10.1186/1748-7188-9-6/TABLES/1

Comfort, N. C. (2001). From controlling elements to transposons: Barbara McClintock and the Nobel Prize. *Trends in Biochemical Sciences*, *26*(7), 454–457. https://doi.org/10.1016/S0968-0004(01)01898-9

Costanzo, M., Giaever, G., Nislow, C., & Andrews, B. (2006). Experimental approaches to identify genetic networks. In *Current Opinion in Biotechnology* (Vol. 17, Issue 5, pp. 472–480). Elsevier Current Trends. https://doi.org/10.1016/j.copbio.2006.08.005

Côté, J. P., French, S., Gehrke, S. S., MacNair, C. R., Mangat, C. S., Bharat, A., & Brown, E. D. (2016). The genome-wide interaction network of nutrient stress genes in Escherichia coli. *MBio*, *7*(6). https://doi.org/10.1128/MBIO.01714-16/SUPPL_FILE/MBO006163075ST4.XLSX

Coward, C., Dharmalingham, G., Abdulle, O., Avis, T., Beisken, S., Breidenstein, E., Carli, N., Figueiredo, L., Jones, D., Khan, N., Malara, S., Martins, J., Nagalingam, N., Turner, K., Wain, J., Williams, D., Powell, D., & Mason, C. (2020). High-density transposon libraries utilising outward-oriented promoters identify mechanisms of action and resistance to antimicrobials. *FEMS Microbiology Letters*, *367*(22), 185. https://doi.org/10.1093/FEMSLE/FNAA185

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6640–6645. https://doi.org/10.1073/PNAS.120163297/ASSET/103C1B8D-D302-4337-8E67-9F9084156407/ASSETS/GRAPHIC/PQ1201632006.JPEG

Davis, G., Kayser, K. J., Marx, A., Seitz, O., Suarez, M. F., Bozhkov, P., Day, J. G., & Stacey, G. (2008). Microbial Gene Essentiality. In *Gene Therapy Protocols Gene Therapy Protocols Organelle Proteomics Post-Transcriptional Gene Regulation Avidin–Biotin Interactions: Methods and Applications Tissue Engineering Microbial Gene Essentiality: Protocols and Bioinformatics Innate Immunity Envir* (Vol. 2, Issue 1).

DeJesus, M. A., Ambadipudi, C., Baker, R., Sassetti, C., & Ioerger, T. R. (2015). TRANSIT - A Software Tool for Himar1 TnSeq Analysis. *PLoS Computational Biology*, *11*(10), e1004401. https://doi.org/10.1371/journal.pcbi.1004401

DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., Sassetti, C. M., & Ioerger, T. R. (2017). Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. *MBio*, *8*(1). https://doi.org/10.1128/MBIO.02133-16/SUPPL_FILE/MBO002173137ST6.XLSX

DeJesus, M. A., & Ioerger, T. R. (2013). A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. *BMC Bioinformatics*, *14*(1), 1–12. https://doi.org/10.1186/1471-2105-14-303/TABLES/6

Díaz-Martín, A., Martínez-González, M. L., Ferrer, R., Ortiz-Leyba, C., Piacentini, E., Lopez-Pueyo, M. J., Martín-Loeches, I., Levy, M. M., Artigas, A., Garnacho-Montero, J., Navas, A., Álvarez, M., Sirvent, J. M., Ulldemolins, S. H., Galdós, P., Balziscueta, G., Marco, P., Azkarate, I., Sierra, R., … Vallejo, L. (2012). Antibiotic prescription patterns in the empiric therapy of severe sepsis: Combination of antimicrobials with different mechanisms of action reduces mortality. *Critical Care*, *16*(6), 1–9. https://doi.org/10.1186/CC11869/TABLES/6

Dicenzo, G. C., Mengoni, A., & Fondi, M. (2017). *Tn-Seq based metabolic modelling*. 1–27. https://doi.org/10.1101/221325

Dlakic, M., & Harrington, R. E. (1995). Bending and torsional flexibility of G/C-rich sequences as determined by cyclization assays. *Journal of Biological Chemistry*, *270*(50), 29945–29952. https://doi.org/10.1074/JBC.270.50.29945

Dower, W. J., Miller, J. F., & Ragsdale, C. W. (1988). High efficiency transformation of E.coli by high voltage electroporation. *Nucleic Acids Research*, *16*(13), 6127–6145. https://academic.oup.com/nar/article/16/13/6127/1046461

Draper, G. C., McLennan, N., Begg, K., Masters, M., & Donachie, W. D. (1998). Only the N-Terminal Domain of FtsK Functions in Cell Division. *Journal of Bacteriology*, *180*(17), 4621. https://doi.org/10.1128/JB.180.17.4621-4627.1998

Eagen, K. P. (2018). Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences*, *43*(6), 469–478. https://doi.org/10.1016/j.tibs.2018.03.006

Elena, S. F., & Lenski, R. E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, *390*(6658), 395–398. https://doi.org/10.1038/37108

Ellis, T., Wang, X., & Collins, J. J. (2009). Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotechnology*, *27*(5), 465–471. https://doi.org/10.1038/nbt.1536

Evans, L. E., Krishna, A., Ma, Y., Webb, T. E., Marshall, D. C., Tooke, C. L., Spencer, J., Clarke, T. B., Armstrong, A., & Edwards, A. M. (2019). *Exploitation of Antibiotic Resistance as a Novel Drug Target: Development of a β-Lactamase-Activated Antibacterial Prodrug*. https://doi.org/10.1021/acs.jmedchem.8b01923

Fasugba, O., Gardner, A., Mitchell, B. G., & Mnatzaganian, G. (2015). Ciprofloxacin resistance in community- and hospital-acquired Escherichia coli urinary tract infections: A systematic review and meta-analysis of observational studies. *BMC Infectious Diseases*, *15*(1), 1–16. https://doi.org/10.1186/S12879-015-1282-4/FIGURES/6

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson, B. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*. https://doi.org/10.1038/MSB4100155

Ferrières, L., Hémery, G., Nham, T., Guérout, A. M., Mazel, D., Beloin, C., & Ghigo, J. M. (2010). Silent mischief: Bacteriophage Mu insertions contaminate products of Escherichia coli random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. *Journal of Bacteriology*, *192*(24), 6418–6427. https://doi.org/10.1128/JB.00621-10

Fourment, M., & Gillings, M. R. (2008). A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics*, *9*(1), 1–9. https://doi.org/10.1186/1471-2105-9-82/TABLES/1

Freed, N. E. (2017). Creation of a Dense Transposon Insertion Library Using Bacterial Conjugation in Enterobacterial Strains Such As Escherichia Coli or Shigella flexneri. *Journal of Visualized Experiments : JoVE*, *2017*(127). https://doi.org/10.3791/56216

Gaio, D., Anantanawat, K., To, J., Liu, M., Monahan, L., & Darling, A. E. (2022). Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microbial Genomics*, *8*(1), 744. https://doi.org/10.1099/MGEN.0.000744

Gale, A. N., Sakhawala, R. M., Levitan, A., Sharan, R., Berman, J., Timp, W., & Cunningham, K. W. (2020). Identification of Essential Genes and Fluconazole Susceptibility Genes in Candida glabrata by Profiling Hermes Transposon Insertions. *G3 Genes|Genomes|Genetics*, *10*(10), 3859–3870. https://doi.org/10.1534/G3.120.401595

Garcia, E. C., Brumbaugh, A. R., & Mobley, H. L. T. (2011). Redundancy and Specificity of Escherichia coli Iron Acquisition Systems during Urinary Tract Infection . *Infection and Immunity*, *79*(3), 1225–1235. https://doi.org/10.1128/iai.01222-10

Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. v., & Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences*, *106*(38), 16422–16427. https://doi.org/10.1073/pnas.0906627106

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. v, Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A.-L., … Osterman, A. L. (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of Bacteriology*, *185*(19), 5673–5684. https://doi.org/10.1128/JB.185.19.5673-5684.2003

Ghomi, F. A., Langridge, G. C., Cain, A. K., Boinett, C., Abd, M., Ghany, E., Pickard, D. J., Kingsley, R. A., Thomson, N. R., Parkhill, J., Gardner, P. P., & Barquist, L. (2022). High-throughput transposon mutagenesis in the family Enterobacteriaceae reveals core essential genes and rapid turnover of essentiality. *BioRxiv*, 2022.10.20.512852. https://doi.org/10.1101/2022.10.20.512852

Giladi, M., Altman-Price, N., Levin, I., Levy, L., & Mevarech, M. (2003). FolM, a new chromosomally encoded dihydrofolate reductase in Escherichia coli. *Journal of Bacteriology*, *185*(23), 7015–7018. https://doi.org/10.1128/JB.185.23.7015-7018.2003

Gillis, J., & Pavlidis, P. (2011). The Impact of Multifunctional Genes on "Guilt by Association. *Analysis. PLoS ONE*, *6*(2), 17258. https://doi.org/10.1371/journal.pone.0017258

Goldford, J. E., George, A. B., Flamholz, A. I., & Segre, D. (2022). Protein cost minimization promotes the emergence of coenzyme redundancy. *Proceedings of the National*

Academy of Sciences of the United States of America, *119*(14), e2110787119. https://doi.org/10.1073/PNAS.2110787119/SUPPL_FILE/PNAS.2110787119.SD01.XLSX

Goodall, E. C. A., Robinson, A., Johnston, I. G., Jabbari, S., Turner, K. A., Cunningham, A. F., Lund, P. A., Cole, J. A., & Henderson, I. R. (2018). The Essential Genome of Escherichia coli K-12. *MBio*, *9*(1). https://doi.org/10.1128/mbio.02096-17

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., & Gordon, J. I. (2009). Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. *Cell Host and Microbe*, *6*(3), 279–289. https://doi.org/10.1016/j.chom.2009.08.003

Goryshin, I. Y., Miller, J. A., Kil, Y. v, Lanzov, V. A., & Reznikoff, W. S. (1998). *Tn5/IS50 target recognition*. *95*, 10716–10721. www.pnas.org.

Goryshin, I. Y., Naumann, T. A., Apodaca, J., & Reznikoff, W. S. (2003). Chromosomal Deletion Formation System Based on Tn5 Double Transposition: Use For Making Minimal Genomes and Essential Gene Analysis. *Genome Research*, *13*(4), 644–653. https://doi.org/10.1101/GR.611403

Green, B., Bouchier, C., Fairhead, C., Craig, N. L., & Cormack, B. P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mobile DNA*, *3*(1), 1–6. https://doi.org/10.1186/1759-8753-3-3/FIGURES/3

Grenier, F., Matteau, D., Baby, V., & Rodrigue, S. (2014). Complete genome sequence of Escherichia coli BW25113. *Genome Announcements*, *2*(5), 1038–1052. https://doi.org/10.1128/genomeA.01038-14

Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C., & Gelbart, W. M. (2000). *Catabolite repression of the lac operon: positive control*. https://www.ncbi.nlm.nih.gov/books/NBK22065/

Grimbs, A., Klosik, D. F., Bornholdt, S., & Hütt, M. T. (2019). A system-wide network reconstruction of gene regulation and metabolism in Escherichia coli. *PLOS Computational Biology*, *15*(5), e1006962. https://doi.org/10.1371/JOURNAL.PCBI.1006962

Guzmán, G. I., Olson, C. A., Hefner, Y., Phaneuf, P. v., Catoiu, E., Crepaldi, L. B., Micas, L. G., Palsson, B. O., & Feist, A. M. (2018). Reframing gene essentiality in terms of adaptive

flexibility. *BMC Systems Biology*, *12*(1), 143. https://doi.org/10.1186/s12918-018-0653-z

Harte, D. (2021). HiddenMarkov: Hidden Markpv Models. In *Statistics Research Associates, Wellington.* (R package version 1.8-13). Statistics Research Associates. https://www.statsresearch.co.nz/dsh/sslib/

Hayes, F. (2003). Transposon-Based Strategies for Microbial Functional Genomics and Proteomics. *Annual Review of Genetics*, *37*(1), 3–29. https://doi.org/10.1146/annurev.genet.37.110801.142807

Hensel, M., Shea, J. E., Gleeson, C., Jones, M. D., Dalton, E., & Holden, D. W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. *Science (New York, N.Y.)*, *269*(5222), 400–403. https://doi.org/10.1126/SCIENCE.7618105

Holden, E. R., Yasir, M., Turner, A. K., Wain, J., Charles, I. G., & Webber, M. A. (2021). Massively parallel transposon mutagenesis identifies temporally essential genes for biofilm formation in Escherichia coli. *Microbial Genomics*, *7*(11), 673. https://doi.org/10.1099/MGEN.0.000673

Hutchison, C. A., Merryman, C., Sun, L., Assad-Garcia, N., Richter, R. A., Smith, H. O., & Glass, J. I. (2019). *Polar Effects of Transposon Insertion into a Minimal Bacterial Genome*. https://doi.org/10.1128/JB

Illumina. (2018). *NextSeq System Denature and Dilute Libraries Guide* . www.illumina.com/company/legal.html.

Illumina. (2019). *Illumina Adapter Sequences (1000000002694)*. www.illumina.com/company/legal.html.

Illumina. (2022). *Illumina DNA Prep Reference Guide (1000000025416)*.

Iwasaki, H., Takahagi, M., Shiba, T., Nakata, A., & Shinagawa, H. (1991). Escherichia coli RuvC protein is an endonuclease that resolves the Holliday structure. *The EMBO Journal*, *10*(13), 4381–4389.

Jackson, N., Czaplewski, L., & Piddock, L. J. V. (2018). *Discovery and development of new antibacterial drugs: learning from experience?* https://doi.org/10.1093/jac/dky019

Jackson, S. A., Fellows, B. J., & Fineran, P. C. (2020). Complete Genome Sequences of the Escherichia coli Donor Strains ST18 and MFD pir. *Microbiology Resource Announcements*, *9*(45). https://doi.org/10.1128/MRA.01014-20

Jacobsson, S., Mason, C., Khan, N., Meo, P., & Unemo, M. (2020). High in vitro activity of DIS-73285, a novel antimicrobial with a new mechanism of action, against MDR and XDR Neisseria gonorrhoeae. *Journal of Antimicrobial Chemotherapy*, *75*(11), 3244. https://doi.org/10.1093/JAC/DKAA322

Jana, B., Cain, A. K., Doerrler, W. T., Boinett, C. J., Fookes, M. C., Parkhill, J., & Guardabassi, L. (2017). The secondary resistome of multidrug-resistant Klebsiella pneumoniae. *Scientific Reports*, *7*. https://doi.org/10.1038/srep42483

Johnson, B., & Chandran, A. S. (2021). COMPARISON BETWEEN PYTHON, JAVA AND R PROGRMMING LANGUAGE IN MACHINE LEARNING. *Www.Irjmets.Com @International Research Journal of Modernization in Engineering*, *3288*. www.irjmets.com

Joyce, A. R., Reed, J. L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S. A., Palsson, B., & Agarwalla, S. (2006). Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli. *Journal of Bacteriology*, *188*(23), 8259. https://doi.org/10.1128/JB.00740-06

Juhas, M., Eberl, L., & Church, G. M. (2012). Essential genes as antimicrobial targets and cornerstones of synthetic biology. In *Trends in Biotechnology* (Vol. 30, Issue 11, pp. 601–607). Elsevier. https://doi.org/10.1016/j.tibtech.2012.08.002

Kato, J. I., & Hashimoto, M. (2007). Construction of consecutive deletions of the Escherichia coli chromosome. *Molecular Systems Biology*, *3*(1), 132. https://doi.org/10.1038/MSB4100174

Kellermann, J., Lottspeich, F., & Bachera, A. (1991). Biosynthesis of tetrahydrofolate. Sequence of GTP cyclohydrolase I from Escherichia coli. *Biological Chemistry Hoppe-Seyler*, *372*(11), 991–998. https://doi.org/10.1515/BCHM3.1991.372.2.991

Kelly, C. L., Taylor, G. M., Hitchcock, A., Torres-Méndez, A., & Heap, J. T. (2018). A Rhamnose-Inducible System for Precise and Temporal Control of Gene Expression in Cyanobacteria. *ACS Synthetic Biology*, *7*(4), 1056–1066.

https://doi.org/10.1021/ACSSYNBIO.7B00435/ASSET/IMAGES/LARGE/SB-2017-00435Z_0008.JPEG

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., … Karp, P. D. (2017). The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, *45*(D1), D543–D550. https://doi.org/10.1093/nar/gkw1003

Keshavjee, K., Pyne, C., & Bognars, A. L. (1991). Characterization of a Mutation Affecting the Function of Escherichia coli Folylpolyglutamate Synthetase-Dihydrofolate Synthetase and Further Mutations Produced in Vitro at the Same Locus*. *Journal of Biological Chemistry*, *266*(30), 19925–19929. https://doi.org/10.1016/S0021-9258(18)54871-7

Khatiwara, A., Jiang, T., Sung, S. S., Dawoud, T., Kim, J. N., Bhattacharya, D., Kim, H. B., Ricke, S. C., & Kwon, Y. M. (2012). Genome scanning for conditionally essential genes in Salmonella enterica serotype typhimurium. *Applied and Environmental Microbiology*, *78*(9), 3098–3107. https://doi.org/10.1128/AEM.06865-11/SUPPL_FILE/AEM-AEM06865-11-S04.ZIP

Khlebnikov, A., Risa, Skaug, T., Carrier, T. A., & Keasling, J. D. (2000). Regulatable Arabinose-Inducible Gene Expression System with Consistent Control in All Cells of a Culture. *Journal of Bacteriology*, *182*(24), 7029. https://doi.org/10.1128/JB.182.24.7029-7034.2000

Kimura, S., Hubbard, T. P., Davis, B. M., & Waldor, M. K. (2016). The nucleoid binding protein H-NS biases genome-wide transposon insertion landscapes. *MBio*, *7*(4). https://doi.org/10.1128/MBIO.01351-16/SUPPL_FILE/MBO004162965ST4.XLS

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., & Mori, H. (2006). Complete set of ORF clones of Escherichia coli ASKA library (A Complete Set of E. coli K-12 ORF Archive): Unique Resources for Biological Research. *DNA Research*, *12*(5), 291–299. https://doi.org/10.1093/dnares/dsi012

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology 2019 37:5*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Konreddy, A. K., Rani, G. U., Lee, K., & Choi, Y. (2019). Recent Drug-Repurposing-Driven Advances in the Discovery of Novel Antibiotics. *Current Medicinal Chemistry*, *26*(28), 5363–5388. https://doi.org/10.2174/0929867325666180706101404

Koo, B. M., Kritikos, G., Farelli, J. D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J. M., Hachmann, A. B., Rudner, D. Z., Allen, K. N., Typas, A., & Gross, C. A. (2017). Construction and Analysis of Two Genome-Scale Deletion Libraries for Bacillus subtilis. *Cell Systems*, *4*(3), 291-305.e7. https://doi.org/10.1016/j.cels.2016.12.013

Krebs, M. P., & Reznikoff, W. S. (1988). Use of a Tn5 derivative that creates Z&Z translational fusions to obtain a transposition mutant (IS50; transposition screen; papillation; transposase inhibitor; ochre suppressor; fusion proteins; recombinant DNA). *Gene*, *63*, 277–285.

Kuan, C. T., & Tessman, I. (1991). LexA protein of Escherichia coli represses expression of the Tn5 transposase gene. *Journal of Bacteriology*, *173*(20), 6406–6410. https://doi.org/10.1128/JB.173.20.6406-6410.1991

Kumar, A., Beloglazova, N., Bundalovic-Torma, C., Phanse, S., Deineko, V., Gagarinova, A., Musso, G., Vlasblom, J., Lemak, S., Hooshyar, M., Minic, Z., Wagih, O., Mosca, R., Aloy, P., Golshani, A., Parkinson, J., Emili, A., Yakunin, A. F., & Babu, M. (2016). Conditional Epistatic Interaction Maps Reveal Global Functional Rewiring of Genome Integrity Pathways in Escherichia coli. *Cell Reports*, *14*(3), 648–661. https://doi.org/10.1016/j.celrep.2015.12.060

Kumar, A., Safdar, N., Kethireddy, S., & Chateau, D. (2010). A survival benefit of combination antibiotic therapy for serious infections associated with sepsis and septic shock is contingent only on the risk of death: a meta-analytic/meta-regression study. *Critical Care Medicine*, *38*(8), 1651–1664. https://doi.org/10.1097/CCM.0B013E3181E96B91

Kuzminov, A. (2011). Homologous Recombination-Experimental Systems, Analysis, and Significance. *EcoSal Plus*, *4*(2). https://doi.org/10.1128/ECOSALPLUS.7.2.6

Kwon, Y. K., Lu, W., Melamud, E., Khanam, N., Bognar, A., & Rabinowitz, J. D. (2008). A domino effect in antifolate drug action in Escherichia coli. *Nature Chemical Biology*, *4*(10), 602. https://doi.org/10.1038/NCHEMBIO.108

Lampe, D. J., Akerley, B. J., Rubin, E. J., Mekalanos, J. J., & Robertson, H. M. (1999). Hyperactive transposase mutants of the Himar1 mariner transposon. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(20), 11428–11433. https://doi.org/10.1073/PNAS.96.20.11428/ASSET/4B1C911E-7ED8-440B-BE17-F5E5CCFCE3F8/ASSETS/GRAPHIC/PQ1993188004.JPEG

Langridge, G. C., Phan, M. D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., & Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, *19*(12), 2308–2316. https://doi.org/10.1101/gr.097097.109

Larivière, D., Wickham, L., Keiler, K., & Nekrutenko, A. (2021). Reproducible and accessible analysis of transposon insertion sequencing in Galaxy for qualitative essentiality analyses. *BMC Microbiology*, *21*(1), 1–15. https://doi.org/10.1186/S12866-021-02184-4/FIGURES/9

Láruson, Á. J., Yeaman, S., & Lotterhos, K. E. (2020). *The Importance of Genetic Redundancy in Evolution*. https://doi.org/10.1016/j.tree.2020.04.009

Lee, S. T., Kim, J. Y., Kown, M. J., Kim, S. W., Chung, J. H., Ahn, M. J., Oh, Y. L., Kim, J. W., & Ki, C. S. (2011). Mutant Enrichment with 3′-Modified Oligonucleotides: A Practical PCR Method for Detecting Trace Mutant DNAs. *The Journal of Molecular Diagnostics : JMD*, *13*(6), 657. https://doi.org/10.1016/J.JMOLDX.2011.07.003

Lee Ventola, C. (2015). *The Antibiotic Resistance Crisis Part 1: Causes and Threats*. *40*(4).

Lepak, A. J., Parhi, A., Madison, M., Marchillo, K., Vanhecker, J., & Andes, D. R. (2015). In Vivo Pharmacodynamic Evaluation of an FtsZ Inhibitor, TXA-709, and Its Active Metabolite, TXA-707, in a Murine Neutropenic Thigh Infection Model. *Antimicrobial Agents and Chemotherapy*, *59*. https://doi.org/10.1128/AAC.01464-15

Lepak, A. J., Wang, W., & Andes, D. R. (2020). Pharmacodynamic Evaluation of MRX-8, a Novel Polymyxin, in the Neutropenic Mouse Thigh and Lung Infection Models against Gram-Negative Pathogens. *Antimicrobial Agents and Chemotherapy*, *64*(11). https://doi.org/10.1128/AAC.01517-20

Lerminiaux, N. A., & Cameron, A. D. S. (2019). Horizontal transfer of antibiotic resistance genes in clinical environments. *Canadian Journal of Microbiology*, *65*(1), 34–44. https://doi.org/10.1139/CJM-2018-0275/ASSET/IMAGES/CJM-2018-0275TAB1.GIF

Li, B., Cao, W., Zhou, J., & Luo, F. (2011). Understanding and predicting synthetic lethal genetic interactions in Saccharomyces cerevisiae using domain genetic interactions. *BMC Systems Biology*, *5*, 73. https://doi.org/10.1186/1752-0509-5-73

Li, H. (2017). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lipszyc, A., Szuplewska, M., & Bartosik, D. (2022). How Do Transposable Elements Activate Expression of Transcriptionally Silent Antibiotic Resistance Genes? *International Journal of Molecular Sciences 2022, Vol. 23, Page 8063*, *23*(15), 8063. https://doi.org/10.3390/IJMS23158063

Liu, D., & Chalmers, R. (2014). Hyperactive mariner transposons are created by mutations that disrupt allosterism and increase the rate of transposon end synapsis. *Nucleic Acids Research*, *42*(4), 2637–2645. https://doi.org/10.1093/NAR/GKT1218

Liu, H., Price, M. N., Waters, R. J., Ray, J., Carlson, H. K., Lamson, J. S., Chakraborty, R., Arkin, A. P., & Deutschbauer, A. M. (2018). Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria. *MSystems*, *3*(1). https://doi.org/10.1128/mSystems.00143-17

Lott, M., Yasir, M., Turner, A. K., Bastkowski, S., Page, A., Webber, M. A., & Charles, I. G. (2022). LoRTIS Software Suite: Transposon mutant analysis using long-read sequencing. *BioRxiv*, 2022.05.26.493556. https://doi.org/10.1101/2022.05.26.493556

Luo, H., Lin, Y., Gao, F., Zhang, C. T., & Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research*, *42*(D1), D574. https://doi.org/10.1093/nar/gkt1131

Luo, H., Lin, Y., Liu, T., Lai, F. L., Zhang, C. T., Gao, F., & Zhang, R. (2021). DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Research*, *49*(D1), D677–D686. https://doi.org/10.1093/nar/gkaa917

Maddamsetti, R. (2021). Selection maintains protein interactome resilience in the long-term evolution experiment with Escherichia coli. *BioRxiv*, 2021.01.20.427477. https://doi.org/10.1101/2021.01.20.427477

Mahadevan, R., & Lovley, D. R. (2008). The Degree of Redundancy in Metabolic Genes Is Linked to Mode of Metabolism. *Biophysical Journal*, *94*(4), 1216–1220. https://doi.org/10.1529/BIOPHYSJ.107.118414

Mahase, E. (2020). *Antibiotic resistance: more companies are withdrawing sales reps or incentives in order to limit use, finds report*. https://doi.org/10.1136/bmj.m235

Mahmutovic, A., Abel zur Wiesch, P., & Abel, S. (2020). Selection or drift: The population biology underlying transposon insertion sequencing experiments. *Computational and Structural Biotechnology Journal*, *18*, 791–804. https://doi.org/10.1016/J.CSBJ.2020.03.021

Malyarchuk, S., Wright, D., Castore, R., Klepper, E., Weiss, B., Doherty, A. J., & Harrison, L. (2007). *Expression of Mycobacterium tuberculosis Ku and Ligase D in Escherichia coli results in RecA and RecB-independent DNA end-joining at regions of microhomology*. https://doi.org/10.1016/j.dnarep.2007.04.004

Mani, R., St Onge, R. P., Hartman, J. L., Giaever, G., Roth, F. P., & Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(9), 3461–3466. https://doi.org/10.1073/pnas.0712255105

Exponential Mutagenesis - A new approach to antibiotic discovery, Unpublished Proposal ___ (2018). https://doi.org/10.22201/fq.18708404e.2004.3.66178

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Martínez-García, E., Aparicio, T., de Lorenzo, V., & Nikel, P. I. (2014). New transposon tools tailored for metabolic engineering of Gram-negative microbial cell factories. *Frontiers in Bioengineering and Biotechnology*, *2*(OCT). https://doi.org/10.3389/FBIOE.2014.00046/ABSTRACT

Martínez-García, E., Aparicio, T., Lorenzo, V. de, & Nikel, P. I. (2014). New Transposon Tools Tailored for Metabolic Engineering of Gram-Negative Microbial Cell Factories. *Frontiers in Bioengineering and Biotechnology*, *2*. https://doi.org/10.3389/FBIOE.2014.00046

Martínez-García, E., Calles, B., Arévalo-Rodríguez, M., & de Lorenzo, V. (2011). pBAM1: an all-synthetic genetic tool for analysis and construction of complex bacterial phenotypes. *BMC Microbiology*, *11*, 38. https://doi.org/10.1186/1471-2180-11-38

Marzan, L. W., Barua, R., Akter, Y., Arifuzzaman, Md., Islam, Md. R., & Shimizu, K. (2017). A single metabolite production by Escherichia coli BW25113 and its pflA.cra mutant cultivated under microaerobic conditions using glycerol or glucose as a carbon source. *Journal of Genetic Engineering and Biotechnology*, *15*(1), 161–168. https://doi.org/10.1016/J.JGEB.2017.01.004

Mashimo, K., Nagata, Y., Kawata, M., Iwasaki, H., & Yamamoto, K. (2004). Role of the RuvAB protein in avoiding spontaneous formation of deletion mutations in the Escherichia coli K-12 endogenous tonB gene. *Biochemical and Biophysical Research Communications*, *323*(1), 197–203. https://doi.org/10.1016/J.BBRC.2004.08.078

Matlock, B. (2015). *Assessment of Nucleic Acid Purity*. www.thermoscientific.com

Matthews, T. C., Bristow, F. R., Griffiths, E. J., Petkau, A., Adam, J., Dooley, D., Kruczkiewicz, P., Curatcha, J., Cabral, J., Fornika, D., Winsor, G. L., Courtot, M., Bertelli, C., Roudgar, A., Feijao, P., Mabon, P., Enns, E., Thiessen, J., Keddy, A., … van Domselaar, G. (n.d.). *The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform*. https://doi.org/10.1101/381830

Mccarthy, A. J., Stabler, R. A., & Taylor, P. W. (2018). *Genome-Wide Identification by Transposon Insertion Sequencing of Escherichia coli K1 Genes Essential for In Vitro Growth, Gastrointestinal Colonizing Capacity, and Survival in Serum*. https://doi.org/10.1128/JB.00698-17

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(6), 344. https://doi.org/10.1073/PNAS.36.6.344

Mccoy, K. M., Antonio, M. L., & van Opijnen, T. (2017). *MAGenTA: a Galaxy implemented tool for complete Tn-Seq analysis and data visualization*. https://doi.org/10.1093/bioinformatics/btx320

Meredith, T. C., Wang, H., Beaulieu, P., Gründling, A., & Roemer, T. (2012). Harnessing the power of transposon mutagenesis for antibacterial target identification and evaluation. *Mobile Genetic Elements*, *2*(4), 171. https://doi.org/10.4161/MGE.21647

Miethke, M., Pieroni, M., Weber, T., Brönstrup, M., Hammann, P., Halby, L., Arimondo, P. B., Glaser, P., Aigle, B., Bode, H. B., Moreira, R., Li, Y., Luzhetskyy, A., Medema, M. H., Pernodet, J.-L., Stadler, M., Rubén Tormo, J., Genilloud, O., Truman, A. W., … Müller,

R. (2021). *Towards the sustainable discovery and development of new antibiotics*. https://doi.org/10.1038/s41570-021-00313-1

Miravet-Verde, S., Burgos, R., Delgado, J., Lluch-Senar, M., & Serrano, L. (2020). FASTQINS and ANUBIS: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies. *Nucleic Acids Research*, *48*(17), 102. https://doi.org/10.1093/nar/gkaa679

Mobegi, F. M., van Hijum, S. A. F. T., Burghout, P., Bootsma, H. J., de Vries, S. P. W., van der Gaast-de Jongh, C. E., Simonetti, E., Langereis, J. D., Hermans, P. W. M., de Jonge, M. I., & Zomer, A. (2014). From microbial gene essentiality to novel antimicrobial drug targets. *BMC Genomics*, *15*(1), 1–11. https://doi.org/10.1186/1471-2164-15-958/FIGURES/3

Mok, W. W. K., Patel, N. H., & Li, Y. (2010). Decoding Toxicity. *The Journal of Biological Chemistry*, *285*(53), 41627. https://doi.org/10.1074/JBC.M110.149179

Monk, I. R., Shah, I. M., Xu, M., Tan, M.-W., & Foster, T. J. (2012). Transforming the untransformable: application of direct transformation to manipulate genetically Staphylococcus aureus and Staphylococcus epidermidis. *MBio*, *3*(2), e00277-11. https://doi.org/10.1128/mBio.00277-11

Morris, E. R., Grey, H., McKenzie, G., Jones, A. C., & Richardson, J. M. (2016). A bend, flip and trap mechanism for transposon integration. *ELife*, *5*(MAY2016). https://doi.org/10.7554/ELIFE.15537

Muñoz-López, M., & García-Pérez, J. L. (2010). DNA Transposons: Nature and Applications in Genomics. *Current Genomics*, *11*(2), 115. https://doi.org/10.2174/138920210790886871

Murlidharan Nair, T. (2010). *Sequence periodicity in nucleosomal DNA and intrinsic curvature*. https://doi.org/10.1186/1472-6807-10-S1-S8

Naorem, S. S., Han, J., Zhang, S. Y., Zhang, J., Graham, L. B., Song, A., Smith, C. v., Rashid, F., & Guo, H. (2018). Efficient transposon mutagenesis mediated by an IPTG-controlled conditional suicide plasmid. *BMC Microbiology*, *18*(1), 1–11. https://doi.org/10.1186/S12866-018-1319-0/FIGURES/4

Nazareno, E. S., Acharya, B., & Dumenyo, C. K. (2021). A mini-Tn5-derived transposon with reportable and selectable markers enables rapid generation and screening of

insertional mutants in Gram-negative bacteria. *Letters in Applied Microbiology*, *72*(3), 283–291. https://doi.org/10.1111/lam.13423

Nlebedim, V. U., Chaudhuri, R. R., & Walters, K. (2021). Probabilistic identification of bacterial essential genes via insertion density using TraDIS data with Tn5 libraries. *Bioinformatics*, *37*(23), 4343–4349. https://doi.org/10.1093/BIOINFORMATICS/BTAB508

O'Dwyer, K., Spivak, A. T., Ingraham, K., Min, S., Holmes, D. J., Jakielaszek, C., Rittenhouse, S., Kwan, A. L., Livi, G. P., Sathe, G., Thomas, E., van Horn, S., Miller, L. A., Twynholm, M., Tomayko, J., Dalessandro, M., Caltabiano, M., Scangarella-Oman, N. E., & Brown, J. R. (2015). Bacterial Resistance to Leucyl-tRNA Synthetase Inhibitor GSK2251052 Develops during Treatment of Complicated Urinary Tract Infections. *Antimicrobial Agents and Chemotherapy*, *59*(1), 289. https://doi.org/10.1128/AAC.03774-14

O'Neill, J. (2014). Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations. *The Review on Antimicrobial Resistance*, *December*. https://amr-review.org/sites/default/files/AMR Review Paper - Tackling a crisis for the health and wealth of nations_1.pdf

Oshima, T., Aiba, H., Masuda, Y., Kanaya, S., Sugiura, M., Wanner, B. L., Mori, H., & Mizuno, T. (2002). Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12. *Molecular Microbiology*, *46*(1), 281–291. https://doi.org/10.1046/j.1365-2958.2002.03170.x

Outterson, K., Rex, J. H., Jinks, T., Jackson, P., Hallinan, J., Karp, S., Hung, D. T., Franceschi, F., Merkeley, T., Houchens, C., Dixon, D. M., Kurilla, M. G., Aurigemma, R., & Larsen, J. (2016). Accelerating global innovation to address antibacterial resistance: introducing CARB-X. *Nature Reviews Drug Discovery 2016 15:9*, *15*(9), 589–590. https://doi.org/10.1038/nrd.2016.155

Page, A. J., Bastkowski, S., Muhammad, Y., Keith, A. T., Le Vietid, T., Savvaid, G. M., Webber, M. A., & Charlesid, I. G. (2020). *AlbaTraDIS: Comparative analysis of large datasets from parallel transposon mutagenesis experiments*. https://doi.org/10.1371/journal.pcbi.1007980

Paris, Ü., Mikkel, K., Tavita, K., Saumaa, S., Teras, R., & Kivisaar, M. (2015). NHEJ enzymes LigD and Ku participate in stationary-phase mutagenesis in Pseudomonas putida. *DNA Repair*, *31*, 11–18. https://doi.org/10.1016/J.DNAREP.2015.04.005

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043. https://doi.org/10.1101/GR.186072.114

Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clinical Microbiology Reviews*, *31*(4). https://doi.org/10.1128/CMR.00088-17/ASSET/808BB10D-7E85-4717-AE02-F26B28D2DBE2/ASSETS/GRAPHIC/ZCM0041826370008.JPEG

Paulsen, I. T., Cain, A. K., & Hassan, K. A. (2017). Physical enrichment of transposon mutants from saturation mutant libraries using the TraDISort approach. *Mobile Genetic Elements*, *7*(3), 1–7. https://doi.org/10.1080/2159256X.2017.1313805

Payne, D. J., Gwynn, M. N., Holmes, D. J., & Pompliano, D. L. (2007). Drugs for bad bugs: Confronting the challenges of antibacterial discovery. In *Nature Reviews Drug Discovery* (Vol. 6, Issue 1, pp. 29–40). Nature Publishing Group. https://doi.org/10.1038/nrd2201

Peraman, R., Sure, S. K., Dusthackeer, V. N. A., Chilamakuru, N. B., Yiragamreddy, P. R., Pokuri, C., Kutagulla, V. K., & Chinni, S. (2021). Insights on recent approaches in drug discovery strategies and untapped drug targets against drug resistance. *Future Journal of Pharmaceutical Sciences*, *7*(1), 1–25. https://doi.org/10.1186/s43094-021-00196-5

Perry, B. J., Akter, M. S., & Yost, C. K. (2016). The use of transposon insertion sequencing to interrogate the core functional genome of the legume symbiont rhizobium leguminosarum. *Frontiers in Microbiology*, *7*(NOV), 1873. https://doi.org/10.3389/FMICB.2016.01873/BIBTEX

Phan, M. D., Peters, K. M., Sarkar, S., Lukowski, S. W., Allsopp, L. P., Moriel, D. G., Achard, M. E. S., Totsika, M., Marshall, V. M., Upton, M., Beatson, S. A., & Schembri, M. A. (2013). The Serum Resistome of a Globally Disseminated Multidrug Resistant Uropathogenic Escherichia coli Clone. *PLoS Genetics*, *9*(10). https://doi.org/10.1371/journal.pgen.1003834

Phillips, P. C. (2008). Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. In *Nature Reviews Genetics* (Vol. 9, Issue 11, pp. 855–867). Nature Publishing Group. https://doi.org/10.1038/nrg2452

Pletz, M. W., Hagel, S., & Forstner, C. (2017). Who benefits from antimicrobial combination therapy? *The Lancet Infectious Diseases*, *17*(7), 677–678. https://doi.org/10.1016/S1473-3099(17)30233-5

Ponsting, H., & Ning, Z. (2012). SMALT - A New Mapper for DNA Sequencing Reads. *F1000Posters*, *3*, 327. https://f1000research.com/posters/327%0Ahttp://cdn.f1000.com/posters/docs/327

Pribat, A., Blaby, I. K., Lara-Núñez, A., Gregory, J. F., de Crécy-Lagard, V., & Hanson, A. D. (2010). FolX and FolM are essential for tetrahydromonapterin synthesis in Escherichia coli and Pseudomonas aeruginosa. *Journal of Bacteriology*, *192*(2), 475–482. https://doi.org/10.1128/JB.01198-09

Pritchard, J. R., Chao, M. C., Abel, S., Davis, B. M., Baranowski, C., Zhang, Y. J., Rubin, E. J., & Waldor, M. K. (2014). ARTIST: High-Resolution Genome-Wide Assessment of Fitness Using Transposon-Insertion Sequencing. *PLoS Genetics*, *10*(11), e1004782. https://doi.org/10.1371/journal.pgen.1004782

Pundir, S., Martin, M. J., & O'Donovan, C. (2016). UniProt Tools. *Current Protocols in Bioinformatics*, *53*(1), 1.29.1-1.29.15. https://doi.org/10.1002/0471250953.BI0129S53

R Core Team. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* https://www.R-project.org/

Renwick, M. J., Brogan, D. M., & Mossialos, E. (2015). A systematic review and critical assessment of incentive strategies for discovery and development of novel antibiotics. *The Journal of Antibiotics 2016 69:2*, *69*(2), 73–88. https://doi.org/10.1038/ja.2015.98

Reznikoff, W. S. (2003). Tn5 as a model for understanding DNA transposition. *Molecular Microbiology*, *47*(5), 1199–1206. https://doi.org/10.1046/j.1365-2958.2003.03382.x

Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K. B., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T., Mori, H., Perna, N. T., Plunkett, G., Rudd, K. E., Serres, M. H., Thomas, G. H., Thomson, N. R., Wishart, D., & Wanner, B. L. (2006). Escherichia coli K-12: A cooperatively developed annotation snapshot - 2005. *Nucleic Acids Research*, *34*(1), 1–9. https://doi.org/10.1093/nar/gkj405

Ross, J. A., Trussler, R. S., Black, M. D., McLellan, C. R., & Haniford, D. B. (2014). Tn5 transposition in Escherichia coli is repressed by Hfq and activated by over-expression

of the small non-coding RNA SgrS. *Mobile DNA*, *5*(1), 27. https://doi.org/10.1186/s13100-014-0027-z

Saebelfeld, M., Das, S. G., Brink, J., Hagenbeek, A., Krug, J., & de Visser, J. A. G. M. (2021). Antibiotic Breakdown by Susceptible Bacteria Enhances the Establishment of β-Lactam Resistant Mutants. *Frontiers in Microbiology*, *12*, 2397. https://doi.org/10.3389/FMICB.2021.698970/BIBTEX

Safdar, N., Handelsman, J., & Maki, D. G. (2004). Does combination antimicrobial therapy reduce mortality in Gram-negative bacteraemia? A meta-analysis. *Lancet Infectious Diseases*, *4*(8), 519–527. https://doi.org/10.1016/S1473-3099(04)01108-9

Salama, N. R., Shepherd, B., & Falkow, S. (2004). Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. *Journal of Bacteriology*, *186*(23), 7926–7935. https://doi.org/10.1128/JB.186.23.7926-7935.2004

Santillán, M., & Mackey, M. C. (2004). Influence of Catabolite Repression and Inducer Exclusion on the Bistable Behavior of the lac Operon. *Biophysical Journal*, *86*(3), 1282–1292. https://doi.org/10.1016/S0006-3495(04)74202-2

Sargison, F. A., & Fitzgerald, J. R. (2021). Advances in Transposon Mutagenesis of Staphylococcus aureus: Insights into Pathogenesis and Antimicrobial Resistance. *Trends in Microbiology*, *29*(4), 282–285. https://doi.org/10.1016/J.TIM.2020.11.003

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2001). Comprehensive identification of conditionally essential genes in mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(22), 12712–12717. https://doi.org/10.1073/pnas.231275498

Sciarretta, K., Røttingen, J. A., Opalska, A., van Hengel, A. J., & Larsen, J. (2016). Economic Incentives for Antibacterial Drug Development: Literature Review and Considerations From the Transatlantic Task Force on Antimicrobial Resistance. *Clinical Infectious Diseases*, *63*(11), 1470–1474. https://doi.org/10.1093/CID/CIW593

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/BIOINFORMATICS/BTU153

Senior, A. E. (1990). The proton-translocating ATPase of Escherichia coli. *Annual Review of Biophysics and Biophysical Chemistry*, *19*, 7–41. https://doi.org/10.1146/ANNUREV.BB.19.060190.000255

Sharma, D., Misba, L., & Khan, A. U. (2019). Antibiotics versus biofilm: an emerging battleground in microbial communities. *Antimicrobial Resistance & Infection Control 2019 8:1*, *8*(1), 1–10. https://doi.org/10.1186/S13756-019-0533-3

Shevchenko, Y., Bouffard, G. G., Butterfield, Y. S. N., Blakesley, R. W., Hartley, J. L., Young, A. C., Marra, M. A., Jones, S. J. M., Touchman, J. W., & Green, E. D. (2002). Systematic sequencing of cDNA clones using the transposon Tn5. *Nucleic Acids Research*, *30*(11), 2469–2477. http://www.phrap.org

Shields, R. C., & Jensen, P. A. (2019). The Bare Necessities: Uncovering Essential and Condition-Critical Genes with Transposon Sequencing. *Molecular Oral Microbiology*. https://doi.org/10.1111/omi.12256

Silver, L. L. (2016). *Appropriate Targets for Antibacterial Drugs*. https://doi.org/10.1101/cshperspect.a030239

Singh, R., Sahore, S., Kaur, P., Rani, A., & Ray, P. (2016). Penetration barrier contributes to bacterial biofilm-associated resistance against only select antibiotics, and exhibits genus-, strain- and antibiotic-specific differences. *Pathogens and Disease*, *74*(6), 56. https://doi.org/10.1093/FEMSPD/FTW056

Sivakumar, R., Ranjani, J., Vishnu, U. S., Jayashree, S., Lozano, G. L., Miles, J., Broderick, N. A., Guan, C., Gunasekaran, P., Handelsman, J., & Rajendhran, J. (2019). Evaluation of InSeq To Identify Genes Essential for Pseudomonas aeruginosa PGPR2 Corn Root Colonization . *G3: Genes|Genomes|Genetics*, *9*(March), g3.200928.2018. https://doi.org/10.1534/g3.118.200928

Solaimanpour, S., Sarmiento, F., & Mrázek, J. (2015). Tn-seq explorer: A tool for analysis of high-throughput sequencing data of transposon mutant libraries. *PLoS ONE*, *10*(5), e0126070. https://doi.org/10.1371/journal.pone.0126070

Subashchandrabose, S., Smith, S. N., Spurbeck, R. R., Kole, M. M., & Mobley, H. L. T. (2013). Genome-Wide Detection of Fitness Genes in Uropathogenic Escherichia coli during Systemic Infection. *PLoS Pathog*, *9*(12), 1003788. https://doi.org/10.1371/journal.ppat.1003788

Tellier, M., & Chalmers, R. (2020). Compensating for over-production inhibition of the Hsmar1 transposon in Escherichia coli using a series of constitutive promoters. *Mobile DNA*, *11*(1), 1–14. https://doi.org/10.1186/S13100-020-0200-5/TABLES/4

Thibault, D., Wood, S., Jensen, P., & Opijnen, T. van. (2018). droplet-Tn-Seq combines microfluidics with Tn-Seq identifying complex single-cell phenotypes. *BioRxiv*, 391045. https://doi.org/10.1101/391045

Thoma, S., & Schobert, M. (2009). *An improved Escherichia coli donor strain for diparental mating*. https://doi.org/10.1111/j.1574-6968.2009.01556.x

Tillotson, G. S., & Blondeau, J. M. (2014). Antimicrobial development and the risk–benefit assessment: recent adverse events and their implications. *Http://Dx.Doi.Org/10.1586/14787210.4.4.515*, *4*(4), 515–517. https://doi.org/10.1586/14787210.4.4.515

Typas, A., Nichols, R. J., Siegele, D. A., Shales, M., Collins, S. R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B. L., Mori, H., Weissman, J. S., Krogan, N. J., & Gross, C. A. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. *Nature Methods*, *5*(9), 781–787. https://doi.org/10.1038/nmeth.1240

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, *6*(10), 767–772. https://doi.org/10.1038/nmeth.1377

van Opijnen, T., & Camilli, A. (2013). Transposon insertion sequencing: A new tool for systems-level analysis of microorganisms. *Nature Reviews Microbiology*, *11*(7), 435–442. https://doi.org/10.1038/nrmicro3033

van Opijnen, T., Lazinski, D. W., & Camilli, A. (2015). *Genome-Wide Fitness and Genetic Interactions Determined by Tn-seq, a High-Throughput Massively Parallel Sequencing Method for Microorganisms*. https://doi.org/10.1002/9780471729259.mc01e03s36

Verma, S. C., Qian, Z., & Adhya, S. L. (2019). Architecture of the Escherichia coli nucleoid. *PLoS Genetics*, *15*(12). https://doi.org/10.1371/JOURNAL.PGEN.1008456

Vestheim, H., & Jarman, S. N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples – a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, *5*, 12. https://doi.org/10.1186/1742-9994-5-12

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/JOURNAL.PONE.0112963

Warr, A. R., Hubbard, T. P., Munera, D., Blondel, C. J., zur Wiesch, P. A., Abel, S., Wang, X., Davis, B. M., & Waldor, M. K. (2019). Transposon-insertion sequencing screens unveil requirements for EHEC growth and intestinal colonization. *PLOS Pathogens*, *15*(8), e1007652. https://doi.org/10.1371/JOURNAL.PPAT.1007652

Weerdenburg, E. M., Abdallah, A. M., Rangkuti, F., el Ghany, M. A., Otto, T. D., Adroub, S. A., Molenaar, D., Ummels, R., ter Veen, K., van Stempvoort, G., van der Sar, A. M., Ali, S., Langridge, G. C., Thomson, N. R., Pain, A., & Bitter, W. (2015). Genome-wide transposon mutagenesis indicates that Mycobacterium marinum customizes its virulence mechanisms for survival and replication in different hosts. *Infection and Immunity*, *83*(5), 1778–1788. https://doi.org/10.1128/IAI.03050-14

Weinreich, M. D., Yigit, H., & Reznikoff, W. S. (1994). Overexpression of the Tn5 transposase in Escherichia coli results in filamentation, aberrant nucleoid segregation, and cell death: analysis of E. coli and transposase suppressor mutations. *Journal of Bacteriology*, *176*(17), 5494–5504. https://doi.org/10.1128/JB.176.17.5494-5504.1994

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio*, *6*(3), e00306-15. https://doi.org/10.1128/mBio.00306-15

Wickham, H. (2016). ggpolt2 Elegant Graphics for Data Analysis. In *Use R! series* (p. 211). Springer.

Wiles, T. J., Norton, J. P., Russell, C. W., Dalley, B. K., Fischer, K. F., & Mulvey, M. A. (2013). Combining Quantitative Genetic Footprinting and Trait Enrichment Analysis to Identify Fitness Determinants of a Bacterial Pathogen. *PLoS Genetics*, *9*(8), e1003716. https://doi.org/10.1371/journal.pgen.1003716

Wilkins, D. (2020). *Draw Gene Arrow Maps in "ggplot2" [R package gggenes version 0.4.1]*. https://CRAN.R-project.org/package=gggenes

World Health Organisation (WHO). (2021). 2021 Antibacterial Agents in Clinical and Preclinical Development: an overview and analysis. *World Health Organization 2021*, *August*, 76. https://www.who.int/publications/i/item/9789240021303

Wu, Y., Aandahl, R. Z., & Tanaka, M. M. (2015). Dynamics of bacterial insertion sequences: can transposition bursts help the elements persist? *BMC Evolutionary Biology*, *15*(1), 288. https://doi.org/10.1186/S12862-015-0560-5

Xu, H., Yiğiter, A., & Chen, J. (2022). onlineBcp: An R package for online change point detection using a Bayesian approach. *SoftwareX*, *17*, 100999. https://doi.org/10.1016/J.SOFTX.2022.100999

Xu, Z., Wang, Y., Chater, K. F., Ou, H. Y., Xu, H. H., Deng, Z., & Tao, M. (2017). Large-Scale Transposition Mutagenesis of Streptomyces coelicolor Identifies Hundreds of Genes Influencing Antibiotic Biosynthesis. *Applied and Environmental Microbiology*, *83*(6). https://doi.org/10.1128/AEM.02889-16

Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., Datsenko, K. A., Nakayashiki, T., Tomita, M., Wanner, B. L., & Mori, H. (2009). Update on the Keio collection of Escherichia coli single-gene deletion mutants. *Molecular Systems Biology*, *5*(1), 335. https://doi.org/10.1038/MSB.2009.92

Yamazaki, Y., Niki, H., & Kato, J. (2008). *Profiling of Escherichia coli Chromosome Database* (pp. 385–389). Humana Press. https://doi.org/10.1007/978-1-59745-321-9_26

Yasir, M., Keith Turner, A., Bastkowski, S., Baker, D., Page, A. J., Telatin, A., Phan, M. D., Monahan, L., Savva, G. M., Darling, A., Webber, M. A., & Charles, I. G. (2020). TRADIS-XPress: A high-resolution whole-genome assay identifies novel mechanisms of triclosan action and resistance. *Genome Research*, *30*(2), 239–249. https://doi.org/10.1101/gr.254391.119

Yasir, M., Turner, A. K., Lott, M., Rudder, S., Baker, D., Bastkowski, S., Page, A. J., Webber, M. A., & Charles, I. G. (2022). Long-read sequencing for identification of insertion sites in large transposon mutant libraries. *Scientific Reports 2022 12:1*, *12*(1), 1–9. https://doi.org/10.1038/s41598-022-07557-x

Yu, X. C., Tran, A. H., Sun, Q., & Margolin, W. (1998). Localization of cell division protein FtsK to the Escherichia coli septum and identification of a potential N-Terminal targeting domain. *Journal of Bacteriology*, *180*(5), 1296–1304. https://doi.org/10.1128/JB.180.5.1296-1304.1998/ASSET/EBEE8701-1812-4335-92C4-C108C9D2EE51/ASSETS/GRAPHIC/JB0581337006.JPEG

Zhang, R., Ou, H. Y., & Zhang, C. T. (2004). DEG: A database of essential genes. *Nucleic Acids Research*, *32*(DATABASE ISS.), D271. https://doi.org/10.1093/nar/gkh024

Zhang, T., Zhou, J., Gao, W., Jia, Y., Wei, Y., & Wang, G. (2022). Complex genome assembly based on long-read sequencing. *Briefings in Bioinformatics*, *23*(5), 1–11. https://doi.org/10.1093/BIB/BBAC305

Zhang, Y., Zhao, D., & Liu, J. (2014). *The Application of Baum-Welch Algorithm in Multistep Attack*. https://doi.org/10.1155/2014/374260

Zhao, L., Anderson, M. T., Wu, W., T. Mobley, H. L., & Bachman, M. A. (2017). TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. *BMC Bioinformatics*, *18*(1), 326. https://doi.org/10.1186/s12859-017-1745-2

Zhu, J., Gong, R., Zhu, Q., He, Q., Xu, N., Xu, Y., Cai, M., Zhou, X., Zhang, Y., & Zhou, M. (2018). Genome-Wide Determination of Gene Essentiality by Transposon Insertion Sequencing in Yeast Pichia pastoris. *Scientific Reports 2018 8:1*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-28217-z

Zomer, A., Burghout, P., Bootsma, H. J., Hermans, P. W. M., & van Hijum, S. A. F. T. (2012). ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PloS One*, *7*(8), e43012. https://doi.org/10.1371/journal.pone.0043012

9 Appendices

9.1 Hybrid reference FASTA sequence file.

File Name: hybrid_reference.fasta

Electronic only.

9.2 Hybrid reference annotation EMBL sequence file.

File Name: hybrid_reference.embl

Electronic only.

9.3 Essential gene determination using the BioTradis pipeline.

Tables of Essential and ambiguous genes reported from analyses through the BioTradis pipeline. Gene lists are for all of the individual and concatenated libraries and the EM Tn*5* analysis. Electronic only.

File Name: Gene Lists.xlsx

Electronic only.

9.4 The Python scripted functions used in this work.

File Name: Tn_analysis.ipynb

```
Set Working Directory

import os
os.chdir('E:\Code test')
os.getcwd()

Define "Insert File Process" function

def insert_file_process():
    from audioop import avg
    import pandas as pd
    import glob
    import gzip
    import shutil
    from Bio import SeqIO
    fasta = SeqIO.read("hybrid_reference.fasta", "fasta")
    from Bio.Seq import Seq
    from Bio import SeqUtils
    files = [f for f in glob.glob("*plot.gz")]
    print(files)
    for m in range(len(files)):
        with gzip.open(files[m], 'rb') as f_in:
            with open(files[m]+".txt", 'wb') as f_out:
                shutil.copyfileobj(f_in, f_out)
    seq = fasta.seq
    print("Reference length:")
    print(len(seq))
    input = [f for f in glob.glob("Stats.txt")]
    print(input)
    df =  pd.read_csv(
        "Stats.txt",
        sep = ',',
        engine = 'python')
    file_list = list(df["File"])
    reads_mapped = df["Reads Mapped"]
    unique_ins = df["Total Unique Insertion Sites"]
    print(file_list)
    for i in range(len(file_list)):
        print(file_list[i])
        unique = str(unique_ins[i])
        mapped = str(reads_mapped[i])
        avg = int(mapped) / int(unique)
        x = 0.1 * avg
```

```
        y = round (x)
        print("10% of average =")
        print(x)
        print ("rounded to ")
        print(y)
        df_2 =  pd.read_csv(
            file_list[i]+".insert_site_plot.gz.txt",
            sep = ' ',
            engine = 'python',
            header = None,
            index_col = False,
            names= [
                "F","R"])
        forward = df_2["F"]
        reverse = df_2["R"]
        col_1 = (forward)
        col_2 = (reverse)
        combined = df_2['F'] + df_2['R']
        print(combined)
        ins_count =
open("Unfiltered_Insertion_total_"+str(file_list[i])+".txt", 'a')
        for j in range (len(seq)):
            ins_count.write(str(combined[j]))
            ins_count.write(" ")
        ins_count.close()
        site = open("Filtered_Insert_plot"+str(file_list[i])+".txt",
'a')
        for k in range (len(seq)):
            if combined[k] >= y:
                site.write(str(forward[k]) + ' ')
                site.write(str(reverse[k]) + '\n')
            else:
                site.write('0 0' + '\n')
        site.close()
        ins_count_filt =
open("Filtered_Insertion_total"+str(file_list[i])+".txt", 'a')
        for l in range (len(seq)):
            if combined[l] >= y:
                ins_count_filt.write(str(combined[l])+ " ")
            else:
                ins_count_filt.write("0 ")
        ins_count_filt.close()
        print(file_list[i]+" Done")
    print("Finished")

Run the insert_file_process function
```

```
Build the "Tn_analysis_test" function. Can be modified for any
dinucleotide (or other)

def Tn_analysis_test():
    print("Importing Files")
    from Bio import SeqIO
    fasta = SeqIO.read("hybrid_reference.fasta", "fasta")
    from Bio.Seq import Seq
    from Bio import SeqUtils
    seq = fasta.seq
    print("Reference length:")
    print(len(seq))
    start = seq[0:12]
    end = seq[(len(seq)-12):len(seq)]
    sequence = start + seq + end
    pattern = Seq("TA")
    TA_list = SeqUtils.nt_search(str(seq),pattern)
    TA_sites = TA_list[1:]
    print("Number of TA sites:")
    print(len(TA_sites))
    import glob
    filelist = [f for f in glob.glob("*.txt")]
    print("Files:")
    print(filelist)
    print("Thinking Counts")
    for i in range(len(filelist)):
        print(filelist[i])
        ins_string = open(filelist[i], 'r').read()
        ins_str_list = list(ins_string.split(" "))
        ins_list = list(map(str,ins_str_list))
        ta = open("TA_ins_"+str(filelist[i]), 'w')
        for k in range(len(TA_sites)):
            l = int(k)
            m = (TA_sites[l])
            n = str(ins_list[m])
            ta.write(n)
            ta.write('\n')
        ta.close()
    print("Thinking Site")
    for i in range(len(filelist)):
        print(filelist[i])
        ins_string = open(filelist[i], 'r').read()
        ins_str_list = list(ins_string.split(" "))
        ins_list = list(map(str,ins_str_list))
        ins_seq = open("site_seq_"+str(filelist[i]), 'w')
        for b in range(len(seq)):
            if int(ins_list[b]) >= 1:
                for j in range(26):
```

```
                    k = int(b) + int(j)
                    ins_seq.write(str(b))
                    ins_seq.write("> ")
                    ins_seq.write(sequence[k]+ " ")
                    ins_seq.write('\n')
        else:
            pass
    ins_seq.close()
print("Done Site :)")
print ("Thinking Fasta")
for i in range(len(filelist)):
    print(filelist[i])
    ins_string = open(filelist[i], 'r').read()
    ins_str_list = list(ins_string.split(" "))
    ins_list = list(map(str,ins_str_list))
    fasta = open("site_seq_"+str(filelist[i])+".fasta", 'w')
    for b in range(len(seq)):
        if int(ins_list[b]) >= 1:
            for j in range(26):
                    k = int(b) + int(j)
                    fasta.write(str(b))
                    fasta.write(">")
                    fasta.write(str(b))
                    fasta.write('\n')
                    fasta.write(sequence[k]+ " ")
                    fasta.write('\n')
            fasta.write(">")
            fasta.write(str(b))
            fasta.write('\n')
        else:
            pass
    fasta.close()
print("Done Fasta :)")
print ("Thinking first base")
for i in range(len(filelist)):
    print(filelist[i])
    ins_string = open(filelist[i], 'r').read()
    ins_str_list = list(ins_string.split(' '))
    ins_list = list(map(str,ins_str_list))
    first_bp = open("first_base_"+str(filelist[i]), 'w')
    for b in range(len(seq)):
        v = int(b+1)
        if int(ins_list[b]) >= 1 :
            first_bp.write(str(v))
            first_bp.write(", " +seq[b])
            first_bp.write('\n')
        else:
            pass
```

```
        first_bp.close()
    print("Done first base :)")
    print("Combining FASTA")
    ta = open("All_sites.fasta", 'w')
    ta.close()
    for p in range(len(filelist)):
        fasta_string = open("site_seq_"+str(filelist[p])+".fasta",
'r').read()
        fasta_str_list = list(fasta_string.split('\n'))
        ta = open("All_sites.fasta", 'a')
        for q in range(len(fasta_str_list)):
            ta.write(fasta_str_list[q])
            ta.write('\n')
        ta.close()
    print ("Done Combined :)")
    print("Thinking Zero Sites")
    for p in range(len(filelist)):
        zero = open("TA_zero"+str(filelist[p])+".fasta", 'w')
        print(filelist[p])
        ins_string = open(filelist[p], 'r').read()
        ins_str_list = list(ins_string.split(' '))
        ins_list = list(map(str,ins_str_list))
        for i in range(len(TA_sites)):
            n = (TA_sites[i])
            m = int(ins_list[n])
            if m == 0:
                zero.write(">")
                zero.write(str(n))
                zero.write('\n')
                for j in range(26):
                    k = n + int(j)
                    zero.write(str(sequence[k]))
                zero.write('\n')
                zero.write('\n')
            else:
                pass
        zero.close()

    print("Done Zero Sites")
    print("Finished")

Run the Tn_analysis_test function
```

9.5 The R scripted functions used in this work

File Name: R functions.R

9.5.1 Import Insertion Count Files

```
setwd ("U:\\Insert plots final seq run\\txt files\\Tn5 Unfiltered")


library(data.table)
library(seqinr)
library(ggplot2)
library(dplyr)


BW25113  <- read.fasta("hybrid_reference.fasta")
BW25113seq <- BW25113[[1]]


inserts2 <- rbindlist(lapply(1:11, function(i){
  filename <- sprintf("Tn5_%d_Both_Trim.fastq.insert_site_plot.gz.t
xt", i)
  bw <- fread(file = filename)
  bw[, suminserts := V1 + V2]
  bw[, pos:=seq_along(V1)]
  bw[, biorep := factor(1+((i-1)%/%3)) ]
  bw[, techrep := factor(1+((i-1)%%3)) ]
  bw[, any := suminserts>0]
  bw[, i:=i]
  bw[, bw:=BW25113seq ]
  bw
}))

## Rep 10 is the no growth library.  Here repeat this library into
a new column so you can standardise or condition on it.
inserts2[, NG:=rep(inserts2[i==10]$suminserts,11)]
inserts2[, WGS:=rep(inserts2[i==11]$suminserts,11)]

## Make a sort of standardised count - adds 1 to *every* count to a
void dividing by zero.

inserts2[, suminsertsNG := (suminserts+1)/(NG+1)]
inserts2[, suminsertsWGS := (suminserts+1)/(WGS+1)]

## Inserts3 is needed for Segmentor
inserts_3<- inserts2[i %in% 1:9 , .(V1=sum(V1), V2=sum(V2), suminse
rts=sum(suminserts)),
                     by=.(pos, NG, bw, totalinserts)]

genes <- read.csv("annotations.csv")
genes$pos_m<- genes$Start
head(genes)
```

```
tab<-merge(inserts3, genes, all.x = TRUE)
head(tab)

inserts3<-tab
```

9.5.2 TraDIS Viewer

```
## Says "if inserts2 object doesn't exist then run the source file
to make it".
if(!exists("inserts2")) source("loadTn5.R")


### This is the function that generates the views of the tradis dat
asets.

tradisViewer <- function(data=inserts2,   # What is the name of the
dataset?  Default is 'inserts2'
                         area=FALSE,      # Do you want the area un
der the curve filled in?
                         sequence=FALSE,  # Do you want the sequenc
e added to the final facet?
                         xstart=0,        # What is the start posit
ion (BP)
                         scales="fixed",  # Do you want to allow th
e y axis scale to vary between facets?
                         length=1000,     # How long do you want th
e sequence to be? (BP)
                         logscale=TRUE,   # Log scale for the y-axi
s?
                         line=FALSE,      # Join the points with a
line?
                         xstop=xstart+length,  # Instead of specify
ing the sequence length, you can specify the end poistion if you li
ke
                         reps=1,          # Which replicates do you
want to plot?
                         threshold=0,     # Only plot positions wit
h at least this many mutations.
                         pch=".",         # Which symbol to use for
the points?  "." is a single pixel, use this if there are many poin
ts to plot otherwise things get very busy and very slow.
                         standardiseNG=FALSE, # Should 'standardise
' the counts against the NG library?
                         standardiseWGS=FALSE,
                         ori=FALSE
                         # Should 'standardise' the counts against
the WGS library?
                         ){
```

```
  replabels=c("Tn5","Mariner", "WGS")
  names(replabels) <- as.character(1:3)
  if(standardiseNG) {target="suminsertsNG"}
  else {if(standardiseWGS) {target="suminsertsWGS"}
  else {target="suminserts"}}
  g = ggplot(data[pos<xstop & pos>xstart & i %in% reps & suminserts
>=threshold], aes(x=pos,y=get(target))) +
    theme_bw() +
    facet_grid(rows=vars(i),cols=NULL, labeller=labeller(i=replabel
s), scales=scales)

  g = g + geom_point(pch=pch) + labs(y="Insert count",x="position")
+
    if(sequence==TRUE) g = g + geom_text(data=inserts2[i==11&pos<xs
top & pos>xstart & i %in% reps ],aes(label=toupper(bw), y=11))
    if(area==TRUE)     g = g + geom_area()
    if(logscale==TRUE) g = g + scale_y_log10()
    if(line==TRUE)     g = g + geom_line()
    if(ori==TRUE)      g = g + geom_vline(xintercept = 286056, line
type = 4, color = "red", size= 1.5)
}

library(stringr)
library(grid)
library(gridExtra)
library(ggplot2)


### Example of using the viewer:

i<- 0
j<-1e5

k<- (i+j)

l<- paste(i, "to ", k, "; no standardisation" )
m<- paste(i, "to ", k, "; standardised to NG" )
n<- paste(i, "to ", k, "; standardised to WGS" )



my_title_1 <- (l)
plot_1<-tradisViewer(reps=1:11, xstart= i,length = j, threshold=1)
+
  labs(title = str_wrap(my_title_1, 60))
```

9.5.3 Segmentor3 Model

```
# Try for the first 100K reads.
# 300 max breakpoints.

### Had to get this from github as its not for this version.
### Seems a bit buggy!  (tends to crash R when used on a very big s
equence)

library(Segmentor3IsBack)
library(ggnewscale)
library(remotes)

## Try it for first 100,000 bp.

segmentPlot <- function(dat){

  ## This functions finds breakpoints with 4 different objects.

  ## x is the sum of all inserts across replicates (should already
exist in the data)
  x <-  dat$suminserts_m
  ## This is the number of replicates with at least one insert
  xT <- dat$totalinserts_m
  ## This one is x truncated at 100 (to avoid problematic distribut
ions)
  x100 <- pmin(x,100)
  ## 'any' insert at the site (for binomial distribution)
  x0 <- as.numeric(x>0)

  ### Now set up segmentor using this vector.
  Seg0<-Segmentor(x0,model=1,Kmax=300)
  SegT<-Segmentor(xT,model=3,Kmax=300)
  Seg100<-Segmentor(x100,model=3,Kmax=300)

  ## Select the number of breakpoints?
  Kchoose100<-SelectModel(Seg100, penalty="oracle")
  KchooseT<-SelectModel(SegT, penalty="oracle")
  Kchoose0<-SelectModel(Seg0, penalty="oracle")

  ## Extract the breakpoints
  v0=data.frame(pos=getBreaks(Seg0)[Kchoose0, 1:Kchoose0],breaks0="
v0")
  vT=data.frame(pos=getBreaks(SegT)[KchooseT, 1:KchooseT],breaksT="
vT")
  v100=data.frame(pos=getBreaks(Seg100)[Kchoose100, 1:Kchoose100],
breaks100="v100")

  # Make a copy of the input for plotting
  inserts4 <- dat
  inserts4$pos <- seq_along(inserts4$pos)

  ## Merge the breakpoints into the dataset.
  inserts4 <- merge(inserts4, v0, all.x=TRUE)
```

```
  inserts4 <- merge(inserts4, v100, all.x=TRUE)
  inserts4 <- merge(inserts4, vT, all.x=TRUE)

  ## Now we need to find the average rate of the mutations within e
ach of the breakpoints.

  ## First make a variable corresponding to which group (cumsum inc
rements every time it sees a breakpoint)
  inserts4[, group0 := cumsum(!is.na(breaks0))]
  inserts4[, group100 := cumsum(!is.na(breaks100))]
  inserts4[, groupT := cumsum(!is.na(breaksT))]

  ## Then find the average by group.
  inserts4[, average0 := mean(suminserts_m>0), by=group0]
  inserts4[, average100 := log(mean(suminserts_m)), by=group100]
  inserts4[, averageNG := log(mean(NG_m)), by=group100]
  inserts4[, averageT := log(mean(suminserts_m)), by=groupT]

  ## Make a palette for colour plotting
  hex <- c("#FF3333", "#FFA500", "#FFFF00", "#008000", "#9999ff", "
#000066")
  bw<- c("black", "grey", "white")
  labs<-c("white", "black", "black")

  ## Now graph (first 1e5 points)
  g <- ggplot(inserts4[1:1e5]) + aes(x=pos, y=suminserts_m) +
    scale_fill_gradientn(colours=hex)+
    geom_line(col="grey") +
    geom_raster(aes(y=1,fill=average100))+
    geom_point(pch=19,size=0.1) +
    labs(fill='Average Insertions')+
    new_scale_fill()+
    scale_fill_manual(values=bw, na.translate=F)+
    scale_colour_manual(values=labs)+
    geom_gene_arrow(aes(xmin = as.numeric(Start), xmax = as.numeric
(End), y = 1, fill=Type))+
    geom_gene_label(aes(xmin = as.numeric(Start), xmax = as.numeric
(End), y = 1 , label = Name, colour = Type))+
    labs(fill='Feature') +
    facet_wrap(~cut(pos, breaks=5), scale="free", ncol=1) + scale_y
_log10() +
    geom_vline(aes(xintercept=ifelse(breaks100=="v100",pos,NA))) +
    theme_bw() +
    theme(strip.background = element_blank(),
        strip.text.x = element_blank()) +
    ylab("Total Insertions") +
    xlab("Genome Location (bp)")


  ## Return a list of the annotated sequence and the graph.
  print(g)

}
```

```
### Now we can run the file.
gSP <- segmentPlot(inserts3[1:1e5])
```

9.5.4 OnlineBcp Model

```
library(onlineBcp)

x<-(inserts3$suminserts[1:1e5])
x100 <- pmin(x,100)
x0 <- as.numeric(x>0)

breakx<-online_cp(x, debug=TRUE)
break100<- online_cp(x100, debug=TRUE)
break0<-online_cp(x0, debug=TRUE)

sumx<-summary(breakx)
sum100<-summary(break100)
sum0<-summary(break0)

dfx<-as.data.frame(sumx$result$segment)
df100<-as.data.frame(sum100$result$segment)
df0<-as.data.frame(sum0$result$segment)

head(dfx)
head(df100)
head(df0)

cgx<- data.frame(pos=dfx$end, meanx=log(dfx$mean), chgx="break")
cg100<- data.frame(pos=df100$end, mean100=log(df100$mean), chg100="
break")
cg0<- data.frame(pos=df0$end, mean0=log(df0$mean), chg0="break")

head(cgx)

data<-merge(inserts3[1:1e5], cgx, all.x = TRUE)
data<-merge(data, cg100, all.x = TRUE)
data<-merge(data, cg0, all.x = TRUE)

head(data)

hex <- c("#FF3333", "#FFA500", "#FFFF00", "#008000", "#9999ff", "#0
00066")
bw<- c("black", "grey", "white")
labs<-c("white", "black", "black")


g <- ggplot(data) + aes(x=pos, y=suminserts) +
  scale_fill_gradientn(colours=hex)+
  geom_line(col="grey") +
  geom_point(pch=19,size=0.1) +
  labs(fill='Average Insertions')+
  geom_vline(aes(xintercept=ifelse(chg100=="break",pos,NA))) +
  geom_raster(aes(y=1,fill=mean100))+
```

```
  new_scale_fill()+
  geom_gene_arrow(aes(xmin = as.numeric(Start), xmax = as.numeric(E
nd), y = 1, fill=Type))+
  geom_gene_label(aes(xmin = as.numeric(Start), xmax = as.numeric(E
nd), y = 1 , label = Name))+
  labs(fill='Feature') +
  facet_wrap(~cut(pos, breaks=5), scale="free", ncol=1) + scale_y_l
og10() +
  theme_bw() +
  theme(strip.background = element_blank(),
        strip.text.x = element_blank()) +
  ylab("Total Insertions") +
  xlab("Genome Location (bp)")

print(g)
```