STREAMLINING MINIMAL BACTERIAL GENOMES

Analysis of the pan bacterial essential genome, and a novel strategy for random genome deletions in *Mycoplasma pneumoniae*

Daniel Shaw

TESI DOCTORAL UPF / 2019

Thesis supervisors

Prof. Luis Serrano Pubul & Dra. Maria Lluch-Senar

EMBL/CRG Systems Biology Research Unit

Centre for Genomic Regulation (CRG)

Universitat Pompeu Fabra Barcelona

Dedication and Acknoledgments

First and foremost, I would like to thank my supervisors Luis Serrano and Maria Lluch-Senar for giving me the opportunity to work in this amazing lab, and undertake this thesis. Your seemingly endless patience and encouragement when things went wrong, which was fairly often, and constant advice and guidance made this thesis possible. Thank you both for your great ideas, helpful criticisms and analytical appraisals in both the design and analysis of the experiments, not to mention the amount of mentorship you were able to provide. No matter where I end up as a scientist, the skills and knowledge I gained from you will be its foundation, and I will be forever grateful.

I would also like to thank the members of my thesis committee; James Sharpe, Fatima Gebauer and Oscar Quijada. The feedback you provided was extremely useful, and our discussions helped me see my projects in a new light.

It would be remiss of me in not mentioning the huge role Carlos Piñero played in the inception through to the outcome of this thesis. Your advice and ideas spawned the original strategies we employed, and while not an official supervisor, you showed me the ropes from day one and were always on hand to bounce ideas off. You were instrumental to the development of the wet lab aspects, and it is not an understatement to say this thesis was made possible in a large part by your help. You also kept my Croatian skills sharp, and your skill at asking grammar questions I couldn't answer was impressive.

Similarly, I owe a huge debt to Toni Hermoso. The bioinformatic services you provided were exemplary, surpassed only by your helpfulness and hard work. I enjoyed our discussions immensely, and you provided a great deal of insight that shaped the project as a whole.

A special thank you should also go out to Jochen Hecht and the CRG Genomics facility. Your knowledge and insight regarding our sequencing dilemmas made our new strategy possible, and the speed and willingness both you and your facility designed and implemented the technique allowed us to collect the data we needed for this thesis to exist.

I must also recognise the huge amount of assistance provided by Samuel Miravet Verde. Your analysis of the sequencing data, along with all the other pieces of extraneous data I asked of you saved me on many occasions. Not to mention that your ability to explain the bioinformatic processes and issues in a way that even a wet-lab Neanderthal like me could understand deserved some form of accreditation by itself.

This thesis was completed due to the funding provided by a FPI-Severo Ochoa Fellowship, and I am incredibly grateful to them for their support.

To the rest of the Serrano Lab, past and present, I offer my heartfelt thanks. For all the help, the advice, the dinners, the parties, the volleyball matches, the lunches, the discussions, the stadium bar soirees, and most importantly the friendships. You took in such an obvious extranjero as me, and made me feel at home here. I honestly couldn't have done it without you, nor would I want to.

There are also a few other rouges and miscreants that deserve a mention. Paul Chammas, Claudia Vivori, Birgit Ritschka, Chris Wyatt and João Frade, congratulations, you made it into the big leagues. Please take comfort that your names will be inscribed forever in one of the great works of the scientific pantheon. In all seriousness however, expressing my thanks for the times we've shared has a ring of finality about it, so instead here's a toast to all the beers, all the burgers, all the game nights, all the laughs and all the memories still to come.

To Moritz Bauer, what can I say? Who would have thought that deciding to live with someone you met briefly for three days could have worked out so well! I hope all the whiskeys, the football, the laughs and the commiserations over your repeatedly terrible life choices makes up for the fact cleaned I cleaned the bathroom about 6 times in the last 4 years... You were the best flatmate a guy could ask for.

Finally, there a four people who deserve special mention. To my family, Lynda, Martin and Jenny, who supported me unconditionally throughout this whole adventure. You helped me out in so many ways over these last 4 years, and I hope some day I can actually explain this who thing to you properly, as you were instrumental to its creation.

And to Elly Crozier, whose constant love and support kept me going, who never once stopped believing in me, and to whom this entire work is dedicated.

ABSTRACT

Understanding what constitutes a true Minimal Cell is a key challenge in synthetic biology. In this work, we present two new tools to aid in this endeavour. i) A novel methodology for minimising the *Mycoplasma pneumoniae* genome via random deletions of genetic material. This protocol utilises the Cre Lox system coupled with random transposon mutagenesis to create a population with random lox sites dispersed around the genome. This allows for a population of cells containing a high variability of large and small-scale deletions ranging from 50bp to 25Kb within *M. pneumoniae.* ii) The first large scale analysis of the essentiality of genes from multiple bacterial species, and how the composition and function of the essential genome of a bacterium changes based on the genome's complexity.

Keywords: Minimal genome, essentiality, Cre Lox, Random deletions, comparative genomics

RESUMEN

Discernir cuales son los componentes que podrían constituir una célula mínima es un desafío clave para la Biología Sintética. En esta tesis, se presentan dos nuevas herramientas para facilitar esta tarea. (i) Una nueva metodología para minimizar el genoma de *Mycoplasma pneumoniae* mediante la deleción aleatoria de material genético. Esta técnica combina el sistema Cre/lox con la mutagénesis aleatoria mediada por transposones para generar poblaciones bacterianas en las que los sitios lox están distribuidos de manera aleatoria a lo largo de su genoma. Esto permite la generación de poblaciones bacterianas en las que el tamaño de las deleciones efectuadas varia desde 50 pb hasta 25 kb. (ii) El primer análisis a gran escala de la esencialidad genética en múltiples especies bacterianas, y cómo la composición y función del grupo de genes esenciales de una bacteria cambia en función de la complejidad de su genoma.

Palabras claves: Genoma mínimo, Esencialidad, Cre/Lox, deleción aleatoria, genómica comparativa

PREFACE

The creation of a true Minimal cell is one of the great challenges of synthetic biology, and methodologies for large-scale genome engineering are coming closer and closer to achieving this goal. However, while our knowledge of genetics has advanced rapidly thanks to landmark advances in DNA sequencing and engineering technologies, it is still far from complete. Bacterial genomes have been sequenced at an exponential pace since *Haemophilus influenzae* was first sequenced in 1995. However, large swathes of genes remain unannotated with their functions unknown. On top of this, novel regulatory systems such as small RNAs and proteins are being revealed, providing further layers of complexity to this already daunting challenge. While rational approaches to create minimal cells have achieved startling successes recently, most notably with the work of the J. Craig Venter Institute and the creation of JCVI-Syn3.0, we are still trying to put together a puzzle for which we not only lack many of the pieces, but are also unsure of exactly how all the pieces fit together.

Regardless, multiple attempts at genome minimisation are underway. These attempts fit into two broad categories, either top down engineering or bottom up engineering. Top down engineering focuses of modifying and minimising pre-existing organisms, removing genes and pathways deemed non-essential to simplify the cell as much as possible. Bottom up engineering on the other hand focuses on building novel organisms from scratch, identifying which pathways and function are essential for a cells survival and attempting to build an organism that contains only the essential functions it needs to survive, and nothing extra.

In this work, we attempt to contribute with tools that can be useful to both schools of thought. With regards to top down engineering, we provide a novel protocol for the genome streamline of the already minimal Mollicutes species *Mycoplasma pneumoniae*. While not a true minimisation technique in its current form, this protocol allows for large scale deletions within the genome, and its repetition could assist in the obtaiing of a minimal cell. The main aspect of novelty to the system is its focus on producing totally random deletions, removing any pre-conceived biases on what genomic regions should or should not be deleted, based on incomplete information regarding genetic essentiality or function. The protocol consists of the addition of lox sites into the genome via transposition, which recombine to create random deletions within the genome. This allows for those cells with the most tolerable deletions to survive, and has the potential to elucidate novel epistatic interactions between genes, based on which regions can and cannot be deleted at the same time.

We also include the first large-scale analysis of the conservation of essential genes across a large and diverse population of bacterial species. We show how the composition of the essential genome of a bacteria changes with regard to the complexity of the genome, and using clustering techniques assign functionality to genes via the COG system. This allows us to elucidate how specific functions change in essentiality as complexity changes, and if there are any specific pathways or genes that become more or less essential as genome complexity changes. We identify which genetic features are highly conserved among even disparate species, and which functions appear to be either essential or not for cellular survival. This information can also be used by bottom up engineering approaches, as it can help identify genes or homologs that are more highly conserved among less complex bacteria, or genes with unknown functions that appear to be non-essential in certain

species, but essential in others. This data can help guide researchers into choosing more appropriate genes to add into circuits, or which areas can be safely removed. Both projects are framed in the systems and synthetic biology fields, through understanding the processes of genome minimisation we hope to be able to rationally engineer minimal cells.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 - INTRODUCTION

1.1. Bacterial Genetics

The genome of an organism contains all of the information needed to create and maintain itself, and it is the main source of heritable information between generations. In bacteria, the genome is not contained within a nucleus like in eukaryotes, but consists of large free-floating chromosomes and smaller plasmids, though in some species the genetic material is semi-segregated in a structure known as the nucleoid, ensuring that the DNA does not occupy the whole cytoplasm (Kleckner et al., 2014). Generally, bacteria exhibit monopartite genome organisation, meaning their genome consists of a single circular chromosome. However, approximately 10% of bacterial species have a multipartite genome (diCenzo and Finan, 2017). This can take the form of either multiple circular chromosomes (Suwanto and Kaplan, 1989), linear chromosomes (Hayakawa et al., 1979) or megaplasmids. These molecules can also display varying characteristics and properties while being retained in the same cell. For instance, they can differ in codon usage, GC content and relative abundances of dinucleotides (diCenzo and Finan, 2017).

Bacterial genomes are generally far smaller than their eukaryotic counterparts. As of 2017, the NCBI database contained 1708 fully sequenced bacterial genomes with an average size of 3.65Mb and a media size of 3.46Mb (diCenzo and Finan, 2017). The largest bacterial species sequenced is currently *Sorangium cellulose*, with a genome of 14.7Mb (Han et al., 2013). In contrast, when looking at eukaryotic cells that exhibit a single-celled lifestyle, their genome sizes tend to be much larger. *Saccharomyces cerevisiae*, a model organism for yeasts has a genome of approximately 12Mb (Ramakrishnan, 2002), and some single celled amoeba can have larger genomes still, with *Acanthamoaeba castellani* containing a 45.1 Mb genome (Clarke et al., 2013) and *Naegleria gruberi* 41Mb (Fritz-Laylin et al., 2010).

To complement their small size in terms of number of base pairs, bacteria have highly compact genomes with protein coding regions comprise on average 88% of the nucleotides present, though this can be as high as 97% (Land et al., 2015). As laid out in the basic dogma of biology, genes in the form of DNA are transcribed into RNA, which are then translated into proteins. These proteins are responsible for the biochemistry of the cell, mediating all enzymatic reactions and creating structures. Other stretches of DNA are transcribed without being translated into proteins, known as non-coding RNAs.

Many of these non-coding RNAs act as regulatory molecules binding to the DNA and activating specific responses. For example non-coding RNA have been found to activate and modulating virulence factors in pathogenic bacteria such as *Salmonella enterica* (Quereda and Cossart, 2017), activate resistance genes when the bacteria enters the presence of certain antibiotic molecules (Dar and Sorek, 2017). These molecules can also act as silencers, recognising specific targets in the mRNA and binding to the ribosome in order to prevent the translation of the target gene (Pfeiffer et al., 2009). Non-coding RNAs are also a key component of the CRISPR-Cas systems, which uses a dedicated guide RNA to target specific loci in the genome for the Cas system to bind to and act upon (Jiang and Doudna, 2017). Others RNAs function in highly essential capacities, such as tRNAs and rRNAs. Those RNAs are key constituents of the translation machinery, either acting as transport vessels to bring the required amino acids to the ribosome (Giegé and Springer,

2016) or as components of the ribosome itself (Nikulin, 2018). They can also affect transcription by interacting with the RNA polymerase, such as is the case with the 10S RNA (Ray and Apirion, 1979) or act as an RNAse, enzymes that degrade RNAs, such as the RNAse E protein that is essential in most bacterial species (Mackie, 2013).

1.1.1. Operons

Multiple genes that share a specific pathway or function are often grouped into single transcription units, known as operons. These operons are groups of genes that share a single transcriptomic regulation, with all components of the operon transcribed at the same time, within a single polycistronic mRNA (Conway et al., 2014). The canonical operon is traditionally the Lac operon from *Escherichia coli*, the first operon to be identified (Jacob and Monod, 1961). The operon regulates the transcription of the three *lac* genes, *lacZ, lacY and lacA*, via the presence or absence of the *lacI* repressor, which itself is regulated by relative levels of glucose and lactose within the cell (Beckwith, 2013; Marbach and Bettenbrock, 2012). This simple yet elegant system allows the cell to activate the three genes it needs to metabolise lactose under a single impetus, and control the levels of activity of all three simultaneously. There are multiple forms of operonic organisation, with a myriad of different regulators (Sáenz-Lahoya et al., 2019), yet the versatility and utility they provide to the cell results in approximately 50% of genes in prokaryotes being grouped into transcriptionally regulated operons (Zhou et al., 2014).

However, while each operon can code for a single polycistronic mRNA, many operons have multiple transcription start sites located within them, capable of creating multiple different mRNAs depending on where the transcription factor binds. This leads to operons being further divided into sub-operons, or "transcriptional units" (Okuda et al., 2007). This can vastly increase the complexity of transcriptional regulation, as not only are the number of potential transcripts increased, but many of them also appear to have regulatory functions. In *Mycoplasma pneumoniae*, a model of a minimal cell, it was found that many of these sub-operons can encode for antisense RNAs of genes within the transcript, which can have a dampening effect of the transcription rate of the targeted gene. In total, 13% of all genes in the *M. pneumoniae* genome contain a potential antisense transcript (Güell et al., 2009a).

The evolution of operons appears to correlate closely with the ability of bacteria to acquire and donate genes between species, known as horizontal gene transfer (HGT). The HGT of operons over genes makes evolutionary sense within the 'selfish gene' theory, as they can be seen as a functional unit, complete with regulation, instead of a single extra protein (Rocha, 2008). This ability of a transferred region to regulate its own expression, along with the acquisition to the cell of a new functionality, drastically lowers the deleterious effects from a misbalance of gene dosages. When a single gene is transferred from one species to another, there is a risk of the new gene being a duplication, in function if not copy, of a previously existing gene. The altered levels of the resultant protein produced from the combination of genes may have deleterious effects on the cell. Indeed, most duplication events that happen naturally are strongly selected against in a population (Hooper and Berg, 2003). However, a self-regulated operon is less likely to be affected by this phenomena, and especially when able to integrate into existing networks, metabolic operons seem to have a high rate of retention when transferred between species (Pál et al., 2005).

1.1.2. Chromosome structure

The structure of the bacterial chromosome also has a regulatory effect on the genes present, as it needs to be highly compacted and organised to fit within the confines of the cell. In *E. coli*, the genome is divided into 40-50 macrodomains, which are generally between 40-100Kb in size (Niki et al., 2000). These domains help with cell division, but also ensure that the same loci inhabit the same area of cytoplasmic space (Badrinarayanan et al., 2015). When the DNA is supercoiled, it is too tightly packed for the RNA polymerase to bind, thus transcription is not possible and all the genes inside are silenced. However when this supercoiling is relaxed, in *M. pneumoniae* operons that are located close to each other spatially are co-expressed to similar levels compared to operons further away in the genome. Chromosome organisation appears to play an important role in bacterial transcription regulation (Güell et al., 2011), not only regulating when modules such as the RNA polymerase can bind to the genome, but also enforcing that as genes with a similar function tend to cluster into operons. Thus, operons of similar functions or in related pathways tend to cluster locally (Junier et al., 2016; Trussart et al., 2017).

1.2. Gene classification

Within bacteria, genes are classified usually as essential or non-essential based on their importance to cellular survival. Essential genes are indispensable to the survival of the cell, and without them the cell is no longer viable. They usually preform roles related to DNA and cellular replication, transcription, translation and core metabolism (Christen et al., 2011). In contrast, non-essential genes encode for functions that are dispensable for cell survival, and can be lost or disrupted without initiating a lethal phenotype (Glass et al., 2006; Jordan et al., 2002). The classification of these genes is normally facilitated via disruption or knock out studies. Individual genes are either removed in a defined and systematic way (Baba et al., 2006), or DNA elements are inserted into the genome at random and disrupt any gene that they land within (French et al., 2008). If the cell can survive and continue to divide after a gene has been disrupted, then that gene is classified as non-essential. If the disruption of the gene is lethal, it is designated as an essential gene.

1.2.1. Quantifying gene functions – The COG classification system

In an attempt to standardise the functions of genes, and make comparisons between genomes easier, the Clusters of Orthologous Genes (COG) database was created. The rationale was that large numbers of genomes were being sequenced, but it was impossible to do functional studies for them all. To try and annotate the genes with functions, the functions of genes from well-studied organisms could be applied to orthologues of those genes in unknown or poorly studied genomes. If in two organisms, the DNA sequence that codes for a protein has a reasonably similar amino acid sequence or conserved domains, it can be inferred that the proteins are orthologues, and thus will undertake the same functions within the cell. These clusters of orthologous genes were split into 26 different categories, allowing for genes to be grouped into specific niches without the classification system becoming too broad to be unable to infer functions, or too narrow to be overcomplicated (Tatusov et al., 2000). Within each category, every specific function is ascribed a COG code. For example, the alanyl-tRNA synthetase is part of the COG category ‘J’ (Translation, ribosomal structure and Biogenesis) and has the COG

code COG0013. Any new organisms sequenced that contain an orthologous sequence to this can reasonably expect to contain that gene, and thus function can be ascribed. As of the most recent update in 2014 (Galperin et al., 2015), the current COG classes are shown in Table 1.

Table 1: COG categories and functions

COG Category	Function
A	RNA processing and modification
B	Chromatin structure and dynamics
C	Energy production and conversion
D	Cell cycle control, cell division and chromosome partitioning
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
G	Carbohydrate transport and metabolism
H	Co-enzyme transport and metabolism
I	Lipid transport and metabolism
J	Translation, ribosome structure and biogenesis
K	Transcription
L	Replication, recombination and repair
M	Cell wall, membrane and envelope biogenesis
N	Cell motility
O	Post-translational modification, protein turnover and chaperones
P	Inorganic ion transport and metabolism
Q	Secondary metabolite biosynthesis, transport and catabolism
R	General function prediction only
S	Function unknown
T	Signal transduction mechanisms
U	Intracellular trafficking, secretion and vesicular transport
V	Defence mechanisms
W	Extracellular structures
X	Mobilome: prophages and transposons
Y	Nuclear structure
Z	Cytoskeleton

1.2.2. Essential functions for a bacterial cell

1.2.2.1. DNA replication and cell division

Arguably the most basic and important capabilities of any lifeform is the ability to replicate itself. Due to this focus, when environmental conditions are appropriate, many bacteria tend to replicate as quickly as possible. The process contains a multitude of highly conserved and essential genes (Koonin, 2003), and generally falls into three stages; DNA replication, chromosome segregation and cytokinesis (den Blaauwen et al., 2017). Generally, bacterial cells contain only one chromosome (Rocha, 2008), and thus only one origin of replication, which is usually flanked by the *dnaA* and often the *dnaN* genes (Wolański et al., 2014). Here, the process starts as the *DnaA* protein binds to the genome and begins the formation of the replisome protein complex, whose key components are listed below in Table 2:

Table 2:Essential genes in DNA replication – Modified from van Eijk et al., (2017)

Gene	Function
Chromosomal replication initiator	Initiates replication of the DNA at the origin of replication
DNA helicase	Unwinds the double stranded DNA at the replication fork
DNA helicase loader	Required for the functional activity of the DNA helicase
Primase	Synthesis of the primers on the lagging strand
DNA polymerase III α	Elongating of the leading and lagging strand during DNA synthesis
DNA polymerase I	Removal of RNA primers and gap filling
DNA Gyrase	Reforming the double stranded DNA after replication
Topoisomerase	Unwinds DNA ahead of the replication fork
DNA Ligase	Ligation of Okazaki fragments in the lagging strand during DNA replication

In bacteria, DNA replication and chromosome segregation occur concurrently. While the DNA is being duplicated, the two chromosomes are segregated, and the two daughter cells formed (Badrinarayanan et al., 2015). In many species, the *parABS* system is used for segregation and partitioning, and are often essential for plasmid maintenance as well as well as cell viability (Gerdes et al., 2010).

1.2.2.2. Gene expression – Transcription

The first stage in the expression of a gene is its transcription, the process of producing mRNA from the DNA template. To initiate transcription, a promoter sequence upstream of the gene of interest needs to be recognised and bound to by the RNA polymerase complex. This complex consists of the two large β subunits (β and β'), two σ subunits and a single small ω subunit. This complex then binds a sigma factor, a multi domain protein able to recognise the promoter sequence and other features in the 5' UTR such as -10 and -35 elements. As shown in Figure 1:Transcription in bacteriaFigure 1, the RNA polymerase complex, with attached sigma factor, binds to the promoter region of the DNA. The double-stranded DNA is then opened to allow access to the single strands. After transcription is initiated, the sigma factor detaches from the complex, and the remaining RNA polymerase complex proceeds along the DNA until a transcription termination sequence is encountered, and the RNA polymerase terminates transcription, releasing the newly formed mRNA and detaches from the DNA (Browning and Busby, 2016). The resultant single stranded mRNA is then released. Coding mRNAs will be recognised by free floating ribosomes in the cell's cytoplasm and be translated, while non-coding RNAs will bind to their targets of interest.

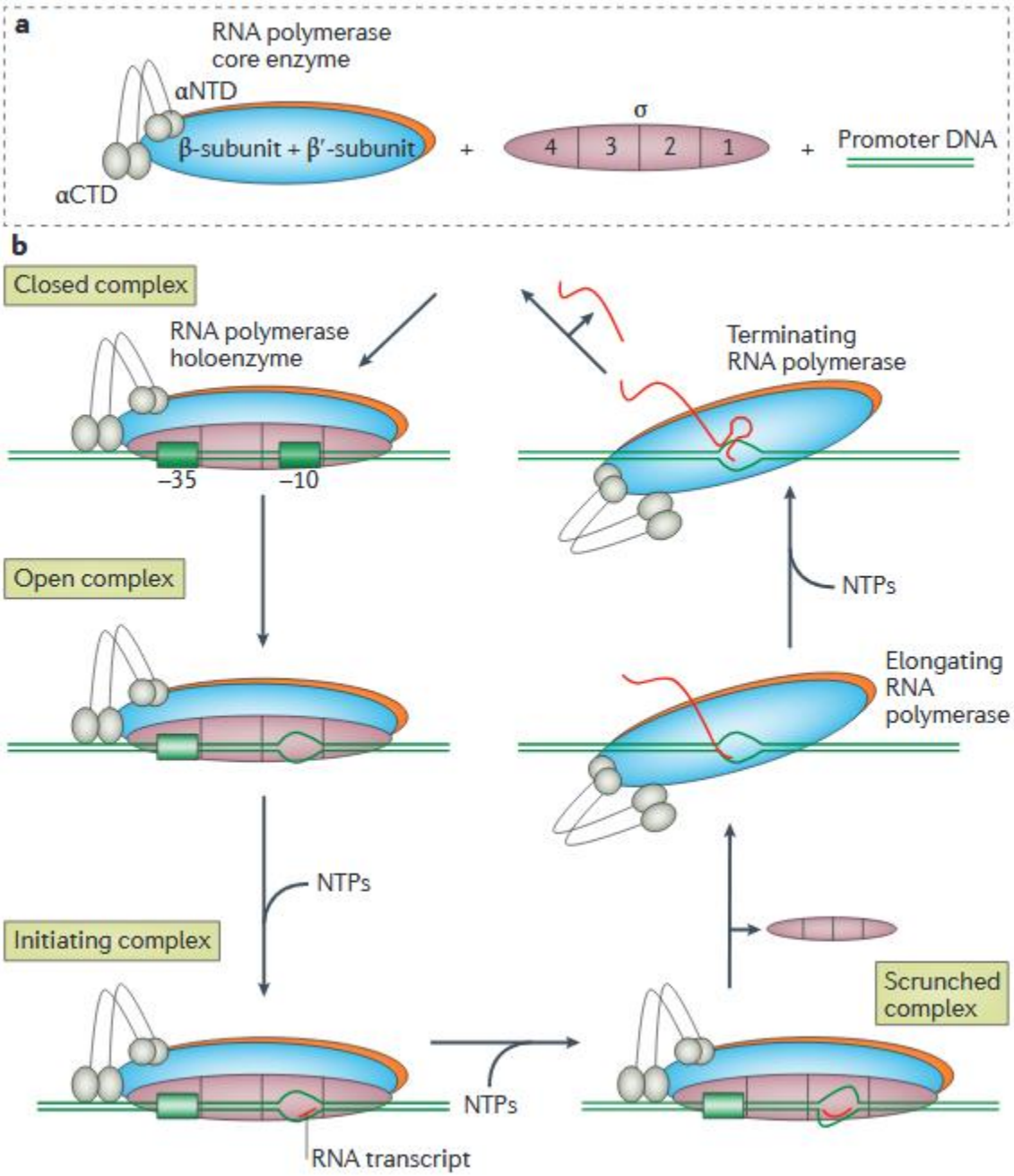


Figure 1:Transcription in bacteria. A: The RNA polymerase complex and structure of the sigma factor. B: The bacterial transcription cycle. RNA polymerase holoenzyme, which comprises the RNA polymerase core enzyme and a sigma factor, interacts with promoter DNA to form the closed complex. The closed complex transitions to the open complex by unwinding the DNA duplex in the region of the transcription start site. The addition of nucleoside triphosphates (NTPs) enables a further transition to the initiating complex, which synthesizes the RNA transcript. Initially, the template strand of the DNA is pulled into the initiating complex, which is a process known as 'scrunching'. The scrunched complex can be held at the promoter, which results in cycles of abortive initiation that only produce small RNA fragments. Alternatively, the RNA polymerase can escape the promoter to enter the elongation phase, leading to the release of the sigma factor and elongation of the RNA transcript using NTPs and elongation factors (not shown). Transcription proceeds until the RNA polymerase encounters a transcriptional terminator, after which the RNA transcript is released and the polymerase dissociates from the DNA template to re-engage with a sigma factor and repeat the cycle. Adapted from Browning and Busby, (2016).

1.2.2.3. Translation

After the mRNA has been transcribed, it is translated into the requisite protein structure using the free-floating ribosomes in the bacterial cytoplasm. The bacterial ribosome is slightly different from its eukaryotic and archaeal counterparts, though of the approximately fifty unique ribosomal proteins found in bacteria, 34 are universally conserved across all domains of life (Yutin et al., 2012). It contains two major subunits, the 30S subunit which is responsible for mRNA binding and initiation, and the 50S subunit which is responsible for tRNA accumulation and elongation. Together, they form a complete 70S bacterial ribosome.

Translation begins when the mRNA binds to the dissociated 30S subunit of the ribosome, in the presence of an initiation factor, usually *infA* (Translation initiation factor IF-1). The start codon is recognised by the CAU anticodon in the P site of the 30S subunit and bound there. The larger 50S subunit then binds and the initiation factor is released (Gualerzi and Pon, 2015). Elongation factors then bind to the 70S ribosome, and allows the integration of an aminoacylated tRNA to enter the A (acceptor) site in the ribosome complex. Here, the anti-codon on the tRNA attempts to bind to the codon of the mRNA, and if successful, the charged amino acid of the tRNA is added to the polypeptide chain. The complex is then moved to the P site to allow the next tRNA to bind in the A site, and the process cycles, and step adding a new amino acid to the polypeptide chain (Ramakrishnan, 2002). After the mRNA has been read and translated, the process is terminated via the reading of a stop codon at the end of the protein coding region of mRNA. Release factors then bind to the ribosome and terminate transcription, generally release factors 1 and 2, releasing the polypeptide chain for folding, and dissociating the mRNA from the ribosome (Baggett et al., 2017).

1.2.2.4. Metabolism

Bacterial cells require multiple metabolic pathways for survival. While certain obligate intracellular parasites and mutualists can take advantage of their hosts biosynthetic pathways to avoid the need to produce their own metabolic substrates (Zientz et al., 2004), most axenic species need to be able to correctly metabolise the needed lipids, carbohydrates, nucleotides, co-enzymes, amino acids for their survival. Metabolism is also the key driver in providing energy for all cellular processes, specifically ATP, and enzymes related to both substrate level phosphorylation and oxidative phosphorylation are well conserved (Chubukov et al., 2014). Despite the vast differences in environment and niche across the bacterial Domain, the fundamental processes and pathways involved are fairly well conserved across all species, as illustrated in Figure 2.

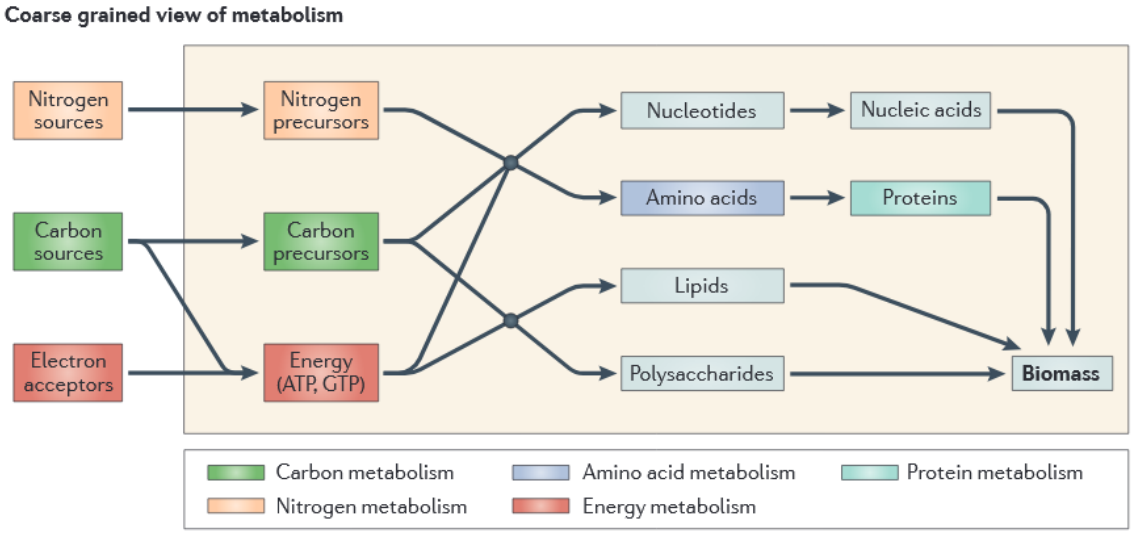


Figure 2: Metabolic tasks in bacteria. Coarse-grained view of different sectors that compose large parts of metabolism in many bacteria. Microorganisms need to carry out a range of metabolic tasks to ensure a supply of metabolic fluxes through the sectors and thus sustain cell maintenance and growth. All organisms must regulate the uptake of nutrients and coordinate carbon, energy and nitrogen metabolism to balance monomer synthesis and macromolecule polymerization. Modified from Chubukov et al., (2014).

The ability to import carbon and nitrogen sources for biosynthesis, and some form of electron acceptor for energy production, is universal. However, the means and molecules used are as varied as the niches supporting the bacteria, ranging from obligate iron-oxidising lithoautotrophs (Summers et al., 2013) to bacteria fixing nitrogen from the air (Mus et al., 2016) to bacteria that can obtain carbon from digesting poly(ethylene terephthalate) (Tanasupawat et al., 2016).

1.2.2.5. Cell wall and membrane biogenesis

The creation and maintenance of a functioning cell wall is essential for the survival of a bacteria, not only as it ensures there is a cell present at all, but it allows for the transport of metabolites, defence against environmental stress and interaction with the environment (Cho et al., 2016). While there are many differences in composition of bacterial cell walls, the divide between gram positive and gram negative bacteria being the most obvious. Almost all species of bacteria contain a cell wall which is composed (at least in part) by peptidoglycan (Errington, 2013), with notable exceptions of the Mollicutes (Trachtenberg, 2005). However, mechanisms of cell wall biogenesis are not strongly conserved, with no individual gene responsible for the creation of the structures appearing in large genomic studies of multiple species of bacteria (Charlebois and Doolittle, 2004; Koonin, 2003, 2000). Therefore, the dependency of the bacterium to adapt its cell wall makeup to its environment and the metabolites present have ensured the convergent evolution of multiple cell membrane biogenesis pathways (Ruiz et al., 2006).

1.2.2.6. Homeostasis

The ability of cells to regulate their internal homeostasis is vital to their survival. There are multiple pathways that are highly conserved across the bacterial domain that ensure that basic cellular functions are not perturbed and the cell can function normally.

One of the key mechanisms to ensure that cellular processes run as intended are the chaperone proteins. These chaperone proteins regulate the cells proteins by folding freshly translated amino acid chains, identifying miss-folded proteins and preventing

aggregation of proteins that still need to provide a function (Santra et al., 2017). One of the most ubiquitous bacterial chaperones is the *groEL/groES* family. It is highly conserved and generally essential across the bacterial domain, and is responsible for the folding of newly translated proteins into their functional forms (Endo and Kurusu, 2007).

Another vital homeostatic network present in bacteria is ribosome rescue after stalling. Ribosome stalling occurs when the ribosome cannot continue with the translation process and is stalled on the mRNA, preventing the translation of gene by it or any further ribosomes that may be bound upstream of the ribosome. This is most commonly due to a lack of the relevant tRNA needed to continue with the translation process, though the presence of truncated transcripts is also a common cause of translation stalling (Buskirk and Green, 2017). As this is both a common and fundamental error involved in the translation process, almost all bacteria utilise a common recovery mechanism. This is mediated by ribosome rescue factors, typically alternative ribosome rescue factor A (*arfA*) or B (*arfB).*

Maintaining an optimal level of NAD is another key homeostatic process for bacteria. The DNA repair mechanisms of the cell are generally dependant on a ready source of NAD to supply the DNA ligase with an adenosine mono-phosphate (AMP) molecule to bind to the 5' phosphate of the nicked strand, allowing the ligase to repair the DNA break. There are also deacetylates such as the *sir2* family of proteins that modulate some aspects of DNA repair and gene silencing that are NAD dependant (Sorci et al., 2014). While there is no universally conserved single NAD synthesis or uptake mechanism, all bacteria do contain some pathway allowing for the generation or procurement of NAD (Gazzaniga et al., 2009).

1.3. Tools to study essentiality in bacteria

1.3.1. Directed Mutagenesis

By manually perturbing or removing a gene from a cell, then observing the resultant phenotype, it is possible elucidate which genes are essential to a cells function and which are not. There are multiple ways of achieving this, either by disrupting the gene or removing it entirely so it cannot be expressed, or by silencing the expression of the gene so no protein is produced.

1.3.1.1. Recombineering

Recombineering is a technique that exploits the ability many bacteria have to repair damaged DNA and take up new segments of genetic information, known as homologous recombination (Pines et al., 2015). Using this system, targeted deletions can be made within the genome of a bacteria, by inserting a new section of DNA, often a selective cassette in the place of a gene by double crossover.

One of the most famous examples of this is the Keio collection, where researchers attempted to knock out every single gene in the *E. coli* K-12 genome. This was done by systematically designing oligos that amplified a kanamycin resistance gene with a 50 nucleotide homology region for the genes directly upstream and downstream of the gene of interest. This cassette would then get integrated into the native *E. coli* genome via homologous recombination. This can theoretically allow the deletion of any region of the genome, thus enabling systematic testing of the how essential the region of the genome

is. Using this system, the researchers identified 303 genes that they could not remove, thus these genes were designated as the essential genes within the *E. coli* K12 genome (Baba et al., 2006).

1.3.2. Random mutagenesis

1.3.2.1. Random Transposon mutagenesis

One of the most common methodologies of determining essential genes is via random transposon mutagenesis. Transposons are mobile DNA elements, consisting of the transposase gene and a transposon cargo, which is inserted into the host genome. Transposons are nearly ubiquitous to all currently sequenced genomes, both prokaryotic and eukaryotic, though their activity is highly variable. In prokaryotes, they appear to be a primary source of genetic rearrangement, and their effects can have large functional effects on the host cell. As such, it is hypothesised that despite their potential to cause large scale genetic disruption, this may also lead to beneficial changes depending on the organisms circumstances, and thus an evolutionary advantage (Hickman and Dyda, 2016). Transposons are broadly assigned into two different classes, depending on the nucleic acid used in the intermediate step. Class 1 transposons, which use an RNA intermediate, are known as retrotransposons. Class 2 transposons, by contrast, use a DNA intermediate and thus are known as DNA transposons (Babakhani and Oloomi, 2018).

1.3.2.2. Retrotransposons

Retrotransposons are most commonly found in eukaryotes, though they appear to have evolved originally in prokaryotes. The typical bacterial retrotransposons used as mobile DNA elements are the group II introns, although types of retrotransposon groups such as diversity generating retroelements and retrons have also been documented (Zimmerly and Wu, 2015). As the name implies, these enzymes use a reverse transcriptase to convert RNA into DNA.

For the standard type II intron, the mechanism of action consists of two main parts. First, a catalytic self-splicing RNA, often under 1Kb, and an intron encoded protein (IEP), are combined as a single operon. When the intron is transcribed, the IEP is then translated. The intron RNA then folds into a tertiary structure than in amenable for splicing. The IEP then binds to the folded RNA and splices out the exons, leaving the IEP bound to the intron lariat, forming a stable ribonuclearprotein (RNP). This complex then inserts the intron into a new random location within the host genome. This is done via cutting the genome and ligating the leading strand. The enzyme then utilises its intrinsic reverse transcriptase activity to reverse transcribe the RNA into DNA. When complete, it relies on the hosts cellular repair mechanisms to ligate the new dsDNA with the genomic DNA (Zimmerly and Wu, 2015). These act as a ‘copy and paste’ mechanism, with the original template DNA retained in the genome and copies of it are transcribed and inserted, often multiple times, allowing the intron to proliferate throughout the genome.

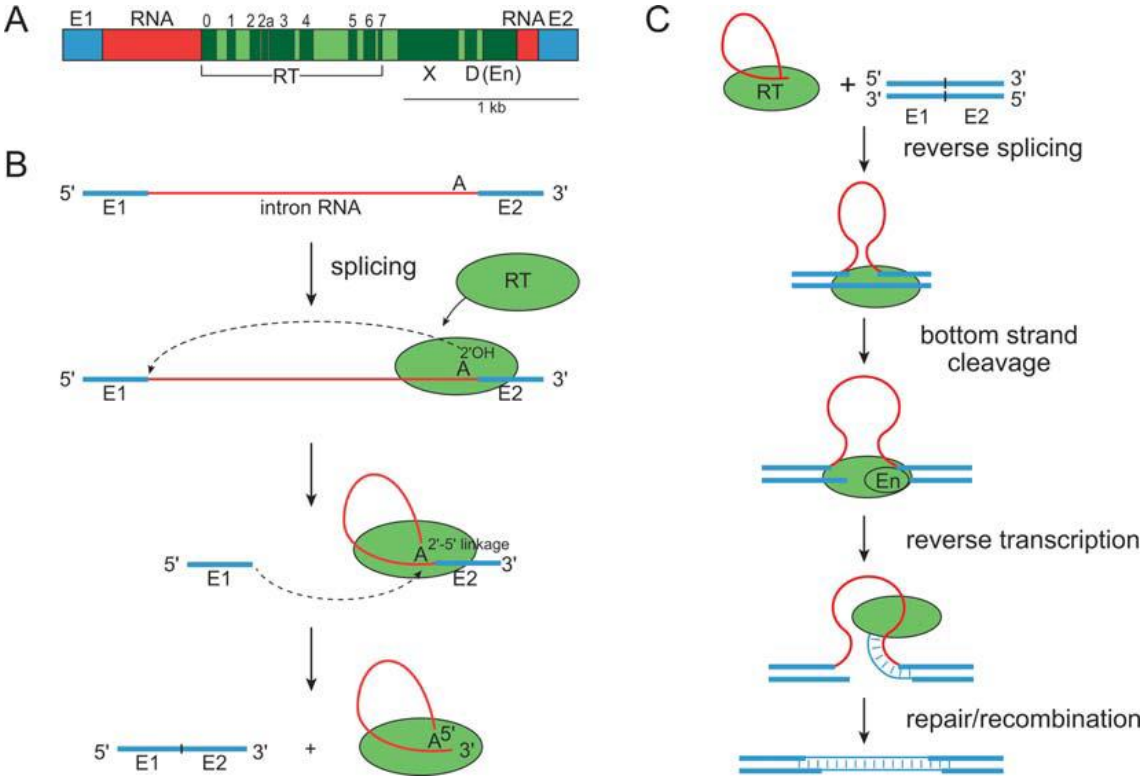


Figure 3: Group II introns. *(A) Genomic structure of a group II intron consists of a sequence for an RNA structure (500-800 bp; red boxes) and an ORF for an intron-encoded protein (green). The protein contains a reverse transcriptase domain (RT) with motifs 0 to 7, and X/thumb domain, a DNA-binding domain (D), and sometimes, an endonuclease domain (EN). The intron is flanked by exons E1 and E2 (blue). (B) After transcription, the intron-encoded protein is translated from the unspliced transcript and binds to the RNA structure to facilitate a two-step splicing reaction, yielding spliced exons and an RNP consisting of the RT and intron lariat RNA. (C) The RNP inserts the intron sequence into new genomic target. To do this, the RNP binds to the double-stranded DNA target, the intron lariat reverse splices into the top strand, and the En domain cleaves the bottom strand to produce a primer that is reverse transcribed by the RT. Cellular repair activates covert the insertion product to dsDNA. Taken from Zimmerly and Wu, (2015).*

Diversity generating retro-elements have been shown to be key factors in generating diversity within bacterial lineages. They work in a similar manner to group II introns, however the reverse transcriptase they contain is typically highly error prone. This lack of fidelity leads to numerous point mutations within defined locations at the target gene of the retrotransposon, allowing for the generation of large number of diverse mutant proteins. As these elements are originally phage based, they were originally targeted towards membrane proteins on the cell wall, which in turn allowed phages to bind to them more easily (Guo et al., 2014; Paul et al., 2017).

1.3.2.3. DNA Transposons

Class 2 transposons, also known as DNA transposons, rely on the physical relocation of the DNA cargo, instead of copying it and pasting it elsewhere, and are thus sometimes known as “cut and paste” transposons as a result. These systems typically include a transposase protein, a gene or genes of interest (often antibiotic resistance markers) and a pair of inverted repeats, that flank the region and that are recognised by the transposase.

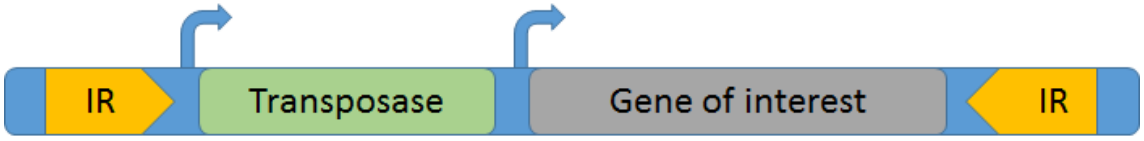


Figure 4: Traditional composition of a DNA transposon, with endogenous transposase and gene of interest flanked by inverted repeats.

For the transposon to act, the transposase protein is translated and binds to an inverted repeat, recognising it via their helix-turn-helix motif. This interaction forms a single-end complex (SEC). One transposase molecule is bound at each end of the transposon, one on each inverted repeat, forming two separate SECs. Both transposases then cleave the 5' end of the inverted repeat via hydrolysis of the phosphodiester bond, freeing the 5' strand. This excises any DNA that was stored between the two inverted repeats. The two SECs then draw together, binging the two opposite ends of the transposon together, and forming the dimer paired end complex, and hydrolyses the 3' end of the DNA strands. The PEC moves through the cell, and binds to the genome in a random location, forming a target capture complex. The 5' end of the host DNA is attacked by the 3'-OH of the inverted repeats via nucleophilic attack, leading to DNA breakage and the integration of the 3' strands of the transposon into the host genome. The 5' gaps are then filled, either by the activity of the transposase or by host ligase activity (Babakhani and Oloomi, 2018; Muñoz-López and García-Pérez, 2010).

The integration of the transposon into the host DNA relies on the ability of the paired end complex to cut the genome at a site specific to its species, with many different species having multiple different preferred sites of integration. For example, the Tc1/*mariner* complex can integrate at any AT site in the genome (Plasterk et al., 1999), whereas the Tn7 transposon is highly site specific, and will only insert in high frequency at the *attTn7* site downstream of the *glmS* gene. While low levels of insertion have been identified in sites with a similar composition to the *attTn7* site, it is still highly target specific (Craig, 1991).

The insertion of the transposon, using strand exchange, is mediated via transesterification reactions. The 3'-OH groups on the opposite strands of the transposon serving as nucleophiles to break the phosphodiester bonds in the DNA backbone, however the size and conformation of the different transposases means different points on the chromosome are attacked. This results in target site duplications (TSDs) being generated either side of the insertion point, though their size can be slightly different depending on the species of transposon used (Muñoz-López and García-Pérez, 2010). Figure 5 shows the mechanisms outlined for the integration of the transposon DNA, and the location of the TSDs. (Hickman and Dyda, 2016).

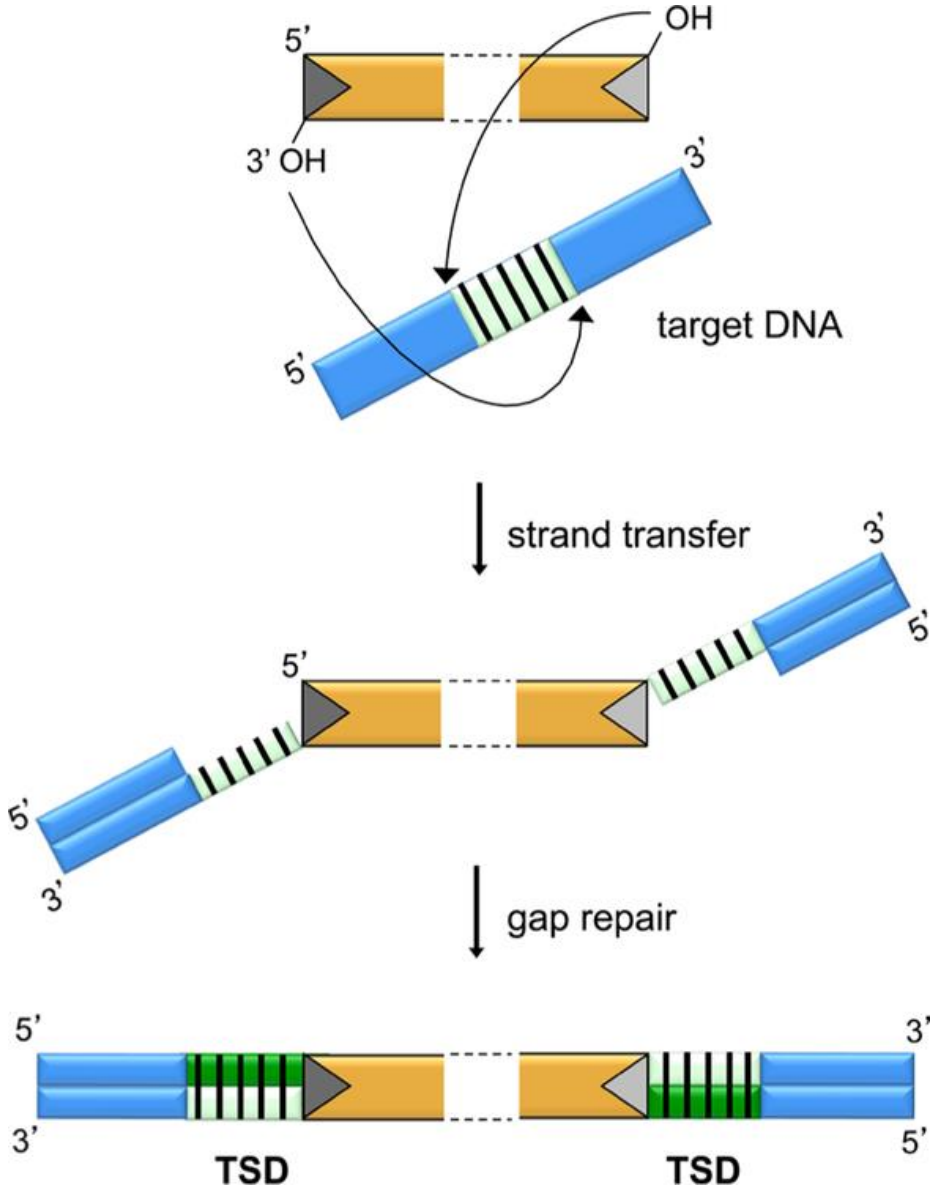


Figure 5: *Class II transposon integration. Generation of target site duplications (TSDs) upon staggered strand transfer of excised transposon DNA integrating into the host genome. Taken from Hickman and Dyda, (2016).*

While transposons are highly useful for manipulating bacterial genomes, their endogenous forms make them less suitable for the generation of long-term stable mutants. The presence of the active transposase gene means transposition could occur at any time, potentially re-activating any genes that were previously silenced, or killing the cell by transposing into a new, essential region. This lead to the development of a *de novo* class of DNA transposons known as mini-transposons. Instead of containing the transposase gene within the transposon, it is instead encoded outside the inverted repeats. This allows the transposase to still act upon the transposon, however the newly inserted DNA does not contain the transposase gene. As such, it becomes silent and cannot move again, leading to a stable insertion over multiple bacterial passages (Christie-Oleza et al., 2013; de Lorenzo and Timmis, 1994).

1.3.3. Tn5

The Tn5 system was one of the first transposases to be characterised in bacteria (Berg et al., 1975), and one of the most commonly used transposons for work in bacterial systems. It is a class II DNA transposon, comprising of two inverted sequences named *IS50L* and *IS50R.* These in turn are flanked by almost identical 19-base pair inverted repeats, OE (outside edge) and IE (inside edge). The *IS50R* region contains the transposase and an inhibitor factor, while *IR50L* encodes for three antibiotic resistance genes, conferring resistance against kanamycin, bleomycin and streptomycin respectively. The *IR50L* also contained inactive copies of the transposase and inhibitor genes, which have been truncated in the C-terminal. The inhibition factor is itself an inactivated version of the transposase gene which is unable to bind to the DNA, due to lacking the initial 55 amino acid residues of the gene. It inhibits the transposase by binding to the gene and forming heterodimeric complexes (Reznikoff, 2008).

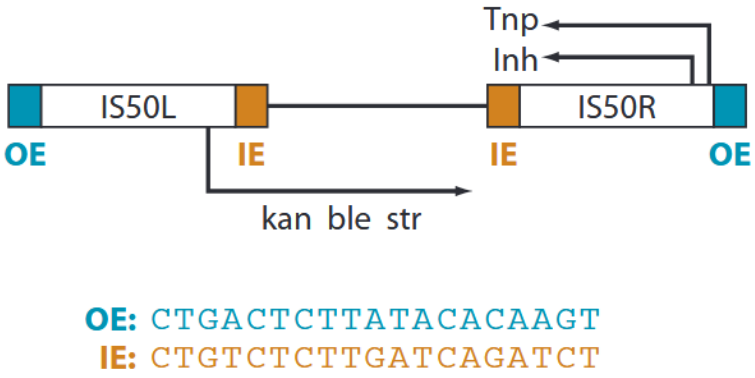


Figure 6: *Structure of Tn5 Transposon. The orientation and layout of the two elements of the Tn5 transposon system; the transposase (Tnp) and inhibition factor (inh), along with the sequence of the inverted repeats. Adapted from Reznikoff, (2008).*

The system in highly inefficient, due to the proximity of the N and C termini in the final protein. Coupled with the effect of the inhibition factor present, the wild type system is highly inactive *in vivo* and totally inactive *in vitro* (Steiniger-White et al., 2004). Indeed, the frequency of Tn5 transposition in *E. coli* is less than 10^{-5} events per cell per generation. This is most likely an evolutionary desirable trait, as too much activity of the transposon could easily initiate lethal phenotypes by often disrupting essential functions, however low level genetic alteration could provide the occasional fitness advantage (Reznikoff, 2008). Mutations in the transposase protein have allowed the efficiency to be raised dramatically. The LP372 and EK345 mutations when applied together raise the efficiency of the system 80-fold. Combined, they form a break in one of the alpha helixes near the C-terminal, allowing the C and N termini to separate, and also block the binding of the inhibition factor, allowing the transposase proteins to form the required homodimers to recognise and excise the OE sequences (Weinreich et al., 1994).

Mini transposon variants of the Tn5 have also been developed (de Lorenzo et al., 1990), allowing for more stable genetic engineering. This inherent stability, along with further mutations increasing the efficacy of the transposase has allowed for the development of hyperactive Tn5 mutants that are effective across not only multiple bacterial species (Lyell et al., 2008; Naorem et al., 2018; Watabe et al., 2014), but also in plant (Wu et al., 2011) and animal (Bright and Veenstra, 2019) species. This ability of hyperactive Tn5 transposases to be able to insert into such a wide range of host genomes has led it to

becoming the main biological component in ATAC-Seq, a protocol used to identify accessible chromatin in multiple eukaryotic species (Buenrostro et al., 2015).

1.3.4. Tc1/Mariner

The Tc1/Mariner superfamily is a class of type II DNA transposons that are almost ubiquitous in nature (Plasterk et al., 1999), named for the transposons of the class first characterised, Tc1 in *Caenorhabditis elegans* (Emmons et al., 1983) and Mariner in *Drosophila mauritiana* (Jacobson et al., 1986). This class of transposons has been extensively well used and studied, and is active across almost all forms of life, such as humans (Robertson and Zumpano, 1997), insects (Coates et al., 1997), protozoa (Gueiros-Filho and Beverley, 1997), bacteria (Cassier-Chauvat et al., 1997), yeast (Zhou et al., 2017) and plants (Jacobs et al., 2004). While there are a myriad of different transposon species in the superfamily, they all share a basic structure. Generally 1300-2400bp in length, they contain a single transposase gene which recognises a single pair of inverted repeats, which can be between 31 and 462bp. The natural form of these transposons, as shown in Figure 7, is as standard transposons, though mini transposon variants have been created (Zhang et al., 2000).

As DNA transposons, they integrate DNA in the same method as described in chapter 2.2.1.2. However, they can be identified by the unique target site duplications that they leave, both in the donor section of DNA and in the ligation site. The inverted repeat regions always contain a TACA motif at their boundary, with the AT inside the genomic DNA and the CA at the 5' of the inverted repeat. As shown in Figure 7, the transposon can only integrate at TA sites inside the genome. The overhangs left by the excision of the transposon cause a TACATA motif to form where the DNA was excised, and the newly integrated transposon contains the TACA motif again at the 5' ends, allowing for recognition again by the transposase (Plasterk et al., 1999).

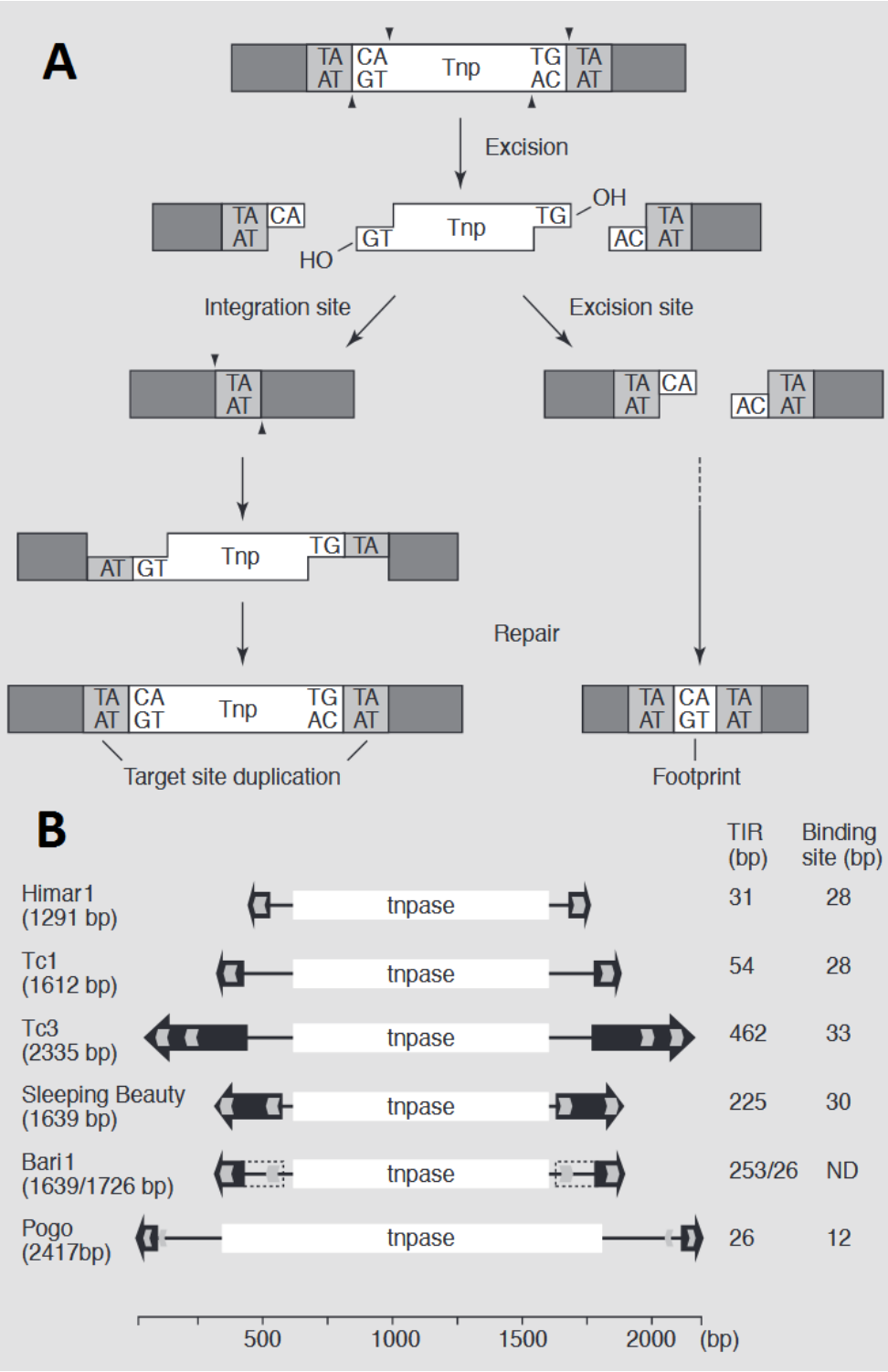


Figure 7: *Tc1/Mariner model. (A) The Tc3 element is excised by transposase-mediated double-stranded breaks at the ends of the inverted repeats. The DNA cut is staggered, which generates single-stranded transposon termini of two overhanging nucleotides with reactive 39-hydroxyl groups (OH) and leaves two nucleotides of the transposon ends at the site of excision. Some other Tc1/mariner elements probably excise via a 3 bp staggered cut. The excised element integrates into a TA dinucleotide site in the target DNA. During integration, another staggered double-stranded DNA break is introduced by the incoming transposon at the TA target site, so that the TA will be duplicated and flank the inserted element after the single-stranded gap in the DNA is sealed by cellular repair processes. The excision site is also subject to DNA repair that can, in some cases, regenerate the terminal nucleotides of the transposon inverted repeats left in the gap, resulting in transposon footprints. (B) The central transposase genes (tnpase) are flanked by terminal inverted repeats (TIR; black arrows) that contain binding sites for the transposase. TIRs come in different lengths and contain binding sites in different numbers and patterns in the Tc1/mariner superfamily. Dotted lines in Bari elements indicate that certain versions of these transposons have long inverted repeats. Actual or putative transposase-binding sites are indicated as grey arrows near the ends of the elements. Figure modified from Plasterk et al., (1999).*

1.4. Analysis of high throughput transposon essentiality studies

1.4.1. Tn-Seq and HITS protocols for analysing transposon data

Transposon mutagenesis is one of the most popular methods for determining gene essentialities, due to the reasons outlined in chapter 1.3.2.1. When coupled with next generation sequencing, it can provide a wealth of data not only on which genes can be disrupted, but also at what frequency compared to the rest of the genome. One of the most common methodologies used to analyse transposon data coupled with next generation sequencing is Tn-Seq (van Opijnen et al., 2009). This protocol uses genomic DNA isolated from a pool of cells transformed with a transposon under a condition of interest (the mariner derived transposon Himar I in the original paper). The DNA is then fragmented using the *MmeI* restriction enzyme. *MmeI* is a type IIS restriction endonuclease that recognises a specific site found within the inverted repeat of the Himar I transposon, and cuts 20bp downstream of the site (Morgan et al., 2009). Adapters are then ligated to the DNA at the cut sites, and DNA primers that bind to the adapter sequence and the inverted repeat are used to amplify the region of DNA between inverted repeat and the adapter. The pool of amplified DNA is sequenced via ultra-sequencing, giving the insertion location of each of the transposons. The protocol also quantifies the number of reads from each location. This allows the identification of areas that are enriched in transposon insertions compared to the overall level, and those areas that are lacking in insertions. By running this analysis on transposon libraries that have been grown in different conditions, it allows for identification of genes that change essentiality based on the growth parameters. Tn-seq also allows for a metric of comparison, by looking at the relative increase or decrease in the number of transposon insertions in a specific region across different conditions (van Opijnen et al., 2009).

A second method, one that did not rely on the usage of the *MmeI* restriction enzyme intrinsic to the Himar I transposon, was developed known as high-throughput insertion tracking by deep sequencing (HITS) (Gawronski et al., 2009). The protocol is similar to that of Tn-Seq, however it uses random DNA shearing instead of restriction digest to break up the DNA. Once the DNA is fragmented, adapters are ligated. DNA primers are then annealed, one in the adapter region that contains the recognition site for the ultra-sequencing primer, and one in the transposon region, which is biotinylated. The PCR is run and fragments are size-selected to allow for efficient sequencing read lengths. PCR fragments containing the biotin are then purified to ensure only regions containing transposon DNA are present, then the DNA is sent to sequence. As with Tn-Seq, this allows for both the location and read density of each transposon to be mapped (Gawronski et al., 2009).

For side by side comparison, both methods are illustrated in Figure 8, along with similar methodologies based on each.

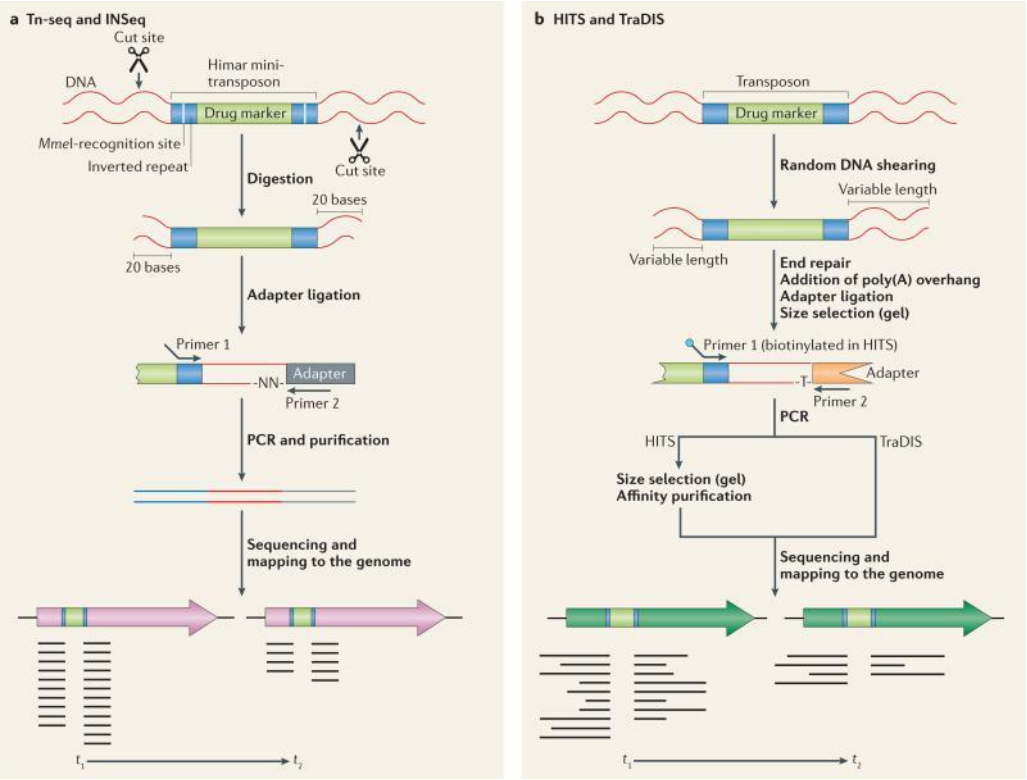


Figure 8: Comparison of Tn-Seq and HITS based transposon sequencing protocols. Taken from (van Opijnen and Camilli, 2013)

1.4.2. Statistical analysis of essentiality

As stated previously, determining if a cell can tolerate the insertion of a transposon inside a specific gene is the benchmark for deciding whether a gene is essential or not. However, given the high sensitivity of next generation sequencing protocols, statistical analysis of the results is a necessity as blanket black-or-white statements about what can and cannot be deemed essential are insufficient (DeJesus et al., 2013; Deng et al., 2013; Liu et al., 2015). As there are multiple confounding factors in these studied (as explained further in the chapter), often a set of 'gold standards', specific genes known to be highly essential or non-essential, are used to help classify the rest of the genes. The specific gold standard used can vary, either the knowledge of the organism is sufficient to ascribe the most or least essential genes to an organism specific list (Lluch-Senar et al., 2015b), or a previously annotated gold standard is taken from a different but related organism and used as a benchmark for analysis (Freed et al., 2016). These standards help set the boundaries for dealing with experimental noise, as it is possible that the genetic material from dead cells is related in the sample and sequenced, showing insertions in essential regions. By analysing the sequence of known highly essential genes, a noise threshold can be established to help overcome spurious results.

1.4.3. Fitness genes

To further complicate the issue of if a gene is essential or not, a third class of genes is becoming more widely reported, known as either "quasi essential" (Hutchison et al., 2016) or "fitness" (Lluch-Senar et al., 2015b). This class of genes will be referred to as fitness genes from hereon in, as that is the most common phrase found in the literature. These genes are technically non-essential, as their loss does not impart a lethal phenotype

upon the cell, however disruption of them does cause significant negative phenotypic effects, traditionally as slower growth or less robust metabolism. These genes are often associated with housekeeping function (Hutchison et al., 2016). As such, it would be valid to view all genes as fitness genes, just at various degrees of effect. An essential gene would have a 100% fitness cost when lost, whereas many non-essential genes only have a small if even negligible fitness cost when lost. As every gene encodes for at least one function with its protein, they must contribute something to the overall fitness of the cell (Juhas et al., 2011a; Koonin, 2000).

1.4.4. Confounding factors in essentiality studies using transposons

1.4.4.1. Location and density of transposons

High throughput omics data on the essentiality of genes in bacteria has become readily available over the last decade, in no small part due to the proliferation of next generation sequencing technologies (Land et al., 2015). Due to this increase in data, much deeper understanding of the nature of which genes are essential and to what extent they are essential has appeared. Due to the random nature of transposon mutagenesis, and the fact that it is by definition a negative screening test for essential genes, there is always the possibility for false positive or false negative results to appear. There are multiple confounding factors as to why the annotated essentiality of a gene may be questionable (Juhas et al., 2011a; Liu et al., 2015).

If the number if transposon mutants generated is too low, then it is difficult to say if the genes with no corresponding insertion are truly essential, or that they were just missed by chance, thereby creating false positive results. This is also a factor for gene length, as if the transposon coverage is too low, then smaller genes are more likely to be missed by accident, thus annotated as falsely essential (Deng et al., 2013; Liu et al., 2015).

Another factor is the number and location of the inserts within the gene. The insertion of transposons in the extreme 3’ or 5’ end of the protein may also be acceptable to the cell, as they may not change the conformation of the protein’s active site, thus creating false negative results. Similarly, very long genes may be able to accommodate insertions in multiple positions along their length that do not affect the active sites or conformation too strongly, thus are given as a false negative result (Deng et al., 2013; Lamichhane et al., 2003). The genetic sequence for one gene may contain multiple ORFs, either as genes overlap or for small proteins with a gene. These different peptides may have different essentialities, and many peptides under 100 amino acids in length are not properly mapped, further confounding the essentiality of a given region (Lluch-Senar et al., 2015b).

Finally, there are downstream effects to consider as well. The insertion of a transposon into a gene contained within an operon will not only effect its expression, but potentially the expression of the entire operon. If the gene that is disrupted is non-essential, but the disruption of the operon is lethal to the cell, then that gene will be falsely labelled as essential (Deng et al., 2013). On the other hand, some transposon species contain promotor like elements within their inverted repeats, which can cause transcriptional dis-balance in the disrupted region (Lluch-Senar et al., 2015b). This can affect the surrounding cells in different ways, as the knockout of a potentially essential regulator

gene for an operon can be masked by the new promoter, or the new promoter could cause over-expression of a non-essential gene or operon, leading to cell death and the miss-labelling of the disrupted region.

From these issues, it is clear that a black-and-white system stating if a gene is disrupted by a transposon it is non-essential, and if it is not disrupted it is essential, is a gross oversimplification, and care with the statistical analysis of the data is paramount for determining which genetic elements are indeed essential to a cell (Deng et al., 2013; Gibson et al., 2010; Glass et al., 2006; Juhas et al., 2011a; Lluch-Senar et al., 2015b).

1.4.4.2. Conditional Essentiality

One of the most obvious confounding factors when it comes to assaying essential vs non-essential genes is the environment the cells are subjected to. As described in the section on conditionally essential genes, the essentiality of a gene is based on the environment the cell inhabits, its niche (Joyce et al., 2006). Large numbers of genes will be essential for the bacteria to survive and thrive in an ecological setting that are dispensable for growth under controlled laboratory conditions. These conditionally essential genes by are only essential in specific circumstances, and are elsewise usually non-essential (Timmermans and Van Melderen, 2009). These genes usually take the role of either metabolic genes or housekeeping genes.

As an example of a conditionally essential housekeeping gene would be the ferric uptake regulator (*Fur*) in *Pseudomonas aeruginosa*. This protein controls cell metabolism in response to the availability of iron, and its loss is lethal to the cell on solid media. However, it becomes a non-essential gene when the cells are grown in liquid media. This conditional essentiality on solid media is due to the lack of regulation of pyochelin siderophore biosynthesis, which appears to have a specific deleterious effect on cells during growth on solid media. Growth was not affected upon supplementation of pyochelin during planktonic growth of *fur* depleted cells, indicating it is the synthesis, not uptake, of pyochelin that kills *fur* depleted cells on solid agar. Thus, the *fur* gene is conditionally essential on solid media (Pasqua et al., 2017).

Conditionally essential metabolic genes are relate to functions that are often non-essential in rich media, but essential in minimal media. For example, the *glpD* and *glpK* genes in *E.* coli are non-essential in standard LB media. However, when glycerol becomes the only carbon source, those genes become essential, along with 23 other genes that are non-essential in glucose containing media (Joyce et al., 2006).

There are examples of genes that reverse this trend too. The *dacA* gene in *Listeria monocytogenes* was found to be conditionally essential in rich media but not in minimal media. The *dacA* gene is responsible for the production of c-di-AMP, which in turn regulates the level of guanosine penta- and tetraphosphate ((p)ppGpp). As the levels (p)ppGpp increase, they repress the activity of *CodY*, a transcriptional regulator, thus killing the cell. However, in minimal media, the loss of the *dacA* gene and the resultant lack of c-di-AMP prevents inhibition of *PycA*, an enzyme that catalyses the conversion of pyruvate to oxaloacetate. This resultant over-activity of the TCA cycle is essential to provide enough energy for the cell to survive in depleted media. In rich media, the loss of transcriptional regulation is lethal to the cell when combined with an overactive TCA cycle draining the cells resources. However, when in minimal media and the TCA cycle

being the only source of ATP, this lack of regulation becomes necessary for the cell to survive, and thus the loss of transcriptional regulation becomes a fitness cost that is acceptable in the light of viable energy production (Whiteley et al., 2015).

Traditionally, conditionally essential genes are simple to screen for, by growing saturated transposon libraries of a specific species on different media and analysing the differences in the results (Sassetti et al., 2001).

Other genes will be essential for survival in the bacterium's natural habitat, but not in the lab. One of the most obvious confounding factors when it comes to assaying essential vs non-essential genes is the environment the cells are subjected to. As described in the section on conditionally essential genes, the essentiality of a gene is based on the environment the cell inhabits, its niche (Joyce et al., 2006). Large numbers of genes will be essential for the bacteria to survive and thrive in an ecological setting that are dispensable for growth under controlled laboratory conditions. For example, genes that confer a pathogenic phenotype, aid in the colonisation of host tissue or evade the immune system will confer a strong fitness advantage to the bacterium, and may even be essential.

This was shown in the pathogenic *P. aeruginosa* PAO1 strain. When a representative library of transposon mutants was grown in complex laboratory media, it contained 434 essential genes. However, when the cells were grown in its 'natural environment', i.e. in sputum taken from cystic fibrosis patients, a further 122 essential genes were identified (Turner et al., 2015). Of these genes, 62% were associated with biosynthesis of amino acids and nucleotides, and the rest with the biosynthesis of various metabolites and co factors, such as biotin, riboflavin and spermidine, all of which were provided in the rich media, and thus were not deemed essential to the cells survival in that context.

Another important human pathogen, *Staphylococcus aureus*, shows a similar pattern. When grown on standardised laboratory media, in this case overnight growth on BHI (Brain Heart Infusion) media, 420 genes were described as essential. However, when transposon libraries were grown in various different ocular fluids, 518 genes were found to be essential, with each condition varying in essential gene composition. The same was true of mutants harvested 48 hours after infection in a subcutaneous lesion in mice. Here, a total of 646 genes were deemed essential (Valentino et al., 2014). In every case, the number of genes that are essential for survival rises.

These observations show that comparing the essential genes of multiple organisms has to take account the environment they were grown in, and the essentiality of genes can only be compared across species if the context those cells were grown in is comparable (Koonin, 2003).

1.5. Genetic redundancy, epistasis and moonlighting proteins

1.5.1. Genetic redundancy

The discussion of essential genes is somewhat of a misnomer, as it is not the gene itself that is essential to the survival of the cell, but the function of the protein it encodes for. As such, a cells' cohort of essential genes could be seen as a list of essential functions the cells must achieve to stay alive, not just the list of components required to do so. As stated

earlier, bacteria require different genes to be able to survive in different niches (Joyce et al., 2006; Turner et al., 2015; Valentino et al., 2014). Therefore, they need to be able to not only encode for enough genes to survive in as many different niches as they are likely to encounter, but to also survive the damage or loss to genetic material (Ghosh and O'Connor, 2017). As the loss of any essential function is lethal to the cell, organisms have evolved multiple strategies for mitigating this outcome. The two most common mechanisms used by bacteria are gene duplication, where a genome contains multiple copies of a certain gene, or function duplication, where different genes or pathways produce the same product independent of each other (Ghosh and O'Connor, 2017).

Altering the copy number of genes or larger segments of DNA is one of the most frequent and ubiquitous DNA mutation events, across all domains of life (Reams and Roth, 2015). By providing multiple copies of a gene, the bacterial cell is able to withstand loss or mutation of the original gene and still produce a viable product, thus retaining the function of the gene with minimal change to the phenotype (Zhang, 2012). This strategy is extensively used among prokaryotes, as their haploid genomes do not contain multiple copies of genes by default, unlike diploid or polyploidy genomes. Analysis of multiple prokaryotic genes shows that up to 16% of the genes they contain have a sequence identity score of at least 80% with another gene in the genome (Yu et al., 2015). In general, duplicated genes do not perform the exact same function, but can fill the same role if needed. Despite this, they still have specialised functionalities that may be non-essential but provide an adaptive advantage in specific circumstances.

An example of this can be found in *Streptomyces ambofaciens*, where the gene coding for a stress response transcription factor (σ^B) was duplicated, creating the *hasR* and *hasL* genes. These two genes share 98% nucleotide identity and 97% amino acid identify, however their transcriptional control is highly divergent. In stationary phase, *hasR* is expressed 100-fold higher than *hasL*, and it appears that the two proteins have a positive auto-regulatory loop (Roth et al., 2004). Both genes are regulators of the stress response, however their respective mutations have begun their divergence from pure duplicates to having discrete functions.

Another example can be found in the duplication of the chaperonin gene *groEL* in many species of cyanobacteria, such as *Chlorogloeopsis fritschii*. This organism has a duplication of the *groEL/groES* bicistronic operon and monocistronic *groEL* gene. The genes translated from both operons able to form hybrid complexes with each other, evidences via protein-protein interaction tests, while the monocistronic copy cannot. All three regions have different transcription patterns that appear to be independent of each other, specifically during diazotrophic conditions, indicating that while all three genes generally perform the same role. Even in *E. coli*, the two bicistronic operons can be substituted with the native genes with no loss of function, whereas the monocistronic gene cannot. Due to the widespread adoption of *groEL/groES* duplications in cyanobacteria, it appears that the duplication allows for multiple sub-functionalisation of the protein complex to evolve (Weissenbach et al., 2017).

This duplication can also impart a strong potential for fitness increase onto the cell by allowing adaptation to new environments. The *P. aeruginosa* strain PAO1 is poorly adapted for growth on adenosine, with a doubling time of >40 hours. However experiments have shown when this strain is passed repeatedly on this medium, multiple fast growing strains can emerge. All of these strains contained duplications of the

PA0148, the adenine deaminase, along with a nucleoside hydrolase (*nuh*) that is regulated via quorum sensing (Toussaint et al., 2017).

There is however, a fitness cost associated with duplicating genes. At a basic level, the more duplicate genes a genome contains, the more energy is required to transcribe, translate and replicate the genome. As such, a trade-off is required; balancing the advantages having backups of every gene that conveys a essential functionality to the cell as a defence against deleterious mutation vs the energetic cost of maintaining and translating the increased genome complement. Experiments looking into quantifying the fitness cost of these additions to the genome have found that even modest increases in genome quantity can have marked effects on the fitness of the organism. In experiments in *E. coli* using a high copy number plasmid to express the antibiotic resistance beta-lactamase (*bla*) in various concentrations of the antibiotic was used as a proxy for increased numbers of genes. As the concertation increased, the copy number of the gene increased proportionately, measured via qPCR. A reduction in cell fitness, measured via change in relative growth rate, was observed as more copies of the gene were created, and a loss of 0.15% relative fitness was recorded for each single copy increase of a 1Kb section of DNA (Adler et al., 2014). These findings were replicated in independently in *Salmonella enterica*, with similar costs for duplication fitness (Pettersson et al., 2009; Reams et al., 2010).

1.5.2. Function duplication

A way to mitigate the costs of having backup copies of multiple genes is via function duplication. Here, separate genes or pathways produce the same product independently of each other. For example, riboflavin is an essential co-factor for bacterial cells. It is not only critical to the formation of flavoproteins, but also used in signalling pathways between cells. Due to its importance, many bacteria contain multiple separate biosynthetic pathways for the creation of riboflavin, as well as the ability to uptake it from the environment. This dual functionality ensures that the cell can both uptake or synthesise riboflavin at any time, and the loss of one of the systems can be mitigated by the presence of the others. Thus, a mutation knocking out the transport pathway can be neutralised by the presence of a biosynthetic pathway. Neither gene is a biological duplicate of the other, but they both produce the same product for the cell. This is frequently found in metabolism (García-Angulo, 2017).

1.5.3. Moonlighting functions

Moonlighting proteins are those that undertake multiple functions other than their primary role in the cell. In bacteria they are generally metabolic enzymes, most commonly glycolytic, or molecular chaperones which can localise to the cell membrane to perform secondary activities (G. Wang et al., 2014). Most commonly, moonlighting proteins that localise on the cell membrane are involved in adhesion, despite often having no known anchoring or adhesion domain within their structure (Jeffery, 2018; Kainulainen and Korhonen, 2014).

Moonlighting proteins also play a large role in pathogenesis for many species. Here, it is worth noting that the majority of the moonlighting proteins are highly conserved genes, traditionally involved in core metabolism such as the TCA cycle, glycolysis, chaperones and proteases. Many of these genes are conserved across multiple domains of life, yet in bacteria can have vastly different functions to their standard properties (Henderson,

2014). Indeed, every single gene involved in bacterial glycolysis, the transformation of glucose to pyruvate, has been shown to have moonlighting functions, as shown in Figure 9 below:

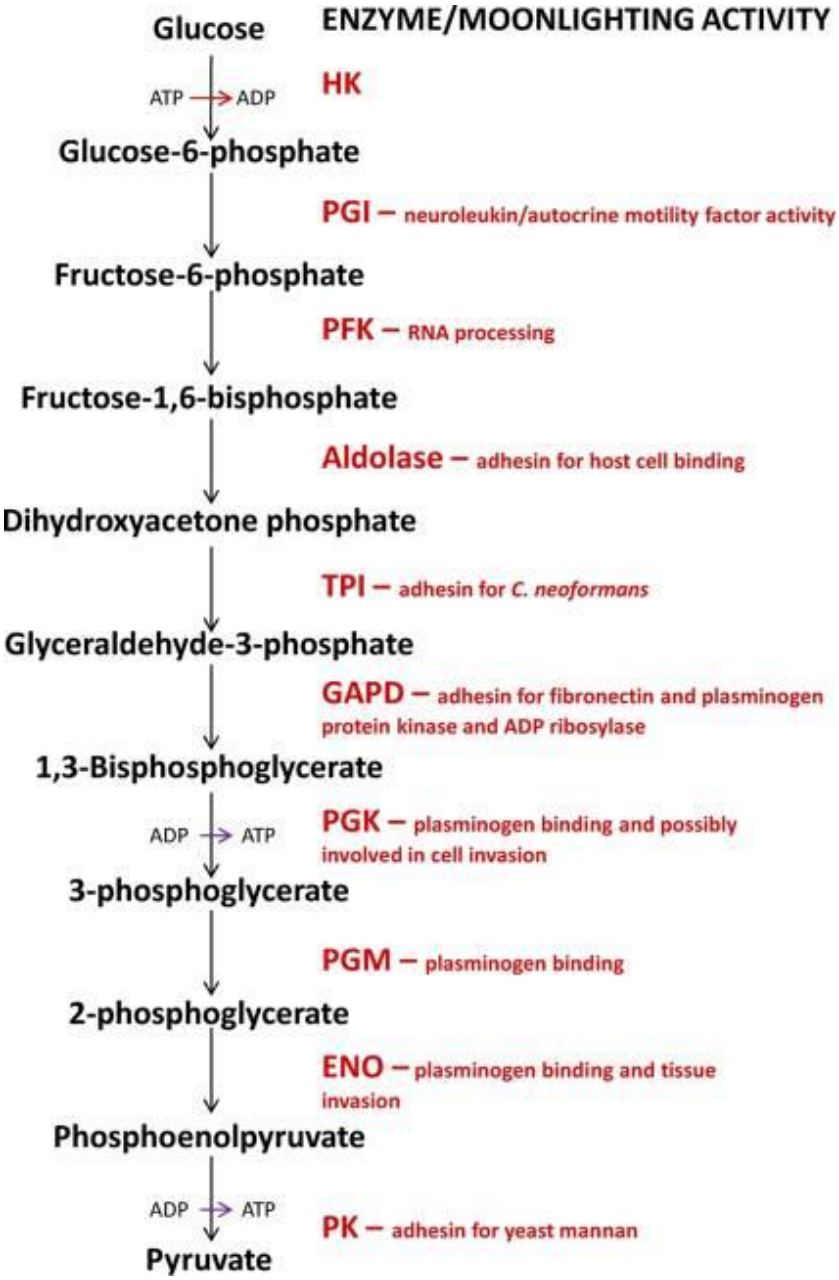


Figure 9: Moonlighting functions of the glycolysis enzymes in bacteria. HK, hexokinase; PGI, phosphoglucose isomerase; PFK, phosphofructokinase; TPI, triose phosphate isomerase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; ENO, enolase; PK, pyruvate kinase. Taken from Henderson and Martin,(2011).

1.5.4. Epistasis in Bacteria

Epistasis is the name given to the interaction of multiple factors onto a single phenotype (Weinreich et al., 2013). Generally, a single mutation will elicit a specific change in fitness for the cell, either positive or negative. When multiple mutations act in combination, the sum total of the fitness change is often not linear. For example, two mutations that impart a negative fitness change into a cell individually could result in a benign or even beneficial double mutation (Sackman and Rokyta, 2018). For example, antibiotic resistance genes impart a large fitness cost on a bacterium, as they often need

to produce large amounts of the protein. As a result, it has been observed that bacteria that retain multiple copies of antibiotic resistance genes compensate by allowing mutations in other pathways. Alone, these mutations are often impart a negative fitness cost onto the cell. However, when paired with the resistance genes, they allow the cell to transfer energy and resources to the production of the antibiotic proteins, and their presence increases cellular fitness (Moura de Sousa et al., 2017).

Most phenotypes are multifactorial, being the product of multiple proteins working in tandem, and many proteins having multiple functions (G. Wang et al., 2014). Therefore, a mutation in a gene will not only affect the phenotype that the gene is directly responsible for, but also any other phenotypes that the gene contributes to (Jerison and Desai, 2015; Weinreich et al., 2013). As such, mutations and deletions can have unintended fitness consequences for the cell unrelated to the main function lost. An extreme example of epistatic interactions are synthetic lethality pairs. These are pairs of genes which under a given condition are both non-essential, and thus can be individually mutated or removed. However, upon the disruption of both genes, a lethal phenotype is initiated (Klobucar and Brown, 2018). Generally, these genes both contribute to the same essential function, and due to the cell being non-viable when that function is lost, it indicates that the function is essential. However, neither of the genes that are responsible for the function are essential, and thus these interactions can give a misleading indication as to how many essential genes there are in a cell (Mori et al., 2015).

1.5.5. Persistent Non-Essential genes

A sub-class within the fitness genes are the persistent non-essential genes. These genes tend to be very conserved across the different classes of bacteria, indicating a strong selective pressure for their retention, yet are almost always indicated as non-essential. These genes tend to have a mix of translational and housekeeping functions, typically functions that have other genes encoding for redundancies or that will impart a strong fitness disadvantage without killing the cell (Fang et al., 2005).

1.6. Prokaryotic pan-genome

1.6.1. LUCA and common decent

Evolution from common descent is one of the cornerstones of modern evolutionary theory, and can trace its way back to Darwin's original theory of evolution. All life on earth shares at least three basic traits; being fully carbon based, the use of a (with a few minor exceptions) a universal genetic code consisting of both DNA and RNA for the storage of information, and a shared core of ribosomal proteins for the decoding of that information (Koskela and Annila, 2012; Weiss et al., 2018; Yutin et al., 2012). This similarity indicates that all life that we currently know of descended from a single source, known as the Last Universal Common Ancestor (LUCA). It is difficult to say whether we would classify whatever LUCA was as an organism by modern standards. All we can currently tell is that it contained some form of proto-ribosome and it translated mRNA into protein, and this can only be inferred from analysing current genomic data (Weiss et al., 2018). However, the fact that a common point of descent exists allows us to see what is still common among all bacterial species, which genes are so vital for cellular survival that they have persisted since life's antiquity. By analysing the pan-genome of prokaryotes for similarities, it allows us to build a blueprint for the functions of life that are indispensable.

1.6.2. Previous studies examining the conserved genes between bacteria

Many studies have looked at which genes are conserved across the genomes of sequenced bacteria, attempting to quantify patterns of conservation in genes across the bacterial domain. The aim has been to identify if there are any pathways or processes that are essential or even conserved across such a huge diversity of life.

Using comparative genomics to look for patterns and similarities in large datasets is a stable of modern bioinformatics analysis, and one of the first papers to do this on a large scale for bacterial genomics was Eugene Koonin's review entitled "Comparative genomics. Minimal gene-sets and the last universal common ancestor" (Koonin, 2003). The review focuses on the makeup of minimal gene complements, such as that of *Mycoplasma genitalium*, and how they allow the cell to survive in their respective niches. He argues that by studying which proteins are found in minimal organisms, and then looking for orthologues in larger bacteria, it is possible to deduce which proteins and functions are necessary to all life.

Koonin asserts that the main sources of the vast genetic variation found in prokaryotes is due to two main factors, horizontal gene transfer (HGT) and non-orthologous gene deletion (NOGD). Indeed, he paraphrases Theodosius Dobzhansky to "Nothing about (at least prokaryotic) evolution makes sense except in the light of horizontal gene-transfer and lineage-specific gene-loss". This claim asserts bacteria gaining new genes through HGT, which out-compete current genes at a certain task, mainly explain genetic variance among species. These less fit genes are subsequently lost, and all future bacterial progeny will contain the new gene instead. From this, by looking at the proteins that are only found in the minimal genomes and accounting for the rates of gene loss and gene gain, a reasonable assumption of the genome complement of LUCA can be made, which Koonin places at around 500-600 genes.

The other main message of the paper is which genes at the time were found in all sequenced bacterial species. Koonin claims to have studied the sequence of "~100 genomes" and found 63 genes that were ubiquitous in all species, shown in Table 3.

Table 3: Ubiquitous genes, taken from Koonin, (2003)

FUNCTION	NUMBER OF GENES
TRANSLATION	
Ribosomal proteins	30
Aminoacyl-transfer-RNA synthetases	15
Translation factors	6
Enzymes involved in RNA and protein modification	3
Signal-recognition-particle components involved in secretion	3
Molecular chaperone/protease	1
TRANSCRIPTION	
RNA-polymerase subunits	2
REPLICATION/REPAIR	
DNA-polymerase subunit, exonuclease, topoisomerase	3
TOTAL	63

Translation is the main conserved function, with a small number of transcription and DNA repair genes present as well, which as Koonin pointed out, supports the hypothesis that the LUCA used RNA and ribosomes. He also argued that the lack of DNA replication machinery in the pan-genome indicates that LUCA may not have had a traditional DNA genome, as it cannot have supported the functions we know are currently required to maintain such a state of being, specifically through the lack of DNA repair and replication genes.

As the number of sequenced bacterial species rose, larger analyses could be run. Charlebois and Doolittle, (2004), looked to see if increasing the sample of bacteria further reduced the core of orthologous genes in the genome, or if Koonin's gene set was stable at larger samples sizes. They also looked to see if adding further prokaryotes, in this case 17 archaea, dramatically altered the conserved core of genes or not. In total, they analysed 130 bacterial genomes, alongside 17 archaeal ones. They also looked to see if the method they used for finding the ubiquitous genes biased the outcome of the results, and to that end if the data they found was only "a statistical illusion".

To see if the core set of genes was shared between multiple search methods, they used two discrete methodologies to acquire orthologous genes from the species available. The first was the reciprocal best match (RBM) method. This used the BLASTP bit score for the protein of interest against the other proteins, to see if there were any other proteins with a highly similar sequence. BLAST, or Basic Local Alignment Tool, is a program that aligns nucleotide or amino acid sequences against the NCBI database to identify them. The more closely two sequences resemble each other, the higher the score attributed to the pair is (Altschul et al., 1990). Any sequences that were above a certain cut-off were analysed and BLASTed against each other. The entry which had the best score when each protein was BLASTed was assigned the orthologue status. This ensures that to assign protein A and B as orthologues, protein B had to have the highest BLASTP score when protein A was queried, and *vice versa*. By varying the cut-off of the BLASTP score, they could make the search more or less stringent in regard to sequence homology.

The other method used was consensus gene name (CGN). Here, the RBM of any ORF in a genome was found, and if it had an annotated gene name, that name was added to a list. For each of the genes, the most common name ascribed to it was chosen as the query ORFs CGN, and searched for in all other species.

The number of conserved genes found within the study is shown in Table 4, and how altering the BLASTP cut-off value changes the results returned is shown in Table 5.

Table 4: Charlebois & Doolittle – Number of genes strictly shared (via RBM, BLASTP cut-off 1.0e-5) within prokaryotes

	Orthologues	Range	No. Species
All prokaryotes	14.82 (sd = 2.55)	10-23	147
Archaea	144.53 (sd = 8.00)	128-156	17
Bacteria	61.53 (sd = 2.71)	56-70	130

Table 5: Charlebois & Doolittle – Number of genes strictly shared by prokaryotes, by simple match and RBM, and various BLASP cut-off expectation values

E-value	147 sp. Simple match	147 sp. RBM	130 bacteria RBM	17 archaea RBM
1.0E-3	101.5	18.0	63.5	153.6
1.0E-4	93.6	16.1	62.6	150.1
1.0E-5	87.2	14.8	61.5	144.6
1.0E-7	78.1	12.1	59.5	133.6
1.0E-10	65.8	10.0	55.3	118.6
1.0E-15	35.2	7.4	46.6	99.9
1.0E-20	12.1	5.5	28.8	83.5
1.0E-30	4.7	2.2	24.5	58.8
1.0E-50	1.5	9.9	12.6	33.5
1.0E-100	0.0	0.0	5.9	10.5

The study also found very close agreement in the gene sets identified via both methods, indicating that the results probably are not just artefacts of the search methodology. Using the CGN method, the following genes were identified in all 147 prokaryotic species:

*Table 6: Charlebois & Doolittle - 34 consensus gene names found in all 147 prokaryotic genomes. * indicates the genes are found in Table 4 from strict sharing analysis.*

	Transcription
**rpoB*	RNA polymerase, β subunit
nusA	Transcription pausing, L factor
nusG	Involved in transcription anti-termination
	Translation
dnaG	DNA primase
**infB*	Translation initiation factor IF-2
**tufa*	Translation elongation factor EF-Tu
**fusA*	Translation elongation factor EF-G
ksgA	S-adenosylmethionine-6-N_,N_-adenosyl (rRNA) Dimethyltransferase
**ychF*	GTP binding protein
**argS*	Arginyl-trna synthetase
**gltX*	Glutamyl-tRNA synthetase
**hisS*	Histidyl-tRNA synthetase
leuS	Leucyl-tRNA synthetase
lysS	Lysyl-tRNA synthetase
**metG*	Methionyl-tRNA synthetase
**pheS*	Phenylalanyl-tRNA synthetase
**proS*	Prolyl-tRNA synthetase
**serS*	Seryl-tRNA synthetase
**thrS*	Threonyl-tRNA synthetase
**trpS*	Tryptophanyl-tRNA synthetase
**valS*	Valyl-tRNA synthetase
**rplA*	Ribosomal protein L1
**rplC*	Ribosomal protein L3
**rplE*	Ribosomal protein L5
rplF	Ribosomal protein L6
**rplK*	Ribosomal protein L11
**rplN*	Ribosomal protein L14
**rpsB*	Ribosomal protein S2
**rpsC*	Ribosomal protein S3
**rpsD*	Ribosomal protein S4
**rpsG*	Ribosomal protein S7
**rpsH*	Ribosomal protein S8
	Intracellular trafficking and secretion
secY	ATPase subunit of translocase
	Post-translational Modification
**gcp*	O-sialoglycoprotein endopeptidase

They also found very close correlation in results of Koonin (2003), both in number of genes shared by the bacterial and their general functions, with the vast majority found to be translation genes. They also agree on the fact that there are so many complexities involved in evolution, and such huge timespans since LUCA existed, that it is probably not advantageous to limit our analysis of the data to genes that are 100% ubiquitous. Both the machinations of evolution, and unavoidable human error in miss-annotations of large genomes, mean that there will always be a level of false negative results that we cannot distinguish.

The study by Luo *et al.,* (2015), looked again at comparative genomics of a pool of different bacterial species, but this time focusing on species where the essentiality of genes was known, along with their function. Twenty-three bacterial species were analysed, and genes were analysed via a Ka/Ks ratio to see how evolutionarily conserved they are. The Ka/Ks ratio looks at the number of non-synonymous mutations per non-synonymous site (Ka) against the number of synonymous mutations per synonymous site (Ks), with the lower the ratio, the more conserved the gene is, a marker for the strength of the evolutionary pressure in retaining the gene.

To assign the Ka/Ks values to the genes, the full sequence of the genomes was subject to an all-to-all BLAST to determine the ORFs. These were aligned using ClustalW2 and mapped to the appropriate amino acid sequence using Pal2Nal. Their dataset contained over 70,000 genes across the 23 species, with approximately 180,000 pairs of proteins that formed an orthologous pair and had valid Ks/Ka rates. They also cross-referenced all of the genes against the COG database, and assigned COG classes as necessary.

The average Ka, Ks and Ka/Ks ratio for each of the essential and non-essential genes in each organism was calculated, and the statistical difference between them was determined using the Mann-Witney U test. This showed that for most organisms, the essential genes had a statistically lower Ka/Ks ratio, as shown in Figure 10.

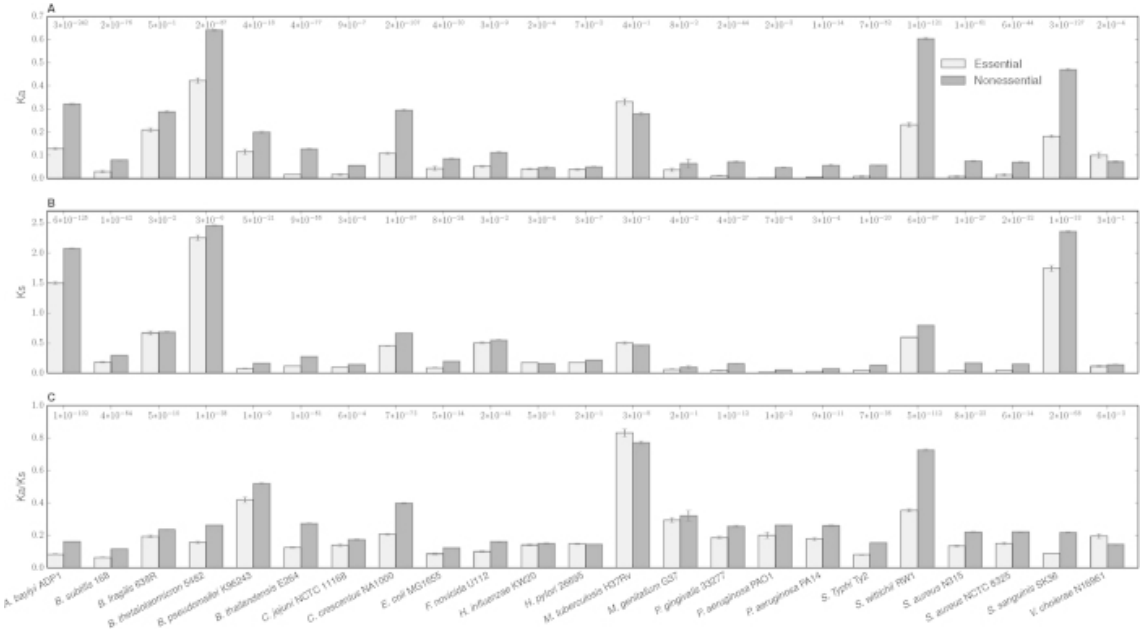


*Figure 10: Luo et al., 2015: The histogram shows the averages for (**A**) Ka, (**B**) Ks and (**C**) Ka/Ks values between the essential and nonessential genes, respectively. The* P*-values calculated by Mann–Whitney U Test are also displayed at the top of figures.*

The reason that the gap is not larger in some species is likely due to persistent none essential genes, as discussed in chapter 1.5.5. *Burkholderia pseudomalliae, Mycobacterium tuberculosis* and *Vibrio cholera* contain large amounts of these genes, and the value of *Sphingomas wittichi* is unknown, which could explain their high values for the non-essential genes. However, this analysis is biased towards larger genes. The greater the amount of the protein used to interface with the environment or complete its function, the more conserved that protein will be. For example, the *groEL* chaperone contains multiple binding domains arranged in ring-like structures, all of which are in contact with the proteins they bind (Ishii, 2017; Ryabova et al., 2013). Therefore, there is

likely to be much less allowance for mutation across the whole sequence of the gene. Other essential genes may have very small active sites in comparison, and can tolerate much more variation in the remaining amino acid sequence, despite their function being just as essential. However, in this analysis the Ka/Ks ratio of the *groEL* gene would be far lower that of a smaller, more variable gene of equal essentiality, thus the results have the potential to bias towards potentially over-representing the essentiality of large genes.

Along with the data from Koonin (2003) and Charlebois & Doolittle (2004), they also found that genes involved in translation, transcription and DNA replication and repair where highly conserved evolutionarily. Genes that were identified with COG classes J (Translation, ribosomal structure and biogenesis), K (Transcription) and L (Replication, recombination and repair) to be highly conserved, along with classes I (Lipid metabolism, G (Carbohydrate transport and metabolism) and H (Coenzyme transport and metabolism).

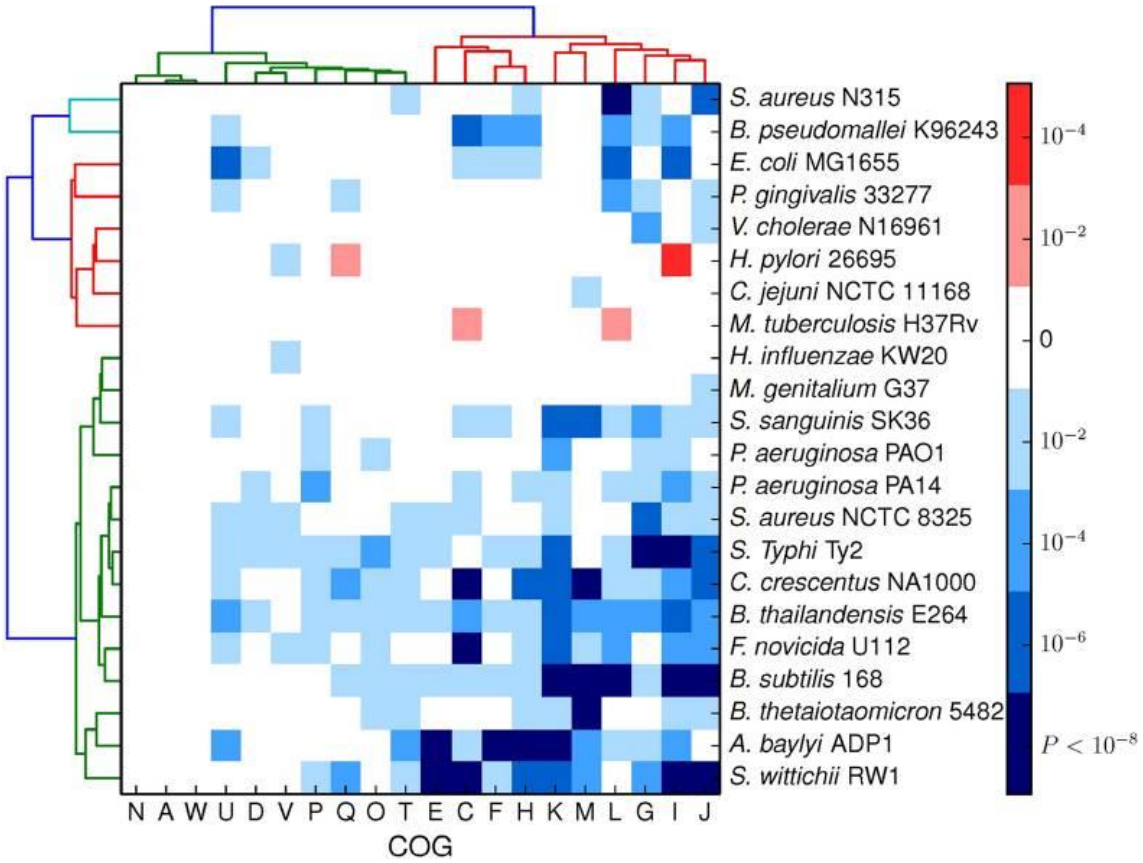


Figure 11: Luo et al., (2015) COG Classes: The hierarchal cluster diagram was constructed by Ward's linkage clustering. The P-values of each COG category are calculated with Mann–Whitney U Test by organism, which reflect the significance of the difference for the Ka/Ks value between the essential and nonessential genes. The blue boxes represent that the COG subcategory in which the essential genes are evolutionarily conserved than the nonessential ones, while the red boxes represent the opposite case.

1.6.3. Sparsity in pan-conserved genome

One of the strongest conclusions of the papers listed above is the paucity of genes that are conserved among all species of bacteria, with between 30 and 60 depending on if you include archaea or not. The vast majority of the genes present are related to translation, with a few transcription and DNA repair genes alongside (Charlebois and Doolittle, 2004; Koonin, 2003; Luo *et al.*, 2015). This core genome is evidently unable to support life, with even the conserved genes being incapable of fulfilling their tasks alone. While a

large amount of genes present code for ribosomal proteins, there are not enough to form a functional translation machine, nor are there enough genes to provide the transcribed mRNA required, or tRNA synthetases to translate the mRNA into proteins. Even if these processes were possible however, the lack of cell division and DNA duplication creates an inherent obstacle to cell division and proliferation. The lack of metabolic genes also renders the cell incapable of producing energy or precursors for any form of biogenesis, and the lack of genes encoding membrane proteins would obviate it from being classified as a lifeform at all (McKay, 2004).

The complete lack of metabolic functions within the sets of conserved genes shows the true vastness of niches that bacteria inhabit, and the level of their adaptation to them. As outlined in chapter 1.2.2.4, there are multiple metabolic pathways bacteria exploit with the difference in aerobic vs anaerobic respiration being a key factor, along with the ability to extract metabolites and biological precursors from the environment vs the need to synthesis them independently (Chubukov et al., 2014; Pál et al., 2005; Passalacqua et al., 2016).

However, studies on metabolomics in large bacterial datasets have shown large levels of similarity between disparate species. In a study of 94 bacterial species, 42 annotated metabolic pathways were found to be conserved among all species (Kolhi and Kolaskar, 2012), shown in Table 7.

Table 7: Conserved metabolic pathways in 94 bacterial species. Adapted from Kolhi and Kolaskar, (2012)

Amino Acid Biosynthesis	**Carbohydrate Biosynthesis**	**Nucleotide Metabolism**
Alanine Biosynthesis I	Gluconeogenesis I	Adenosine nucleotide *De Novo* Biosynthesis
Arginine Biosynthesis II (Acetyl cycle)	Glycolysis I	Guanosine nucleotide *De Novo* Biosynthesis
Cystenine Biosynthesis I	Pentose Phosphate Pathway (Non oxidative)	Pyrimidine Deoxyribonucleotide *De Novo* Biosynthesis
Glutamine Biosynthesis I	TCA Cycle	Pyrimidine ribonucleotide interconversion
Glycine Biosynthesis I		Uridine-5'-phosphate biosynthesis
Histidine Biosynthesis	**Cofactor Biosynthesis**	
Isoleucine Biosynthesis I (from Threonine)	Coenzyme A Biosynthesis	**Others**
Leucine Biosynthesis I	S-Adenosylmethionine Bioyhtesis	tRNA charging pathway
Methionine Biosynthesis I		Cardiolipin Biosynthesis I
Phenylalanine Biosynthesis I	**Lipid Biosythesis**	Chorismate Biosynthesis I
Proline Biosynthesis I	Fatty acid & Beta;-Oxidation I	Flavin Biosynthesis I
Serine Biosynthesis I	Fatty acid Elongation (Saturated)	Tetrahydrofolate Biosynthesis I
Threonine Biosynthesis (From Homoserine)	Fatty acid Biosynthesis Inititation III	PRPP Biosynthesis I
Trpytophan Biosynthesis		UDP-N-Acetyl-D-Glucosamine Biosynthesis I
Valine Biosynthesis		Thioredoxin pathway
Tyrosine Biosynthesis I		Acetyl CoA Biosynthesis (from pyruvate)
Homoserine Biosynthesis		Superoxide Radicals Degradation

The largest subset of these were pathways involved in amino acid biosynthesis, 17 unique pathways, which mirrors the preponderance of genes related to translation in the other analyses. However, while all of the bacteria shared many standard biological pathways such as glycolysis, nucleotide synthesis and the production of standard cofactors such as riboflavin, none of the genes involved in these pathways were found in the gene level analysis. The analysis included 50 aerobic, 38 facultative and 6 anaerobic bacteria, with no archaea present (Kolhi and Kolaskar, 2012), indicating that the results were not biased by including just a single metabolic class of bacteria.

1.6.4. Function vs gene in conservation

One of the most probable reasons that bacterial genomes contain so few conserved genes is due to the phenomena discussed in chapter 1.5 regarding genetic redundancies. All bacteria need to be able to fulfil the basic functions of DNA transcription and translation, basic metabolic function and replication, as outlined in chapter 1.2.2. However the ability of proteins to moonlight to other functions, and after 4.5 billion years since LUCA (Cantine and Fournier, 2018), evolution has provided a plethora of alterative proteins for

use in all pathways. Therefore, it is appropriate to conclude that it is the functionality, not the gene itself, which is preserved.

1.7. The minimal genome concept

1.7.1. Hypothetical Minimal Genome

The idea of a Minimal genome is one that has interested researchers for decades. At its core, the Minimal genome is one that codes for only the genes that are essential to its function, and no other superfluous functions (Glass et al., 2017). However, as mentioned above, it is very difficult to define in the real world what genes are necessary or not. It depends entirely on the context the organism is found in (Koonin, 2003). The minimal genome required for a cell to survive in a nutrient poor environment will be very different to that required for the cell to survive in rich medium, which will be in turn very different from that of an organism that forms a parasitic or mutualistic relationship with a host. The cell in the nutrient poor environment will require a vastly different metabolic gene set, as it will need to synthesise all of its own metabolites from very basic starting materials. However, the cell in the rich medium will need different genes to deal with increased osmotic pressures, and transporters to uptake all of the varying metabolites it can find, but need fewer metabolic genes as it can source metabolites from the environment. The mutualist/parasite may be able to extract key metabolites from its host, thus lose many genes related to key metabolic functions, yet require genes for immune evasion or production of molecules of interest to the host cell. All of these organisms can be defined as Minimal genomes, yet will have markedly different sizes and genetic compositions (Choe et al., 2016; Koonin, 2000; Mushegian, 1999).

This concept of the Minimal genome also begins to blur the lines on what we can define as life, a topic that already is not without significant controversy. If an organism is so reliant on its host that it cannot provide the metabolic functions needed to survive on its own, or is incapable of replication outside a host, then can it be defined as alive? The general consensus on whether viruses and phages are alive is that they are not, although again this is not without significant controversy (Forterre, 2016, 2010; Lwoff, 1967; Moreira and López-García, 2009; van Regenmortel, 2016). However many Minimal genomes, both theoretical and naturally occurring, display many of the traits that some believe should exclude viruses and phages from being classed as alive. For example, on the basis of size and number of genes, the detection of "giant viruses" has uncovered viral 'species' with genomes sizes larger than some known bacteria. The 1.57Mb genome of the Klosneuvirus also contains RNA synthetases for all 20 amino acids, along with multiple transcription factors and tRNA modifying enzymes, a larger complement than many minimal bacteria (Schulz et al., 2017). This makes the genome of the virus, and many like it, far larger and with a far greater coding capacity than many endosymbionts, yet generally these minimal bacteria are defined as living organisms and viruses are not (Tamames et al., 2007). As with viruses, endosymbionts are generally also incapable of cell division outside of their host, relying on them for either the provision of tRNAs they are unable to synthesise or even components of the requisite translation machinery (Uchiumi et al., 2019). The dependency on foreign replication machinery is often cited as a key factor in the denial of life to viruses (Moreira and López-García, 2009; Ruiz-Saenz and Rodas, 2010), yet many Minimal bacteria blur this line.

It is worth defining at this point that from hereon in, 'Minimal genome' (with capitalisation) refers to the hypothetical construct just explained, while 'minimal genome' denotes actual organisms with highly reduced genomes.

1.7.2. Naturally occurring minimal bacteria

1.7.2.1. Axenic genus: *Mycoplasma*

Mycoplasmas, a bacterial genus within the Mollicutes class, are among the smallest and simplest self-replicating lifeforms on Earth. With an average genome size of 1 megabase (Mb) (Lin and Zhang, 2011), they are the product of large scale reductive evolution. Originally believed to have descended from the Firmicutes class, mycoplasmas are Gram-positive bacteria, which lack a cell wall (this forms the Latin root of mollicute; *mollis* meaning soft and *cutis* meaning skin). Typically, they also have a very low G+C content and employ the TGA codon non-canonically to encode for tryptophan instead of the usual STOP codon (Weisburg et al., 1989).

Due to their diminutive size, mycoplasmas were originally believed to be viruses, known as Eaton's agent, as they can pass through standard bacteriological 0.22µm filters. However, after researchers discovered they could be cultured and displayed a 'fried egg' morphology similar to L-form bacteria, this became their new designation. It wasn't until the 1960's with the advent of DNA hybridisation techniques that it became clear that the mycoplasmas were clear and separate family of bacteria (Razin and Hayflick, 2010; Waites and Talkington, 2004).

Due to their highly reduced genome, mycoplasmas tend to have severely limited biosynthetic capabilities compared to other free-living bacteria, necessitating their obligate parasitic lifestyle. Most mycoplasmas cannot synthesize their own lipids or nucleic acids and lack the enzymatic pathways to allow for a functional tricarboxylic acid cycle, depending glycolysis for ATP production and acquiring what they cannot produce from their hosts (Dybvig and Voelker, 1996). Another result of this lack of synthetic capabilities is the requirement for sterols. Due to the lack of cell wall, the osmotic and mechanical pressures on the plasma membrane are high and therefore sterols such as cholesterol or phosphatidylcholine are required to maintain rigidity (Waites and Talkington, 2004). Indeed, due to this dependence on external lipid and sterol sources, to date there are no known abiotic mycoplasma species occurring naturally outside the laboratory (Blötz and Stülke, 2017). As a result, each member of the of the mycoplasma genus has adapted to a highly specific environmental niche, with at least 16 species documented as parasites in humans and a further 35 in other birds and mammals (Trachtenberg, 2005).

Despite their obligate parasitic nature, many species of mycoplasmas can be grown independently as a pure culture under laboratory conditions (Beier et al., 2018; Herrmann and Reiner, 1998; Razin, 1985; Robertson et al., 1975; Watanabe, 1994), making them the smallest axenic bacteria currently known. The title of smallest naturally occurring axenic bacteria goes to *Mycoplasma genitalium*, with a genome size of only 580Kb (Blanchard and Bébéar, 2011; Hutchison et al., 1999).

Bacteria belonging to the genus *Mycoplasma* are ideal candidates to study minimal bacterial organisms, not just because they contain the current smallest known axenic bacterial species, but because they have evolved into these forms from more complex

predecessors. As a member of the Tenericutes, they are closely reated to another class in that phyla, the Firmicutes. As such, they are closely related to the *Bacilli* class, such as *Bacillus subtilis* (Davis et al., 2013; Weisburg et al., 1989; Woese et al., 1980). They are not, therefore, fossils that have stayed genetically static over the eons, but the product of large scale gene loss through evolution, showing us exactly what genes are necessary for survival in their specific niches, and what biological functions are superfluous (Glass et al., 2017).

As explained previously, *M. genitalium* has the smallest genome of any known axenic bacteria, containing only ≈480 protein coding genes (Blanchard and Bébéar, 2011). It is an obligate human pathogen, colonising the urogenital tract of both males and females, causing urethritis and pelvic floor inflammatory disease, amongst other illnesses (Hamasuna, 2013; Lis et al., 2015). It was first isolated in 1980 from the urethral swabs of men with non-gonococcal urethritis (Tully et al., 1981), and has since been the subject of intense study, due to its ability to survive despite its diminutive protein coding capacity.

The Mycoplasmas as a genus were first put forward as model organisms for studying minimal genomes by Harold Morowitz, a physicist at Yale, in 1984 (Morowitz, 1984). He pointed out that we were able to grow *Mycoplasma mycoides* in a defined media, and the steps he envisioned were not far removed from the analyses that were performed as technology progressed. It was not until the 1990's however that molecular biology techniques had advanced to the point of large-scale interrogation of the mycoplasmas became possible, and the discovery of *M. genitalium* in 1981 shifted the focus of the investigations to it as the model organism of choice.

The genome of *M. genitalium* was first sequenced by Fraser et al., (1995) at the J. Craig Venter Institute (JCVI), and found to be 580070 base pairs in length, with a G+C content of 32%. 670 putative ORFs were assigned, discarding any ORF that was less than 100 amino acids in length, and of the 96 were found to have no known orthologs in other known bacteria at the time. Of the genes that could be identified, ≈45% of them were directly related to transcription, translation or DNA replication (Fraser et al., 1995). The first major investigation into the essentiality of these genes also took place at the JCVI, using a global transposon mutagenesis of the genome (Hutchison et al., 1999). The rationale was that by identifying all of the non-essential genes within the already minimal genome, a closer approximation of the minimal number of essential genes needed to support cellular life could be found. The cultures were passed for over 30 generations to ensure the retention of the transposon, and thus the loss of function of the gene it disrupted. After the passages, the location of the transposon insertions was identified via sequencing. The study identified 685 unique transposon insertion sites throughout the genome, matching to 484 locations within genes and 199 intergenic regions, indicating strong preference for intergenic regions. Coupling these disruption locations with the rationale that insertions within the extreme 5' or 3' end of a gene may not destroy the proteins function, they created an estimate of between 180-215 non-essential genes in *M. genitalium* (Hutchison et al., 1999).

As a model organism for a minimal genome, its pattern of essential genes follows generally what was predicted that a Minimal genome should have, with a large portion of the essential genes encoding for basic cellular processes such as transcription and translation, DNA repair and replication, and basic energy metabolism. However, there were also many essential genes that had unknown functions, over 100 when the study was

published, that indicated there is more complexity to cellular survival than anticipated (Hutchison et al., 1999).

Mycoplasma pneumoniae is the species most closely related to *M. genitalium*, and is also an obligate human pathogen. It colonises the upper and lower respiratory tract, and causes atypical or community acquired pneumoniae (Broulette et al., 2013; Cunha and Pherez, 2009), though neurological sequela have been reported (Poddighe, 2018; Tsiodras et al., 2005), which is true in many other mycoplasma species (Rosales et al., 2017). The majority of the virulence of *M. pneumoniae* is caused by the production an exotoxin named community-acquired respiratory distress (CARDS) toxin, along with the ability to excrete hydrogen peroxide (Waites et al., 2017).

It has a genome size 816Kb, containing ≈700 genes (Lluch-Senar et al., 2015a) and orthologs for almost every gene found in *M. genitalium* (Hutchison et al., 1999), making it a useful point of comparison. Of the genes present, approximately 45% of them are classified as essential, comprising of 33% of the total genome (Lluch-Senar et al., 2015b). *M. pneumoniae* cells are typically spindle shaped cells, with a diameter of 0.3 µm and a length of between 1-2µm, with a prominent terminal organelle at one end (Krause and Balish, 2004).

M. pneumoniae is highly specialised and adapted to its niche, specifically through its prominent terminal organelle. This allows for cytadherance to the epithelium of the host cells, along with allowing *M. pneumoniae* to exhibit a gliding motility (Krause and Balish, 2004; Waites et al., 2017). This complex structure is an extension of the membrane, and is comprised of 37 different proteins. It is comprised of an electron dense core surrounded by adhesins and attachment proteins, anchored to the rest of the cell membrane (Hasselbring et al., 2006a; Krause and Balish, 2004; Nakane et al., 2015), and in combination with the attached P30 adhesin, is responsible for cell division as well as movement (Lluch-Senar et al., 2010).

1.7.2.2. Non-axenic genus: *Buchnera*

There are organisms with smaller genomes than the Mycoplasmas, however these bacteria cannot be cultured on independently, and rely on a host organism for survival. The obligate insect endosymbiont genus *Buchnera* contain many examples of such organisms. This genus of bacteria contain genome sizes ranging from 416 to 644Kb, and are endosymbionts of the aphid fly, where they colonise the insect mid-gut (Chong and Moran, 2018). These organisms are classified as endosymbionts, not pathogens, as they provide a mutually beneficial relationship with their host. The relationship between the two species is highly interdependent, as the *Buchnera* cannot survive outside of the aphid host, and similarly the aphid cannot reproduce without the *Buchnera* (Zhang et al., 2016). The relationship between the two species relies on mutually assisted nutrition, as the *Buchnera* provides the aphid with amino acids that are deficient in its diet (Douglas, 1998), along with purines that the aphid cannot metabolise (Ramsey et al., 2010). In return, the aphid provides the *Buchnera* with a safe environment to grow, and their location in the mid-gut allows them access to the nutrients ingested by the aphid (Chong and Moran, 2018).

An analysis of three closely related *Buchnera aphidicola* strains isolated from different aphid species showed an average genome size of 641Kb, with a GC content of 26% and

between 504 and 560 protein-coding genes. When looking at the minimization of the genome, the authors noted that all of the genes involved in the amino acid biosynthesis pathways were conserved, however genes involved in transport were extensively lost, along with many DNA recombination and repair genes, such as the homologous recombination and DNA methylation pathways. The species has also seen major reductions in its DNA replication machinery, with large truncations in the dnaX and polA genes removing the subunits that encode for the DNA polymerase III holoenzyme and DNA polymerase I respectively. This truncation of the DNA pol I gene means the loss of 3' to 5' exonuclease proofreading, and forcing the cell to reply purely on the DNA polymerase III enzyme as the sole active polymerase. This highly reduced repair and replication machinery may have had an important role in the cell's genome reduction process (van Ham et al., 2003).

Buchnera spp. are among the larger of the endosymbiotic bacteria however. Due to the ability to take up nutrients and metabolites directly from the host, some species have reduced their genome size considerably further. The current smallest known is *Candidatus Carsonella ruddii*, with a genome size between 160Kb (Nakabachi et al., 2006) and 174Kb (Katsir et al., 2018), depending on the host organism. They are endosymbionts of psyllids, and like *Buchnera*, supply the hosts with amino acids. However, with a genetic coding capacity of only 182 genes, it is missing the genes needed for functional histidine, phenylalanine and tryptophan biosynthesis. On top of this, the organism lacks almost all the genes needed for DNA replication, transcription and translation, with only degraded copies of the DNA helicase, DNA primase and RNA polymerase present. There are no DNA polymerases or gyrases, no DNA repair mechanisms, all tRNA synthetsases are missing apart from degraded or probable non-functional copies of the phenylalanine or valine synthetases, and only 2 mutated versions of ribosomal proteins (Tamames et al., 2007). The lack of replication and translation machinery indicates these functions are undertaken by external factors, and thus raises the question of whether *C. ruddii* can even be considered a true living organism or not.

1.7.3. Artificial minimal bacteria

As mentioned previously, the concept of the minimal genome has intrigued scientists for decades. While many analyses have been run on naturally occurring minimal genomes (Hutchison et al., 1999; Koonin, 2003; Lluch-Senar et al., 2015b; van Ham et al., 2003), modern sequencing and biosynthesis technologies, specifically buoyed on by the development of the synthetic biology field, have allowed researchers to enter a new paradigm of discovery with regard minimal genomes: the creation of artificial minimal genomes (Abil et al., 2015; MacDonald and Deans, 2016).

Within this field, there are two major approaches that are being taken towards the creation of an artificial Minimal bacteria. The top down approach focuses on engineering and exploiting already available genetic resources, using the knowledge gained from pre-existing systems to reduce current ones down to their bare-bones necessities. The bottom up approach by contrast aims to create new sequences and organisms *de novo*, engineering genetic circuits to work together to form a viable lifeform (Ausländer et al., 2017, p.).

1.7.3.1. Bottom up engineering –JCVI-syn1.0 and JCVI-syn3.0

The team at the J. Craig Venter Institute (JCVI) has been the most high profile and the most successful at utilising the bottom up approach to creating minimal bacterial cells. Their first major success was the creation of JCVI-syn1.0, the world's first bacterial cell with a genome that was fully synthesised and assembled *de novo* (Gibson et al., 2010). The cell contains a synthetic version of the *M. mycoides* genome, containing 19 unique single nucleotide polymorphisms, an intentionaly deleted 4Kb region and four larger watermark regions. These watermarks contain non-coding DNA to prove that the genome within the cell is indeed the engineered one and not the WT. The entire genome was synthesised as 1080 base pair fragments, containing 80bp overlap regions. The composite genome was then built in a hierarchical manner, with sets of ten 1080 bp fragments producing 109 ≈10Kb fragments. Ten of these larger fragments were then combined to produce 11 ≈100Kb fragments, which in turn were combined to form the full genome, as shown in Figure 12.

For the recombination of the synthetic fragments, each of the first 1080 bp fragments contained a *NotI* restriction site, and the pools of 10 fragments were recombined in the presence of a cloning vector in yeast cells, before being transferred to *E. coli*. Plasmid DNA was then isolated and screened for the requisite 10Kb band, and successful ligations were sequenced to ensure no errors had been included during replication. Correct ligations then had the same protocol applied, pools of 10 fragments were recombined in the presence of a cloning vector and grown in yeast cells. Plasmid DNA was then extracted from the yeast cells and tested for to see that the ligations were successful. The 100Kb fragments were purified to remove and linear yeast chromosomal material, and the final 11 fragments were cloned into yeast cells. Of the 48 colonies screened via PCR, one was found to have the correct ligation. The synthetic genome was then transplanted into a donor *Mycoplasma. capricolum* cell which had its native genetic material removed (Lartigue et al., 2009), and resultant clones were tested via PCR for both the absence of the *M. capricolum* genome and the presence of the synthetic *M. mycoides* genome, wherein the successful cell was named JCVI-syn 1.0 (Gibson et al., 2010).

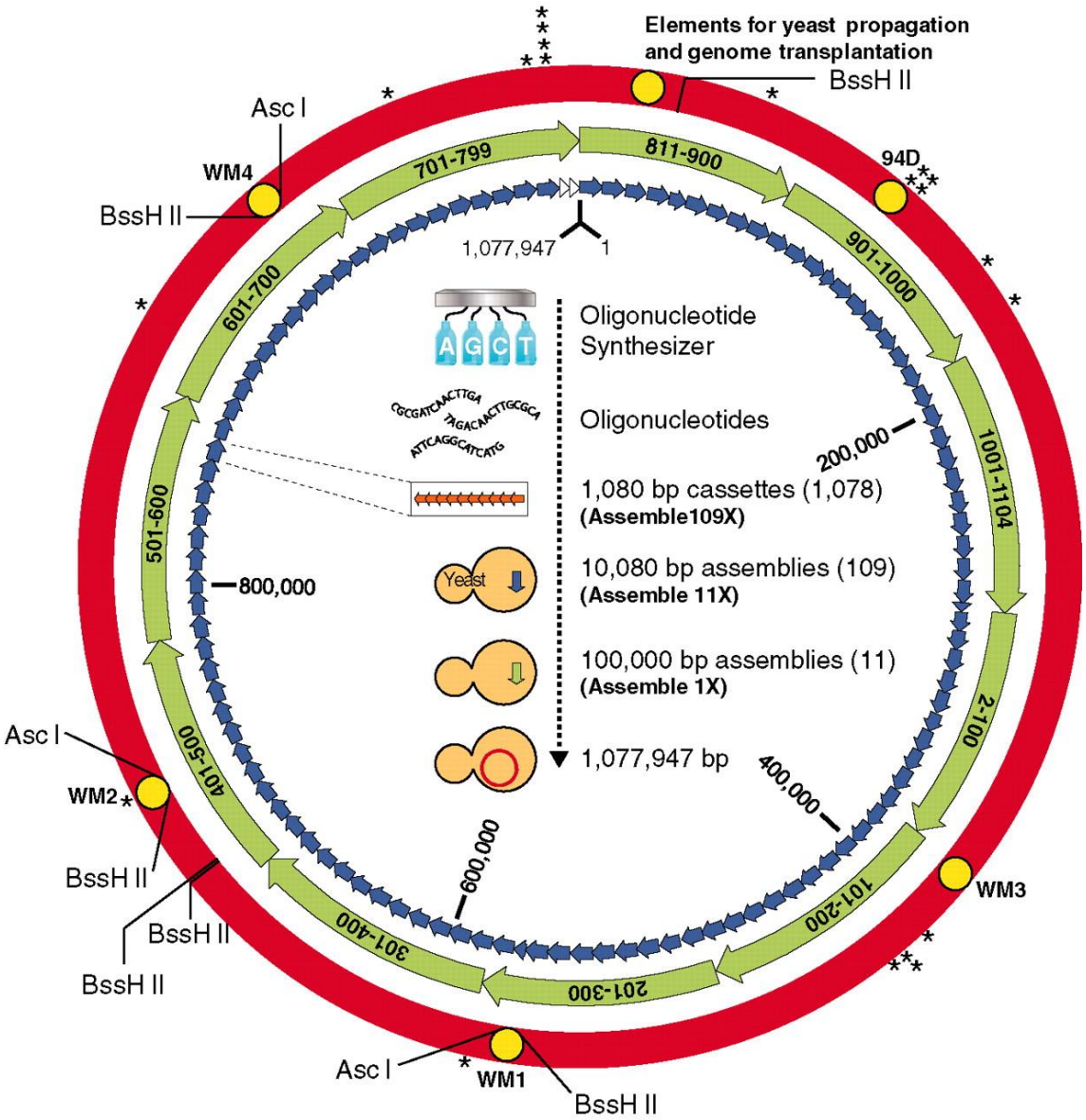


Figure 12: JCVI Syn1.0. The design scheme for the hierarchical construction of the genome, showing the locations of the different sized genomic bands and where they correspond. The yellow circles denote major variations from the WT genome, either as watermarks (WM1-4), the deleted 4Kb region (94D) and the yeast transplantation machinery. The asterisks denote SNPs. Taken from Gibson et al., (2010).

After the success of JCVI-syn 1.0, the same team at JCVI went on to further experiment with the cell-line they had created, in an attempt to minimise the genome as much as possible (Hutchison et al., 2016). By running large transposon mutagenesis studies on the JCVI-syn 1.0, they identified 440 non-essential genes within the genome. Discarding these genes, they designed a genome containing 432 protein genes and 39 RNAs that were deemed to be essential to cellular survival. Using the methods outlined in the creation of the JCVI-syn1.0, a new 438Kb genome was designed.

The original JCVI-syn1.0 genome was split into eight sections, and for each section a new reduced version was created with the non-essential genes removed. Each of the eight sections was then tested separately, with the new minimised section and the seven other sections unchanged. Of these eight cell permeations, only one produced a viable cell, indicating that at least some of the non-essential genes that had been lost were in fact needed for cell viability, and given that 7/8 of the minimisation attempts failed, there was

probably many more than expected. To gather more data, further high quality transposon mutagenesis experiments were run, with both *Tn4001* and *Tn5* transposons. Over 30,000 unique insertion sites in the JCVI-syn1.0 genome revealed many non-essential genes were in fact fitness genes, and while on their own were not essential, losing too many of them was causing the cell to become non-viable (see chapters 1.4.2 and 1.5.4 for further details).

As rationally designing the genome did not work, largely due to the presence of genes of unknown function and lack of knowledge regarding higher epistatic functions, a design/build/test cycle was developed for each of the eight sections of the JCVI-syn1.0 genome, as outlined in Figure 13. Each of the eight sections was mutated via random transposon mutagenesis, and the non-essential genes were removed. The new section was then built and transformed into a new cell. The process was then repeated until the smallest possible size for each of the eight sections was found that still gave robust growth in the recipient cell. Non-essential genes that had important biochemical functions were retained, or those that were surrounded by large regions of essential genes, but roughly 90% of the non-essential genes from JCVI-syn 1.0 were removed.

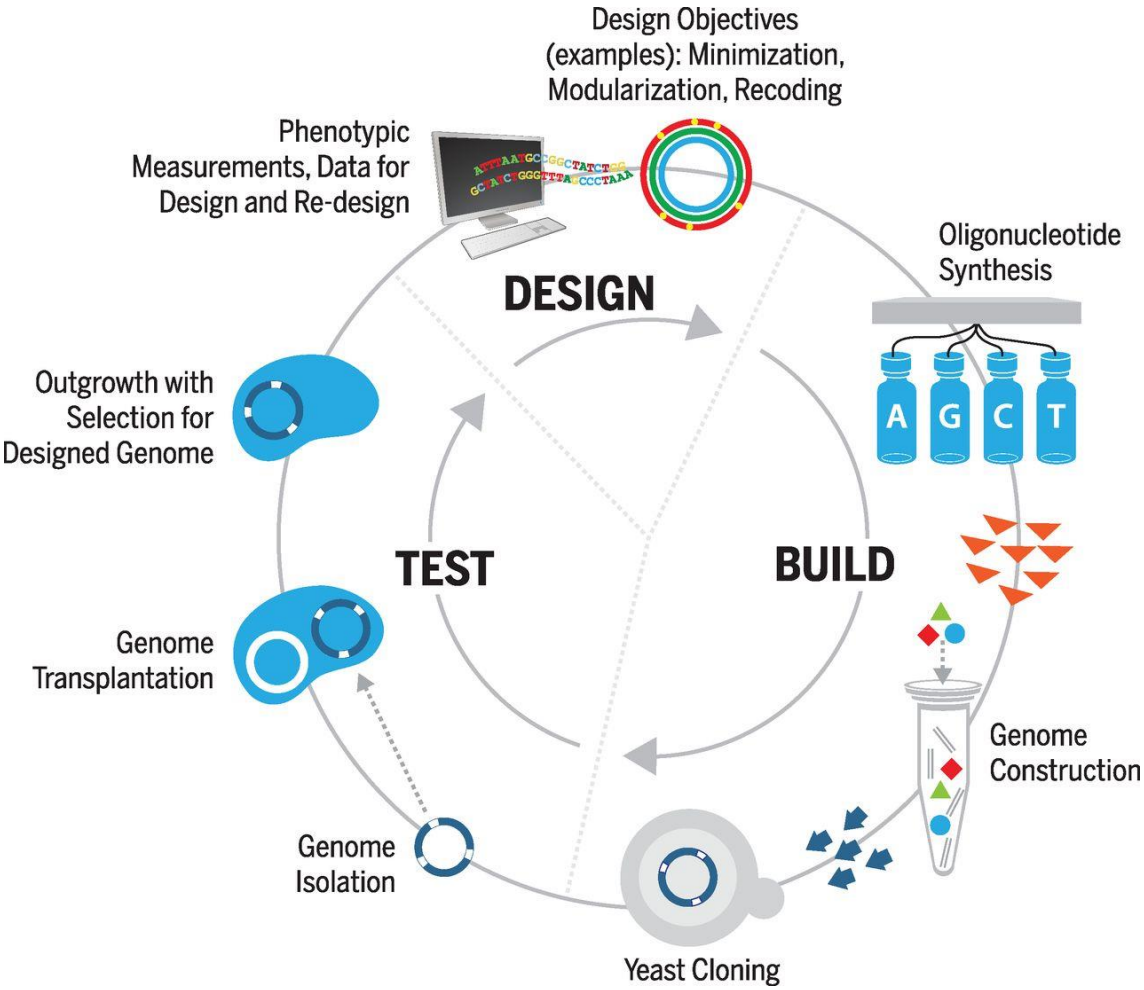


Figure 13: JCVI-syn3.0 Design Build Test cycle. Taken from Hutchison et al., (2016).

The new fragments, known as reduced genome design (RGD) fragments were then assembled in a variety of ways in yeast cells, with the combination of RGD fragments 2, 6, 7 and 8, and JCVI-syn1.0 fragments 1, 3, 4 and 5 giving a strong growth phenotype, known as RGD2678. This RGD2678 genome was then subjected to another round of

transposon mutagenesis, and many previously non-essential genes in JCVI-syn1.0 had turned to fitness or essential genes in this new context. This analysis lead to 26 genes being added back into the remaining RGD fragments 1, 3, 4 and 5, which when created, the combination of the eight RGD fragments produced a viable cell, named JCVI-syn2.0, containing 478 genes and 38 RNAs in a 578 Kb genome.

Finally, the JCVI-syn2.0 genome was again analysed via transposon mutagenesis, and again, previously non-essential genes had become fitness or essential, while many fitness genes became essential. Of these genes, 37 that had previously been classified as non-essential in other experiments, and were still non-essential in JCVI-syn2.0 were removed, along with the beta-lactamase, lacZ and ribosomal RNA genes from the cloning vector. The eight new RGD fragments were then created and cloned together. This process created a viable cell with 438 genes and a genome size of 531Kb, smaller than that of *M. genitalium* and currently the smallest axenic lifeform known, known as JCVI-Syn3.0 (Hutchison et al., 2016).

Despite the large number of minimisation steps the genome went through to get to the JCVI-syn3.0, there are still twelve genes within the genome that are non-essential, and 149 (34%) that have an unknown function. Of the genes that were deleted from JCVI-syn1.0 to JCVI-syn3.0, 65% were either of unknown function, mobile elements or lipoproteins. Due to the rich media the cells were grown in, specifically with the plentiful supply of glucose, many metabolic genes were also lost, with 34 of the 36 genes involved in transport or catabolism of carbon sources other than glucose being removed, yet all 15 glucose related genes persisting. As expected from previous computational models and comparative analyses (Charlebois and Doolittle, 2004; Koonin, 2003) and pre-existing minimal genomes (Glass et al., 2006; Lin and Zhang, 2011; Lluch-Senar et al., 2015b), the largest conserved category of genes belongs to those involved in transcription, translation and DNA replication and repair. In this study they were split into two categories known as “Expression of genome information” and “Preservation of genome information”, which combine to consist of 48% of the genes retained (Hutchison et al., 2016).

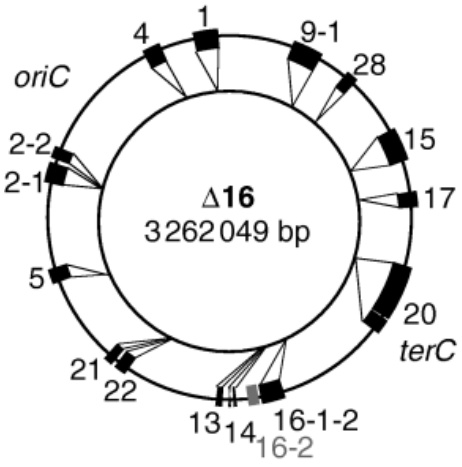
With regard to cell physiology, the original JCVI-syn1.0 retained a highly similar cell morphology to *M. mycoides* (Gibson et al., 2010), and JCVI-syn3.0 retains this similarity. The main difference between the two is in growth speed, JCVI-syn1.0 doubled approximately every 60 minutes, whereas JCVI-syn3.0 has a doubling time of approximately 180 minutes (Hutchison et al., 2016). However, this is still remarkably faster than *M. genitalium*, with its doubling speed of every 960 minutes (Jensen et al., 1996).

1.7.3.2. Top down engineering

While there are a few examples of top down engineering to create minimized genomes from pre-existing cells, they have not been as successful at creating a Minimal genome as the bottom up methods outlined earlier (Xavier et al., 2014a). As the concept of top down engineering relies on removing functions from a pre-existing cell, many of the existing attempts are in well characterised cells.

One of the best examples is in *E. coli*, where 1,377,172 base pairs were removed, accounting for 29.7% of the genome (Hashimoto et al., 2005). Each gene was classified as either essential or non-essential, and from this, large areas of non-essential genes,

known as LD regions, were assigned. The deletions were completed in a step-wise manner, with sixteen separate regions deleted in the order shown in Figure 14. When looking at the physiology of the strains, there is also variation depending on the number of deletions. The parental strain had a doubling time of 26.2 mins, however none of the subsequent deletion strains had an equal or faster growth speed. The strains with ten or fewer deletions all had doubling times of 30 mins ± 2, while after ten deletions the doubling time increases, with the final strain containing sixteen deletions has a doubling time of 45.4 mins. With regard to cell shape, after the fifth deletion, the standard bacilli shape of the cells became shorter and wider, until deletion three where the cells became longer and thinner again (Hashimoto et al., 2005).



Mutant	Deletion units removed (Number)	(bp)	Cumulative deletion (bp)	(%)
Δ1	1	111 734	111 734	2.4
Δ2	5	53 250	164 984	3.6
Δ3	28	52 245	217 229	4.7
Δ4	4	82 372	299 601	6.5
Δ5	2-2	46 153	345 754	7.5
Δ6	2-1	99 304	445 058	9.6
Δ7	16-2	46 517	491 575	10.6
Δ8	16-1-2	109 112	600 687	12.9
Δ9	17	72 554	673 241	14.5
Δ10	22	57 798	731 039	15.8
Δ11	21	42 636	773 675	16.7
Δ12	14	47 675	821 350	17.7
Δ13	13	41 411	816 244	17.6
Δ14	20	300 703	1 116 947	24.1
Δ15	9-1	125 568	1 242 515	26.8
Δ16	15	134 657	1 377 172	29.7

Figure 14: Locations and sizes of deletions in the E. coli genome. Taken from Hashimoto et al., (2005).

A similar example of reduction in the *E. coli* genome by 20% after 69 step-wise deletions (Karcagi et al., 2016a). Here, instead of targeting specifically non-essential regions, they removed genes that are often involved in horizontal gene transmission. In total, 965 genes were deleted, and similar to the study by Hashimoto *et al.*, (2005), there were clear but moderate reductions in fitness after the deletions of many regions, with reductions in both gene size and increase in doubling times (Karcagi et al., 2016a).

Another well-characterised bacterial species is *Bacillus subtilis*, and a similar genome reduction experiment to those performed in *E. coli* have been done (Ara et al., 2007), though the focus of this work was to produce a minimal *B. subtilis* cell that could be used as a protein production platform. Therefore, in the context of minimal genomes, overall size of the genome was weighed against functionality. This is shown clearly when the authors identify the essential and non-essential genes, as the genes were not just assayed for if they caused a lethal phenotype when lost, but also how they effected the cell's overall ability to produce protein. For all genes where the effect of a knockout on protein production was not known, knockouts were generated, and the level of cellulose produced by the culture was assayed compared to the wild type (WT). As a result, 271 genes that were involved in cell growth, protein production and secretion were classified as putative-essential, along with 92 genes whose deletion increase the levels of protein production.

From this data, fourteen regions were identified that could be deleted, totalling 991Kb, or approx. 24% of the total genome. The regions were deleted in a sequential manner, similar

to the *E. coli* experiments, using homologous recombination. However due to the authors aim of the cell being capable of robust protein production, growth rate, and protein production stayed very constant as the deletions progressed, instead of declining, as shown in Figure 15.

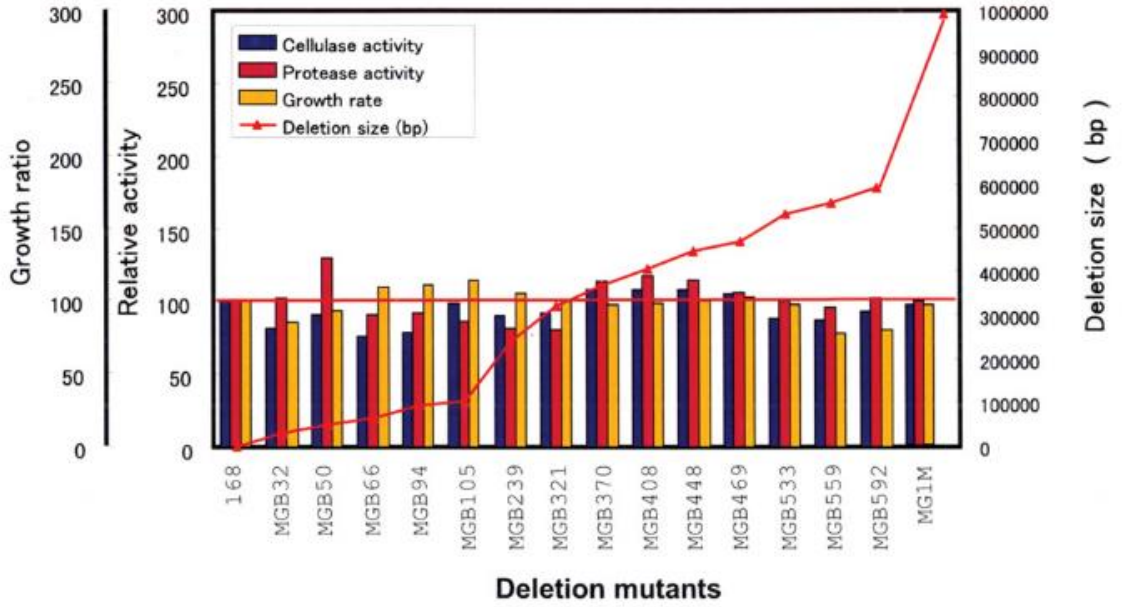


Figure 15: Size and productivity of B. subtilis deletion mutants. The B. subtilis 168 strain and strains with a reduced genome (sequentially deleted multiple regions) were transfected with pHYS237 (for cellulase productivity evaluation) and pHP237-K16 for subtilisin-like alkaline protease (B. clausii-KSM-K16-strain-derived) productivity evaluations. Cultures were shake-cultured in modified 2×L-Mal medium at 30◦C for 75 h, evaluated for cellulase (blue bars) and protease (red bars) productivity, and measured for the degree of growth (yellow bars) at 600 nm. Red lines indicate the length of gene deletions. Taken from Ara et al., (2007).

Other studies in *B. subtilis* have gone even further, with 36% of the genome being removed (Reuß et al., 2017). Here, the authors produced two similar strains, with genomes of 2.76 and 2.68Mb, deleting 1553 and 1605 genes respectively over 88 and 94 separate deletions. The deletions were based on the group's previous work identifying the all of the genes necessary for the growth of *B. subtilis* at 37°C. Due to the relatively low number of essential genes in *B. subtilis*, the authors focused on important pathways and reactions alongside essentiality, similar to the method used to created the JCVI-syn3.0 cell (Hutchison et al., 2016; Reuß et al., 2016). Large regions containing non-essential genes involved in sporulation, antibiotic resistance, unknown functions, non-glucose carbon sources and motility were targeted, and cells were evaluated on fitness after every deletion. Those deletions that either increased the fitness burden considerably or were not tolerated were either modified or retained in the genome (Reuß et al., 2017).

In the two cell lines produced, proteomic analysis showed that despite the fact that essential genes accounted for only 9% of the total genes present in each species, they accounted for 50% and 51% of the proteome. Predictably, a large fraction of these proteins, approximately 33%, relate to ribosomal proteins, aminoacyl-tRNA synthetases and transcription factors (Reuß et al., 2017), this pattern of strong retention of transcription and translation machinery an emerging theme across multiple orthology and genome reduction studies (Charlebois and Doolittle, 2004; Gibson et al., 2010; Glass et al., 2017; Hutchison et al., 2016; Koonin, 2003; Lluch-Senar et al., 2015b).

CHAPTER 2: RANDOM DELETIONS IN MYCOPLASMA PNEUMONIAE

Genome reduction is an important strategy in the synthetic biology toolbox, specifically in reagard to building a minimal chassis cell. It is one that allows us to both study the function and interaction of genetic systems, but also to develop cells that take advantage of the knowledge gleaned to improve their functionality. This chapter focuses on the development of a new strategy for genome deletions in *Mycoplasma pneumoniae*, and will explain the development and iterations of the protocol as it evolved, along with its validations.

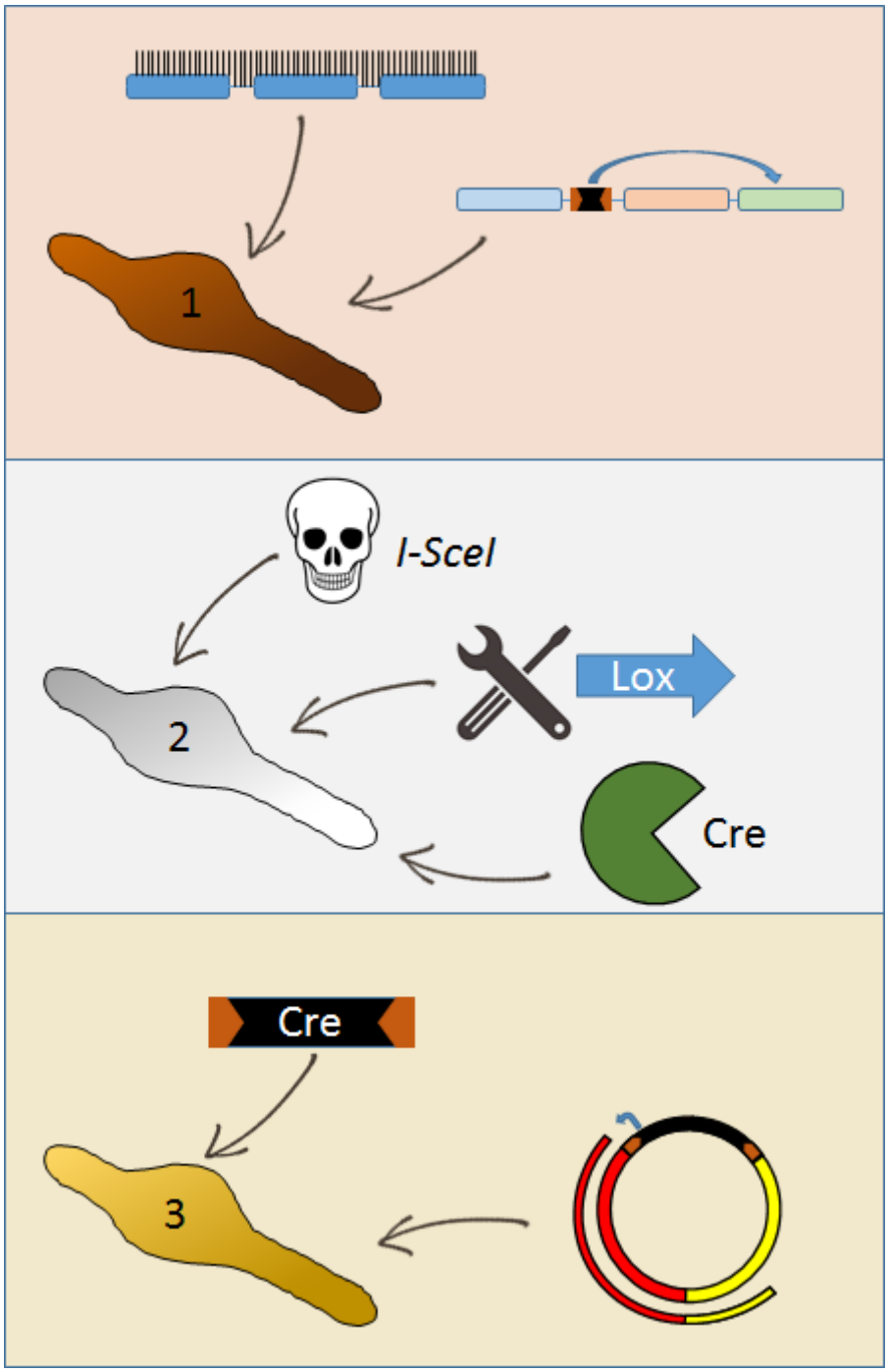


Figure 16: Graphical overview of the main iteration steps leading to each of the three protocols used in this chapter

2.1. Background and rationale

Many methods have been used to reduce the genome in bacteria and eukaryotes (Hutchison et al., 2016; Karcagi et al., 2016b; Reuß et al., 2017; Shen et al., 2016). However most of them have relied on essentiality maps built by transposon mutagenesis, which contain issues with cconfounding factors (as explained in chapters 1.4.1and 1.4.4). Mainly, the problem with this approach is that it gives a static view of which genes are dispensible to growth in a population of cells. Despite the broad view of which genes are essential or not, factors such as epistasis cannot be fully accounted for. This can explain the failure of Hutchison *et al* to produce a viable cell when the annotated non-essential genes were deleted in a logical manner (Hutchison et al., 2016). Ideally, having a map with the different roads that could lead to a minimal cell would facilitate finding the most feasible trajectory. To address this issue, we decided to create a protocol allowing for the random deletions of large genomic regions from the genome reduced bacteria *Mycoplasma pneumoniae.* The protocol involved the fusion of two molecular biology techniques, random transposon mutagenesis and the Cre-Lox recombinase system. For an overview of random transposon mutagenesis, see chapter 1.3.2.1. This study utilised the DNA transposon Tn4001 (Reddy et al., 1996), as it has shown to work with high efficiency in *M. pneumoniae* (Hedreyda et al., 1993).

M. pneumoniae was chosen as the model organism for this genome reduction study over the smaller *Mycoplasma genitalium* for three main reasons. First, there is extensive information of the gene identity (Fadiel et al., 2007; Halbedel and Stülke, 2007; Hasselbring et al., 2006b; Himmelreich et al., 1996; Reddy et al., 1996), metabolism and '-omics' (Blötz and Stülke, 2017; Breuer et al., 2019; Güell et al., 2009b; Kühner et al., 2009; Lluch-Senar et al., 2015a; Maier et al., 2013; Schmidl et al., 2010; van Noort et al., 2012; Wodke et al., 2015; Yus et al., 2019, 2009), genome architecture (Trussart et al., 2017), cell biology (Balish, 2014; Dybvig and Voelker, 1996; Hasselbring et al., 2006a; Parrott et al., 2016; Razin, 1985; Razin et al., 1998; Razin and Hayflick, 2010; Waites and Talkington, 2004) and genetic essentiality (Lluch-Senar et al., 2015b). This compendium of knowledge is not available in the same level of detail in *M. genitalium*, and thus the characterisation of deletion mutants can be better understood against a background of higher knowledge.

The second reason *M. pneumoniae* was used was while it is a true minimal genome, its larger size comes with benefits for a reduction study. *M. genitalium* has a genome of approximately 482 genes of which 382 are essential, giving just 21% of the genome as non-essential (Glass et al., 2006). In contrast, *M. pneumoniae* has a genome of approximately 700 genes of which 342 are essential, giving 52% of the genome as non-essential (Lluch-Senar et al., 2015b). The fact that *M. pneumoniae* contains both a much larger contingent of non-essential genes, and a large genome in general implies that there are more possibilities for combinatorial deletion. Due to their highly evolved nature, it is highly unlikely that drastic gene loss is possible in either species, however *M. pneumoniae's* increased genetic repertoire may allow it to produce a higher variation of deletion mutants, and with that the potential for clones that have an acceptably robust metabolism and growth speed, and specific properties.

The final reason was due to the fact that *M. pneumoniae* has a faster growth speed, dividing on average every 8 hours compared to *M. genitalium* which has a doubling speed of every 16 hours (Jensen et al., 1996). As the main endeavour of this chapter was to

create and validate a new methodology, it was felt that working with *M. pneumoniae* would allow for faster data collection and error correction. On top of that, many step-wise genome deletion studies have shown a gradual loss of fitness to the cell, especially in regard to growth speed, as deletion occur (Hashimoto et al., 2005; Karcagi et al., 2016b). In light of this, it was felt that further moderate to severe loss of fitness and growth speed could more tolerated from a practical point of view in *M. pneumoniae* than it would be in *M. genitalium*.

The rationale behind using a random process to delete non-essential regions of the genome instead of a rational approach was two-fold. First, when the project began we lacked the tools to create targeted insertions or deletions within the *M. pneumoniae* genome. While this was clearly critical, our aim was always to use a random manner regardless. Employing a random methodology removes any biases towards the deletion of specific regions brought about by our incomplete knowledge of the function of every gene, and more importantly the effects of epistasis on gene maintenance and redundancy. Due to the lack of information on double or triple knock out mutants, there may be many genes that are non-essential in the wild type genome, but as other regions get removed these genes impart a larger and larger fitness to the cell, eventually becoming strong fitness genes or even essential. By allowing for a protocol that produces multiple random deletions within a population, we not only can identify all the large and small genomic regions that are amenable to deletion, but also via competition select for those cells that have the strongest or most robust growth characteristics.

While the deletion of individual or groups of genes will clearly have a large fitness effect, as has been shown by multiple deletion experiments (Hutchison et al., 2016; Karcagi et al., 2016b; Reuß et al., 2017), another key cause of fitness loss could be disruption to the genome architecture. The genome of *M. pneumoniae* is arranged in two halves, with the origin and mid-point of replication at polar ends of the genome. Disruption to this well defined layout, especially if it causes changes to the patterns of supercoiling present, could lead to large changes in the level of transcription of many of the genes near the deletions site (Trussart et al., 2017). If larger regions are deleted, this could also cause a noticeable miss-balance in the amount of genetic material in each ‘half’ of the genome between the origin of replication and the mid-point, potentially causing changes to the genome replication efficiency.

While this methodology, being top down deletion strategy, is unlikely to produce a minimal genome on par with the JCVI-syn3.0 in terms of pure genetic minimisation (Glass et al., 2017), we can characterise the effects and implications of certain deletions on the viability and fitness of subsequent cells. By completing multiple deletions, we can locate specific genes, operons or regions on the chromosome that allow for subsequent deletions to occur, or in contrast non-essential regions whose deletion has a strong epigenetic impact and subsequent deletions show poor fitness. By tracking the order of which the deletions take place, we can also potentially identify synthetic lethality pairs, and begin to unravel the effects of epistasis on the organisation and composition of a minimal genome. These insights can in turn help us when rationally designing future minimal genomes.

2.1.1. Cre Lox system

The Cre Lox system is a recombinase system isolated from the P1 bacteriophage (Sternberg et al., 1981). It consists of three components, the Cre recombinase protein and two 34 base pair lox sites within the DNA. The Cre recombinase is a 34kDa protein that mediates the reaction between the two lox sites. These are 34 base pair sites contain an eight base pair central spacer region, flanked by 13 base pair palindromic sequences. This central spacer region gives the lox sites their directionality, and the interactions between themselves and the Cre recombinase is dependant on this (Yan et al., 2008).

The directionality that the lox sites have in relation to each other predicates the outcome of the reaction with the Cre. As shown in Figure 17, when the lox sites are in *cis* orientation, the DNA between the lox sites is deleted, becoming circularised with lox site retained in the genome and one lox site within the circularised DNA. This reaction can happen in the reverse direction as well, though the efficiency is far lower. When the lox sites are in *trans*, then the DNA is inverted between the two sites (Kühn and Torres, 2002).

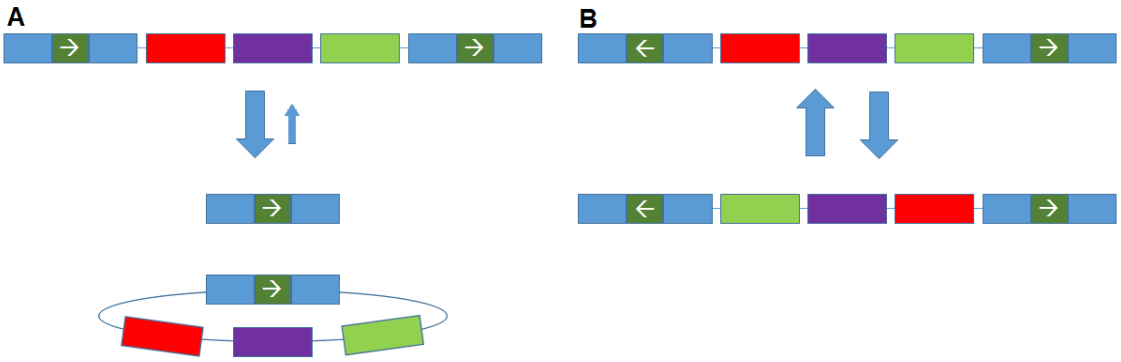


Figure 17: Effect of Lox site orientation on the product of the Cre recombinase reaction. (A) Lox sites in cis give a circularised deletion product and a genomic lox site. (B) Lox sites in Trans invert the DNA between Lox sites

The WT configuration has the palindromes of the lox arms unaltered, known as a LoxP site. However, there are many known mutations of the lox sites that can alter their activity. The Cre recombinase can recognise and act upon a lox site if only one of the two arms is mutated, however if both arms of the lox site are non-canonical the Cre can no longer recognise the site, making it functionally silent (Oh-McGinnis et al., 2010). The current literature shows many, often contradictory, labels for these mutant lox sites, so for sake of brevity and simplicity, I will refer to them with the names and orientations given in Figure 18. The Left Element lox site (LE-Lox) contains a five base pair mutation in the 5' end of the site, and conversely the Right Element Lox (RE-Lox) contains the same mutation in the 3' arm of the lox site. The double mutant lox site (lox72) contains both mutations, and is functionally silent within the genome.

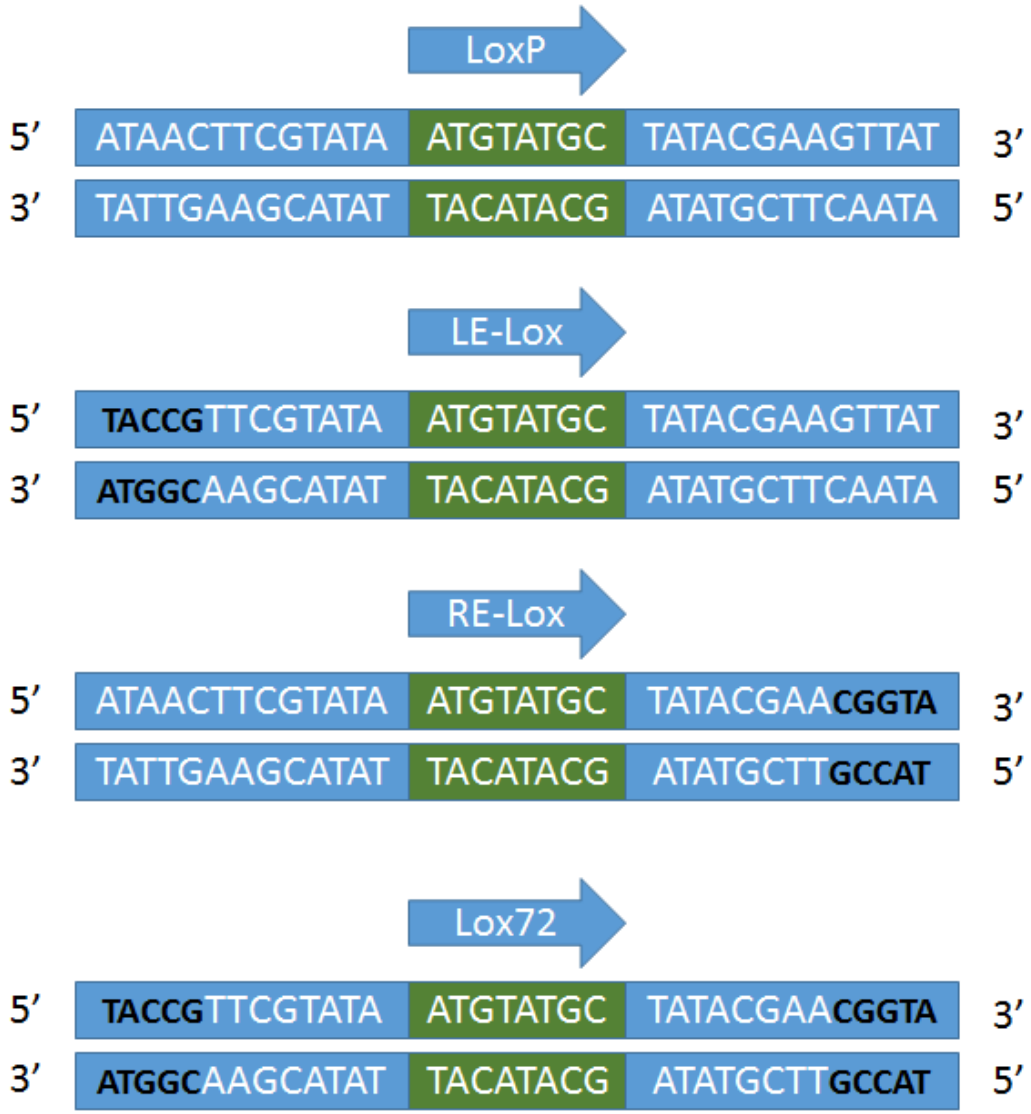


Figure 18: Sequence and orientation of the lox sites used in this work. The core spacer region is shown in green and the palindromic arms in blue. Black text indicates mutations from the WT loxP site, and the arrow above each lox site indicates its orientation.

The reaction pathway of the Cre Lox deletion can be found in molecular detail here (Van Duyne, 2001). In summary, Figure 19 shows the outline of the process, along with how two single mutant lox sites, in this case LE-Lox and RE-Lox can delete a section of genomic DNA and form a silent double mutant Lox72 site within the genome. It requires four Cre molecules to complete the reaction, as each independent palindromic arm is bound by a single Cre protein. These then act as nickases, cutting the spacer region of the leading strand at the central AT dinucleotide. This exposes a hydroxyl group and a phosphotyrosine, which are recombined via a Holliday junction-mediated recombination, with the leading strand of the first lox site binding to the leading strand of the second. This process then repeats for the lagging strand, and the DNA between the Lox sites is circularised, containing a Lox site that consists of the 3’ arm of the first Lox site, and the 5’ arm of the second Lox site. The genome also contains a lox site, consisting of the 5’ end of the first Lox site and the 3’ end of the second (Kühn and Torres, 2002; Van Duyne, 2001).

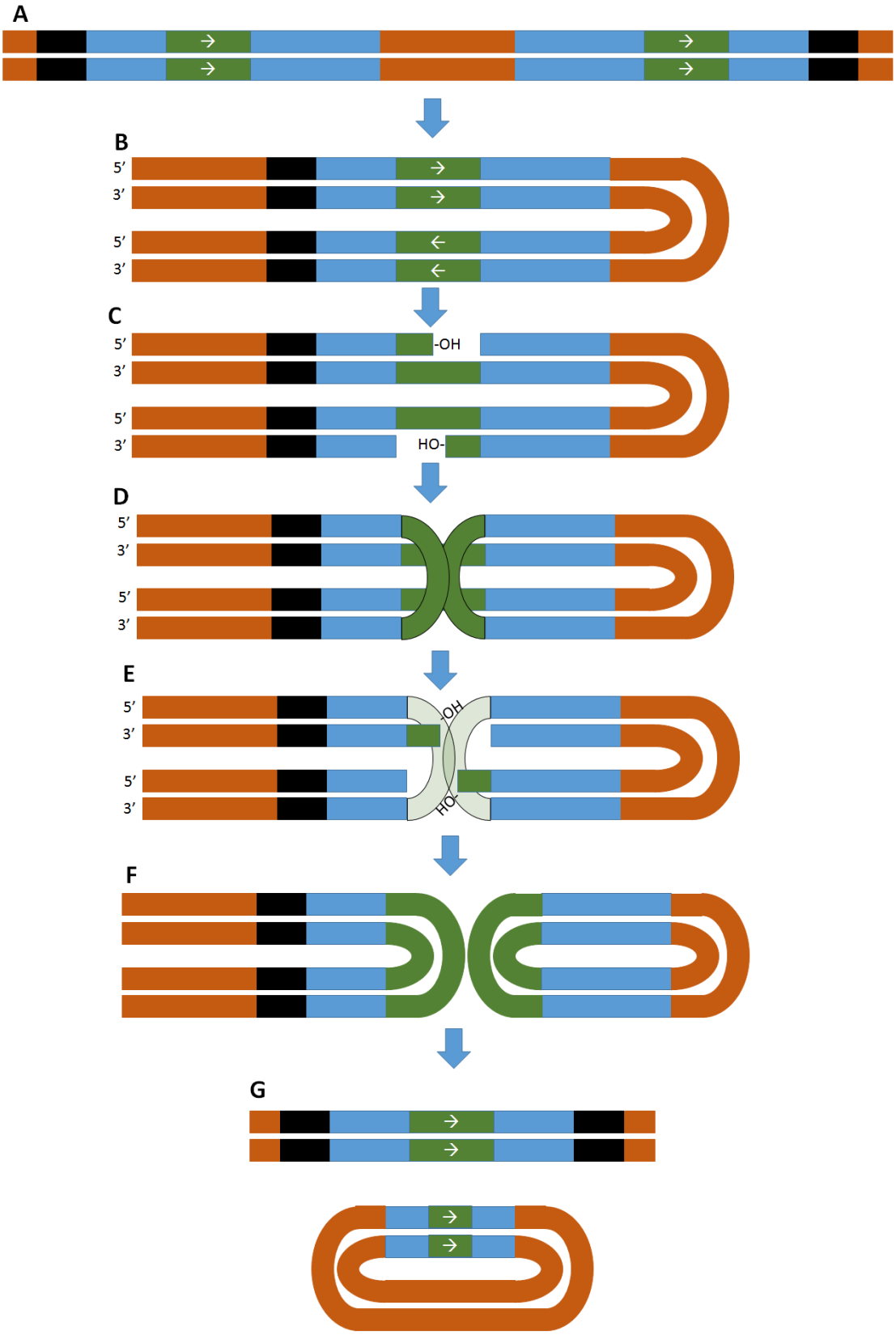


Figure 19: Cre Lox deletion and creation of a double mutant Lox site. (A) The genome, in orange, contains a LE-Lox and RE-Lox in cis orientation. Using the same configuration as Figure 18, the central spacer regions are in green, with an arrow indicating orientation, the palindromic arms in blue and the mutant regions in black with the mutant arms shown in black. (B) The Cre recombinase brings the two Lox sites together. (C) The Cre recombinase cuts the leading strand at the 5′ strand of the spacer region, exposing the OH groups. (D) The hydroxyl groups bond to the exposed phophotyrosine groups on the opposite strand, forming a Holliday junction. (E) The Cre recombinase repeats for the

lagging strand. (F) The host DNA ligase repair the Holliday junctions. (G) A double mutant lox72 site is created in the chromosome, and the DNA between the original two lox sites in circularised, containing a LoxP site.

2.1.2. Tn4001

The Tn4001 is a DNA transposon that is a close relative of the Tn5. It was first characterised in *Staphylococcus aureus* conferring resistance to gentamicin, tobramycin and kanamycin (Lyon et al., 1984). The WT Tn4001 has a similar structure to the Tn5, with two inverted sequences flanking an *aacA-aphD* gene, responsible for the gentamicin resistance phenotype. Unlike the Tn5 however, where only one of the inverted regions contains an active transposase, both of the inverted flanking regions contain an active transposase gene, making this system highly active. Tn4001 transposons are active in a wide range of Gram-positive species, and most importantly for this study have been shown to be active in a wide range of Mycoplasmas (Cao et al., 1994; Hedreyda et al., 1993; Prudhomme et al., 2002; Reddy et al., 1996). It is not the only transposon that is shown to be active in Mycoplasma, the Tn916 has been shown to be effective in a wide range of Mollicutes (Cao et al., 1994, p. 916; Dybvig and Alderete, 1988; Dybvig and Cassell, 1987). However, its 19Kb size makes it an unnecessary burden in molecular cloning.

Tn4001 transposons inserted more preferentially into AT sites within the genome, allowing for good potential coverage of the *M. pneumoniae* genome. They also allow for stable insertion, as mini-transposons variants have been created and demonstrated in Mycoplasma (Pour-El et al., 2002). Therefore, the basis of all transformation protocols in this chapter that use transposition rely on the mini-Tn4001 transposons.

2.1.3. Rationale for Protocol 1

This protocol relied on the delivery of a lox site into two random loci in the *M. pneumoniae* genome via transposon mutagenesis. The two transposons would deliver the lox sites, which through chance would be in either a *cis* or *trans* orientation. Each transposon contained a selective antibiotic marker and an *I-Sce I* restriction site. The *I-SceI* enzyme is a yeast derived mega nuclease with a recognition site of 'TAGGGATAACAGGGTAAT', which is not found in the WT *M. pneumoniae* genome. The enzyme has been shown to effectively cut double stranded DNA in spiroplasmas (Breton et al., 2011), and is used as a counter selective agent. A suicide vector with an antibiotic resistance is then added containing the Cre recombinase and *I-SceI* proteins, which act upon the genome in the as shown in Figure 20:

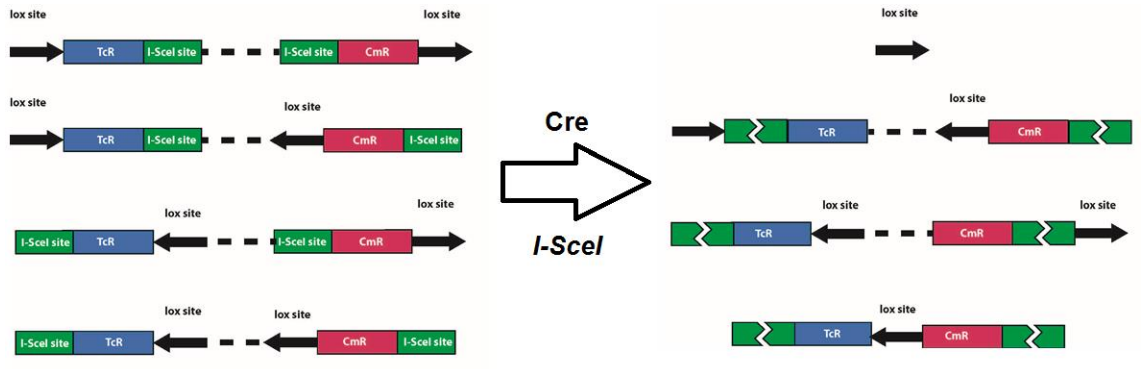


Figure 20: Possible recombination events in Protocol 1

If the lox sites are in a *trans* orientation, then the DNA between the sites becomes inverted. The *I-SceI* sites are still present, and the action of the mega nuclease causes two double strand breaks in the DNA, which are lethal to the cell. Therefore, after transforming the cells with the suicide vector that contains the Cre and *I*-SceI proteins, the only cells to survive the process are those that had their lox sites integrated in a *cis* orientation and caused a deletion, removing the *I-SceI* recognition sites. This does also remove half of the successful deletions, but ensures the resultant pool of cells are pure deletions. The resultant lox72 site is functionally silent, as it has a drastically lower affinity to the Cre protein (Kühn and Torres, 2002; Suzuki and Nakayama, 2011, p. 72), thus will not interfere with future rounds of deletions. As the antibiotic resistance markers are removed from the cells, the system can then be recycled.

This protocol was the first attempt at creating deletions in the WT_{M129} cell line. As described above, two transposons were used to integrate separate lox sites and *I-SceI* restriction sites into the genome. Plasmid $pMTnCm^{66}.2$ contained a transposon with a chloramphenicol selectable marker along with a lox66 site in the 3' of the transposon, and the $pMTnTc^{71}$ contained a transposon with a tetracycline resistance and a lox71 at the 5'. Deletion of the region between the lox sites was instigated via the use of suicide plasmid pBSK_pM438_Cre_Sce_Puro containing the Cre recombinase gene, the *I-SceI* gene as a selection agent against Cre mediated inversions and selected for by a puromycin resistance gene.

2.1.3.1. Ramifications of jumping transposons

To test if this method was viable, first we had to ensure that the transposons were stable inside the genome after multiple exposures to the transposase gene. As the *Tn4001* keeps the sequence of the inverted repeats intact when it transposes its DNA cargo (Dybvig et al., 2000), it is possible that the transposase from one of the plasmids could recognise an *in situ* transposon that had been inserted in a previous transposition event, and act upon it. This could potentially move the pre-existing transposon to a new region of the chromosome.

If this phenomena occurs, then it could allow for a simplified deletion protocol. Both lox sites in the protocol could be standard loxP sites, as they would not need to be silenced after a deletion event. If a deletion occurred, but the lox sites could jump throughout the genome, then the integration of the new lox site after the deletion would move the previous one. This prevents bottlenecking from lox sites that integrate near an essential gene, as they have the potential to be able to move to a new non-essential region of the genome. It would also increase the amount of random movement within the genome, allowing for less bias towards the initial insertion site. Therefore, the ability of the Tn4001 transposons present in the genome to be affected by both other transposons and suicide vectors containing the transposase gene was queried before protocol one was begun.

2.1.3.2. Transposition density

The other main efficacy barrier we predicted was the transformation efficiency of the transposons. The higher the transformation efficiency of the protocol, the more likely it is to have insertions in the regions of the genome amenable to deletions, and the more variation in deletions we can obtain. Previous work had indicated that high transformation efficiencies in *M. pneumoniae* could be achieved using the Tn4001 transposon, with the

original work by Hedreyda et al., (1993) giving yields of $1x10^7$ to $1x10^8$ colony forming units per transformation. This protocol uses 30µg of DNA per transformation however, which makes day-to-day cultivation of plasmid DNA far more laborious than ideal. Recent experiments in our lab have shown that $1x10^7$ CFU per transformation can be achieved with just 1.5µg of plasmid DNA (Montero-Blay et al., n.d.). For our two transposons, 1pMole of DNA equates to 3 – 3.8µg of DNA, depending on the size. It has also been shown in the lab that transformations of 5µg of DNA or above drastically increase the chance of multiple transposons entering the same cell, with ≈10% of cells containing two insertions (Burgos. R, personal communication). As this would innitate a lethal phenotype after Cre exposure, we decided not to increase the volume of DNA further.

2.1.4. Rationale for Protocol 2

In the second iteration of the protocol, we focussed on fixing the lox sites and ensuring that the individual components of the system were operating as expected. This involved testing the activity of the Cre recombinase and *I-SceI* gene to ensure that both were functioning properly, and trying to use the new combination of lox sites to achieve deletions within the genome.

We also decided to allow for the selection of truly non-essential insertions during the transposon mutagenesis phase. The hypothesis was if we allow multiple passages after each transposon is inserted, then those cells that grow the best would contain insertions that have the least fitness deficit, thus are more amenable to deletion. As such, we allowed for three passages after the insertion of each lox transposon to allow this selection to occur.

Due to the mis-labelling of the lox sites in Protocol 1 (see chapter 2.4 for specifics), new vectors were designed with lox sites in the correct orientations. To avoid the confusion of different terminologies that caused the initial confusion, the mutant lox sites were labelled as Left Element lox (LE_Lox) and Right Element lox (RE_Lox). The double mutant retained the lox72 label for easy distinction from the other two sites. The orientation of the lox sites was also adjusted to ensure that the deletion that formed the inactive lox72 site removed the antibiotic resistance and *I-SceI* sites. This resulted in the creation of two new transposon vectors, labelled LE_Cm and Tc_RE. In order to track the progression of the deletions if multiple rounds are possible, each LE_Cm vector contained a generational barcode, consisting of the four bases immediately upstream of the LE_Lox site.

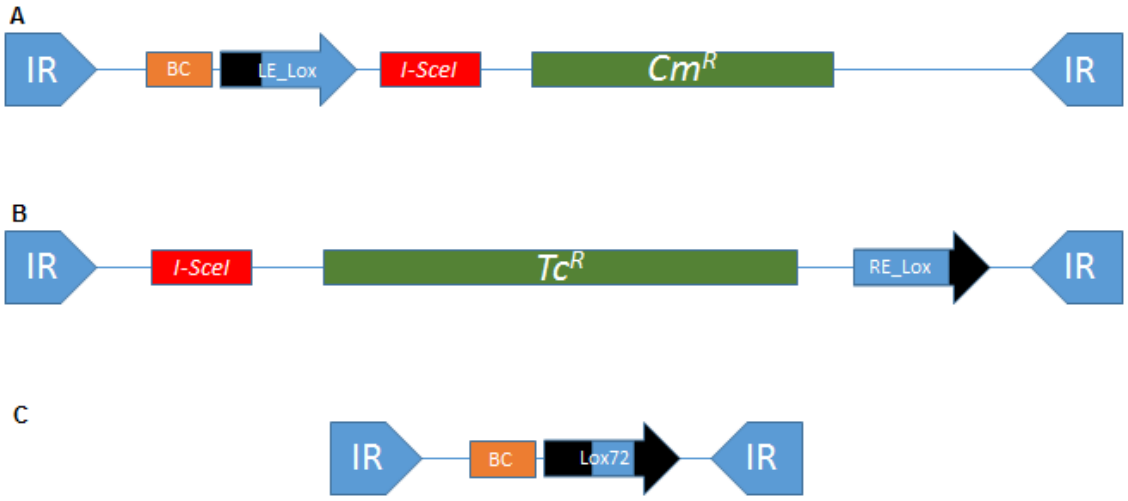


Figure 21: Location and orientation of transposons used in Protocol 2. A: LE_Cm transposon containing chloramphenicol resistance (CmR), I-sceI site, Left Element lox site and generation barcode (BC). B: Tc_RE transposon containing tetracycline resistance gene (TcR), I-SceI site and Right Element Lox site. C: Scar formed when action of the Cre recombinase of the two sites in cis causes a deletion, resulting in an inactive lox72 site and generational barcode

The barcode show in in orange in Figure 21 consists of the four bases directly upstream of the left element lox site. As such, they will be retained in the genome after a successful deletion as part of the scar, as shown in Figure 21 C. As each round of deletions progresses, the barcode will be altered, following the pattern in Table 8. These barcode will allow the temporal tracking of deletions within genome of each cell, as the barcode in each scar will give the order each deletion occurred in, and the number of cells that harbour the deletion.

Table 8: Generation deletion barcodes

Deletion Round	Barcode
1st	ATCG
2nd	CCGG
3rd	AATT
4th	GGAA

2.1.4.1. Cre and *I-SceI* testing

To ensure that the Cre recombinase and *I-SceI* proteins are active within the system, suicide vectors containing each enzyme independantly were designed. The efficacy of the *I-SceI* at killing the mycoplasma strains containing restriction site but not the WT_{M129} needed to be tested, though had been reported in spiroplasmas (Breton et al., 2011).

It was also hypothesised that if the *I-SceI* is very effective at killing the Mycoplasma cells, the cells could be dying before the Cre can act, thus lowering the amount of potential deletions. Therefore, as there already existed a suicide vector in the lab that containined the Cre recombinase alone (pBSK_pM438_Cre_Gm), its ability to induce deletions and inversions was tested.

2.1.5. Rationale for Protocol 3

With the successful deletion of two regions in protocol 2, it was clear we were able to delete both small and large genomic regions using the Cre recombinase. However, the efficiency of the system was still low, and thus modifications to the protocol were needed if this was to become a viable tool for large scale genomic streamlining.

The first major alteration was changing the Cre delivery system from a suicide plasmid to via its own transposon. Having the Cre constitutively active within the genome will hopefully ensure that the lethal effect imposed upon the cell when a single lox site is present is amplified, instead of potentially being lost as a suicide plasmid, allowing for a more powerful counter selective agent. As by itself the Cre appeared to be as powerful as the *I-SceI* as a counter-selective agent, we elected to continue with using just the Cre, as it can evidently both induce deletions and select for them. However, including the Cre gene in the chromosome means that it needs to be removed before a second round of deletions can begin, as the activity of the Cre will kill cells with a single lox site, which would occur in the first round of transformations for the second round.

As such, the transposon containing the Cre recombinase and gentamicin resistance contained mutant VLox sites. The Vlox/VCre system is a paralogue of the Cre/Lox system, but the two do not share any cross-talking ability (Suzuki and Nakayama, 2011), thus the Cre recombinase cannot recognise the Vlox sites, and the VCre cannot recognise lox sites. Therefore, when the deletions have occurred and surviving cells propagated under gentamicin selection, they can then remove the Cre cassette with induction from the VCre suicide vector. As the cells have propagated already, this should drastically reduce the bottleneck effect of the suicide vector.

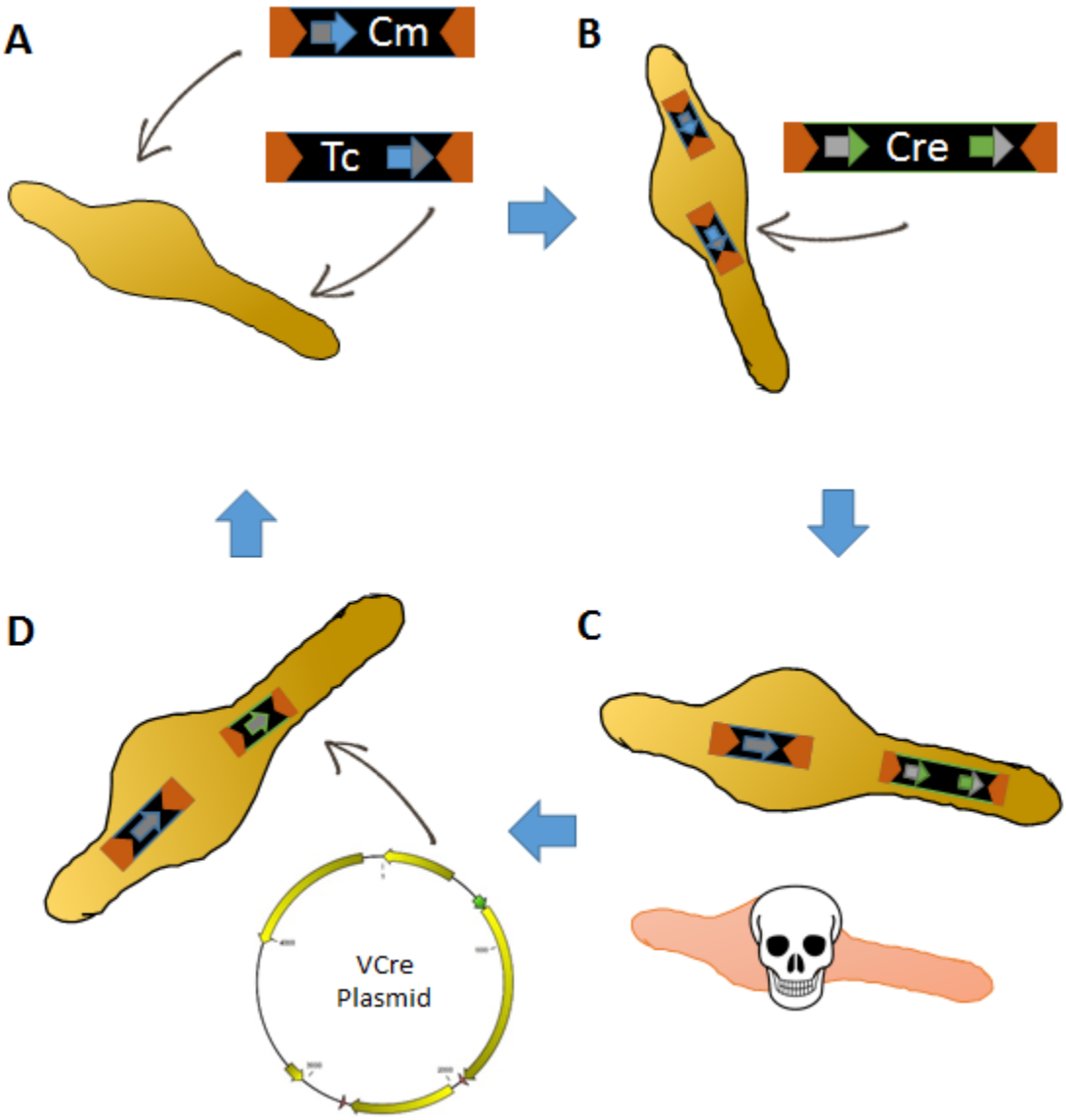


Figure 22: Protocol 3 outline. (A) WT_{M129} is transformed with the LE_Cm and Tc_RE vectors in series, with only one passage between the transformations. (B) Cells resistant to both chloramphenicol and tetracycline are transformed with the Cre transposon, flanked by left element and right element VLox sites (in green), allowing for deletions and inversions to occur. (C) Due to the lethal activity of the Cre on single lox sites, only cells that successfully deleted a genomic region to form an inactive lox 72 (in grey) survive. (D) Surviving cells that have a deletion are then grown out and are transformed with a suicide vector containing the VCre gene, to remove the Cre from the genome. This ensures the bottleneck happens after the deletions have occurred.

To ensure the highest possible variation of potential deletion targets, we also removed the extra passaging steps of the cells between transformations. While this will include more non-viable permutations, it is also introducing an unacceptable level of bias to the protocol. As the aim of the project is to develop a method that allows for unbiased genome deletions, selecting for cells that have the least fitness loss over multiple insertions inevitably biases against the possibility of strong positive epistatic effects. There are chances that knocking out one gene in a non-essential operon could cause strong loss of fitness, but removing the whole operon could have a net positive result. By biasing the selection after the first transformation event, we drastically lower the chances of seeing these events. Therefore, as soon as cells were grown to mid-log phase post transformation, the next transformation was initiated.

2.1.5.1. Custom Next-generation sequencing protocol

While current standard protocol such as HITs and Tn-Seq are highly adept at identifying insertion sites of transposons, (see chapter 1.4.1), they are more limited in regards to identifying deletions. Standard next-generation sequencing of the pools could be done in two ways, either i) using a standard Mi-Seq protocol (Bronner et al., 2009) and extracting reads that happen to contain the deletion scar, or ii) amplifying the DNA containing a deletion via PCR then sequencing the amplified DNA. While the first method would give the DNA sequence of the genome upstream and downstream of the deletion scar, thus giving an accurate read of the deletion size, the lack of specificity would ensure large numbers of deletions go un-mapped. The second method allows for much higher coverage of the deletions due to the PCR amplification, but this amplification ensures that the sequencing can give us the location of only one side of the deletion. This means that it is impossible to deduce the size of any deletions, as you cannot match which read came from which cell, thus which read matches to the other half of the deletion.

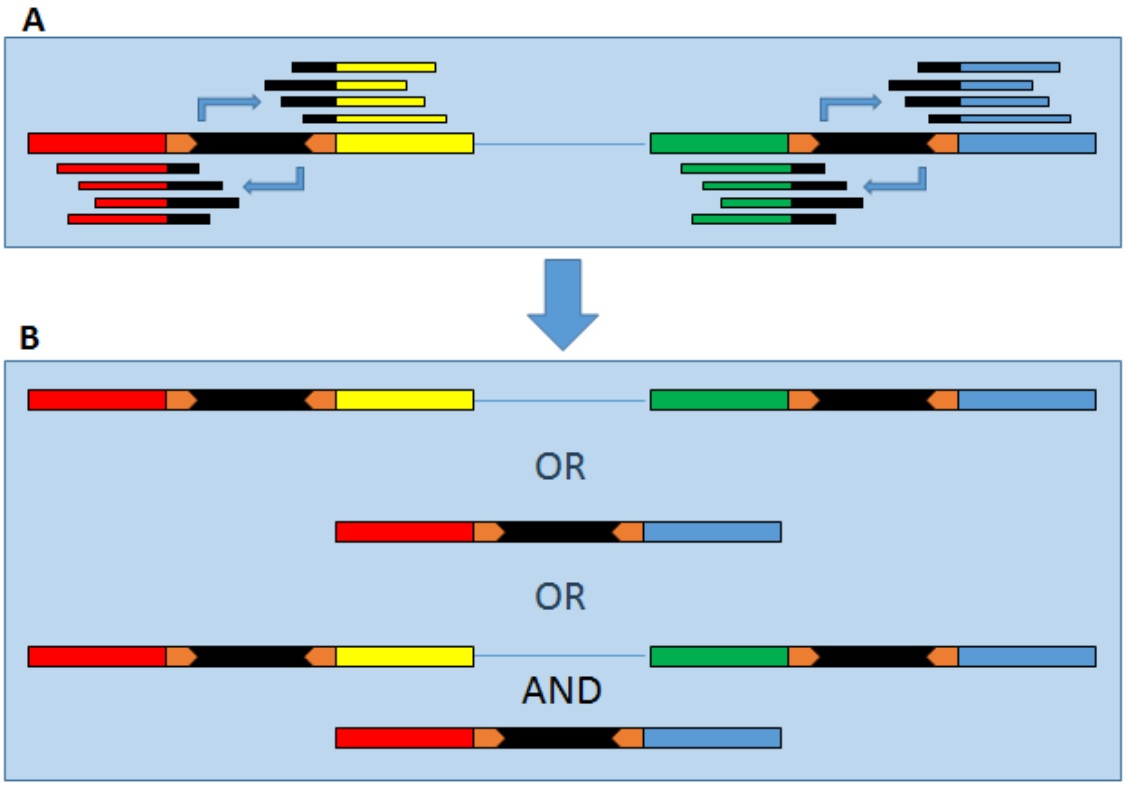


Figure 23: Issues with traditional sequencing methods in regard to deletion identification. (A) The location of two separate deletions, with the reads typically that would be generated. Black sections represent the deletion scar, orange arrows the inverted repeats and coloured regions genomic DNA. (B) The three possible ways to recombine the data, giving the two deletions, one large deletion encompassing both or a combination of the two.

As shown in Figure 23, if you have two deletions in two separate cells, you cannot tell from the reads generated where the true deletion is, and spurious deletions could be mapped.

As such, we designed a high-throughput sequencing protocol for the identification of deletions with a pool or homogenous mutants. This method relied on circularising the DNA before sequencing, to ensure that DNA from both sides of the scar was present. Briefly, genomic DNA from the pool of random deletions was isolated, then sonicated to 300bp. The DNA was then circularised and amplified using an oligo specific to the lox72

site in a circular PCR to enrich for fragments containing a deletion scar. The amplified DNA was then sequenced via a standard 125bp paired-end sequencing protocol using an Illumina Hi-Seq 2500 using an oligo inside the deletion scar. As the DNA was circularised from both sides of the scar, the sequencing should read the point where the circularisation occurred, thus contain DNA regions from both sides of the scar, allowing the size and scope of the deletion to be observed, as shown in Figure 24:

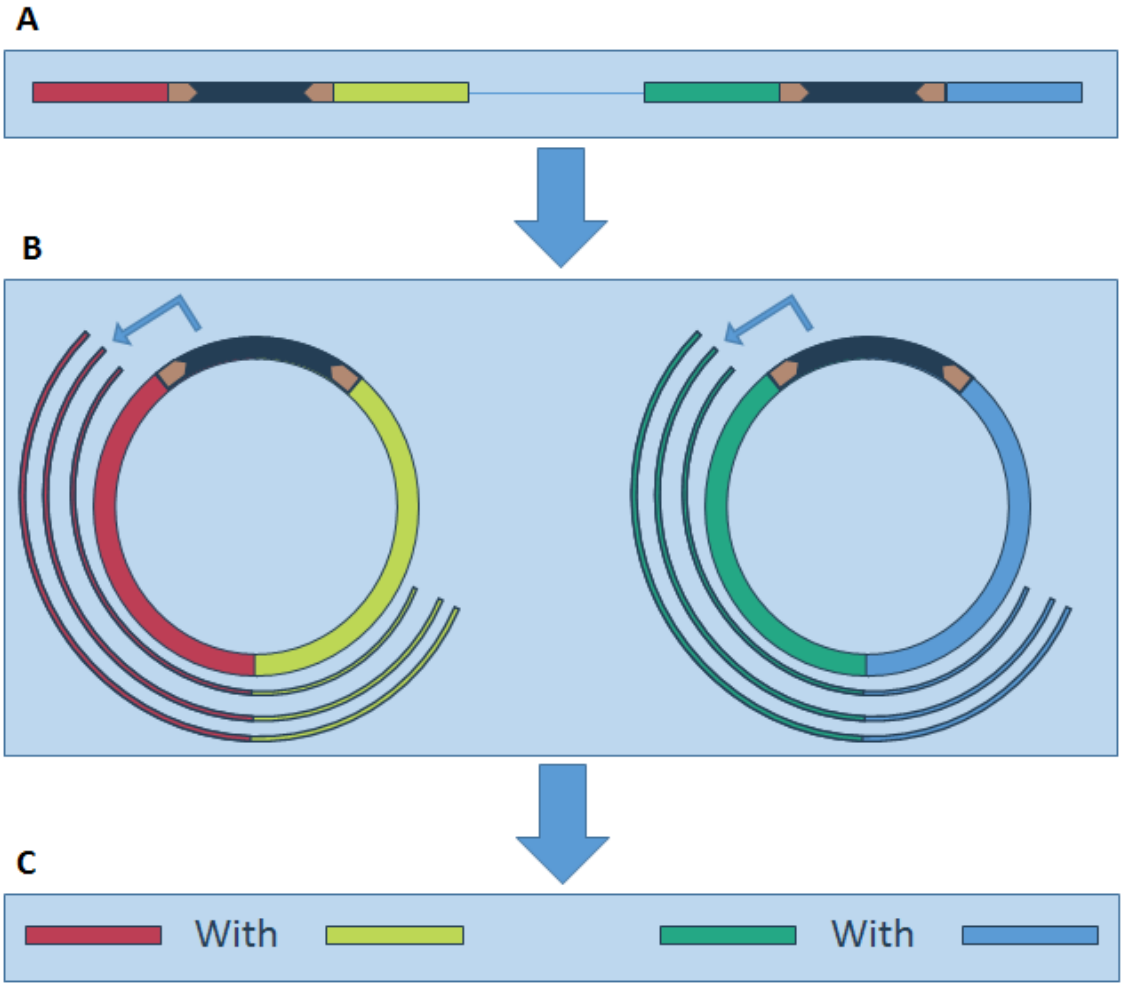


Figure 24: Circularisation protocol for sequencing deletion regions. (A) The location of two separate deletions, with the reads typically that would be generated. Black sections represent the deletion scar, orange arrows the inverted repeats and coloured regions genomic DNA. (B) The genomes are fragmented to 300bp and circularised. An oligo inside the deletion scar amplifies the genomic DNA in a circular PCR, which is then sequenced. (C) Fragments showing two disparate regions of the genome can be assumed to have been brought together via a deletion event. Therefore, even if both IRs are not present, the approximate size and scope of the deletion can be elucidated.

2.2. Materials and Methods

2.2.1. Strains and culture methods

NEB 5-alpha Competent *E. coli* cells (New England Biolabs, Catalogue number C2987H) were used for plasmid amplification and cloning. They were grown at 37°C in Lysogeny Broth (LB) at 200RPM or static on LB agar plates, supplemented with 100µg/ml ampicillin.

The WT *M. pneumoniae* strain used was M129 (ATCC 29342, subtype 1, broth passage no. 35), as described by Regula et al., (2000), hereon in referred to as WT_{M129}. Cultures were grown in Hayflick media at 37°C, previously described by Hayflick, (1965) and Yus et al., (2009), supplemented with 2µg/ml tetracycline, 3.3µg/ml puromycin, 200µg/ml gentamicin or 20µg/ml chloramphenicol as appropriate.

Mpn_A37 is an isolated clone from a *M. pneumoniae* M129 strain transformed with a cassette containing genes coding for the Non-homologues end joining machinery isolated from *Bacillus subtilis* and a chloramphenicol resistance marker. The culture was grown in Hayflick media at 37°C, supplemented with 20µg/ml chloramphenicol.

2.2.2. DNA manipulations

Plasmid DNA was isolated via the QIAprep® Spin Miniprep kit (QIAGEN, Catalogue number 27106). Genomic DNA from the Mycoplasma strains was isolated via the MasterPure™ Complete DNA and RNA Purification Kit (Lucigen, Catalogue number MC85200). DNA fragments were amplified via use of the Phusion™ Hot Start II High-Fidelity DNA Polymerase (ThermoFisher, Catalogue number F549S). Digested DNA samples and PCR products were isolated via the QIAquick® PCR Purification Kit (QIAGEN, Catalogue number 28106). Purified PCR fragments were isolated from agarose gels via the QIAquick® Gel Extraction Kit (QIAGEN, Catalogue number 28706). Visualisation of the DNA within the gel was accomplished using a GelRed stain (Biotium, Catalogue number 41003). All DNA oligonucleotides were synthesised via Sigma-Aldrich, and purified using reverse phase, shipped in water and at a concentration of 50µM. Sequencing of DNA samples was undertaken via Sanger sequencing using GATC Biotech. Next generation sequencing was performed at the Genomics Facility at the CRG.

2.2.2.1. Molecular cloning

All plasmids were constructed using the Gibson assembly method (Gibson et al., 2009), with the master mix provided by the Biomolecular Screening Protein Technologies Facility at the CRG, unless otherwise stated.

The pBSK_Gent_TPA plasmid was constructed using restriction digestion and T4 ligation, with enzymes purchased from New England Biolabs (*XhoI*: catalogue number R0146S, *PtsI*: catalogue number R0140S, T4 DNA ligase: catalogue number M0202S, T4 DNA ligase buffer: catalogue number B0202S).

pBSK_Gent_TPA The pMTnCat plasmid was amplified with oligos 47 & 48 creating a 1.2Kb fragment. The pBSK_Cre_Gm plasmid and PCR fragments were digested with *XhoI* and *PstI*, creating a 4.4Kb fragment and 1.2Kb fragment. Both fragments were

digested with *DpnI* and isolated via gel electrophoresis. The fragments were annealed using a T4 ligase reaction, creating the pBSK_Gent_TPA plasmid.

pMTnCm^{66}.2 Plasmid pMTnCm^{66} was amplified with oligos 75 & 76, and oligos 77 & 78 to create two fragments of 4.2Kb and 0.7K respectively, which were then purified via *DpnI* digestion. The samples were isolated via gel electrophoresis and added ligated via Gibson assembly to create plasmid pMTnCm^{66}.2

LE_Cm Re-named version of the original pMTnCm^{66}, to make nomenclature more standardised.

Tc_RE The Tc^{71} plasmid was amplified with oligs 189 & 206, and oligos 188 & 207, creating fragments ≈ 4.2Kb and 2Kb respectively. Theses were digested with *DpnI* and isolated via gel electrophoresis. The two fragments were ligated via Gibson ligation to create the Tc_RE plasmid.

pBSK_pM438_Gm The pBSK_pM438_Cre_Gm vector was amplified with oligos 200 & 201, creating a 4.3Kb fragment. The fragment was digested with *DpnI* and isolated via gel electrophoresis. It was then self-annealed via Gibson ligation.

pBSK_pM438_Puro The pBSK_pM438_Cre_Sce_Puro vector was amplified with oligos 202 & 203, creating a 3.5Kb fragment. The fragment was digested with *DpnI* and isolated via gel electrophoresis. It was then self-annealed via Gibson ligation.

pBSK_pM438_Sce_Puro The pBSK_pM438_Cre_Sce_Puro was amplified with oligos 139 & 140, creating a 4.2Kb fragment. The fragment was digested with *DpnI* and isolated via gel electrophoresis. It was then self-annealed via Gibson ligation.

pMTn_VL_Cre_Gm_VL The pBSK_pM438_Cre_Gm plasmid was amplified via oligos 305 & 306 and the Tc_RE plasmid was amplified with oligos 303 & 304, giving 2.7Kb and 4.2Kb fragments respectively. The fragments were digested with *DpnI* and isolated via electrophoresis. They were then ligated via Gibson ligation to form intermediary plasmid pMTn_Cre_Gm_VL. This plasmid was then amplified with oligos 307 & 308 to give a 7Kb fragment. This was digested with *DpnI* and isolated via electrophoresis, then self-annealed via Gibson Ligation to form pMTn_VL_Cre_Gm_VL.

pBSK_VCre_Sce_Puro The pBSK_pM438_Sce_Puro was amplified via oligos 309 & 310 and the pBSK_Genta_VCre_Mut was amplified via oligos 311 & 312, giving 1.3Kb and 4.2Kb fragments respectively. The fragments were digested with *DpnI* and isolated via electrophoresis. They were then ligated via Gibson ligation to form pBSK_VCre_Sce_Puro.

Table 9: Plasmids used in this study

Plasmid Name	Description	Reference
pMTnTetM438	pMTn4001 containing a tetracycline resistance gene under the control of the 22 bp pM438 promoter	Pich et al., 2006

pMTnCat	A derivative of the pMTnTetM438, with the tetracycline resistance gene swapped for a chloramphenicol resistance gene	Burgos and Totten, 2014
pMTnGm	pMTn4001 containing the *aac(6')-aph(2'')* gentamicin resistance gene	Pich et al., 2006
pMTnCat_NHEJ	Kindly provided by Dr Piñero from our lab, a derivative of the pMTnCat transposon containing the non-homologous end joining (NHEJ) genes *ykoU* & *ykoV* taken from *Bacillus subtilis* along with a chloramphenicol resistance gene.	This study
pBSK_pM438_Gent_TPA	Derivative of the pBSK_pM438_Cre_Gm with the Cre recombinase gene swapped for the transposase gene from pMtnCat	This study
pMTnCm^{66}	Mini *Tn4001* transposon based on the pMTnCat, containing a chloramphenicol resistance gene and a lox66 site in the 5' end of the transposon	This study
pMTnCm^{66}.2	Mini *Tn4001* transposon based on the pMTnCat, containing a chloramphenicol resistance gene and a lox66 site in 3' end of the transposon	This study
pMTnTc^{71}	Mini *Tn4001* transposon containing a tetracycline resistance gene and a lox71 site in the 5' end of the transposon	This study
pBSK_pM438_Cre_Gm	pBSK derived suicide plasmid containing the Cre recombinase gene, and *aac(6')-aph(2'')* gentamicin resistance gene	This study
pBSK_pM438_Cre_Puro	pBSK derived suicide plasmid containing the Cre recombinase gene, and puromycin resistance gene	This study
pBSK_pM438 _Sce_Puro	Derived from pBSK_pM438_Cre_Sce_Puro, has the Cre recombinase gene removed to leave just the *I-SceI* and Puromycin resistance	This study
pBSK_pM438_Cre_Sce_Puro	pBSK derived suicide plasmid containing the Cre recombinase gene, *I-SceI* mega nuclease gene and puromycin resistance gene	This study
LE_Cm	Mini *Tn4001* transposon based on the pMTnCat, containing a	This study

	chloramphenicol resistance gene and a LE_Lox site, *I-SceI* site and 1st generation barcode in 3' end of the transposon	
Tc_RE	Mini *Tn4001* transposon containing a tetracycline resistance gene and a RE_Lox site in the 3' end of the transposon	This study
pBSK_pM438_Puro	Empty suicide vector based on the pBSK_pM438_Cre_Sce_Puro, containing only the puromycin resistance gene	This study
pMTn_VL_Cre_Gm_VL	Transposon derived from the pMTnTc backbone, containing the Cre recombinase and gentamicin resistance inside the inverted repeats. Containing a left element mutant VLox site at the 5' and Right Element mutant Vlox site in the 3'.	This study
pBSK_pM438_VCre_Sce_Puro	Derived from the pBSK_pM438_Sce_Puro, also containing the VCre gene as a suicide vector.	This study

Table 10: Oligonucleotides used in this study

Primer Name	Sequence (5' to 3')
Oligo 11	GGCCGTAATATCCAGCTGAA
Oligo 47	TGCTCTCGAGAATTGTGTAAAAGTAAAAAGG
Oligo 48	GCACCTGCAGCTAGTCTACTTATCAAAATTGATG
Oligo 66	ATGAATTACAACAGTACTGC
Oligo 73	CACGAAGAGAAGAAGGAAGC
Oligo 74	TGCAGGCCTTATTATTTTCC
Oligo 75	TCGTATAATGTATGCTATACGAAGTTAT CGCTTTTACACAATTATACG
Oligo 76	CTATTCTATGTACCTGAATC ATATCAAGCTTATCGATACCG
Oligo 77	GATTCAGGTACATAGAATAG TAGGGATAACAGGGTAATTAGTATTTAG
Oligo 78	CACGAAGAGAAGAAGGAAGC
Oligo 74	AGCATATCGTATGTAATATGCTTGCCAT GTGGATCGGATCCTTACG
Oligo 98	TAATTGTGTAAAAGGGCC
Oligo 99	GTATAATTGTGTAAAAGCGTACC
Oligo C50	TACATGCATCTTACCACCCG
Oligo C51	GGTTGATCTAAATTGTGGCG
Oligo 139	CGGCCAGTGAATTGTAATAC GGAAGATGGCGATTAGATCG
Oligo 140	GTATTACAATTCACTGGCCG
Oligo 188	GGTATAGGGATAACAGGGTAATTAG

Oligo 189	TACCCTGTTATCCCTATACC
	TCAAGCTTATCGATACCGTC
Oligo 200	CTGCAAGGCGATTAAGTTGG
	TAGATCGAATTCCTGCAGC
Oligo 201	CCAACTTAATCGCCTTGC
Oligo 202	CGGCCAGTGAATTGTAATAC
	TAACGAATTCCTGCAGCC
Oligo 203	GTATTACAATTCACTGGCCG
Oligo 206	ATAACTTCGTATAATGTATGCTATACGAACGGTA
	ATCCACTAGTTCTAGAGCGG
Oligo 207	TACCGTTCGTATAGCATACATTATACGAAGTTAT
	GATCCCTAAGTTATTTTATTGAAC
Oligo 212	GATAAAGTCCGTATAATTGTGTAAAA
Oligo 213	TTTTACACAATTATACGGACTTTATC
Oligo 307	CGTGATTCTGAGAACTGTCATTCTCGGAAATTGA
	CGGCCAGTGAATTGTAATACG
Oligo 308	TCAATTTCCGAGAATGACAGTTCTCAGAATCACG
	TCAAGCTTATCGATACCGTCG
Oligo 309	GGCGTAATCATGGTCATAGC
Oligo 310	CGAAATTAACCCTCACTAAAGGG
Oligo 311	TTTAGTGAGGGTTAATTTCG
	ATACGACTCACTATAGGGCG
Oligo 312	GCTATGACCATGATTACGCC
	TAAATACTAGGATCCCCCCG
LE-RE_Lox72	CCCTCGAGGTCGAC*G*G*T
Loxloop-6	GCATA*C*A*T
Loxloop-7	GCAT*A*C*A
Loxloop-8	GCA*T*A*C

2.2.3. Transformation of *M. pneumoniae*

WT_{M129} cells were transformed in line with the protocol described by Hedreyda et al., (1993). Cells grown to mid-log phase, indicated by the change of colour in Hayflick media from red to orange. The media was decanted and the flask was washed 3x with 10ml chilled electroporation buffer (EB: 8mM HEPES, 272nM sucrose, pH 7.4). Cells were scraped into 500µl chilled EB and homogenised via 10x passages through a 25-gauge syringe needle.

Aliquots of 50µl of the homogenised cells were mixed with a pre-chilled 30µl EB solution containing the required plasmid DNA. Samples were then kept on ice for 15 mins. Electroporation was done using a Bio-Rad Gene Pulser set to 1250 V, 25 µF and 100Ω. After electroporation, cells were incubated on ice for 15 mins, then recovered into a total of 500µl Hayflick media and incubated at 37°C for 4 hours. 125µl of transformed cells were then inoculated into T75 culture flasks containing 20ml Hayflick and supplemented with the required antibiotic.

2.2.4. Recovery of transformation mutants

To ensure both planktonic and attached cells were recovered, the following centrifugation protocol was employed to recover all transformation mutants. The media from each flask

was decanted into a 50ml flacon tube to recover any planktonic cells. The cells attached to the base of the flask were scraped into 500µl Hayflick media and added to the Falcon tube. The cultures were centrifuged at 10,000RPM for 10 mins at 4°C. The supernatant was discarded and the pellet was re suspended in 500µl fresh EB. The cells were then transformed following the protocol described in Chapter 2.2.3.

2.2.5. Purification of Mpn_A37

The Mpn_A37 transformation pool was provided by lab member Carlos Pinero after transformation of WT_{M129} with the pMTnCat_NHEJ transposon. The culture was serial diluted to 10^{-6} in Hayflick media and plated onto Hayflick agar plates, supplemented with chloramphenicol. Cultures were grown for seven days, and isolated colonies were picked and inoculated into a T25 culture flask containing 5ml Hayflick media at 37°C, supplemented with 2µg/ml chloramphenicol.

Cultures were grown to mid-exponential phase, and cells were harvested via scraping into 500ml fresh Hayflick media. A 100µl aliquot was taken and genomic DNA was isolated which was sent for Sanger sequencing with Oligo 66, annealing inside the transposon, to identify the location of the insertion site.

2.2.6. Quantification of transposon jumping

A culture of Mpn_A37 was transformed with the 500fMoles of the pMTnGm, pBSK_pM438_Cre_Gm and pBSK_pM438_Gent_TPA plasmids, as described in Chapter 2.3. 125µl of transformed cells were then inoculated into T75 culture flasks containing 20ml Hayflick and supplemented with 200µg/ml gentamicin. The pBSK_pM438_Cre_Gm and pBSK_pM438_Gent_TPA transformations were incubated at 37°C for 5 days, and the pMTnGent transformation was grown at 37°C until mid-log phase.

At each end-point, the media from each flask was decanted into a 50ml flacon tube to recover any planktonic cells. The cells attached to the base of the flask were scraped into 500µl Hayflick media and added to the Falcon tube. The cultures were centrifuged at 10,000RPM for 10 mins at 4°C. The supernatant was discarded and the pellet was re suspended in 500µl fresh Hayflick media. All cultures were serial diluted to 10^{-6} and plated on Hayflick plates supplemented with gentamicin to isolate single clones, and a 50µl aliquot from each transformation had its genomic DNA isolated. Five isolated clones from each transformation were picked and inoculated into a T25 culture flask containing 5ml Hayflick supplemented with gentamicin. They were grown to mid-log phase and genomic DNA was isolated from each.

2.2.7. Quantification of transposon density

A culture of WT_{M129} was transformed with the 1pMole of the pMTnCm^{66}.2 plasmid, as described in Chapter 2.2.3. 125µl of transformed cells were then inoculated into a T75 culture flask containing 20ml Hayflick and supplemented with chloramphenicol. The culture was grown to mid-log phase then cells were isolated via the centrifugation protocol outlined in Chapter 2.2.4. Genomic DNA was isolated and sent for sequence using a standard 125bp paired-end read library preparation protocol for an Illumina Mi-Seq. Insertion sites for the transposon were identified using the Oligo 11 which bound to

a region directly downstream of the inverted repeat sequence, and identified using BLAST against the *M. pneumoniae* M129 genome.

2.2.8. Genome deletion using Protocol 1

WT_{M129} cells were transformed via electroporation, as described previously in Chapter 2.3. WT_{M129} cells were grown to mid-log phase and transformed with 1pMole of pMTnCm[66].2. The culture was incubated at 37°C for 4 hours post-transformation in 500µl antibiotic-free Hayflick media. 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with chloramphenicol. The culture was labelled R0.1 and incubated at 37°C.

When the R0.1 culture reached mid-log growth phase, cells were harvested via the scraping and centrifugation method described in Chapter 2.4 to capture both attached and planktonic cells. Cells were then transformed via the standard protocol with 1pMole of pMTnTc[71]. After 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with tetracycline and chloramphenicol. The culture was labelled R0.3 and incubated at 37°C.

When the R0.3 cells were grown to mid-log phase, the harvesting and transformation protocol above was repeated with 1pMole of the pBSK_pM438_Cre_Sce_Puro. After 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with puromycin. The culture was labelled R0.3 and incubated at 37°C for 5 days, and the culture was labelled R1.0.

After 5 days of selection under puromycin, the cells were isolated via the centrifugation protocol. The resultant pellet was suspended in 500µl antibiotic-free Hayflick media, and serial diluted down to 10^{-5}. 100µl of the original stock was inoculated into a T75 tissue culture flask containing 20ml antibiotic free Hayflick media and incubated at 37°C. The serial dilutions were plated onto Hayflick plates and incubated at 37°C. Both the liquid and plate cultures were labelled R1.0.

When the R1.0 culture reached mid-log phase, an aliquot was taken and genomic DNA was isolated. 20 individual clones were picked from the plates and inoculated into T25 tissue culture flasks containing 5ml Hayflick and incubated at 37°C. When they were grown to mid-log phase, aliquots of each were taken for genomic DNA extraction.

2.2.8.1. Identification of strains harbouring a deletion

Both the extracted genomic DNA from the R1.0 pool, and the genomic DNA from the 20 isolated clones were tested via PCR for the presence of a 132 base pair scar left behind by the recombination of the two lox sites by the Cre. Oligos 98 and 99 anneal just inside of the inverted repeats of the transposons and used to amplify the region, giving a 109bp band if positive.

As a control, all samples were tested via PCR reactions that amplified the chloramphenicol resistance gene from the pMTnCm[66].2 transposons, as well as amplifying the *glpD* gene (MPN051) to ensure genomic amplification was successful.

2.2.9. Protocol 2

2.2.9.1. Genome deletions using protocol 2

WT_{M129} cells were transformed via electroporation, as described previously in Chapter 2.2.3. WT_{M129} cells were grown to mid-log phase and transformed with 1pMole of LE_Cm. The culture was incubated at 37°C for 4 hours post-transformation in 500µl antibiotic-free Hayflick media. 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with chloramphenicol. The culture was labelled P0.1 and incubated at 37°C.

When the P0.1 culture reached mid-log growth phase, cells were harvested via the scraping and centrifugation method described in Chapter 2.4 to capture both attached and planktonic cells. The cells were suspended in 500µl Hayflick and 100µl was inoculated into a tissue culture flask containing 20ml Hayflick supplemented with chloramphenicol. This process was repeat twice more to allow for 3 passages of cells.

When the third passage reached mid-log phase, cells were isolated by the described centrifugation method and were transformed via the standard protocol with 1pMole of Tc_RE. After 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with tetracycline. The culture was labelled P0.3 and incubated at 37°C. As with the P0.1 cells, P0.3 cells were passed three times in Hayflick containing tetracycline and chloramphenicol.

When the P0.3 third passage cells were grown to mid-log phase, the harvesting and transformation protocol above was repeated with 1pMole of the pBSK_pM438_Cre_Gm. After 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with gentamicin. The culture was labelled P0.3_Cre and incubated at 37°C for 5 days.

After 5 days of selection under gentamicin, the cells were isolated via the centrifugation protocol. The resultant pellet was suspended in 500µl antibiotic-free Hayflick media, and serial diluted down to 10^{-5}. 100µl of the original stock was inoculated into a T75 tissue culture flask containing 20ml antibiotic free Hayflick media and incubated at 37°C. The serial dilutions were plated onto Hayflick plates and incubated at 37°C. Both the liquid and plate cultures were labelled P0.3_Cre. When the P0.3_Cre culture reached mid-log phase, an aliquot was taken and genomic DNA was isolated.

2.2.9.2. *I-SceI* efficacy test protocol

A 50µl aliquot of P0.1 cells was grown in a 75 tissue culture flask containing 20ml Hayflick media supplemented with chloramphenicol until mid-log phase. Cells were isolated via the described centrifugation method, and transformed with 1pMole of the plasmids pBSK_pM438_Cre_Puro, pBSK_pM438_Sce_Puro and pBSK_pM438_Puro ad an empty control. The cultures were incubated at 37°C for 4 hours post-transformation in 500µl antibiotic-free Hayflick media. 125µl aliquot of cells were passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with puromycin. The cultures were incubated at 37°C for 5 days.

Cells were then isolated via the described centrifugation method, and the pellets were suspended in 500µl Hayflick media. They were serial diluted and plated onto Hayflick plates supplemented with chloramphenicol, and incubated at 37°C.

2.2.9.3. Cre efficacy test protocol

From the P0.3_Cre agar plates, 100 colonies were picked and added into a 96 well plate containing 200µl antibiotic free Hayflick media. The cells were homogenised and 2x 50µl aliquots were passed to the adjacent two wells, containing 150µl of Hayflick media supplemented with 1.25x chloramphenicol and 1.25x tetracycline, and plain Hayflick respectively. 100µl of plain Hayflick was then added to the original seed well. Plates were incubated at 37°C and checked for growth via change in media colour.

2.2.9.4. Protocol 2 deletion validation

After 6 days incubation at 37°C, the 96 well plate containing isolated clones were assayed for growth by eye for the colour change from red to yellow. If the cells contained an inversion, all three wells should allow for growth as the cells retain the two antibiotic resistance genes, whereas deletions should not show growth in the antibiotic well. Six deletion cultures from the 96 well plate test were chosen at random and had their genomic DNA isolated, along with a sample from the P0.3_Cre pool. They were then assayed with oligos 212 & 213, which bind the inverted repeat regions of the transposons to amplify the 150bp deletion scar.

2.2.10. Genome deletion using protocol 3

WT_{M129} cells were transformed via electroporation, as described previously in chapter 2.2.3. WT_{M129} cells were grown to mid-log phase and transformed with 1pMole of LE_Cm. The culture was incubated at 37°C for 4 hours post-transformation in 500µl antibiotic-free Hayflick media. 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with chloramphenicol. The culture was labelled P0.1 and incubated at 37°C.

When the cells reached mid-log phase, they were isolated by the described centrifugation method and were transformed via the standard protocol with 1pMole of Tc_RE. After 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with tetracycline. The culture was labelled P0.3 and incubated at 37°C.

When the P0.3 cells were grown to mid-log phase, the harvesting and transformation protocol above was repeated with 1pMole of pMTnVL_Cre_Gm_VL. 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with gentamicin, and incubated at 37°C, labelled P0.3_VCV.

When the P0.3_VCV cells were grown to mid-log phase, they were isolated via the described centrifugation protocol and suspended in 500µl EB. A 50µl aliquot was transformed with the 1pMole pBSK_Sce_VCre_Puro. 4 hours post transformation, 125µl aliquot of cells was passed into a T75 tissue culture flask containing 20ml of Hayflick, supplemented with puromycin and incubated at 37°C for 5 days. A second 50µl aliquot

of P0.3_VCV cells was serial diluted and plated on Hayflick plates supplemented with gentamicin and grown at 37°C.

When the plates had grown, 100 colonies were picked and passed into the 96-well plate screen, described in chapter 2.2.9.3, and incubated at 37°C.

2.2.11. Custom circularised Next-generation Sequencing protocol

Specific step-by-step protocols can be found in Supplementary Materials A.

Genomic DNA was fragmented to 300bp via Covaris sonication. 5' phosphorylation was undertaken to allow for adapter binding, then 3' overhangs were filled to create blunt ends. These were then ligated using T4 ligase to create circular fragments, and linear DNA was removed via digest with exonuclease I and lambda exonuclease. Circular DNA was then denatured and amplified using primer mix containing LE-RE_Lox72, loxloop-6, loxloop-7 and loxloop-8 and a phi29 polymerase to amplify DNA containing a deletion scar. This amplifed product was then fragmented again using Covaris to 300bp and NEBNext Adaptor for Ilumina were annealed to the linearised DNA. This DNA was then sequenced using paired end reads of 150bp in an Illumina Hi-Seq 2500.

2.3. Results

2.3.1. Identification of Mpn_A37 clone for jumping transposon test

To test if the transposase could act on a transposon within the genome, a clone with a known transposon insertion site needed to be isolated to see if it had changed after a second transformation. Therefore, a clone from a transformation of the WT_{M129} with the Non-homologous end joining plasmid was isolated from Hayflick agar plates. The sequencing of the isolated clone via Sanger sequencing showed the location of the transposon as within the 5' end of the MPN162 gene. To ensure the purity the clone, it was re-passed on plates and three fresh clones were picked. All three had their genomic DNA amplifed via PCR using oligos 73 and 74, which flank the insertion site. The PCRs of the region with these oligos revealed a ≈4Kb band in the Mpn_A37 strain and a ≈500bp band in the WT_{M129}, confirming the insertion site, as shown in Figure 25, and Mpn_A37-1 was designated as Mpn_A37.

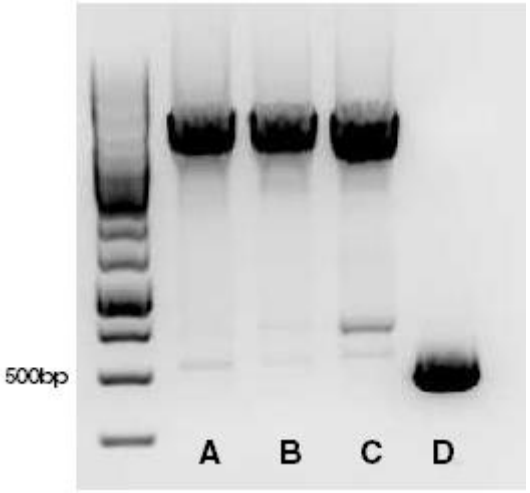


Figure 25: Identification of the Mpm_A37 clone. (A) Mpn_A37-1 amplified with oligos 73 & 74. (B) Mpn_A37-2 amplified with oligos 73 & 74. (C) Mpn_A37-1 amplified with oligos 73 & 74. (D) WT_{M129} amplified with oligos 73 & 74.

2.3.2. Quantifying if transposons can jumping after being inserted into the genome

Using the freshly isolated Mpn_A37 clone as a known standard of a cell line with a transposon in a known location, the colony was then transformed with three separate vectors, as explained in chapter 2.2.6. Genomic DNA from the resultant transformants was isolated, along with individual clones, and analysed via PCR for a change in size, indicating if the transposon had jumped, as outlined in Figure 26.

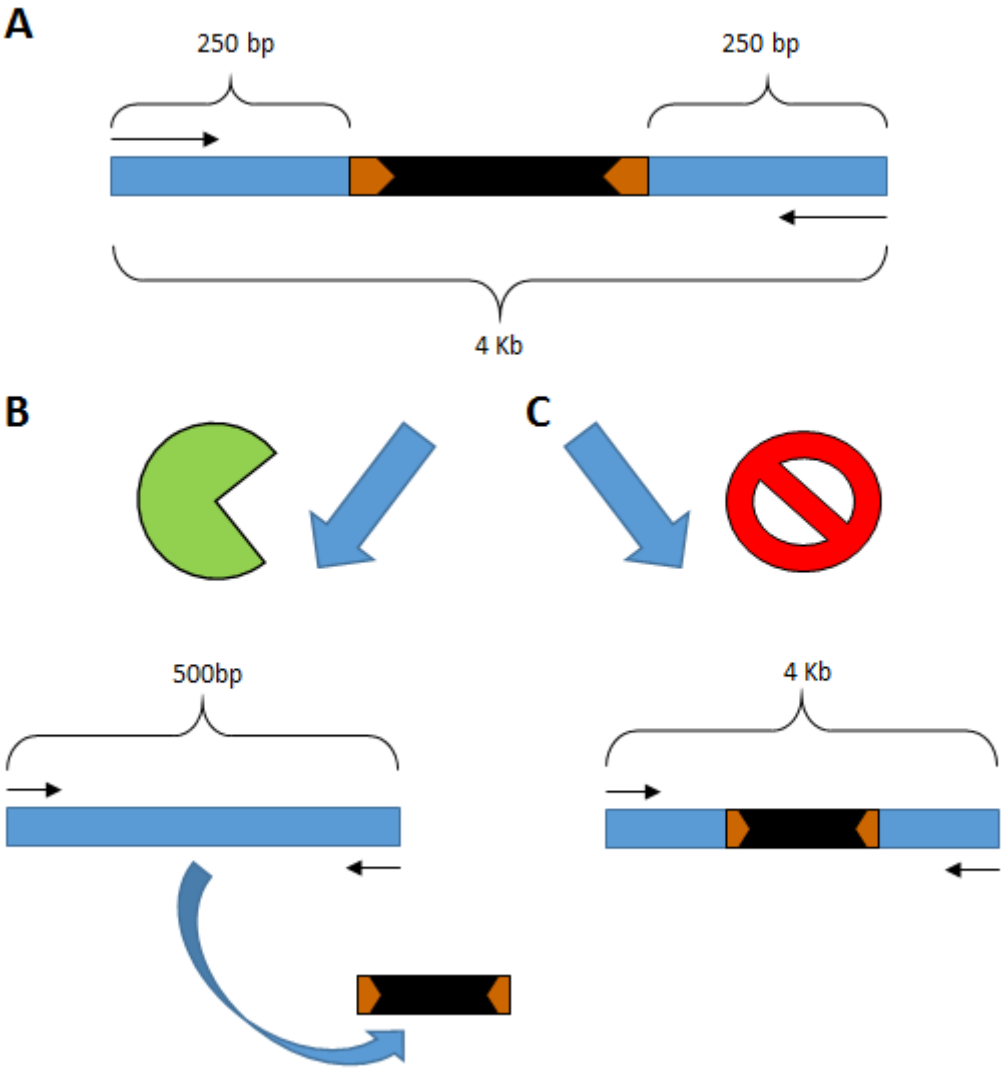


Figure 26: Outline of jumping transposon experiment. (A) Scheme of Mpn_A37, with a transposon (black) containing Inverted Repeats (orange arrows), with oligos 250bp upstream and downstream of the insertion site, giving a total size of 4Kb. (B) If the transposase can act on the genomic transposon, the transposon will be removed, and the PCR using the same oligos as in (A) will give a 500 bp band. (C) If the transposase cannot act on a genomic transposon, it will remain within the genome and the PCR will show the same 4Kb band as in (A).

Across all of the transformation pools and individual clones, a ≈4Kb amplification was seen, identical to the M129_A37 control. No cultures contained the 500bp band seen in the WT_{M129} condition to indicate that the transposase had been able to move the primary NHEJ transposon from its original location. Figure 27 shows the result of the PCR reaction with the genomic DNA pools from each reaction.

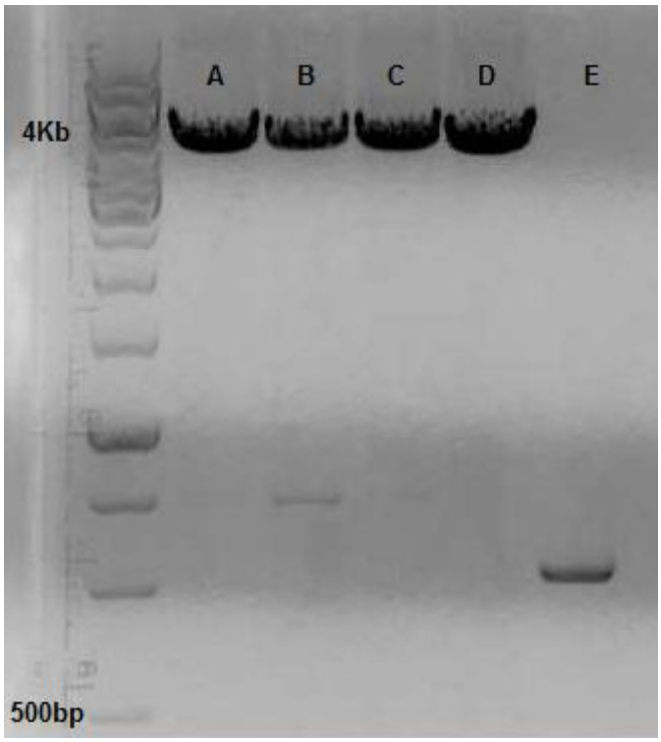


Figure 27: Amplification with oligos C50 & C51 of transformation of M129_A37 with different plasmids. (A) pMTnGent (B) pBSK_pM438_Gent_TPA (C) pBSK_pM438_Cre_Gm (D) Empty control (E) WT_{M129}

2.3.3. Quantification of transposon density

The WT_{M129} strain was transformed with 1pMoles of the pMTnCm[66].2 vector, creating R0.1 population (see chapter 2.2.8). The genomic DNA from this pool was sequenced, showing 201891 unique insertion sites, with a mean of 209 reads per insertion. This gives an average insertion of 1 transposon per 4 bases, accounting for a genome size of 819Kb. Looking at the distribution of reads across the genome, we find that the profile matches our expectations for a high-density transposon insertion yield. Table 11 shows the breakdown of the distribution of insertions by genomic context.

Table 11: Read counts for the transposon density study

	Non-Coding	Essential	Fitness	Non-Essential	Total
No. bases	86216	373778	86631	269769	81639
No. Unique Insertions	35513	41660	23690	101028	201891
No. reads	9377455	534846	4720620	27278317	41911238
No. Reads/Insertion	264.06	12.84	199.27	270.01	186.54
% total insertions	17.59	20.63	11.73	50.04	100.00
% total reads	22.37	1.28	11.26	65.09	100.00

It should be noted that the data representing the non-coding regions is approximate. Non-coding status was ascribed to everything that did not have a MPN annotation as annotated in the MyMPN database (Wodke et al., 2015). Therefore, it could contain small proteins or RNAs not identified in this database (Lluch-Senar et al., 2015b).

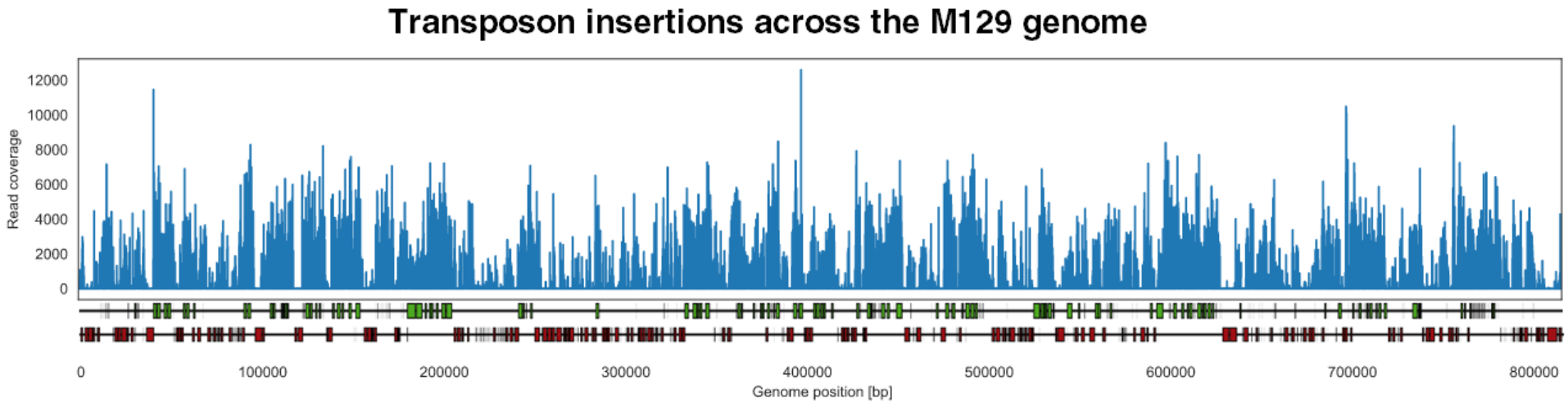


Figure 28: Plot showing the insertion density of the R0.1 population in blue. Green bars indicate non-essential genes and red bars essential genes on a scaled representation of the WT_{M129} genome

Figure 28 shows the distribution of the insertions across the genome, indicating the location of essential and non-essential genes. There is good agreement between peaks of insertions and generally non-essential regions, along with troughs in insertion frequency corresponding to higher levels of essential genes being present.

From this, using one pMole of plasmid appears to give as high a transformation efficiency as we are likely to get, thus increasing the concentration higher would be of no benefit. As mentioned in chapter 2.1.3.2, increasing the volume of DNA past this point will likely result in multiple transposon inserting into a single cell, thus being counter-productive to overall efficiency.

2.3.4. Results from Protocol 1

From the R1.0 population generated using Protocol 1 (see chapter 2.2.8) and the 20 isolated clones taken from it, genomic DNA was extracted. The PCR of both the pool DNA and 20 clones failed to show the desired 109bp after PCR. However, the amplification of the chloramphenicol gene was present in every colony tested, while absent from the WT, and the amplification of *glpD* gene was present as well. The protocol was repeated and still no positive indications of a deletion were found. This showed the original pMTnCm[66].2 transposon was still present, and no deletions had occurred.

2.3.5. Results from Protocol 2

Figure 29 shows a schematic overview of protocol 2, with the sequential addition of the lox-containing transposons to the genome, then selection with a suicide vector containing the Cre and *I-SceI* proteins, and the expected results from the selection.

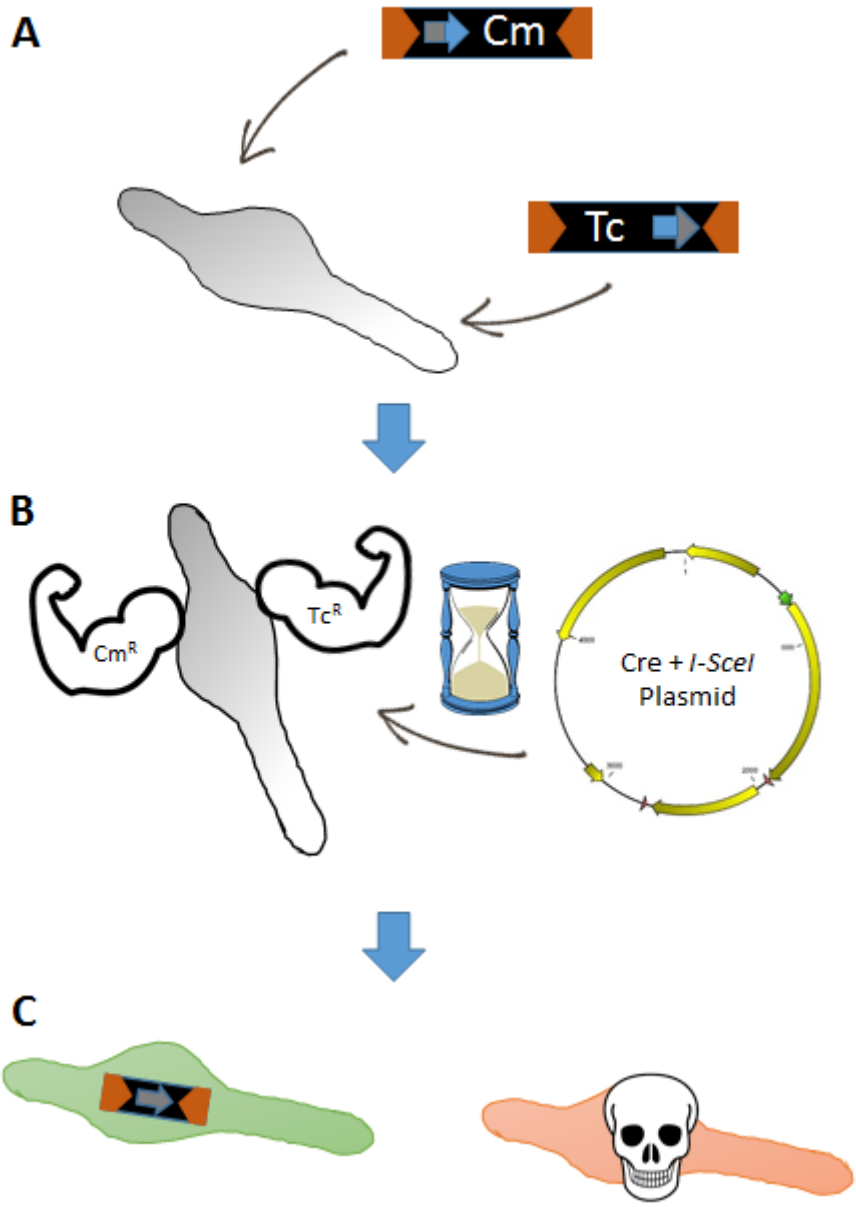


Figure 29: Overview of protocol 2. (A) WT_{M129} is transformed with the LE_Cm and Tc_RE vectors, both containing a left element and right element lox site, and chloramphenicol and tetracycline resistance genes respectively. (B) Cultures that are resistant to both antibiotics are transformed with a suicide vector containing the Cre recombinase, I-SceI mega-nuclease and an antibiotic resistance marker, and incubated under selection for 5 days. (C) Surviving cells will have allowed for a deletion that results in a lox72 site being formed. All other permutations will have been killed by either the antibiotic, action of the Cre recombinase or action of the I-SceI mega-nuclease.

2.3.5.1. Results of the *I-SceI* efficacy test

For the validation of the *I-SceI*, P0.1 cells were transformed with puromycin-resistant suicide vectors containing either the Cre recombinase, the *I-SceI* restriction enzyme or just the puromycin resistance for 5 days, then plated on plain Hayflick agar (see chapter 2.2.9.2). After 10 days of incubation, the CFUs on each plate were counted.

Table 12: CFU counts of P0.1 cells transformed with Cre and Sce containing suicide vectors

Transformation	Average CFU/10µl
P0.1 + pBSK_pM438_Cre_Puro	$2.35x10^2$
P0.1 + pBSK_pM438_Sce_Puro	$2.4\ x10^2$
P1.0 + pBSK_pM438_Puro	$3.65\ x10^4$
WT_{M129} + pBSK_pM438_Cre_Puro	$2.8\ x10^4$
WT_{M129} + pBSK_pM438_Sce_Puro	$3.2\ x10^4$
WT_{M129} + pBSK_pM438_Puro	$3.2\ x10^4$

While there are still survivors of the *I-SceI* treatment, there is a 2-order of magnitude difference between the survival rate of the cells transformed with the empty puromycin vector and the cells transformed with the *I-SceI* vector. The same effect is also seen in the Cre containing plasmid, indicating that the Cre also has a lethal effect on cells with a single lox site. There was no effect of the WT_{M129} cells, indicating that the Cre and *I-SceI* proteins do not have significant off-target activity.

2.3.5.2. Results of the Cre efficacy test

To test that the Cre could cause deletions, the P0.3 cell line (containing both lox sites and antibiotic resistances) was transformed with a suicide vector containing only the Cre recombinase and gentamicin resistance. After selection with the antibiotic, surviving clones were isolated and tested for their antibiotic resistance profiles in the 96 well plate test (see chapter 2.2.9.3). After 6 days incubation, the 96 well plate containing isolated clones were assayed for the ratio between deletion and inversion mutants. As *M. pneumoniae* acidifies the media, growth can be assayed by a change in light absorbancies in the 430 and 560 nm wavelengths, which correlates with a change in media colour from red to yellow. If the cells contained an inversion, all three wells should allow for growth as the cells retain the two antibiotic resistance genes, whereas deletions should not show growth in the antibiotic well. Of the 100 clones picked, 61 showed the deletion phenotype and 34 the inversion phenotype, with the remaining five not growing.

2.3.5.3. Protocol 2 random deletion validation

As they had already been transformed with the LE_Cm and Tc_RE transposons, then been subjected to the Cre recombinase already as part of the Cre efficacy test, six of the deletion cultures from the 96 well plate test designed to test for the Cre efficiency were chosen at random and had their genomic DNA isolated, along with a sample from the P0.3_Cre pool. They were then assayed with oligos 212 & 213, which bind the inverted repeat regions of the transposons, shown in Figure 30. The 150bp band indicating a deletion was present in all six clones and the pool. Larger bands accounting for the transposons were present in the pool, signifying the presence of the inversion clones, but not in the six deletion clones.

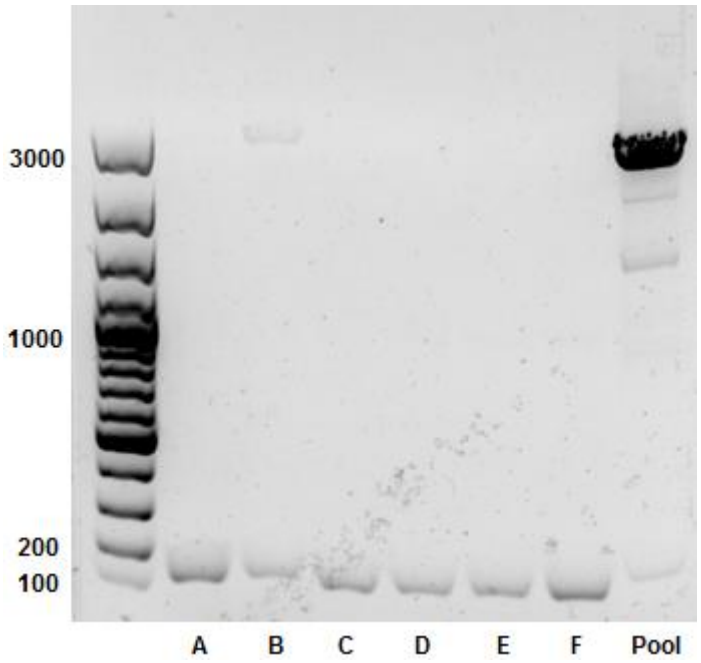


Figure 30: Deletion scars for Protocol 2, with all six clones analysed via PCR with oligos 212 & 213, along with the pool DNA

To identify the deleted region in the six sample, genomic DNA from the six clones was sent for Sanger sequencing using oligo 207, which anneals at the 5' end of the scar. However, of the six clones, only the sequencing of clone D gave a positive result. This showed the lox72 site was located upstream of the *glpD* gene (MPN051).

Oligos flanking the region, C50 and C51 that had been used as a positive control for genomic amplification in protocol 1, were used, along with oligos 212 & 213 binding to the inverted repeats. The results of the PCRs are shown in Figure 31:

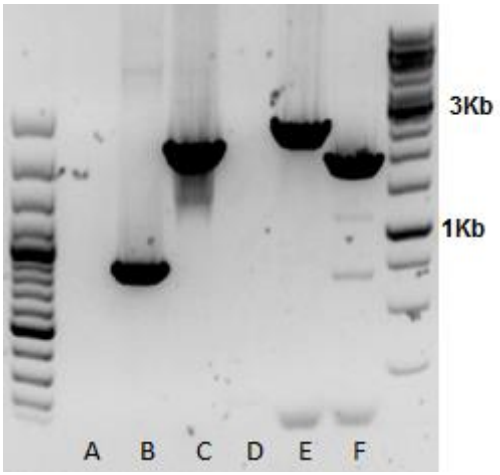


Figure 31: Analysis of P1.0_D. (A) WT_{M129} + Oligos 212 & 213. (B) LE_Cm + Oligos 212 & 213 (C) Tc_RE + Oligos 212 & 213. (D) P1.0_D + Oligos 212 & 213. (E) WT_{M129} + Oligos C50 & C51. (F) P1.0_D + Oligos C50 & C51.

The P1.0_D DNA shows no amplification of the transposon bands, further indicating that the deletion has occurred. It also shows a decrease in band size compared to the WT_{M129} when amplified by the C50 & C51 oligos, around the *glpD* gene. The 2Kb band from lane F was isolated and sent for Sanger sequencing for confirmation.

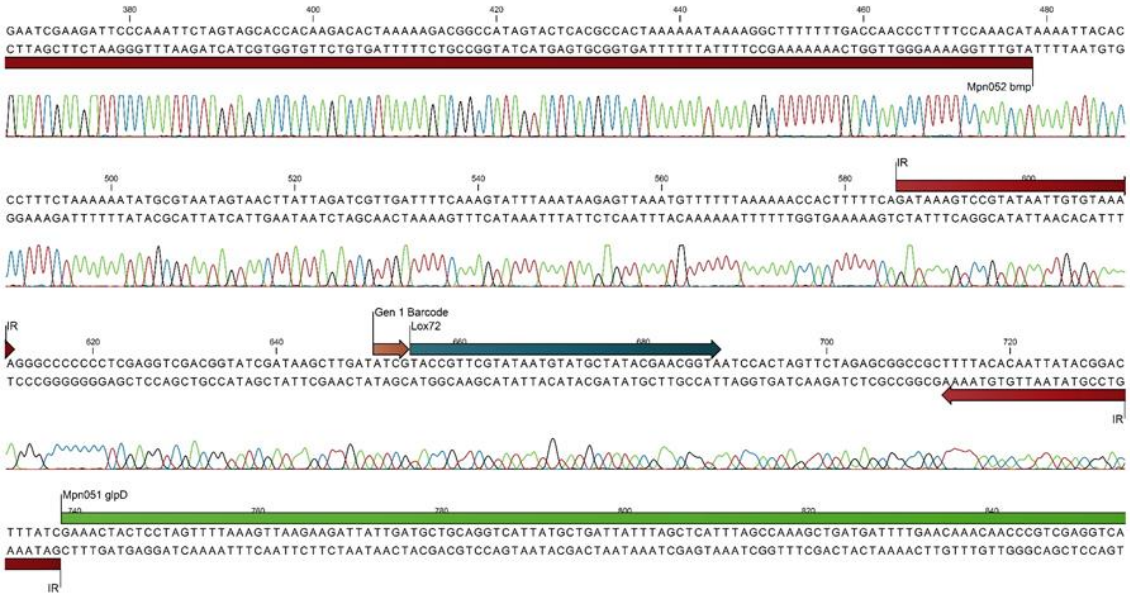


Figure 32: Result of Sanger sequencing of P1.0_D deletion

Figure 32 shows the expected deletion scar containing the double-mutant Lox72 site, generation 1 barcode and surrounding genomic DNA belonging to the Mpn051 and Mpn052 genes. In total 671 bases were deleted, including the first 582 bases of the *glpD* gene.

To ascertain the location of the deletions in the other five clones, genomic DNA from all six were sent for whole genome sequencing. The standard library preparation for a 125bp paired-end read via Illumina Mi-Seq was performed. Of these, five of the six clones (P1.0_A, P1.0_C, P1.0_D, P1.0_E & P1.0_F) contained an identical deletion, and only P1.0_B varied, with a 6.7Kb deletion. This deletion removed five non-essential genes (MPN368, MPN369, MPN370, MPN371 and MPN372).

2.3.6. Protocol 3

Figure 33 shows a schematic overview of protocol three, highlighting the change of Cre delivery from a suicide vector to a transposon based system.

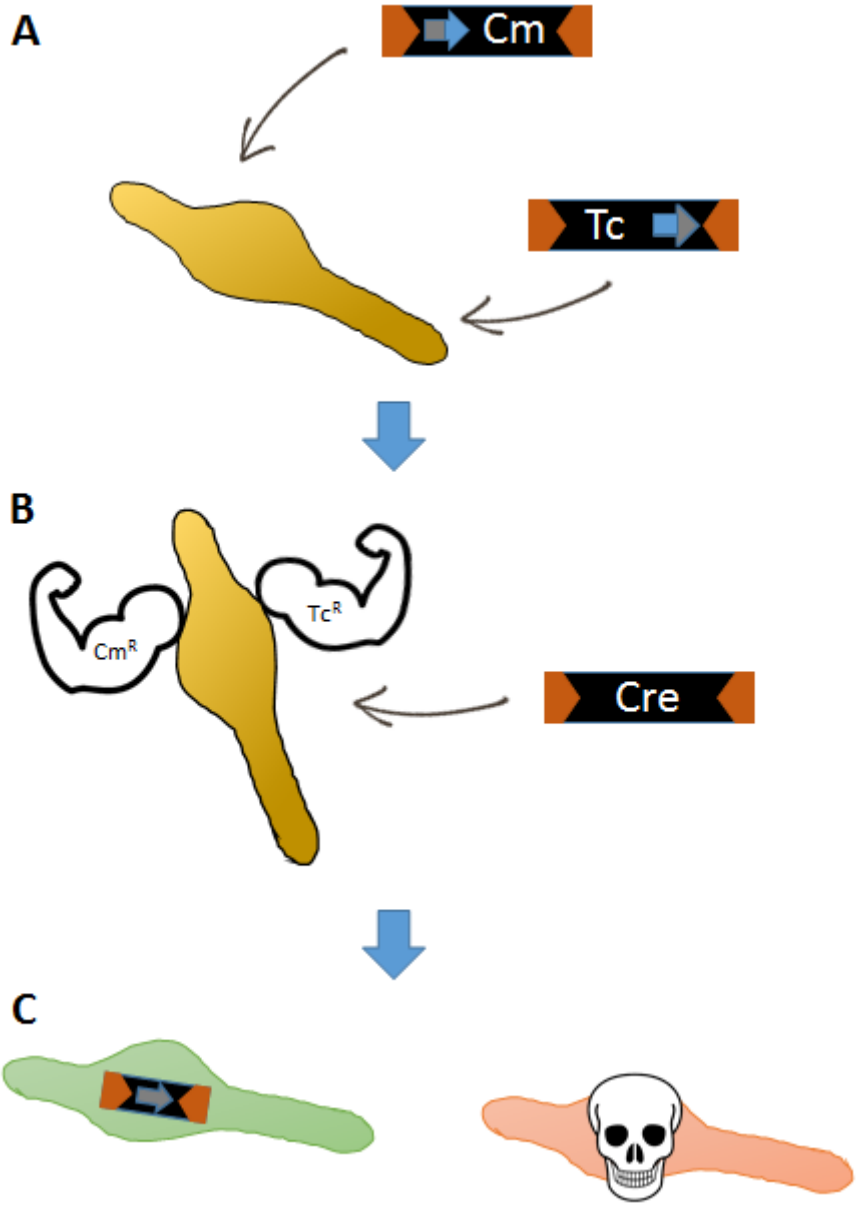


Figure 33: Overview of genome deletions via protocol 3. (A) WT_{M129} is transformed with the LE_Cm and Tc_RE vectors, both containing a left element and right element lox site, and chloramphenicol and tetracycline resistance genes respectively. (B) Cultures that are resistant to both antibiotics are transformed a third transposon containing the Cre recombinase (C) Surviving cells will have allowed for a deletion that results in a lox72 site being formed. All other permutations will have been killed by either deleting an essential gene or by the action of the Cre recombinase on a remaining lox site.

2.3.6.1. 96-well plate test to assay deletion vs inversion ratio

In protocol three, the P0.3 cells (containing both lox transposons and antibiotic resistances) were transformed with the third transposon containing the Cre recombinase flanked by mutant left element and right elelment Vlox sites, and a gentamicin resistance. From the surviving culture, 100 colonies were picked and placed into the 96 well plate screening method (see chapter 2.2.9.3). After incubation, all 100 cultures grown in the 96 well plate test showed the deletion phenotype, with growth in the plain Hayflick wells and no growth in the well containing chloramphenicol and tetracycline. A representative plate is shown in Figure 34:

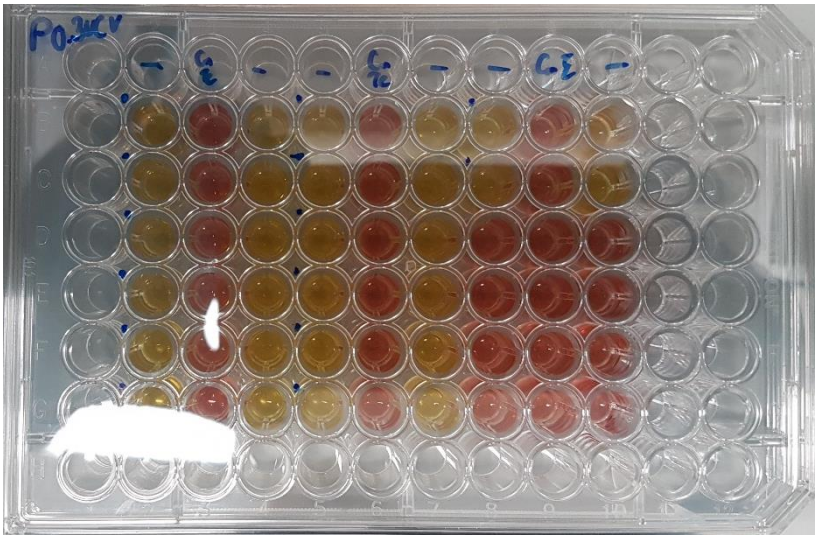


Figure 34: Example of P0.3_VCV 96 well screening test

Figure 34 shows a representative plate containing fourteen of the clones isolated, along with four bank sterility controls. The antibiotic free media in columns 2, 4, 5, 7, 8 and 9 (indicated by the "-" marker) have fully acidified, indicating growth. However, the media containing antibiotic, columns 3, 6, and 9 (indicated by the $^{Cm}_{Tc}$ notation) show no acidification, thus no growth. The penultimate three wells in rows D to G are empty controls for sterility. It is clear that every single sample isolated shows a clear deletion phenotype.

2.3.6.2. Custom Next-generation sequencing results

DNA extracted from the pool of P0.3_VCV cells was sequenced using the custom circularisation protocol described in chapter 2.2.11. The results from this were analysed for reads containing disparate regions of the WT_{M129} genome separated by the adaptor sequence. As not all reads shows both of the inverted repeats, the exact location of both transposon sites was unknown. As such, the genome was divided into 16388 bins of 50bp. Any site that matched the sequence in one of those bins was added to it, thus giving insertion location accurate to 50bp. From this, 285 unique deletions were discovered which contained no essential genes, and 1365 that did contain essential genes. The composition of the two populations is shown in Table 13.

Table 13: Breakdown of putative deletions from the custom DNA circularisation protocol

	Containing Essential genes	**Containing no Essential genes**
No. Reads	41335	1250377
No. unique deletions	1365	285
Average no. reads per deletion	30	4387
Average size of deletion (Kb's)	279.7	7.7

This gives us a stringent cut-off rate of over 30 reads per deletion to filter out any false positives generated from the sequencing reactions. The results were mapped onto the WT_{M129} genome, giving the deletion profile shown in Figure 35. Each bar indicates a unique insertion site identified via the custom sequencing protocol. It is worth noting that while the average number of reads per deletion in the population that contain no essential genes is 4387, this is biased by a small number of highly prolific deletions. The number of unique deletions with a representation of over 30 reads is 42. Figure 35 shows the different deletions filtered by number of reads:

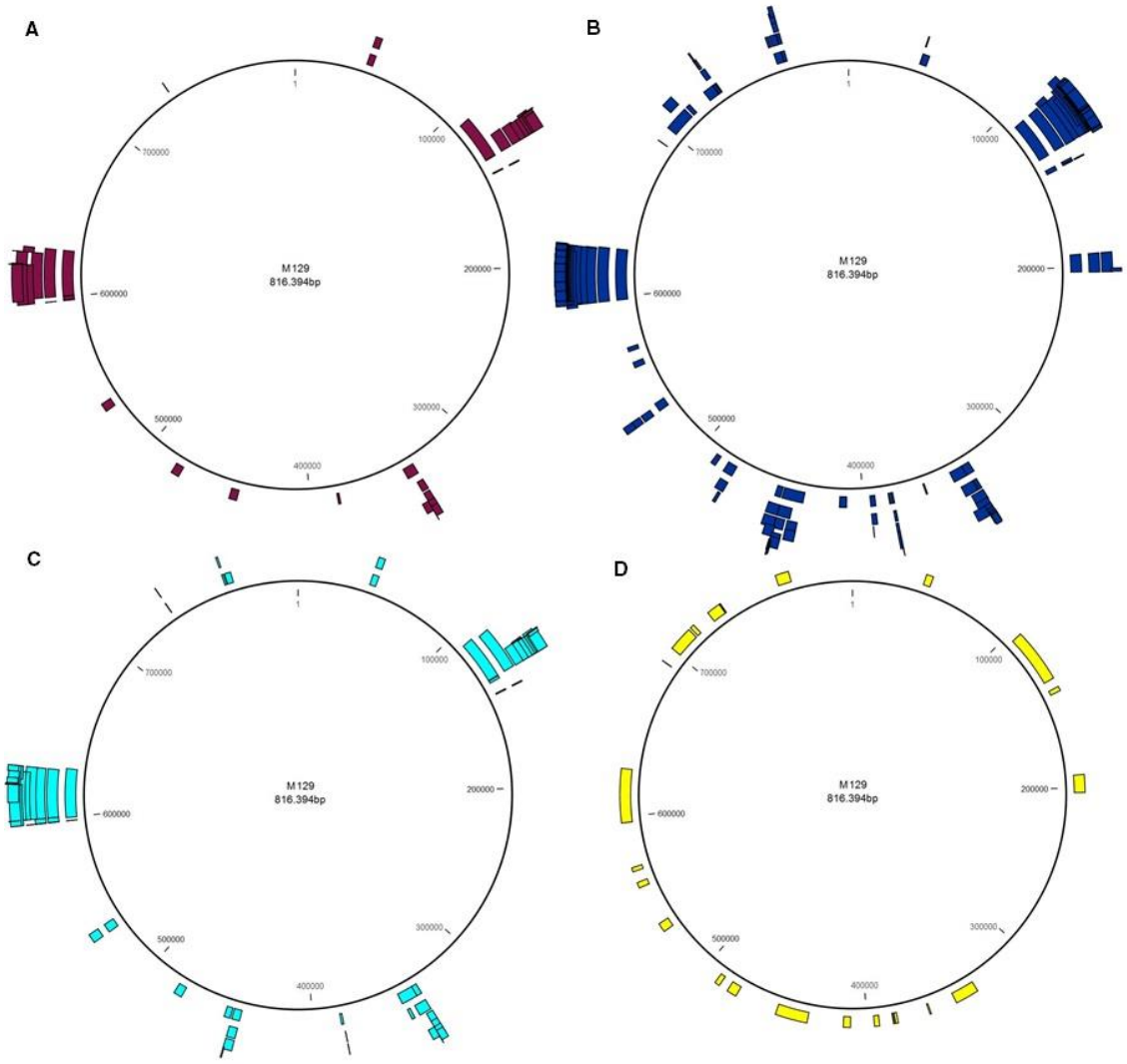


Figure 35: Deleted genomic regions identified via custom sequencing protocol of the P0.3_VCV pool. (A) All regions that do not contain an essential gene and are represented by over thirty reads. (B) All regions that do not contain an essential gene and are represented by over 10 reads. (C) All regions that do not contain an essential gene and are represented by less than ten reads. (D) All regions that have been deleted at least once in the previous maps.

There is a high degree of similarity in the maps generated in Figure 35. Indeed, there is only one region that is found deleted in the group of deletions with less than 10 reads that is not found in the group of reads with over 30 reads. This lends credence to the idea that the deletions with lower reads may also be accurate.

In all cases, there are clear hotspots of deletion, specifically at 6000000 bp mark and the 120000 bp mark. Both sites show large deletion where the entire non-essential region has been removed, and also numerous smaller deletions throughout the region. This indicates that it is not just due to one fortunate lox site integration, but multiple stable deletions are possible at this point across multiple cell lines

2.3.6.3. Validations

To validate the deletion mapping had worked, three regions were tested for the presence of deletions in the P0.3_VCV pool DNA and lack of in the WT_{129} genome. Region 1 contained seven non-essential genes (MPN096 to MPN102) over ≈10Kb, region 2 contained four non-essential genes (MPN397 to MPN400) over ≈5Kb and region 3 contained nineteen non-essential genes and one fitness gene (MPN493 to MPN512) over ≈25Kb. As shown in Figure 36, the WT_{M129} amplifications do not contain the smaller ≈500bp band present in the deletions.

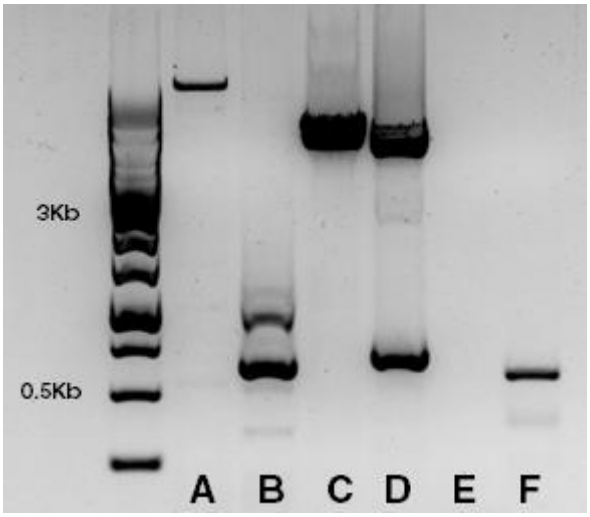


Figure 36: Validation of P0.3_VCV pool. From left to right; (A) WT_{M129} amplified with oligos 346 & 347, (B) P0.3_VCV pool amplified with 346 & 347, (C) WT_{M129} amplified with oligos 352 & 353, (D) P0.3_VCV pool amplified with oligos 352 & 353, (E) WT_{M129} amplified with oligos 360 & 355, (F) P0.3_VCV pool amplified with oligos 352 & 353

The bands from the P0.3_VCV samples that showed a deletion were cut and sequenced via Sanger sequencing. All three showed the correct lox72 deletion scar, and the genomic regions upstream and downstream mapped onto the *M. pneumoniae* in the expected regions. Figure 37 shows the sequencing data from region 3, with its genomic DNA context. This region showed a 25Kb deletion.

Figure 37:Validation of P03_VCV region 3.(A) Sanger sequencing data of the ≈500bp deletion band of region 3, showing the inverted repeats (IR) in red, lox72 site in purple and M. pneumoniae homology regions in pink and blue. (B) The genome of M. pneumoniae M129, non-essential genes shown in green, fitness genes shown in dark blue, homology regions from panel (A) shown using the same colours, deleted region in yellow totalling 25571bp.

The genes deleted from this sample are listed below, along with their known functions:

Table 14: Genes deleted from the 25Kb M. pneumoniae deletion

Gene	Function	Gene	Function
MPN493 (*ulaD*)	Probable 3-keto-L-gulonate-6-phosphate decarboxylase	**MPN503**	Putative mgpC-like protein
MPN494 (*ulaC*)	Ascorbate-specific phosphotransferase enzyme II Component A	**MPN504**	Uncharacterized protein
MPN495 (*ulaB*)	Ascorbate-specific phosphotransferase enzyme II Component B	**MPN505**	Uncharacterized protein
MPN496 (*ulaA*)	Ascorbate-specific permease II Component C	**MPN506**	Conserved hypothetical lipoprotein
MPN497 (*ulaG*)	Probable L-ascorbate-6-phosphate lactonase	**MPN507**	Putative type-1 restriction enzyme
MPN498 (*araD*)	Probable L-ribulose-5-phosphate 4-epimerase	**MPN508**	Putative membrane export protein
MPN499	Uncharacterized protein	**MPN509**	Uncharacterized protein
MPN500	Putative adhesin P1-like protein	**MPN510**	Uncharacterized protein
MPN501	Uncharacterized protein	**MPN511**	Uncharacterized protein
MPN502	Uncharacterized protein	**MPN512**	Uncharacterized protein

Looking more generally at the data, if we take all of the regions that could have been deleted, as shown in Figure 35 D, there were 147 genes that were either partially or fully deleted at least once according to the sequencing results. The full list of all deleted genes can be found in Supplementary data B, however Table 15 contains a breakdown of their class and function:

Table 15: All genes deleted in P0.3_VCV transformation

Function	Number different genes deleted
Uncharacterized protein	37
Conserved hypothetical protein	30
Conserved hypothetical lipoprotein	16
Putative *mgpC*-like protein	9
Uncharacterized lipoprotein	7
Putative type-1 restriction enzyme specificity protein	6
Putative adhesin P1-like protein	5
Uncharacterized amino acid permease	3
Putative type-1 restriction enzyme	1
UPF0134 protein	1
Uncharacterized adenine-specific methylase	1
Probable DNA helicase I homolog	1
Ribonucleoside-diphosphate reductase subunit beta	1
Protein *nrdI*	1
Ribonucleoside-diphosphate reductase subunit alpha	1
Putative ABC transport system permease protein	1
Putative ABC transporter ATP-binding protein	1
Putative type I restriction enzyme *hsdM*	1
Putative type-1 restriction enzyme mpnORFDP R protein part 2	1
ADP-ribosylating toxin CARDS	1
Probable guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	1
Predicted lipase	1
Protein *recA* recombinase	1
Membrane nuclease A	1
Probable L-ribulose-5-phosphate 3-epimerase *ulaE*	1
Probable 3-keto-L-gulonate-6-phosphate decarboxylase	1
Ascorbate-specific phosphotransferase enzyme IIA component	1
Ascorbate-specific phosphotransferase enzyme IIB component	1
Ascorbate-specific permease IIC component *ulaA*	1
Probable L-ascorbate-6-phosphate lactonase *ulaG*	1
Probable L-ribulose-5-phosphate 4-epimerase *araD*	1
Putative membrane export protein	1
Hemolysin-type ABC transporter	1
Putative protease	1
Probable ribose-5-phosphate isomerase B	1
Negative regulator of *FtsZ* ring formation	1
Phosphate import ATP-binding protein *pstB*	1
Phosphate transport system permease protein *pstA* homolog	1
Phosphate-binding protein *pstS*	1
PTS system mannitol-specific EIICB component	1
Mannitol-1-phosphate 5-dehydrogenase	1
Mannitol-specific phosphotransferase enzyme IIA component	1
Total	147

The vast majority of the genes deleted code for completely unknown functions, with 90 out of 147 being annotated as "Uncharacterised proteins", "Conserved hypothetical proteins", "Uncharacterised lipoproteins" or "Conserved hypothetical lipoproteins". Of the genes that do have a known function, only 29 have an ascribed gene name, with the rest being "putative" or "probable" proteins.

The breakdown of essentiality is similarly skewed, with genes annotated as non-essential representing 139 of the 147 of the deleted genes, with the remaining 8 being fitness genes. As expected, there were no essential genes deleted.

Combined, the deletions totalled approximately 171.2Kb, 21% of the genome.

2.4. Discussion

Protocol three showed that we can achieve a large scale deletion program within the *M. pneumoniae* genome, being able to delete a wide variety of targets both large and small. However, it took many iterations of the protocol to achieve a working solution. This discussion section is intended as a guide to the decision making that led to each iteration, as well as a discussion of the results gathered.

The repeated failure of the first protocol to produce viable deletions within the *M. pneumoniae* genome indicated that our system was not working as intended. The issue was not due to lack of proper transposon coverage, as the first transposon pool had a transposon inserted on average every four bases, as well as correlating strongly with the essentiality profile of the genome.

Therefore, to identify the problems, we looked at the design of the system again. The first issue we encountered was a discrepancy in the description of the mutant lox sites across the scientific literature, with the identities and orientations of the lox66 and lox71 sites varying from paper to paper, as shown in Figure 38. The early papers describing the WT loxP sites and their mutants, such as Albert et al., (1995), give the orientation as:

LoxP: 5' - ATAACTTCGTATA ATGTATGC TATACGAAGTTAT – 3'
Lox66: 5' - <u>TACCG</u>TTCGTATA ATGTATGC TATACGAAGTTAT – 3'
Lox71: 5' - ATAACTTCGTATA ATGTATGC TATACGGAA<u>CGGTA</u> – 3'

The mutant sites are indicated by underscores and the directionality of the spacer region is from 5' to 3'. However, in more recent papers there seems to be no consensus between the various authors on this orientation scheme. Some claim the left element mutant is identified as lox71 (Missirlis et al., 2006), some as lox66 (Leibig et al., 2008). There is even disagreement on the directionality conferred by the central spacer region, with some authors showing the spacer region acting in the opposite direction (Chatterjee et al., 2010). Figure 38 shows a selection of papers, all with a slightly different variation o the lox nomenclature.

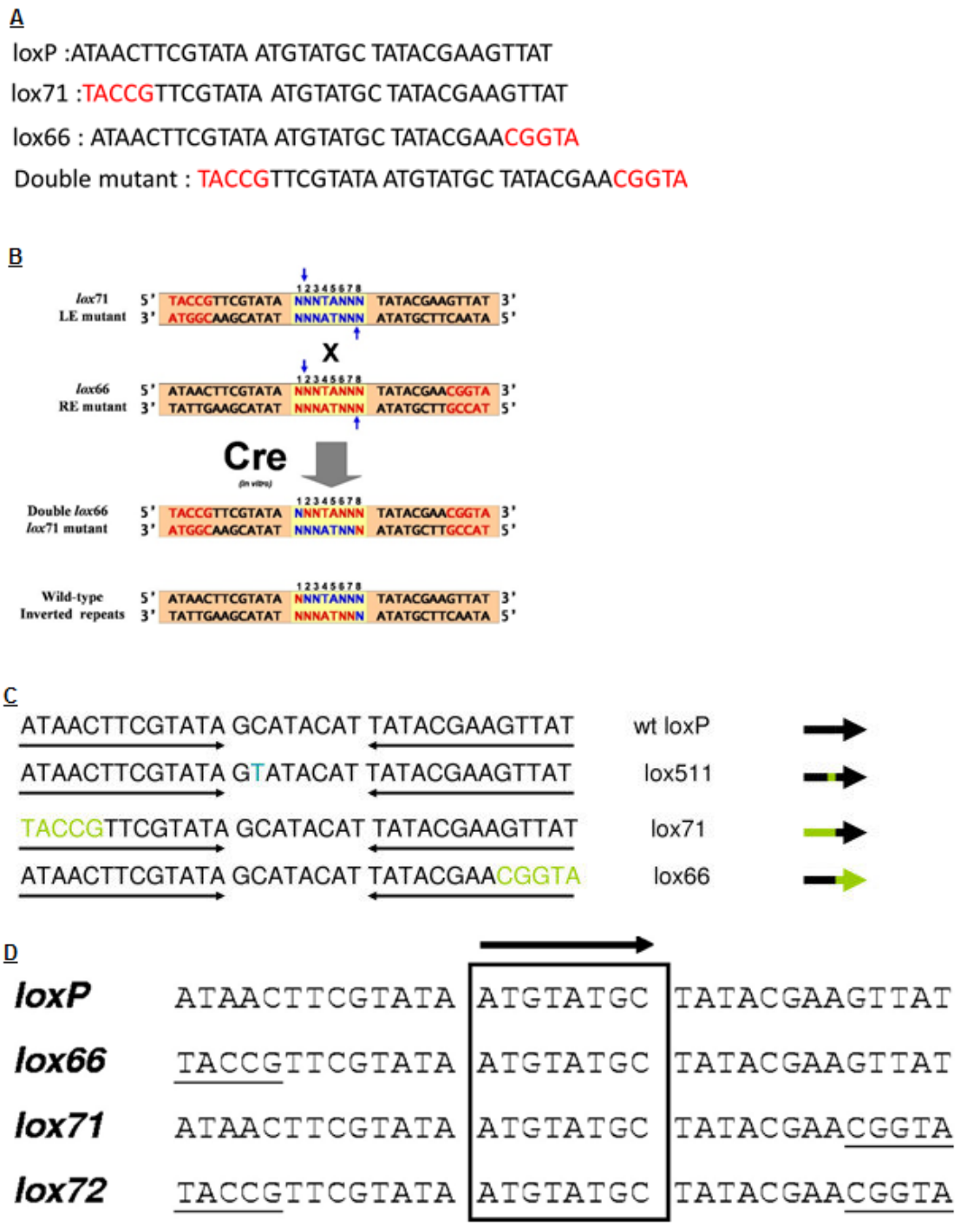


Figure 38: Variation in lox site identifications and annotations. (A) Lox sites described with the left element mutant as lox71 and right element mutant as lox66, taken from IGEM 2014. (B) Lox sites indicating that the left element mutant in lox71 and right element mutant as lox66, with ambiguous directionality. Taken from (Missirlis et al., 2006). (C) Lox sites indicating that the left element mutant in lox71 and right element mutant as lox66, however the directionality is reversed in sequences yet not in the diagrams. Taken from (Chatterjee et al., 2010). (D) Lox sites that the left element mutant in lox66 and right element mutant as lox71, taken from (Leibig et al., 2008).

Given the differences in lox orientations reported in the literature, we decided to use the original orientations we found in the older paper by Albert et al., (1995). Looking back at our plasmid maps with this lox orientations as standard, we realised that the pMTnCm^{66}.2 contained a left element lox site at the 3' end of the transposon instead of a right element lox, and conversely the pMTnTc^{71} contained a right element lox at the 5' end of the transposon instead of a left element lox. This meant that instead of creating an inactive

Lox72 when the Cre acted, a WT LoxP site was formed instead. Given that we later found the action of the Cre on a single active lox site to be toxic to the cell, this could explain why we never found a successful deletion using this protocol. Any cells that underwent a deletion that removed the *I*-SceI counter selection would contain a loxP site, and thus be killed by the Cre. Conversely, any that formed the inactive lox would still contain the *I-SceI* site and would be killed that way instead.

We therefore re-designed the plasmids to fit with the lox scheme from Albert et al., (1995) to ensure that a lox72 site was formed. The original plasmids, pMTnCm[66] and pMTnTc[71] both had lox sites at the 5' end of the transposon. Originally, the lox site in the pMTnCm[66] plasmid was moved to the 3' end to allow deletion of the whole cassette, creating the pMTnCm[66].2. However, sequencing of the original pMTnCm[66] plasmid showed it contained the left element lox at the 5'. It was therefore renamed LE_Cm to avoid further confusion. The second vector was then cloned, with right element lox cloned into the 3' end and the original lox site removed, thus creating the vectors for protocol two.

The second protocol showed the first proof of concept that the protocol is viable to obtain random deletions within the *M. pneumoniae* genome, using the two new vectors with the correct lox sites, and induction through a suicide vector containing the Cre recombinase and *I-SceI* restriction enzyme. The two regions that were deleted had very different attributes and genetic environments. The P1.0_D deletion showed the system is capable of deleting small regions surrounded by essential genes. The *glpD* gene is flanked by two essential genes (MPN050 and MPN052), yet the system had high enough transposon coverage to allow for a deletion within the small ≈1Kb non-essential region between the two genes.

The deletion in the P1.0_B cell showed that the system is also capable of deleting large regions of non-essential genes, in this case 6.7Kb. While there are larger regions of non-essential genes with in *M. pneumoniae* genome, this is a good representation of one of the largest contiguous regions available. The genes it contains are two uncharacterised protein (MPN368 and MPN371), and uncharacterised lipoprotein (MPN369), a putative P1 adhesin (MPN370) and the *ptxA* gene, also known as the CARDS toxin (MPN372), and an ADP-ribosyltransferase Pertussis toxin. Knowing that large stretches of non-essential genes can be deleted in a single attempt means there are probably other large regions of non-essential genes that can also be deleted.

However, the downside of this protocol is there seems to be very little variation within the deletions. Indeed, of the six colonies picked, five not only had deletions in the same area, but the deletions were identical. This indicates that it is not just that the *glpD* gene is highly amenable to deletion, but that either this was the only region to survive a deletion, or the number of different successful deletions after the suicide vector test was exceptionally low. Given that five genes were deleted in the P1.0_B test, it is highly unlikely that only that combination of the five genes could have been deleted successfully. Instead, it is far more likely that any of those individually could also have been deleted, or shorter regions within the deletion be viable in other cell lines, yet this area was not over represented. When a random sample of the population was tested, 86% were identical. The identical nature of the five *glpD* genes instead indicates that they came from a single cell that propagated, or protocol selected for only a very few possible deletion regions.

Given the transformation rate of *M. pneumoniae* is approximately one transformed cell in one thousand (Montero-Blay et al., n.d.), every transformation is a bottleneck event that drastically reduces the number of potentially successful lox combinations. Given that the cells were passaged three times after each transposition event, it is possible that the methodology was more effective than expected at promoting cell lines with a low fitness loss. Indeed, the overrepresentation of the *glpD* deletion could be explained by the fact the second transposon insertion caused no further fitness loss to the cell, as the gene had already been knocked out. However, other regions that may have been equally amenable to deletion contained knock outs of two genes, and thus the fitness loss prevented their propagation to the same extent, and thus were not represented.

The other highly interesting result was the identification of the Cre acting on a single lox site produces a lethal phenotype at almost identical levels to the *I-SceI* enzyme. This could explain the selective advantage deletion mutants appeared to have over inversion in the 96-well plate screen. Statistically, the number of deletions should be very low when the counter-selective *I-SceI* is not present. If an even distribution of insertions is assumed, then only one in four cells should have a pair of lox sites in the correct orientation to produce a deletion and remove the antibiotic resistances, giving the deletion growth phenotype on the plate. Of these 25% of cells, only a tiny fraction will contain two lox sites that can produce a viable deletion with no essential genes between them. However, 61% of colonies picked with no counter-selection showed the deletion phenotype, indicating a large selective pressure towards deletions.

When an inversion takes place between a mutant LE_lox site and RE_Lox site, the DNA between the sites invert and a double mutant lox72 and WT loxP are formed (Suzuki and Nakayama, 2011; Van Duyne, 2001). However, the Cre is still being expressed on the suicide plasmid and thus acting on the now formed LoxP site. This therefore acts as its own strong counter-selection, allowing the deletion clones with only the inactive lox72 site to survive. As the cultures were grown under suicide vector selection with the Cre for 5 days, then allowed to grow freely on plates, this indicates the action of the Cre is bactericidal instead of bacteriostatic, as growth under non-selective conditions did not allow for the cells to rebound as would be the case if they were only suppressed instead of killed.

The precise nature of the lethal activity of the Cre is unknown, although our current hypothesis is the activity is linked to the nickase activity it utilises for recombination. The Cre forms a quadramer complex, with a single Cre molecule bound to each arm of each lox site. Each Cre molecule has nickase ability to cut the stand it is bound to. The Cre molecule at the 5’ end of the lox site (in regard to its directionality) binds the free phosphotyrosine exposed by the nickase cut to itself as an intermediary before transferring the strand to its destination on the opposite lox site. However, as no other lox site is present, there is no DNA for it to bind to (Pinkney et al., 2012; Van Duyne, 2001). This breakage of the DNA, along with the binding of the Cre molecules to genome potentially inhibiting the translation and replication machinery, could cause the lethal phenotype seen in the cells.

However, even with this lethal phenotype displayed by the Cre and *I-SceI* molecules, each were not effective at completely removing inversions. Both had a similar level of background colonies that survived the suicide vector treatments in Table 12, with a drop in viable cells of two orders of magnitude, however still a significant surviving cell

population. In the 96-well plate test, 34% of colonies were inversions and survived the exposure to the Cre. This may be due to the exposure of the proteins was mediated via suicide vector, and those cells that were exposed to a vector which contained a mutant in either the Cre or *I-SceI* gene but not the puromycin resistance were selected for. Another possibility is that the cells lost the plasmid early through cell division (as the plasmids are not replicative, only one daughter cell will retain it) to mitigate the effects of the Cre or *I-SceI* but survived due having produced enough of the puromycin resistance protein to allow for survival in the media. This becomes more likely as time progresses as more and more of the puromycin is metabolised by the cell and thus removed from the media.

Taking these factors account however, we designed a new Cre delivery system for protocol three, relying on its expression via a transposon. Having the Cre expressed from within the genome clearly allows for much greater levels of expression, or at least more stable levels. As a result, the lethal activity of the Cre on a single lox site in the genome was enough to act as a full counter-selective system on its own, without the need for the *I-SceI* system as a reserve. The inclusion of Vlox sites surrounding the Cre and gentamicin resistance genes allowed for the removal of the genes before further rounds of transformation, to prevent the cells from being immediately killed by the Cre's activity on the newly introduced lox site.

By introducing the Cre induction on a transposon, the third protocol was able to produce a promising variation of deletions, with the protocol capable of initiating the deletion of genomic regions from 50bp up to 25Kb, and has enough coverage to allow for multiple variations of deletion within a single large non-essential region.

This variation within the larger regions is important, as it shows that the variation of insertion sites for the original lox site transposons is as high as we expected. The concentration of deletions in hotspots is not due to a bottleneck caused by poor transformation efficiency in either of the lox insertion stages, as the variation in the hotspots shows multiple integrations, with a vast range in the size of the deletions across the general region. If the hotspot was caused by the fact that only a small number of transposons were present in one of stages, the vast majority of the deletions would share a common end or starting point, which is not what we observe. Due to the uncertainty inherent to the data generated from the sequencing protocol, i.e. the majority of reads not containing both inverted repeat regions, we decided that splitting the genome into 50bp bins gave us specificity enough to map the deletions as accurately as possible. Due to this aggregation method however, there could be many more deletions that are similar to each other by fewer than 50 bases, and thus are missed from the analysis by being grouped with the other reads.

With regard to previous protocols, we did not find a deletion of the *gplD* gene within the results from protocol three. This indicates that it was indeed due to chance that the specific cell line with that deletion propagated as heavily as it did. Interestingly, the P1.0_B deletion was found in the third protocol results, as it would have been located in the cluster found at the 450,000 bp mark, further showing that the deletions at these hotspot areas are reproducible ad not a sequencing artefact.

While there is a large variation in the deletions we achieved in the third protocol, they do indicate that there were others that we lost. The deletion of the *glpD* gene in protocol two showed that fitness genes can be lost, and indeed we do see fitness genes among those

deleted in protocol three. However, it must be recognised that not only do they represent a tiny minority (8/147 genes), but the genes that were deleted were exactly the ones that would be expected to be deleted. Due to the fact that during growth after the Cre transposon was transformed, the cells were in competition with each other. While only for one passage, those cells who contain a deletion that has as low a fitness loss as possible will certainly try to outcompete other cells for nutrients. It is notable that 15 separate adhesion proteins (out of a total of 22) were among those deleted, as were nine restriction putative enzyme proteins. These would have no metabolic cost to the cell through their loss (assuming a lack of moonlighting functionality), and thus given the cells a large growth advantage compared to any that lost even non-essential metabolic functions.

The selection for faster growing mutants is inevitable, and while we have tried to minimise this as much as possible by allowing only a single passage between transformations, it remains an inherent property of bacterial life that the fast-growers will proliferate at the expense of the slow growers. As a protocol to produce a 'Minimal Genome', this is an issue, as it means potential deletions with a large fitness defect are lost from the population over time. However, as a methodology of genome streamlining, it can be seen as an advantage. This protocol allows for the partial selection of those deletion strains with the most robust growth, while also allowing for a large range of mutants to be created. The fact that the main virulence factor in *M. pneumoniae*, the CARDS toxin (Parrott et al., 2016; Waites and Talkington, 2004), was also among those genes deleted, this indicates that the streamlining process can work in tandem as an attenuation process as well.

Now that we have a working protocol, the future steps of the project will be to create multiple rounds of deletion, to see how much genomic material we can remove from the cell, and if there are any deletions that pre-dispose others, in the form of an epistatic interaction. With regard to the multiple deletion steps, this protocol does contain one factor that can be seen as both a positive and a negative. The introduction of the third transposon means that there is an extra knock-out stage in the protocol per round, as the transposon has a chance to disrupt a gene every time it is added. While our data from the initial transposon frequency experiments does show a proportionality high increase of transposition into non-coding regions (see Table 11), this is most likely due to those reads having the least fitness effect on the cell, and thus propagated the most during the experiment. The insertion of the extra transposon will inevitably cause the loss of some cells with lox sites in a viable deletion configuration due to the insertion of the Cre transposon into an essential gene, however the low efficiency of the suicide vector based system in protocol two means that this lower efficiency is still far higher than the alternatives.

The downside is a double-edged sword however, as the Cre transposon has the potential to disrupt non-essential genes as much as essential ones. It also has the potential to disrupt genes that are unlikely to be removed via the standard deletion system; for example small non-essential genes surrounded by essential or fitness genes. While protocol two showed that our system is capable of removing these genes, with the removal of the 1.1 Kb *glpD* gene, the closest similar event in the third protocol was the deletion of the MPN457 gene, a 3 Kb non-essential gene surrounded by essential genes. This should hopefully allow for a much larger mix of large and small deletions to work together in tandem.

2.5. Conclusion

We have presented here a novel methodology for the streamlining of the *M. pneumoniae* genome, and shown the steps taken to iterate the protocol into a workable tool for bio-engineering. We also found that the activity of the Cre recombinase on a single lox site is lethal in *M. pneumoniae*, and as such could have the potential to be used as a counter selective marker or kill switch in future projects. Our next steps will be to continue iterating the protocol to achieve multiple rounds of deletion, and to attempt to use it in other Mycoplasma species.

CHAPTER 3: COMPARISON AND ANALYSIS OF A PAN-BACTERIAL ESSENTIAL GENOME

This chapter describes the creation of, and results gathered from, a database designed to compile the essential genes from as many bacterial species as possible, and ascribe functions to the genes where appropriate. The aim was to collate as many diverse bacteria as possible, and identify any trends or changes in the essentiality of genes or functions as the complexity of the genome changed.

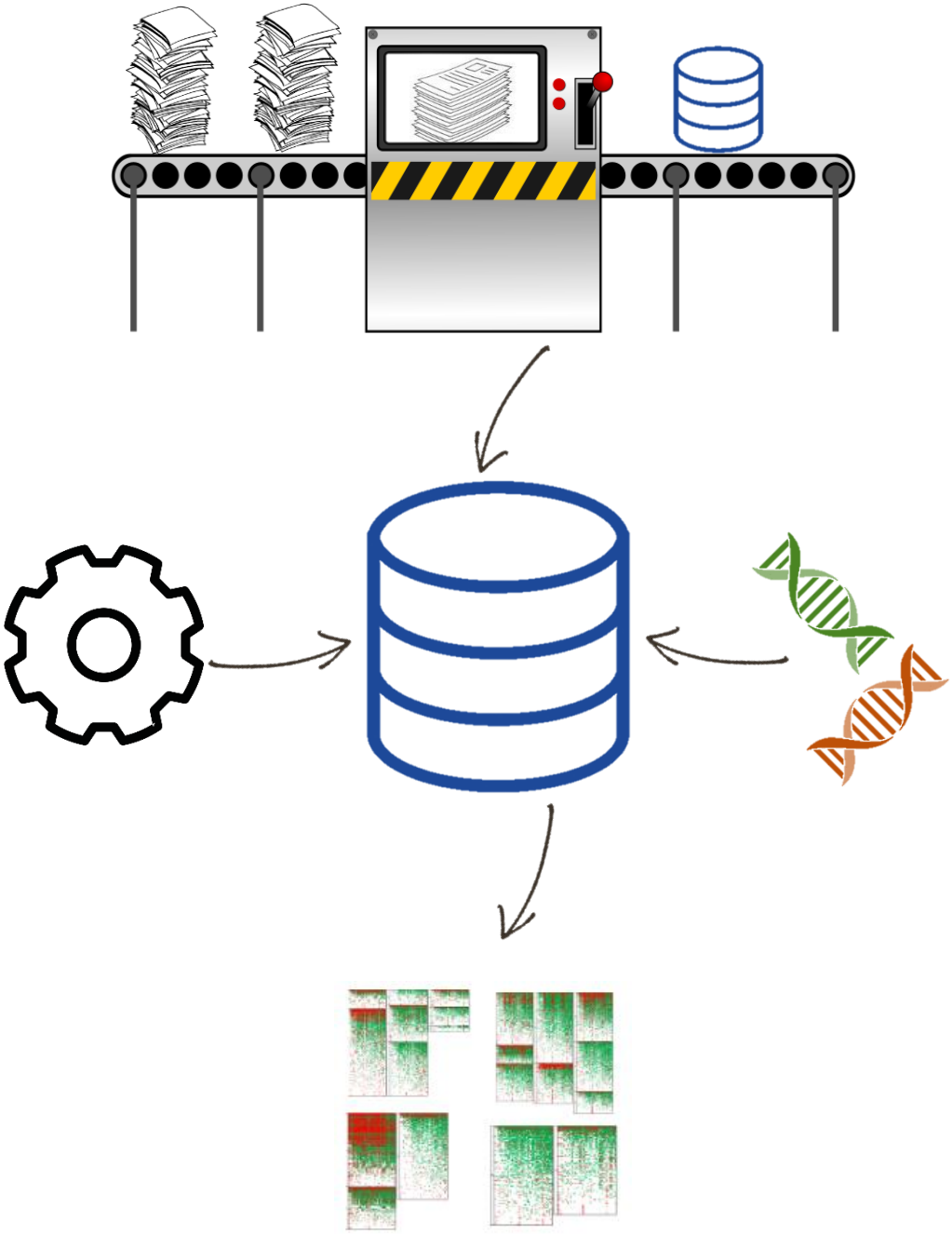


Figure 39: Graphical overview of Chapter 3, highlighting the key aspects of the database creation, annotation and analysis

3.1. Introduction

While there have been studies comparing the makeup of multiple bacterial genomes, and which genes are shared between them (Charlebois and Doolittle, 2004; Grazziotin et al., 2015; Juhas et al., 2011b; Klasson and Andersson, 2004; Koonin, 2003; Luo et al., 2015), no one has yet looked at this information in the context of essential genes. By looking at which genes are shared between multiple disparate species, we can begin to trace evolutionary lineages and infer core processes that are shared among many bacteria. However, by looking at which genes multiple organisms rely on as essential to their survival, we can see which processes become more or less important as genome composition changes. If a gene is generally essential in simple organisms but non-essential in more complex species, we can infer that its function is quickly backed up by the acquisition of new pathways, leading to redundancy (Yu et al., 2015). However, if a gene is non-essential in simple organisms but generally becomes more essential as the genome complexity increases, then it may act as a nexus or metabolite producing step for a function that has become essential. By looking at the trends in how genes become more or less essential as genomes become more or less complex, we can begin to understand: i) which processes these different organism rely more heavily on and ii) if there are any individual genes that seem to be retained for their essential function, or if that function is instead served by a plethora of different genes across different bacteria.

The advent of next generation sequencing has revolutionised the study of bacterial genomics, and the vast amount of sequenced data allows us to investigate the makeup and function of bacterial genomes in great depth (Land et al., 2015). Alongside this, the use of 16S rRNA sequencing has allowed much greater specificity in bacterial taxonomy, allowing us to construct far more accurate evolutionary lineages and species identifications (Parks et al., 2018). While the total number of bacterial species is both unknown and probably unknowable, estimates based on current knowledge of species diversity and distribution put the possible number of different bacterial species at upwards of 10^{12} (Locey and Lennon, 2016). The definition of a bacterial species is non-trivial by itself, with an early definition being a collection of strains sharing at least one diagnostic phenotypic trait and a minimum 70% cross-hybridisation rate in a DNA-DNA hybridisation test (Wayne et al., 1987). This indicated the need for sub-species and strains below the traditional species level. Due to the often rapid generation time and strong environmental factors encouraging variation, defining individual species in the way we do with multi-cellular organisms may not be universally applicable to the microbial world (Konstantinidis et al., 2006). Given this potential enormity of variation within the domain, analysing what similarities remain between them could shed light on how they were able to produce such diversity in the first place, and which processes are fundamentally important to single cellular life, regardless of environmental niche or lifestyle.

The main similarity, and reason that comparing disparate species may still allow for reasonable inferences to be drawn between them is the near universality of the genetic code. DNA is the genetic basis of all known forms of life, and this implies a common descent from a Last Universal Common Ancestor (LUCA), whose biochemistry is still foundational for all species (Mushegian, 2008). While the debate over the precise nature and traits of this organism is extensive (Di Giulio, 2011; Forterre, 2015; Koskela and Annila, 2012), this evolution from a common source implies there may be a basal level of similarity between their respective genomes (Grazziotin et al., 2015). By analysing the

genes that are present in bacterial species across the tree of life, and specifically which of these genes are essential to the survival of these bacteria, we can infer which processes are most strongly preserved, and which processes are ancillary to the overall survival of a bacterial cell.

Analysis of genes conserved among different species has been done previously, with 147 bacterial and archaeal genomes (130 bacteria, 17 archaea) showing a core of 34 genes conserved between all species (Charlebois and Doolittle, 2004). In this list of genes, the vast majority are related to information processing and biogenesis. Only a glycoprotein endopeptidase (*gcp*), ATPase subunit (*secY*) and GTP binding protein (*ychF*) are not directly involved in transcription or translation, compared to twelve tRNA synthetases, eleven ribosomal proteins, two translation elongation factors and various enzymes modulating DNA priming, RNA polymerisation, transcription initiation, pausing and termination (See Chapter 1.6.2 for more details). While the focus on transcription and translation is interesting, it is obvious that this conserved list is not enough by itself to sustain either process (Charlebois and Doolittle, 2004).

The lack of metabolic genes in this list is interesting, and can most likely be explained by niche disparity within the domain. Bacteria inhabit almost all known niches on the planet, from intracellular parasites to the Atacama desert (Finstad et al., 2017; Locey and Lennon, 2016; Zientz et al., 2004). As such, establishing a common metabolome under such vast differences in nutrient availability and composition is unlikely (Juhas et al., 2011a). When only 100 bacterial species were compared, the number of conserved genes shared between all present was 63 (Koonin, 2003), and when only 7 minimal endosymbionts and parasites were compared the value rises to 156 (Klasson and Andersson, 2004). In both cases, the number of conserved metabolic and structural genes rises comparatively to the decrease in diversity of the pool of organisms studied, as they are either closer evolutionarily, or have similarities in lifestyle or niche (endosymbionts and parasites).

This core-conserved genome is generally made up of genes that are indispensable for the survival of the organism. Typically, bacterial genes are classified as either essential or non-essential (Christen et al., 2011). Essential genes cannot be disrupted or removed without killing the bacteria, and are by definition non-dispensable to the survival of the organism. Usually, these genes are involved in core metabolism, biosynthesis and cell division. In contrast, non-essential can be deleted without initiating a lethal phenotype (Glass et al., 2017; Zomer et al., 2012).

The studies above revealed very little similarity between the core genomes of the studied bacteria. However, genomic architecture and the level of genetic redundancy in bacteria may play an important factor in this diversity. Genetic redundancy allows bacteria to survive the loss or disruption of an essential gene by containing genes that either duplicate the needed function (Lal et al., 2014) or can provide a moonlighting function to recover the lost phenotype, despite it not being the proteins original purpose (Kumar et al., 2015). Indeed, the idea of genetic redundancy for important systems is reinforced by the retention and persistence of many no-essential genes. In a study of 55 Firmicutes and gamma-proteobacteria, a class of genes defined as “persistent non-essential” was identified. These genes showed strong rates of retention across species and generally encoded for important cellular function related to stress response and cell maintenance (Fang et al., 2005). This raises an interesting philosophical question of why do bacteria contain essential genes at all? While the functions the proteins supply are unquestionably

essential, why would evolution not favour replication of essential genes to survive the loss of a single copy, or even favour a diploid genome? It is after all, the functions that the genes enable that are essential, not the genes themselves. Many genes encoding for essential functions have different essentiality profiles, depending on the context they are in (Zhang and Ren, n.d.).

Compared to eukaryotic genomes, bacteria have much smaller genomes concerning size and number of genes, but a much greater coding density. It is traditionally believed that as bacterial life is mainly predicated on growth, thus any excess genetic sequences will result in a fitness cost (Sela et al., 2016). As a result, it is beneficial for the organism to be as genetically streamlined as possible with regard to genome size and number of genes (Sela et al., 2016). While this is clearly only meant as a generic simplification, experimental data has shown that the ratio of synonymous vs non-synonymous nucleotide substitutions in bacteria is lower for essential genes than non-essential genes, despite this not being the case in studied eukaryotes (Jordan et al., 2002). This indicates some level of purifying selection among essential genes where their function appears highly important and evolutionarily retained, however the fitness cost of duplicating the genes appears too high.

3.1.1. Rationale

Herein, we attempted to collate information on essential genes from as multiple separate papers. From these, we tried to standardise the data as much as possible, retaining and analysing the data from the genomes of 47 diverse species of bacteria, and explore the relationship between the size of the genome with regard to coding capacity and the composition of the cells essential genome. To standardise the analysis of gene function, we chose to use the COG (Cluster of Orthologous Genes) classification system. This system is used to group genes into broad functional categories (Tatusov et al., 2000), and will allow us to look both at the individual genes conserved between species, but also the functions that they share and how the functionality of the essential genome changes across species.

We hypothesise that the size of the genome predicts the number of essential genes within it. While minimal genomes by their definition contain a very high proportion of essential genes (Glass et al., 2017), as complexity within the genome increases, redundancies in essential functions allow for fewer genes to be labelled essential (Mendonça et al., 2011). However, as genome size gets larger and more modules and functions are added, there is a proportional increase in the number of essential genes encoded. This relationship has been explored analogously by Basler (2016), showing that metabolic networks of different sizes demonstrate the same pattern with regard to the number of "driver reactions" (Basler et al., 2016). As the complexity of the system increases, more points of failure becoming present, and certain non-essential functions become integrated into new essential circuits, changing their original essentialities.

We aim to explore the relationship between the complexity of the bacterial genome and the function of its essential genes. We will study specifically: i) how the number of these genes changes, ii) which non-dispensable functions are added or removed as complexity increases, and iii) if there is any universally conserved aspects to the essential genomes, such as any universally essential genes that are conserved.

We hypothesise that as new essential functions area added, they interact with pre-existing pathways and change the essentiality profile of the pathways. Certain genes are nodes that were non-essential before may become hubs for new, essential pathways and thus the number of essential genes overall rises.

Discerning which non-essential pathways or genes are vital for the functions of other essential genes, and which essential genes are highly conserved vs which are niche specific, could help us explore the idea of how any evolutionarily conserved minimal genome is formed, and also help guide efforts to rationally design and produce minimal genomes *de novo*.

3.2. Material and Methods

3.2.1. Selection of candidate species

A PubMed search was initiated using the search terms "Essential genes" and "Bacteria". 107 separate entries were found listing the essential genes of a specific bacterial species, corresponding to 84 different bacteria across 68 papers.

These results were filtered to allow for a standardised comparison across the different data sets. As such, the studies were filtered via three categories.

One: They must have used mini-transposons to disrupt the genome.

Two: Insertions and thus gene essentiality were determined via a Next Generation Sequencing methodology.

Three: The paper must provide a list of genes deemed essential for the organism being studied.

3.2.2. Database creation and standardisation of genome annotations

47 entries matched the inclusion criteria and were thus analysed further. For each species, a genome assembly for the strain specified was downloaded for use within the database to map the essential genes against. Wherever possible, RefSeq datasets were preferred, or GenBank files if RefSeq was not available. If the species had a .gff annotation, this was then downloaded, as the positional information supplied could be used in tandem with the ProteinOrtho program (Lechner et al., 2011) for identification purposes. For any species without at .gff file, gene identities were established using the EDirect Entrez Programmining Utilites (Kans, 2019). These lists were parsed to remove all pseudogenes and RNAs from the analysis, ensuring that only protein coding sequences were analysed. For each species, the list of essential genes was extracted from the appropriate paper.

3.2.3. Assigning Essentiality status to each gene

From every paper selected, the list of genes deemed essential to the specific species was downloaded and an identifying factor chosen, either a provided RefSeq ID, genomic loci, protein ID or gene name. These were queried against the database, and if a match was

found that entry was marked essential. Entries that did not automatically match to the database were annotated manually. Pseudogenes and RNAs were automatically excluded, ensuring the database contained only protein coding genes. Genes with no matches were deemed non-essential. This resulted in a database containing every gene from every species linked to a RefSeq ID and essentiality status.

3.2.4. Assigning COG classes

For each gene in the database, its RefSeq ID was queried against the COG database. All genes with a RefSeq ID or GenBank ID that matched an entry within the COG database was ascribed with the relevant COG class.COGs were also grouped into four Super-COGs for general analysis. These Super-COGs consisted of the following COG classes:

Table 16:Composition of Super-COG classes

Cellular Processes and Signalling	Information Storage and processing	Metabolism	Unknown Function
[B] Chromatin structure and dynamics	[J] Translation, ribosomal structure and biogenesis	[C] Energy production and conversion	[R] General function prediction only
[D] Cell cycle control, cell division, chromosome partitioning	[K] Transcription	[E] Amino acid transport and metabolism	[S] Function unknown
[M] Cell wall/membrane/envelope biogenesis	[L] Replication, recombination and repair	[F] Nucleotide transport and metabolism	No Assigned COG
[N] Cell motility		[G] Carbohydrate transport and metabolism	
[O] Post-translational modification, protein turnover, and chaperones		[H] Coenzyme transport and metabolism	
[T] Signal transduction mechanisms		[I] Lipid transport and metabolism	
[U] Intracellular trafficking, secretion, and vesicular transport		[P] Inorganic ion transport and metabolism	
[V] Defence mechanisms		[Q] Secondary metabolites biosynthesis, transport, and catabolism	
[W] Extracellular structures			
[X] Mobilome: prophages, transposons			
[Z] Cytoskeleton			

3.2.5. Gene clustering

All genes within the database were ran through the ProteinOrtho program to identify homologues. Each set of homologous genes was clustered together, and if any containing a COG classification, this was ascribed to all genes within that homologous cluster. If a gene contained multiple COG classes, and one of which was class "S" (Unknown function), the S classification was removed from the gene, as by definition it was no longer appropriate. This allowed the database to update and contain both the essentiality and COG class of every gene in every species it contained.

By parsing this clustering, the database could then be queried to identify how many genes in each cluster were Essential or not, how many unique clusters were ascribed to each COG, and how many different species were in each cluster.

3.3. Results

3.3.1. Phylogeny of analysed species

From the 107 entries identified initially, 47 entries fit the inclusion criteria to be included in the study. The 47 species analysed are listed in
Table 17. All 47 species were selected as outlined in 2.1, with two exceptions. *E. coli* and *B. subtilis* were included using the datasets from Baba et al., (2006) and Kobayashi et al., (2003) respectively. These studies did not use transposon mutagenesis but systematic knockouts of all genes.

*Table 17: Subset of 47 species that were used for the analysis. Given are the internal code used throughout the analysis for each species and their full strain identification (where possible) with NCBI taxonomy ID. * Average number of bases per transposon insertion*

Internal code	Species	NCBI Taxonomy	No. E genes	No. genes	% E genes	Transposon Coverage*	Reference
1	*Porphyromonas gingivalis* ATCC 33277	431947	463	2155	21.48	43.61	Klein *et al.,* 2012
2	*Burkholderia pseudomallei* K96243	272560	505	5807	8.70	30.20	Moule *et al.,* 2014
9	*Vibrio cholerae* O1 biovar El Tor str. N16961	243277	343	3625	9.46	8.42	Chao *et al.,* 2013
17	*Caulobacter cresentus* NA1000	565050	480	4077	11.77	7.65	Christen *et al.,* 2011
20	*Mesoplasma florum* L1	265311	290	715	40.56	280.00	Baby *et al.,* 2018
24	*Acinetobacter baumannii* ATCC 17978	400667	458	3887	11.78	26.68	Wang *et al.,* 2014
26	*Agrobacterium tumefaciens* C58	176299	372	5430	6.85	6.23	Curtis and Brun, 2015
27	*Brevundimonas subvibrioides*	633149	447	3383	13.21	2.95	Curtis and Brun, 2015
29	*Bacillus thuringiensis* BMB171	714359	516	5495	9.39	N/A	Bishop *et al.,* 2014
30	*Bacteroides fragilis* 638R	862962	550	4382	12.55	53.73	Veeranagouda *et al.,* 2014
31	*Bacteroides thetaiotaomicron* VPI-5482	226186	325	4902	6.63	179.81	Goodman *et al.,* 2009
39	*Mycobacterium tuberculosis* H37Rv	83332	742	3976	18.66	58.88	Zhang *et al.,* 2012
45	*Rhodopseudomonas palustris* CGA009	258594	552	4882	11.31	31.00	Pechter *et al.,* 2016
47	*Sphingomonas wittichii* RW1	392499	535	5401	9.91	134.44	Roggo *et al.,* 2013
50	*Streptococcus agalactiae* GBS strain A909	205921	317	2094	15.14	15.57	Hooven *et al.,* 2016

52	*Streptococcus pyogenes* M49 NZ131	471876	227	1766	12.85	21.36	Le Breton *et al.*, 2015
57	*Methylobacterium extorquens* PA1	419610	590	4904	12.03	10.80	Ochsner *et al.*, 2017
58	*Aggregatibacter actinomycetemcomitans* 624	714	413	1912	21.60	89.00	Narayanan *et al.*, 2017
60	*Proteus mirabilis* HI4320	529507	436	3767	11.57	50.90	Armbruster *et al.*, 2017
61	*Herbaspirillum seropedicae* SmR1	757424	395	4799	8.23	95.00	Rosconi *et al.*, 2016
64	*Rubrivivax gelatinosus*	983917	388	4756	8.16	8.89	Curtis, 2016
65	*Liberibacter crescens* BT-1	1215343	314	1432	21.93	354.04	Lai *et al.,* 2016
68	*Azospirillum brasilense* Sp245	1064539	340	7673	4.43	71.08	Price *et al.,* 2018
69	*Burkholderia phytofirmans* PsJN	398527	409	7323	5.59	95.76	Price *et al.,* 2018
71	*Cupriavidus basilensis* 4G11	68895	474	7751	6.12	43.66	Price *et al.,* 2018
72	*Dechlorosoma suillum* PS	640081	584	3507	16.65	15.74	Price *et al.,* 2018
74	*Dinoroseobacter shibae* DFL-12	398580	535	4244	12.61	36.91	Price *et al.,* 2018
75	*Dyella japonica* UNC79MFTsu3.2	1380365	371	4390	8.45	32.65	Price *et al.,* 2018
76	*Echinicola vietnamensis*	390884	492	4606	10.68	20.15	Price *et al.,* 2018
79	*Kangiella aquimarina* SW-154T	523791	399	2514	15.87	28.50	Price *et al.,* 2018
81	*Marinobacter adhaerens* HP15	225937	555	4470	12.42	54.27	Price *et al.,* 2018
83	*Phaeobacter gallaeciensis* DSM 26640	1423144	538	4389	12.26	20.88	Price *et al.,* 2018
84	*Pontibacter actiniarum*	323450	472	4322	10.92	26.05	Price *et al.,* 2018
85	*Pseudomonas fluorescens* FW300-N1B4	294	426	6169	6.91	23.57	Price *et al.,* 2018
90	*Pseudomonas simiae* WCS417	321846	473	5624	8.41	56.01	Price *et al.,* 2018
91	*Pseudomonas stutzeri* RCH2	644801	430	4348	9.89	27.64	Price *et al.,* 2018
92	*Shewanella amazonensis* SB2B	326297	394	3775	10.44	11.06	Price *et al.,* 2018
93	*Shewanella loihica* PV-4	323850	289	3983	7.26	90.08	Price *et al.,* 2018
96	*Sinorhizobium meliloti* 1021	266834	559	6288	8.89	47.95	Price *et al.,* 2018
99	*Mycoplasma pneumoniae* M129	272634	342	694	49.28	11.66	Lluch-Senar *et al.,* 2015
100	*Mycoplasma agalactiae* 5632	347258	303	689	43.98	4.17	Montero-Blay *et al.,*(in preparation)
101	*Bacillus subtilis* subsp. subtilis str. 168	224308	269	4352	6.18	N/A	Kobayashi *et al.,* 2003
103	*Escherichia coli* str. K-12 substr. MG1655	511145	300	4419	6.79	N/A	Baba *et al.,*2006
104	*Francisella tularensis* subsp. novicida U112	401614	395	1767	22.35	115.70	Gallagher *et al.*, 2007
105	Pseudomonas aeruginosa PAO1	208964	336	5678	5.92	62.64	Turner *et al.*, 2015
106	*Rhizobium leguminosarum* bv. viciae 3841	216596	292	7131	4.09	66.51	Perry & Yost, 2014
107	*Synechococcus elongatus* PCC 7942	1140	718	2714	26.46	11.00	Rubin *et al.,* 2015

From the NCBI Taxonomy IDs, a phylogenetic tree of all 47 strains was constructed:

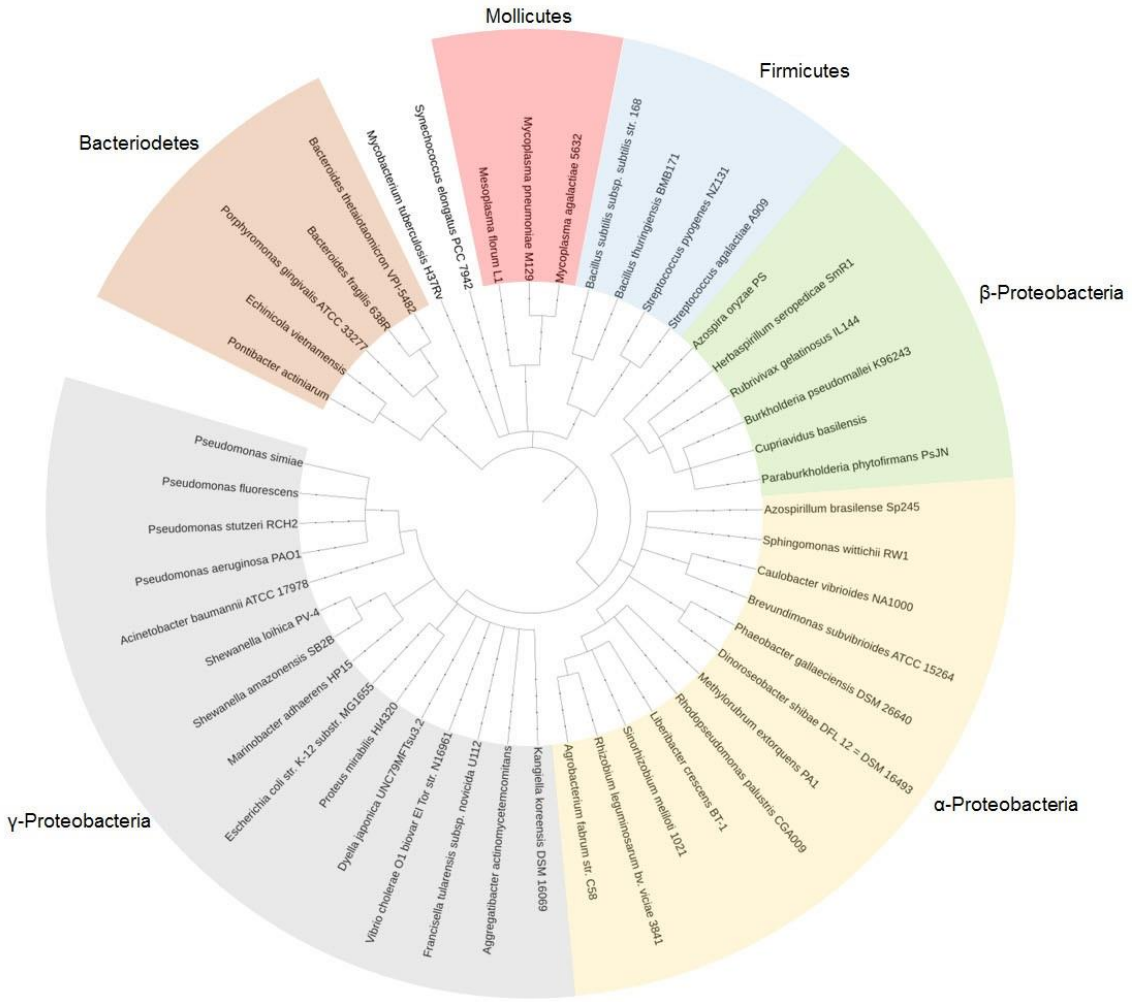


Figure 40: Phylogenetic tree of the 47 analysed species, denoting the major phyla found within the study. The tree was generated using NCBI Taxonomy IDs for each species via PhyloT (https://phylot.biobyte.de) and visualised using iTOL (https://itol.embl.de). The major Phyla are highlighted in unique colours, with the actinobacteria and cyanobacteria left blank on account of having only one species represented.

While the bias towards proteobacteria is clear, there is also representation from the other important phyla such as the Bacteriodetes, Firmicutes and Mollicutes. Phyla that have only one representative species include the Cyanobacteria (*S. elongatus*) and Actinobacter (*M. tuberculosis*).

Table 18: Phyla represented in the 47 analysed species

Phlyum	No. Species	% of Species
γ-proteobacteria	15	31.91
α-proteobacteria	12	25.53
β-proteobacteria	6	12.77
Bacteriodetes	5	10.64
Firmicutes	4	8.51
Mollicutes	3	6.38
Cyanobacteria	1	2.13

Actinobacteria	1	2.13

3.3.2. Genome Size vs Essentiality

To study the relationship between genome size and essentiality, we first studied the correlation between 107 lists of essential genes, covering 84 different bacterial species. The relationship between the number of genes present in the genome and the number of essential genes is shown in Figure 41.

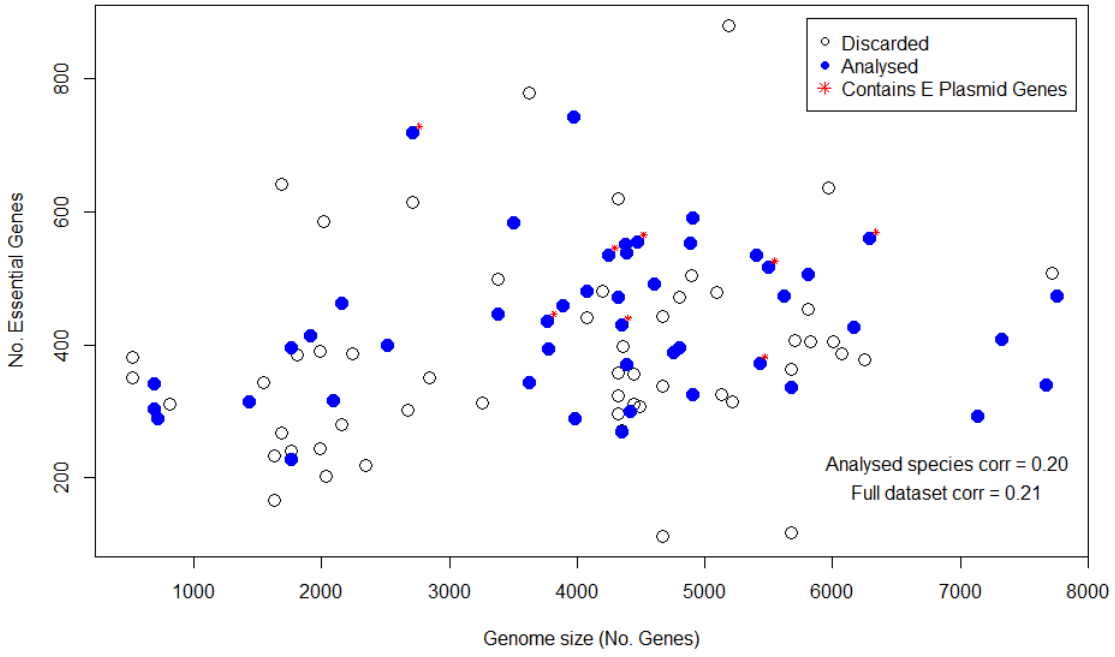


Figure 41: Relationship between number of geens in the genome and the number of essential genes. Species denoted with a blue dot were retained in the study. For inclusion critera, see chapter 3.2.1

The correlation between the genome size and complement of essential genes is low but present, with a Pearson's Product-Moment correlation 0.2 for both the full set of 107 species, and the further analysed subset of 47. This indicates that the subset is representative of the overall trend in regard to the genome size vs essentiality paradigm. When the percentage of genes that are essential were plotted against the total genome, a much clearer pattern emerged:

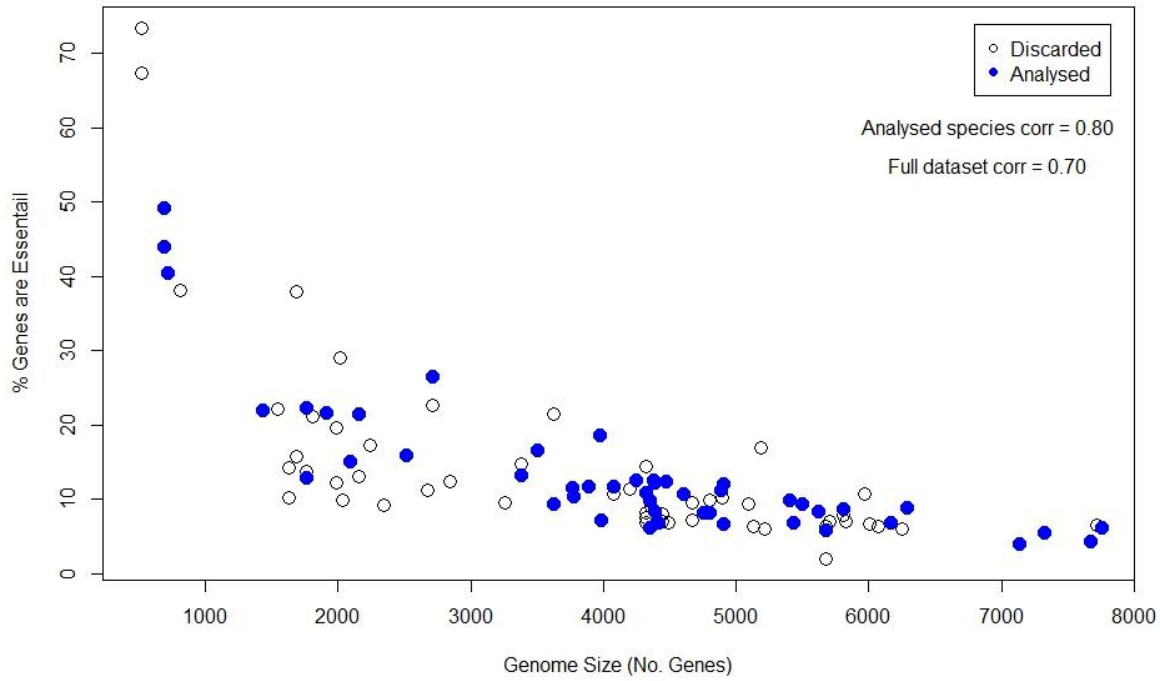


Figure 42: Genome size vs Percentage of genes that are essential, showing the general trend of increasing bacterial size correlating with fewer essential genes as an overall percentage of the genome.

Again, the subset of analysed species are representative of the overall trend. However, there is a clear trend of as the number of genes increases, the overall percentage of those genes which are essential drops, although their number increases as shown in Figure 41.

3.3.3. Database composition by COG class

Analysis of our database showed the following distribution of each COG Class (Table 19). This refers to the number of times each COG class was represented. As some genes were classified into multiple COG classes that gene would be represented in all of the COG classes it is classified in. For example, the *ftsP* cell division protein is annotated as belonging to the COG classes D, M and P, as its function fits into all three categories. The database is also non-redundant, so as the *ftsP* is present in 12 species, its functions are accounted for 12 times in the database. Therefore, Table 19 represents the total number of functions within all species the database, not specifically the number of genes.

Table 19: Database composition organised via COG class

COG Class	No. functions	No. Essential Functions	% Essential Functions
[J] Translation, ribosomal structure and biogenesis	8916	3603	40.4
[D] Cell cycle control, cell division, chromosome partitioning	1347	484	35.9
[F] Nucleotide transport and metabolism	2969	755	25.4
[H] Coenzyme transport and metabolism	6177	1515	24.5
[M] Cell wall/membrane/envelope biogenesis	8345	1764	21.2
[L] Replication, recombination and repair	4582	911	19.9
[U] Intracellular trafficking, secretion, and vesicular transport	2039	405	19.7
[I] Lipid transport and metabolism	5691	971	17.1
[C] Energy production and conversion	7349	1189	16.2
[O] Post-translational modification, protein turnover, and chaperones	5417	688	12.7
[E] Amino acid transport and metabolism	11853	1122	9.5
[G] Carbohydrate transport and metabolism	7117	529	7.4
[K] Transcription	10829	712	6.6
[Q] Secondary metabolites biosynthesis, transport, and catabolism	3643	239	6.6
[V] Defence mechanisms	3180	206	6.5
[S] Function unknown	6679	364	5.4
NO COG	82899	4396	5.3
[P] Inorganic ion transport and metabolism	8117	392	4.9
[X] Mobilome: prophages, transposons	1327	64	4.8
[T] Signal transduction mechanisms	8229	377	4.6
[N] Cell motility	2665	116	4.4
[R] General function prediction only	12794	537	4.2
[W] Extracellular structures	771	25	3.2
[B] Chromatin structure and dynamics	28	0	0.0
[Z] Cytoskeleton	20	0	0.0
All functions	212983	21365	10.0

As a whole, the database contains 191341 distinct genes, with 19856 classified as essential in at least one species, and the remaining 171493 as non-essential. These can be further split into 63923 clusters of orthologs.

3.3.4. Conserved genes across all 47 species

Analysis of the genomes of all 47 species revealed 92 universally conserved genes. Of these, only one (the chromosome replication initiator *dnaA*) was classified as essential in every species.

Table 20: 92 genes conserved in all 47 species.

**RefSeq ID of M. pneumoniae M129 homolog*

†Most commonly applied gene name

‡Percentage of species where the gene is essential

• Percentage of species that have at least one replicate of the gene

RefSeq ID*	Gene†	Function	% ESSENTIALITY‡	COG CATEGORY	% species with replicas•
NP_110375.1	*dnaA*	Chromosomal replication initiator protein DnaA	100	L	0
NP_109793.1	*pheS*	Phenylalanine--tRNA ligase alpha subunit	98	J	0
NP_109733.1	*hisS*	Histidine--tRNA ligase	96	J	2
NP_109856.1	*rplB*	50S ribosomal protein L2	96	J	0
NP_109872.1	*secY*	Protein translocase subunit SecY	96	U	2
NP_109691.1	*gyrB*	DNA gyrase subunit B	94	L	0
NP_109734.1	*aspS*	Aspartate--tRNA ligase	94	J	0
NP_109853.1	*rplC*	50S ribosomal protein L3	94	J	2
NP_109879.1	*rpoA*	DNA-directed RNA polymerase subunit alpha	94	K	2
NP_110066.1	*dnaE*	DNA polymerase III subunit alpha	94	L	0
NP_110204.1	*rpoB*	DNA-directed RNA polymerase subunit beta	94	K	0
NP_109693.1	*serS*	Serine--tRNA ligase	91	J	0
NP_109859.1	*rpsC*	30S ribosomal protein S3	91	J	4
NP_109868.1	*rplF*	50S ribosomal protein L6	91	J	0
NP_110072.1	*leuS*	Leucine--tRNA ligase	91	J	2
NP_110107.1	*alaS*	Alanine--tRNA ligase	91	J	0
NP_109692.1	*gyrA*	DNA gyrase subunit A	89	L	0
NP_109748.1	*metK*	S-adenosylmethionine synthase	89	H	0
NP_109854.1	*rplD*	50S ribosomal protein L4	89	J	0
NP_110113.1	*ftsY*	Signal recognition particle receptor FtsY	89	U	0
NP_110168.1	*valS*	Valine--tRNA ligase	89	J	0
NP_110242.1	*thrS*	Threonine--tRNA ligase	89	J	4
NP_109860.1	*rplP*	50S ribosomal protein L16	87	J	0
NP_109870.1	*rpsE*	30S ribosomal protein S5	87	J	0
NP_109865.1	*rplE*	50S ribosomal protein L5	85	J	0
NP_109877.1	*rpsM*	30S ribosomal protein S13	85	J	0
NP_109914.1	*rpsG*	30S ribosomal protein S7	85	J	0
NP_109915.1	*fusA*	Elongation factor G	85	J	11
NP_109934.1	*gmk*	Guanylate kinase	85	F	2
NP_110005.1	*ftsZ*	Cell division protein FtsZ	85	D	6
NP_110041.1	*dnaG*	DNA primase	85	L	0
NP_110045.1	*ligA*	DNA ligase	85	L	0
NP_110227.1	*rplJ*	50S ribosomal protein L10	85	J	2
NP_109867.1	*rpsH*	30S ribosomal protein S8	83	J	0
NP_110289.1	*atpA*	ATP synthase subunit alpha	83	C	0
NP_110306.1	*rplM*	50S ribosomal protein L13	83	J	0
NP_110049.1	*prfA*	Peptide chain release factor 1	81	J	0

NP_110257.1	*era*	GTPase Era	81	J	0
NP_109842.1	*nusA*	Transcription termination/antitermination protein NusA	79	K	0
NP_109878.1	*rpsK*	30S ribosomal protein S11	79	J	0
NP_109953.1	*trpS*	Tryptophan--tRNA ligase	79	J	15
NP_110360.1	*ftsH*	ATP-dependent zinc metalloprotease FtsH	79	O	2
NP_109871.1	*rplO*	50S ribosomal protein L15	77	J	0
NP_109873.1	*adk*	Adenylate kinase	77	F	6
NP_109898.1	*secA*	Protein translocase subunit SecA	77	U	0
NP_109907.1	*rplK*	50S ribosomal protein L11	77	J	0
NP_109908.1	*rplA*	50S ribosomal protein L1	77	J	0
NP_109913.1	*rpsL*	30S ribosomal protein S12	77	J	0
NP_110110.1	*mnmA*	tRNA-specific 2-thiouridylase MnmA	77	J	6
NP_109805.1	*rplT*	50S ribosomal protein L20	74	J	0
NP_109863.1	*rplN*	50S ribosomal protein L14	72	J	4
NP_109869.1	*rplR*	50S ribosomal protein L18	72	J	0
NP_110305.1	*rpsI*	30S ribosomal protein S9	72	J	0
NP_109852.1	*rpsJ*	30S ribosomal protein S10	70	J	12
NP_109858.1	*rplV*	50S ribosomal protein L22	70	J	4
NP_109862.1	*rpsQ*	30S ribosomal protein S17	70	J	4
NP_109880.1	*rplQ*	50S ribosomal protein L17	70	J	0
NP_110122.1	*dnaK*	Chaperone protein DnaK	70	O	4
NP_110252.1	*obg*	GTPase Obg	70	D,L	2
NP_109864.1	*rplX*	50S ribosomal protein L24	68	J	0
NP_110348.1	*trmD*	tRNA (guanine-N(1)-)-methyltransferase	68	J	0
NP_110232.1	*fmt*	Methionyl-tRNA formyltransferase	66	J	0
NP_109857.1	*rpsS*	30S ribosomal protein S19	62	J	0
NP_109918.1	*rpsR*	30S ribosomal protein S18	62	J	4
NP_109874.1	*map*	Methionine aminopeptidase	60	J	32
NP_110015.1	*rpmA*	50S ribosomal protein L27	60	J	0
NP_109875.1	*infA*	Translation initiation factor IF-1	55	J	23
NP_110117.1	*pgk*	Phosphoglycerate kinase	55	G	0
NP_110347.1	*rplS*	50S ribosomal protein L19	55	J	0
NP_109939.1	*rpe*	Probable ribulose-phosphate 3-epimerase	53	G	6
NP_110311.1	*rpsO*	30S ribosomal protein S15	53	J	2
NP_109717.1	*efp*	Elongation factor P	51	J	12
NP_109709.1	*dnaJ*	Chaperone protein DnaJ	49	O	2
NP_109804.1	*rpmI*	50S ribosomal protein L35	49	J	0
NP_110106.1	*ruvX*	Holliday junction resolvase	49	K,L	0
NP_110003.1	*rsmH*	Ribosomal RNA small subunit methyltransferase H	45	J,M	0
NP_110230.1	*rpsT*	30S ribosomal protein S20	45	J	0
NP_110313.1	*rpmB*	50S ribosomal protein L28	45	J	2
NP_110371.1	*rpmH*	50S ribosomal protein L34	43	J	0
NP_110265.1	*glyA*	Serine hydroxymethyltransferase	40	E	23
NP_110354.1	*tuf*	Elongation factor Tu	38	J	28
NP_109876.1	*rpmJ*	50S ribosomal protein L36	36	J	19
NP_109762.1	*smpB*	SsrA-binding protein	34	O	0

NP_110048.1	*rpmE*	50S ribosomal protein L31	34	J	15
NP_110368.1	*rsmA*	Ribosomal RNA small subunit methyltransferase A	19	J	0
NP_110224.1	*ruvA*	Holliday junction ATP-dependent DNA helicase RuvA	15	L	0
NP_109919.1	*rplI*	50S ribosomal protein L9	13	J	0
NP_109980.1	*rluA*	RluA family RNA pseudouridine synthase	9	J	0
NP_109714.1	*ychF*	Ribosome-binding ATPase YchF	4	J	0
NP_109759.1	*rsmI*	Ribosomal RNA small subunit methyltransferase I	4	J	0
NP_109899.1	*uvrB*	UvrABC system protein B	2	L	2
NP_110308.1	*uvrA*	UvrABC system protein A	2	L	0

As predicted by other studies, these conserved genes have a predominantly focus on functions related to transcription, DNA recombination/repair and translation, consisting of 83% of the functions present. The remaining 17% are split between nine COG classes, which unlike previous studies include metabolism genes. For all genes with multiple COG classes, such as GTPase Obg (*obg*), both functions were treated separately. Therefore while the list contains 92 genes, the 95 separate functions present are used to calculate the overall percentage of functionality.

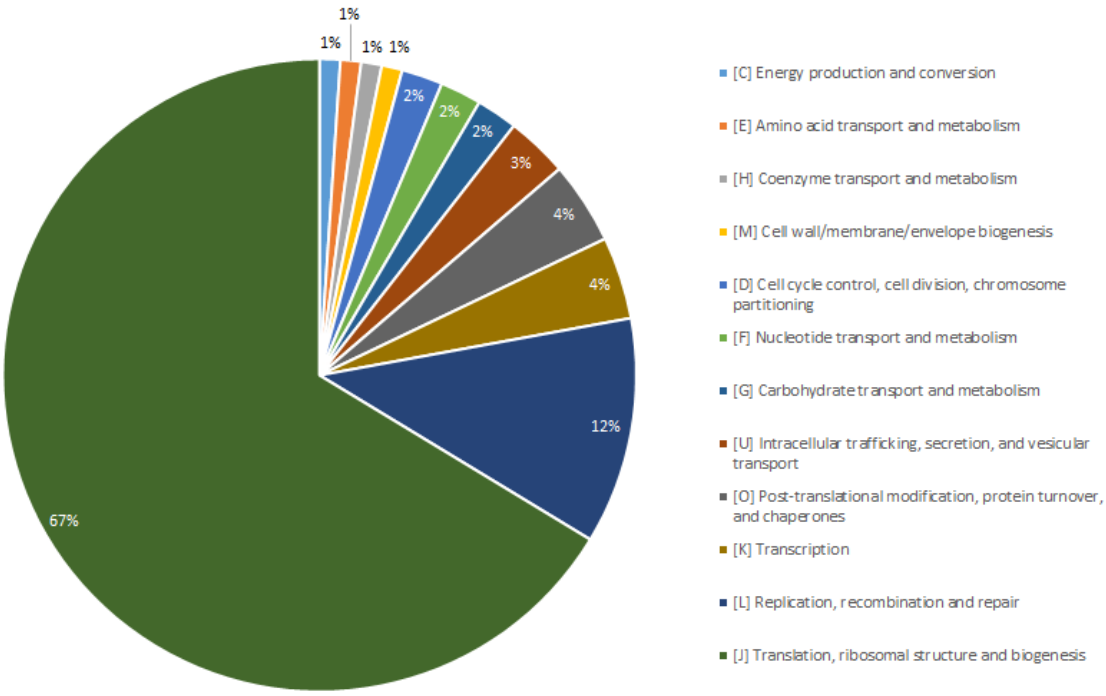


Figure 43: COG Class composition of genes conserved in all 47 species

Interestingly, the number of conserved genes that have replicates is very low, with only 34 of the 92 genes having a duplicate copy in at least one species of the bacteria studied. Furthermore, the presence of duplicate copies does not explain the loss of essentiality across the board, as none of the genes can explain the loss of essentiality by the presence of duplicates alone. While many come close, for example the tryptophan tRNA ligase is essential in 79% of the species, and the gene is duplicated in another 15% where it is non-essential. This still leaves three species where the gene is no duplicated and non-essential.

3.3.5. Conserved genes across all species compared to previous data

In order to test if out search and clustering algorithms were working, we looked at how our list of conserved genes compared to other known lists. Previous analysis of 100 bacterial genomes (Charlebois and Doolittle, 2004) revealed a list of 34 genes that were conserved among all species. Comparing our list of universally conserved genes to theirs, there is a broad consensus between the two:

Table 21: List of universally conserved bacterial genes from Charlebois & Doolittle (2004), with the percentage of species in our analysis we find that gene present in.

RefSeq ID	Gene	Function	% Species conserved
NP_110245.1	*argS*	Arginyl-tRNA synthetase	98
NP_109965.1	*lysS*	Lysyl-tRNA synthetase	81
NP_110367.1	*gltX*	Glutamyl-tRNA synthetase	98
NP_109711.1	*metG*	Methionyl-tRNA synthetase	91
NP_385392.1	*proS*	Prolyl-tRNA synthetase	28
NP_110090.1			19
WP_011866393.1			57
NP_109843.1	*infB*	Translation initiation factor IF-2	98
NP_385449.1	*nusG*	Transcription antitermination	98*
NP_109747.1	*gcp*	O-sialoglycoprotein endopeptidase (tsaD)	98
NP_109716.1	*rpsB*	Ribosomal protein S2	98
NP_110134.1	*rpsD*	Ribosomal protein S4	94
NP_109872.1	*secY*	ATPase subunit of translocase	100
NP_110041.1	*dnaG*	DNA Primase	100
NP_109714.1	*ychF*	GTP binding protein (engD)	100
NP_109733.1	*hisS*	Histidyl-tRNA synthetase	100
NP_110072.1	*leuS*	Leucyl-tRNA synthetase	100
NP_109793.1	*pheS*	Phenylalanyl-tRNA synthetase	100
NP_109693.1	*serS*	Seryl-tRNA synthetase	100
NP_110242.1	*thrS*	Threonyl-tRNA synthetase	100
NP_109953.1	*trpS*	Tryptophanyl-tRNA synthetase	100
NP_110168.1	*valS*	Valyl-tRNA synthetase	100
NP_109908.1	*rplA*	Ribosomal protein L1	100
NP_109907.1	*rplK*	Ribosomal protein L11	100
NP_109863.1	*rplN*	Ribosomal protein L14	100
NP_109853.1	*rplC*	Ribosomal protein L3	100
NP_109865.1	*rplE*	Ribosomal protein L5	100
NP_109868.1	*rplF*	Ribosomal protein L6	100
NP_109859.1	*rpsC*	Ribosomal protein S3	100
NP_109914.1	*rpsG*	Ribosomal protein S7	100
NP_109867.1	*rpsH*	Ribosomal protein S8	100
NP_110204.1	*rpoB*	RNA polymerase, β subunit	100
NP_110368.1	*ksgA*	S-adenosylmethionine-6-N',N'-adenosyl (rRNA) dimethyltransferase	100
NP_109842.1	*nusA*	Transcription pausing, L factor	100
NP_109915.1	*fusA*	Translation elongation factor EF-G	100
NP_110354.1	*tufA*	Translation elongation factor EF-Tu	100

* *M. pneumoniae* is the only species not to contain the NP_385449.1 version of the *nusG* gene that is shared by all other species in this list, instead containing a different version, NP_109755.1, which according to a BLAST of the sequence against the NCBI database, it shares only with *M. genitalium*. Therefore, technically all species present contain a *nusG* gene.

24/34 of the universal genes identified by Charlebois & Doolittle were also universal in our dataset. Those genes that were not universally conserved were only missing in a very small number of species.

The large exception is the Prolyl-tRNA synthetase. While a version of the gene was found in every species, there appear to be three distinct paralogs of the gene located in our database. They fall broadly within phylogenetic families, as shown in Figure 44, and there are only two species that contain copies of two different proS genes. The first is *B. thuringiensis*, a Firmicute which contains the NP_110090.1 and WP_011866393.1 versions, both of which are classified as non-essential. The second species is *B. phytofirmans*, a β-proteobacteria which contains a copy of the NP_385392.1 and WP_011866393.1 genes. In *B. phytofirmans*, the NP_385392.1 copy is essential while the WP_011866393.1 is not, making *B. phytofirmans* the only species not an α-proteobacteria to contain the NP_385392.1 gene, and the only species to have two copies and have one be essential.

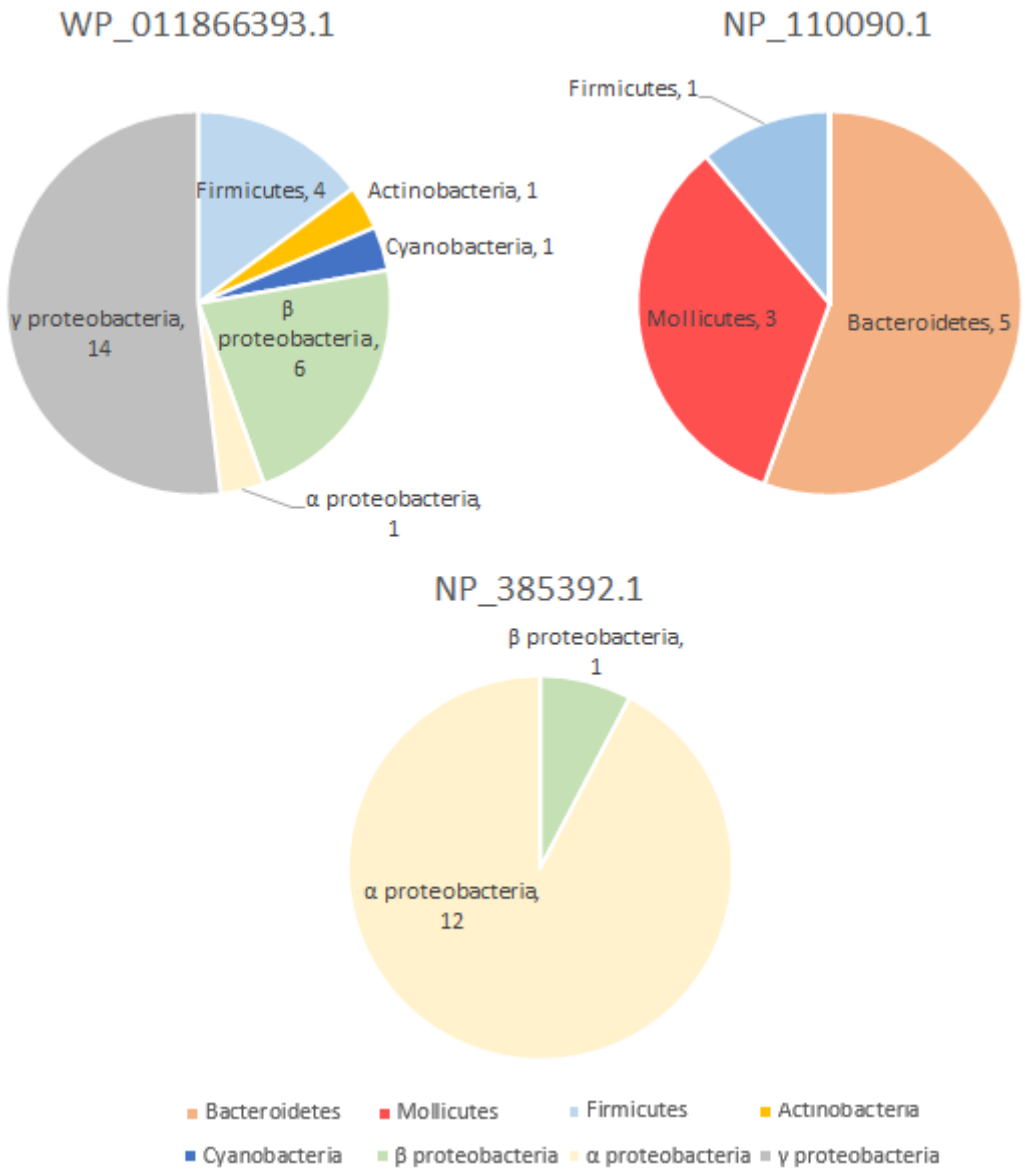


Figure 44: Distribution of how the three different proS genes are grouped in our database, represented on a phyla level, using the same colour codes as shown in Figure 40. Each pie represents a different proS gene, with relevant RefSeq ID above each.

3.3.6. Ribosomal protein internal control

In order to test if our search and clustering algorithms were working, we looked at how our list of conserved genes compared to other known lists, specifically for the ribosomal proteins. These were chosen as a large number of different ribosomal proteins are supposedly conserved across all bacterial species. As such, the ribosomal proteins we found to be fully conserved were tested against the list produced by Yutin et al., (2012) who compared the sequences of 995 completely sequenced bacteria and 87 archaea. This comparison included multiple sequence alignements for each protein, which were used to generate a Hidden Markov Model (HMM) based search algorithm to query our database for any ribosomal proteins. Using both search methods, our standard search function and the HMM search function, we split the results into two lists, one detailing all the ribosomal proteins that are conserved across all domains of life (bacteria, archaea and eukaryotes), along with the ribosomal proteins conserved in all bacterial species:

Table 22: Comparison between lists of conserved ribosomal proteins.

Universal Ribosomal Proteins		
Yutin *et al.*	Original List	HMM List
50S Ribosomal protein L1	50S ribosomal protein L1	50S ribosomal protein L1
50S Ribosomal protein L2	50S ribosomal protein L2	50S ribosomal protein L2
50S Ribosomal protein L3	50S ribosomal protein L3	50S ribosomal protein L3
50S Ribosomal protein L4	50S ribosomal protein L4	50S ribosomal protein L4
50S Ribosomal protein L5	50S ribosomal protein L5	50S ribosomal protein L5
50S Ribosomal protein L6	50S ribosomal protein L6	50S ribosomal protein L6
50S Ribosomal protein L10	50S ribosomal protein L10	50S ribosomal protein L10
50S Ribosomal protein L11	50S ribosomal protein L11	50S ribosomal protein L11
50S Ribosomal protein L12		
50S Ribosomal protein L13	50S ribosomal protein L13	50S ribosomal protein L13
50S Ribosomal protein L14	50S ribosomal protein L14	50S ribosomal protein L14
50S Ribosomal protein L15	50S ribosomal protein L15	30S ribosomal protein L15
50S Ribosomal protein L18	50S ribosomal protein L18	50S ribosomal protein L18
50S Ribosomal protein L22	50S ribosomal protein L22	50S ribosomal protein L22
50S Ribosomal protein L23		
50S Ribosomal protein L24	50S ribosomal protein L24	50S ribosomal protein L24
50S Ribosomal protein L29		
30S Ribosomal protein S2		
30S Ribosomal protein S3	30S ribosomal protein S3	30S ribosomal protein S3
30S Ribosomal protein S4		
30S Ribosomal protein S5		**30S ribosomal protein S5**
30S Ribosomal protein S7	30S ribosomal protein S7	30S ribosomal protein S7
30S Ribosomal protein S8		**30S ribosomal protein S8**
30S Ribosomal protein S9	30S ribosomal protein S9	30S ribosomal protein S9
30S Ribosomal protein S10	30S ribosomal protein S10	30S ribosomal protein S10
30S Ribosomal protein S11	30S ribosomal protein S11	30S ribosomal protein S11
30S Ribosomal protein S12	30S ribosomal protein S12	30S ribosomal protein S12
30S Ribosomal protein S13	30S ribosomal protein S13	30S ribosomal protein S13
30S Ribosomal protein S14		
30S Ribosomal protein S15	30S ribosomal protein S15	30S ribosomal protein S15
30S Ribosomal protein S17	30S ribosomal protein S17	30S ribosomal protein S17
30S Ribosomal protein S19		**30S ribosomal protein S19**
32/32	23/32	26/32
Ribosomal proteins found in Bacteria only		
Yutin *et al.*	Original List	HMM List
50S Ribosomal protein L9	50S ribosomal protein L9	50S ribosomal protein L9
50S Ribosomal protein L16	50S ribosomal protein L16	50S ribosomal protein L16
50S Ribosomal protein L17	50S ribosomal protein L17	50S ribosomal protein L17
50S Ribosomal protein L19	50S ribosomal protein L19	50S ribosomal protein L19
50S Ribosomal protein L20	50S ribosomal protein L20	50S ribosomal protein L20
50S Ribosomal protein L21		
50S Ribosomal protein L27	50S ribosomal protein L27	50S ribosomal protein L27
50S Ribosomal protein L28	50S ribosomal protein L28	50S ribosomal protein L28

50S Ribosomal protein L31	50S ribosomal protein L31	50S ribosomal protein L31
50S Ribosomal protein L32		
50S Ribosomal protein L33		
50S Ribosomal protein L34		**50S ribosomal protein L34**
50S Ribosomal protein L35		**50S ribosomal protein L35**
50S Ribosomal protein L36		**50S ribosomal protein L36**
30S Ribosomal protein S6		
30S Ribosomal protein S16		
30S Ribosomal protein S18	30S ribosomal protein S18	30S ribosomal protein S18
30S Ribosomal protein S20	30S ribosomal protein S20	30S ribosomal protein S20
18/18	10/18	13/18

Table 22 shows the results of the two database queries alongside the Universal Ribosome and Bacterial Ribosome lists generated by Yutin *et al.*, (2012). Our initial identification and clustering program returned the majority of both lists, but still with plenty of missing entries. Adding in the HMM search improved the outcomes in both lists, but only marginally, returning three previously unidentified proteins to each list. Given that the HMM model was generated from a large database of multiple alignments of the different proteins, we expected it to do slightly better than the standard homology based methods. However, the increase was modest and not nearly enough to fill out the list from Yutin et al. Given the similarities in the two lists we generated, we can be confident that our search algorithms are returning the vast majority of the data we require.

3.3.7. Conservation of a gene's essentiality across different genomes

As a whole, the database contains 191341 genes. When this is filtered by essential genes, there are 19856 that are classified as essential in at least one species. However, the majority of these have homologs in other species, meaning that there are 57,798 occurrences of a gene being present in a genome which is classified essential in at least one genome within the database. On average, each essential gene has a homolog in 16.4 species, and is essential 35.8% of the time it is included in a genome.

3.3.8. Essential gene variation across different genome sizes

To elucidate how the size and complexity of a genome affected the composition of the essential genome of a species, the bacteria were split into four size categories; Minimal, Small, Medium and Large.

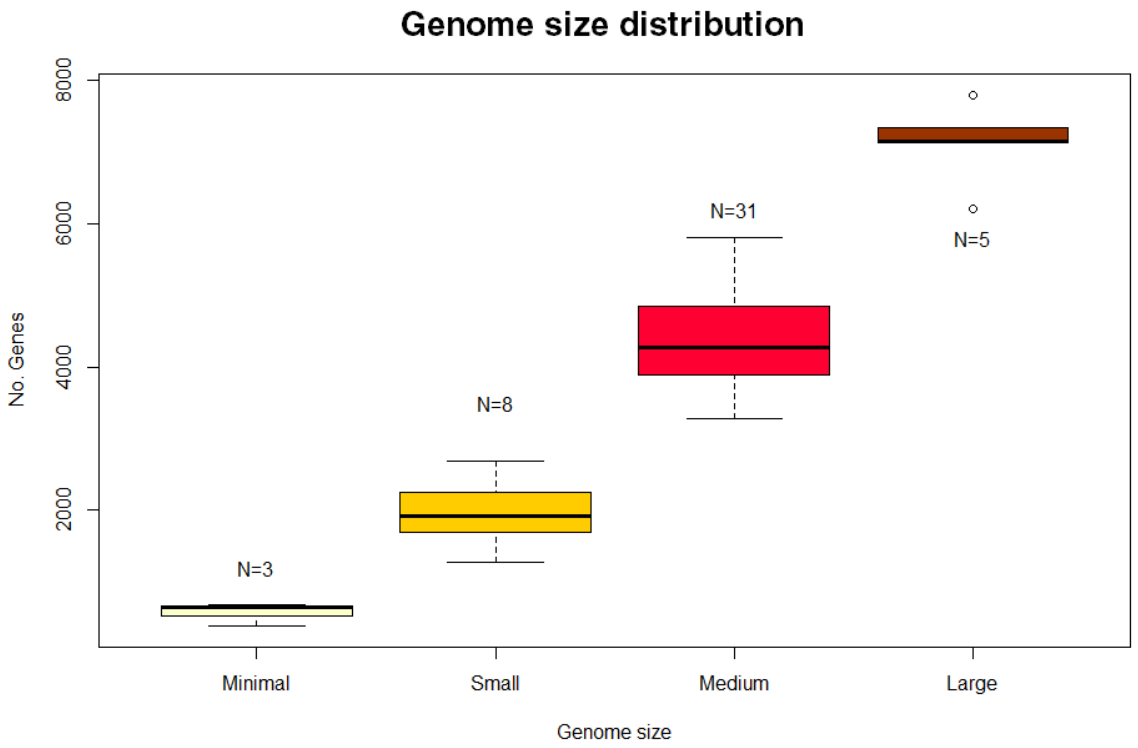


Figure 45: Grouping of analysed species into size categories

Due to the nature of the organisms that were included in the analysis, along with the general trend of bacteria to have a genome averaging at ≈4.5Mb (Land et al., 2015), there is a clear bias towards the Medium group in terms of numbers.

When the four size categories had their genes split into the four Super-COGs, a clearer picture of how the composition of the genome changes over size is shown. Figure 46 shows which percentage of the total number of essential genes each Super-COG consists of for each genome size. There is a clear trend of genes in the Super-COGs of Cellular Processes & Signalling and Metabolism to become more essential as the genome size increases, while genes belonging to Information Storage & Processing and those that are poorly characterised decrease in overall essentiality.

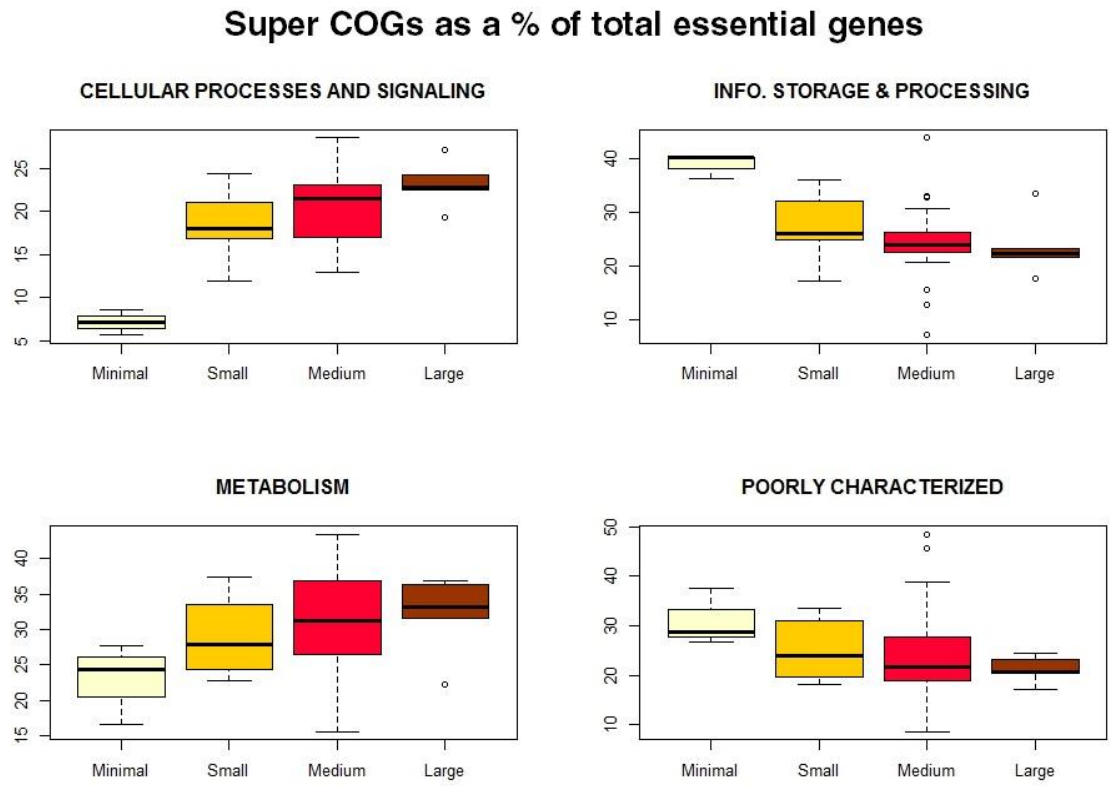


Figure 46: Super COGs as a percentage of total Essential genes across different genome sizes

This trend is mirrored to an extent when the raw number of essential genes is compared for each Super-COG across different genome sizes, as shown in Figure 47. While the general increase in number of essential genes with regard to Cellular Processes & Signalling and Metabolism holds true in both raw number and percentage of each cells essential genome, the raw number of Information Storage & Processing and those that are poorly characterised stay fairly constant instead of decreasing with an increase in genome size.

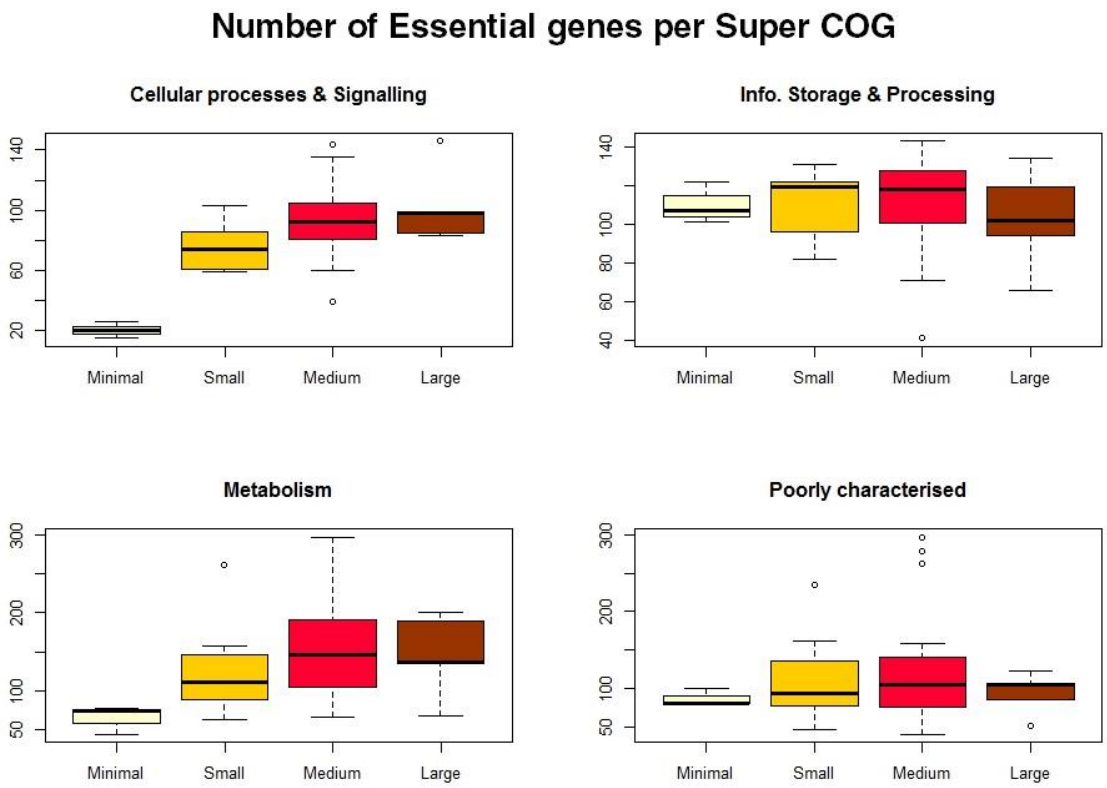


Figure 47: Total number of essential genes in each Super-COG across different genome sizes

We hypothesised that as the complexity of the genome increased, the amount of essential genes related to metabolism would increase. Therefore, the individual COG categories within the Metabolism Super-COG were analysed across the different genome sizes. Figure 48 shows how the levels of each metabolism COG changes with genome size. There is a general trend towards the COGs making up a larger percentage of the overall essential genome as the size increases, with this being most noticeable in the genes relating to energy production and amino acid metabolism. There are however many different trends in the data set depending on the metabolite, such as the comparative lack of essential genes in Minimal genomes related to co-enzyme or lipid metabolism compared to the relatively stable number of each by both raw and percentage counts in the other genome sizes.

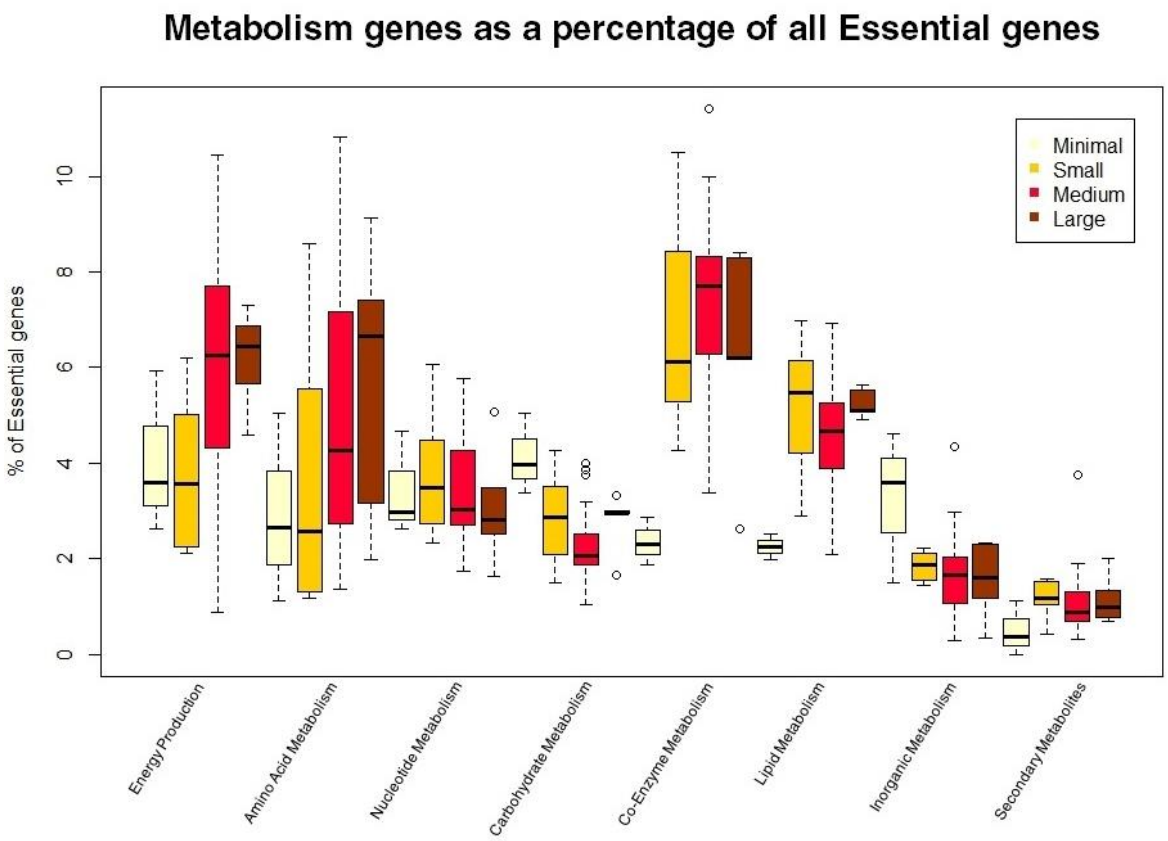


Figure 48: Metabolism COGs as a percentage of essential genes, across genome sizes

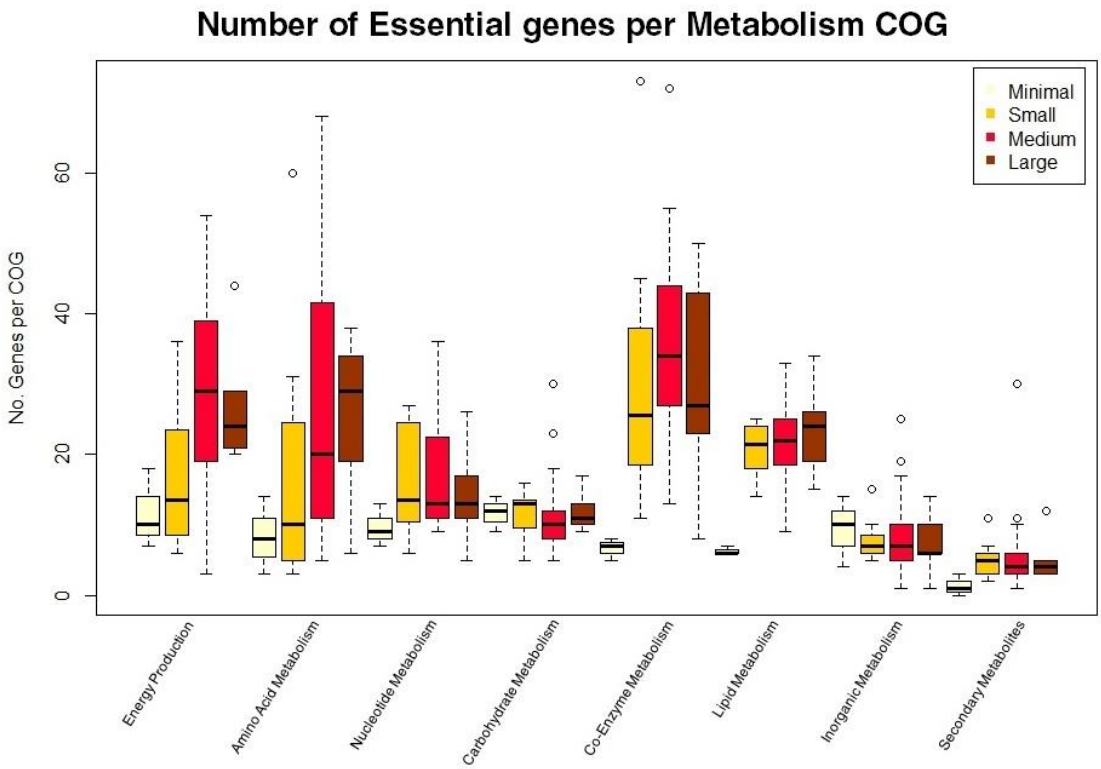


Figure 49: Metabolism COGs as the raw number of essential genes across genomes sizes

With regard to the genes involved in the Cellular Processes & Signalling Super-COG, the positive correlation between number of essential genes in a COG and genome size is even more apparent. Figure 50 shows the percentage each COG contributes to the cells' essential genome while Figure 51 shows the raw numbers for each COG category. The most striking results show that there is a huge reduction in both the number and percentage of genes that are essential for cell division and cell wall biosynthesis.

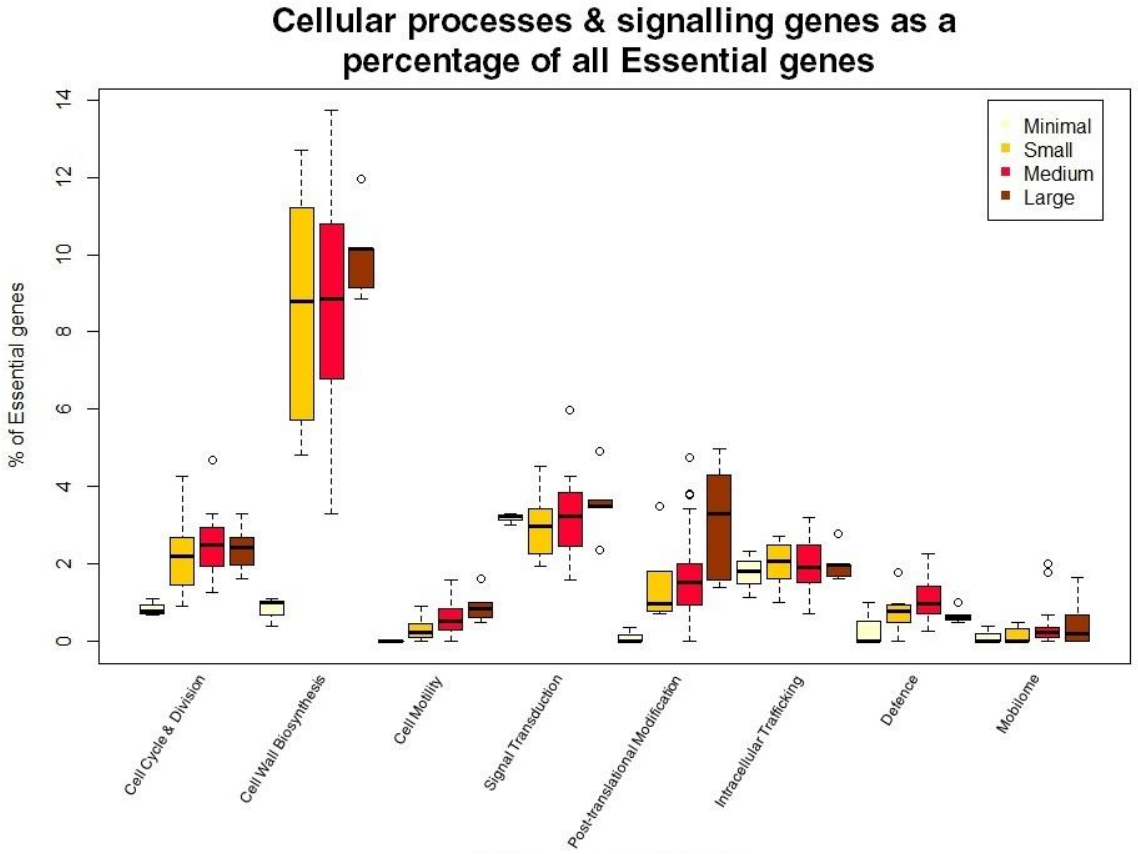


Figure 50: Cellular Processes & Signalling COGs as a percentage of essential genes, across genome sizes

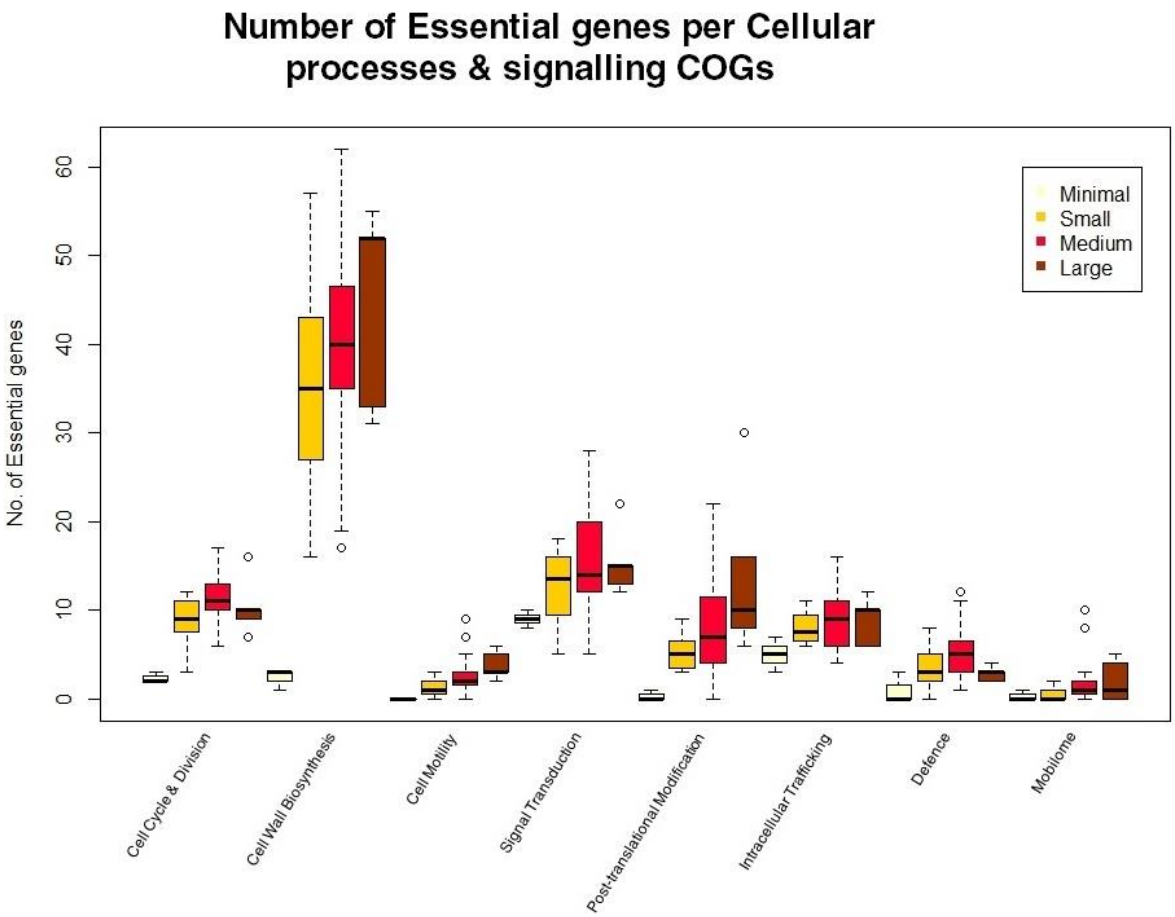


Figure 51: Cellular Processes & Signalling COGs as the raw number of essential genes across genomes sizes

3.3.9. Querying the change in gene essentiality in regard to by genome size

To evaluate if there is a specific trend where genes become more essential as genome-wide complexity increase, every gene in our database that was essential in at least one species was extracted. These genes were then queried against every other species in the database to see if it was present in that species, and if so, if it was essential or not. They were then plotted by COG class on the y axis and genome size on the x axis, with the smallest genomes to the left and largest to the right. This generated a heat-map, with essential genes in red, non-essential gens in green and absent genes in white. The heat-map for each Super-COG can be seen in Figures 52-55.

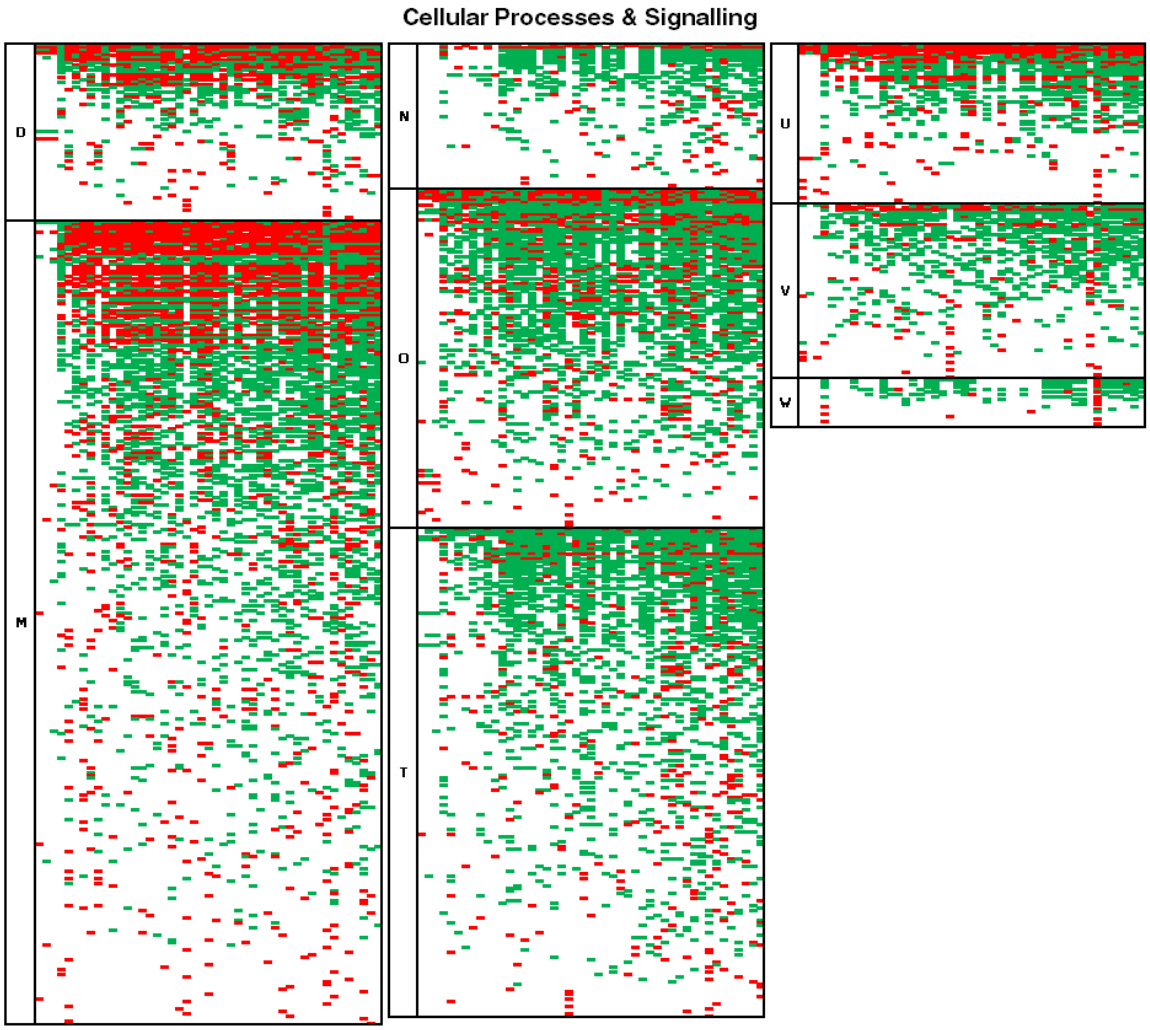


Figure 52: Essential genes in the Cellular processing & signalling Super-COG, arranged by genome size

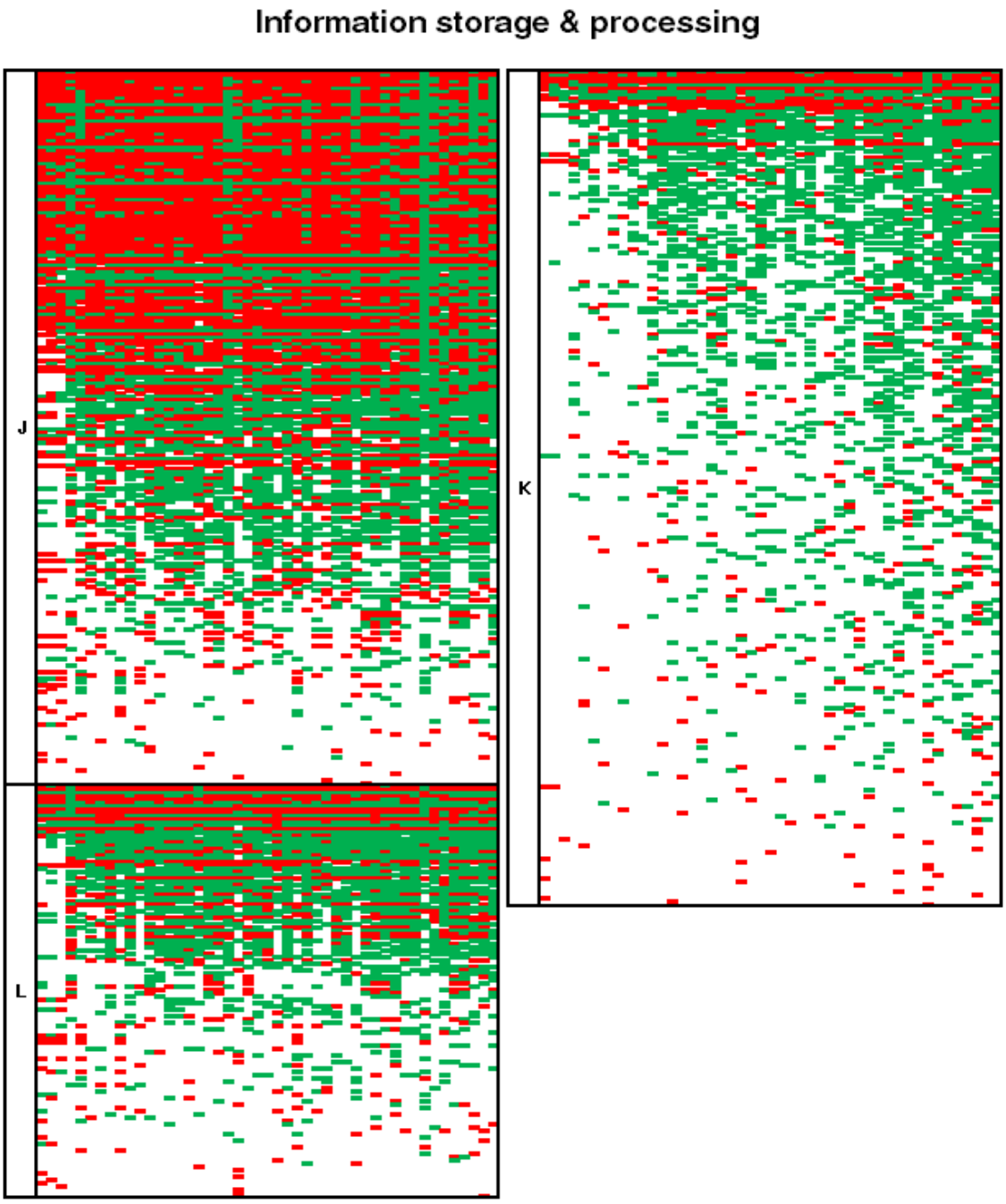


Figure 53: Essential genes in the Information storage & processing Super-COG, arranged by genome size

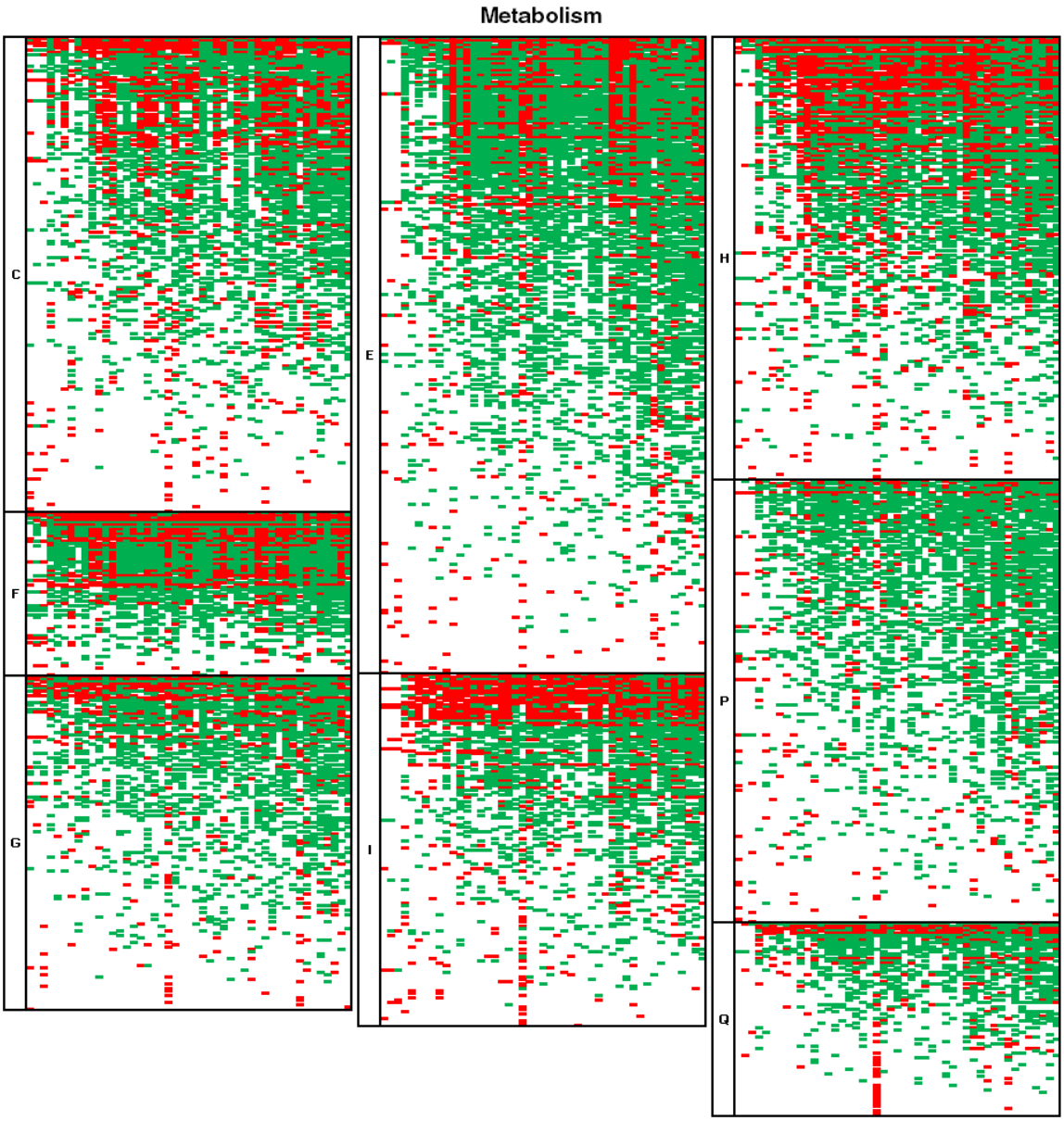


Figure 54: Essential genes in the Metabolism Super-COG, arranged by genome size

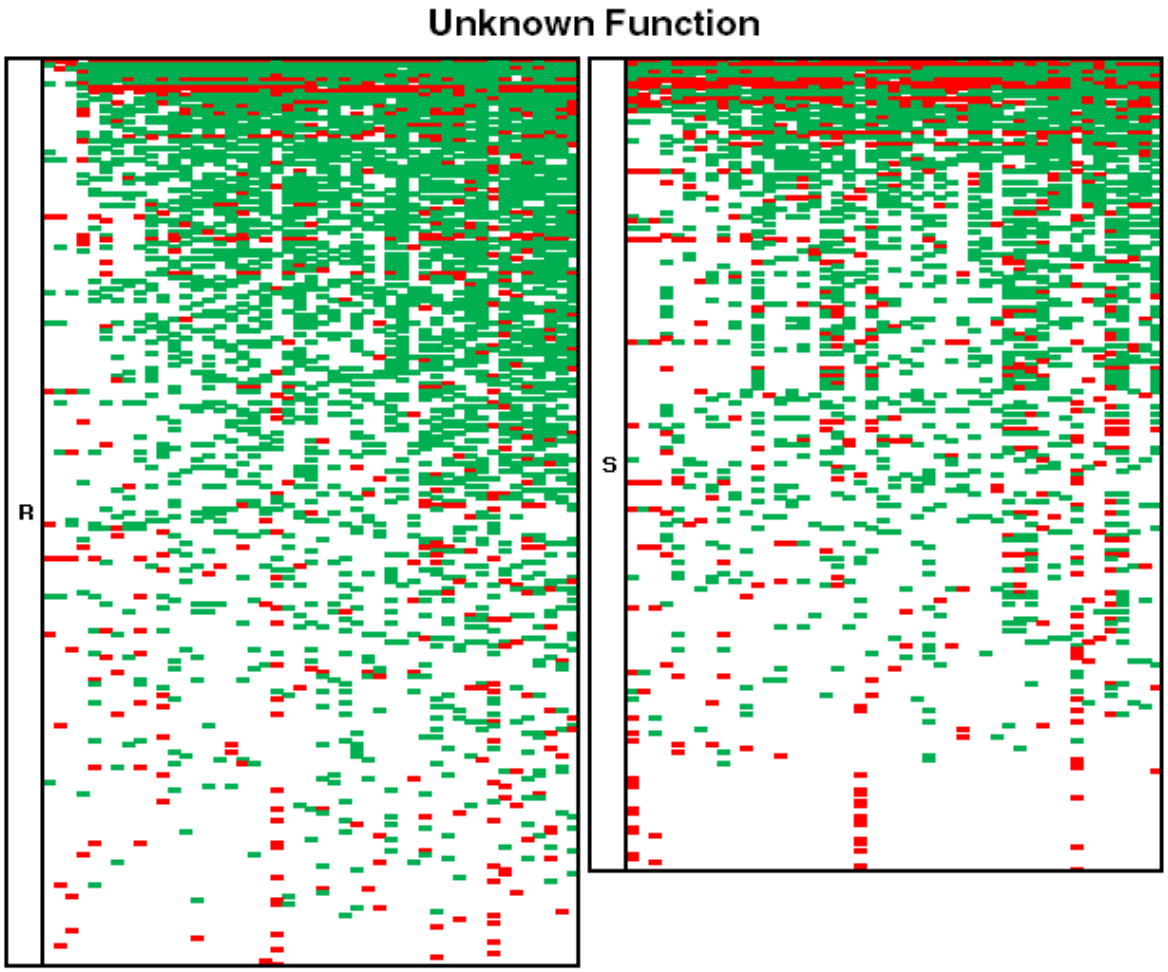


Figure 55: Essential genes in the Unknown function Super-COG, arranged by genome size

Looking at the bacterial population as a whole, there seemed to be little evidence of specific genes becoming more essential as genome size increased. However, there does seem to be a trend of genes that are on the extreme ends of conservation, either the genes that are present in almost all or very few species, are generally more essential. A good example of this is shown in the genes regarding the formation and maintenance of the cell wall and membrane, COG class M. Many of the highly conserved genes are essential, and those that are specific to very few species are also highly essential. The genes were split into five even-sized groups based on the number of homologs each gene had in the data set. Group 1 therefore contained the 20% of genes which were found in the most species, while group 5 contained the 20% of genes found in the least species:

Table 23: Changes in essentiality across number of homologs

	Total COG	Group 1	Group 2	Group 3	Group 4	Group 5
Total genes	327	66	65	65	65	67
Essential genes	1764	1194	276	129	92	105
Non-Essential genes	2585	1150	893	387	152	34
% Essentiality	40.56	50.94	23.61	25.00	37.70	75.54

Looking more broadly at the population, we calculated the percentage essentiality for each gene included in the heat maps as a function of how many homologs they had in other species, and how many of those were essential. Figure 56 shows clearly that the trend of genes being more essential at the polar ends of the homolog inclusion scale holds well across the entire population, not just COG class M, with genes quickly becoming less essential as the number of homologs increases. Then, once a gene is represented in over 30 species, the overall essentiality of that gene begins to rise again.

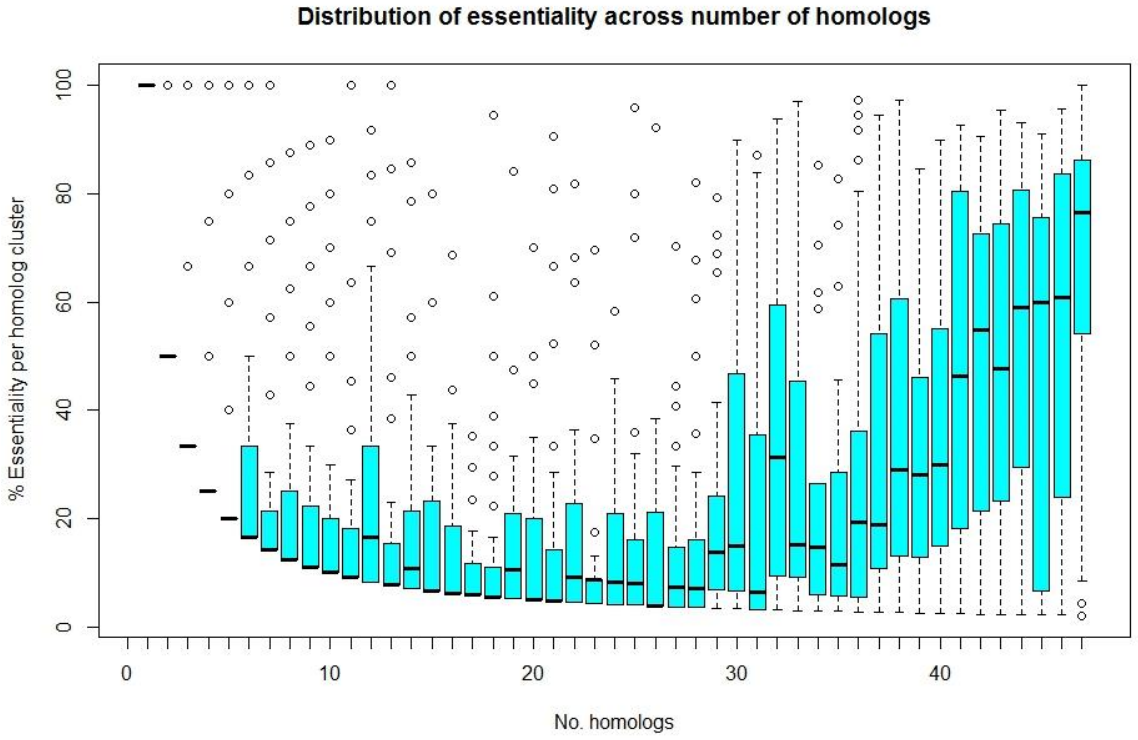


Figure 56: How essentiality changes based on number of homologs a gene possesses

3.4. Discussion

With regard to the hypothesis that the larger a genome becomes, the more essential genes are added, our data showed this is true to an extent, but not as clear as we expected. While there is a small correlation between genome size and number of essential genes (Pearson's Product-Moment correlation of 0.2), it is hardly indicative of a robust relationship between the two factors. However, there is a much clearer relationship between the genome size and the overall percentage of essential genes. Which starts of high in the 40-50% range for minimal genomes then decreases to level off at around 10% once genome size reaches 3000 genes. This is generally in line with our hypothesis that the number of essential genes increase with genome size.

As the complexity of the genome increases, new reaction pathways are added, and thus become integrated into pre-existing circuits. This can make some genes become non-essential, as new genes bring with them redundancies for pre-existing ones, but it can also make pre-existing genes essential. This can be due to the fact the substrate they produce is now vital to the proper functioning of a new pathway, or is somehow involved in its regulation, thus leading to an increase in the essential genes.

A good example of a specific case of this is within the lipid metabolism pathways. The cells that made up the 'minimal cells' population were all mollicutes (*M. pneumoniae, M. agalactiae* and *M. florum*). As such, they contain a vastly reduced ability to synthesise their own lipids (Dybvig and Voelker, 1996). Isoprenoid biosynthesis is one such molecule that mollicutes cannot synthesise. The 1-deoxy-D-xylulose 5-phosphate reductoisomerase gene *dxr* is the first step in the isopentenyl diphosphate biosynthesis pathway. It is present in 41 of the 47 species and is essential in 38, absent in only the three mollicute species, the two streptococci and *L. crescens,* all of which are in the 'minimal' or 'small' genome categories.

Conversely, as the number of genes within a genome increases, more and more pathways are added and thus pathways can be replaced, increasing genetic redundancy. A good example of this can be seen in the genes related to carbohydrate metabolism in Figure 54. Among the highly conserved genes, there is a clear tendency for genes to be essential in the smaller genomes and non-essential in the larger ones. This is also seen in the genes involved in translation (COG category J, Figure 53), though to a much smaller extent.

This can be observed in the metabolism genes shown in Figure 54. In the COGs related specifically to Energy metabolism [C] and Amino Acid Metabolism [E], in the bottom left of each panel there are clusters of essential genes. These genes tend to be essential in the smaller bacteria, and if they are found in larger species they are likely to be non-essential. This indicates one of two possibilities, that either as the genome gets more complex these genes quickly become non-essential or are lost from the genome entirely, or that as genomes reduce in size novel pathways emerge to replace or re-work existing ones.

With regard to the list of 92 genes that were universally conserved, as shown in Table 20, there is a generally high level of essentiality compared to the rest of the database. Essential genes are only classified as essential on average 35.8% of the time they occur. The list of conserved genes, all of which were essential in at least one species, have an average essentiality of 67.7%. This is significantly above the average value of the essential genes, as is expected from the universally conserved genes. Looking at the highly essential genes, it is interesting that there is only one gene that is universally considered essential, the chromosomal replicator protein *dnaA.* This is most likely due to the accumulation of experimental error or noise, indicated by the fact that none of the tRNA synthetases or ribosomal proteins are universally essential.

While 34 of the genes in Table 20 have replicates in some bacteria, in none of the cases where this occurs are there enough can explain all cases of non-essentiality. For example, the 50S ribosomal protein L2 is essential in 96% of species, and has a replicate in 2%. This leaves a further 2% of species where the has no duplicate but is still non-essential, which given the vital role the protein plays in the binding of the tRNAs to both the A and P sites, along with mediating the association of the 50S and 30S subunit (Diedrich et al., 2000), seems unlikely to be the case. Therefore, given to the nature of this analysis as a compilation of multiple methodologies, there must be an assumed error rate inherent to the analysis.

We initially looked at the gene present in Table 21, those gene that both Charlebois & Doolittle (2004) and we found universally conserved as a gold standard for essentiality. While the majority of those genes are highly essential, the list also contains *ychF*, *ksgA*,

and *fusA*, which are essential in only 4, 19 and 38 percent of the species respectively, further indicating that just because a gene is highly conserved, does not mean it is highly essential. The genes *uvrA* & *uvrB* genes are found in every species we included, but are essential in only *M. tuberculosis* and *M. pneumoniae* respectively, despite only *uvrB* having a single species containing a homolog, and *uvrA* none.

As such, while it may be feasible or useful to conclude that there is a strong chance that genes with a percentage essentiality of 85% or higher can be assumed to be universally essential; this cut-off contains a large majority of presumed vital transcription, translation & DNA replication machinery, there are certain to be individual cases that do not conform. For example, the S-adenoylmethionine gene *metX* is found in all species, and is a major component of the methylation systems of both DNA, RNA and proteins (Grillo and Colombatto, 2008). It is essential in 89% of species, but is non-essential in many of the larger bacteria. For example, *A. tumefaciens* contains a secondary adenine methyltransferase known as *CcrM*, which is essential (Kahng and Shapiro, 2001). This, gene acts in a similar way as *metX*, and could explain its loss of essentiality. Due to the inherent chance that there are paralogs or moonlighting functions for these highly conserved genes, it is probably not feasible to ascribe a hard error rate to the sample. Instead, it may be more accurate to classify the functions a highly conserved gene attributes to the cell as essential, regardless of the specific gene's essentiality, or treat the percentage essentiality of a gene as a confidence level that it is truly essential across a disparate population of bacteria.

Looking at the comparison between the dataset generated in this study and that by Charlebois & Doolittle (2004), there is a high degree of overlap. We identified more proteins that are universal, 92 vs 34, however they were using a larger species pool, and as a rule the more species you compare the smaller the universal overlap becomes (Juhas et al., 2011a). The agreement between the two sets is also high, with 24/34 genes universally conserved in both studies, and the vast majority of those that were not universally conserved in our dataset only being lost in only a handful of species.

B. thuringiensis appears to be an outlier from the rest of the species analysed in regard to pattern of essential genes, specifically within the Information Storage & Processing Super-COG. As the eighth largest organism by genome size, its bias towards non-essentiality in Figure 53 stands out clearly in COG category J, and to a lesser extent in categories K and L. This deviation from the standard pattern of essentiality can be explained by the fact there are no ribosomal proteins classified as essential, according to the paper investigating *B. thuringiensis* (Bishop et al., 2014). This lack of ribosomal genes is not mentioned in the original paper, and analysis of the organism at the genome level does not indicate multiple copies of the proteins to provide redundancy. Therefore, either they were excluded deliberately or lost to experimental noise.

The other large takeaway from this analysis is that while there is a degree of conservation between genes across the species, there are only 92 genes represented in every species. The vast majority of these are in COG class J, translation and biosynthesis, which is in line with previous analyses (Charlebois and Doolittle, 2004; Koonin, 2003). The remaining COG classes have very few genes shared between all species. This is most evident in the genes regarding transcription (COG class K) and the genes involved in metabolism. Transcription initiation is a vital cellular processes, yet there are only four genes that are conserved across all species: the Holliday junction resolvase *ruvX*,

transcription termination/anti-termination protein *nusA*, and the two DNA directed RNA polymerase subunits A & B, *rpoA* and *rpoB*. These four proteins are responsible for the process of transcription, and functions relating to its termination, but there are no universal transcription initiation proteins.

The lack of universal transcription factors is interesting, as is the general lack of essentiality within the class. Of the 10829 genes in the database that belong to COG class K, only 712 are essential. While this is not the lowest percentage of essentiality for a COG at 6.6%, compared to the other COGs in its cluster (J has 40.4% essentiality and L has 19.9%), it is a significant change. By contrast, the transcription genes are far more diverse than the genes relating to translation and DNA replication and repair, as transcription in bacteria is a highly diverse process. Many bacteria relying on a vast array of different, niche specific transcription factors, along with other factors such as supercoiling DNA and nucleoid assisted proteins (Browning and Busby, 2016; Güell et al., 2009a; Visweswariah and Busby, 2015), all of which are far more species specific than the fundamental DNA repair & replication and translation machinery.

The bi-modal trend of the essentiality of the genes changing with the number of homologs present fits well with our hypothesis of increasing complexity and gene utilisation. Genes only found in a single organisms are likely there as a response to some form of environmental stress specific to the niche that bacteria inhabits. This trait is especially true in pathogenic bacteria, which tend to evolve similar orphan genes when dealing with similar pathogenic niches. These genes are rarely if ever found in non-pathogenic species, even within the same genus, implying that there is a strong evolutionary pressure stemming from their niche which these genes help alleviate (Entwistle et al., 2019). However, as genes become present in more and more species, this implies that their functionality becomes useful to a wider range of niches. The more widely conserved a gene is, the more likely it is to encode for a protein that assists in a general or housekeeping function instead of a niche specific one, thus the more likely it is to have some level of genetic redundancy (Ghosh and O'Connor, 2017; Mendonça et al., 2011) . Finally, genes that are nearly universally conserved, by definition must play a role in a fundamental cell process. While there will be some level of redundancy in its functionality, the fact that none of the species analysed has replaced the original protein with a redundant or modified one implies that the function it provides is still vital for cellular function at a fundamental level.

This spectra of essentiality and conservation can be seen well in the COG class M, relating to cell wall biogenesis and maintenance. For the genes that are specific to a specific organism due to its individual biology or niche, a good example is the *glfT2* gene in *M. tuberculosis.* This gene is involved in the polymerisation of arabinogalactan, region of the mycolylarabinogalactan-peptidoglycan (mAGP) complex, an essential component of the mycobacteria cell wall (Mikusová et al., 2000). It is essential to the growth of *M. tuberculosis* but is not found in any other species due to *M. tuberculosis's* unique cell wall configuration compared to the others.

An example of moderately conserved genes that are generally non-essential would be the glycosyltransferase family 2 protein, which are found in 21 species and are only essential in 6. These proteins are involved in cell wall biogenesis and maintenance, specifically by the production of multiple polysaccharides molecules via the transfer of nucleotide-diphosphate sugars. While their function is essential to cell wall maintenance, there are

many families of glycosyltransferases, and thus a level of redundancy is present in most genomes (Campbell et al., 1998), causing it to be essential in only a few species despite playing an important housekeeping role in the cell.

Similarly, an example of genes that are generally useful to cell survival but rarely essential is the *SMc02856* gene from *Sinorhizobium meliloti.* This gene is a penicillin binding protein, and is found in 24 out of the 47 species. It is only essential in one however, *Pseudomonas stutzeri*. This specific strain (*P. stutzeri* RCH2) was isolated from contaminated ground water (Chakraborty et al., 2017), thus containing antibiotic resistance genes with essential characteristics makes sense due to its competitive environment.

Finally, the *murF* gene (UDP-N-acetylemuramoyl-tripeptide D-alanyl-D-alanine ligase) is found in every non-mollicute, and is essential in 39 of the 44 species studied. It is vital to the formation of peptidoglycan, attaching the dipeptide to the UDP-*N*-acetylmuramic acid (MurNAc)-tripeptide to complete the synthesis of the molecule (Sobral et al., 2006). Due to its fundamental importance to cell wall formation, it is understandable why the gene is universally conserved among bacteria containing a cell wall, and why it is similarly essential.

There are however, many confounding factors in this study that need to be addressed. The first, and probably largest, confounding factor this study faces is the diversity of the species being analysed. The phyla belonging to the Proteobacteria comprise 33 of 47 species in the analysis, and thus bias towards this phylas' genetic predispositions is inevitable. This is not just an issue specific to this study, but found across microbiology in general. A review of the GenBank entries regarding sequenced bacterial genomes in 2015 found that just six bacterial phyla comprised 95% of all sequenced bacterial genomes, and 46% of the total sequences were from Proteobacteria (Land et al., 2015). Of the remaining phyla, in order of number of genomes sequenced they were the Firmicutes (31%), Actinobacteria (13%), Bacteroidetes (3%), Spirochaetes (2%), Cyanobacteri (1%) and all other phyla (5%). Therefore, while our ratios are slightly different, this study did analyse data that is generally representative of the overall state of sequenced bacteria.

Why the Proteobacteria are so enriched in experimental data is not fully known, however there are a few important factors worth noting. First, all known sub-groups of Proteobacteria (α, β, γ, δ and ε) are amenable to transformation, whereas there are phyla with no known transformable species, such as the Acidobacteria, Spirochaetes, Chlamydiae, Chloroflexi and Aquificae (Mell and Redfield, 2014). Almost all Proteobacteria are Gram negative, and the lack of thick peptidoglycan wall coupled with well documented transport proteins for the uptake of DNA in these species make them well suited to importing DNA (Johnsborg et al., 2007; Mell and Redfield, 2014). This ability to be studied in greater detail under laboratory settings is probably one of the key reasons why they command such a high percentage of the species we have taken the time to sequence and investigate.

Because of the bias towards Proteobacteria, there is a second implicit bias towards Gram-negative bacteria. This cannot be wholly attributed to the Proteobacteria however, as there are only four Gram-positive bacteria within the study, the Firmicutes. This could be due to the fact that Gram-negative bacteria are intrinsically more receptive to transformation

due to their lack of peptidoglycan cell wall, thus studies on them are easier to perform. The thick peptidoglycan layer acts as a natural barrier for transformation and makes methods such as electroporation less effective, though polythethylene glycol has been shown to be effective (Rattanachaikunsopon and Phumkhachorn, 2009). For the sake of this discussion point, I am including the Mollicutes and *M. tuberculosis* within the sphere of Gram-negative bacteria. While the Mollicutes did evolve from Ternicute (thus gram positive) ancestors (Trachtenberg, 2005) and *M. tuberculosis* is classified as neither Gram-positive or Gram-negative but instead as an acid-fast bacteria (Koch and Mizrahi, 2018), neither contain the requisite peptidoglycan cell wall, and thus do not present the same barrier to transformation that true Gram-positive bacteria exhibit. Thus, in the context of amenability to transformation as a function of having a peptidoglycan cell wall, I feel it fair to group them with the 'true' Gram-negative bacteria.

Another source of bias, as eluded to above, is that running essentiality screens on bacteria is not a trivial task, and requires cultivation, efficient transformation and next-generation sequencing of each species. The data generated from such experiments is also usually designed to gain a deeper understanding of a specific bacteria, such as elucidation of novel drug targets (Moule et al., 2014), understanding and identifying pathogenic determinants (Goodman et al., 2009; Hasegawa et al., 2017; Lin et al., 2014; N. Wang et al., 2014), ascribing gene functions (Deutschbauer et al., 2011; Price et al., 2018) or identifying which genes are necessary when grown in *In vivo* vs *in vitro* environments (Bachman et al., 2015; Bishop et al., 2014; Turner et al., 2015). As such, due to the preliminary need to study the these bacteria in the first place, they tend to fall within the human 'sphere of interest'. Sixteen of the species (34%) included in this study are direct human pathogens or commensals with pathogenic potential, and another ten (21%) are important plant pathogens or nitrogen-fixing bacteria, both involved heavily in human agriculture. They may not therefore be representative of the Bacterial Domain as a whole, but just those bacteria that are either useful or dangerous to us as a species, and have thus merited our attention.

This is certainly a point of bias within the study, but not one that I believe invalidates the results, just the framing of them. While the trends in specific COG categories, such as highly conserved genes related to carbohydrate metabolism becoming less essential as genome complexity increases, may not hold true across the entirety of bacteria, they are representative for the species here. As these species cover phyla and classes of bacteria that fall within the 'human sphere of interest', it is likely that the information here is most useful and most relevant to the bacteria that humans as a species are most interested in.

The standardisation of the data accumulated was one of the most important aspects of setting up this study, and another source of bias within the analysis. To try and keep everything as standardised as possible, we attempted to use the same genome annotation format throughout and cluster genes with homologs. This could allow annotated species to infer and double check annotations from less well defined species. In this spirit, we deliberately decided to focus on protein coding genes only, excluding RNAs from the analysis. This was due to the desire to focus on looking for specific changes in the protein coding capacity of cells as genome complexity changed, but also as not all studies included essential RNAs in their essentiality maps. For example, *Brevundimonas subvibrioides* has two non-coding RNAs established as essential (Curtis and Brun, 2014), as does *Rubrivivax gelatinosus* (Curtis, 2016). However, most other species have no records of essential RNAs. This could be due to specific papers looking only for essential

protein coding genes themselves, or essential RNAs are are not as widespread. Given that when looked for, large number of essential non-coding RNAs can be found (Lluch-Senar et al., 2015b), it is likely they were excluded from initial results in many of the papers analysed. Therefore, to standardise the factors being compared, we looked at only protein coding genes.

Within the issue of standardisation, we encountered a wide variety in the quality and completeness of the annotations used. While we tried to standardise all of the genes from each organisms in our database to be linked to a RefSeq ID, this was not always possible. As such, GenBank IDs and ProteoIDs were also used to ensure that there were no gaps in our records for each species. On top of this, matching the IDs given for the essential genes to our database for the species often non-trivial. This was due in part to the variety of reporting methodologies used by each author, and in part to a lack of synchronisation between the genome annotations and the list of essential genes provided.

For each list of essential genes generated, an essentiality indicator was established. For example, the list of essential genes provided for *Synechoccus elongatus* contained the RefSeq ID for each essential gene (Watabe et al., 2014), so matching this against our database was easy to do automatically, and allowed us to easily annotate which genes in the database relating to *S. elongatus* were essential. Most lists gave a locus ID that matched to the genome annotation, such as those given for *Streptococcus pyogenese* (Le Breton et al., 2015), and some such provided genomic loci for each gene, along with other identifiers, such as *Herbaspirillum seropedicae* (Rosconi et al., 2016). These in general were fairly simple to match, however there were ambiguous cases. For those lists that contained other identifying information, such as genetic loci, resolving these miss-matches was much easier. Some papers provided only common genes names, like the list provided for *Bacillus subtilis* (Kobayashi et al., 2003). This was most problematic, as many genes have multiple common names, such as the Ribosomal RNA small subunit methyltransferase A being referred to as either *rsmA* or *ksgA* interchangeably (Kyuma et al., 2015).

Due to the large amount of variation in the input data, this meant that many essential genes had to be identified manually, as they did not match directly to the database. In such cases, we generally ran a pBLAST to identify if there were any obvious homologs in our dataset already if the sequence of the protein was known. If this was not the case, descriptions of the proteins function were used to see if it could be mapped to a protein in that species from the database.

These efforts were hampered further by a mix of miss-annotation of essential genes and changes to the genome annotation files after the papers had been published. For example, in the list of essential genes provided for *Bacillus thuringiensis*, there are five typos in the gene names (Bishop et al., 2014). The essential Asparginase in this organism is labelled with the incomplete locus tag "BMB171_C1". Looking into the genome, *B. thuringiensis* contains two asparginases, BMB171_C2086 and BMB171_C1329. As no other information on the gene is provided, it was annotated as BMB171_C1329 on the basis of the partial locus tag. Other errors, such as underscores (_) being replaced with hyphens (-) were less ambiguous to correct, but still required manual curation.

Other times, the genome annotation had been updated or modified since the paper relying on it had been published. A good example of this can be found in the annotations of *M.*

tuberculosis, where the genes Rv3021c and Rv1784 are designated as essential (Zhang et al., 2012). However, in the genome annotation we downloaded of *M. tuberculosis*, Rv3021c was annotated as a pseudogene and Rv1784 no longer existed, as it had been determined that it was actually a part of Rv1783, not a unique protein itself. As such, both annotations were discarded, as we discarded all pseudogenes from the analysis (on the basis that they are not genes) and Rv1783 was already annotated as essential.

Due to a combination of the aforementioned errors, almost all species we analysed had some level of miss-identification, and thus manual curation of essential genes was required. This in turn may have introduced our own errors, as when working with sometimes limited information on the specifics of certain genes, we may have added or discarded genes incorrectly. While our automatic annotation programs did map over 99% of the essential genes accurately, it is still worth mentioning that errors from manual curation may be present.

Accounting for experimental noise is a key issue with this analysis. Analysing transposon data for essentiality is inherently prone to many confounding factors, and eventually based on statistical probabilities instead of empiric observation (Deng et al., 2013; Zomer et al., 2012). As a result of this, by combining multiple different results, each using large variations in methodology, such as choice of transposon, transformation method, growth condition(s) and analysis pipeline, the noise will propagate throughout this study. This makes discerning the level of trust we can place in the specific data points difficult. The overall agreement between our universal genes and the set from Charlebois & Doolittle (2004) indicates that our results are not fully spurious, and the filtering steps we took initially to restrict the data we analysed to broadly similar methodologies probably contributed a great deal to this.

This brings up the issue of how to differentiate the actual trends from the experimental noise. The case of *B. thuringiensis* shows that there is a level of noise generated by the described issues in reporting and annotation of essential genes. However, due to the huge diversity in bacterial niches and metabolisms, there will also be legitimate reasons why genes that are essential in the vast majority of species are not essential in others. For example, as mentioned in chapter 1.7.2.2, extreme endosmbionts often rely on their host's DNA replication machinery (Chong and Moran, 2018; Feldhaar and Gross, 2009). As a result, candidatus *Carsonella ruddii* would buck the trend in our data, as it does not contain a *dnaA* gene, let alone it be essential (Tamames et al., 2007).

Finally, we encountered an issue regarding gene paralogs. All of our clustering techniques were based on protein identity, grouping genes with a similar protein structure with each other and assuming homology of structure equals homology of function. However, this ignores the fact that there is often more than one gene responsible for the same phenotype in different bacteria [ref]. As a standard, we used the genes in *M. pneumoniae* as the basis for the initial clustering, as they are well described and annotated (Wodke et al., 2015). However, just because a gene performs specific function in *M. pneumoniae*, does not mean all other bacteria that contain that function will contain that specific gene. While the *M. pneumoniae* genes were used as a base, we clustered the entire database by homology, so all gene groups that are homologs are successfully clustered. The issue comes when querying the database about information for a specific gene. A clear example of this was found in the genes coding for the proline tRNA synthetase. According to Charlebois & Doolittle (2004) and Koonin *et al.*, (2003), we should find the proline tRNA

synthetase in every species. However, the copy of the gene found in *M. pneumoniae* only had homologs in eight other species. We therefore had to manually query every species to see if it contained an annotated proline tRNA synthetase, and we found that there were three distinct paralogs of the gene that were generally split along phylogenetic lines. Having different classes of tRNAs has been well documented (Eriani et al., 1990), and serves as a reminder that evolution has allowed for multiple different paths to the same outcome.

It should therefore be noted that just because a specific gene is not represented in a genome, it does not mean that the function that gene provides is missing. Similarly to the issue faced by the multiple proS genes, Charlebois & Doolittle, (2004), stated that there should be the transcription antiterminination gene *nusG* present in every species. Using the *M. pneumoniae* copy of *nusG* to search the database, we found no other species that contained the gene. Further interrogation of the databse revealed that every other species did contain the *nusG* gene, just a paralog of the copy found in *M. pneumoniae*. Indeed, according to a BLASTp of the sequence of the *M. pneumoniae nusG* gene against the NCBI database, it is only shared with *M. genitalium*, its closest relative, indicating that this is a novel and fairly recent evolutionary acquisition. This issue of search results being biased due to the expectation of preserved homologs gives further credence to the secondary search strategy employed by Charlebois & Doolittle, where they ascribed a function and gene name to each homolog cluster they collated. Searching databases via protein sequence or a RefSeq ID alone will inevitably bias the results towards that specific homolog, ignoring important paralogs. While there is a huge variance in the gene name annotations given (see earlier in this discussion), the ability to search via function instead of a specific gene would be a huge help in identifying common features among multiple species.

Therefore, it is probable that the potentially unclassifiable variation in bacteria (Locey and Lennon, 2016) means that there are no universal rules in regard to any specific essential genes in the pan-bacterial genome. The abilities of genes to moonlight to other functions (Jeffery, 2018; Kainulainen and Korhonen, 2014; G. Wang et al., 2014), and especially in pathogenic bacteria (Henderson, 2014; Henderson and Martin, 2013, 2011) mean the levels of redundancy allowed within the bacterial genome allows them to specialise their functionality to perfectly suit their niche (Ghosh and O'Connor, 2017).

3.5. Conclusion

In summary, we presented here the first large-scale analysis of the effects of genome size and complexity on the composition of a bacterium's essential genome. We outlined a methodology for compiling and annotating data from multiple sources with the focus of standardising them as much as possible, and with the hope that the database can be searched, used and expanded on in the future by the field at large. We validated our methodology by producing a list of universally conserved genes that were broadly in line with previous analyses, and showed that while we could not capture every protein that shared homology with a target, more advanced search strategies such as Hidden Markoff Model-based searches only showed a moderate gain in success.

In terms of the conservation of essential genes, while we did find the genes that were conserved among all species were more essential than average, only one gene (*dnaA*) was universally essential, and on average these conserved genes were only essential in 68% of the species. As with other analyses, the gens in this list consisted mainly of translation related functions, with genes ribosomal proteins making the majority. There were however larger number of metabolism genes conserved, along with genes related to basic cellular processes.

Our analysis of the relationship between genome size and essential gene content revealed a modest positive relationship between the two, though as with all the analysis here decoupling the noise generated through inherently noisy protocols and then the merging of that data means we can only state in terms of generalisations. Similarly, while there is evidence to support the hypothesis that as genomes become more complex, both the raw number of essential genes and the percentage of essential genes involved in Super COGs Metabolism and Cellular Processes increases, there are few identifiable instances of this trend bearing out at the individual gene level.

We also found a strong trend in enrichment of essential genes at the polar ends of the conservation spectrum. Genes that were essential in at least one species which found in almost none or almost all species were more likely to be essential than those with a more median number of homologs. This is most likely related to niche specific functionality vs fundamental utility to the cell, with genes found in a moderate number of species clearly imparting some benefit to the cell, but also likely to have redundancies or paralogs, whereas genes found in all species impart such a useful phenotype to the cell that they are retained universally.

While our dataset does contain an inherent bias towards the essentialities and gene complements of those bacteria within the ‘human sphere of interest’, and more specifically the proteobacteria, this is not necessarily a detriment. By collating which genes appear to be essential in so many species, we can begin to filter these by more and more specific factors such as niche or core metabolism. In doing so, this database may be useful in the rational design of new synthetic biology projects, allowing for more insights into which genetic combinations could be best suited for inclusion to a cell designed to thrive in a specific niche, or which circuits would be detrimental to remove.

CHAPTER 4: DISCUSSION AND CONCLUDING REMARKS

As our knowledge of biological systems deepen, our attempts to both forward and reverse engineer these systems grows ever more powerful. The combination of ever advancing specificity in wet lab protocols, such as Gibson ligation (Gibson et al., 2009) and CRISPR editing (Cong et al., 2013), with ever increasing refinement and power in bioinformatics analysis and "–omics" techniques have afforded us the ability to edit genomes with base pair specific precision and intent. However, while systems biology approaches have increased our understanding of living systems, we are still far from acheiving one of the main aims of systems biology, reliable rational engineering of those systems to develop novel applications. One of the main goals in this discipline has been the obtaining of a minimal chassis, either rationally designed or not, but this is still unmet. A minimal chassis cell would have countless applications, and there has been a lot of attention on developing a bacteria that can act as a smart pill or theraputic agent (Braff et al., 2016; Claesen and Fischbach, 2015; Haellman and Fussenegger, 2016; Piñero-Lambea et al., 2015; Weber and Fussenegger, 2011). However, there are many other permutations a minimal chassis could have, and as discussed previously, the genes a bacteria requres are entirly enviroment specific (Joyce et al., 2006). Therefore, tools to create multiple versions of minimal chassis cells, either from top down or bottom up approaches, will be highly useful (Ausländer et al., 2017; Danchin, 2012; Moe-Behrens et al., 2013; Vickers et al., 2010; Xavier et al., 2014b).

The closest attempt so far, the JCVI's Syn-3.0, is a landmark achievement in the field, yet even this organism retains 79 genes of completely unknown function, of which 24 are non-essential (Hutchison et al., 2016). This challenge in annotating functions to all genes, and deciphering the role of organisms specific genes may be one of the final steps we need to take as a field, with various attempts ongoing (Price et al., 2018; Yang and Tsui, 2018).

This thesis aims to address two aspects that are important in the concept of a minimal chassis; epistasis and the importance of different networks through comparative analysis of essentiality, and the concept of the minimal genome through the development of of novel genetic tools that can be used to randomly deplete large genomic regions. These two chapters represent top down and bottom up approaches in the characterisation of these concepts repectivly.

In Chapter 2, we outlined a novel mechanism for reducing the size and composition of a bacterial genome in a random, non-biased manner, and have shown that we can deplete large stretches of DNA up to 25Kb in a single step. While not a full 'genome minimisation' technique yet, we believe it can be useful in tandem with other top down genome deletion strategies to achieve a minimal chassis. Other studies using the Cre loxP system to delete specific regions of DNA have shown that it is capable of much greater deletion sizes, with a study in delta-proteobacteria *Myxococcus xanthus* reporting a deletion of 466Kb (Yang et al., 2018). This bacteria has a genome size of 9.14Mb however, and was the largest known bacterial species in regard to genome size until fellow myxobacterim *S. cellulosum* was sequenced (Han et al., 2013; Yang et al., 2018), thus not endearing *M. Xanthus* as an attractive minimisation candidate. However, the knowledge that the Cre lox system has the ability to delete any potentially non-essential region in a genome, regardless of size, further boosts the utility of this methodology.

We have applied the technique without isolating or growing cells individually to be able to obtain the maximum number of mutants and variability between modified genomes. However, cells in the population can compete against each other, or even cooperate in growth by complementation or supplying factors or nutrients that are depleted in mutants located nearby. This can make obtaining single clones with streamlined genomes difficult, but by deliberately modifying the selection conditions, it can also become a method to create streamlined cells with specific traits. While the phenotypes cannot be created *de novo*, by growing cells at different temperatures, salinities, metabolites etc., this technique can be used to select for those cells which are best adapted for the environment. If the technique is amenable to multiple rounds of deletion, then both fine-tuning of a desired phenotype along with large scale genome minimisation is possible.

The technique also allows for identification of epigenetic networks within bacterial genomes. By observing the temporal progression of deletions alongside the genes that are removed, it may be possible to identify progenitor deletions that predispose a cell to a specific essentiality profile, or prohibit the deletion of other presumably non-essential regions. If a specific deletion occurs in a cell line, but is never observed in other cell lines that contain a specific previous deletion, we can infer an epistatic interaction between the two regions. If enough rounds of deletions are possible, it can also be used to see if there are any 'paths of least resistance' genome minimisation can occur in. Do, for example, genomes tend towards a similar deletion phenotype over time, with the same regions being removed in similar orders across the population. Alternatively, are their multiple different minimised genomes that can be obtained, potentiated by specific driver deletions that prevent certain genes from being removed and forcing the cell down a specific deletion pathway? If these genetic predispositions exist, then their identification will allow us to better rationally design networks which work robustly.

The technique also provided evidence of a useful counter-selective agent when working with *M. pneumoniae*¸ that of the Cre itself. The lethal effect of Cre recombinase was unknown in *M. pneumoniae*, and we demonstrated that the Cre recombinase binding to a single active lox site causes a lethal phenotype to the cell, at a similar level that that caused by the introduction of a restriction mega-nuclease. This lethality in turn allowed our causative agent of genome deletions, the Cre recombinase, to also be a self-selective agent. Any cell line that did not undergo a deletion that resulted in an inactive lox72 site was killed, either by the antibiotic after not being transformed, or by the action of the Cre on the loxP site that was formed. Due to the presence of the two initial lox sites as LE_lox or RE_lox sites, any combination between the two would cause the formation of a loxP and lox72 (Pinkney et al., 2012; Van Duyne, 2001), and thus only the cell line with a lox72 site in the genome, and the loxP in the excised DNA will survive.

The emphasis of chapter two was certainly on the iterations taken to build the final protocol, and the points of failure are also noteworthy in their own regard. The lack of consistency in the annotation of the lox sites in the general scientific literature is worrying, and as was the case frequently in chapter three, incorrect annotations take a lot of time and effort to manually curate and correct. Re-designing the lox sites solved this issue, and hopefully the future use of annotations such as left element and right element mutants can alleviate this in the future. The effect of the transitory expression of the Cre recombinase when encoded on a suicide vector was also a factor we had not anticipated, and as such forced a re-design of the system to utilising another transposon to express the Cre. While this did add an extra step to the protocol, which inevitably increases bias brought about

by growth speeds, it also brings the benefit of having an extra deletion step targeted towards single genes or smaller regions that have a higher chance of being missed via the main deletion steps.

While chapter two focussed on a novel top-down engineering system, chapter three looked at the issue of genome minimisation from the bottom up perspective. As important as being able to delete non-essential regions is, we also need a comprehensive understanding of which regions in a genome are essential and why. This chapter introduced the first large-scale analysis of gene essentiality across multiple bacterial species, and the database that was created as a result of it can hopefully be used as a tool to compare which genes are essential in one bacteria to a wide range of others.

One of the key messages from this chapter was the focus on the function, instead of the gene. By assigning COG categories to as many genes as possible, we can see how the composition of the essential genome changes with complexity in a more general view. As shown in Figure 46, we can see clear trends in how the composition of the essential genome changes with added complexity, and the heat maps generated can show us which functions are the causes of this shift. While at the global level there are few clear trends, there are certainly some indicators of functions becoming more essential as complexity increases. At a population level, we can see that the Super COG of Cellular processes and signalling becomes a larger percentage of the essential genome as complexity increases. In Figure 52, we can see subsets of genes in the cell motility [N] COG and signal transduction [T] COG that are more essential at the larger genome sizes, specifically in those genes that are conserved among half or fewer of the species.
One of the most powerful tools we have to compare genomes is searching for orthologous genes that are shared between species, with the supposition that if two gene sequences share a generally similar amino acid sequence, then they will encode for a protein that performs a similar task (Pearson, 2013; Tatusov et al., 1997). While this assumption has worked well for protein identification, indeed it is the basis of the COG system used to assign functions in this study (Tatusov et al., 2000, 1997), it cannot account as well for unrelated proteins that share a function.

This was demonstrated well by the proline tRNA synthetases (*proS*) genes. While the three *proS* genes we identified did indeed share two main homology regions, they were dissimilar enough so that they did not cluster with each other using our standard algorithm. They were only identified via a manual search of the database, using gene names linked to RefSeq IDs. There will inevitably be further cases similar to this within the database, namely for two reasons. The main reason behind this however may be an inevitable facet of biology, evolution and adaptation. The three *proS* genes were distributed logically across the phylogenetic space, with one gene found almost exclusively in the alpha-proteobacteria, another in only the Mollicutes, Firmicutes and Bacteroidetes, and a third with representation from all clades excluding the Mollicutes and Bacteroidetes. This illustrates that even for highly conserved and fundamental functions, there is at least one *proS* gene in every cell, there can be significant variation of gene utilisation across the genome. With the hypothetical minimal gene complement required to sustain the most basic forms of life estimated to be around the 300-400 gene mark (Gil et al., 2004; Huang et al., 2013; Juhas et al., 2011a; Koonin, 2003), our conserved list consists of just 92.

This indicates that many of the functions of these genes must be encoded for by genes that are different enough to be disregarded by traditional searches of amino acid sequence looking for homologs. With the phenomena of moonlighting proteins becoming more and more well known (Henderson, 2014; G. Wang et al., 2014), the complexity of gene functions and interactions is only increasing. By analysing how the essentiality of a gene changes with the complexity of the genome it is located in, we can start to untangle where some of these divergent genes begin to occur. We can investigate if they are a function of a pre-existing gene that gains a new function as it interacts with more pathways within the genome or metabolome, or if new genes start appearing that have no homologs in simpler bacteria.

With regards to assisting in bottom up genome engineering, this tool can help by highlighting how certain genes present across a population of bacteria, and in which cases they are essential or not. As with the *proS* case, it can help identify why certain genes are present or not, and may be able to guide further engineering efforts by highlighting if a specific gene of interest appears to become more or less essential as genomes increase or decrease in complexity. By matching the correct gene with the correct functionality for the genome size, we can potentially remove extraneous or inappropriate genes in our circuits. This could be even more useful when combined with more stringent filtering, such as grouping the bacteria by niche or desired trait, and looking at which genes rise or fall in essentiality as a result.

Similarly, it can act as a tool to further include genes of unknown function into the rational design toolbox, and help identify the niches they are contribute to survival in the most. While it may not be able to accurately predict the function of the gene, identifying genes of unknown function that are shared between bacteria of a similar complexity or niche could improve the fitness of cells engineered to share that purpose. This could therefore indicate either regions to be included in a bottom up engineering effort, or regions to avoid deleting via top down methods.

Taken as a whole, we believe this thesis provides two useful components that can be added to the systems biology toolbox. A methodology to allow for bias-free deletion of both large and small genomic regions with a high degree of variation, creating a genomic streamlining tool, and the first large-scale investigation of what essential genes and functions are shared across the bacterial domain, and how these essentialities change with complexity. These tools can complement both top down and bottom up engineering attempts, and with further iteration can hopefully be useful additions to a wide range of synthetic biology approaches.

SUPPLEMENTARY MATERIAL A – CUSTOM CIRULARISATION SEQUENCING PROTOCOL

NEBNext Ultra II Sample Prep for TN-seq_Custom Adaptor Ligation
NEBNext Ultra II DNA Library Prep kit for Illumina (Ref.: # E7645)
NEBNext Index sequences correspond to Illumina Index sequences.

NOTE: NEBNext Singleplex and Multiplex Oligos for Illumina (NEB #E7350, #E7335 and #E7500) have new concentrations (10µM).

Fragmentation (Covaris)

1. Prepare **up to 250 ng of samples** in **50 ul** of water.
2. Cool down the water bath and degas Covaris (takes about 1h).
3. Shear DNA with the appropriate conditions to the desired bp.

Size	Temperature	Duty Cycle	Intensity	Cycle / Burst	Time	Covaris tubes
400-500bp	6-8ºC (Set 4ºC)	10	5	200	50”	Microtubes

Custom Blunt-end End Repair

5´ Phosphorylation (for subsequent Adaptor Ligation)

1. In a separate tube, prepare a Mix combining the following components (on ice). Add the T4 PNK at the very end:

Water	**32 ul**
CutSmart Buffer, 10X (NEB, REF B7204S)	**7 ul**
100 mM Dithiothreitol [DTT] (Invitrogen, REF Y00147)	**7 ul**
100 mM ATP, Tris buffered (Thermo Scientific, REF R1441)	**1 ul**
T4 Polynucleotide Kinase [PNK] (NEB, 10000 U/ml, REF M0201L)	**1 ul**

2. Mix well by pipetting. Short spin.

3. Transfer **48 ul** of Mix to a new PCR tube and add **50 ul** of fragmented sample.

4. Mix well by pipetting. Short spin.

5. Incubate in a thermal cycler using the program:

    - Choose the pre-heat lid option and set to 100°C.
    - 37°C for 30 minutes.
    - Hold at 4ºC forever.

6. Centrifuge the PCR tube and continue immediately with the blunt-ending step.

Blunt-ending (3´ Overhang Removal & 5´ Overhang Filling)

7. Add the following components (on ice) to the 5´ Phosphorylation reaction:

25 mM dNTP Mix (Illumina, REF 11318102)	**1 ul**
T4 DNA Polymerase (NEB, 3000 U/ml, REF M0203S)	**0.5 ul**

8. Mix well by pipetting. Short spin.

9. Incubate in a thermal cycler using the following program:

   - Choose the pre-heat lid option and set to 100°C.
   - 12°C for 15 minutes.
   - Hold at 4ºC forever.

10. Centrifuge the PCR tube and continue immediately with the AMPure XP bead cleanup step.

Cleanup of Blunt-ended DNA WITHOUT size selection (1x ratio)

1. Vortex AMPure XP Beads to resuspend.

2. Add **100 ul** of beads to the blunt end reaction. Mix well by pipetting up and down at least 10 times.
3. Incubate for 5 min at room temperature.

4. Quickly spin (not more than 2000 rpm) and place it on a magnetic stand to separate the beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Do not disturb the beads.

5. Add **200 ul** of freshly prepared 80% ethanol to each sample while in the magnetic stand. Incubate at room temperature for 30 seconds, then carefully remove and discard the supernatant.

6. Repeat step 5 twice, for a total of 3 washes.

7. Air dry the beads for 5-10 min while the tube is on the magnetic stand with the lid open.

8. Elute the DNA adding **27 ul** of Elution buffer (QIAgen).

9. Mix well by pipetting up and down. Incubate for 2 min at room temperature.

10. Quickly spin the samples and place them on the magnetic stand.

11. After the solution is clear (about 5 min), transfer **25 ul** to a new 1.5 ml tube.

Qubit HS Quantification

Measure the concentration of the blunt-ended samples with the Qubit HS assay.
See Qubit HS protocol.

Custom Blunt-end Circularization Ligation

1. Prepare **up to 50 ng of blunt-ended samples**, adding water to a final volume of **268 ul** in a new 1.5 ml tube.

2. In a separate tube, prepare a Mix combining the following components (on ice):

T4 DNA Ligase Buffer, 10X (Thermo Scientific, Part of EL0011)	**30 ul**

3. Mix by vortexing. Short spin.

4. At the very end, add **2 ul** of **T4 DNA Ligase, 5 Weiss U/ul (Thermo Scientific, EL0011)** to the ligation reaction.

5. Mix well by pipetting. Short spin.

6. Incubate in a Thermomixer block

    - 16°C overnight.

Exonuclease Treatment for Linear DNA Removal

7. Add 15 ul Exonuclease I buffer
8. Add 1 ul lambda exonuclease to the ligation reaction.
9. Add 1 ul Exonuclease I to the ligation reaction

10. Mix well by pipetting. Short spin.

11. Incubate in a Thermomixer block

    - 37°C for 20 minutes.

Cleanup of Adaptor-ligated DNA WITHOUT size selection (1.0x ratio)

1. Vortex AMPure XP Beads to resuspend.
2. Add **320 ul** of beads to the ligation reaction. Mix well by pipetting up and down at least 10 times.
3. Incubate for 5 min at room temperature.
4. Quickly spin (not more than 2000 rpm) and place it on a magnetic stand to separate the beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Do not disturb the beads.
5. Add 4**00 ul** of freshly prepared 80% ethanol to each sample while in the magnetic stand. Incubate at room temperature for 30 seconds, then carefully remove and discard the supernatant.
6. Repeat step 5 twice, for a total of 3 washes.
7. Air dry the beads for 5-10 min while the tube is on the magnetic stand with the lid open.

8. Elute the DNA adding **27 ul** of Elution buffer (QIAgen).
9. Mix well by pipetting up and down. Incubate for 2 min at room temperature.
10. Quickly spin the samples and place them on the magnetic stand.
11. After the solution is clear (about 5 min), transfer **25 ul** to a new PCR tube for the first PCR amplification.

RCA whole genome amplification

1. Denature the DNA and the primers

    a. Mix:

2.1 µl DNA
0.6 µl 10nM primer mix
LE-RE_Lox72 CCCTCGAGGTCGAC*G*G*T
loxloop-8 GCATA*C*A*T
loxloop-7 GCAT*A*C*A
loxloop-6 GCA*T*A*C
0.3 µl Binding Buffer 10x (200 mM TrisHCl, 200 Mm KCl, 1 Mm EDTA)

	1 ml Binding Buffer 10x
1 M Tris-HCl (pH 7.5)	200 µl
2 M KCl	100 µl
0.5 M EDTA	2 µl
H_2O	698 µl

    b. Incubate at 95ºC for 1 min and cool down to 25ºC at 0.1ºC/sec

2. Amplify the circular DNA with phi29 polymerase
    a. Mix:

Denatured sample	3 µl
H_2O	11.1 µl
25mM dNTP	3.2 µl
10 U/µl Phi29 polymerase	0.3 µl
10x Phi29 Buffer	2 µl
100x BSA	0.2 µl
0.1 U/µl Pyrophosphatase	0.2 µl

    b. Incubate at 30ºC for ≈ 24h
    c. Inactivate the enzyme at 65ºC for 10 minutes

Cleanup the RCA product WITHOUT size selection (1x ratio)

3. Vortex AMPure XP Beads to resuspend.
4. Add **20 ul** of beads to the blunt end reaction. Mix well by pipetting up and down at least 10 times.
5. Incubate for 5 min at room temperature.
6. Quickly spin (not more than 2000 rpm) and place it on a magnetic stand to separate the beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Do not disturb the beads.
7. Add **200 ul** of freshly prepared 80% ethanol to each sample while in the magnetic stand. Incubate at room temperature for 30 seconds, then carefully remove and discard the supernatant.
8. Repeat step 5 twice, for a total of 2 washes.
9. Air dry the beads for 5 min while the tube is on the magnetic stand with the lid open.
10. Elute the DNA adding **24 ul** of Elution buffer (QIAgen).

11. Mix well by pipetting up and down. Incubate for 2 min at room temperature.
12. Quickly spin the samples and place them on the magnetic stand.
13. After the solution is clear (about 5 min), transfer **22 ul** to a new 1.5 ml tube.

Qubit HS to check the amplification

Fragmentation (Covaris)

1. Add TE up to **50 ul** of water.
2. Cool down the water bath and degas Covaris (takes about 30min).
3. Shear DNA with the appropriate conditions to the desired bp.

Size	Temperature	Duty Cycle	Intensity	Cycle / Burst	Time	Covaris tubes
300 bp	6-8ºC (Set 4ºC)	10	5	200	80”	Microtubes

NEBNext Ultra II

Continue only with the 2 samples (not the Phi29 reaction control)

End Prep

1. Prepare **500pg – 1ug of sample** in a final volume of **50ul** in a PCR tube.
2. In a separate tube, prepare a Mix combining the following components (on ice) and mixing by pipette:

End Prep enzyme Mix	3 µl
End Prep Reaction Buffer	7 µl

3. **Add 10ul of Mix** and mixing by pipette. Spin.
4. Incubate in a thermal cycler using the program:
   - Choose the pre-heat lid option and set to 100°C
   - 20°C for 30 minutes
   - 65°C for 30 minutes
   - Hold at 4ºC

Samples can be stored at -20ºC, however, a slight loss in yield (20%) may observed. It is recommended continue with adaptor ligation before stopping.

Adaptor Ligation

NOTE: If DNA input is < 100ng, dilute the NEBNext Adaptor following the table below:

INPUT	ADAPTOR DILUTION (volume of adaptor : total volume)	WORKING ADAPTOR CONCENTRATION
1ug – 101ng	**No dilution**	**15 uM**
100ng – 5ng	**1:10**	**1.5uM**
less than 5ng	**1:25**	**0.6uM**

5. In a separate tube, prepare a Mix combining the following components (on ice) and mixing by pipette:

Ligation Master Mix	30ul
NEBNext Adaptor for Illumina*	2.5ul
Ligation Enhancer	1ul

*Provided in NEBNext singleplex or multiplex oligos for Illumina

6. **Add 33.5ul of Mix** and mix by pipette. Spin.
7. Incubate in a thermal cycler using the program:
   - Choose the pre-heat lid option and set to 100°C
   - 20°C for 15 minutes (keep samples at 20ºC)
   - **Add 3ul of USER Enzyme** to the ligation mixture. Mix well and incubate.
   - 37º for 15 minutes
8. Centrifuge the tubes

Samples can be stored at -20ºC

NOTE: A precipitate can form upon thawing of the NEBNext Q5 Hot Start HiFi PCR Master Mix. To ensure optimal performance, place the master mix at room temperature while performing size selection/cleanup of adaptor-ligated DNA. Once thawed, gently mix by inverting the tube several times.

Cleanup Adaptor-ligated DNA WITHOUT Size selection

9. Vortex AMPure XP Beads to resuspend.
10. **Add 87ul (0.9X) of beads** to the ligation reaction. Mix well by pipetting up and down at least 10 times.
11. **Incubate for 5 min** at room temperature.
12. Quickly spin the tube (no more 2000rpm) and place it on a magnetic stand to separate the beads from the supernatant. After the solution is clear (5 min), carefully remove and discard the supernatant. Do not disturb the beads.
13. **Add 200ul of 80% freshly ethanol** to each sample while in the magnetic stand. Incubate at room temperature for 30seconds, then carefully remove and discard the supernatant.
14. Repeat step 62, for a **total of 2 washes**.
15. **Air dry the beads for 5min** while the tube is on the magnetic stand with the lid open.
16. Elute the DNA **adding 17 ul of Elution buffer** (QIAgen).
17. Mix well by pipetting up and down. **Incubate for 2 minutes** at room temperature.
18. Quickly spin the tube and place it on the magnetic stand.
19. After the solution is clear (about 5 min), **transfer 15 ul** to a new PCR tube.

PCR Amplification

20. Mix the following components:

NEBNext Ultra II Q5 Master Mix	25 µl
Index Primer / i7 Primer	5 µl
Universal PCR Primer / i5 Primer	5 ul

*The primers are provided in the NEBNExt Singleplex or Multiplex oligos for Illumina.

21. PCR Cycling conditions:

CYCLE STEP	TEMP	TIME	CYCLES
Initial Denaturation	98ºC	30 seconds	1
Denaturation	98ºC	10 seconds	3 – 15 cycles **10 cycles**
Annealing/Extension	65ºC	75 seconds	
Final Extension	65ºC	5 minutes	1
Cooling	4ºC	∞	1

22. Vortex AMPure XP Beads to resuspend.
23. **Add 45 ul (0.9x) of beads** to the PCR reaction. Mix well by pipetting up and down at least 10 times.
24. **Incubate for 5 min** at room temperature.
25. Quickly spin the tube (no more 2000rpm) and place it on a magnetic stand to separate the beads from the supernatant. After the solution is clear (5 min), carefully remove and discard the supernatant. Do not disturb the beads.
26. **Add 200 ul of 80% freshly ethanol** to each sample while in the magnetic stand. Incubate at room temperature for 30 seconds, then carefully remove and discard the supernatant.
27. Repeat previous step, for a **total of 2 washes**.
28. **Air dry the beads for 5 min** while the tube is on the magnetic stand with the lid open.
29. Elute the DNA **adding 52 ul of Elution buffer** (QIAgen).
30. Mix well by pipetting up and down. **Incubate 2min** at room temperature.
31. Quickly spin the tube and place it on the magnetic stand.
32. After the solution is clear (about 5 min), **transfer 50ul to a new tube**.
33. Vortex AMPure XP Beads to resuspend.
34. **Add 45 ul of beads** to the new tube for a **second purification**. Mix well by pipetting up and down at least 10 times.
35. **Incubate for 5 min** at room temperature.
36. Quickly spin the tube (no more 2000 rpm) and place it on a magnetic stand to separate the beads from the supernatant. After the solution is clear (5 min), carefully remove and discard the supernatant. Do not disturb the beads.
37. **Add 200 ul of 80% freshly ethanol** to each sample while in the magnetic stand. Incubate at room temperature for 30 seconds, then carefully remove and discard the supernatant.
38. Repeat previous step, for a **total of 2 washes**.
39. **Air dry the beads for 5 min** while the tube is on the magnetic stand with the lid open.
40. Elute the DNA **adding 33 ul of Elution buffer** (QIAgen).
41. Mix well by pipetting up and down. **Incubate 2min** at room temperature.
42. Quickly spin the tube and place it on the magnetic stand.
43. After the solution is clear (about 5 min), **transfer 30 ul to a new tube**.

Assess the library quality on a Bioanalyzer

Check that the electropherogram shows a narrow distribution with the appropriated peak size. DNA 1000 assay.

SUPPLEMENTARY MATERIAL B – ALL DELETED GENES FROM PROTOCOL 3

MPN ID	Gene	Function	Essentiality
MPN035		Conserved hypothetical protein	NE
MPN036		Conserved hypothetical protein	NE
MPN037		Uncharacterized protein	NE
MPN083		Uncharacterized lipoprotein	F
MPN084		Conserved hypothetical lipoprotein	NE
MPN085		Uncharacterized protein	NE
MPN086		Uncharacterized protein	NE
MPN087		Uncharacterized protein	NE
MPN088		Uncharacterized protein	NE
MPN089	*hsdS*	Putative type-1 restriction enzyme specificity protein	NE
MPN090		Uncharacterized protein	NE
MPN091		Conserved hypothetical protein	NE
MPN092		Putative mgpC-like protein	NE
MPN093		Putative mgpC-like protein	NE
MPN094		UPF0134 protein	NE
MPN095		Uncharacterized amino acid permease	NE
MPN096		Uncharacterized amino acid permease	NE
MPN097		Conserved hypothetical lipoprotein	NE
MPN098		Conserved hypothetical lipoprotein	NE
MPN099		Putative adhesin P1-like protein	NE
MPN100		Uncharacterized protein	NE
MPN101		Uncharacterized protein	NE
MPN102		Putative mgpC-like protein	NE
MPN103		Uncharacterized protein	NE
MPN104		Uncharacterized protein	NE
MPN108		Uncharacterized adenine-specific methylase	NE
MPN109		Uncharacterized protein	NE
MPN110		Conserved hypothetical protein	NE
MPN145		Uncharacterized protein	NE
MPN146		Conserved hypothetical protein	NE
MPN147		Conserved hypothetical protein	NE
MPN148		Conserved hypothetical protein	NE
MPN149		Putative mgpC-like protein	NE
MPN150		Putative mgpC-like protein	NE
MPN151		Uncharacterized protein	NE
MPN152		Uncharacterized lipoprotein	NE
MPN153	*uvrD*	Probable DNA helicase I homolog	NE
MPN281		Conserved hypothetical lipoprotein	NE
MPN282		Conserved hypothetical protein	NE
MPN283		Uncharacterized protein	NE
MPN284		Uncharacterized lipoprotein	NE
MPN285	*prrB*	Putative type-1 restriction enzyme specificity protein	NE

MPN286		Putative adhesin P1-like protein	NE
MPN287		Uncharacterized protein	NE
MPN288		Conserved hypothetical lipoprotein	NE
MPN289		Putative type-1 restriction enzyme specificity protein	NE
MPN290		Putative type-1 restriction enzyme specificity protein	NE
MPN308		Uncharacterized amino acid permease	NE
MPN322	*nrdF*	Ribonucleoside-diphosphate reductase subunit beta	NE
MPN323	*nrdI*	Protein nrdI	NE
MPN324	*nrdE*	Ribonucleoside-diphosphate reductase subunit alpha	NE
MPN333		Putative ABC transport system permease protein	NE
MPN334	*bcrA*	Putative ABC transporter ATP-binding protein	NE
MPN343		Putative type-1 restriction enzyme specificity protein	NE
MPN344		Uncharacterized protein	NE
MPN345		Putative type-1 restriction enzyme mpnORFDP R protein part 2	NE
MPN364		Conserved hypothetical protein	NE
MPN365		Putative type-1 restriction enzyme specificity protein	NE
MPN366		Putative mgpC-like protein	NE
MPN367		Putative mgpC-like protein	NE
MPN368		Uncharacterized protein	NE
MPN369		Uncharacterized lipoprotein	NE
MPN370		Putative adhesin P1-like protein	NE
MPN371		Uncharacterized protein	NE
MPN372	*ptxA*	ADP-ribosylating toxin CARDS	NE
MPN373		Uncharacterized protein	NE
MPN374		Uncharacterized protein	NE
MPN375		Uncharacterized protein	NE
MPN376		Uncharacterized protein	NE
MPN397	*spoT*	Probable guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	NE
MPN398		Uncharacterized protein	NE
MPN399		Conserved hypothetical protein	NE
MPN400		Conserved hypothetical protein	NE
MPN407		Predicted lipase	NE
MPN438		Uncharacterized protein	NE
MPN439		Uncharacterized lipoprotein	NE
MPN440		Uncharacterized protein	NE
MPN441		Conserved hypothetical protein	NE
MPN457		Uncharacterized protein	NE
MPN458		Conserved hypothetical protein	NE
MPN459		Conserved hypothetical lipoprotein	NE
MPN465		Conserved hypothetical protein	NE
MPN466		Conserved hypothetical protein	NE
MPN490	*recA*	Protein recA	F
MPN491	*mnuA*	Membrane nuclease A	F
MPN492	*ulaE*	Probable L-ribulose-5-phosphate 3-epimerase	NE
MPN493	*ulaD*	Probable 3-keto-L-gulonate-6-phosphate decarboxylase	NE

MPN494	*ulaC*	Ascorbate-specific phosphotransferase enzyme IIA component	F
MPN495	*ulaB*	Ascorbate-specific phosphotransferase enzyme IIB component	NE
MPN496	*ulaA*	Ascorbate-specific permease IIC component	NE
MPN497	*ulaG*	Probable L-ascorbate-6-phosphate lactonase	NE
MPN498	*araD*	Probable L-ribulose-5-phosphate 4-epimerase	NE
MPN499		Uncharacterized protein	NE
MPN500		Putative adhesin P1-like protein	NE
MPN501		Uncharacterized protein	NE
MPN502		Uncharacterized protein	NE
MPN503		Putative mgpC-like protein	NE
MPN504		Uncharacterized protein	NE
MPN505		Uncharacterized protein	NE
MPN506		Conserved hypothetical lipoprotein	NE
MPN507		Putative type-1 restriction enzyme specificity protein	NE
MPN508		Putative membrane export protein	NE
MPN509		Uncharacterized protein	NE
MPN510		Uncharacterized protein	NE
MPN511		Uncharacterized protein	NE
MPN512		Uncharacterized protein	NE
MPN513		Conserved hypothetical protein	NE
MPN577		Conserved hypothetical protein	NE
MPN578		Conserved hypothetical protein	NE
MPN579		Conserved hypothetical protein	NE
MPN580		Putative protease	NE
MPN581		Conserved hypothetical protein	NE
MPN582		Conserved hypothetical lipoprotein	NE
MPN583		Conserved hypothetical protein	NE
MPN584		Conserved hypothetical protein	NE
MPN585		Conserved hypothetical lipoprotein	NE
MPN586		Conserved hypothetical protein	NE
MPN587		Conserved hypothetical lipoprotein	NE
MPN588		Uncharacterized lipoprotein	NE
MPN589		Conserved hypothetical protein	NE
MPN590		Conserved hypothetical lipoprotein	NE
MPN591		Conserved hypothetical protein	NE
MPN592		Conserved hypothetical lipoprotein	NE
MPN593		Conserved hypothetical protein	NE
MPN594		Conserved hypothetical protein	NE
MPN595	*lacA*	Probable ribose-5-phosphate isomerase B	F
MPN596	*erzA*	Negative regulator of FtsZ ring formation	NE
MPN609	*pstB*	Phosphate import ATP-binding protein	F
MPN610	*pstA*	Phosphate transport system permease protein	F
MPN611	*pstS*	Phosphate-binding protein	F
MPN612		Conserved hypothetical protein	NE
MPN613		Conserved hypothetical protein	NE

MPN614	Conserved hypothetical protein	NE
MPN647	Conserved hypothetical lipoprotein	NE
MPN648	Conserved hypothetical lipoprotein	NE
MPN649	Uncharacterized protein	NE
MPN650	Uncharacterized lipoprotein	NE
MPN651	PTS system mannitol-specific EIICB component	NE
MPN652	Mannitol-1-phosphate 5-dehydrogenase	NE
MPN653	Mannitol-specific phosphotransferase enzyme	NE
MPN654	Conserved hypothetical lipoprotein	NE
MPN655	Uncharacterized protein	NE

REFERENCES

Abil, Z., Xiong, X., Zhao, H., 2015. Synthetic Biology for Therapeutic Applications. Mol. Pharm. 12, 322–331. https://doi.org/10.1021/mp500392q

Adler, M., Anjum, M., Berg, O.G., Andersson, D.I., Sandegren, L., 2014. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. Mol. Biol. Evol. 31, 1526–1535. https://doi.org/10.1093/molbev/msu111

Albert, H., Dale, E.C., Lee, E., Ow, D.W., 1995. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. Plant J. Cell Mol. Biol. 7, 649–659.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ara, K., Ozaki, K., Nakamura, K., Yamane, K., Sekiguchi, J., Ogasawara, N., 2007. Bacillus minimum genome factory: effective utilization of microbial genome information. Biotechnol. Appl. Biochem. 46, 169–178. https://doi.org/10.1042/BA20060111

Ausländer, S., Ausländer, D., Fussenegger, M., 2017. Synthetic Biology-The Synthesis of Biology. Angew. Chem. Int. Ed Engl. 56, 6396–6419. https://doi.org/10.1002/anie.201609229

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H., 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. 2, 2006.0008. https://doi.org/10.1038/msb4100050

Babakhani, S., Oloomi, M., 2018. Transposons: the agents of antibiotic resistance in bacteria. J. Basic Microbiol. 58, 905–917. https://doi.org/10.1002/jobm.201800204

Bachman, M.A., Breen, P., Deornellas, V., Mu, Q., Zhao, L., Wu, W., Cavalcoli, J.D., Mobley, H.L.T., 2015. Genome-Wide Identification of Klebsiella pneumoniae Fitness Genes during Lung Infection. mBio 6, e00775. https://doi.org/10.1128/mBio.00775-15

Badrinarayanan, A., Le, T.B., Laub, M.T., 2015. Bacterial chromosome organization and segregation. Annu. Rev. Cell Dev. Biol. 31, 171–199. https://doi.org/10.1146/annurev-cellbio-100814-125211

Baggett, N.E., Zhang, Y., Gross, C.A., 2017. Global analysis of translation termination in E. coli. PLoS Genet. 13, e1006676. https://doi.org/10.1371/journal.pgen.1006676

Balish, M.F., 2014. Mycoplasma pneumoniae, an Underutilized Model for Bacterial Cell Biology. J. Bacteriol. 196, 3675–3682. https://doi.org/10.1128/JB.01865-14

Basler, G., Nikoloski, Z., Larhlimi, A., Barabási, A.-L., Liu, Y.-Y., 2016. Control of fluxes in metabolic networks. Genome Res. 26, 956–968. https://doi.org/10.1101/gr.202648.115

Beckwith, J., 2013. Fifty years fused to lac. Annu. Rev. Microbiol. 67, 1–19. https://doi.org/10.1146/annurev-micro-092412-155732

Beier, L.S., Siqueira, F.M., Schrank, I.S., 2018. Evaluation of growth and gene expression of Mycoplasma hyopneumoniae and Mycoplasma hyorhinis in defined medium. Mol. Biol. Rep. 45, 2469–2479. https://doi.org/10.1007/s11033-018-4413-3

Berg, D.E., Davies, J., Allet, B., Rochaix, J.D., 1975. Transposition of R factor genes to bacteriophage lambda. Proc. Natl. Acad. Sci. U. S. A. 72, 3628–3632. https://doi.org/10.1073/pnas.72.9.3628

Bishop, A.H., Rachwal, P.A., Vaid, A., 2014. Identification of genes required by Bacillus thuringiensis for survival in soil by transposon-directed insertion site sequencing. Curr. Microbiol. 68, 477–485. https://doi.org/10.1007/s00284-013-0502-7

Blanchard, A., Bébéar, C., 2011. The evolution of Mycoplasma genitalium. Ann. N. Y. Acad. Sci. 1230, E61-64. https://doi.org/10.1111/j.1749-6632.2011.06418.x

Blötz, C., Stülke, J., 2017. Glycerol metabolism and its implication in virulence in Mycoplasma. FEMS Microbiol. Rev. 41, 640–652. https://doi.org/10.1093/femsre/fux033

Braff, D., Shis, D., Collins, J.J., 2016. Synthetic biology platform technologies for antimicrobial applications. Adv. Drug Deliv. Rev. 105, 35–43. https://doi.org/10.1016/j.addr.2016.04.006

Breton, M., Duret, S., Béven, L., Dubrana, M.-P., Renaudin, J., 2011. I-SceI-mediated plasmid deletion and intra-molecular recombination in Spiroplasma citri. J. Microbiol. Methods 84, 216–222. https://doi.org/10.1016/j.mimet.2010.11.020

Breuer, M., Earnest, T.M., Merryman, C., Wise, K.S., Sun, L., Lynott, M.R., Hutchison, C.A., Smith, H.O., Lapek, J.D., Gonzalez, D.J., de Crécy-Lagard, V., Haas, D., Hanson, A.D., Labhsetwar, P., Glass, J.I., Luthey-Schulten, Z., 2019. Essential metabolism for a minimal cell. eLife 8. https://doi.org/10.7554/eLife.36842

Bright, A.R., Veenstra, G.J.C., 2019. Assay for Transposase-Accessible Chromatin-Sequencing Using Xenopus Embryos. Cold Spring Harb. Protoc. 2019, pdb.prot098327. https://doi.org/10.1101/pdb.prot098327

Bronner, I.F., Quail, M.A., Turner, D.J., Swerdlow, H., 2009. Improved Protocols for Illumina Sequencing. Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al 0 18. https://doi.org/10.1002/0471142905.hg1802s62

Broulette, J., Yu, H., Pyenson, B., Iwasaki, K., Sato, R., 2013. The Incidence Rate and Economic Burden of Community-Acquired Pneumonia in a Working-Age Population. Am. Health Drug Benefits 6, 494–503.

Browning, D.F., Busby, S.J.W., 2016. Local and global regulation of transcription initiation in bacteria. Nat. Rev. Microbiol. 14, 638–650. https://doi.org/10.1038/nrmicro.2016.103

Buenrostro, J., Wu, B., Chang, H., Greenleaf, W., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al 109, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Burgos, R., Totten, P.A., 2014. Characterization of the operon encoding the Holliday junction helicase RuvAB from Mycoplasma genitalium and its role in mgpB and mgpC gene variation. J. Bacteriol. 196, 1608–1618. https://doi.org/10.1128/JB.01385-13

Buskirk, A.R., Green, R., 2017. Ribosome pausing, arrest and rescue in bacteria and eukaryotes. Philos. Trans. R. Soc. B Biol. Sci. 372. https://doi.org/10.1098/rstb.2016.0183

Campbell, null, Davies, null, Bulone, null, Henrissat, null, 1998. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem. J. 329 (Pt 3), 719. https://doi.org/10.1042/bj3290719

Cantine, M.D., Fournier, G.P., 2018. Environmental Adaptation from the Origin of Life to the Last Universal Common Ancestor. Orig. Life Evol. Biosphere J. Int. Soc. Study Orig. Life 48, 35–54. https://doi.org/10.1007/s11084-017-9542-5
Cao, J., Kapke, P.A., Minion, F.C., 1994. Transformation of Mycoplasma gallisepticum with Tn916, Tn4001, and integrative plasmid vectors. J. Bacteriol. 176, 4459–4462. https://doi.org/10.1128/jb.176.14.4459-4462.1994
Cassier-Chauvat, C., Poncelet, M., Chauvat, F., 1997. Three insertion sequences from the cyanobacterium Synechocystis PCC6803 support the occurrence of horizontal DNA transfer among bacteria. Gene 195, 257–266.
Chakraborty, R., Woo, H., Dehal, P., Walker, R., Zemla, M., Auer, M., Goodwin, L.A., Kazakov, A., Novichkov, P., Arkin, A.P., Hazen, T.C., 2017. Complete genome sequence of Pseudomonas stutzeri strain RCH2 isolated from a Hexavalent Chromium [Cr(VI)] contaminated site. Stand. Genomic Sci. 12, 23. https://doi.org/10.1186/s40793-017-0233-7
Charlebois, R.L., Doolittle, W.F., 2004. Computing prokaryotic gene ubiquity: Rescuing the core from extinction. Genome Res. 14, 2469–2477. https://doi.org/10.1101/gr.3024704
Chatterjee, P.K., Shakes, L.A., Stennett, N., Richardson, V.L., Malcolm, T.L., Harewood, K.R., 2010. Replacing the wild type loxP site in BACs from the public domain with lox66 using a lox66 transposon. BMC Res. Notes 3, 38. https://doi.org/10.1186/1756-0500-3-38
Cho, H., Wivagg, C.N., Kapoor, M., Barry, Z., Rohs, P.D.A., Suh, H., Marto, J.A., Garner, E.C., Bernhardt, T.G., 2016. Bacterial cell wall biogenesis is mediated by SEDS and PBP polymerase families functioning semi-autonomously. Nat. Microbiol. 1, 16172. https://doi.org/10.1038/nmicrobiol.2016.172
Choe, D., Cho, S., Kim, S.C., Cho, B.-K., 2016. Minimal genome: Worthwhile or worthless efforts toward being smaller? Biotechnol. J. 11, 199–211. https://doi.org/10.1002/biot.201400838
Chong, R.A., Moran, N.A., 2018. Evolutionary loss and replacement of Buchnera, the obligate endosymbiont of aphids. ISME J. 12, 898–908. https://doi.org/10.1038/s41396-017-0024-6
Christen, B., Abeliuk, E., Collier, J.M., Kalogeraki, V.S., Passarelli, B., Coller, J.A., Fero, M.J., McAdams, H.H., Shapiro, L., 2011. The essential genome of a bacterium. Mol. Syst. Biol. 7, 528. https://doi.org/10.1038/msb.2011.58
Christie-Oleza, J.A., Brunet-Galmés, I., Lalucat, J., Nogales, B., Bosch, R., 2013. MiniUIB, a Novel Minitransposon-Based System for Stable Insertion of Foreign DNA into the Genomes of Gram-Negative and Gram-Positive Bacteria. Appl. Environ. Microbiol. 79, 1629–1638. https://doi.org/10.1128/AEM.03214-12
Chubukov, V., Gerosa, L., Kochanowski, K., Sauer, U., 2014. Coordination of microbial metabolism. Nat. Rev. Microbiol. 12, 327–340. https://doi.org/10.1038/nrmicro3238
Claesen, J., Fischbach, M.A., 2015. Synthetic Microbes As Drug Delivery Systems. ACS Synth. Biol. 4, 358–364. https://doi.org/10.1021/sb500258b
Clarke, M., Lohan, A.J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T.R., Frech, C., Turcotte, B., Kopec, K.O., Synnott, J.M., Choo, C., Paponov, I., Finkler, A., Heng Tan, C.S., Hutchins, A.P., Weinmeier, T., Rattei, T., Chu, J.S., Gimenez, G., Irimia, M., Rigden, D.J., Fitzpatrick, D.A., Lorenzo-Morales, J., Bateman, A., Chiu, C.-H., Tang, P., Hegemann, P., Fromm, H., Raoult, D., Greub, G., Miranda-Saavedra, D., Chen, N., Nash, P., Ginger, M.L., Horn, M., Schaap, P., Caler, L., Loftus,

B.J., 2013. Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol. 14, R11. https://doi.org/10.1186/gb-2013-14-2-r11
Coates, C.J., Turney, C.L., Frommer, M., O'Brochta, D.A., Atkinson, P.W., 1997. Interplasmid transposition of the mariner transposable element in non-drosophilid insects. Mol. Gen. Genet. MGG 253, 728–733.
Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science 339, 819–823. https://doi.org/10.1126/science.1231143
Conway, T., Creecy, J.P., Maddox, S.M., Grissom, J.E., Conkle, T.L., Shadid, T.M., Teramoto, J., San Miguel, P., Shimada, T., Ishihama, A., Mori, H., Wanner, B.L., 2014. Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. mBio 5. https://doi.org/10.1128/mBio.01442-14
Craig, N.L., 1991. Tn7: a target site-specific transposon. Mol. Microbiol. 5, 2569–2573. https://doi.org/10.1111/j.1365-2958.1991.tb01964.x
Cunha, B.A., Pherez, F.M., 2009. Mycoplasma pneumoniae community-acquired pneumonia (CAP) in the elderly: Diagnostic significance of acute thrombocytosis. Heart Lung J. Crit. Care 38, 444–449. https://doi.org/10.1016/j.hrtlng.2008.10.005
Curtis, P.D., 2016. Essential Genes Predicted in the Genome of Rubrivivax gelatinosus. J. Bacteriol. 198, 2244–2250. https://doi.org/10.1128/JB.00344-16
Curtis, P.D., Brun, Y.V., 2014. Identification of essential alphaproteobacterial genes reveals operational variability in conserved developmental and cell cycle systems. Mol. Microbiol. 93, 713–735. https://doi.org/10.1111/mmi.12686
Danchin, A., 2012. Scaling up synthetic biology: Do not forget the chassis. FEBS Lett. 586, 2129–2137. https://doi.org/10.1016/j.febslet.2011.12.024
Dar, D., Sorek, R., 2017. Regulation of antibiotic-resistance by non-coding RNAs in bacteria. Curr. Opin. Microbiol. 36, 111–117. https://doi.org/10.1016/j.mib.2017.02.005
Davis, J.J., Xia, F., Overbeek, R.A., Olsen, G.J., 2013. Genomes of the class Erysipelotrichia clarify the firmicute origin of the class Mollicutes. Int. J. Syst. Evol. Microbiol. 63, 2727–2741. https://doi.org/10.1099/ijs.0.048983-0
de Lorenzo, V., Herrero, M., Jakubzik, U., Timmis, K.N., 1990. Mini-Tn5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in gram-negative eubacteria. J. Bacteriol. 172, 6568–6572.
de Lorenzo, V., Timmis, K.N., 1994. Analysis and construction of stable phenotypes in gram-negative bacteria with Tn5- and Tn10-derived minitransposons. Methods Enzymol. 235, 386–405. https://doi.org/10.1016/0076-6879(94)35157-0
DeJesus, M.A., Zhang, Y.J., Sassetti, C.M., Rubin, E.J., Sacchettini, J.C., Ioerger, T.R., 2013. Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. Bioinforma. Oxf. Engl. 29, 695–703. https://doi.org/10.1093/bioinformatics/btt043
den Blaauwen, T., Hamoen, L.W., Levin, P.A., 2017. The divisome at 25: the road ahead. Curr. Opin. Microbiol. 36, 85–94. https://doi.org/10.1016/j.mib.2017.01.007

Deng, J., Su, S., Lin, X., Hassett, D.J., Lu, L.J., 2013. A statistical framework for improving genomic annotations of prokaryotic essential genes. PloS One 8, e58178. https://doi.org/10.1371/journal.pone.0058178
Deutschbauer, A., Price, M.N., Wetmore, K.M., Shao, W., Baumohl, J.K., Xu, Z., Nguyen, M., Tamse, R., Davis, R.W., Arkin, A.P., 2011. Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet. 7, e1002385. https://doi.org/10.1371/journal.pgen.1002385
Di Giulio, M., 2011. The last universal common ancestor (LUCA) and the ancestors of archaea and bacteria were progenotes. J. Mol. Evol. 72, 119–126. https://doi.org/10.1007/s00239-010-9407-2
diCenzo, G.C., Finan, T.M., 2017. The Divided Bacterial Genome: Structure, Function, and Evolution. Microbiol. Mol. Biol. Rev. MMBR 81. https://doi.org/10.1128/MMBR.00019-17
Diedrich, G., Spahn, C.M.T., Stelzl, U., Schäfer, M.A., Wooten, T., Bochkariov, D.E., Cooperman, B.S., Traut, R.R., Nierhaus, K.H., 2000. Ribosomal protein L2 is involved in the association of the ribosomal subunits, tRNA binding to A and P sites and peptidyl transfer. EMBO J. 19, 5241–5250. https://doi.org/10.1093/emboj/19.19.5241
Douglas, A.E., 1998. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. Annu. Rev. Entomol. 43, 17–37. https://doi.org/10.1146/annurev.ento.43.1.17
Dybvig, K., Alderete, J., 1988. Transformation of Mycoplasma pulmonis and Mycoplasma hyorhinis: transposition of Tn916 and formation of cointegrate structures. Plasmid 20, 33–41.
Dybvig, K., Cassell, G.H., 1987. Transposition of gram-positive transposon Tn916 in Acholeplasma laidlawii and Mycoplasma pulmonis. Science 235, 1392–1394. https://doi.org/10.1126/science.3029869
Dybvig, K., French, C.T., Voelker, L.L., 2000. Construction and Use of Derivatives of Transposon Tn4001 That Function in Mycoplasma pulmonis and Mycoplasma arthritidis. J. Bacteriol. 182, 4343–4347.
Dybvig, K., Voelker, L.L., 1996. Molecular biology of mycoplasmas. Annu. Rev. Microbiol. 50, 25–57. https://doi.org/10.1146/annurev.micro.50.1.25
Emmons, S.W., Yesner, L., Ruan, K.S., Katzenberg, D., 1983. Evidence for a transposon in Caenorhabditis elegans. Cell 32, 55–65.
Endo, A., Kurusu, Y., 2007. Identification of in vivo substrates of the chaperonin GroEL from Bacillus subtilis. Biosci. Biotechnol. Biochem. 71, 1073–1077. https://doi.org/10.1271/bbb.60640
Entwistle, S., Li, X., Yin, Y., 2019. Orphan Genes Shared by Pathogenic Genomes Are More Associated with Bacterial Pathogenicity. mSystems 4. https://doi.org/10.1128/mSystems.00290-18
Eriani, G., Delarue, M., Poch, O., Gangloff, J., Moras, D., 1990. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347, 203–206. https://doi.org/10.1038/347203a0
Errington, J., 2013. L-form bacteria, cell walls and the origins of life. Open Biol. 3. https://doi.org/10.1098/rsob.120143
Fadiel, A., Eichenbaum, K.D., El Semary, N., Epperson, B., 2007. Mycoplasma genomics: tailoring the genome for minimal life requirements through reductive evolution. Front. Biosci. J. Virtual Libr. 12, 2020–2028.

Fang, G., Rocha, E., Danchin, A., 2005. How essential are nonessential genes? Mol. Biol. Evol. 22, 2147–2156. https://doi.org/10.1093/molbev/msi211
Feldhaar, H., Gross, R., 2009. Genome degeneration affects both extracellular and intracellular bacterial endosymbionts. J. Biol. 8, 31. https://doi.org/10.1186/jbiol129
Finstad, K.M., Probst, A.J., Thomas, B.C., Andersen, G.L., Demergasso, C., Echeverría, A., Amundson, R.G., Banfield, J.F., 2017. Microbial Community Structure and the Persistence of Cyanobacterial Populations in Salt Crusts of the Hyperarid Atacama Desert from Genome-Resolved Metagenomics. Front. Microbiol. 8. https://doi.org/10.3389/fmicb.2017.01435
Forterre, P., 2016. To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. Stud. Hist. Philos. Biol. Biomed. Sci. 59, 100–108. https://doi.org/10.1016/j.shpsc.2016.02.013
Forterre, P., 2015. The universal tree of life: an update. Front. Microbiol. 6, 717. https://doi.org/10.3389/fmicb.2015.00717
Forterre, P., 2010. Defining Life: The Virus Viewpoint. Orig. Life Evol. Biosph. 40, 151–160. https://doi.org/10.1007/s11084-010-9194-1
Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., Venter, J.C., 1995. The minimal gene complement of Mycoplasma genitalium. Science 270, 397–403.
Freed, N.E., Bumann, D., Silander, O.K., 2016. Combining Shigella Tn-seq data with Gold-standard E. coli Gene Deletion Data Suggests Rare Transitions between Essential and Non-essential Gene Functionality. bioRxiv 038869. https://doi.org/10.1101/038869
French, C.T., Lao, P., Loraine, A.E., Matthews, B.T., Yu, H., Dybvig, K., 2008. Large-scale transposon mutagenesis of Mycoplasma pulmonis. Mol. Microbiol. 69, 67–76. https://doi.org/10.1111/j.1365-2958.2008.06262.x
Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J., Shu, S., Neupane, R., Cipriano, M., Mancuso, J., Tu, H., Salamov, A., Lindquist, E., Shapiro, H., Lucas, S., Grigoriev, I.V., Cande, W.Z., Fulton, C., Rokhsar, D.S., Dawson, S.C., 2010. The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell 140, 631–642. https://doi.org/10.1016/j.cell.2010.01.032
Galperin, M.Y., Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43, D261-269. https://doi.org/10.1093/nar/gku1223
García-Angulo, V.A., 2017. Overlapping riboflavin supply pathways in bacteria. Crit. Rev. Microbiol. 43, 196–209. https://doi.org/10.1080/1040841X.2016.1192578
Gawronski, J.D., Wong, S.M.S., Giannoukos, G., Ward, D.V., Akerley, B.J., 2009. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proc. Natl. Acad. Sci. U. S. A. 106, 16422–16427. https://doi.org/10.1073/pnas.0906627106
Gazzaniga, F., Stebbins, R., Chang, S.Z., McPeek, M.A., Brenner, C., 2009. Microbial NAD Metabolism: Lessons from Comparative Genomics. Microbiol. Mol. Biol. Rev. MMBR 73, 529–541. https://doi.org/10.1128/MMBR.00042-08

Gerdes, K., Howard, M., Szardenings, F., 2010. Pushing and pulling in prokaryotic DNA segregation. Cell 141, 927–942. https://doi.org/10.1016/j.cell.2010.05.033
Ghosh, S., O'Connor, T.J., 2017. Beyond Paralogs: The Multiple Layers of Redundancy in Bacterial Pathogenesis. Front. Cell. Infect. Microbiol. 7, 467. https://doi.org/10.3389/fcimb.2017.00467
Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.-Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A., Smith, H.O., Venter, J.C., 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329, 52–56. https://doi.org/10.1126/science.1190719
Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345. https://doi.org/10.1038/nmeth.1318
Giegé, R., Springer, M., 2016. Aminoacyl-tRNA Synthetases in the Bacterial World. EcoSal Plus 7. https://doi.org/10.1128/ecosalplus.ESP-0002-2016
Gil, R., Silva, F.J., Peretó, J., Moya, A., 2004. Determination of the Core of a Minimal Bacterial Gene Set. Microbiol. Mol. Biol. Rev. 68, 518–537. https://doi.org/10.1128/MMBR.68.3.518-537.2004
Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O., Venter, J.C., 2006. Essential genes of a minimal bacterium. Proc. Natl. Acad. Sci. U. S. A. 103, 425–430. https://doi.org/10.1073/pnas.0510013103
Glass, J.I., Merryman, C., Wise, K.S., Hutchison, C.A., Smith, H.O., 2017. Minimal Cells-Real and Imagined. Cold Spring Harb. Perspect. Biol. https://doi.org/10.1101/cshperspect.a023861
Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R., Gordon, J.I., 2009. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe 6, 279–289. https://doi.org/10.1016/j.chom.2009.08.003
Grazziotin, A.L., Vidal, N.M., Venancio, T.M., 2015. Uncovering major genomic features of essential genes in Bacteria and a methanogenic Archaea. FEBS J. 282, 3395–3411. https://doi.org/10.1111/febs.13350
Grillo, M.A., Colombatto, S., 2008. S-adenosylmethionine and its products. Amino Acids 34, 187–193. https://doi.org/10.1007/s00726-007-0500-9
Gualerzi, C.O., Pon, C.L., 2015. Initiation of mRNA translation in bacteria: structural and dynamic aspects. Cell. Mol. Life Sci. CMLS 72, 4341–4367. https://doi.org/10.1007/s00018-015-2010-3
Gueiros-Filho, F.J., Beverley, S.M., 1997. Trans-kingdom transposition of the Drosophila element mariner within the protozoan Leishmania. Science 276, 1716–1719.
Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., Serrano, L., 2009a. Transcriptome complexity in a genome-reduced bacterium. Science 326, 1268–1271. https://doi.org/10.1126/science.1176951
Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., Serrano, L., 2009b. Transcriptome

complexity in a genome-reduced bacterium. Science 326, 1268–1271. https://doi.org/10.1126/science.1176951
Güell, M., Yus, E., Lluch-Senar, M., Serrano, L., 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nat. Rev. Microbiol. 9, 658–669. https://doi.org/10.1038/nrmicro2620
Guo, H., Arambula, D., Ghosh, P., Miller, J.F., 2014. Diversity-generating Retroelements in Phage and Bacterial Genomes. Microbiol. Spectr. 2. https://doi.org/10.1128/microbiolspec.MDNA3-0029-2014
Haellman, V., Fussenegger, M., 2016. Synthetic Biology--Toward Therapeutic Solutions. J. Mol. Biol. 428, 945–962. https://doi.org/10.1016/j.jmb.2015.08.020
Halbedel, S., Stülke, J., 2007. Tools for the genetic analysis of Mycoplasma. Int. J. Med. Microbiol. IJMM 297, 37–44. https://doi.org/10.1016/j.ijmm.2006.11.001
Hamasuna, R., 2013. Mycoplasma genitalium in male urethritis: diagnosis and treatment in Japan. Int. J. Urol. Off. J. Jpn. Urol. Assoc. 20, 676–684. https://doi.org/10.1111/iju.12152
Han, K., Li, Z., Peng, R., Zhu, L., Zhou, T., Wang, L., Li, S., Zhang, X., Hu, W., Wu, Z., Qin, N., Li, Y., 2013. Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Sci. Rep. 3, 2101. https://doi.org/10.1038/srep02101
Hasegawa, N., Sekizuka, T., Sugi, Y., Kawakami, N., Ogasawara, Y., Kato, K., Yamashita, A., Takeuchi, F., Kuroda, M., 2017. Characterization of the Pathogenicity of Streptococcus intermedius TYG1620 Isolated from a Human Brain Abscess Based on the Complete Genome Sequence with Transcriptome Analysis and Transposon Mutagenesis in a Murine Subcutaneous Abscess Model. Infect. Immun. 85. https://doi.org/10.1128/IAI.00886-16
Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., Ote, T., Yamakawa, T., Yamazaki, Y., Mori, H., Katayama, T., Kato, J., 2005. Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. Mol. Microbiol. 55, 137–149. https://doi.org/10.1111/j.1365-2958.2004.04386.x
Hasselbring, B.M., Jordan, J.L., Krause, R.W., Krause, D.C., 2006a. Terminal organelle development in the cell wall-less bacterium Mycoplasma pneumoniae. Proc. Natl. Acad. Sci. U. S. A. 103, 16478–16483. https://doi.org/10.1073/pnas.0608051103
Hasselbring, B.M., Page, C.A., Sheppard, E.S., Krause, D.C., 2006b. Transposon Mutagenesis Identifies Genes Associated with Mycoplasma pneumoniae Gliding Motility. J. Bacteriol. 188, 6335–6345. https://doi.org/10.1128/JB.00698-06
Hayakawa, T., Tanaka, T., Sakaguchi, K., Otake, N., Yonehara, H., 1979. A LINEAR PLASMID-LIKE DNA IN STREPTOMYCES SP. PRODUCING LANKACIDIN GROUP ANTIBIOTICS. J. Gen. Appl. Microbiol. 25, 255–260. https://doi.org/10.2323/jgam.25.255
Hayflick, L., 1965. Tissue cultures and mycoplasmas. Tex. Rep. Biol. Med. 23, Suppl 1:285+.
Hedreyda, C.T., Lee, K.K., Krause, D.C., 1993. Transformation of Mycoplasma pneumoniae with Tn4001 by electroporation. Plasmid 30, 170–175. https://doi.org/10.1006/plas.1993.1047
Henderson, B., 2014. An overview of protein moonlighting in bacterial infection. Biochem. Soc. Trans. 42, 1720–1727. https://doi.org/10.1042/BST20140236

Henderson, B., Martin, A., 2013. Bacterial moonlighting proteins and bacterial virulence. Curr. Top. Microbiol. Immunol. 358, 155–213. https://doi.org/10.1007/82_2011_188
Henderson, B., Martin, A., 2011. Bacterial virulence in the moonlight: multitasking bacterial moonlighting proteins are virulence determinants in infectious disease. Infect. Immun. 79, 3476–3491. https://doi.org/10.1128/IAI.00179-11
Herrmann, R., Reiner, B., 1998. Mycoplasma pneumoniae and Mycoplasma genitalium: a comparison of two closely related bacterial species. Curr. Opin. Microbiol. 1, 572–579.
Hickman, A.B., Dyda, F., 2016. DNA Transposition at Work. Chem. Rev. 116, 12758–12784. https://doi.org/10.1021/acs.chemrev.6b00003
Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., Herrmann, R., 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. 24, 4420–4449.
Hooper, S.D., Berg, O.G., 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol. Biol. Evol. 20, 945–954. https://doi.org/10.1093/molbev/msg101
Huang, C.H., Hsiang, T., Trevors, J.T., 2013. Comparative bacterial genomics: defining the minimal core genome. Antonie Van Leeuwenhoek 103, 385–398. https://doi.org/10.1007/s10482-012-9819-7
Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science 351, aad6253. https://doi.org/10.1126/science.aad6253
Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., Venter, J.C., 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. Science 286, 2165–2169.
Ishii, N., 2017. GroEL and the GroEL-GroES Complex. Subcell. Biochem. 83, 483–504. https://doi.org/10.1007/978-3-319-46503-6_17
Jacob, F., Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356.
Jacobs, G., Dechyeva, D., Menzel, G., Dombrowski, C., Schmidt, T., 2004. Molecular characterization of Vulmar1, a complete mariner transposon of sugar beet and diversity of mariner- and En/Spm-like sequences in the genus Beta. Genome 47, 1192–1201. https://doi.org/10.1139/g04-067
Jacobson, J.W., Medhora, M.M., Hartl, D.L., 1986. Molecular structure of a somatically unstable transposable element in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 83, 8684–8688. https://doi.org/10.1073/pnas.83.22.8684
Jeffery, C., 2018. Intracellular proteins moonlighting as bacterial adhesion factors. AIMS Microbiol. 4, 362–376. https://doi.org/10.3934/microbiol.2018.2.362
Jensen, J.S., Hansen, H.T., Lind, K., 1996. Isolation of Mycoplasma genitalium strains from the male urethra. J. Clin. Microbiol. 34, 286–291.
Jerison, E.R., Desai, M.M., 2015. Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. Curr. Opin. Genet. Dev. 35, 33–39. https://doi.org/10.1016/j.gde.2015.08.008
Jiang, F., Doudna, J.A., 2017. CRISPR-Cas9 Structures and Mechanisms. Annu. Rev. Biophys. 46, 505–529. https://doi.org/10.1146/annurev-biophys-062215-010822

Johnsborg, O., Eldholm, V., Håvarstein, L.S., 2007. Natural genetic transformation: prevalence, mechanisms and function. Res. Microbiol. 158, 767–778. https://doi.org/10.1016/j.resmic.2007.09.004
Jordan, I.K., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12, 962–968. https://doi.org/10.1101/gr.87702
Joyce, A.R., Reed, J.L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S.A., Palsson, B.Ø., Agarwalla, S., 2006. Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli. J. Bacteriol. 188, 8259–8271. https://doi.org/10.1128/JB.00740-06
Juhas, M., Eberl, L., Glass, J.I., 2011a. Essence of life: essential genes of minimal genomes. Trends Cell Biol. 21, 562–568. https://doi.org/10.1016/j.tcb.2011.07.005
Juhas, M., Eberl, L., Glass, J.I., 2011b. Essence of life: essential genes of minimal genomes. Trends Cell Biol. 21, 562–568. https://doi.org/10.1016/j.tcb.2011.07.005
Junier, I., Unal, E.B., Yus, E., Lloréns-Rico, V., Serrano, L., 2016. Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium. Cell Syst. 2, 391–401. https://doi.org/10.1016/j.cels.2016.04.015
Kahng, L.S., Shapiro, L., 2001. The CcrM DNA methyltransferase of Agrobacterium tumefaciens is essential, and its activity is cell cycle regulated. J. Bacteriol. 183, 3065–3075. https://doi.org/10.1128/JB.183.10.3065-3075.2001
Kainulainen, V., Korhonen, T.K., 2014. Dancing to Another Tune—Adhesive Moonlighting Proteins in Bacteria. Biology 3, 178–204. https://doi.org/10.3390/biology3010178
Kans, J., 2019. Entrez Direct: E-utilities on the UNIX Command Line. National Center for Biotechnology Information (US).
Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O., Balikó, G., Szappanos, B., Györfy, Z., Fehér, T., Bogos, B., Blattner, F.R., Pál, C., Pósfai, G., Papp, B., 2016a. Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. Mol. Biol. Evol. 33, 1257–1269. https://doi.org/10.1093/molbev/msw009
Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O., Balikó, G., Szappanos, B., Györfy, Z., Fehér, T., Bogos, B., Blattner, F.R., Pál, C., Pósfai, G., Papp, B., 2016b. Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. Mol. Biol. Evol. 33, 1257–1269. https://doi.org/10.1093/molbev/msw009
Katsir, L., Zhepu, R., Piasezky, A., Jiang, J., Sela, N., Freilich, S., Bahar, O., 2018. Genome Sequence of "Candidatus Carsonella ruddii" Strain BT from the Psyllid Bactericera trigonica. Genome Announc. 6. https://doi.org/10.1128/genomeA.01466-17
Klasson, L., Andersson, S.G.E., 2004. Evolution of minimal-gene-sets in host-dependent bacteria. Trends Microbiol. 12, 37–43.
Kleckner, N., Fisher, J.K., Stouf, M., White, M.A., Bates, D., Witz, G., 2014. The Bacterial Nucleoid: Nature, Dynamics and Sister Segregation. Curr. Opin. Microbiol. 22, 127–137. https://doi.org/10.1016/j.mib.2014.10.001
Klein, B.A., Duncan, M.J., Hu, L.T., 2015. Defining essential genes and identifying virulence factors of Porphyromonas gingivalis by massively-parallel sequencing of transposon libraries (Tn-seq). Methods Mol. Biol. Clifton NJ 1279, 25–43. https://doi.org/10.1007/978-1-4939-2398-4_3

Klobucar, K., Brown, E.D., 2018. Use of genetic and chemical synthetic lethality as probes of complexity in bacterial cell systems. FEMS Microbiol. Rev. 42. https://doi.org/10.1093/femsre/fux054

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Débarbouille, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Le Coq, D., Masson, A., Mauël, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F.M.L., Sekiguchi, J., Sekowska, A., Séror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., van Dijl, J.M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U., Ogasawara, N., 2003. Essential Bacillus subtilis genes. Proc. Natl. Acad. Sci. U. S. A. 100, 4678–4683. https://doi.org/10.1073/pnas.0730515100

Koch, A., Mizrahi, V., 2018. Mycobacterium tuberculosis. Trends Microbiol. 26, 555–556. https://doi.org/10.1016/j.tim.2018.02.012

Kolhi, S., Kolaskar, A.S., 2012. Categorization of metabolome in bacterial systems. Bioinformation 8, 309–315. https://doi.org/10.6026/97320630008309

Konstantinidis, K.T., Ramette, A., Tiedje, J.M., 2006. The bacterial species definition in the genomic era. Philos. Trans. R. Soc. B Biol. Sci. 361, 1929. https://doi.org/10.1098/rstb.2006.1920

Koonin, E.V., 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. 1, 127–136. https://doi.org/10.1038/nrmicro751

Koonin, E.V., 2000. How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. 1, 99–116. https://doi.org/10.1146/annurev.genom.1.1.99

Koskela, M., Annila, A., 2012. Looking for the Last Universal Common Ancestor (LUCA). Genes 3, 81–87. https://doi.org/10.3390/genes3010081

Krause, D.C., Balish, M.F., 2004. Cellular engineering in a minimal microbe: structure and assembly of the terminal organelle of Mycoplasma pneumoniae. Mol. Microbiol. 51, 917–924.

Kühn, R., Torres, R.M., 2002. Cre/loxP recombination system and gene targeting. Methods Mol. Biol. Clifton NJ 180, 175–204. https://doi.org/10.1385/1-59259-178-7:175

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castaño-Diez, D., Chen, W.-H., Devos, D., Güell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Böttcher, B., Frangakis, A.S., Russell, R.B., Serrano, L., Bork, P., Gavin, A.-C., 2009. Proteome organization in a genome-reduced bacterium. Science 326, 1235–1240. https://doi.org/10.1126/science.1176343

Kumar, C.M.S., Mande, S.C., Mahajan, G., 2015. Multiple chaperonins in bacteria--novel functions and non-canonical behaviors. Cell Stress Chaperones 20, 555–574. https://doi.org/10.1007/s12192-015-0598-8

Kyuma, T., Kizaki, H., Ryuno, H., Sekimizu, K., Kaito, C., 2015. 16S rRNA methyltransferase KsgA contributes to oxidative stress resistance and virulence in Staphylococcus aureus. Biochimie 119, 166–174. https://doi.org/10.1016/j.biochi.2015.10.027

Lal, P.B., Schneider, B.L., Vu, K., Reitzer, L., 2014. The redundant aminotransferases in lysine and arginine synthesis and the extent of aminotransferase redundancy in Escherichia coli. Mol. Microbiol. 94, 843–856. https://doi.org/10.1111/mmi.12801

Lamichhane, G., Zignol, M., Blades, N.J., Geiman, D.E., Dougherty, A., Grosset, J., Broman, K.W., Bishai, W.R., 2003. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. U. S. A. 100, 7213–7218. https://doi.org/10.1073/pnas.1231432100

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., Ussery, D.W., 2015. Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genomics 15, 141–161. https://doi.org/10.1007/s10142-015-0433-4

Lartigue, C., Vashee, S., Algire, M.A., Chuang, R.-Y., Benders, G.A., Ma, L., Noskov, V.N., Denisova, E.A., Gibson, D.G., Assad-Garcia, N., Alperovich, N., Thomas, D.W., Merryman, C., Hutchison, C.A., Smith, H.O., Venter, J.C., Glass, J.I., 2009. Creating bacterial strains from genomes that have been cloned and engineered in yeast. Science 325, 1693–1696. https://doi.org/10.1126/science.1173759

Le Breton, Y., Belew, A.T., Valdes, K.M., Islam, E., Curry, P., Tettelin, H., Shirtliff, M.E., El-Sayed, N.M., McIver, K.S., 2015. Essential Genes in the Core Genome of the Human Pathogen Streptococcus pyogenes. Sci. Rep. 5, 9838. https://doi.org/10.1038/srep09838

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J., 2011. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics 12, 124. https://doi.org/10.1186/1471-2105-12-124

Leibig, M., Krismer, B., Kolb, M., Friede, A., Götz, F., Bertram, R., 2008. Marker Removal in Staphylococci via Cre Recombinase and Different lox Sites. Appl. Environ. Microbiol. 74, 1316–1323. https://doi.org/10.1128/AEM.02424-07

Lin, T., Troy, E.B., Hu, L.T., Gao, L., Norris, S.J., 2014. Transposon mutagenesis as an approach to improved understanding of Borrelia pathogenesis and biology. Front. Cell. Infect. Microbiol. 4. https://doi.org/10.3389/fcimb.2014.00063

Lin, Y., Zhang, R.R., 2011. Putative essential and core-essential genes in Mycoplasma genomes. Sci. Rep. 1. https://doi.org/10.1038/srep00053

Lis, R., Rowhani-Rahbar, A., Manhart, L.E., 2015. Mycoplasma genitalium infection and female reproductive tract disease: a meta-analysis. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. 61, 418–426. https://doi.org/10.1093/cid/civ312

Liu, X., Wang, B., Xu, L., 2015. Statistical Analysis of Hurst Exponents of Essential/Nonessential Genes in 33 Bacterial Genomes. PloS One 10, e0129716. https://doi.org/10.1371/journal.pone.0129716

Lluch-Senar, M., Cozzuto, L., Cano, J., Delgado, J., Llórens-Rico, V., Pereyre, S., Bebear, C., Serrano, L., 2015a. Comparative “-omics” in Mycoplasma

pneumoniae Clinical Isolates Reveals Key Virulence Factors. PloS One 10, e0137354. https://doi.org/10.1371/journal.pone.0137354
Lluch-Senar, M., Delgado, J., Chen, W.-H., Lloréns-Rico, V., O'Reilly, F.J., Wodke, J.A., Unal, E.B., Yus, E., Martínez, S., Nichols, R.J., Ferrar, T., Vivancos, A., Schmeisky, A., Stülke, J., van Noort, V., Gavin, A.-C., Bork, P., Serrano, L., 2015b. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Mol. Syst. Biol. 11. https://doi.org/10.15252/msb.20145558
Lluch-Senar, M., Querol, E., Piñol, J., 2010. Cell division in a minimal bacterium in the absence of ftsZ. Mol. Microbiol. 78, 278–289. https://doi.org/10.1111/j.1365-2958.2010.07306.x
Locey, K.J., Lennon, J.T., 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. U. S. A. 113, 5970–5975. https://doi.org/10.1073/pnas.1521291113
Luo, H., Gao, F., Lin, Y., 2015. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci. Rep. 5, 13210. https://doi.org/10.1038/srep13210
Lwoff, A., 1967. Principles of classification and nomenclature of viruses. Nature 215, 13–14. https://doi.org/10.1038/215013a0
Lyell, N.L., Dunn, A.K., Bose, J.L., Vescovi, S.L., Stabb, E.V., 2008. Effective mutagenesis of Vibrio fischeri by using hyperactive mini-Tn5 derivatives. Appl. Environ. Microbiol. 74, 7059–7063. https://doi.org/10.1128/AEM.01330-08
Lyon, B.R., May, J.W., Skurray, R.A., 1984. Tn4001: a gentamicin and kanamycin resistance transposon in Staphylococcus aureus. Mol. Gen. Genet. MGG 193, 554–556. https://doi.org/10.1007/bf00382099
MacDonald, I.C., Deans, T.L., 2016. Tools and applications in synthetic biology. Adv. Drug Deliv. Rev. 105, 20–34. https://doi.org/10.1016/j.addr.2016.08.008
Mackie, G.A., 2013. RNase E: at the interface of bacterial RNA processing and decay. Nat. Rev. Microbiol. 11, 45–57. https://doi.org/10.1038/nrmicro2930
Maier, T., Marcos, J., Wodke, J.A.H., Paetzold, B., Liebeke, M., Gutiérrez-Gallego, R., Serrano, L., 2013. Large-scale metabolome analysis and quantitative integration with genomics and proteomics data in Mycoplasma pneumoniae. Mol. Biosyst. 9, 1743–1755. https://doi.org/10.1039/c3mb70113a
Marbach, A., Bettenbrock, K., 2012. lac operon induction in Escherichia coli: Systematic comparison of IPTG and TMG induction and influence of the transacetylase LacA. J. Biotechnol. 157, 82–88. https://doi.org/10.1016/j.jbiotec.2011.10.009
McKay, C.P., 2004. What Is Life—and How Do We Search for It in Other Worlds? PLoS Biol. 2. https://doi.org/10.1371/journal.pbio.0020302
Mell, J.C., Redfield, R.J., 2014. Natural Competence and the Evolution of DNA Uptake Specificity. J. Bacteriol. 196, 1471–1483. https://doi.org/10.1128/JB.01293-13
Mendonça, A.G., Alves, R.J., Pereira-Leal, J.B., 2011. Loss of Genetic Redundancy in Reductive Genome Evolution. PLoS Comput. Biol. 7. https://doi.org/10.1371/journal.pcbi.1001082
Mikusová, K., Yagi, T., Stern, R., McNeil, M.R., Besra, G.S., Crick, D.C., Brennan, P.J., 2000. Biosynthesis of the galactan component of the mycobacterial cell wall. J. Biol. Chem. 275, 33890–33897. https://doi.org/10.1074/jbc.M006875200
Missirlis, P.I., Smailus, D.E., Holt, R.A., 2006. A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in Cre-

mediated recombination. BMC Genomics 7, 73. https://doi.org/10.1186/1471-2164-7-73
Moe-Behrens, G.H.G., Davis, R., Haynes, K.A., 2013. Preparing synthetic biology for the world. Front. Microbiol. 4. https://doi.org/10.3389/fmicb.2013.00005
Montero-Blay, A., Miravet-Verde, S., Lluch-Senar, M., Piñero-Lambea, C., Serrano, L., n.d. SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes. DNA Res. https://doi.org/10.1093/dnares/dsz012
Moreira, D., López-García, P., 2009. Ten reasons to exclude viruses from the tree of life. Nat. Rev. Microbiol. 7, 306–311. https://doi.org/10.1038/nrmicro2108
Morgan, R.D., Dwinell, E.A., Bhatia, T.K., Lang, E.M., Luyten, Y.A., 2009. The MmeI family: type II restriction-modification enzymes that employ single-strand modification for host protection. Nucleic Acids Res. 37, 5208–5221. https://doi.org/10.1093/nar/gkp534
Mori, H., Baba, T., Yokoyama, K., Takeuchi, R., Nomura, W., Makishi, K., Otsuka, Y., Dose, H., Wanner, B.L., 2015. Identification of essential genes and synthetic lethal gene combinations in Escherichia coli K-12. Methods Mol. Biol. Clifton NJ 1279, 45–65. https://doi.org/10.1007/978-1-4939-2398-4_4
Morowitz, H.J., 1984. The completeness of molecular biology. Isr. J. Med. Sci. 20, 750–753.
Moule, M.G., Hemsley, C.M., Seet, Q., Guerra-Assunção, J.A., Lim, J., Sarkar-Tyson, M., Clark, T.G., Tan, P.B.O., Titball, R.W., Cuccui, J., Wren, B.W., 2014. Genome-Wide Saturation Mutagenesis of Burkholderia pseudomallei K96243 Predicts Essential Genes and Novel Targets for Antimicrobial Development. mBio 5. https://doi.org/10.1128/mBio.00926-13
Moura de Sousa, J., Balbontín, R., Durão, P., Gordo, I., 2017. Multidrug-resistant bacteria compensate for the epistasis between resistances. PLoS Biol. 15. https://doi.org/10.1371/journal.pbio.2001741
Muñoz-López, M., García-Pérez, J.L., 2010. DNA Transposons: Nature and Applications in Genomics. Curr. Genomics 11, 115–128. https://doi.org/10.2174/138920210790886871
Mus, F., Crook, M.B., Garcia, K., Garcia Costas, A., Geddes, B.A., Kouri, E.D., Paramasivan, P., Ryu, M.-H., Oldroyd, G.E.D., Poole, P.S., Udvardi, M.K., Voigt, C.A., Ané, J.-M., Peters, J.W., 2016. Symbiotic Nitrogen Fixation and the Challenges to Its Extension to Nonlegumes. Appl. Environ. Microbiol. 82, 3698–3710. https://doi.org/10.1128/AEM.01055-16
Mushegian, A., 2008. Gene content of LUCA, the last universal common ancestor. Front. Biosci. J. Virtual Libr. 13, 4657–4666.
Mushegian, A., 1999. The minimal genome concept. Curr. Opin. Genet. Dev. 9, 709–714.
Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314, 267. https://doi.org/10.1126/science.1134196
Nakane, D., Kenri, T., Matsuo, L., Miyata, M., 2015. Systematic Structural Analyses of Attachment Organelle in Mycoplasma pneumoniae. PLoS Pathog. 11, e1005299. https://doi.org/10.1371/journal.ppat.1005299
Naorem, S.S., Han, J., Zhang, S.Y., Zhang, J., Graham, L.B., Song, A., Smith, C.V., Rashid, F., Guo, H., 2018. Efficient transposon mutagenesis mediated by an IPTG-controlled conditional suicide plasmid. BMC Microbiol. 18, 158. https://doi.org/10.1186/s12866-018-1319-0

Niki, H., Yamaichi, Y., Hiraga, S., 2000. Dynamic organization of chromosomal DNA in Escherichia coli. Genes Dev. 14, 212–223.
Nikulin, A.D., 2018. Structural Aspects of Ribosomal RNA Recognition by Ribosomal Proteins. Biochem. Biokhimiia 83, S111–S133. https://doi.org/10.1134/S0006297918140109
Oh-McGinnis, R., Jones, M.J., Lefebvre, L., 2010. Applications of the site-specific recombinase Cre to the study of genomic imprinting. Brief. Funct. Genomics 9. https://doi.org/10.1093/bfgp/elq017
Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M., Goto, S., 2007. Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli. BMC Genomics 8, 48. https://doi.org/10.1186/1471-2164-8-48
Pál, C., Papp, B., Lercher, M.J., 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat. Genet. 37, 1372–1375. https://doi.org/10.1038/ng1686
Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004. https://doi.org/10.1038/nbt.4229
Parrott, G.L., Kinjo, T., Fujita, J., 2016. A Compendium for Mycoplasma pneumoniae. Front. Microbiol. 7, 513. https://doi.org/10.3389/fmicb.2016.00513
Pasqua, M., Visaggio, D., Lo Sciuto, A., Genah, S., Banin, E., Visca, P., Imperi, F., 2017. Ferric Uptake Regulator Fur Is Conditionally Essential in Pseudomonas aeruginosa. J. Bacteriol. 199. https://doi.org/10.1128/JB.00472-17
Passalacqua, K.D., Charbonneau, M.-E., O'Riordan, M.X.D., 2016. Bacterial Metabolism Shapes the Host-Pathogen Interface. Microbiol. Spectr. 4. https://doi.org/10.1128/microbiolspec.VMBF-0027-2015
Paul, B.G., Burstein, D., Castelle, C.J., Handa, S., Arambula, D., Czornyj, E., Thomas, B.C., Ghosh, P., Miller, J.F., Banfield, J.F., Valentine, D.L., 2017. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat. Microbiol. 2, 17045. https://doi.org/10.1038/nmicrobiol.2017.45
Pearson, W.R., 2013. An introduction to sequence similarity ("homology") searching. Curr. Protoc. Bioinforma. Chapter 3, Unit3.1. https://doi.org/10.1002/0471250953.bi0301s42
Peng, C., Lin, Y., Luo, H., Gao, F., 2017. A Comprehensive Overview of Online Resources to Identify and Predict Bacterial Essential Genes. Front. Microbiol. 8, 2331. https://doi.org/10.3389/fmicb.2017.02331
Pettersson, M.E., Sun, S., Andersson, D.I., Berg, O.G., 2009. Evolution of new gene functions: simulation and analysis of the amplification model. Genetica 135, 309–324. https://doi.org/10.1007/s10709-008-9289-z
Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J.C.D., Vogel, J., 2009. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. Nat. Struct. Mol. Biol. 16, 840–846. https://doi.org/10.1038/nsmb.1631
Pich, O.Q., Burgos, R., Planell, R., Querol, E., Piñol, J., 2006. Comparative analysis of antibiotic resistance gene markers in Mycoplasma genitalium: application to studies of the minimal gene complement. Microbiol. Read. Engl. 152, 519–527. https://doi.org/10.1099/mic.0.28287-0

Piñero-Lambea, C., Ruano-Gallego, D., Fernández, L.Á., 2015. Engineered bacteria as therapeutic agents. Curr. Opin. Biotechnol. 35, 94–102. https://doi.org/10.1016/j.copbio.2015.05.004
Pines, G., Freed, E.F., Winkler, J.D., Gill, R.T., 2015. Bacterial Recombineering: Genome Engineering via Phage-Based Homologous Recombination. ACS Synth. Biol. 4, 1176–1185. https://doi.org/10.1021/acssynbio.5b00009
Pinkney, J.N.M., Zawadzki, P., Mazuryk, J., Arciszewska, L.K., Sherratt, D.J., Kapanidis, A.N., 2012. Capturing reaction paths and intermediates in Cre-loxP recombination using single-molecule fluorescence. Proc. Natl. Acad. Sci. U. S. A. 109, 20871–20876. https://doi.org/10.1073/pnas.1211922109
Plasterk, R.H.A., Izsvák, Z., Ivics, Z., 1999. Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet. 15, 326–332. https://doi.org/10.1016/S0168-9525(99)01777-1
Poddighe, D., 2018. Extra-pulmonary diseases related to Mycoplasma pneumoniae in children: recent insights into the pathogenesis. Curr. Opin. Rheumatol. 30, 380–387. https://doi.org/10.1097/BOR.0000000000000494
Pour-El, I., Adams, C., Minion, F.C., 2002. Construction of mini-Tn4001tet and its use in Mycoplasma gallisepticum. Plasmid 47, 129–137. https://doi.org/10.1006/plas.2001.1558
Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V., Melnyk, R.A., Lamson, J.S., Suh, Y., Carlson, H.K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G.M., Rubin, B.E., Wall, J.D., Visel, A., Bristow, J., Blow, M.J., Arkin, A.P., Deutschbauer, A.M., 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature 557, 503–509. https://doi.org/10.1038/s41586-018-0124-0
Prudhomme, M., Turlan, C., Claverys, J.-P., Chandler, M., 2002. Diversity of Tn4001 transposition products: the flanking IS256 elements can form tandem dimers and IS circles. J. Bacteriol. 184, 433–443. https://doi.org/10.1128/jb.184.2.433-443.2002
Quereda, J.J., Cossart, P., 2017. Regulating Bacterial Virulence with RNA. Annu. Rev. Microbiol. 71, 263–280. https://doi.org/10.1146/annurev-micro-030117-020335
Ramakrishnan, V., 2002. Ribosome structure and the mechanism of translation. Cell 108, 557–572. https://doi.org/10.1016/s0092-8674(02)00619-0
Ramsey, J.S., MacDonald, S.J., Jander, G., Nakabachi, A., Thomas, G.H., Douglas, A.E., 2010. Genomic evidence for complementary purine metabolism in the pea aphid, Acyrthosiphon pisum, and its symbiotic bacterium Buchnera aphidicola. Insect Mol. Biol. 19 Suppl 2, 241–248. https://doi.org/10.1111/j.1365-2583.2009.00945.x
Rattanachaikunsopon, P., Phumkhachorn, P., 2009. Glass bead transformation method for gram-positive bacteria. Braz. J. Microbiol. 40, 923–926. https://doi.org/10.1590/S1517-838220090004000025
Ray, B.K., Apirion, D., 1979. Characterization of 10S RNA: a new stable rna molecule from Escherichia coli. Mol. Gen. Genet. MGG 174, 25–32. https://doi.org/10.1007/bf00433301
Razin, S., 1985. Molecular biology and genetics of mycoplasmas (Mollicutes). Microbiol. Rev. 49, 419–455.
Razin, S., Hayflick, L., 2010. Highlights of mycoplasma research--an historical perspective. Biol. J. Int. Assoc. Biol. Stand. 38, 183–190. https://doi.org/10.1016/j.biologicals.2009.11.008

Razin, S., Yogev, D., Naot, Y., 1998. Molecular Biology and Pathogenicity of Mycoplasmas. Microbiol. Mol. Biol. Rev. 62, 1094–1156.
Reams, A.B., Kofoid, E., Savageau, M., Roth, J.R., 2010. Duplication frequency in a population of Salmonella enterica rapidly approaches steady state with or without recombination. Genetics 184, 1077–1094. https://doi.org/10.1534/genetics.109.111963
Reams, A.B., Roth, J.R., 2015. Mechanisms of Gene Duplication and Amplification. Cold Spring Harb. Perspect. Biol. 7. https://doi.org/10.1101/cshperspect.a016592
Reddy, S.P., Rasmussen, W.G., Baseman, J.B., 1996. Isolation and characterization of transposon Tn4001-generated, cytadherence-deficient transformants of Mycoplasma pneumoniae and Mycoplasma genitalium. FEMS Immunol. Med. Microbiol. 15, 199–211.
Regula, J.T., Ueberle, B., Boguth, G., Görg, A., Schnölzer, M., Herrmann, R., Frank, R., 2000. Towards a two-dimensional proteome map of Mycoplasma pneumoniae. Electrophoresis 21, 3765–3780. https://doi.org/10.1002/1522-2683(200011)21:17<3765::AID-ELPS3765>3.0.CO;2-6
Reuß, D.R., Altenbuchner, J., Mäder, U., Rath, H., Ischebeck, T., Sappa, P.K., Thürmer, A., Guérin, C., Nicolas, P., Steil, L., Zhu, B., Feussner, I., Klumpp, S., Daniel, R., Commichau, F.M., Völker, U., Stülke, J., 2017. Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism. Genome Res. 27, 289–299. https://doi.org/10.1101/gr.215293.116
Reuß, D.R., Commichau, F.M., Gundlach, J., Zhu, B., Stülke, J., 2016. The Blueprint of a Minimal Cell: MiniBacillus. Microbiol. Mol. Biol. Rev. MMBR 80, 955–987. https://doi.org/10.1128/MMBR.00029-16
Reznikoff, W.S., 2008. Transposon Tn5. Annu. Rev. Genet. 42, 269–286. https://doi.org/10.1146/annurev.genet.42.110807.091656
Robertson, H.M., Zumpano, K.L., 1997. Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. Gene 205, 203–217.
Robertson, J., Gomersall, M., Gill, P., 1975. Mycoplasma hominis: growth, reproduction, and isolation of small viable cells. J. Bacteriol. 124, 1007–1018.
Rocha, E.P.C., 2008. The organization of the bacterial genome. Annu. Rev. Genet. 42, 211–233. https://doi.org/10.1146/annurev.genet.42.110807.091653
Rosales, R.S., Puleio, R., Loria, G.R., Catania, S., Nicholas, R.A.J., 2017. Mycoplasmas: Brain invaders? Res. Vet. Sci. 113, 56–61. https://doi.org/10.1016/j.rvsc.2017.09.006
Rosconi, F., de Vries, S.P.W., Baig, A., Fabiano, E., Grant, A.J., 2016. Essential Genes for In Vitro Growth of the Endophyte Herbaspirillum seropedicae SmR1 as Revealed by Transposon Insertion Site Sequencing. Appl. Environ. Microbiol. 82, 6664–6671. https://doi.org/10.1128/AEM.02281-16
Roth, V., Aigle, B., Bunet, R., Wenner, T., Fourrier, C., Decaris, B., Leblond, P., 2004. Differential and cross-transcriptional control of duplicated genes encoding alternative sigma factors in Streptomyces ambofaciens. J. Bacteriol. 186, 5355–5365. https://doi.org/10.1128/JB.186.16.5355-5365.2004
Ruiz, N., Kahne, D., Silhavy, T.J., 2006. Advances in understanding bacterial outer-membrane biogenesis. Nat. Rev. Microbiol. 4, 57–66. https://doi.org/10.1038/nrmicro1322
Ruiz-Saenz, J., Rodas, J.D., 2010. Viruses, virophages, and their living nature. Acta Virol. 54, 85–90.

Ryabova, N.A., Marchenkov, V.V., Marchenkova, S.Y., Kotova, N.V., Semisotnov, G.V., 2013. Molecular chaperone GroEL/ES: unfolding and refolding processes. Biochem. Biokhimiia 78, 1405–1414. https://doi.org/10.1134/S0006297913130038

Sackman, A.M., Rokyta, D.R., 2018. Additive Phenotypes Underlie Epistasis of Fitness Effects. Genetics 208, 339–348. https://doi.org/10.1534/genetics.117.300451

Sáenz-Lahoya, S., Bitarte, N., García, B., Burgui, S., Vergara-Irigaray, M., Valle, J., Solano, C., Toledo-Arana, A., Lasa, I., 2019. Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. Proc. Natl. Acad. Sci. U. S. A. 116, 1733–1738. https://doi.org/10.1073/pnas.1812746116

Santra, M., Farrell, D.W., Dill, K.A., 2017. Bacterial proteostasis balances energy and chaperone utilization efficiently. Proc. Natl. Acad. Sci. U. S. A. 114, E2654–E2661. https://doi.org/10.1073/pnas.1620646114

Sassetti, C.M., Boyd, D.H., Rubin, E.J., 2001. Comprehensive identification of conditionally essential genes in mycobacteria. Proc. Natl. Acad. Sci. U. S. A. 98, 12712–12717. https://doi.org/10.1073/pnas.231275498

Schmidl, S.R., Gronau, K., Pietack, N., Hecker, M., Becher, D., Stülke, J., 2010. The phosphoproteome of the minimal bacterium Mycoplasma pneumoniae: analysis of the complete known Ser/Thr kinome suggests the existence of novel kinases. Mol. Cell. Proteomics MCP 9, 1228–1242. https://doi.org/10.1074/mcp.M900267-MCP200

Schulz, F., Yutin, N., Ivanova, N.N., Ortega, D.R., Lee, T.K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G.J., Kyrpides, N.C., Koonin, E.V., Woyke, T., 2017. Giant viruses with an expanded complement of translation system components. Science 356, 82–85. https://doi.org/10.1126/science.aal4657

Sela, I., Wolf, Y.I., Koonin, E.V., 2016. Theory of prokaryotic genome evolution. Proc. Natl. Acad. Sci. U. S. A. 113, 11399–11407. https://doi.org/10.1073/pnas.1614083113

Shen, Y., Stracquadanio, G., Wang, Y., Yang, K., Mitchell, L.A., Xue, Y., Cai, Y., Chen, T., Dymond, J.S., Kang, K., Gong, J., Zeng, X., Zhang, Y., Li, Y., Feng, Q., Xu, X., Wang, Jun, Wang, Jian, Yang, H., Boeke, J.D., Bader, J.S., 2016. SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. Genome Res. 26, 36–49. https://doi.org/10.1101/gr.193433.115

Sobral, R.G., Ludovice, A.M., de Lencastre, H., Tomasz, A., 2006. Role of murF in Cell Wall Biosynthesis: Isolation and Characterization of a murF Conditional Mutant of Staphylococcus aureus. J. Bacteriol. 188, 2543–2553. https://doi.org/10.1128/JB.188.7.2543-2553.2006

Sorci, L., Ruggieri, S., Raffaelli, N., 2014. NAD homeostasis in the bacterial response to DNA/RNA damage. DNA Repair 23, 17–26. https://doi.org/10.1016/j.dnarep.2014.07.014

Steiniger-White, M., Rayment, I., Reznikoff, W.S., 2004. Structure/function insights into Tn5 transposition. Curr. Opin. Struct. Biol. 14, 50–57. https://doi.org/10.1016/j.sbi.2004.01.008

Sternberg, N., Hamilton, D., Hoess, R., 1981. Bacteriophage P1 site-specific recombination. II. Recombination between loxP and the bacterial chromosome. J. Mol. Biol. 150, 487–507. https://doi.org/10.1016/0022-2836(81)90376-4

Summers, Z.M., Gralnick, J.A., Bond, D.R., 2013. Cultivation of an Obligate Fe(II)-Oxidizing Lithoautotrophic Bacterium Using Electrodes. mBio 4, e00420-12. https://doi.org/10.1128/mBio.00420-12

Suwanto, A., Kaplan, S., 1989. Physical and genetic mapping of the Rhodobacter sphaeroides 2.4.1 genome: presence of two unique circular chromosomes. J. Bacteriol. 171, 5850–5859. https://doi.org/10.1128/jb.171.11.5850-5859.1989
Suzuki, E., Nakayama, M., 2011. VCre/VloxP and SCre/SloxP: new site-specific recombination systems for genome engineering. Nucleic Acids Res. 39, e49. https://doi.org/10.1093/nar/gkq1280
Tamames, J., Gil, R., Latorre, A., Peretó, J., Silva, F.J., Moya, A., 2007. The frontier between cell and organelle: genome analysis of Candidatus Carsonella ruddii. BMC Evol. Biol. 7, 181. https://doi.org/10.1186/1471-2148-7-181
Tanasupawat, S., Takehana, T., Yoshida, S., Hiraga, K., Oda, K., 2016. Ideonella sakaiensis sp. nov., isolated from a microbial consortium that degrades poly(ethylene terephthalate). Int. J. Syst. Evol. Microbiol. 66, 2813–2818. https://doi.org/10.1099/ijsem.0.001058
Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33–36.
Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637. https://doi.org/10.1126/science.278.5338.631
Timmermans, J., Van Melderen, L., 2009. Conditional essentiality of the csrA gene in Escherichia coli. J. Bacteriol. 191, 1722–1724. https://doi.org/10.1128/JB.01573-08
Toussaint, J.-P., Farrell-Sherman, A., Feldman, T.P., Smalley, N.E., Schaefer, A.L., Greenberg, E.P., Dandekar, A.A., 2017. Gene Duplication in Pseudomonas aeruginosa Improves Growth on Adenosine. J. Bacteriol. 199. https://doi.org/10.1128/JB.00261-17
Trachtenberg, S., 2005. Mollicutes. Curr. Biol. CB 15, R483-484. https://doi.org/10.1016/j.cub.2005.06.049
Trussart, M., Yus, E., Martinez, S., Baù, D., Tahara, Y.O., Pengo, T., Widjaja, M., Kretschmer, S., Swoger, J., Djordjevic, S., Turnbull, L., Whitchurch, C., Miyata, M., Marti-Renom, M.A., Lluch-Senar, M., Serrano, L., 2017. Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae. Nat. Commun. 8, 14665. https://doi.org/10.1038/ncomms14665
Tsiodras, S., Kelesidis, I., Kelesidis, T., Stamboulis, E., Giamarellou, H., 2005. Central nervous system manifestations of Mycoplasma pneumoniae infections. J. Infect. 51, 343–354. https://doi.org/10.1016/j.jinf.2005.07.005
Tully, J.G., Taylor-Robinson, D., Cole, R.M., Rose, D.L., 1981. A newly discovered mycoplasma in the human urogenital tract. Lancet Lond. Engl. 1, 1288–1291. https://doi.org/10.1016/s0140-6736(81)92461-2
Turner, K.H., Wessel, A.K., Palmer, G.C., Murray, J.L., Whiteley, M., 2015. Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum. Proc. Natl. Acad. Sci. U. S. A. 112, 4110–4115. https://doi.org/10.1073/pnas.1419677112
Uchiumi, Y., Ohtsuki, H., Sasaki, A., 2019. Evolution of self-limited cell division of symbionts. Proc. Biol. Sci. 286, 20182238. https://doi.org/10.1098/rspb.2018.2238
Valentino, M.D., Foulston, L., Sadaka, A., Kos, V.N., Villet, R.A., Santa Maria, J., Lazinski, D.W., Camilli, A., Walker, S., Hooper, D.C., Gilmore, M.S., 2014. Genes Contributing to Staphylococcus aureus Fitness in Abscess- and Infection-Related Ecologies. mBio 5. https://doi.org/10.1128/mBio.01729-14

Van Duyne, G.D., 2001. A structural view of cre-loxp site-specific recombination. Annu. Rev. Biophys. Biomol. Struct. 30, 87–104. https://doi.org/10.1146/annurev.biophys.30.1.87
van Eijk, E., Wittekoek, B., Kuijper, E.J., Smits, W.K., 2017. DNA replication proteins as potential targets for antimicrobials in drug-resistant bacterial pathogens. J. Antimicrob. Chemother. 72, 1275–1284. https://doi.org/10.1093/jac/dkw548
van Ham, R.C.H.J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J.M., Jiménez, L., Postigo, M., Silva, F.J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Morán, F., Moya, A., 2003. Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. U. S. A. 100, 581–586. https://doi.org/10.1073/pnas.0235981100
van Noort, V., Seebacher, J., Bader, S., Mohammed, S., Vonkova, I., Betts, M.J., Kühner, S., Kumar, R., Maier, T., O'Flaherty, M., Rybin, V., Schmeisky, A., Yus, E., Stülke, J., Serrano, L., Russell, R.B., Heck, A.J.R., Bork, P., Gavin, A.-C., 2012. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. Mol. Syst. Biol. 8, 571. https://doi.org/10.1038/msb.2012.4
van Opijnen, T., Bodi, K.L., Camilli, A., 2009. Tn-seq; high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods 6, 767–772. https://doi.org/10.1038/nmeth.1377
van Opijnen, T., Camilli, A., 2013. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat. Rev. Microbiol. 11. https://doi.org/10.1038/nrmicro3033
van Regenmortel, M.H.V., 2016. The metaphor that viruses are living is alive and well, but it is no more than a metaphor. Stud. Hist. Philos. Biol. Biomed. Sci. 59, 117–124. https://doi.org/10.1016/j.shpsc.2016.02.017
Vickers, C.E., Blank, L.M., Krömer, J.O., 2010. Grand challenge commentary: Chassis cells for industrial biochemical production. Nat. Chem. Biol. 6, 875–877. https://doi.org/10.1038/nchembio.484
Visweswariah, S.S., Busby, S.J.W., 2015. Evolution of bacterial transcription factors: how proteins take on new tasks, but do not always stop doing the old ones. Trends Microbiol. 23, 463–467. https://doi.org/10.1016/j.tim.2015.04.009
Waites, K.B., Talkington, D.F., 2004. Mycoplasma pneumoniae and its role as a human pathogen. Clin. Microbiol. Rev. 17, 697–728, table of contents. https://doi.org/10.1128/CMR.17.4.697-728.2004
Waites, K.B., Xiao, L., Liu, Y., Balish, M.F., Atkinson, T.P., 2017. Mycoplasma pneumoniae from the Respiratory Tract and Beyond. Clin. Microbiol. Rev. 30, 747–809. https://doi.org/10.1128/CMR.00114-16
Wang, G., Xia, Y., Cui, J., Gu, Z., Song, Y., Chen, Y.Q., Chen, H., Zhang, H., Chen, W., 2014. The Roles of Moonlighting Proteins in Bacteria. Curr. Issues Mol. Biol. 16, 15–22.
Wang, N., Ozer, E.A., Mandel, M.J., Hauser, A.R., 2014. Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung. mBio 5, e01163-01114. https://doi.org/10.1128/mBio.01163-14
Watabe, K., Mimuro, M., Tsuchiya, T., 2014. Development of a high-frequency in vivo transposon mutagenesis system for Synechocystis sp. PCC 6803 and Synechococcus elongatus PCC 7942. Plant Cell Physiol. 55, 2017–2026. https://doi.org/10.1093/pcp/pcu128
Watanabe, T., 1994. Effects of manganese on growth of Mycoplasma salivarium and Mycoplasma orale. J. Clin. Microbiol. 32, 1343–1345.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, Rge., Stackebrandt, E., 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int. J. Syst. Evol. Microbiol. 37, 463–464.
Weber, W., Fussenegger, M., 2011. Emerging biomedical applications of synthetic biology. Nat. Rev. Genet. 13, 21–35. https://doi.org/10.1038/nrg3094
Weinreich, D.M., Lan, Y., Wylie, C.S., Heckendorn, R.B., 2013. Should evolutionary geneticists worry about higher-order epistasis? Curr. Opin. Genet. Dev. 23, 700–707. https://doi.org/10.1016/j.gde.2013.10.007
Weinreich, M.D., Gasch, A., Reznikoff, W.S., 1994. Evidence that the cis preference of the Tn5 transposase is caused by nonproductive multimerization. Genes Dev. 8, 2363–2374. https://doi.org/10.1101/gad.8.19.2363
Weisburg, W.G., Tully, J.G., Rose, D.L., Petzel, J.P., Oyaizu, H., Yang, D., Mandelco, L., Sechrest, J., Lawrence, T.G., Van Etten, J., 1989. A phylogenetic analysis of the mycoplasmas: basis for their classification. J. Bacteriol. 171, 6455–6467.
Weiss, M.C., Preiner, M., Xavier, J.C., Zimorski, V., Martin, W.F., 2018. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. PLoS Genet. 14, e1007518. https://doi.org/10.1371/journal.pgen.1007518
Weissenbach, J., Ilhan, J., Bogumil, D., Hülter, N., Stucken, K., Dagan, T., 2017. Evolution of Chaperonin Gene Duplication in Stigonematalean Cyanobacteria (Subsection V). Genome Biol. Evol. 9, 241–252. https://doi.org/10.1093/gbe/evw287
Whiteley, A.T., Pollock, A.J., Portnoy, D.A., 2015. The PAMP c-di-AMP Is Essential for Listeria monocytogenes Growth in Rich but Not Minimal Media due to a Toxic Increase in (p)ppGpp. [corrected]. Cell Host Microbe 17, 788–798. https://doi.org/10.1016/j.chom.2015.05.006
Wodke, J.A.H., Alibés, A., Cozzuto, L., Hermoso, A., Yus, E., Lluch-Senar, M., Serrano, L., Roma, G., 2015. MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae. Nucleic Acids Res. 43, D618. https://doi.org/10.1093/nar/gku1105
Woese, C.R., Maniloff, J., Zablen, L.B., 1980. Phylogenetic analysis of the mycoplasmas. Proc. Natl. Acad. Sci. U. S. A. 77, 494–498. https://doi.org/10.1073/pnas.77.1.494
Wolański, M., Donczew, R., Zawilak-Pawlik, A., Zakrzewska-Czerwińska, J., 2014. oriC-encoded instructions for the initiation of bacterial chromosome replication. Front. Microbiol. 5, 735. https://doi.org/10.3389/fmicb.2014.00735
Wu, J., Du, H., Liao, X., Zhao, Y., Li, L., Yang, L., 2011. Tn5 transposase-assisted transformation of indica rice. Plant J. Cell Mol. Biol. 68, 186–200. https://doi.org/10.1111/j.1365-313X.2011.04663.x
Xavier, J.C., Patil, K.R., Rocha, I., 2014a. Systems Biology Perspectives on Minimal and Simpler Cells. Microbiol. Mol. Biol. Rev. MMBR 78, 487–509. https://doi.org/10.1128/MMBR.00050-13
Xavier, J.C., Patil, K.R., Rocha, I., 2014b. Systems Biology Perspectives on Minimal and Simpler Cells. Microbiol. Mol. Biol. Rev. MMBR 78, 487–509. https://doi.org/10.1128/MMBR.00050-13
Yan, X., Yu, H.-J., Hong, Q., Li, S.-P., 2008. Cre/lox System and PCR-Based Genome Engineering in Bacillus subtilis. Appl. Environ. Microbiol. 74, 5556–5562. https://doi.org/10.1128/AEM.01156-08

Yang, Y.-J., Singh, R.P., Lan, X., Zhang, C.-S., Li, Y.-Z., Li, Y.-Q., Sheng, D.-H., 2018. Genome Editing in Model Strain Myxococcus xanthus DK1622 by a Site-Specific Cre/loxP Recombination System. Biomolecules 8. https://doi.org/10.3390/biom8040137

Yang, Z., Tsui, S.K.-W., 2018. Functional Annotation of Proteins Encoded by the Minimal Bacterial Genome Based on Secondary Structure Element Alignment. J. Proteome Res. 17, 2511–2520. https://doi.org/10.1021/acs.jproteome.8b00262

Yu, J.-F., Chen, Q.-L., Ren, J., Yang, Y.-L., Wang, J.-H., Sun, X., 2015. Analysis of the multi-copied genes and the impact of the redundant protein coding sequences on gene annotation in prokaryotic genomes. J. Theor. Biol. 376, 8–14. https://doi.org/10.1016/j.jtbi.2015.04.002

Yus, E., Lloréns-Rico, V., Martínez, S., Gallo, C., Eilers, H., Blötz, C., Stülke, J., Lluch-Senar, M., Serrano, L., 2019. Determination of the Gene Regulatory Network of a Genome-Reduced Bacterium Highlights Alternative Regulation Independent of Transcription Factors. Cell Syst. 9, 143-158.e13. https://doi.org/10.1016/j.cels.2019.07.001

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., Wodke, J.A.H., Güell, M., Martínez, S., Bourgeois, R., Kühner, S., Raineri, E., Letunic, I., Kalinina, O.V., Rode, M., Herrmann, R., Gutiérrez-Gallego, R., Russell, R.B., Gavin, A.-C., Bork, P., Serrano, L., 2009. Impact of genome reduction on bacterial metabolism and its regulation. Science 326, 1263–1268. https://doi.org/10.1126/science.1177263

Yutin, N., Puigbò, P., Koonin, E.V., Wolf, Y.I., 2012. Phylogenomics of prokaryotic ribosomal proteins. PloS One 7, e36972. https://doi.org/10.1371/journal.pone.0036972

Zhang, J., 2012. Genetic redundancies and their evolutionary maintenance. Adv. Exp. Med. Biol. 751, 279–300. https://doi.org/10.1007/978-1-4614-3567-9_13

Zhang, J.K., Pritchett, M.A., Lampe, D.J., Robertson, H.M., Metcalf, W.W., 2000. In vivo transposon mutagenesis of the methanogenic archaeon Methanosarcina acetivorans C2A using a modified version of the insect mariner-family transposable element Himar1. Proc. Natl. Acad. Sci. U. S. A. 97, 9665–9670. https://doi.org/10.1073/pnas.160272597

Zhang, Y.-C., Cao, W.-J., Zhong, L.-R., Godfray, H.C.J., Liu, X.-D., 2016. Host Plant Determines the Population Size of an Obligate Symbiont (Buchnera aphidicola) in Aphids. Appl. Environ. Microbiol. 82, 2336–2346. https://doi.org/10.1128/AEM.04131-15

Zhang, Y.J., Ioerger, T.R., Huttenhower, C., Long, J.E., Sassetti, C.M., Sacchettini, J.C., Rubin, E.J., 2012. Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. PLoS Pathog. 8, e1002946. https://doi.org/10.1371/journal.ppat.1002946

Zhang, Z., Ren, Q., n.d. Why are essential genes essential? - The essentiality of Saccharomyces genes. Microb. Cell 2, 280–287. https://doi.org/10.15698/mic2015.08.218

Zhou, C., Ma, Q., Li, G., 2014. Elucidation of operon structures across closely related bacterial genomes. PloS One 9, e100999. https://doi.org/10.1371/journal.pone.0100999

Zhou, M.-B., Hu, H., Miskey, C., Lazarow, K., Ivics, Z., Kunze, R., Yang, G., Izsvák, Z., Tang, D.-Q., 2017. Transposition of the bamboo Mariner-like element Ppmar1 in yeast. Mol. Phylogenet. Evol. 109, 367–374. https://doi.org/10.1016/j.ympev.2017.02.005

Zientz, E., Dandekar, T., Gross, R., 2004. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol. Mol. Biol. Rev. MMBR 68, 745–770. https://doi.org/10.1128/MMBR.68.4.745-770.2004

Zimmerly, S., Wu, L., 2015. An Unexplored Diversity of Reverse Transcriptases in Bacteria. Microbiol. Spectr. 3. https://doi.org/10.1128/microbiolspec.MDNA3-0058-2014

Zomer, A., Burghout, P., Bootsma, H.J., Hermans, P.W.M., van Hijum, S.A.F.T., 2012. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. PloS One 7, e43012. https://doi.org/10.1371/journal.pone.0043012