The Development and Application of a Transposon Insertion Sequencing Methodology in *Escherichia coli* BW25113

by

Ashley Robinson

A thesis submitted to

The University of Birmingham

For the degree of

Doctor of Philosophy

College of Medical and Dental Sciences

September 2016

UNIVERSITY OF BIRMINGHAM

ABSTRACT

Escherichia coli is one of the most studied model organisms in biology. Even with decades of research, there are a substantial number of genes with an as yet unknown function. Previously, to determine the link between gene function and phenotype took significant experimental effort. However, newer methods are capable of providing large amounts of biological data in short timeframes. One such method, transposon insertion sequencing, is a powerful research tool, which couples transposon mutagenesis and next generation sequencing to identify genes that have important or essential functions.

Here, three transposon insertion sequencing methods were compared. The techniques were adapted from previously published literature. Based on a number of metrics one technique was shown to be superior for data generation. This method was chosen for application in further transposon-insertion sequencing experiments. Subsequently, the optimised method was used to assess which genes were essential for the viability of the model organism *E. coli* K12. The results of this work were compared with the literature and other databases of gene essentiality. A high degree of concordance was observed between our datasets and those generated previously through other methods. Indeed, the method described here was shown to have several benefits over previously used approaches. Finally, genes involved with maintenance of the outer membrane were identified by using markers for membrane permeability in tandem with the chosen method. In keeping with previous literature multiple genes involved with many aspects of the cell envelope were reported. Many of the reported genes were shown to be involved with metabolic processes related to the biogenesis and maintenance of the cell envelope.

ACKNOWLEDGEMENTS

This collection of ordered letters, by one monkey with one (admittedly modern) typewriter, would not have been completed without the friendship, guidance and support of so many people. I fear this will not do you justice, but here goes!

Magic Rainbow Five® - you have made my time in the lab incredible! I may never be a part of such a friendly, hilarious and supportive microcosm again, and for your friendship I am truly grateful. O Hai Chris – the laughs we have had! I am indebted to you for so much - your graffiti skills, taste in cinema and your incisive wit amongst other things. Thank you, and keep on searching for those triangles bro. Jacky B, Caddaby and Dairylea – your astounding knowledge of science, music and arabica alike always made for an entertaining discussion. Thank you for always being downstairs and available for caffeination, contemplation and celebration. James – my musical brother from an unrelated matriarch! Whenever I hear anything even remotely Latin in origin, I will fondly remember our improvs. Thank you so much for the music and your support. Pete – you have helped me for quite a few years now, and I am indebted – thank you so much for your ever helpful knowledge and insight, as well as Christmas festive cheer. Amanda, Faye, Josh, Sara, Rachael, Tamar, and likely many others that have fallen through the sieve between my ears – I have you all to thank for so much!

Ian and Cathy – you have been wonderful to me, and really made me feel like part of the lab family. It has been a privilege –I am eternally grateful for the opportunity you provided, and I hope that what I have done will always be of use. If you ever need computer help…

Finally, my beautiful family – Kayleigh, Erin, Niamh and Isla. You enrich my life with an unspeakable quality, and make my days joyous beyond description. Thank you for making life so worthwhile.

TABLE OF CONTENTS

List of Figures

Abbreviations

ATP – adenosine triphosphate

BAM complex – beta barrel assembly machinery

COG – cluster of orthologous groups

DNA – deoxyribonucleic acid

GTP – guanosine triphosphate

IM – Inner membrane of the bacterial cell envelope

IPTG - Isopropyl 1-thio-b-D-galactopyranoside

KEIO – reference to Baba *et al.* (2006) work, also Japanese era name after Genji and before Meiji

LB – Luria broth, also denotes growth experiments in Luria broth

LPS – lipopolysaccharide

NTL – neat transposon library

OM – outer membrane of the bacterial cell envelope

OMP – outer membrane protein

PCR – polymerase chain reaction

RNA – ribonucleic acid

S4.8 – SDS at 4.8%

SDS – sodium dodecyl sulphate

SRP – signal recognition particle

V100 – vancomycin at 100 µg per ml

CHAPTER 1

GENERAL INTRODUCTION

1.1 The model organism *Escherichia coli*

Throughout the history of scientific endeavour, a significant proportion of research effort has focused upon a relatively small number of organisms. These organisms were chosen for study for a variety of reasons, including ease of handling, genetic tractability and ethical considerations. As the knowledge base increases for a specific organism, this sets off a self-fulfilling cycle; as an increased amount of literature becomes available for a given organism, the more attractive the organism becomes for other researchers.

One such example of a model organism is *Escherichia coli*, which is one of, if not the most, widely studied bacterial species in history (Cronan, 2014). *E. coli* was first discovered and reported by Theodore Escherich, a microbiologist who studied the microbial component of the infant gut in 1885 (Blount, 2015; Escherich, 1988). Through the course of his research, he isolated this fast-growing rod-shaped bacterium that has, through the work of multiple scientists years later, become a mainstay of modern science. As testament to its impact, *E. coli* was the organism in which many biochemical pathways were elucidated and the fundamental workings of DNA were discovered along with the related processes of transcription, translation and replication. In addition to the more academic perspective, *E. coli* has been inextricably linked with industrial and pharmaceutical progress, due to its ease of genetic manipulation and metabolic versatility.

1.2 The cell envelope

In contrast to the wealth of information available today, decades ago the simple identification and differentiation of microbes presented a challenge. One early method of differentiating microbes centred upon the use of the Gram stain (Bartholomew and Mittwer, 1952), and this method of classification is still in use today. In practice, a microbial sample is

fixed onto a microscope slide, to which a stain or stains are added. Microbes are then generally described as either Gram negative or Gram positive. Gram-negative bacteria do not take up the deep purple crystal violet stain, whereas Gram-positive bacteria do.

The simple binary classification of either Gram-negative or Gram-positive bacteria belies a complex but fundamental structural difference between the cell envelopes of the two groups. Gram positive organisms possess a single membrane surrounded by a thick, externally facing layer of peptidoglycan (Lee and Schneewind, 2001). In contrast, Gram negative organisms are comprised of two membranes separated by the periplasm, an aqueous compartment that contains a layer of peptidoglycan, albeit much thinner than that found in Gram-positive bacteria (Silhavy, Kahne and Walker, 2010). Gram-positive organisms are stained much more deeply than Gram-negative bacteria, and this aids in their differentiation under a microscope.

It is important to note the evolutionary significance of the Gram-negative cell envelope as an adaptation to increase the chance of survival. The increase in fitness from possession of this structure is clearly demonstrated by phylogenetic analysis, which suggests the dominance of the dual membrane envelope structure across 17 of 24 bacterial phyla (Sutcliffe, 2010). However, this simple binary distinction between Gram-negative and -positive organisms is not truly representative of the variation in cellular envelope structure. This is hinted at by the existence of Gram-variable organisms belonging to members of the *Actinomyces* and *Clostridium* spp. (Beveridge, 1990). Discussion regarding the usage of these terms lies outside the scope of this introduction.

The Gram-negative cell envelope is a highly complex organelle with a great variety of tightly regulated proteins, enzymes and macromolecules (Fig. 1.1). In consideration of the whole cell, the envelope is of paramount importance. It is a crucial cellular component which

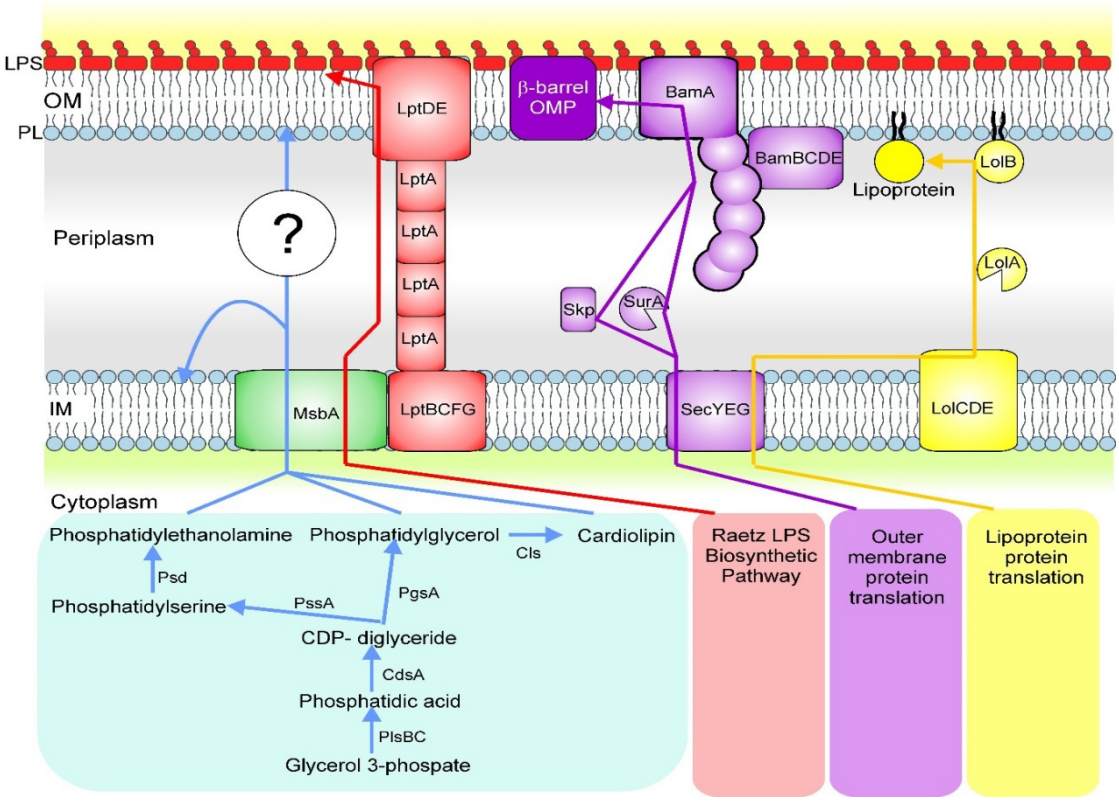

Figure 1.1. The *E. coli* cell envelope. This figure depicts the Gram negative envelope. Phospholipids are shown in blue across both membranes. In the inner membrane (IM), SecYEG (purple) transports unfolded protein into the periplasm, where it is bound by periplasmic chaperones such as Skp and SurA. These chaperones deliver proteins to the BAM complex (BamABCDE) which straddles the outer membrane (OM). The BAM complex facilities the insertion of beta barrel outer membrane proteins (OMPs) into the OM. The Lpt system (LptABCDEFG in red) is responsible for transporting lipopolysaccharide from the cytosol to the outer leafleft of the OM. The Lol pathway (LolABCDE in yellow) traffics OM destined lipoproteins across the IM. MsbA is embedded in the IM, and transports phospholipids and other molecules across the IM into the periplasm.

plays multiple important roles. One key role is the maintenance of internal homeostasis. The delineation of internal and external environments is fundamental in the very definition of a cell, and also the ability to control the influx and efflux of molecules into and out of the cell (Krulwich, Sachs and Padan, 2011). Another vital role concerns the protection of the cell from potentially damaging stresses. The presence of the envelope physically occludes certain molecules from entering into the cell and disrupting metabolism (Ruiz, Kahne and Silhavy, 2006). Furthermore, the envelope can also serve as a platform for molecular machineries that are involved with the active influx or efflux of desired or undesired molecules. This allows for a finer grained control of molecule exchange which, for example, is important in the ability to react to changes in the environment. With regards to the initiation of pathogenesis, the envelope is the point of first contact between host and pathogen (Lee and Schneewind, 2001). By necessity, the envelope is the location of protein complexes that are required for pathogens to attach to and invade host cells.

1.2.1 Inner membrane. The inner plasma membrane is the direct boundary of the cytosol. It is a symmetrical phospholipid bilayer, with the inner layer in contact with the cytosol and the outer layer in contact with the aqueous periplasm (Bos, Robert and Tomassen, 2007). In addition to the fundamental role of cellular delineation, this membrane plays host to a number of integral proteins that are involved with key cellular processes (Weiner and Li, 2008; Silhavy, Kahne and Walker, 2010).

1.2.1.1 Protein translocation. The inner membrane forms a barrier the cell has to overcome to transport proteins into the periplasm, the outer membrane and the environment. Additionally, the insertion of proteins into a plasma membrane poses a

thermodynamic challenge for the cell, in that this process requires energy expenditure. As such, *E. coli* contains three translocation systems that are essential for the insertion of proteins into and across the inner membrane. The most commonly utilised of these systems is the Sec pathway, which transports unfolded proteins and is minimally comprised of the inner membrane SecYEG channel along with the SecA motor protein (Driessen and Nouwen, 2008). The importance of this transport system is illustrated by its high conservation across all three domains of life (Pohlschröder *et al*., 1997). Sec substrates synthesized in the cytosol generally contain an N terminal signal sequence, and are termed preproteins. Preproteins are targeted to the Sec apparatus in two ways, either by SecB and the trigger factor (TF) protein or by the signal recognition particle (SRP). The route taken is determined by the hydrophobicity of the signal sequence in the preprotein (Du Plessis, Nouwen and Driessen, 2011). The signal sequence is exposed as part of the nascent polypeptide emerging from the ribosome, which is competitively recognised by both TF and SRP. Highly hydrophobic signal sequences are preferentially bound by SRP, which leads to their co-translational translocation. The binding of SRP to the nascent polypeptide slows translational activity, which allows SRP to dimerise with FtsY which is associated with SecYEG. This brings the ribosome and the Sec apparatus into close proximity. The FtsY/SRP dimerisation stimulates the hydrolysis of GTP, leading to the nascent chain being transferred to the Sec apparatus. The remaining polypeptide chain then continues to be synthesized, and this elongation provides the energy for the cotranslational insertion of the protein. Less hydrophobic signal sequences are preferentially bound by TF, which leads to post translation translocation. The binding of TF prevents the binding of SRP, in turn allowing the full translation of the preprotein. SecB then associates with this full length polypeptide and keeps the preprotein

in an unfolded state ready for translocation. Additionally, SecB delivers the unfolded preprotein to SecA, an ATP dependent motor protein (Zhou and Xu, 2005). SecB interacts with and transfers the preprotein to SecA, which then goes on to interact with SecYEG and provide the requisite energy for translocation of the preprotein.

A second pathway of protein translocation across the inner membrane involves YidC (Xie and Dalbey, 2008). YidC directly contacts hydrophobic regions of its substrates, and it is thought that YidC uses hydrophobic force to facilitate protein insertion into the inner membrane. Interestingly, YidC plays a dual role. It can assist Sec-dependent translocation and it can transport substrate proteins across the membrane in a Sec-independent manner. However, only a small number of the latter category have been reported (Dalbey *et al*., 2014).

The third method of translocation, the twin arginine transport (Tat) pathway, is comprised of at least three proteins (TatABC). The activity of these proteins is coordinated to transport fully-folded proteins across the inner membrane (Palmer and Berks, 2012). Proteins governed by the Tat system have a characteristic N-terminal signal peptide containing a twin-arginine motif, which is recognised by the TatBC complex embedded in the inner membrane. Upon the interaction of TatBC and a protein substrate, TatA is then able to associate with the protein complex, and the substrate is moved across the membrane into the periplasm, at which point the signal peptide is cleaved by a signal peptidase. This process is dependent upon energy from the proton motive force.

1.2.1.2 Energy generation and proton motive force. One such key process concerns the production of ATP. The F_0F_1 ATP synthase catalyses the production of ATP, the universal cellular energy currency (Yoshida *et al*., 2001; Senior, Nadanaciva and Weber, 2002). This

multi protein enzyme is embedded within the inner membrane, and harnesses the proton motive force to create ATP by channelling the movement of H+ ions from the periplasm into the cytoplasm, where they are actively pumped back into the periplasm. ATP synthase has a tripartite structure; a rotary motor integrated into the inner membrane (F_0) is directly linked to a cytoplasmic 'headpiece' containing three catalytic sites (F_1), with a rotor stalk connecting the two functional units (Fig. 1.2). When there are many protons in the periplasm, an electrochemical potential gradient is formed, also known as the proton motive force. Protons move down this gradient through the F_0 subunit. This proton flow induces F_0 rotation, the energy of which is transferred to the F_1 subunit via the rotor stalk. This energy transmission prompts conformational changes in the F_1 catalytic sites, leading to the synthesis of ATP from ADP and Pi. The action of ATP synthase is dependent upon the maintenance of the electrochemical gradient of protons between the periplasm and the cytoplasm. Protons are actively pumped into the periplasm through the electron transport chain (Hosler, Ferguson-Miller and Mills, 2006). Electrons, donated from reduced molecules such as NADH and succinate, move sequentially through multiple proteins. Some of these proteins are proton pumps, which, upon electron transmission, pump protons from the cytoplasm into the periplasm.

1.2.1.3 Environment sensing and response. Integral inner membrane proteins are also linked to sensing and reacting to the environment. Bacteria are capable of assessing their immediate environment, for example, in terms of nutrient availability, temperature, pH, and toxicity among other environmental conditions (Blair, 1995). This in turn allows for organisms to respond specifically to the detected stimuli and to coordinate an appropriate response, for example, in the movement towards more nutrient rich areas. Often, such

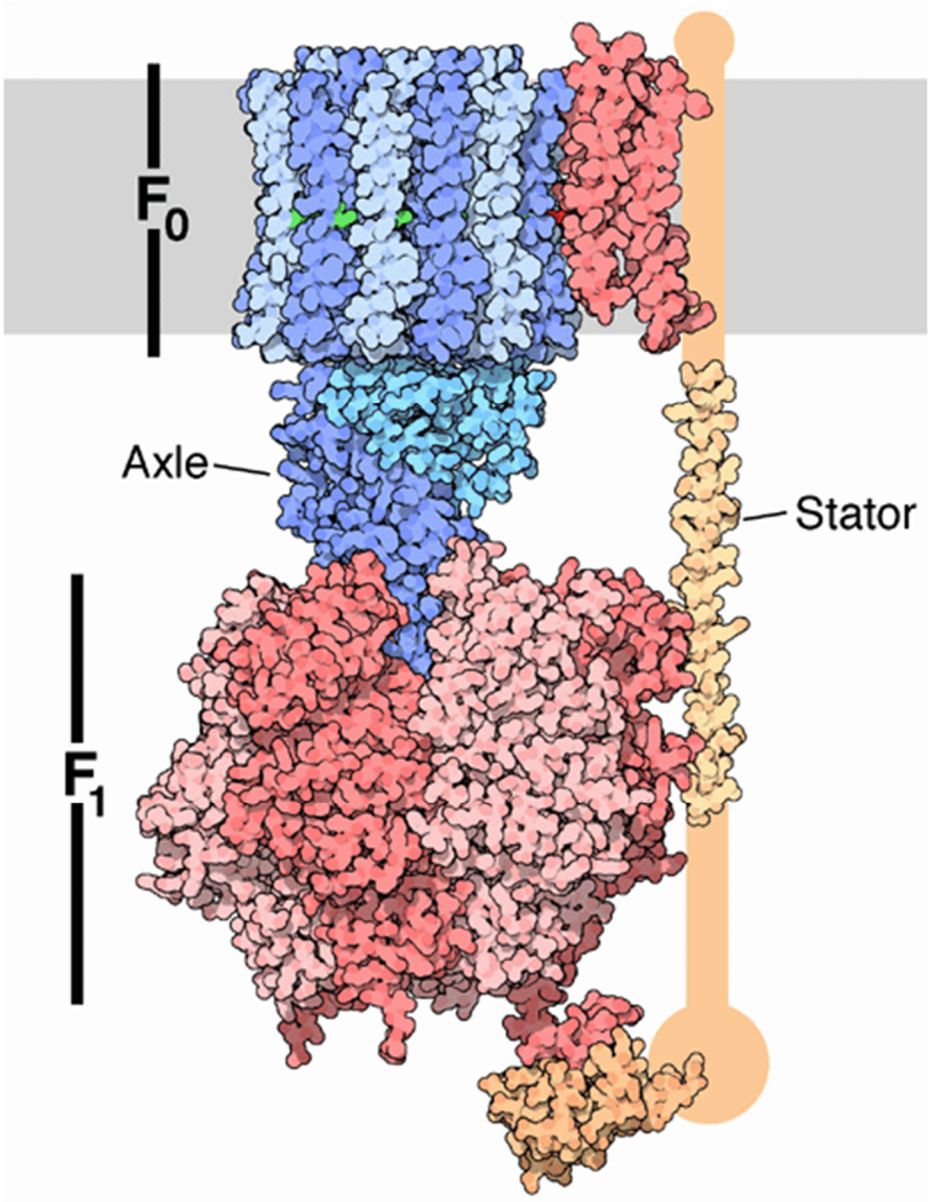

Figure 1.2. ATP Synthase (Berman *et al.*, 2000). ATP synthase is anchored within the inner membrane, shown in grey. The F_0 rotor is embedded in the membrane, and the rest of the complex is in the cytoplasm. The rotor stalk (here labelled as the axle) connects the F_0 and F_1 subunits, with F_1 containing the catalytic sites where ADP and inorganic phosphate are catalysed into ATP.

sensing and response is coordinated by two component regulatory systems (Bourret and Silversmith, 2010; Yamamoto *et al*., 2005). Two component systems are signal transduction cascades that allow for adaptation to external stimuli through the direct regulation of specific genes, and *E. coli* is thought to contain approximately 30 of these systems. One example of such a system is the EnvZ/OmpR two component system (Fig. 1.3). EnvZ, a dimeric histidine kinase embedded in the inner membrane, senses changes in environmental osmolarity (Cai and Inouye, 2002; Feng et al., 2003). OmpR, the DNA binding cognate response regulator to EnvZ, is directly controlled through phosphorylation by EnvZ, which in turn moderates its transcriptional factor activity on downstream genes. EnvZ reacts to a high osmolarity environment by autophosphorylation, and this phophoryl group is then transferred to OmpR. OmpR, upon activation by phosphorylation, subsequently regulates the expression of the outer membrane porins OmpC and OmpF. The transcription of OmpC is upregulated while OmpF is repressed, which restricts the movement of water and solutes out of the cell by virtue of the smaller pore size of OmpC. In low osmolarity, EnvZ does not autophosphorylate and this leads to an abundance of OmpF over OmpC, and the increased pore size of OmpF leads to the greater inwards movement of water and solutes.

which then facilitate the inward movement of small hydrophilic molecules.

1.2.2 Periplasm. The outer leaflet of the inner membrane borders the periplasm, an aqueous compartment between the inner and outer membranes. The periplasm is a viscous, oxidising environment which contains a large number of proteins as well as a peptidoglycan layer (Ruiz, Kahne and Silhavy, 2006).

1.2.2.1 Peptidoglycan. The peptidoglycan layer (also known as murein) is a structurally ordered component of the envelope which acts to maintain cellular shape, provide rigidity, and prevent cellular lysis (Silhavy, Kahne and Walker, 2010; Fig. 1.4). The layer can be described as a porous, covalently polymerised mesh (Gumbart *et al*., 2014; Vollmer, Blanot and de Pedro, 2008); long, linear glycan strands composed of *N*-acetylglucosamine (Glc*N*ac) and *N*-acetylmuramic acid (Mur*N*ac) are cross linked by short oligopeptides. As to be expected with such a structure, there are a considerable number of tightly regulated enzymes which are responsible for peptidoglycan synthesis, which occurs in both the cytoplasm and the periplasm (Vollmer and Bertsche, 2007). The first step in the pathway is done in the cytoplasm, where a UDP-Mur*N*ac-pentapeptide monomer precursor is assembled sequentially through several UDP based precursors and lipid intermediates (van Heijenoort, 2001). This process contains 6 enzymatic steps mediated by MurABCDEF. The phospho-Mur*N*ac-pentapeptide group of the precursor is then transferred to the inner membrane associated carrier undecaprenyl phosphate by MraY, resulting in lipid I. MurG then transfers a GlcNac molecule to lipid I, to result in lipid II. MurJ is the flippase that then transfers lipid II from the inner leaflet of the inner membrane to the outer leaflet, at which point it becomes available for incorporation into the peptidoglycan layer. (Sham *et al*., 2014). Several steps in the pathway then occur periplasmically. Lipid II is then polymerised to form glycan strands, and this step is catalysed by periplasmic peptidoglycan synthases anchored to the inner membrane (Lovering, Safadi and Strynadka, 2012; Derouaux, Sauvage and Terrak, 2013; Vollmer and Bertsche, 2008). Specifically, glycosyltransferases catalyse the synthesis of the glycan strands, and transpeptidases assemble the peptide cross links between the glycan strands (van Heijenoort, 2001). In addition to its synthesis, the

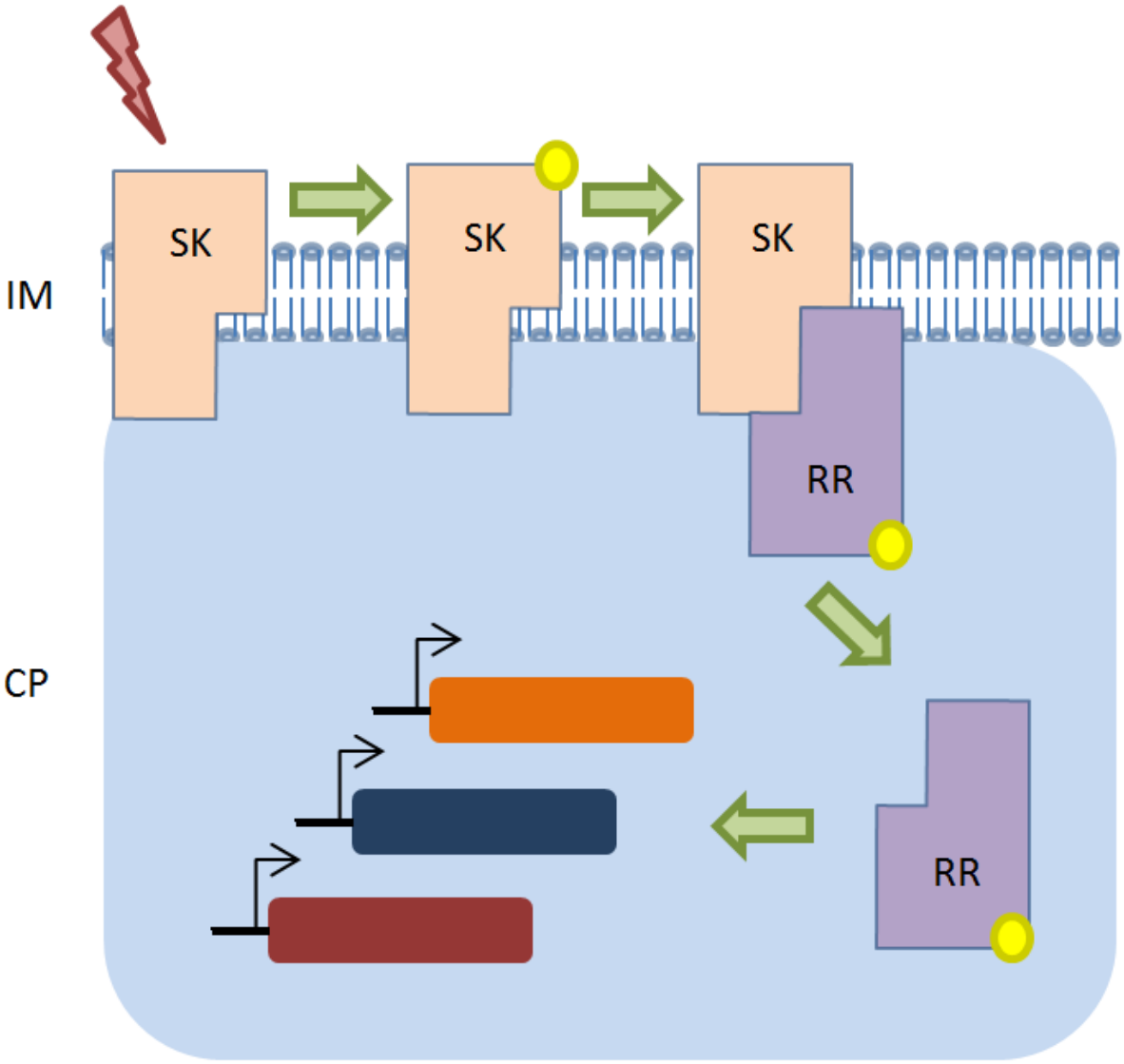

Figure 1.3. Two component systems. A sensor kinase (SK) in the inner membrane (IM) detects a specific environmental change. This leads to autophosphorylation (yellow circle) and subsequent transphosphorylation of a cognate response regulator (RR). The RR, now active and free in the cytoplasm, goes on to regulate genes involved in the response to the initial environmental stimulus.

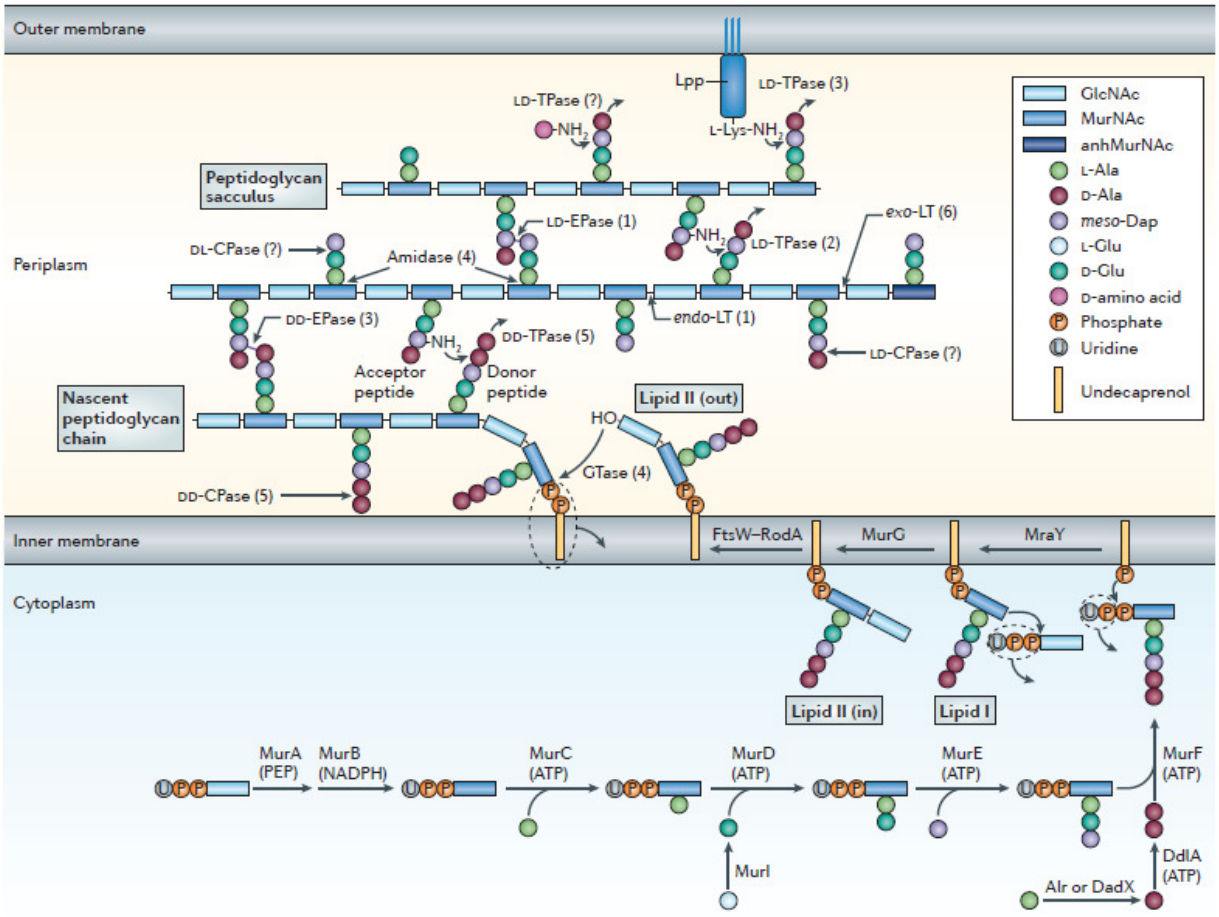

Figure 1.4. Peptidoglycan biosynthesis (Typas et al., 2011). Peptidoglycan is found in the periplasm of *E. coli*, and undergoes multiple enzymatic steps to get there. In the cytoplasm, MurABCDEF act sequentially to form a UDP-Mur*N*ac-pentapeptide precursor. MraY then transfers the phospho-Mur*N*ac-pentapeptide group of this precursor to undecaprenyl phosphate in the inner membrane, to form lipid I. MurG transfers a GlcNac unit to lipid I to form lipid II, which is then flipped to the periplasmic face of the inner membrane by the MurJ flippase. Here, lipid II can be incorporated into growing peptidoglycan chains by glycosyltransferases. Transpeptidases act to form peptide cross links between glycan chains.

breakdown of peptidoglycan is necessary, for example in the adaptation of the cell to changing environments and also for the separation of daughter cells after division (Vollmer *et al*., 2008). Peptidoglycan hydrolases act to cleave the bonds both within the glycan strands and also within the peptide crosslinks. Different hydrolases exhibit different specificities; for example, amidases (including AmiABC) cleave bonds between glycan and peptide, whereas endopeptidases cleave bonds within peptide cross links.

1.2.2.2 Chaperones. Despite the periplasm being devoid of ATP, multiple enzymes function to regulate protein folding (Silhavy, Kahne and Walker, 2010; Goemans, Denoncin and Collet, 2013). Approximately 20% of all proteins produced in *E. coli* are destined for the cell envelope. However, proteins are only produced in the cytoplasm, meaning the cell must have strategies to a) transport proteins to their correct final destination, and b) ensure that they are folded to function correctly.

As previously discussed, unfolded proteins are transported to the periplasm primarily through the Sec translocon. To regulate the folding of these proteins, and to help traffic them to their correct destination in any part of the envelope, there are an array of protein chaperones which directly interact with the unfolded polypeptides. Chaperones also act upon misfolded proteins, which can also occur for multiple reasons. These include environmental stress (for example heat leading to protein denaturation), protein overexpression and genetic mutation (Miot and Betton, 2004). Unfolded and misfolded proteins, unless protected by chaperone binding, are then at risk of proteolytic degradation or aggregation into inclusion bodies, which subsequently activate stress response pathways in the cell.

Outer membrane proteins (OMPs) form a large subset of proteins that are trafficked through the envelope, with the outer membrane as their final destination (discussed in more detail later in the chapter). Two chaperone folding pathways have been outlined in *E. coli* (Goemans, Denoncin and Collet, 2013). The primary pathway relies upon SurA, a dual functioning chaperone and peptidyl prolyl isomerase (PPIase). SurA was first found to be essential for survival in the stationary phase of growth (Tormo, Almiron and Kolter, 1990). More recent evidence suggests that SurA is in fact the primary chaperone of *E. coli*, due to the fact that the depletion of SurA leads to a marked decrease in the outer membrane proteome, in contrast to the lack of effect seen upon depletion of other chaperones (Denoncin *et al*., 2012; Sklar *et al*., 2007a). The second, lesser chaperone pathway is fulfilled by Skp and DegP. Sklar *et al*. (2007a) found that upon the depletion of Skp and DegP individually, there was no change to the density of the outer membrane. Combination depletions of SurA/Skp and SurA/DegP, however, resulted in envelope defects. Furthermore, apparently no proteins have a preference for Skp/DegP over SurA. Even with this evidence, other findings that demonstrate chaperone activity by Skp and DegP have led to this pathway being thought of as partially redundant, and as being involved with "rescuing" proteins that fall off of the central SurA pathway (Goemans, Denoncin and Collet, 2013). Another chaperone, LolA, acts to traffic lipoproteins that are destined for the outer membrane (Okuda and Tokuda, 2009). The LolCDE complex moves lipoproteins from the cytoplasm to the periplasm, at which point LolA binds. LolA transports the lipoprotein to LolB in the outer membrane, which then facilitates the insertion of the lipoprotein.

1.2.3 Outer membrane. Beyond the inner membrane and the periplasm is the outer membrane, which is in direct contact with the external environment. This lipid bilayer is the physical difference between Gram negative and Gram-positive bacteria. This membrane acts as a selectively permeably barrier, which simultaneously prevents the entry of damaging agents into the cell while allowing the entry of nutrients (Ruiz, Kahne and Silhavy, 2006; Bos, Robert and Tommassen, 2007).

1.2.3.1 Lipopolysaccharide (LPS). In contrast to the inner membrane, the outer membrane is asymmetrical (Silhavy, Kahne and Walker, 2010). The periplasmically facing inner leaflet is comprised of phospholipid, whereas the environmentally facing outer leaflet is primarily composed of lipopolysaccharide (LPS), a glycolipid containing lipid A, a core oligosaccharide and an O antigen polysaccharide. LPS is central to the barrier function of the outer membrane (Ruiz, Kahne and Silhavy, 2006). In wild type cells the LPS is highly compacted, which physically occludes the entry of compounds such as antibiotics or detergents (Snyder and McIntosh, 2000). Correspondingly, mutants with defects in LPS biogenesis pathways have an increased outer membrane permeability and susceptibility to external agents.

1.2.3.2 Lipoproteins. The outer membrane also hosts its own complement of proteins. Generally these proteins are either lipoproteins or integral outer membrane proteins (Ruiz, Kahne and Silhavy, 2006). Lipoproteins are those containing both protein and lipid regions within their structure, allowing for proteins to be anchored to a plasma membrane (Okuda and Tokuda, 2011). On the periplasmic face of the inner membrane, lipid modifications are added to the N terminal cysteine residue of a lipoprotein by the phospholipid transacylase Lnt. This results in mature lipoprotein that is anchored to the

inner membrane. The final destination of lipoproteins can be in either membrane of the Gram negative envelope. The transfer of lipoproteins from the inner to the outer membrane is mediated by the Localisation of lipoprotein (Lol) transport system (Fig. 1.5). LolA, a periplasmically located chaperone, delivers outer membrane destined lipoproteins to LolB, another lipoprotein anchored within the inner leaflet of the outer membrane. It is uncertain how LolB then acts to insert the lipoprotein into the outer membrane.

One major function of lipoproteins is to help anchor protein complexes to the envelope. One example of this is the peptidoglycan layer, which is anchored within the periplasm by multiple proteins. Most notably, Braun's lipoprotein (Lpp), the most abundant protein within *E. coli*, has been shown to be important for this anchoring (Vollmer and Bertsche, 2007; Cowles *et al*., 2011). Lpp, in the outer membrane, exists in a free and a bound form; bound Lpp is covalently linked to peptidoglycan, whereas free Lpp is not. The ratio between the free and bound forms is approximately 2:1. These two forms are spatially separated in the envelope, with free Lpp being surface exposed in the outer membrane, in contrast to the presence of bound Lpp in the periplasm. Pal is another protein that interacts with the peptidoglycan layer (Cascales *et al*., 2002; Parsons, Lin and Orban, 2006). Pal (peptidoglycan associated lipoprotein) forms part of the Tol-Pal complex, an envelope spanning multi protein complex which is involved with the constriction of the envelope during cell division (Egan and Vollmer, 2013; Gerding *et al*., 2007). This lipoprotein is anchored to the outer membrane, and noncovalently binds to the peptidoglycan, which helps to maintain the structure of the peptidoglycan layer.

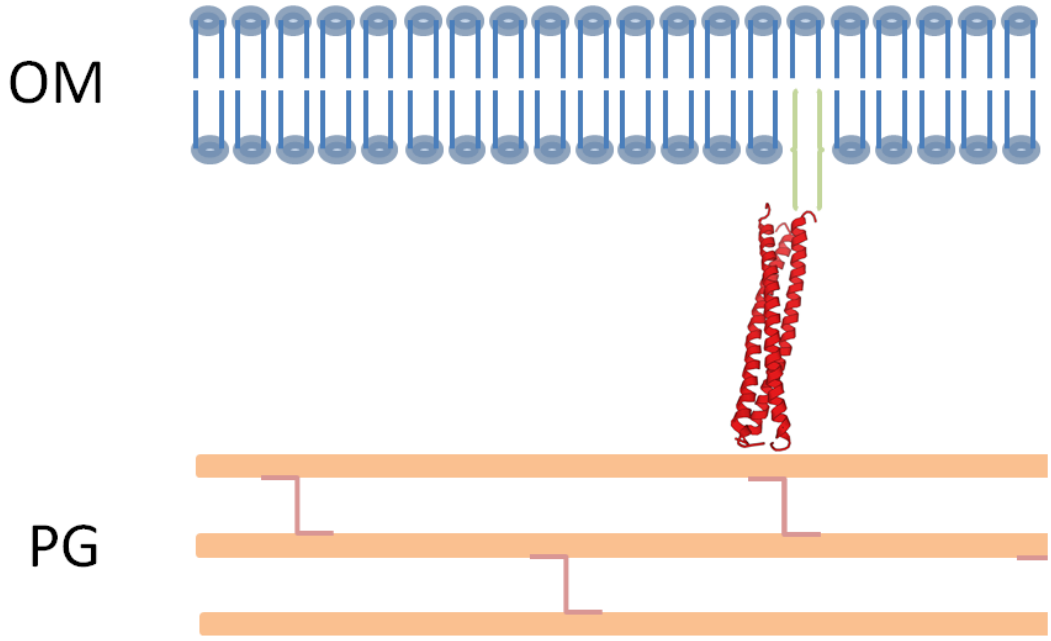

Figure 1.5. Lipoprotein anchoring. Lpp is shown in red and green (Shu et al., 2000). The green region represents an N terminal lipid modification which inserts into the outer membrane (OM). The polypeptide region of the protein shown in red is in the periplasm, and covalently attaches to the peptidoglycan layer (PG) at the C terminal end of the protein. These physical interactions act to stabilise the cell envelope.

1.2.3.3 Outer membrane proteins (OMPs). In addition to the lipoprotein complement, the outer membrane hosts many outer membrane proteins. A key difference between the integral proteins of the inner and outer membranes is their structure; whereas inner membrane proteins contain α helical transmembrane domains, outer membrane proteins (OMPs) contain antiparallel β strands. These antiparallel strands form a hydrophobic surface which facilitates the embedding of the OMP by physically spanning the OM (Tamm, Hong and Liang, 2004). This allows OMPs to form β barrel structures which are embedded within the OM, and it is thought that the folding of the OMPs occurs upon insertion into the OM. Such β barrel containing structures can serve as passages of entry into the cell, as enzymes and also as adhesins (Ruiz, Kahne and Silhavy, 2006).

Proteins in the outer membrane are associated with a number of functions. OmpA is one of the most abundant OMPs in *E. coli*, and forms a non-specific pore through which small solutes diffuse (Smith *et al*., 2007; Sugawara and Nikaido, 1992). Other major porins include OmpC, OmpF and PhoE (Hancock, 1987). In addition to the channels involved with the influx of solutes into the cell, there are dedicated efflux pumps to actively remove toxic substrates and prevent damage to the cell (Webber and Piddock, 2003). A wide variety of structurally differing substrates are recognised and exported by efflux pumps, including detergents, dyes, antibiotics and biocides (Piddock, 2006). In Gram negatives, these pumps are envelope spanning multi protein complexes, which form a direct channel between the cytosol and the external environment. There are five efflux pump families, each with their own substrate specificities and structural composition. One of the most well-known pumps in *E. coli* is the AcrAB-TolC system (Tikhonova and Zgurskaya, 2004; Du *et al*., 2014; Fig. 1.6). TolC forms a channel in the outer membrane, AcrB is a proton-substrate antiporter and AcrA

is a periplasmic lipoprotein which physically interacts with AcrB and TolC and bridges the two proteins. This efflux pump has multiple substrates, including β-lactams, fluoroquinolones, bile salts and detergents (Piddock, 2006). Efflux pumps have profound medicinal importance; it is well understood that these systems can confer multidrug resistance upon many bacterial strains, and antimicrobial resistance has been recognised as an ever growing threat to the clinical treatment of infection (Cole, 2016).

Another similarity between the inner and outer membranes is the presence of dedicated protein machinery to insert proteins into the membranes. The β-barrel assembly machinery (BAM) complex is responsible for inserting folded proteins into the outer membrane, and is comprised of BamABCDE (Knowles *et al*., 2009; Hagan, Silhavy and Kahne, 2011). BamA, an integral β-barrel in the outer membrane, is the core component of the complex. As an indication of its importance, it is an essential protein conserved across all Gram negatives. In addition to the channel forming barrel, the protein has 5 periplasmic POTRA domains which receive substrate proteins to be inserted into the outer membrane.

The periplasmic chaperone SurA has been shown to directly contact the POTRA1 domain of BamA where Skp has not, reinforcing the idea of SurA being the primary chaperone pathway (Kim, Aulakh and Paetzel, 2012). The POTRA domains are also contacted by all four of the other lipoprotein BAM components, including the essential BamD. This protein has two domains; the N terminal domain is thought to directly bind to proteins to be inserted, while the C terminal is important for maintaining interactions with BamBCE (Misra, 2012). As such, it is likely to function in the delivery of substrate proteins to BamA. BamBCE are non-essential genes, although the biogenesis of outer membrane proteins is negatively affected upon their individual deletion (Rigel and Silhavy, 2012).

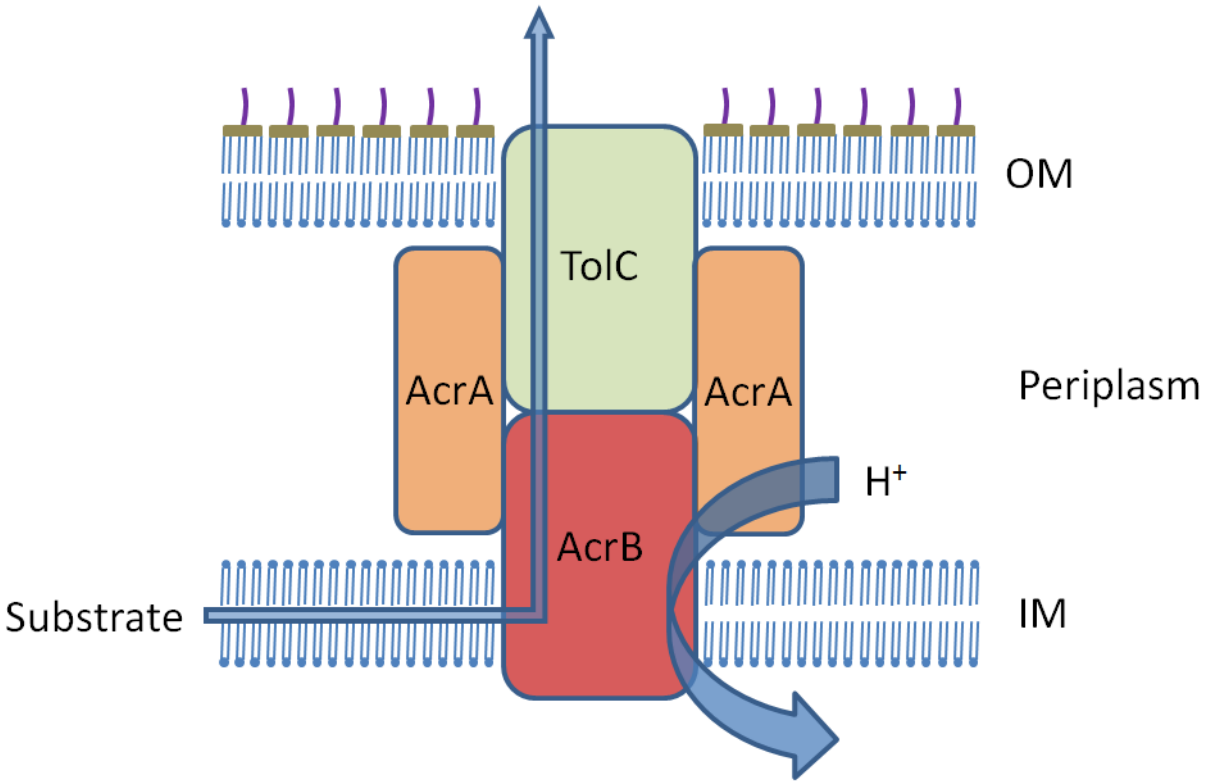

Figure 1.6. Schematic of the AcrAB/TolC efflux pump. The three components form a complex that spans the envelope of the cell. The movement of ions from the periplasm into the cytoplasm through AcrB (due to the proton motive force) provides the energy required for substrate efflux (Müller and Pos, 2015). The disruption of any one of these proteins leads to the lack of complex formation and the cease of efflux.

Evidence suggests that, while BamB is not essential for protein loading onto the BAM complex, it is important in making the process efficient (Hagan, Silhavy and Kahne, 2011).

1.2.4 Envelope integrity and permeability. As discussed previously, the outer membrane of *E. coli* is key in protecting the cell from the ingress of toxic agents. The combination of tightly packed LPS, the peptidoglycan layer, dual lipid membranes and active efflux together comprise an effective barrier to all manner of compounds. Disruption of the biogenesis or the maintenance of the outer membrane leads to a suboptimal structure that is less densely packed, which in turn weakens the cell by increasing the membrane permeability. There is much experimental evidence to support this. BamB, SurA, TolQRAB, Pal, and AmiA are all examples of genes that result in increased outer membrane permeability upon deletion (Ruiz *et al*., 2005; Justice *et al*., 2005; Lazzaroni *et al*., 1999; Heidrich *et al*., 2002). One method by which deletions are tested for their effects on membrane integrity involves the use of molecules that the wild type strain is normally resistant to. This includes many different classes of antibiotics, detergents, dyes and other molecules (Nikaido and Vaara, 1985). Upon perturbation of the outer membrane, these molecules are able to pass into the cytoplasm of the cell and act to stop growth.

1.2.4.1 Vancomycin. Vancomycin is one antibiotic often used to assess outer membrane integrity (Tamae *et al*., 08; Liu *et al*., Lazdunski and Shapiro, 1972). Initially discovered and purified from *Streptomyces orientalis* in 1952 (Levine, 2006), vancomycin is the archetypal member of the glycopeptide antibiotics (Loll and Axelson, 2000; Pootoolal, Neu and Wright, 2002). The structure consists of a covalently linked core of seven amino acids, along with glycosylative and other amino acid modifications. Vancomycin's mode of

action is to inhibit the synthesis of peptidoglycan (Fig. 1.7). By associating with the terminal D-ala-D-ala residues of the glycan strand subunits, the extension of the glycan strands (through transglycosylation) and the crosslinking between strands (through transpeptidation) are physically impeded, leading to the cessation of peptidoglycan maturation. Because of this mode of action, the outer membrane of *E. coli* provides a natural resistance to vancomycin (Shlaes *et al*., 1989; Reimer, Stratton and Reller, 1981). The outer membrane porins physically occlude the entry of the hydrophilic vancomycin due to the large size of the molecule, and the tightly packed LPS layer in the outer leaflet of the outer membrane provides another physical barrier. It therefore follows that, with a suboptimally maintained outer membrane, vancomycin is able to pass through into the periplasm, where it can affect peptidoglycan synthesis and exhibit its bactericidal activity.

1.2.4.2 Sodium dodecyl sulphate (SDS). The anionic surfactant sodium dodecyl sulphate (SDS) is another molecule used to investigate membrane integrity (Bernstein, Rolfe and Onodera, 1972; Lazdunski and Chapiro, 1972). In the laboratory setting, SDS is generally used to denature proteins in preparation for polyacrylamide gel electrophoresis. In the same manner as with vancomycin, *E. coli* possesses a natural resistance to SDS, mediated by LPS in the outer membrane (Rajagopal, Sudarsan and Nickerson, 2002; Nikaido and Vaara, 1985). LPS and SDS are both negatively charged, and this charge repellence counteracts the ability of SDS to move across the hydrophobic OM. However, as opposed to completely preventing its entry into the cell, the OM is weakly permeable to SDS, but the cell is able to tolerate its presence in low quantities. Active efflux (including the AcrAB-TolC pump) works in addition to the barrier function of the outer membrane to pump SDS into the environment and to keep it out of the cytoplasm (Yu, Aires and Nikaido, 2003).

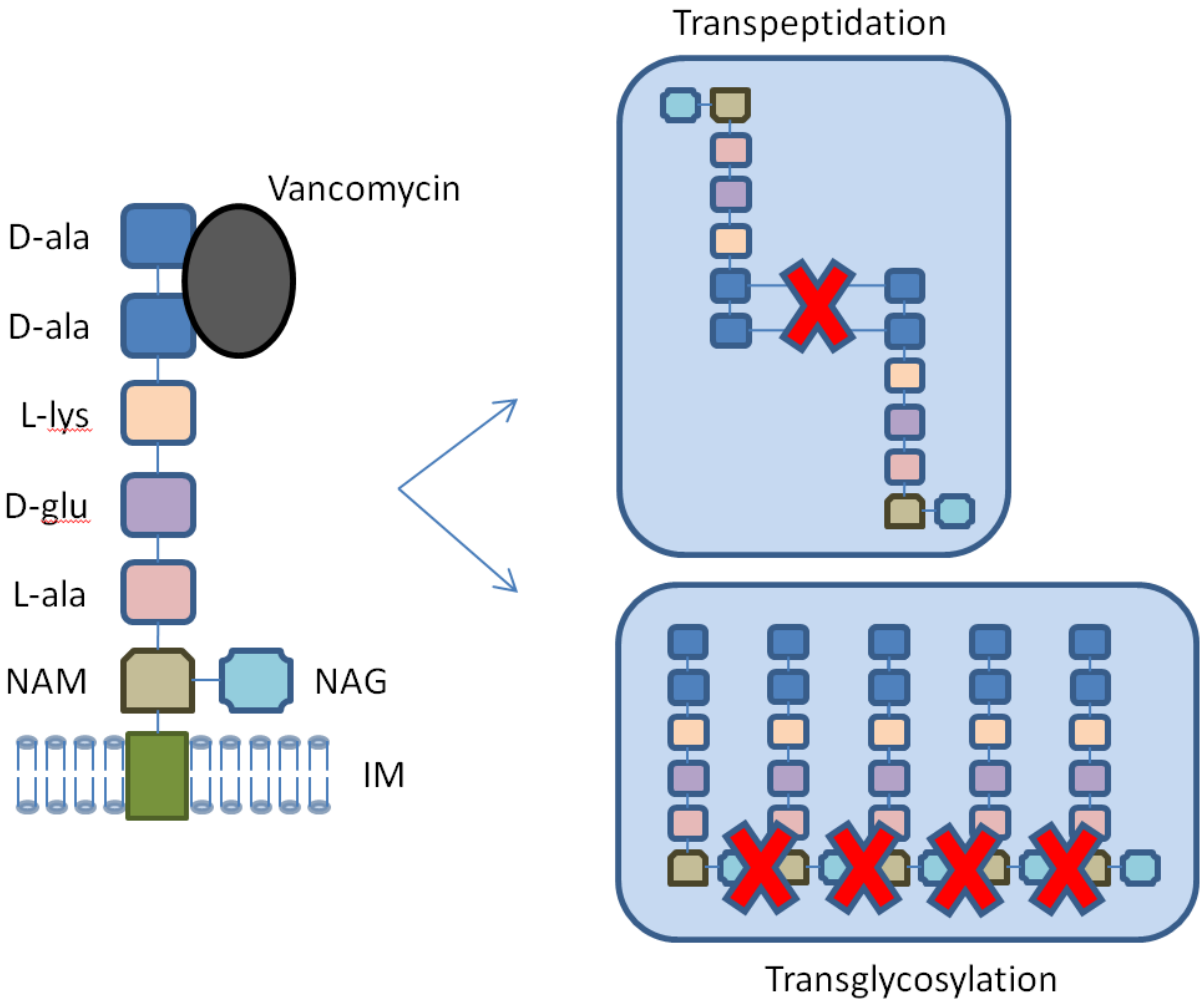

Figure 1.7. Vancomycin and its mechanism of action. Adapted from Pootoolal, Neu and Wright (2002). Vancomycin binds to the terminal D-ala-D-ala residues of the glycan strand in the periplasm. This physically occludes further extension of the glycan (transglycosylation) and the crosslinking of adjacent strands (transpeptidation), which in turn prevents the formation of the peptidoglycan layer, which then weakens the structural integrity of the cell.

1.2.4.3 Industrial application. In addition to the academic virtues of research into envelope integrity, a more practical application exists from an industrial perspective. Knowledge of the genes involved with these aspects of the envelope may be applied towards the goal of recombinant protein production, which is the manipulation of microbes to produce specific proteins of greater interest. In previous years, the isolation and purification of a particular protein would have required huge amounts of biomass, of which only a tiny percentage was actually desirable. Technical development in the production of recombinant proteins using bacterial systems has greatly facilitated academic, industrial and medical endeavour (Rosano and Ceccarelli, 2014). From a research perspective, the relatively simple production of large amounts of protein has meant that biochemical, structural and enzymatic studies are far easier and quicker to do. Industrially speaking, several avenues of research with societal impact have been enabled through the use of microbes, including bioremediation and the production of enzymes for household use (Karigar and Rao, 2011; Basketter *et al*., 2008). Arguably most importantly of all, medical advances have been greatly facilitated by using microbes for to produce therapeutic treatments. For example, the bacterial production of insulin has obviated the dependence of diabetic individuals upon crude porcine pancreatic material, while also enabling the modification of insulin for better disease management (Johnson, 1983; Kamionka, 2011).

E. coli is a commonly used microbiological platform used to manufacture recombinant proteins, due to its robust growth, ease of manipulation and well understood biological underpinnings (Baneyx, 1999). Generally, protein production systems are designed to be wholly contained in the cytoplasm. One downside to this method is that substantial processing is required after growth, to isolate the desired protein from the other cellular

components. An alternative approach to protein production is to engineer a system in which proteins are secreted or "leaked" from the cell (Le and Trotta, 1991; Rinas and Hoffman, 2004). Due to the desired protein product being extricated from the cytoplasmic complement of macromolecules, simpler and more effective purification is facilitated (Mergulhao, Summers and Monteiro, 2005). Additionally, for some protein families with specific structural features, such as disulphide bridges, the reducing environment of the cytoplasm is suboptimal for protein production (de Marco, 2009; Ke and Berkmen, 2014). As such, non-cytoplasmically based production strains could be improved by the deletion of specific genes involved with envelope homeostasis. The decrease in envelope integrity would physically allow for the easier movement of proteins either into the periplasmic space or into the extracellular environment.

1.3 Gene essentiality

Fundamentally, the existence of each and every organism is dependent upon its genetic underpinnings. The genes contained within an organism define the entirety of its ability, throughout every aspect of the lifecycle. This includes but is not limited to metabolic capacity, reproductive capability, pathogenic strategy and environmental adaptation. Microorganisms inhabit a vast range of ecological niches, and organisms necessarily have to adapt to conditions to survive. This brings us to the consideration of gene function. For a given microbial species, there will be genes absolutely indispensable for survival, and there will be genes that are only required for growth under certain conditions, ie conditionally essential (Zhang and Lin, 2009; Juhas, Eberl and Glass, 2011).

In the discussion of Indispensable genes, hereafter referred to as essential genes, care must be taken from a more philosophical perspective. Due to the vast expanse of biological complexity, a singular, all-inclusive definition of life is controversial. One significantly confounding factor that illustrates this complexity is the existence of unusual organisms such as viruses, which contain genetic material but are utterly reliant upon the molecular machinery of other organisms. Due to the interplay of an organisms genetic complement and its environment, the defining of a gene as essential can be seen as tenuous and dependent upon many factors. For example, one gene may appear essential in one growth media and inessential in another, due to the presence of a particular metabolite. In the literature, essentiality has generally been assessed in either minimal media or LB. In the following work growth in LB has been used as a proxy for life, in order to be able to make comparisons with previous high quality and well recognised literature.

Essential genes are interesting for multiple reasons. From an academic perspective, knowledge of the minimal genetic requirements for life is of key importance in the very definition and classification of life itself (Gustafson *et al*, 2006). Additionally, knowledge of the core genome versus the pan genome of an organism can be of great utility in the descriptive and evolutionary comparison of species (Medini *et al*., 2005). More practically, a minimal gene set is central to the concept of the minimal genome, which in turn has the potential to impact upon multiple areas of science, most crucially the area of biotechnology and the production of societally important macromolecules. Theoretically, microbes can be manipulated into miniscule production factories producing potentially any type of macromolecule, by inserting the genes necessary for production into the minimal genome of an engineered organism. Knowledge of essential genes can also be used in targeted drug

development. Due to their importance, essential genes and the pathways they are involved with are excellent antimicrobial targets, as exemplified by the penicillin family of drugs which specifically interfere with peptidoglycan synthesis.

1.3.1 Research methodologies. Given *E. coli*'s position as a model organism, there is much research that has gone towards the definition of the essential genes it contains. Historically, the effort required to delete even a small part of a gene was substantial. However, the publication of a recombination based gene deletion protocol by Datsenko and Wanner (2000) greatly facilitated the deletion of genes in *E. coli* (Fig. 1.8) Briefly, PCR is used to amplify an antibiotic resistance gene. The forward and reverse primers used in this reaction are designed with homology to the ends of the gene in question. This results in a single linear fragment containing the resistance gene, flanked by homology to a particular gene at both ends. This fragment is recombined with the native gene through recognition of the flanking homologous regions, and successfully recombined cells are then selected for on agar plates supplemented with antibiotics. Later, by utilisation of the Datsenko and Wanner method, Baba *et al*. (2006) published a paper detailing the KEIO library, a collection of single gene deletions for every non-essential gene in *E. coli* BW25113. Conversely, the creation of this library also led to the first definition of an *E. coli* essential gene list. There were 303 genes that could not be successfully deleted, which suggested that they were essential for cellular function. This list included new candidate essential genes of unknown function, in addition to genes previously shown to be essential. The KEIO library has since become the "gold standard" dataset for essential and non-essential genes in *E. coli*, due to the breadth of its scope and the reliability of the deletion method it employed. However, there are flaws in

this approach. The technique has no way of identifying whether there are multiple copies of a particular gene. Yamamoto *et al*. (2009) published evidence of several genes that were duplicated in the initial construction of the library (Baba *et al*. 2006). This suggested their deletion had a limited effect and they were assigned as non-essential genes. Additionally, several of the mutants in the KEIO collection were found to have acquired second-site compensatory mutations elsewhere on the chromosome; these mutants alleviated the lethal effect of loss of the essential gene suggesting the gene was not essential. Furthermore, this method of assessing essentiality is both highly resource and labour intensive, requiring many experimental steps and a substantial amount of manual preparation.

Recombination based gene deletion is not the only method of assessing gene essentiality. Another approach that can be taken is to deplete proteins encoded by genes, through the use of inducible promoters. In these experiments, a gene of interest is cloned onto a plasmid, and the chromosomal copy is deleted. The plasmid borne gene is placed under the transcriptional control of an inducible promoter, in turn meaning that the gene is only transcribed in the presence of an inducer, for example lactose. The strain is plated on growth media containing the inducer. Upon confirmation of growth, the strain is then transferred to two sets of media, containing and lacking the inducer respectively. If the strain grows on both sets of media, the gene in question is non-essential. This is because the strain survives, even when plasmid mediated transcription ceases. If the strain is only viable in the media containing the inducer, this shows the gene to be essential. Depletion studies such as this can give information of other aspects of the gene in question. For example, particular deletions may lead to short or long survival times in media lacking inducer. This is informative in relation to the level of the protein encoded by the gene in the cell.

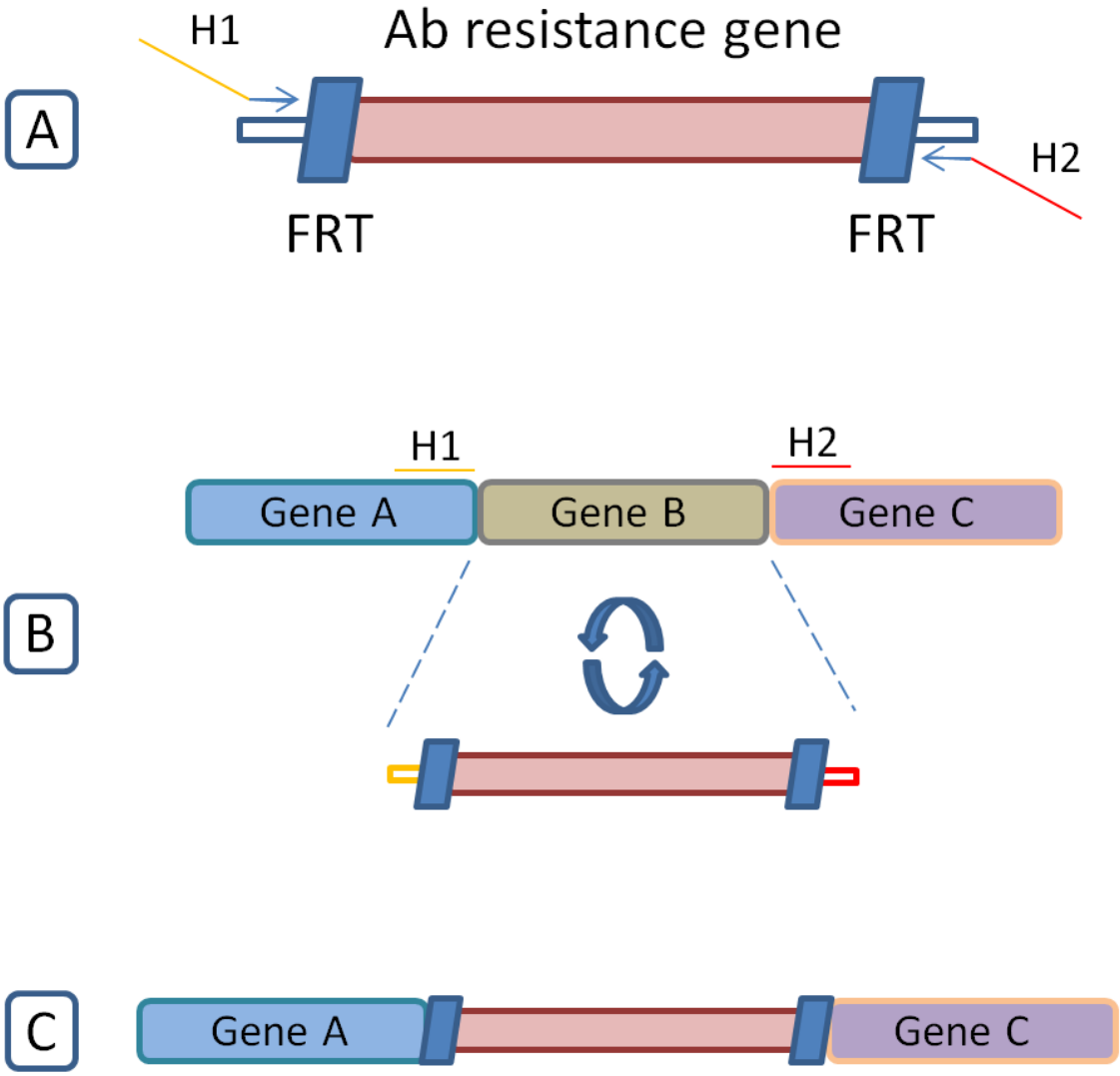

Figure 1.8. The Datsenko and Wanner (2000) gene deletion method. A) PCR is used to amplify an antibiotic resistance gene. The primers used contain two differing sites of homology, in this example on either side of gene B (H1 and H2). B) The amplicon from step A) is recombined with the genome. This "flips" out gene B between H1 and H2, and replaces it with the antibiotic resistance gene. This process is carried out in a λ red recombinase background, with the recombinase performing the essential recombination. C) The FRT sites can be used to excise the resistance gene, leaving a defined genomic "scar" in its place.

Gerdes *et al*. (2003) employed transposon mutagenesis to investigate essentiality. Transposons are small, mobile genetic elements that can semi randomly integrate into the chromosome, which can in turn lead to the disruption the regulation and expression of genes (Reznikoff, 2003; Hamer *et al*., 2001). Transposition into a coding sequence disrupts the subsequent translated protein by the insertion of non-native amino acids, which in turn changes the conformation of the protein. If the function of a disrupted protein is essential to the cell, then it will cease to grow and divide. If the disrupted protein is not essential to the cell, then the cell will continue to grow and multiply, while passing down the disruption through successive generations. Gerdes *et al*. (2003) mutagenized cells with transposons, grew the surviving cells and then used a nested PCR approach to map the insertions to their location in the genome. This work led to the estimation of 620 essential genes, in contrast to the 303 predicted by Baba *et al*. (2006). The major issue with this approach lies in the experimental effort to characterise each and every insertion, and in the number of insertions that are needed to fully survey the genome.

1.4. Advances in DNA sequencing and applications.

The discovery of DNA, and the realisation of its function as the primary data storage medium of life, is still only recent in human history. In a relatively short time frame, DNA has moved from being a macromolecule of unknown function to a cellular component of fundamental importance. DNA has a variable length, double helical chain composed of four nucleotides (adenine, thymine, guanine and cytosine) ordered in a specific sequence (Lewis *et al*., 2007). DNA is first transcribed into RNA, another polynucleotide chain similar to DNA. RNA strands are then translated into proteins comprised of amino acid subunits. Proteins

then go on to fulfil biological functions and, in concert with other cellular components and molecules, allow for the existence of life. DNA is the basal container of information; the nucleotide sequence subsequently transcribed to RNA is read in groups of three, with each triplet combination of the four nucleotides being called a codon. Each codon is recognised by specific tRNA molecules, in turn associated with particular amino acids, which are then used to build up a peptide chain of a defined sequence. This hierarchy of encoding means that knowledge of DNA sequences is crucial to the greater understanding of life.

The ability to sequence DNA came some years after definition of its role in heredity. Dideoxy DNA sequencing, also known as Sanger sequencing, was the technical development which facilitated the beginnings of modern genomics (Heather and Chain, 2016). This technique was based upon the use of radioactively or fluorescently labelled dideoxy nucleotide analogues in a polymerase chain reaction. These molecules lack the 3' hydroxyl group necessary for chain extension, and in combination with polyacrylamide gel electrophoresis (or in later years capillary electrophoresis), the sequence of nucleotides in a DNA sequence could be determined. Notably, commercial DNA sequencers using this technique were used to produce the first draft of the human genome (Lander *et al*., 2000).

1.4.1 Next generation sequencing. Technical innovation continued throughout and after the era of Sanger sequencing, and a number of "next generation" sequencing methodologies were commercialised. Of them all, one particular sequencing platform has come to dominate the field, and as such will be the focus from here onwards. This platform, now the Illumina sequencing platform, has its roots in the mid-1990s under the name of Solexa (Bentley *et al*., 2008; Bio-IT World, 2010; Goodwin, McPherson and McCombie, 2016; Fig. 1.9). The

underlying sequencing technique, termed sequencing by synthesis (SBS), was developed in Cambridge. In a sequencing machine, single stranded DNA is washed over a flow cell, to which small oligonucleotides are covalently attached. The linear ssDNA contains regions complementary to the oligonucleotides, allowing the two to bind. These strands are then used as seeds to generate 'clusters' of DNA containing ~1000 identical linear fragments, through bridging amplification. The process of clustering acts to increase the fluorescent signal eventually created, to make the imaging of base incorporation easier. Sequencing occurs cyclically; first, engineered DNA polymerase and all four fluorescently labelled nucleotides are washed over the flowcell. These nucleotides are also reversible terminators, in that they contain removable 3' azidomethyl groups that prevent extension after their incorporation into the nascent strand. The clusters are then imaged using laser excitation of the fluorophores, enabling identification of the incorporated base. Fluorophores and 3' groups are then excised, allowing for the next cycle of incorporation. As these cycles continue, a DNA sequencing read is generated of the ancestral linear ssDNA. As this technique has been refined over the years, there have been great improvements in read quality, length and number. The earliest iterations of the sequencers resulted in 10-12 bp reads, whereas now, with the latest sequencing chemistry, up to 600 bp reads can be generated. Furthermore, the newer sequencers can output up to 5 billion paired end reads in a single run.

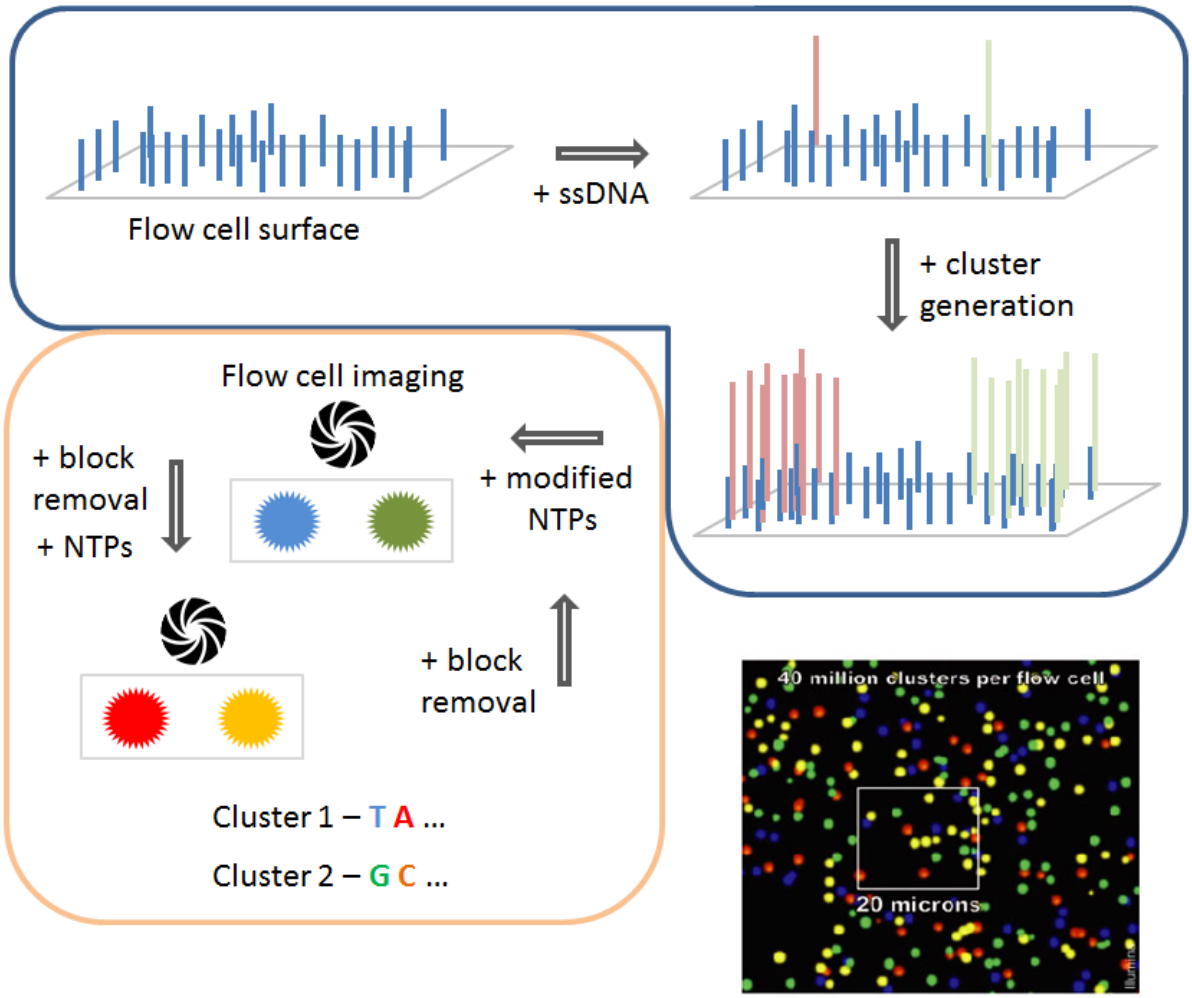

Figure 1.9. A simplified depiction of Illumina sequencing by synthesis. Starting from the top left of the image, is a flow cell with many covalently attached oligonucleotides. The surface of the flow cell is an aqueous environment. ssDNA is washed over the flow cell, which binds to the oligonucleotides through complementary sequence. Local bridging amplification is used to generate clonal clusters of ssDNA. NTPs are then incorporated into the nascent strands. Laser excitation is used to excite a fluorophore attached to the base and is then imaged. A 3' block on the incorporated nucleotide is then removed, and this continues in cycles leading to the generation of sequence reads (lower left panel). The bottom right image is of an actual flow cell during imaging (Chi, 2008).

1.4.2 Transposon insertion sequencing. In microbes, DNA sequence information is often used to look for (or confirm) single nucleotide polymorphisms, deletions or rearrangements. These are now standard uses of sequencing data, and as such, there is widely available, free to use analytical software. However, there are other research aims and approaches that have been enabled by the mainstreaming of next generation sequencing. Most relevant to this work is the advent of transposon insertion sequencing, in which large scale transposon libraries are coupled to next generation sequencing (van Opijnen and Camilli, 2013).

Transposons have been a significant driver of genetic diversity in all areas of life (Munoz-Lopez and Garcia-Perez, 2010). Generally, they can be described as mobile genetic elements that can move throughout a genome via a "cut and paste" mechanism, in that DNA is excised from one genomic location and inserted into another (Reznikoff, 2003). Transposons are defined by their terminal inverted repeats which flank a linear DNA sequence. In scientific application this middle region can be engineered to contain any DNA sequence, and is often made to contain antibiotic resistance encoding genes. In nature, this middle region generally contains the coding sequence for a transposase, which is the machinery that physically inserts the linear transposon into a genome. Two transposase molecules recognise and bind to the inverted repeats of the transposon. The two transposases then interact with each other to form a synapsis, which causes the middle transposon region to loop out. Then, the 3' strands are nicked by nucleophilic attack which requires the presence of magnesium and oxygen. This ultimately leads to the excision of the transposon, after which the remaining DNA is rejoined and repaired by the host.

There are multiple methodologies which can be used for transposon insertion sequencing, although all work on the same principle, which is to assess where in the genome the

transposons have been inserted. The technique works as follows (Fig. 1.10). First, a high density transposon library is created in the organism of interest. To do so, multiple aliquots of a single preparation of competent cells are subjected to single rounds of transposition and antibiotic selection, followed finally by the pooling of viable mutants. Successfully recombined, viable cells are selected by the resistance encoded within the transposon sequence. Second, DNA is isolated from the transposon library and prepared in such a manner as to allow compatibility with the sequencing platform of choice. This preparation is specifically designed to result in the generation of sequence reads that start from within the transposon immediately prior to either 5′ or 3′ end, across the transposon/chromosome junction and into genomic DNA. Finally, these sequence reads are processed to remove the transposon sequence and leave reads that can be mapped to the reference genome. Insertion sites are then calculated alongside a number of other metrics, which allows the description of where insertions have occurred with respect to the boundaries of genome features.

Transposon insertion sequencing data can be used in a number of ways. Primarily, it is a useful method in the determination of essential genes. Christen *et al*. (2011) used the technique to define the essential genes in *Caulobacter crescentus*, an important model organism in the study of the cell cycle. Another usage of this data is in the assessment of genes and their effect on fitness. Langridge *et al*. (2009), in one of the landmark transposon insertion sequencing papers, compared a *Salmonella Typhi* transposon library before and after passage in nutrient broth. There were examples of genes with either a decreased or an increased number of reported insertions, allowing their definition as genes either advantageous or disadvantageous for growth respectively. Similarly, transposon libraries can

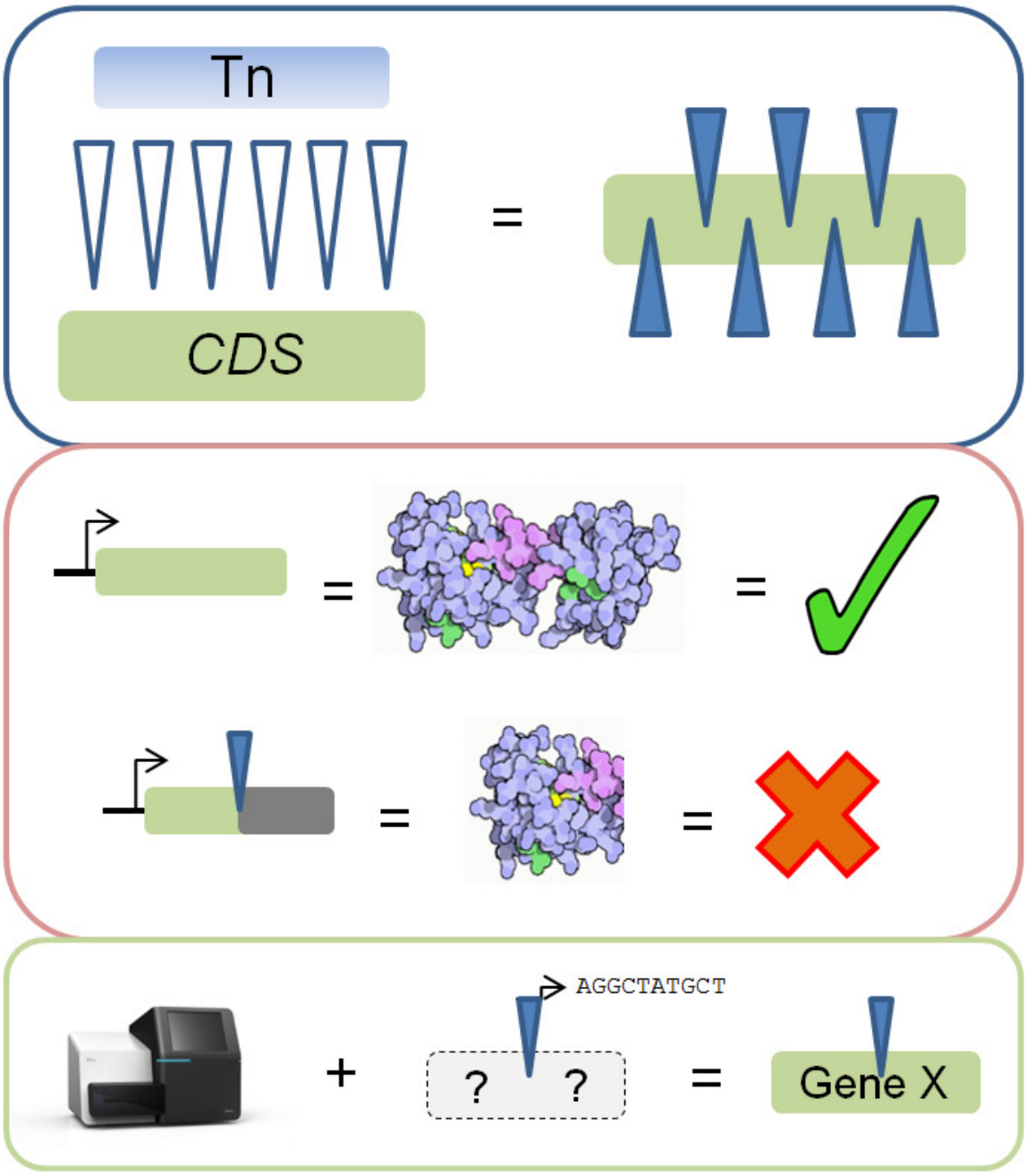

Figure 1.10. A simplified depiction of transposon insertion sequencing. (Top panel) Transposon mutagenesis can happen at any point along a coding sequence, in either orientation. (Middle panel) A wild type coding sequence encodes for a functional protein. A disrupted coding sequence produces a truncated, non-functional protein. (Lower panel) Reads are generated that read out of the transposon. Alignment of these reads to the genome allows for the precise location of the insertion.

be grown with and without the presence of environmental pressures and compared, to find which genes are related to the condition employed. For example, Phan *et al*. (2013) used human serum as a stress condition with *E. coli* ST131, a uropathogenic strain of clinical importance. From comparing growth with and without serum, they were able to define a serum resistome of 56 genes.

1.5 Aims

Despite all of the previously discussed studies, and the importance of *E. coli* K12 as a laboratory strain and a model organism, there has been no transposon insertion sequencing study undertaken in this lineage. Furthermore, the approaches used previously to investigate gene essentiality and link genotype to phenotype have well known flaws which we are now, due to recent advancements, able to be overcome. Most notably is the fact that knock out approaches have missed particular essential genes. As such, the broad aim of this work was to develop a complete transposon insertion sequencing methodology that covered almost every aspect of the workflow required, from the initial wet lab work, through to the preliminary data preparation and finally the requisite interpretation of the results. To enable direct comparisons between the insertion sequencing work and the KEIO library, *E. coli* strain BW25113 was the strain used to create the transposon library.

The work presented in this thesis can be split into three sections. The first aim was to adopt and develop a methodology for transposon insertion sequencing with application to *E. coli* BW25113. This work entails the development of both wet lab protocols as well as *in silico* data analysis from the ground up, and is utterly necessary as no protocols or data analysis frameworks were available to work from. The second aim was to use the chosen

methodology to generate datasets and use them to investigate the essential gene complement of BW25113, and compare the results to those of the KEIO study (Baba *et al*., 2006). Because the KEIO study is seen as the gold standard for the essential gene set of *E. coli*, it is logical to use as a comparator. The third section aimed to assess which genes are important for envelope structure, by using the methodology to compare insertional representation of the library before and after growth in the presence of vancomycin and SDS. The cell envelope is a crucial structure to the cell, and understanding which of the genes underpins it is important for advancements in many areas, most notably antimicrobial development.

CHAPTER 2

MATERIALS AND METHODS

2.1 Bacterial strains and primers.

An *E. coli* K12 strain designated BW25113, the parent strain of the KEIO library (Baba *et al*., 2006), was used as the host strain for the mutant library. All primers used in these methods are detailed in Table 2.1 below.

2.2 Transposon library creation

The transposon library used in this work was created by collaborators in Discuva, Cambridge. The library was created based on a method described by Langridge *et al*. (2009), in which transposomes are electroporated into the strain of interest (Fig. 2.1). The linear DNA fragment was amplified so as to contain the chloramphenicol resistance cassette from pACYC184 (Chang and Cohen, 1978). In this reaction, overhanging primers were used to introduce the inverted repeats at the terminal ends required for recognition and binding to the transposase (5'-CTGTCTCTTATACACATCTTTGGCGAAAATGAGACGTTG and 5'-CTGTCTCTTATACACATCTACCGGGTCGAATTTGCTTTCG). Upon electroporation of the transposomes into cells, the transposases then act to insert the linear transposon DNA sequence into the host cell genome at a random position. Successful mutants were selected for on chloramphenicol containing agar plates. Multiple rounds of electroporation and selection were done, and the successful mutants were pooled to form the transposon library used in the following experiments.

2.3 Two-PCR library preparation method

Christen *et al*. (2011) detailed the use of a wholly PCR-based library preparation method with a *Caulobacter crescentus* transposon library. Two sequential PCRs were used to

Table 2.1. Primers used in this work.

Primer name	Sequence (5’ -> 3’)	Description
TTc-nIx-seq1.1	TCTCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAA	Custom sequencing primer required for the 2-PCR method.
TTc-nIx-P1	TTATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAA	The forward primer of the 1st PCR used in the 2-PCR method.
TTc-nIx-P2a	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCGCCA	One of four reverse primers of the 1st PCR used in the 2-PCR method.
TTc-nIx-P2b	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCCAGC	One of four reverse primers of the 1st PCR used in the 2-PCR method.
TTc-nIx-P2c	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTGATG	One of four reverse primers of the 1st PCR used in the 2-PCR method.
TTc-nIx-P2d	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTGCTG	One of four reverse primers of the 1st PCR used in the 2-PCR method.
TTc-nIx-P3	AATGATACGGCGACCACCGAGATCTCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAA	The forward primer of the 2nd PCR used in the 2-PCR method.
TTc-nIx-P4	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	The forward primer of the 2nd PCR used in the 2-PCR method.
TTc-slx.6.1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTACGGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.7.4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTAGGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.8.2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATGCATGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.9.2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGATCGAGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.

TTc-slx.8.3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGCATGGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.9.3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGATCGATGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.6.3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTACGTAGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.7.2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAGCTGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.8.4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCATGCGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.9.4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATCGATCGTCTCATTTTCGCCAAAGATGTGTA	A forward primer used in the shearing method, or the 2nd PCR of the hybrid method.
TTc-slx.P1.F1	TCTTACGTGCCGATCAACGTCTCATTTTCGCC	The forward primer of the 1st PCR used in the hybrid method.
TTc-slx.P1.R	GATCGGAAGAGCACACGTCTGAACTCCAGTC	The reverse primer of the 1st PCR used in the hybrid method.

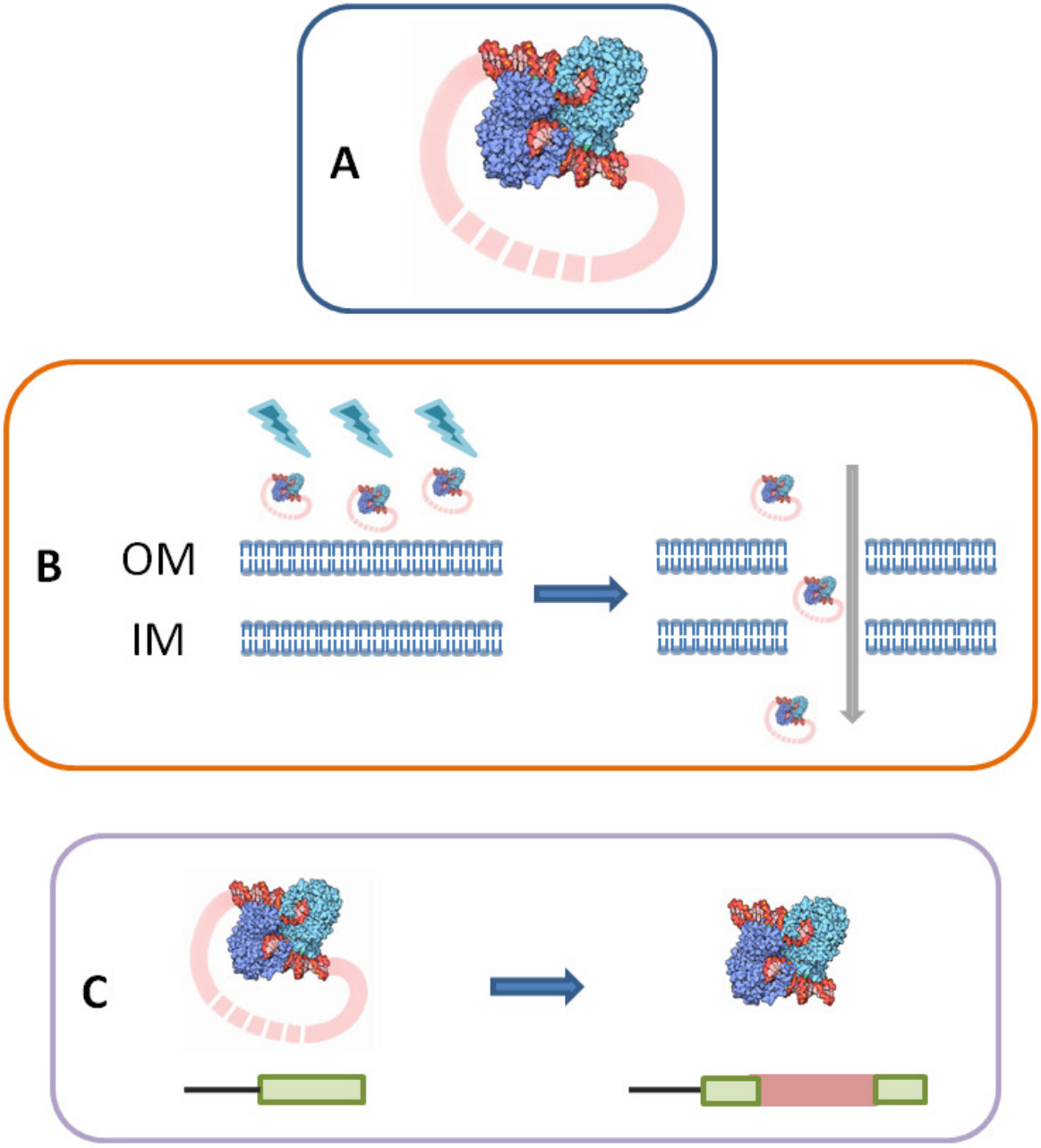

Figure 2.1. Transposome mediated mutagenesis. (A) A transposome. The linear DNA is shown in red, and the dimeric transposase is shown in blue. (B) Electroporation of cells allows the movement of transposomes to pass into the cytosol. (C) When inside a cell, the transposase inserts its attached DNA fragment into the host genome.

generate sequencing ready DNA libraries (Fig. 2.2). The first PCR used a forward primer that is complementary to transposon sequence, and amplifies outwards into the flanking genomic DNA. Within the amplified sequence an Illumina compatible PE 1.0 sequence was engineered into the transposon immediately before the 5′ terminal 2.1end. To complement the first primer a semi-arbitrarily random reverse primer is used. This primer is consisted of a 3′ pentanucleotide sequence, followed by a random 10-bp spacer and then a 5′ Illumina PE2.0 adapter sequence. Three variants of the reverse primers included differing pentanucleotide sequences, which were designed to bind to the genome every ~300 bp and so theoretically complement PCR fragments originating throughout the genome. These fragments were then amplified a second time. The forward primer in this reaction contained a 3′ PE 1.0 complementary sequence and a 5′ Illumina compatible adapter P5 incorporated into the fragments. The reverse primer contained 3′ sequence complementary to the PE 2.0 sequence previously incorporated into the fragments, and a 5′ Illumina compatible adapter P7.

2.3.1 Two PCR method adaptation. The protocol described above was adapted for use with our BW25113 transposon library (Fig. 2.2). Broadly the same steps were retained, but with changes at several steps. From here, the forward primers in the first PCR are known as P1 and P2A-D. The second round PCR primers are known as P3 and P4 (forward and reverse respectively in each case).

First, while Christen *et al*. use bacterial culture to provide the DNA template in the first PCR, genomic DNA was used instead. Next, given that the transposon used to create the transposon library used did not contain any Illumina-compatible sequence, the forward

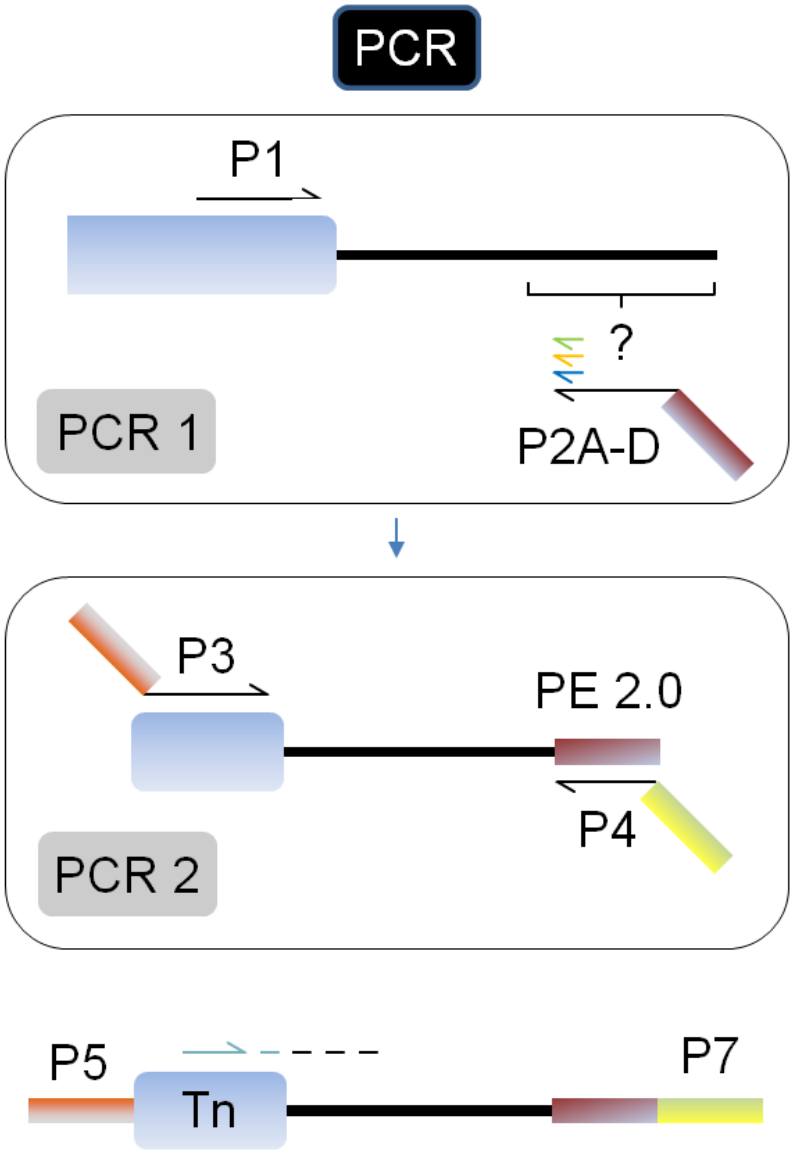

Figure 2.2. The adapted 2 PCR method used in this work. A transposon (blue) is inserted into genomic DNA. In the first PCR, the forward primer (P1) is complementary to the transposon, and the reverse primers (P2A-D) are semi arbitrarily random. In the second PCR, P3 recognises the transposon and P4 recognises PE 2.0 introduced in the first PCR. The final fragment organisation is shown at the bottom of the image. Each sequence read begins with 18 bases of transposon sequence (shown in dotted black line, with custom sequencing primer in dotted blue)

primer in the second PCR was designed to be complementary to transposon sequence at the 3' end, while still containing the necessary 5' P5 sequence. Additionally, this change means that a custom sequencing primer must be used during the sequencing runs. PE 1.0 is complementary to primers included in the sequencing kits, and the binding of these primers to PE 1.0 primes a read by allowing nucleotide incorporation by a DNA polymerase. With no such sequence between P5 and the transposon sequence, there is nowhere for a polymerase to initiate a sequence read. A custom primer, wholly complementary to transposon sequence, was designed to be added to the sequencing cartridges to allow read initiation during a run (seq 1.1 in Fig. 2.2). Use of this primer results in sequence reads starting 18 bases before the end of the transposon, followed by genomic DNA. In another change to the primer design, Illumina-compatible indexes were introduced into the primers used in the second PCR. The indexes were placed in between P5 and the transposon sequence in P3, and in between P7 and PE 2.0 in P4. By using indexes, multiple samples can be sequenced simultaneously, allowing for greater throughput and utility. To enable the use of the i7 index, another custom primer (index 1) was added into the sequencing cartridge. Where Christen *et al*. (2013) use 3 variants of the reverse primer in the first PCR, four variants were used in this work, corresponding to the four most common pentanucleotides present in the BW25113 genome. In contrast to Christen *et al*. (2013), who perform a gel based size selection and clean only after the second PCR, a 1.5x SPRI clean was used after both PCRs.

2.3.2 Two PCR method protocol. An overview is shown in Fig. 2.2. Genomic DNA was isolated from the BW25113 transposon library using a Qiagen QIAamp DNA Blood Mini Kit, according to the manufacturer's specifications. All primers were used at 20 µM. This

genomic DNA was then used in four separate PCRs, each identical except for the reverse primers used. 5 µl of genomic DNA, 1 µl of primer P1, 1 µl of primer P2A/B/C/D, 25 µl of MyTaq polymerase (Bioline) and 18 µl of deionised water were used in each reaction. PCRs were run on a Mastercycler Pro (Eppendorf) The cycling conditions for these reactions were as follows; 94 °C for 3 minutes, 6 cycles of 94 °C for 30 seconds, 42 °C for 30 seconds (with a slope of -1 °C per cycle), 72 °C 60 seconds followed by 25 cycles of 94 °C for 30 seconds, 58 °C for 30 seconds, 72 °C for 60 seconds and then finally followed by 72 °C for 3 minutes. All four PCRs were then pooled and cleaned using an Ampure XP SPRI bead based clean up step (Beckman Coulter). In this cleanup, a 2:3 ratio of PCR volume to bead volume was used, to remove DNA fragments shorter than 150 bp as per the manufacturer's instructions. From the SPRI cleaned pool, 2 µl was taken forward into the second PCR, along with 1 µl of primer P3, 1 µl of primer P4, 25 µl of MyTaq polymerase and 21 µl of deionised water. This was cycled as follows; 94 °C for 3 minutes, followed by 30 cycles of 94 °C for 30 seconds, 64 °C for 30 seconds, 72 °C for 60 seconds, and then finally followed by 72 °C for 3 minutes. Another SPRI clean was used after this PCR, with the same ratios as previously.

Samples were loaded on the Miseq (Illumina) to aim for an optimal cluster density of 800 clusters per mm^2. Qubit (Thermo Fisher Scientific) was used to quantify the sample concentration, and estimate sample loading volumes. Immediately prior to the sequencing run, 4 µl of the custom sequencing primer seq 1.1 at 100 µM was added to the 500 cycle V2 sequencing cartridges. The single read lengths of each run were set to 250 bp.

2.4 Shearing-based library preparation method

The two PCR method tested previously relies wholly upon PCR to generate fragments for sequencing that contain transposon/chromosome junctions. Another way of creating these sequencing libraries includes the use of mechanical shearing through ultrasonication. This process has been used in several transposon sequencing publications, notably Phan *et al*. (2013; Fig. 2.3). Genomic DNA is quantified and standardised to a given amount, and then subjected to ultrasonication. The next step is to repair the sheared fragments. Sonication leaves the DNA with 5′ and 3′ overhangs that are repaired to leave blunt ended, 5′ phosphorylated fragments. After repair, the newly formed blunt ends are A tailed at the 3′ ends to facilitate the next step of adaptor ligation, in which Illumina read one and two sequence-containing adapters are ligated to the A tailed fragments. A PCR step is then used to enrich fragments containing transposon/chromosome junctions at the same time as introducing the necessary P5 and P7 flow cell binding Illumina sequences. A long forward primer is then used to enrich junction containing fragments of a defined structure. From 5′ to 3′, the primer consists of the Illumina P5 and read 1 sequences, followed by an in-line barcode and 25 bases complementary to the transposon. During the enrichment, the 3′ transposon complementary end of the primer binds to the transposon and subsequently introduces the prior sequences into the fragments.

2.4.1 Adaptation of the shearing-based method. The method outlined in Phan *et al*. (2013) was adapted for use with our BW25113 library. As with the Christen *et al*. (2011) method, some slight modifications were made. An overview is shown in Fig. 2.3. The previous method used a Covaris ultrasonicator to break down genomic DNA. In this work, the Bioruptor platform was used. While the two technologies have slight technical differences, the

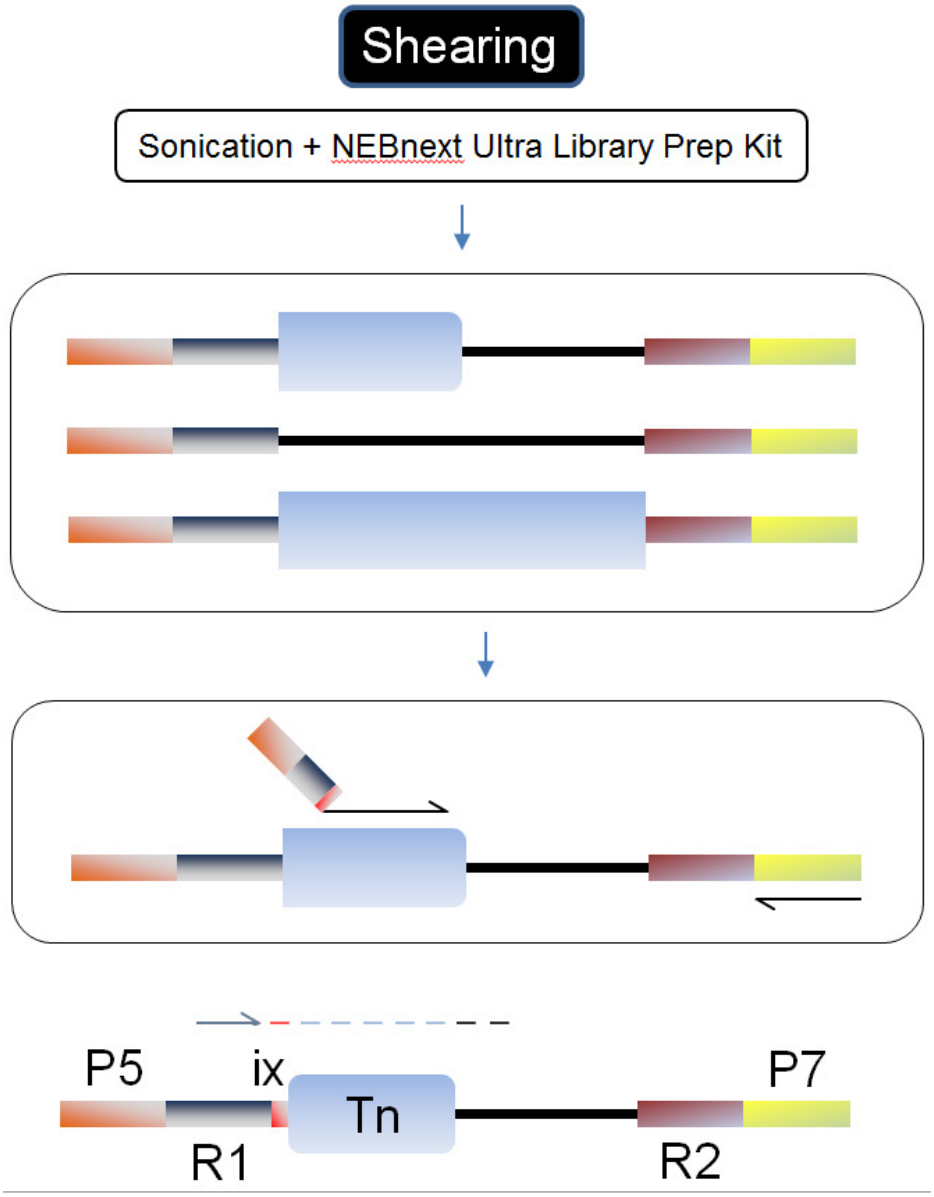

Figure 2.3. The adapted shearing based method used in this work. DNA is first sonicated and processed using the NEBnext kit. An enrichment PCR is done to specifically amplify only the fragments containing the 3' end of the transposon (shown in blue). The final fragment structure after processing is shown at the bottom of the image.

principle of mechanical shearing of the DNA is the same in both. In the Bioruptor, ultrasound waves are pulsed through an ice bath in which the gDNA samples are immersed. The propagation of these waves, through cavitation, creates mechanical stresses that break apart the DNA into smaller fragments. This process can be tuned to give a reproducible range of DNA fragments of a given average length. An average fragment length of between 200-300 base pairs was targeted. Phan *et al*. used an Illumina Truseq kit for the next steps of end repair, A tailing and adapter ligation. In contrast, an NEB NEBNext Ultra kit was used in this work. In this kit, the steps taken are identical, albeit with a slight difference in adapter ligation. The adapters included in the NEBNext kit have a hairpin loop structure with a uracil at the centre of the loop. An extra step in the kit protocol is to excise the uracil to leave linear DNA ligated to the gDNA fragments.

In the enrichment PCR, two adaptations have been made. First, in the previous method the inline barcodes in the forward primers are all the same length. Here, the forward primers were designed to have staggered inline barcodes. The purpose of this primarily is to increase base diversity during sequencing. Immediately after the inline barcode is the expected transposon sequence. While barcodes of a given length can be designed to have different sequence, and so have maximal diversity during sequencing, the following base calling of the transposon sequence will be identical during the imaging of each cluster on the flow cell. This low diversity makes it harder for the sequencer to differentiate between clusters and subsequently negatively impacts cluster definition, base calling and read quality. The staggering of the inline indexes then leads to the staggering of the transposon sequence immediately after, increasing the base diversity at every cycle. This theoretically leads to better cluster definition and higher quality base calling. Another

adaptation made to the previous method is to improve the multiplexing potential of the technique. Previously, the only way to multiplex samples on a single run was by using the inline barcodes. Here, Illumina compatible indexes have been introduced into the fragments through the enrichment PCR by using NEBNext reverse primers. When used in conjunction with the custom enrichment forward primers, the dual indexing of samples is facilitated. This is another major benefit to the use of staggered inline barcodes, in that the capacity for multiplexing is greater with a wider variety of inline barcode lengths and complexities available.

2.4.2 Shearing method protocol. Genomic DNA was isolated from the BW25113 transposon library using a Qiagen QIAamp DNA Blood Mini Kit, according to the manufacturer's specifications. Following isolation, the DNA was quantified using the Qubit platform. 1 µg of DNA in a volume of 500 µl was then sheared to an average fragment length of 250 bp using the Bioruptor sonication device (Diagenode), following the manufacturer's instructions. 15 shearing cycles, consisting of 30 seconds on at the low setting, following by 90 seconds off, were used. The 500 µl sheared volume of DNA was concentrated down to approximately 55.5 µl using a Concentrator 5301 (Eppendorf). At this point, the concentrated DNA was processed using the NEBnext DNA library preparation kit (New England Biolabs). The steps of end repair, 5' phosphorylation, adapter ligation and USER excision were done following the instructions provided. In the following amplification step, custom designed primers were used to specifically enrich fragments containing transposon/chromosome junctions. These reactions contained 25 µl 2X Hifi polymerase mix (KAPA Biosystems), 2.5 µl of custom enrichment forward primer at 10 µM, 2.5 µl of the standard NEBnext Illumina reverse primer

and 20 µl of the NEBnext processed DNA. Indexes were present within both primers of this reaction, and each sample used a different variant of each primer to give uniquely identifiable indexes for each sample. This reaction was temperature cycled for the following; 98 °C for 48 seconds, followed by 22 cycles of 98 °C for 15 seconds, 60 °C for 30 seconds and 72 °C for 30 seconds, followed by 72 °C for 1 minute. This reaction was also SPRI cleaned using a 2:3 reaction volume to bead ratio. The resulting cleaned DNA was quantified using qPCR with a SYBR FAST kit (KAPA), following the manufacturer's instructions.

The processed, quantified samples were loaded on the Miseq to aim for an optimal cluster density of 800 clusters per mm^2. 150 cycle V3 sequencing cartridges were used for these sequencing runs.

2.5 Hybrid shearing-based library preparation method

In addition to the 2 PCR and shearing methodologies tested, a hybrid of the two was also evaluated (Fig. 2.4). This method is centred upon the use of DNA ultrasonication as in Phan *et al*. (2013), but instead of only a single PCR enrichment, two PCRs were used as in Christen *et al*. (2011). The rationale behind this design is that of increased specific transposon enrichment. After the ligation of adapters to the repaired, sheared genomic DNA, two PCRs are done. In the first of the 2 PCRs, the forward primer is entirely complementary to transposon sequence, and the reverse primer is entirely complementary to Illumina specific sequence incorporated through the ligation reaction. This is in contrast to the 2PCR method, in which semi-arbitrarily random reverse primers are used to complement the transposon specific forward primer. This step should raise the number of fragments containing transposon/chromosome junctions against the background of fragments that do

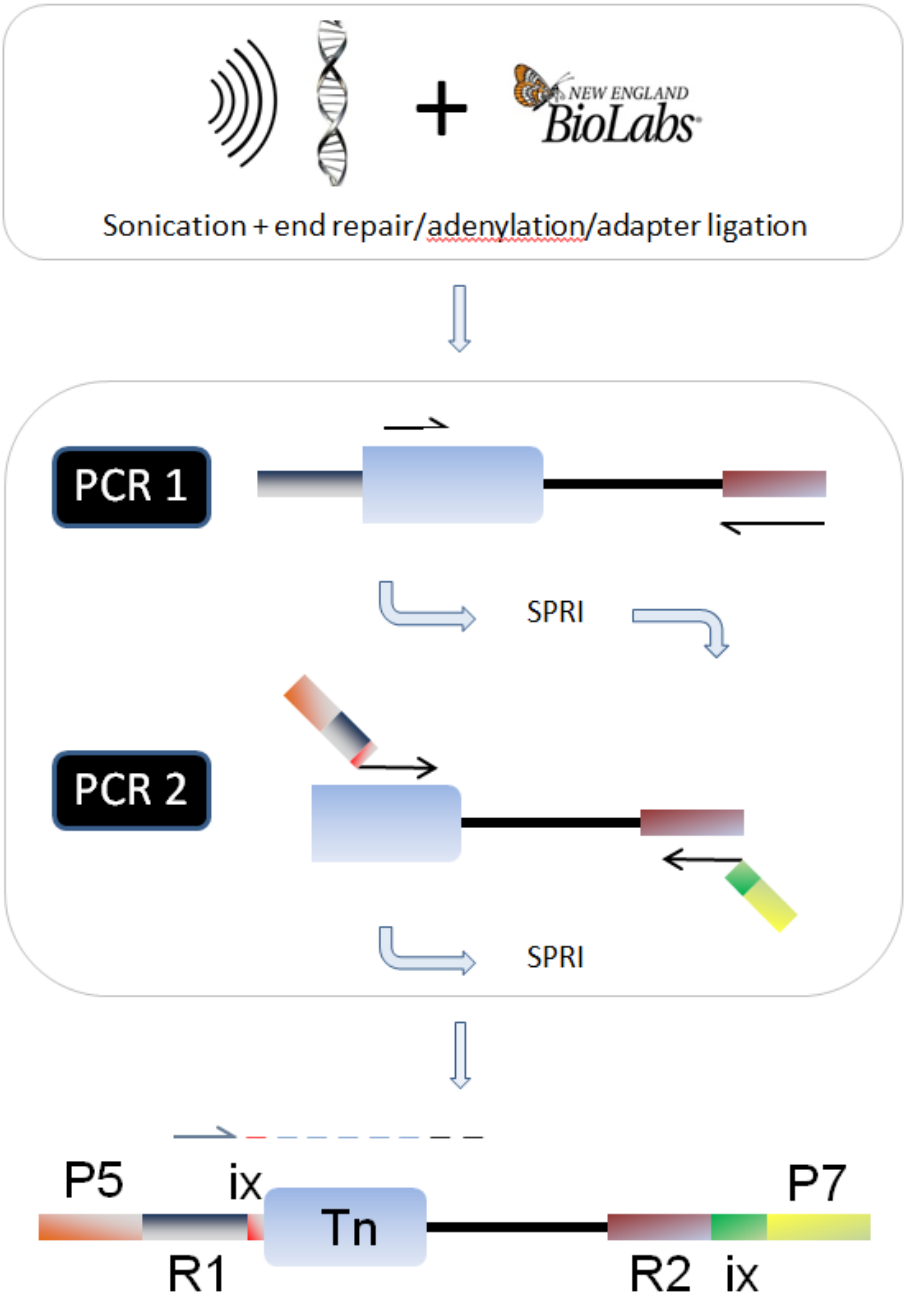

Figure 2.4. The hybrid based method used in this work. DNA is first sonicated and processed using the NEBnext kit. The first PCR uses primers specific to the transposon (blue) and Illumina specific sequence introduced through the NEBnext kit. After SPRI cleaning, the second PCR introduces more requisite Illumina compatible sequences. The final fragment structure is shown at the bottom of the image.

not contain them, and so act to improve the effectiveness of the second PCR. The second PCR is then the same as the enrichment PCR used in the shearing method previously tested, resulting in Illumina-compatible fragments.

2.5.1 Adaptation of the hybrid method. The genomic DNA shearing, end repair, A-tailing, adapter ligation and magnetic bead cleaning steps are the same in the hybrid method as in the shearing method. One slight change has been made at the shearing stage: whereas previously there were 15 cycles, there were 13 in this protocol. This is to try and reduce the number of reads lost due to being shorter than the minimum read length of 20. After these steps, the first PCR is done. The primers for this reaction were designed to have melting temperatures as close together as possible. Additionally, the forward transposon specific primer was designed to anneal to the transposon 11 bases upstream of where the 2nd PCR forward primer does, thus nesting the second reaction. In the second PCR, the forward and reverse primers used previously were used again, meaning that inline indexing was still available alongside Illumina indexing. An additional 0.75x SPRI bead cleanup was used inbetween the two PCRs.

2.5.2 Hybrid method protocol. The hybrid methodology is identical to the shearing one, up until the final steps of library amplification. Immediately after the final step of the NEBnext library preparation, the first of two PCRs was done. This reaction contained 25 µl of Hifi polymerase, 2.5 µl of TTc-sIx.P1.F1 at 10 µM, 2.5 µl of TTc-sIx.P1.R at 10 µM, 14 µl of NEBnext processed sample and 6 µl of deionised water. This reaction was cycled as follows, at 98 °C for 48 seconds, followed by 10 cycles of 98 °Cfor 15 seconds, 65 °C for 30 seconds,

72 °C for 30 seconds and then followed by 72 °C for 60 seconds. The resulting PCR mixture was SPRI cleaned with a 1:0.8 ratio of PCR to beads. The second PCR then consisted of 25 µl of Hifi polymerase, 2.5 µl of custom forward enrichment primer at 10 µM (the same primers used in the shearing methodology), 2.5 µl of the standard NEBnext Illumina reverse primer, 15 µl of the SPRI cleaned first PCR sample and 6.5 µl of deionised water. As used previously in the shearing method, different primer combinations were used to allow for multiplexing. This reaction was cycled with the same temperatures as the first PCR, but for 20 cycles instead of 10. A 1:0.8 SPRI clean was repeated on the samples after the second PCR.

At this point, samples were quantified using the KAPA qPCR kit as per the manufacturer's instructions. Quantified libraries were then loaded on the Miseq to aim for an optimal cluster density of 800 clusters per mm^2. 150 cycle V3 sequencing cartridges were used for these sequencing runs.

2.6 Sequence read analysis

For the analysis detailed in 2.5, Ubuntu 12.04 was used as the host environment.

2.6.1 Preliminary read processing. The raw sequencing reads produced in this work, by design, contained transposon specific sequence at their beginning. Specifically, these bases were from the very 3' end of the transposon. As such, these reads required processing to assess and remove these sequences. Between the three preparatory methods tested, there were differences in the structures of the reads generated. Reads from the 2-PCR method had 18 bp of 5' transposon sequence. The shearing method leads to reads with 35 base pairs of 5' transposon sequence. The hybrid method leads to 35 base pairs of transposon sequence, but with an additional variable length inline index upstream of it. The Fastx barcode splitter

and trimmer tools, as part of the Fastx toolkit (Pearson *et al*., 1997), were used to assess and trim the sequences. For the 2-PCR method, reads were only retained for further processing with 1 mismatch to the expected sequence of 5'-GATGTGTATAAGAGACAG allowed. For the shearing and hybrid methods reads were first filtered by their inline indexes, and no mismatches were allowed. Then, the transposon similarity matching was done in two parts. For the first 25 bases from the 5' end, 3 mismatches were allowed, at which point the 25 bases were trimmed. Then, 1 mismatch was allowed for the remaining 10 bases, prior to trimming of the 10 bases.

2.6.2 Primary read processing. Individually, all three sets of reads, from each preparatory method, were brought forward from preliminary processing and run through the same set of analytical steps in a script. Reads less than 20 bases long were removed using Trimmomatic (Bolger, Lohse and Usadel, 2014). Length filtered reads were then aligned to the reference sequence for *E. coli* W3110 (NC_007779), the parent strain of BW25113, which was obtained from the NCBI genome repository (Tatusova *et al*., 2014). The aligner bwa was used, with the mem algorithm (0.7.8-r455, Li and Durbin, 2009). Next, the aligned reads were filtered to remove any soft clipped reads. The subsequent steps of conversion from sam files to bam files, and the requisite sorting and indexing, were done using samtools (0.1.19-44428cd, Li *et al*., 2009). Next, the bedtools suite (Quinlan and Hall, 2010) was used to, from the bam files created previously, create bed files and then intersect them against the coding sequence boundaries defined in general feature format (.gff) files obtained from NCBI. Custom python scripts were then used to ensure that only reads that correctly emanated from within a coding sequence were retained, along with multiple other steps of sorting and processing.

The metrics reported in each chapter were obtained from all of the files created during this analysis.

2.7 Essential gene prediction

This is done as described in Langridge *et al*. (2009). Briefly, the distribution of insertion indices is bimodal with the mode containing insertion index 0 corresponding to an essential model. The cut-off between the two modes is chosen to be the minimum bin in the appropriate range of insertion indices, in general between 0 and 0.02. Gamma distributions are fitted to each mode in each data set using the R MASS library (R Development Core Team, 2016). Log2-likelihood ratios are calculated between the two distributions for each gene. We call a gene essential if it has a log2-likelihood ratio of less than -3.6, corresponding to the gene being at least 12 times more likely to belong to the essential distribution than the non-essential distribution. A gene is deemed non-essential if it has log2-likelihood greater than 3.6.

2.8 Differential representation calculation

DESeq2 was used to detect the differential representation of genes in the insertion sequencing datasets, with and without the presence of SDS or vancomycin (Love, Huber and Anders, 2014). The numbers of insertion sites in each gene, and also the numbers of reads emanating from within each gene, were individually compared between the control and test condition datasets. DESeq2 is replicate aware, and so each replicate of the control and test datasets were used in the calculation of $\log_2$ fold change values (L_2FCs) and also adjusted *p* values. Genes with a less than two fold change in either direction, and/or an adjusted *p*

value greater than 0.05, were removed from further analysis. In brief, DESeq2 assesses the variability between datasets in addition to the variability between replicates, to be able to report differentially represented genes that are more likely to be genuine. To assess the differential expression of genes, negative binomial linear models are used. The data used in these models are subjected to normalisation to account for differences between datasets. Additionally, variance within the replicates is accounted for in the final analysis.

CHAPTER 3

A COMPARISON OF TRANSPOSON SEQUENCING LIBRARY PREPARATION METHODS

3.1 Introduction

The fundamental aim of transposon sequencing is to generate sequence reads originating from within a transposon insertion and continuing across into adjacent genomic DNA. Subsequent processing of these reads then allows the precise inference of where insertions are located in the genome. In order to generate a DNA fragment library that is ready for sequencing, from the genomic DNA of a bacterial insertion library, two requirements must be met. First, fragments must contain transposon/chromosome junctions, and secondly, fragments must be compatible with the sequencing technology to be employed. To meet these requirements, the preparation of the gDNA must be highly specific and suitably designed.

There is no single methodology to achieve this aim: multiple publications detail different library preparation methods (as reviewed by van Opijnen and Camilli, 2013). Between them, no one methodology is distinguished in terms of performance, and furthermore no method comparisons are available in the literature. As such, it was decided to test multiple methodologies and to assess and compare their outputs. The following sections compare three library preparation strategies that were tested by applying them to an *E. coli* BW25113 transposon library.

3.2 Results

3.2.1 A two-PCR based library preparation method. Christen *et al*. (2011) demonstrated a library preparation based upon the use of PCR to generate Illumina-compatible fragments spanning transposon/chromosome junctions. Their method was adapted as detailed in chapter 2.

In order to assess the two-PCR preparation method, the transposon library was tested with two types of sample, each with two biological replicates. The first set of samples were gDNA derived from the neat transposon library (NTL) without any further growth. To produce the second set of samples (LB), 50 ml of LB inoculated with 10 µl of transposon library (to a starting OD_{600} of ~0.05) were grown at 37°C to an OD_{600} of 1, at which point gDNA was extracted. The reason for using two types of sample was to determine whether the addition of a growth step results in better representation of insertion sites. It was plausible that the neat transposon library might contain non-viable mutants that, while not capable of growth under the test conditions, were still present at the point of harvesting. The presence of such mutants could lead to insertion sites being erroneously reported in essential genes.

The four genomic DNA samples were processed using the adapted two-PCR method, and the resulting libraries were sequenced using the Illumina MiSeq. After sequencing, the reads were analysed using the analytical pipeline outlined in the materials and methods (Table 3.1). As a general figure, we aimed for approximately 10 million reads per individual sample. The numbers of reads obtained varies widely in different studies, from between 7-11 million reads per sequencing run from Langridge *et al*. (2009) and over 100 million raw reads in Christen *et al*. (2011). For the NTL replicates, 14 million (NTL1) and 12 million (NTL2) raw reads were obtained, respectively. For the LB replicates, 10 million (LB1) and 8.9 million (LB2) reads were obtained, respectively. Broadly, across all four samples a relatively small proportion of the raw reads could be included in the final dataset. The attrition of reads occurred at nearly every individual analytical step. After the first step, in which the similarity of the first 18 bases of each read to the expected transposon sequence was tested, between

Table 3.1. Dataset metrics from the two-PCR method.

Condition	Total raw reads	Reads with 18 bp tn sequence [1]	Reads < 20 bases	Mapped reads before clipped read filtering	Genome wide insertions	CDS insertions	Mapped reads after clipped read filtering
NTL1	14696851	6086629 (41.4% raw reads)	0	5211644 (35.5% raw reads)	240137	208866 (87% of total insertions)	2294800 (15.6% raw reads)
NTL2	12248551	7555700 (61.7% raw reads)	0	6917556 (56.5% raw reads)	298688	260843 (87.3% of total insertions)	2611531 (21.3% raw reads)
LB1	10600951	5774146 (64.5% raw reads)	0	5336935 (50.3% raw reads)	276880	241303 (87.2% of total insertions)	2389117 (22.5% raw reads)
LB2	8951965	4359408 (48.7% raw reads)	0	3549342 (39.6% raw reads)	227424	199422 (87.7% of total insertions)	1439687 (16.1% raw reads)

[1] 1 mismatch allowed

~35 and ~59% of reads were filtered out because they did not match the expected transposon sequence. After reads passing this filter were trimmed of the transposon sequence, they were filtered to ensure a minimum read length of 20 bases. At this point no reads were removed from any data set. Then, the reads were mapped to the *E. coli* W3110 genome. After mapping, a small percentage of reads were discarded because they did not align to the reference genome. This is in line with observations from similar studies, such as Langridge *et al*. (2009). During mapping, the bwa read aligner removes (clips) poor quality bases from some reads, to allow the better mapping of the rest of the read. For stringency, any clipped reads were removed from the datasets, which reduced the final number of mapped reads by more than half in each case.

Following read processing and mapping, the position of each insertion site on the chromosome was determined. In terms of unique insertion sites, the four datasets gave broadly similar results. Approximately, 230-300 thousand unique insertions were reported across the genome. The majority of these insertions (circa 87%) reside in coding sequences.

The next step in the analysis was to estimate the number of essential genes. One method to predict gene essentiality from transposon sequencing data is to calculate and manipulate insertion indexes, as employed by Langridge *et al*. (2009) and Phan *et al*. (2013). An insertion index is the frequency of unique insertions in a coding sequence normalised for its length. An insertion index of 0 indicates that no insertions were found in a coding sequence, and the greater the index, the greater the frequency of insertions. Histograms were generated for the insertion indexes of each coding sequence. Intervals were created over the range of insertion indexes, and the insertion indexes are then separated into each interval (hereafter referred to as bin). The bimodal distributions of the histograms correspond to essential and non-essential coding sequences, with smaller indexes indicating more likely essentiality.

Insertion indexes were calculated for each of the four datasets, and plotted in histograms (Fig. 3.1). In both plots, bimodality can be observed. The first peak at an insertion index of 0 corresponds to the essential coding sequences and, in both LB and NTL plots, contained the largest number (between ~170 and ~250) of coding sequences out of all the bins. The distribution decreased drastically towards an increased insertion index value. The rightmost second mode was less similar between the LB and NTL plots. In the LB datasets, the mode for the second distribution is at approximately 0.03, whereas in the NTL datasets the mode lies at approximately 0.04 (Fig. 3.1). These second modes are similar in both plots in that they tail off to greater insertion indexes less dramatically than the first distribution, although this tailing occurs at smaller insertion indexes in the LB plot than in the NTL plot. For both types of sample, the replicates were broadly similar. Between the LB replicates, there was a slightly higher number of coding sequences with the smallest insertion index in the second replicate along with a greater proportion of coding sequences in bins between 0 and ~0.04. Past this point, increased bin frequencies were found for the first replicate relative to the second replicate. Between the NTL replicates, a similar pattern was observed, with an increased representation in the smallest insertion indexes, and additionally between 0 and ~0.03. Beyond this point, the bins for the second replicate contained more coding sequences.

The reproducibility of the insertion indexes obtained for both replicates was assessed. For this purpose, a coefficient of determination (R^2) can be calculated. This statistic is a measure of how well two datasets correlate with each other, and it is normally used to compare how well data fit to a model. In this case, the R^2 value was calculated for each of the replicated sets of insertion indexes for each sample (Fig. 3.2). The R^2 values for the LB

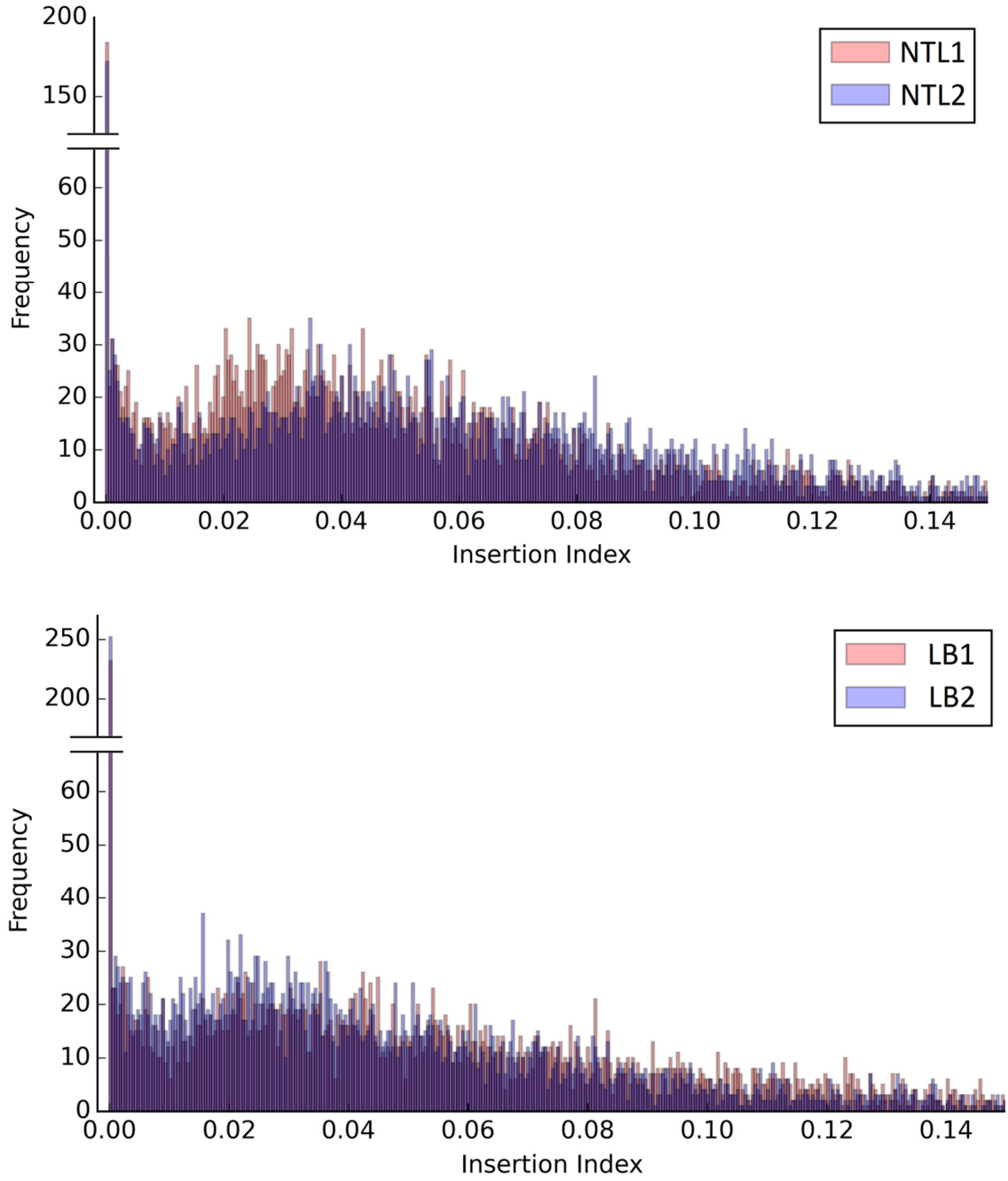

Figure 3.1. Histograms of insertion indexes calculated from datasets generated using the two-PCR methodology. Insertion indexes greater than 0.15 were omitted. The two panels show the insertion indexes for the two replicate samples of the neat transposon library (NTL, upper panel) and for the two replicate samples after growth in LB (LB, lower panel). In each panel, the first replicates are shown in pink and the second replicates are shown in blue. Each vertical bar on the x axis represents a different bin containing insertion indexes of equal, defined intervals. The y axis denotes the frequency of each of these bins.

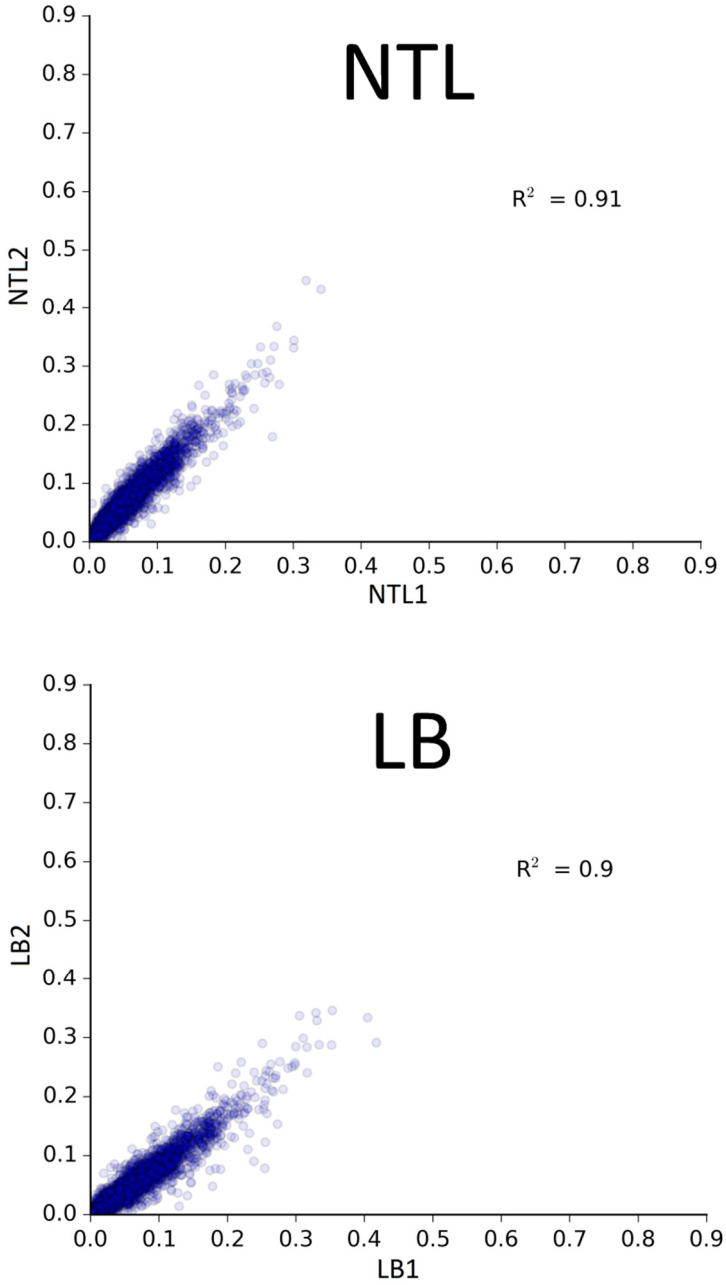

Figure 3.2. Insertion index correlation scatterplots for the datasets generated using the two-PCR method. For each sample, the insertion indexes calculated for every W3110 coding sequence for each replicate were plotted against each other, and a coefficient of determination (R^2) was calculated. The max R^2 value is 1, which would indicate a perfect positive correlation.

and NTL datasets were 89.7% and 91.3%, respectively. These high values indicate that the replicates correlate well with each other in terms of insertion indexes.

Insertion index histograms can be used to statistically assess gene essentiality. This was an approach taken by Langridge *et al*. (2009) and Phan *et al*. (2013). Essentiality predictions were made as described in the materials and methods. 486 coding sequences were predicted to be essential from the combined LB1 and LB2 datasets. 467 essential coding sequences were predicted from the combined NTL1 and NTL2 datasets. These numbers of predicted essential genes are substantially higher than other estimates; for example, the KEIO library originally outlined 303 essential gene candidates (Baba *et al*., 2006). Furthermore, this method identified genes as essential that have been proven not to be essential e.g. *cspBEHI*. As such, it was necessary to evaluate other library preparation methods to see if other techniques could provide more accurate estimations of essential genes.

3.2.2 A shearing based library preparation method. Phan *et al*. (2013) previously used transposon insertion sequencing to predict the serum resistome of *E. coli* ST131. They used a shearing-based library preparation method to create Illumina compatible sequencing libraries. The same gDNA samples used for the assessment of the two-PCR method, described in the previous section, were used to assess the shearing-based methodology. This was adapted as detailed in the materials and methods. Sequencing libraries were prepared for two biological replicates of the transposon library after growth in LB and were compared with a single sample of the neat transposon library itself. (Table 3.2). As observed in the analysis of the two-PCR data, the number of usable reads decreased at each stage of processing. However, there were key differences between the two techniques. First, note

Table 3.2. Metrics from the shearing method dataset.

Condition	Reads with matching inline barcode [1]	Reads with 1st 25 bp tn sequence [2]	Reads with 2nd 10 bp tn sequence [3]	Reads < 20 bases	Mapped reads before clipped read filtering	Genome wide insertions	CDS insertions	Mapped reads after clipped read filtering
ntl	13801288	8922192 (64.6% raw reads)	615729 (4.5% raw reads)	68994 (0.5% raw reads)	423753 (3.1% raw reads)	162396	139560 (86% of total insertions)	411187 (3% raw reads)
lb1	11416046	7607065 (66.6% raw reads)	736402 (6.5% raw reads)	86660 (0.76% raw reads)	568275 (5% raw reads)	202610	173814 (85.8% of total insertions)	554417 (4.9% raw reads)
lb2	10193801	8202090 (80.5% raw reads)	895680 (8.8% raw reads)	93688 (0.92% raw reads)	683565 (6.7% raw reads)	291594	250922 (88.1% of total insertions)	663654 (6.5% raw reads)

[1] 0 mismatches

[2] 3 mismatches allowed

[3] 1 mismatch allowed

that due to differences in the primer design, the data from the two methods were not processed identically. The first test assessed similarity over the first 25 bases of transposon sequence, corresponding to the sequence present in the forward primer of the enrichment PCR. The second test was used for the final 10 bases of transposon sequence immediately after the previous 25. There was a large difference in attrition between the two transposon similarity tests. The first test resulted in the loss of ~65-80% of the total. The second step resulted in only 4.5-9% of the sequences being taken forward to the next processing step. This was in stark contrast to the datasets generated by the two-PCR method, in which ~41-65% of the raw reads were carried through. Additionally, more reads were lost after the minimum read length filter in the shearing data than in the two-PCR data. Whereas none were lost at this step in the two-PCR data, between ~69 and ~94 thousand reads were too short to be carried forward in the shearing data. Only ~3-7% of the raw reads were mapped prior to being filtered for clipping. The numbers of unique insertion sites between the three datasets are broadly similar. Between ~1.6 and ~2.9 x 10^5 insertion sites were reported across the genome, of which ~86-88% arose in coding sequences. The numbers of unique insertions reported for the shearing method datasets were generally lower than those reported for the two-PCR datasets. The datasets from both preparation techniques show a very similar proportion of insertion sites within coding sequences. There was a clear difference in the number of clipped reads generated from both preparatory techniques: during processing, more than 75% were clipped and removed in each two-PCR dataset. In contrast, only ~3% were removed at the same step in the shearing datasets.

Insertion indexes were calculated for the three shearing datasets and plotted in histograms (Fig. 3.3). Slightly different insertion index profiles can be seen in the replicate LB datasets (lower panel of Fig. 3.3). While bimodality can be observed in each replicate,

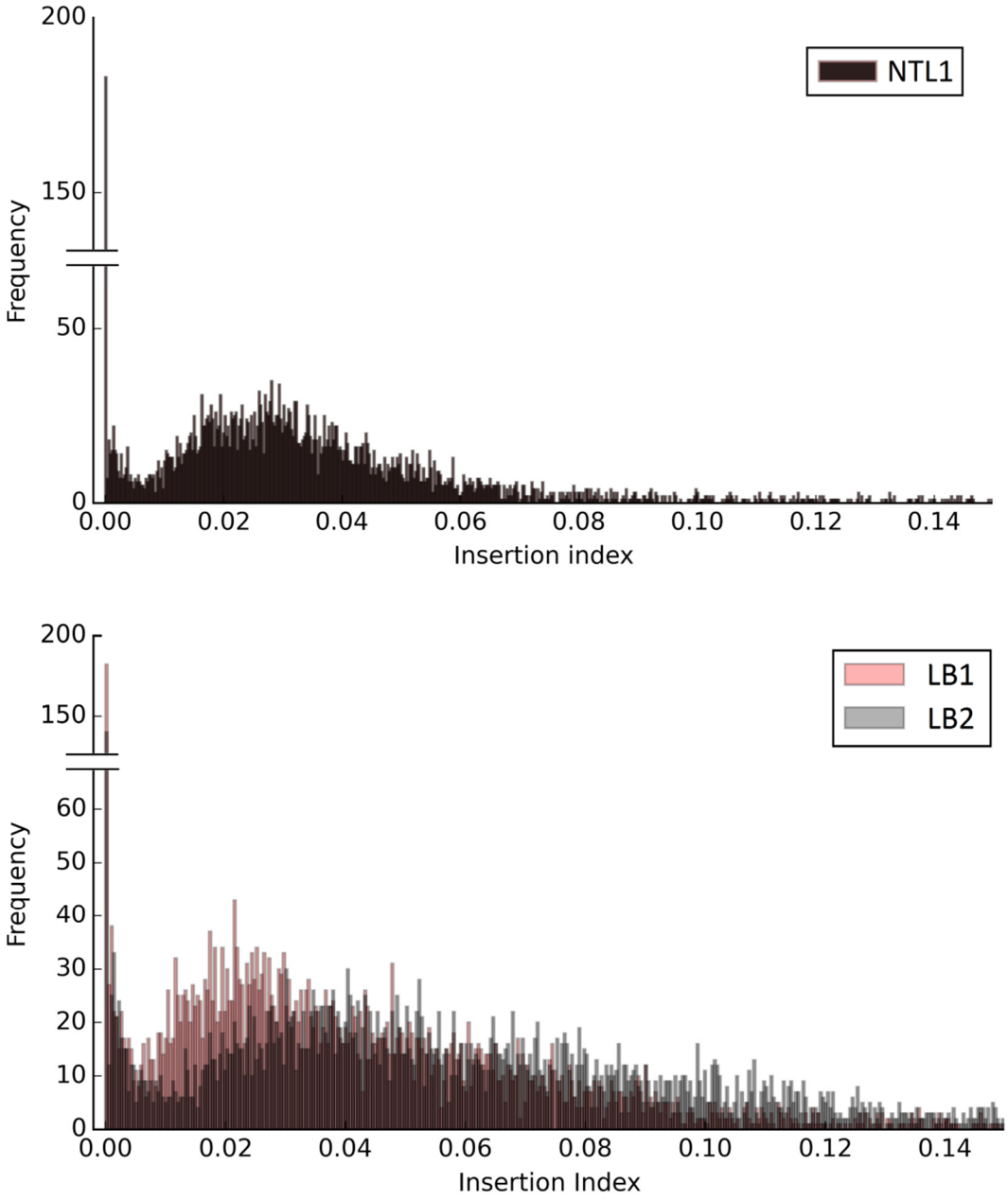

Figure 3.3. Histograms of insertion indexes calculated from datasets generated using the shearing methodology. Insertion indexes greater than 0.15 were omitted. The upper panel shows the insertion indexes for NTL dataset, and the lower panel shows the two LB1 and LB2 replicate samples.

between the replicates there are differences in each mode. The first peak at an insertion index of 0 can be observed in both replicates. There are ~180 coding sequences in the leftmost bin for the first replicate and approximately 140 in the second replicate. In the first replicate (LB1) the second mode is found at a smaller insertion index (centring at ~0.02 as opposed to ~0.04) and contains bins with a greater frequency of coding sequences. These differences between the replicates can be explained simply by the numbers of reads for each replicate. The second replicate generated over 110,000 more reads than the first. With an increasing number of reads in a dataset, there is a greater chance of finding more unique insertion sites. This decreases the number of coding sequences without insertions (barring essential genes) and increases the insertion indexes of the non-essential coding sequences. This corresponds to what is observed in the histogram: fewer coding sequences were collected in the leftmost bin of the LB2 dataset, with a greater spread of coding sequences across higher insertion indexes in the right mode.

Several observations can be made by comparing the LB histograms from the two-PCR and shearing methods. First, the two-PCR data do not appear to be as distinctly bimodal as in the shearing data. The split between the two modes is much easier to discern in both LB replicates of the shearing data. Second, a greater frequency in the leftmost bin of the two-PCR data can be seen when compared to the shearing data. This is especially important when considering the differences in the numbers of reads, with approximately 5×10^5 reads in the shearing data and then 1.5×10^6 reads and upwards in the two-PCR datasets. Third, the right modes are broadly similar in the data from both methods, centring at an insertion index between 0.02 and 0.04.

Although there is only a single dataset for the NTL sample, the insertion index histogram produced is very similar to those produced for the LB samples (upper panel of Fig.

3.3). This dataset contained fewer reads (~400,000 reads) than the two LB replicates. In keeping with the idea that an increased number of reads will move the right mode towards increased insertion indexes, the right mode is closer to that of the first LB replicate than the second. In comparison to the NTL histograms from the two-PCR data, the insertion index profiles look very similar, with similar coding sequence frequencies in each mode. Again, this is notable given the discordance in the numbers of reads in each dataset (between 2.2 and 2.6 x 10^6 for the two-PCR datasets and ~400,000 for the shearing dataset).

The reproducibility of insertion indexes between each LB replicate was assessed. The coefficient of determination was calculated for the replicates and the resulting plot is shown in Figure 3.4. The insertion indexes for the coding sequences in each replicate are very similar, as shown by the high coefficient of determination (R^2 = 0.95). This value is higher than those found from the LB and NTL two-PCR datasets.

The two LB replicate datasets were combined and the number of essential genes was predicted. The 374 coding sequences predicted to be essential after this analysis is ~100 less than the number predicted from the two-PCR method, but still contained some genes known to be non-essential, e.g. *aceEF*.

Having considered the high level of read attrition, observed after the second transposon similarity testing step, we considered this technique unsuitable for wider use. As such, another technique was tested to create sequenceable fragments spanning the transposon/chromosome junction.

3.2.3 A hybrid two-PCR/shearing-based library preparation method. Elements of both the two-PCR and shearing methods were used in the final preparation method tested. DNA was first sheared and then amplified using two PCR enrichment steps. From here onwards this

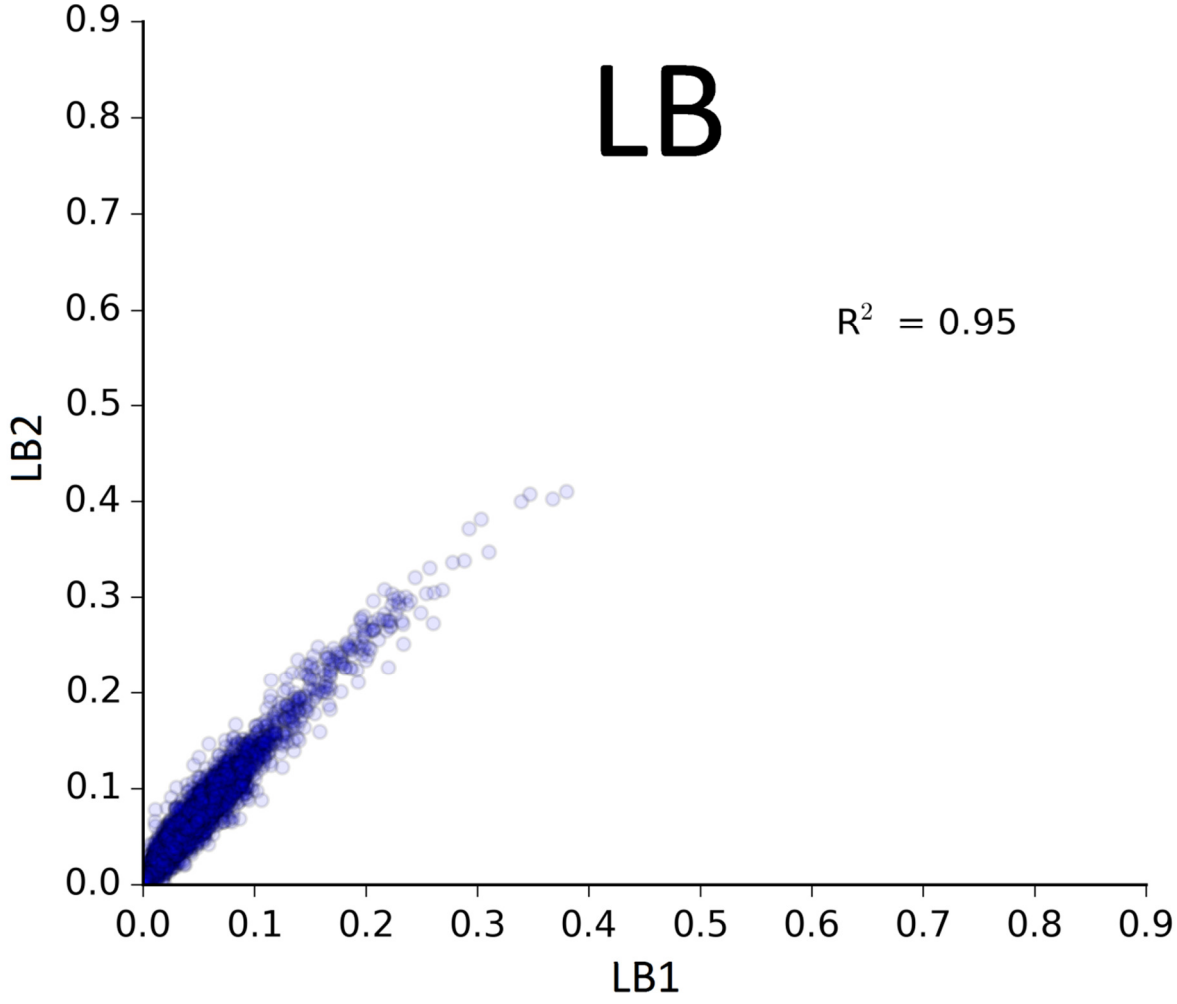

Figure. 3.4. Correlation of LB datasets derived from the shearing method. For each sample, the insertion indexes calculated for every W3110 coding sequence for each replicate were plotted against each other, and a coefficient of determination (R^2) was calculated. The max R^2 value is 1, which would indicate a perfect positive correlation

method will be referred to as the hybrid method. Sequencing libraries were prepared for the transposon library after growth in LB and the neat transposon library itself, each with two biological replicates. The results from sequencing these libraries are shown in Table 3.3. Read attrition can be seen at each processing step in each of the four datasets. However, in contrast to the previous two-PCR and shearing based methods, a far greater proportion of reads were retained after the final step of minimum read length filtering: 74-91% were retained at this point, in comparison to 3-7% in the first shearing method and 36-57% in the two-PCR method. The key difference between the hybrid method and the previous shearing method was in the efficiency of the two step transposon matching. In the shearing method, the vast majority of reads (between 91 and 95%) were rejected at the second part of this step. Far fewer reads were lost at the same stage in the hybrid method, with between 4% and 19% being removed. This is also the case when comparing the hybrid method with the two-PCR method, in that more reads were retained after transposon sequence matching. After the minimum read length filtering, the proportion of reads retained in the hybrid method was increased in comparison to the shearing method but not in comparison with the two-PCR method. This is true except for the NTL2 dataset, in which 1% of reads are shorter than 20 bases.

The number of clipped reads generated in the hybrid datasets was assessed. Between 2% and 4% were clipped and removed from the datasets. In comparison, the shearing method and the two-PCR methods resulted in between 0.1-0.2% and 20-35% clipped reads, respectively. Clearly, the two-PCR method is by far the least efficient in this regard. Although clipping is more common in the hybrid method, the total proportion lost is acceptable in light of inefficiencies in the other methods. The number of unique insertion sites was calculated for each of the four hybrid datasets. In the datasets there were between

Table 3.3. Dataset metrics from the hybrid method

Condition	Reads with matching inline barcode [1]	Reads with 1st 25 bp tn sequence [2]	Reads with 2nd 10 bp tn sequence [3]	Reads < 20 bases	Mapped reads before clipped read filtering	Genome wide insertions	CDS insertions	Mapped reads after clipped read filtering
NTL1	4818864	4606798 (95.6% raw reads)	4368206 (90.6% raw reads)	6270 (0.14% raw reads)	4061395 (84.3% raw reads)	502131	431608 (86% of total insertions)	3894330 (80.8% raw reads)
NTL2	6189409	5650877 (91.3% raw reads)	5023141 (81.2% raw reads)	50226 (1% raw reads)	4582674 (74% raw reads)	818674	706397 (86.3% of total insertions)	4391724 (71% raw reads)
LB1	5908163	5780360 (97.8% raw reads)	5636257 (95.4% raw reads)	7771 (0.14% raw reads)	5367487 (90.8% raw reads)	402025	344734 (85.7% of total insertions)	5205339 (88.1% raw reads)
LB2	6403324	6268530 (97.9% raw reads)	6141374 (95.9% raw reads)	12685 (0.21% raw reads)	5646798 (88.2% raw reads)	421778	361523 (85.7% of total insertions)	5387542 (84.1% raw reads)

[1] 0 mismatches

[2] 3 mismatches allowed

[3] 1 mismatch allowed

4-8 x 10^5 unique insertion sites across the whole genome, of which ~86% were found in coding sequences. The insertion numbers reported were higher than those in the two-PCR and shearing methods. However, these numbers are not directly comparable due to the differing numbers of reads in each dataset between the 3 preparation methods. However, the proportion of insertions that were found in coding sequences was very similar across the three methods.

Insertion indexes for all four datasets were calculated and plotted in histograms (Fig. 3.5). Very similar bimodal insertion index profiles can be seen in the two NTL replicates. The leftmost peak is again found at an insertion index of 0, and the leftmost bin at this position contains ~75 and ~35 coding sequences in either replicate. The right mode centres at approximately 0.09 in each replicate, although generally there are higher frequencies across most of the mode in the second NTL replicate.

There were clear differences between the hybrid histograms and the corresponding histograms of the two-PCR and shearing datasets. A much increased frequency can be seen in the leftmost bin of the two-PCR NTL histogram when compared to the hybrid NTL histogram. Additionally, the right mode is much closer to the zero mode in the two-PCR histogram. The same observations hold true in the histogram of the single shearing NTL dataset: when compared to the hybrid NTL histogram, the leftmost bin contains a greater number of coding sequences, and the right mode is much closer to the left.

There were also differences in the profile between the histograms for the hybrid NTL and LB datasets. In the LB histogram the leftmost bin contains ~100 coding sequences. The right mode centres at an insertion index of approximately 0.06, which is closer to the left mode than that seen in the NTL histogram. The distribution of this right mode is also broader than in the NTL histogram. A more clearly defined bimodal distribution can be observed in

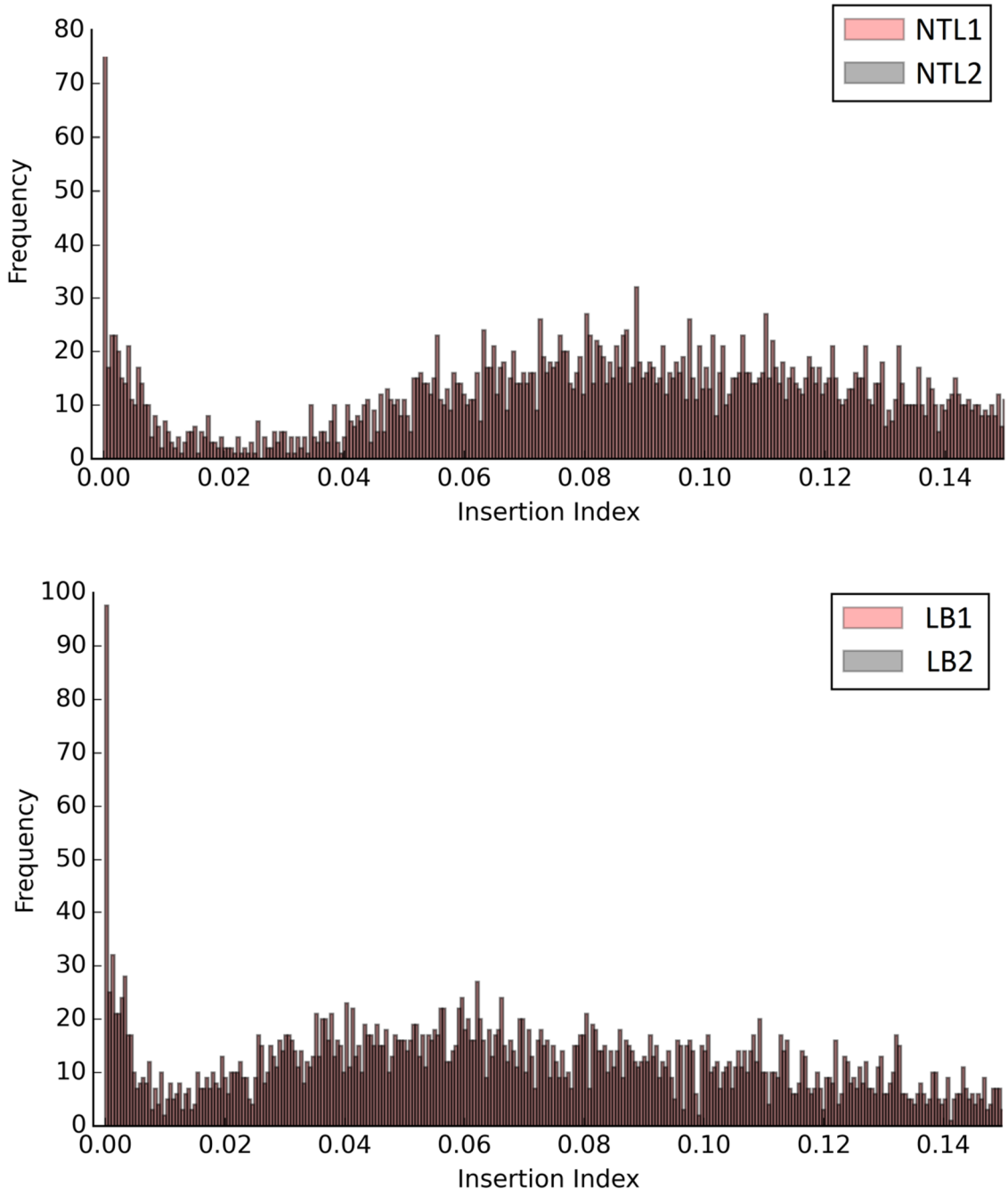

Figure 3.5. Histograms of insertion indexes calculated from datasets generated using the hybrid methodology. Insertion indexes greater than 0.15 were omitted. The upper panel shows the two NTL1 and NTL2 replicate sample datasets, and the lower panel shows the two LB1 and LB2 replicate sample datasets.

the hybrid NTL histogram in comparison to the two-PCR LB histogram, with a greater distance between the two modes. When compared to the shearing method LB histogram, the right mode of the hybrid LB histogram centres at an increased insertion index. Additionally, the leftmost bin contains fewer coding sequences.

The reproducibility of the insertion indexes generated from each replicate for both NTL and LB samples was tested (Fig. 3.6). The coefficient of determination was calculated to be 0.96 and 0.97 for the NTL and LB replicates, respectively. These values indicate that the insertion indexes generated in both replicates in each sample are highly reproducible. Interestingly, there appears to be a slight skew towards the second replicate in the NTL correlation plot, possibly corresponding to the generally higher insertion index frequencies seen in the right mode in Figure 3.5.

The insertion index histograms were used to predict statistically essential genes. After merging the two replicates for each sample, and using the same analysis as used with the previous datasets, 317 and 356 coding sequences were predicted to be essential in the NTL and LB datasets, respectively. These numbers were smaller than reported for the two-PCR and shearing datasets but are more consistent with reports of gene essentiality from other studies.

3.3 Discussion

In summary, three preparation techniques to produce transposon sequencing libraries have been adapted, applied and assessed using an *E. coli* BW25113 transposon library. The aim of this work was to compare the methods and to assess the data generated by each, in order to choose a technique to use in further work. At this point, the shearing method will be removed from further discussion. This is due to the huge read attrition seen

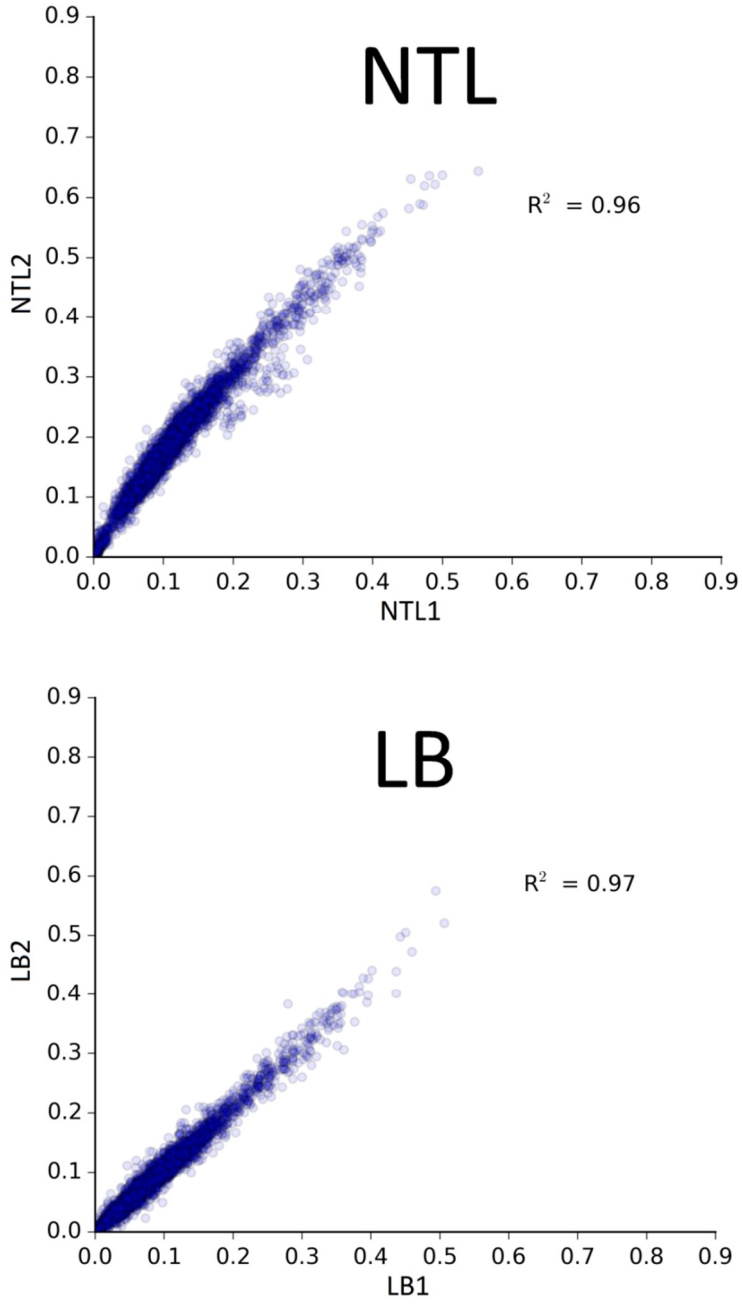

Figure 3.6 Insertion index correlation scatterplots for the hybrid datasets. For each sample, the insertion indexes calculated for every W3110 coding sequence for each replicate were plotted against each other, and a coefficient of determination (R^2) was calculated. The max R^2 value is 1, which would indicate a perfect positive correlation.

at the second transposon sequence similarity testing step, which removed the vast majority of the reads in each dataset. In most other steps the read retention rates were acceptable, but as a whole the methodology is considered unusable because it is so inefficient.

Arguably the most important comparator between the remaining two-PCR and hybrid methods is the proportion of reads that were not filtered out through the processing steps and that can be subsequently mapped to the W3110 genome. From this perspective, the hybrid method outperformed the two-PCR method substantially through an accumulation of improved retention at each processing step.

Another important consideration is in the number of essential genes that were predicted from each methodology using the same analytical process. Without accounting for the differences in read number between the two-PCR and hybrid datasets, the hybrid method appears to be the most promising, in that this method predicted the smallest number of essential genes for both the neat library and the library after growth in LB. The hybrid data revealed many examples where, over specific genomic areas, insertions could be observed which were not identified in the two-PCR datasets (Fig. 3.7). To compare the methodologies on an equal level, the program seqtk was used to subsample 3.8×10^6 reads randomly from the combined hybrid LB replicate datasets. When this subsampled dataset was analysed, 382 coding sequences were predicted to be essential, in comparison to 486 predicted from the combined two-PCR LB replicate datasets. Taken together these differences indicate that the hybrid methodology resulted in higher quality data, in turn enhancing the quality of essential gene analyses.

There are other considerations to be made alongside the raw data metrics. From a practical perspective, the hybrid method is more flexible because of the combination of Illumina indexing and the inline indexes. After the preparation of sequencing libraries and

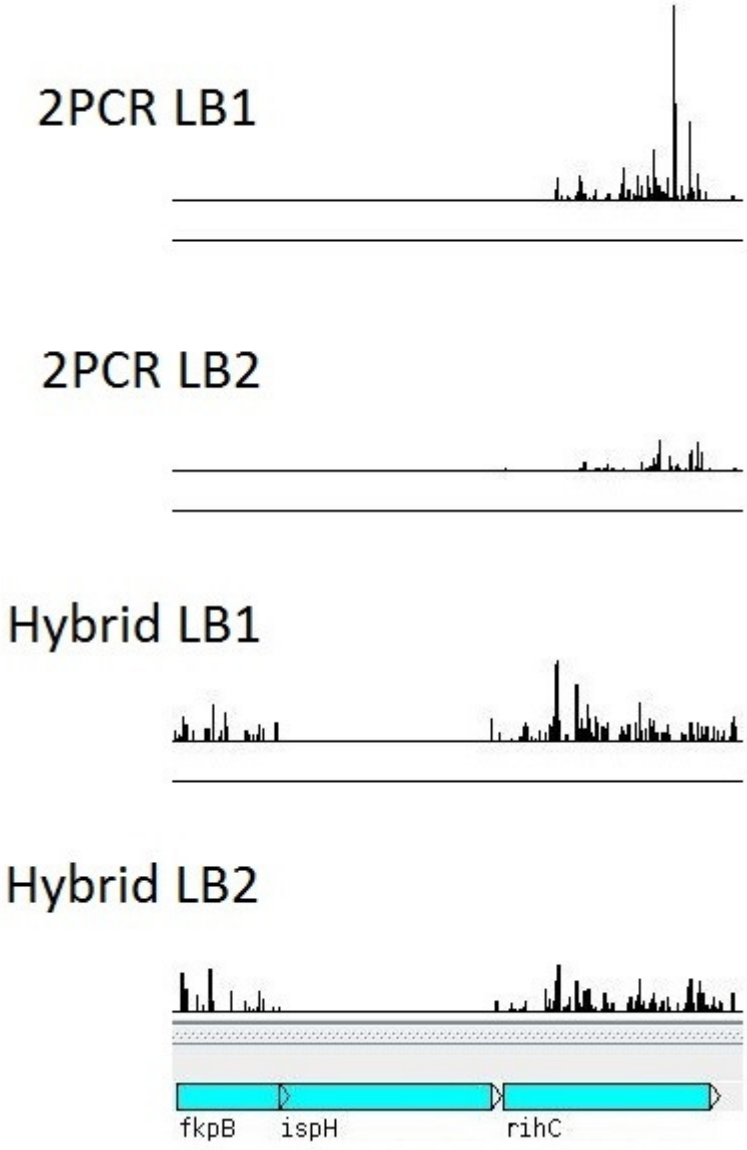

Figure 3.7. Differences in insertion representation between the two-PCR and hybrid methods. Artemis was used to look at the region containing 3 genes. Insertions can be seen across the whole of the *fkpB* coding sequence in the hybrid datasets. In contrast, none can be seen in the two-PCR datasets. The hybrid datasets also have a greater coverage over the 5' sequence coding for *rihC*.

during their sequencing, indexing is used to sample multiple libraries simultaneously. Both the two-PCR and hybrid methods utilise standard Illumina compatible sequencing, but the inline indexes are key. The usage of inline indexes in the hybrid method has two benefits. First, they can be easily designed to allow the multiplexing of a great number of samples, especially in combination with the standard Illumina indexes. However, perhaps their most impactful feature is in the prevention of low diversity issues. This issue, where the sequencer cannot properly determine the DNA sequence because of too many areas fluorescing across the imaging surface, is unavoidable through the use of the two-PCR method. However, inline indexes can be designed to include different numbers of bases. This effectively staggers the DNA fragments during sequencing, greatly increasing the diversity. This in turn allows for greater productivity, by enabling of sample multiplexing during sequencing. The sequencing of multiple samples at once, has the dual effects of increasing data quality and increasing the cost effectiveness of each sequencing run. In addition to this, the two-PCR method requires two custom sequencing primers to be added to the sequencing cartridge, to allow for the second Illumina compatible index read and to prime the sequence reads themselves.

In summary, of three tested, the hybrid methodology delivers the greatest amount and highest quality data, and so will be used in future transposon sequencing experiments.

CHAPTER 4

ESCHERICHIA COLI BW25113 ESSENTIAL GENE ANALYSIS

4.1 Introduction

One application of transposon insertion sequencing is in the determination of genes that are essential for growth. Essential genes are defined as being absolutely required for cell survival (Juhas, Eberl and Glass, 2011). However, this definition can be tempered by context dependence (Acevedo-Rocha *et al*., 2013): some genes may only appear essential under certain conditions, and so may not be ultimately essential to the cell.

During the creation of the transposon library, transposons will insert into the coding sequence of essential genes. These insertions physically disrupt the coding sequence, which in turn equate to disrupted polypeptides. The disruption of these proteins then leads to loss of viability, and a lack of propagation in the culture. At this point, when the transposon library is sampled to isolate genomic DNA, insertions within essential genes should not be present amongst the other genome wide insertions that do not affect viability.

The aim of the work presented in this chapter is to elucidate the essential genes of *E. coli* BW25113 through the use of our transposon library. To do so, the transposon insertion sequencing data will be compared with the gold standard database of *E. coli* essential genes, the KEIO library (Baba *et al*., 2006). In the work of Baba *et al*., precise gene deletions (Datsenko and Wanner, 2000) were used to investigate essentiality. There are multiple known issues with the creation of deletions in this way, including second site mutations, gene duplications, and cross contamination. Theoretically, insertion sequencing should not be prone to these issues. Additionally, the transposon library with and without growth will be tested, to assess which sample condition provides the best granularity.

4.2 Results

4.2.1 Datasets and essential gene prediction. In the previous chapter, the hybrid methodology was used to generate four datasets using our BW25113 transposon library: two biological replicates each from the neat transposon library (NTL) and from the library after growth in LB (LB). These datasets were used as the basis for this chapter.

For both LB and NTL samples, the raw reads from both biological replicates were combined. Insertion indexes were calculated and plotted in histograms (Fig. 4.1). The histograms were then used to predict which coding sequences were likely to be essential (see materials and methods for the calculation of gene essentiality). The numbers of coding sequences predicted to be essential from the combined NTL and LB datasets were 317 and 356, respectively. These essential gene lists were compared with the essential gene list from the KEIO library (Baba *et al*., 2006). This list, as initially published, contained 303 essential gene candidates. Since then 3 candidates have been shown to be spurious open reading frames (ORFs) and so were removed from the list, leaving 300 remaining candidates (*Z*hou and Rudd, 2013).

After correlating the three essential gene lists, there were 404 unique candidate essential genes. The Venn diagram in Figure 4.2 shows the overlap of essential genes between the three datasets. The largest subset of these genes is the set of 248 genes that were reported to be essential in all three datasets. The second largest subset (64) is that of essential genes reported in both LB and NTL lists, but not in the KEIO list. The third largest subset of 43 genes were reported only in the KEIO essential gene list, and 38 genes were uniquely present in the LB essential gene list. The remaining three subsets were small in comparison: 6 genes were reported in both the KEIO and LB gene lists; 3 genes were reported in both the KEIO and NTL lists and 2 genes were reported in the NTL list alone.

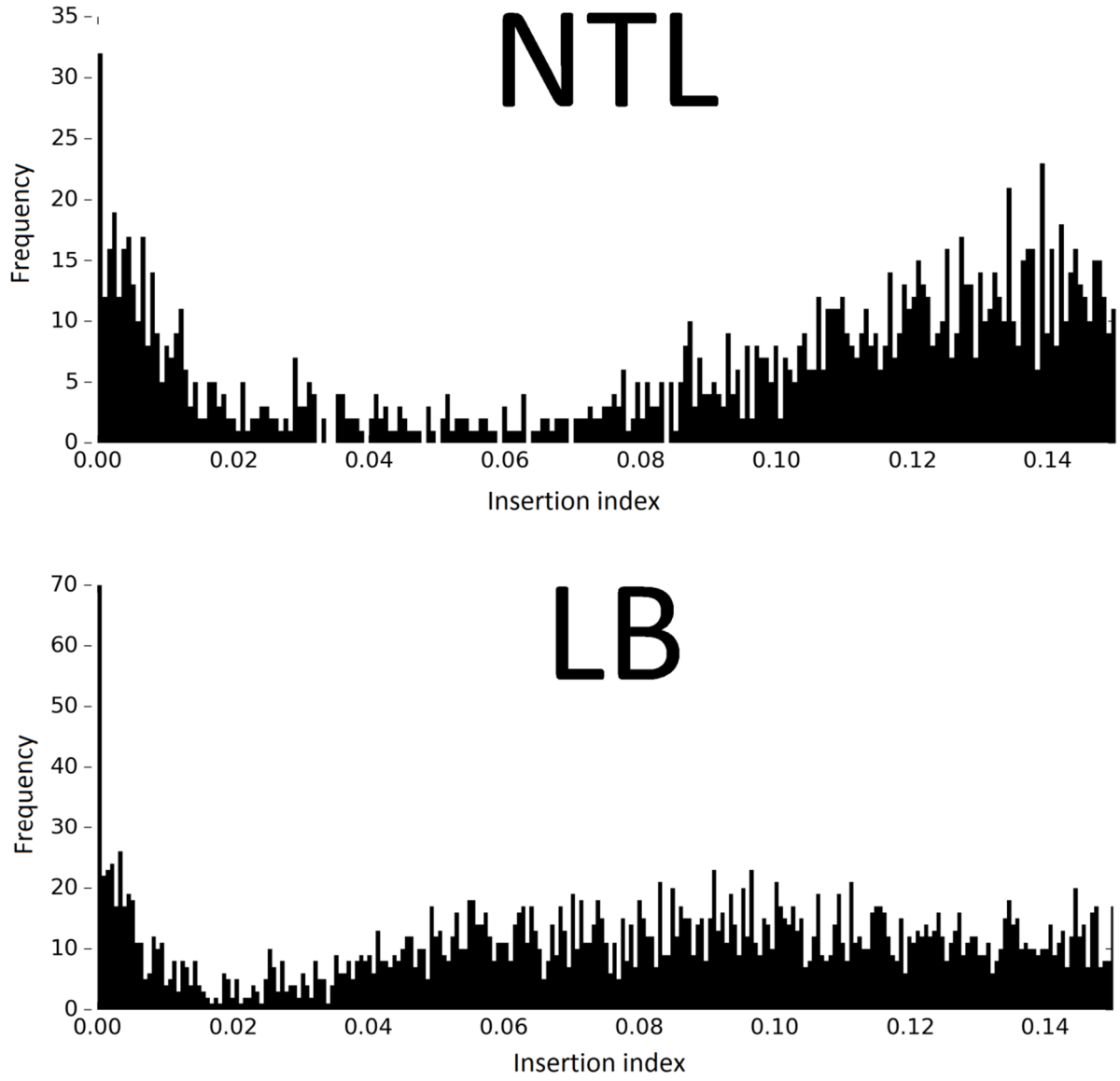

Figure 4.1 Insertion index histograms for the combined NTL and LB datasets. Insertion indexes greater than 0.15 are omitted. These histograms show the combined data from both replicates for each of the NTL and LB samples. The insertion indexes for each coding sequence are tallied in each bin, with indexes close to zero more likely to be essential.

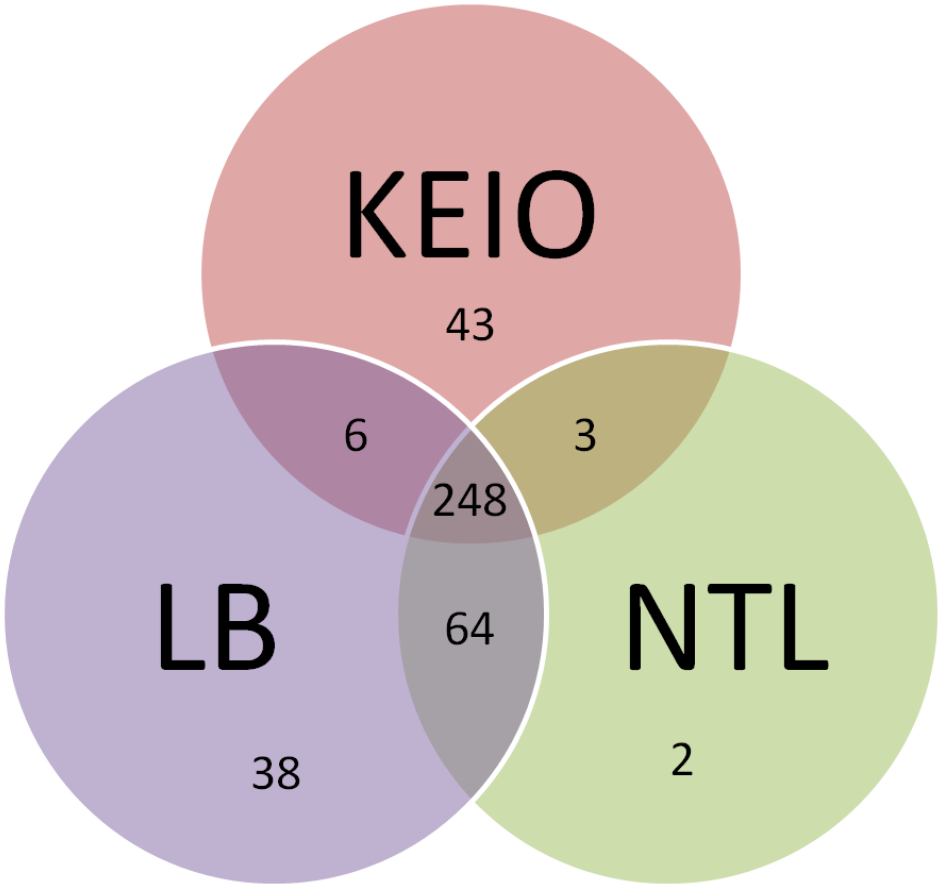

Figure 4.2 Comparison of the KEIO, LB and NTL essential gene lists. This Venn diagram shows the overlap of essential genes from the KEIO library (Baba et al., 2006) and predicted essential genes from the NTL and LB datasets produced in chapter 3. There are 404 unique candidates in total. Out of the 300 essential KEIO genes analysed, 248 (approx. 83%) were also reported in both NTL and LB datasets.

4.2.2 Manual inspection of essential genes. For the purposes of this chapter, we accept that the 248 coding sequences identified at the intersect of all three datasets are truly essential (Fig. 4.2). These genes will be the basis of what is termed the core essential gene list. These will not be considered further here.

Each of the remaining 156 coding sequences, which were not consistently reported in the 3 datasets, were manually inspected using the Artemis genome browser (Rutherford *et al*., 2000). The coding sequences were assessed for the number of insertion sites, the position of the insertions within the coding sequence, and the frequency of insertions. The coding sequences were defined as likely or not likely to be essential. The lists were then updated accordingly.

From the manually inspected 156 genes, 60 genes were defined as likely to be non-essential. This list was split into two sets: 26 genes that were essential in the KEIO dataset but not in either LB or NTL datasets, and 34 that were predicted to be essential from the LB and NTL datasets but not in the KEIO dataset. In the first set of 26 genes reported to be essential in the KEIO dataset, manual inspection revealed insertions throughout these coding sequences; multiple examples of these genes are shown in Figure 4.3. This, in addition to the presence of insertions after growth in LB, is strong evidence of the non-essentiality of these coding sequences. For all but one (*yceQ*) of these 26 genes, literature evidence supports their non-essentiality, with the relevant citations shown in Table 4.1 and section 4.2.3 below. Of the 34 coding sequences that were predicted to be essential in the LB and NTL datasets, the majority (24) show a pattern in which they are predicted to be essential in the LB dataset but not in the NTL dataset. In these 24 (*aceE, aceF, cmk, crr, gnsB, guaB, hscA, icd, ihfA, lpcA, ptsH, ptsI, rpe, seqA, sucA, tonB, ubiF, yccK, yciS, yddL, ydhR,*

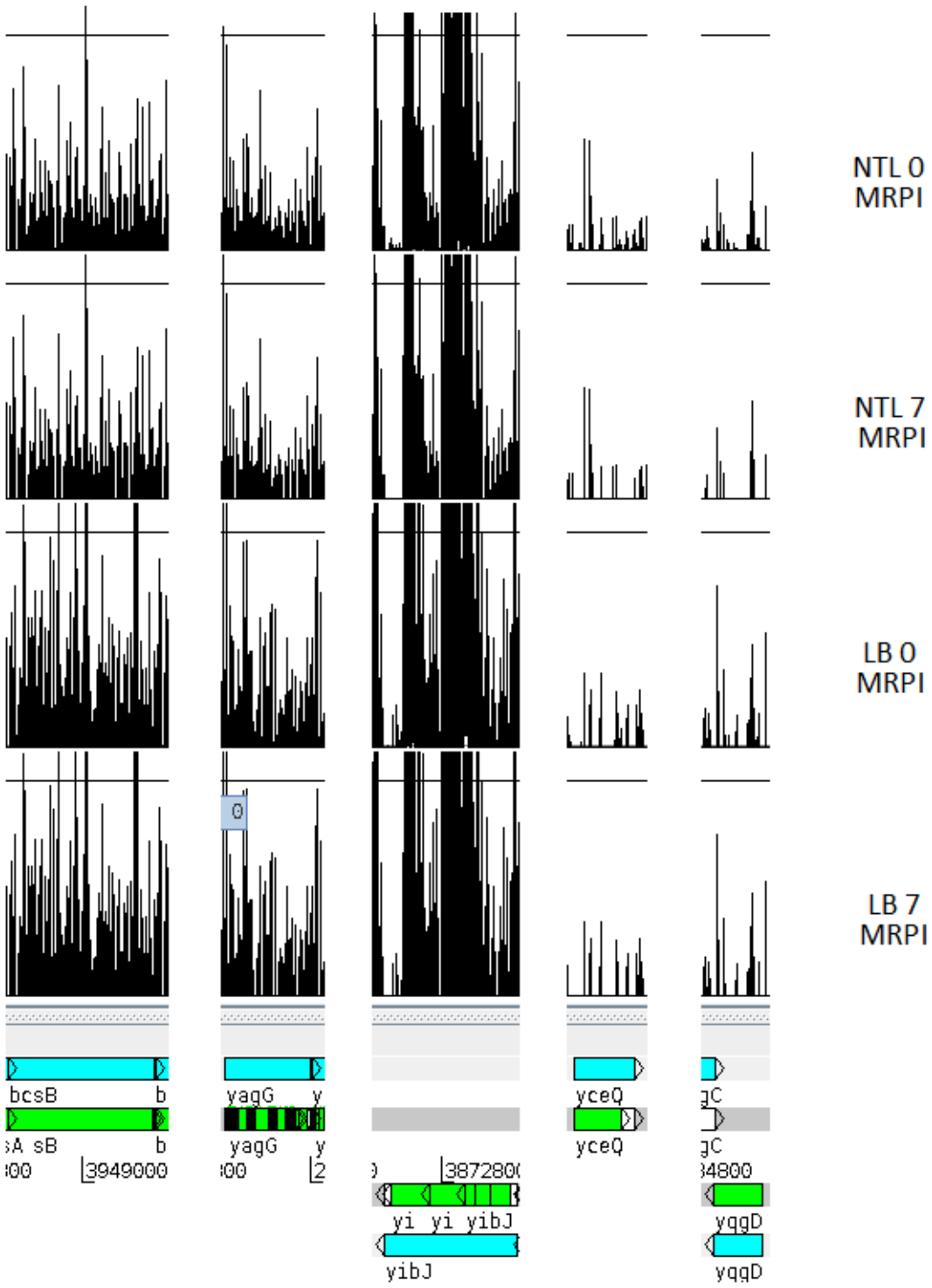

Figure 4.3 Insertions throughout genes previously predicted to be essential. In each of the genes shown above (*bcsB*, *yagG*, *yibJ*, *yceQ* and *yqgD*), insertions with high read frequencies can be seen throughout the coding sequences. This indicates that they are non-essential.

Table 4.1. Genes defined as likely to be non-essential after manual inspection.

Gene	Description[1]	COG number	COG category	Evidence	Datasets[2]
aceE	Pyruvate dehydrogenase, decarboxylase component E1; acetate requirement	2609	C	Ito et al., 2005	xxL
aceF	Pyruvate dehydrogenase, dihydrolipoamide acetyltransferase E2; acetate requirement	0508	C	Ito et al., 2005	xxL
alsK	D-allose kinase	1940	G	Gerdes and Osterman, 2008	Kxx
bcsB	Cellulose synthase, regulatory subunit; binds cyclic-di-GMP; periplasmic, membrane-anchored	ENOG410XNNB	M	Gerdes and Osterman, 2008	Kxx
chpS	ChpS antitoxin, toxin is ChpB	2336	K	Gerdes and Osterman, 2008	Kxx
cmk	Cytidylate kinase; multicopy suppressor of UMP kinase mutations	0283	F	Fricke et al., 1995	xxL
crr	EIIA(Glc), phosphocarrier for glucose PTS transport; negative control of rpoS	2190	G	Guo et al., 2015	xxL
entD	Enterochelin synthase, component D; EntB(ArCP)/EntF-CoA phosphopantetheinyltransferase; facilitates secretion of enterobactin peptide; enterobactin biosynthesis	2977	q	Coderre and Earhart, 1984	Kxx
ftsE	Cell division ATP-binding protein; associated with the inner membrane via FtsX; null mutant has filamentous growth and requires high salt for viability	2884	d	Leeuw et al., 1999	Kxx
ftsX	Integral membrane protein invoved in cell division; binds FtsE to the inner membrane	2177	d	Reddy, 2006	Kxx
gnsB	Multicopy suppressor of secG(Cs) and fabA6(Ts), Qin prophage; overexpression increases unsaturated fatty acid content of phospholipids; gnsA paralog	ENOG410Y8R8	s	Sugai et al., 2001	xxL
guaB	Inosine-5'-monophosphate (IMP) dehydrogenase	0516	F	Kang et al., 2004	xxL
hscA	DnaK-like chaperone Hsc66, IscU-specific chaperone HscAB; involved in FtsZ-ring formation	0443	O	Jang and Imlay, 2010	xxL
icd	Isocitrate dehydrogenase, NADP(+)-specific; e14 attachment site; tellurite reductase	0538	C	Okamoto et al., 2014	xxL
ihfA	Integration Host Factor (IHF), alpha subunit; host infection, mutant phage lambda; site-specific recombination; sequence-specific DNA-binding	0776	L	Gopel et al., 2011	xxL

	transcriptional activator				
lpcA	Phosphoheptose isomerase; D-sedoheptulose 7-phosphate isomerase; GDP-heptose biosynthesis; T-phage resistance	0279	G	Brooke and Valvano, 1996	xxL
mazE/chpR	MazE antitoxin, toxin is MazF	2336	K	Gerdes and Osterman, 2008	Kxx
minD	Inhibitor of FtsZ ring polymerization; chromosome-membrane tethering protein; membrane ATPase that activates MinC	2894	D	Gerdes and Osterman, 2008	Kxx
mlaB/yrbB	Probable phospholipid ABC transporter, quinolone resistance; peripheral membrane protein, cytoplasmic; maintains OM lipid asymmetry; STAS subunit	3113	s	Malinverni and Silhavy, 2009	Kxx
priB	Primosomal protein n; ssDNA-binding protein	2965	L	Bubunenko, Baker and Court, 2007	xNL
ptsH	PTS system histidine phosphocarrier protein HPr; phosphohistidinoprotein-hexose phosphotransferase	1925	G	Gershanovitch et al., 1977	xxL
ptsI	Phosphoenolpyruvate-protein phosphotransferase; phosphotransferase system, enzyme I; E1; PEP-dependent autokinase	1080	G	Hernandez-Montalvo et al., 2003	xxL
rnc	RNase III; cleaves double-stranded RNA	571	K	Bubunenko, Baker and Court, 2007	Kxx
rpe	D-ribulose-5-phosphate 3-epimerase	0036	G	Ito et al., 2005	xxL
rsgA	Ribosome-stimulated GTPase, 30S subunit assemby; low adundance protein; putative RNA binding protein	1162	s	Hase et al., 2009	xNL
rsmI/yraL	16S rRNA C1402 2'-O-ribose methyltransferase, SAM-dependent	0313	s	Dassain et al., 1999	Kxx
secM	Secretion monitor controlling secA expression	ENOG4111GJA	K	Rajapandi, Dolan and Oliver, 1991	Kxx
seqA	Multi-faceted genome stability factor; negative modulator of initiation of replication; replication fork tracking protein required for chromosome segregation; chromosome cohesion protein; hemimethylated GATC binding protein	3057	L	Waldminghaus and Skarstad, 2010	xxL
sucA	2-oxoglutarate dehydrogenase, E1 component; yields succinyl-CoA and CO(2); also known as alpha-ketoglutarate dehydrogenase	0567	C	Nishio et al., 2013	xxL
sucB	2-oxoglutarate dehydrogenase, E2 component; dihydrolipoamide	0508	C	Kohanski et al., 2007	xNL

	succinyltransferase; acid-inducible; yields succinyl-CoA and CO(2); also known as alpha-ketoglutarate dehydrogenase				
tdcF	Putative reactive intermediate deaminase, UPF0076 family; trimeric; reaction intermediate detoxification	0251	J	Gerdes and Osterman, 2008	Kxx
tnaB	Tryptophan:H+ symport permease, low affinity	0814	E	Yanofsky, Horn and Gollnick, 1991	Kxx
tonB	Uptake of chelated Fe(2+) and cyanocobalamin; works in conjunction with OM receptors; energy transducer; sensitivity to T1, phi80, and colicins; forms a complex with ExbB and ExbD	0810	M	Kohanski et al., 2007	xxL
ubiF	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase; produces 2-octaprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinol; required for ubiquinone synthesis; mutation confers resistance to bleomycin, phleomycin and heat	0654	CH	Ito et al., 2005	xxL
yabQ	Pseudogene reconstruction, pentapeptide repeats-containing	ENOG410XV6S	S	Gerdes and Osterman, 2008	Kxx
yafF	Pseudogene, C-terminal fragment, H repeat-associated protein	5433	L	Gerdes and Osterman, 2008	Kxx
yagG	Putative sugar symporter, function unknown, CP4-6; putative prophage remnant	2211	g	n/a	Kxx
ybbD	Pseudogene reconstruction, novel conserved family	1472	G	n/a	xNL
yccK	mnm(5)-s(2)U34-tRNA 2-thiolation step sulfurtransferase; binding partner linking TusBCD to MnmA; may transfer sulfur first to MnmA or directly to tRNA	2920	P	Ikeuchi et al., 2006	xxL
yceQ	Function unknown	ENOG410YYPH	S	n/a	Kxx
yciS	DUF1049 family inner membrane protein	3771	S	Mahalakshmi et al., 2014	xxL
ydaS	Putative Cro-like repressor, Rac prophage	2261	S	n/a	xNL
yddL	Pseudogene, OmpCFN porin family, N-terminal fragment	na	na	n/a	xxL
ydfB	Expressed protein, function unknown, Qin prophage	ENOG4111SFN	S	Gerdes and Osterman, 2008	Kxx
ydfO	DUF1398 family protein, Qin prophage	5562	S	n/a	xNL

ydhR	Predicted monooxygenase, function unknown; dimeric	ENOG4111VBS	S	n/a	xxL
ydiL	Putative HTH domain DNA-binding protein; lambda repressor-like protein	ENOG41120Y0	s	Gerdes and Osterman, 2008	Kxx
yedM	Pseudogene reconstruction, IpaH/YopM family	4886	S	n/a	xNL
yefM	Antitoxin for YoeB toxin; binds YoeB RNase-like domain	2161	D	Gerdes and Osterman, 2008	Kxx
ygeL	Pseudogene reconstruction, part of T3SS PAI ETT2 remnant; response regulator family	na	na	n/a	xNx
ygeM	Pseudogene reconstruction, orgB homolog; part of T3SS PAI ETT2 remnant	na	na	n/a	xNx
yhbV	U32 peptidase family protein, function unknown,	0826	O	Yu et al., 2008	Kxx
yheM	2-thiolation step of mnm(5)-s(2)U34-tRNA synthesis; sulfur relay system; required for swarming phenotype	2923	P	Ikeuchi et al., 2006	xxL
yhhQ	DUF165 family inner membrane protein	1738	s	Gerdes and Osterman, 2008	Kxx
yibJ	Pseudogene, Rhs family	3209	m	Gerdes and Osterman, 2008	Kxx
yigP	Aerobic ubiquinone synthesis protein, SCP2 family protein	3165	S	Aussel et al., 2014	Kxx
ynfN	Cold shock-induced protein, function unknown, Qin prophage	ENOG410Y031	S	n/a	xxL
ypjC	Pseudogene reconstruction	1284	s	n/a	xNL
yqgD	n/a	ENOG410Y8M8	S	Gerdes and Osterman, 2008	Kxx
zwf	Glucose-6-phosphate 1-dehydrogenase	0364	G	Sandoval et al., 2011	xxL

[1]The descriptions for each gene were obtained from Ecogene (Zhou and Rudd, 2012). The COG categories were obtained from eggNOG (Huerta-Cepas et al., 2015). The evidence column refers to papers which provide evidence of non-essentiality

[2]K - Essential in KEIO. N - Essential in neat transposon library. L - Essential after growth in LB.

yheM, ynfN, zwf), a number of low frequency insertions can be seen. These insertions are likely to be a form of background noise in the data, and can be seen in the top row displaying the insertions in the NTL dataset with 0 minimum reads per insertion required (as shown later in Figure 4.6). These likely spurious low frequency insertion sites act to increase the insertion index of the coding sequences and cause them to be predicted as non-essential. In contrast, the lack of such background in the LB dataset decreases the insertion index and increases the likelihood of being predicted as essential. Eight (*priB, rsgA, sucB, ybbD, ydaS, ydfO, yedM, ypjC*) of the remaining 10 genes were predicted to be essential in both NTL and LB datasets, and these generally contain a small number of low frequency insertions. The final 2 genes (*ygeL* and *ygeM*) were predicted to be essential only in the NTL dataset. It is highly unlikely that genes would be essential during the construction of the library and non-essential afterwards, indeed to be present after growth such insertions would have to be present in the original library. Therefore, these genes are predicted to be non-essential after growth and their predicted essentiality in the NTL dataset is thought to be anomalous. Furthermore, for 24 of the 34 genes, literature evidence can be found supporting their non-essentiality, shown in Table 4.1 and section 4.2.3. Additionally, the chromosomal position of these genes were investigated, and the genes appeared to be spread evenly throughout the chromosome.

After removal of the 60 non-essential genes from the 156 manually inspected, 96 remain. Twenty six of these genes were added to the core essential gene list after manual inspection. Seventeen of the 26 genes were previously predicted to be essential in the KEIO library, but were not highlighted in either the NTL or LB dataset. Upon manual inspection, all but one of these coding sequences were observed to have one of two specific patterns of insertion. Ten of the 17 genes (*ftsK, ftsN, grpE, lptC/yrbK, minE, mqsA/ygiT, rne, spoT,*

waaU/rfaK, yejM) appeared to contain essential regions within their coding sequences (Fig. 4.4). These are regions in which no transposons are found, in contrast to adjacent regions which could be inserted into. These regions are of a variable size in the genes containing them. All but one of these regions were found at the 5’ end of the gene. The one example of a 3' essential region, in *grpE*, is shown in Figure 4.4. For each of these genes, the read alignments were checked to see in which orientation the transposons had inserted. Four of the genes (*grpE, lptC/yrbK, minE, mqsA/ygiT*) had insertions only in the reverse strand with respect to the 5' - 3' direction of the gene, including *grpE*. The remaining 7 genes had insertions in both strands. These findings make sense when considering how transposons disrupt genes. For the genes with a 5' essential region, forward and reverse strand insertions after the essential region are permissible because the insertion does not affect the transcription and translation of the essential region itself. However, it is not understood why insertions only occurred in the reverse strand for four of the genes with 5’ essential regions. For *grpE*, which contains a 3' essential region, no insertions can be found in the forward strand prior to the essential region because insertion here would affect the transcription and translation of the essential region. However, the fact that insertions are permissible at the 5’ end of the gene in the reverse strand suggest that there are characteristics of either the transposon or the gene that allow these insertions to occur, for example a promoter in the transposon facilitating transcription outwards of the transposon. For 6 of the 17 genes predicted to be essential in the KEIO dataset but neither NTL or LB dataset, insertions were observed in either or both their very 5’ and 3’ ends but not over the majority of the central coding sequence. Upon closer inspection, 4 of these (*folK, ftsL, psd, rnpA*) have insertions in the relative reverse strand only, while the remaining 2 (*ribB* and *secF*) have insertions in both strands. Interestingly, in one of the genes the genes with this pattern of insertion, *rnpA*, the

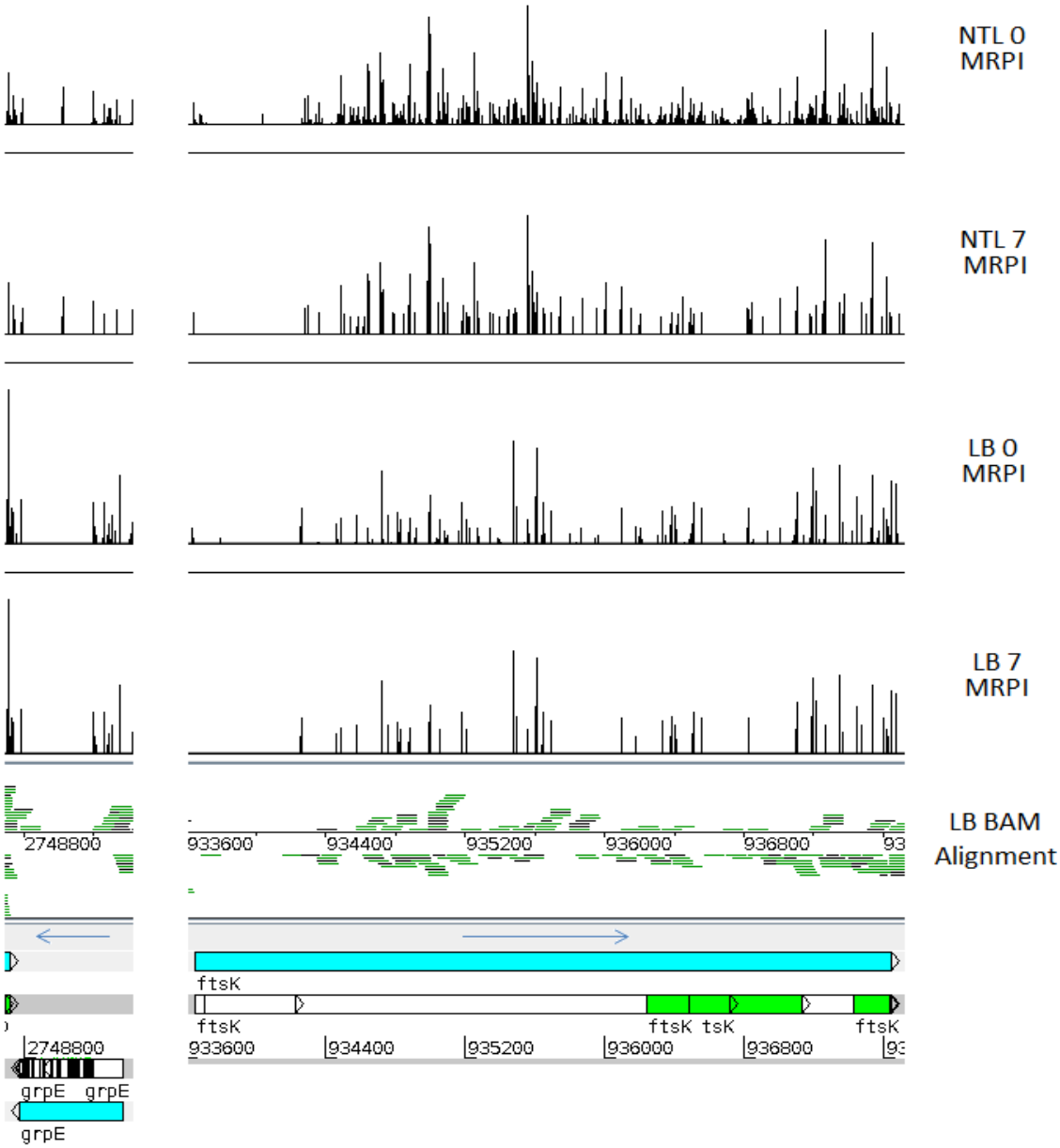

Figure 4.4 Essential gene regions of *grpE* and *ftsK*. *grpE* and *ftsK* contain 3' and 5' essential regions respectively. The BAM alignment in the fifth row shows the full read alignments for the LB dataset, with zero minimum required reads per insertion site. In this row, reads mapping to the forward strand are shown on the top half, and reads aligning to the reverse strand are shown on the bottom half. The top half of the fifth row shows the forward strand and the bottom half shows the reverse strand. Within the 5' region of *grpE*, insertions can only be observed in the reverse strand relative to the 5' - 3' direction of the coding sequence. In *ftsK*, after the 5' essential region, insertions can be seen in either strand

.

3’ of the coding sequence overlaps with the 5’ of another coding sequence (*yidD:* Fig. 4.5). Insertions can be found throughout the *yidD* coding sequence, indicating that it is non-essential. The insertions found in the 3’ end of *rnpA* occur exclusively in the overlapping part of the coding sequences. This example serves to highlight the importance of manual inspection: from the insertion index essentiality prediction *rnpA* was predicted to be non-essential, and only after manual inspection could it be said that *rnpA* is likely essential. In other cases, the presence of insertion sites at the 3’ of the coding sequence might indicate that it is only the 3’ of the gene that is non-essential for function. The presence of insertions at the very 5’ end may suggest an incorrectly labelled translational start site. Alternatively, this could be explained by a promoter in the transposon initiating transcription as previously described. For the single gene remaining of the 17, *cydC*, a particular pattern of insertions could be observed whereby it appeared only insertions at particular positions were viable and the majority of the coding sequences contained no insertions (Fig. 4.6). In *cydC*, there appear to be two clusters in which insertions are relatively frequent. In the NTL data, the insertions are of a low frequency which increases after growth. This observation might suggest that *cydC* contains more than one region of essentiality. In total, of the 17 genes discussed, supporting evidence of essentiality in the literature could be found for 11 of them. For another four genes, no evidence of essentiality in addition to Baba *et al*. could be found. For the remaining two genes, literature was found detailing context dependent essentiality, which is discussed further below.

The remaining 9 of the 26 genes added to the core essential gene list were predicted to be essential in the KEIO dataset and also in either the NTL or LB dataset. These genes could be grouped by the same patterns of insertion as discussed previously. Three of the 9

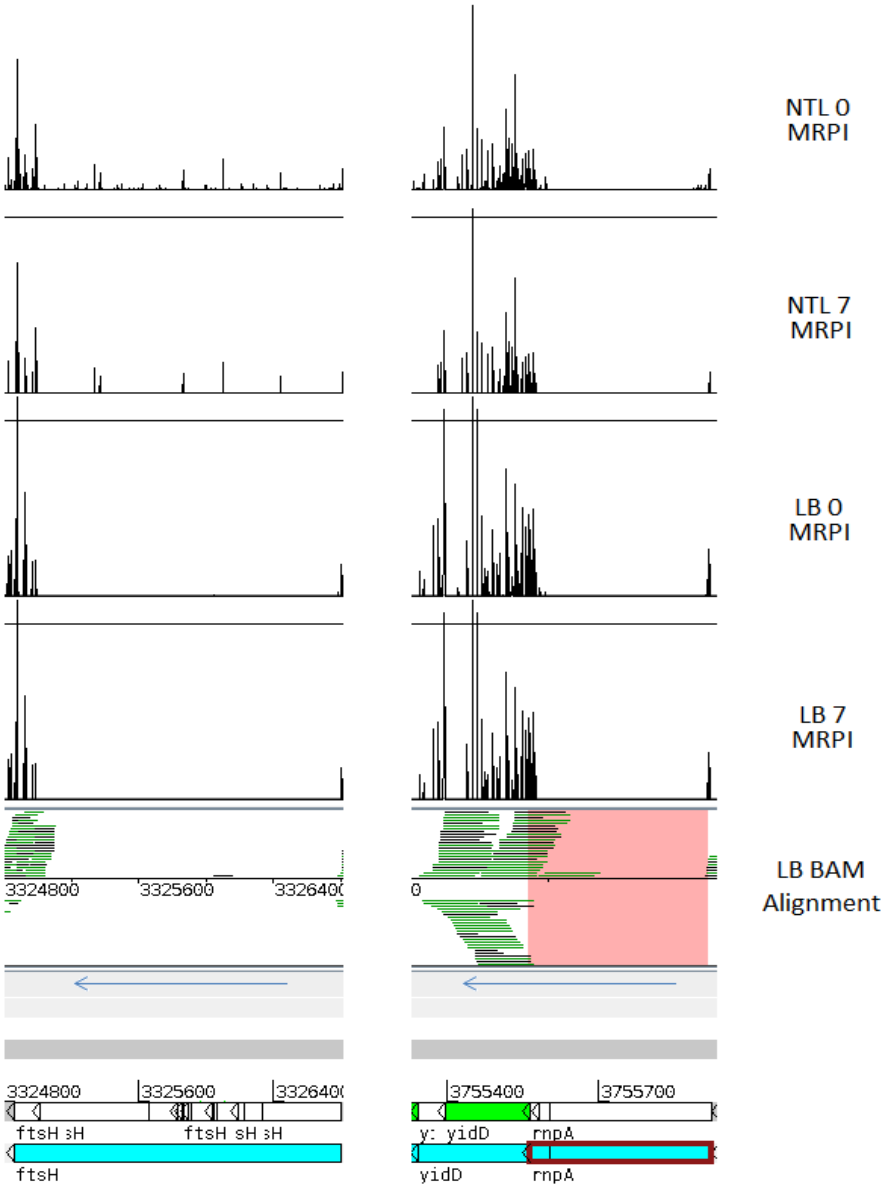

Figure 4.5 Insertions into the 5' and 3' regions of *ftsH* and *rnpA*. Within the very 5' of *ftsH*, a small cluster of closely spaced insertions can be seen. The reads aligning to these insertions have all mapped to the reverse strand with respect to the 5' to 3' direction of the coding sequence. Across approximately 10% of the coding sequence at the 3' end, a larger cluster of insertions can be seen. The vast majority of these reads are mapped to the reverse strand, with a negligible few mapping to the forward strand. Across *rnpA* (highlighted in red for clarity), every read is mapped to the reverse strand. Insertions can be seen at the 5' and 3' ends. The coding sequence of *rnpA* overlaps with that of *yidD*. Each insertion at the 3' end of rnpA is located within this overlapping section.

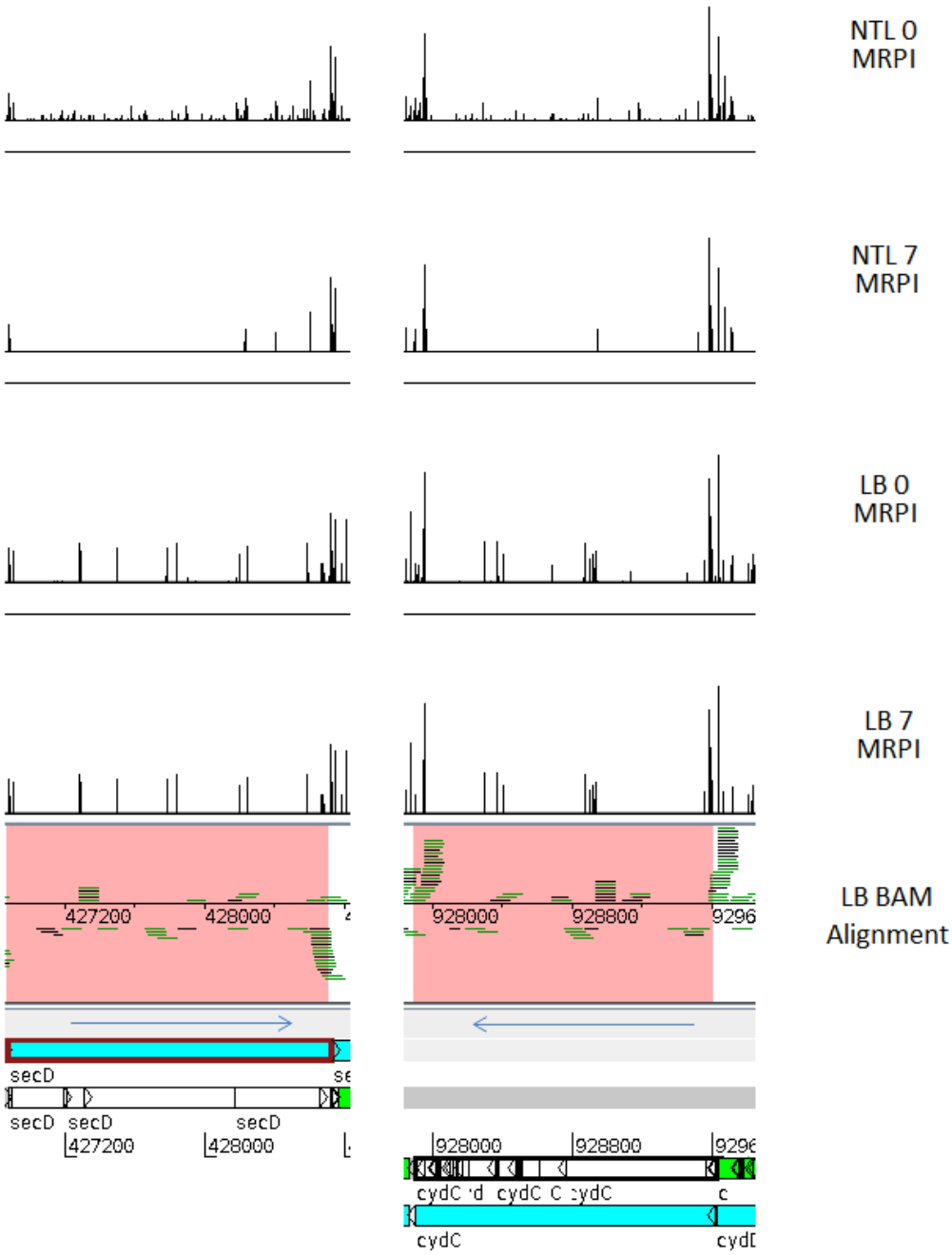

Figure 4.6 Insertions into *cydC* and *secD*. The insertions within the *secD* and *cydC* coding sequences show a different pattern in comparison to the majority of other coding sequences. The insertions appear more tightly grouped at defined regions in the coding sequences, as opposed to a more even representation throughout.

genes (*degS, lptA, mreC*) appeared to contain essential regions, and 5 more (*csrA, ftsH, rseP, tadA, ftsB*) had insertions at the very 5' or 3' ends. The one remaining gene (*secD*) shared the pattern of insertions discussed previously for *cydC*, whereby small clusters of insertions could be seen, with the majority of the coding sequence uninterrupted. Seven of these 9 genes had literature supporting their essentiality. For one other gene (*secD*) no evidence for essentiality other than Baba *et al*. was found, and for the final gene evidence for context dependent essentiality was found.

The 26 genes discussed in this section were not predicted to be essential from the LB and NTL datasets due to the insertions within the non-essential regions. Insertions in these regions increase the insertion indexes of the coding sequences, meaning that the statistical analysis would predict the coding sequence to be non-essential.

After the consideration of these 86 genes out of the 156, the 70 remaining were predicted to be essential in the LB and NTL datasets but not in the KEIO dataset. Manual inspection of these coding sequences suggested that they are likely to be either essential or at least important for growth. Amongst these candidate essential genes, 25 were either completely free or almost free of insertions (Fig. 4.7). In a further 42 of the 70 coding sequences, insertions could be seen in either or both their very 5' and 3' ends but not over the majority of the coding sequence. This is the same pattern as seen in 11 of the genes manually defined as essential that were also defined as essential in the KEIO data. Furthermore, 3 of the 70 coding sequences appear to contain essential regions, another pattern observed in the core essential gene list.

In summary, after the correlation of the KEIO essential gene list and the predicted essential gene lists from the LB and NTL datasets, there were 404 essential gene candidates. After manual inspection and re-analysis, there are now three lists. The first is the core

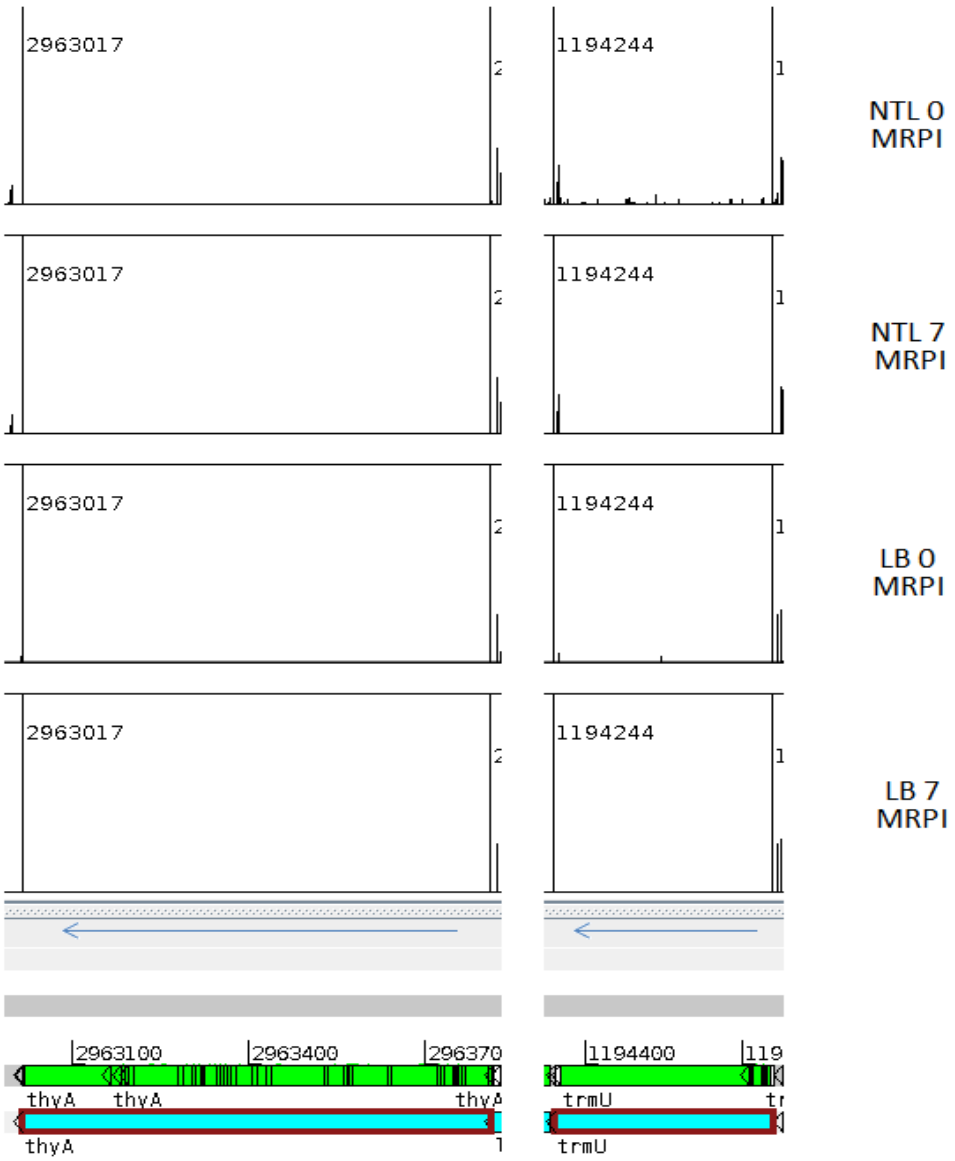

Figure 4.7 Lack of insertions in genes not reported to be essential from the KEIO library. Each of the four rows show the position of insertions (black lines) across particular coding sequences (blue boxes with the name shown underneath). Each black line represents only the first base of an insertion, and the rest of the aligned read is removed from view. Both NTL and LB datasets are shown each with 0 and 7 minimum reads per insertion required (MRPI), as indicated along the right. The blue arrows at the bottom indicate the direction of the coding sequence 5' to 3'. The vertical black lines in each row show the gene boundaries. The coding sequences are not to scale in width. The vertical scale for each row is limited to a maximum depth of 80 reads. This information applies to each following figure in the chapter. In the NTL 0 MRPI row across *trmU*, it is possible that the low frequency insertions are noise in the data.

essential gene list of 274 genes, which contains the certainly essential genes. The second list of 70 genes consists of genes that are either essential gene candidates or genes that are important for the normal growth of the cell. The third list of 60 genes consists of genes that are unlikely to be essential.

4.2.3 Supporting evidence for manually inspected genes. The KEIO essential gene list used to compare against the LB and NTL datasets was taken from Baba *et al*. (2006). An update to this paper was published by the same research group (Yamamoto *et al*., 2009). In this work, several genes that were originally listed as non-essential from Baba *et al*. (2006) were re-analysed and found to be essential. The strains containing these gene deletions were found to contain duplications of the target gene. Duplication of genes in this manner would allow an apparently authentic deletion alongside a remaining copy of the wild type gene, when in fact the gene would be essential for growth. Twenty five genes were found to contain such duplications, and 14 of these were listed as new essential gene candidates. Interestingly, all but one of the 14 new candidates were present in the candidate essential gene list of 70 genes and in each case, the genes were predicted to be essential in both LB and NTL datasets. The single gene (*polA*) not found in the correlated list of 404 genes encodes DNA polymerase I. Upon manual inspection, this coding sequence appeared to contain an essential region, explaining why this gene was not predicted to be essential in either LB or NTL dataset. As such, *polA* was added to the list of 404 genes to make 405, and subsequently added to the core essential gene list now containing 275 genes.

The remaining 11 of the 25 genes containing duplications were listed as genes with uncertain essentiality. Seven of these were found in the original list of 404 genes: 6 were

predicted as essential in both LB and NTL datasets (*hemE, priB, rplK, rplY, rpsO, rpsU*) and the single gene remaining (*folP*) was predicted to be essential from the LB dataset. In this single gene, a low frequency of reads likely to be background was present in the NTL data, explaining why it was not predicted to be essential in both LB and NTL datasets. After manual inspection, one of the 7 genes (*priB*) had been defined as likely non-essential, due to a low frequency of insertion. However, the other 6 genes were retained in the candidate essential gene list. The remaining 4 genes (*btuB, djlB, tpr, yiaD*) were not present in the intersected gene list. Each of these 4 coding sequences had a high frequency of insertion in both LB and NTL datasets, indicating they are unlikely to be essential.

In the list of 70 genes defined as likely essential or important for growth, literature could be found for 26 genes (*crp, cydB, cydD, dnaK, efp, fabH, holD, iscS, iscU, lpd, lpxL, nusB, rimM, rluD, rnt, rplA, rpmJ, rpsF, rpsT, rrmJ, ubiE, ubiG, ubiH, ubiX, ybeD, ybeY*) which disruption of the gene led to a slower growth rate. Notably, this set included genes such as *crp*. The first evidence of the non-essentiality of *crp* was published in 1975 by Dennis Sabourin and John Beckwith. D'Ari *et al*. (1988) later published evidence of the slower growth of a Δ*crp* mutant in comparison to the wild type. During steady state growth in medium supplemented with glucose and cas-amino acids, the wild type strain had a doubling time of 27 minutes. However, in the same medium, the Δ*crp* mutant doubled every 44 minutes. In addition to *crp*, genes encoding ribosomal proteins were predicted to be essential from the transposon sequencing data. As examples, *rplA*, *rplY*, *rpsO* and *rpsT*, the genes encoding for the ribosomal proteins L1, L25, S15 and S20 respectively, were predicted to be essential from both LB and NTL datasets. In unicellular organisms such as *E. coli*, protein synthesis is a rate limiting factor for growth (Paier *et al*., 2014). Disruption of each of

the five example genes above leads to a slower growth rate, likely explaining the low insertion representation seen in the LB and NTL datasets.

There is published evidence of essentiality for two more genes in the candidate essential gene list, one of which is *yciM*. Mahalakshmi *et al*. (2014) created a strain of *E. coli* in which the chromosomal copy of *yciM* was replaced with a kanamycin resistance cassette. The strain also contained a plasmid containing *yciM* under the control of an IPTG inducible promoter. In the presence of IPTG the strain grew normally in LB and on minimal A agar. In the absence of IPTG, the strain grew poorly on LB agar and not at all on minimal A agar. Additionally, this strain was shown to lyse after approximately 3 hours of growth and exhibit morphological aberrations when grown without IPTG. Another gene for which evidence of essentiality can be found is *hda*. Kato and Katayama (2001) did complementation studies which suggested that *hda* was essential for cell viability. The chromosomal *hda* coding sequence was deleted from an *E. coli* strain containing a wild type *hda* copy on a plasmid. P1 phage transduction was attempted from this strain to strains with or without the *hda* containing plasmid, and transduction only occurred successfully into cells containing the plasmid.

For the remaining 23 of the 70 genes in the candidate essential gene list, no evidence for either growth defects or essentiality upon disruption could be found in the literature. As such, these genes need further characterisation and experimentation to establish whether they are truly essential or important for cellular growth.

Out of the 60 genes manually inspected and defined as likely non-essential, 26 genes were originally reported as essential in the KEIO data. After searching the literature for more information on these genes, evidence of non-essentiality was found for all but one of them.

This evidence varies from studies with mutants with large genome deletions, transposon disruption and gene deletion. An example gene from this list is *entD.* Coderre and Earhart (1989) reported that cells were still viable even when containing an inactivating Tn5 insertion within the *entD* coding sequence. Several deletion mutants of *entD* were also viable. Both of these findings suggest that *entD* is non-essential. There is very little literature available for the remaining gene of which there is no evidence of non-essentiality, *yceQ*.

Literature evidence of non-essentiality can be found for 24 of the remaining 34 genes out of the 60 defined as likely to be non-essential. The remaining 10 genes for which no further evidence can be found are all uncharacterised Y genes of unknown function.

To summarise this analysis, the list of 405 candidate essential genes have been finalised into three lists. What will be known as the core essential gene list is shown in Table 4.1. This list of 290 genes includes the 274 genes from section 4.3, as well as the 14 extra essential genes from Yamamoto *et al*. (2009) and the *yciM* and *hda* genes for which literature evidence of essentiality was found. Genes that are unlikely to be essential are shown in Table 4.1, and genes likely to impact growth are shown in Table 4.2.

4.2.4 Cluster of orthologous groups (COG) analysis. The cluster of orthologous group (COG) categories were determined for each gene in the summarised gene lists by using the eggNOG database (Huerta-Cepas *et al*., 2015). The counts for each category were tallied and shown in Figure 4.8. COGs are a method of functionally classifying proteins by comparison with proteins from multiple phylogenies. A single COG consists of a group of orthologous proteins in multiple organisms across multiple phylogenies. Each COG then corresponds to a generalised function. As expected for the core essential gene list, the majority of genes were

shown to be involved with central cellular processes. Over 50% of genes in this list were found in the three categories of translation, envelope maintenance and coenzyme metabolism. Genes involved with translation comprised the largest single category, containing over a quarter of all genes in the list. This indicates the fundamental importance of protein synthesis to the cell. This pattern was the same in the list of genes important for growth, the majority of which were related to translation. Otherwise, the genes in this list were broadly split over multiple categories. For the list of non-essential genes, the majority of genes were in the unknown function category. The rest of the genes were evenly spread across categories.

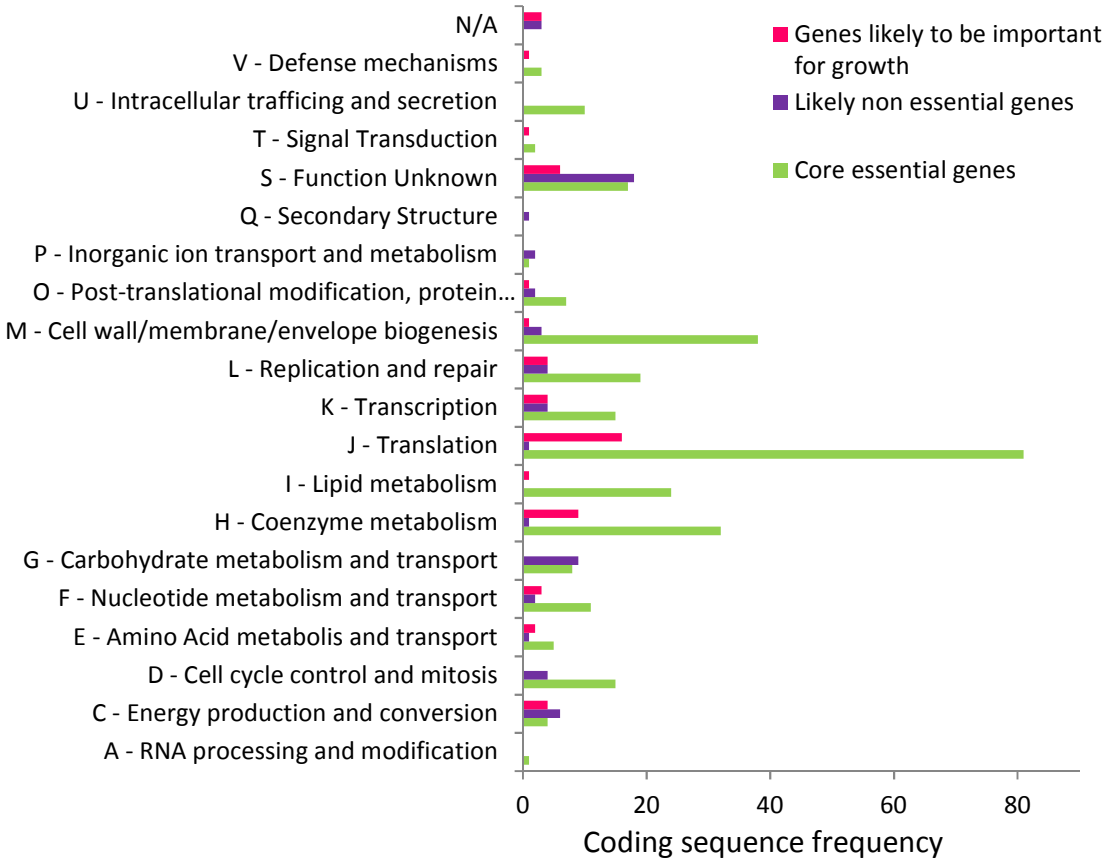

Figure 4.8 Distribution of COG categories in the summarised gene lists. Each Cluster of orthologous group (COG) category contains genes with a related biological function. Each of the 405 genes were classified and separated into their COG categories (Huerta-Cepas et al., 2015), which are shown in the histogram above.

Table 4.1. Core essential genes.

Gene	Description	COG number	COG category	Notes
accA	Acetyl-CoA carboxylase, carboxyltransferase alpha subunit	0825	I	All 3 datasets
accB	Acetyl-CoA carboxylase, biotin carboxyl carrier protein; BCCP; homodimeric	0511	I	All 3 datasets
accC	Acetyl-CoA carboxylase, biotin carboxylase (BC) subunit	0439	I	All 3 datasets
accD	Acetyl-CoA carboxylase, carboxyltransferase beta subunit	0777	i	All 3 datasets
acpP	Acyl carrier protein ACP	0236	I	All 3 datasets
acpS	ACP-CoA phosphopantetheinyltransferase; Holo-ACP synthase; 4'-phosphopantetheinyl transferase	0736	I	All 3 datasets
adk	Adenylate kinase; weak nucleoside diphosphate kinase activity; pleiotropic effects on glycerol-3-phosphate acyltransferase activity; monomeric	0563	F	All 3 datasets
alaS	Alanine--tRNA ligase, autorepressor	0013	J	Extra copy in Yamamoto et al., 2009
argS	Arginine--tRNA ligase	0018	J	All 3 datasets
asd	Aspartate semialdehyde dehydrogenase	0136	E	All 3 datasets
asnS	Asparagine--tRNA ligase	0017	J	All 3 datasets
aspS	Aspartate--tRNA ligase	0173	J	All 3 datasets
bamA/yaeT	Outer membrane protein required for OM biogenesis; in BamABCDE complex; forms pores; PORTA repeats	4775	M	All 3 datasets
bamD/yfiO	TPR-repeat lipoprotein required for OM biogenesis; in BamABCDE complex	4105	M	All 3 datasets
birA	Bifunctional biotin protein ligase, biotin operon repressor; biotin-[acetyl-CoA carboxylase] holoenzyme synthase; monomeric	0340	H	All 3 datasets
can	Carbonic anhydrase, beta class	0288	P	All 3 datasets
cca	tRNA nucleotidyltransferase, repairs terminal CCA of tRNAs	0617	J	All 3 datasets
cdsA	CDP-diglyceride synthase, integral membrane protein with eight transmembrane helices; also known as phosphatidate cytidylyltransferase	0575	I	All 3 datasets

coaA	Pantothenate kinase	1072	H	Extra copy in Yamamoto et al., 2009
coaD	Phosphopantetheine adenylyltransferase	0669	H	All 3 datasets
coaE	Dephospho-CoA kinase; final step in CoA synthesis	0237	H	Extra copy in Yamamoto et al., 2009
cohE/ymfK	CI-like repressor, e14 prophage	1974	K	All 3 datasets
csrA	Global regulator of carbon source metabolism; RNA binding protein	1551	T	3' insertions
cydA	Cytochrome d (bd-I) ubiquinol oxidase subunit 1; upregulated in biofilms and microaerobic conditions; aerobically repressed by H-NS; anaerobically repressed by FNR	1271	C	All 3 datasets
cydC	Glutathione/cysteine ABC transporter permease/ATPase; exports glutathione and cysteine to the periplasm as required for cytochrome assembly	4987	V	Essential regions
cysS	Cysteine--tRNA ligase; binds Zn(II)	0215	J	All 3 datasets
dapA	Dihydrodipicolinate synthase	0329	E	All 3 datasets
dapB	Dihydrodipicolinate reductase	0289	E	All 3 datasets
dapD	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase; mutations suppress growth defects of strains lacking superoxide dismutase	2171	E	All 3 datasets
dapE	N-succinyl-diaminopimelate desuccinylase, DAP/lysine biosynthesis, contains Zn(2+)/Co(2+)	0624	E	All 3 datasets
def	Peptide deformylase; N-formylmethionylaminoacyl-tRNA deformylase; PDF	0242	J	All 3 datasets
degS	Serine protease, degrades periplasmic RseA, activating RpoE; multicopy suppressor of prc; periplasmic stress sensor for unfolded or misfolded OMPs	0265	o	5' essential region
der	Multicopy suppressor of ftsJ, GTPase, ribosome biogenesis; depleted cells form filaments with defective chromosome segregation; Der-YhiI complex	1160	S	All 3 datasets
dfp	Coenzyme A biosynthesis, bifunctional enzyme; phosphopantothenoylcysteine decarboxylase (N) and phosphopantothenoylcysteine synthase (C)	0452	H	All 3 datasets
dicA	Transcriptional repressor for dicB, Qin prophage	1396	K	All 3 datasets

dnaA	DNA synthesis initiator and global transcription regulator; binds DNA at DnaA boxes, binds cardiolipin and other acidic phospholipids, binds ATP	0593	L	All 3 datasets
dnaB	Replicative DNA helicase; DNA-dependent ATPase involved in DNA synthesis; binds DNA contrahelicase termination protein Tus at Ter sites; possibly involved in DNA recombination	0305	L	All 3 datasets
dnaC	DNA biosynthesis, helicase DnaB loader; dual ATP/ADP switch protein	0305	L	All 3 datasets
dnaE	DNA polymerase III, alpha subunit; suppressor of dnaG-Ts	0587	L	All 3 datasets
dnaG	Primase for DNA replication; primer synthesis for leading- and lagging-strand synthesis; binds Zn(II)	0358	L	Extra copy in Yamamoto et al., 2009
dnaN	DNA polymerase III sliding clamp beta subunit; required for high processivity; required for regulatory inactivation of DnaA	0592	L	All 3 datasets
dnaX	DNA polymerase III holoenzyme, tau and gamma ATPase subunits; gamma chain (aa 1-431) is main subunit of the clamp loader complex	2812	L	All 3 datasets
dut	dUTP pyrophosphatase; dUTPase	0756	F	All 3 datasets
dxr	1-deoxy-D-xylulose 5-phosphate (DXP) reductoisomerase, NAPDH-dependent; also called 2-C-methyl-D-erythritol 4-phosphate (MEP) synthase; alternative nonmevalonate (DXP) pathway for terpenoid biosynthesis; dimeric	0743	I	All 3 datasets
dxs	DXP synthase; DXP is precursor to isoprenoids, thiamine, pyridoxol	1154	H	All 3 datasets
eno	Enolase; phosphoprotein; component of RNA degradosome	0148	G	All 3 datasets
era	Ribosome-associated GTPase essential for growth; also required for a normal adaptation response to thermal stress; GTP-dependent autophosphorylating protein kinase activity; membrane-associated, 16S rRNA-binding protein; cell cycle arrest	1159	S	All 3 datasets
erpA/yadR	Iron-sulfur cluster insertion protein; A-type Fe-S protein; essential for respiratory growth	0316	S	All 3 datasets
fabA	3R-3-hydroxydecanoyl acyl carrier protein (ACP) dehydratase; also called beta-hydroxydecanoylthioester dehydrase	0764	I	All 3 datasets
fabB	3-oxoacyl-[acyl-carrier-protein] synthase I; beta-Ketoacyl-ACP synthase I; KAS I; homodimeric	0304	I	All 3 datasets

fabD	Malonyl-CoA-acyl carrier protein transacylase	0331	I	All 3 datasets
fabG	Beta-ketoacyl-ACP reductase	ENOG410XNW1	S	All 3 datasets
fabI	Enoyl-ACP reductase, NADH dependent	0623	I	All 3 datasets
fabZ	3R-hydroxymyristoyl acyl carrier protein (ACP) dehydratase	0764	I	All 3 datasets
fbaA	Fructose 1,6-bisphosphate aldolase, class II; binds Zn(II); homodimeric	0191	G	All 3 datasets
ffh	Signal Recognition Particle (SRP) protein, with 4.5S RNA; GTPase involved in co-translational protein translocation into and through membranes	0541	U	All 3 datasets
fldA	Flavodoxin I	0716	C	All 3 datasets
fmt	Methionyl-tRNA formyltransferase	0223	J	All 3 datasets
folA	Dihydrofolate reductase; trimethoprim resistance	0262	H	All 3 datasets
folC	Dihydrofolate:folylpolyglutamate synthase	0285	H	All 3 datasets
folD	Methenyltetrahydrofolate dehydrogenase/cyclohydrolase	0190	H	All 3 datasets
folE	GTP cyclohydrolase I	0302	H	All 3 datasets
folK	6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase; monomeric	0801	H	3' insertions
frr	Ribosome recycling factor (RRF); dissociates ribosomes from mRNA after termination of translation; tRNA mimic	0233	J	All 3 datasets
ftsA	Cell division and septation protein, specific role unknown; recruited to FtsZ ring	0849	D	All 3 datasets
ftsB	Membrane protein required for cell division; septum localization dependent on Ftsl and FtsQ	2919	D	5' and 3' insertions
ftsH	ATP-dependent membrane protease, complexed with HflCK; regulates lysogeny; mutants are defective in cell growth, septum formation and phage lambda development; mutants rescued by divalent cations; binds Zn(II); hexameric	0465	O	3' insertions
ftsI	Transpetidase, PBP3; penicillin-binding protein 3 involved in septal peptidoglycan synthesis	0768	M	All 3 datasets
ftsK	DNA translocase at septal ring sorting daughter chromsomes	1674	D	5' essential region
ftsL	Cell division and growth, membrane protein	3116	D	5' insertions

ftsN	Cell division and growth; multicopy suppresses ftsA12	3087	D	5' essential region
ftsQ	Divisome assembly protein; cell division and growth of wall at septum	1589	M	All 3 datasets
ftsW	Putative lipid II flippase; divisome protein recruiting FtsI; SEDS protein	0772	D	All 3 datasets
ftsY	Signal recognition particle (SRP) receptor, GTPase	0552	U	All 3 datasets
ftsZ	Septal ring GTPase required for cell division and growth; initiation of septation; tubulin-like protein	0206	D	All 3 datasets
fusA	Elongation Factor EF-G; GTPase required for translocation from the A-site to the P-site in the ribosome; fusidic acid resistance	0480	J	All 3 datasets
gapA	Glyceraldehyde 3-P dehydrogenase A	0057	G	All 3 datasets
glmM	Phosphoglucosamine mutase; UDP-GlcNAc pathway, peptidoglycan, lipopolysaccharide synthesis; mRNA stability effects	1109	G	Extra copy in Yamamoto et al., 2009
glmS	Glucosamine-6-phosphate synthase; glucosamine--fructose-6-phosphate aminotransferase; C-terminal F6P-binding domain has isomerase activity	0449	M	All 3 datasets
glmU	Bifunctional glucosamine-1-phosphate acetyltransferase and N-acetylglucosamine-1-phosphate uridylyltransferase, hexameric	1207	M	All 3 datasets
glnS	Glutamine--tRNA ligase	0008	J	All 3 datasets
gltX	Glutamate--tRNA ligase	0008	J	All 3 datasets
glyQ	Glycine--tRNA ligase, alpha-subunit	0752	J	All 3 datasets
glyS	Glycine--tRNA ligase, beta-subunit	0751	J	Extra copy in Yamamoto et al., 2009
gmk	Guanylate kinase	0194	F	All 3 datasets
gpsA	sn-Glycerol-3-phosphate dehydrogenase [NAD(P)+]	0240	C	All 3 datasets
groL	Chaperonin Cpn60; phage morphogenesis; GroESL large subunit GroEL, weak ATPase; binds Ap4A	0459	O	Extra copy in Yamamoto et al., 2009

groS	Chaperonin Cpn10; GroESL small subunit GroES; phage morphogenesis	0234	O	All 3 datasets
grpE	Nucleotide exchange factor for the DnaKJ chaperone; heat shock protein; mutant survives lambda induction; stimulates DnaK and HscC ATPase	0576	O	3' essential region
gyrA	DNA gyrase, subunit A; nalidixic acid resistance; cold shock regulon	0188	L	All 3 datasets
gyrB	DNA gyrase, subunit B; novobiocin, coumermycin resistance	0187	L	All 3 datasets
hda	Required for regulatory inactivation of DnaA; multicopy suppressor of dnaN(ts)	0593	L	Kato and Katayama, 2001
hemA	Glutamyl-tRNA reductase, hemin biosynthesis; neomycin sensitivity	0373	h	All 3 datasets
hemB	5-Aminolevulinate dehydratase; also known as porphobilinogen synthase; binds Zn(II)	0113	H	All 3 datasets
hemC	Porphobilinogen deaminase; neomycin sensitivity	0181	H	All 3 datasets
hemD	Uroporphyrinogen III synthase; neomycin sensitivity	1587	H	All 3 datasets
hemG	Protoporphyrinogen oxidase; neomycin sensitivity; flavodoxin-like	4635	H	All 3 datasets
hemH	Ferrochelatase	0276	H	All 3 datasets
hemL	Glutamate-1-semialdehyde aminomutase	0001	h	All 3 datasets
hisS	Histidine--tRNA ligase	0124	J	All 3 datasets
holA	DNA polymerase III, delta subunit; part of the DnaX clamp loader complex; acts as a wrench to open the sliding clamp	1466	I	All 3 datasets
holB	DNA polymerase III, delta' subunit; part of the DnaX clamp loader complex, the stator protein	0470	L	All 3 datasets
ileS	Isoleucine--tRNA ligase	0060	J	Extra copy in Yamamoto et al., 2009
infA	Translation initiation factor IF-1	0361	J	All 3 datasets
infB	Translation initiation factor IF-2	0532	J	All 3 datasets
infC	Translation initiation factor IF-3; unusual AUU start codon	0290	J	All 3 datasets
ispA	Farnesyl diphosphate synthase, isoprenoid biosynthesis	0142	H	All 3 datasets

ispB	Octaprenyl diphosphate synthase, isoprenoid biosynthesis	0142	H	All 3 datasets
ispD	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; alternative nonmevalonate (DXP) pathway for terpenoid biosynthesis; essential gene	1211	I	All 3 datasets
ispE	4-diphosphocytidyl-2-C-methylerythritol kinase; isopentenyl phosphate kinase; alternative nonmevalonate (DXP) pathway for terpenoid biosynthesis; essential gene	1947	I	All 3 datasets
ispF	2-C-methyl-D-erythritol 2,4-cyclodiphosphate (MECP) synthase; alternative nonmevalonate (DXP) pathway for terpenoid biosynthesis; essential gene; trimeric	0245	I	All 3 datasets
ispG	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase; alternative nonmevalonate (DXP) pathway for terpenoid biosynthesis; [4Fe-4S] protein	0821	I	All 3 datasets
ispH	4-hydroxy-3-methylbut-2-enyl diphosphate reductase; last, branched, step of isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) synthesis from 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate; alternative nonmevalonate (DXP) pathway for terp	0761	I	All 3 datasets
ispU	Undecaprenyl pyrophosphate synthase; dimeric	0020	I	All 3 datasets
kdsA	3-deoxy-D-manno-octulosonate 8-phosphate (KDO8-P) synthase; LPS biosynthesis	2877	M	All 3 datasets
kdsB	3-deoxy-manno-octulosonate cytidylyltransferase; CMP-KDO synthase (CKS); LPS biosynthesis	1212	M	All 3 datasets
lepB	Signal peptidase I; SPI; responsible for type I signal cleavages of periplasmic, OM, some IM, and extracellular proteins	0681	U	All 3 datasets
leuS	Leucine--tRNA ligase	0495	J	All 3 datasets
lexA	Global regulator (repressor) for SOS regulon; dimeric	1974	K	All 3 datasets
lgt	Phosphatidylglycerol:prolipoprotein diacylglycerol transferase	0682	M	All 3 datasets
ligA	DNA ligase A, NAD(+)-dependent	0272	L	All 3 datasets
lnt	Apolipoprotein N-acyltransferase; copper sensitivity	0815	M	All 3 datasets
lolA	Periplasmic protein responsible for sorting and transporting lipoproteins to outer membrane	2834	M	All 3 datasets
lolB	OM lipoprotein required for localization of lipoproteins	3017	M	All 3 datasets
lolC	LolA-dependent release of lipoproteins from inner membrane; essential gene	4591	M	All 3 datasets

lolD	LolA-dependent release of lipoproteins from inner membrane; essential gene	1136	V	All 3 datasets
lolE	LolA-dependent release of lipoproteins from inner membrane; essential gene	4591	M	All 3 datasets
lptA/yhbN	LPS export ABC transporter periplasmic binding protein; Lipid A binding protein; LPS export and assembly protein	1934	s	5' essential region
lptB/yhbG	LPS export ABC transporter ATPase	1137	S	Extra copy in Yamamoto et al., 2009
lptC/yrbK	Periplasmic membrane-anchored LPS-binding protein; LPS export	3117	s	5' essential region
lptD/imp	LPS assembly OM complex LptDE, beta-barrel component	1452	M	All 3 datasets
lptE/rlpB	LPS assembly OM complex LptDE, LPS-binding lipoprotein component	2980	M	All 3 datasets
lptF/yjgP	LPS export ABC transporter permease	0795	S	All 3 datasets
lptG/yjgQ	LPS export ABC transporter permease	0795	S	All 3 datasets
lpxA	Lipid A synthesis, UDP-N-acetylglucosamine acyltransferase	1043	M	All 3 datasets
lpxB	Lipid A disaccharide synthase	0763	M	All 3 datasets
lpxC	Lipid A synthesis, UDP-3-O-(R-3-hydroxymyristoyl)-N-acetylglucosamine deacetylase; zinc metalloamidase; cell envelope and cell separation	0774	M	All 3 datasets
lpxD	Lipid A synthesis, UDP-3-O-(R-3-hydroxymyristoyl)-glucosamine N-acyltransferase	1044	M	All 3 datasets
lpxH	Lipid A synthesis, UDP-2,3-diacylglucosamine pyrophosphohydrolase	2908	S	All 3 datasets
lpxK	Lipid A 4' kinase	1663	M	All 3 datasets
lspA	Prolipoprotein signal peptidase, signal peptidase II; SPII	0597	mu	All 3 datasets
map	Methionine aminopeptidase	0024	J	All 3 datasets
metG	Methionine--tRNA ligase	0143	J	All 3 datasets
metK	S-adenosylmethionine synthase; methionine adenosyltransferase; ethionine sensitivity; essential gene	0192	H	All 3 datasets
minE	Blocks MinCD inhibition of FtsZ polymerization at cell center; forms membrane-associated coiled arrays in a ring at the cell center	0851	d	5' essential region

mqsA/ygiT	Antitoxin for MqsR toxin; transcriptional repressor	1396	k	5' essential region
mraY	UDP-N-acetylmuramoyl-pentapeptide:undecaprenyl-PO4 phosphatase	0472	M	All 3 datasets
mrdA	Penicillin-binding protein PBP2; transpeptidase recruited by cognate SEDS protein MrdB; mecillinam resistance	0768	M	All 3 datasets
mrdB	Affects cell shape, mecillinam sensitivity; recruits cognate transpetidase MrdA; SEDS protein	0772	D	All 3 datasets
mreB	Cell wall structural actin-like protein in MreBCD complex; mecillinam resistance protein	1077	D	All 3 datasets
mreC	Cell division and growth; mecillinam resistance; rod shape-determining protein	1792	M	5' essential region
mreD	Mecillinam resistance; rod shape-determining protein	2891	M	All 3 datasets
msbA	Lipid exporter, fused permease and ATPase components; exports LPS, phospholipids, and lipid A to the outer membrane outer leaflet; drug export and resistance; ABC family transporter; flippase; biogenesis of outer membrane; lipid-activated ATPase	1132	V	All 3 datasets
mukB	Chromosome condensin MukBEF, ATPase and DNA-binding subunit; SMC-related protein	3096	D	All 3 datasets
mukE	Chromosome condensin MukBEF, MukE localization factor	3095	D	All 3 datasets
mukF	Chromosome condensin MukBEF, kleisin-like subunit, binds calcium	3006	D	All 3 datasets
murA	UDP-N-acetylglucosamine enoylpyruvyl transferase; fosfomycin resistance	0766	m	All 3 datasets
murB	UDP-N-acetylenolpyruvoylglucosamine reductase, FAD-binding	0812	m	All 3 datasets
murC	UDP-N-acetylmuramate:L-alanine ligase; L-alanine adding enzyme	0773	m	All 3 datasets
murD	D-glutamic acid adding enzyme; UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase	0771	m	All 3 datasets
murE	meso-diaminopimelate adding enzyme; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate:meso-diaminopimelate ligase	0769	m	All 3 datasets
murF	D-alanyl:D-alanine adding enzyme; UDP-N-acetylmuramoyl-tripeptide:D-alanyl-D-alanine ligase	0770	M	All 3 datasets
murG	N-acetylglucosaminyl transferase; UDP-N-acetylglucosamine:N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase; murein synthesis peripheral membrane protein interacting with cardiolipin	0707	M	All 3 datasets

murI	Glutamate racemase, D-glutamate synthesis	0796	M	All 3 datasets
murJ/mviN	Putative lipid II flippase; required for murein synthesis	0728	S	All 3 datasets
nadD	Nicotinate mononucleotide adenylyltransferase, NAD(P) biosynthesis	1057	H	All 3 datasets
nadE	NAD synthase, ammonia dependent	0171	H	All 3 datasets
nadK/yfjB	ATP-NAD kinase	0061	G	All 3 datasets
nrdA	Ribonucleoside-diphosphate reductase 1, alpha subunit; class Ia aerobic ribonucleotide reductase; B1 protein, R1 subunit	0209	F	All 3 datasets
nrdB	Ribonucleoside-diphosphate reductase 1, beta subunit; class Ia aerobic ribonucleotide reductase; B2 protein, R2 subunit	0208	f	All 3 datasets
nusA	Transcription termination/antitermination L factor; mutant survives lambda induction	0195	K	All 3 datasets
nusG	Stabilizes phage lambda protein N-NusA-RNAP antitermination complex	0250	K	All 3 datasets
obgE	DNA-binding GTPase involved in cell partioning and DNA repair; involved in ribosome assembly; GTP-bound form associates with 50S ribosomal subunits; ribosome-associated SpoT ppGpp-degradation stimulator	0536	s	All 3 datasets
orn	3' to 5' oligoribonuclease; mutants accumulate oligoribonucleotides that are 2-5 residues long	1949	A	All 3 datasets
parC	Topoisomerase IV, subunit A, ATP-dependent, type II; chromosome decatenase; relaxes both positive and negative supercoils; DNA unknotting activity; heterotetrameric	0188	L	Extra copy in Yamamoto et al., 2009
parE	Topoisomerase IV, subunit B, ATP-dependent, type II; chromosome decatenase; relaxes positive supercoils much faster than negative supercoils; DNA unknotting activity; heterotetrameric	0187	I	All 3 datasets
pgk	Phosphoglycerate kinase	0126	g	All 3 datasets
pgsA	Phosphatidylglycerophosphate synthase	0558	i	All 3 datasets
pheS	Phenylalanine--tRNA ligase, alpha-subunit	0016	J	All 3 datasets
pheT	Phenylalanine--tRNA ligase, beta-subunit	0072	j	All 3 datasets
plsB	Glycerol-3-phosphate acyltransferase	2937	i	All 3 datasets
plsC	1-Acyl-n-glycerol-3-phosphate acyltransferase; affects partitioning	0204	I	All 3 datasets

polA	DNA polymerase I; required for plasmid replication; translesion synthesis; synthetic lethal with ygdG	0258	L	Yamamoto et al., 2009
ppa	Inorganic pyrophosphatase; binds Zn(II); homohexameric, dimer of trimers	0221	C	All 3 datasets
prfA	Peptide chain release factor 1, RF-1; translation termination factor recognizes UAG and UAA.	0216	j	All 3 datasets
prfB	Peptide chain release factor 2, RF-2; translation termination factor recognizes UGA and UAA; slightly defective allele	1186	J	Extra copy in Yamamoto et al., 2009
prmC	Release factor (RF1, RF2) glutamine methyltransferase	2890	J	All 3 datasets
proS	Proline--tRNA ligase	0442	J	All 3 datasets
prsA/prs	Phosphoribosylpyrophosphate synthase	0462	F	All 3 datasets
psd	Phosphatidylserine decarboxylase, phospholipid biosynthesis	0688	I	3' insertions
pssA	Phosphatidylserine synthase	1183	I	All 3 datasets
pth	Peptidyl-tRNA hydrolase; required for phage lambda growth	0193	J	All 3 datasets
purB	Adenylosuccinate lyase, purine synthesis	0015	f	All 3 datasets
pyrG	CTP synthase; CtpS	0504	F	All 3 datasets
pyrH	Uridylate kinase; hexameric	0528	F	All 3 datasets
racR	Rac prophage repressor	ENOG41126JH	K	All 3 datasets
rho	Transcription termination factor Rho; hexameric; RNA-dependent ATPase; ATP-dependent RNA helicase; bicyclomycin target	1158	K	Extra copy in Yamamoto et al., 2009
ribA	GTP cyclohydrolase II, riboflavin biosynthesis	0807	H	All 3 datasets
ribB	3,4-dihydroxy-2-butanone 4-phosphate synthase; riboflavin biosynthesis; acid-inducible; homodimeric	0108	h	3' insertions
ribC	Riboflavin synthase; homotrimer; associated with RibE 60-mer	0307	H	All 3 datasets
ribD	Bifunctional enzyme for second and third steps in riboflavin biosynthesis; 2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5'-phosphate deaminase; ribosyl reductase	1985	H	All 3 datasets

ribE	Lumazine (6,7-dimethyl-8-ribityllumazine) synthase; 60-mer capsid; penultimate step in the biosynthesis of riboflavin; binds RibC homotrimer	0307	H	All 3 datasets
ribF	Riboflavin kinase and FAD synthase	0196	H	All 3 datasets
rne	RNase E; component of RNA degradosome; mRNA turnover; 5S and 16S RNA maturation	1530	J	5' essential region
rnpA	RNase P, C5 protein component; involved in tRNA and 4.5S RNA-processing	0594	J	5' and 3' insertions
rplB	50S ribosomal subunit protein L2; binds Zn(II)	0090	J	All 3 datasets
rplC	50S ribosomal subunit protein L3	0087	J	All 3 datasets
rplD	50S ribosomal subunit protein L4; erythromycin sensitivity	0088	J	All 3 datasets
rplE	50S ribosomal subunit protein L5; 5S rRNA-binding	0094	J	All 3 datasets
rplF	50S ribosomal subunit protein L6; gentamicin sensitivity	0097	J	All 3 datasets
rplJ	50S ribosomal subunit protein L10; streptomycin resistance	0244	J	All 3 datasets
rplL	50S ribosomal subunit protein L7/L12	0222	J	All 3 datasets
rplM	50S ribosomal subunit protein L13; binds Zn(II)	0102	J	All 3 datasets
rplN	50S ribosomal subunit protein L14	0093	J	All 3 datasets
rplO	50S ribosomal subunit protein L15	0200	J	All 3 datasets
rplP	50S ribosomal subunit protein L16	0197	J	All 3 datasets
rplQ	50S ribosomal subunit protein L17	0203	J	All 3 datasets
rplR	50S ribosomal subunit protein L18; 5S rRNA-binding	0256	J	All 3 datasets
rplS	50S ribosomal subunit protein L19	0335	J	All 3 datasets
rplT	50S ribosomal subunit protein L20	0292	J	All 3 datasets
rplU	50S ribosomal subunit protein L21	0261	J	All 3 datasets
rplV	50S ribosomal subunit protein L22; erythromycin sensitivity	0091	J	All 3 datasets
rplW	50S ribosomal subunit protein L23	0089	J	All 3 datasets
rplX	50S ribosomal subunit protein L24	0198	J	All 3 datasets
rpmA	50S ribosomal subunit protein L27	0211	J	All 3 datasets

rpmB	50S ribosomal subunit protein L28	0227	J	All 3 datasets
rpmC	50S ribosomal subunit protein L29	0255	J	All 3 datasets
rpmD	50S ribosomal subunit protein L30	1841	J	All 3 datasets
rpmH	50S ribosomal subunit protein L34	0230	J	All 3 datasets
rpoA	RNA polymerase, alpha subunit; binds Zn(II)	0202	K	All 3 datasets
rpoB	RNA polymerase, beta subunit; binds Zn(II)	0085	K	All 3 datasets
rpoC	RNA polymerase, beta' subunit; binds Zn(II)	0086	K	All 3 datasets
rpoD	RNA polymerase subunit, sigma 70, initiates transcription; housekeeping sigma	0568	K	Extra copy in Yamamoto et al., 2009
rpoE	RNA polymerase sigma E factor; role in extracytoplasmic, high temperature and oxidative stress responses; sigma 24 initiation factor	1595	K	All 3 datasets
rpoH	RNA polymerase subunit, sigma 32, heat shock transcription	0568	K	All 3 datasets
rpsA	30S ribosomal subunit protein S1; subunit of RNA phage Q beta replicase; binds and stimulates RNAP	0539	J	All 3 datasets
rpsB	30S ribosomal subunit protein S2; binds Zn(II)	0052	J	All 3 datasets
rpsC	30S ribosomal subunit protein S3	0092	J	All 3 datasets
rpsD	30S ribosomal subunit protein S4; NusA-like antitermination factor	0522	J	All 3 datasets
rpsE	30S ribosomal subunit protein S5	0098	J	All 3 datasets
rpsG	30S ribosomal subunit protein S7, mutated stop codon	0049	J	All 3 datasets
rpsH	30S ribosomal subunit protein S8	0096	J	All 3 datasets
rpsI	30S ribosomal subunit protein S9	0103	J	All 3 datasets
rpsJ	30S ribosomal subunit protein S10	0051	J	All 3 datasets
rpsK	30S ribosomal subunit protein S11	0100	J	All 3 datasets
rpsL	30S ribosomal subunit protein S12; RNA chaperone	0048	J	All 3 datasets
rpsM	30S ribosomal subunit protein S13	0099	J	All 3 datasets
rpsN	30S ribosomal subunit protein S14	0199	J	All 3 datasets

rpsP	30S ribosomal subunit protein S16; endonuclease	0228	J	All 3 datasets
rpsQ	30S ribosomal subunit protein S17	0186	J	All 3 datasets
rpsR	30S ribosomal subunit protein S18	0238	J	All 3 datasets
rpsS	30S ribosomal subunit protein S19	0185	J	All 3 datasets
rseP/yaeL	Inner membrane zinc RIP metalloprotease; activates RpoE by degrading RseA; multicopy rpoE suppresses rseP mutation	0750	M	5' insertions
secA	Preprotein translocase secAYEG receptor/ATPase subunit; autogenous translational repressor; ATP-dependent helicase activity on secMA mRNA; homodimeric/monomeric	0653	u	All 3 datasets
secD	SecDFyajC inner membrane secretion protein complex subunit; assists the SecYEG translocon to interact with SecA and export proteins	0342	u	
secE	SecYEG inner membrane translocon core subunit; preprotein translocase secAYEG subunit; core translocon secYE subunit	0690	U	All 3 datasets
secF	SecDFyajC inner membrane secretion protein complex subunit; assists the SecYE core translocon to interact with SecA and export proteins	0341	U	3' essential region
secY	SecYEG inner membrane translocon core subunit; preprotein translocase secAYEG subunit; core translocon secYE subunit	0201	u	All 3 datasets
serS	Serine--tRNA ligase; serine hydroxamate resistance	0172	J	All 3 datasets
spoT	ppGpp 3'-pyrophosphohydrolase and ppGpp synthase II; guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	0317	KT	5' essential region
ssb	Single-stranded DNA-binding protein; alkali-inducible; homotetramer	0629	l	All 3 datasets
suhB	Inositol-1-monophosphatase; mutation suppresses TS growth phenotype of rpoH15, dnaB121, and secY24; suhB mutations confers CS growth	0483	G	All 3 datasets
tadA	A34-tRNA adenosine deaminase; point mutation confers resistance to HokC(Gef)-mediated cell killing; essential gene; homodimeric	0590	FJ	5' insertions
thiL	Thiamine monophosphate kinase, involved in thiamine salvage	0611	h	All 3 datasets
thrS	Threonine--tRNA ligase, autogenously regulated; binds Zn(II)	0441	j	All 3 datasets
tilS	tRNA(Ile) lysidine (L34) synthase, ATP-dependent; solely responsible for tRNA(Ile) lysidine 34 (L34) formation	0037	d	All 3 datasets

tmk	Thymidylate kinase	0125	f	All 3 datasets
topA	Topoisomerase I; omega protein	0550	l	All 3 datasets
trmD	tRNA m(1)G37 methyltransferase, SAM-dependent	0336	j	All 3 datasets
trpS	Tryptophan--tRNA ligase	0180	j	All 3 datasets
tsaB/yeaZ	tRNA(NNU) t(6)A37 threonylcarbamoyladenosine modification; binding partner and protease for TsaD	1214	O	All 3 datasets
tsaC/yrdC	tRNA(NNU) t(6)A37 threonylcarbamoyladenosine modification; threonine-dependent ADP-forming ATPase	0009	J	All 3 datasets
tsaD/ygjD	tRNA(NNU) t(6)A37 threonylcarbamoyladenosine modification; glycation binding protein	0533	o	All 3 datasets
tsaE/yjeE	tRNA(NNU) t(6)A37 threonylcarbamoyladenosine modification; ADP binding protein	0802	S	All 3 datasets
tsf	Translation elongation factor EF-Ts; exchanges GDP for GTP in EF-Tu-GDP complex; binds Zn(II); subunit of RNA phage Q beta replicase	0264	J	All 3 datasets
tyrS	Tyrosine--tRNA ligase	0162	J	All 3 datasets
ubiA	4-Hydroxybenzoate polyprenyltransferase	0382	H	All 3 datasets
ubiB	Regulator of octaprenylphenol hydroxylation, ubiquinone synthesis; regulator of 2'-N-acetyltransferase; putative ABC1 family protein kinase	0661	S	All 3 datasets
ubiD	3-octaprenyl-4-hydroxybenzoate carboxylyase; ubiquinone biosynthesis, third step; UbiX isozyme	0043	H	All 3 datasets
valS	Valine--tRNA ligase	0525	J	All 3 datasets
waaA/kdtA	3-deoxy-D-manno-octulosonate(Kdo)-lipid A transferase	1519	M	All 3 datasets
waaU/rfaK	Adds terminal GlcNac side branch to the lipopolysaccharide core prior to attachment of the O antigen; not the same as Salmonella rfaK	0859	M	5' essential region
wzyE	Wzy protein involved in ECA polysaccharide chain elongation; involved in polymerization of the UDP-linked ECA trisaccharide repeat unit of cyclic enterobacterial common antigen ECA(CYC)	ENOG410XT3V	M	All 3 datasets
yciM	LPS regulatory protein; putative modulator of LpxC proteolysis; EnvC-interacting protein; N-terminally anchored cytoplasmic protein; rubredoxin-type redox-sensitive iron center; TPR-repeats-containing protein	2956	g	Mahalakshmi et al., 2014

yejM	Essential inner membrane DUF3413 domain protein; lipid A defect; membrane permeability defect	3083	S	5' essential region
yidC	Membrane protein insertase; inner membrane protein integration factor; binds TM regions of nascent IMPs; required for Sec-independent IMP integration; associated with the Sec translocase	0706	U	All 3 datasets
yihA	GTP-binding protein required for normal cell division; predicted GTPase; also binds GDP	0218	S	All 3 datasets
yqgF	Putative anti-termination factor for Rho-dependent terminators	0816	L	All 3 datasets
yrfF	Putative RcsCDB-response attenuator; inner membrane protein	ENOG410XNRW	s	All 3 datasets
zipA	FtsZ stabilizer; septal ring structural protein for cell division and growth	3115	D	all 3 datasets

[1] Gene descriptions obtained from Ecogene (Zhou and Rudd, 2012). The COG categories were obtained from eggNOG (Huerta-Cepas et al., 2015).

Table 4.2. Genes defined as likely important for growth.

Gene	Description[1]	COG number	COG category	Evidence	Datasets[2]
crp	cAMP-activated global transcription factor; mediator of catabolite repression; CRP; CAP	0664	T	Perrenoud and Sauer, 2005: D'Ari et al., 1988	xL
cydB	Cytochrome d (bd-I) ubiquinol oxidase subunit 2; upregulated in biofilms and microaerobic conditions; aerobically repressed by H-NS; anaerobically repressed by FNR	1294	C	Mempin et al., 2013	NL
cydD	Glutathione/cysteine ABC transporter permease/ATPase; exports cysteine to periplasm as required for cytochrome assembly	4988	V	Pittman et al., 2002: Sezonov, Joselau-Petit and D'Ari, 2007	xL
dcd	dCTP deaminase; deoxycytidine triphosphate deaminase; mutants suppress lethal dut mutants	0717	F	n/a	NL
dnaK	Hsp70 molecular chaperone, heat-inducible; bichaperone with ClpB for protein disaggregation	0443	o	Bukau and Walker, 1989	xL
dnaT	Primasomal protein i	ENOG410ZNDQ	L	n/a	NL
efp	Polyproline-specific translation elongation factor EF-P	0231	j	Yanagisawa et al., 2010	xL
fabH	Beta-ketoacyl-ACP synthase III; KAS III; monomer	0332	I	Yao et al., 2012	NL
folB	Dihydroneopterin aldolase	1539	H	n/a	NL
folP	Dihydropteroate synthase	0294	H	n/a	xL
glyA	Serine hydroxymethyltransferase; binds Zn(II)	0112	E	n/a	NL
guaA	GMP synthase	0518	F	n/a	NL
hemE	Uroporphyrinogen decarboxylase	0407	H	n/a	NL
hipB	Antitoxin of HipAB TA pair; transcriptional repressor of the hipBA operon; role in persister formation	1396	k	n/a	NL
holD	DNA polymerase, psi subunit, clamp loader complex subunit	3050	L	Duigou et al., 2014	NL
iscS	Cysteine desulfurase, PLP-dependent; used in synthesis of Fe-S clusters and 4-thiouridine; ThiI transpersulfidase; TusA transpersulfidase; YnjE transpersulfidase; MoaD transpersulfidase; pyridoxal phosphate cofactor linked to Lys206	1104	E	Lauhon, 2002	NL

iscU	Iron-sulfur cluster assembly scaffold protein	0822	C	Barras, Loiseau and Py, 2005	NL
lipA	Lipoyl synthase, iron-sulfur protein; SAM-dependent chemistry	0320	H	n/a	NL
lpd	Dihydrolipoyl dehydrogenase, NADH-dependent; E3 component of pyruvate and 2-oxoglutarate dehydrogenases complexes; glycine cleavage system L protein; dihydrolipoamide dehydrogenase; binds Zn(II)	1249	C	Takeuchi et al., 2014	NL
lpxL	Lipid A synthesis, KDO2-lipid IVA lauroyl-ACP acyltransferase; not under heat shock regulation; membrane protein affecting cell division, growth, and high-temperature survival	1560	m	Vorachek-Warren et al., 2002	NL
lysS	Lysine--tRNA ligase, constitutive	1190	j	n/a	NL
nusB	Transcription termination/antitermination factor; mutant survives lambda induction	0781	K	Quan et al., 2005	xL
pdxH	Pyridoxine/pyridoxamine phosphate (PNP/PMP) oxidase; isoniazid resistance	0259	h	n/a	NL
relB	Antitoxin for RelE, Qin prophage; transcriptional repressor of relB operon; mutants have a delayed relaxed regulation of RNA synthesis and slow recovery from starvation	3077	L	n/a	NL
rimM	Ribosome maturation factor; 30S subunit maturation factor; S19 binding protein	0806	J	Hase et al., 2013	NL
rluD	23S rRNA pseudouridine(1911,1915,1917) synthase; mutation suppresses ftsH(Ts) mutants; null mutants grow very poorly in K-12 only	0564	J	Schaub and Hayes, 2011	NL
rnt	RNase T; exoribonuclease T; structured DNA DNase; RNA processing; DNA repair	0847	L	Hsiao et al., 2014	NL
rplA	50S ribosomal subunit protein L1	0081	J	Takeuchi et al., 2014	NL
rplK	50S ribosomal subunit protein L11; kasugamycin sensitivity	0080	J	n/a	NL
rplY	50S ribosomal subunit protein L25; 5S rRNA-binding	1825	J	Aseev, Bylinkina and Boni, 2015	NL
rpmF	50S ribosomal subunit protein L32	0333	J	n/a	NL
rpmI	50S ribosomal subunit protein A (L35)	0291	J	n/a	NL

rpmJ	50S ribosomal subunit protein X (L36)	0257	j	Ikegami et al., 2005	xL
rpsF	30S ribosomal subunit protein S6; suppressor of dnaG-Ts	0360	J	Hase et al., 2013	NL
rpsO	30S ribosomal subunit protein S15	0184	J	Bubunenko et al., 2006	NL
rpsT	30S ribosomal subunit protein S20	0268	J	Bubunenko, Baker and Court, 2007	NL
rpsU	30S ribosomal subunit protein S21	0828	J	n/a	NL
rrmJ	23S rRNA U2552 2'-O-ribose methyltransferase, SAM-dependent; involved in cell division and growth; heat inducible; suppressed by cloned ObgE and Der	0293	J	Hase et al., 2013	xL
thyA	Thymidylate synthase; aminopterin, trimethoprim resistance; homodimer	0207	F	n/a	NL
trmU	tRNA(Gln,Lys,Glu) U34 2-thiouridylase; first step in mnm(5)-s(2)U34-tRNA synthesis; TusE binding partner; antisuppressor	0482	J	n/a	NL
ubiE	Ubiquinone/menaquinone biosynthesis methyltransferase; SAM-dependent; (1) Ubiqinone synthesis, 2-octaprenyl-6-methoxy-1,4-benzoquinone methyltransferase; (2) Menaquinone synthesis, 2-demethylmenaquinone (DMK) methyltransferase	2226	H	Takeuchi et al., 2014	xL
ubiG	SAM:OMHMB methyltransferase; Reactions: 2-octaprenyl-6-hydroxylphenol to 2-octaprenyl-6-methoxyphenol; 2-octaprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinone to ubiquinone 8	2227	H	Takeuchi et al., 2014	NL
ubiH	2-octaprenyl-6-methoxyphenol hydroxylase; produces 2-octaprenyl-6-methoxy-1,4-benzoquinone	0654	CH	Takeuchi et al., 2014	xL
ubiX	3-octaprenyl-4-hydroxybenzoate carboxylyase; UbiD isozyme	0163	H	Takeuchi et al., 2014	NL
ybeD	UPF0250 family protein; required for swarming and phage lambda growth	2921	s	Takeuchi et al., 2014	xL
ybeY	ssRNA-specific endoribonuclease; co-endoribonuclease working with RNase R in 16S rRNA 3' end maturation and quality control; rRNA transcription antitermination factor	0319	S	Takeuchi et al., 2014	NL
ydaE	Metallothionein, function unknown, Rac prophage	ENOG410Z75V	S	n/a	NL
ydcD	Putative immunity protein for RhsE	0864	K	n/a	NL

yddK	Pseudogene, frameshifted, leucine-rich protein	4886	s	n/a	NL
ygeF	Pseudogene reconstruction, part of T3SS PAI ETT2 remnant	n/a	n/a	n/a	NL
ygjM	Antitoxin for HigB toxin	5499	k	n/a	xL
ykiB	n/a	n/a	n/a	n/a	NL
ymfE	Predicted membrane protein, function unknown, e14 prophage	ENOG410Y9BP	S	n/a	NL
ymgB	Connector protein for RcsB regulation of biofilm and acid-resistance	ENOG410Y1T2	s	n/a	xL
yncH	IPR020099 family protein required for swarming, function unknown	n/a	n/a	n/a	NL

[1] Gene descriptions obtained from Ecogene (Zhou and Rudd, 2012). The COG categories were obtained from eggNOG (Huerta-Cepas et al., 2015). The evidence column refers to papers which provide evidence of essentiality.

[2]N - Essential in neat transposon library. L - Essential after growth in LB.

Table 4.3. Genes defined as likely non-essential after manual inspection.

Gene	Description[1]	COG number	COG category	Evidence	Datasets[2]
aceE	Pyruvate dehydrogenase, decarboxylase component E1; acetate requirement	2609	C	Ito et al., 2005	xxL
aceF	Pyruvate dehydrogenase, dihydrolipoamide acetyltransferase E2; acetate requirement	0508	C	Ito et al., 2005	xxL
alsK	D-allose kinase	1940	G	Gerdes and Osterman, 2008	Kxx
bcsB	Cellulose synthase, regulatory subunit; binds cyclic-di-GMP; periplasmic, membrane-anchored	ENOG410XNNB	M	Gerdes and Osterman, 2008	Kxx
chpS	ChpS antitoxin, toxin is ChpB	2336	K	Gerdes and Osterman, 2008	Kxx
cmk	Cytidylate kinase; multicopy suppressor of UMP kinase mutations	0283	F	Fricke et al., 1995	xxL
crr	EIIA(Glc), phosphocarrier for glucose PTS transport; negative control of rpoS	2190	G	Guo et al., 2015	xxL
entD	Enterochelin synthase, component D; EntB(ArCP)/EntF-CoA phosphopantetheinyltransferase; facilitates secretion of enterobactin peptide; enterobactin biosynthesis	2977	q	Coderre and Earhart, 1984	Kxx
ftsE	Cell division ATP-binding protein; associated with the inner membrane via FtsX; null mutant has filamentous growth and requires high salt for viability	2884	d	Leeuw et al., 1999	Kxx
ftsX	Integral membrane protein invoved in cell division; binds FtsE to the inner membrane	2177	d	Reddy, 2006	Kxx
gnsB	Multicopy suppressor of secG(Cs) and fabA6(Ts), Qin prophage; overexpression increases unsaturated fatty acid content of phospholipids; gnsA paralog	ENOG410Y8R8	s	Sugai et al., 2001	xxL
guaB	Inosine-5'-monophosphate (IMP) dehydrogenase	0516	F	Kang et al., 2004	xxL
hscA	DnaK-like chaperone Hsc66, IscU-specific chaperone HscAB; involved in FtsZ-ring formation	0443	O	Jang and Imlay, 2010	xxL
icd	Isocitrate dehydrogenase, NADP(+)-specific; e14 attachment site; tellurite reductase	0538	C	Okamoto et al., 2014	xxL
ihfA	Integration Host Factor (IHF), alpha subunit; host infection,	0776	L	Gopel et al., 2011	xxL

	mutant phage lambda; site-specific recombination; sequence-specific DNA-binding transcriptional activator				
lpcA	Phosphoheptose isomerase; D-sedoheptulose 7-phosphate isomerase; GDP-heptose biosynthesis; T-phage resistance	0279	G	Brooke and Valvano, 1996	xxL
mazE/chpR	MazE antitoxin, toxin is MazF	2336	K	Gerdes and Osterman, 2008	Kxx
minD	Inhibitor of FtsZ ring polymerization; chromosome-membrane tethering protein; membrane ATPase that activates MinC	2894	D	Gerdes and Osterman, 2008	Kxx
mlaB/yrbB	Probable phospholipid ABC transporter, quinolone resistance; peripheral membrane protein, cytoplasmic; maintains OM lipid asymmetry; STAS subunit	3113	s	Malinverni and Silhavy, 2009	Kxx
priB	Primosomal protein n; ssDNA-binding protein	2965	L	Bubunenko, Baker and Court, 2007	xNL
ptsH	PTS system histidine phosphocarrier protein HPr; phosphohistidinoprotein-hexose phosphotransferase	1925	G	Gershanovitch et al., 1977	xxL
ptsI	Phosphoenolpyruvate-protein phosphotransferase; phosphotransferase system, enzyme I; E1; PEP-dependent autokinase	1080	G	Hernandez-Montalvo et al., 2003	xxL
rnc	RNase III; cleaves double-stranded RNA	571	K	Bubunenko, Baker and Court, 2007	Kxx
rpe	D-ribulose-5-phosphate 3-epimerase	0036	G	Ito et al., 2005	xxL
rsgA	Ribosome-stimulated GTPase, 30S subunit assemby; low adundance protein; putative RNA binding protein	1162	s	Hase et al., 2009	xNL
rsmI/yraL	16S rRNA C1402 2'-O-ribose methyltransferase, SAM-dependent	0313	s	Dassain et al., 1999	Kxx
secM	Secretion monitor controlling secA expression	ENOG4111GJA	K	Rajapandi, Dolan and Oliver, 1991	Kxx
seqA	Multi-faceted genome stability factor; negative modulator of initiation of replication; replication fork tracking protein required for chromosome segregation; chromosome cohesion protein; hemimethylated GATC binding protein	3057	L	Waldminghaus and Skarstad, 2010	xxL
sucA	2-oxoglutarate dehydrogenase, E1 component; yields succinyl-CoA and CO(2); also known as alpha-ketoglutarate dehydrogenase	0567	C	Nishio et al., 2013	xxL

sucB	2-oxoglutarate dehydrogenase, E2 component; dihydrolipoamide succinyltransferase; acid-inducible; yields succinyl-CoA and CO(2); also known as alpha-ketoglutarate dehydrogenase	0508	C	Kohanski et al., 2007	xNL
tdcF	Putative reactive intermediate deaminase, UPF0076 family; trimeric; reaction intermediate detoxification	0251	J	Gerdes and Osterman, 2008	Kxx
tnaB	Tryptophan:H+ symport permease, low affinity	0814	E	Yanofsky, Horn and Gollnick, 1991	Kxx
tonB	Uptake of chelated Fe(2+) and cyanocobalamin; works in conjunction with OM receptors; energy transducer; sensitivity to T1, phi80, and colicins; forms a complex with ExbB and ExbD	0810	M	Kohanski et al., 2007	xxL
ubiF	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase; produces 2-octaprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinol; required for ubiquinone synthesis; mutation confers resistance to bleomycin, phleomycin and heat	0654	CH	Ito et al., 2005	xxL
yabQ	Pseudogene reconstruction, pentapeptide repeats-containing	ENOG410XV6S	S	Gerdes and Osterman, 2008	Kxx
yafF	Pseudogene, C-terminal fragment, H repeat-associated protein	5433	L	Gerdes and Osterman, 2008	Kxx
yagG	Putative sugar symporter, function unknown, CP4-6; putative prophage remnant	2211	g	n/a	Kxx
ybbD	Pseudogene reconstruction, novel conserved family	1472	G	n/a	xNL
yccK	mnm(5)-s(2)U34-tRNA 2-thiolation step sulfurtransferase; binding partner linking TusBCD to MnmA; may transfer sulfur first to MnmA or directly to tRNA	2920	P	Ikeuchi et al., 2006	xxL
yceQ	Function unknown	ENOG410YYPH	S	n/a	Kxx
yciS	DUF1049 family inner membrane protein	3771	S	Mahalakshmi et al., 2014	xxL
ydaS	Putative Cro-like repressor, Rac prophage	2261	S	n/a	xNL
yddL	Pseudogene, OmpCFN porin family, N-terminal fragment	na	na	n/a	xxL
ydfB	Expressed protein, function unknown, Qin prophage	ENOG4111SFN	S	Gerdes and Osterman, 2008	Kxx
ydfO	DUF1398 family protein, Qin prophage	5562	S	n/a	xNL

ydhR	Predicted monooxygenase, function unknown; dimeric	ENOG4111VBS	S	n/a	xxL
ydiL	Putative HTH domain DNA-binding protein; lambda repressor-like protein	ENOG41120Y0	s	Gerdes and Osterman, 2008	Kxx
yedM	Pseudogene reconstruction, IpaH/YopM family	4886	S	n/a	xNL
yefM	Antitoxin for YoeB toxin; binds YoeB RNase-like domain	2161	D	Gerdes and Osterman, 2008	Kxx
ygeL	Pseudogene reconstruction, part of T3SS PAI ETT2 remnant; response regulator family	na	na	n/a	xNx
ygeM	Pseudogene reconstruction, orgB homolog; part of T3SS PAI ETT2 remnant	na	na	n/a	xNx
yhbV	U32 peptidase family protein, function unknown,	0826	O	Yu et al., 2008	Kxx
yheM	2-thiolation step of mnm(5)-s(2)U34-tRNA synthesis; sulfur relay system; required for swarming phenotype	2923	P	Ikeuchi et al., 2006	xxL
yhhQ	DUF165 family inner membrane protein	1738	s	Gerdes and Osterman, 2008	Kxx
yibJ	Pseudogene, Rhs family	3209	m	Gerdes and Osterman, 2008	Kxx
yigP	Aerobic ubiquinone synthesis protein, SCP2 family protein	3165	S	Aussel et al., 2014	Kxx
ynfN	Cold shock-induced protein, function unknown, Qin prophage	ENOG410Y031	S	n/a	xxL
ypjC	Pseudogene reconstruction	1284	s	n/a	xNL
yqgD	n/a	ENOG410Y8M8	S	Gerdes and Osterman, 2008	Kxx
zwf	Glucose-6-phosphate 1-dehydrogenase	0364	G	Sandoval et al., 2011	xxL

[1] Gene descriptions obtained from Ecogene (Zhou and Rudd, 2012)

[2] K - Essential in KEIO. N - Essential in neat transposon library. L - Essential after growth in LB.

4.3 Discussion

Through a combination of transposon sequencing data, statistical analysis, manual inspection and literature searching, a thorough assessment of the essential genes of *E. coli* BW25113 has been made. After processing of the essential gene lists, a set of 290 core essential genes was generated, of which 248 genes were reported from Baba *et al*. (2006). These were included without any further manual inspection. After manual inspection, 26 genes were found to contain specific patterns of insertions which explained their lack of predicted essentiality from the transposon sequencing data. These genes, when added to the core essential list, brought the number of genes in the list to 274, 94% of the total list. This congruence between the essential gene candidates from this work and that of Baba *et al*. (2006), in addition to the categorisation of these genes as being largely involved with central cellular processes would suggest that these genes are highly likely to be truly essential to the cell. One of the genes not previously found to be essential, *yciM*, was only recently shown to be essential by Mahalakshmi *et al*. (2014). This example demonstrates the capability of transposon sequencing.

In addition to the investigation of essentiality, another benefit of transposon sequencing is in the depth of information provided. Due to the base pair precision with which insertions are defined, interesting and potentially previously unknown information can be learned about particular genes. Most notable from our data is the visualisation of apparently essential regions within single genes. Such regions would not have been visible from the knockout strategy used by Baba *et al*. (2006), in which deletions were made across the majority of the coding sequences. The same can be said for the examples of insertions into 5’ and 3’ regions seen in the LB and NTL datasets. Essential gene regions have been observed in previous work: in *Salmonella enterica* serovars Typhimurium and Typhi, Canals

et al (2012) noted that *yejM* and *ftsN* contained regions that could not be inserted into. Upon closer inspection, the region of *yejM* in which insertion does not occur overlaps entirely with a transmembrane region (Fig. 4.9). Although further work would be required to prove a conclusive link between these findings, it is highly likely that the lack of insertional representation across the transmembrane domain is a biologically relevant finding.

There are also disadvantages in the use of transposon sequencing for essential gene analysis. Possibly the greatest issue with analysing transposon sequencing data is the intensive, in depth analysis required to end up with essential gene lists. More specifically, manual inspection is essential even after the statistical gene prediction, due to the properties of the data. Without manual inspection, the core essential gene list would have been incomplete. The most obvious example of this is *polA*: without manual inspection, this gene would not have been revealed as essential. As of yet, no experimental tool or process is available for the assessment of essential gene regions, which would go some way in automating parts of the manual inspection. While manual inspection is laborious and time consuming, it is undoubtedly important. An example of this lies in *yejM*. This gene was found to be essential in this work after manual inspection, and in other work (Canals *et al*., 2012). Without manual inspection, this gene would have been classed as non-essential. Interestingly, *yejM* is also found in *E. coli* ST131, studied by Phan *et al* (2013). In their study, no manual inspection is undertaken after statistical prediction of gene essentiality, and as a result *yejM* was classified as non-essential. Even though, in this work, it is only a relatively small number of genes that apparently possess essential regions, or even insertions within only their very 5' and/or 3' ends, this specific example serves to illustrate the importance of more in depth investigation.

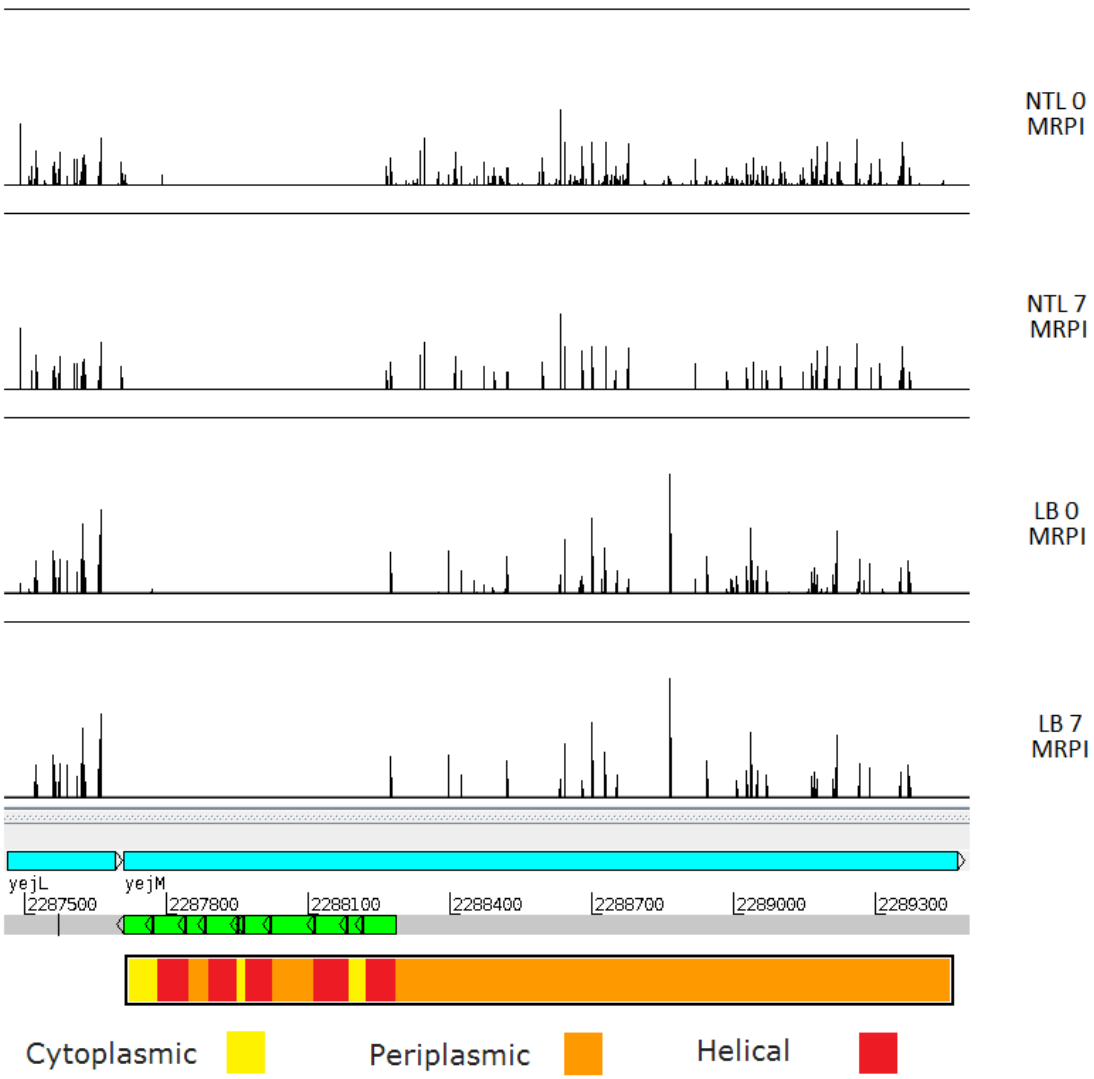

Figure 4.9 The overlap of an essential gene region with the transmembrane domain of *yejM*. The details of the transmembrane domain were obtained from Uniprot (The Uniprot Consortium, 2014). Over the 5' region of *yejM* no insertions can be seen. This whole region maps to the transmembrane domain of the gene.

A subset of 31 of the 405 candidate genes were initially predicted to be essential from the LB and NTL datasets, whereas upon manual inspection they were defined as unlikely to be as such. This is only a small proportion of the total coding sequences of the genome (31/4213). This means that while transposon sequencing is a vastly effective tool for essential gene scanning, in that the vast majority of the coding sequences were found to be non-essential, it is not perfect. Given the position of *E. coli* as a model organism, literature was available to support non-essentiality for the majority of these genes. However, in non-model organisms, there may not be such a wealth of literature available.

One reason for using the NTL and LB samples was to test whether there was any selection imparted by growing the library in LB. This would appear to be true from looking at the number of genes predicted to be essential from each condition: 317 and 356 genes were predicted to be essential from the NTL and LB datasets, respectively. The increased number of genes predicted to be essential after growth would suggest that the growth had imparted a selection on the transposon library.

One final point of discussion lies with the usage of the term “essential”. The essential gene analyses conducted here are specific to *E. coli* during aerobic growth at 37 °C. While some genes might be important for cellular growth under any growth condition, others might only appear to be essential under specific environmental conditions. However, given that the samples tested here are of the transposon library without an extensive amount of growth, it can be assumed that the core essential gene list outlined is representative of genes required for growth even in rich media.

CHAPTER 5

ESCHERICHIA COLI BW25113 CONDITIONAL GENE ANALYSIS IN RESPONSE TO MARKERS FOR OUTER MEMBRANE PERMEABILITY

5.1 Introduction

The physical delineation of the cell is a defining feature of life itself. *E. coli* is a Gram-negative organism, meaning that it has a cell envelope consisting of an inner and an outer lipid membrane (Silhavy, Kahne and Walker, 2010). These are separated by an aqueous compartment called the periplasm which contains the peptidoglycan cell wall. The outer membrane is essential for the viability of *E. coli*, demonstrating its importance.

Cell envelopes exist in a spectrum of integrity. Optimally maintained envelopes prevent nearly all but the controlled movement of molecules from the environment into the cell, preventing the influx of molecules that would not normally enter the cell. The presence of an envelope also allows the control over efflux of molecules out of the cytoplasm. Mutations or insults that completely compromise the integrity of the OM are lethal. However, there are many proteins, which are non-essential under standard laboratory growth, that contribute to OM homeostasis. Mutants lacking these proteins possess cell envelopes that are more permeable which allows molecules larger than the diffusion limit to enter the cell. Mutants exhibiting envelope defects can be selected through the use of markers such as the glycoside antibiotic vancomycin and the anionic surfactant sodium dodecyl sulphate (SDS: Nikaido, 2003; Lazdunski and Shapiro, 1972). Normally, Gram-negative cell envelopes repel both molecules, by the physical occlusion of the sizeable vancomycin and by the charge based repellance of SDS. Mutants which give rise to disrupted membranes allow ingress of these molecules into the cell where they slow or prohibit growth.

Much is already known about the genetic determinants of cell envelope function and maintenance. Multiple experimental strategies have been used to identify these genes (Tamae *et al*., 2008; Liu *et al*., 2010; Nichols *et al*., 2011). However, there are inconsistencies

between these studies. The aim of the work in this chapter was to use TRADIS to provide a global picture of which nonessential genes in *E. coli* are responsible for cell envelope homeostasis.

5.2 Results

5.2.1 Sample Datasets. To produce the conditional samples, 10 µl of transposon library was added into 50 ml of LB (to a starting OD_{600} of ~0.05) containing either 4.8% SDS or 100 µg/ml vancomycin; samples derived from these experiments are designated S4.8 and V100 respectively. Genomic DNA was extracted after the samples were grown to an OD_{600} of 1, and processed using the hybrid methodology as defined in the materials and methods (see Chapter 2). For both conditions there were two biological replicates. Metrics of these datasets are shown in Table 5.1.

For each dataset, insertion indexes were calculated. The reproducibility of the insertion indexes between the replicates was assessed for each condition (Fig. 5.1). For the S4.8 and V100 datasets respectively, R^2 values of 0.96 and 0.97 were reported, indicating that the insertion indexes of each replicate in each sample were highly correlated.

The raw reads from both replicates in each sample were then combined and re-analysed as in the materials and methods. Insertion indexes were calculated and plotted in histograms (Figure 5.2). The bimodal profiles are similar, especially in the right mode. The leftmost bin contains a slightly lower frequency in the V100 profile.

5.2.2 Differential Representation Analysis. After growing the transposon library in the presence of a selective condition, it is expected that some transposon mutants will become

Table 5.1. Dataset metrics for the S4.8 and V100 datasets.

Condition	Reads with matching inline barcode (0 mismatches)	Reads with 1st 25bp tn sequence (3 mismatches allowed)	Reads with 2nd 10bp tn sequence (1 mismatch allowed)	Reads shorter than < 20 bases		Genome wide insertions	CDS insertions	Mapped reads after clipped read filtering
S4.8.1	5171502	4934874 (95.4% raw reads)	4651774 (90% raw reads)	3192 (0.07% raw reads)		396586	340301 (85.8% of total insertions)	4089053 (79.1% raw reads)
S4.8.2	6118243	5609431 (91.7% raw reads)	5268286 (86.1% raw reads)	25192 (0.48% raw reads)		495603	424686 (85.7% of total insertions)	4967477 (81.2% raw reads)
V100.1	5699814	4993723 (87.6% raw reads)	4309993 (75.6% raw reads)	11331 (0.26% raw reads)		420098	360578 (85.8% of total insertions)	3346576 (58.7% raw reads)
V100.2	4285755	4085325 (95.3% raw reads)	3945587 (92.1% raw reads)	14298 (0.36% raw reads)		609122	524628 (86.1% of total insertions)	3721646 (86.8% raw reads)

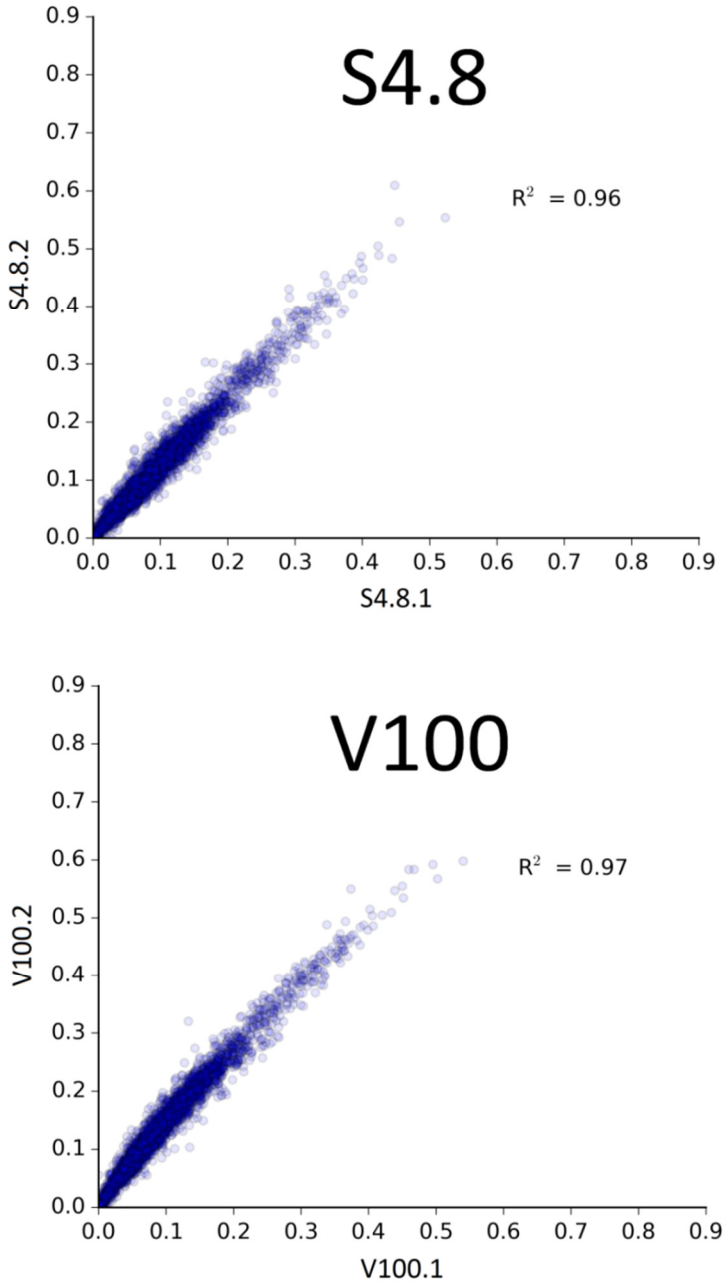

Figure 5.1. Insertion index correlation scatterplots for the biological replicates of the S4.8 and V100 samples. For each sample, the insertion indexes calculated for every W3110 coding sequence for each replicate were plotted against each other, and a coefficient of determination (R^2) was calculated. The max R^2 value is 1, which would indicate a perfect positive correlation.

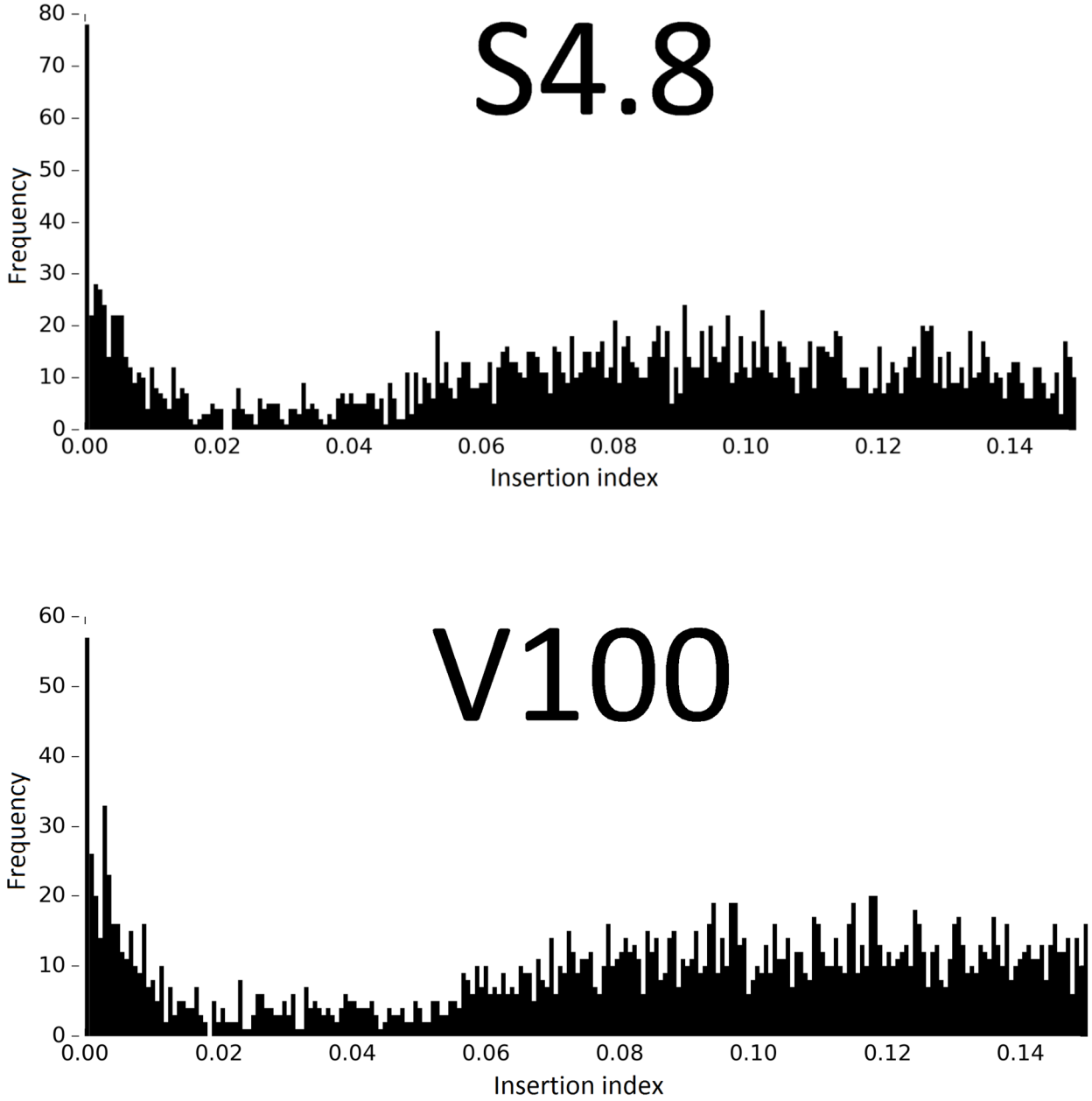

Figure 5.2. Insertion index histograms for the combined S4.8 and V100 datasets. Distinct bimodality can be seen in both plots. Both histograms display clearly bimodal distributions as seen in the previous chapters.

under represented if the mutated genes are required for growth under the selection pressure. The opposite is also expected, that some mutants will become over represented, where loss of the gene allows increased fitness for growth under the selective pressure. Thus, non-essential genes responsible for cell envelope maintenance will become underrepresented upon inactivation, and non-essential genes whose inactivation leads to increased resistance to the vancomycin and/or SDS will become overrepresented. As a preliminary measure, a visual inspection was undertaken for a small number of genes known to be affected by growth under these selective pressures (Fig. 5.3). Thus, inserts in *yejM* were underrepresented after growth in the presence of vancomycin and SDS, whereas inserts in *galU*, *rfaG* and *rfaP* were underrepresented only in the presence of SDS.

Whilst visual inspection confirmed that the underlying experimental approach was sound, statistical rigour was required to be able to properly analyse and compare the datasets. To discern the differentially represented genes from the test and control datasets, the program DESeq2 was used (Love, Huber and Anders, 2014). DESeq2 is more commonly used for the analysis of RNA-seq data. However, the data produced by RNA-seq and transposon sequencing are fundamentally identical, given that it can be reduced to the numbers of reads aligned across coding sequences. Additionally, DESeq2 has previously been used for analysing transposon sequencing data (Christiansen *et al*., 2014). The output of DESeq2 gives a $\log_2$ fold change (L_2FC) value and an adjusted *p* value for each coding sequence. L_2FC values can be either positive, indicating a representational increase in the test condition, or negative, indicating a representational decrease in the test condition.

When discussing RNA-Seq experiments, genes are normally assessed in terms of their differential expression. While transposon sequencing data is fundamentally identical to RNA-Seq data, differential expression is not a technically accurate term to use.

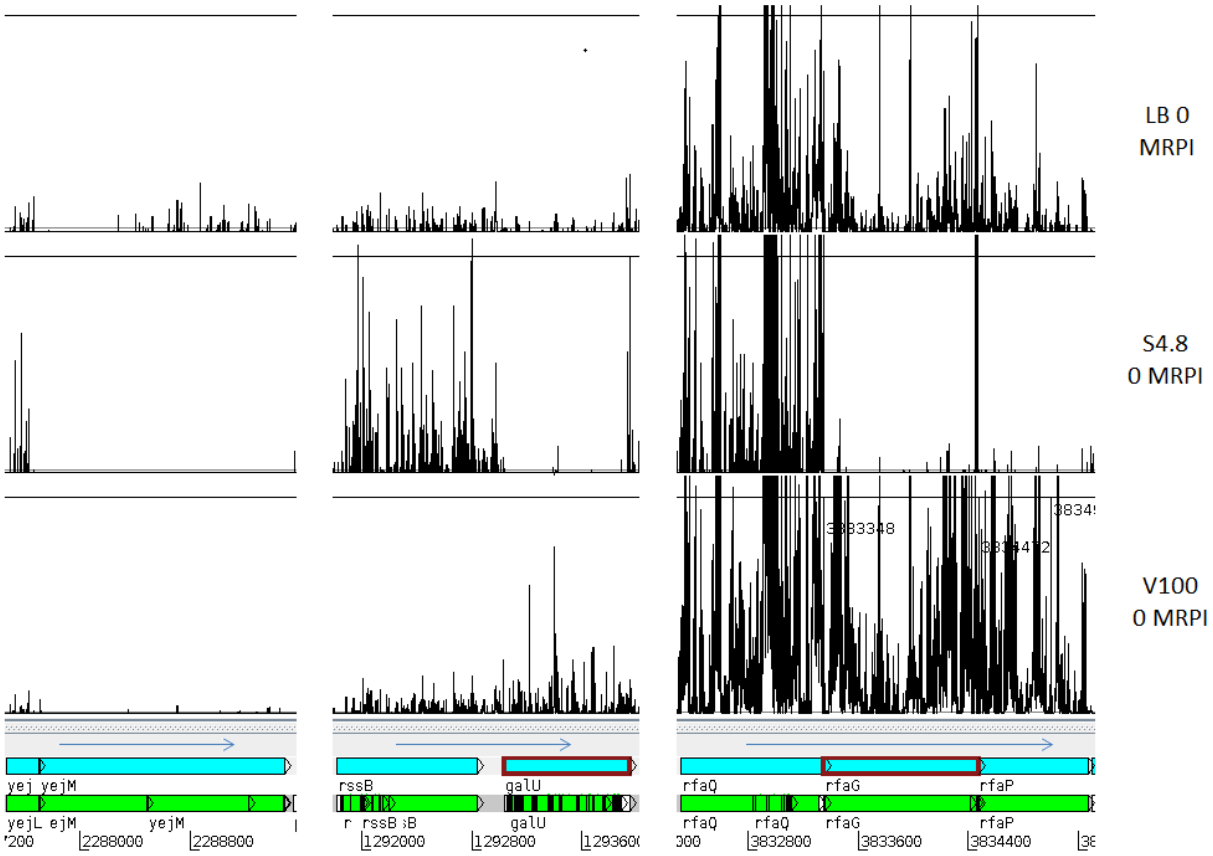

Figure 5.3. Insertion profiles for genes known to be involved with response to vancomycin and SDS. *yejM* mutants are not able to grow in the presence of vancomycin or SDS. *rssB* mutants are advantaged in the presence of SDS, but grow no differently in vancomycin. *galU* mutants are advantaged in the presence of vancomycin, but disadvantaged in SDS. Compared to wild type, *rfaG* and *rfaP* mutants grow better in vancomycin and worse in SDS.

From here, onwards the data will be discussed in terms of the differential representation of insertions in coding sequences with and without the presence of SDS or vancomycin.

Here, DESeq2 was used to determine the coding sequences which were differentially represented between the control and test condition datasets. For each comparison (growth in LB vs. growth in LB supplemented with SDS or vancomycin), two separate analyses were made. First the numbers of unique insertion points were compared with control populations. Second, the numbers of reads aligned to a particular gene were compared with control populations. The resulting lists were then filtered by their L_2FC and adjusted *p* values. Thus, genes with positive or negative L_2FC values < +1 or > -1 were removed from the list, meaning that only genes with at least a 2-fold change (in either direction) as a result of treatment were retained. Any genes with a *p* value of greater than 0.05 were removed, meaning that there was a false discovery rate of less than or equal to 5% in the remaining genes. The two lists for each comparison were then merged into one, and subsequently split into positive and negative $\log_2$ fold changes.

5.2.3 Genes differentially represented after growth in the presence of SDS. After filtering the gene lists, 45 genes were differentially represented after growth in LB with 4.8% SDS (Table 5.2). Strikingly only one of these genes (*rssB*) had a positive L_2FC value, and so was overrepresented in the test dataset. The *rssB* gene encodes for a response regulator which governs the σ^S subunit (also known as RpoS) of RNA polymerase (Muffler *et al*., 1996: Pratt and Silhavy, 1996). Through direct and specific binding to RpoS, RssB facilitates the proteolytic degradation of RpoS by presentation to the proteolytic ClpXP machinery (Becker, Klauck and Hengge-Aronis., 1999: Zhou *et al*., 2001). Additionally, RssB negatively regulates RpoS levels by modulating polyadenylation and mRNA stability of the

Table 5.2. Differentially represented genes after treatment with SDS.

Gene	Description[1]	+/- represented
pal	Lipoprotein associated with peptidoglycan; involved in maintaining cell membrane integrity	-
tolA	Tolerance to group A colicins, single-stranded filamentous DNA phage; required for OM integrity; membrane protein; bacteriocin tolerant	-
plpa/yraP	OM lipoprotein, function unknown, mutant is SDS-sensitive	-
surA	Periplasmic OM porin chaperone, has PPIase activity; required for stationary-phase survival	-
yfgL	Beta-propeller lipoprotein in OM biogenesis BamABCDE complex; WD40/PQQ repeats; mutant has pleiotropic envelope defects; required for swarming phenotype	-
yrbE	Probable phospholipid ABC transporter permease; MlaFEDB phospholipid ABC transporter; maintains OM lipid asymmetry	-
atpA	ATP synthase subunit alpha, membrane-bound, F1 sector	-
atpD	ATP synthase subunit beta, membrane-bound, F1 sector	-
atpG	ATP synthase subunit gamma, membrane-bound, F1 sector	-
amiA	N-acetylmuramyl-L-alanine amidase, periplasmic; role in septal cleavage during cell division; activated by EnvC	-
amiC	N-acetylmuramyl-L-alanine amidase, periplasmic; recruited to the septal ring by FtsN during cell division; overexpression causes lysis; activated by NlpD	-
envC/yibP	Activator of AmiB,C murein hydrolases, septal ring factor	-
nlpD	Activator of AmiC murein hydrolase activity, lipoprotein	-
crl	Pseudogene, sigma factor-binding protein; stimulates RNAP holoenzyme formation; stimulates RpoS activity during stationary phase; mutants display rpoS mutant phenocopy; mutant does not have reduced amount of RpoS protein	-
dksA	RNAP-binding protein modulating ppGpp and iNTP regulation; reduces open complex half-life on rRNA promoters; removes transcriptional roadblocks to replication	-
fadR	Repressor/activator for fatty acid metabolism regulon; fatty acid-responsive transcription factor; fabAB, iclR activator (regulates aceBAK, glyoxylate shunt); fad repressor; homodimeric	-
fepD	Ferrienterobactin ABC transporter permease	-
greA	Transcript cleavage factor	-
hfq	Global regulator of sRNA function; host factor for RNA phage Q beta replication; HF-I; DNA- and RNA-binding protein; RNA chaperone; binds ATP and RNAP	-
nhaA	Na+/H+ antiporter 1, strongly pH-dependent; helps regulate intracellular pH and extrude lithium; nhaA_P1 activated by NhaR, repressed by H-NS and stimulated by Na(+)	-
oxyR	Oxidative and nitrosative stress transcriptional regulator	-

qseC	Quorum sensing two-component sensor kinase; cognate to QseB response regulator; regulates flagella synthesis and motilty by activating transcription of flhDC; responds to AI-3 and	-
rbsR	Regulatory gene for rbs operon	-
yhdP	DUF3971-AsmA2 domains protein	-
yraO	DnaA-binding protein; involved in the timing of the initiation of DNA replication; dnaA(Cs) suppressor; homodimer	-
dacA	D-alanine D-alanine carboxypeptidase PBP5, cell morphology; penicillin-binding protein 5; beta-lactamase activity	-
galU	Glucose-1-P uridylyltransferase; also called UDP-glucose pyrophosphorylase	-
mrcA	Murein polymerase, PBP1A; bifunctional murein transglycosylase and transpeptidase; penicillin-binding protein 1A; dimeric	-
pgm	Phosphoglucomutase	-
rfaD	ADP-L-glycero-D-manno-heptose-6-epimerase; heat-inducible, LPS; allows high-temperature growth	-
rfaE	Heptose 7-P kinase/heptose 1-P adenyltransferase; LPS core prescursor synthesis: bifunctional enzyme involved in both D-glycero-D-manno-heptose-1-phosphate and ADP-D-glycero-D-manno-heptose synthesis	-
rfaF	ADP-heptose:LPS heptosyltransferase II	-
rfaG	UDP-glucose:(heptosyl)LPS alpha-1,3-glucosyltransferase; LPS core biosynthesis protein; glucosyltransferase I	-
rfaH	Transcription antitermination factor, LPS biosynthesis genes; negatively controls expression and surface presentation of Ag43 (Flu), reducing adhesion and biofilm; also regulates F-factor sex pilus and hemolysin genes	-
rfaP	Lipopolysaccharide kinase; LPS core biosynthesis; phosphorylation of core	-
rfe	UDP-GlcNAc:undecaprenylphosphate GlcNAc-1-P transferase; ECA and O-antigen synthesis, tunicamycin sensitivity	-
yraM	OM lipoprotein stimulator of MrcA transpeptidase	-
acrA	AcrAB-TolC multidrug efflux pump; additionally dye, detergent, solvent resistance; membrane-fusion lipoprotein	-
acrB	AcrAB-TolC multidrug efflux pump; additionally dye, detergent and solvent resistance; RND-type transporter	-
tolC	Outer membrane factor (OMF) of tripartite efflux pumps; channel-tunnel spanning the outer membrane and periplasm; trimeric; ColE1 tolerance	-
rpoS	RNA polymerase subunit, stress and stationary phase sigma S; sigma 38	-
tatA	Protein translocase, Sec-independent; mediates export of folded and ligand-bound proteins	-
tatC	Protein translocase, Sec-independent; mediates export of folded and ligand-bound proteins	-
yejM	Essential inner membrane DUF3413 domain protein; lipid A defect; membrane permeability defect	-
rssB	Response regulator binding RpoS to initiate proteolysis by ClpXP; required for the PcnB-degradosome interaction during stationary phase; major cognate sensor kinase is ArcB	+

[1]Gene descriptions obtained from Ecogene (Zhou and Rudd, 2012)

Table 5.3. Differentially represented genes after treatment with vancomycin.

Gene	Description	+/- represented
asmA	Suppressor of OmpF assembly mutants; inner membrane-anchored periplasmic protein; putative outer membrane protein assembly factor; required for swarming phenotype	-
aspA	L-aspartate ammonia-lyase; L-aspartase	-
astE	Succinylglutamate desuccinylase, arginine catabolism	-
envC/yibP	Activator of AmiB,C murein hydrolases, septal ring factor	-
greA	Transcript cleavage factor	-
oxyR	Oxidative and nitrosative stress transcriptional regulator	-
rfe	UDP-GlcNAc:undecaprenylphosphate GlcNAc-1-P transferase; ECA and O-antigen synthesis, tunicamycin sensitivity	-
smpA	Lipoprotein stabilizer of BamABCDE OM biogenesis complex	-
sucA	2-oxoglutarate dehydrogenase, E1 component; yields succinyl-CoA and CO(2); also known as alpha-ketoglutarate dehydrogenase	-
tatC	Protein translocase, Sec-independent; mediates export of folded and ligand-bound proteins	-
yejM	Essential inner membrane DUF3413 domain protein; lipid A defect; membrane permeability defect	-
yfgC	Periplasmic metalloprotease and chaperone; outer membrane protein maintenance and assembly	-
yfgL	Beta-propeller lipoprotein in OM biogenesis BamABCDE complex; WD40/PQQ repeats; mutant has pleiotropic envelope defects; required for swarming phenotype	-
yhdP	DUF3971-AsmA2 domains protein	-
yhjK	Cyclic-di-GMP phosphodiesterase associated with cellulose production; dual domain protein; defective cyclase domain	-
envZ	Osmosensor histidine protein kinase/phosphatase; regulates production of outer membrane proteins; dimeric	+
ompR	Response regulator for osmoregulation; regulates production of outer membrane proteins	+
nlpI	Lipoprotein involved in osmotic sensitivity and filamentation	+
ompC	Outer membrane porin C	+
rseA	Anti-RpoE sigma factor, spans inner membrane	+
mdoH	OPG biosynthetic UDP-glucose beta-1,2 glycosyltransferase; transmembrane, ACP-dependent; nutrient-dependent cell size regulator; FtsZ assembly antagonist	+
prc	Periplasmic carboxy-terminal protease with specificity for non-polar C-termini	+
yhcB	DUF1043 family inner membrane-anchored protein; biofilm-related	+
wzzE	Enterobacterial common antigen ECA chain length determination; also involved in cyclic enterobacterial common antigen ECA(CYC) synthesis	+

yhiO	Universal stress protein B, confers ethanol resistance in stationary phase; sigma S-regulated gene divergent from uspA	+
ycbC	Envelope biogenesis factor; DUF218 superfamily protein	+
dacA	D-alanine D-alanine carboxypeptidase PBP5, cell morphology; penicillin-binding protein 5; beta-lactamase activity	+
ldcA	Murein tetrapeptide carboxypeptidase; LD-carboxypeptidase A; cytoplasmic protease that cleaves the terminal D-alanine from cytoplasmic muropeptides	+
fbp	Fructose-1,6-bisphosphatase; allosteric: inhibited by AMP	+
galU	Glucose-1-P uridylyltransferase; also called UDP-glucose pyrophosphorylase	+
pgm	Phosphoglucomutase	+
rfaG	UDP-glucose:(heptosyl)LPS alpha-1,3-glucosyltransferase; LPS core biosynthesis protein; glucosyltransferase I	+
rfaH	Transcription antitermination factor, LPS biosynthesis genes; negatively controls expression and surface presentation of Ag43 (Flu), reducing adhesion and biofilm; also regulates F-factor sex pilus and hemolysin genes	+
rfaP	Lipopolysaccharide kinase; LPS core biosynthesis; phosphorylation of core	+
rfaQ	Glycosyltransferase needed for heptose region of LPS core	+
clpP	Proteolytic subunit of ClpXP and ClpAP ATP-dependent proteases; protease Ti	+
chpR	MazE antitoxin, toxin is MazF	+
ihfB	Integration Host Factor (IHF), beta subunit; host infection, mutant phage lambda; site-specific recombination; sequence-specific DNA-binding transcriptional activator	+
mtlR	Mannitol operon repressor	+
ytfK	DUF1107 family protein	+
hfq	Global regulator of sRNA function; host factor for RNA phage Q beta replication; HF-I; DNA- and RNA-binding protein; RNA chaperone; binds ATP and RNAP	+

[1]Gene descriptions obtained from Ecogene (Zhou and Rudd, 2012)

RpoS transcript (Carabetta *et al*., 2009). In *E. coli*, RpoS is the central regulator of the general stress response (Battesti, Madjalani and Gottesman, 2011). Levels of RpoS are very low during exponential growth and during optimal growth conditions, due to RssB and many other regulatory factors. On approach to stationary phase, or in response to adverse environmental conditions, RpoS levels increase and allow the cell to alter the expression of genes and pathways linked to resistance towards a number of stresses. The transposon sequencing data indicates that inactivation of *rssB* leads to increased fitness (Fig. 5.4). This is logical according to what is known: the inactivation of *rssB* would lead to constitutively increased levels of RpoS in the cell, meaning that the general stress response would be mounted throughout every stage of growth. Indeed, a 10-fold increase in RpoS was observed during the exponential growth of an *rssB* mutant (Muffler *et al*, 1996). In the presence of SDS, the activation of the general stress response is likely to lead to increased fitness, and this would serve to explain the increased representation observed here. In keeping with these findings, previous work has shown that *rssB* mutants are more resistant to osmotic stress, oxidative stress and heat stress (Fontaine *et al*., 2008). Given the mechanics of how RssB is involved with RpoS regulation, it would be expected to see *rpoS* in the list of negatively represented genes. This was exactly the case. The disruption of *rpoS* meant that the general stress response could not be mounted, and so these cells were more susceptible to SDS.

All three components of the RND family AcrAB-TolC multidrug efflux system were present in the negatively represented gene list. This tripartite protein complex forms a pump that spans the whole of the cell envelope, forming a channel between the cytoplasm of a cell and its external environment (Symmons *et al*., 2009). This machinery is recognised as being the main multidrug resistance mechanism within *E. coli*, and a wide variety of

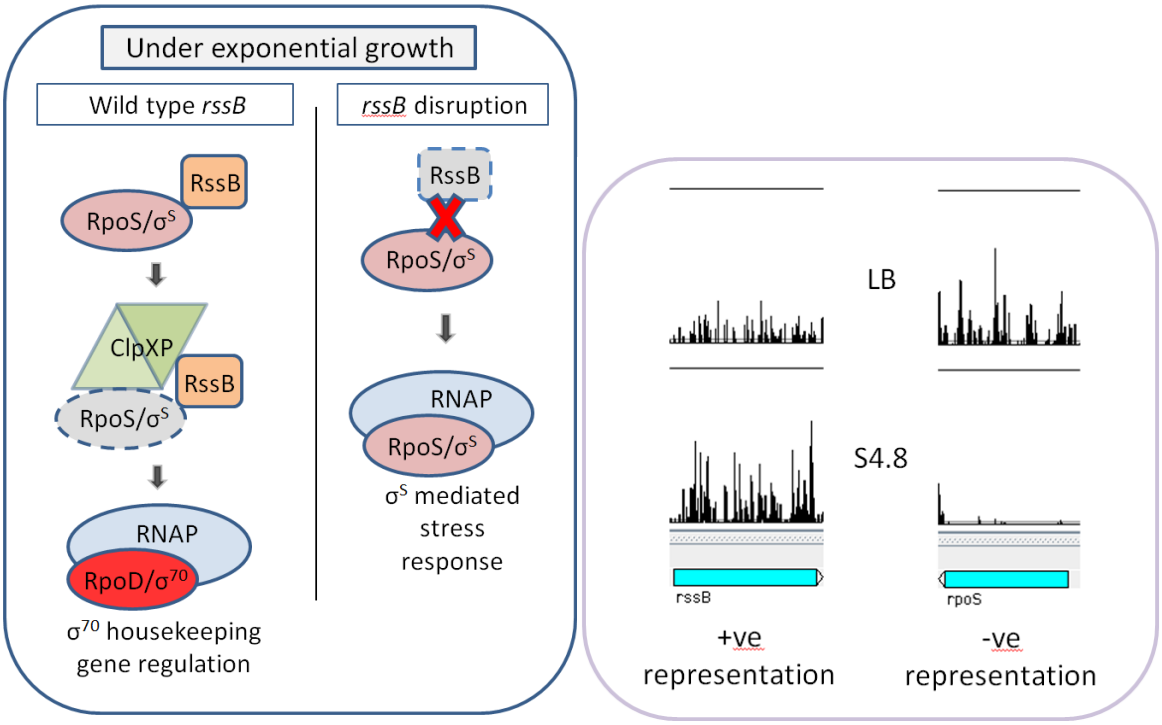

Figure 5.4. Schematic showing the RssB regulation of RpoS. (Upper panel) Wild type RssB binds to RpoS and facilitates its proteolysis by the ClpXP machinery. This leads to low levels of RpoS, meaning RNA polymerase partners mainly with the housekeeping σ^{70}. Upon disruption of RssB, RpoS is not degraded by ClpXP leading to its accumulation. It then binds to RNA polymerase and begins the regulation of the RpoS mediated stress response regulon. This graphic doesn't take into account the post transcriptional regulation of *rpoS* expression by RssB. (Lower panel) After growth in 4.8% SDS, there is increased insertional representation throughout *rssB*. The opposite is true for *rpoS*, which shows a decreased representation after growth in SDS.

substrates are recognised and pumped out by this system, including antibiotics, dyes and detergents (Elkins and Nikaido, 2002: Ma *et al*., 1995). When any of the three components are absent the complex cannot be formed and the cell becomes hypersusceptible to toxic compounds. SDS is a known substrate of the system (Nikaido, 1998). When the activity of the efflux pump is compromised SDS cannot be pumped out of the cytosol. This leads to accumulation of SDS within the cell, subsequent adverse effects on the cell and results in cell lysis. As such, the finding that *acrA, acrB* and *tolC* are negatively represented in the library exposed to SDS is consistent with previous observations.

Examination of the data revealed that TolA and Pal, two components of the Tol-Pal cell envelope complex, are also negatively represented. The Tol-Pal system consists of five proteins (TolBRAQ and Pal) which interact to form a complex that spans the envelope (Gerding *et al*., 2007). While the precise function of this system is unknown, it has been implicated with maintenance of the outer membrane and the control of envelope constriction during division. Mutations in the Tol-Pal genes lead to a variety of phenotypic effects including increased outer membrane vesiculation (Bernadac *et al*., 1998), sensitivity to antibacterial agents (Lazzaroni *et al*., 1999) and the leakage of periplasmic proteins (Cascales *et al*., 2002). Disruption of *tolA* and *pal* in the transposon library is likely to disrupt the formation of the Tol-Pal complex, in turn weakening the envelope, allowing SDS to move into the cytoplasm and weakening the efforts of any functional efflux pumps. Interestingly, the *tolBRQ* genes that make up the remainder of the system all had negative L_2FC values. However, they were not included in the negatively represented gene list as the L_2FC values did not pass the threshold for inclusion.

Two peptidoglycan hydrolases, *amiA* and *amiC*, are negatively represented after growth in SDS. These genes encode N-acetylmuramyl-L-alanine amidases (Vollmer *et al*.,

2008). These amidases, along with a third (*amiB*), cleave crosslinks in the periplasmic peptidoglycan layer. Individual deletion of the three amidases resulted in long chains of unseparated cells, and in a strain with all three amidases deleted this effect was multiplied (Heidrich *et al*., 2001). Additionally, cells containing these deletions had uncleaved septa. The presence of uncleaved septa and the chaining of cells indicates that in cells lacking amidase activity, while the formation of the septum is unimpeded, it is specifically the cleavage of the septum that is affected. AmiA and AmiC were both found to be localised to the periplasm, after being trafficked through the Tat system (van Heijenoort, 2011). Furthermore, AmiC was found to be specifically localised to the septal ring during cell division in contrast to AmiA, which was diffused among the periplasm. Another finding was that amidase lacking chain forming mutants had impaired outer membrane integrity (Heidrich *et al*., 2002). Cells containing amidase deletions became sensitive to several agents that normally do not affect growth, including vancomycin and the detergent Triton X-100. Interestingly, although Heidrich *et al*. (2002) find that *amiB* is involved with septal cleavage and that *amiABC* have overlapping roles, in this work *amiB* is not negatively represented. In contrast to *amiA* and *amiC*, *amiB* had a positive L^2FC value, although below the threshold set. This might suggest that *amiB* is functionally divergent from the other amidases.

AmiA and AmiC are transported into the periplasm by the Tat system (Ize *et al*., 2003). All three components of the Tat system, *tatA*B*C*, were negatively represented after growth in SDS. However, only *tatA* and *tatC* were retained past the L_2FC and *p* value thresholds. Both of these components are integral to the inner membrane, forming the translocation machinery (Robinson *et al*., 2011). Tat mutants have been previously shown to be sensitive to SDS (Ize *et al*., 2003). In *E. coli*, there is experimental evidence for 27 Tat system substrates, of which AmiA and AmiC are two (Palmer and Berks, 2012). Interestingly,

in the remaining 25 proteins there were no significant positive or negative L_2 fold changes. This suggests that the inclusion of *tatA* and *tatC* in the negatively represented list is solely due to their involvement with the transport of AmiA and AmiC.

Two more genes that localise to the septal ring apparatus were also negatively represented; *envC* and *nlpD*. These two proteins were previously found to localise to the division site and to be required for the separation of daughter cells after division (Uehara, Dinh and Bernhardt, 2009). Furthermore, EnvC directly activates AmiA and AmiB, and that NlpD directly activates AmiC (Uehara *et al*., 2010). EnvC and NlpD did not appear to have catalytic activity when incubated with high concentrations of peptidoglycan *in vitro*. This is in contrast to other work in which EnvC was shown to have peptidoglycan hydrolytic activity (Bernhardt and de Boer, 2004). Presumably the disruption of *envC* and *nlpD* leads to SDS sensitivity in the same manner as observed for disruptions of *amiA* and *amiC*.

Continuing with the theme of peptidoglycan structure, the negatively represented *dacA* (also known as penicillin binding protein 5) encodes a carboxypeptidase involved with the final stages of peptidoglycan biogenesis (Ghosh, Chowdhury and Nelson, 2008). In other work, the deletion of penicillin binding proteins (PBPs) including *dacA* was found to lead to morphological aberration (Nelson and Young, 2000). Additionally, deletion of *dacA* was found to increase susceptibility to beta lactam antibiotics (Sarkar, Chowdhury and Ghosh, 2010). Another PBP, *mrcA* (PBP1A), was negatively represented. *mrcA* encodes a bifunctional enzyme that has transglycosylation and transpeptidation activities (Born, Breukink and Vollmer, 2005). Additionally, the activator of *mrcA*, *lpoA*, is also negatively represented. LpoA is an outer membrane lipoprotein that directly binds to MrcA and stimulates its transpeptidase activity (Typas *et al*., 2010). Without the activities of LpoA/MrcA, the

peptidoglycan layer will not fully mature and in turn is likely to be weaker, which might explain the apparent sensitivity to SDS.

E. coli are naturally resistant to a multitude of hydrophobic agents, and this resistance is mediated by lipopolysaccharide (LPS) present in the outer leaflet of the outer membrane (Nikaido and Vaara, 1985). Eight genes involved with LPS biosynthesis were negatively represented (Fig. 5.5); *rfaDEFGHP, galU and pgm. rfaE* encodes a bifunctional enzyme involved with two steps of the synthesis of the LPS core precursor ADP-d-*glycero*-d-*manno*-heptose (Valvano *et al*., 2000), and *rfaD* encodes an epimerase that catalyses the last step in the pathway (Kneidinger *et al*., 2002). Two other genes encode proteins that catalyse steps in the synthesis of this precursor, *lpcA* and *gmhB*, also had negative L_2FC values, although they did not meet the required threshold. *lpcA* encodes a D-sedoheptulose 7-phosphate isomerase, and *gmhB* encodes a heptose bisphosphatase phosphatase. *rfaFGP* are all involved in the synthesis of the lipid A core (Yethon and Whitfield, 2000: Gronow, Brabetz and Brade, 2000: Roncero and Casadaban, 1992). The remaining *rfa* gene, *rfaH*, is a transcriptional anti terminator that is required for the expression of multiple LPS biosynthetic genes, including *rfaGP* (Bailey, Hughes and Koronakis, 1996; Pradel and Schnaitman, 1991). The formation of lipid A core also relies on the incorporation of UDP-α-d-glucose at 3 steps in the biosynthetic pathway. Two enzymes in the UDP-α-d-glucose biosynthetic pathway, *pgm* and *galU*, are negatively represented. *pgm* mutants were found to be sensitive to SDS by Phan *et al*. (2013). Taken together, mutations in these genes lead to a lack of mature LPS, which in turn leads to susceptibility to hydrophobic agents (Nikaido and Vaara, 1985).

In addition to LPS, the enterobacterial common antigen is also present on the cell surface (Kajimura, Rahman and Rick, 2005). The negatively represented gene *rfe* is an

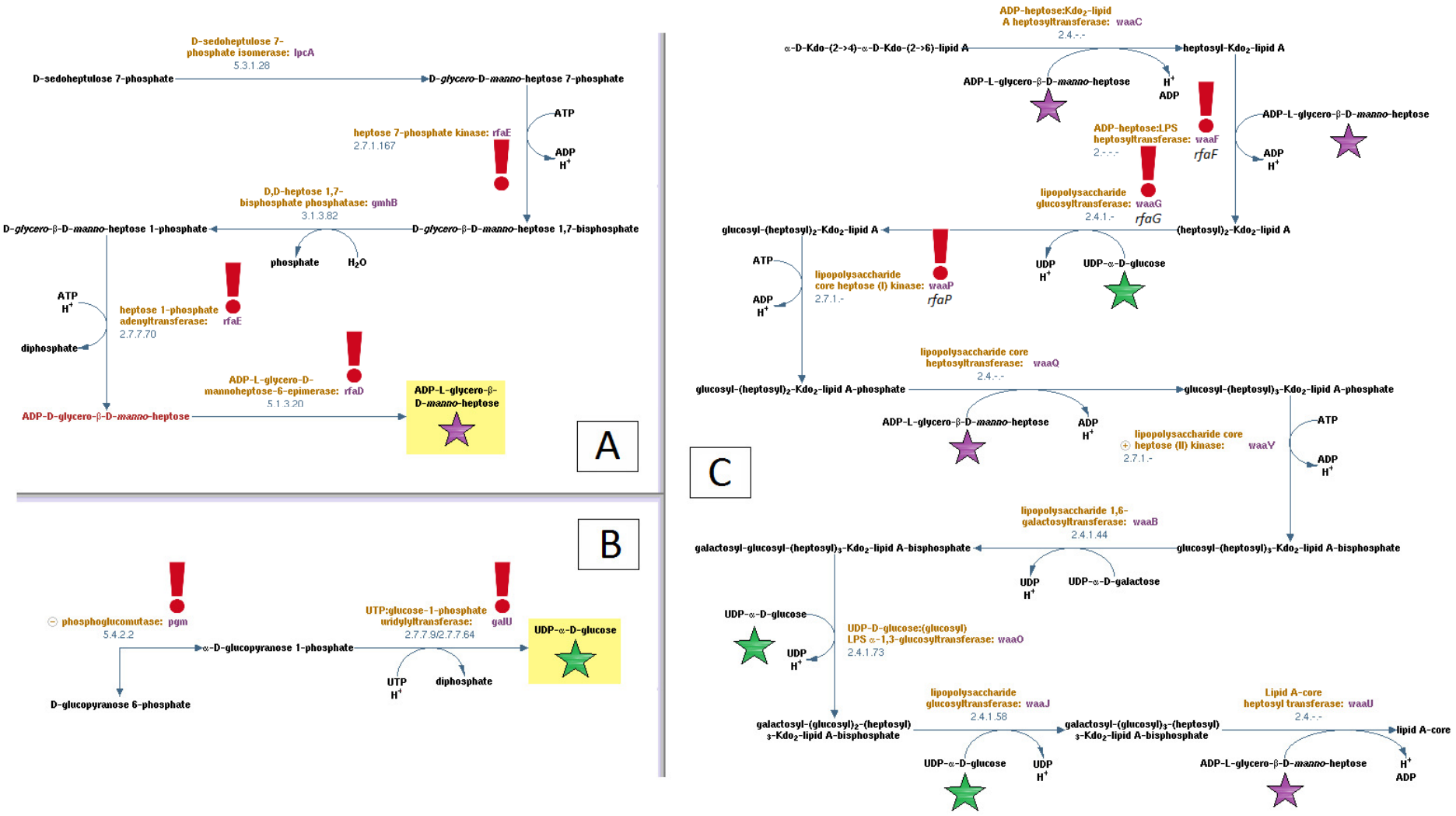
D-sedoheptulose 7-phosphate isomerase: lpcA
5.3.1.28
D-sedoheptulose 7-phosphate
D-glycero-D-manno-heptose 7-phosphate
ATP
ADP
H+
heptose 7-phosphate kinase: rfaE
2.7.1.167
D,D-heptose 1,7-bisphosphate phosphatase: gmhB
3.1.3.82
D-glycero-β-D-manno-heptose 1-phosphate
D-glycero-β-D-manno-heptose 1,7-bisphosphate
phosphate
H2O
ATP
H+
diphosphate
heptose 1-phosphate adenyltransferase: rfaE
2.7.7.70
ADP-L-glycero-D-mannoheptose-6-epimerase: rfaD
5.1.3.20
ADP-D-glycero-β-D-manno-heptose
ADP-L-glycero-β-D-manno-heptose
A
B
phosphoglucomutase: pgm
5.4.2.2
α-D-glucopyranose 1-phosphate
D-glucopyranose 6-phosphate
UTP:glucose-1-phosphate uridylyltransferase: galU
2.7.7.9/2.7.7.64
UTP
H+
diphosphate
UDP-α-D-glucose
C
ADP-heptose:Kdo2-lipid A heptosyltransferase: waaC
2.4.-.-
α-D-Kdo-(2->4)-α-D-Kdo-(2->6)-lipid A
heptosyl-Kdo2-lipid A
ADP-L-glycero-β-D-manno-heptose
H+
ADP
ADP-heptose:LPS heptosyltransferase: waaF
2.-.-.-
rfaF
ADP-L-glycero-β-D-manno-heptose
ADP
H+
lipopolysaccharide glucosyltransferase: waaG
2.4.1.-
rfaG
glucosyl-(heptosyl)2-Kdo2-lipid A
(heptosyl)2-Kdo2-lipid A
ATP
H+
ADP
H+
lipopolysaccharide core heptose (I) kinase: waaP
2.7.1.-
rfaP
UDP
H+
UDP-α-D-glucose
lipopolysaccharide core heptosyltransferase: waaQ
2.4.-.-
glucosyl-(heptosyl)2-Kdo2-lipid A-phosphate
glucosyl-(heptosyl)3-Kdo2-lipid A-phosphate
ADP-L-glycero-β-D-manno-heptose
ADP
H+
lipopolysaccharide core heptose (II) kinase: waaY
2.7.1.-
ATP
ADP
H+
lipopolysaccharide 1,6-galactosyltransferase: waaB
2.4.1.44
galactosyl-glucosyl-(heptosyl)3-Kdo2-lipid A-bisphosphate
glucosyl-(heptosyl)3-Kdo2-lipid A-bisphosphate
UDP
H+
UDP-α-D-galactose
UDP-α-D-glucose
UDP
H+
UDP-D-glucose:(glucosyl) LPS α-1,3-glucosyltransferase: waaO
2.4.1.73
lipopolysaccharide glucosyltransferase: waaJ
2.4.1.58
galactosyl-(glucosyl)2-(heptosyl)3-Kdo2-lipid A-bisphosphate
galactosyl-(glucosyl)3-(heptosyl)3-Kdo2-lipid A-bisphosphate
UDP-α-D-glucose
UDP
H+
Lipid A-core heptosyl transferase: waaU
2.4.-.-
lipid A-core
ADP-L-glycero-β-D-manno-heptose
H+
ADP

Figure 5.5. Negatively represented genes in the LPS biosynthetic pathway after growth in SDS. Pathway information obtained from Ecocyc (Karp *et al.*, 2013). Panels A and B show the synthetic pathway information for two requisite precursors to lipid A core biosynthesis. Each precursor is labelled, with green stars representing UDP-α-D-glucose and purple stars representing ADP-L-glycero-beta-D-manno-heptose. In panel C, the stars indicate where the precursors are used in the synthesis of lipid A. In all three panels, red exclamation marks denote genes that are significantly negatively represented past the thresholds set after growth in SDS. Gene names in italics are the reference names used in this chapter. For every gene in this figure without an exclamation mark, a negative L_2FC value was reported, although the genes did not pass the thresholds set for fold change and significance.

enzyme that catalyses the transfer of N-acetylglucosamine-1-phosphate to undecaprenyl phosphate, and the resulting precursor is then used in the synthesis of the enterobacterial common antigen and LPS O antigens (Rush, Dick and Waechter, 1997). There is little literature available relating ECA with resistance to external agents. However, Ramos-Morales *et al*. (2003) found that mutation of *wecD* and *wecA* in *Salmonella enterica* serovar Typhimurium, two genes involved with the synthesis of ECA, resulted in increased sensitivity to deoxycholate suggesting a compromised cell envelope.

One component of the β barrel assembly machinery (BAM) complex, *bamB*/*yfgL*, was negatively represented. The BAM complex is a collection of proteins that function to fold and insert outer membrane proteins (OMPs) into the outer membrane (Knowles *et al*., 2009). These OMPs play key roles in multidrug resistance, virulence and in the maintenance of the envelope. Specifically, work published by Ruiz *et al*. (2005) and Wu *et al*. (2005) implicated *yfgL/bamB* with the maintenance of the outer membrane. Confusingly, the other components of the BAM complex (*smpA*/*bamE* and *nlpB*/*bamC*) do not have significantly variable L_2FC values after growth in SDS. Regarding *bamC*, this is in contrast to findings reported by Sklar *et al*. (2007b). Regarding *bamE*, this is in contrast to Knowles *et al*. (2011).

The periplasmic chaperone *surA* was also found to be negatively represented. Along with *skp* and *degP*, *surA* interacts with proteins translocated across the inner membrane by the Sec apparatus and assists in their folding and delivery to the BAM complex (Sklar *et al*., 2007a). *surA* deletion mutants were previously found to be sensitive to a variety of detergents, antibiotics and dyes (Justice *et al*., 2005). In contrast to *surA*, the were no significant differences in the number or frequency of transposon insertions in the *skp* and *degP* genes when data from the SDS treated library was compared with data derived from the control experiment.

Interestingly, one of the negatively represented genes, *yejM*, was discussed in chapter 4. *yejM* encodes an essential cardiolipin binding protein (De Lay and Cronan, 2008; Dalebroux *et al*., 2015). However, this gene has a nonessential 3' region. The presence of a non-essential region in *yejM* explains how it can be negatively represented after growth in SDS; there are fewer insertions in the non-essential region after growth in SDS. Given that the non-essential region has been implicated with resistance to SDS, it may be the case that the essential and non-essential regions of YejM have different functions.

A further three negatively represented genes, *atpADG*, encode components of ATP synthase F_1 complex. The likely explanation of this finding is that ATP is required for efflux of SDS, and so disruption of these genes leads to less energy in the form of ATP, and in turn increased susceptibility to the molecule. The other components of ATP synthase, *atpBCEFH*, were also negatively represented, but not past the required significance thresholds.

For the remaining 12 negatively represented genes, no obvious or immediate explanation can be given for their involvement with resistance to SDS. As such, these genes might represent as of yet unknown genetic links to the maintenance of the outer membrane.

5.2.4 Genes differentially represented after growth in the presence of vancomycin. Following the filtering criteria outlined earlier, 41 genes were differentially represented after growth in vancomycin (Table 5.3). Of these, 15 were negatively represented. in contrast to the majority of the genes negatively represented after growth in SDS. Eight of the 15 were also found to be negatively represented after growth in SDS, including *yibP/envC*, *greA*, *oxyR*, *rfe*, *tatC*, *yejM*, *yfgL/bamB* and *yhdP*. Earlier, it was concluded that the negative representation of *tatC* was likely due to its involvement with the transport of *amiA* and *amiC*. After growth in vancomycin *amiA, amiC* and *tatB* are negatively represented, but fall

short of the *p* value and L_2FC filtering thresholds. The representation of other substrates for the Tat secretion system does not different significantly after growth in the presence of vancomycin. This, in addition to the negative representation of *envC*, suggests the previous conclusion is likely to apply here. Supporting evidence in the literature was found and discussed for *rfe*, *bamB* and *yejM*, but no obvious explanation could be provided for the identification of *greA*, *oxyR* and *yhdP* in these screens. Supporting literature can be found for some of the remaining 7 genes of the 15 negatively represented genes. *asmA*, which localises to the inner membrane, has been loosely implicated with envelope-associated function (Deng and Misra, 1996). *bamE*, which forms part of the beta barrel assembly machinery (BAM) complex, is also located in the outer membrane (Sklar *et al*., 2007b). Upon deletion of *bamE*, cells show slight defects in membrane permeability and susceptibility to antibiotics including vancomycin (Rigel *et al*., 2012; Browning *et al*., 2013). Additionally, *bamA* is more susceptible to protease treatment in cells with no *bamE*, indicating a more permeable membrane. The coding sequence for the periplasmic protein BepA is also negatively represented. Deletion of this coding sequence has been implicated with increased susceptibility to multiple antibiotics including vancomycin (Tamae *et al*., 2008; Girgis *et al*., 2009). For the remaining negatively represented genes *aspA*, *astE*, *sucA* and *yhjK*, no obvious link to the outer membrane can be found.

In addition to the 15 genes negatively represented, 26 genes were positively represented, that is to say that the insertion frequency of these genes was increased after growth in vancomycin. Some of these 26 genes, *pgm, galU* and *rfaGHP*, had previously been shown to be negatively represented after growth in SDS. In addition to these genes, *rfaQ* was also positively represented These genes are all involved with the formation of LPS present in the outer membrane. Their positive representation after growth in vancomycin is

completely at odds with their negative representation after growth in SDS, when considering that SDS and vancomycin are both indicators of cell envelope defects. However, this finding may be explained by charge. SDS is negatively charged, and vancomycin contains both negatively and positively charged residues. The disruption of *pgm, galU* and *rfaGHPQ* would ultimately lead to less phosphorylation of the lipid A core, in turning making it more positively charged. In the presence of SDS, such a charge change would act to facilitate the entry of the negatively charged SDS, meaning disruption of these genes would increase susceptibility to SDS. In the presence of vancomycin, the charge change might act to repel the entry of the zwitterionic vancomycin, leading to greater fitness in relation to other cells in the culture without these disruptions. Even though this is conjecture, the fact that this grouping of genes with a related function was shown to be positively represented is a strong indicator of biological significance.

A further 4 genes from the list of 26 are linked to peptidoglycan synthesis, including *dacA*, *fbp, ldcA* and *elyC*. *dacA* was identified as a negatively represented gene after growth in SDS, and has been discussed previously. As well as opposing the negative representation of *dacA* in our SDS dataset, the positive representation of *dacA* in this dataset directly conflicts with previous literature. Zeevi *et al*. (2013) reported that, in *Listeria monocytogenes*, the upregulation of *dacA* reduced susceptibility to vancomycin. Additionally, Turner *et al*. (2013) observed the increased labelling of *ΔdacA E. coli* with fluorescent vancomycin, due to an increased number of D-ala motifs caused by the lack of *dacA*. This would suggest that the inactivation of *dacA* would make cells more susceptible to vancomycin. As is the case with *dacA*, the positive representation of *fbp* after growth in vancomycin conflicts with experimental evidence. Saito *et al*. (2014), working with *Staphylococcus aureus*, reported that vancomycin resistance was in line with the

upregulation of *fbp*. *ldcA* encodes a carboxypeptidase which is involved with the recycling of peptidoglycan (Templin, Ursinus and Holtje, 1999). *elyC* is an inner membrane associated protein that has been linked to peptidoglycan synthesis and maintenance of the cell envelope (Paradis-Bleau *et al*., 2014). Given that opposing evidence can be found for two of the positively represented genes linked to peptidoglycan, further work is necessary to confirm whether or not these findings are erroneous. One potential explanation may be that, by disrupting genes involved with peptidoglycan synthesis, the amount of peptidoglycan available for vancomycin to bind to is reduced.

Several genes relating to osmolarity are positively represented: *envZ*, *ompR*, *nlpI*, *prc*, *ompC*, *rseA* and *opgH*. EnvZ and OmpR form a two component system that responds to changes in osmolarity (Cai and Inouye, 2002; Yamamoto *et al*., 2005). EnvZ senses osmolarity and phosphorylates the response regulator OmpR. In turn, OmpR alters the expression of multiple genes including *ompF* and *ompC*, resulting in the abundance of OmpC in high osmolarity and the abundance of OmpF in low osmolarity. Interestingly, *ompF* is also positively represented after growth in vancomycin, but below the L_2FC threshold imposed in this study. OmpC is a transmembrane osmoporin in the outer membrane which allows the movement of nutrients across the cell envelope (Nikaido, 2003; Maeda *et al*., 1991). The positive representation of these genes, that together comprise a regulatory pathway, is highly indicative of biological significance. The disruption of these porins would lead to decreased movement of water and solutes out of the cell, which would provide a growth advantage in conditions of high osmolarity. Given that OmpC is downstream of EnvZ and OmpR in the activation pathway, it is likely that the presence of *envZ* and *ompR* in the positively represented list of genes is solely due to their role in positively regulating expression of porins. These findings also suggest that native OmpC predisposes the cell to

vancomycin susceptibility, which is support by findings reported by Tran *et al*. (2014). *nlpI* is a predicted outer membrane bound lipoprotein (Wilson, Kajander and Regan, 2005; O'hara *et al*., 1999). *In vivo*, NlpI acts as an adaptor protein to facilitate the degradation of the peptidoglycan endopeptidase MepS by the protease Prc, which is also in the positively represented gene list (Singh *et al*., 2015). However, other research demonstrates that the inactivation of *nlpI* leads to the overproduction of outer membrane vesicles, a phenotype which has been linked to increased fitness under stress conditions (McBroom and Kuehn, 2007). *rseA* encodes an anti-sigma factor that inhibits σ^E (Missiakas *et al*., 1997). The binding of RseA to σ^E physically prevents it from regulating genes involved with multiple environmental changes, one of which is osmolarity (Bianchi and Baneyx, 1999). The inactivation of *rseA* would lead to the constitutive activation of σ^E, which in turn is likely to lead to the regulation of genes that reduce cellular susceptibility to vancomycin. *opgH* is a glycosyltransferase that is embedded in the inner membrane (Bohin, 2000; Bontemps-Gallo and Lacroix, 2015). This enzyme catalyses a key step in the formation of multiple osmoregulated periplasmic glucans (OPGs). These molecules have been linked to several functions, including envelope structure and maintenance, virulence and pathogenicity. It has been previously shown that the inactivation of *opgH* leads to the increased expression of colanic acid, which might lead to increased vancomycin resistance (Ebel *et al*., 1997).

Of the ten remaining positively represented genes, 3 are predicted to be associated with the cell envelope. *yhcB* encodes a protein associated with the inner membrane (Stenberg *et al*., 2005). Little is known about the function of *yhcB*, although it has previously been found to be synthetically lethal with *rodZ* (Li, Hamamoto and Kitakawa, 2012). *uspB* is a protein predicted to be membrane associated (Farewell, Kvint and Nystrom, 1998). In this work, *uspB* was implicated with ethanol resistance in stationary phase, although there is

very little other experimental evidence of function. The last of the 3 genes, *wzzE*, encodes a protein located in the periplasm that regulates the polysaccharide chain length of the enterobacterial common antigen (Barr, Klena and Rick, 1999).

For the remaining 6 positively represented genes, there is no obvious evidence to explain their presence in this list. *clpP* encodes a serine protease which is a part of multiple protease complexes (Alexopoulos, Guarne and Ortega, 2012). *mazE* encodes an anti-toxin to MazF, which exhibits ribonuclease activity and in turn leads to global translation inhibition (Zhang *et al*., 2003). *ihfB* is one of the two subunits of the integration host factor, which is a transcriptional regulator (Goosen and van de Putte, 1995). *mtlR* encodes the mannitol repressor, which regulates the *mtlA* and *mtlD* mannitol utilisation genes (Figge, Ramseier and Saier Jr, 1994). Very little literature is available for *ytfK*, although it has been previously implicated as part of the phosphate regulon (Yoshida *et al*., 2011). *hfq* encodes an RNA binding protein that has been implicated in global gene regulation (Sobrero and Valverde, 2012).

5.2.5 Comparison with other studies. Through the use of the KEIO library, Tamae *et al*. (2008) investigated the susceptibility of *E. coli* to multiple classes of antibiotics, including vancomycin. Thirty one genes were reported to increase susceptibility to vancomycin upon deletion, and these genes were compared with the genes reported to be involved with vancomycin resistance from this work. For seventeen of these 31 genes, $\log_2$ fold changes of less than 0.3 in either positive or negative direction were reported, translating to less than a 1.23 fold change in representation. Seven genes had negative L_2FC values of over 0.3 but below the previously chosen threshold of 1, and 6 genes had negative L_2FC values over 1, and were present in the gene lists discussed previously. Intriguingly there is a single example

of a gene, namely *envZ*, in which Tamae *et al*. (2008) reported sensitivity to vancomycin in contrast to this work, in which a positive L_2FC value was reported beyond the set threshold. As such, 18 of the 31 genes could be said to be incongruent with the findings reported here, while the remaining 13 are reported similarly in both datasets. Liu *et al*. (2010) similarly investigated the effects of vancomycin on the KEIO collection, and subsequently published a list of 52 genes that led to vancomycin sensitivity upon deletion. In the same manner as above, the log_2-fold changes from the dataset generated here were collected for each of the 52 vancomycin sensitive genes reported by Liu *et al*. (2010). Using the same thresholds as set to compare against Tamae *et al*. (2008), 34 genes had $\log_2$ fold changes of less than 0.3 in either direction, 12 genes had negative L_2FC values between 0.3 and 1, and 5 genes had negative L_2FC values of over 1. Additionally, *envZ* was reported to negatively impact growth, in contrast to the positive L_2FC reported from this work.

From both comparisons of Tamae *et al*. (2008) and Liu *et al*. (2010), the majority of genes shown to be important for vancomycin resistance in these papers were not reported to be substantially affected after growth in vancomycin in this work. One key consideration is a difference in experimental design. In both studies the KEIO collection was grown in microtitre plates containing LB, followed by plating onto LB agar containing vancomycin. The plating of mutants on agar containing different concentrations of antibiotics allowed for the calculation of minimum inhibitory concentrations, which were then used to define genes linked to antibiotic sensitivity. In the hybrid transposon sequencing method adopted in this work, no growth on solid media is required. As such, the findings discussed are not directly comparable. Furthermore, when in the microtitre plates only a growth period of 3-4 hours is referred to, as opposed to the defined growth permitted through the transposon sequencing method used here. As such, it could be argued that the more rigorous application of the

transposon sequencing experiments would have resulted in more robust results than those of Liu *et al*. and Tamae *et al*. Finally, considering that the same experimental techniques and antibiotic concentrations are used in Liu *et al*. and Tamae *et al*., there is incongruence in the genes reported from either study. Out of the 52 genes reported by Liu *et al*., and the 31 genes reported from Tamae *et al*., only 16 are present in each, with 15 genes being found exclusively in the Tamae *et al*. list of genes and 36 genes being reported exclusively by Liu *et al*. Arguably, this disparity casts doubt on the veracity of these findings.

As a comparator for the genes resulting from the SDS dataset, fitness scores calculated by Nichols *et al*. (2010) were used. The experimental design employed by Nichols *et al*. is similar to that of Tamae *et al*. and Liu *et al*., in that the KEIO library was used. In contrast, however, the cultures were grown in a 1536 well format on solid LB agar plates, from which colony size was digitally measured and used to create fitness scores from comparing LB plates with and without antibiotic or chemical. Across the 45 differentially represented genes from the SDS dataset, the range of fitness scores varies from -23.29 to 1.58, with negative values indicating impacted growth rate in the presence of antibiotic and positive values indicating an increased growth rate. Out of the 44 negatively represented genes, 35 genes have negative fitness scores, 6 have positive scores and for 3 genes no score was available. Of the 35 genes with negative fitness scores, approximately two thirds have scores between 0 and -4, where the remaining genes have scores spread between approximately -5 and -23. Interestingly, these genes with the lowest negative fitness scores include the *rfa* and *acr* genes, for which there is much corroborating evidence in the literature. In consideration of the 6 genes with positive fitness scores, they are spread from approximately 0.12 to 1.59, which indicates that these genes do not lead to appreciably different growth rates in the presence of SDS, especially in light of the negative fitness

scores reported previously. For the single gene that was positively represented in the transposon sequencing data, *rssB*, Nichols *et al*. give a fitness score of ~0.59. Again, this fitness score is not strong evidence of a change in fitness.

It is important to keep in mind that because of the techniques used in these studies the data are not precisely comparable with the data generated in this study. Additionally, there is a key difference in the usage of fitness scores and $\log_2$-fold changes, which are not directly comparable metrics. Furthermore, for the fitness scores reported by Nichols *et al*. no statistical likelihood values were available, in contrast to the analysis undertaken in this work. Another key difference between the use of transposon sequencing versus traditional knockout libraries is that the positive growth effects of disrupted genes are quantifiable, where this was not possible from Tamae *et al*. and Liu *et al*. Even so, when considering that there was an overlap of reported genes whose involvement with resistance to SDS is well understood, the utility of both techniques have been shown.

5.3 Discussion

The hybrid transposon sequencing method from chapter 3 has been used in this work to investigate which genes are involved with resistance to vancomycin and SDS. Of the genes that passed the thresholds chosen for significance and fold change, literature supporting their involvement in resistance to these chemicals could be found for them, and many of the genes are known to be involved in related cellular functions, such as envelope maintenance.

One potentially glaring omission from this chapter is the lack of additional wet lab experimentation to confirm these findings; none of the candidate genes that were positively or negatively represented were singularly deleted and then tested to confirm any change in growth rate. It could be argued that with a new technique such as transposon sequencing

such confirmation is of the utmost importance. However, given that many genes empirically proven to be involved with resistance have been shown to be differentially represented is strong evidence that at least some of the findings here are robust and meaningful. The problem is more likely to lie with genes for which no literature evidence could be found. Further experimentation is crucial to fully assess these genes.

Another point of consideration in this work is with regard to the choice of thresholds used for significance and fold change. As is well known among researchers using RNAseq as an experimental technique, a gene with a L_2FC of 1 is arguably just as interesting as a gene with a L_2FC of 0.99. Even with this criticism, the use of thresholds is justifiable when considering the number of genes that may result from such a study. It makes sense to target the search for genes that are maximally affected by the growth conditions used. Over time, as the analysis of such data becomes more commonplace there may be other methods developed which make better use of and allow for wider consideration of the data.

CHAPTER 6

GENERAL DISCUSSION

The work presented here has demonstrated the testing and comparison of three transposon insertion sequencing methodologies, and the application of the chosen technique in the assessment of genes and their phenotypic relevance to envelope homeostasis. Discussion will be presented in the context of each chapter below. This work represents the establishment of an investigatory technique in a working laboratory, from the very beginnings of growth experimentation, through to the raw data generation and subsequent end result of data analysis. It is expected that this work will form a seed from which further work will grow, with varied experimental aims and outcomes.

The ultimate aim of the work in chapter 3 was to select an insertion sequencing technique with which to do follow on studies. To date, despite 7 years of insertion sequencing publications (using Langridge *et al*. (2009) as a benchmark), there are no literature references that specifically compare insertion sequencing methodologies. As such, the work presented in this chapter is uniquely informative of methodological differences, given that the same transposon library was used throughout.

Although the hybrid methodology was chosen for further use, this is by no means the "best" insertion sequencing technique available; merely, the best of the three tested. It is expected that, in time, there will be protocol changes and entirely different approaches that will surpass this method in every metric. Improvements are likely to occur in every aspect of the technique. For example, in the library preparation steps, there is much potential for refinement and optimisation of the current protocol, and even the adoption of different experimental techniques (for example, different ways to quantify and size select sequencing libraries). In the sequencing steps, there may, for example, be improvements enabled by the recent advances in nanopore sequencing by companies such as Oxford Nanopore Technologies. Analytically, there are a number of approaches that can be taken in the

assessment of gene essentiality and gene fitness, and it is certain that there will be new approaches that will provide better statistically predictive capabilities.

Chapter 4 details the use of the hybrid insertion sequencing technique to predict the essential genes of *E. coli*, and to compare the list generated with the literature and specifically the findings from the KEIO library. The use of insertion sequencing in this way is a definitive strength of the technique, and is arguably more efficient and informative in comparison to the previously used approach of creating single deletion libraries. The most striking demonstration of this is in the presence of genes with essential regions. Single deletion libraries may correctly assess a particular gene as essential due to the lack of derivable knock out strain, but insertion sequencing also has the potential to illuminate exactly which regions of a gene are indeed essential. The work presented here also identified essential genes that were missed in the first iteration of the KEIO library work, and only corrected upon further investigation. This shows the robustness and sensitivity that insertion sequencing can achieve.

As ever, there are caveats. In contrast to single deletion libraries, the data analysis is much more intensive. Where deletions can be confirmed via PCR and partially assessed via growth on an agar plate, insertion sequencing data must be heavily processed and manipulated. Although such bioinformatic analysis is becoming ever more widespread, it still poses an obstacle for researchers in its execution and interpretation. Furthermore, in this work not all of the analysis was totally objective. Indeed, laborious manual inspection was used to look at candidate essential genes in greater detail. However, it is almost certain that newer software packages and programs will be released that can at the very least undertake aspects of the analysis done here, along with the potential to perform as yet unused analyses. Additionally, there is much consideration required when discussing genes in terms

of essentiality. The interpretation of insertion sequencing data is wholly based upon factors such as the "age" of a culture at the point of DNA sampling. For example, if there was a particular gene that, while not ultimately essential for growth, was important for the rapidity of cell growth, this gene may be erroneously predicted to be essential based on when the culture was sample after initial transposon library creation. This suggests that more established lab techniques such as single gene deletions will still be required in future, as a confirmatory practice. Here lies another issue with the work presented, in that no further wet lab confirmation of essential gene candidates was undertaken. However, given that a wealth of literature was present to support the assertions made, this is not considered to be an issue.

The aim of chapter 5 was to use the hybrid sequencing technique to define genes important for the maintenance of the cell envelope. The linking of gene functionality with environmental conditions is another incredibly useful application of insertion sequencing, especially when considering that, in a single experiment, a whole genome is assessed in response to a particular condition. In addition to the volume of data that can be gleaned from such experiments, insertion sequencing was demonstrated to show not only the negative impacts of gene disruption, but also positive effects on growth fitness. This extra layer of contextual information provided is in contrast to previously adopted single gene deletion approaches, where only the lack of growth was the central metric of assessment. This in turn allows for the deeper understanding of underlying genetic networks and the greater integration of functional knowledge. More broadly speaking, insertion sequencing should be applicable to any selective pressure or growth condition in any organism that is tolerant of insertions. This in turn suggests the application of this technique in multiple

areas, for example, in the search for the next generation of antimicrobial molecules, or in the search for new as yet unknown genes that perform societally useful functions.

Perhaps the most difficult issue to resolve when discussing this work is in the assessment of representation with and without selective pressure. Here, the approach used was to calculate and compare fold-change differences in representation for each gene. This is the same approach used in RNA-seq and other sequencing based techniques. The statistical cut-offs used here were also taken from standard RNA-seq protocols. However, the selection of these thresholds could be said to lead to the dismissal of biologically relevant information. For example, consider two genes with $\log_2$ fold changes of 0.99 and 1. The two are both likely to be important or represent at the least interesting results, but the 0.01 difference in $\log_2$ fold change would mean that only one gene would actually be taken forward. This issue lies with any technique that utilises statistical analysis, and will continue to do so. It is almost certain that other more powerful and applicable statistical methodologies exist, and it would be wise to test them to find the strongest. When considering the changing of the current analytical pipeline, it would also be beneficial to try and incorporate more features, for example in the automatic outlining of genes containing essential genomic regions. This would greatly reduce the manual labour required in analysis. Another key concern is that for all the examples of differentially represented genes, none have been confirmed through knock outs or complementation studies. This would be the next logical step in terms of practical work. Additionally, it would be beneficial to test the hybrid methodology in application to a different insertion library in *E. coli*, or to a library in a completely different organism. Ideally it would be beneficial to validate the technique for use with multiple transposons used to make insertion libraries. The next two discussion points regarding this work relate to more practical concerns. The analysis made looked only

at the coding sequences within the genome, with no attention paid to any other genomic feature. Given the genome wide insertion of transposons, there may be other genomic features relating to envelope integrity that have simply been glossed over. While it is certainly possible for this to be done, it would require more analytical steps to be coded for and incorporated into the scripts. This leads to the second practical concern, in the assessment of the differentially represented genes reported by the analysis. Here, a literature search was undertaken for each and every gene passing the thresholds set in the analysis. This was time consuming and laborious, and the utility of the technique would be greatly improved if future software could be designed to automate this process. As a more distant future goal, the generation of a matrix with data concerning each gene in a multitude of growth conditions would be fantastic. This data would be highly informative across many fields of study, and could become a cornerstone reference for many researchers.

CHAPTER 7

BIBLIOGRAPHY

Acevedo-Rocha, C.G., Fang, G., Schmidt, M., Ussery, D.W. and Danchin, A. (2013) 'From essential to persistent genes: A functional approach to constructing synthetic life', Trends in Genetics, 29(5), pp. 273–279.

Alexopoulos, J.A., Guarné, A. and Ortega, J. (2012) 'ClpP: A structurally dynamic protease regulated by AAA+ proteins', *Journal of Structural Biology*, 179(2), pp. 202–210.

Aseev, L.V., Bylinkina, N.S. and Boni, I.V. (2015) 'Regulation of the *rplY* gene encoding 5S rRNA binding protein L25 in *Escherichia coli* and related bacteria', *RNA*, 21(5), pp. 851–861.

Aussel, L., Loiseau, L., Hajj Chehade, M., Pocachard, B., Fontecave, M., Pierrel, F. and Barras, F. (2013) '*ubiJ*, a new gene required for aerobic growth and proliferation in Macrophage, is involved in Coenzyme Q Biosynthesis in *Escherichia coli* and *Salmonella enterica* serovar Typhimurium', *Journal of Bacteriology*, 196(1), pp. 70–79.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. and Mori, H. (2006) 'Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection', *Molecular Systems Biology*, 2.

Bailey, M.J.A., Hughes, C. and Koronakis, V. (1996) 'Increased distal gene transcription by the elongation factor RfaH, a specialized homologue of NusG', *Molecular Microbiology*, 22(4), pp. 729–737.

Baneyx, F. (1999) 'Recombinant protein expression in *Escherichia coli*', Current Opinion in Biotechnology, 10(5), pp. 411–421.

Barr, K., Klena, J. and Rick, P.D. (1999) 'The Modality of Enterobacterial Common Antigen Polysaccharide Chain Lengths Is Regulated by *o349* of the *wec* Gene Cluster of *Escherichia coli* K-12', *Journal of Bacteriology*, 181(20), p. 65646568.

Barras, F., Loiseau, L. and Py, B. (2005) 'How *Escherichia coli* and *Saccharomyces cerevisiae* build Fe/S proteins', Advances in Microbial Physiology, 50, pp. 40–101.

Bartholomew, J.W. and Mittwer, T. (1952) 'The gram stain', *Bacteriology Reviews*, 16(1), pp. 1–29.

Basketter, D.A., English, J.S.C., Wakelin, S.H. and White, I.R. (2008) 'Enzymes, detergents and skin: Facts and fantasies', British Journal of Dermatology, 158(6), pp. 1177–1181.

Battesti, A., Majdalani, N. and Gottesman, S. (2011) 'The RpoS-Mediated general stress response in *Escherichia coli**', *Annual Review of Microbiology*, 65(1), pp. 189–213.

Becker, G., Klauck, E. and Hengge-Aronis, R. (1999) 'Regulation of RpoS proteolysis in *Escherichia coli*: The response regulator RssB is a recognition factor that interacts with the turnover element in RpoS', *Proceedings of the National Academy of Sciences*, 96(11), pp. 6439–6444.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J. and et, al (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(6), pp. 53–59.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) 'The protein data bank', Nucleic Acids Research, 28(1), pp. 235–242.

Bernadac, A., Gavioli, M., Lazzaroni, J.-C., Raina, S. and Lloubes, R. (1998) '*Escherichia coli tol-pal* Mutants Form Outer Membrane Vesicles', *Journal of Bacteriology*, 180(18), pp. 4872–4878.

Bernhardt, T.G. and De Boer, P.A.J. (2004) 'Screening for synthetic lethal mutants in *Escherichia coli* and identification of EnvC (YibP) as a periplasmic septal ring factor with murein hydrolase activity', *Molecular Microbiology*, 52(5), pp. 1255–1269.

Bernstein, A., Rolfe, B. and Onodera, K. (1972) 'Pleiotropic properties and genetic organization of the *tolAB* locus of *Escherichia coli* K-12*'*, Journal of Bacteriology, 112(1), pp. 74–83.

Beveridge, T.J. (1989) 'Mechanism of Gram Variability in Select Bacteria', *Journal of Bacteriology*, 172(3), pp. 1609–1620.

Bianchi, A.A. and Baneyx, F. (1999) 'Hyperosmotic shock induces the sigma32 and sigmaE stress regulons of *Escherichia coli*', *Molecular Microbiology*, 34(5), pp. 1029–1038.

Blair, D.F. (1995) 'How bacteria sense and swim', *Annual Review of Microbiology*, 49(1), pp. 489–522.

Blount, Z.D. (2015) 'The unexhausted potential of *E. coli'*, *eLife*, 4.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120.

Bontemps-Gallo, S. and Lacroix, J.-M. (2015) 'New insights into the biological role of the osmoregulated periplasmic glucans in pathogenic and symbiotic bacteria', *Environmental Microbiology Reports*, 7(5), pp. 690–697.

Born, P., Breukink, E. and Vollmer, W. (2006) 'In Vitro synthesis of cross-linked Murein and its attachment to Sacculi by PBP1A from *Escherichia coli*', Journal of Biological Chemistry, 281(37), pp. 26985–26993.

Bos, M.P., Robert, V. and Tommassen, J. (2007) 'Biogenesis of the gram-negative bacterial outer membrane', *Annual Review of Microbiology*, 61(1), pp. 191–214.

Bourret, R.B. and Silversmith, R.E. (2010) 'Two component signal transduction', *Current Opinion in Microbiology*, 13(2), pp. 113–115.

Brooke, J.S. and Valvano, M.A. (1996) 'Biosynthesis of inner core Lipopolysaccharide in Enteric bacteria identification and characterization of a conserved Phosphoheptose Isomerase', *Journal of Biological Chemistry*, 271(7), pp. 3608–3614.

Browning, D.F., Matthews, S.A., Rossiter, A.E., Sevastsyanovich, Y.R., Jeeves, M., Mason, J.L., Wells, T.J., Wardius, C.A., Knowles, T.J., Cunningham, A.F., Bavro, V.N., Overduin, M. and

Henderson, I.R. (2013) 'Mutational and Topological analysis of the *Escherichia coli* BamA protein', *PLoS ONE*, 8(12), p. e84512.

Bubunenko, M., Baker, T. and Court, D.L. (2007) 'Essentiality of ribosomal and transcription Antitermination proteins analyzed by systematic gene replacement in *Escherichia coli*', *Journal of Bacteriology*, 189(7), pp. 2844–2853.

Bubunenko, M., Korepanov, A., Court, D.L., Jagannathan, I., Dickinson, D., Chaudhuri, B.R., Garber, M.B. and Culver, G.M. (2006) '30S ribosomal subunits can be assembled in vivo without primary binding ribosomal protein S15', *RNA*, 12(7), pp. 1229–1239.

Bukau, B. and Walker, G.C. (1989) 'Cellular Defects Caused by Deletion of the *Escherichia coli dnaK* Gene Indicate Roles for Heat Shock Protein in Normal Metabolism', *Journal of Bacteriology*, 171(5), pp. 2337–2346.

Cai, S.J. and Inouye, M. (2002) 'EnvZ-OmpR interaction and Osmoregulation in *Escherichia coli*', *Journal of Biological Chemistry*, 277(27), pp. 24155–24161.

Canals, R., Xia, X.-Q., Fronick, C., Clifton, S.W., Ahmer, B.M., Andrews-Polymenis, H.L., Porwollik, S. and McClelland, M. (2012) 'High-throughput comparison of gene fitness among related bacteria', *BMC Genomics*, 13(1), p. 212.

Carabetta, V.J., Mohanty, B.K., Kushner, S.R. and Silhavy, T.J. (2009) 'The response regulator SprE (RssB) Modulates Polyadenylation and mRNA stability in *Escherichia coli*', *Journal of Bacteriology*, 191(22), pp. 6812–6821.

Cascales, E., Bernadac, A., Gavioli, M., Lazzaroni, J.. and Lloubes, R. (2002) 'Pal Lipoprotein of *Escherichia coli* plays a major role in outer membrane integrity', *Journal of Bacteriology*, 184(3), pp. 754–759.

Chang, A.C.Y. and Cohen, S.N. (1978) 'Construction and Characterization of Amplifiable Multicopy DNA Cloning Vehicles Derived from the P15A Cryptic Miniplasmid', *Journal of Bacteriology*, 134(3), pp. 1141–1156.

Chi, K.R. (2008) 'The year of sequencing', *Nature Methods*, 5(1), pp. 11–14. doi: 10.1038/nmeth1154.

Christen, B., Abeliuk, E., Collier, J.M., Kalogeraki, V.S., Passarelli, B., Coller, J.A., Fero, M.J., McAdams, H.H. and Shapiro, L. (2014) 'The essential genome of a bacterium', *Molecular Systems Biology*, 7(1), pp. 528–528.

Christiansen, M.T., Kaas, R.S., Chaudhuri, R.R., Holmes, M.A., Hasman, H. and Aarestrup, F.M. (2014) 'Genome-wide high-throughput screening to investigate essential genes involved in Methicillin-Resistant *Staphylococcus aureus* sequence type 398 survival', *PLoS ONE*, 9(2), p. e89018.

Coderre, P.E. and Earhart, C.F. (1989) 'The *entD* gene of the *Escherichia coli* K12 Enterobactin gene cluster', *Microbiology*, 135(11), pp. 3043–3055.

Cole, J. (2016) ‘Antimicrobial resistance – a “rising tide” of national (and international) risk’, *Journal of Hospital Infection*, 92(1), pp. 3–4.

Cowles, C.E., Li, Y., Semmelhack, M.F., Cristea, I.M. and Silhavy, T.J. (2011) ‘The free and bound forms of Lpp occupy distinct subcellular locations in *Escherichia coli*’, *Molecular Microbiology*, 79(5), pp. 1168–1181.

Cronan, J.E. (2014) ‘*Escherichia coli* as an Experimental Organism’, *eLS*, pp. 1–7.

Dalbey, R.E., Kuhn, A., Zhu, L. and Kiefer, D. (2014) ‘The membrane insertase YidC’, *Biochimica et Biophysica Acta*, 1843(8), pp. 1489–1496.

Dalebroux, Z.D., Edrozo, M.B., Pfuetzner, R.A., Ressl, S., Kulasekara, B.R., Blanc, M.-P. and Miller, S.I. (2015) ‘Delivery of Cardiolipins to the salmonella outer membrane is necessary for survival within host tissues and virulence’, *Cell Host & Microbe*, 17(4), pp. 441–451.

D’Ari, R., Jaffe, A., Bouloc, P. and Robin, A. (1988) ‘Cyclic AMP and Cell Division in *Escherichia coli*’, *Journal of Bacteriology*, 170(1), pp. 65–70.

Dassain, M., Leroy, A., Colosetti, L., Carolé, S. and Bouché, J.-P. (1999) ‘A new essential gene of the “minimal genome” affecting cell division’, *Biochimie*, 81(8-9), pp. 889–895.

Datsenko, K.A. and 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*, 97(12), pp. 6640–6645.

Davies, K. (2010) Bio-iT world. Available at: http://www.bio-itworld.com/2010/issues/sept-oct/solexa.html (Accessed: 13 September 2016).

De Lay, N.R. and Cronan, J.E. (2008) ‘Genetic interaction between the *Escherichia coli* AcpT Phosphopantetheinyl Transferase and the YejM inner membrane protein’, *Genetics*, 178(3), pp. 1327–1337.

Deng, M. and Misra, R. (1996) ‘Examination of AsmA and its effect on the assembly of *Escherichia coli* outer membrane proteins’, *Molecular Microbiology*, 21(3), pp. 605–612.

Denoncin, K., Schwalm, J., Vertommen, D., Silhavy, T.J. and Collet, J.-F. (2012) ‘Dissecting the *Escherichia coli* periplasmic chaperone network using differential proteomics’, *Proteomics*, 12(9), pp. 1391–1401.

Derouaux, A., Sauvage, E. and Terrak, M. (2013) ‘Peptidoglycan Glycosyltransferase substrate mimics as Templates for the design of new antibacterial drugs’, *Frontiers in Immunology*, 4, 78.

Driessen, A.J.M. and Nouwen, N. (2008) ‘Protein Translocation across the bacterial Cytoplasmic membrane’, *Annual Review of Biochemistry*, 77(1), pp. 643–667.

Du, D., Wang, Z., James, N.R., Voss, J.E., Klimont, E., Ohene-Agyei, T., Venter, H., Chiu, W. and Luisi, B.F. (2014) ‘Structure of the AcrAB–TolC multidrug efflux pump’, *Nature*, 509(7501), pp. 512–515.

Duigou, S., Silvain, M., Viguera, E. and Michel, B. (2014) 'Ssb gene duplication restores the viability of *ΔholC* and *ΔholD Escherichia coli* Mutants', *PLoS Genetics*, 10(10), p. e1004719.

Ebel, W., Vaughn, G.J., Peters III, H.K. and Trempy, J.E. (1997) 'Inactivation of *mdoH* Leads to Increased Expression of Colanic Acid Capsular Polysaccharide in *Escherichia coli*', *Journal of Bacteriology*, 179(21), pp. 6858–6861.

Egan, A.J.F. and Vollmer, W. (2012) 'The physiology of bacterial cell division', *Annals of the New York Academy of Sciences*, 1277(1), pp. 8–28.

Elkins, C.A. and Nikaido, H. (2002) 'Substrate specificity of the RND-Type Multidrug Efflux pumps AcrB and AcrD of *Escherichia coli* is determined predominately by Two large Periplasmic loops', *Journal of Bacteriology*, 184(23), pp. 6490–6498.

Escherich, T. (1988) 'The intestinal bacteria of the Neonate and breast-fed infant', *Review of Infectious Disease*, 10(6), pp. 1220–1225.

Farewell, A., Kvint, K. and Nystrom, T. (1998) '*uspB*, a New sigmaS-Regulated Gene in *Escherichia coli* Which Is Required for Stationary-Phase Resistance to Ethanol', *Journal of Bacteriology*, 180(23), pp. 6140–6147.

Feng, X., Oropeza, R., Walthers, D., Kenney, L.J. (2003) 'OmpR Phosphorylation and Its Role in Signaling and Pathogenesis', *ASM News,* 69(8), pp. 390-395

Figge, R.M., Ramseier, T.M. and Saier Jr, M.H. (1994) 'The Mannitol Repressor (MtlR) of *Escherichia coli*', *Journal of Bacteriology*, 176(3), pp. 840–847.

Fontaine, F., Stewart, E.J., Lindner, A.B. and Taddei, F. (2007) 'Mutations in two global regulators lower individual mortality in *Escherichia coli*', *Molecular Microbiology*, 67(1), pp. 2-14

Fricke, J., neuhard, J., Kelln, R.A. and Pedersen, S. (1995) 'The *cmk* Gene Encoding Cytidine Monophosphate Kinase Is Located in the *rpsA* Operon and Is Required for Normal Replication Rate in *Escherichia coli*', *Journal of Bacteriology*, 177(3), pp. 517–523.

Gerdes, S. and Osterman, A.L. (eds.) (2008) *Microbial Gene Essentiality: Protocols and Bioinformatics*. United States: Springer-Verlag New York.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, 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., Barabasi, A.., Oltvai, Z.N. and Osterman, A.L. (2003) 'Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655', *Journal of Bacteriology*, 185(19), pp. 5673–5684.

Gerding, M.A., Ogata, Y., Pecora, N.D., Niki, H. and de Boer, P.A.J. (2007) 'The trans - envelope Tol-Pal complex is part of the cell division machinery and required for proper outer-membrane invagination during cell constriction in *E. coli*', *Molecular Microbiology*, 63(4), pp. 1008–1025.

Gershanovitch, V., Ilvina, T., Rusina, O., Yourovitskaya, N. and Bolshakova, T. (1977) 'Repression of inducible enzyme synthesis in a mutant of *Escherichia coli* K12 deleted for the *ptsH* gene', MGG Molecular & General Genetics, 153(2).

Ghosh, A.S., Chowdhury, C. and Nelson, D.E. (2008) 'Physiological functions of D-alanine carboxypeptidases in *Escherichia coli*', Trends in Microbiology, 16(7), pp. 309–317.

Girgis, H.S., Hottes, A.K. and Tavazoie, S. (2009) 'Genetic architecture of intrinsic antibiotic susceptibility', *PLoS ONE*, 4(5), p. e5629.

Goemans, C., Denoncin, K. and Collet, J.-F. (2014) 'Folding mechanisms of periplasmic proteins', *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1843(8), pp. 1517–1528.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) 'Coming of age: Ten years of next-generation sequencing technologies', *Nature Reviews Genetics*, 17(6), pp. 333–351.

Goosen, N. and Putte, P. (1995) 'The regulation of transcription initiation by integration host factor', *Molecular Microbiology*, 16(1), pp. 1–7.

Gopel, Y., Luttmann, D., Heroven, A.K., Reichenbach, B., Dersch, P. and Gorke, B. (2010) 'Common and divergent features in transcriptional control of the homologous small RNAs GlmY and GlmZ in Enterobacteriaceae', *Nucleic Acids Research*, 39(4), pp. 1294–1309.

Gronow, S., Brabetz, W. and Brade, H. (2000) 'Comparative functional characterization in vitro of heptosyltransferase I (WaaC) and II (WaaF) from *Escherichia coli*', *European Journal of Biochemistry*, 267(22), pp. 6602–6611.

Gumbart, J.C., Beeby, M., Jensen, G.J. and Roux, B. (2014) '*Escherichia coli* Peptidoglycan structure and mechanics as predicted by atomic-scale simulations', *PLoS Computational Biology*, 10(2), p. e1003475.

Guo, M., Wang, H., Xie, N. and Xie, Z. (2015) 'Positive effect of carbon sources on natural transformation in *Escherichia coli*: Role of low-level Cyclic AMP (cAMP)-cAMP receptor protein in the Derepression of *rpoS'*, *Journal of Bacteriology*, 197(20), pp. 3317–3328.

Gustafson, A.M., Snitkin, E.S., Parker, S.C.J., DeLisi, C. and Kasif, S. (2006) 'Towards the identification of essential genes using targeted genome sequencing and comparative analysis', BMC Genomics, 7(265).

Hagan, C.L., Silhavy, T.J. and Kahne, D. (2011) 'β-barrel membrane protein assembly by the bam complex', *Annual Review of Biochemistry*, 80(1), pp. 189–210.

Hamer, L., DeZwaan, T.M., Montenegro-Chamorro, M.V., Frank, S.A. and Hamer, J.E. (2001) 'Recent advances in large-scale transposon mutagenesis', *Current Opinion in Chemical Biology*, 5(1), pp. 67–73.

Hancock, R.E.W. (1987) 'Role of Porins in Outer Membrane Permeability', *Journal of Bacteriology*, 169(3), pp. 929–933.

Hase, Y., Tarusawa, T., Muto, A. and Himeno, H. (2013) 'Impairment of Ribosome maturation or function confers salt resistance on *Escherichia coli* cells', *PLoS ONE*, 8(5), p. e65747.

Hase, Y., Yokoyama, S., Muto, A. and Himeno, H. (2009) 'Removal of a ribosome small subunit-dependent GTPase confers salt resistance on *Escherichia coli* cells', *RNA*, 15(9), pp. 1766–1774.

Heather, J.M. and Chain, B. (2015) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*, 107(1), pp. 1-8

Heidrich, C., Templin, M.F., Ursinus, A., Merdanovic, M., Berger, J., Schwarz, H., De Pedro, M.A. and Höltje, J.-V. (2001) 'Involvement of N-acetylmuramyl-l-alanine amidases in cell separation and antibiotic-induced autolysis of *Escherichia coli*', *Molecular Microbiology*, 41(1), pp. 167–178.

Heidrich, C., Ursinus, A., Berger, J., Schwarz, H. and Holtje, J.. (2002) 'Effects of multiple deletions of Murein Hydrolases on viability, septum cleavage, and sensitivity to large toxic molecules in *Escherichia coli*', *Journal of Bacteriology*, 184(22), pp. 6093–6099.

Heijenoort, J. v. (2001) 'Formation of the glycan chains in the synthesis of bacterial peptidoglycan', *Glycobiology*, 11(3), pp. 25R–36R.

van Heijenoort, J. (2011) 'Peptidoglycan Hydrolases of *Escherichia coli*', *Microbiology and Molecular Biology Reviews*, 75(4), pp. 636–663.

Hernández-Montalvo, V., Martínez, A., Hernández-Chavez, G., Bolivar, F., Valle, F. and Gosset, G. (2003) 'Expression of galP and glk in a *Escherichia coli* PTS mutant restores glucose transport and increases glycolytic flux to fermentation products', *Biotechnology and Bioengineering*, 83(6), pp. 687–694.

Hosler, J.P., Ferguson-Miller, S. and Mills, D.A. (2006) 'Energy transduction: Proton transfer through the respiratory complexes', *Annual Review of Biochemistry*, 75(1), pp. 165–187.

Hsiao, Y., Fang, W., Lee, C., Chen, Y. and Yuan, H. (2014) 'Structural Insights Into DNA Repair by RNase T—An Exonuclease Processing 39 End of Structured DNA in Repair Pathways', *PLOS Biology*, 12(3), e1001803.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., von Mering, C. and Bork, P. (2015) 'EggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic Acids Research*, 44(D1), pp. D286–D293.

Ikegami, A., Nishiyama, K., Matsuyama, S. and Tokuda, H. (2005) 'Transposome insertional mutagenesis and direct sequencing of microbial genomes', *Bioscience Biotechnology and Biochemistry*, 69(8), pp. 1595–1602.

Ikeuchi, Y., Shigi, N., Kato, J., Nishimura, A. and Suzuki, T. (2006) 'Mechanistic insights into sulfur relay by multiple sulfur mediators involved in Thiouridine Biosynthesis at tRNA wobble positions', *Molecular Cell*, 21(1), pp. 97–108.

Ito, M., Baba, T. and Mori, H. (2005) 'Functional analysis of 1440 *Escherichia coli* genes using the combination of knock-out library and phenotype microarrays', *Metabolic Engineering*, 7(4), pp. 318–327.

Ize, B., Stanley, N.R., Buchanan, G. and Palmer, T. (2003) 'Role of the *Escherichia coli* tat pathway in outer membrane integrity', *Molecular Microbiology*, 48(5), pp. 1183–1193.

Jang, S. and Imlay, J.A. (2010) 'Hydrogen peroxide inactivates the *Escherichia coli* Isc ironsulfur assembly system, and OxyR induces the Suf system to compensate', *Molecular Microbiology*, 78(6), pp. 1448–1467.

Johnson, I. (1983) 'Human insulin from recombinant DNA technology', *Science*, 219(4585), pp. 632–637. doi: 10.1126/science.6337396.

Juhas, M., Eberl, L. and Glass, J.I. (2011) 'Essence of life: Essential genes of minimal genomes', Trends in Cell Biology, 21(10), pp. 562–568.

Justice, S.S., Hunstad, D.A., Harper, J.R., Duguay, A.R., Pinkner, J.S., Bann, J., Frieden, C., Silhavy, T.J. and Hultgren, S.J. (2005) 'Periplasmic Peptidyl Prolyl cis-trans Isomerases are not essential for viability, but SurA is required for Pilus Biogenesis in *Escherichia coli*', *Journal of Bacteriology*, 187(22), pp. 7680–7686.

Kajimura, J., Rahman, A. and Rick, P.D. (2005) 'Assembly of Cyclic Enterobacterial common antigen in *Escherichia coli* K-12', *Journal of Bacteriology*, 187(20), pp. 6917–6927.

Kamionka, M. (2011) 'Engineering of therapeutic proteins production in *Escherichia coli*', *Current Pharmaceutical Biotechnology*, 12(2), pp. 268–274.

Kang, Y., Durfee, T., Glasner, J.D., Qiu, Y., Frisch, D., Winterberg, K.M. and Blattner, F.R. (2004) 'Systematic Mutagenesis of the *Escherichia coli* genome', *Journal of Bacteriology*, 186(15), pp. 4921–4930.

Zeevi, M., Shafir, N.S., Shaham, S., Friedman, S., Sigal, N., Nir Paz, R., Boneca, I.G. and Herskovits, A.A. (2013) 'Listeria monocytogenes Multidrug resistance transporters and Cyclic di-aMP, which contribute to type I interferon induction, play a role in cell wall stress', *Journal of Bacteriology*, 195(23), pp. 5250–5261.

Karigar, C.S. and Rao, S.S. (2011) 'Role of microbial enzymes in the Bioremediation of pollutants: A review', *Enzyme Research*, 2011, pp. 1–11.

Karp, P.D. and et, al (2002) 'The EcoCyc database', Nucleic Acids Research, 30(1), pp. 56–58.

Kato, J. and Katayama, T. (2001) 'Hda, a novel DnaA-related protein, regulates the replication cycle in *Escherichia coli*', *The EMBO Journal*, 20(15), pp. 4253–4262.

Ke, N. and Berkmen, M. (2014) 'Production of Disulfide-Bonded Proteins in *Escherichia coli*', *Current Protocols in Molecular Biology*, .

Kim, K.H., Aulakh, S. and Paetzel, M. (2012) 'The bacterial outer membrane β-barrel assembly machinery', *Protein Science*, 21(6), pp. 751–768.

Kneidinger, B., Marolda, C., Graninger, M., Zamyatina, A., McArthur, F., Kosma, P., Valvano, M.A. and Messner, P. (2002) 'Biosynthesis pathway of ADP-L-glycero- -D-manno-Heptose in *Escherichia coli*', *Journal of Bacteriology*, 184(2), pp. 363–369.

Knowles, T.J., Scott-Tucker, A., Overduin, M. and Henderson, I.R. (2009) 'Membrane protein architects: The role of the BAM complex in outer membrane protein assembly', *Nature Reviews Microbiology*, 7(3), pp. 206–214.

Kohanski, M.A., Dwyer, D.J., Hayete, B., Lawrence, C.A. and Collins, J.J. (2007) 'A common mechanism of cellular death induced by Bactericidal antibiotics', *Cell*, 130(5), pp. 797–810.

Krulwich, T.A., Sachs, G. and Padan, E. (2011) 'Molecular aspects of bacterial pH sensing and homeostasis', *Nature Reviews Microbiology*, 9(5), pp. 330–343.

Lander, E.S., International Human Genome Sequencing Consortium and et, al (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409, pp. 860–921.

Langridge, G.C., Phan, M.., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., Wain, J., Parkhill, J. and Turner, A.K. (2009) 'Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants', *Genome Research*, 19(12), pp. 2308–2316.

Lauhon, C.T. (2002) 'Requirement for IscS in Biosynthesis of all Thionucleosides in *Escherichia coli*', *Journal of Bacteriology*, 184(24), pp. 6820–6829.

Lazdunski, C. and Shapiro, B.M. (1972) 'Isolation and Some Properties of Cell Envelope Altered Mutants of *Escherichia coli*', *Journal of Bacteriology*, 111(2), pp. 495–496.

Lazzaroni, J.C., Germon, P., Ray, M.-C. and Vianney, A. (1999) 'The Tol proteins of *Escherichia coli* and their involvement in the uptake of biomolecules and outer membrane stability', *FEMS Microbiology Letters*, 177(2), pp. 191–197.

Le, H.V. and Trotta, P.P. (1991) 'Purification of secreted recombinant proteins from *Escherichia coli*', 12, pp. 163–181.

Lee, V.T. and Schneewind, O. (2001) 'Protein secretion and the pathogenesis of bacterial infections', *Genes & Development*, 15(14), pp. 1725–1752.

de Leeuw, E., Graham, B., Phillips, G.J., ten Hagen-Jongman, C.M., Oudega, B. and Luirink, J. (1999) 'Molecular characterization of *Escherichia coli* FtsE and FtsX', *Molecular Microbiology*, 31(3), pp. 983–993.

Levine, D.P. (2006) 'Vancomycin: A history', *Clinical Infectious Diseases*, 42(Supplement 1), pp. S5–S12.

Lewis, J., Alberts, B., Johnson, A. and Walter, P. (2007) Molecular biology of the cell. 5th edn. New York: Garland Publishing

Li, G., Hamamoto, K. and Kitakawa, M. (2012) 'Inner membrane protein YhcB interacts with RodZ involved in cell shape maintenance in*Escherichia coli*', *ISRN Molecular Biology*, 2012, pp. 1–8.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 'The sequence alignment/map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079.

Liu, A., Tran, L., Becket, E., Lee, K., Chinn, L., Park, E., Tran, K. and Miller, J.H. (2010) 'Antibiotic sensitivity profiles determined with an *Escherichia coli* gene knockout collection: Generating an antibiotic bar code', *Antimicrobial Agents and Chemotherapy*, 54(4), pp. 1393–1403.

Loll, P.J. and Axelsen, P.H. (2000) 'The structural biology of molecular recognition by Vancomycin', *Annual Review of Biophysics and Biomolecular Structure*, 29(1), pp. 265–289.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550.

Lovering, A.L., Safadi, S.S. and Strynadka, N.C.J. (2012) 'Structural perspective of Peptidoglycan Biosynthesis and assembly', *Annual Review of Biochemistry*, 81(1), pp. 451–478.

Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H. and Hearst, J.E. (1995) 'Genes acrA and acrB encode a stress-induced efflux system of *Escherichia coli*', *Molecular Microbiology*, 16(1), pp. 45–55.

Maeda, S., Takayanagi, K., Nishimura, Y., Maruyama, T., Sato, K. and Mizuno, T. (1991) 'Activation of the Osmoregulated ompC Gene by the OmpR Protein in *Escherichia coli*: A Study Involving Synthetic OmpR-Binding Sequences', *Journal of Biochemistry*, 110, pp. 324–327.

Mahalakshmi, S., Sunayana, M.R., SaiSree, L. and Reddy, M. (2013) 'YciM is an essential gene required for regulation of lipopolysaccharide synthesis in *Escherichia coli*', *Molecular Microbiology*, 91(1), pp. 145–157.

Malinverni, J.C. and Silhavy, T.J. (2009) 'An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane', *Proceedings of the National Academy of Sciences*, 106(19), pp. 8009–8014.

de Marco, A. (2009) 'Strategies for successful recombinant expression of disulfide bond-dependent proteins in *Escherichia coli*', *Microbial Cell Factories*, 8(1), p. 26.

McBroom, A.J. and Kuehn, M.J. (2006) 'Release of outer membrane vesicles by gram-negative bacteria is a novel envelope stress response', *Molecular Microbiology*, 63(2), pp. 545–558.

Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R. (2005) 'The microbial pan-genome', *Current Opinion in Genetics & Development*, 15(6), pp. 589–594.

Mempin, R., Tran, H., Chen, C., Gong, H., Kim Ho, K. and Lu, S. (2013) 'Release of extracellular ATP by bacteria during growth', *BMC Microbiology*, 13(1), p. 301.

Mergulhão, F.J.M., Summers, D.K. and Monteiro, G.A. (2005) 'Recombinant protein secretion in *Escherichia coli*', *Biotechnology Advances*, 23(3), pp. 177–202.

Miot, M. and Betton, J. (2004) 'Protein quality control in the bacterial periplasm', *Microbial Cell Factories*, 3:4.

Misra, R. (2012) 'Assembly of theβ-barrel outer membrane proteins in gram-negative bacteria, Mitochondria, and Chloroplasts', *ISRN Molecular Biology*, 2012, pp. 1–15.

Missiakas, D., Mayer, M.P., Lemaire, M., Georgopoulos, C. and Raina, S. (1997) 'Modulation of the *Escherichia coli* sigmaE (RpoE) heat-shock transcription-factor activity by the RseA, RseB and RseC proteins', *Molecular Microbiology*, 24(2), pp. 355–371.

Muffler, A., Fischer, D., Altuvia, S., Storz, G. and Hengge-Aronis, R. (1996) 'The response regulator RssB controls stability of the sigmas subunit of RNA polymerase in *Escherichia coli*', *EMBO*, 15(6), pp. 1333–1339.

Müller, R.T. and Pos, K.M. (2015) 'The assembly and disassembly of the AcrAB-TolC three-component multidrug efflux pump', Biological Chemistry, 396(9-10), pp. 1083-1089.

Munoz-Lopez, M. and Garcia-Perez, J. (2010) 'DNA Transposons: Nature and applications in Genomics', Current Genomics, 11(2), pp. 115–128. doi: 10.2174/138920210790886871.

Nelson, D.E. and Young, K.D. (2000) 'Penicillin binding protein 5 affects cell diameter, contour, and morphology of *Escherichia coli*', *Journal of Bacteriology*, 182(6), pp. 1714–1721.

Nichols, R.J., Sen, S., Choo, Y.J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K.M., Lee, K.J., Wong, A., Shales, M., Lovett, S., Winkler, M.E., Krogan, N.J., Typas, A. and Gross, C.A. (2011) 'Phenotypic landscape of a bacterial cell', *Cell*, 144(1), pp. 143–156.

Nikaido, H. (1998) 'Multiple antibiotic resistance and efflux', *Current Opinion in Microbiology*, 1(5), pp. 516–523.

Nikaido, H. (2003) 'Molecular basis of bacterial outer membrane Permeability revisited', *Microbiology and Molecular Biology Reviews*, 67(4), pp. 593–656.

Nikaido, H. and Vaara, M. (1985) 'Molecular Basis of Bacterial Outer Membrane Permeability', *Microbiological Reviews*, 49(1), pp. 1–32.

Nishio, Y., Ogishima, S., Ichikawa, M., Yamada, Y., Usuda, Y., Masuda, T. and Tanaka, H. (2013) 'Analysis of l-glutamic acid fermentation by using a dynamic metabolic simulation model of *Escherichia coli*', *BMC Systems Biology*, 7(1), p. 92.

O'Hara, M., Wu, H.C., Sankaran, K. and Rick, P.D. (1999) 'Identification and Characterization of a New Lipoprotein, NlpI, in *Escherichia coli* K-12', *Journal of Bacteriology*, 181(14), pp. 4318–4325.

Okamoto, S., Chin, T., Hiratsuka, K., Aso, Y., Tanaka, Y., Takahashi, T. and Ohara, H. (2014) 'Production of itaconic acid using metabolically engineered *Escherichia coli*', *The Journal of General and Applied Microbiology*, 60(5), pp. 191–197.

Okuda, S. and Tokuda, H. (2009) 'Model of mouth-to-mouth transfer of bacterial lipoproteins through inner membrane LolC, periplasmic LolA, and outer membrane LolB', *Proceedings of the National Academy of Sciences*, 106(14), pp. 5877–5882.

Okuda, S. and Tokuda, H. (2011) 'Lipoprotein sorting in bacteria', *Annual Review of Microbiology*, 65(1), pp. 239–259.

van Opijnen, T. and Camilli, A. (2013) 'Transposon insertion sequencing: A new tool for systems-level analysis of microorganisms', *Nature Reviews Microbiology*, 11(7), pp. 435–442.

Paier, A., Leppik, M., Soosaar, A., Tenson, T. and Maiväli, Ü. (2015) 'The effects of disruptions in ribosomal active sites and in intersubunit contacts on ribosomal degradation in *Escherichia coli*', *Scientific Reports*, 5, p. 7712.

Palmer, T. and Berks, B.C. (2012) 'The twin-arginine translocation (tat) protein export pathway', *Nature Reviews Microbiology*, 10(7), pp. 483–496.

Paradis-Bleau, C., Kritikos, G., Orlova, K., Typas, A. and Bernhardt, T.G. (2014) 'A genome-wide screen for bacterial envelope Biogenesis Mutants identifies a novel factor involved in cell wall precursor metabolism', PLoS Genetics, 10(1), p. e1004056.

Parsons, L.M., Lin, F. and Orban, J. (2006) 'Peptidoglycan recognition by pal, an outer membrane Lipoprotein † , ‡', *Biochemistry*, 45(7), pp. 2122–2128.

Pearson, W.R., Wood, T., Zhang, Z. and Miller, W. (1997) 'Comparison of DNA sequences with protein sequences', Genomics, 46(1), pp. 24–36.

Perrenoud, A. and Sauer, U. (2005) 'Impact of global Transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose Catabolism in *Escherichia coli*', *Journal of Bacteriology*, 187(9), pp. 3171–3179.

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. and Schembri, M.A. (2013) 'The serum Resistome of a globally disseminated Multidrug resistant Uropathogenic *Escherichia coli* clone', *PLoS Genetics*, 9(10), p. e1003834.

Piddock, L.J.V. (2006) 'Multidrug-resistance efflux pumps — not just for resistance', *Nature Reviews Microbiology*, 4(8), pp. 629–636.

Pittman, M.S., Corker, H., Wu, G., Binet, M.B., Moir, A.J.G. and Poole, R.K. (2002) 'Cysteine is exported from the *Escherichia coli* cytoplasm by CydDC, an ATP-binding cassette-type

transporter required for Cytochrome assembly', *Journal of Biological Chemistry*, 277(51), pp. 49841–49849.

du Plessis, D.J.F., Nouwen, N. and Driessen, A.J.M. (2011) 'The Sec translocase', *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1808(3), pp. 851–865.

Pohlschröder, M., Prinz, W.A., Hartmann, E. and Beckwith, J. (1997) 'Protein Translocation in the Three domains of life: Variations on a theme', *Cell*, 91(5), pp. 563–566.

Pootoolal, J., Neu, j and Wright, G.D. (2002) 'Glycopeptide antibiotic resistance', *Annual Review of Pharmacology and Toxicology*, 42, pp. 381–408.

Pradel, E. and Schnaitman, C.A. (1991) 'Effect of rfaH (sfrB) and Temperature on Expression of rfa Genes of *Escherichia coli* K-12', *Journal of Bacteriology*, 173(20), pp. 6428–6431.

Pratt, L.A. and Silhavy, T.J. (1996) 'The response regulator SprE controls the stability of RpoS', *Proceedings of the National Academy of Sciences*, 93(6), pp. 2488–2492.

Quan, S., Zhang, N., French, S. and Squires, C.L. (2005) 'Transcriptional polarity in rRNA Operons of *Escherichia coli* nusA and nusB mutant strains', *Journal of Bacteriology*, 187(5), pp. 1632–1638.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rajagopal, S., Sudarsan, N. and Nickerson, K.W. (2002) 'Sodium Dodecyl Sulfate Hypersensitivity of clpP and clpB Mutants of *Escherichia coli*', *Applied and Environmental Microbiology*, 68(8), pp. 4117–4121.

Rajapandi, T., Dolan, K.M. and Oliver, D.B. (1991) 'The First Gene in the *Escherichia coli secA* Operon, Gene X, Encodes a Nonessential Secretory Protein', *Journal of Bacteriology*, 173(22), pp. 7092–7097.

Ramos-Morales, F., Prieto, A.I., Beuzon, C.R., Holden, D.W. and Casadesus, J. (2003) 'Role for salmonella enterica Enterobacterial common antigen in bile resistance and virulence', *Journal of Bacteriology*, 185(17), pp. 5328–5332.

Reddy, M. (2006) 'Role of FtsEX in cell division of *Escherichia coli*: Viability of ftsEX Mutants is dependent on functional SufI or high osmotic strength', *Journal of Bacteriology*, 189(1), pp. 98–108.

Reimer, L.G., Stratton, C.W. and Reller, L.B. (1981) 'Minimum inhibitory and bactericidal concentrations of 44 antimicrobial agents against three standard control strains in broth with and without human serum', *Antimicrobial Agents and Chemotherapy*, 19(6), pp. 1050–1055.

Reznikoff, W.S. (2003) 'Tn5 as a model for understanding DNA transposition', *Molecular Microbiology*, 47(5), pp. 1199–1206.

Rigel, N.W. and Silhavy, T.J. (2012) 'Making a beta-barrel: Assembly of outer membrane proteins in gram-negative bacteria', *Current Opinion in Microbiology*, 15(2), pp. 189–193.

Rinas, U. and Hoffmann, F. (2004) 'Selective leakage of host-cell proteins during high-cell-density cultivation of recombinant and non-recombinant *Escherichia coli*', *Biotechnology Progress*, 20(3), pp. 679–687.

Robinson, C., Matos, C.F.R.O., Beck, D., Ren, C., Lawrence, J., Vasisht, N. and Mendel, S. (2011) 'Transport and proofreading of proteins by the twin-arginine translocation (tat) system in bacteria', *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1808(3), pp. 876–884.

Roncero, C. and Casadaban, M.J. (1992) 'Genetic Analysis of the Genes Involved in Synthesis of the Lipopolysaccharide Core in *Escherichia coli* K-12: Three Operons in the rfa Locus', *Journal of Bacteriology*, 174(10), pp. 3250–3260.

Rosano, G.L. and Ceccarelli, E.A. (2014) 'Recombinant protein expression in *Escherichia coli*: Advances and challenges', *Frontiers in Microbiology*, 5.

Ruiz, N., Falcone, B., Kahne, D. and Silhavy, T.J. (2005) 'Chemical Conditionality', *Cell*, 121(2), pp. 307–317.

Ruiz, N., Kahne, D. and Silhavy, T.J. (2006) 'Advances in understanding bacterial outer-membrane biogenesis', *Nature Reviews Microbiology*, 4(1), pp. 57–66.

Rush, J.S., Rick, P.D. and Waechter, C.J. (1997) 'Polyisoprenyl phosphate specificity of UDP-GlcNAc: Undecaprenyl phosphate N-acetylglucosaminyl 1-P transferase from *E. coli*', *Glycobiology*, 7(2), pp. 315–322.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.. and Barrell, B. (2000) 'Artemis: Sequence visualization and annotation', Bioinformatics, 16(10), pp. 944–945.

Sabourin, D. and Beckwith, J. (1975) 'Deletion of the *Escherichia coli* crp Gene', *Journal of Bacteriology*, 122(1), pp. 338–340.

Saito, M., Katayama, Y., Hishinuma, T., Iwamoto, A., Aiba, Y., Kuwahara-Arai, K., Cui, L., Matsuo, M., Aritaka, N. and Hiramatsu, K. (2014) '"Slow VISA," a novel Phenotype of Vancomycin resistance, found in vitro in heterogeneous Vancomycin-Intermediate staphylococcus aureus strain Mu3', *Antimicrobial Agents and Chemotherapy*, 58(9), pp. 5024–5035.

Sandoval, C.M., Baker, S.L., Jansen, K., Metzner, S.I. and Sousa, M.C. (2011) 'Crystal structure of BamD: An essential component of the β-barrel assembly machinery of gram-negative bacteria', *Journal of Molecular Biology*, 409(3), pp. 348–357.

Sarkar, S.K., Chowdhury, C. and Ghosh, A.S. (2010) 'Deletion of penicillin-binding protein 5 (PBP5) sensitises *Escherichia coli* cells to β-lactam agents', *International Journal of Antimicrobial Agents*, 35(3), pp. 244–249.

Schaub, R.E. and Hayes, C.S. (2010) 'Deletion of the RluD pseudouridine synthase promotes SsrA peptide tagging of ribosomal protein S7', *Molecular Microbiology*, 79(2), pp. 331–341.

Senior, A.E., Nadanaciva, S. and Weber, J. (2002) 'The molecular mechanism of ATP synthesis by F1F0-ATP synthase', *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1553(3), pp. 188–211.

Sezonov, G., Joseleau-Petit, D. and D'Ari, R. (2007) '*Escherichia coli* physiology in Luria-Bertani broth', *Journal of Bacteriology*, 189(23), pp. 8746–8749.

Sham, L.., Butler, E.K., Lebar, M.D., Kahne, D., Bernhardt, T.G. and Ruiz, N. (2014) 'MurJ is the flippase of lipid-linked precursors for peptidoglycan biogenesis', *Science*, 345(6193), pp. 220–222.

Shlaes, D.M., Shlaes, J.H., Davies, J. and Williamson, R. (1989) '*Escherichia coli* susceptible to glycopeptide antibiotics', *Antimicrobial Agents and Chemotherapy*, 33(2), pp. 192–197.

Shu, W., Liu, J., Ji, H. and Lu, M. (2000) 'Core structure of the outer membrane lipoprotein from *Escherichia coli* at 1.9 å resolution', *Journal of Molecular Biology*, 299(4), pp. 1101–1112.

Silhavy, T.J., Kahne, D. and Walker, S. (2010) 'The bacterial cell envelope', *Cold Spring Harbor Perspectives in Biology*, 2(5), pp. a000414–a000414.

Singh, S.K., Parveen, S., SaiSree, L. and Reddy, M. (2015) 'Regulated proteolysis of a cross-link–specific peptidoglycan hydrolase contributes to bacterial morphogenesis', *Proceedings of the National Academy of Sciences*, 112(35), pp. 10956–10961.

Sklar, J.G., Wu, T., Kahne, D. and Silhavy, T.J. (2007a) 'Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in *Escherichia coli*', *Genes & Development*, 21(19), pp. 2473–2484.

Sklar, J.G., Wu, T., Gronenberg, L.S., Malinverni, J.C., Kahne, D. and Silhavy, T.J. (2007b) 'Lipoprotein SmpA is a component of the YaeT complex that assembles outer membrane proteins in *Escherichia coli*', *Proceedings of the National Academy of Sciences*, 104(15), pp. 6400–6405.

Smith, S.G.J., Mahon, V., Lambert, M.A. and Fagan, R.P. (2007) 'A molecular Swiss army knife: OmpA structure, function and expression', *FEMS Microbiology Letters*, 273(1), pp. 1–11.

Snyder, D.S. and McIntosh, T.J. (2000) 'The Lipopolysaccharide barrier: Correlation of antibiotic susceptibility with antibiotic Permeability and fluorescent probe binding Kinetics †', *Biochemistry*, 39(38), pp. 11777–11787.

Sobrero, P. and Valverde, C. (2012) 'The bacterial protein Hfq: Much more than a mere RNA-binding factor', Critical Reviews in Microbiology, 38(4), pp. 276–299.

Stenberg, F., Chovanec, P., Maslen, S.L., Robinson, C.V., Ilag, L.L., von Heijne, G. and Daley, D.O. (2005) 'Protein complexes of the *Escherichia coli* cell envelope', *Journal of Biological Chemistry*, 280(41), pp. 34409–34419.

Sugai, R., Shimizu, H., Nishiyama, K. and Tokuda, H. (2001) 'Overexpression of *yccL* (*gnsA*) and *ydfY* (*gnsB*) increases levels of unsaturated fatty acids and suppresses both the temperature-sensitive *fabA6* mutation and cold-sensitive *secG* null mutation of *Escherichia coli*', *Journal of Bacteriology*, 183(19), pp. 5523–5528.

Sugawara, E. and Nikaido, H. (1992) 'Pore-forming Activity of OmpA Protein of Escherichia coZi', *The Journal of Biological Chemistry*, 267(4), pp. 2507–2511.

Sutcliffe, I.C. (2010) 'A phylum level perspective on bacterial cell envelope architecture', *Trends in Microbiology*, 18(10), pp. 464–470.

Symmons, M.F., Bokma, E., Koronakis, E., Hughes, C. and Koronakis, V. (2009) 'The assembled structure of a complete tripartite bacterial multidrug efflux pump', *Proceedings of the National Academy of Sciences*, 106(17), pp. 7173–7178.

Takeuchi, R., Tamura, T., Nakayashiki, T., Tanaka, Y., Muto, A., Wanner, B.L. and Mori, H. (2014) 'Colony-live —a high-throughput method for measuring microbial colony growth kinetics— reveals diverse growth effects of gene knockouts in *Escherichia coli*', *BMC Microbiology*, 14(1), p. 171.

Tamae, C., Liu, A., Kim, K., Sitz, D., Hong, J., Becket, E., Bui, A., Solaimani, P., Tran, K.P., Yang, H. and Miller, J.H. (2008) 'Determination of antibiotic Hypersensitivity among 4, 000 single-gene-knockout Mutants of *Escherichia coli*', *Journal of Bacteriology*, 190(17), pp. 5981–5988.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. and Tolstoy, I. (2015) 'RefSeq microbial genomes database: New representation and annotation strategy', *Nucleic Acids Research*, 43(7), pp. 3872–3872.

Templin, M.F., Ursinus, A. and Holtje, J. (1999) 'A defect in cell wall recycling triggers autolysis during the stationary growth phase of *Escherichia coli*', *The EMBO Journal*, 18(15), pp. 4108–4117.

The Uniprot Consortium (2014) 'UniProt: A hub for protein information', *Nucleic Acids Research*, 43(D1), pp. D204–D212.

Tikhonova, E.B. and Zgurskaya, H.I. (2004) 'AcrA, AcrB, and TolC of *Escherichia coli* form a stable Intermembrane Multidrug Efflux complex', *Journal of Biological Chemistry*, 279(31), pp. 32116–32124.

Tormo, A., Almiron, M. and Kolter, R. (1990) '*surA*, an *Escherichia coli* Gene Essential for Survival in Stationary Phase', *Journal of Bacteriology*, 172(8), pp. 4339–4347.

Tran, Q.-T., Pearlstein, R.A., Williams, S., Reilly, J., Krucker, T. and Erdemli, G. (2014) 'Structure-kinetic relationship of carbapenem antibacterials permeating through *E. coli* OmpC porin', *Proteins: Structure, Function, and Bioinformatics*, 82(11), pp. 2998–3012.

Turner, R.D., Hurd, A.F., Cadby, A., Hobbs, J.K. and Foster, S.J. (2013) 'Cell wall elongation mode in gram-negative bacteria is determined by peptidoglycan architecture', *Nature Communications*, 4, p. 1496.

Typas, A., Banzhaf, M., van den Berg van Saparoea, B., Verheul, J., Biboy, J., Nichols, R.J., Zietek, M., Beilharz, K., Kannenberg, K., von Rechenberg, M., Breukink, E., den Blaauwen, T., Gross, C.A. and Vollmer, W. (2010) 'Regulation of Peptidoglycan synthesis by outer-membrane proteins', *Cell*, 143(7), pp. 1097–1109.

Typas, A., Banzhaf, M., Gross, C.A. and Vollmer, W. (2011) 'From the regulation of peptidoglycan synthesis to bacterial growth and morphology', *Nature Reviews Microbiology*, 10(2), pp. 123-136.

Uehara, T., Dinh, T. and Bernhardt, T.G. (2009) 'LytM-Domain factors are required for daughter cell separation and rapid Ampicillin-Induced Lysis in *Escherichia coli*', *Journal of Bacteriology*, 191(16), pp. 5094–5107.

Uehara, T., Parzych, K.R., Dinh, T. and Bernhardt, T.G. (2010) 'Daughter cell separation is controlled by cytokinetic ring-activated cell wall hydrolysis', *The EMBO Journal*, 29(8), pp. 1412–1422.

Valvano, M.A., Marolda, C.L., Bittner, M., Glaskin-Clay, M., Simon, T.L. and Klena, J.D. (2000) 'The *rfaE* gene from *Escherichia coli* Encodes a Bifunctional protein involved in Biosynthesis of the Lipopolysaccharide core precursor ADP-L-glycero-D-manno-Heptose', *Journal of Bacteriology*, 182(2), pp. 488–497.

Vollmer, W. and Bertsche, U. (2008) 'Murein (peptidoglycan) structure, architecture and biosynthesis in *Escherichia coli*', *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1778(9), pp. 1714–1734.

Vollmer, W., Blanot, D. and De Pedro, M.A. (2008) 'Peptidoglycan structure and architecture', *FEMS Microbiology Reviews*, 32(2), pp. 149–167.

Vollmer, W., Joris, B., Charlier, P. and Foster, S. (2008) 'Bacterial peptidoglycan (murein) hydrolases', *FEMS Microbiology Reviews*, 32(2), pp. 259–286.

Vorachek-Warren, M.K., Ramirez, S., Cotter, R.J. and Raetz, C.R.H. (2002) 'A triple mutant of *Escherichia coli* lacking secondary Acyl chains on lipid A', *Journal of Biological Chemistry*, 277(16), pp. 14194–14205.

Waldminghaus, T. and Skarstad, K. (2010) 'ChIP on chip: Surprising results are often artifacts', *BMC Genomics*, 11(1), p. 414.

Webber, M.A. and Piddock, L.J.V. (2003) 'The importance of efflux pumps in bacterial antibiotic resistance', *Journal of Antimicrobial Chemotherapy*, 51(1), pp. 9–11.

Weiner, J.H. and Li, L. (2008) 'Proteome of the *Escherichia coli* envelope and technological challenges in membrane proteome analysis', *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1778(9), pp. 1698–1713.

Wilson, C.G.M., Kajander, T. and Regan, L. (2004) 'The crystal structure of NlpI', *FEBS Journal*, 272(1), pp. 166–179.

Wu, T., Malinverni, J., Ruiz, N., Kim, S., Silhavy, T.J. and Kahne, D. (2005) 'Identification of a Multicomponent complex required for outer membrane Biogenesis in *Escherichia coli*', *Cell*, 121(2), pp. 235–245.

Xie, K. and Dalbey, R.E. (2008) 'Inserting proteins into the bacterial cytoplasmic membrane using the Sec and YidC translocases', *Nature Reviews Microbiology*, 6, pp. 234–244.

Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R. and Ishihama, A. (2005) 'Functional characterization in vitro of all Two-component signal transduction systems from *Escherichia coli*', *Journal of Biological Chemistry*, 280(2), pp. 1448–1456.

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. and Mori, H. (2009) 'Update on the Keio collection of *Escherichia coli* single-gene deletion mutants', *Molecular Systems Biology*, 5.

Yanagisawa, T., Sumida, T., Ishii, R., Takemoto, C. and Yokoyama, S. (2010) 'A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P', *Nature Structural & Molecular Biology*, 17(99), pp. 1136–1143.

Yanofsky, C., Horn, V. and Gollnick, P. (1991) 'Physiological Studies of Tryptophan Transport and Tryptophanase Operon Induction in *Escherichia coli*', *Journal of Bacteriology*, 173(19), pp. 6009–6017.

Yao, Z., Davis, R.M., Kishony, R., Kahne, D. and Ruiz, N. (2012) 'Regulation of cell size in response to nutrient availability by fatty acid biosynthesis in *Escherichia coli*', *Proceedings of the National Academy of Sciences*, 109(38), pp. E2561–E2568. doi: 10.1073/pnas.1209742109.

Yethon, J.A. and Whitfield, C. (2000) 'Purification and characterization of WaaP from *Escherichia coli*, a Lipopolysaccharide Kinase essential for outer membrane stability', *Journal of Biological Chemistry*, 276(8), pp. 5498–5504. doi: 10.1074/jbc.m008255200.

Yoshida, M., Muneyuki, E. and Hisabori, T. (2001) 'ATP Synthase - a marvellous rotary engine of the cell', *Nature Reviews Molecular Cell Biology*, 2, pp. 669–677.

Yoshida, Y., Sugiyama, S., Oyamada, T., Yokoyama, K., Kim, S.-K. and Makino, K. (2011) 'Identification of PhoB binding sites of the *yibD* and *ytfK* promoter regions in *Escherichia coli*', *The Journal of Microbiology*, 49(2), pp. 285–289. doi: 10.1007/s12275-011-0360-6.

Yu, B.J., Kang, K.H., Lee, J.H., Sung, B.H., Kim, M.S. and Kim, S.C. (2008) 'Rapid and efficient construction of markerless deletions in the *Escherichia coli* genome', Nucleic Acids Research, 36(14), pp. e84–e84. doi: 10.1093/nar/gkn359.

Yu, E.W., Aires, J.R. and Nikaido, H. (2003) 'AcrB Multidrug Efflux Pump of *Escherichia coli*: Composite Substrate-Binding Cavity of Exceptional Flexibility Generates Its Extremely Wide Substrate Specificity', *Journal of Bacteriology*, 185(19), pp. 5657–5664.

Zhang, R. and Lin, Y. (2009) 'DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes', *Nucleic Acids Research*, 37(Database), pp. D455–D458.

Zhang, Y., Zhang, J., Hoeflich, K.P., Ikura, M., Qing, G. and Inouye, M. (2003) 'MazF Cleaves cellular mRNAs specifically at ACA to block protein synthesis in *Escherichia coli*', *Molecular Cell*, 12(4), pp. 913–923.

Zhou, J. and Rudd, K.E. (2012) 'EcoGene 3.0', *Nucleic Acids Research*, 41(D1), pp. D613–D624.

Zhou, J. and Xu, Z. (2005) 'The structural view of bacterial translocation-specific chaperone SecB: Implications for function', *Molecular Microbiology*, 58(2), pp. 349–357.

Zhou, Y., Gottesman, S., Hoskins, J.R., Maurizi, M.R. and Wickner, S. (2001) 'The RssB response regulator directly targets sigmaS for degradation by ClpXP', *Genes & Development*, 15(5), pp. 627–637.