Engineering a Feedback-based Synthetic Gene Circuit for Targeted Continuous Evolution of a Gene in *E. coli*

Haris Mallick

Department of Life Sciences

Imperial College London

Submitted in accordance with the requirements for the degree of

DOCTOR OF PHILOSOPHY

2020

Declarations

Declaration of Originality

I certify that unless otherwise stated, the work within this document is my own. Furthermore, where data or material from other sources has been used, the original author has been credited and all the required measures have been taken to ensure third party copyrights are not infringed.

Copyright declaration

Abstract

Directed evolution is an invaluable technique for engineering proteins to possess desired physical and chemical properties when very little structural and functional information is known. It is divided into two sequential steps: generating a library of protein variants using mutagenic techniques; and applying a screening or selection strategy to scan the library for variants displaying desired properties. Library generation is performed using either *in vitro* or *in vivo* techniques, while screening or selection typically occurs in a suitable host cell. Currently, *in vitro* methods like error-prone PCR are popular for library generation. However, these techniques can be labour intensive, prone to mutation biases, and generate limited library sizes for screening. *In vivo* mutagenic techniques overcome these limitations by enabling simultaneous library generation and selection within cells. By generating random mutations in the gene-of-interest within one cell cycle, each cell in a batch culture potentially represents a library variant. Such a continuous evolution system can run for weeks with minimal human intervention, greatly expanding the genetic search space for protein engineering. The challenge lies in developing a mutator system that specifically generates mutations in the target gene, while maintaining the cell's genomic fidelity. With this goal in mind, a mutator system was engineered in *E. coli* that introduces targeted cytidine deamination damage and subsequently performs error-prone DNA repair by hijacking the base excision repair pathway. The targeted damage occurs via activation induced cytidine deaminase fused to T7 RNA polymerase, while the error-prone DNA repair is performed by a three-protein fusion comprising a 5'-3'-exonuclease, an AP-endonuclease and an error-prone DNA polymerase. The mutagenic characteristics of this system was tested by knocking out GFP expression and analysing the mutant library using next generation sequencing techniques. The system was also experimentally shown to generate functionally active mutations that reverted inactivated β-lactamase gene variants to confer ampicillin resistance.

Acknowledgements

I would like to thank the Biotechnology and Biological Sciences Research Council (BBSRC) for funding this research. Equally, I would like to thank my supervisor, Prof Geoff Baldwin, for the immense amount of support he has provided over the years. Geoff, your guidance and wisdom not only helped to improve my experimental techniques and scientific reasoning, but also helped me to develop a strong work ethic. The skills and sense of professionalism that I have gained from working for you is something that will benefit me for the rest of my career, and for that I am truly grateful. I would also like to thank my second supervisor, Prof Guy-Bart Stan and his entire research group. Guy, the bi-weekly meetings with your group were invaluable as everyone provided me with constructive criticism on my work and helped brainstorm ideas on how I could resolve experimental issues I had along the way. Plus, the Stan Group social events were an enjoyable way to take a break from lab work.

I would also like to thank the members of the Baldwin lab group that have guided me along this four-year journey. Dr Marko Storch, your advice and guidance really helped me become accustomed to working in a laboratory environment, as I had limited lab experience before starting my PhD research. Along the way, you always provided constructive criticism, ways to improve the quality of my work, and taught me to always validate the work I have done, before starting another big experiment. All this training from a sensei like you has resulted in the work that I was able to present in this thesis. I would also like to thank my fellow cohort of doctoral students, Ari, Rebekah and Matt. Sharing all the laughs, stories and experiences along this journey with you guys is definitely something I will cherish. I would also like to thank all the undergraduate and postgraduate students that contributed to this work in their short stints in the Baldwin lab. Alicia G, Nur Y, Joshua L, Eliza A and Lucas C, thank-you for all your hard work, especially with all the characterisation and plating experiments!

Finally, I would like to thank my friends and family that provided the pillar of emotional support and encouragement I needed to keep at it and stay positive regardless of how stressful the work may have gotten. I would especially like to thank my best friend Rohan. When I was struggling to write the Python code needed to analyse my data, he came through for me and offered his pearls of wisdom to get the script working. Abda, Wahab, Humma, Lukshi and Rumsha, a big thanks to all of you for the laughs, words of encouragement, and always having my best interests at heart. Last and certainly not least, I would like to thank you, Mum and Dad. You two have always been there for me, nurtured and motivated me to always push myself, and taught me to fully dedicate myself to the work I am doing. You both always go above and beyond to provide me with the tools I need to succeed. For all the love and support, I will always be grateful.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

+ve – Positive

5’-3’Pol-I Exo – 5’-3’ exonuclease domain from DNA polymerase I

AID – Activation Induced Cytidine Deaminase

Amp^{R} – Ampicillin/Carbenicillin resistance

aTc – Anhydrotetracycline

BER – Base Excision Repair Pathway

bp – Base Pair (unit referring the length of a di-nucleotide sequence)

Cam^{R} – Chloramphenicol resistance

CE – Continuous Evolution

CRISPR – Clustered Regularly Interspersed Short Palindromic Repeats

DE – Directed Evolution

DNA – Deoxyribose Nucleic Acid

DNAse – Deoxyribonuclease

dsDNA – Double stranded DNA

EP-DNA-repair - Error-prone DNA repair

EP-PCR – Error-prone PCR

EP-Pol-I – Error-prone DNA polymerase I

EP-DNA-polymerase – Error-prone DNA polymerase

Exo-III – *E. coli* Exonuclease III

FACS – Fluorescence Activated Cell Sorting

Gen^{R} – Gentamycin resistance

GOI – Gene of interest

IVC – *In Vitro* Compartmentalisation

Kan^{R} – Kanamycin resistance

Kbp – Kilobase pair

MAGE – Multiplex Automated Genome Engineering

MMR – Mismatch DNA Repair Pathway

MS – Mass Spectrometry

NAPE – *N. Meningitis* AP Endonuclease

NMR – Nuclear Magnetic Resonance

nt – nucleotide

PACE – Phage Assisted Continuous Evolution

PAM – Protospacer Adjacent Motif

Pol-I – *E. coli* DNA Polymerase I

Pol-III – *E. coli* DNA polymerase III

Pol-IV – *E. coli* DNA Polymerase IV

P_{T3} – T3 Promoter

P_{T7} – T7 Promoter

RBS – Ribosome Binding Site

Rifamp – Rifampicin antibiotic

RNA – Ribose Nucleic Acid

sbp^{-1} – per sequenced base-pair

$sbp^{-1}g^{-1}$ – per sequenced base-pair per generation

SM – Somatic Hypermutations

spb – Substitutions per base

ssDNA – Single stranded DNA

SSM – Sequence Saturation Mutagenesis

T7pol – T7 RNA polymerase

TetO – Tet Operator

TetR – Tet Repressor

UGI – Uracil DNA Glycosylase Inhibitor

UNG – Uracil-N-glycosylase

WT – Wildtype

-ve – Negative

Definition of Key Terms

1. Biopart: A term used in Synthetic Biology to describe the nucleotide sequence of a gene used for developing a synthetic gene expression circuit.

2. Biological Chassis: A term used in Synthetic Biology used to describe an organism capable of housing and supporting artificial genetic components and circuits, given the necessary resources.

3. Burden: The maintenance and expression of synthetic expression systems places an unnatural load on host cells, commonly referred to as burden. Most burden placed on cells arises from the consumption of finite cellular resources during the expression of synthetic genes.

4. Neutral drift: Neutral drift explores accessible sequence space by repeated rounds of mutagenesis and selection to maintain wild-type function. Mutations that are largely neutral for the native function accumulate, and those that are highly detrimental are purged, yielding a protein library that is highly diverse.

5. Transition: A nucleotide substitution event where a purine is substituted with another purine or a pyrimidine is substituted with a pyrimidine.

6. Transversion: A nucleotide substitution event where a purine is substituted with a pyrimidine or vice-versa.

7. Insertion: A form of genetic manipulation where one or more nucleotides are inserted into a given DNA sequence.

8. Deletion: A form of genetic manipulation where one or more nucleotides are deleted from a given DNA sequence.

9. Sign Epistasis: This occurs when one mutation has the opposite effect in the presence of another mutation. An example of this is when a mutation that is deleterious on its own can enhance the effect of another mutation that is beneficial, providing a growth advantage to the organism.

10. Loss of Function Assay: Genetic manipulation experiments designed to knock out the expression of a detectable phenotype from cells.

11. Gain of Function Assay: Genetic manipulation experiments designed to revert non-functional genes back to the functional, detectable phenotype.

12. Patch Repair: A mechanism of DNA repair where a short patch of nucleotides is excised upstream or downstream of the damaged nucleotide, generating a long single-stranded DNA template for gap filling by a DNA polymerase

13. Directed Evolution: A method of protein engineering which involves generating a library of protein variants using mutagenic techniques and subsequently screening the library for the desired protein function.

14. Phasing: A phenomenon witnessed in sequence-by-synthesis next generation sequencing platforms, where some molecules being sequenced simultaneously in a cluster lose sync with others. Phasing is when molecules are sequenced slower than the remaining clusters, typically due to high GC-content in the DNA.

15. Pre-Phasing: This is the opposite of phasing, where the sequencing of some molecules is faster than the remaining cluster.

16. Circular Consensus Sequence (CCS): During real-time single molecule sequencing on PacBio platforms, a circularised DNA sequence can be polymerised multiple times, generating numerous sub-reads of the same DNA molecule. These subreads are aligned to reduce sequencing error, and generate a consensus sequence, called CCS.

Chapter 1: Introduction

1.1 Boom of Biotechnology and Protein Engineering

Biotechnology is a field centred around studying and utilising living organisms or biological materials for the purpose of producing desirable products. Living cells were being utilised in this fashion for centuries once the usefulness of single-celled organisms like yeast and bacteria was discovered. The ancient Egyptians, for example, used yeast to brew beer and to bake bread. Some 7,000 years ago in Mesopotamia, people used bacteria to convert wine into vinegar. Ancient civilisations rotated crops in agricultural fields with leguminous crops (unaware of their nitrogen fixing properties) to increase future crop yields.

It would not be until the 19th century, that researchers would begin to elucidate the biochemical processes occurring withing the bacteria and yeasts cells, which resulted in such practical uses of these organisms. Louis Pasteur’s research on fermentation showed that biochemical processes occurring inside yeast cells allowed it to convert sugar into ethanol[1,2]. This discovery was extended by Hans and Eduard Buchner, who showed that an enzyme inside yeast cells, called zymase, was responsible and that such enzymes could also function independently in a cell-free extract[3]. Also, in the 19th century, Hermann Hellriegel discovered that leguminous plants can convert atmospheric nitrogen into ammonia by the process of nitrogen fixation[4]. Martinus Willem Beijerinck furthered this research by discovering that the root nodules contain bacteria known as rhizobium which performed the nitrogen fixation[5]. Eventually in the 1960s, it was shown that the enzymes responsible for nitrogen fixation could be extracted from rhizobia and made to function in a cell-free extract.

Vital discoveries made in the mid-20th century enabled researchers to finally understand the intra-cellular biological mechanisms and how these unique functions can be recreated *in vitro* or transferred to another organism. The discovery of genetic information being encoded in discrete units of heredity called DNA[6]; the discovery of DNA’s double helical structure[7]; and the mechanism of translating the genetic code into functional proteins and enzymes, paved the way for understanding gene expression, regulation and protein synthesis[8,9]. In the 1960s the complete library of 64 codons was identified, which translates the genetic information from mRNA into a polypeptide via ribosomes[10–12]. This was followed in the 70s by Frederick Sanger, who developed a DNA sequencing technique to confirm that the ordered sequence of amino acids in a polypeptide is the result of an ordered sequence of nucleotides in the DNA sequence of the corresponding gene[13,14].

During the same time-period, efforts to elucidate the structure and function of DNA was mirrored by the discovery of techniques to manipulate it. Enzymes called restriction endonucleases were identified, possessing the ability to recognise a unique DNA sequence, and cut the double-stranded DNA (dsDNA) to generate sticky ends[15,16]. Boyer and Cohen demonstrated that if two different DNA sequences are cut using the same restriction enzyme, they can anneal to each other via the complimentary sticky ends and produce a single molecule of dsDNA[17].

The discoveries mentioned above played a key role in the development of recombinant DNA technologies, genetic engineering and gave birth to the Biotech industry[18]. DNA recombination is the process by which the genes encoding desired proteins are isolated from their natural environments, modified *in vitro* using molecular cloning techniques and injected into host cells to express the desired protein. Recombinant DNA techniques had an immediate impact in the pharmaceutical and agriculture industries. Gene expression cassettes were assembled in plasmids to produce human insulin[19], antifungal peptides[20], viral antigens[21] in bacterial hosts; and insect- or herbicide-resistant crops were created via genetic modifications[22,23].

As DNA and protein manipulation techniques improved overtime, so did the commercial ambitions of the industry. Researchers began applying genetic and metabolic engineering approaches to transfer entire gene clusters from one organism to another[24,25]; engineering artificial metabolic pathways in host cells for producing desired chemicals, like biofuels and bioplastic[26,27]; and engineering cells to become biosensors for detecting harmful chemicals[28]. A major challenge in these engineered pathways where multiple proteins are expressed, is optimising the expression and catalytic activity of each protein to function in a foreign cellular environment, where the enzyme needs to adjust to new stress factors[29]. For example, in microalgae-based biofuel production, the activity of enzymes in the Kennedy Pathway for generating triacylglycerol (component of biofuel) is affected by stress factors like nitrogen accumulation, nutrient deprivation, ambient temperature and pH[30]. Similarly, when the enzymes are placed in a foreign environment, they may not recognise local substrate molecules or cofactors. In the pathway for producing isobutanol in *E. coli*, the pathway enzymes, ketol-acid reducto-isomerase and alcohol dehydrogenase, require NADPH as a cofactor, but *E. coli* produces NADH via the normal glycolysis metabolic pathway[31]. In other instances, the enzyme needed to perform a chemical reaction within the designed pathway may not exist in nature and must be artificially synthesised. Overcoming these challenges to successfully programme cells to perform the desired function required engineering the proteins themselves[32,33].

Protein engineering involves modifying the peptide sequence of a protein via amino-acid substitutions, deletions, or insertions to achieve a variant protein with the desired properties[34]. These desired properties may include, improved catalytic activity, a change in the substrate specificity, and increased stability to temperature and pH changes. Novel proteins can also be engineered that possess enzymatic properties that do not exist in nature. There are two approaches to protein engineering: a rational approach involving *in silico* protein modelling and site-directed mutagenesis of the protein's DNA sequence; and directed evolution, where a library of mutant DNA sequences are generated and subsequently screened to identify protein variants possessing the desired chemical and physical properties[29].

1.2 Rational Approach to Protein Engineering

The rational approaches rely on utilising the structural and functional information that is known about the protein-of-interest, along with information about other similar proteins, to understand the conserved structure–function relationship and introduce very specific changes in the amino acid sequence of the protein[35–37]. These amino acid substitutions can be performed in the active site or the periphery of the tertiary structure to alter the substrate binding pocket and other structural domains to achieve the desired physical and chemical properties. This rational approach to protein design requires structural data of the protein, acquired via X-ray crystallography, NMR or electron microscopy. This 3D structure is then analysed by a host of software programmes software, like CAVER[38], B-fit[39], FRESCO[40], CASCO[41] and Rosetta[42]. These software programmes create molecular dynamic simulations of the protein and assist in analysing how amino acid substitutions affect protein characteristics, like thermostability, substrate specificity, stereoselectivity, enantioselectivity, cofactor binding and protein folding[36].

These *in silico* simulations generate a computer-aided design of the improved protein. Computational engineering creates large virtual libraries of variants *in silico*. Designs are then evaluated and ranked automatically, e.g., by energy scoring functions or geometric restraints, and only a few hits (10-100 library members) are manually inspected and tested in the lab[43]. Different tools that introduce several mutations at once are used for the creation of the libraries, with mutations often concentrated in the flexible regions of the protein identified via molecular dynamics simulations[36]. Mutations in these flexible regions can affect protein stability and activity. Computational enzyme design allows for engineering highly enantioselective catalysts for a chemical reaction already catalysed by the enzyme and for a complete redesign of active sites to fit substrate structures that are very different from the natural ones[44,45].

These rational approaches to protein design have been applied to engineer numerous proteins, like converting a phenylalanine aminomutase into a phenylalanine ammonia lyase via a single Arg92Ser mutation[46]; engineering a thermostable and efficient PETase enzyme for the degradation of polyethylene terephthalate (PET), which is the most widely used form of plastic[35]; and increasing the catalytic efficiency of the XynCDBFV xylanase at temperatures lower than the normal required temperature of 60°C for enzyme activity[47]. All these examples demonstrate the strength of rational protein design, how strategically changing only one or two amino acids drastically changes the characteristics of a protein.

The major limitation of this approach to protein design is that extensive knowledge of the 3D structure of a protein and its structure–function relationship is needed. For most proteins that have not yet had their 3D structures resolved, rational engineering approaches cannot be applied. Also, rational protein engineering can sometimes yield undesirable results. For example, when trying to extend the range of DNA sequences recognised by the EcoRV restriction endonuclease, mutations introduced resulted in either no functional divergence from the wildtype protein or extreme bias to cleave DNA at TA-flanked sites[37]. Both these experimental results were different from the enzymatic activity predicted from *in silico* designs. Authors concluded that this was due to the lack of 3D structures of the protein in its transition state, preventing the molecular dynamic simulations from accurately predicting conformational changes[48]. This example demonstrates how the predictive power of rational design is limited by our knowledge of protein structures. As more 3D structures are elucidated, of proteins in different conformational states, it will help to improve the scoring function of the design models. But this iterative improvement of algorithms will take time as the number of new 3D structures added to the Protein Data Bank is only 10,000 per year and the number of novel protein folds identified is much lower[49].

1.3 Advent of Directed Evolution

Directed evolution involves performing mutations to a gene or set of genes which are then expressed to create a library of protein variants. These protein variants are screened for the desired function using appropriate screening or selection strategies. After numerous cycles of library generation and screening, once a protein possessing the desired phenotype are identified, it must be linked to the corresponding genotype (i.e, linking peptide sequence to DNA sequence). When a gene evolves naturally within an organism, the phenotype and genotype are intrinsically coupled within each organism, based on Darwinian principles of evolution[50]. If the mutation provides a selective advantage to the organism, it is maintained within the genome of the organism and passed on to the progeny. If the mutation produces deleterious effects, the organism perishes, and the mutant genotype is lost.

Applying this directed evolution technique to protein engineering alleviates the need for 3D structures of proteins and complex computational modelling to predict mutational hot spots to evolve protein function. A method to introduce mutations in the DNA sequence of target proteins and a method to screen the mutants are the only requirements to engineer a protein using this method. Emulating this process of evolution in a laboratory does comes with its own set of challenges.

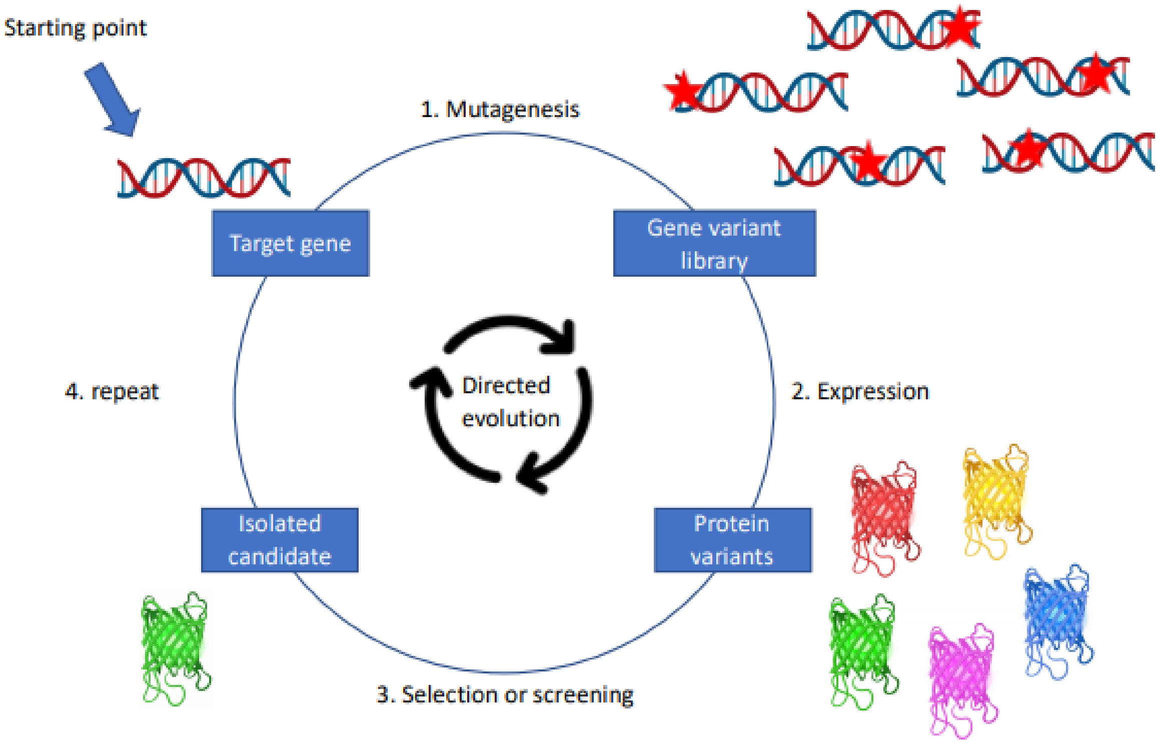


***Figure 1.1**: **General directed evolution workflow**. 1) Directed evolution introduces random mutations (red stars) into the target gene to produce a library of gene variants. 2) The gene variants are expressed in a host. 3) The resulting protein variants undergo screening or selection to isolate variants with desired function. 4) The cycle of mutation and selection is iterated until an isolated variant possesses the desired phenotypic properties. (Figure adapted from Zeymer & Hilvert, 2018)*

1.3.1 First Challenge: Designing a Screening and Selection Method to Identify Functional Proteins

The first challenge when mutating genes in a laboratory environment, is developing a method to identify the desired phenotype from a pool of potentially millions of protein variants tested simultaneously. This requires creating a screening or selection criteria that enables high throughput filtration of the variants based on the desired protein function. Screening of a protein library involves investigating each variant in the library for the desired phenotype. Selection, on the other hand, only looks for functional mutants that improve host cell fitness and the non-functional variants are discarded[51]. Linking the enzymatic function to the fitness of cells is not always possible, therefore screening is the preferred strategy in most cases.

The main hurdle in developing a screening strategy is being able to detect the enzymatic activity of the protein variants. Visual markers, like luminescence, fluorescence or change in optical density as the enzymatic reaction ensues is considered the best strategy for rapidly screening through large protein libraries. Such a screening system is easily set-up for proteins whose enzymatic function results in depletion of substrate or creation of product molecules that can act as fluorophores or chromophores. However, most proteins do not interact with such substrate molecules or generate products with detectable visual signals. Such proteins require designing screening strategies where the enzyme function can be linked to fluorescence or luminescence. If a fluorescent signal is generated, the cells can be sorted using fluorescence activated cell sorting (FACS)[52].

1.3.1.2 High-throughput screening with microfluidics, cell surface display and FACS

FACS coupled with droplet microfluidics[53] or cell surface display[54] greatly increased the throughput of the screening process. These techniques enabled larger libraries of ~ 10^7 variants to be screened for enzymatic activity within a relatively short period of time. Droplet microfluidics involves producing microscale diameter (10-900 µm) droplets of one fluid within a second immiscible carrier fluid (water-oil compartmentalisation). Such femto-, pico- or nanolitre aqueous compartments are called a microreactor[55]. In terms of protein library screening, each microreactor is analogous to a cell expressing one version of the protein, encapsulated with the relevant biomolecules (proteins, DNA and metabolites) to elicit an enzymatic reaction. Within each droplet, a unique enzymatic reaction takes place, which may generate a fluorescent signal. This fluorescent signal enables cell sorting via FACS to identify functional protein variants at rates of around 10^4 cells per second[56].

Cell surface display is a technique that involves creating a protein fusion between the library of protein variants (called passenger protein) with an anchoring motif or transmembrane protein (called carrier protein), such that the protein-of-interest is displayed outside the cell membrane[57]. Once substrate molecules are added to the solution containing the cell-surface displayed protein library, the functional variants perform the desired enzymatic reaction, where either the substrate or product is linked to a fluorescent signal, enabling isolation of functional variants via FACS. Bacterial phage display was applied for screening a library of esterases for enhanced enantioselective properties[58]. The library of esterases displayed on the *E. coli* surface were incubated in solution with tyramide esters of (R)- and (S)-2-methyldecanoic acid, each enantiomer tagged with a different indicator group (2,4 dinitrophenyl and biotin, respectively). Target-enzyme-mediated hydrolysis resulted in tyramide derivaties, which were covalently attached to the cell surface by peroxidase activity. The indicator groups attached to the derivatives, 2,4 dinitrophenyl and biotin, interact with fluorophore-tagged antibodies to generate green and red fluorescence, respectively. Analysing the ratio of the two fluorophores led to identification of mutant esterase with a 15-fold bias towards generating the (R)-2-methyldecanoic acid. The popularity of this screening method prompted development of other protein display techniques, where cells are lysed and instead to anchoring the protein-of-interest to the cell surface, it is anchored to the encoding plasmid[59] (plasmid display), its transcribed mRNA[60] (mRNA display), or to its mRNA and the ribosome[61,62] (ribosome display).

Ribosome display is a powerful cell-free technique where the translated protein is anchored to ribosome and its corresponding mRNA by removal of the stop codon and by the lack of ribosome recycling factors. When immobilised antigens are used to select for desired proteins, the corresponding mRNA can be isolated, reverse-transcribed into DNA and applied to another cycle of evolution. Overall, 10^{14} mRNA molecules can be screened simultaneously using this technique[63].

Both droplet microfluidics and cell surface display, coupled with FACS, greatly increased the throughput of the enzyme library screening process, with 10^7 variants analysed per day. However, both methods have their limitations. In both cases, screening can only be performed if enzyme function is linked to fluorescence output. Designing techniques to create such phenotypic linkage is a laborious process. With microfluidics, once a water-oil droplet is formed, exchanging growth media or performing washing steps is impossible. This restricts the survival time of cells and makes multi-step screening assays challenging[64]. With cell surface display, proteins anchored to the cell surface have shown reduced enzymatic activity. The steric hindrance, incomplete exposure, unfolded or misfolded structure and repulsion of substrate by the hydrophobicity of the cell wall are thought to cause this problem[65,66]. Such factors affect the quality of the protein library that is displayed on the cell surface. The complexities of these high-throughput screening methods can be mitigated by using modern feedback-based selection to identify functional variants from a protein library.

1.3.1.3 Selection Methods: Designing Genetic Circuits for Conditional Expression of Reporters

As protein libraries can contain as many as of 10^7-10^9 variants after each mutagenesis cycle, screening each variant in the library becomes a laborious process. Selection techniques offer the advantage of filtering out the protein variants that are non-functional and do not generate the desired phenotype, returning only the functional variants to be analysed. Selection is typically performed *in vivo* by linking the enzymatic activity of the target protein to the fitness of the host cells. If enzyme activity confers resistance to an antibiotic, or if the enzyme product can be used as a metabolic source by an auxotrophic strain, it allows the host cell to survive[51]. This is exemplified by a xylose auxotrophic strain of *S. cerevisiae* used to select for xylose isomerases with improved catalytic efficiency[67]; and using aspartate auxotrophic *E. coli* to select for lipases with improved enantioselectivity, where producing the undesired enantiomer yielded a poison (fluoroacetic acid), killing the cell[68].

This necessity for desired enzyme function to be directly linked to cell fitness greatly limited the use of selection assays in directed evolution experiments. Instead, the desired enzymatic activity can be linked to the transcriptional activation of a reporter molecule[69,70], demonstrated by *in vivo* directed evolution techniques developed in recent years. Once the desired enzyme activity is achieved, it will cause the expression of the reporter protein, which is visibly or functionally detectable in the experimental assay[71].

Such a conditional selection system has so far been demonstrated using phage-assisted continuous evolution (PACE)[i]. In the case of PACE, the 'reporter' protein is the M13-phage pIII coat protein; its expression is essential for producing functional M13 phage (pIII knocked out from the phage genome). If the protein library generates the desired phenotype, it will enable expression of pIII, which results in production of virulent phage that can infect other bacterial cells in the culture, enriching the phage population with the functional protein. Such conditional phage production has been used to evolve the specificity of T7 RNA polymerase (T7pol) from a T7 promoter to a T3 promoter[72]; evolving the λ cI repressor to recognise different promoter sequences to act as a repressor, activator or dual activator-repressor[73]; evolving hepatitis C virus protease to recognise and cleave desired cleavage sites, where proteolytic activity cleaves the T7 lysozyme fused to T7pol, activating the RNA polymerase to transcribe pIII[74]; and evolving recombinases to recognise different attachment sites[72,75].

These selection strategies can be translated beyond the PACE system. Instead of expression of pIII, the enzymatic activity can be linked to the conditional expression of a fluorescent reporter, which will enable high-throughput selection of the desired phenotype by using FACS. Similarly, desired enzymatic activity can be conditionally linked to the expression of an antibiotic resistance gene. Plating the protein-library-containing cells on agar with the appropriate antibiotic would enable rapid identification of protein variants displaying the desired phenotype. In Chapter 6, we discuss strategies for how such a selection system can be applied with a library generation technique in *E. coli*.

[i] Refer to 1.5.1 for detailed overview of the PACE technique

1.3.2 Second Challenge: Method to Generate a Library of Mutant Proteins

Designing a screening or selection technique to identify and quantitate the enzymatic activity of desired proteins is half the battle. It needs to be coupled with a method that introduces mutations (called mutator system) in the DNA sequence encoding the target protein to generate a variant library. The quality of the mutant library generated by a mutator system determines the efficiency of the directed evolution process. Ideally, a mutator system should produce an unbiased mutation spectrum to generate a genotypically diverse protein library, to explore a large evolutionary search space[76]. An ideal mutator system should also possess a controllable mutation frequency, which can be increased or decreased to achieve an optimal rate of mutation for evolution.

In the early years of protein engineering, chemical and physical mutagens were being used to induce damage in the gene-of-interest placed inside a host cell. The lack of targeting meant that the entire genome of the cell was subjected to random mutations, affecting host cell viability after long periods of mutagenesis. The chemical agents used include alkylating compounds such as ethyl methanesulfonate (EMS)[76], deaminating compounds such as nitrous acid[77,78], base analogues such as 2-aminopurine, and ultraviolet irradiation[79] to induce DNA damage. Chemical mutagenesis is not commonly used for directed evolution because of biases in the mutational spectrum and a low mutation frequency due to the lack of targeting, and DNA repair pathways fixing the damaged nucleotides[80].

Eventually, researchers developed mutator strains of bacteria that are capable of inducing genome-wide mutations. These strains were designed to introduce mutations during the DNA replication process for cell division. Normally in bacteria, DNA polymerase III (pol-III) replicates the two DNA strands of the bacterial genome with an error rate of 10^{-10} base^{-1}. This error rate is 10^{2}-10^{4}-fold higher in mutator strains as DNA proofreading and repair enzymes mutS (mismatch DNA repair protein), mutL, mutT (prevent A:G mispairs during DNA replication) and mutD (ε subunit of pol-III with 3’-5’ exonuclease activity) are knocked out[81]. Due to the lack of targeting, these strains not only mutate the target gene but also induce deleterious mutations in the host genome. Host intolerance to a high degree of genomic mutation places an upper limit on the mutation rate that can be applied for directed evolution experiments with mutator strains (mutation densities of only 1 point mutation per 2-5 kbp[82]).

In Vitro Mutator Systems

The ability to target mutations to the gene-of-interest and having a controllable mutation rate became essential qualities for a mutator system. In the early years of protein engineering, there was a lack of laboratory techniques capable of reliably manipulating targeted DNA sequences *in vivo*. The focus, therefore, shifted to using molecular cloning and recombinant DNA technologies for performing such tasks *in vitro* to generate libraries of mutated DNA sequences. These DNA sequences would subsequently be transformed into cells for screening or selection. The *in vitro* mutator systems fall into three categories: PCR-based, semi-rational design and DNA Shuffling.

Error prone PCR (EP-PCR)

Researchers began utilising EP-PCR to generate a mutant library of target genes as the method is simplistic, allows for controllable amplification of the target gene, and can potentially generate randomised mutations, which means a vast evolutionary search space can theoretically be explored[83]. The technique involves using Taq polymerase and enhancing its error rate by the addition of Mn^{2+} in the reaction mixture and by modifying the ratio of the dNTPs. Performing EP-PCR under these conditions displays a strong bias for mutations occurring at A:T base-pairs and biased for transitions (A → G, T → C)[84,85]. Eliminating such biases requires using a mixture of different polymerases, such that the mutational bias of one is offset by another polymerase. Optimising an EP-PCR reaction to increase the diversity of the mutation spectrum can be a challenging process. Also, a diverse mutations spectrum in the DNA sequence doesn't always correspond to diversity in the peptide sequence, due to the degeneracy of the codons[84,86]. Due to the random nature of incorporating mutations, many library members contain silent or deleterious mutations, resulting in smaller populations of the variant library being functionally diverse.

To address these limitations of EP-PCR, a method called Sequence Saturation mutagenesis (SeSaM) was developed. This method can potentially eliminate mutational 'hot-spots' and increase the frequency of mutations per mutation cycle[87]. It is a 4-step process involving chemical fragmentation of DNA; PCR with phosphorothioate nucleotide and biotinylated primers to generate different length single-stranded DNA (ssDNA) of the target gene; and using universal DNA analogues to promote nucleotide exchange with the fragmented DNA. The phosphorothioester bonds within the resulting PCR products can be cleaved selectively in the presence of iodine and under alkaline conditions. These amplicon fragments are tagged using universal base analogues, which are then elongated to the full length of the gene sequence. In the final step, a PCR reaction is performed to replace the base analogues with standard nucleotide bases. The entire process takes 2-3 days for one library of mutants to be generated. The base-analogues used dictates the mutations spectrum that can be achieved with method[87]. Avoiding a mutation bias requires using multiple base analogues to cover all transitions and transversions, adding to the complexity of this method for generating a mutant DNA library.

The main advantages of these PCR-based methods for generating a mutant DNA library is that they are relatively fast, economical and can generate library sizes of 10^9 variants relatively quickly. These advantages are offset by the degree of bespoke optimisation that is required to achieve a diverse library of mutants, to achieve a low mutation bias and a high mutation coverage. Creating the right reaction conditions to achieve all these goals is challenging.

Semi-rational Design to Generate DNA sequence Libraries

A semi-rational approach combines rational design with directed evolution to reduce the sequence search space explored for evolving the gene. Unlike EP-PCR, where 10^6-10^{10} variants are screened per mutation cycle, number of variants screened using semi-rational approaches are significantly lower[36]. Using structural information about the target protein and analysing features such as its active site, only specific DNA nucleotides, encoding the amino acids forming the active site or other unique features of the protein, are identified and mutated[88,89]. This helps to significantly reduce the sequence space being explored. Another advantage of semi-rational approaches is that two consecutive DNA nucleotides can be mutated, increasing the probability of a codon change and reducing the probability of silent mutations[90].

Site Saturation Mutagenesis (SSM) involves iterative or combinatorial incorporation of mutations by adding degeneracy into the three nucleotides at a codon position. During translation, this provides access to any of the 20 canonical amino acids to be incorporated at the codon position. The experimental process involves performing overlap extension PCR to generate the DNA library, which is transformed into cells for subsequent screening or selection with an appropriate method[89]. In iterative SSM, the mutants from one round are carried over to the next round of PCR to create specific combinations of mutations in the library[88]. An advantage of SSM is that it helps elucidate the structure-function relationship of proteins as the mutational tolerance at specific amino acid positions can be probed readily.

Combinatorial SSM at multiple amino acid sites simultaneously helps elucidate mutation combinations that may exhibit epistatic interactions. The two mutations might be silent or deleterious on their own, but when combined provide a fitness advantage. Such combinations cannot be explored using iterative SSM[91]. Iterative SSM however, offers the advantage of mapping out an evolutionary tree to determine how each sequential mutation affected the protein's structure and function[88,92].

Overall, semi-rational approaches to protein engineering suffer the same limitations as rational approaches — limited to proteins with detailed information about their structure and function. The reduced search space also means SSM techniques may not uncover the best variant of the target protein[88].

DNA Shuffling Techniques:

Restriction Enzyme-Directed Shuffling

DNA shuffling techniques are typically coupled with EP-PCR to generate a greater genetic variety in the library of mutants. The technique involves DNAse-I mediated digestion of the mutant library generated using EP-PCR. These digested fragments then randomly anneal to one another based on sequence homology to generate full-length genes, where mutations have been incorporated from multiple EP-PCR templates. A limitation of recombination methods is that they cannot generate genetic diversity on their own. An initial mutant library with genetic diversity is required, where the genetic diversity can be further amplified by such shuffling techniques. The diversity may be artificially introduced by using synthetic oligonucleotides, designed to contain point mutations[93]. Another limitation is that for successful DNA shuffling events, the parental DNA sequences must maintain a 60-70% sequence homology. These restricts the scale of genetic diversity created by shuffling DNA sequences of divergent proteins together.

Assembly of Designed Oligonucleotides

This DNA shuffling technique involves generating the target DNA sequence as small synthetic oligonucleotides with degenerative nucleotide bases, that anneal via homology and are extended via overlap extension PCR[94]. The first step involves a primer-free, polymerase cycling assembly or overlap extension PCR step using carefully designed overlapping synthetic oligonucleotides. The second step is a PCR amplification of the overlapped assemblies using outer 5'-binding primers, creating a high quality and bias-free dsDNA library that can be shuffled with other gene fragments. This is an example of homology-dependent recombination, where the sense and anti-sense synthetic oligonucleotides must have overlapping regions to maintain sequence homology and preserve the gene structure of recombinants. Mutations cannot be incorporated in these overlapping regions[95].

Homology Independent Shuffling

These techniques were developed to enable recombination of two gene sequences or DNA fragments that share no sequence homology to each other. One such technique, Sequence Homology Independent Protein Recombination (SHIPREC) enables the shuffling of disparate DNA sequences. This is particularly useful for recombining families of proteins with similar functions but possessing low sequence homology. The technique involves random fragmentation (using a DNase or an exonuclease) of the different gene sequences, followed by sequence-independent ligation of the fragments[96]. Tuning the fragmentation conditions helps to alter the average number of ligation events, and electrophoresis is used to isolate ligated products of the desired length. The electrophoresis step helps to minimize isolation of inactive library members that are too short or too long. Despite such a rigorous filtration process, many of the variants obtained after homology independent shuffling tend to display domain disruption and folding instability[95].

Incremental Truncation for the Creation of Hybrid Enzymes (ITCHY) generates different sized fragments of two genes by performing exonuclease-III-mediated digestion. These ftagments are then joined by blunt-end ligation to generate a chimera of the two proteins. Different sized fragments produce chimera proteins of different sizes, which can be screened for protein function[97]. With the ITCHY method, two-thirds of the mutant library generated tend to contain frameshifts. This generates many unstable proteins and restricts the library's use in downstream DNA shuffling techniques (like SCRATCHY[98]) to enhance genetic diversity.

Multiplex Automated Genome Engineering (MAGE)

This method involves the use of oligonucleotides to introduce multiple mutations simultaneously into the target bacterial genome. These oligonucleotides interact with the lagging strand during DNA replication by mimicking Okazaki fragments. Mechanistically, with the help from heterologously expressed RecE/RecT or λ-red system, single-stranded DNA oligonucleotides harbouring deletions, insertions, or mismatches anneal to the lagging strand during DNA replication, resulting in the incorporation of customized sequence modifications in the newly synthesized genome. This method has been limited to model bacterial strains due to difficulty of generating homologous recombination with short oligos in other organisms. Owing to the short length of the oligos, only short 20 bp long modifications can be introduced in a particular DNA sequence per cycle of MAGE[99]. For every cycle, new oligos would have to be designed, which would require knowledge of the structure and function of the protein to efficiently evolve the corresponding gene sequence.

Each of the *in vitro* methods discussed here for generating protein libraries have their advantages and disadvantages. Identifying the appropriate technique to use for library generation depends primarily on the protein to be evolved; the available screening or selection strategy for the protein; and the amount of evolutionary search space that needs to be explored to achieve the desired protein function.

1.4 Evolutionary Search Space and the Impact of Neutral Drift Libraries

John Maynard described protein evolution as the process of walking from one functional protein to another in a vast space of all possible protein sequences. Each functional protein can be envisioned as a mountain, where the tip of the mountain represents the ideal protein sequence, providing the best function and fitness to the host organism, while other regions of the mountain represent sequences that are less efficient versions of the protein (Figure 1.2)[100]. The valleys in between the different protein mountains represent non-functional protein sequences that must be 'traversed' to get from one mountain to another. During directed evolution, when protein sequences are mutated, the mutation can either be beneficial and make the protein sequence climb up a fitness mountain, or non-beneficial and move it further down the hill[101]. Climbing to the top of the mountain requires successive beneficial mutations to be accumulated in the protein sequence. Therefore, for improving the qualities of existing proteins, rational and semi-rational design approaches are ideal. By applying structural and functional knowledge of the protein, only a few amino acid substitutions can result in vastly improved enzyme properties[88,102,103]. However, due to the small evolutionary search space explored by these methods, evolving a novel protein function or converting one enzyme to another can be challenging.

Changing a protein's enzyme activity requires moving down one fitness mountain and climbing onto another. This process requires both beneficial and non-beneficial mutations to traverse across mountains[104,105]. This endeavour requires exploring a large sequence space and numerous sequential mutations are required to traverse across this search space. High throughput screening coupled with a mutator system to generate randomly mutated libraries is currently the best way of exploring such vast sequence spaces. Error-prone PCR can generate libraries large enough to explore such vast sequence spaces; however, the number of beneficial mutations generated by EP-PCR tend to be only 0.01-1.0% of the total mutations[101]. Most random mutations generated are silent or deleterious. This low frequency of functional mutations greatly increases the iterative cycles of mutation and screening needed to identify functional phenotypes. This challenge is even tougher when working with protein sequences that are reluctant to evolve[106].

One way to increase the evolvability of proteins and traverse the mountains and valleys faster is to generate neutral drift libraries of proteins. In certain cases, the starting protein sequence may not be suited to evolution due to a high tolerance to mutations. In these instances, cycles of library generation with selection pressure can be performed, selecting for the wild-type function. This enables the accumulation of variants that have different amino acid sequences but retain their native function to provide cells with a growth advantage. Therefore, the starting sequence at the tip of the mountain is converted to a library of protein sequences spread across different fitness landscapes.

This creates many unique evolutionary trajectories to evolve the ‘promiscuous’ activity of a protein[104,105]. This process is called neutral drift and it serves as an ideal starting point to evolve new protein functions. For example, Bloom et al. showed that after 25-rounds of neutral drift, from a library of only 100 variants, mutant cytochrome P450 enzymes exhibited up to four-fold changes in its ability to perform the hydroxylation of five different promiscuous substrates[107].

Generating neutral drift libraries and downstream cycles of mutation and screening can be performed using EP-PCR. But the constant cycles of transforming DNA libraries into cells for screening, then isolation of valuable targets from cells for another round of mutation and screening can be laborious and time consuming. The transformation efficiency of the host cells also places an upper limit on the library sizes that can be screened per cycle (~ 10^9 variants). Such limitations of *in vitro* mutator systems for studying long evolutionary trajectories and efficiently evolving promiscuous protein function led researchers to engineer targeted *in vivo* mutator systems.

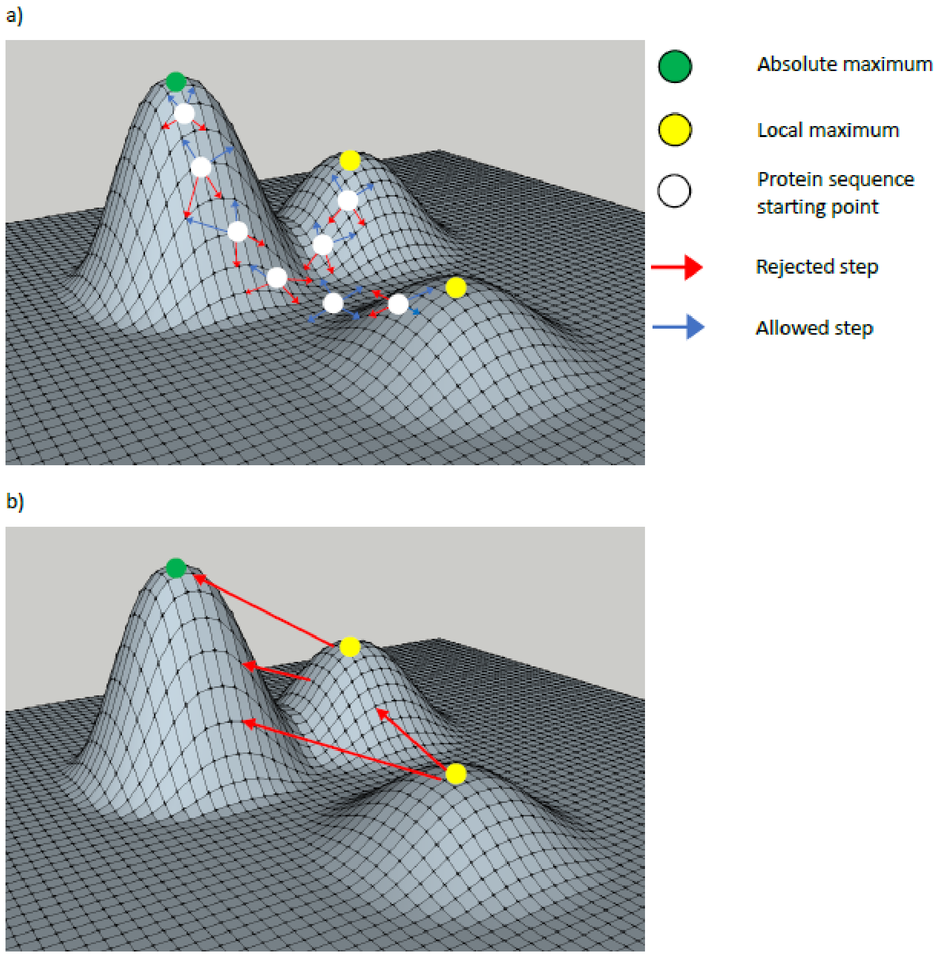


Figure 1.2: ***The three-dimensional protein fitness landscape***. *Each point on the grid represents a different protein sequence, such that neighbouring points differ by only one amino acid. a)*

Accumulation of point mutations on a starting DNA sequence, followed by selection or screening of variants allows for identification of mutants that moved up a fitness peak. Eventually, the protein may arrive at a local maximum. b) Depending on the evolutionary path taken, a variant can be trapped at a local maximum, unable to reach the absolute fitness maximum. Larger jumps that cross valleys to reach neighbouring peaks require numerous iterative or simultaneous mutations to occur, allowing the protein to walk down one fitness peak, and climb another. (Figure adapted from Packer & Liu, 2015)

1.5 *In Vivo* Mutator Systems

An *in vivo* mutator system involves placing a single copy of the gene encoding the target protein inside a cellular host. This gene is targeted for mutations, generating a library of mutants directly inside the cell. By coupling the desired enzyme function to the fitness of the host cell or to the conditional expression of a reporter, variants with the desired properties providing a fitness advantage will automatically be selected by the cells and be enriched in the cell culture. As both, library generation and selection, are occurring stochastically inside the cells, targeted *in vivo* mutator systems can evolve desired proteins with minimal human intervention. When setup in a chemostat or turbidostat, such evolutionary systems have been demonstrated to run for up to 500 hours without human intervention[108]. As the target gene is continually evolving inside replicating cells, it is allowed to traverse through many different routes in the fitness landscape until the desired protein is identified via the *in vivo* selection system. This technique of evolving proteins inside naturally growing and replicating cells is known as continuous evolution. The discovery of orthogonal plasmid-polymerase pairs, CRISPR-Cas9 gene editing technologies, and the targeted use of DNA damaging enzymes has resulted in efforts to develop *in vivo* continuous evolution systems where mutations are targeted to a gene-of-interest, while shielding the host cell genome from mutagenic activity[108,109,110,111].

1.5.1 Phage-assisted Continuous Evolution (PACE)

PACE separates the library generation and phenotypic selection process between two organisms: *E. coli* and a M13 bacteriophage[72,75]. The PACE system does not use a targeted mutator system; it uses DNA damaging proteins that disrupt DNA replication, DNA proofreading, base excision repair and mismatch DNA repair pathways, which are expressed from a mutator plasmid (MP). The gene-of-interest is integrated into a modified phage genome where gene-III knocked out. The bacterial cells also express an accessory plasmid (AP), which contains an expression cassette for gene-III. Gene-III encodes the pIII coat protein that forms the phage capsid. PIII enables the virulence of phage; its N2 domain binds to the F-pilus of bacteria, freeing the N1 domain to interact with a TolA protein on the surface of the bacterium and promoting subsequent invasion into the cell cytoplasm[112]. In the PACE system, the expression of functional protein variants is linked to the conditional expression of gene-III.

When evolving T7 RNA polymerase to recognise T3 promoter rather than P_{T7}, gene-III was placed downstream of P_{T3} to conditionally link its expression to T7-RNA-Pol evolution. The library generation process begins once the phage has infected the host cell; the mutator plasmid randomly introduces mutations to the selection phage's genome containing the T7-RNA-Pol gene. If T7-RNA-Pol has not evolved to recognise P_{T3}, phage molecules will be packaged without pIII, and will lack virulence as a result. These phage molecules will be washed out from the system as they cannot infect bacterial cells. If a T7-RNA-Pol variant capable of initiating transcription via P_{T3} is generated via continuous evolution, pIII will be expressed and virulent phage molecules will be produced. These phage molecules will continue to infect cells in the culture, replicate and be retained in the batch culture. The enriched functional phenotype can subsequently be detected using bacterial plaques. The protein variant is subsequently isolated for characterisation.

This continuous evolution system was used to evolve the λ cI repressor[71,73], hepatitis C virus protease[74] and TEV protease[113]. The system has a few limitations. The lack of targeted mutagenesis by the DNA damaging proteins expressed from the mutator plasmid results in a lack of genetic diversity at the start of the evolution experiment. Lack of genetic diversity means most phage molecules at the start remain non-virulent and get washed out. Genetic diversity needs to be artificially generated using EP-PCR to prevent this from happening. The lack of targeting by the mutagenic proteins also means that mutations accumulate in the host cell genome, ultimately affecting its fitness. To run the PACE system for long periods, the population of bacterial cells needs to be replenished constantly via outflow of damaged cells and inflow of fresh cells into the bacteria-phage culture.

1.5.2 Targeted DNA Damage using CRISPR or T7 RNA Polymerase

DNA Glycosylase and CRISPR

Cas9 and deactivated cas9 (dCas9) are endonucleases that use RNA molecules as guides (called gRNA) to find and bind to DNA sequences that are complimentary to the gRNA[114,115]. This targeting system can be implemented for introducing mutations to a gene-of-interest. This was achieved by fusing the Cas9 or dCas9 protein to DNA damaging enzymes such as glycosylases. In one example, dCas9 was fused to the cytidine deaminase, AID (activation-induced cytidine deaminase), to generate C → T mutations in the target DNA sequence and evolve wildtype GFP to have higher fluorescence[116]. The current limitation of CRISPR-mediated directed evolution is the small mutational window. Guide RNA are typically 20-nucleotides long and AID-dCas9 mediated damage was only shown to occur -12bp to -16bp upstream of the PAM sequence[117], and +12bp to +32bp downstream relative to the PAM[116]. Due to this restricted mutational window, extensive knowledge about the structure-function relationship of the protein is required to target the base editing machinery to amino acid residues relevant for catalytic activity. Also, this targeting system is not suitable for continuous evolution as multiple rounds of library generation will require multiple unique gRNA sequences to bind to the evolving gene. Generating such a redundant library of gRNA for continuous evolution is not rationally or economically feasible.

EvolvR: Error-prone DNA Polymerase I and CRISPR

EvolvR fuses a nickase mutant of Cas9 (nCas9) to EP-Pol-I for introducing semi-random mutations in a 58-bp window for any genomic or plasmid site that can be targeted by nCas9[118,119]. First, nCas9 creates a single-strand break at the DNA site complimentary to the gRNA, and subsequently disassociates from the DNA. Then EP-Pol-I binds at the nick and polymerises through the target DNA sequence, while its native 5'-3' exonuclease domain degrades the displaced strand. Fusing nCas9 and EP-Pol-I together allows for substantially enhanced mutagenesis. The technology is an improvement over AID-dCas9 as it can incorporate mutations other than C → T. However, EvolvR has been shown to exhibit substitution biases, with dA and dT comprising more than 80% of the introduced mutations. Also, multiple guide RNAs would be needed to mutate multiple regions of the gene[95].

DNA Glycosylase and T7 RNA polymerase (T7pol)

To address the small mutational window limitation of CRISPR-based DNA damaging devices, Moore et al fused a cytidine deaminase (rApo1) to T7 RNA polymerase. They demonstrated that any DNA sequence, regardless of length, placed between a T7 promoter and a T7 terminator can be subjected to mutations by the glycosylase-T7pol fusion protein[120]. However, the system naturally showed a bias for C → T mutations. Similar research was performed where the cytidine deaminase, AID, was fused to T7-RNA-Pol. This version of the glycosylase-T7pol mutator generated C → T or G → A mutations at the site of U:G lesions generated by AID. Despite the lack of genetic diversity, the rate of targeted mutation was significantly higher than off-target mutation for this fusion protein[121].

1.5.3 OrthoRep: Utilising orthogonal plasmid-polymerase pairs for targeted mutagenesis

OrthoRep is a yeast-based continuous evolution system that utilises the p1 and p2 linear cytoplasmic plasmids isolated from *Kluyveromyces lactis*[109,122]. These plasmids are flanked by terminal proteins (TP) that act as origin of replications for TP-DNA polymerase to bind and replicate the DNA. TP-DNA polymerase-I (TP-DNAP1) can only interact with p1 terminal proteins, which creates orthogonality. The OrthoRep system works by integrating the gene-of-interest into p1 and using an error-prone version of TP-DNAP1 to introduce mutations into the target gene. The orthogonality of this system means that the yeast genome is protected from mutagenic activity and is able to maintain a natural mutation rate of ~10^{-10} per base, while the mutation rate for the target gene is 100,000-fold higher. By linking cell fitness to the expression of the target protein, selection for the desired phenotype was performed to evolve PfDHFR's resistance to pyrimethamine after 90 yeast replicative cycles[109,122].

OrthoRep exemplifies an *in vivo* continuous evolution system that is designed for long time-course evolutionary experiments to explore a vast evolutionary search space for improving enzyme function and discovering novel proteins. The system can specifically target mutations to the gene-of-interest at mutation rates of ~10^{-5} substitutions per base for numerous cell cycles lasting for days; it is easy to setup and minimal human intervention is required until the selection step to isolate functional mutants; it is scalable with the possibility of multiple continuous evolution experiments running in tandem; and it can be combined with fitness-based selection strategies or FACS-based screening methods to identify protein variants with the desired properties[108,109,122].

A potential limitation of this method is the accumulation of mutations at the promoter and RBS sequences, which are required for the expression of the evolving protein. If a protein cannot be expressed, it cannot be screened for the functional phenotype. The authors reported 24 out of 78 instances of evolved PfDHFR where mutations were incorporated into the promoter sequence[109]. While the promoter remained functional in these 78 cases, there could also have been instances where the promoter became non-functional, preventing expression and subsequent screening of potential PfDHFR candidates. Losing the ability to screen functional variants can potentially result in the loss of useful protein sequences and information about the evolutionary trajectory taken by them.

1.6 Designing a Targeted Mutator System for Continuous Evolution of target genes in *E. coli*

In vivo continuous evolution techniques like OrthoRep and PACE that generate protein libraries via random mutations enable researchers to gain deep insights into the mechanism of protein evolution[123]. These methods allow for traversal through long mutational pathways, and by applying different selection pressure, study the impact it has on the evolutionary trajectory adopted by the evolving protein. With OrthoRep being limited to yeast strains and PACE lacking a targeted mutator system, there is scope for a targeted continuous evolution system to be developed in *E. coli*, that contributes to the collective effort of protein engineering and towards better understanding the mechanism of protein evolution.

In this doctoral thesis, the aim was to develop a mutator system that combines some of the useful features of the published methods discussed above to specifically induce damage in target DNA sequences placed in plasmids or integrated into the *E. coli* genome. We also developed a novel method for performing targeted error-prone DNA repair of genes in *E. coli*. Combining the targeted DNA damage system with the error-prone DNA repair complex resulted in producing a library of mutants for the gene-of-interest, at mutation rates comparable to the OrthoRep system (Chapter 5). The designed mutator system was also shown to target mutations to the gene-of-interest at a 10^2-10^3-fold higher frequency than producing off-target mutations in the RpoB gene, which can confer resistance to rifampicin[124] (Shown in Chapter 3 and Chapter 4). The mutator system has been designed to comprise two fusion proteins, one for performing DNA damage and the other for error-prone DNA repair of the damage. Both fusion proteins are expressed via an engineered J23101$_{\text{TetO}}$ inducible promoter employing the Tet repression system. This creates a switch-like mechanism to regulate the expression of the mutagenic proteins.

The first fusion protein is the activation-induced cytidine deaminase (AID) fused to T7 RNA polymerase (AID-T7pol), which has been named the DNA damage device. With the gene-of-interest placed downstream of a T7 promoter, this fusion protein can generate U:G lesions randomly across the gene sequence. The second fusion protein is designed to contain three domains, a 5'-3' exonuclease, an AP-endonuclease and an error-prone DNA polymerase (5'-3'Exo—AP-Endo—EP-DNA-Polymerase), called the error-prone DNA repair complex. This mutator system is designed to emulate the process of somatic hypermutations[125] in *E. coli*.

Mechanism of the Designed Mutator System for Generating Genetic Diversity

The mutator system is designed to hijack the base excision repair pathway (BER)[ii]. First, T7-RNA-pol of the AID-T7pol fusion transcribes the target DNA. The transcription bubble created by the RNA polymerase provides AID with single-stranded DNA template it needs to perform C → U deamination

[ii] Biological mechanism of the mutator system described in detail in Chapter 2

at random sites along the gene[126]. The U:G lesion generates an SOS response, activating the BER pathway[127]. Uracil-N-glycosylase (UNG) scans the DNA for the lesion and cleaves the pyrimidine base of deoxyuridine to generate an AP-site. This is where the error-prone DNA repair complex hijacks the traditional BER pathway. The AP-endonuclease within the 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion protein nicks the DNA at the AP-site, creating a gap. The error-prone DNA polymerase then performs gap-filling. The 5’-3’ exonuclease excises nucleotides upstream of the gap to create longer ssDNA template for gap-filling by the EP-DNA-polymerase, increasing the chance of errors being incorporated downstream of the site of U:G lesion (Figure 1.3).

In this mutator system, the large mutational window of AID-T7pol enables U:G lesions to be generated across the entire length of the target gene, which was verified using next generation sequencing (Chapter 5). With a 1000-fold lower off-target mutation rate, cells expressing the mutator module maintained their genetic fidelity and displayed minimal loss of fitness in mutagenesis experiments performed for library generation over 144-hours.

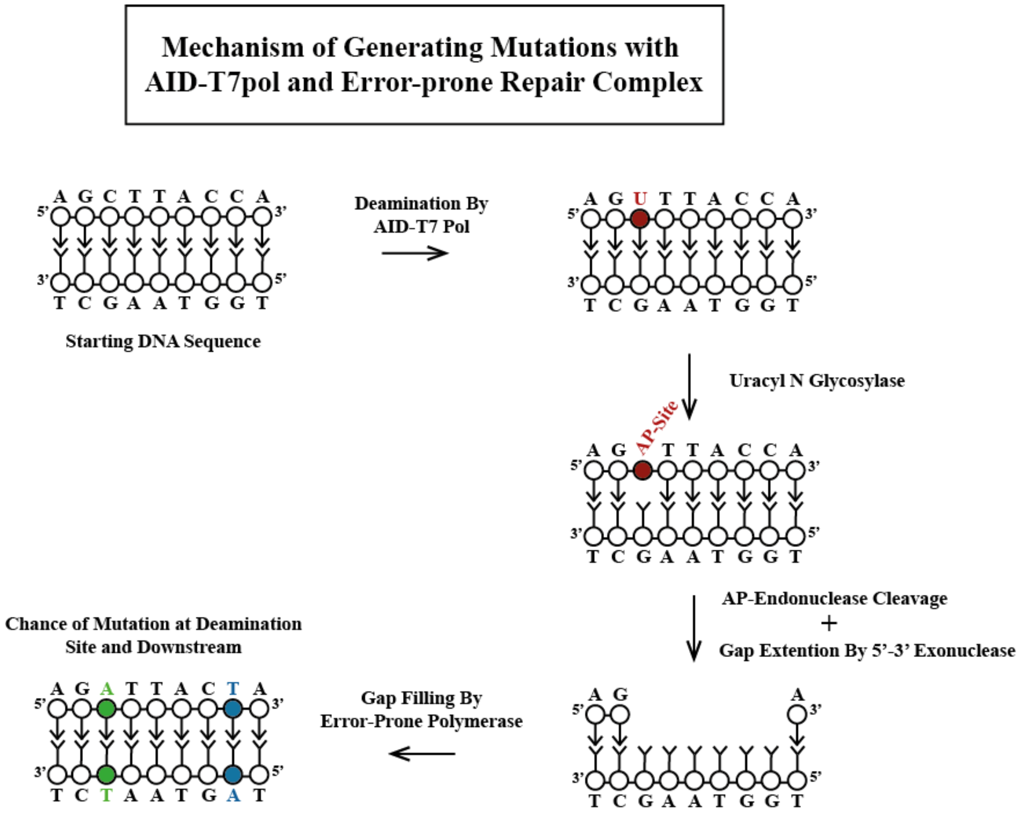

Mechanism of Generating Mutations with AID-T7pol and Error-prone Repair Complex
A G C T T A C C A
5’ 3’
3’ 5’
T C G A A T G G T
Starting DNA Sequence
Deamination By AID-T7 Pol
A G U T T A C C A
T C G A A T G G T
Uracyl N Glycosylase
AP-Site
A G T T A C C A
T C G A A T G G T
AP-Endonuclease Cleavage + Gap Extention By 5’-3’ Exonuclease
A G A
T C G A A T G G T
Gap Filling By Error-Prone Polymerase
Chance of Mutation at Deamination Site and Downstream
A G A T T A C T A
T C T A A T G A T

***Figure 1.3**: **Mechanism of emulating somatic hypermutations in E. coli**. The AID-T7pol fusion protein deaminates dC to dU, generating a U:G mismatch. Damage caused to one strand of DNA activates the BER pathway, where uracil-N-glycosylase recognises and cleaves the uridine to create an apyrimidinic site (AP-site). The EP-DNA-repair pathway hijacks the repair process at this step via the AP-endonuclease in the 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion. The AP-endo nicks the backbone*

to generate a gap. This gap is filled in by the EP-DNA-polymerase, with a chance to incorporate nucleotide substitutions.

1.7 Aims of the Doctoral Thesis

The research goal in this doctoral thesis was three-fold: (1) optimise the expression of the AID-T7pol DNA damage device to reduce fitness burden, toxicity, and enable long time-course mutator experiments; (2) build, test and validate 5’-3’Exo—AP-Endo—EP-DNA-Polymerase EP-DNA-repair complexes assembled with different biological parts for their ability to perform targeted error-prone DNA repair of damaged DNA; and (3) Perform a long time-course mutagenesis experiment with candidate mutator systems on a gene-of-interest, sequence the mutant gene library using NGS platforms and analyse the diversity of mutations that the mutator modules can generate.

In Chapter 2, the biological mechanism of how the mutator system is design to work is presented. Emulating the mechanism of somatic hypermutations required an understanding of the different proteins that are involved in the process and how the function of these proteins could be localised to the gene-of-interest. An AP-endonuclease, error-prone DNA polymerase and a 5-3’ exonuclease were identified as protein classes that are essential to somatic hypermutation. Different candidate proteins for each class were identified that could be used in the EP-DNA-repair complex. The main experimental strategies used to validate the mutagenic properties of the mutator modules are also presented. The design of a loss of function experiment with GFP and a gain of function experiment with inactive versions of the β-lactamase antibiotic resistance gene are outlined.

In Chapter 3, the process of optimising the expression of AID-T7pol in *E. coli* is presented. This DNA damage device was developed in previous research, but its expression level was toxic to the host cells[121]. Optimising the expression rate of this fusion protein to minimise toxicity was crucial for developing a mutator system that can evolve target genes over multiple cell cycles. The expression of this fusion protein was optimised using a library of degenerative RBS sequences. RBS sequences were identified that enabled high, medium and low expression levels of AID-T7pol in cells, with minimal effect on the host cell fitness. The ability of AID-T7pol to only generate C → T and G → A mutations was validated with the loss-of-function and gain-of-function experiments.

In Chapter 4, the process of engineering the error-prone DNA repair complex is presented. Each of the candidate 5’-3’ exonuclease, AP-endonuclease and error-prone DNA polymerase parts for building the EP-DNA-repair complex were characterised for their expression in *E. coli*. A set of RBS sequences were identified that enabled expression of these candidate proteins at low, medium and high expression rates, while causing minimal fitness burden to the cellular host. These candidate proteins were assembled into EP-DNA-repair complexes containing 2-protein (AP-Endo—EP-DNA-Polymerase) or 3-protein (5’-3’Exo—AP-Endo—EP-DNA-Polymerase) fusions. A library of mutator modules was subsequently assembled by combining the expression cassettes for AID-T7pol and the EP-DNA-repair complex onto a single plasmid. The mutagenic properties of this library of mutator modules was assessed using the loss-of-function and gain-of-function experiments outlined in Chapter 2. Mutator modules possessing the highest mutation frequencies and the ability to perform specific nucleotide substitutions were identified.

In Chapter 5, assessment of the mutational characteristics of candidate mutator modules, shortlisted from Chapter 4, using next generation sequencing is presented. The characteristics that were assessed include the mutation frequency, diversity of mutations, and the spread of mutations across the target gene ORF. The candidate mutator modules performed cycles of library generation for the GFP-mut3b gene in a continuous evolution experiment lasting 144-hours. The mutant GFP library was sequenced using the Pacific Biosciences Sequel and the Illumina iSeq100 platforms. The sequencing reads were analysed using bespoke Python scripts utilising the Biopython API.

In Chapter 6, a conclusion is derived from all the data presented in this thesis. Potential future work where this mutator system can be linked to a selection system to perform directed evolution of the LuxR transcription factor is presented. Evolving LuxR to recognise butanoyl-homoserine lactone (C4-HSL) as the activating ligand, instead of C6-HSL, would serve as a proof-of-principle experiment to test the applicability of the designed mutator system in a continuous evolution system for library generation.

Chapter 2: Biological Principles and Experimental Design for Engineering an *In Vivo* Gene-targeted Mutator System

The aim of this doctoral research was to develop an *in vivo* mutation system capable of targeting DNA damage to a specific gene-of-interest and performing subsequent error-prone DNA repair of the damage to generate genetic diversity, which would fulfil the requirements for an efficient *in vivo* platform for directed evolution. To accomplish this task, some challenges needed to be overcome. It was important that the mutator system can direct mutation to a specific gene, whilst not causing a significant increase in background mutation of the host genome. It was also necessary to ensure that the expression of the mutator system is not toxic to cells or incurs a heavy fitness burden. In addition, an efficient mutator system needs to generate a diverse range of mutations in the gene-of-interest.

The ability to both target DNA damage and generate genetic diversity were recognised as requiring separate solutions. DNA damage can result from enzymes that chemically modify nucleotides in a DNA sequence, while the opportunity to generate genetic diversity lies at the point of DNA repair, when new DNA is synthesised. We therefore devised a two-component mutator system that is capable of both directing DNA damage and performing error prone DNA repair for creating genetic diversity. All these qualities should result in a mutator module that can perform continuous evolution of a target gene *in vivo*. In this chapter, we discuss the strategies that were applied for developing a mutator system suitable for library generation in an *in vivo* continuous evolution platform in *E. coli*.

2.1 Applying the Principle of Somatic Hypermutations in the Design of the Mutator Module

It was crucial in the early stages of the research to identify and understand natural biological means by which the process of evolution can be accelerated in certain cells or organisms. Attempting to adapt such a biological method to function in *E. coli* cells was the experimental goal. Attention was immediately drawn to the process somatic hypermutations (SHM) — a method used to generate antibody diversity for eliciting an adaptive immune response[126].

Generating these hypermutations relies on two key steps: inducing DNA damage on the genes encoding the variable regions of the antibody; and DNA repair performed by an error-prone DNA polymerase[125,126,128]. During B- and T-cell maturation, the variable (V), joining (J) and diversity (D) gene segments encoding the variable regions of an antibody undergo random rearrangements called VDJ recombination. The process of somatic hypermutations is initiated after this step, by an enzyme called activation induced cytidine deaminase (AID). It induces DNA damage in the exons of genes encoding the VDJ segments by deaminating a deoxycytidine to deoxyuridine and creating a U:G mismatch.

Mutations are generated around the site of U:G mismatch in three ways: (i) during DNA replication, DNA polymerase pairs dA to dU, generating a C → T and a G → A transition in one of the daughter DNA strands; (ii) error-prone base excision repair (BER) pathway to excise the deoxyuridine and replace it with another nucleotide; (iii) Recruiting the mismatch repair machinery to incorporate mutations at the site of U:G mismatch and upstream, potentially at A:T base pairs[125,128].

2.1.1 Understanding the SHM Error-Prone BER Pathway for Application in Bacteria

U:G lesions are typically repaired by the base excision repair pathway (BER). In both prokaryotes and eukaryotes, this mismatch is recognised by the uracil N-glycosylase (UNG) enzyme, which breaks the glycosyl bond linking the nitrogenous base to the sugar-phosphate backbone to generate an apurinic/apyrimidinic site (AP-site). AP-sites and U:G lesions can block or slow down DNA replication via a normal proof-reading polymerase, resulting in stalling of DNA replication[126]. As a result, the AP-sites are recognised and cleaved by an AP-endonuclease (APE1) and subsequent repair by a proof-reading DNA polymerase is performed to prevent such replicative slowdown.

During SHM however, the expression of APE1 is downregulated[129]. This allows B-cells to maintain the AP-site during maturation and during DNA replication, a lesion bypass polymerase needs to be recruited to read through lesions like a U:G mismatch. Lesion bypass polymerases typically lack a proof-reading domain and have a high error-rate[130]. Both these qualities allow such error-prone DNA polymerases to generate nucleotide substitutions at a rate which is ~ 10^6-fold higher than the rate of natural evolution[131]. This two-step process of inducing DNA damage and subsequent error-prone DNA repair enables rapid evolution of antibodies for building an adaptive immune response.

To emulate this natural phenomenon of directed evolution exhibited by B-cells within bacteria, two crucial components were required — a DNA damaging device and an error-prone DNA repair complex.

2.2 A Targeted DNA Damaging Device with AID for Generating Random Deamination Events

2.2.1 Background Research

In recent years, the strive to develop efficient *in vivo* directed evolution system led many to implement natural base editors in their mutator system design. A key challenge with the implementation of such base editors has been to limit their global mutagenic activity and prevent deleterious mutations from accumulating in the host cell genome. Research done on mice showed that when AID acts on non-immunoglobulin loci in B- and -T-cells, it strongly promotes carcinogenesis[132]. Therefore, a fine balance must be struck between generating targeted mutations and maintaining genomic integrity of the host for developing a robust *in vivo* directed evolution system.

In recent years, researcher have attempted to limit the activity of deaminases, like AID, to the target DNA by fusing them to Cas9 and deactivated Cas9 (dCas9) endonucleases utilised in the CRISPR gene-editing system[114,115,133]. In each instance, a guide RNA (gRNA) is used to target mutations to a site complimentary to the gRNA sequence within the gene-of-interest. Typically, gRNA that are 17-20 nucleotides long are used to achieve efficient editing of the target gene[134]. This greatly limits the size of the mutational window for incorporating mutations, requiring prior knowledge of the relationship between the target protein structure and its function to alter specific amino acid residues for improving protein function. Also, multiple gRNAs are needed to subject the entire target gene sequence to mutations by the AID-dCas9 fusion protein. However, this is not ideal for a continuous evolution system as new gRNA sequences would have to be designed for each round of library generation, which is impossible and not economically feasible.

2.2.2 Targeted Deamination Events with AID and T7 Polymerase

Considering the limitations of the techniques described above, a targeted DNA damaging device where AID is fused to the T7 RNA polymerase (T7pol) was developed in previous reseach[121]. This research was aimed at fusing AID to T7pol and testing its ability to target deamination events to a fluorescent reporter gene placed downstream of the T7 promoter (P_{T7}) in *E. coli*[121]. The deaminase was fused at the N-terminal end of T7pol, as the C-terminal amino acid residues are involved in recognising and binding to the P_{T7}[135]. Once the transcription bubble is formed by unwinding the target DNA sequence, AID is provided with a ssDNA template for deaminating dC to dU[136]. As the T7pol continually unwinds DNA until it encounters a transcriptional terminator, the entire DNA sequence enclosed within the P_{T7} and the T7-terminator can be subjected to dC deamination events.

This was proven experimentally by preventing DNA repair of the U:G lesions generated in the GFP gene (integrated into the bacterial genome). The uracil glycosylase inhibitor (UGI) was expressed constitutively along with AID-T7pol (Figure 2.1). UGI blocks UNG’s ability to bind to DNA at the point of mismatch, preventing the formation of an AP-site. During DNA replication, DNA pol-III recognises dU as dT and adds a deoxyadenosine nucleotide to the complimentary strand, resulting in a C → T and

G → A mutation being acquired by a daughter cell after cell division. These mutations were screened using flow cytometry and Sanger Sequencing of the mutant GFP[121].

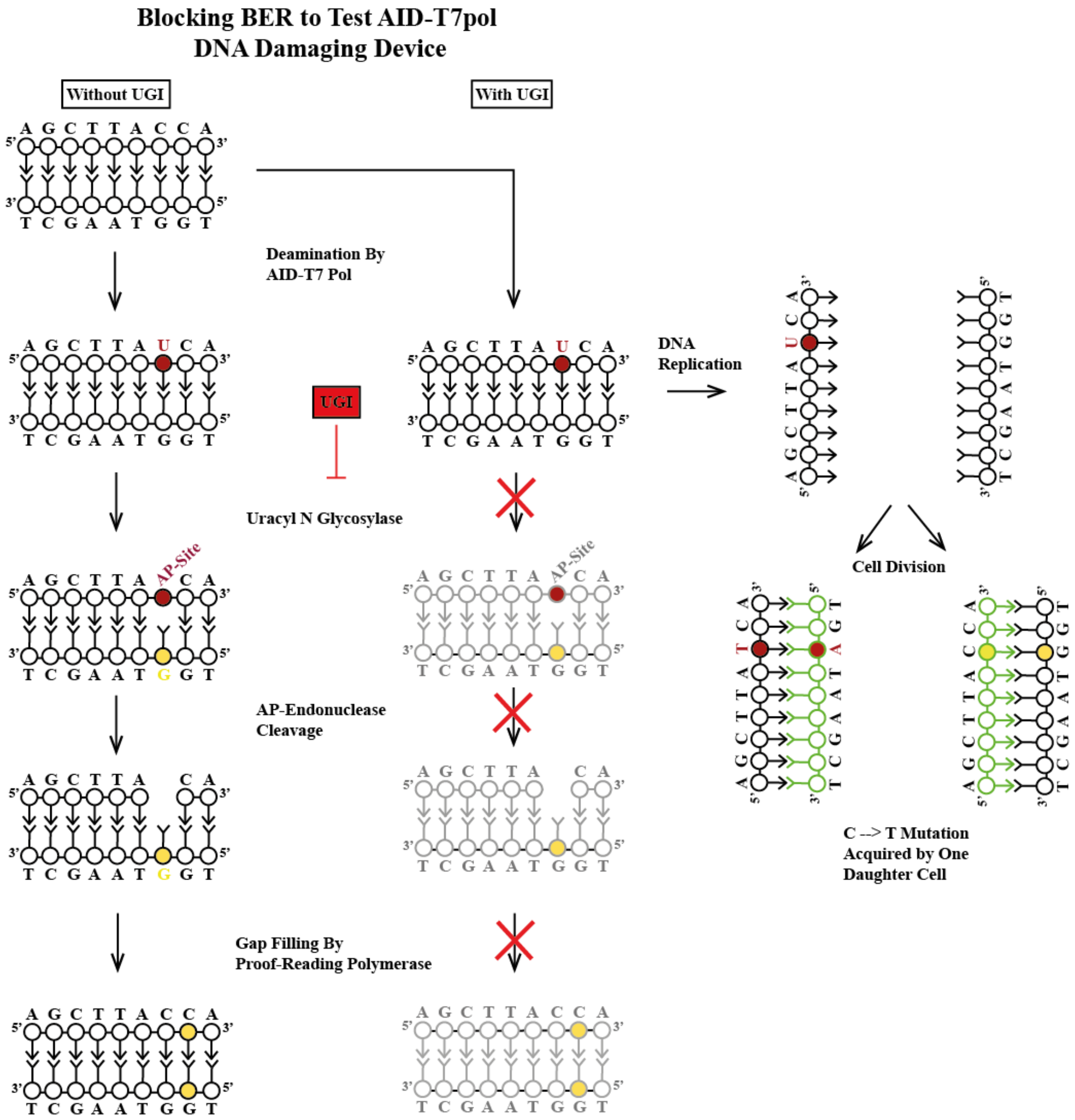


__Figure 2.1: Biological mechanism for testing the AID-T7pol DNA damaging device.__ The ability of AID-T7pol to introduce C → T was tested by blocking the BER DNA repair pathway. The pathway involves the use of DNA glycosylases to remove the damaged base from the site of mismatch. In this instance, Uracil N-glycosylase is blocked by expressing UGI along with AID-T7pol in the host cell. The U:G mismatch created by AID-T7pol in the target gene is prevented from being repaired. During cell division, the DNA polymerase compliments dU with dA, meaning the mutation is passed on to one of the daughter cells.

2.2.3 Research Objective: Optimising the AID-T7pol Targeted DNA Damaging Device

The research discussed in 2.2.2 provided confirmation of AID-T7pol's ability to target mutations to the gene-of-interest via the orthogonal P_{T7} promoter-T7 RNA-polymerase pair. Before implementing this DNA damaging device in a mutator system for continuous evolution, key challenges needed to be addressed.

The first challenge was the toxicity from expressing AID-T7pol and T7 RNA polymerase in the bacterial host cell. Experimentation showed severe growth defects in cells when these proteins were expressed using a high expression RBS and a high-copy number plasmid backbone (pMB1). This limitation was resolved by applying an engineering approach to optimising the expression of these genes in *E. coli*. Characterisation experiments were performed by assembling expression circuits for AID-T7pol and T7 RNA polymerase with a randomised library of RBS sequences in a low copy number plasmid backbone (pSC101).

The second challenge was the lack of understanding involving the DNA damaging device's ability to generate functional mutations. Generating a mutation that creates functional phenotype is more challenging than creating a mutation that knocks-out a phenotype. The device's ability to generate functional mutations was tested using an antibiotic gain of function assay (described in section 2.5.2). The β-lactamase gene was inactivated by a specific point mutation in the ATG start codon, to generate ACG. The device's ability to create a C → T mutation resulted in a functional start codon, which was selected for on LB-agar plates containing carbenicillin.

The third challenge to overcome was to reduce the AID-T7pol device's reliance on UGI to generate mutations. In the absence of the UNG blocker, the U:G mismatches are repaired by the cell's native BER pathway. The presence of UGI blocks DNA repair via BER for all U:G lesions in the cell. This can create deleterious effects in the host cell by increasing the global mutation rate. This is not ideal for a continuous evolution system.

The fourth challenge was the lack of a diverse mutation spectrum. Simply generating C → T of G → A mutations will not result in a wide range of amino acid substitutions, thus restricting the evolutionary search-space that can be explored for protein engineering. The third and fourth limitations of the system were addressed by engineering an error-prone DNA repair complex to synergise with the DNA damaging device in the host cell.

2.3 Engineering an Error-prone DNA repair complex

2.3.1 Background

It is generally accepted in the field that developing a robust continuous evolution system will require the use of an error-prone DNA polymerase to incorporate nucleotide substitutions in the target gene sequence[109,137,138]. However, there is no consensus on how the error-prone DNA polymerase should be implemented in the system. In the PACE system, mutagenic proteins are expressed freely in the bacterial cell and not only induce mutations in the target gene, but also the bacterial genome[139]. OrthoRep utilises an orthogonal error-prone DNA polymerase-Plasmid pair in yeast for targeted mutation of the gene-of-interest every yeast replication cycle. While this mutator system has no mutagenic activity on the host cell genome, its use is limited to yeast strains[140,108].

2.3.2 Bioparts for an Error-prone DNA repair complex to Hijack the BER Pathway in *E. coli*

The aim to develop an error-prone DNA repair complex that would spatially localise the error-prone DNA polymerase to the site of the U:G mismatch generated by AID-T7pol and shield the host cell genome from mutations. As discussed in Section 2.1.1, generating genetic diversity via somatic hypermutation relies on error-prone DNA repair. This requires recognition of the deoxyuridine and its excision to generate an AP-site and subsequent hydrolysis of the abasic site by an AP-endonuclease to create a gap that is filled by an error-prone DNA polymerase — which is the agent of genetic diversity creation. To emulate this in *E. coli*, AP-endonucleases and error-prone DNA polymerases that are functional in bacteria were screened for their ability to work synergistically and reduce the fidelity of BER preferentially for the target gene. Exonuclease-III (Exo-III) and the *N. Meningitidis* AP-endonuclease (NAPE) were tested as potential AP-endonuclease candidates, while DNA polymerase IV (Pol-IV) and error-prone mutants of DNA Polymerase-I[138,141] were tested as potential polymerase candidates.

The main challenge was shielding the host cell genome from error-prone DNA polymerase activity, while maximising the probability of the EP-DNA-polymerase binding to the gaps generated in the target gene by a candidate AP-endonuclease. This was achieved in two ways: creating a fusion of the AP-endonuclease and EP-DNA-polymerase (AP-Endo—EP-DNA-Polymerase); and in the case of DNA Pol-IV, removing amino-acid residues from the C-terminus to prevent it from binding to the β-clamp at the replication fork[142,143].

The AP-endonuclease was fused to the N-terminus of the EP-DNA-polymerase. Within the mutator system, the two fusion proteins (AID-T7pol and AP-Endo—EP-DNA-Polymerase) should work synergistically to induce deamination events in the gene-of-interest, by excising the AP-sites and subsequently performing error-prone DNA repair by hijacking the native BER pathway. The process is identical to native BER until the polymerisation step. Instead of a proofreading polymerase filling in the gap generated by the AP-endonuclease, it will be either DNA Pol-IV or EP-DNA pol-I. AID on its own allows for G:C → A:T transitions but the incorporation of an AP-Endo—EP-DNA-Polymerase fusion protein has the potential to increase the mutational diversity at the U:G lesions (Figure 2.2).

2.3.3 Bioparts for Extending the Mutational Window for Activity of the Error-prone DNA repair complex

Another key objective for developing an efficient error-prone DNA repair complex was the ability to extend mutational window for the AP-Endo—EP-DNA-Polymerase fusion beyond the site of U:G lesions. When generating somatic hypermutations during B- and T-cell maturation, the mutational window is expanded by the mismatch repair pathway recruiting the 5'-3'-Exonuclease-I (Exo-I) and the error-prone DNA polymerase η (Pol-η)[125]. The gap generated by the excision of the AP-site is extended in the 5'-3' direction using Exo-I. This extended gap is filled-in by Pol-η and this process is collectively known as patch repair. Patch repair also introduces a likelihood of mutations being incorporated at A:T sites that are downstream of the U:G mismatch (Figure 2.2).

Three different 5'-3' exonucleases that are functional in bacteria were selected as candidates for building an error-prone DNA repair complex capable of patch repair: RecJ, RecE and the 5'-3' Exonuclease domain from DNA pol-I (5'-3'Pol-I-Exo). To ensure spatial localisation of the exonuclease at the U:G lesions generated by AID-T7pol, the 5'-3' exonucleases were fused to the N-terminus of the AP-Endo—EP-DNA-Polymerase fusion protein to generate a three protein fusion, 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase.

Experimental testing with mutator modules comprising AID-T7pol and 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase confirmed their ability to perform patch repair[iii]. This was tested using a gain of function assay with the inactivated β-lactamase gene as the target, where the mutator module successfully performed an A → T mutation to create functional β-lactamase molecules.

[iii] Results in Chapter 4

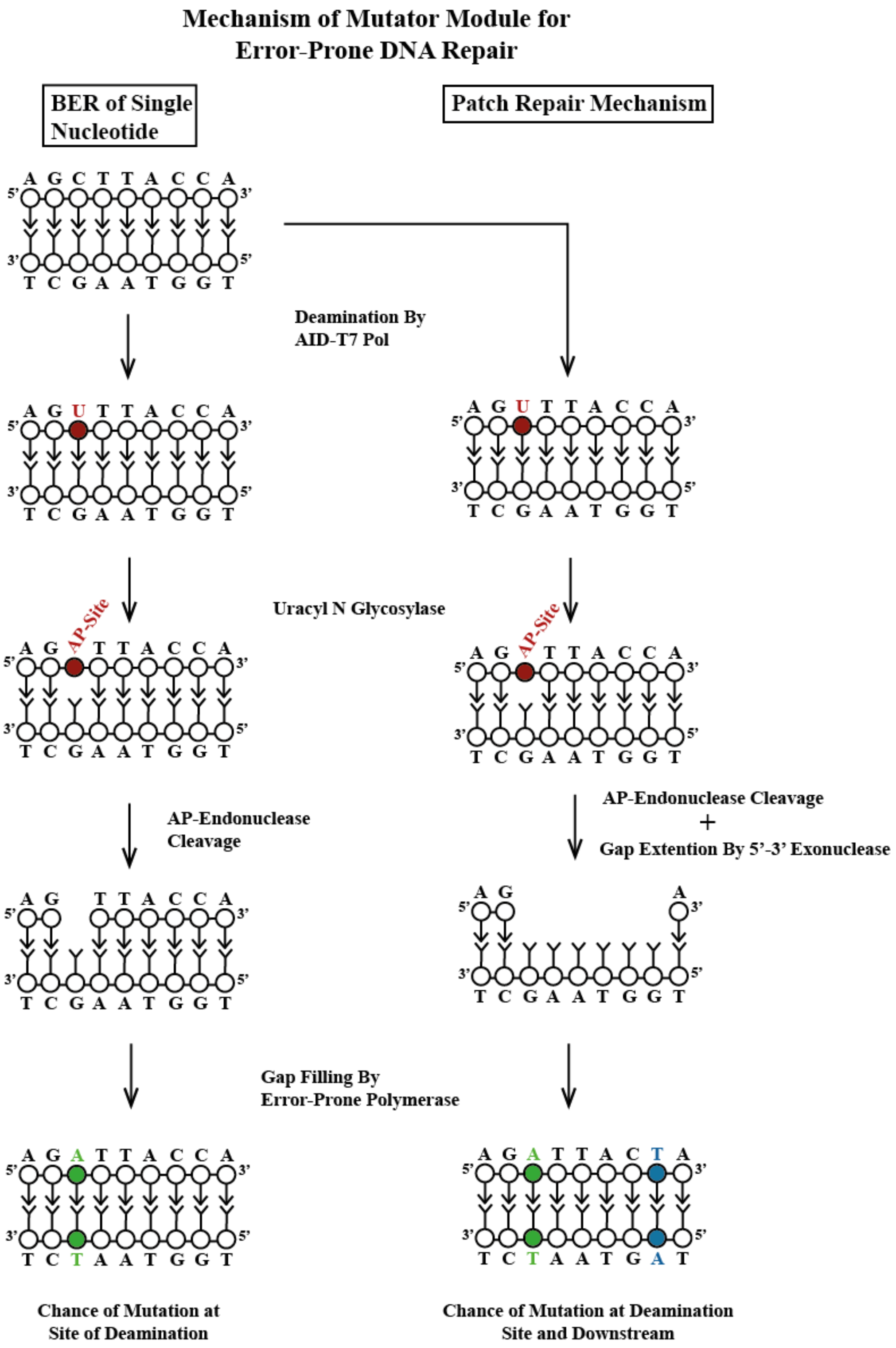


Figure 2.2: Mutator module's mechanism of error-prone DNA repair. *Utilising the AID-T7pol DNA damaging device, U:G lesions are introduced into the target gene. This lesion activates the base excision repair pathway, where an AP-site is generated by UNG. This is where the error-prone DNA repair complex hijacks the pathway to introduce mutations. The AP-endonuclease domain of the 5'-*

3’Exo—AP-Endo—EP-DNA-Polymerase fusion protein excises the abasic site to generate a gap, which is filled in by the error-prone polymerase domain. The error-prone DNA repair complex can also perform patch repair by utilising the 5’-3’ exonuclease domain after AP-endonuclease cleavage to extend the gap. This extended gap is then filled by the error-prone polymerase, increasing the likelihood of mutations being incorporated at A:T sites.

2.3.4 Research Objective: Optimising Biopart Expression, Assembling Error-prone DNA Repair Complexes, and Testing for Mutagenic Activity

The first objective was synthesising and characterising the expression of each Biopart in *E. coli* cells. Expression of complex synthetic gene circuits imposes a heavy burden on cells if poorly optimised[,144,145]. The process of characterisation involves building expression systems for bioparts with a library of promoters and RBS sequences possessing different expression strengths. The expression of the target biopart with each promoter-RBS combination is assessed in correlation to the cell growth rate overtime. Such characterisation allowed us to identify promoter-RBS combinations that enabled adequate expression of the bioparts, while maintaining cell fitness over time. All the biopart candidates for building the error-prone DNA repair complex were characterised in this fashion.

The second objective was assembling an array of mutator modules containing the AID-T7pol DNA damaging device and different versions of the error-prone DNA repair complex. The mutator modules were assembled using the BASIC assembly method[iv], which uses orthogonal linker sequences and a type-IIs restriction enzyme (BsaI) for creating easily modifiable plasmids[146]. A total of 11 different mutator modules were assembled using Pol-IV as the error-prone DNA polymerase, while 8 were assembled with the error-prone polymerase domain of DNA Pol-I.

The third objective was screening the 19 different mutator modules for their ability to induce mutations in the target gene. Two different experimental assays were used for this objective: a loss of function assay with GFP-mut3b as the target gene[v]; and a gain of function assay with inactivated β-lactamase as the target gene.

The final objective was to elucidate the types of mutations, mutation spread and the mutation rate that can be achieved with the mutator modules shortlisted from the loss-of-function and gain-of-function experiments. The next generation sequencing (NGS) methods, Illumina iSeq100 and Pacific Biosciences (PacBio) Sequel were used for achieving this goal.

[iv] Method discussed in Section 2.5.1

[v] Discussed in Section 2.5.2

2.4 An Inducible Expression System for the Mutator Module

Having an inducible expression system for the mutator module is essential for developing a continuous evolution system. Continuous evolution involves multiple cycles of generating a library of mutants and subsequently screening or selecting for functional variants. With an inducible expression system, if a functional variant is not identified, the mutator module can remain active and continue to induce mutations on the target gene. However, once a variant displaying the desired phenotype has been identified, there needs to be a feedback mechanism in place to shut off the mutator module to prevent further mutation of the corresponding genotype before it is isolated from the host cell. Otherwise, the ability to create a link between genotype and the displayed phenotype is lost.

The promoter used to regulate the expression of the DNA damaging device and the error-prone DNA repair complex is a J23101 promoter from the Anderson library, modified to contain tet operator sites (called J23101$_{\text{TetO}}$). This placed the AID-T7pol and 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase fusion proteins under the TetR repression system[147] (Figure 2.3). This creates an inducible expression system, which is switched on using the inducer molecule, anhydrotetracycline (aTc). When aTc is present in the media, it will bind to TetR molecules, alleviating the repression and allowing the two fusion proteins to be expressed. The system will keep inducing deamination events in the GOI with AID-T7pol, while the 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase would introduce mutations through error-prone DNA repair. Once variants with the desired phenotype have been identified using appropriate selection systems, aTc can be removed from the media to shut off the mutator module.

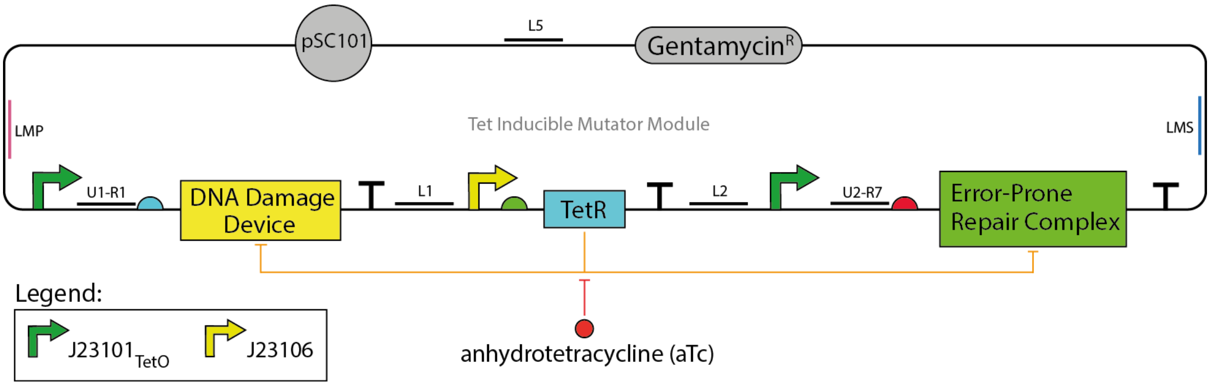


***Figure 2.3: Inducible expression system for the mutator module**. The expression of the DNA damage device and the error-prone DNA repair complex are regulated by the TetR repression system. TetR undergoes conformational change once bound to aTc, and can no longer bind to the operator sites of the modified J23101 Anderson promoter with tet operator[147]. With aTc in the cell growth media, the*

mutator module is activated and can induce deamination events and perform error-prone DNA repair of the target gene.

2.5 Key Experimental Techniques for Building, Testing and Validating the Different Mutator Modules

In this doctoral thesis, three experimental methodologies have been applied throughout: (1) A DNA assembly method to assemble all the expression plasmids used in this study; (2) A loss-of-function assay with GFP-mut3b where the loss of fluorescence is used to monitor the mutagenic activity of the assembled mutator modules; and (3) A gain-of-function experiment with inactivated β-lactamase variants to test the mutator modules for their ability to generate functional mutations. An overview of these methodologies is presented in this section.

2.5.1 BASIC DNA Assembly Method

BASIC is a DNA assembly format where the DNA parts are joined to one another via orthogonal linker pairs[146,148]. Each linker pair is split into a prefix and a suffix half linker, which anneal to one another via a unique 21-bp overhang (Figure 2.5). When the prefix and suffix half linkers of each linker pair are ligated to two different parts, the 21-bp overhangs drive the two parts to anneal to one another in an assembly reaction. This assembly format utilises the type-IIs restriction enzyme, BsaI, and a T4 ligase to attach the linker sequences to the bioparts. The BASIC DNA assembly format enabled simultaneous construction of multiple expression cassettes for the characterisation of all bioparts to be used in building the error-prone DNA repair complex.

Assembly Methodology

The bioparts are synthesised with BsaI recognition and cutting sites flanking the complete sequence of the DNA part. The BsaI site flanking the 5’-end of the part (called iP) produces a different 4-bp overhang to the BsaI site flanking the 3’ end (called iS) (Figure 2.4).

Each half linker in the linker pair is designed to possess different overhangs at its 5’ and 3’ ends: a 21-bp overhang at one end to anneal to its linker pair; and a 4-bp overhang at the other end that is complimentary to either the iP or iS overhang. Within the BASIC linker pairs, the 4-bp overhang of the prefix linker is complimentary to that of the iP site, while the suffix linker is complimentary to the iS site.

The assembly process is divided into three steps: (i) Individual clip reactions for each Biopart to be assembled with appropriate prefix and suffix half linkers using BsaI and T4 ligase; (ii) A DNA clean-up reaction to purify linker-ligated bioparts from the reaction mixture; and (iii) a one-pot annealing reaction where all bioparts are assembled together by the orthogonal half linkers of each linker pair annealing to one another via the 21-bp overhangs[vi].

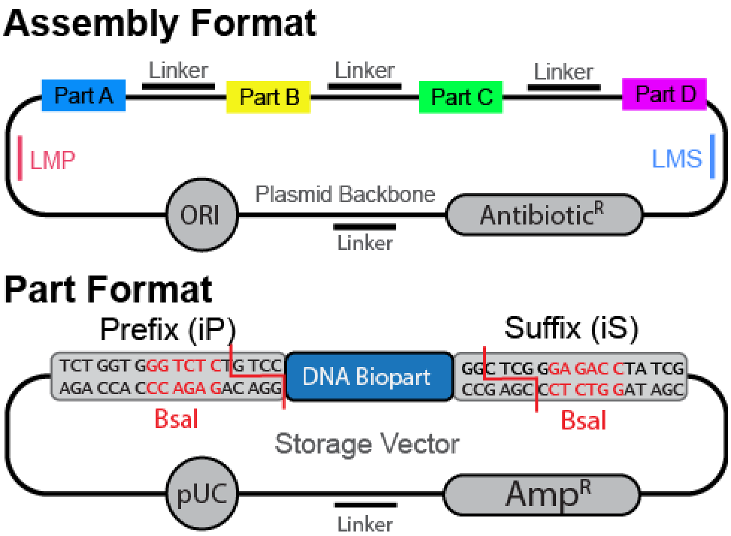


Figure 2.4: Graphic representation of expression plasmids assembled using BASIC.

Biologically Functional linkers

The BASIC linkers can be encoded to perform certain biological functions. RBS linkers contain functional RBS sequences in the 5'-UTR of the prefix half linkers. Fusion linkers were designed for creating fusion proteins. The nucleotide sequence of some fusion linkers translates to a glycine-serine repeat, enabling linkage of two protein domains via a flexible motif. Other fusion linkers translate to rigid motifs containing lysine and proline[vii]. Finally, there are methylated linkers, which contain a functional BsaI restriction enzyme recognition and cutting site. These linkers enable modularity in the assembly process. Assembled expression cassettes that are flanked by the LMP and LMS methylated linkers can be cut from their assembled plasmid and reused for further DNA assemblies downstream.

[vi] Refer to Section 7.3 for the complete protocol

[vii] Refer to Appendix 9.1 for list of BASIC linker sequences

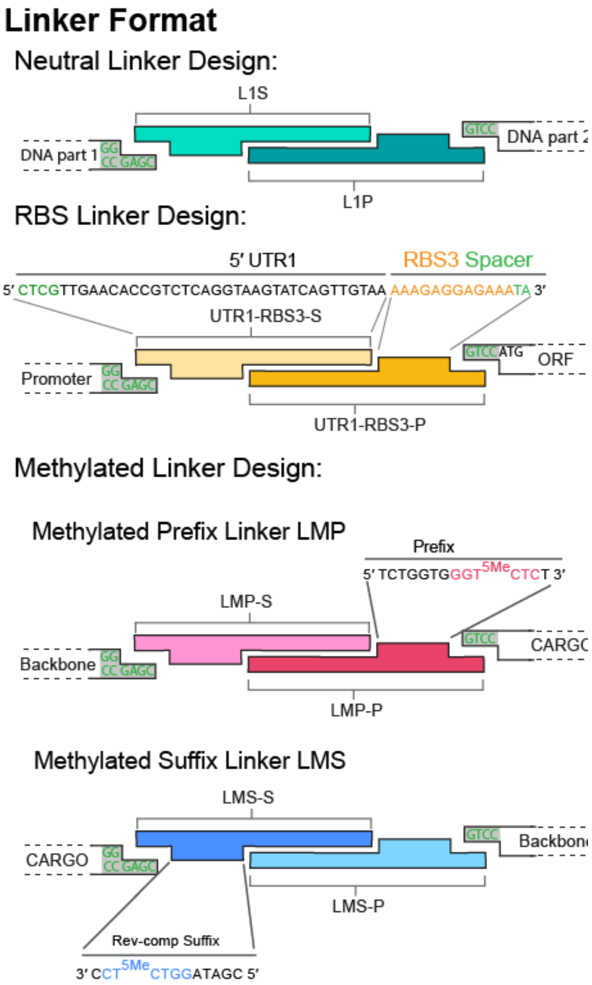


Figure 2.5: Design of BASIC assembly linkers*. Each orthogonal linker pair consists of a prefix and a suffix half linker. Each half linker consists of a 4-bp overhang at one end to anneal to the DNA bioparts and a 21-bp overhang at the other end to anneal to the complimentary half linker during the assembly reaction. Neutral linkers serve no biological function. RBS linkers are designed to contain a functional RBS sequence in the prefix half linker to enabled translation initiation of the downstream open reading frame. Methylated linkers contain a functional BsaI recognition and cutting site. This enables modularity in the assembly process when the assembled DNA construct is flanked by the methylated linkers at the 5′ and 3′ ends.*

Textual Representation of BASIC Assemblies

While SBOL glyphs can be used to represent assembled genetic circuits as diagrams, a textual method of representing DNA assemblies was also required. To represent assembled plasmids in text, the DNA part name written in bold is flanked with BASIC linker abbreviations in subscript[viii]. This allows the reader to easily interpret the parts and the BASIC linkers utilised in an assembly reaction. It follows the following format:

$_{\text{Linker}}$ **Biopart** $_{\text{linker}}$ **Biopart** $_{\text{linker}}$ **Biopart**

Example:

$_{\text{LMP}}$ **Promoter** $_{\text{U1-R1}}$ **GFP** $_{\text{L1}}$ **Terminator** $_{\text{LMS}}$ **Amp$^{\mathbf{R}}$** $_{\text{L2}}$ **pMB1**

This example represents a simple expression cassette for the GFP fluorescent protein (Figure 2.6). The written schematic would enable the reader to setup the appropriate linker ligation reactions with the prefix and suffix half linkers of the linker pairs flanking the bioparts. This written schematic is used to represent assembled plasmids throughout the written report.

Graphical Representation of BASIC Assemblies

The Synthetic Biology Open Language (SBOL) is an open standard created for the representation of *in silico* biological designs[149]. The is an array of globally recognised schematic glyphs that can be used to depict genetic circuits graphically.

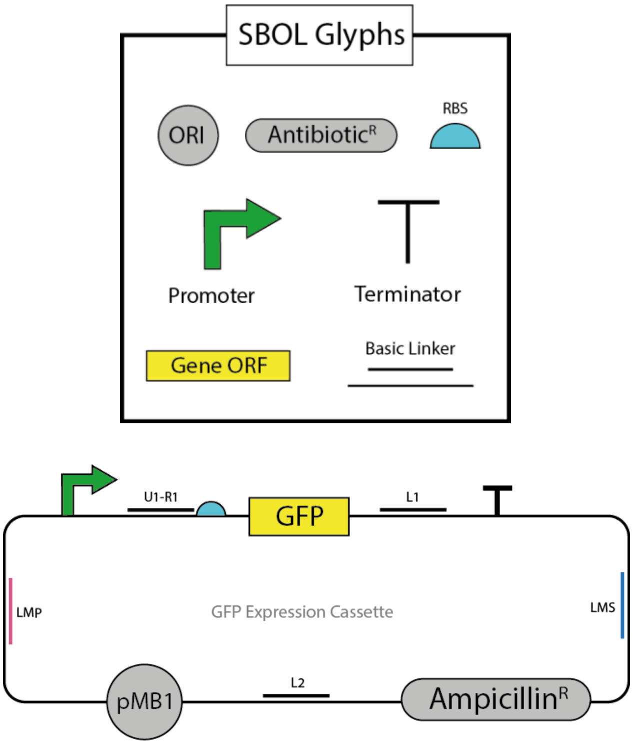


Figure 2.6: Graphical design for representing assembled expression circuits in SBOL format

[viii] Linker abbreviation can be found in Appendix 9.1

2.5.2 Gain of Function and Loss of Function Experiment to Screen Error-prone DNA Repair Complexes for Mutagenic Activity

Once the expression of candidate bioparts was characterised in *E. coli* and the 19 different mutator modules were assembled using BASIC assembly, a screening strategy was needed to assess the mutagenic capability of the 19 different mutator modules and identify the ones capable of performing targeted DNA damage on the gene-of-interest and subsequently generating genetic diversity with the EP-DNA-polymerase.

Loss of Function Assay with GFP

An expression cassette for GFP-mut3b with the P_{T7} promoter and a T7-terminator was integrated into the genome of GM31 strain of *E. coli*[121]. This strain was initially selected for testing the mutator modules as most uracil-initiated DNA repair in GM31 is regulated by UNG, which can be blocked by UGI. The GM31^{GFP} cells were transformed with expression plasmids of the different mutator modules. The degree of fluorescence loss achieved using each mutator module was monitored using flow cytometry. Subsequently, the presence of mutations in the GFP gene was confirmed using Sanger sequencing (Figure 2.7).

This workflow was ideal for a qualitative comparison of the mutator modules by identifying ones displaying strong mutagenic activity from the weak ones. The assessment is qualitative as the limited sample size of Sanger Sequencing restricted the ability to identify all unique nucleotide substitutions that are possible using specific mutator modules. In Chapter 5, a version of this workflow is combined with next generation sequencing to achieve an in-depth analysis of the mutation characteristics of candidate mutator modules.

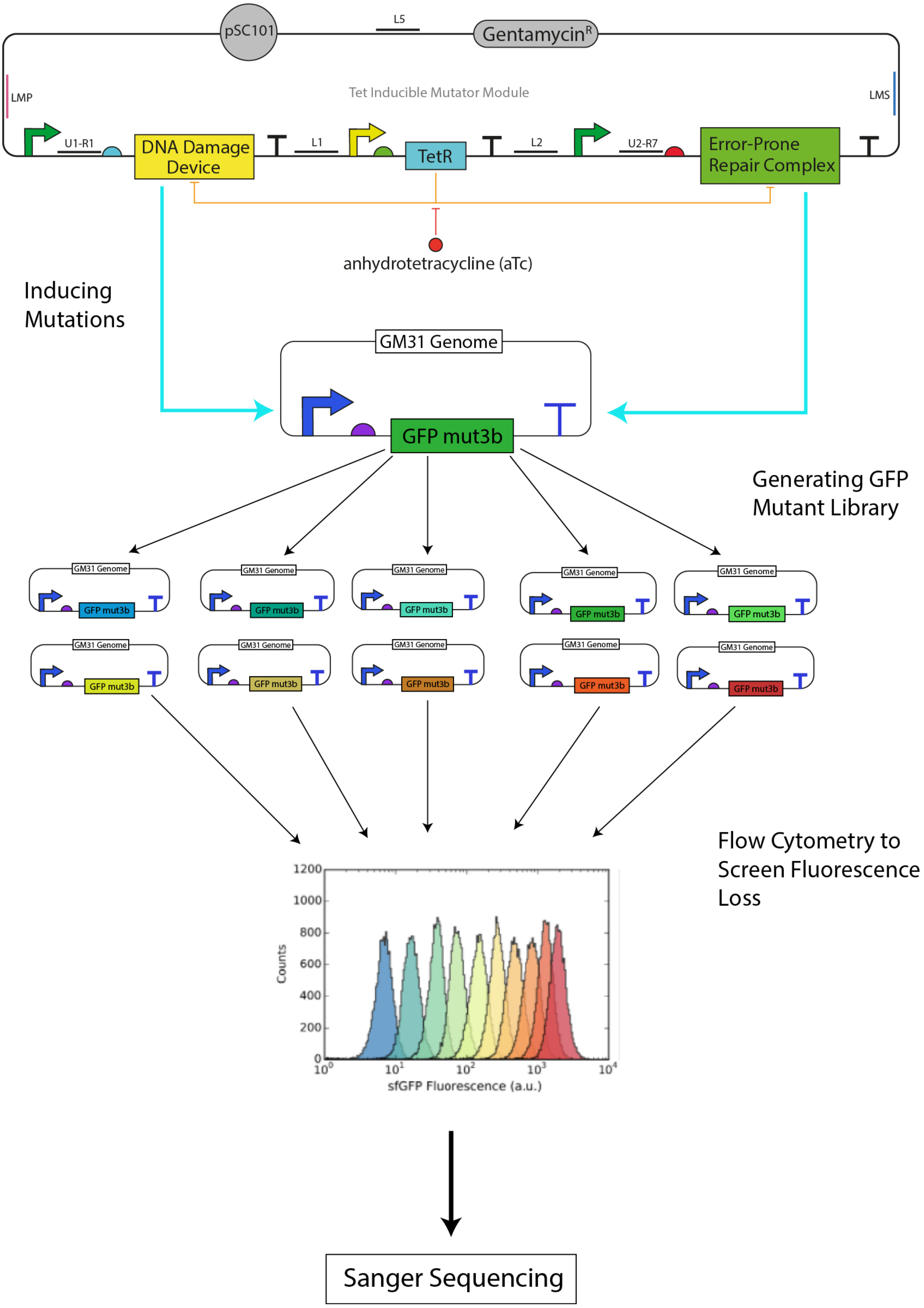


***Figure 2.7: Outline of the loss of function experiment**. The gene for GFP is integrated into the genome of the GM31 bacterial host. The library of mutator modules are transformed into GM31^{GFP} cells and*

activated by adding aTc in the cell growth media. The library of mutant GFP are screened for fluorescence activity using a flow cytometer. The degree of fluorescence loss within a cell population is used to assess the mutagenic strength of individual mutator modules. The mutant GFP ORF resulting from select mutator modules were sequenced to confirmed mutations were the cause of fluorescence loss.

Gain of Function Assay with Inactivated β-lactamase

A gene can be inactivated via single nucleotide substitutions in the start codon or a substitution within the open reading frame (ORF) to generate a premature stop codon. Unique nucleotide substitutions were used to create a library of inactive gene variants. The mutator modules were subsequently used to target mutations to these inactive gene variants. Cells displaying the functional phenotype enabled easy screening of mutator modules that could generate particular nucleotide substitutions.

The target gene that we selected was β-lactamase (amp^R), which confers resistance to ampicillin and carbenicillin antibiotics. To inactivate the expression of β-lactamase, a dC residue was inserted at each position of the ATG start codon, creating ATC, ACG and CTG start codons that are 500-1000 times less likely to initiate mRNA translation compared to ATG[150] (Figure 2.8). AID-T7pol in the mutator modules can only deaminate deoxycytidine, therefore dC was added at each position in the start codon to ensure DNA damage can be generated at the targeted site. This helped to localise the 5'-3'Exo—AP-Endo—EP-DNA-Polymerase activity to the inactivated start codon.

We also wanted to test the mutator system's ability to perform patch repair by synergising 5'-3'-exonuclease activity with the EP-DNA-polymerase function during error-prone DNA repair. The ability to excise nucleotides upstream of the U:G mismatch and create a larger gap to be filled-in by the EP-DNA-polymerase would increase the likelihood of mutations being incorporated at A:T sites. Testing the ability to perform patch repair was achieved by modifying a TTA codon in a GC-rich region of the β-lactamase gene to a TAA stop codon. For successful amp^R phenotype, the 5'-3'-Exo needs to bind to the gap generated by the AP-Endo at a U:G mismatch and excise nucleotides reading into the TAA stop codon. The EP-DNA-polymerase can subsequently add nucleotides from the U:G mismatch site to the TAA stop codon, with a chance of an A → T transversion event occurring.

Cells expressing these inactivated amp^R genes were plated on carbenicillin-containing (50 µgml^{-1}) LB-agar plates and no cell growth was seen, confirming the loss of phenotype because of these nucleotide substitutions. The cells were transformed with different versions of the mutator modules and after 24-hours or greater periods of the mutator being active, the cells would be plated on carbenicillin plates again. The formation of colonies confirmed a mutator modules ability to generate functional mutations and enabled screening for the following nucleotide substitutions:

ACG → ATG   CTG → ATG   ATC →ATG   TAA → TTA

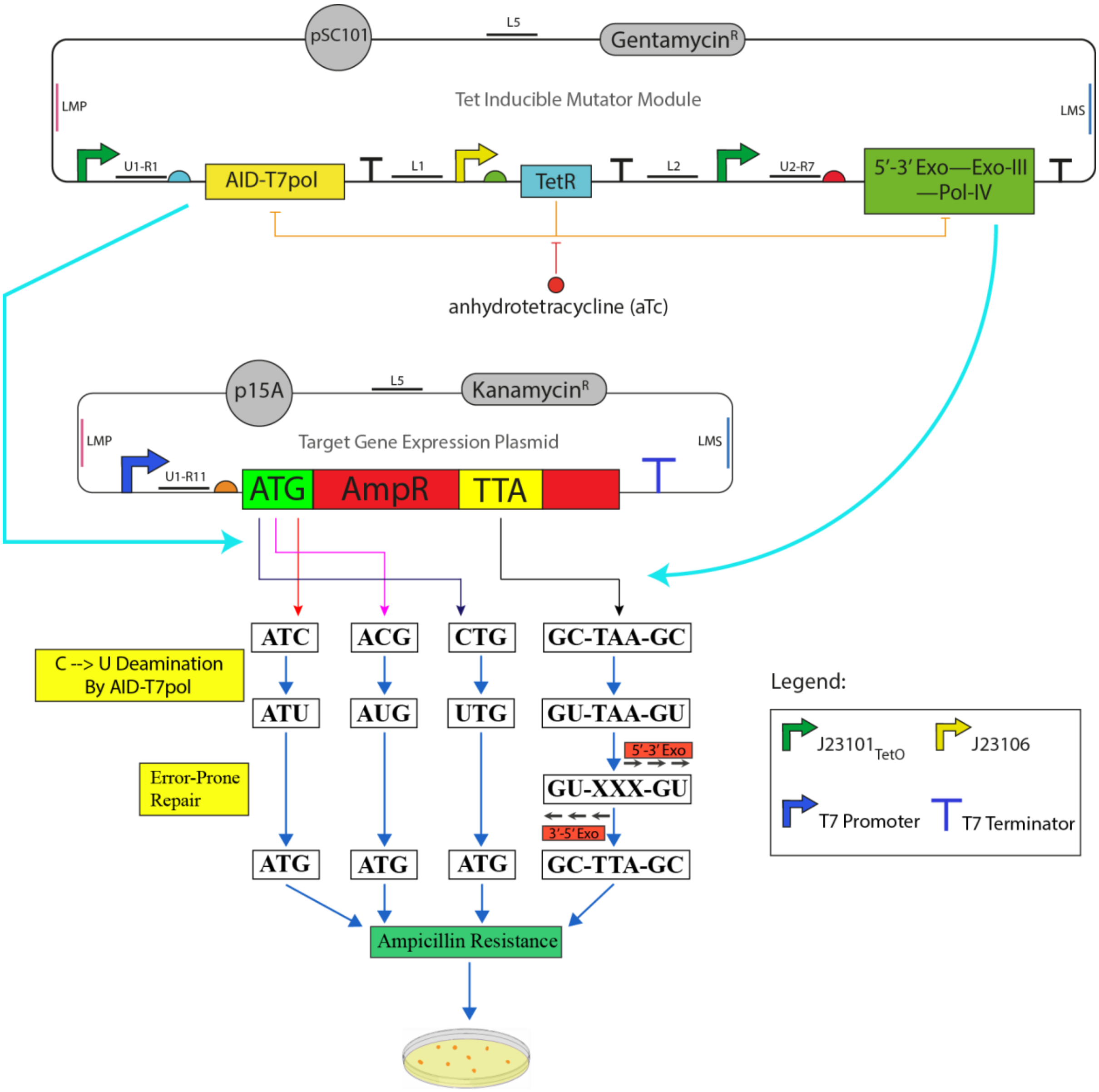


***Figure 2.8: Outline of the gain of function assay**. The β-lactamase gene is made non-functional via point mutations in the ATG start codon and by converting a TTA into a TAA stop codon within the gene. These non-functional β-lactamase genes will be targeted for mutations using the mutator system. If a functional mutant is achieved, it will be selected on carbenicillin plates. Such a selection assay enabled screening for mutagenic activity. ATC, ACG and CTG will require error-prone DNA polymerase activity,*

while TAA will require a 5'-3'-Exonuclease synergising with the error-prone DNA polymerase to perform patch repair for achieving a functional gene.

2.5.3 Next Generation Sequencing to Elucidate the Complete Mutagenic Characteristics of a Candidate Mutator Module

Figure 2.9: Nucleotide Substitutions Investigated with the Gain of Function Assay

Initial Nucleotide	**New Nucleotide**			
	A	G	T	C
A	▨	🟥	🟩	🟥
G	🟥	▨	🟥	🟥
T	🟥	🟥	▨	🟥
C	🟩	🟩	🟩	▨

Figure 2.9: *There are 12 possible nucleotides substitutions that can occur via a mutation. With the gain of function assay, we investigated 4 of the 12, confirming the mutator module can generate a 33% genetic diversity. Next generation sequencing enabled us to identify all possible substitutions that the mutator module can produce.*

Using the GFP loss-of-function and β-lactamase gain-of-function experiments, the mutator module candidates were narrowed down from nineteen to one. The complete spectrum of mutations that can be generated using this mutator module was analysed using two different NGS methods, Illumina iSeq100[151] and PacBio Sequel[152,153].

For NGS analysis, the target GFP-mut3b gene was integrated into the genome of DH5α cells with a dual promoter expression system (J23116 and P_{T7}). The DH5α^{GFP} cells were transformed with a strong, medium and weak version of the candidate mutator module and the mutagenesis of GFP-mut3b was performed for 144-hours. The mutant GFP sequences were isolated using Phusion polymerase high-fidelity PCR and prepped using standard sample preparation protocols for iSeq100 and PacBio Sequel. The acquired sequencing data was analysed using bespoke python scripts implementing the Biopython API, and via 3rd party software[ix].

[ix] Chapter 5 for the complete NGS analysis

2.6 Conclusion

The natural phenomenon of somatic hypermutations found in maturing B- and T-cells to generate antibody diversity is a clear example of an *in vivo* continuous evolution system. The mechanism of generating mutations in the target genes involves inducing DNA damage using the base editor AID, followed by error-prone DNA repair of the damage using an error-prone DNA polymerase. This natural phenomenon provided a possible strategy for developing a continuous evolution system in *E. coli*. To emulate this process in *E. coli*, candidate DNA parts had to be identified and workflows needed to be established to assess the designed mutator systems for their mutagenic characteristics.

Biopart candidates for building an error-prone DNA repair complex were identified and their expression was characterised in the DH5α strain of *E. coli* using different promoter-RBS combinations. This characterisation resulted in expression systems for the AID-T7pol DNA damaging device and 5’-3’-Exo—AP-Endo—EP-DNA-Polymerase EP-DNA-repair complex, where sufficient protein is expressed for mutagenic activity without imposing a heavy fitness burden or being toxic to the host cells.

The loss-of-function and gain-of-function experiments enabled screening of the mutator modules containing AID-T7pol and different versions of the error-prone DNA repair complex based on their mutagenic activity and the ability to generate four specific nucleotide substitutions.

Finally, next generation sequencing was used to build a complete mutation profile that was generated by the short-listed mutator module containing AID-T7pol and 5’-3’Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$ as the error-prone DNA repair complex.

Chapter 3: Optimising, Testing and Validating the AID-T7pol DNA Damage Device

3.1 Introduction

In vivo directed evolution techniques are designed on the principle of inducing DNA damage, which can result in mutations via biological mechanisms such as error-prone DNA repair. Long standing *in vivo* mutagenesis techniques have involved using chemical mutagens or UV-radiation to damage the DNA[154]. Such methods do not reliably introduce mutations in the target gene and have deleterious effects on the host cell genome. Reduced host cell viability restricts the ability to accumulate multiple mutations in the target gene over numerous cell cycles and study long evolutionary trajectories.

The same limitation exists when using mutagenic cell strains, which are achieved by knocking out enzymes involved in DNA repair and upregulating expression of error-prone DNA polymerases[155–157]. Due to the lack of targeting, the global mutation rate in mutator strains of bacteria and yeast is increased drastically, affecting host cell viability for long time-course directed evolution experiments lasting multiple cell cycles.

The search for a targeted DNA damaging device has led many researchers in the field to combine the CRISPR-based targeting system with DNA base editors, like deaminases. By creating a protein fusion of deaminases to dCas9, many examples have been shown for nucleotide substitutions being successfully introduced to the gene of interest, while maintaining the integrity of the host cell genome[158–160]. As discussed in Chapter 1, complete gene coverage with this technique requires the use of multiple gRNA sequences to target the deaminase-dCas9 fusion protein to multiple regions of the target DNA. Generating novel gRNA for every cycle of mutation and selection is not feasible.

This limitation of restricted mutational window was overcome by using the orthogonal T7 RNA polymerase-P_{T7} promoter pair to express the target gene and fusing a deaminase to T7pol[161]. T7pol is a highly processive RNA polymerase which can unwind and read through thousands of kilobases of DNA until it encounters a T7pol-specific terminator sequence[162,163]. This enables entire gene sequences to be subjected to deamination by the deaminase-T7pol fusion protein within one cycle of mRNA transcription[161].

In previous research, a deaminase-T7pol fusion protein was developed by fusing AID to the N-terminal of T7 RNA polymerase. This AID-T7pol fusion protein was shown to successfully perform C → T mutations on a GFP gene, expressed via P_{T7}[121]. However, the expression of T7 RNA polymerase and

AID-T7pol was toxic to GM31 and DH5α cells[121]. In their study, T7-RNA-pol and AID-T7pol were being expressed using the tet inducible J23101$_{\text{TetO}}$ promoter, the BBa_B0030 RBS from the iGEM parts collection and the expression cassette was placed on a high copy number plasmid (pMB1 ORI). Such high expression levels of these proteins stunted the growth rate of cells.

In this chapter, the steps taken to optimise the expression of T7pol and AID-T7pol in *E. coli* are discussed. Optimal RBS sequences were identified for these proteins that enabled sufficient expression for functional activity, while allowing the host cells to maintain a stable growth rate that was comparative to wild-type *E. coli* cells. The optimised expression cassette for AID-T7pol was used to perform a 144-hr mutagenesis experiment on GFP-mut3b targets[x] and assess the accumulation of mutations over time.

In addition to this, the ability of AID-T7pol to induce DNA damage in a target gene was further explored. The gain of function assay described in Section 2.5 was used to assess AID-T7pol's ability to generate C → T mutations and create functional β-lactamase for ampicillin resistance. Finally, the targeting ability of AID-T7pol was assessed using rifampicin antibiotic selection to compare on-target mutations to off-target mutations.

3.2 Using a degenerative RBS library for the optimisation of T7 RNA polymerase and AID-T7pol expression

Synthetic biology involves the forward engineering of biological systems, a process where well characterized, predefined biological DNA parts (bioparts) are combined in a rational manner to achieve a desired behaviour or function[164]. However, even well characterized bioparts can create unpredictable behaviour when introduced into new cellular contexts and very often, when a recombinant pathway is introduced into a cellular host, it can cause metabolic imbalances, which leads to impaired growth and reduced product yields[165]. In their natural context, gene expression cassettes and metabolic pathways possesses regulatory mechanisms to ensure optimal balance between expression of the target genes and the cell's metabolic flux. Such mechanisms are absent in synthetic cellular contexts, which makes optimisation of synthetic gene expression cassettes a prerequisite before building complex genetic pathways in a new cellular context.

A promising technique to optimising the expression rate of target proteins is to alter the translation rate by engineering the RBS sequence[166–168]. As RBS sequences are only 10-15 nucleotides long, a large range of protein expression levels can be explored by making few nucleotide substitutions to generate

[x] Loss of function experiment described in Section 2.5

a library of such RBS sequences. A randomised RBS library can be synthesised by encoding degeneracy into the RBS sequence. This degeneracy is introduced by replacing the standard four base identifiers (A, G, C and T) with IUPAC recognised code for degenerate bases[xi]. Replacing the complete RBS sequence with random bases (series of N's) is not ideal as it can generate up to 1.2x10^{18} RBS combinations to be tested experimentally, which is not possible. Also, many of the RBS sequences from such a vast library tend to be redundant, producing identical expression levels of the target protein. This is overcome by using specially designed algorithms, like RedLibs, which iterates though such vast RBS sequence search space by combining *in silico* and experimental data to generate an optimised degenerate RBS library containing 4-24 unique sequences that can explore a broad range of protein expression levels[169,170].

These techniques enabled the development of linkers used in the BASIC assembly format with a degenerate RBS library encoded into the prefix half-linker (Figure 3.1). Linkers containing a degenerate library of 12, 24 and 36 RBS sequences were designed using RedLibs[146,169]. This means that performing a single DNA assembly reaction with the degenerate RBS linker, a promoter and the target gene created 12, 24 or 36 unique expression systems for this gene. By fusing the target gene to a fluorescent protein (sfGFP), the expression rate achieved by each RBS-Biopart-sfGFP combination could be monitored by picking multiple random colonies and measuring fluorescence using a plate reader. The cell culture growth rate was also monitored over time using the spectrophotometer feature of plate readers. The RBS-Biopart-sfGFP combinations that displayed ideal expression levels of the target protein, without imposing significant burden on cells were identified. The DNA sequence of the RBS in these expression cassettes was elucidated using DNA sequencing methods. This workflow enabled us to optimise the expression rate of T7 RNA polymerase and AID-T7pol in the DH5α strain of *E. coli*.

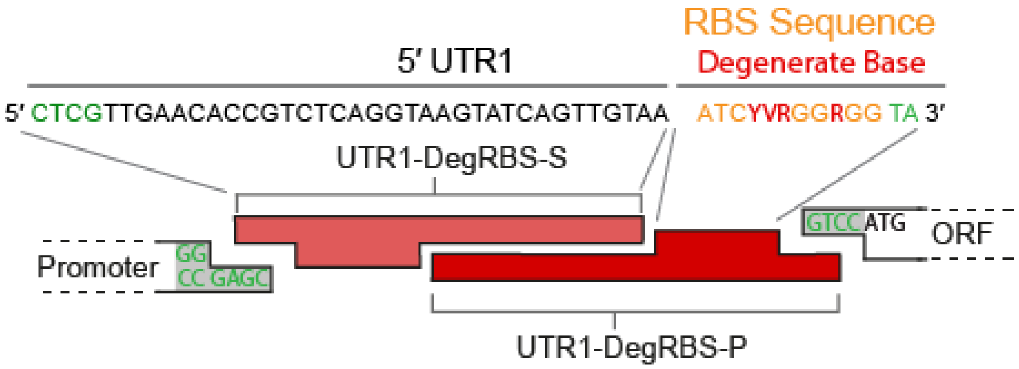


***Figure 3.1**: Degenerative RBS sequence used to characterise T7 RNA polymerase and AID-T7pol Expression*. *Y = C/T, V = A/C/G, R = A/G*

[xi] Table of IUPAC Codes for degenerate bases can be found in appendix 9.2

3.3 Optimisation of T7 polymerase and AID-T7pol Expression using Degenerate RBS Library

3.3.1 Background and Experimental Design

As discussed in section 3.1, expressing T7 RNA polymerase using a strong RBS and on a high copy number plasmid was toxic to host cells. Transferring the expression cassettes to a lower copy number plasmid with the p15A ORI (~15 copies per cell) still resulted in a certain degree of toxicity. Cells expressing T7 RNA polymerase from such plasmids resulted in cell cultures achieving stationary phase at OD$_{600}$ ~ 0.3. Under similar conditions, healthy *E. coli* cells generally achieve an OD$_{600}$ ~ 1.0. This meant that the expression cassettes for the T7 RNA polymerase and AID-T7pol fusion had to be further optimised to alleviate stress imposed on host cells when these proteins are expressed.

The expression of T7pol and AID-T7pol in DH5α cells was characterised using a degenerate RBS BASIC linker, UTR1-DegRBS24. DegRBS24 is was synthesised to contain a common 21-bp overhang for linker annealing and the standard 4-bp sticky ends to ligate to bioparts, with degenerate nucleotides used in the RBS region of the linker to generate 24 unique RBS sequences (Figure 3.1). To visualise the expression of the target protein, the C-terminal ends of T7pol and AID-T7pol were fused to sfGFP. This enabled easy qualitative analysis of the real-time expression of the target protein in a liquid culture using a plate reader. The expression cassette uses the TetR repression system to further alleviate stress on host cells by switching off the expression of T7pol and AID-T7pol when not needed[xii]. The following expression cassettes were assembled to characterise the expression of T7pol and AID-T7pol:

Textual Representation:

1: $_{\text{LMP}}$ **[J23101**$_{\textbf{TetO}}$ $_{\text{UTR1-Deg24}}$ **T7 RNA polymerase** $_{\text{FL2}}$ **sfGFP]** $_{\text{L1}}$ **[TetR]** $_{\text{LMS}}$ **pSC101 Gen**$^{\textbf{R}}$

2: $_{\text{LMP}}$ **[J23101**$_{\textbf{TetO}}$ $_{\text{Deg24}}$ **AID-T7pol** $_{\text{FL2}}$ **sfGFP]** $_{\text{L1}}$ **[TetR]** $_{\text{LMS}}$ **pSC101 Gen**$^{\textbf{R}}$

[xii] Discussed in Section 2.4

Graphical Representation:

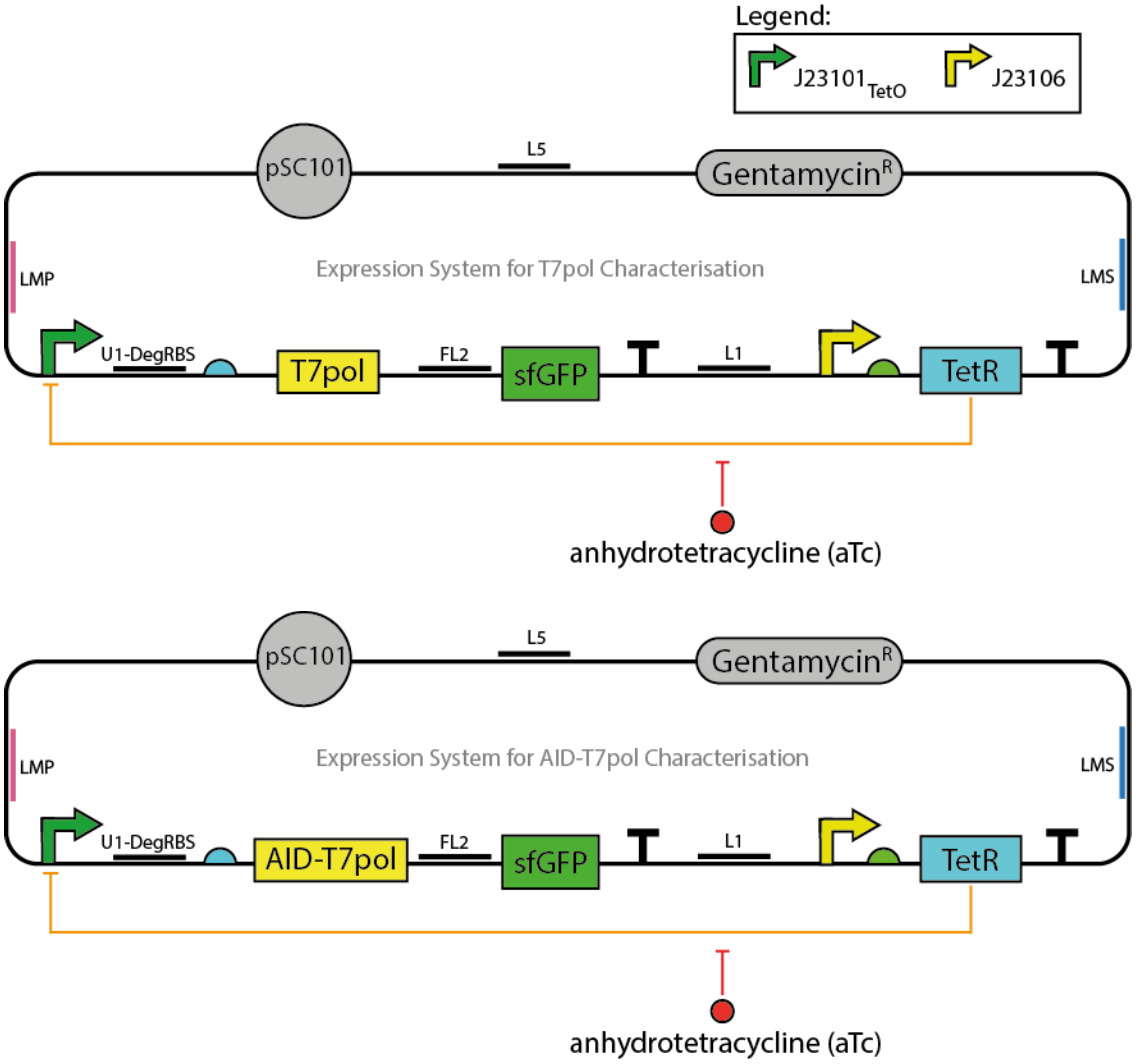

Legend:
J23101TetO
J23106
pSC101
L5
GentamycinR
Expression System for T7pol Characterisation
LMP
LMS
U1-DegRBS
T7pol
FL2
sfGFP
L1
TetR
anhydrotetracycline (aTc)
pSC101
L5
GentamycinR
Expression System for AID-T7pol Characterisation
LMP
LMS
U1-DegRBS
AID-T7pol
FL2
sfGFP
L1
TetR
anhydrotetracycline (aTc)

***Figure 3.2: SBOL schematic of the expression cassette used to characterise the expression of T7 RNA polymerase and AID-T7pol** using the degenerative linker library. The target protein was fused to sfGFP to monitor expression level based on fluorescence output.*

The expression cassettes were characterised in the presence and absence of the inducer molecule, aTc. This enabled monitoring of the stress imposed on the cell when T7pol and AID-T7pol expression is switched on or off and assess the J23101$_{\text{TetO}}$ promoter for leakiness. The burden was assessed by monitoring the real-time change in optical density of the liquid cultures at 600nm. The cells with the degenerative RBS library were placed in 100 ul of LB at a starting OD$_{600}$ of 0.05. The changing OD$_{600}$ and fluorescence were monitored in a plate reader for 6-8 hours, until the cell cultures reached stationary phase. The fluorescence (FL) recorded at each time point is presented relative to the corresponding optical density of the cell culture at the same time point as FL/OD[171].

3.3.2 Results of characterising T7-polymerase expression

The assembled plasmids expressing T7 RNA polymerase via the degenerate RBS linker were transformed in DH5α cells and plated on LB-agar containing gentamycin (25 µgml^{-1}). Eight random colonies were picked from the plate (each containing a different RBS-T7pol-sfGFP combination labelled Deg-RBS A – Deg-RBS H) and cultured with three technical replicates in a 96-well plate overnight. The overnight cultures were diluted to an OD_{600} of 0.05 to start monitoring the fluorescence and absorbance in the plate reader.

Monitoring the absorbance overtime revealed that all the RBS combinations tested exhibited a small burden on the host cells, with cell cultures achieving stationary phase at an OD_{600}~ 0.6 when expression of T7-RNA-pol is switch on with aTc (10 ngul^{-1}) (Figure 3.3). In the absence of the inducer molecule, cells reached stationary phase at a slightly higher OD_{600}~ 0.65. The control cells (DH5a cells with an empty pSC101-Gen^{R} backbone) achieved an OD_{600}~ 0.8. This difference in optical density of the control cells and cells with T7pol expression cassette switched off can possibly be attributed to slight leakiness of the J23101$_{TetO}$ promoter, imposing some burden even when T7-RNA-pol expression is switched off (Figure 3.3, D). The FL/OD of the control cells was ~ 300 au for the complete 360 minutes of the experiment. For cells with T7-RNA-pol expression switched off, the FL/OD was 20,000 au at the beginning of the experiment, indicating a basal level of GFP expression. However, as the cell culture's optical density increased overtime, the FL/OD decreased hyperbolically. By the end of the time course, the FL/OD for all the RBS-T7pol-sfGFP combinations averaged at ~ 5000 au. Witnessing this trend confirmed slight leaky expression of the T7pol-sfGFP fusion via the J23101$_{TetO}$ promoter.

Despite the leaky expression, these expression cassettes are still able to maintain a strong switch-like behaviour with the 8 degenerative RBS linkers tested. When aTc is present in the growth medium, the T7pol-sfGFP fusion is released from TetR repression with a 6-fold to 9-fold increase in fluorescence intensity compared to the samples with no aTc in the media (Figure 3.3). The highest fluorescence was exhibited by the cells during the log phase of bacterial growth (200-minutes into the cell growth assay), with an FL/OD ~ 90,000 au for the strongest RBS (RBS-B) and ~ 20,000 for the weakest RBS in the library (RBS-A). The peak fluorescence decreased linearly by 10,000 au from the log phase to the stationary phase for each RBS-T7pol-sfGFP combination. This behaviour could possibly be the result of ribosome content increasing faster than cell mass when cells exhibit fast growth rates[172–174]. More ribosomes mean increased translation of the T7pol mRNA and therefore higher expression is recorded. After log phase, even though bacterial cells continue to replicate, no more resources can be allocated to target gene expression, leading to a gradual drop in fluorescence relative to the optical density of the cell cultures.

Identifying the DNA sequence of selected RBS via Sanger Sequencing

Analysing the effects on cell culture growth and the fluorescence patterns produced by each RBS enabled identification of expression systems that produced the T7-RNA-pol at high, medium, and low levels without being toxic to the host cells. The shortlisted RBS that enabled these varying ranges of T7pol expression were Deg-RBS-B, Deg-RBS-H and Deg-RBS-F respectively (Table 3.1). The expression plasmids with these RBS were purified and sequenced. The nucleotide sequences of RBS-B, RBS-H and RBS-F were subsequently synthesised as BASIC RBS linkers and added to the BASIC linker library for building future expression cassettes with T7-RNA-pol.

Table 3.1: RBS sequences identified from Degenerative RBS screen for optimised T7 RNA polymerase Expression in DH5α

Name	RBS sequence	RBS Strength
Deg RBS-B	TCCGGGGAGG	Strong
Deg RBS-C	TCCCGGGGGG	Weak
Deg RBS-F	TCTGGGGGGG	Weak
Deg RBS-H	TCCCGGGAGG	Intermediate

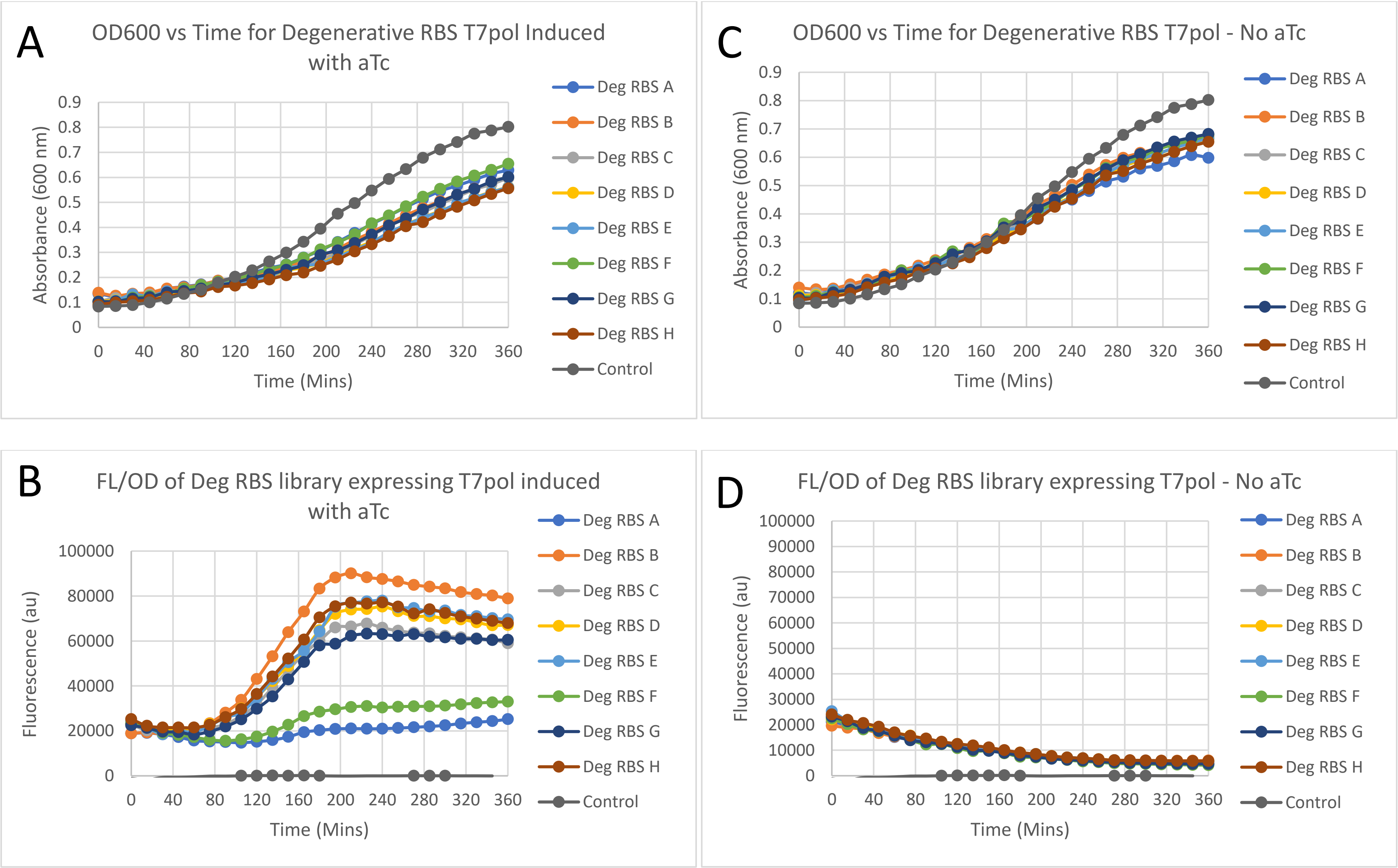

A
OD600 vs Time for Degenerative RBS T7pol Induced with aTc
Absorbance (600 nm)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
40
80
120
160
200
240
280
320
360
Time (Mins)
Deg RBS A
Deg RBS B
Deg RBS C
Deg RBS D
Deg RBS E
Deg RBS F
Deg RBS G
Deg RBS H
Control
C
OD600 vs Time for Degenerative RBS T7pol - No aTc
Absorbance (600 nm)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
40
80
120
160
200
240
280
320
360
Time (Mins)
Deg RBS A
Deg RBS B
Deg RBS C
Deg RBS D
Deg RBS E
Deg RBS F
Deg RBS G
Deg RBS H
Control
B
FL/OD of Deg RBS library expressing T7pol induced with aTc
Fluorescence (au)
0
20000
40000
60000
80000
100000
0
40
80
120
160
200
240
280
320
360
Time (Mins)
Deg RBS A
Deg RBS B
Deg RBS C
Deg RBS D
Deg RBS E
Deg RBS F
Deg RBS G
Deg RBS H
Control
D
FL/OD of Deg RBS library expressing T7pol - No aTc
Fluorescence (au)
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
0
40
80
120
160
200
240
280
320
360
Time (Mins)
Deg RBS A
Deg RBS B
Deg RBS C
Deg RBS D
Deg RBS E
Deg RBS F
Deg RBS G
Deg RBS H
Control

***Figure 3.3: Characterisation of T7 RNA polymerase expression in DH5α using a degenerative RBS library**. The expression cassette utilised the Tet-repression system, where anhydrotetracycline (aTc) inactivated the repressor and allowed expression of the T7pol—sfGFP protein. A and B – Cell culture growth rate was monitored over time and the average OD_{600} of biological triplicates is plotted at each time point. Whether the expression system was in the on- or off-state, the cell cultures reached stationary phase at an OD_{600} ~ 0.65, achieving 20% lower optical density than the cultures containing wild-type DH5α cells. C and D – The relative fluorescence (FL/OD) indicating the level of T7pol expression achieved with each RBS is shown. The average fluorescence for biological triplicates was divided by the average OD_{600} at each timepoint to obtain the relative fluorescence. $J23101_{TetO}$ promoter enabled a switch-like mechanism in the log and stationary phases, where in the presence of aTc, AID-T7pol was expressed and in the absence of aTc, the expression system was switched off. But during the lag phase, leaky expression was seen from the $J23101_{TetO}$ promoter.*

3.3.3 Results of characterising AID-T7pol expression:

Similar to the T7-RNA-pol characterisation workflow, the assembled expression systems for AID-T7pol-sfGFP with UTR1-Deg24 RBS linker were transformed in DH5α cells and plated on LB-agar containing gentamycin antibiotic. Twelve random colonies were picked from the plate (each containing a different RBS-T7pol-sfGFP combination labelled Deg-RBS-1 – Deg-RBS-12) and cultured with three technical replicates in a 96-well plate overnight. The overnight cultures were diluted to an OD_{600} of 0.05 to monitor the fluorescence and absorbance of growing cell cultures in the plate reader.

Whether the expression cassette for AID-T7pol-sfGFP was switched on with aTc or in an off state, the cells reached stationary phase in 6-hours at an OD_{600}~ 0.8, similar to the control cells expressing an empty pSC101-Cam backbone (Figure 3.4). This indicates that expressing the AID-T7pol fusion protein was less toxic to the host cells, as compared to expressing native T7-RNA-pol. The cells expressing AID-T7pol via Deg-RBS-1 exhibited a better growth rate than the control (Figure 3.4).

The expression cassettes for AID-T7pol displayed significantly less leaky expression in the off state compared to T7-RNA-pol, with the basal fluorescence intensity (FL/OD ~ 2000 au) being closer to the control cells (FL/OD ~ 200). This could be the result of differences in the context dependency between the promoter, RBS and AID-T7pol ORF, compared to context dependency of the T7pol ORF in an identical expression system[175,176]. Interactions between the T7pol ORF, the RBS and the 5'-untranslated region (5'-UTR) upstream of the promoter are factors that can affect promoter strength and stringency[177]. Context dependency could be enabling easier transcription and translation initiation of the T7-RNA-pol ORF, while the process is more tightly controlled for AID-T7pol. Overall, this lower leaky expression means the expression of the AID-T7pol DNA damaging device can be tightly regulated by the TetR system.

Presence of inducer molecules enabled a 10-fold to 20-fold increase in fluorescence intensity compared to the cells where the expression system was switched off. Like the T7pol-sfGFP expression cassette, the maximum FL/OD was achieved in the log phase for all samples around 150-minutes into the growth assay. The peak FL/OD recorded for each RBS-AID-T7pol combination tapered off at 40,000 au when the samples reached stationary phase. This means the greatest deviation in RBS strength was witness in the log phase, while all RBS sequences achieved similar expression levels once the cell cultures achieved stationary phase.

Identifying DNA sequence of RBS via Sanger Sequencing

Like the characterisation experiment for T7-RNA-pol, RBS sequences resulting in high, medium and low expression levels of AID-T7pol were purified, sequenced and synthesised as BASIC RBS linkers (Table 3.2). These RBS linkers would be utilised in building and tuning the expression of AID-T7pol in the mutator module. The different expression levels would not only help to regulate the burden imposed on cells, but also alter the mutation rate that can be achieved with the mutator system[xiii]. Higher expression of AID-T7pol would mean more deamination events on the target gene over time.

Table 3.2: RBS sequences identified from RBS screen for optimised AID-T7pol Expression in DH5α

Name	RBS sequence	RBS Strength
Deg RBS-1	TCCCGGGGGG	Weak
Deg RBS-6	TCAGGGGGG	Intermediate
Deg RBS-7	TCCGGGGGGG	High

[xiii] Point explored further in Chapter 4, section 4.6.1

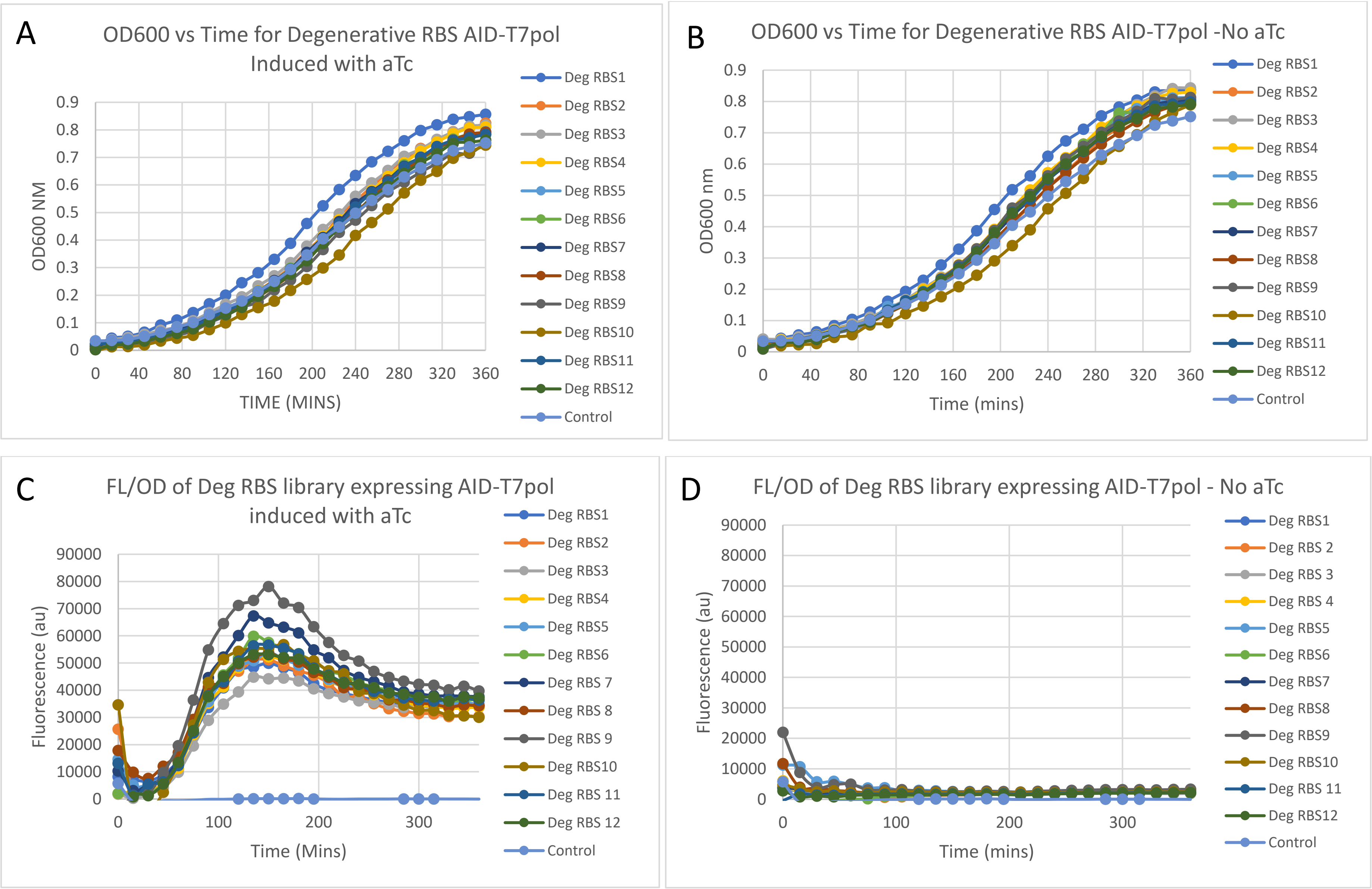

A
OD600 vs Time for Degenerative RBS AID-T7pol Induced with aTc
OD600 NM
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
40
80
120
160
200
240
280
320
360
TIME (MINS)
Deg RBS1
Deg RBS2
Deg RBS3
Deg RBS4
Deg RBS5
Deg RBS6
Deg RBS7
Deg RBS8
Deg RBS9
Deg RBS10
Deg RBS11
Deg RBS12
Control
B
OD600 vs Time for Degenerative RBS AID-T7pol -No aTc
OD600 nm
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
40
80
120
160
200
240
280
320
360
Time (mins)
Deg RBS1
Deg RBS2
Deg RBS3
Deg RBS4
Deg RBS5
Deg RBS6
Deg RBS7
Deg RBS8
Deg RBS9
Deg RBS10
Deg RBS11
Deg RBS12
Control
C
FL/OD of Deg RBS library expressing AID-T7pol induced with aTc
Fluorescence (au)
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
0
100
200
300
Time (Mins)
Deg RBS1
Deg RBS2
Deg RBS3
Deg RBS4
Deg RBS5
Deg RBS6
Deg RBS 7
Deg RBS 8
Deg RBS 9
Deg RBS10
Deg RBS 11
Deg RBS 12
Control
D
FL/OD of Deg RBS library expressing AID-T7pol - No aTc
Fluorescence (au)
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
0
100
200
300
Time (mins)
Deg RBS1
Deg RBS 2
Deg RBS 3
Deg RBS 4
Deg RBS5
Deg RBS6
Deg RBS7
Deg RBS8
Deg RBS9
Deg RBS10
Deg RBS 11
Deg RBS12
Control

***Figure 3.4: Characterisation of AID-T7pol expression in DH5α using a degenerative RBS library**. The expression cassette utilised the Tet-repression system, where anhydrotetracycline (aTc) inactivated the repressor and allowed expression of the AID-T7pol—sfGFP protein. A and B – Cell culture growth rate was monitored over time and the average OD_{600} of biological triplicates is plotted at each time point. Whether the expression system was in the on- or off-state, the cell cultures reached stationary phase at an OD_{600} ~ 0.8, similar to the DH5α control cells. C and D – The relative fluorescence (FL/OD) indicating the level of AID-T7pol expression achieved with each RBS is shown. The average fluorescence for biological triplicates was divided by the average OD_{600} at each timepoint to obtain the relative fluorescence. $J23101_{TetO}$ promoter enabled a switch-like mechanism, where in the presence of aTc, AID-T7pol was expressed. In the absence of aTc, the expression system was switched off.*

3.3.4 Summary of T7pol and AID-T7pol Characterisation Experiments

The characterisation workflow allowed us to identify multiple RBS sequences for achieving stable expression of T7-RNA-pol and AID-T7pol in a low-copy-number plasmid expression system (pSC101 ORI). In the case of both, T7-RNA-pol and AID-T7pol, the RBS selected for building expression cassettes allowed for high, medium and low expression levels of the genes, while imposing minimal fitness buden. Optimisation of the expression of mutagenic proteins was essential for developing a continuous evolution system, which involves keeping the mutator system active for multiple cell cycles until protein variants with the desired phenotype are achieved. In our experimentation, the RBS-tuned expression systems allowed the cell cultures to consistently reach an OD_{600} ~ 1.5 after 24 hours of induction with aTc in 1 ml cultures. Assuming there are ~ $8x10^{8}$ *E. coli* cells in a 1 ml liquid culture at OD_{600} of 1.0, this means that a library of over a billion protein variants could potentially be generated within 24-hours (considering each cell in the culture contains a unique variant of the wildtype gene). Therefore, using characterisation principles, the expression of the DNA damage device could be optimised for DH5α cells, which meant generating a larger mutant library of the target gene within 24-hours. Additionally, the improved cell fitness allows cells to grow for multiple cell cycles, increasing genetic diversity further over multiple cell generations. Furthermore, reducing the toxicity means that escape mutants are less likely to accumulate since knocking out T7pol or AID-T7pol expression will no longer be needed for a fitness advantage[178].

3.4 Assessing optimised AID-T7pol expression system for targeted mutagenesis

A characteristic feature of continuous evolution systems is the ability to target mutations to the gene-of-interest over long time-periods, lasting weeks, even months[108,137,179]. This enables researchers to investigate long evolutionary trajectories and also study the combinatorial effect of different mutations on the target protein, i.e., sign epistasis[180]. To achieve this objective, the mutator system with AID-T7pol needs to continually induce damage in the gene-of-interest overtime. More events of DNA damage should result in more chances for the error-prone DNA repair complex to generate genetic diversity[181]. AID-T7pol's ability to accumulate dC → dU deamination events was tested by performing a 144-hour mutagenesis assay with GFP-mut3b as the target gene.

Methodology:

The loss of function experiment with GFP-mut3b[xiv] was used to assess the mutagenic activity of AID-T7pol over 144-hours. UGI was added to the expression system to ensure the U:G lesions are not

[xiv] Loss-of-function workflow described in Section 2.5.2

repaired by the cells native BER pathway[xv]. The GFP-mut3b ORF was integrated into the DH5α genome in an expression cassette with a J23116 + P_{T7} dual promoter and a double T7-terminator (Figure 3.5). The constitutive Anderson promoter would ensure expression of GFP-mut3b variants when the mutator system is shut off. These cells were transformed with the mutator (AID-T7pol expressed via Deg RBS-6, plus UGI) and control plasmids (T7 RNA polymerase expressed via Deg RBS-H) and grown in LB medium containing aTc for 144-hours. Cell cultures were diluted in fresh media every 24-hours and analysed for loss of fluorescence using flow cytometry. For cells that displayed a loss of fluorescence, the mutant GFP-mut3b ORF was isolated and sequenced at the 24-hour and 144-hour time-points to assess for an accumulation of C → T and G → A mutations.

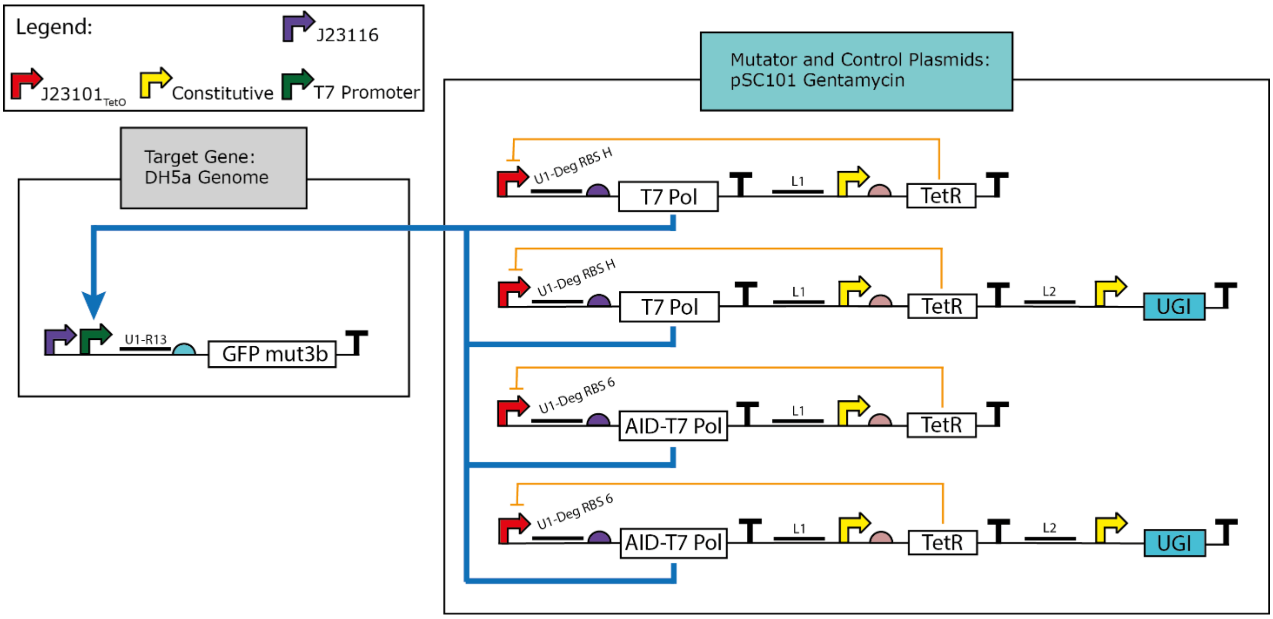


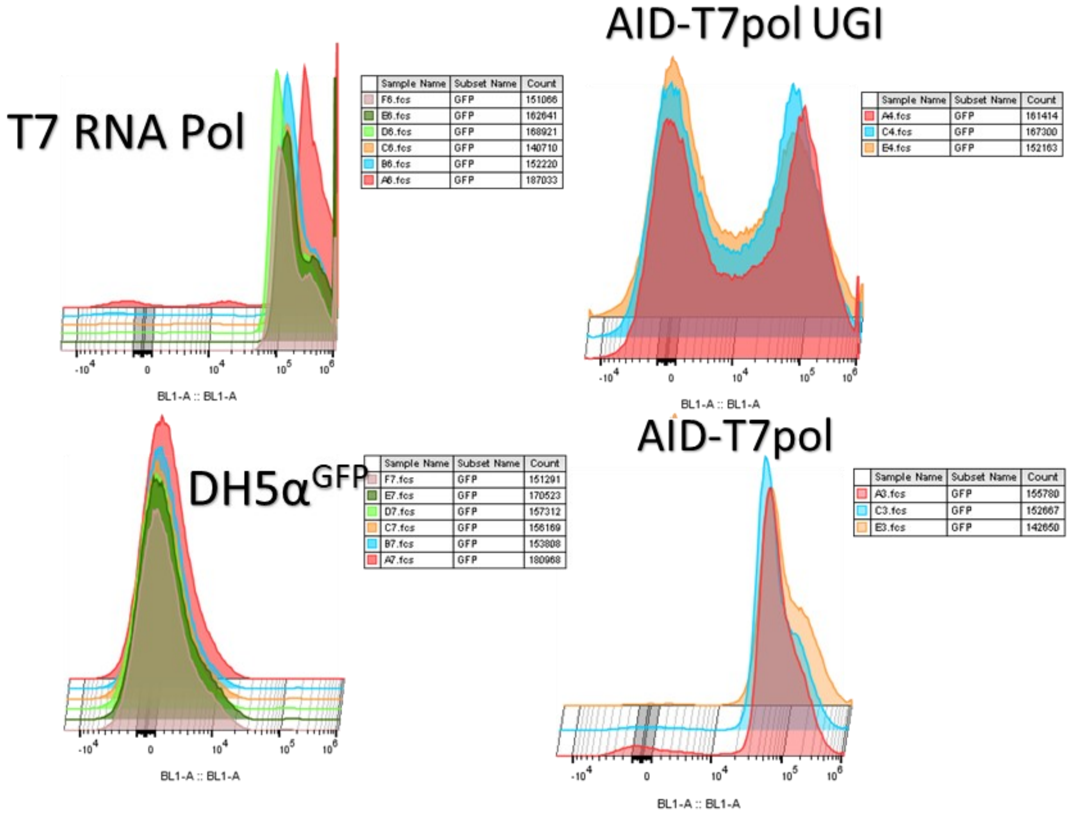
Figure 3.5: SBOL schematic of the long time-course loss-of-function experiment with AID-T7pol+UGI and selected controls.

Results:

Analysing the cell culture samples using flow cytometry revealed vital information about the inducible protein expression system used to express AID-T7pol in the cells. Flow cytometry data was recorded at 24-,96- and 144-hours of the mutagenic assay being active. At all the time points, when no UGI is expressed in the system, there is no loss of fluorescence. This indicates that the cell's native BER pathway is active and repairing U:G lesions generated by AID-T7pol (Figure 3.6).

[xv] Refer to Figure 2.1 in Chapter 2

The GFP phenotype was observed in over 99% of the cell populations for samples not expressing UGI. With UGI in the system along with AID-T7pol, the cell population was split into fluorescent and non-fluorescent. 144-hours into the mutagenic assay, nearly 50%-60% of the cell populations expressing AID-T7pol UGI exhibited loss of the GFP phenotype. This ratio of GFP$^-$ - GFP$^+$ cells was maintained throughout the course of the mutagenic assay, which made it difficult to assess if mutations were being accumulated overtime at a population level. As a result, GFP mutants were amplified via high-fidelity PCR and cloned into a high-copy number backbone (pUC-Amp^R) for DNA sequencing.



	Sample Name	Subset Name	Count
▨	F6.fcs	GFP	151066
🟩	B6.fcs	GFP	162641
🟩	D6.fcs	GFP	168921
🟧	C6.fcs	GFP	140710
🟦	B6.fcs	GFP	152220
🟥	A6.fcs	GFP	187033

	Sample Name	Subset Name	Count
🟥	A4.fcs	GFP	161414
🟦	C4.fcs	GFP	167300
🟧	E4.fcs	GFP	152163

	Sample Name	Subset Name	Count
▨	F7.fcs	GFP	151291
🟩	E7.fcs	GFP	170523
🟩	D7.fcs	GFP	157312
🟧	C7.fcs	GFP	156169
🟦	B7.fcs	GFP	153808
🟥	A7.fcs	GFP	180968

	Sample Name	Subset Name	Count
🟥	A3.fcs	GFP	155780
🟦	C3.fcs	GFP	152667
🟧	E3.fcs	GFP	142650

***Figure 3.6: Flow cytometry analysis of mutator activity at three time-points**. The loss of fluorescence in a population of 50,000 cells was analysed at 72-, 96-, and 144-hours into the mutator assay with AID-T7pol + UGI and the AID-T7pol, T7-RNA-Pol controls. Loss of fluorescence was only witnessed in cell populations expressing AID-T7pol with UGI, which blocks the BER DNA repair pathway. The three time points are shown in red, cyan and orange, respectively.*

Sequencing GFP-mut3b after 24-hours of the mutagenic assay:

To investigate the mutator module’s ability to accumulate mutations in the target gene over time, it needed to be sequenced at a minimum of two time points during the mutagenic assay. 24-hrs and 144-hrs were selected as the timepoints as the sequenced reads should display the least and most mutations at these timepoints, respectively. Five GFP-mut3b mutants were selected at random and sequenced using a forward and a reverse primer — designed to prime upstream of the J23116 + P_{T7} dual promoter and downstream of the T7 terminator, respectively. These primers enabled sequencing of the GFP expression cassette from both ends for full coverage. After 24-hours, four of the mutants lost GFP fluorescence due to a single C → T or G → A nucleotide substitution in the open reading frame. The mutations were identified in both the forward and reverse sequencing primers, which validates the nucleotide substitution as a mutation and not an error generated by the sequencing platform.

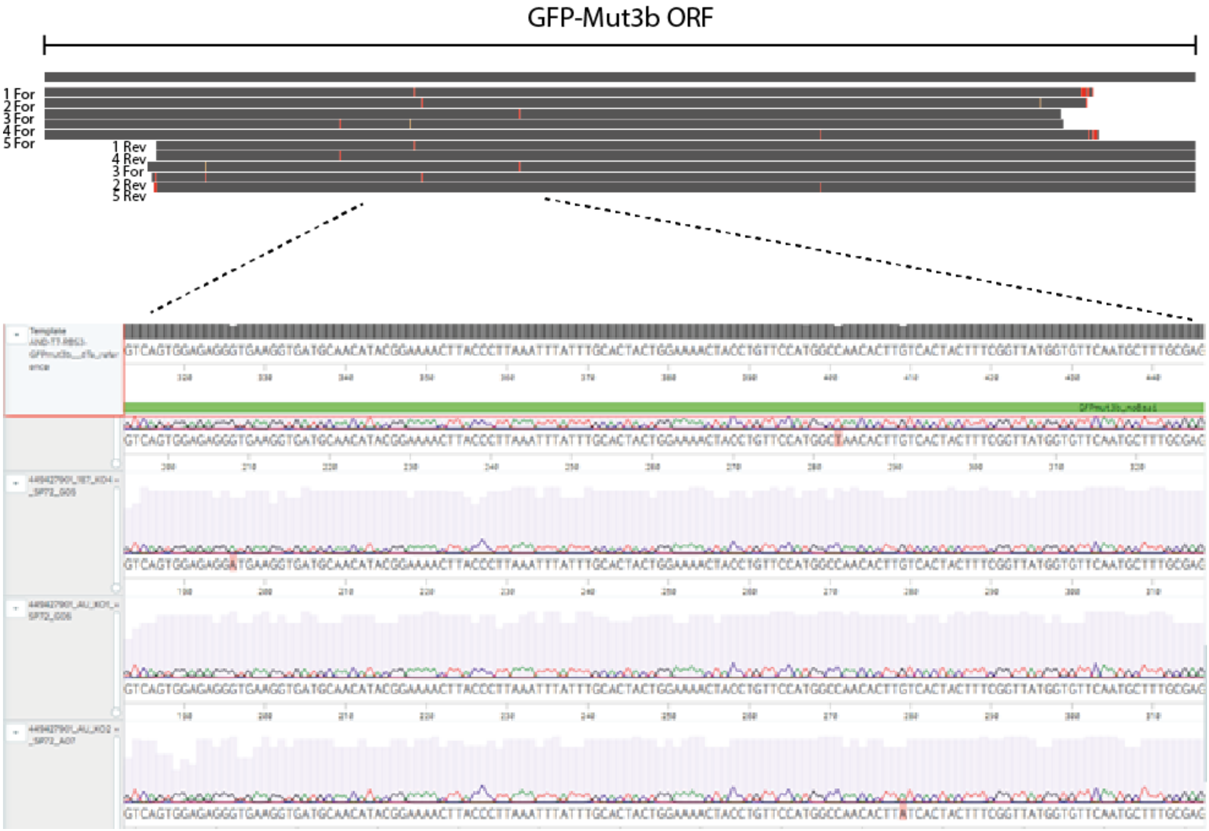
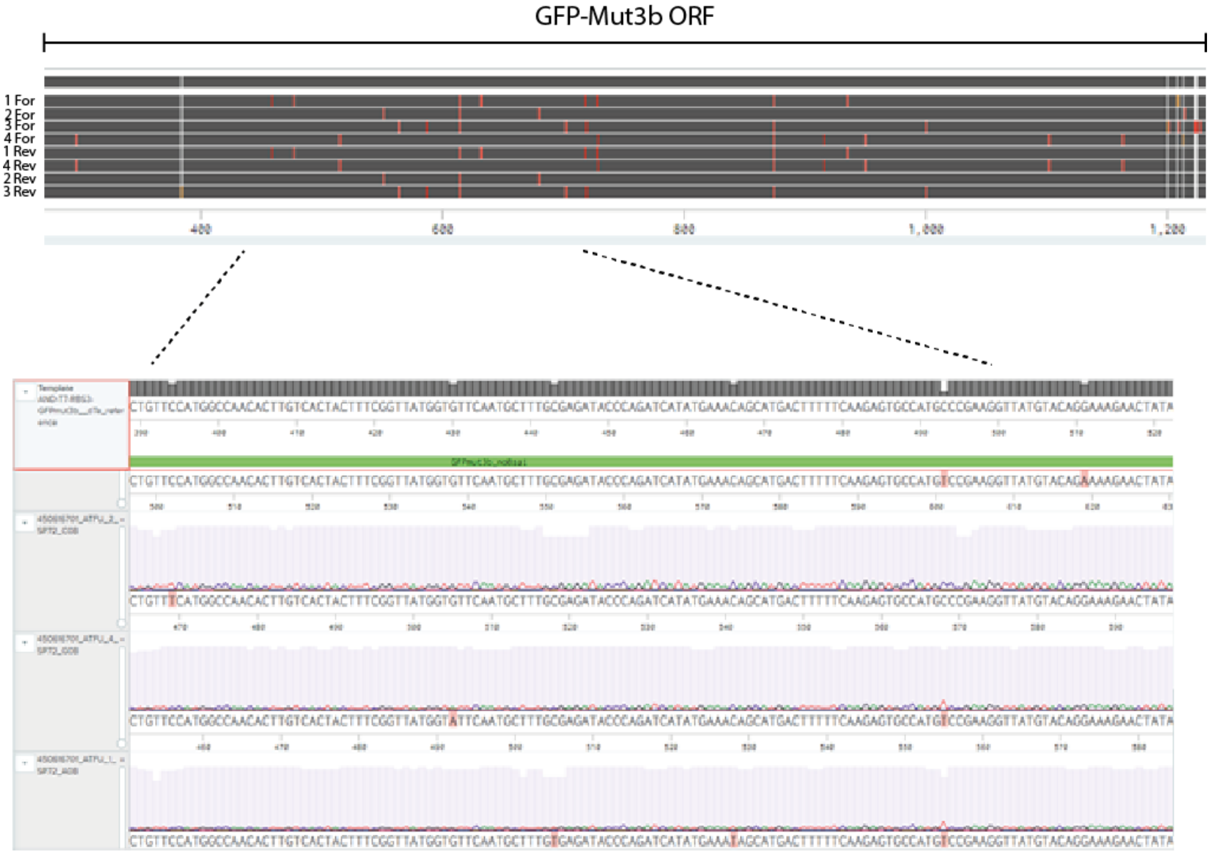

GFP-Mut3b ORF
1 For
2 For
3 For
4 For
5 For
1 Rev
4 Rev
3 For
2 Rev
5 Rev
GTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGTCACTACTTTCGGTTATGGTGTTCAATGCTTTGCGAG

Figure 3.7: Mutation count after 24-hour mutator assay. After 24-hours of the loss-of-function experiment with AID-T7pol + UGI, five mutant GFP-mut3b open reading frames were sequenced using forward and reverse primers. The grey bars represent the alignment to the reference, while the red lines represent C → T or G → A substitutions. After 24-hours, roughly one or two mutation events occurred on each ORF. Alignments were generated on Benchling using the ClustalW algorithm.

Mutations after 144-hours of the Mutagenic Assay:

Four GFP-mut3b mutants were selected at random and sequenced after 144-hours of the mutagenic assay. The mutations were found in both the forward and reverse primers, validating them as nucleotide substitutions resulting from the AID-T7pol + UGI mutator module and not as sequencing errors. One of the mutants had accumulated three C → T mutations, while the other three mutants displayed 6-8 mutations in the GFP ORF. This confirmed that AID-T7pol can perform successive deamination events in the target DNA sequence over multiple cell cycles. Within this small sample size, the mutations were found to be spread across the entire GFP-mut3b ORF, which confirmed the AID being fused to T7-RNA-Pol can target long stretches of DNA downstream of a T7 promoter for deamination of deoxycytidine.

Within this small sample size, no mutations were found in the DNA sequence for the promoter and one instance of a C → T mutation in the RBS. No mutations were witnessed downstream of the T7 terminator, which could mean the activity of the AID-T7pol fusion is isolated to the target gene sequence. The complete analysis of the mutation spread generated by AID-T7pol+UGI was performed using sequencing reads from Illumina iSeq100 and PacBio Sequel[xvi].


GFP-Mut3b ORF
1 For
2 For
3 For
4 For
1 Rev
4 Rev
2 Rev
3 Rev
400
600
800
1,000
1,200
Template
CTGTTCCATGGCCAACACTTGTCACTACTTTCGGTTATGGTGTTCAATGCTTTGCGAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATA

[xvi] Refer the Chapter 5 for NGS analysis of AID-T7pol UGI mutagenic activity

***Figure 3.8: Mutation count after 144-hour mutator assay**. After 144-hours, four GFP-mut3b ORFs were sequenced using forward and reverse primers on the Sanger platform. The grey bars represent alignments to the reference gene, while red lines in the bars represent substitutions. After 144-hours, roughly 8-10 mutations had accumulated on each ORF. Alignments were generated on Benchling using the ClustalW algorithm.*

3.5 Confirming AID-T7pol's Ability to Generate Functional C → T Mutations with UGI and the Gain of Function Experiment

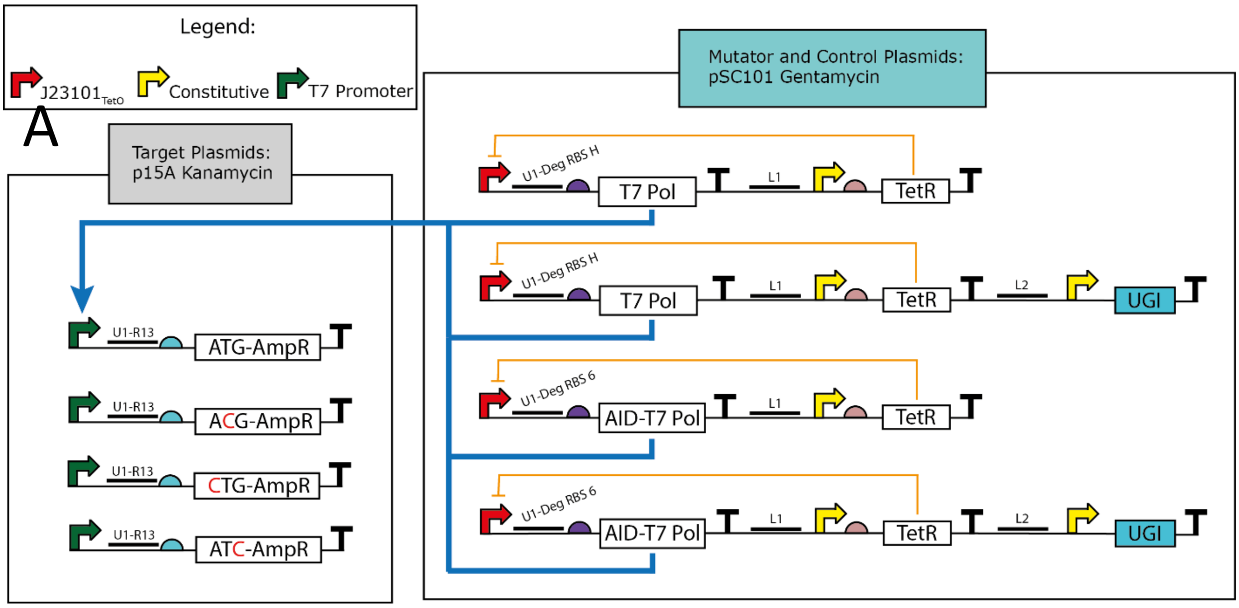
As demonstrated by the loss-of-function experiment above, phenotypic loss can result from any random nucleotide substitution event or indel resulting in a frame shift, and most random mutations tend to be deleterious or produce non-functional mutants[182]. For AID-T7pol to be viable as a DNA damaging device in a continuous evolution system, it needs to possess a wide mutational window so mutations can occur at specific positions along the gene ORF to create functional mutants. Gain-of-function experiments, where one starts with a non-functional gene variant, enabled screening of AID-T7pol's to generate functional mutations.

Methodology:

For this experiment, the inactivated variants of β-lactamase described in Section 2.5 were used as the target gene. The gene was inactivated by changing the ATG start codon to ACG, CTG and ATC. The inactive β-lactamase variants were assembled in an expression cassette with P_{T7} on a p15A-Kanamycin^{R} backbone. The AID-T7pol UGI mutator (expressed via medium strength Deg RBS-6) and three control modules were assembled on the pSC101-Gentamycin^{R} backbone for this experiment. The control modules AID-T7pol without UGI and T7-RNA-Pol with and without UGI. UGI blocks the BER pathway for U:G lesions, which is essential for being able to screen the C → T mutations generated by AID-T7pol. The 4 target gene plasmids and 4 mutator/control plasmids were co-transformed into DH5α cells. The samples were grown for 72-hours with the inducer molecule, aTc in LB media to switch-on the mutator system. Samples were transferred to fresh media every 24-hours to avoid an elongated stationary phase. After the 72-hour period, serial dilutions of the samples (at OD_{600} ~ 1.0) were spotted (2.5 µl spots) onto LB-agar plates containing carbenicillin to screen for functional mutations resulting in resistance to carbenicillin.

Results:

The results confirmed that fusing AID to T7-RNA-Pol provided the deaminase with the single-stranded template it required to perform deoxycytidine deamination. This ssDNA template was provided by the transcription bubble created by T7-RNA-pol, specifically targeting the β-lactamase gene downstream of a T7-promoter[161]. By inhibiting UNG (uracil N-glycosylase) activity through an inhibitor (UGI), base excision repair of the U:G lesions in β-lactamase gene was blocked. This meant the U:G mismatch was conserved in the gene until DNA replication. DNA pol-III recognises dU as dT and adds dA to the elongating DNA strand, enabling one of the daughter cells to acquire a C → T and G → A mutation. Of the three mutant ATG start codons, only ACG was reverted to functional ATG start codon, and corresponding cells gained resistance to carbenicillin. Cells expressing ATC- and CTG-inactivated β-lactamase did not gain carbenicillin resistance, confirming that only C → T substitutions occurred and no C → G or C → A. In the absence of UGI, the native BER pathway remains active in the host cells and constantly repairs the deamination events. As a result, no carbenicillin-resistant cells resulted from the control module expressing only AID-T7pol. The control modules expressing T7pol with and without UGI displayed no carbenicillin$^+$ cells. These controls validate the experimental data obtained for AID-T7pol + UGI, confirming that only C → T mutations occurred on the target gene due to deamination events induced by AID.



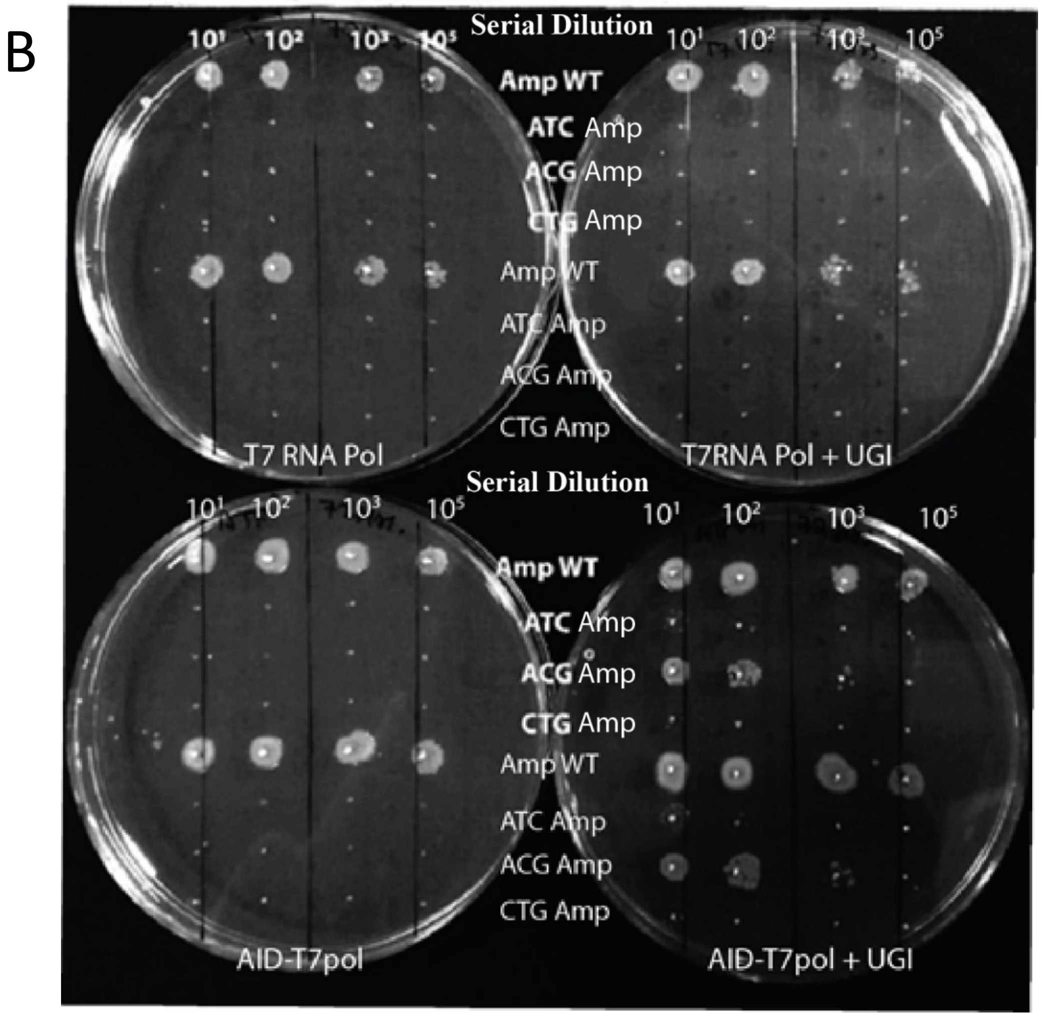


Figure 3.9: Gain of function experiment with AID-T7pol + UGI. *A – The SBOL schematic of the expression circuits used to verify the activity of AID-T7pol are shown. By adding an UNG blocker (UGI) into the mutator module, the cell's native BER pathway for deoxyuridine based damage was blocked. B – Cells with the inactivated β-lactamase genes were subjected to mutations using T7-pol, T7pol+UGI, AID-T7pol and AID-T7pol + UGI in a 72-hour gain-of-function experiment. Amp$^+$ colonies were only displayed for cells containing β-lactamase genes with an ACG start codon. A C → T transition created a functional ATG start codon. B-lactamase variants with CTG and ATC start codons were not reverted to functional phenotype, confirming AID-T7pol can only perform C → T and G → A substitutions.*

3.6 Global Mutagenic Activity of AID-T7pol Tested Using Rifampicin Resistance Assays

The ideal continuous evolution system should elevate the mutation rate for the gene-of-interest, while maintaining the global mutation rate close to natural levels (~ 10^{-10} mutations per base per generation[183]). If AID-T7pol is able to generate mutations in the bacterial genome at the same rate as the target gene, the high rate of accumulating global mutations can have deleterious effects on the host cell[184]. This reduces host cell viability for continuous evolution experiments performed over long time periods. Consequently, a qualitative analysis of the global mutagenic activity of AID-T7pol was done using rifampicin selection assays. Mutations in the RpoB gene encoding the β-subunit of bacterial RNA polymerase confers resistance to the rifampicin antibiotic[185,186]. Research shows that a single C → T mutation generates a Ser531Leu amino acid substitution in RpoB, enabling it to confer resistance to rifampicin[124,187]. Therefore, the C → T mutations generated by AID-T7pol UGI in ACG-β-lactamase and RpoB provide a direct comparison of the targeted vs off-target mutation frequency with AID-T7pol.

Methodology:

Cells expressing AID-T7pol + UGI and the three control modules were grown in LB with aTc for 72-hrs, being transferred into fresh growth medium every 24-hrs. After 72-hrs, 0.8 ml of cells at an OD_{600} ~ 1.2 (832 million cells) were plated on LB-agar containing rifampicin (50 µgml^{-1}). The number of rifamp^{+} cells resulting from AID-T7pol + UGI expression can be compared to rifamp^{+} cells from a cell culture containing wildtype DH5α cells. A rise in rifamp^{+} colonies indicates the mutator system is capable of damaging genomic DNA other than the target gene.

Table 3.3: Rifampicin+ Colony Count to Assess Global Mutator Activity

	CON-3[xvii]	**CON-4**	**Mutator**	**CON-6**	**CON-7**	**Control Cells**
Circuit Design	**T7 Pol**	**AID-T7pol**	**AID-T7pol + UGI**	**T7pol + UGI**	**UGI**	**Dh5α**
Run 1	74	73	490	390	97	30
Run 2	43	37	256	178	147	37
Run 3	20	35	284	85	153	3
Average	46	48	343	218	132	23

[xvii] Refer to Tables 4.3 and 4.4 for detailed description of controls.

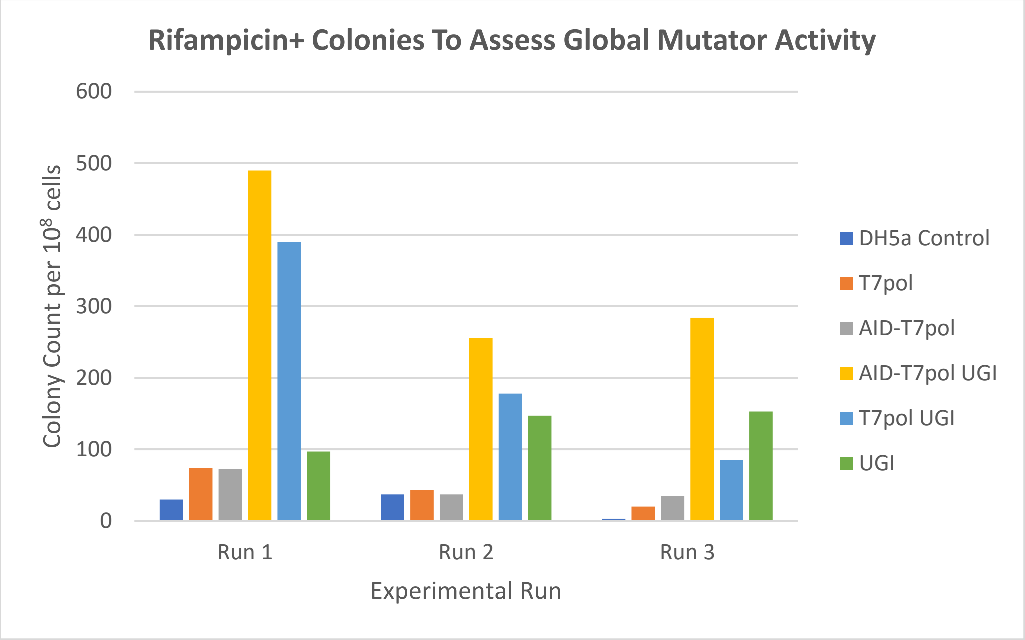


***Figure 3.10: Rifampicin$^+$ colony count indicating global mutagenic activity**. AID-T7pol and T7 RNA polymerase expression caused a two-fold increase in rifampicin+ cells resulting from mutations in the RpoB gene, compared to wildtype DH5α control cells. With UGI being expressed the genomic mutation rate was increased 10-fold.*

Results:

In the case of DH5α control cells, an average of 23 rifamp$^+$ colonies were spotted. This means the chance of naturally gaining a mutation conferring rifampicin-resistance is 2.8x10^{-8} per cell (23/832,000,000). This is comparable to the documented natural mutation rate of *E. coli*[183,124].

T7-RNA-pol and AID-T7pol displayed low levels of global mutagenic activity, comparable to the wildtype cells (Figure 3.10). This confirms that majority of the U:G lesions introduced by AID would be repaired by the cell's native DNA repair pathways. This most likely results in the overall low global mutagenic activity witnessed. Once UGI is expressed in the mutator module, the global mutagenic activity rises 100-fold. As this rise is seen in both AID-T7pol + UGI and T7pol + UGI expression systems, the major cause for the rise is UGI expression and not off-target deamination events caused by AID. UGI blocks all U:G lesions introduced into the DNA. The accumulation of deoxyuridine resulting from random mutations in the RpoB gene is the most likely the cause of the 100-fold rise in rifamp$^+$ cells expressing AID-T7pol + UGI and T7pol + UGI.

3.6.1 Targeted vs Off-target Mutation Frequency by AID-T7pol UGI

Table 3.4: Targeted vs Off-target Mutation frequency of AID-T7pol UGI

	AID-T7pol + UGI	
Target Gene	**ACG β-lactamase**	**RpoB**
Carb+/Rifamp+ cells	80000	211
Total cells plated OD_{600} ~ 1.0	20000000	80000000
Mutation frequency (per cell)	0.004	2.6×10^{-6}

To calculate the targeted vs off-target activity of AID-T7pol, the frequency of generating carbenicillin-resistant colonies was compared to the frequency of generating rifampicin-resistance bacterial colonies. At 1000x serial dilution in Figure 3.9B, 4 $carb^{+}$ colonies resulted from ACG-β-lactamase reversion. As serial dilution was performed after 72-hrs with 10 µl of the total 200 µl cell culture (OD_{600} ~ 1.0), this can be extrapolated to a total of 80,000 $carb^{+}$ cells in the total cell culture (colony count x dilution factor x total cell culture volume). To estimate the number of true global mutation events resulting from AID-T7pol activity, the average mutation events recorded for UGI in Table 3.3 was subtracted from mutation events for AID-T7pol UGI. Using these adjusted values, AID-T7pol's targeted mutational activity resulted in C → T mutations at a frequency of 4×10^{-3} per cell, while the off-target activity was a 1000-fold lower at 2.6×10^{-6}. Therefore, AID-T7pol is 1000-times more likely to generate U:G lesions in the gene-of-interest than random genes in the bacterial genome. This would help maintain the genomic fidelity of cells during continuous evolution experiments lasting numerous *E. coli* cell cycles.

3.7 Conclusion

Using a degenerative library of RBS sequences, the expression of T7-RNA-pol and AID-T7pol was optimised to reduce translational burden and toxicity on host DH5α cells. RBS sequences enabling low, medium and high levels of expression of these proteins were identified via DNA sequencing. These mutagenic proteins could now be expressed in the host cell for mutagenic assays performed over long periods of time, currently experimentally tested to 144-hours.

These optimised expression cassettes were then used to target mutations on GFPmut3b integrated into the DH5α genome. Analysing the mutant library using flow cytometry showed that GFP fluorescence was only lost when the inducer molecule, aTc, was present in the growth media. This confirmed a switch-like activation and repression of the AID-T7pol DNA damage device via the tet repressor. This is ideal for a continuous evolution system, where shutting of the mutator system is crucial to isolating and identifying the genotype that generated a functional protein variant. Using the gain of function experiment, it was confirmed that AID-T7pol + UGI is only capable of performing C → T and G → A mutations, as only the inactive ACG start codon versions of β-lactamase reverted to confer resistance to carbenicillin.

Finally, using the rifampicin antibiotic, the global mutagenic activity of the AID-T7pol device was assessed. There was no significant rise in rifamp$^+$ cells compared to control cells when only AID-T7pol is expressed. This is because off-target deamination events introduced by the DNA damaging device would be repaired by the native BER pathway. Once UGI is introduced in the mutator system, the global mutagenic activity increases 100-fold. Also, the targeted mutagenic activity was estimated to be 1000-fold higher than its off-target activity. This finding means a mutator system implementing AID-T7pol will most likely have a 1000-fold lower global mutation rate than the targeted mutation rate for the gene-of-interest. To avoid high global mutation rates, and to increase the diversity of nucleotide substitutions that can be introduced to the target gene, UGI needed to be replaced in the mutator system with a targeted error-prone DNA repair complex. This change ensures the native BER pathway remains active to repair most off-target U:G lesions generated by AID-T7pol.

Chapter 4: Engineering an Error-prone DNA Repair Complex for Generating Greater Genetic Diversity in a Target Gene

4.1 Introduction

In the introductory chapter and Chapter 2, the importance of using an error-prone DNA polymerase in an *in vivo* continuous evolution system for introducing mutations in a target gene is exemplified. The lack of proof-reading by EP-DNA-polymerases creates opportunities for nucleotides to be incorporated into the target DNA sequence without conforming to Watson-Crick base pairing[188]. EP-DNA-polymerases can potentially copy one of the four nucleotide bases (A, C, T, G) into each position in the target DNA sequence, creating a high genetic diversity for evolving a gene[189]. The main challenge for an *in vivo* mutator system is concentrating the mutagenic activity of EP-DNA-polymerases to the target gene or set of genes, while minimising off-target mutations.

In methods like PACE[xviii], mutator strains of bacteria are used. The EP-DNA-polymerase activity is not targeted, resulting in the accumulation of mutations in the host cell genome, which reduces the cell's viability over time. This makes it necessary to add a fresh pool of bacterial cells into the evolving lagoon of bacteria and bacteriophage over time[71,139,179]. The lack of targeting also means the rate of mutating the target gene cannot be regulated to speed-up or slow down the evolutionary process.

EvolvR utilises a fusion protein of Cas9 and an error-prone DNA polymerase-I[118]. This CRISPR-based method can target mutations to any DNA sequence complimentary to the gRNA sequences used, however the mutational window is limited to a few 100 base-pairs[118,119]. Multiple gRNAs are required to target entire gene sequences to mutations; also, after each successive round of mutation, new gRNA sequences would need to be synthesised to bind to the mutant variants of the target gene. This is not feasible as one cannot predict the evolutionary path taken by the evolving gene over time.

OrthoRep utilises an orthogonal plasmid—EP-DNA-polymerase pair in yeast to perform random mutations on the cargo DNA placed on linear p1 plasmids containing a specific terminal protein (TP1)[108,109]. As the TP-DNA polymerase can only initiate replication via TP1, no off-target mutations occur on the yeast genome. This orthogonal plasmid-polymerase pair enables mutation of the target gene at 10^5-fold higher rates than the native mutation rate of *S. cerevisiae*. However, the use of this continuous evolution system is limited to yeast strains.

[xviii] Refer to Section 1.4.1 for full description

As the existing continuous evolution systems utilising EP-DNA-polymerases have certain limitations, we aimed to engineer a novel error-prone DNA repair complex that utilises an EP-DNA-polymerase to specifically create genetic diversity in the target gene, while generating minimal off-target mutations. The aim was to design a mechanism by which EP-DNA-polymerases in the cytosol would be localised to the U:G DNA lesions created in the target gene by AID-T7pol. Such localisation of the EP-DNA-polymerase was achieved by fusing it to an AP-endonuclease, creating an AP-Endo—EP-DNA-Polymerase fusion protein. As described in Chapter 2, AP-endonucleases are involved in the base excision repair pathway for nicking the AP-site and generating a gap[190]. The C → U deamination by AID-T7pol would initiate the BER pathway and an AP-site is generated by the uracil-N-glycosylase (UNG). This AP-site would be recognised by the AP-Endo—EP-DNA-Polymerase fusion protein; the AP-endo would create the gap and the localised EP-DNA-polymerase would subsequently fill the single-nucleotide gap with a chance of mismatches being generated. This AP-Endo—EP-DNA-Polymerase fusion would therefore enable mutations to be generated at C:G base-pairs within the target DNA.

For inducing mutations at A:T base pairs and increasing the mutational window beyond single nucleotides, a 5’-3’ exonuclease was fused to the N-terminus of AP-Endo—EP-DNA-Polymerase, to create a 5’-3’-Exo—AP-Endo—EP-DNA-Polymerase three-protein fusion. This three-protein error-prone DNA repair complex would now be able to perform patch repair, as hypothesised in Section 2.3. In this instance, when the damaged base is excised by the AP-endonuclease, the 5’-3’ exonuclease would unwind the DNA helix and excise nucleotides downstream of the U:G lesion to generate a larger gap. The EP-DNA-polymerase fills the larger gap, with a chance for mismatches to occur along the way. Thus, the AID-T7pol DNA damaging device and the 5’-3’-Exo—AP-Endo—EP-DNA-Polymerase EP-DNA-repair complex would work synergistically to induce targeted DNA damage and generate genetic diversity in the gene-of-interest.

In this chapter, we present AP-endonuclease, EP-DNA-polymerase and 5’-3’ exonuclease candidates that are functional in *E. coli* for building an error-prone DNA repair complex. These candidate bioparts were characterised using a library of 15 RBS sequences to identify expression systems that enabled high, medium and low levels of expression for these proteins in DH5α cells. Unique error-prone DNA repair complexes were subsequently assembled using different combinations of the candidate bioparts to create 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion proteins. Expression cassettes for the different EP-DNA-repair complexes were combined with the AID-T7pol expression system to generate a library of mutator modules capable of inducing targeted DNA damage and subsequently generating genetic diversity in the gene-of-interest. To test the mutator modules for their ability to perform the hypothesised targeted error-prone DNA repair activity, the loss-of-function and gain-of-function workflows described in Section 2.5 were utilised. The gain-of-function experiments with inactive β-lactamase gene variants enabled direct assessment of the mutator systems for their ability to perform nucleotide substitutions and patch repair. C → (A, C, G) and A → T mutations were investigated with this workflow.

The rate of targeted mutagenesis compared to off-target mutations generated by the mutator modules was also investigated using rifampicin antibiotic selection assays. Using these workflows, the library of mutator modules that were assembled, were tested and narrowed down to a single version that performed all four nucleotide substitutions investigated using inactive β-lactamase; generated minimal off-target mutations; and imposed minimal burden on the host cell when it was actively expressed.

4.2 Experimental objectives for this chapter

Objective 1. Identifying, isolating, and characterising the expression of bioparts intended for building the error-prone DNA repair complex in *E. coli*.

Objective 2. Assembly of candidate 5’-3’-Exonucleases, AP-endonucleases and EP-DNA-polymerases into the 3-protein fusion, creating a library error-prone DNA repair complexes. This is followed by modular assembly of the AID-T7pol expression cassette with the library of EP-DNA-repair complexes to generate complete mutator modules capable of inducing targeted DNA damage and creating genetic diversity.

Objective 3. Using the loss of function experiment described in Section 2.5 to screen different AID-T7pol + EP-DNA-repair complex combinations (called a mutator module) for their mutagenic activity on a target gene. The mutator modules with the highest mutagenic activity were shortlisted further characterisation.

Objective 4. Testing the shortlisted mutator modules for their ability to generate nucleotide substitutions beyond C → T and G → A, something that can be achieved by AID-T7pol activity alone. The gain-of-function workflow with inactivated β-lactamase enabled identification of mutator modules capable of performing four specific nucleotide substitutions to convert inactive β-lactamase into functional proteins.

Objective 5. Testing the impact of applying a selection pressure in the gain-of-function experiments. This experiment was to explore if applying a selection pressure increases the frequency at which functional mutations are incorporated into the β-lactamase gene.

Objective 6. Validating the mutagenic activity of the mutator module with AID-T7pol and an EP-DNA-repair complex. Various control modules were assembled with different combinations of AID, T7-RNA-Pol, UGI and EP-DNA-repair complexes to show that mutations are only incorporated into the gene-of-interest when AID-T7pol creates targeted U:G lesions, and the EP-DNA-repair complex hijacks the BER to generate mismatches via an error-prone DNA polymerase.

Objective 7. Testing the global mutation frequency generated by shortlisted mutator modules using rifampicin selection assays and comparing the frequency of targeted mutations to the frequency of off-target mutations.

4.3 Objective 1: Characterising and Optimising the Expression of Bioparts used for Assembling Different Error-prone DNA Repair Complexes

4.3.1 List of Bioparts Selected

The error-prone DNA repair complex was designed to comprise either the 2-protein AP-Endo—EP-DNA-Polymerase fusion protein or the 3-protein 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion. The candidate parts chosen for building these fusion proteins are naturally functional in bacterial species.

Candidate AP-endonucleases to excise damaged nucleotide in BER pathway:

The BER pathway is initiated with the recognition of the U:G lesion by UNG; this glycosylase cleaves the nitrogenous base of deoxyuridine to generate an apyrimidinic site (AP-site). Such AP-sites are recognized and excised by AP-endonucleases. Our mutator system therefore required an AP-endonuclease to ensure the enzyme is expressed sufficiently in the cell to recognise the artificially induced deamination events by AID-T7pol. Two different AP-endonucleases, *E. coli* exonuclease-III (Exo-III) and the *N. Meningitidis* AP-endonuclease (NAPE) were chosen as suitable candidates for the mutator circuit (Table 4.1).

Candidate DNA polymerases to perform error-prone DNA repair:

The gap-filling process in the BER pathway is naturally carried out by a proof-reading DNA polymerase (DNA pol-I in *E. coli*[191]). This gap-filling process needs to be hijacked and performed by an error-prone DNA polymerase for genetic diversity to be created. To limit the global activity of the EP-DNA-polymerase, the mutator design involves fusing the DNA polymerase to the AP-endonuclease. Hypothetically, this should concentrate the EP-DNA-polymerase activity to the AP-sites. The EP-DNA-polymerase candidates chosen for testing were *E. coli*’s native DNA polymerase IV (DNA Pol-IV) and the polymerase domain of error-prone versions of DNA Polymerase-I, developed by Loeb and colleagues[141].

Candidate 5'-3' exonucleases to extend the area of effect for the EP-DNA-polymerase:

To increase the mutation window and to ensure mutations can be induced at nucleotides other than C:G base pairs, we had to find a way to increase the area of effect for the EP-DNA-polymerase. DNA polymerases require single-stranded DNA (ssDNA) as template to synthesise a complimentary strand in the 5' to 3' direction. To continually unwind DNA and generate longer ssDNA template beyond the nicked gap at AP-sites, a 5'-3' exonuclease was added to the AP-Endo—EP-DNA-Polymerase repair complex. The 5'-3' exonucleases that were selected for testing were RecJ, RecE and the 5'-3' exonuclease domain of DNA pol-I[192] from *E. coli*.

Table 4.1: List of Candidate Bioparts to Build Mutator Modules for a Continuous Evolution System

Part	Function	Description
AID-T7 RNA Polymerase fusion protein	Targeted Cytosine Deaminase	Performs dC → dU deamination in the DNA of a target gene, placed downstream of a T7 promoter[121].
T7 RNA Polymerase	Orthogonal transcription system	The T7 RNA polymerase performs transcription of genes placed downstream of its orthogonal T7-promoter (P_{T7})[193,194]. This targeting system ensures AID's mutagenic activity is localised to the gene-of-interest placed downstream of a P_{T7}.
Exonuclease-III	AP endonuclease 3'-5' exonuclease	Cleave the sugar-phosphate backbone at the AP-site to generate a nick in the DNA strand. Exo-III also exhibits 3'-5' exonuclease activity[190].
NAPE	AP endonuclease	Break the phosphodiester bond at the AP site to generate a 3' hydroxyl group[195].
DNA Polymerase IV	Error prone polymerase	An error-prone polymerase encoded by the DinB gene in *E. coli*[196]. Three versions of the polymerase are tested in the mutator circuit: the wildtype, 5-amino acid truncated mutant (Pol-IV$^{\Delta5}$) and a 12-amino acid truncated mutant (Pol-IV$^{\Delta12}$). The amino acid residues are truncated from the C-terminal to reduce Pol-IV's global mutagenic activity by eliminating its ability to interact with the β-clamp at replication forks[143,197].
EP-DNA Polymerase-I (polymerase domain only)	Error prone polymerase	Four DNA Pol-I mutants were selected from the mutant library generated by Loeb and colleagues. These mutants were 5x (Pol-I^{5}), 46x (Pol-I^{46}), 150x (Pol-I^{150}) and 1100x (Pol-I^{1100}) more error prone than the wildtype polymerase[198].
RecJ	5'-3' exonuclease	RecJ requires ssDNA as template for exonuclease activity[192,199].
RecE	5'-3' exonuclease	RecE is a large exonuclease protein that requires dsDNA as template. It froms a ring around dsDNA and cleaves DNA in the 5'-3' direction, even in the absence of a 5'-phosphate[192,200].
5'-3' Exo Domain of DNA Pol-I	5'-3' exonuclease	The 5'-3' exonuclease domain that excises nucleotides from gaps generated at DNA lesions during DNA repair, creating a patch of ssDNA to be filled-in by DNA pol-I[201]. We isolated two versions of this exonuclease from MG1655. This version is 882 bp long, containing the nucleotides annotated as the 5'-3'-exonuclease of the polA gene in the Uniprot database, plus another 120 bp of unannotated bases downstream.
5'-3' Exo Domain of DNA Pol-I (Shorter version)	5'-3' exonuclease	This is the short version and contains only the 762 bp annotated as the 5'-3' exonuclease domain in polA from the Uniprot database.

4.3.2 Experimental Design of the Characterisation Workflow

Unlike the characterisation of AID-T7pol and T7-RNA-Pol with a degenerative RBS library, the bioparts for building the error-prone DNA repair complex where characterised using a defined set of 15 RBS sequences. This greatly improved the throughput of the characterisation workflow by eliminating the need to purify each of the characterised expression cassettes and sequencing them to identify the nucleotide sequence of the RBS. The BASIC DNA assembly workflow[xix] enabled simultaneous assembly of 15 different expression cassettes for six candidate bioparts highlighted in Table 4.1. T7 RNA polymerase and AID-T7pol expression were characterised previously so they were not included in this study. The use of a defined set of RBS sequences also enabled investigating how context dependency between the upstream element, the RBS and the gene coding sequence can affect the expression strength of RBS sequences[202]. An RBS enabling high expression of one protein may express another protein at low rates. The defined set if RBS allowed for such context dependency trends to be identified. The polymerase domain of EP-DNA Polymerase-I was also not characterised as its expression would not be impacted by RBS context. The AP-Endo and 5'-3'-Exonuclease domains would always be upstream of the polymerase domain.

To visualise expression rates via fluorescence, biopart candidates for the EP-DNA-repair complex were fused to splitGFP (Figure 4.1). SplitGFP is created by removing the 11th β-strand (β11) from the protein structure, which inactivates the fluorescent protein. β11 and splitGFP can be expressed independently in the cell, and a fluorescence signal is only produced when both components interact with one another[203,204]. The β11 is C-terminally fused to the Biopart being characterised, while the remaining splitGFP domain is expressed constitutively. The advantage of splitGFP is that it does not impede with the tertiary structure of the target protein to which β11 is fused. The assay also generates less noise and even signal amplification can be achieved by fusing multiple β11 sequences in tandem to the gene-of-interest[203].

Characterisation plasmid design:

Characterisation: $_{\mathrm{LMP}}$ **J23101$_{\mathbf{TETO}}$** $_{\mathrm{U1\text{-}RBSX}}$ **Biopart** $_{\mathrm{FL2}}$ **β11 [splitGFP]** $_{\mathrm{L1}}$ **TetR** $_{\mathrm{LMS}}$ **pSC101** $_{\mathrm{L5}}$ **Kan**[xx]

Where, X = An RBS from a 15 RBS library

Control: $_{\mathrm{LMP}}$ **J23101$_{\mathbf{TETO}}$** $_{\mathrm{U1\text{-}RBS15}}$ **mCherry** $_{\mathrm{FL2}}$ **β11 [splitGFP]** $_{\mathrm{L1}}$ **TetR** $_{\mathrm{LMS}}$ **pSC101** $_{\mathrm{L5}}$ **Kan**

[xix] Refer to Sections 2.5.1 and 7.3 for details on BASIC assembly

[xx] Bioparts shown in bold and BASIC linker sequences shown in subscript

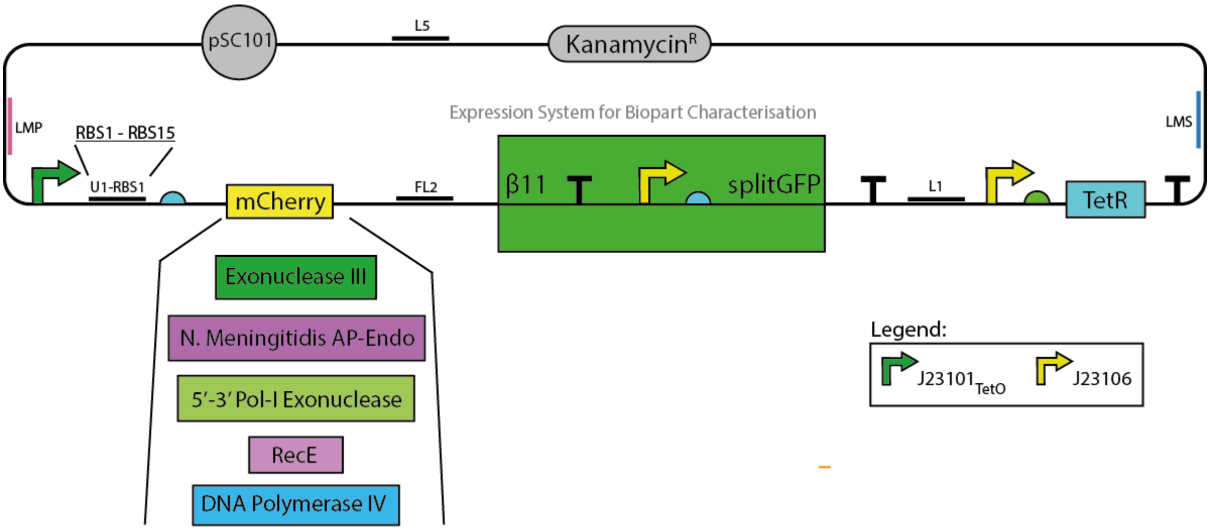


Figure 4.1: SBOL schematic of the expression plasmid designed for the characterisation of parts used for building the error-prone DNA repair complex in E. coli. *To analyse the expression rate of the proteins using different RBS sequences, the bioparts were fused to the 11th B-sheet of GFP. The remaining GFP domain was expressed constitutively. When β11 and the larger GFP domain interact, a fluorescence signal is generated. RecJ could not be characterised with splitGFP and the data is not shown. Analysing the crystal structure of RecJ revealed the C-terminus is surrounded by α-helices, which most likely impedes splitGFP's β11 interaction with the remaining protein domain*[205].

Cell Growth and Fluorescence Measurements Methodology

After DNA assembly of the 15 RBS—biopart—splitGFP combinations in the pSC101-Kanamycin backbone (Figure 4.1), the plasmids are transformed into competent DH5α cells. Three colonies are picked from LB-agar plates for each RBS—biopart—splitGFP combination and grown overnight in a 96-well plate. The overnight cultures were diluted in fresh LB media to an OD_{600} of 0.05. aTc (10 ngul^{-1}) was added to the media to activate the expression of the biopart—splitGFP fusion via J23101$_{TetO}$. Absorbance and fluorescence measurements were recorded every 15-minutes until the cell cultures reached stationary phase at 6-hours.

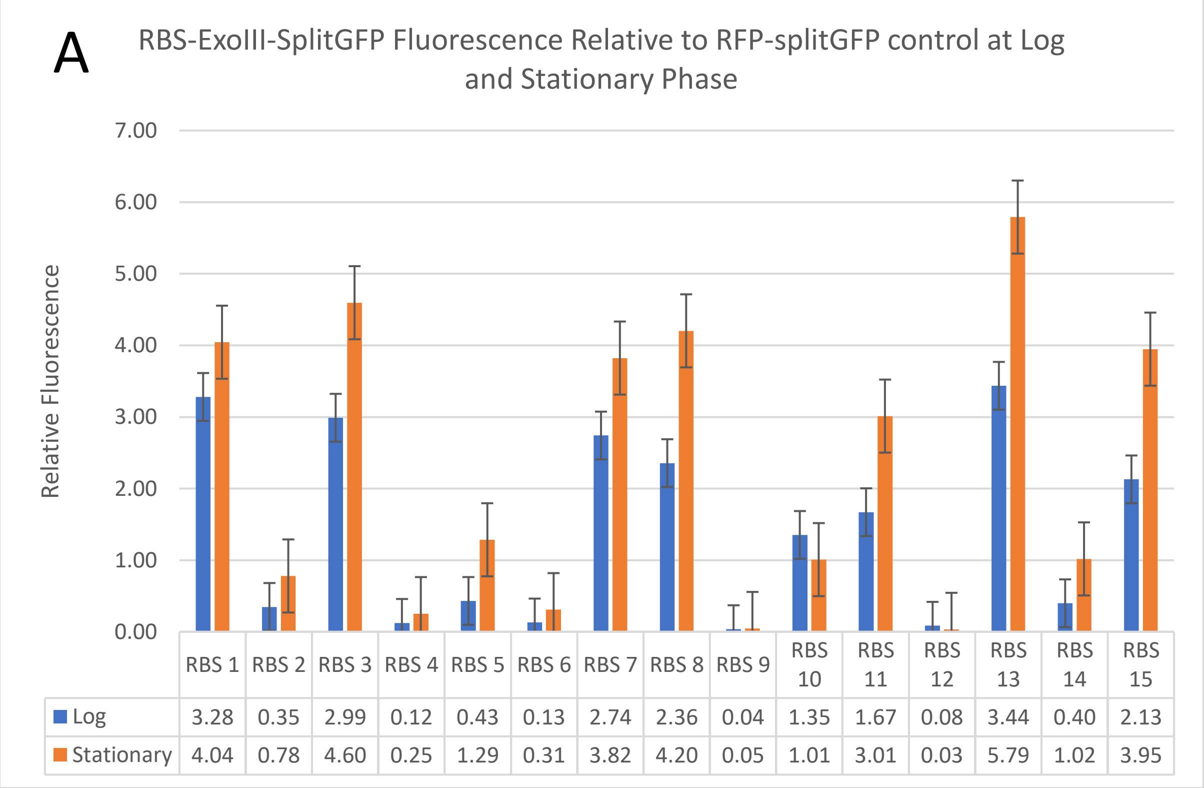
For visualising the data, the fluorescence relative to the optical density (FL/OD) of each experimental culture was calculated. This was divided by the FL/OD of the RBS15—mCherry—splitGFP control to obtain a comparable fluorescence value for each RBS—biopart combinations. The relative fluorescence was calculated for the cell cultures in log phase and stationary phase to assess which phase of bacterial growth is ideal for the expression of the target gene.

$$Relative\ Fluorescence = \frac{\{\frac{FL}{OD}(Sample)]}{\{\frac{FL}{OD}(mCherry)\}}$$

4.3.3 Results: Characterisation of the AP-endonuclease, Exo-III

The growth curve analysis of cell cultures expressing RBS—Exo-III—splitGFP combinations revealed great variance in the fitness level of the cells. Expression via only RBS8 and RBS15 enabled cells to maintain fitness levels comparative to the control (DH5α cells expressing an empty pSC101-Kanamycin plasmid). Cell cultures expressing Exo-III via RBS 4, 5, 6, 9, 11, 12 and 14 exhibited some degree of burden throughout log phase, but achieved stationary phase at an OD_{600}~ 0.7, comparative with the control (Figure 4.2). RBS 1, 3, 7, 10 and 14 resulted in a clear fitness burden on the cells, achieving stationary phase at OD_{600} of 0.2-0.4. Such RBS sequences were eliminated as candidates for building expression cassettes of the AP-endonuclease—EP-DNA-polymerase error-prone DNA repair complex with Exo-III.

When observing the relative fluorescence, it becomes clear why RBS 1, 3, 7, 10 and 14 impose a severe fitness burden, with cells expressing 4-fold to 6-fold more Exo-III relative to RBS15-mCherry. Expressing an AP-endonuclease at such high levels resulted in a clear loss of fitness. For building expression cassettes with Exo-III, RBS8 and RBS15 were selected for achieving high expression levels of the protein; RBS11 for achieving medium expression and RBS5 for low expression of levels of Exo-III in the cell.



	RBS 1	RBS 2	RBS 3	RBS 4	RBS 5	RBS 6	RBS 7	RBS 8	RBS 9	RBS 10	RBS 11	RBS 12	RBS 13	RBS 14	RBS 15
🟦 Log	3.28	0.35	2.99	0.12	0.43	0.13	2.74	2.36	0.04	1.35	1.67	0.08	3.44	0.40	2.13
🟧 Stationary	4.04	0.78	4.60	0.25	1.29	0.31	3.82	4.20	0.05	1.01	3.01	0.03	5.79	1.02	3.95

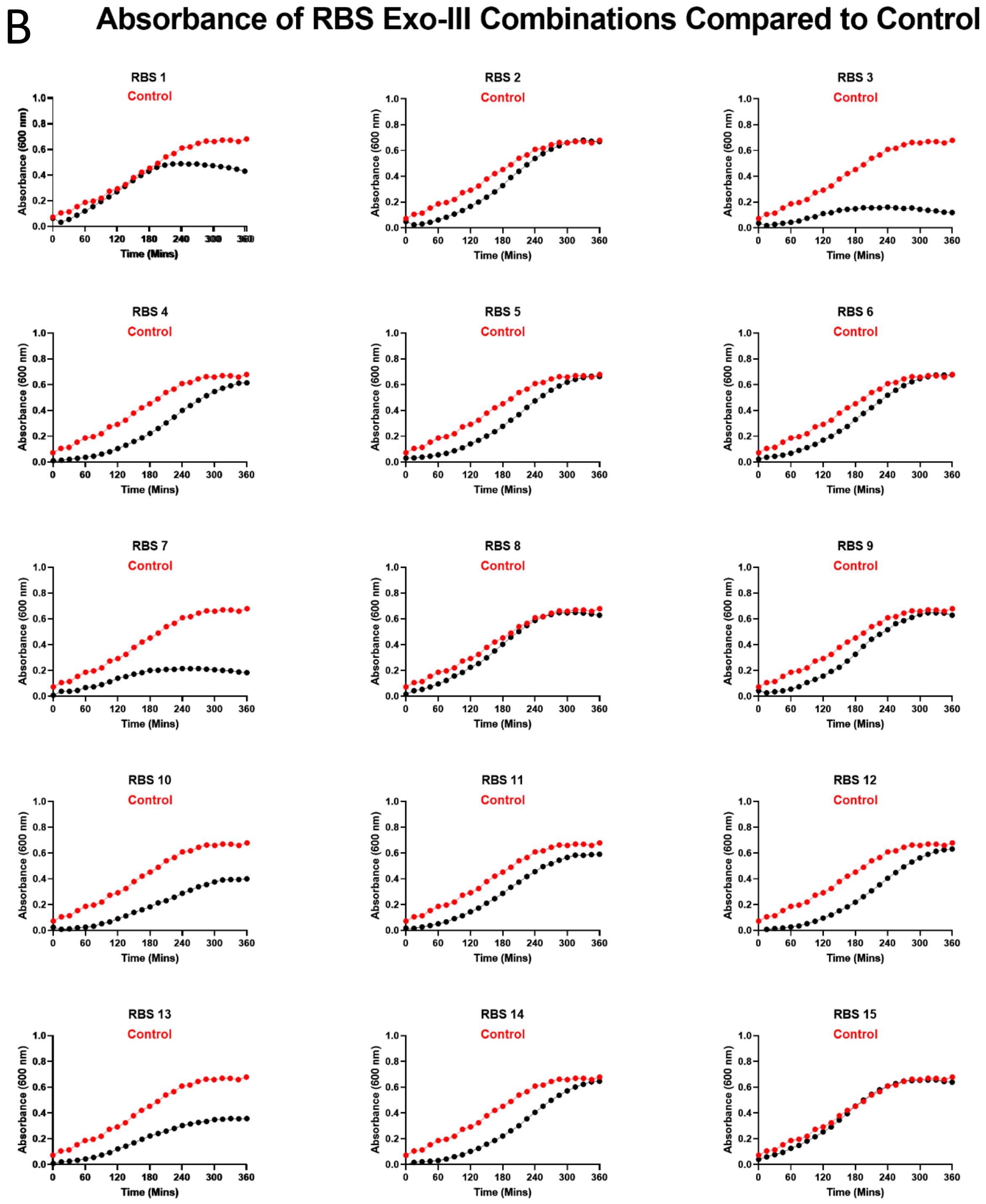


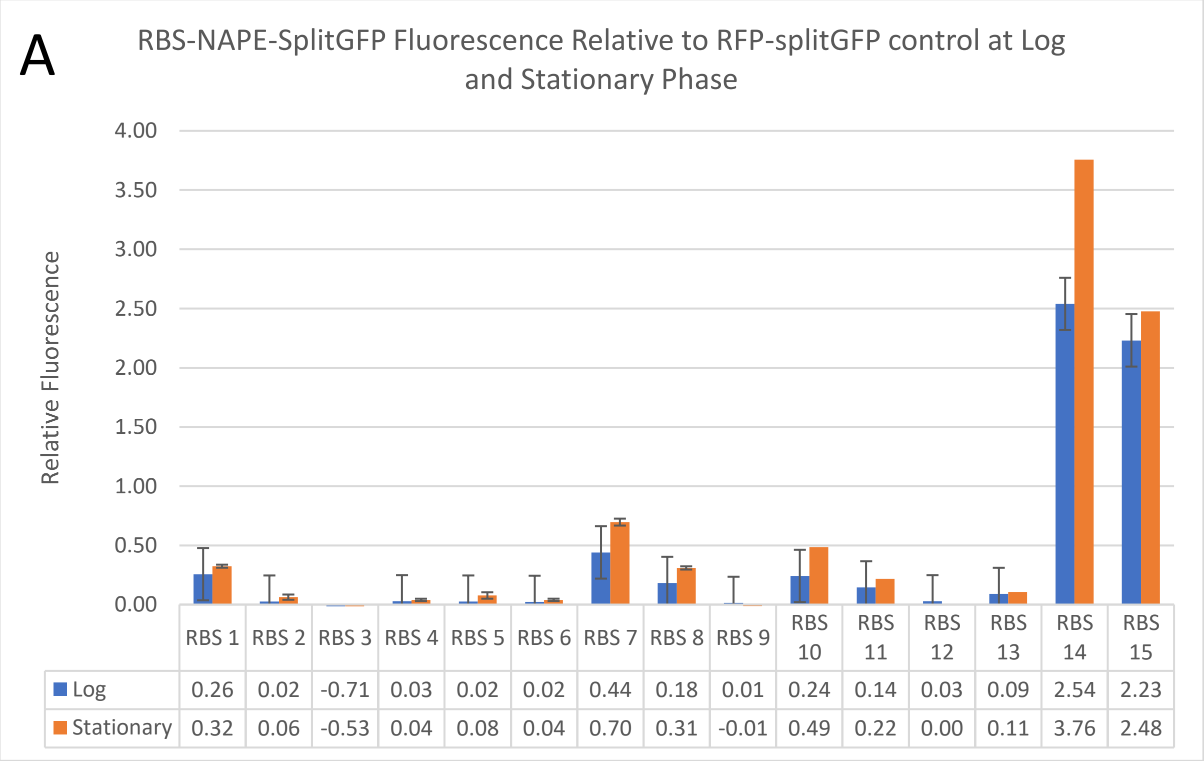
Figure 4.2: Characterising Exonuclease-III expression in DH5a cells using a library of RBS sequences. *A – The fluorescence output indicates the level of expression of exonuclease-III in the cell. The fluorescence output from biological triplicates was averaged and is represented via the error-bars. This fluorescence is presented relative to the GFP fluorescence displayed by the mCherry-splitGFP control fusion protein. The relative fluorescence is presented for two time-points, one with cell cultures in the log phase and one where cultures were in the stationary phase. B – The growth curve over 6-hours for*

cells expressing exonuclease-III with 15 different RBS sequences is shown in black. The control cells expressing an empty antibiotic-ORI vector are shown in red. The absorbance data from biological triplicates was averaged at each timepoint and plotted on the curve.

4.3.4 Results: Characterisation of AP-endonuclease NAPE

Analysing the growth curves for the 15 RBS—NAPE—splitGFP combinations revealed protein expression via RBS14 and RBS15 to impose a burden on the cells. The cells cultures for other RBS-NAPE combinations exhibited a higher optical density than the negative control during the log phase and achieved stationary phase at OD_{600}~ 0.8 (Figure 4.3).

Analysing the relative fluorescence of the samples reveals RBS 14 and 15 enabled 2-fold to 3.75-fold higher expression of NAPE than mCherry in the DH5α cells. Such high expression levels caused a significant fitness burden. By analysing the growth rates and the relative fluorescence data, RBS7 was characterised as a high expression RBS for NAPE, RBS10 as a medium expressor and RBS11 for achieving low expression levels of NAPE in DH5α cells. RBS 14 and 15 were disregarded for building expression cassettes of NAPE due to fitness burden, even though they achieve significantly higher expression of the target protein.



	RBS 1	RBS 2	RBS 3	RBS 4	RBS 5	RBS 6	RBS 7	RBS 8	RBS 9	RBS 10	RBS 11	RBS 12	RBS 13	RBS 14	RBS 15
■ Log	0.26	0.02	-0.71	0.03	0.02	0.02	0.44	0.18	0.01	0.24	0.14	0.03	0.09	2.54	2.23
■ Stationary	0.32	0.06	-0.53	0.04	0.08	0.04	0.70	0.31	-0.01	0.49	0.22	0.00	0.11	3.76	2.48

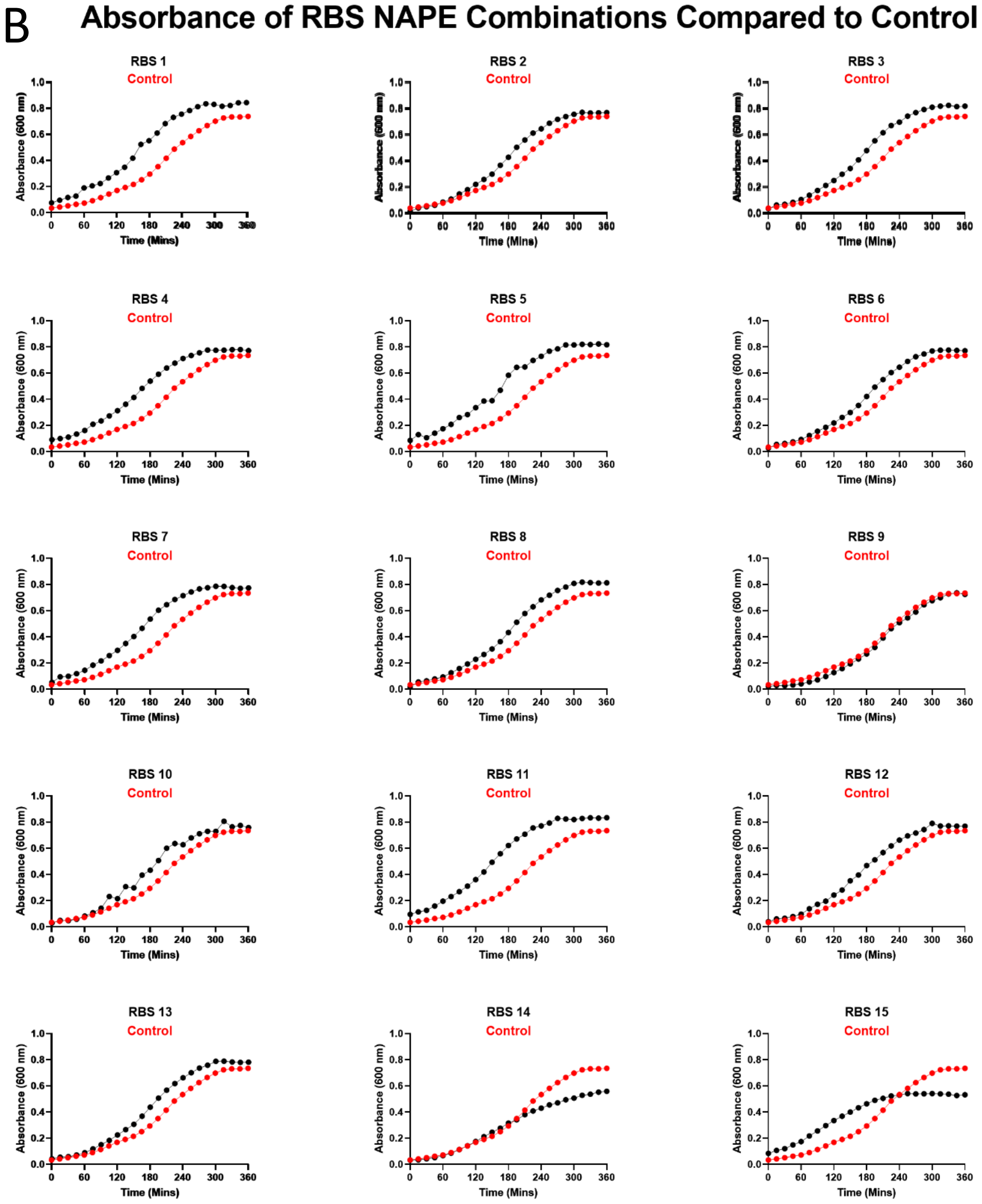
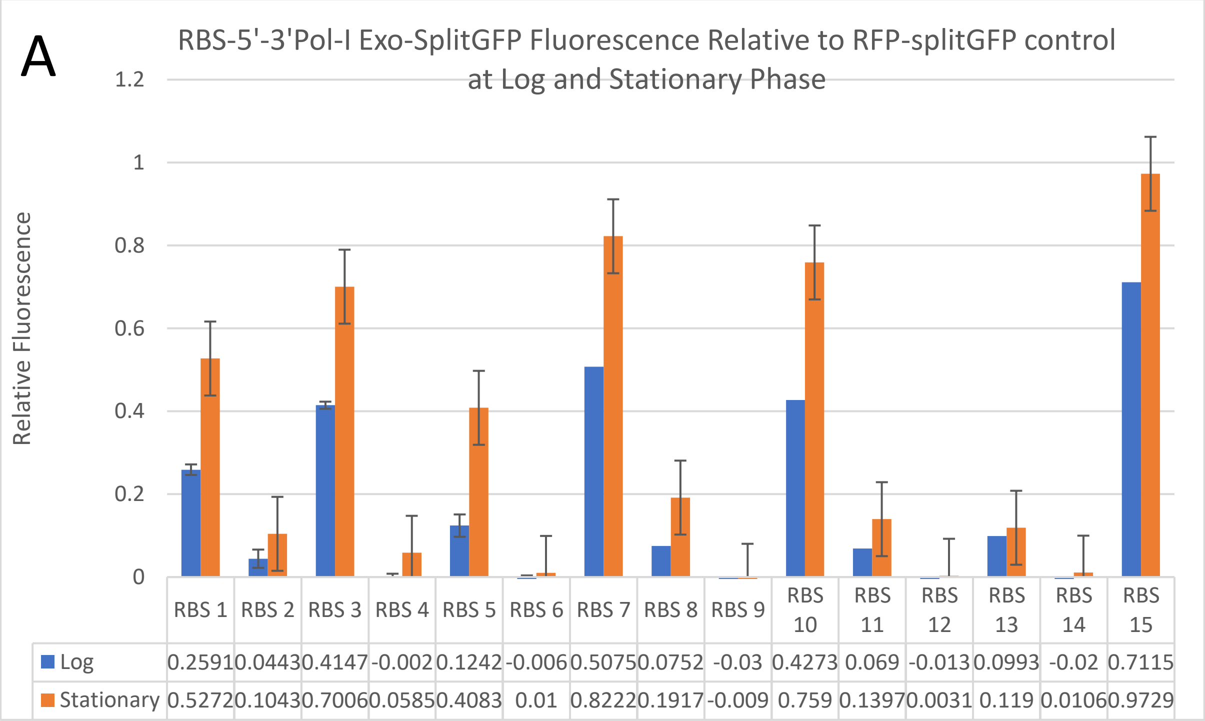


Figure 4.3: Characterising N. Meningitidis AP-Endonuclease expression in DH5a cells using a library of RBS sequences. *A – The fluorescence output indicates the level of expression of NAPE in the cell. The fluorescence output from biological triplicates was averaged and is represented via the error-bars. This fluorescence is presented relative to the GFP fluorescence displayed by the mCherry-splitGFP control fusion protein. The relative fluorescence is presented for two time-points, one where cell cultures were in the log phase and one where cultures were in the stationary phase. B – The growth curve over 6-hours for cells expressing NAPE with 15 different RBS sequences is shown in black. The control cells*

expressing an empty antibiotic-ORI vector are shown in red. The absorbance data from biological triplicates was averaged at each timepoint and plotted on the curve.

4.3.5 Results: Characterisation of 5’-3’ Exonuclease from DNA Polymerase-I

This 5’-3’ exonuclease domain of DNA pol-I (5’-3’Pol-I-Exo) was expressed in cells with all 15 RBS combinations without any significant burden or toxicity (Figure 4.4). All the cell cultures exhibited identical or higher fitness at log phase than the control cells and achieved stationary phase at an OD_{600}~ 0.8. The relative fluorescence data reveals that the 5’-3’Pol-I-Exo is expressed significantly less than the RBS15-mCherry control. Only RBS 1, 3, 5, 7, 10 and 15 enabled expression of the exonuclease at 40-80% of the expression rate of the mCherry fluorescent protein in cells. The remaining RBS sequences expressed the exonuclease at only 20% of the RBS15-mCherry expression level. RBS7 and RBS15 were shortlisted for achieving high expression levels of the 5’-3’Pol-I-Exo domain, RBS5 for a medium expression level and RBS2 for a low expression level in DH5α cells.



	RBS 1	RBS 2	RBS 3	RBS 4	RBS 5	RBS 6	RBS 7	RBS 8	RBS 9	RBS 10	RBS 11	RBS 12	RBS 13	RBS 14	RBS 15
🟦 Log	0.2591	0.0443	0.4147	-0.002	0.1242	-0.006	0.5075	0.0752	-0.03	0.4273	0.069	-0.013	0.0993	-0.02	0.7115
🟧 Stationary	0.5272	0.1043	0.7006	0.0585	0.4083	0.01	0.8222	0.1917	-0.009	0.759	0.1397	0.0031	0.119	0.0106	0.9729

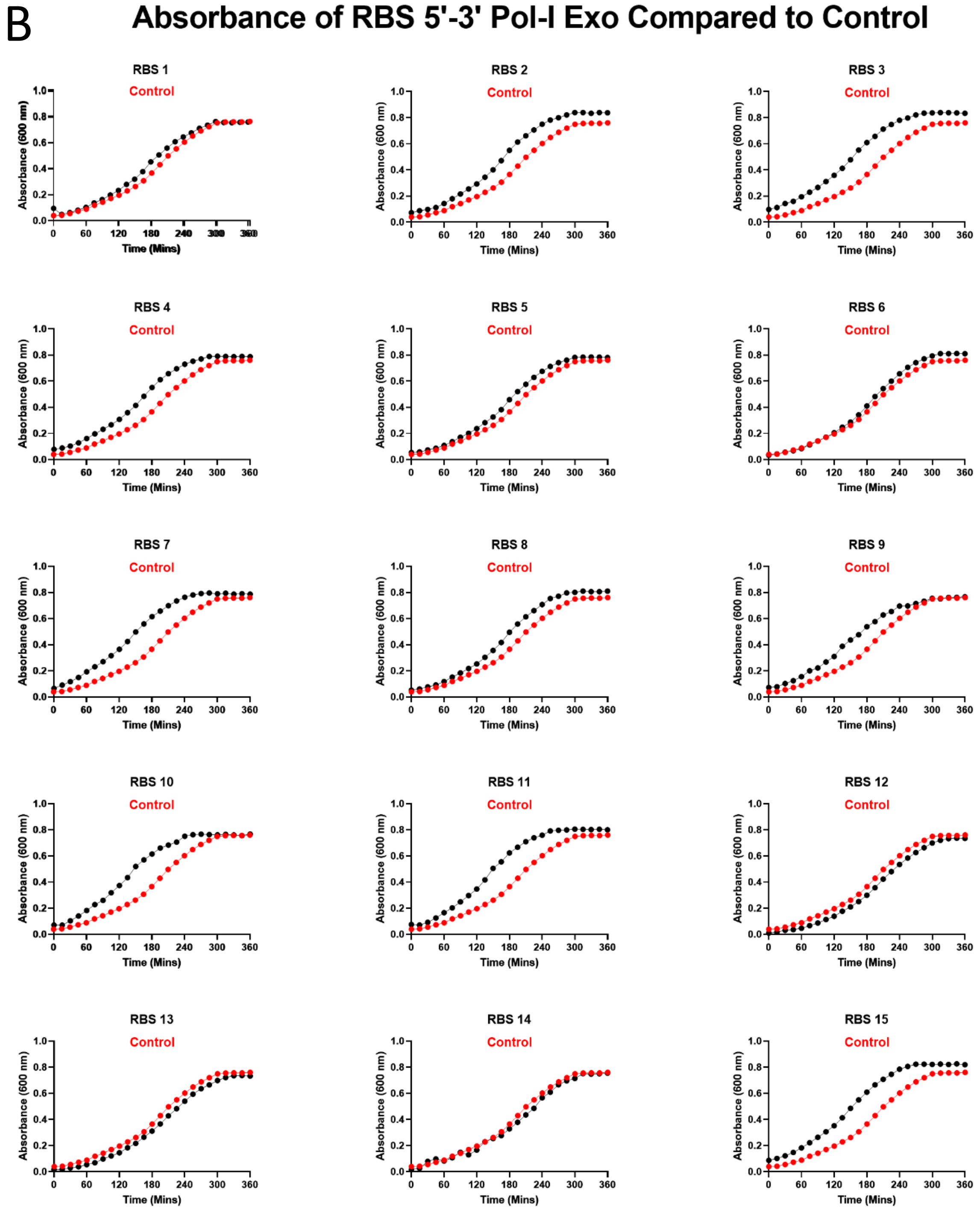
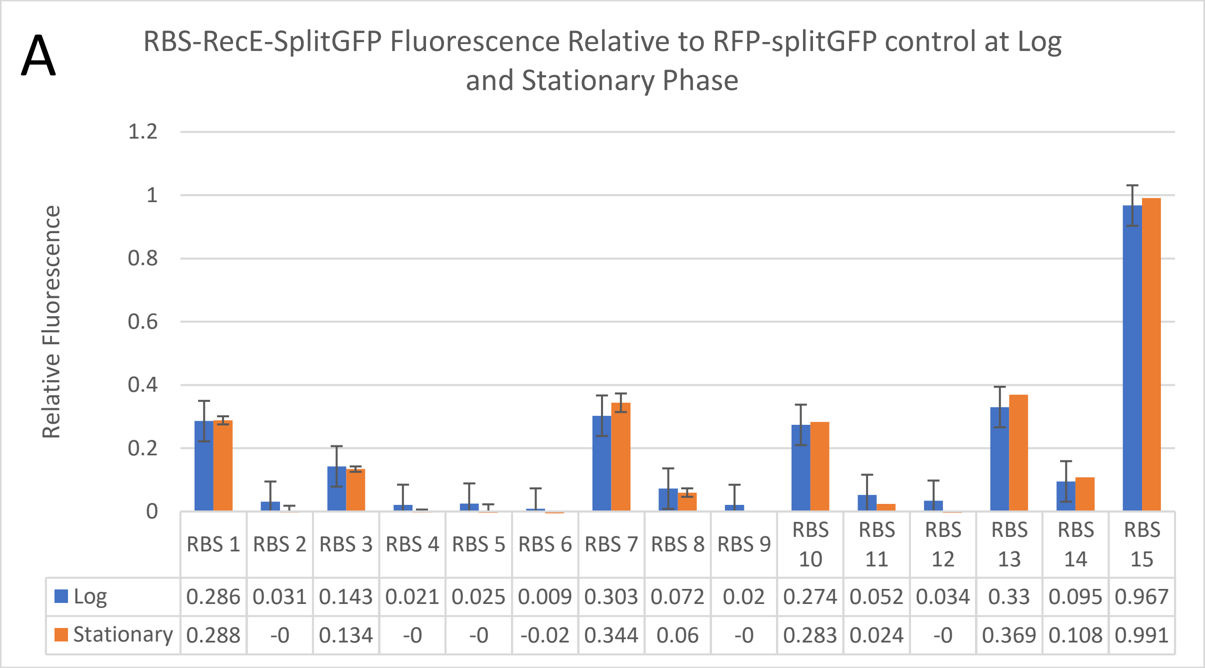


Figure 4.4: ***Characterising expression of the 5’-3’ exonuclease domain of DNA Pol-I in DH5a cells using a library of RBS sequences.*** *A – The fluorescence output indicates the level of expression of 5’-3’Pol-I-Exo in the cell. The fluorescence output from biological triplicates was averaged and is represented via the error-bars. This GFP fluorescence displayed by experimental samples is presented relative to the GFP fluorescence displayed by the mCherry-splitGFP control fusion protein. The relative fluorescence is presented for two time-points, one where cell cultures were in the log phase and one*

where cultures were in the stationary phase. B – The growth curve over 6-hours for cells expressing 5'-3'Pol-I-Exo with 15 different RBS sequences is shown in black. The control cells expressing an empty antibiotic-ORI vector are shown in red. The absorbance data from biological triplicates was averaged at each timepoint and plotted on the curve.

4.3.6 Results: Characterisation of the 5'-3' Exonuclease RecE

With RBS 1, 3, 7, 10, 13, 14 and 15, the cells expressing RecE had a similar growth rate to the control cells expressing the empty pSC101-Kanamycin backbone (Figure 4.5). All other RBS combinations allowed the cells to maintain a higher growth rate than the control. The relative fluorescence analysis reveals that all RBS sequences, except RBS15, expressed RecE at less than 20% of the expression level of mCherry. RBS 15 expressed RecE at the same rate as mCherry during both log and stationary phase of the growth curve. As a result, RBS15 was selected for building high expression systems for RecE, RBS7 for medium expression and RBS3 for low-level expression cassettes.



	RBS 1	RBS 2	RBS 3	RBS 4	RBS 5	RBS 6	RBS 7	RBS 8	RBS 9	RBS 10	RBS 11	RBS 12	RBS 13	RBS 14	RBS 15
■ Log	0.286	0.031	0.143	0.021	0.025	0.009	0.303	0.072	0.02	0.274	0.052	0.034	0.33	0.095	0.967
■ Stationary	0.288	-0	0.134	-0	-0	-0.02	0.344	0.06	-0	0.283	0.024	-0	0.369	0.108	0.991

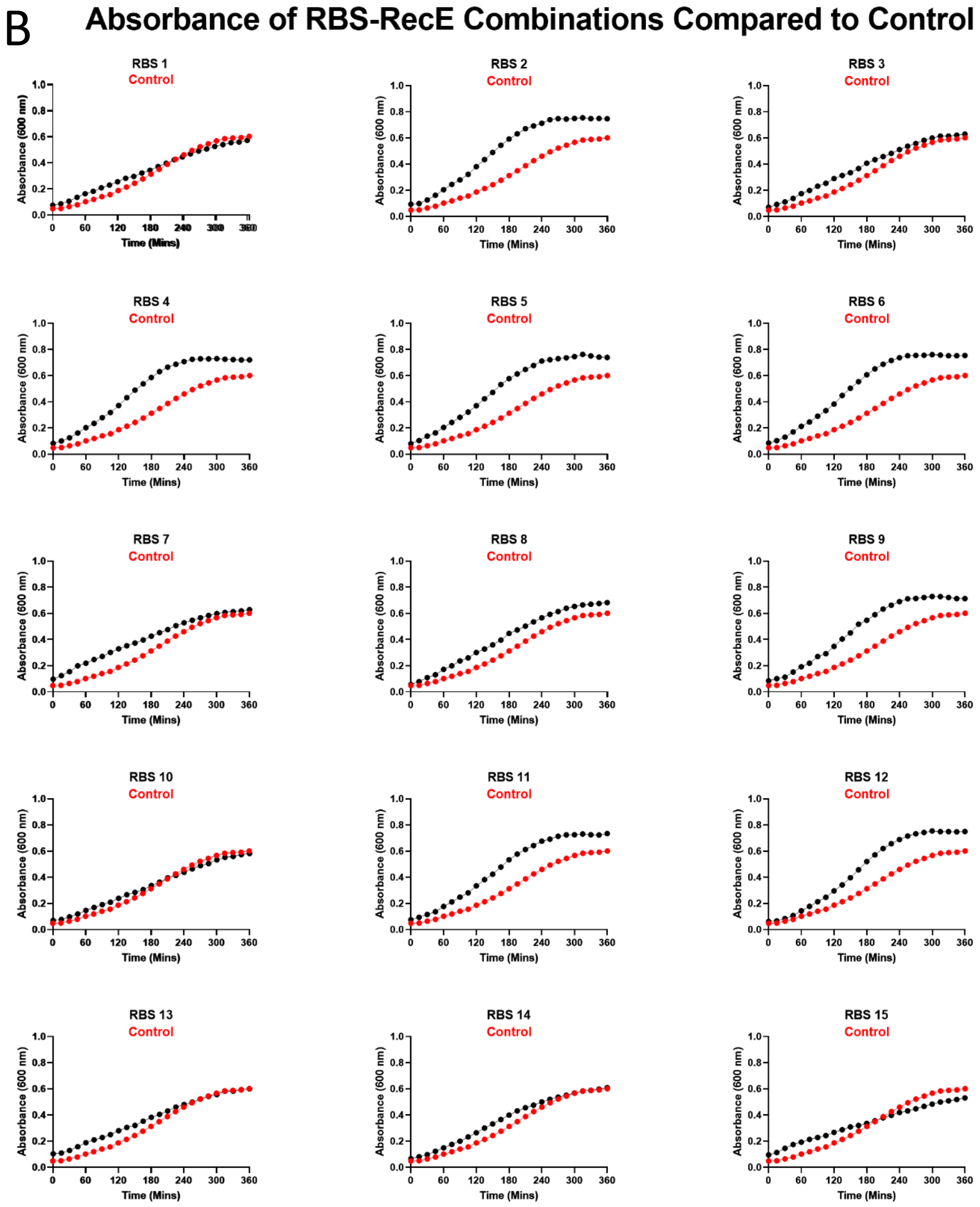


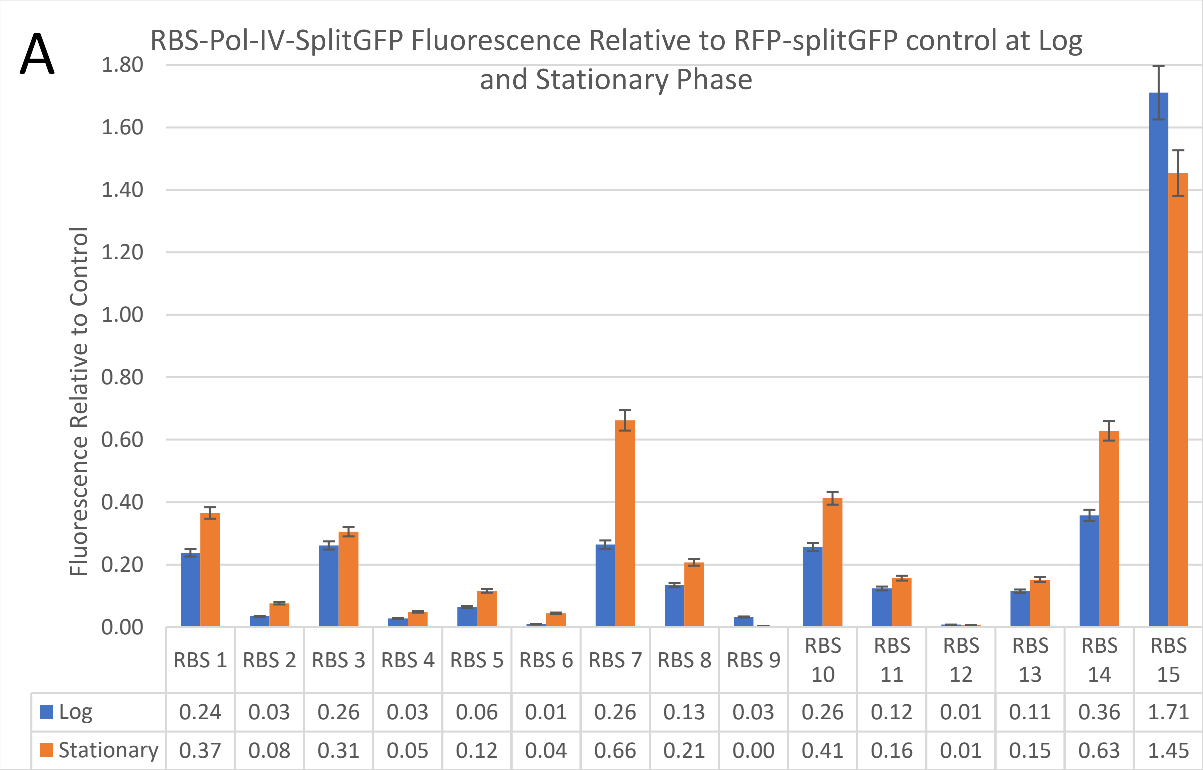
Figure 4.5: Characterising RecE expression in DH5a cells using a library of RBS sequences. *A – The fluorescence output indicates the level of expression of RecE in the cell. The fluorescence output from biological triplicates was averaged and is represented via the error-bars. This GFP fluorescence displayed by experimental samples is presented relative to the GFP fluorescence displayed by the mCherry-splitGFP control fusion protein. The relative fluorescence is presented for two time-points, one where cell cultures were in the log phase and one where cultures were in the stationary phase. B – The growth curve over 6-hours for cells expressing RecE with 15 different RBS sequences is shown in*

black. The control cells expressing an empty antibiotic-ORI vector are shown in red. The absorbance data from biological triplicates was averaged at each timepoint and plotted on the curve.

4.3.7 Results: Characterisation of the EP-DNA-polymerase, DNA Pol-IV:

The absorbance data for all RBS—Pol-IV—splitGFP expression cassettes, except RBS 14 and 15, revealed a stable growth rate of cells, with all cell cultures achieving stationary phase at an OD_{600} in the range of 0.8 (Figure 4.6). Except for RBS14 and RBS15, all other RBS sequences enabled Pol-IV expression with a lower fitness burden on cells at log phase, compared to cells expressing the RBS15—mCherry—splitGFP control. The cell cultures expressing RBS-Pol-IV displayed a faster growth rate in log phase.

Analysing the growth curve and relative fluorescence data, RBS7 was identified as a stable high expression RBS for building Pol-IV expression cassettes, which imposes minimal burden on the host cells. RBS10 was categorised as a medium strength RBS, and RBS3 would be used for building low expression level cassettes.



	RBS 1	RBS 2	RBS 3	RBS 4	RBS 5	RBS 6	RBS 7	RBS 8	RBS 9	RBS 10	RBS 11	RBS 12	RBS 13	RBS 14	RBS 15
🟦 Log	0.24	0.03	0.26	0.03	0.06	0.01	0.26	0.13	0.03	0.26	0.12	0.01	0.11	0.36	1.71
🟧 Stationary	0.37	0.08	0.31	0.05	0.12	0.04	0.66	0.21	0.00	0.41	0.16	0.01	0.15	0.63	1.45

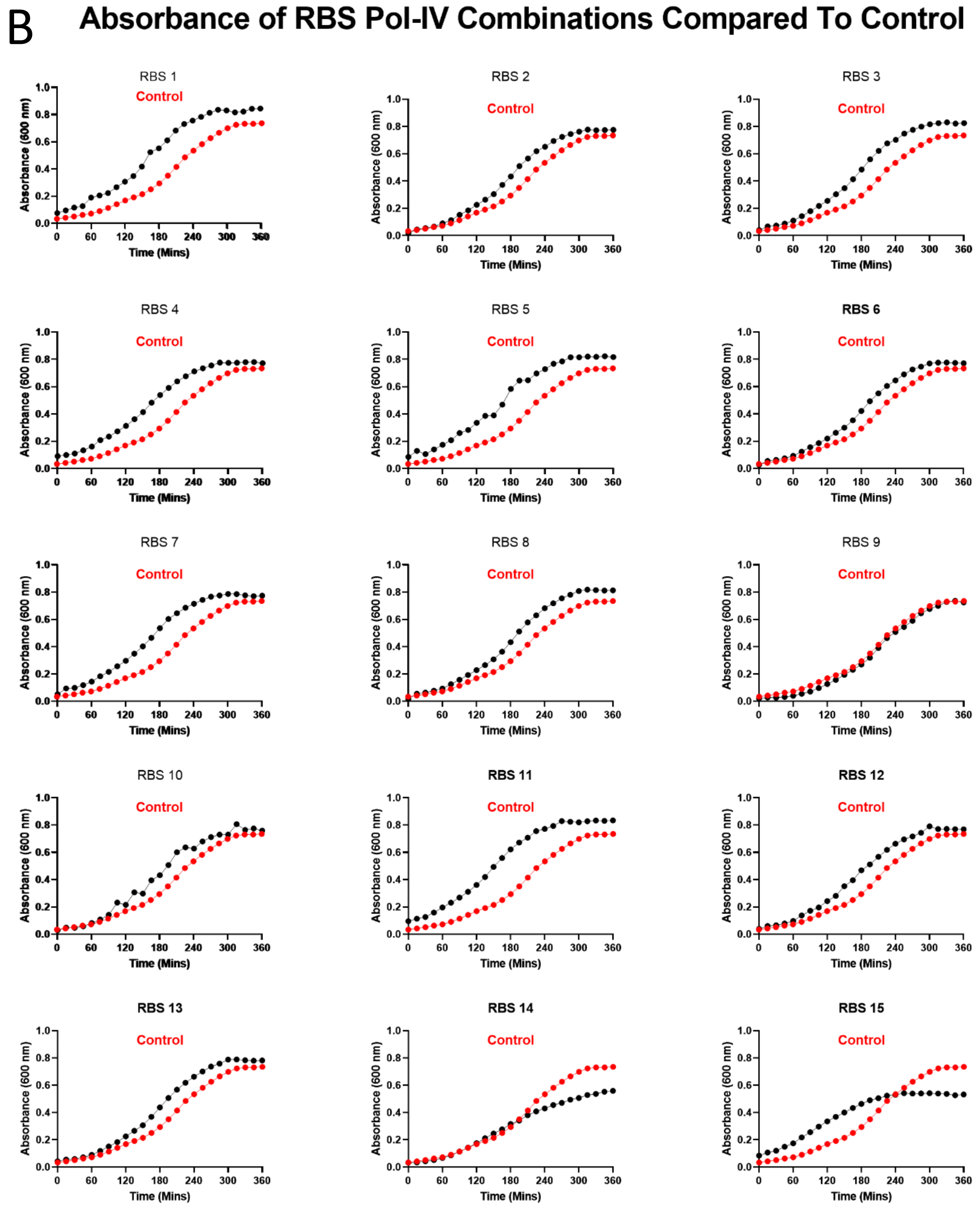


Figure 4.6: Characterising DNA Polymerase IV expression in DH5a cells using a library of RBS sequences. *A – The fluorescence output indicates the level of expression of DNA Pol-IV in the cell. The fluorescence output from biological triplicates was averaged and is represented via the error-bars. This GFP fluorescence displayed by experimental samples is presented relative to the GFP fluorescence displayed by the mCherry-splitGFP control fusion protein. The relative fluorescence is presented for two time-points, one where cell cultures were in the log phase and one where cultures were in the*

stationary phase. B – The growth curve over 6-hours for cells expressing DNA Pol-IV with 15 different RBS sequences is shown in black. The control cells expressing an empty antibiotic-ORI vector are shown in red. The absorbance data from biological triplicates was averaged at each timepoint and plotted on the curve.

4.3.8 Overall Trends Witnessed from the Characterising Data of Multiple Bioparts

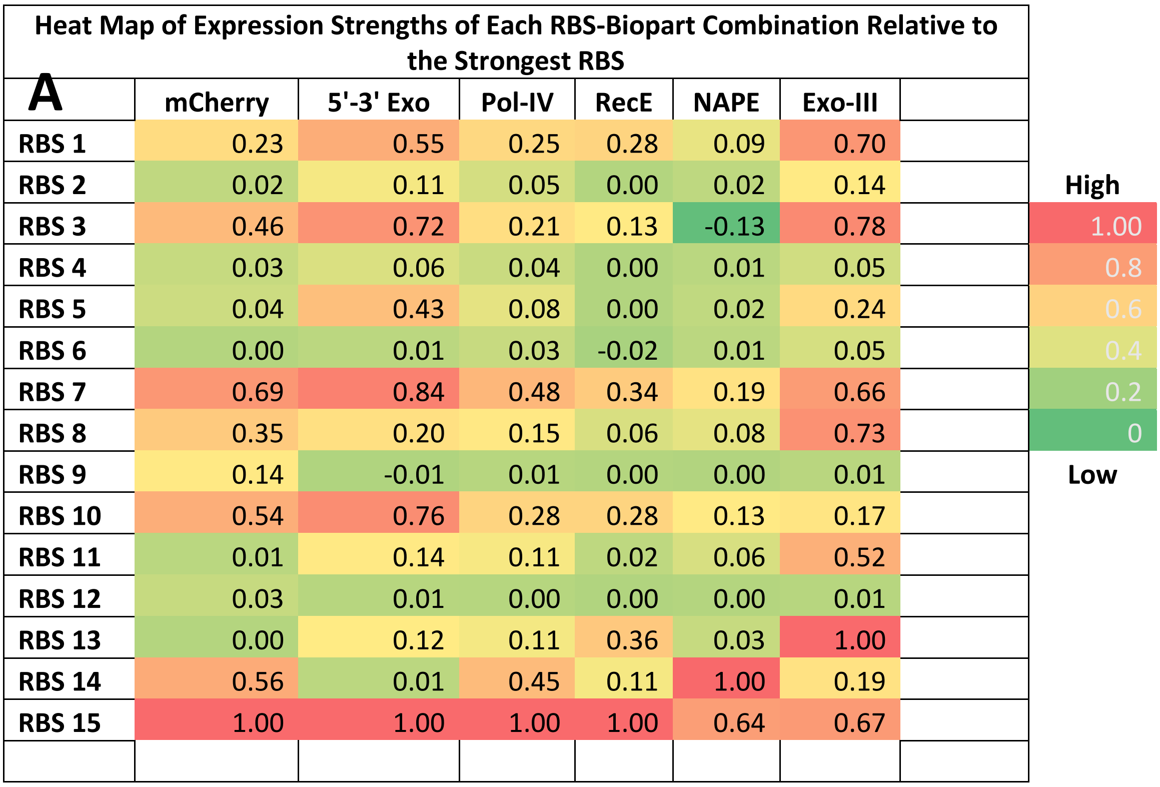
The relative fluorescence measured for each characterised RBS-biopart combination were mapped onto a heat map (Figure 4.7). This allowed for trends to be identified, in terms of the extent to which context dependency affects an RBS sequence from initiating mRNA translation[202]. While RBS sequences are known to be context dependent and the translation rate of the gene can differ depending on the 5'-UTR upstream of the open reading frame, the expression level of 4 out of the 6 bioparts achieved with each RBS sequence was comparable. Except for Exo-III and NAPE, all other bioparts displayed strongest expression with RBS 15 (BB0034 iGEM repository). RBS 1, 3, 7, 8, 10 and 14 displayed high to intermediate expression levels of these bioparts and the remaining RBS sequences displayed weak to minimal expression.

Another heat map was generated to illustrate the growth rate of the RBS-bioparts combinations relative to RBS-mCherry ($Relative\ Growth\ Rate = \frac{Growth\ Rate\ of\ Biopart\ with\ RBS}{(Growth\ Rate\ of\ mCherry\ with\ RBS}$). The relative growth rate of cells expressing the 6 different bioparts showed great variance amongst the 15 RBS-biopart combinations, except for 5'-3'Pol-I-Exo, whose expression did not impose fitness burden on the cell with any of the RBS sequences used (Figure 4.7, B). Even though RBS15 displayed the strongest expression level for most bioparts, it affected the fitness of cells when used to express RecE, NAPE, Exo-III and DNA Pol-IV. In these cases, the cells displayed a significantly lower OD_{600} at stationary phase. Observing the two heat maps, an inverse relationship between target protein expression and cell fitness can be seen. The higher the protein expression achieved, the lower was the relative growth rate of the cells. These findings are in line with published literature which have shown that the expression of synthetic genes in host cells competes for the finite resources in the cell for protein expression[144].

Overall, these characterisation experiments enabled us to accurately tune the expression of each candidate gene for the assembly of an error-prone DNA repair complex. Having such a diverse library of characterised expression rates for the different proteins created a great toolbox for iterative engineering of the mutator module expressing the AID-T7pol DNA damaging device and an error-prone DNA repair complex. If the expression of the two fusion proteins imposed a fitness burden during long mutagenic assays, the module could be edited to express AID-T7pol and the 5'-3'-Exo—

AP-Endo—EP-DNA-Polymerase fusions from a weaker RBS to alleviate the stress on the host cells and enable them to replicate efficiently and not be impeded by burden or toxicity.

This usefulness of this characterisation toolbox is exemplified in Section 4.8.1. When performing mutagenic assays with a high expression version of the mutator module, the cell fitness was reduced greatly, affecting cell growth as cultures achieved stationary phase at OD_{600} ~ 0.3. Having the characterisation data enabled quick troubleshooting and assembly of mutator modules where the EP-DNA-repair complex is expressed via a weaker RBS. The lowered expression level of the 3-protein fusion enabled the cells to maintain better fitness and cell cultures reaching an OD ~ 0.8, which was comparable to control cells (Figures 4.24 and 4.25). Therefore, having performed this characterisation experiments greatly benefitted the design-build-test-learn cycle for assembling functional mutator modules that imposed minimal fitness burden on host cells when the expression of AID-T7pol and the EP-DNA-repair complex was active.



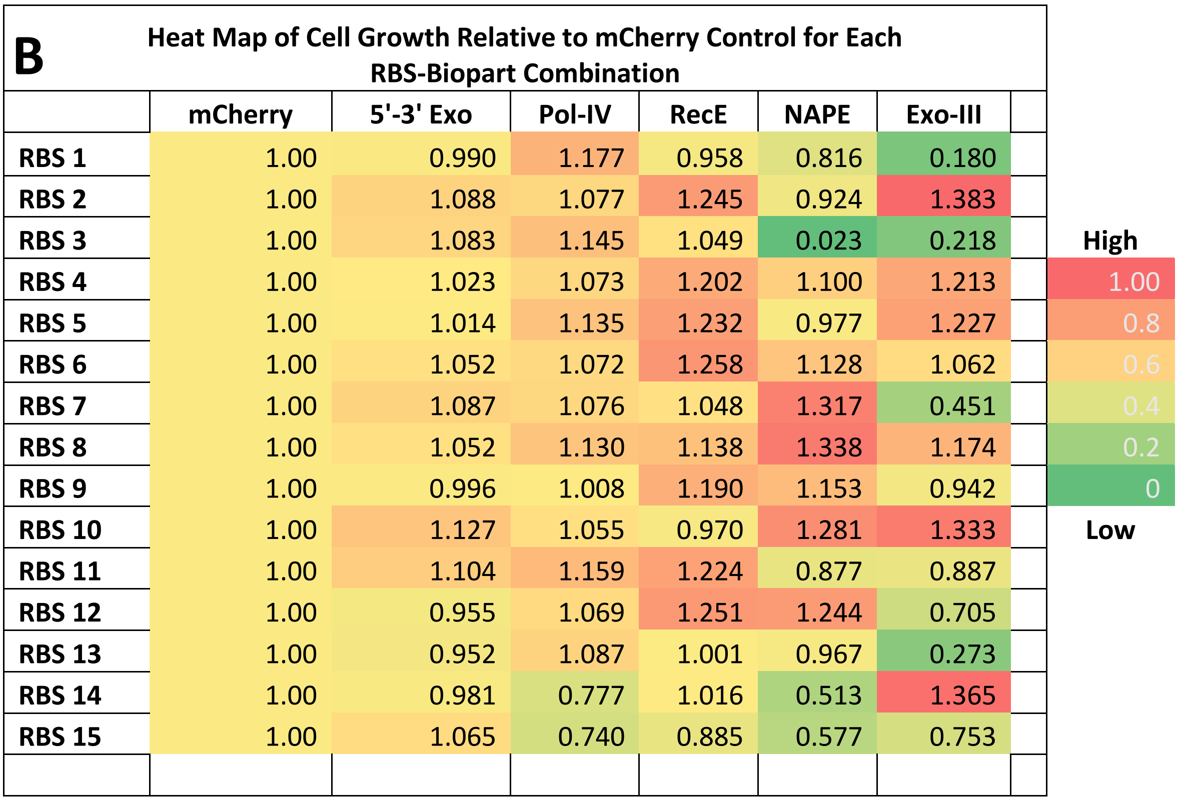
Heat Map of Expression Strengths of Each RBS-Biopart Combination Relative to the Strongest RBS

A	mCherry	5'-3' Exo	Pol-IV	RecE	NAPE	Exo-III	
RBS 1	0.23	0.55	0.25	0.28	0.09	0.70	
RBS 2	0.02	0.11	0.05	0.00	0.02	0.14	
RBS 3	0.46	0.72	0.21	0.13	-0.13	0.78	
RBS 4	0.03	0.06	0.04	0.00	0.01	0.05	
RBS 5	0.04	0.43	0.08	0.00	0.02	0.24	
RBS 6	0.00	0.01	0.03	-0.02	0.01	0.05	
RBS 7	0.69	0.84	0.48	0.34	0.19	0.66	
RBS 8	0.35	0.20	0.15	0.06	0.08	0.73	
RBS 9	0.14	-0.01	0.01	0.00	0.00	0.01	
RBS 10	0.54	0.76	0.28	0.28	0.13	0.17	
RBS 11	0.01	0.14	0.11	0.02	0.06	0.52	
RBS 12	0.03	0.01	0.00	0.00	0.00	0.01	
RBS 13	0.00	0.12	0.11	0.36	0.03	1.00	
RBS 14	0.56	0.01	0.45	0.11	1.00	0.19	
RBS 15	1.00	1.00	1.00	1.00	0.64	0.67	

High
1.00
0.8
0.6
0.4
0.2
0
Low

Figure 4.7: Heat map summarising the target protein expression rate and cell culture growth rate achieved with each RBS sequence. *A – The RBS strengths for each RBS-biopart are presented relative to the strongest RBS for said biopart. Except for Exo-III and NAPE, RBS 15 enabled the highest level of expression for each biopart. B – The heat map represents the optical density at stationary phase for each RBS-Biopart combination relative to the optical density achieved by cell cultures expressing RBS-*

mCherry. While RBS 15 enabled the highest levels of expression, it also resulted in burden which reduced the growth rate of the cells, as indicated by the lower optical density at 600 nm.



B Heat Map of Cell Growth Relative to mCherry Control for Each RBS-Biopart Combination

	mCherry	5'-3' Exo	Pol-IV	RecE	NAPE	Exo-III
RBS 1	1.00	0.990	1.177	0.958	0.816	0.180
RBS 2	1.00	1.088	1.077	1.245	0.924	1.383
RBS 3	1.00	1.083	1.145	1.049	0.023	0.218
RBS 4	1.00	1.023	1.073	1.202	1.100	1.213
RBS 5	1.00	1.014	1.135	1.232	0.977	1.227
RBS 6	1.00	1.052	1.072	1.258	1.128	1.062
RBS 7	1.00	1.087	1.076	1.048	1.317	0.451
RBS 8	1.00	1.052	1.130	1.138	1.338	1.174
RBS 9	1.00	0.996	1.008	1.190	1.153	0.942
RBS 10	1.00	1.127	1.055	0.970	1.281	1.333
RBS 11	1.00	1.104	1.159	1.224	0.877	0.887
RBS 12	1.00	0.955	1.069	1.251	1.244	0.705
RBS 13	1.00	0.952	1.087	1.001	0.967	0.273
RBS 14	1.00	0.981	0.777	1.016	0.513	1.365
RBS 15	1.00	1.065	0.740	0.885	0.577	0.753

High
1.00
0.8
0.6
0.4
0.2
0
Low

4.4 Engineering DNA Polymerase IV to Reduce its Global Mutagenic Activity

DNA polymerase IV is a lesion bypass polymerase that is naturally expressed in *E. coli* cells to overcome blockages in DNA replication[206]. The C-terminal residues of DNA Pol-IV interact with the β-clamp at the stalled replication fork and read over the DNA mismatch to continue the replication process[143,196,207]. DNA pol-IV has a larger DNA binding pocket than proofreading polymerases, which along with DNA Pol-IV possessing a low fidelity and no proofreading capability allows for skipping over the mismatch or lesion to continue DNA replication. Kuban et al demonstrated that DNA Pol-IV does not contribute significantly to the normal chromosomal error rate, by studying mutational frequencies in DinB$^-$ strains[xxi], compared to DinB+ cells[208]. However, DNA Pol-IV has an error rate of 10^{-3} base^{-1}, meaning there is a chance of a mismatch being incorporated every 1000 bases processed by this DNA polymerase. This high-error rate will naturally have a deleterious effect in cells if the dinB gene is overexpressed via a synthetic expression system, like the mutator module. Therefore, it was necessary to supress DNA pol-IV's function as a lesion bypass polymerase.

[xxi] DinB is the gene encoding DNA polymerase IV

To supress this function, a small chain of amino acid residues at the C-terminus, responsible for interacting with the β-clamp at stalled replication forks, were removed. Two different versions of DNA Pol-IV were generated in previous research by truncating five amino acids (Pol-IV$^{\Delta5}$) and truncating 12 amino acids (Pol-IV$^{\Delta12}$)[121].

4.4.1 Assessing Fitness Burden and Toxicity from Expressing Pol-IV variants in *E. coli*

The first step was to assess the fitness burden imposed on the host cell by the expression of Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ variants, compared to expressing Pol-IV^{WT}. Tet-inducible expression cassettes were assembled for Pol-IV^{WT}, Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ with the J23101$_{TetO}$ promoter and RBS3, RBS7 and RBS15 in a pSC101-Kanamycin backbone (Figure 4.8). These three RBS sequences from the characterisation of DNA Pol-IV enabled low, medium and high expression systems respectively for the Pol-IV variants. The array of nine expression plasmids were transformed into DH5α cells and grown overnight. The overnight cultures were diluted to an OD$_{600}$ of 0.05 in fresh LB media containing aTc to activate expression of the DNA polymerases. The growth rate of cell cultures was monitored using a plate reader. The expression circuits looked like the following:

[**J23101$_{\textbf{TetO}}$** $_{\text{U1-RBS15}}$ **Pol-IV** $^{\textbf{WT/Δ5/Δ12}}$] – [**TetR**] – pSC101 Kan^R	(Strong expression)
[**J23101$_{\textbf{TetO}}$** $_{\text{U1-RBS7}}$ **Pol-IV** $^{\textbf{WT/Δ5/Δ12}}$] – [**TetR**] – pSC101 Kan^R	(Intermediate expression)
[**J23101$_{\textbf{TetO}}$** $_{\text{U1-RBS3}}$ **Pol-IV** $^{\textbf{WT/Δ5/Δ12}}$] – [**TetR**] – pSC101 Kan^R	(Weak expression)

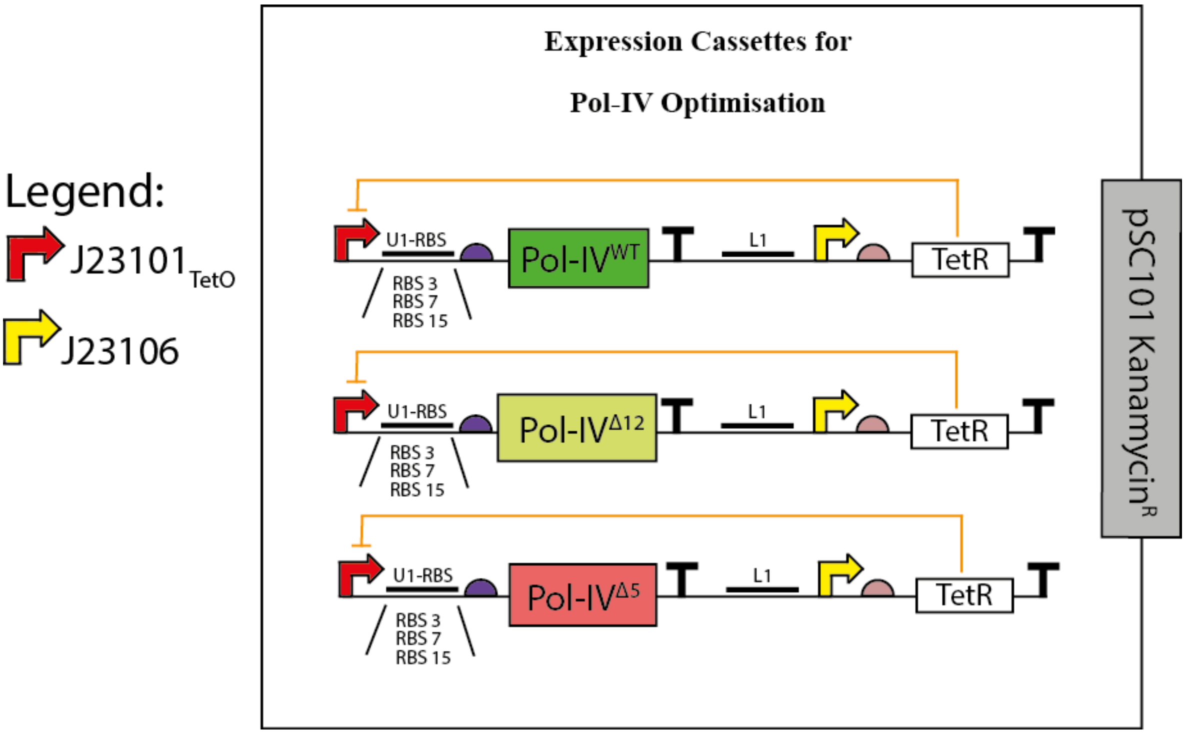


***Figure 4.8: The SBOL schematic of the expression circuits used to analyse the burden imposed by DNA Pol-IV variants**. Three different expression cassettes were assembled for each of the 3 different versions of DNA Pol-IV. Based on the characterisation data of Pol-IV, RBS 3, RBS 7 and RBS 15 enabled low, medium and high levels of expression of Pol-IV, respectively. The Tet repressor was placed in the expression circuit for inducible expression of the protein.*

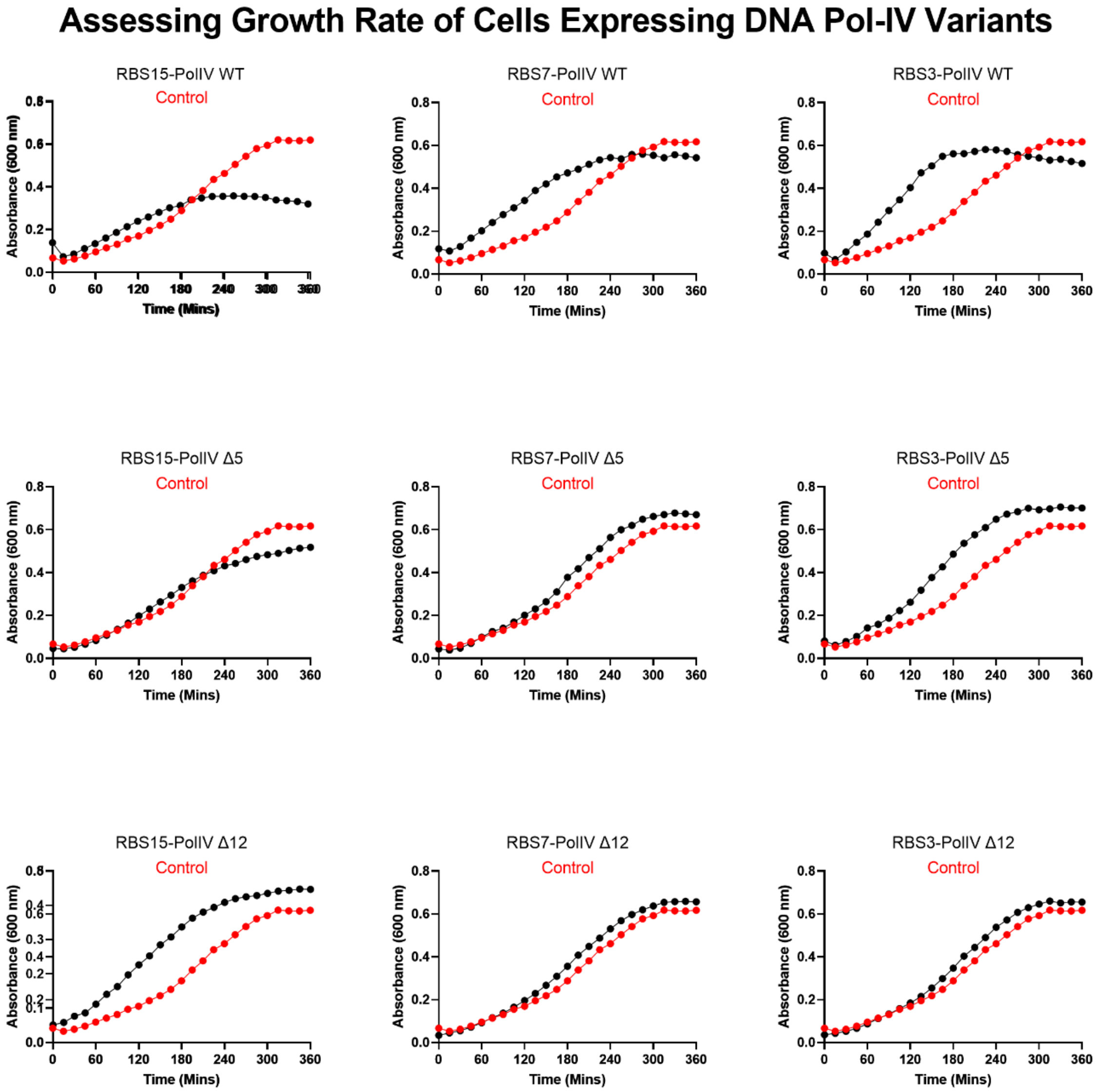


***Figure 4.9: The growth curves for cells expressing DNA Pol-IV WT, Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ using a strong (RBS15), medium (RBS7) and weak RBS (RBS3) are shown**. The black line represents the experimental sample, while the red line is for the negative control (cells expressing empty pSC101-Kan^{R} backbone). The absorbance data of biological triplicates was averaged at each time-point and plotted on the curve. The data shows that expressing Pol-IV with RBS-15 imposes a heavy burden on the cells. When being expressed via the medium and weak RBS, the cultures were able to achieve an OD~0.8 for*

Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$. Cultures expressing Pol-IV^{WT} displayed a non-sigmoidal growth pattern and achieved lower optical densities at stationary phase compared to Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$.

Analysing the growth curve over the 6-hour time course experiment showed that Pol-IVWT and Pol-IV$^{\Delta5}$ versions of the polymerase resulted in decreased fitness for the cells when RBS-15 was used in the expression system, resulting in high expression levels of the DNA Pol-IV variants (Figure 4.9). In the medium and low expression systems with RBS7 and RBS3, only cells expressing Pol-IV^{WT} displayed a decrease in fitness. The level of burden seen in cells expressing Pol-IV^{WT} in this experiment is different to what was observed in the characterisation experiment, where expression via RBS-7 and RBS-3 imposed minimal fitness burden[xxii]. This observed difference likely results from the absence of a β-11 of splitGFP being fused to the C-terminus of DNA Pol-IV in this experiment. This means the C-terminal resides of Pol-IV were free to interact with the β-clamp at replication forks, and possibly generated mutations in the genomic DNA[143]. The toxicity of active DNA Pol-IV most likely caused the decrease in fitness.

The cell cultures actively expressing Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ achieved an average OD_{600} of 0.7 at stationary phase (Figure 4.9), which was comparable to the control (DH5α cells expressing empty pSC101-Kan^{R} backbone). This data shows that Pol-IV^{WT} would not be an ideal EP-DNA-polymerase for use in the error-prone DNA repair complex due to its expression imposing fitness burden on the host cells, and due to the higher global mutagenic activity. Cells expressing PolIV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ via medium and low expression systems achieved similar growth rates to the control during log and stationary phase. The next step was to identify if these C-terminal truncations of the DNA Pol-IV protein resulted in reduced translesion replication activity in the cells. The rifampicin antibiotic selection assay was utilised for this assessment.

4.4.2 Assessing Global Mutation Rate of the Pol-IV Variants

Bacterial cells are not naturally resistant to the rifampicin antibiotic. However nucleotide substitutions in three distinct motifs of the RpoB gene confers rifampicin resistance[207,209]. The RpoB gene normally encodes the β-subunit of the RNA polymerase found in bacteria. Selecting for rifampicin resistance serves as a simple experiment to qualitatively analyse the effect of mutagenic proteins and chemicals on the genome. Cells expressing Pol-IV^{WT}, Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ with RBS-7 were induced with aTc and grown for 24-hours in 1-ml cultures. After the growth period, 800 µl of the cell cultures at an OD600 ~ 1.0 (640 million cells) from each sample were plated on LB-agar plates containing rifampicin (50 µgml^{-1}), and colonies were counted after 24-hours. The positive control for this experiment were

[xxii] Refer to Figure 4.6 – B for characterisation data of DNA Pol-IV

cells expressing AID-T7pol+UGI. As shown in Chapter 3, this control is highly mutagenic as UGI inhibits UNG-mediated glycosylation of deoxyuridine, consequently blocking repair of U:G lesions. The negative control were DH5α cells expressing the pSC101-Kan^{R} backbone.

Results:

Among the three DNA Pol-IV variants, the data showed Pol-IV^{WT} to have the highest global mutagenic activity with an average of 600 rifampicin^{+} colonies being spotted. Pol-IV$^{\Delta5}$ and Pol-IV$^{\Delta12}$ averaged 177 and 150 rifampicin^{+} colonies, respectively. Therefore, the C-terminal truncation led to a 4-fold decrease in the global mutagenic activity by DNA polymerase IV (Figure 4.10). This global mutation frequency was 2-fold higher than wildtype DH5α cells, which displayed ~ 70 rifampicin^{+} colonies. Overall, based on the growth curve analysis and rifampicin^{+} data, DNA Pol-IV$^{\Delta12}$ was selected for building the 2-protein AP-Endo—EP-DNA-Polymerase and 3-protein 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase fusion proteins of the EP-DNA-repair complex. This version of the DNA Pol-IV enabled cells to express the mutator module with minimal fitness burden and a significantly lower global mutation frequency, compared to DNA Pol-IV^{WT}.

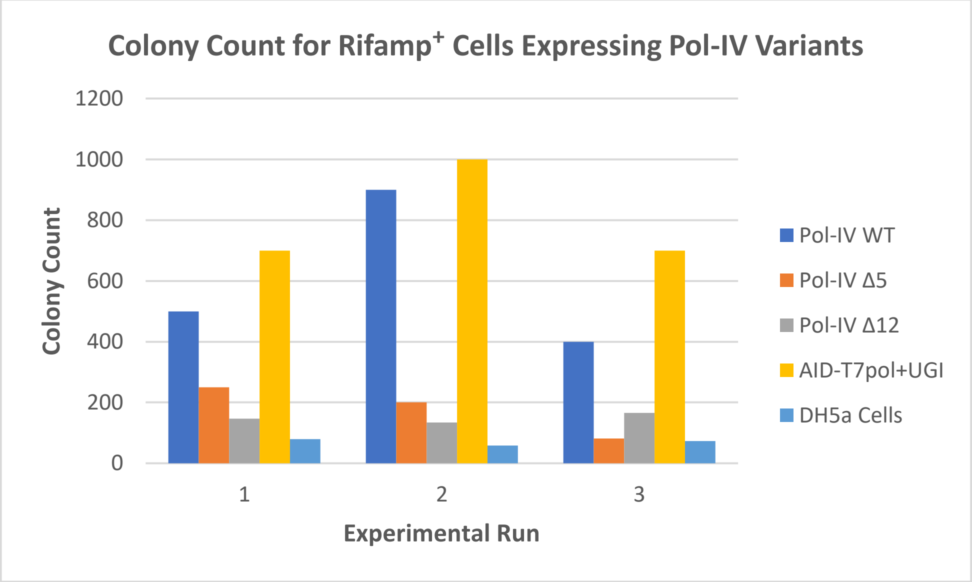


Figure 4.10: Assessing Global Mutation Rate of the three DNA Pol-IV variants using rifampicin reversion assays*. After 24-hrs of growth at 31°C, 800 µl of cell cultures were plated on LB-agar containing 50 µgml^{-1} of rifampicin. The number of colonies observed on the plate are shown. AID-T7pol+UGI was used as a positive control for global mutations, while native DH5α cells were used as a negative control to ascertain the natural mutation rate. The 12 amino acid truncated version of Pol-IV*

displayed 5-fold lower mutagenic activity than the wildtype polymerase and generated mutations at 2-fold higher frequencies than native DH5α cells.

4.5 Objective 2: Assembling a Library of Mutator Modules with DNA Pol-IV or EP-Pol-I in the Error-prone DNA Repair Complex

4.5.1 Step 1: Assembly of the Error-prone DNA Repair Complexes

Order of the protein domains for creating 5'-3'Exo—AP-Endo—EP-DNA-Polymerase fusion protein

To determine the order in which the 5'-3'-exonuclease, AP-endonuclease and the EP-DNA-polymerase should be fused to one another and still maintain functional activity, we looked into the order of functional domains in DNA polymerases encoded within a single gene. Rolf Daniel *et a*l conducted metagenomic library screens to identify homologs of DNA Pol-I in organisms such as *Algoriphagus, Pedobacter, Microscilla, Thermus, Acinetobacter, and Rhodococcus*[210]. Their work revealed that the structure of single gene DNA polymerases is highly conserved among different bacterial strains, such that it can be used as a phylogenetic marker. Similar work by Yi-Ping Huang showed conserved domain order and conserved amino acid residues in DNA Pol-I from four different species of bacteria[211]. In such single-gene encoded polymerases, the 5'-3' exonuclease is always the N-terminal domain, with the 3'-5' proofreading domain in the middle and the polymerase domain at the C-terminal end. This principle was followed in building the 5'-3'Exo—AP-Endo—EP-DNA-Polymerase fusion protein for the error-prone DNA repair complex.

With the expression of different bioparts having been characterised in the DH5α strain, the next step was to assemble the 2-protein and 3-protein fusion of AP-Endo—EP-DNA-Polymerase and 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase, respectively. Two different groups of EP-DNA-repair complexes were assembled, one group utilising Pol-IV$^{\Delta12}$ as the error-prone DNA polymerase, and the other group utilising the polymerase domains of error-prone DNA Pol-I variants generated by Loeb et al[141]. The expression cassettes for all the EP-DNA-repair complexes were assembled using J23101$_{\text{TetO}}$ as the tet-inducible promoter. The expression plasmids also contained an expression cassette with a constitutive promoter for the Tet repressor. Both expression cassettes were assembled on the pSC101-chloramphenicol^{R} backbone.

Assembly of EP-DNA-repair complexes with DNA Pol-IV$^{\Delta12}$

The first EP-DNA-polymerase tested for its mutagenic activity within the mutator module was DNA Pol-IV$^{\Delta12}$. Using the knowledge gained from RBS characterisation and the structure of conserved single-gene DNA polymerases[210], eleven different versions of the EP-DNA-repair complex were assembled using Pol-IV$^{\Delta12}$ in the following fusion protein design:

LMP **[J23101TETO** U1-RBSX **5'-3' Exonuclease** FL2 **AP-Endonuclease** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

Where colour coding is used to differentiate the different biopart functions used in the 2-part and 3-part fusion protein:

Green = 5'-3' Exonuclease, Yellow = AP-Endonuclease, Grey = EP-DNA-polymerase

The EP-DNA-polymerase domain was the C-terminal protein in all the EP-complexes assembled[210,211]. The AP-endonuclease was placed in the middle, while 5'-3' exonuclease was the N-terminal domain in the 3-protein EP-DNA-repair complexes. The RBS sequence used for expressing the 2-protein or 3-protein fusion protein depended on which biopart was the N-terminal domain in the protein fusion. The expression rate of the fusion protein would be dependent on the context between the RBS and the biopart forming the N-terminal domain. As a result, for the AP-Endo—EP-DNA-Polymerase versions of the EP-DNA-repair complex, the RBS sequence enabling highest expression of Exo-III and NAPE, while maintaining host cell fitness were chosen. These were RBS-8 for the expression of Exo-III—Pol-IV$^{\Delta12}$ and RBS-7 for expressing NAPE—Pol-IV$^{\Delta12}$. For the 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase, the RBS sequence selected for achieving high expression with low burden was RBS-7, based on the characterisation data for RecE and 5'-3'-Pol-I-Exo. The eleven different EP-DNA-repair complexes (EPRC) assembled were:

EPRC1: LMP **[J23101TETO** U1-RBS8 **Exo-III** FL2 **Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC2: LMP **[J23101TETO** U1-RBS7 **RecJ** FL2 **Exo-III** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC3: LMP **[J23101TETO** U1-RBS7 **RecE** FL2 **Exo-III** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC4: LMP **[J23101TETO** U1-RBS7 **NAPE** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC5: LMP **[J23101TETO** U1-RBS7 **RecJ** FL2 **NAPE** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC6: LMP **[J23101TETO** U1-RBS7 **RecE** FL2 **NAPE** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC7: LMP **[J23101TETO** U1-RBS7 **5'-3'Pol-I Exo** FL2 **Exo-III** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC8: LMP [**J23101**TETO U1-RBS7 **5’-3’Pol-I Exo** FL2 **NAPE** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC9: LMP [**J23101**TETO U1-RBS7 **5’-3’Pol-I Exo(s)** FL2 **Exo-III** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC10: LMP [**J23101**TETO U1-RBS7 **5’-3’Pol-I Exo(s)** FL2 **NAPE** FL3 **Pol-IV$^{\Delta12}$]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC11: LMP [**J23101**TETO U1-RBS8 **Exo-III** FL3 **Pol-IV^{WT}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

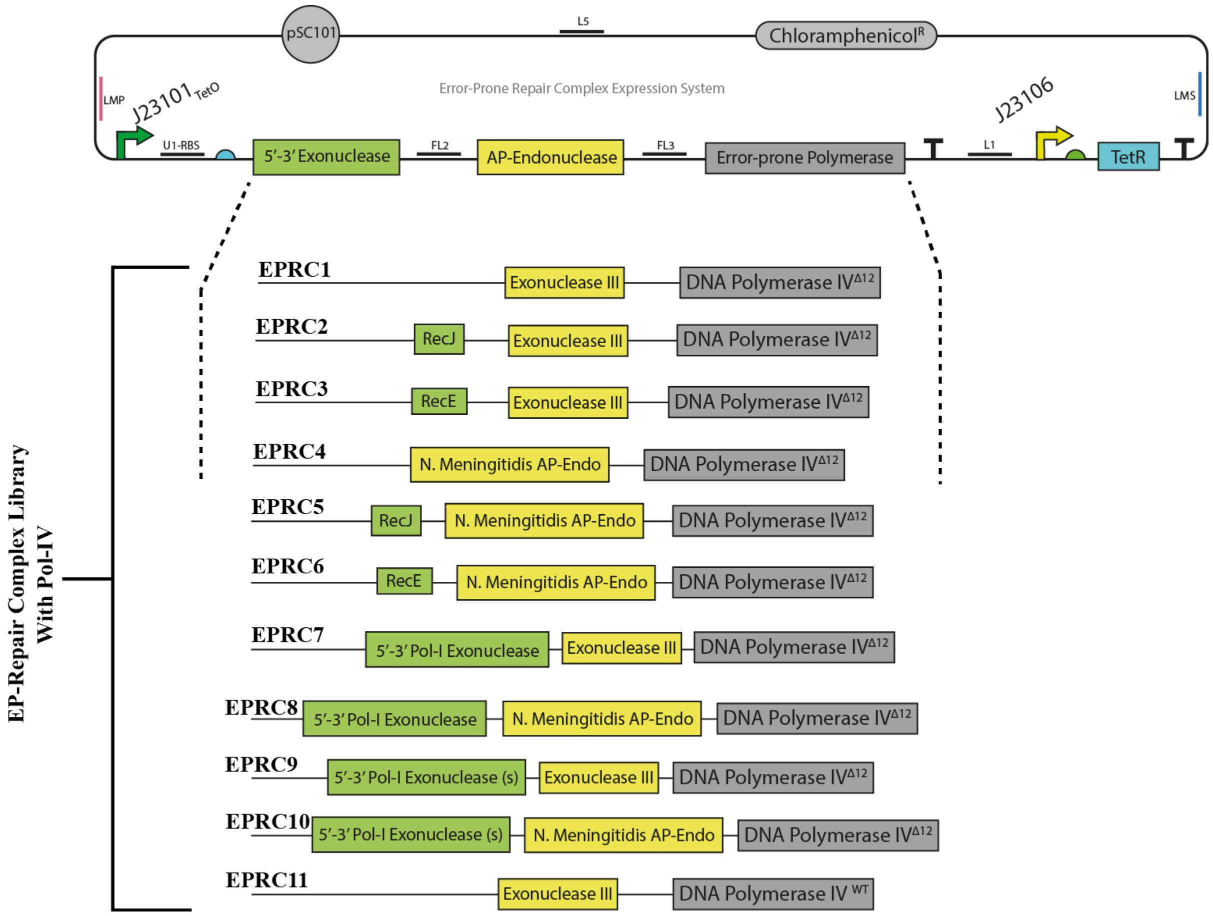

pSC101
L5
Chloramphenicol^{R}
Error-Prone Repair Complex Expression System
LMP
J23101 TetO
J23106
LMS
U1-RBS
5’-3’ Exonuclease
FL2
AP-Endonuclease
FL3
Error-prone Polymerase
L1
TetR
EP-Repair Complex Library With Pol-IV
EPRC1
Exonuclease III
DNA Polymerase IV$^{\Delta12}$
EPRC2
RecJ
Exonuclease III
DNA Polymerase IV$^{\Delta12}$
EPRC3
RecE
Exonuclease III
DNA Polymerase IV$^{\Delta12}$
EPRC4
N. Meningitidis AP-Endo
DNA Polymerase IV$^{\Delta12}$
EPRC5
RecJ
N. Meningitidis AP-Endo
DNA Polymerase IV$^{\Delta12}$
EPRC6
RecE
N. Meningitidis AP-Endo
DNA Polymerase IV$^{\Delta12}$
EPRC7
5’-3’ Pol-I Exonuclease
Exonuclease III
DNA Polymerase IV$^{\Delta12}$
EPRC8
5’-3’ Pol-I Exonuclease
N. Meningitidis AP-Endo
DNA Polymerase IV$^{\Delta12}$
EPRC9
5’-3’ Pol-I Exonuclease (s)
Exonuclease III
DNA Polymerase IV$^{\Delta12}$
EPRC10
5’-3’ Pol-I Exonuclease (s)
N. Meningitidis AP-Endo
DNA Polymerase IV$^{\Delta12}$
EPRC11
Exonuclease III
DNA Polymerase IV ^{WT}

Figure 4.11: SBOL schematic outlining the 11 error-prone DNA repair complexes assembled with the error-prone polymerase, DNA Pol-IV$^{\Delta12}$.

Assembly of EP-DNA-repair complexes with the error-prone polymerase domain from DNA Pol-I mutants

Lawrence Loeb *et al* did significant work in converting the proof-reading polymerase DNA Pol-I into an error-prone one by screening a library of mutants generated via EP-PCR[198]. The library of Pol-I mutants possessed a varying degree of processive fidelity compared to the wildtype polymerase[141]. The error rate of the mutants they generated ranged from 6-fold to 1100-fold greater than the wildtype (wildtype error rate is 1.8x10^{-5})[212]. We selected three such mutants from their mutant library and built

EP-DNA-repair complexes by replacing the Pol-IV$^{\Delta12}$ domain with the polymerase domain of these DNA Pol-I variants to create 5'-3'Exo—AP-Endo—EP-Pol-I fusion proteins. Eight EP-DNA-repair complexes were assembled using the four DNA Pol-I variants shown in Table 4.2, Exo-III as the AP-endonuclease and the long and short versions of the 5'-3' DNA Pol-I exonuclease.

Table 4.2: Mutations to the Polymerase domain of DNA Pol-I to achieve error-prone activity

Polymerase version	Error Rate (Relative to WT)	Mutations
Pol-I^{WT}	1.8x10^{-5}	-
Pol-I^{150}	150x more error prone	I709N
Pol-I^{46}	46x	F742Y, P796H
Pol-I^{1100}	1100x	I709N, A579R

EPRC12: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo** FL2 **Exo-III** FL3 **Pol-I^{WT}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC13: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo** FL2 **Exo-III** FL3 **Pol-I^{150}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC14: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo** FL2 **Exo-III** FL3 **Pol-I^{46}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC15: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo** FL2 **Exo-III** FL3 **Pol-I^{1100}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC16: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo(s)** FL2 **Exo-III** FL3 **Pol-I^{WT}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC17: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo(s)** FL2 **Exo-III** FL3 **Pol-I^{150}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC18: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo(s)** FL2 **Exo-III** FL3 **Pol-I^{46}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

EPRC19: LMP [**J23101TETO** U1-RBS7 **5'-3'Pol-I Exo(s)** FL2 **Exo-III** FL3 **Pol-I^{1100}]** L1 **[TetR]** LMS **pSC101** L5 **Cam^{R}**

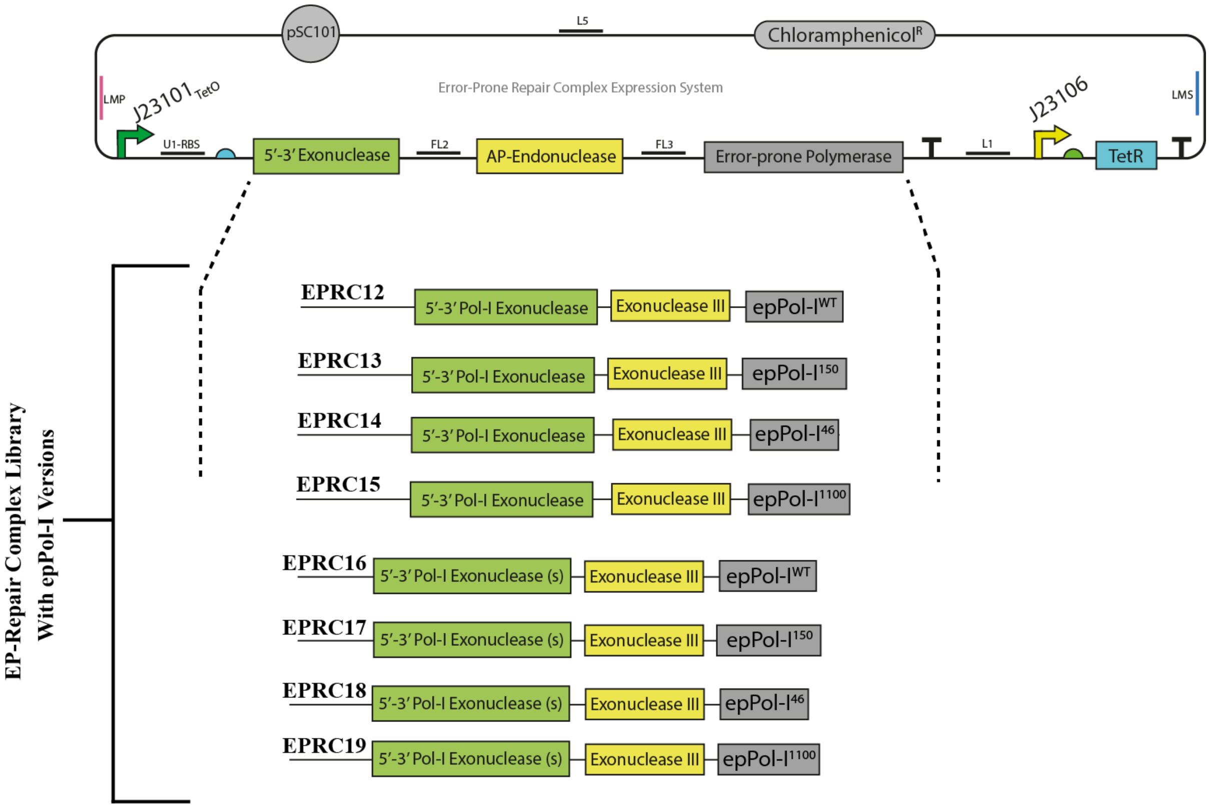


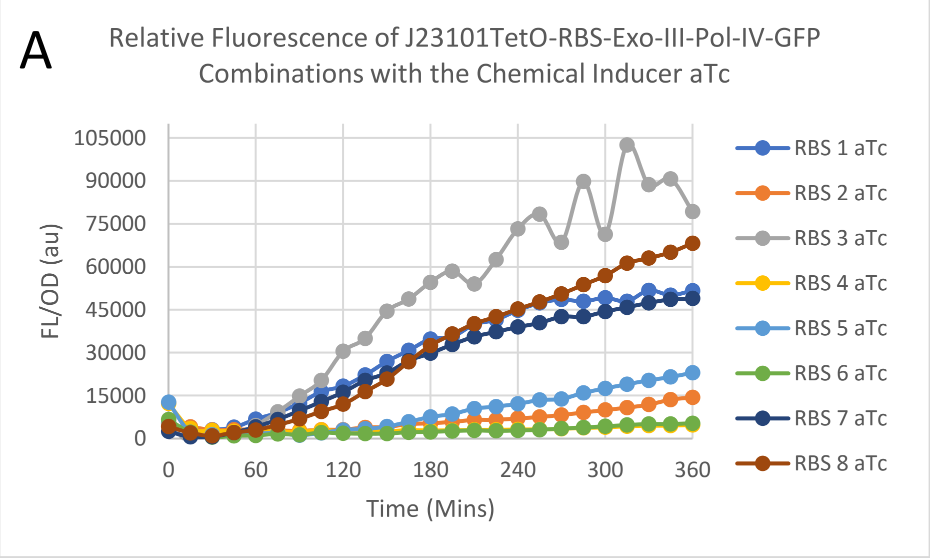
Figure 4.12: SBOL schematic outlining the 11 error-prone DNA repair complexes assembled with mutant polymerase domains of DNA Polymerase I. *Key amino acid substitutions made in the polymerase domain by Loeb and colleagues (shown in Table 4.3) resulted in DNA Pol-I mutants that were 46x to 1100x more error-prone than wildtype. These mutant polymerase domains were tested for their mutagenic capability in the EP-DNA-repair complex.*

4.5.2 Testing the switch-like activation for the error-prone DNA repair complex using TetR:

The mutator module is composed of two components, the AID-T7pol DNA damaging device and an error-prone DNA repair complex to generate genetic diversity. In a continuous evolution system, once the desired phenotype has been identified via selection, it is vital to prevent further mutation of the corresponding genotype. Therefore, the expression of AID-T7pol and the EP-DNA-repair complex is regulated by the TetR repression system, using a bespoke tet-inducible promoter (J23101$_{\text{TetO}}$). J23101$_{\text{TetO}}$ is the J23101 promoter from the Anderson library of promoters[213] modified to contain a tet operator (TetO). The expression of proteins in the mutator system is switched on by adding anhydrotetracycline in the media, which binds to TetR molecules and suppresses their DNA binding activity via a conformational change[147]. The expression can be subsequently shut off by removing aTc from the growth media. TetR is constitutively expressed in the mutator system, ensuring the expression of AID-T7pol and the EP-DNA-repair complex is repressed under normal conditions.

In chapter 3, we presented data confirming robust switch-like expression of the AID-T7pol using the Tet repression system[xxiii]. Similar testing was performed on the EPRC1 EP-DNA-repair complex, to verify that the expression of the 2-protein fusion can be switched on or off by the presence or absence of aTc in the growth medium for the cells. EPRC1 expression cassettes were assembled using eight different RBS sequences, and Exo-III—Pol-IV$^{\Delta12}$ was fused to splitGFP to monitor protein expression via fluorescence detected in a plate reader. The relative fluorescence (FL/OD) for the cell cultures expressing the different RBS—Exo-III—Pol-IV$^{\Delta12}$—splitGFP constructs with and without aTc was calculated.

The FL/OD values indicate a clear switch-like mechanism in the expression of the EPRC1 EP-DNA-repair complex (Figure 4.13). With aTc in the growth medium, each of the RBS—Exo-III—Pol-IV$^{\Delta12}$—splitGFP combinations achieved a higher relative fluorescence than their uninduced counterparts. Similar to the repression characteristics witnessed with AID-T7pol[xxiv], the J23101$_{\text{TetO}}$ promoter displayed low-levels of leaky expression. At the start of the growth assay, the FL/OD for the uninduced samples is relatively high, but as the cell cultures entered log phase, the relative fluorescence tapered off to zero. For cells cultures with aTc to activate expression of EPRC1, the fluorescence relative to optical density of the cell culture increased linearly until stationary phase. The clear switch-like behaviour makes the Tet-repression system ideal for building a feedback-based mutator system with AID-T7pol and an error-prone DNA repair complex.



[xxiii] Refer to Section 3.3.3

[xxiv] Figure 3.4 – D

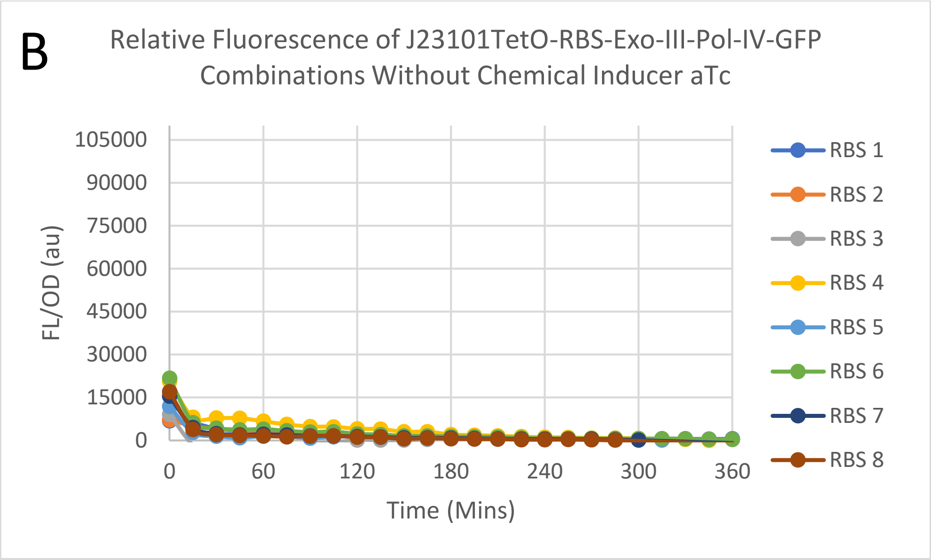


***Figure 4.13: Characterisation experiment to test the switch-like activation and repression of the EP-DNA-repair complex**. The RBS—Exo-III—Pol-IV$^{\Delta12}$—splitGFP combinations were grown in cultures with and without anhydrotetracycline. The recorded fluorescence values at each timepoint for biological triplicates were averaged and plotted on the graph for all RBS-Protein combinations. A – With aTc in the media, transcriptional repression due to the tet-repressor was released and Exo-III—Pol-IV$^{\Delta12}$ expression occurred with all RBS sequences tested. B – Without aTc, tet-repressor remains bound to the tet operator, repressing transcription via the J23101$_{TetO}$ promoter.*

4.5.3 Step 2: Assembly of the Mutator Modules by Combining the AID-T7pol and EPRC Expression Cassettes

The mutator module is designed to comprise an expression cassette for AID-T7pol to generate targeted DNA damage; an error-prone DNA repair complex to ensure mutations can be incorporated into the target gene around the site of DNA damage; and the Tet repressor to regulate the expression of the mutagenic proteins based on the presence of the activating ligand, aTc. The three expression cassettes were assembled into an expression plasmid using the BASIC DNA assembly methodology.

The process of combining the expression cassettes for AID-T7pol and EP-DNA-repair complexes was simplified by the modularity of the BASIC assembly technique[148]. The methylated linkers LMP and LMS contain functional recognition and cutting sites for the type-IIs restriction enzyme BsaI. DNA cargo enclosed within these two linker sequences can be cut from the initial assembled plasmids and used as parts in a subsequent assembly reaction. The expression cassette for AID-T7pol was excised from the expression plasmid discussed in Section 3.4; expression cassettes for the EP-DNA-repair complexes were excised from the EPRC1-19 plasmids[xxv]; and the two sequences of DNA were assembled onto a pSC101-Gentamycin^R backbone to create the complete mutator modules[xxvi].

AID-T7pol expression cassette designed in Chapter 3:

LMP **[J23101**TetO Deg RBS1 **AID-T7pol** L1 **Terminator]** LMS **[TetR]** L1 **pSC101** L5 **Cam^R**

Error-prone DNA repair complex cassette design:

LMP **[J23101**TETO U1-RBSX **5'-3' Exonuclease** FL2 **AP-Endonuclease** FL3 **EP-DNA-polymerase]** L1 **[TetR]** LMS **pSC101** L5 **Cam^R**

Assembled mutator module:

MUT-X: LMP **[AID-T7pol]** L1 **[5'-3' Exo—AP-Endo—EP-DNA-Polymerase]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^R**

Where [ ] = Complete expression cassette with promoter, RBS and terminator

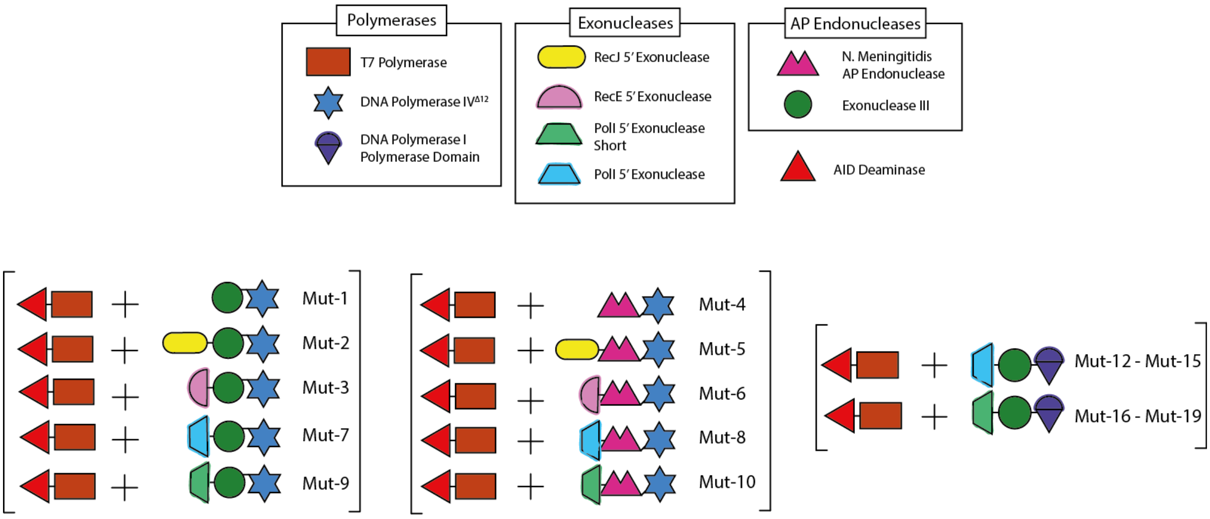


***Figure 4.14: Schematic of the EP-DNA-repair complexes combined with the AID-T7pol DNA damage device**. Nineteen different mutator modules (Mut-1 – Mut-19) were assembled for testing with the loss-of-function and gain-of-function experiments. All mutator modules were assembled in a pSC101-Gentamycin^R backbone.*

[xxv] Section 4.5.1 for plasmid design

[xxvi] Each assembled mutator module is designated an ID number in the format: MUT-X, where X is the unique ID number

Mutator Modules with DNA Pol-IV$^{\Delta12}$ EP-DNA-repair Complexes:

Eleven different mutator modules assembled with DNA Pol-IV$^{\Delta12}$ in the EP-DNA-repair complex, by combining AID-T7pol with EPRC1-EPRC11. The design of the assembled mutator modules was as follows:

MUT-1: LMP **[AID-T7pol]** L1 **[Exo-III— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-2: LMP **[AID-T7pol]** L1 **[RecJ—Exo-III— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-3: LMP **[AID-T7pol]** L1 **[RecE—Exo-III— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-4: LMP **[AID-T7pol]** L1 **[NAPE— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-5: LMP **[AID-T7pol]** L1 **[RecJ—NAPE— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-6: LMP **[AID-T7pol]** L1 **[RecE—NAPE— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-7: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo —Exo-III— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-8: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo —NAPE— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-9: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo(s)—Exo-III— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-10: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo(s)—NAPE— Pol-IV$^{\Delta12}$]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-11: LMP **[AID-T7pol]** L1 **[Exo-III— Pol-IV^{WT}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-1 and MUT-4 were assembled without a 5’-3’ exonuclease to investigate if this exonuclease is indeed necessary for performing patch repair. Exo-III possesses 3’-5’ exonuclease activity along with being an AP-endonuclease, which could potentially perform both the functions of generating gaps by nicking AP-sites and extending the gap in the 3’-5’ direction. Also, to investigate if the 12-amino acid truncation may have inactivated DNA Pol-IV’s polymerase activity, mutator module MUT-11 was assembled using the DNA Pol-IV^{WT}. If mutations occur in the target gene due to MUT-11 and not the mutator modules containing Pol-IV$^{\Delta12}$, it will indicate the truncation made the polymerase non-functional.

Mutator Modules with EP-DNA-Pol-I Error-prone DNA repair complexes

Eight different mutator modules were assembled expressing AID-T7pol and the EP-DNA-repair complexes EPRC12-EPRC19[xxvii] containing the error-prone polymerase domain of mutant DNA Pol-I. The plasmids expressing these mutator modules were:

MUT-12: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo —Exo-III— Pol-I^{WT}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-13: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo —Exo-III— Pol-I^{150}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-14: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo —Exo-III— Pol-I^{46}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-15: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo —Exo-III— Pol-I^{1100}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-16: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo(s)—Exo-III— Pol-I^{WT}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-17: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo(s)—Exo-III— Pol-I^{150}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-18: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo(s)—Exo-III— Pol-I^{46}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

MUT-19: LMP **[AID-T7pol]** L1 **[5’-3’Pol-I Exo(s)—Exo-III— Pol-I^{1100}]** L2 **[TetR]** LMS **pSC101** L5 **Gentamycin^{R}**

Once the Pol-IV and EP-Pol-I versions of the mutator modules were assembled, the next step was to analyse their mutagenic capability using the loss-of-functions and gain-of-function experiments. The mutator modules that displayed a high mutation frequency at the target gene and performed each of the four nucleotide substitutions investigated in the gain-of-function experiment would be shortlisted as candidates to be used for library generation in a continuous evolution system.

[xxvii] Section 4.5.1 for the plasmid design

4.6 Objective 3: Using Loss of Fluorescence Assays[xxviii] to Screen Mutator Modules

Based on Their Mutagenic Activity

The GM31 strain of *E. coli* with P_{T7}—GFP-mut3b—T7-Terminator integrated into its genome was used for screening the mutational capabilities of the mutator modules utilising DNA Pol-IV in the EP-DNA-repair complex. Before the complete library of 11 mutator modules could be analysed, an appropriate expression strength for the EP-DNA-repair complexes needed to be identified that generated detectable mutagenic activity. If the expression level of the EP-DNA-repair complex is low, fewer molecules of the 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase in the cytosol would result in fewer interactions with the target DNA at AP-sites and potentially fewer mutations being incorporated in the gene-of-interest per bacterial cell cycle. If this mutation frequency is too low, the mutator system would fail to generate a diverse library of mutants within a feasible amount of time. Therefore, to identify an appropriate expression strength for the mutator module, MUT-1 was assembled with 15 different RBS sequences expressing the Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex (Figure 4.15). The 15 versions of MUT-1 were used to knock out GFP fluorescence in GM31^{GFP}, and the degree of fluorescence loss within the cell populations was used to assess the relative strength of each version of MUT-1.

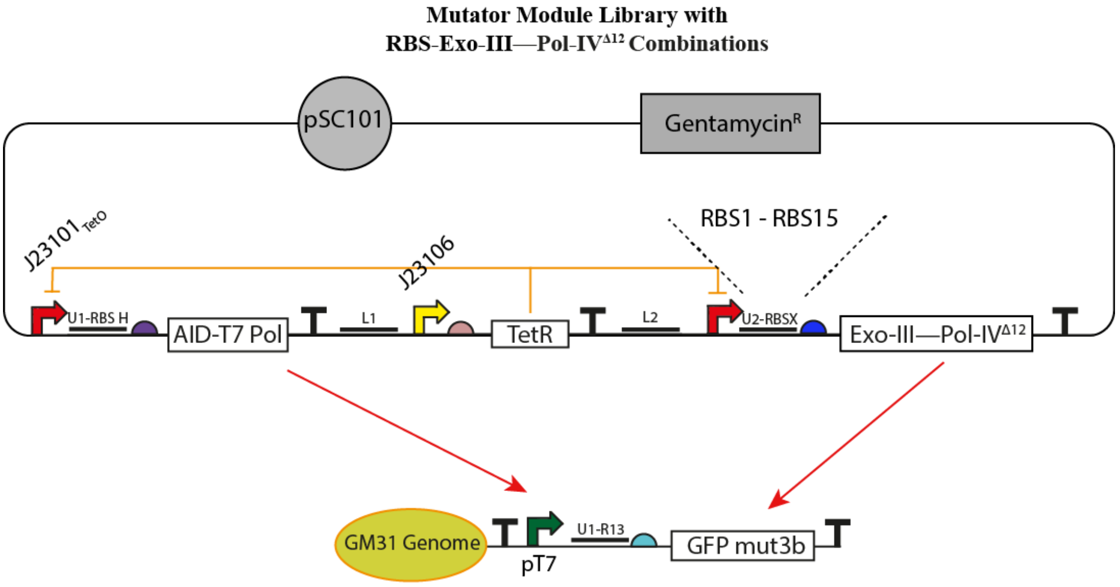


Figure 4.15: SBOL schematic of loss-of-function experiment with different RBS combinations of MUT-1 (Exo-III— Pol-IV$^{\Delta12}$). *The experiment was done to elucidate the expression level of the EP-DNA-repair complex to achieve high mutator activity, while imposing minimal burden on cells.*

[xxviii] Section 2.5.2 for an overview of the loss-of-function assay workflow

4.6.1 Screening MUT-1 with the complete RBS Library to Identify Appropriate Expression Strengths for the Mutator Module

After 24-hours of growth with aTc in the LB media to activate the expression of AID-T7pol and Exo-III— Pol-IV$^{\Delta 12}$, the GFP expression profiles were monitored using flow cytometry.

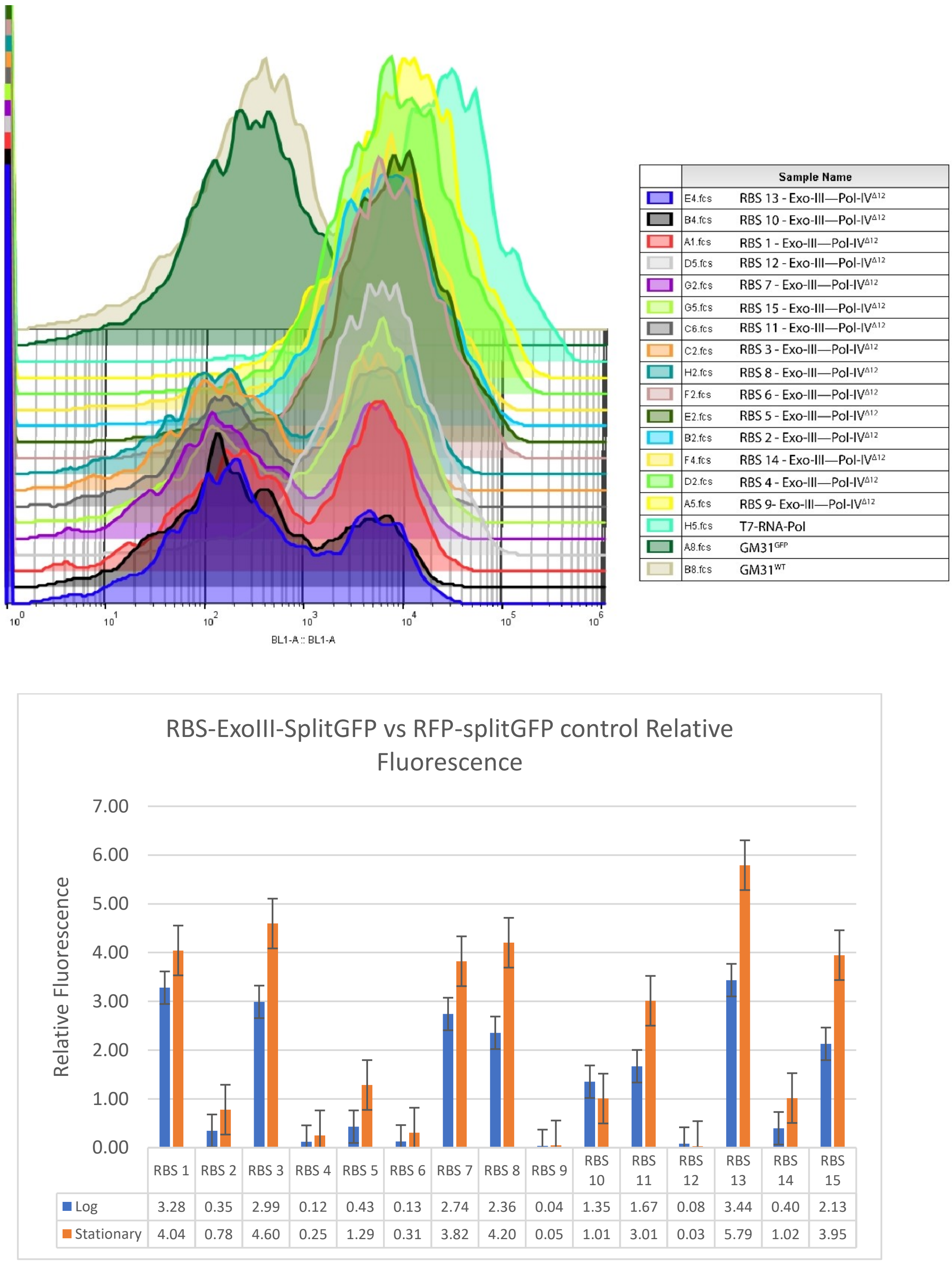
The level of fluorescence loss correlated strongly with the expression level of the Exo-III—Pol-IV$^{\Delta 12}$ EP-DNA-repair complex. RBS sequences that were characterised as strong expressors for Exo-III (RBS1, RBS3, RBS7, RBS8, RBS13 and RBS15) displayed the greatest loss of fluorescence (Figure 4.16). EP-DNA-repair complex expression via these RBS resulted in an average fluorescence that was two orders of magnitude lower in intensity compared to GFP expression via T7-RNA-Pol. Roughly 60-100% of these cell populations exhibited a loss of fluorescence, indicating strong mutator activity within a 24-hour period. The RBS sequences characterised as weak in context to Exo-III resulted in lesser portions of the population losing GFP fluorescence. The intermediate strength RBS, RBS11, resulted in a split population, with half the cells having lost fluorescence. RBS10 and RBS12 were two exceptions where a high degree of fluorescence loss was seen, but in the characterisation data, the two RBS sequences were classified as weak expressors for Exo-III.

To confirm that the loss of fluorescence was the result of error-prone DNA repair of the GFP open reading frame (ORF) and not because of other biological factors, two mutant GFP expression cassettes resulting from RBS8-Exo-III—PolIV$^{\Delta 12}$ were selected and random and sequenced. Sanger Sequencing of the samples confirmed the presence of C → T and G → A mutations in the GFP-mut3b ORF (Figure 4.17). This sequencing data provided insight into the following characteristics about the MUT-1 mutator module:

1. The U:G lesions created by AID-T7pol in the GFP-mut3b ORF were converted into C → T and G → A mutations by MUT-1. These mutations were generated without UGI being expressed in the system, meaning the cell's native BER pathway was active during the course of the mutator experiment. This indicates that the EP-DNA-repair complex is functional and is the most likely cause for nucleotide substitutions being incorporated at the U:G lesions.

2. The occurrence of mutations indicates that the fusion of Exo-III—Pol-IV$^{\Delta 12}$ with the flexible BASIC linker FL2 allows both proteins to fold into functional conformations and perform their designated functions in the EP-DNA-repair complex.

These findings encouraged us to assemble MUT-1 – MUT-11 with a high expression RBS to screen the mutator modules with DNA Pol-IV at their highest possible mutagenic strength.



		Sample Name
🟦	E4.fcs	RBS 13 - Exo-III—Pol-IV$^{\Delta 12}$
⬛	B4.fcs	RBS 10 - Exo-III—Pol-IV$^{\Delta 12}$
🟥	A1.fcs	RBS 1 - Exo-III—Pol-IV$^{\Delta 12}$
⬜	D5.fcs	RBS 12 - Exo-III—Pol-IV$^{\Delta 12}$
🟪	G2.fcs	RBS 7 - Exo-III—Pol-IV$^{\Delta 12}$
🟩	G5.fcs	RBS 15 - Exo-III—Pol-IV$^{\Delta 12}$
⬛	C6.fcs	RBS 11 - Exo-III—Pol-IV$^{\Delta 12}$
🟧	C2.fcs	RBS 3 - Exo-III—Pol-IV$^{\Delta 12}$
🟦	H2.fcs	RBS 8 - Exo-III—Pol-IV$^{\Delta 12}$
🟫	F2.fcs	RBS 6 - Exo-III—Pol-IV$^{\Delta 12}$
🟩	E2.fcs	RBS 5 - Exo-III—Pol-IV$^{\Delta 12}$
🟦	B2.fcs	RBS 2 - Exo-III—Pol-IV$^{\Delta 12}$
🟨	F4.fcs	RBS 14 - Exo-III—Pol-IV$^{\Delta 12}$
🟩	D2.fcs	RBS 4 - Exo-III—Pol-IV$^{\Delta 12}$
🟨	A5.fcs	RBS 9- Exo-III—Pol-IV$^{\Delta 12}$
🟩	H5.fcs	T7-RNA-Pol
🟩	A8.fcs	GM31^{GFP}
⬜	B8.fcs	GM31^{WT}

	RBS 1	RBS 2	RBS 3	RBS 4	RBS 5	RBS 6	RBS 7	RBS 8	RBS 9	RBS 10	RBS 11	RBS 12	RBS 13	RBS 14	RBS 15
🟦 Log	3.28	0.35	2.99	0.12	0.43	0.13	2.74	2.36	0.04	1.35	1.67	0.08	3.44	0.40	2.13
🟧 Stationary	4.04	0.78	4.60	0.25	1.29	0.31	3.82	4.20	0.05	1.01	3.01	0.03	5.79	1.02	3.95

***Figure 4.16: Flow cytometry analysis of the loss of GFP fluorescence at a population level achieved with different RBS—Exo-III—Pol-IV$^{\Delta12}$ combinations**. Cell populations were divided into fluorescent (10^4 au) and non-fluorescent (10^2 au) sub-populations comparative to the GM31 control cells (Dark Green). The RBS—Exo-III—Pol-IV$^{\Delta12}$ combinations were arranged based on mutagenic strength (Purple to Cyan). The cell populations with the largest degree of fluorescence loss expressed Exo-III—Pol-IV$^{\Delta12}$ via RBS 1, 2, 7, 8 and 13. Except RBS2, these RBS sequences resulted in the highest expression levels of*

Exo-III in the characterisation assay presented in Section 4.3.3. This means there is a clear correlation between the level of expression of the EP-DNA-repair complex and the degree of targeted mutagenic activity on the GFP-mut3b ORF.

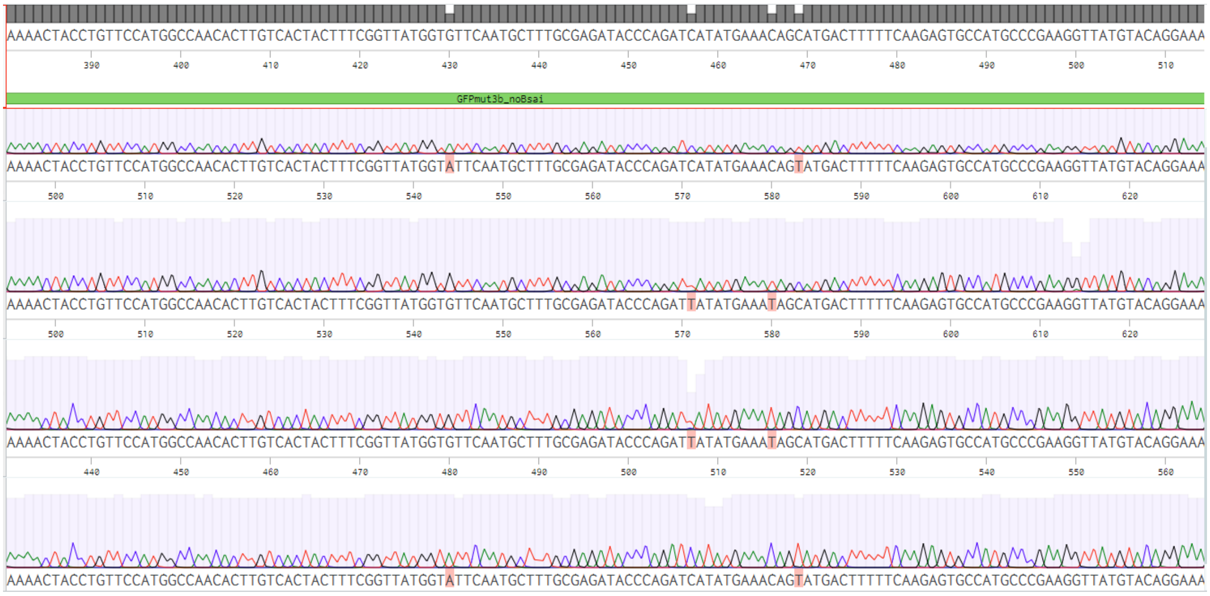


***Figure 4.17: Sanger sequencing of GFP-mut3b ORF after loss-of-function assay with MUT-1**. After 24-hours of generating mutations on the GFP-mut3b ORF by MUT-1, two mutant ORF were isolated from the genome of non-fluorescent cells and sent for sequencing with appropriate forward and reverse primers. The loss of fluorescence resulted from C → T and G → A mutations in the ORF. Due to the small sample size sequenced here, a spectrum of diverse nucleotide substitutions was not detected. To elucidate the full mutation profile that can be generated by a mutator module comprising AID-T7pol and an EP-DNA-repair complex, NGS analysis was performed in Chapter 5.*

4.6.2 Screening the 11 Pol-IV Mutator Modules for Targeted Mutagenic Activity with the Loss of Function Experiment

The mutator modules (MUT-1-MUT-11) were assessed for their mutagenic activity by assembling them into high expression cassettes with a strong contextual RBS for the N-terminal protein in the AP-Endo—EP-DNA-Polymerase and 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion proteins[xxix]. Comparing the different mutator modules at their highest expression levels enabled strong mutator systems to be easily differentiated from the weaker ones. These strong mutator systems could be

[xxix] Strong contextual RBS identified from Heatmaps shown in Section 4.3.8. An RBS that enabled high expression while imposing lowest possible fitness burden was selected for expressing the EP-DNA-repair complex.

optimised in the future with the characterisation toolbox to achieve the ideal balance between mutation frequency and host cell fitness during a continuous evolution experiment. If a mutator module displayed weak mutagenic activity in its strongest viable expression system, it would not be ideal for library generation as it would fail to create a genetically diverse library within each cell replication cycle. Such mutator systems displaying weak targeted mutagenic activity were identified and discarded from the pool of candidate mutator modules for further downstream testing.

MUT-1-MUT-11 were transformed into GM31 *E. coli* cells containing a genome integrated P_{T7}-GFP expression cassette. Biological triplicates were induced with aTc (10 ngµl^{-1}) and cell cultures were grown for 24-hours. The cell cultures were subsequently diluted in PBS and analysed in a flow cytometer. Of the two different AP-endonucleases tested, mutators using Exo-III generated mutations at a significantly higher frequency (Figure 4.18). MUT-1, MUT-2, MUT-3, MUT-7 and MUT-9 resulted in a complete loss of fluorescence compared to Control 3, where GFP is being expressed using T7-RNA-Pol. With these mutator modules active, 70%-100% of the cell population had lost the fluorescence phenotype (Figure 4.18).

With NAPE as the AP-endonuclease, the mutator modules only resulted in a small decrease in fluorescence, indicating weak mutagenic activity on the GFP-mut3b ORF. This weak activity could be the result of NAPE being inefficient in a foreign cellular environment. The low mutagenic activity could also result from the low expression level of NAPE with RBS7. This RBS expressed NAPE at 20% of the expression level with the strongest RBS, RBS14. RBS14 and RBS15, which provided higher expression levels of NAPE than RBS7, but could not be used in this loss-of-function experiment due to the fitness burden witnessed in cell cultures expressing these RBS-NAPE combinations[xxx]. Expressing a 5'-3'-exonucleases with NAPE—Pol-IV$^{\Delta12}$ did not impact the mutagenic activity of the NAPE-based EP-DNA-repair complexes.

A key finding was the difference in the mutagenic activity of Control 1 and Control 2 (Figure 4.18). In Control 1, AID-T7pol was expressed with DNA Pol-IV$^{\Delta12}$ and in Control 2, AID-T7pol was expressed with DNA Pol-IV^{WT}. Control 1 displayed significantly lower mutagenic activity than Control 2, which indicates that the 12-amino acid truncation made to DNA Pol-IV lowered its global mutagenic activity by potentially reducing DNA Pol-IV's ability to interact with the β-clamp at replication forks. Also, comparing Control 1 to mutator modules where Pol-IV$^{\Delta12}$ is fused to Exo-III shows that the EP-DNA-repair complex is able to localise the error-prone DNA polymerase to the site of U:G lesions, as the EP-

[xxx] Refer to Figure 4.3

DNA-Polymerase in these instances displayed significantly higher targeted mutagenic activity on the GFP-mut3b ORF than when the EP-DNA-Polymerase was freely expressed in Control 1.

The low mutation frequency displayed by NAPE-based mutator modules can only be tolerated if these systems can generate a diverse range of nucleotide substitutions in the target DNA. So even if the mutation frequency is low, at least the mutator system can generate a wide genetic diversity of the gene-of-interest. To investigate the Exo-III and NAPE-based mutator modules for their ability to generate specific nucleotide substitutions, the gain-of-function experiments with inactivated β-lactamase genes was performed.

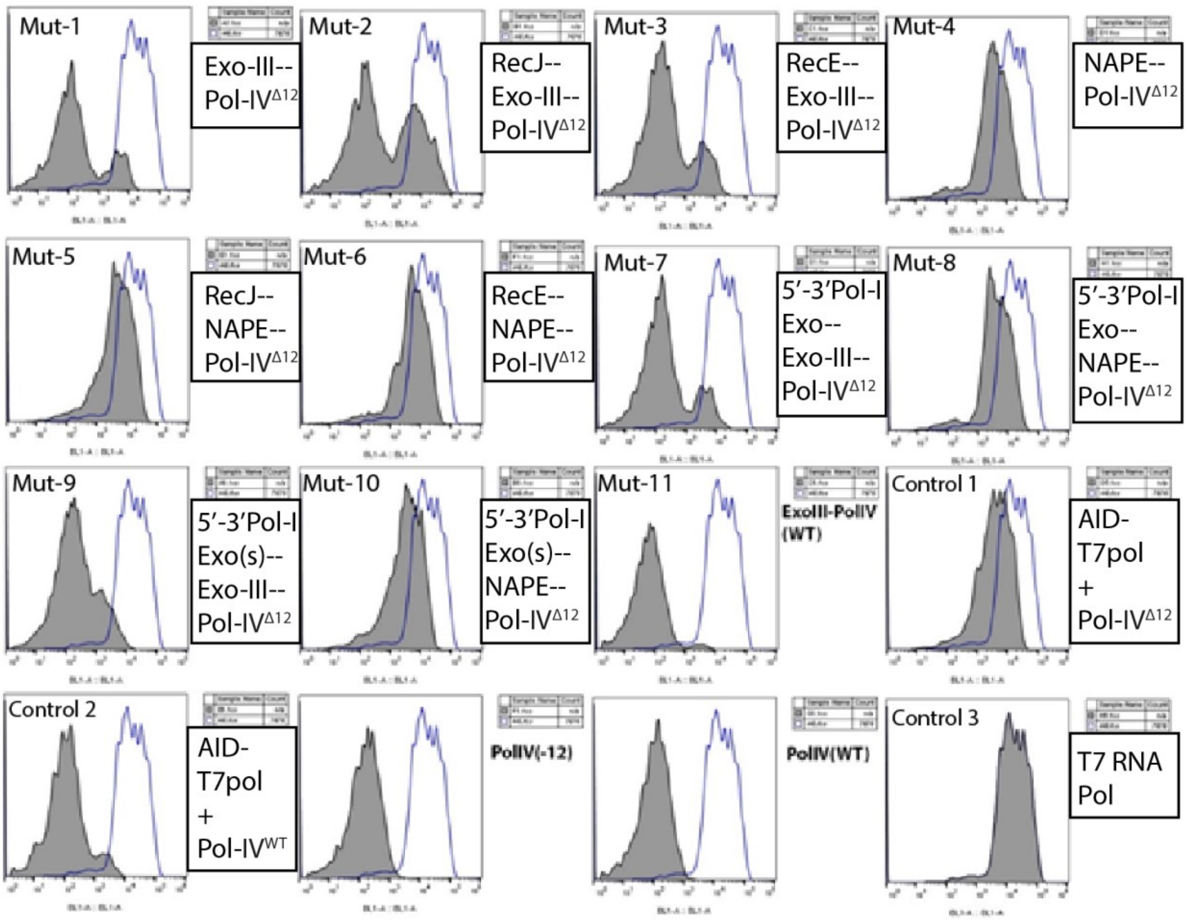


Figure 4.18: Results of the loss-of-function experiment with MUT-1 – MUT-11 analysed using flow cytometry*. The mutator modules utilising Exo-III as the AP-endonuclease resulted in a greater loss of fluorescence than mutator modules implementing NAPE. This possibly indicates that Exo-III is more active in E. coli cells than the foreign N. meningitidis AP-endonuclease. The loss of fluorescence is represented relative to Control 3, where GFP-mut3b is being expression with T7 RNA polymerase. Controls 1 and 2 express AID-T7pol with only Pol-IV$^{\Delta 12}$ and Pol-IV^{WT}, respectively. GFP loss seen in these controls is most likely the result of polymerase IV global mutagenic activity resulting from binding to*

β-clamp at replication forks. Control 1 caused significantly less GFP loss than Control 2, indicating the 12-amino acid truncation at the C-terminus greatly reduced DNA Pol-IV's global mutagenic activity.

4.7 Objective 4: Gain of Function Experiment to Identify Mutator Modules Capable of Performing Specific Nucleotide Substitutions

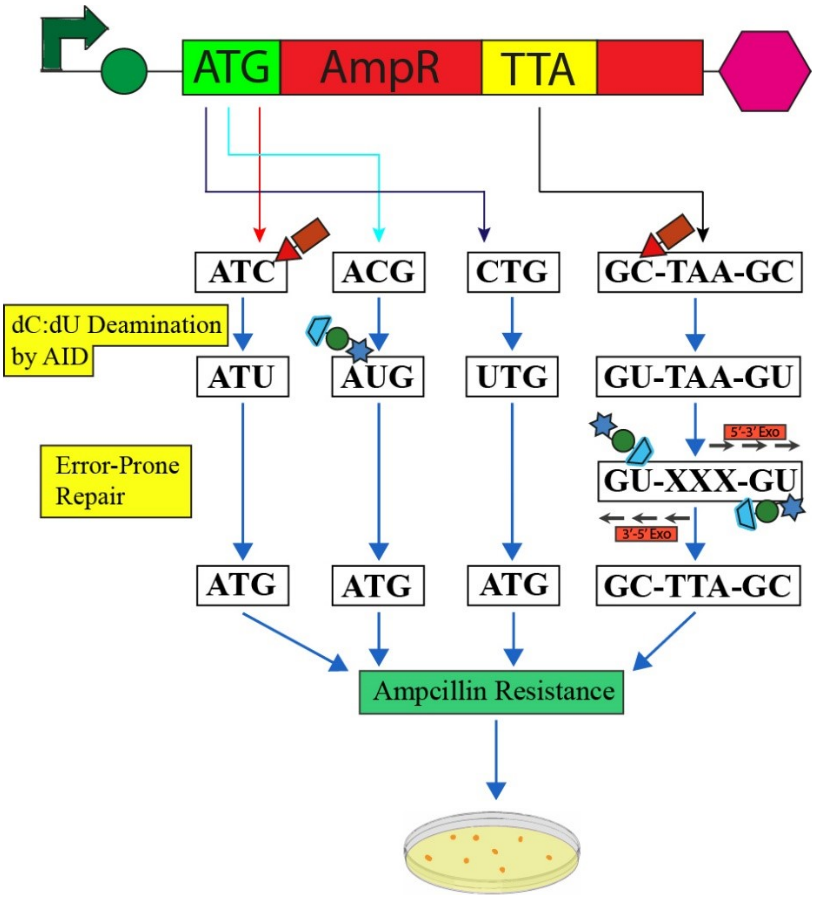


***Figure 4.19: Outline of the gain of function assay with inactivated β-lactamase gene variants**. The β-lactamase gene is made non-functional via point mutations in the ATG start codon and by converting a TTA into a TAA stop codon within the gene. These non-functional β-lactamase genes were targeted for mutations using the mutator modules. If mutator activity results in functional β-lactamase, the phenotype will be selected on carbenicillin plates. Such a selection assay provided a qualitative indication of the genetic diversity capable using the mutator modules. ATC, ACG and CTG required error-prone DNA polymerase activity, while reverting TAA required a 5'-3'Exonuclease synergising with the error-prone DNA polymerase to perform patch repair, for achieve a functional gene.*

Analysing the complete mutation profile generated by the mutator modules would require extensive DNA sequencing to elucidate the types of mutations that are generated (substitutions, insertions, deletions); the frequency with which they occur; and the nucleotide positions in the target gene ORF where they occurred. Next generation sequencing (NGS) would enable such a high-throughput analysis of the complete mutation characteristics of the mutator system[214,215]. However, due to the cost of using NGS platforms and due to the intricate sample preparation techniques, it was not feasible to analyse all 19 assembled mutator modules using NGS techniques. We therefore adopted a gain-of-function antibiotic selection assay[xxxi] to qualitatively assess the 11 DNA Pol-IV-based and 8 EP-Pol-I based mutator modules for their ability to perform C → A, T, G, T → A nucleotide substitutions and patch repair. Performing these nucleotide substitutions at specific nucleotide positions was the required condition to gain resistance to the chosen antibiotic. The mutator modules that were shown to reliably performed these nucleotide substitutions and patch repair were then shortlisted for comprehensive analysis using the Illumina iSeq100 and PacBio Sequel NGS platforms (See Chapter 5).

The ability to perform the four nucleotide substitutions was investigated using an antibiotic selection assay with β-lactamase (amp^R), which confers resistance to ampicillin and carbenicillin antibiotics. The translation of β-lactamase mRNA was inactivated by creating A → C, T → C and G → C mutations in the ATG start codon (Figure 4.19). As the AID-T7pol DNA damage device deaminates deoxycytidine, therefore dC was added at each position in the start codon to ensure the mutator module's activity can be directly targeted to the inactivated translation start site.

Another key characteristic of the EP-DNA-repair complex that required investigation was the ability to synergise 5'-3' exonuclease activity with the EP-DNA-Polymerase to perform error-prone patch repair. This would involve excising nucleotides upstream of the U:G lesion and creating a larger gap to be filled-in by the EP-DNA-Polymerase, increasing the likelihood of mutations being incorporated at A:T sites. The ability to perform patch repair was investigated by changing a TTA codon within a GC-rich region of the β-lactamase gene to a TAA stop codon (Leucine at position 205 changed to stop). In this instance, to gain the amp^R phenotype, the AID-T7pol must deaminate deoxycytidine in the G:C rich regions flanking the TAA stop codon; then the gap generated by either Exo-III or NAPE at the AP-site must be extended by the 5'-3' exonuclease to the TAA site; and finally, the EP-DNA-Polymerase must add nucleotides sequentially from the AP-site to the TAA stop codon, with a chance of an A → T transversion even occurring, converting TAA back to TTA encoding leucine. Cells displaying the amp^R phenotype in this instance would confirm that all three proteins in the 5'-3'-Exo—AP-Endo—EP-DNA-Polymerase fusion retain their native function and work synergistically to create genetic diversity in the target gene.

[xxxi] Section 2.5.2 for a detailed overview of the gain-of-function methodology

4.7.1 Gain of Function Experiment with MUT-1 – MUT-11 DNA Pol-IV Mutator Modules

From the GFP loss-of-function experiment, we identified the relative mutagenic strength of the 11 Pol-IV mutator modules, and confirmed via DNA sequencing that the loss of GFP fluorescence resulted from C → T and G → A substitutions introduced in the target gene ORF by the EP-DNA-repair complex. The next goal was to investigate if these mutator modules are capable of performing nucleotide substitutions other than C → T.

Methodology:

Competent DH5α cells were transformed with expression plasmids the four inactivated Amp^R mutants (assembled on p15A-Kan^R). The 11 different mutator modules were then transformed into the cells containing each inactivated Amp^R mutant shown in Figure 4.19. Individual colonies were picked into a 96-well plate and grown overnight and subsequently diluted in fresh LB media with and without the inducer molecule, aTc, to activate the expression of AID-T7pol and the EP-DNA-repair complexes. After 24-hours of growth, cell cultures with the mutator modules switched on were plated as 3 µl spots of cells (roughly 2.5 million cells plated for each Amp^R-mutator-module combination) on a LB-agar plate containing carbenicillin (50 µgml^{-1}) and gentamycin (25 µgml^{-1}). The cell cultures with the mutator modules switched off were plated under identical conditions on another LB-agar plate with carbenicillin and gentamycin.

Results:

TetR regulates mutator module expression with a switch-like mechanism

Mutation events resulting in the Amp^R phenotype was only witnessed for the cells expressing the mutator modules with aTc in the growth medium. For the cell cultures lacking aTc, no colonies were seen on the carbenicillin + gentamycin plate (Figure 4.20). We can conclude from this data that the TetR repression system successfully shuts off the expression of AID-T7pol and the EP-DNA-repair complexes to the extent that no visible functional mutations were generated in the pool of 2.5 million cells spotted onto the agar plate. In terms of the wild-type β-lactamase control, carbenicillin-resistant cells were displayed when the expression of AID-T7pol was switched off with no aTc in the media. This could result from the slight leakiness of the J23101$_{TetO}$ promoter[xxxii], where AID-T7pol is expressed in miniscule amounts and transcribes the β-lactamase gene downstream of the P$_{T7}$. This low level of leakiness was sufficient to express wild-type β-lactamase, but insufficient to generate mutations and revert inactivated β-lactamase.

[xxxii] Refer to Figure 3.3 and 4.13, where J23101TetO was shown to have very low levels of leaky activity

Nucleotide substitutions resulting from the different mutator modules

This spotting test showed that for AT(C) → AT(G) and (C)TG → (A)TG codon changes, MUT-4 with the NAPE—PolIV$^{\Delta12}$ EP-DNA-repair complex displayed the ability to perform these nucleotide substitutions. The latter mutation was also performed by MUT-11, comprising the Exo-III—PolIV^{WT} 2-protein fusion (Figure 4.20). A(C)G → A(T)G substitution was performed by the mutator expressing 5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$. Overall, none of the 11 mutator modules tested seemed to perform all the four nucleotide substitutions that were investigated in a sample size of 2.5 million plated cells.

The crucial finding in this experiment was that five out of six mutator modules utilising Exo-III as the AP-endonuclease displayed the ability to perform A → T transversions, which required patch repair. The meant that these mutator modules excised nucleotides flanking a U:G lesion via the 5'-3' exonuclease or using the 3'-5' exonuclease activity of Exo-III, generating a larger gap for DNA Pol-IV$^{\Delta12}$ to read into the premature stop codon and mutate A → T to produce functional β-lactamase (Figure 4.20). The 2-protein EP-DNA-repair complexes (MUT-1 and MUT-11) lacking a 5'-3' exonuclease domain were also able to perform patch repair. This finding may have resulted from Exo-III being a bi-functional protein and possessing 3'-5' exonuclease activity along with being an AP-endonuclease[216]. Exo-III possibly makes incisions 5' to a damaged nucleotide and then extends the resulting single-strand break to generate a gap using its nonspecific 3'-5' exonuclease activity[217]. MUT-7 (5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$) and MUT-9 (5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$) displayed the most colonies resulting from patch repair + A → T transversion. This suggests 5'-3'-Pol-I exonuclease possesses a strong affinity for excising nucleotides downstream of the gap generated at the AP-site by Exo-III, creating more ssDNA template for Pol-IV$^{\Delta12}$ to copy.

None of the mutators using NAPE as the AP-endonuclease displayed the A → T transversion, meaning RecJ, RecE or 5'-3'Pol-I-Exo domains were not able to excise a stretch of nucleotides for patch repair in these EP-DNA-repair complexes. This may result from improper folding of the exonuclease domains in the 3-protein fusion. Of the mutator modules using NAPE, only MUT-4 displayed the ability to perform nucleotide substitutions, based on this sample size of 2.5 million cells.

Overall, individual mutator modules seemed to lack diversity in the mutation spectrum they generated, with no single mutator module performing all four nucleotide substitutions being investigated. The mutator modules using Exo-III (MUT-1-MUT-3, MUT-7, MUT-9 and MUT-11) as the AP-endonuclease were the most successful, displaying 3 out of 4 mutations being investigated. As a result, these Exo-III mutator modules were shortlisted for further testing by plating larger culture volumes. In the spotting experiment, only 3 µl of cells were plated for each sample. A search space of only 2.5 million bacterial cells was explored, which is possibly too low to reliably identify unique point mutations that occur at low frequencies. The search space, therefore, needed to be increased to identify nucleotide substitution events with such a low frequency of occurrence.

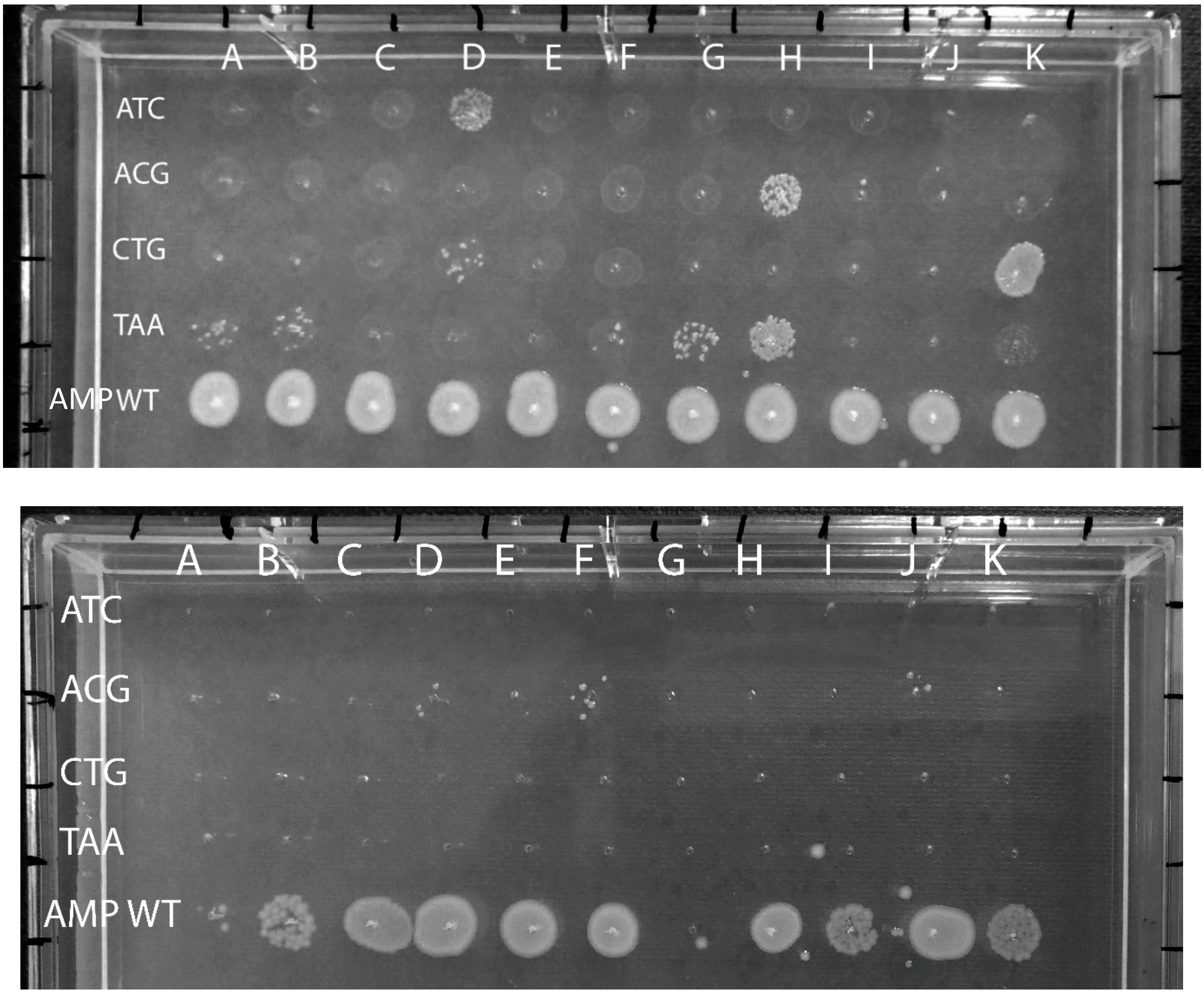


***Figure 4.20: Results of the gain-of-function experiment with MUT-1 – MUT-11. TOP –** The 11 different mutator circuits using DNA Pol-IV as the error-prone DNA polymerase were transformed into cells containing the 4-non-functional amp^R genes, creating a matrix of 44 different mutator activity screens. ATC, ACG and CTG are amp^R mutants generated by changing the start codon, ATG. TAA represents amp^R mutants with a premature stop codon in the ORF. After 24hours of induction with aTc, MUT-1, MUT-2, MUT-7 and MU-9 utilising Exo-III displayed the ability to revert TAA to TTA, making cells resistant to carbenicillin. This key result shows that the mutators using Exo-III as the AP-endonuclease are capable of excising nucleotides downstream of the AP-site using a 5'-3'Exo and filling the gap using DNA Pol-IV. **Bottom** – The cells with the mutator turned off were also plated on a carbenicillin agar plate. No carbenicillin-resistant colonies are seen on the plate, confirming the switch-like behaviour of the circuit. Without aTc, the mutator circuit is switched off and does not perform mutations on the β-*

lactamase gene. However, some leaky expression from the J23101TetO promoter possibly resulted in AID-T7pol expression, which expressed wildtype B-lactamase even with the mutator system turned off.

***Figure Legend**: A = MUT-1; B = MUT-2; C = MUT-3; D = MUT4; E = MUT-5; F = MUT-6; G = MUT-7; H = MUT-9; I = MUT-8; J = MUT-10; and K = MUT-11*

4.7.2 Performing Gain-of-Function Experiment in Larger Culture Volumes

Using larger culture volumes allowed for significantly more cells to be plated on LB-agar plates containing carbenicillin and gentamycin. Instead of 3 µl spots, 800 µl of cell cultures were spun down and plated on individual agar plates for each mutator module that was investigated. From the previous experiment, the mutator module library was reduced to the ones expressing Exo-III as the AP endonuclease, as MUT-1, MUT-2, MUT-7 and MUT-9 exhibited the ability to perform patch repair, a process which is essential for generating mutations at A:T base pairs in the target DNA.

Nucleotide Substitutions with Exo-III—Pol-IV$^{\Delta 12}$ mutators in 1 ml cell cultures

1 ml cultures of cells expressing MUT-1 – MUT-3, MUT-7 and MUT-9 were induced with aTc and after 24-hours, 800 µl of cell culture at an OD_{600} ~ 1.0 (640 million cells) were spun down and plated on carbenicillin + gentamycin agar plates. This 250-fold increase in the number of cells plated greatly expanded the β-lactamase mutant library screened for functional activity. Nucleotide substitutions occurring at frequencies as low as ~ $1.6x10^{-9}$ per cell could now be detected ($Mutation\ Frequency = \frac{Carbenicillin\ Resistant\ Cells}{Total\ Cells\ Plated}$). The experiment was repeated in four different 24-hour iterations. The average number of carbenicillin-resistant colonies resulting from each nucleotide substitution investigated is shown in Figure 4.21. The control DH5α cells only contained an expression plasmid for the inactive β-lactamase genes and no mutator module. Therefore, if no control cells grew on the carbenicillin plates, it would verify that any carbenicillin resistant colony witnessed resulted from targeted mutagenesis by the mutator module, creating functional β-lactamase.

MUT-1 and MUT-3 showed the least diversity in mutations, performing only two of the four substitutions investigated. MUT-1 resulted in 1000 events of C → T and only 7 events of C → A mutations out of 640 million cells. MUT-3 only performed C → G and C → T with a low mutation frequency as only 100 and 10 of these mutations occurred, respectively. As these mutator modules did not consistently perform patch repair in the four experimental runs, these modules were disregarded for future testing to shortlist a functionally diverse mutator module for continuous evolution.

MUT-2 was able to perform 3 of the 4 mutations, one of them being patch repair with A → T. However, less than 10 colonies were counted in each case, indicating that MUT-2 has a low targeted mutation frequency. Due to the low mutation frequency displayed, MUT-2 was also disregarded for downstream testing.

MUT-7 (5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta 12}$) and MUT-9 (5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$) were able to perform each of the four nucleotide substitutions being investigated. Based on the data in Figure 4.21, MUT-7 was significantly better at generating C → T and C → A mutations, with 100-fold greater carbenicillin^{+} colonies than MUT-9. Conversely, MUT-9 performed patch repair and A → T 10-fold more frequently than MUT-7. It also was 100-fold more likely to perform C → G mutations. Overall, both mutator modules reliably performed each of the four nucleotide substitutions with evidence of patch repair. Therefore, MUT-7 and MUT-9 were shortlisted for further testing and optimisation to be used in long time-course mutagenic assays.

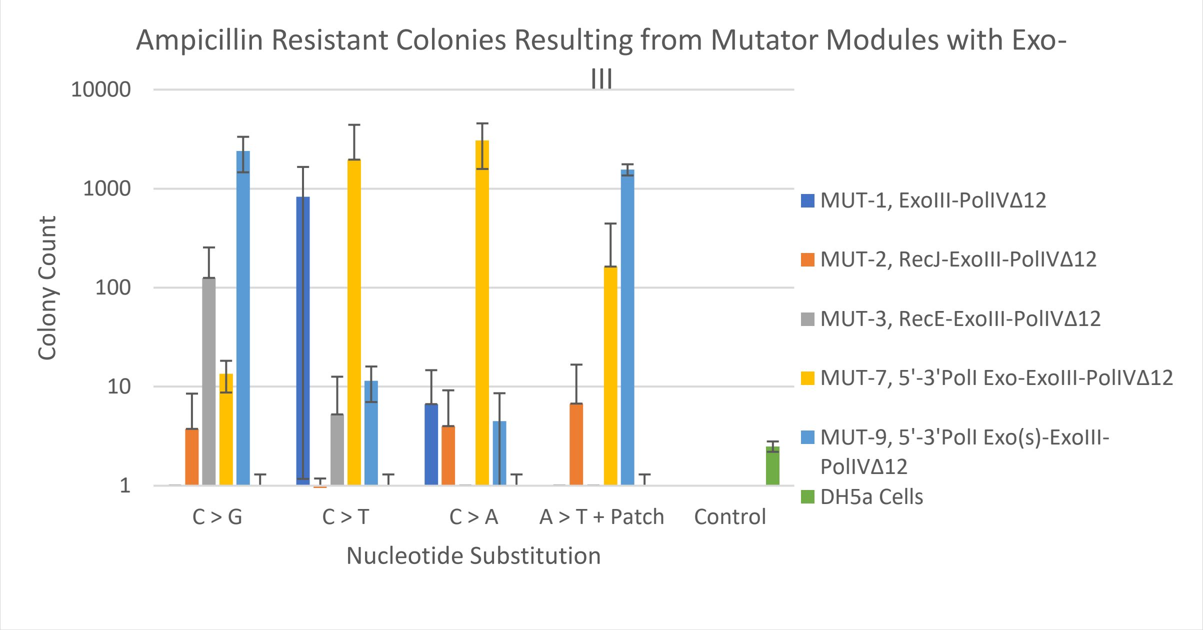


***Figure 4.21: Results of performing large volume gain-of-function assays with mutator modules comprising 5'-3'Exo—Exo-III—Pol-IV$^{\Delta 12}$**. When screening a larger volume of cells for functional β-lactamase phenotype, MUT-7 and MUT-9 were shown to consistently generate all four nucleotide substitutions being investigated. MUT-7 and MUT-9 utilising versions of the 5'-3'Pol-I exonuclease were the only mutator modules to consistently perform patch repair to convert TAA premature stop codon back to TTA, encoding leucine. Control cells lacking carbenicillin resistance were plated to assess*

the occurrence of false positives. The experiment was repeated three times and the average number of colonies counted from the three instances is plotted on the graph.

4.7.3 Gain-of-function experiments with mutator modules comprising EP-Pol-I domains

Performing the Amp^R selection assay with the EP-Pol-I mutator modules revealed a strong bias for C → T mutations. MUT-12-MUT-19 generated 50-400 C → T transition events for 8x10^8 cells plated on carbenicillin-containing plates after 24-hours of induction with aTc. This means there was no difference between using the wildtype Pol-I polymerase domain (MUT-12, MUT-16) and using EP-Pol-I^{1100} (MUT-15, MUT-19), designed to be 1100-fold more mutagenic. No C → G or C → A transversions occurred and less than 10 colonies resulted from an A → T transversion. This meant the 5'-3'Pol-I-Exo—Exo-III—EP-Pol-I variants of the EP-DNA-repair complex were able to perform patch repair, but the mutation frequency was significantly lower than that displayed by DNA Pol-IV based EP-DNA-repair complexes. Overall, this mutational pattern displayed by MUT-12 – MUT-19 surprisingly coincided with the mutational pattern of the DNA Pol-I variants created by Loeb et al. The Loeb library of EP-Pol-I mutants also generated no C → G and C → A transversions, while showing a strong bias for C → T transitions and A → T transversions[138,141]. This same phenomenon held true with our experiments even though only the polymerase domains of their error-prone DNA polymerase I variants were used.

Unfortunately, a mutator system that is heavily biased towards certain types of nucleotide substitutions is not ideal for continuous evolution, as it tends to generate a higher proportion of non-functional library members[218]. After each successive rounds of mutations, the GC% of the target gene would diminish due to the strong bias for C → T mutations. As a result, the EP-Pol-I versions of the mutator modules were disregarded for further testing.

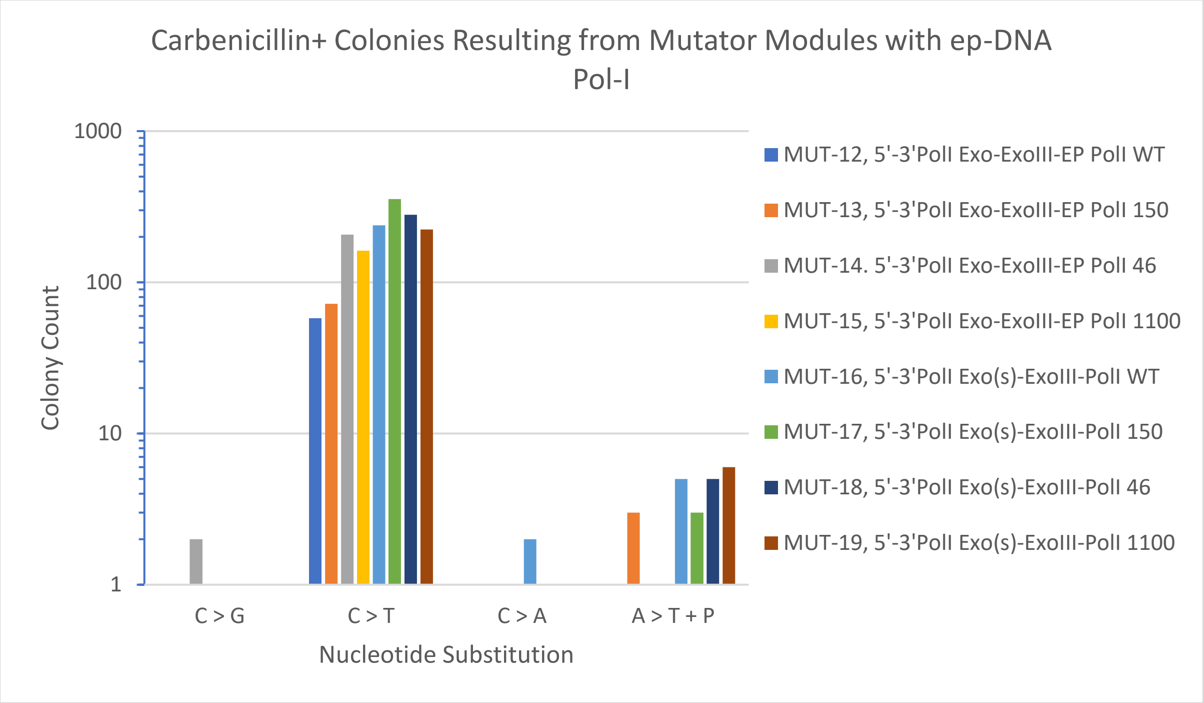


Figure 4.22: Gain of function experiments with mutator modules comprising the polymerase domains from Ep-Pol-I in the error-prone repair complex. *Mutator modules were built with Pol-I^{WT}, Pol-I^{150}, Pol-I^{46} and Pol-I^{1100}. The plasmids containing the mutator modules and the reporter plasmids with β-lactamase variants were transformed into DH5α three different times and grown overnight to start the gain-of-function experiment. The colony count from the three repeats were averaged and plotted in the graph. Subjecting the non-functional β-lactamase genes to mutations with these mutator circuits showed a strong bias for C → T mutations. Some instances of TAA→TTA were seen but this activity was 50-fold less than C → T.*

4.8 Optimisation of MUT-7 and MUT-9 Expression for Long Time-Course Mutagenic Assays

Continuous evolution experiments are designed to run for weeks, until the desired protein is achieved. This requires the mutator system to run stably inside cells for long periods of time, allowing the cells to replicate for numerous cell cycles. Consequently, numerous cycles of mutation and selection can occur, enabling identification of protein variants that only result from long evolutionary trajectories. If the expression of the mutator system is too toxic to cells, or imposes heavy burden, the growth-rate and replication of cells can become stunted[219,220]. So far, the mutator modules were only being tested in mutagenic assays that were run for 24-hours. Attempting to use the high expression versions of MUT-7 and MUT-9 (via RBS7) beyond 72-hours could result in some cell viability issues. The fitness burden imposed by the expression of complete mutator modules, where AID-T7pol and 5’-3’Pol-I-Exo—Exo-III— Pol-IV$^{\Delta12}$ are expressed simultaneously, needed to be tested by performing a cell culture growth assay and monitoring the growth rate when mutator module expression is switched on (+aTc), compared to when its switched off (-aTc). Depending on the degree to which mutator module expression impacted cell growth, the RBS characterisation toolbox[xxxiii] could be used to tune the expression of either the AID-T7pol or the EP-DNA-repair complex.

In the current experiment to monitor cell burden, the expression level of the 5’-3’Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex in MUT-7 and MUT-9 was tuned. New mutator modules were assembled identical to MUT-7 and MUT-9 but using a medium and weak RBS to initiate the translation of the EP-DNA-repair complex. In the context of the 5’-3’-Pol-I-Exo biopart, RBS7, RBS5 and RBS2 achieved high, medium and low levels of expression, respectively.

These new mutator modules and controls were assembled in pSC101-Gen^{R} in the same format as MUT-1-MUT-11, with only the RBS used to express 5’-3’Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$ being changed as follows:

MUT-20. **[AID-T7pol]** $_{\text{L1}}$ **[J23101(TetO)** $_{\text{U1-RBS2}}$ **(5’-3’Pol-I Exo—Exo-III—PolIV$^{\Delta12}$)]**

MUT-21. **[T7 pol]** $_{\text{L1}}$ **[J23101(TetO)** $_{\text{U1-RBS2}}$ **(5’-3’Pol-I Exo—Exo-III—PolIV$^{\Delta12}$)]**

MUT-22. **[AID-T7pol]** $_{\text{L1}}$ **[J23101(TetO)** $_{\text{U1-RBS5}}$ **(5’-3’Pol-I Exo(short)—Exo-III—PolIV$^{\Delta12}$)]**

MUT-23. **[T7 pol]** $_{\text{L1}}$ **[J23101(TetO)** $_{\text{U1-RBS5}}$ **(5’-3’Pol-I Exo(short)—Exo-III—PolIV$^{\Delta12}$)]**

MUT-24. **[AID-T7pol]** $_{\text{L1}}$ **[J23101(TetO)** $_{\text{U1-RBS5}}$ **(5’-3’Pol-I Exo—Exo-III—PolIV$^{\Delta12}$)]**

MUT-25. **[AID-T7pol]** $_{\text{L1}}$ **[J23101(TetO)** $_{\text{U1-RBS2}}$ **(5’-3’Pol-I Exo(short)—Exo-III—PolIV$^{\Delta12}$)]**

[xxxiii] Characterisation data presented in Sections 3.3 and 4.3

4.8.1 Testing Burden Imposed by Mutator Modules on *E. coli* Cells

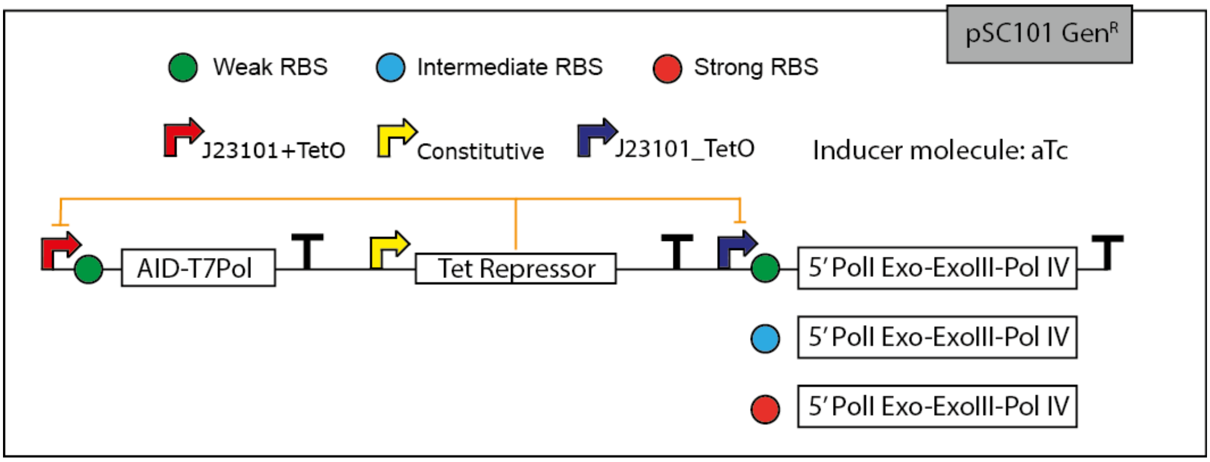


Figure 4.23: SBOL schematic of the mutator modules with 5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$ as the EP-DNA-repair complex*. The EP-DNA-repair complex was assembled into an expression cassette with a strong (RBS7), medium (RBS5) and weak (RBS2) RBS sequence identified from Figures 4.4 and 4.7.*

Methodology:

MUT-7, MUT-9, MUT-20, MUT-22, MUT-24 and MUT-25 were transformed in cells containing the 4 different inactive versions of the β-lactamase gene. The growth rate of these cells was monitored as 100 µl cultures in 96-well plates over time using a plate reader with aTc in the media. To test the fitness burden with the mutator modules in the OFF state, growth rate was also monitored with no aTc in the media.

Results: Burden When Expressing 5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$ at Different Expression Levels

High expression levels of this EP-DNA-repair complex were achieved via RBS7, while AID-T7pol was being expressed via Deg-RBS-1[xxxiv]. The circuit in the ON state exerted a clear fitness burden on the host cells. The optical density of cell cultures at stationary phase (OD_{600}) was ~ 0.3 (Figure 4.24). This is significantly lower than the natural fitness of the host *E. coli* cells, displayed by the mutator module in the OFF state. Cell cultures achieved stationary phase at an OD_{600} ~ 0.7 in this instance. Such significant fitness burden means MUT-7 would not be ideal for a continuous evolution experiment designed to run for long periods of time.

MUT-24 involved expressing the EP-DNA-repair complex via a medium-strength RBS. The cell cultures with MUT-24 expression in the ON and OFF states had comparable growth rates, with cultures

[xxxiv] Refer to Section 3.4, AID-T7pol characterisation data

achieving stationary phase in the OD_{600} range of 0.6-0.75. Cells expressing the MUT-24 with the ACG and CTG β-lactamase expression plasmids grew at a rate greater than or equal to the cells with the mutator switched off. Cells expressing ATC and TAA inactivated β-lactamase grew at a similar rate to the control during the initial periods of log phase period but achieved stationary phase at a lower absorbance value.

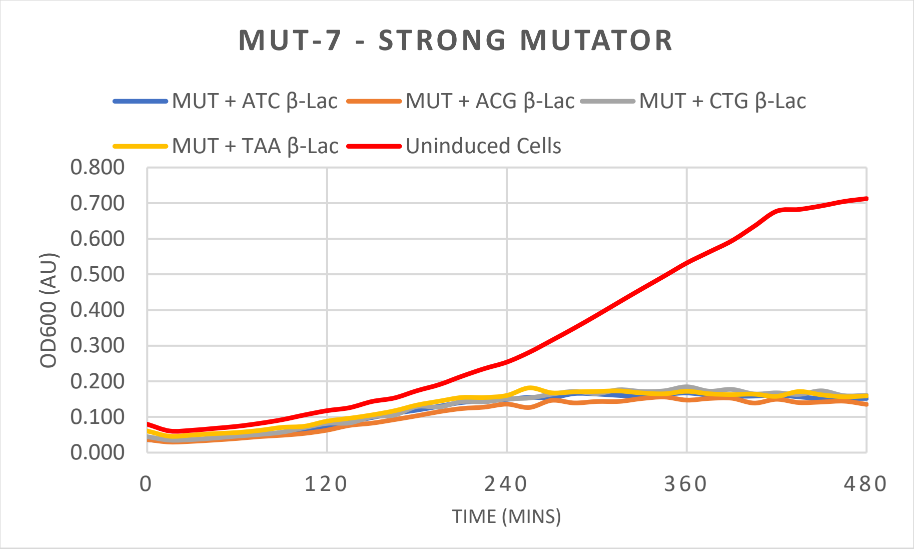
In MUT-20, the EP-DNA-repair complex was expressed via a weak RBS. In this instance, the cells with the mutator module expression being switched on or off grew at identical rates. All the experimental samples achieved stationary phase at an OD_{600} ~ 0.7.

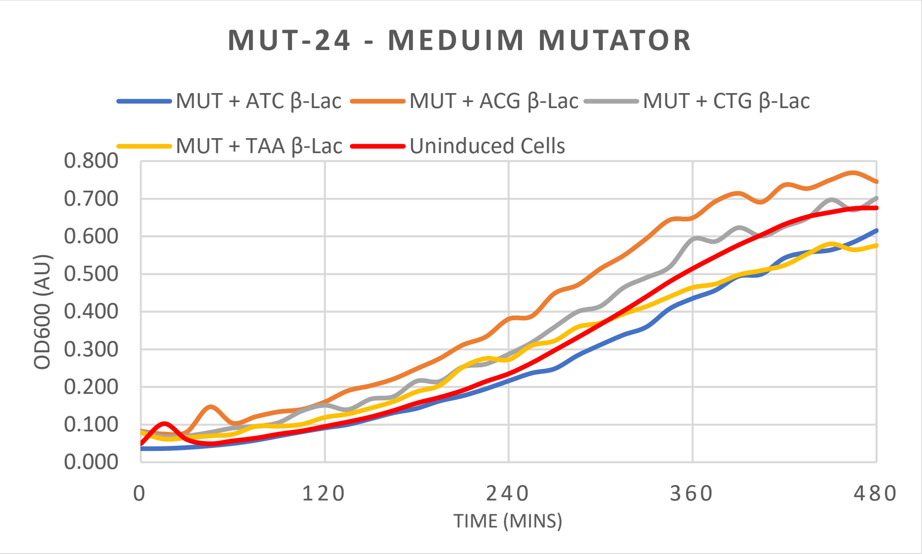
Results: Burden When Expressing 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ at Different Expression Levels

The EP-DNA-repair complexes assembled with the shorter version of the 5'-3'Pol-I-Exo domain exhibited a similar pattern of fitness burden to MUT-7, MUT-20 and MUT-24. With a high expression RBS, a clear fitness burden was imposed on cells with MUT-9 expression switched on (Figure 4.25). Cell cultures with the mutator turned off achieved stationary phase at an OD_{600} ~0.7, while an OD_{600} ~0.3 was achieved with it switched on.

MUT-22 was designed to express the EP-DNA-repair complex via a medium-strength RBS. Apart from cells expressing MUT-22 and expression plasmid of TAA-inactivated β-lactamase, all other cultures grew at comparable rates whether the mutator system was switch ON or OFF. MUT-25 is the low-expression version of this mutator circuit. Most cell cultures displaying a high growth rate with this mutator module whether it was switched on or off.

Overall, with RBS tuning, the mutator system could be expressed in *E. coli* cells without imposing a heavy fitness burden. This should enable cells to grow at stable rates for multiple cell cycles in a continuous evolution experiment, allowing for multiple cycles of library generation and selection to occur and potentially explore long mutational trajectories taken by the evolving protein. Before utilising the RBS-optimised mutator systems in such an endeavour, their ability to perform nucleotide substitutions at detectable frequencies needed to be tested. As shown in Section 4.6.1, the mutagenic activity of the mutator module correlates to the expression level of the EP-DNA-repair complex. Therefore, gain-of-function experiments needed to be performed with MUT-22 and MUT-25 to verify their ability to generate functional mutations at a reasonable mutation frequency.


MUT-7 - STRONG MUTATOR


MUT-24 - MEDUIM MUTATOR

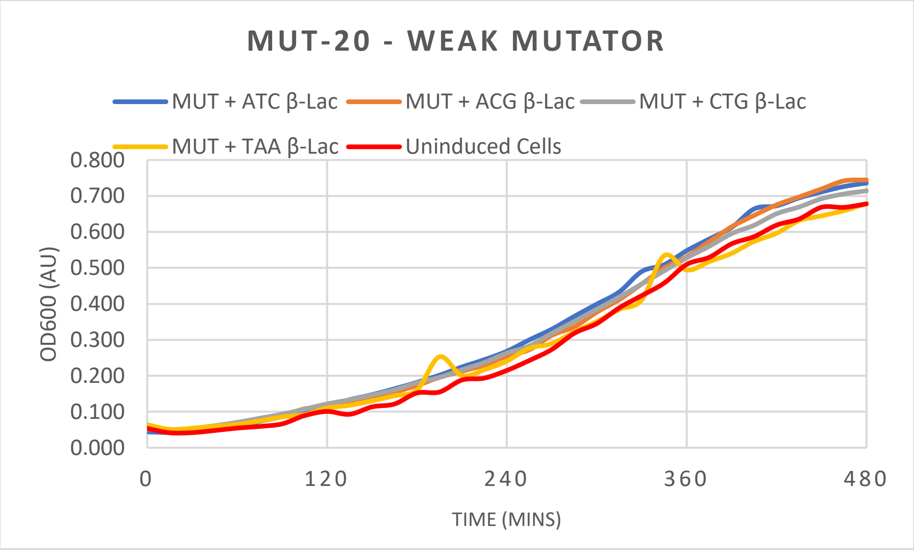
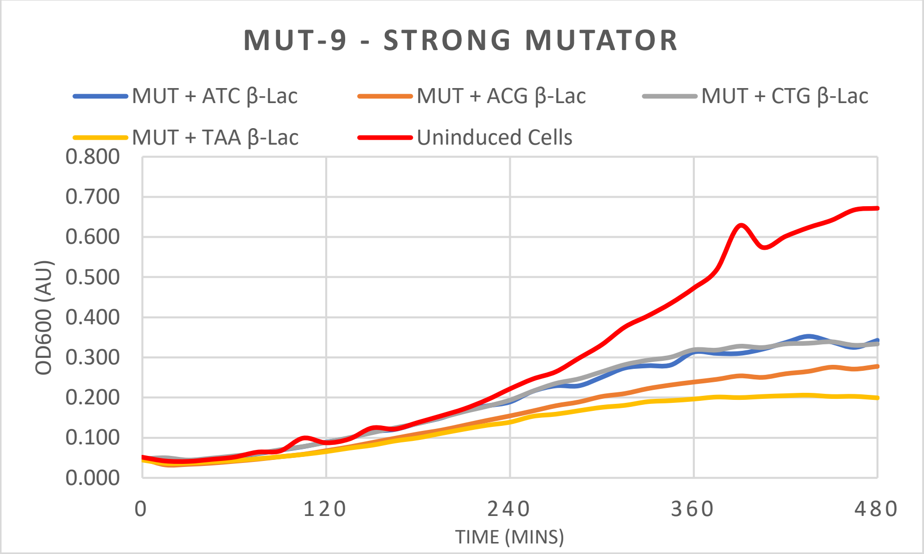
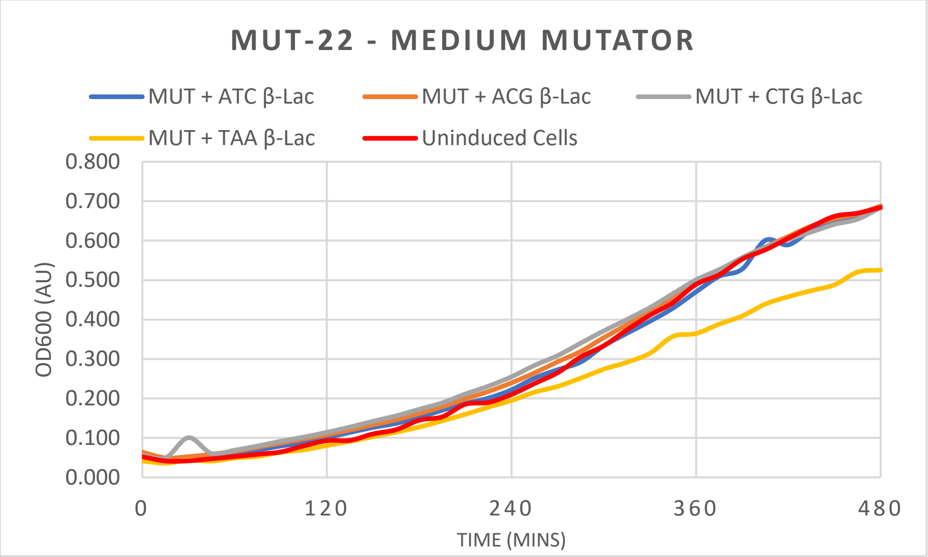

MUT-20 - WEAK MUTATOR

Figure 4.24: Burden analysis with mutator modules containing strong, medium or weak expression cassettes of 5'-3'PolI-Exo—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex*. With aTc in the growth medium to activate the expression of the EP-DNA-repair complex, high levels of expression imposed a clear burden on the cells, in the case of Mut-7. Low and medium levels of expression enabled the cells to grow at comparable rates when the EP-DNA-repair complex expression is switch on or off. Biological triplicates were picked after transformation and grown overnight, before running the growth assay. The absorbance data for the triplicates was averaged at each timepoint and plotted on the curve.*





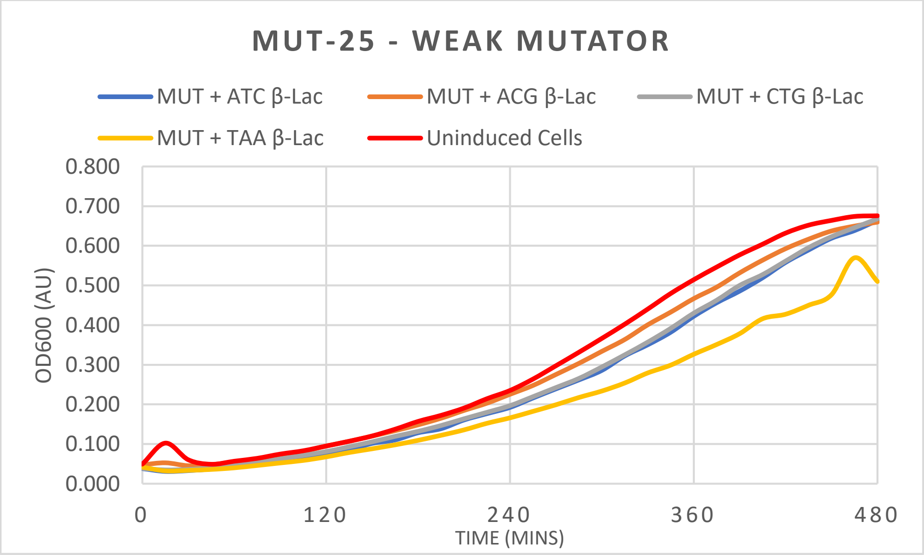


Figure 4.25: ***Burden analysis with mutator modules containing strong, medium or weak expression cassettes of 5'-3'PolI-Exo(short)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex.*** *The EP-DNA-repair complexes comprising the shorter-length 5'-3'PolI-Exonuclease domain caused minimal burden on cells when expressed in medium or low amounts in the cell. High levels of expression with RBS-7 caused burden, restricting cell cultures to an OD600 ~ 0.3. Biological triplicates were picked after transformation and grown overnight, before running the growth assay. The absorbance data for the triplicates was averaged at each timepoint and plotted on the curve.*

4.8.2 Testing MUT-22 and MUT-25 and the Stochasticity of the Mutator System

The mutagenic capability of AID-T7pol + 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$, with the EP-DNA-repair complex being expressed via a medium and weak RBS, was assessed with the gain-of-function workflow. In this workflow, MUT-22 (Medium mutator) and MUT-25 (Weak mutator) were co-transformed with the inactivated β-lactamase expression plasmids into competent DH5α cells and the cell cultures were induced with aTc for 24-hours in a deep-well plate. After 24-hrs 800 ul of the culture (~ 640 million cells) from each sample were plated on LB-agar containing carbenicillin and gentamycin to elucidate the number of functional nucleotide substitutions that occurred. Where appropriate, serial dilution was performed for a more accurate colony count. This workflow was repeated on six occasions and each time, MUT-22 and MUT-25 displayed a different frequency of generating carbenicillin-resistant cells. This finding provided insights into the stochastic behaviour of the mutator modules for generating targeted mutations (Figure 4.26).

C → G Transversion

This transversion event showed the greatest degree of variation throughout the six experimental runs with MUT-25. Colony counts were in the range of 0-10,000 per 10^8 cells (Figure 4.26). With MUT-22, this transversion event generated only 5 carbenicillin-resistant cells on four experimental runs, while 1000 colonies were generated in the other two.

C → T Transition

This transition event displayed the second-most degree of variation for MUT-25. In two experimental runs, less than 10 mutation events generated a functional β-lactamase gene; 200 Amp$^+$ colonies were achieved in another two experimental runs; and ~1000 Amp$^+$ colonies in the remaining two runs. MUT-22 again showed the least degree of stochasticity, producing 150-500 instances of functional C → T mutations that generated carbenicillin-resistant cells.

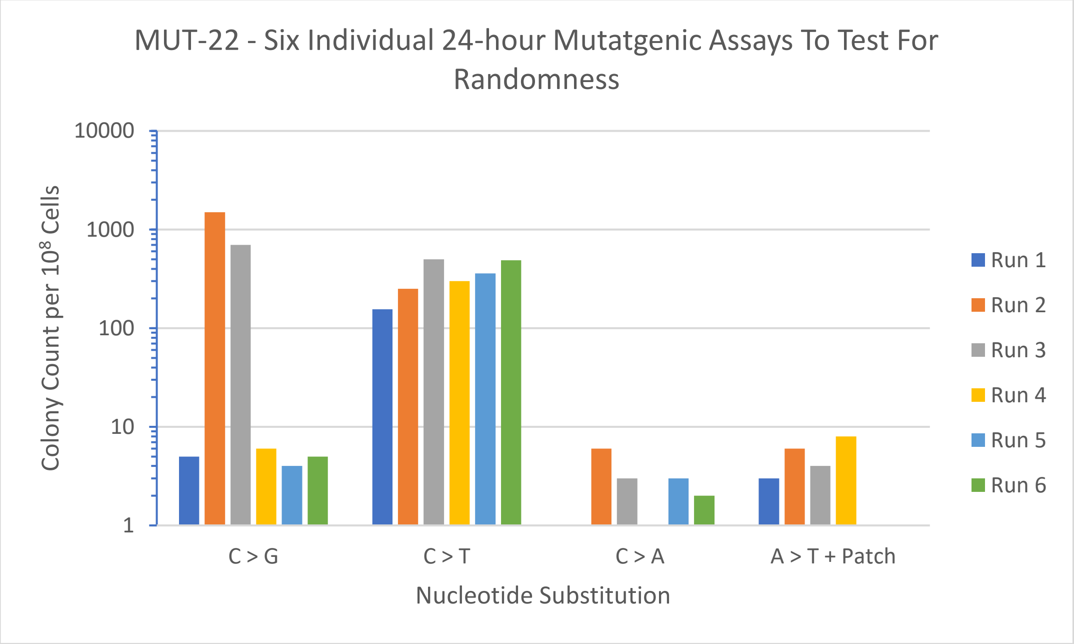
C → A Transversion

The frequency of occurrence for this transversion event was 100-fold lower than C → T and C → G mutations, for both MUT-22 and MUT-25. There was also little variation in the Amp$^+$ colonies seen across the six experimental runs with a range of 0-18 colonies being spotted per 10^8 cells.

A → T Transversion + Patch Repair

The frequency of occurrence for this mutation was the lowest, with 0-6 Amp$^+$ colonies spotted for every 640 million cells plated. Contrastingly, in previous experiments with the strong expression version of the mutator module MUT-9, up to 1500 ampicillin^{+} colonies per 10^{8} plated cells were witnessed after a 24-hour experiment (Figure 4.21).

Overall, performing six iterations of the gain-of-function workflow resulted in a wide variance in the number of functional nucleotide substitutions being generated by the weak and medium mutators, MUT-25 and MUT-22, respectively. This provides an indication that the activity of the mutator module is random. The probability of generating specific mutations at specific nucleotide positions within the target gene ORF possibly depends on stochastic factors, like the mutation bias of DNA Pol-IV, and the affinity for the EP-DNA-repair complex to interact with DNA at the AP-sites.



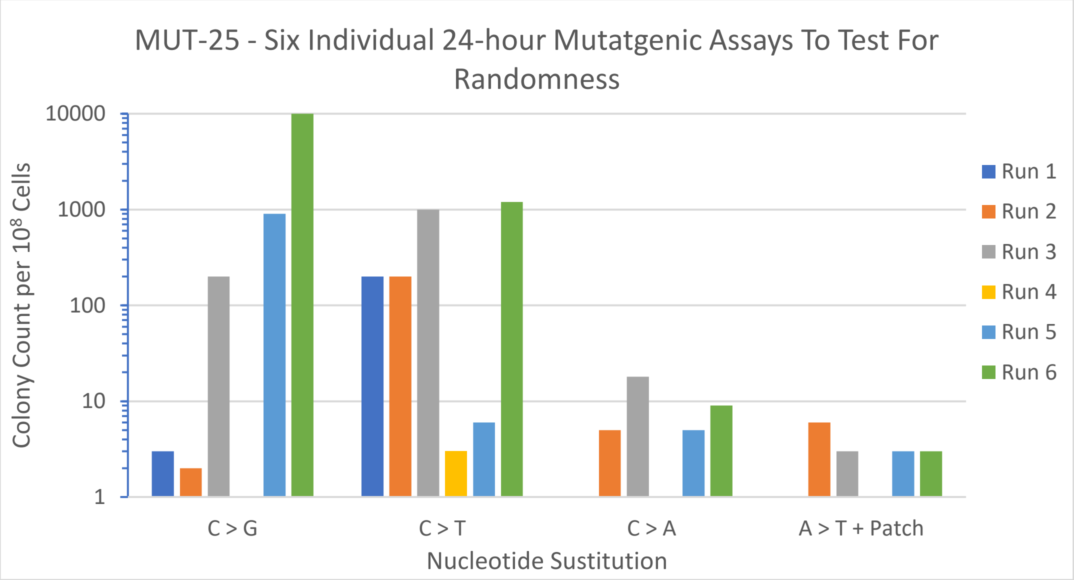


Figure 4.26: Testing the randomness of mutagenic activity with inactivated β-lactamase. *The cells containing inactivated B-lactamase genes were transformed with MUT-22 and MUT-25 on six different occasions and the number of carbenicillin-resistant cells were counted after 24-hours. MUT-25 displayed the greatest degree of variability in its mutagenic activity between iterations. Up to a 1000-fold difference was seen in MUT-25's ability to generate C → G and C → T substitutions. MUT-22 was generated mutations at a consistent rate across different iterations.*

Mutational Stochasticity Results from the Activity of the EP-DNA-repair Complex

Running the gain-of-function experiment with AID-T7pol+UGI showed a different trend to the reversion experiments with MUT-22 and MUT-25. Throughout the 6 experimental runs, ~1000 Amp$^+$ colonies per 10^8 cells were generated where an A(C)G → A(T)G mutation resulted in functional β-lactamase (Figure 4.27). This illustrates that AID-T7pol generates targeted U:G lesions at a relatively constant rate. Therefore, the stochasticity of the mutator system likely does not result from AID-T7pol activity.

As there is no UGI being expressed in MUT-22 and MUT-25, the C → T mutations witnessed here result from the activity of the EP-DNA-repair complex. MUT-22 emulated the trend produced by AID-T7pol+UGI but generated fewer Amp$^+$ cells. MUT-22 expressing 5'-3'PolI-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ generated 50-75% lower colony counts than AID-T7pol+UGI from C → T transitions. This finding possibly results from the fact that with the EP-DNA-repair complex, generating mutations is dependent on the error-rate of DNA Pol-IV$^{\Delta12}$ (upto 10^{-2} base^{-1})[221]. This is likely the cause of stochasticity in the mutation process as not every deoxycytidine deaminated by AID-T7pol would be mutated by the EP-DNA-repair complex. The stochasticity could also result from the EP-DNA-repair complex having to complete with the natively expressed DNA Pol-I to fill-in the gap generated at U:G lesions.

With MUT-25, a wider range of 3-1200 Amp$^+$ cells generated from C → T transitions were seen. If AID-T7pol consistently generates ~1000 C → U deamination events at the ACG mutant start codon, DNA Pol-IV$^{\Delta12}$ needs to incorporate dT in all 1000 instances to create the functional phenotype every time. Due to the high error-rate of DNA Pol-IV$^{\Delta12}$, it can incorporate nucleotides other than dT at the DNA site, resulting in the gene remaining non-functional. Therefore, less than 1000 Amp$^+$ colonies can be seen in some of the experimental runs.

The randomness in the rate of occurrence of each mutation can also possibly result from 'jackpot' mutations. In some instances, the mutation creating functional β-lactamase might be generated early in the cell population (during lag or early log phase). Cells possessing these jackpot mutations can divide numerous times until the cell culture reaches stationary phase. Consequently, a relatively large proportion of the population will possess the amp^R phenotype. If the functional mutation occurs late in the cell growth cycle, a smaller proportion of the population will possess the phenotype. Even though AID-T7pol deaminates the target protein at a relatively fixed rate, the factors discussed here most like result in the great variation witnessed for the number of mutations resulting in functional phenotype being witnessed at the end of the 24-hour mutagenic assay. The next step was to perform gain-of-function workflows for longer than 24-hours and with or without selection pressure to see how these factors affect the stochastic behaviour of the MUT-22 and MUT-25 mutator modules.

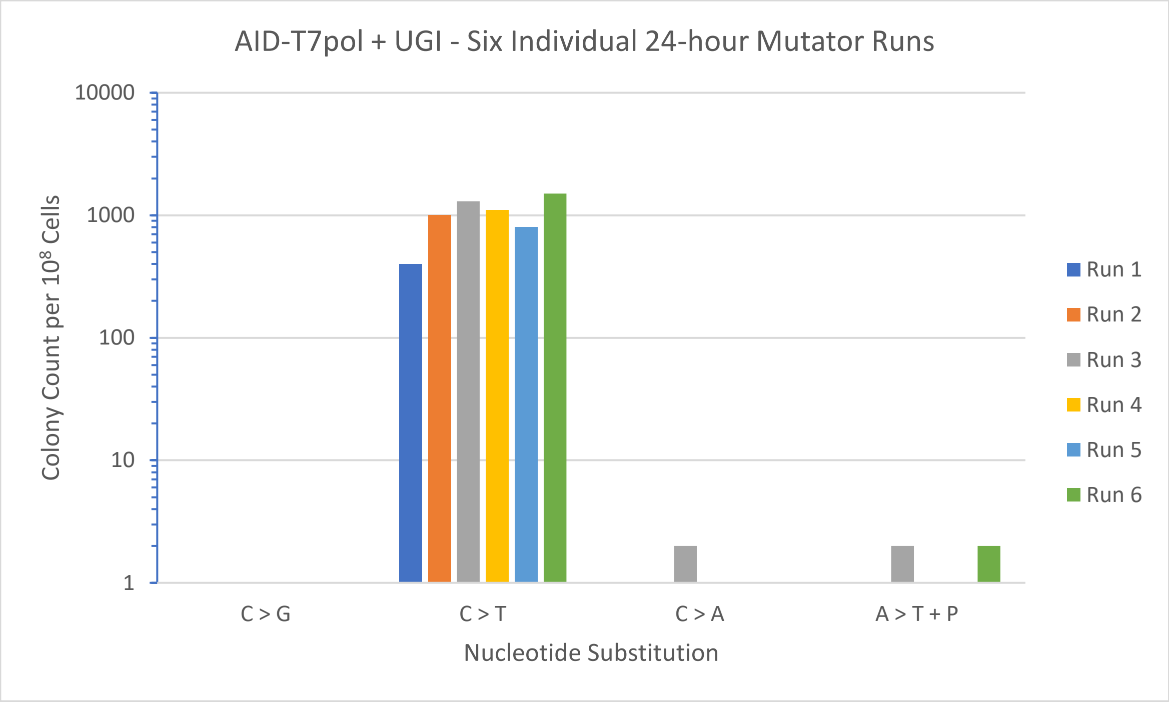


Figure 4.27: Gain of function experiments with AID-T7pol + UGI shows fixed rate of generating C → T mutations. *On six different iterations, ~1000 carbenicillin-resistant colonies were generated by the inactive ACG start codon being reverted to ATG. This implies that AID-T7pol deaminates deoxycytidine in the target gene at a relatively fixed rate.*

4.9 Objective 5: Long Time Course Mutator Assays and the Effect of Applying a Selection Pressure on the Frequency of Functional Mutations

Extensive characterisation and optimisation of the mutator modules performed in this chapter resulted in the assembly of MUT-22 and MUT-25, mutator modules that can remain active in the cell while imposing minimal burden on the cellular host. The next step was to investigate if these optimised mutator modules can generate more functional mutations at a population level, the longer the mutator system is left active. This should hypothetically result in more Amp$^+$ cells being generated as the gain-of-function assay is ran for longer than 24-hours. The ability to accumulate functional mutations within a cell population was investigated with and without the application of a selection pressure on the cell cultures expressing the mutator module. Without selection pressure, the stochastic behaviour of the mutator system should still exist and both functional and non-functional mutations would be incorporated into the β-lactamase ORF[222]. Conversely, selection pressure is applied by adding low concentrations of carbenicillin in the growth media, while the mutator system is active. This artificial environmental stress should force cells to adopt any evolutionary trajectory that results in functional β-lactamase to provide cells with a growth advantage[222,223]. In this section we investigate which system better enables the accumulation of functional mutations in the inactivated β-lactamase target gene.

4.9.1 96-hour Mutator Assay without Selection Pressure

Methodology:

DH5α cells were co-transformed with the inactive β-lactamase target plasmids and the MUT-25 expression plasmid. The overnight cultures of the transformed cells were resuspended in 1 ml of LB media with aTc to a starting OD_{600} of 0.05. Every 24-hours, the cells were transferred into fresh media to a starting OD_{600} of 0.05 (~ 40 million cells transferred to keep the mutagenic assay running). In this fashion, the MUT-25 mutator module was left active in the host cells for 96-hours. The cells were plated on LB-agar containing carbenicillin (50 $\mu g ml^{-1}$) every 24-hours throughout the experiment to monitor the change in carbenicillin-resistant cells over time.

Results:

In the absence of selection pressure, there is no visible accumulation of functional mutations overtime (Figure 4.28). The frequency of occurrence of functional mutations seems to be offset by the frequency of occurrence of non-functional mutations, resulting in a consistent number of Amp$^+$ colonies being seen for three of the four nucleotide substitutions being investigated. By the end of the 96-hours, MUT-25 had generated ~1000 Amp$^+$ colonies resulting from a C → T mutation in the start codon. Only 0 - 10 colonies possessing functional β-lactamase genes resulted from C → A, C → G and A → T mutations every 24-hours. From this data, we could draw the conclusion that in the absence of selection pressure, there is no increase in the number of functional mutants over time.

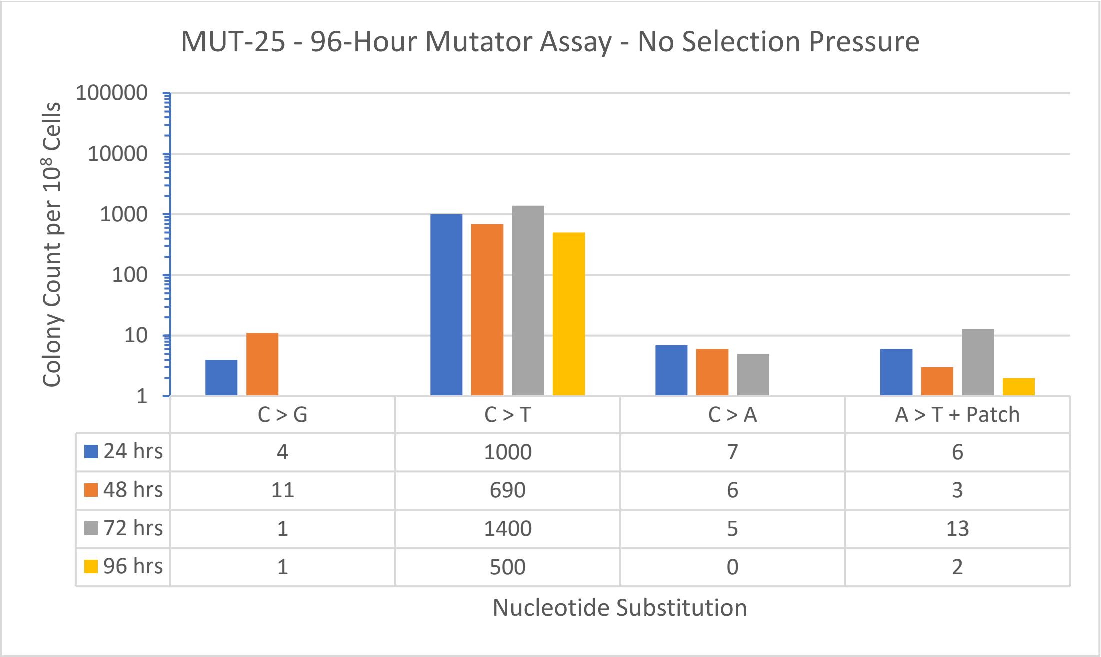


Figure 4.28: Carbenicillin-resistant colonies after a 96-hour mutator assay with MUT-25*. Cells were spread on LB-agar plates containing 50 µgml-1 of carbenicillin every 24-hrs for 96-hrs. In the absence of any selection pressure, the number of nucleotide substitutions of each type remained consistent. The functional phenotype was not enriched in the cultures over time. The frequency of C → T mutations was 100-fold higher than the other nucleotide substitutions being investigated.*

4.9.2 96-hour Mutator Assay with Selection Pressure

Step 1: Finding the ideal concentration of carbenicillin to apply as selection pressure:

Applying a selection pressure meant adding carbenicillin to the growth medium from the start of the mutator assay when the entire cell population is expressing inactivated versions of the β-lactamase gene. An appropriate concentration of carbenicillin needed to be identified that showed some degree of toxicity to cells but allowed them to grow and divide at slower growth rates.

To identify the ideal selection pressure concentration of carbenicillin, the growth of DH5α cells lacking resistance to the antibiotic was monitored in LB containing a range of carbenicillin concentrations (0.5 µgml^{-1} – 100 µgml^{-1}). The cells were grown in 100 µl cultures in a 96-well plate and the growth rate of the cell cultures was monitored in a plate-reader (Figure 4.29). From this growth assay, a concentration range of 2.5 - 5 µgml^{-1} was identified as ideal for allowing cells to grow and replicate, but the burden imposed by environmental stresses meant stationary phase was achieved at a lower OD_{600} ~0.6.

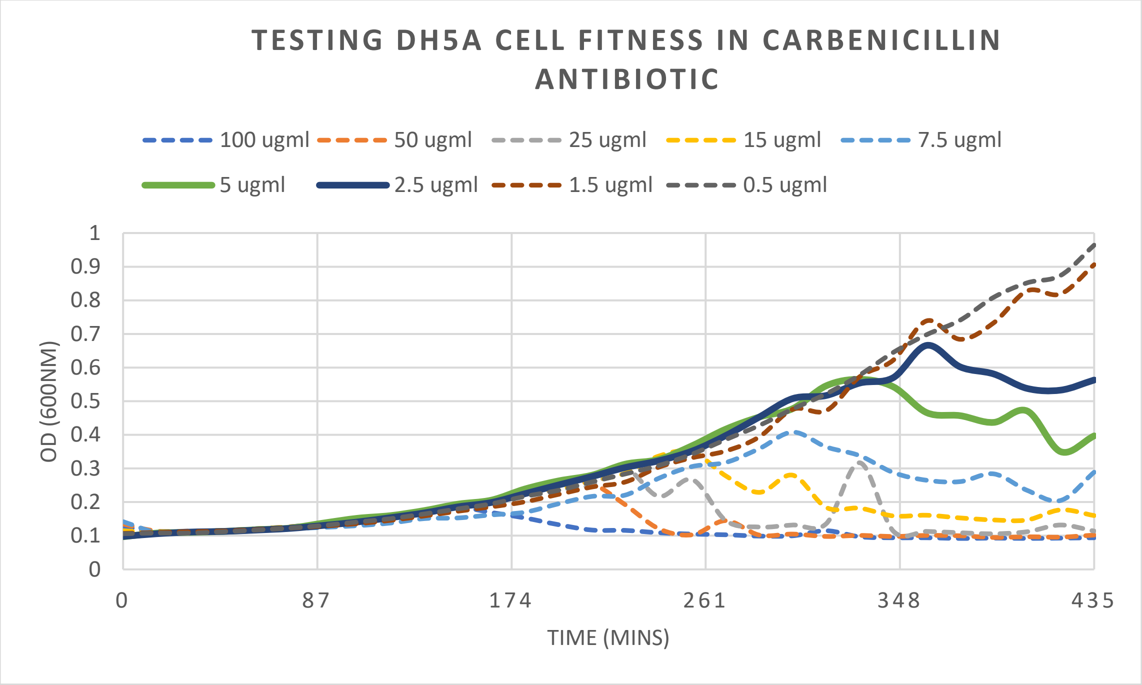
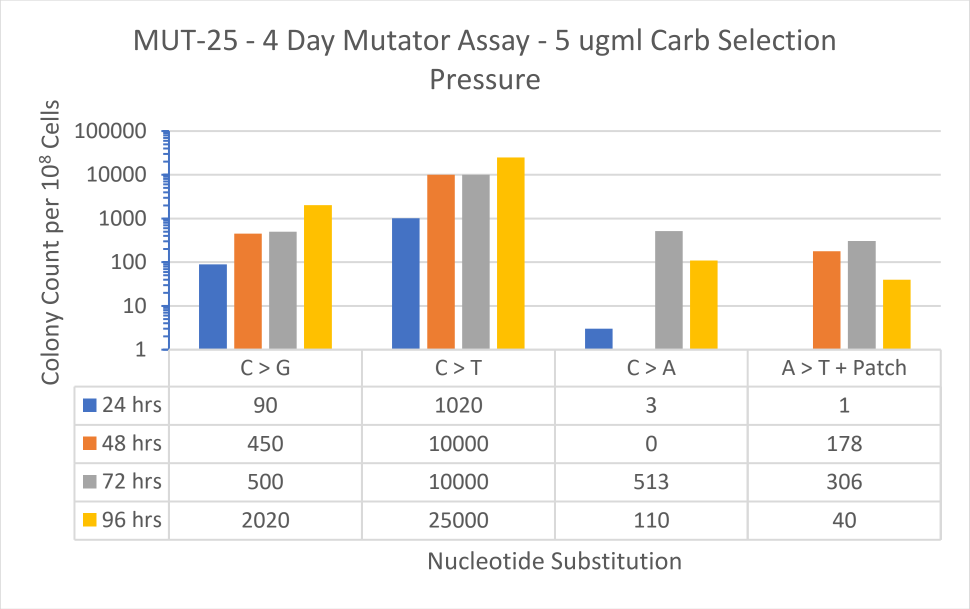


***Figure 4.29: Finding the ideal concentration of carbenicillin to use as selection pressure in mutator assays**. Native DH5a cells lacking resistance to carbenicillin were grown in liquid cultures containing a range of concentrations of the antibiotic. The aim was to identify a concentration where cells grew, but with a clear stress on the growth rate. 2.5 and 5.0 ugml^{-1} were identified as appropriate concentrations of carbenicillin to use as selection pressure in a mutator assay with inactive β-lactamase.*

Effect of Applying a Selection Pressure in the Gain of Function Experiment

The DH5α cells containing the inactive β-lactamase expression plasmids and MUT-25 were inoculated in LB media containing aTc and 2.5 μgml^{-1} of carbenicillin. Every 24-hours, the cells were diluted in 1 ml of fresh LB media in a deep-well plate to keep the mutator assay active. The experiment was performed for 96-hours and cells were plated on LB-agar containing carbenicillin, every 24-hours, to check for carbenicillin-resistant cells (Figure 4.30).

At the end of the 96-hour period with a selection pressure being applied, we witnessed a 10- to 100-fold increase in the number of Amp+ cells resulting from all four nucleotide substitutions being investigated, compared to Figure 4.29, where no selection pressure was applied. Functional mutations progressively accumulated overtime for the C → G and C → T substitutions, resulting in more Amp^{+} cells every 24-hours. In these cases, the lowest number of Amp^{+} cells were counted at the 24-hour mark and for each subsequent time-point, a higher number of cells expressing functional β-lactamase was recorded. With C → A and A → T, the number of amp^{+} colonies increased until 72-hours, but after another 24-hours, the functional phenotype was lost by a percentage of the population. Overall, this experiment highlights the impact of applying a selection pressure in continuous evolution experiments, which helps to drive the mutations being incorporated in the target gene towards a functional phenotype via evolutionary trajectories that provide a fitness advantage to the cell[222].



	C > G	C > T	C > A	A > T + Patch
🟦 24 hrs	90	1020	3	1
🟧 48 hrs	450	10000	0	178
⬜ 72 hrs	500	10000	513	306
🟨 96 hrs	2020	25000	110	40

Figure 4.30: The effect of selection pressure on the enrichment of functional β-lactamase mutants*.* *Applying 5 ugml-1 carbenicillin as selection pressure resulted in the enrichment of the functional phenotype over time. The number of carb-resistant cells, resulting from a C → G or C→ T mutation in the B-lactamase start codon, increased overtime for the 96-hrs. Cells with C → A and patch repair + A → T mutations saw enrichment until 72-hrs, followed by a decline at 96-hrs.*

4.10 Objective 6: Unique Control Modules to Validate the Mutagenic Properties of the Error-prone DNA Repair Complex

The mutator module consists of three expression cassettes placed within a pSC101-gentamycin^{R} backbone. Both AID-T7pol and 5'-3'Exo—AP-Endo—EP-DNA-Polymerase are expressed using individual J23101$_{\text{TetO}}$ promoters. The system also contains a constitutive Tet repressor (TetR) expression cassette for regulating the expression of the DNA damage device and the EP-DNA-repair complex with a switch-like mechanism.

To validate that any targeted mutagenic activity found on GFP-mut3b or inactive β-lactamase resulted from this mutator module design and not due to other biological factors, certain control modules were assembled (Table 4.3). For assembling the control modules, the TetR cassette and the pSC101-Gen^{R} backbones were kept constant while different combinations of the DNA damaging device and error-prone DNA repair complexes were assembled. 14 different control modules expressing different combinations of AID, AID-T7pol, T7-RNA-Pol, UGI, and the Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex were created. Based on the biological design of the mutator module, mutations should only be generated in the target gene with protein combinations where AID, T7pol and Exo-III—Pol-IV$^{\Delta12}$ are expressed simultaneously. Such systems should possess the ability to generate U:G lesions and perform EP-DNA repair. If one of these components is missing, no mutations should hypothetically be generated in the target gene. The control modules in Table 4.3 enabled testing of this hypothesis and validating the biological mechanism of the mutator modules expressing AID-T7pol with an EP-DNA-polymerase.

Table 4.3: List of Control Modules and Their Function

ID	DNA Damage Device	Error-Prone DNA Repair Complex	Hypothesised Biological Activity
CON-1	AID	--	AID expressed freely without UGI should result in minimal targeted or global mutations in the cell. dC deamination events would be repaired using native DNA repair pathways.
CON-2	AID + UGI	--	In this instance, as BER is being blocked, C→T mutations should be seen. In the absence of targeting via T7 polymerase, the mutations should occur globally. This control should indicate a high global mutation frequency in rifampicin antibiotic selection assays.
CON-3	T7 polymerase	--	With the lack of a deaminase, there should be no mutations being introduced on the target gene or the bacterial genome. T7 RNA Pol should act as a negative control for mutations in the cell.
CON-4	AID-T7pol	--	While AID function can now be specifically targeted to the gene of interest placed downstream of P_{T7}, the cell's native BER pathway should repair the U:G lesions. The global mutation rate should also remain low with active DNA repair.
CON-5	AID-T7pol + UGI	--	C → T mutations should be spotted on the target gene. A(C)T → A(T)G mutations will result in amp^{R} resistance. The global mutation rate would increase due to UGI blocking all UNG-mediate repair of U:G mismatches in the bacterial genome.
CON-6	T7pol + UGI	--	The lack of a deaminase means there is no artificial means of inducing DNA damage. However, this control would still increase the global mutation rate in the cell due to the expression of UGI.
CON-7	UGI	--	This control can be used to assess the natural mutation rate of the cells as BER pathway is blocked for UNG-mediated DNA repair.
CON-8	AID + T7pol unfused	--	These two proteins being unfused should result in a higher global mutation rate than when they are fused together. The cell's native DNA repair pathway should repair U:G mismatches generated by AID in the bacterial genome and target gene.

CON-9	AID + T7pol + UGI	--	This control should show a higher global mutation rate than the AID-T7pol fusion protein, validating the need for fusing the two genes for targeted mutagenesis and reducing its global mutagenic activity.
CON-10	AID-T7pol + UGI	Exo-III—Pol-IV$^{\Delta12}$	This control is the complete mutator module plus the UNG blocker. UGI should blocker BER of U:G lesions so the Exo-III—Pol-IV$^{\Delta12}$ repair complex would be unable to act on the target gene due to the lack of AP-sites. This means no functional mutations should be spotted in gain of function experiments for all inactivated amp^{R} mutants, except ACG. The C → T mutation being conserved by UGI should result in expression of functional β-lactamase.
CON-11	T7pol	Exo-III—Pol-IV$^{\Delta12}$	Without AID activity to create U:G lesions, Exo-III—Pol-IV$^{\Delta12}$ should be unable to generate mutations. No functional mutations should occur in the Amp^{R} gain of function experiments. This control would validate the requirement of AID for inducing DNA damage, which is then repaired by the EP-DNA-repair complex.
CON-12		Exo-III—Pol-IV$^{\Delta12}$	Another control to validate the requirement of AID for creating the U:G mismatch before Exo-III—Pol-IV$^{\Delta12}$ can target the gene for error-prone DNA repair. The control module should have a low global mutation rate as Pol-IV$^{\Delta12}$ cannot bind to the replication fork and perform lesion bypass DNA replication.
CON-13	T7pol + UGI	Exo-III—Pol-IV$^{\Delta12}$	Expressing UGI should increase the global mutation rate, but the absence of AID means no targeted mutations should occur in the gene-of-interest.
CON-14	UGI	Exo-III—Pol-IV$^{\Delta12}$	Same validation achieved as the previous control. No means to induce damage on the target gene means no error-prone DNA repair.

4.10.1 Gain of Function Experiment with the Control Modules

The control modules contained unique combinations of AID, T7 RNA polymerase, UGI and the Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex, creating different cellular conditions to assess how the mutations are being generated in the target DNA. We assessed if functional β-lactamase could be achieved expressing only the EP-DNA-repair complex and no AID to create U:G lesions; by expressing only T7pol with the EP-DNA-repair complex; or by expressing AID-T7pol + Exo-III—Pol-IV$^{\Delta12}$ + UGI. The expected biological responses to the different combinations are shown in Table 4.4.

Results of the gain-of-function assay with the different control modules:

As expected, CON-3 and CON-4 did not produce any nucleotide substitutions in the β-lactamase ORF (Figure 4.31). The lack of UGI or an EP-DNA-repair complex meant that any deoxyuridine residues produced by AID in CON-4 were repaired by the cell's DNA repair pathways. CON-11-CON-14 also did not generate any nucleotide substitutions (Figure 4.31). These control modules were expressing the Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex, but no AID to induce DNA damage. In the absence of a DNA damaging mechanism, the BER repair pathway cannot be initiated for the Exo-III—Pol-IV$^{\Delta12}$ to hijack and perform error-prone DNA repair of the DNA.

Table 4.4: Control Modules with different combinations of DNA damaging enzymes and error-prone DNA repair complexes, and their effect on target and off-target DNA

Control Module	Enzyme to Damage DNA	Enzymes to block or Perform error-prone DNA repair	Effect on Target DNA		Effect on Off-target DNA	
			Cytidine Deamination	Mutation in DNA	Cytidine Deamination	Mutation in DNA
CON-3	T7pol	None	No	No	No	No
CON-4	AID-T7pol	None	Yes	No	Yes	No
CON-5	AID-T7pol	UGI	Yes	C → T	No	No
CON-6	T7pol	UGI	No	Low*	No	Low*
CON-7	None	UGI	No	Low*	No	Low*
CON-1	AID	None	Yes	No	Yes	No
CON-2	AID	UGI	Yes	C → T	Yes	C → T
CON-10	AID-T7pol	UGI and Exo-III—Pol-IV$^{\Delta12}$	Yes	C → T**	Yes	C → T**
CON-12	None	Exo-III—Pol-IV$^{\Delta12}$	No	No	No	No
CON-8	AID and T7pol	None	Yes	No	Yes	No
CON-9	AID and T7pol	UGI	Yes	C → T	Yes	C → T
CON-11	T7pol	Exo-III—Pol-IV$^{\Delta12}$	No	No	No	No
CON-13	T7pol	UGI and Exo-III—Pol-IV$^{\Delta12}$	No	Low*	No	Low*
CON-14	None	UGI and Exo-III—Pol-IV$^{\Delta12}$	No	Low*	No	Low*

Low* - Random mutations can be introduced at the natural mutation rate of the cell, due to UGI blocking UNG-mediated DNA repair.

C → T** - CON-10 expressed both AID-T7pol and Exo-III—Pol-IV$^{\Delta12}$, but should only be capable of introducing C → T mutations as activity of the EP-DNA-repair complex is inhibited by UGI blocking the BER pathway

With CON-10, only C → T mutations resulted in functional β-lactamase. Even with the expression of Exo-III—Pol-IV$^{\Delta12}$ in the system, other nucleotide substitutions were not generated. The expression of UGI blocks UNG activity, which means the damaged deoxyuridine base cannot be excised to generate an AP-site. The means the Exo-III—Pol-IV$^{\Delta12}$ complex cannot interact with the damaged DNA and perform error-prone DNA repair. The U:G mismatch results in a T:A mutation being incorporated at the damaged site during semiconservative DNA replication[224]. Therefore, CON-10 created A(C)G → A(T)G mutations resulting in the Amp^R phenotype.

Overall, performing the gain-of-function experiments with these controls validated the mechanism of our mutator module for introducing mutations into the target gene. The AID-T7pol DNA damaging device is needed to deaminate deoxycytidine in the target gene ORF, which initiates the BER pathway within the cell. Once UNG has generated an AP-site, the repair pathway can then be hijacked by the 5’-3’Exo—AP-Endo—EP-DNA-Polymerase EP-DNA-repair complex to nick the DNA at the AP-site and subsequent gap filling occurs via the error-prone DNA polymerase. If either AID-T7pol or the EP-DNA-repair complex are not present in the system, mutations will not be introduced in the target gene.

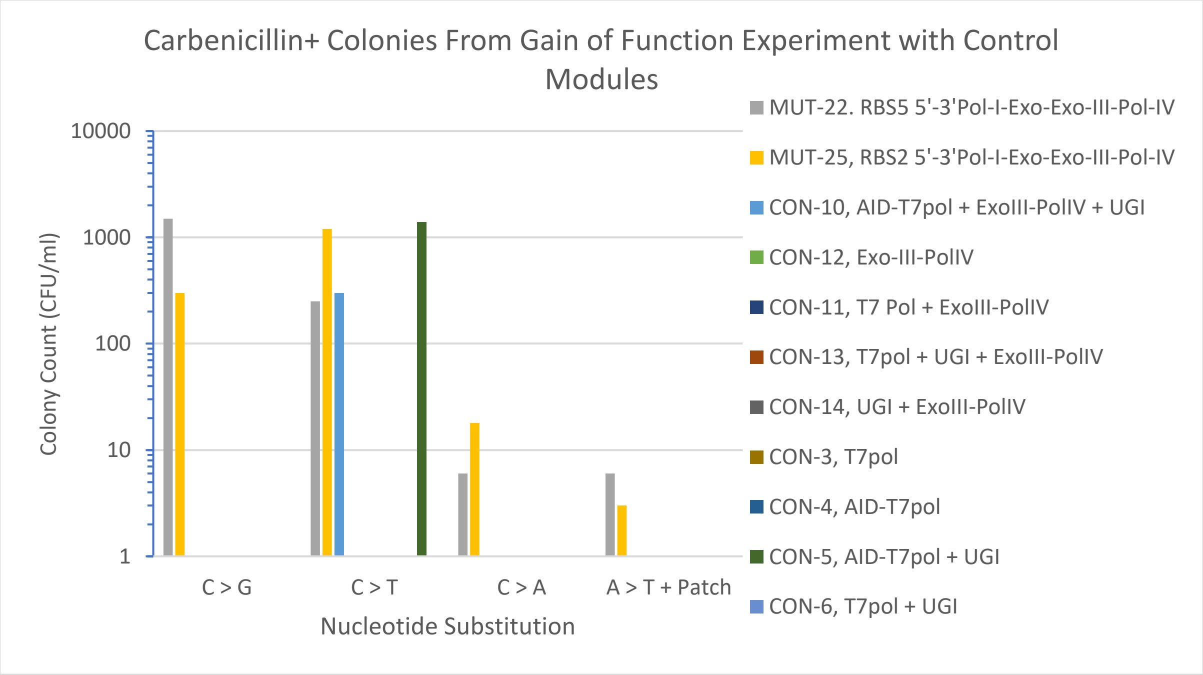


***Figure 4.31: β-lactamase gain of function assays with the control modules**. Of all the control modules tested, AID-T7pol+UGI and AID-T7pol + Exo-III—Pol-IV + UGI generated C → T mutations. The other control modules did not generate functional B-lactamase from any of the four nucleotide substitutions. All 4 mutator modules tested did perform nucleotide substitutions.*

4.11 Objective 7: Assessing Targeted vs Off-target Mutation Frequency

The key qualities expected from the mutator module are to perform targeted mutations, while maintaining fitness of the host cells and the fidelity of their genome. So far, we’ve demonstrated that the mutator modules with AID-T7pol and 5’-3’Pol-I-Exo— Exo-III—Pol-IV$^{\Delta 12}$ as the EP-repair complex can perform targeted mutations, while being minimally toxic to the host cells. Now the off-target mutation frequency of the mutator modules needed to be analysed. For assessing the difference in the rate of targeted vs off-target mutations using the mutator modules, cells expressing MUT-1, MUT-9 and eleven of the control modules mentioned in Table 4.3 were transformed into DH5α cells. These cells were gown in media containing aTc for 24-hours, to switch on the expression of the mutator and control modules. After 24-hours, 8x10^8 cells from the liquid cultures were spread on agar plates containing rifampicin to count the number of cells that became resistant to the antibiotic.

Table 4.5: Rifampicin+ selection assay with Control Modules. 850 ul of cell plated at OD ~1.2

ID	Circuit Design	Run 1	Run 2	Run 3
CON-3	**T7 Pol**	74	43	20
CON-4	**AID-T7pol**	73	37	35
CON-5	**AID-T7pol + UGI**	490	256	284
CON-6	**T7pol + UGI**	390	178	85
CON-7	**UGI**	97	147	153
CON-1	**AID**	63	9	15
CON-2	**AID + UGI**	439	251	145
MUT-1	**AID-T7pol + ExoIII-PolIV**	31	14	4
MUT-9	**AID-T7pol + 5'PolI Exo-ExoIII-PolIV**	13	10	10
CON-10	**AID-T7pol + ExoIII-PolIV + UGI**	34	151	23
CON-8	**AID + T7pol**	111	102	19
CON-9	**AID + T7pol + UGI**	360	350	310
CON-11	**T7pol + ExoIII-PolIV**	5	1	0
Control Cells	**Dh5α**	30	37	3

As demonstrated in the rifampicin selection assay performed in Section 3.6, CON-1 displayed very few rifamp$^+$ cells suggesting that deamination events caused by freely expressed AID are repaired by the cell's native BER pathway. When the BER blocker, UGI, is added to the circuit (CON-2), the number of rifamp$^+$ colonies increased 9-fold, due to U:G lesions not being repaired. Overall, freely expressed AID in bacteria is only mutagenic if the cell's DNA repair pathways are blocked.

The AID-T7pol fusion protein (CON-4) also displayed low global activity, comparable to freely expressed AID. Fusing AID to T7pol did not significantly reduce global deamination events to the levels seen in control DH5α cells. This level of global mutation may not necessarily be indicative of a lack of targeting by the AID-T7pol fusion protein but could be the result of overexpressing a deaminase in the cells. Expressing UGI with this fusion protein (CON-5) had similar global mutator activity as CON-2, due to glycosylase-mediated DNA repair being blocked for U:G lesions. T7 RNA polymerase (CON-3) expressed freely in the cell displayed slightly higher global mutation rate than the control cells and was comparable to freely expressed AID.

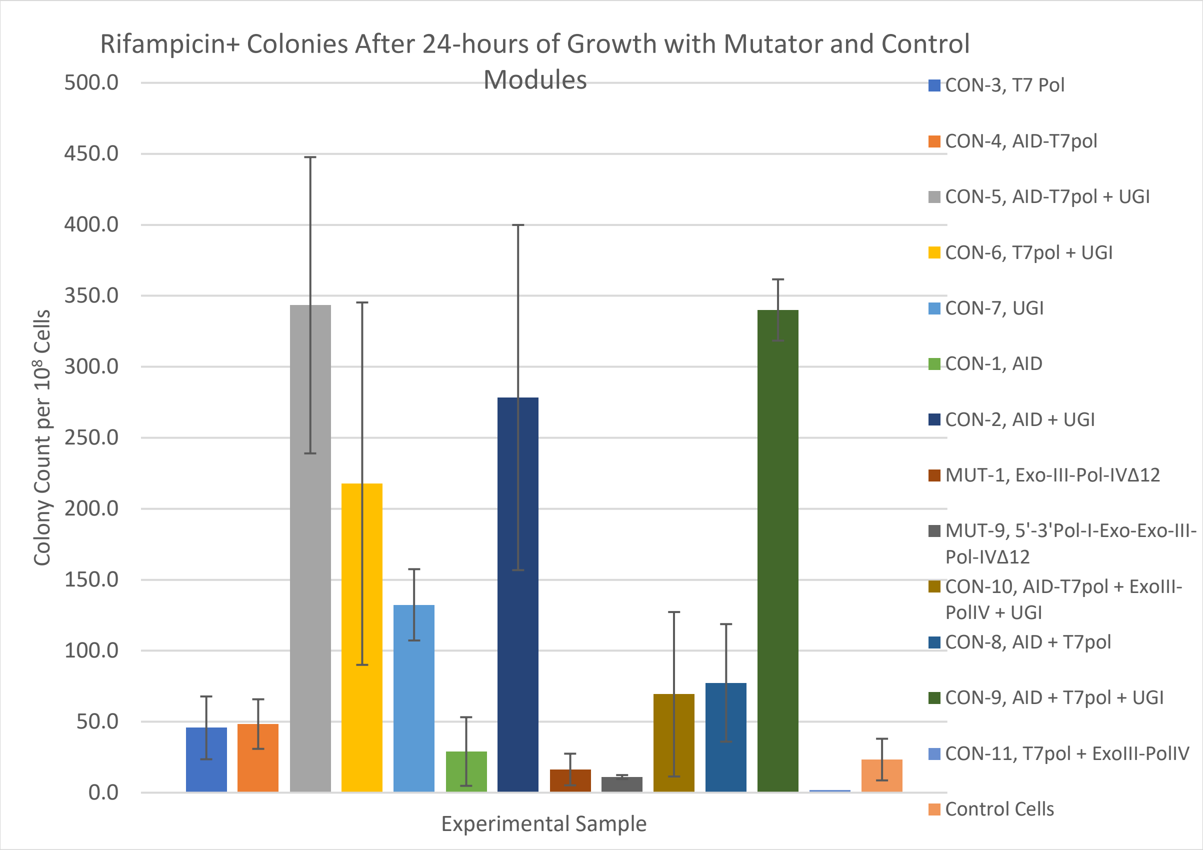


Figure 4.32: Rifampicin reversion assay with mutator and control modules to assess off-target mutagenic activity*. The control modules were transformed into DH5α cells three separate times, grown overnight for 24-hours and plated on rifampicin plates. The average colonies counted from the triplicates was plotted on the chart. Controls with UGI in the expression system resulted in 100- to 200-fold higher rifampicin^{+} colonies than the control DH5α cells. Expressing AID, T7-RNA-Pol, AID-T7pol and the mutator modules resulted in comparable amounts of rifampicin^{+} colonies to the control cells. This indicates that if the cell's native DNA repair pathways are not blocked, the proteins used in the mutator system seem to display low mutagenic activity.*

This trend of lower global mutation rate is also seen with the two mutator modules tested (MUT-1, MUT-9). Plating these cell cultures revealed similar number of rifamp^{+} colonies to the control cells. From the gain-of-function experiments with MUT-1 and MUT-9, we identified that these mutator modules generate up to ~3000 functional C → T mutations for every 10^{8} cells that are screened for the phenotype[xxxv]. Rifampicin resistance can result from a single C → T mutation creating a Ser531Leu substitution in the RpoB gene[124]. This means the rate of targeted C → T mutations was 100- to 1000-fold higher than the rate of off-target C → T introduced by the mutator modules, when comparing Amp$^+$ cells to Rifamp$^+$ cells generated after 24-hrs. A limitation of gain of function experiments is that it is difficult to calculate how many total mutations are performed on the target DNA sequence. Only the functional mutations can be selected for; information about the silent and deleterious mutations is lost. As a result, the mutator modules probably perform targeted mutations at much higher frequencies than what is displayed in carbenicillin antibiotic selection assay. The true targeted mutation frequency of MUT-9, MUT-22 and MUT-25 was calculated using NGS analysis of mutant gene libraries in Chapter 5.

Overall, the findings from the rifampicin selection assay suggest that the MUT-1 and MUT-9 mutator modules perform targeted mutations at 100- to 1000-fold higher frequencies than off-target mutations, and their global mutation frequency is comparable to wildtype DH5α cells. The global mutation frequency increases 10- to 20-fold once UGI is expressed in the mutator system, resulting in all DNA repair of deoxyuridine lesions being blocked. The mutator modules perform C → T mutations at comparable frequencies to AID-T7pol+UGI, while displaying 40-fold lower global mutation frequency (Figure 4.32). This means that cells expressing the mutator modules during long time course mutagenic assays will be able maintain their genomic fidelity, while targeted mutations accumulate in the gene-of-interest. The cells remaining fit and viable for longer means the mutator assays can potentially be run for weeks, where hundreds of successive cycles of library generation with the mutator module plus selection assays for identifying the desired protein function can be performed (we tested the mutator system for up to 144-hours).

4.12 Conclusion

Applying biological engineering principles, key biological parts were characterised for their expression rate in *E. coli* with a library of 15 RBS sequences. Using these bioparts, a library of 2-protein (AP-Endo—EP-DNA-Polymerase) and 3-protein (5’-3’Exo—AP-Endo—EP-DNA-Polymerase) fusions were assembled, designed to perform targeted error-prone DNA repair in *E. coli*. EP-DNA-repair complexes were divided into two groups, one using DNA Pol-IV as the error-prone DNA polymerase and the other group using the polymerase domains from EP-DNA-Polymerase-I designed by Loeb et al[141].

[xxxv] Refer to Figure 4.21

The library of EP-DNA-repair complexes were assembled into mutator modules with the AID-T7pol DNA damaging device. Combining the two devices resulted in a mutator system that is capable of inducing targeted dC → dU DNA damage with AID-T7pol and performing subsequent error-prone DNA repair using the 2-protein or 3-protein fusion to create genetic diversity in the target gene. The loss-of-function and gain-of-function experimental methodologies (described in Chapter 2.5) enabled screening of the library of mutator modules based on their targeted mutagenic activity and the ability to perform functional nucleotide substitutions, respectively.

Analysing the data from these two workflows, versions of the mutator module using *N. Meningitidis* AP-endonuclease or EP-DNA-Pol-I polymerase domains were shown to possess low mutagenic strength and only performed two of the four nucleotide substitutions investigated with the gain-of-function experiments. The mutator module versions expressing Exo-III—Pol-IV$^{\Delta12}$ as the AP-Endo—EP-DNA-Polymerase combination displayed a much higher mutation frequency and performed all four nucleotide substitutions in the gain-of-function assays. The five mutator modules with Exo-III—Pol-IV$^{\Delta12}$ were shortlisted for further optimisation.

Successive 24-hour rounds of the gain-of-function assay with inactive β-lactamase revealed 5’-3’Pol-I Exo—Exo-III—PolIV$^{\Delta12}$ (MUT-7) and 5’-3’Pol-I Exo(s)—Exo-III—PolIV$^{\Delta12}$ (MUT-9) as error-prone DNA repair complexes that can reliably perform C → (A, G, T) substitutions and patch repair + A → T. Patch repair is essential for creating longer ssDNA template, which increases the mutational window of activity for errors incorporated by PolIV$^{\Delta12}$. Patch repair is the mechanism by which mutations can be incorporated at A:T base-pairs. MUT-7 and MUT-9 were further optimised for use in *E. coli* by creating a low- and medium-expression cassette for 5’-3’Pol-I Exo—Exo-III—PolIV$^{\Delta12}$ (MUT-20 and MUT-24) and 5’-3’Pol-I Exo(s)—Exo-III—PolIV$^{\Delta12}$ (MUT-22 and MUT-25). These optimised mutator systems were shown to be active and generate targeted mutations in host cells for up to 96-hours without being toxic to the host cells.

In rifampicin antibiotic selection assays, MUT-1 and MUT-9 displayed a similar global mutation rate to wildtype DH5α. The rate of occurrence of amp+ cells with MUT-9 was 100-fold to 1000-fold higher than the rate of occurrence of rifampicin+ cells, indicating the mutator module’s activity is localised at the target DNA sequence placed downstream of a T7 promoter. The loss-of-function, gain-of-function and rifampicin selection assays provided a qualitative look into the mutation characteristics of MUT-9, MUT-22 and MUT-25. The next step was to perform a more quantitative analysis of the targeted mutation frequency of the 5’-3’Pol-I-Exo(s)—Exo-III—PolIV$^{\Delta12}$ mutator modules, analyse the full spectrum of mutations that they can generate and how these mutations are spread across the ORF of the gene-of-interest.

Chapter 5: Using Next Generation Sequencing to Analyse the Mutation Characteristics of the Mutator Modules

5.1 Introduction

In Chapter 4, a library of mutator modules were assembled with the AID-T7pol DNA damaging device and an error-prone DNA repair complex comprising a 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion protein. After assessing the library of mutator modules for their targeted mutagenic capability using gain-of-function and loss-of-function experiments, the mutator system utilising 5’-3’Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ as the EP-DNA-repair complex (Mut-9) was shortlisted for developing a continuous evolution system. Before doing so, Mut-9 had to be thoroughly assessed for its mutational characteristics, like the complete diversity of mutations it can generate, where it creates these mutations and at what frequency. NGS platforms were utilised to perform this thorough analysis.

The goal with NGS was to obtain enough coverage of mutant DNA library, to quantitively assess the diversity of mutations, mutations rate, and the mutational spread that can be achieved using the mutator modules with 5’-3’Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ as the EP-DNA-repair complex. The goal was also to evaluate if high (MUT-9), medium (MUT-22) and low (MUT-25) levels of expression of the error-prone DNA repair complex directly correlated to a change in the observed mutation rate on the target gene. Also, in the gain-of-function assays with β-lactamase, deletions and insertions could not be analysed. NGS allowed us to investigate all the possible genetic alterations resulting from the MUT-9, MUT-22 and MUT-25 mutator modules. To generate a library of mutant DNA sequences for analysis, the mutator modules were used to target mutations on the GFP-mut3b ORF, integrated into the DH5α genome. The library of mutant GFP-mut3b with the promoters and downstream terminator were amplified via PCR and prepped for sequencing. Two different sequencing platforms were utilised for this endeavour: Pacific Biosciences Sequel and Illumina iSeq100.

PacBio Sequel Platform

The PacBio Sequel platform is based on real-time, single molecule sequencing. It offers the advantage of sequencing very long DNA sequences (1 – 10 kbp) as a single read, rather than requiring the sequences to be fragmented into short 100-200 base-pair lengths[225]. Sequencing such long reads is achieved by ligating hairpin-shaped barcoded adapters to the ends of the DNA sequence library. This creates ssDNA sites flanking the target DNA sequence that are primed by DNA polymerases anchored in arrays of zero-mode waveguides (ZMW)[226]. ZMW consists of small holes in a metal film deposited on a microscope coverslip, where the film acts as cladding, and the contents of the hole compose the core of the waveguide. Millions of such holes can be made on a single coverslip, resulting in massive parallelism in the DNA sequencing process. For direct observation of single-molecule enzymatic activity, the DNA polymerase is adsorbed onto the bottom of the waveguides in a solution containing the fluorescently tagged dNTP molecules. The addition of these tagged nucleotides to the elongating DNA strand is detected, enabling sequencing.

The single read error-rate of PacBio Sequel is 13–15%, which is significantly higher than the 0.1% single read error-rate of Illumina and Ion Torrent sequencing platforms[227,228]. However, this high single-read error-rate is offset by PacBio's ability to generated circular consensus sequence (CCS) reads for each individual DNA molecule. In the PacBio chemistry, the DNA polymerase binds to one of the hairpin adapters and polymerises the target DNA in the 5'-3' direction. For short length DNA targets (< 2.0kbp), the processivity of the DNA polymerase enables it to sequence across the sense strand, loop around to the antisense strand via the second hairpin adapter and continue polymerising the same target DNA sequence. This generates multiple reads of the same DNA molecule, which are all aligned to one another to generate the CCS read. CCS reads increase the accuracy of PacBio Sequel to 97%-99%, while the remaining error-rate may result from the sequencing platform, DNA sample preparation, or the sequencing-library preparation procedures[229]. Overall, as the GFP-mut3b expression cassette is ~ 1.2kbp, PacBio Sequel was able to read through each DNA molecule 10–50 times to generate accurate CCS reads (Figure 5.3). These CCS reads enabled accurate calling of nucleotide substitutions across the complete GFP-mut3b sequence to assess the diversity of mutations, the mutational spread across the whole sequence, and the targeted mutation rate.

Illumina iSeq Platform

Illumina is a high-throughput sequence-by-synthesis sequencing platform, where up to $8x10^9$ individual reads can be generated per run. This is 10^5-fold higher than the number of reads that are obtained from one run of PacBio Sequel[225]. Illumina offers this depth of sequencing information at a single read error-rate of 0.1%. Due to the high number of reads generated, this can equate to ~ 10^8 incorrectly called bases during sequencing. Rectifying these incorrectly called bases involves *in silico* error-correction by aligning the sense and antisense subreads of the same DNA molecule and correcting the base calls with a low Phred score in the overlapped region[230–232]. The caveat is that there must be significant overlap between the sense and antisense reads of a DNA molecule, otherwise error-correction cannot occur at a single-molecule level[230,232–235].

Error-correction at a population level involves aligning multiple reads from different sequenced DNA molecules to a designated reference sequence. During the sequencing process, target sequences are barcoded, and unique hybridisation tags are attached to the 5'- and 3'-ends. Complimentary hybridisation tags are found anchored inside the sequencing flowcell[236]. Once the library of barcoded DNA is passed through the flowcell, hybridisation tags on the DNA sequences anneal to the complimentary anchored tag in the flowcell, and clusters of such DNA molecules are sequenced simultaneously using fluorescently labelled reversible terminator-bound dNTPs. Millions of 150bp reads are generated, which are subsequently aligned to one another and a reference for population-level error-correction. The high sequencing coverage of each 150-300 bp segment and subsequent alignment results in incorrectly called bases being removed.

Such error-correction techniques where sequenced reads from different DNA molecules are aligned to generate a consensus cannot be used in the context of a mutagenesis experiment, as valuable mutations generated by the mutator module will get eliminated. As the mutator module introduces random mutations, it would create great variance between individual 150-bp reads, preventing error-correction algorithms from accurately distinguishing mismatches generated by mutator activity from mismatches caused by sequencing errors. Low frequency SNPs and indels would be discarded, resulting in a loss of mutational data[237–239]. Due to the inability to accurately distinguish sequencing error from true mutations generated by the mutator modules, the data from Illumina iSeq100 was used for a qualitative assessment of the mutator modules, and to validate the mutation characteristics observed in the PacBio data, where the ability to sequence the same DNA molecule 20-50 times resulted in a base-calling accuracy where sequencing errors occurred once every million sequenced bases[240,241].

In this Chapter, the complete mutational profile resulting from the strong, medium and weak mutator modules is presented. Their mutation profile was compared to the profile generated by T7-RNA-Pol and native DH5α^{GFP} host cells, which provided a baseline for the background mutation rate and helped to validate the mutation profile obtained for each of the three mutator modules.

5.2 Experimental Design GFP-mut3b Mutagenesis and Downstream Sample Preparation for Sequencing

Setting up the Continuous Evolution Experiment

The mutagenesis experiment was setup with a genome-integrated GFP-mut3b expression cassette as the target. The gene was subjected to mutations via three mutator modules, a strong (MUT-9), medium (MUT-22) and weak (MUT-25) mutator expressing AID-T7pol as the DNA damage device and 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ as the error-prone DNA repair complex via a strong, medium and weak RBS, respectively (Figure 5.1). The mutagenic capability of these mutator modules was assessed against AID-T7pol+UGI. In Chapter 3, it was confirmed that expression of AID-T7pol+UGI only generates C → T and G → A mutations. A control module expressing only T7 RNA polymerase was used as a negative control where the GFP-mut3b library should exhibit mutation frequencies comparable to wildtype DH5α^{GFP} cells. T7-RNA-Pol and native DH5α^{GFP} cells provided the baseline to accurately call mutations generated by the mutator modules when analysing the NGS sequencing data.

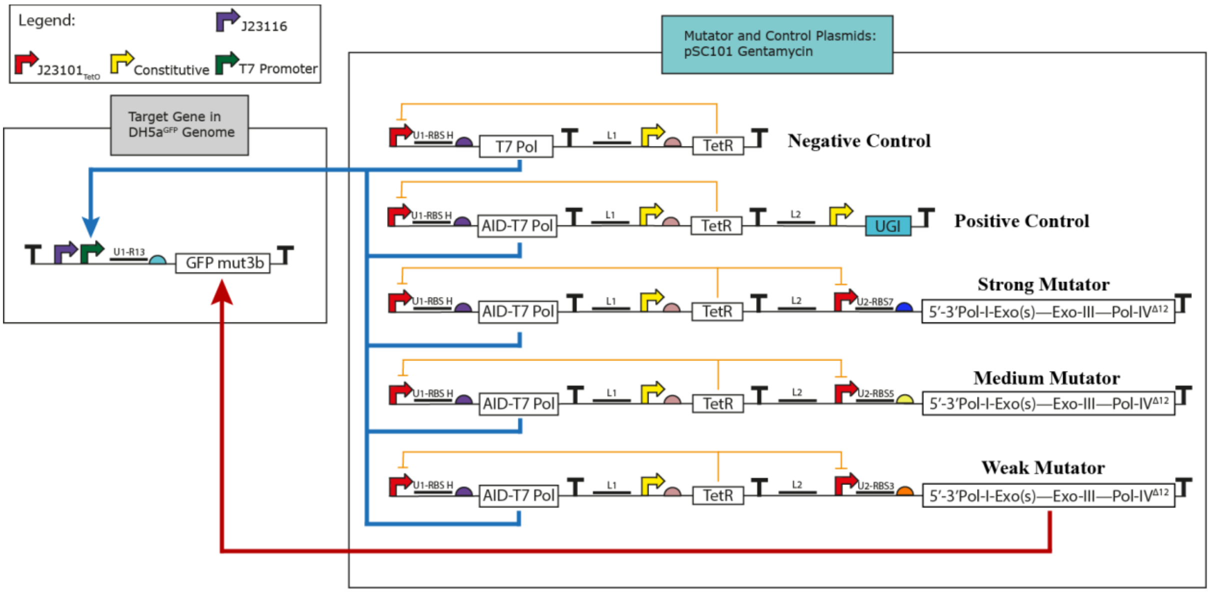


Figure 5.1: SBOL schematic of the loss-of-function experiment performed for 144-hours on GFP-mut3b.

Expressing these mutator and control modules in DH5α^{GFP} cells, libraries of DNA sequences were generated by targeting mutations to the GFP-mut3b ORF. An expression cassette for GFP-mut3b with a J23116 + P_{T7} dual promoter and two T7 terminator sequences downstream was integrated into the

DH5α genome using the pOSIP integrated method[xxxvi]. Plasmids expressing the mutator and control modules were transformed into DH5α^{GFP} cells and plated on LB-agar plates containing gentamycin (25 μgml^{-1}). Three colonies for each experimental condition were picked and diluted into 200 μl of LB, with gentamycin, in a 96-well plate for overnight growth. The following day, the overnight growth cultures were diluted again to a normalised starting OD$_{600}$ of 0.05 in a 2ml deep-well microtiter plate. Anhydrotetracycline (aTc) was added to switch-on the expression of the AID-T7pol DNA damage device and the error-prone DNA repair complexes.

144-hours of targeted mutagenesis

In Chapter 3, it was shown that targeted deamination by AID-T7pol + UGI accumulates more mutations in GFP-mut3b the longer the mutator module is left active, by blocking the cell's BER pathway[xxxvii]. After 24-hours of the mutator assay, only 1 or 2 mutations were identified in the sequenced GFP-mut3b ORF, while 6-10 mutations were witnessed after 144-hours. To test if the mutator modules with an error-prone DNA repair complex could do the same, this mutator assay was performed with the strong, medium and weak mutator systems for 144-hours. The evolving pool of cells were transferred into fresh growth medium with aTc every 24-hours. Roughly 5% (~ 40 million cells at OD$_{600}$ of 1.0) of the cell culture was transferred into fresh media every 24-hours. Transferring a large population of cells ensured no artificial genetic drift was introduced in the population, which could have resulted in the loss of some of the different evolving pathways taken by GFP-mut3b in the evolutionary space[107,242]. Flow cytometry was performed on samples every 24-hours to verify the mutator modules are active. DH5α^{GFP} expressing T7 RNA polymerase were used as a control to monitor for the loss of fluorescence activity in the DH5α^{GFP} cells expressing the mutator modules. After 144-hours, a genomic library from each mutagenised strain was created by isolating the genomes from the cell populations using an appropriate DNA prep kit[xxxviii].

Preparing the GFP-mut3b expression cassettes for sequencing

The mutant GFP-mut3b expression cassettes were amplified from the genomes via PCR with Phusion polymerase and appropriate primers. In the case of sample preparation for PacBio Sequel, primers were used to amplify the complete 1196-bp expression cassette and add BsaI restriction enzyme sites

[xxxvi] Refer to section 7.3.3

[xxxvii] Chapter 3, Section 3.4

[xxxviii] Section 7.2 for details

upstream and downstream of the cassette. The PCR amplicons were subsequently ligated to the barcoded SMRTbell adapters using an adapted version of the BASIC assembly protocol[xxxix]. The concentration of DNA in each barcoded sample was quantified and the samples were pooled in equimolar amounts for the sequencing process.

For Illumina iSeq100, appropriate primers were used to amplify the first 245-bp of the GFP-mut3b ORF and flank these sequences with Illumina-specific DNA tags. The library of tagged PCR amplicons were carried over into a second PCR reaction to attach the Nextera XT pair-end hybridisation tags[243]. The Nextera prefix and suffix tags contain a unique barcode to categorise each sequenced read; they also contain hybridisation tags to anneal to complimentary sequences in the Illumina flowcell. Pair-end sequencing is performed, where 135-nt from the 5'-end were sequenced via the prefix nextera tag (R1 read) and 135-nt of the 3'-end were sequenced via the suffix tag (R2 read), with an overlap of 13-nt between R1 and R2.

For both PacBio and Illumina, the sequencing data was obtained in the 'fastq' format. In the case of PacBio data, the sequencing company determined the single-molecule consensus read sequences (CCS) and filtered the reads based on a Phred quality assessment score. Illumina sequencing was performed in-house and a third party programme, AfterQC[234], was used for assessing the quality of the sequenced reads, for filtering the sequences based on Phred-scores and other quality control parameters. The R1 and R2 subreads from each paired-end reads were combined using FLASH[244] and PandaSeq[230] to obtain sequencing reads representing the complete 245-bp DNA inserts. Subsequently, for both PacBio and Illumina reads, bespoke Python scripts with the Biopython API were used to align the reads to appropriate reference sequences and assess the alignments for the following parameters: types of mutations that occurred; total number of each mutation; the mutational spread across the GFP-mut3b ORF; and the mutation frequency achieved with the strong-, medium- and weak-strength mutator modules.

[xxxix] Section 7.13 for the detailed protocol

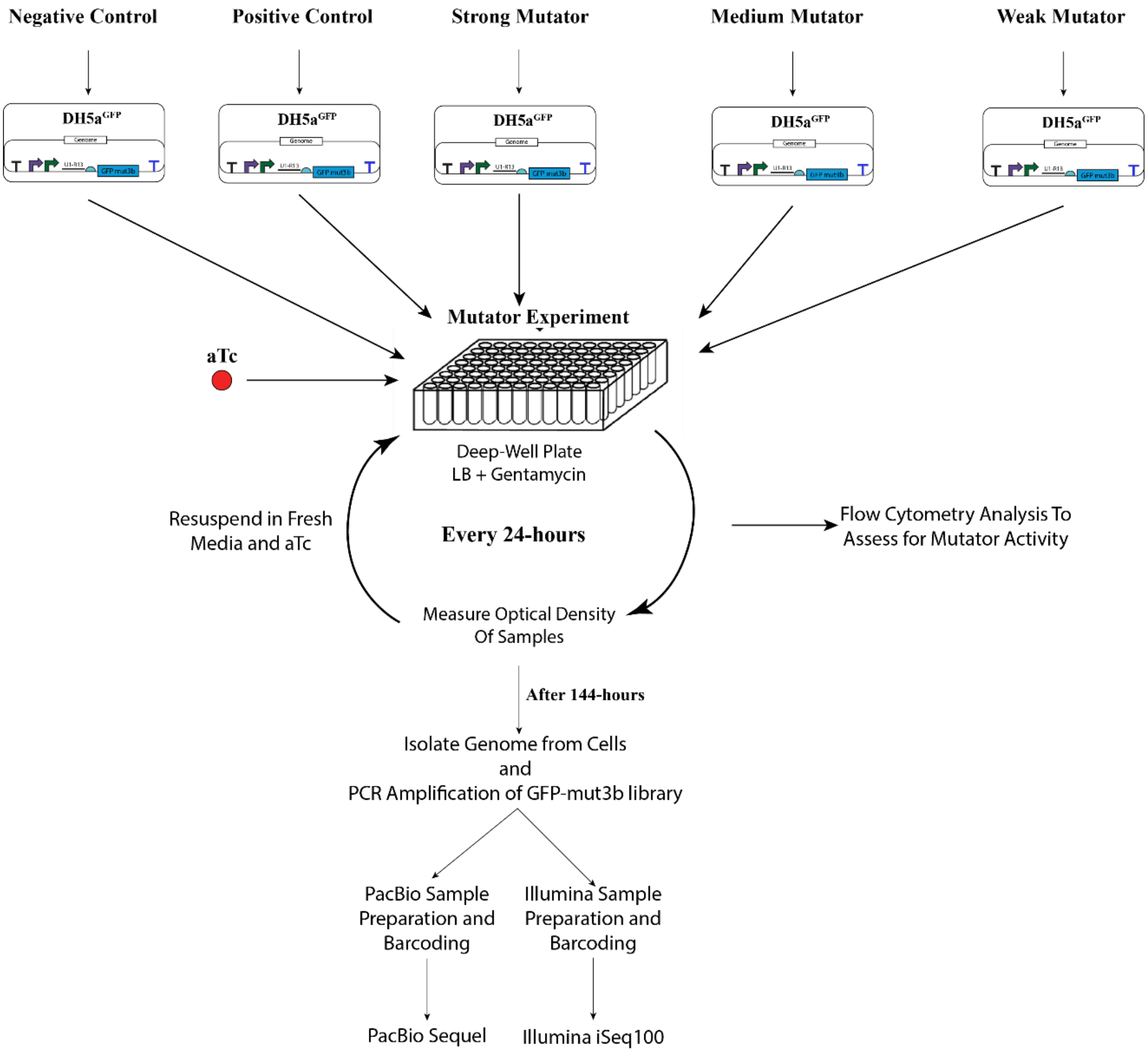


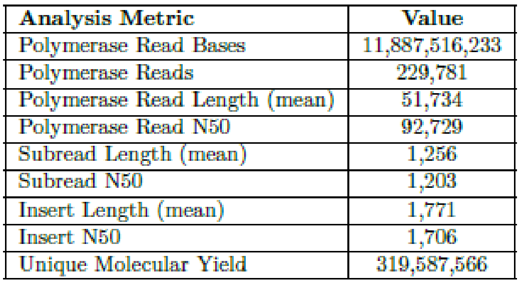
Figure 5.2: Outline of the workflow involving a 144-hour mutator experiment followed by sample preparation for NGS on the PacBio Sequel and Illumina iSeq100 platforms.

5.3 Analysing the Sequencing Data Obtained from PacBio Sequel

5.3.1 Quality of the Reads and Alignment Scores

The individual reads comprising the SMRTbell hairpin adapters and the target DNA sequence were 1256-nt in length. The average processivity of the polymerases in the ZMW was ~ 51,000-nt (Figure 5.3). This means that each circularised DNA molecule was sequenced roughly 40-times to generate multiple subreads. These subreads were subsequently aligned to generate the consensus sequence (CCS). Majority of these CCS reads passed the Phred quality control threshold of 60 (QC_{60}). This means the probability of an SNP resulting from sequencing error was $1x10^{-6}$ per base. The QC_{60} filtered reads were compiled into one file and provided by the company in the 'fastq' format. 50% of the CCS reads were provided of the sense strand, while the remainder sequences were of the antisense strands. Out of 96,000 provided CCS reads, 48,000 reads of the sense strand were used for the mutation analysis. This means that for each of the eight differently barcoded sequences, there were ~ 3000 - 6000 CCS reads to be analysed.

A



Analysis Metric	Value
Polymerase Read Bases	11,887,516,233
Polymerase Reads	229,781
Polymerase Read Length (mean)	51,734
Polymerase Read N50	92,729
Subread Length (mean)	1,256
Subread N50	1,203
Insert Length (mean)	1,771
Insert N50	1,706
Unique Molecular Yield	319,587,566

B

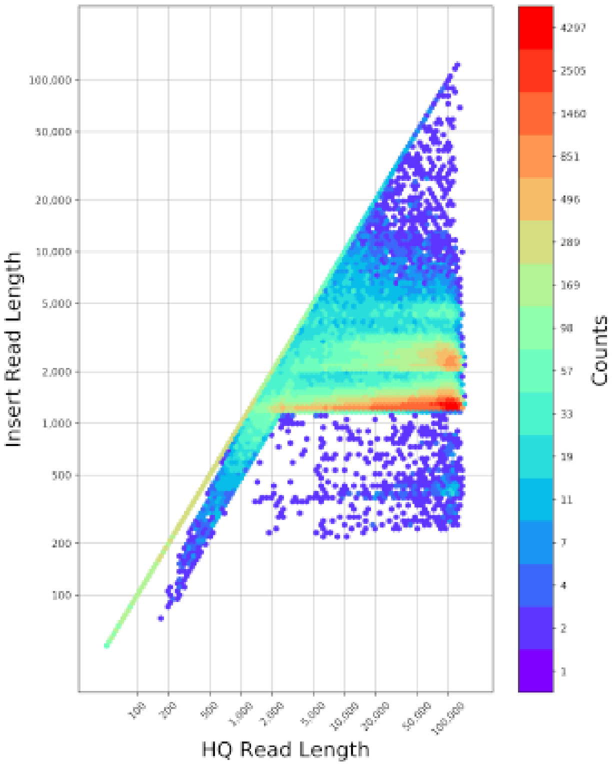
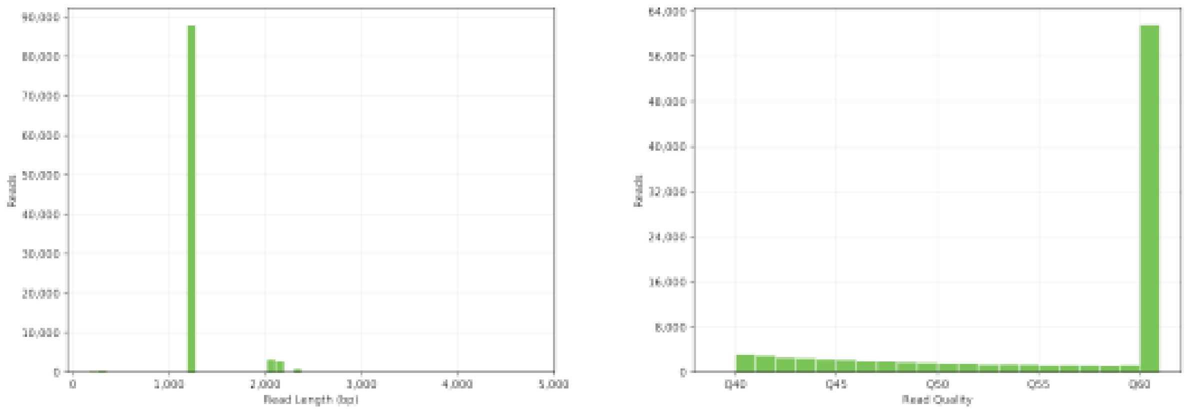


***Figure 5.3: Information regarding the sequencing reads generated from PacBio Sequel**. A – Table shows that a total of ~230,000 different barcoded molecules were sequenced in the Zero Mode Waveguide. The average polymerase read length was 51,000, which indicates that the platform generated 40x coverage of each molecule. The subreads can be aligned to generate a circular consensus sequence (CCS) for each barcoded read. B – A heat map showing the population spread of the length of DNA inserts that were sequenced versus the number of nucleotides sequenced by each polymerase. Majority of the DNA molecules sequenced were ~1200-nt long, with a 20x – 50x sequence coverage of each molecule.*



(a) Read Length Distribution

(b) Read Quality Distribution

***Figure 5.4: Quality control of the PacBio reads using Phred scores**. Left – The GFP-mut3b expression cassette combined with the prefix and suffix barcodes is 1202-nucleotides long. Nearly 90% of the reads were of the correct length, while a few doublets were seen. Right - ~90,000 total CCS reads were generated. 60,000 of these reads passed a Phred score of 60, ie, the chance of an incorrectly called base is 1x10^{-6} per sequenced base.*

Filtering Nonsense and False Positives using Alignment Score Filters

Biopython's Pairwise2 alignment algorithm was used to align the sequencing reads to the reference comprising the J23116 + P_{T7} dual promoter, RBS-15, the GFP-mut3b ORF and the double T7-terminator. The global alignment was performed using four parameters: a score for nucleotide matches; a score for mismatches; a score for creating a gap in the reference or sequencing read; and a score for extending such gaps. The values used for each of these parameters were 1, 0, -1, -0.1, respectively. These parameters enabled generating the best possible alignments between the reference and the sequencing reads, while being tolerant to nucleotide substitutions. Mismatched bases in the alignment were called as nucleotide substitutions by the script. The gap creation and extension values used resulted in gaps only being introduced when the sequencing read was shorter than the reference (indicating deletion) or when the read was longer than the reference (indicating insertion). The highest alignment score of 1202 was assigned to reads that had no mutations and therefore, perfectly aligned with the reference.

Calling mutations with the alignment condition that 0 ≤ alignment score ≤ 1202 produced significant amount of noise in the data. The number of mutation events observed for the negative controls (T7-RNA-Pol and DH5α^{GFP}) was similar to the mutator modules and AID-T7pol+UGI. To eliminate this background noise, a more stringent alignment score filter was applied. Mutations were only called from the reads that aligned to the reference with an alignment score ≥ 1197, thus allowing for up to 5 mismatches. Applying this filtration condition resulted in only 0.0% – 2.0% of the reads being discarded from the total pool of reads for T7-RNA-Pol and DH5α^{GFP} but this this filtration significantly reduced the noise (Tables 5.1 & 5.2). It was a similar case for the medium and weak mutator modules; less than 1% of the reads were discarded. However, for the strong mutator and AID-T7pol+UGI, 60-90% of the reads were discarded at the alignment score filter of 1197. This indicates that the high mutagenic activity of the strong mutator and the AID-T7pol+UGI control resulted in more than 5 mutation events occurring per read. This is corroborated by the Sanger sequencing data in Section 3.4, where 6-10 mutations were seen in each sequenced GFP-mut3b ORF after 144-hrs. However, for a fair comparison between the different mutator modules in the absence of background noise, the same alignment score filter was used for all experimental samples. Owing to this fair comparison, mutations called from the alignment-score-filtered reads could confidently be called as substitutions or insertions-deletions (indels) caused by mutator activity and were not false positives.

Table 5.1 Filtering for Alignment Score > 1197: Its Effect on Total Population of Reads Per Sample

	MUT-9 **<u>Strong</u>**	**MUT-22** **<u>Medium</u>**	**MUT-25-1** **<u>Weak</u>**	**MUT-25-2** **<u>Weak</u>**	**ATF + UGI 1**	**ATF + UGI 2**	**T7pol**	**DH5α^{GFP}**
Total Sequences Before Filtering	3513	6310	6910	5958	2954	2782	5360	5092
Total Sequences After Filtering	1099	6303	6895	5955	300	238	5292	5002
Percentage of Sequences Discarded	68.72	0.11	0.22	0.05	89.84	91.45	1.27	1.77

Table 5.2: Mutation Count for Filtered vs Unfiltered Reads for the negative controls

	Mutation	Unfiltered		Alignment Score > 1197	
Sample Name		T7pol	DH5α^{GFP}	T7pol	DH5α^{GFP}
Substitutions	A --> T	300	464	5	12
	A --> C	355	505	6	6
	A --> G	352	559	21	36
	T --> A	322	510	1	5
	T --> C	309	451	19	30
	T --> G	268	380	3	6
	C --> T	317	500	70	127
	C --> A	328	502	12	14
	C --> G	162	247	1	1
	G --> T	249	453	9	14
	G --> C	157	224	0	1
	G --> A	381	561	83	104
Deletions	T --> -	1899	5052	34	44
	A --> -	2212	5915	12	11
	G --> -	1605	4255	67	56
	C --> -	1317	3685	1	6
Insertions	- --> T	300	426	253	358
	- --> A	301	398	252	349
	- --> G	75	228	59	202
	- --> C	85	76	72	65

5.3.2 Analysing the Mutation Diversity Created by MUT-9, MUT-22 and MUT-25

5.3.2.1 Assessing the strength of the mutator modules

In the context of this experiment, the expression level of the EP-DNA-repair complex in the three mutator modules dictates the relative strength of each mutator system. Two different metrics were used to assess the mutagenic strength of each module: mutations introduced at a population level; and mutations introduced per DNA sequence.

The population level strength of each mutator module was assessed by comparing the number of sequences containing substitutions or indels to the number of sequences that remained unchanged at the end of the continuous evolution experiment. AID-T7pol+UGI, the positive control for mutations, generated U:G lesions in 99.0% of the sequenced reads and only 3 sequences out of 300 remained unchanged (Table 5.3). The strong mutator module displayed a similar strength to AID-T7pol+UGI. In this case, 80.0% of DNA sequences had substitutions or indels, with only 219 sequences remained unchanged. The medium and weak mutator systems performed mutations at a significantly lower frequency than MUT-9. Only 17% - 19% of the sequenced reads for these samples had mutations, while most DNA sequences remained unchanged. Their population-level mutagenic strength was comparable to that displayed by the negative controls, T7-RNA-Pol and native DH5α^{GFP} cells.

Assessing the mutator strength based on the frequency of mutations per read provided a different picture. The average mutation count per DNA sequence was calculated using the formula:

$$Frequency\ of\ Mutations\ per\ Read = \frac{Total\ Mutations\ Called}{Total\ DNA\ Sequences}$$

After 144-hours of mutagenesis, AID-T7pol+UGI generated ~ 3-4 mutations per DNA sequence (Table 5.3). This mutation count per DNA sequence is lower than what was witnessed in the Sanger sequencing data presented in Chapter 3 (Refer to Section 3.4), where 6-10 nucleotide substitutions were found in the analysed reads after the 144-hour mutagenic experiment. For Sanger sequencing, only 6 reads were sampled. Applying an aggressive alignment score filter where only 5 mismatches are allowed and 90% of the reads were discarded, probably resulted in this disparity. The strong mutator, MUT-9, displayed a similar mutation frequency to AID-T7pol+UGI. This finding indicates that using the error-prone DNA repair complex can result in targeted mutations being generated at the same frequency as blocking DNA repair of all U:G lesions.

The medium and weak mutator modules, MUT-22 and MUT-25 displayed significantly lower numbers of mutations on the target gene. Both these mutators averaged ~ 1 mutation per DNA sequence. The large disparity in mutation count could be the result of the different expression levels of the EP-DNA-repair complex between the three mutator systems. This could possibly create a large difference in the number of 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ protein complexes available in the cytosol and the EP-DNA-repair protein complex having to compete with the cell's native DNA repair proteins to bind to the site of damage on the DNA sequence.

Error-prone DNA Repair Complex Having to Compete with Cell's Native BER Pathway

The medium and weak mutator modules displayed a low population-level mutator strength comparative to negative controls, while the mutation count per read sequenced was 3-fold higher indicating higher targeted mutation activity. This phenomenon possibly results from the EP-DNA-repair complex having to compete with the cell's native DNA repair pathway. Once AID-T7pol generates U:G lesions, uracil N-glycosylase scans the DNA sequences for the mismatch and cleaves the deoxyuridine to generate an AP-site. The AP-site can be nicked to create a gap by either natively expressed AP-endonucleases or by Exo-III in the 5’-3’Pol-I-Exo—Exo-III—Pol-IV$^{\Delta 12}$ EP-DNA-repair complex. The final step of the BER pathway could involve DNA Pol-IV$^{\Delta 12}$ having to compete with natively expressed DNA pol-I to repair the gap. The large number of unchanged DNA sequences after the continuous evolution cycle with MUT-22 and MUT-25 indicates that there may not be enough 5’-3’Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ protein molecules in the cytosol to compete with native proof-reading DNA polymerases, resulting in the DNA being properly repaired at a much higher frequency. If the EP-DNA-repair complex is expressed at high levels (MUT-9), this issue seems to be avoided, with significantly higher population-level mutator strength and more mutations being incorporated overall. The caveat with using MUT-9 is that it imposes a higher burden on the cells (Refer to section 4.8), compared to MUT-22 and MUT-25. While the higher burden had low impact on cell growth in this continuous evolution experiment lasting 144-hours, the effect of MUT-9 on host cell viability for longer experimental times is unknown.

To conclude, MUT-9 displayed the strongest affinity for performing mutations, at rates comparable to the AID-T7pol+UGI control, but its expression can be toxic to host cells when activated by the inducer molecule, aTc. MUT-22 and MUT-25 were significantly weaker, performing mutations at 2.5-fold lower frequencies than MUT-9.

Table 5.3: Properties of the Sequencing Reads from Mutator and Control Modules After 144-hour Mutagenesis of GFP-mut3b

	MUT-9 <u>Strong</u>	**MUT-22 <u>Medium</u>**	**MUT-25-1 <u>Weak</u>**	**MUT-25-2 <u>Weak</u>**	**ATF + UGI 1**	**ATF + UGI 2**	**T7pol**	**DH5α^{GFP}**
Reads with Insertions	361	645	684	632	67	59	460	640
Reads with deletions	23	153	156	138	3	0	114	117
Reads with substitutions	496	311	375	326	227	178	211	306
Unchanged Reads	219	5194	5680	4859	3	1	4507	3939
Total Reads	1099	6303	6895	5955	300	238	5292	5002
Total mutations called	2881	5783	4694	4678	992	822	980	1447
Avg. no. of mutations per sequenced read	2.6	0.9	0.7	0.8	3.3	3.5	0.2	0.3

5.3.2.2 Assessing Mutation Frequency and Diversity

Along with mutagenic strength, another important characteristic desired from a mutator system is the ability to generate a diverse mutation spectrum. There are 12 possible nucleotide substitutions that can occur. Most substitutions provide the possibility of converting a three-nucleotide codon into another. More diverse mutations mean a greater degeneracy of the genetic code can be accessed to explore a plethora of evolutionary pathways in the fitness landscape[245]. Apart from nucleotide substitutions, insertions and deletions can also be a good way to explore the search space, only if deleterious frameshifts can be avoided. This requires either tandem addition+deletion of single nucleotides, or addition+deletion of complete codons to ensure the protein coding frame remains intact[246].

To assess the mutation diversity generated by the mutator modules, the sequencing reads were aligned to the appropriate reference and if the alignments had a score of greater than 1197, the nucleotide substitutions, insertions and deletions within the reads were identified. Insertions and deletions were called based on the gaps introduced in the reference or sequenced read by the alignment algorithm. If the sequenced read was longer than the reference, then gaps generated in the reference sequence were counted as insertions. If the sequenced read was shorter than the reference, gaps introduced in the read were counted as deletion events.

For comparability between the mutator modules and the controls, the mutation counts were converted to mutation frequency per sequenced base-pair (sbp^{-1}) using the formula:

$$Mutation\ Frequency\ per\ bp = \frac{Total\ Mutations\ Counted}{Total\ Reads\ x\ Length\ of\ Each\ Read}$$

Assessment of the Control Modules

The negative controls, T7-RNA-Pol and DH5α^{GFP}, displayed the lowest mutation frequencies for 10 of the 12 nucleotide substitutions at ~ 2x10^{-6} sbp^{-1}. The rate of C → T and G → A mutations were 10-fold higher, which could likely be the result of human error while running the continuous evolution experiment (Table 5.4). Cross-contamination of cell cultures during transfer into fresh media could have resulted in cells expressing AID-T7pol+UGI or a mutator module populating the T7-RNA-Pol and DH5α^{GFP} cell cultures. The higher mutation frequency for these substitutions can also result from the natural mechanism of bacterial evolution[247]. In naturally evolving cells, C → T, C → A and G → A are the most frequently occurring substitutions (~ 1x10^{-5} per base-pair)[247]. Overall, in the absence of foreign device to generate mutations, the negative controls displayed a low mutation frequency as expected.

The positive control, AID-T7pol+UGI, only performed C → T and G → A nucleotide substitutions. AID-T7pol generates the U:G lesion, while UGI blocks its repair, resulting in a U → T substitution during DNA replication and a corresponding G → A mutation in the opposite strand. In this experiment, all mutations were counted in the context of the sense strand and AID-T7pol+UGI performed twice as many C → T mutations compared to G → A. This indicates that AID has a higher preference for deaminating the coding strand as opposed to the non-coding strand of the target DNA sequence. In terms of mutation frequency, AID-T7pol+UGI is highly mutagenic and performed C → T at $1.6x10^{-3}$ sbp^{-1}, which is 200-fold higher than the negative controls. This highlights the high activity rate of the AID-T7pol DNA damage device. Somatic hypermutation is transcription dependent, with a higher transcription rate increasing the rate of deamination by AID[248]. AID requires ssDNA as template to hydrolyse the amide bond to convert deoxycytidine to deoxyuridine. The 200-fold higher mutation rate for C → T and G → A compared to sequenced reads from T7-RNA-Pol shows that T7 RNA polymerase is continually transcribing the GFP-mut3b ORF and generating ssDNA in the transcription bubble for AID to perform cytidine deamination.

AID-T7pol+UGI also displayed a high rate of insertion for thymine and adenosine nucleotides. Research by Bowers and colleagues has demonstrated that heterologous expression of AID can generate indels *in vivo* and *in vitro* for evolving the complementarity determining regions (CDR) of antibodies[249]. Research done by Shimatani and colleagues showed that AID-Cas9 fusions generated indels in 60% of the samples when editing the FTIP1e locus to generate transgenic rice[250]. In contrast, this AID-T7pol fusion only displayed indels in 23% of the samples. Overall, AID-T7pol fusion protein generates nucleotide substitutions at a higher rate than indels, which should result in the creation of relatively more functional variants of the target protein in a library than non-functional deleterious aggregates.

Assessment of the Mutator Modules

The mutator modules expressing AID-T7pol and the 5'-3'Pol-Exo(s)—Exo-III—Pol-$IV^{\Delta12}$ EP-DNA-repair complexes were able to generate a wide diversity of mutations. Instances of all 12 possible nucleotide substitutions were counted for MUT-9, MUT-22 and MUT-25. The mutation frequency for each nucleotide substitution was independent of the rate of expression of the EP-DNA-repair complex, except for C → T and G → A. Frequency of occurrence of other substitutions ranged between $2.0x10^{-5}$ and $8.4x10^{-5}$ sbp^{-1} (Table 5.4). C → G and G → C transversions were the least occurring substitution, while C → T and G → A transitions were favoured. Of all the mutations counted, transversion accounted for 49% of the mutations generated by MUT-22 and 45% of the total generated by MUT-25 (Table 5.5). The balanced rate of transitions to transversion makes this system ideal for directed evolution as both classes of DNA mutations create differing ranges of amino acid substitutions. Being able to access both ranges provides the ability to explore a broader spectrum of evolutionary pathways in the protein fitness landscape[251]. The high rate of transversion is also an advantage over *in vitro* mutator systems like EP-PCR, which are transition biased[245].

All three mutator modules seem to generate indels, with insertions being favoured over deletions, like AID-T7pol UGI. This indicates that the insertion events may result from AID activity on the target gene and not the error-prone DNA repair complexes. The deletion events caused by the mutator modules most likely resulted from DNA polymerase IV. Kobayashi and colleagues demonstrated that when a dNTP enters the active site of DNA polymerase IV, as it is bound to a primer-DNA complex, the dNTP can be incorporated at a complementary base downstream due to DNA slippage, which generates a -1 frameshift. In literature, it has been shown that ~ 40% of the mutations generated by DNA Pol-IV are frameshift deletions[252]. However, the mutator modules expressing the truncated Pol-IV$^{\Delta12}$ displayed a significantly lower frequency of deletion. This could be an indication of a reduced affinity to bind to replication forks due to cleavage of the C-terminal amino acids that interact with the β-clamp at DNA replication forks[197,143]. It is important to note that even after generating CCS reads, the PacBio sequencing process can incorrectly call 1.3% of SNPs and indels, with some studies indicating that PacBio has a preference for generating insertions errors[253,227]. One study noted 97% of PacBio generated errors to be indels[254]. So, the high frequency of insertion seen in all the samples could be an artifact of the PacBio sequencing process.

With the strong mutator, 20-fold higher frequencies for C → T and G → A mutations indicates a possible bottleneck in the BER pathway. The U:G mismatch generated by AID-T7pol generates an SOS signal, causing the expression of uracil-N-glycosylase to scan the DNA sequences for the mismatch[255]. However, if the frequency of generating U:G lesions is higher than the frequency with which natively expressed UNG can identify the mismatch and generate AP-sites, it will bottleneck the error-prone DNA repair pathway. 5'-3'Pol-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ requires AP-sites to generate mutations. Low expression of UNG could mean that a portion of U:G lesions are not converted to AP-sites and no error-prone DNA repair is performed to generate mutation diversity. Instead, the U:G mismatch is converted to a C → T or G → A mutation during semi-conservative DNA replication. This hypothesised bottlenecking issue can be tested by expressing UNG in the mutator module to increase the frequency of converting U:G lesions to AP-sites.

Overall, the strong, medium and weak mutator modules displayed a wide diversity of mutations. The ability to generate roughly equal ratios of transitions and transversions means a wide spectrum of amino acid substitutions can be explored. The mutators may also generate insertions at a relatively high frequency, but this could be an artifact of the PacBio sequencing chemistry. The frequency of deletion was 10- to 40-fold lower than the frequency of insertions (Table 5.4).

Table 5.4: Frequency of Each Mutation Type per sbp per read (x10^{-6}) Corrected for PacBio CCS Error-Rate (1.3%)

	Mutation	Strong Mutator	Medium Mutator	Weak Mutator 1	Weak Mutator 2	ATF + UGI 1	ATF + UGI 2	T7pol	DH5α^{GFP}
Substitutions	A --> T	40	52	33	42	0	0	1	2
	A --> C	36	61	42	46	3	0	1	1
	A --> G	45	64	42	49	3	0	3	6
	T --> A	41	59	41	46	0	0	0	1
	T --> C	30	45	32	40	0	7	3	5
	T --> G	33	48	29	37	0	0	0	1
	C --> T	922	54	47	49	1686	1751	11	21
	C --> A	38	58	40	46	0	0	2	2
	C --> G	20	25	18	23	0	3	0	0
	G --> T	42	55	36	44	0	0	1	2
	G --> C	21	22	16	22	0	3	0	0
	G --> A	475	70	50	57	804	811	13	17
Deletions	T --> -	4	8	7	7	3	0	5	7
	A --> -	2	3	3	2	0	0	2	2
	G --> -	8	10	9	10	3	0	10	9
	C --> -	4	0	0	0	3	0	0	1
Insertions	- --> T	238	45	49	50	148	207	39	58
	- --> A	104	53	44	50	83	81	38	56
	- --> G	26	13	11	14	3	0	9	33
	- --> C	37	15	13	15	5	0	11	10

Table 5.5: Ratio of Transitions to Transversion Generated by the Different Mutator Modules

	MUT-9 Strong	MUT-22 Medium	MUT-25-1 Weak	MUT-25-2 Weak
Total Mutations	2621	917	681	786
Total Transitions	1771	279	205	236
Total Transversions	327	456	307	366
Percent Transitions	67.5	30.5	30.1	30.0
Percent Transversions	12.5	49.7	45.1	46.6

5.3.3 Targeted Mutation Frequency Resulting from the Mutator Modules

Mutation frequency is a valuable tool that provides an indication of the likelihood with which the mutator modules will generate mutations in the target gene, per base-pair per generation (sbp^{-1}g^{-1}) of the bacterial cell cycle, during a continuous evolution experiment. In literature, it is reported that for generating a diverse and functional library of protein variants with *in vitro* techniques, 1-5 mutations should occur per mutation cycle[242]. If the mutation rate is too high (~ 5-10 mutations per cycle with

EP-PCR), then majority of the protein variant library will be deleterious or non-functional[256]. In terms of *in vivo* continuous evolution, the orthogonal plasmid-DNA polymerase mutator system of OrthoRep reported a mutation frequency of 1x10^{-5} substitutions per base-pair (s.p.b). At this mutation frequency, the OrthoRep system was able to generate protein variants with the desired properties after 90 yeast generations[109]. This provided us with an indication an appropriate mutation rate for continuous evolution that produces functional variants.

The continuous evolution of GFP-mut3b was performed over 144-hours in *E. coli* DH5α cells. As bacterial cells replicate every 30-minutes, this means ~ 288 cycles of targeted mutation occurred on the GFP ORF. To calculate the targeted mutation frequency generated by the mutator modules, the following formula was used:

$$Mutation\ Frequency = \frac{Total\ Mutations\ Called}{Total\ Sequenced\ Bases\ x\ Total\ Bacterial\ Generations}$$

The mutation frequency is represented per sequenced base-pair per bacterial generation (sbp^{-1}g^{-1}). With the sequencing reads filtered to the alignment score condition of greater than 1197, up to 5 mutational events could be recorded per sequencing read (reference DNA sequence is 1202-bp). At this filtration condition, the strong mutator module displayed a mutation frequency of 0.8x10^{-5} sbp^{-1}g^{-1} (Table 5.6). The medium and weak mutators generated mutations at roughly a 3-fold lower frequency of 2.5x10^{-6} sbp^{-1}g^{-1}. The AID-T7pol+UGI control displayed the highest mutation frequency at 1x10^{-5} sbp^{-1}g^{-1}. AID-T7pol+UGI possessing the highest mutation frequency was expected as the base-excision repair of all U:G lesions is blocked by UGI and these deamination events are passed on to the next generation. The mutation frequency displayed by the strong mutator was only slightly weaker than the AID-T7pol+UGI control, and both of these circuits performed mutations at nearly same frequency as the OrthoRep system (~ 1x10^{-5} s.p.b)[109,122]. These mutation frequencies were calculated for reads filtered at the strict alignment score condition allowing for only 5 mismatches, which resulted in 68%-90% of the reads being discarded for the strong mutator and AID-T7pol+UGI. As these systems can generate greater than 5 mutations per DNA molecule within 144-hours[xl], the true mutation frequency of AID-T7pol+UGI and the strong mutator is potentially higher than 1x10^{-5} sbp^{-1}g^{-1}.

The calculated mutation frequency for MUT-9 and AID-T7pol+UGI is theoretically 5-fold higher than the spontaneous mutation rate in *E. coli*[257,258]. However, the negative controls used in the 144-hr mutator experiment do not reflect this low spontaneous error-rate of *E. coli*. Native DH5α^{GFP} cells and cells expressing T7 RNA polymerase displayed only a 3-fold lower mutation frequency than the weak mutator module, at 8x10^{-7} sbp^{-1}g^{-1}. This could be the result of cross-contamination between cell

[xl] Refer to Sanger sequencing data in Section 3.4

cultures during the transfer step, where cells after 24-hours of a mutation cycle were transferred to fresh media to keep the evolution experiment running. Another possible explanation could be the physiology of bacterial cells in stationary phase. Once cell cultures reach stationary phase, the lack of nutrients results in competition between individual cells and the cell cycle is arrested. Bacterial cells undergo various physiological changes that can alter the proteomic composition of the cells, DNA repair proteins are downregulated and the translation process becomes more error-prone[259,260]. However, under the physiological stress of stationary phase, the natural mutation rate of E. coli only goes up by one order of magnitude to 1x10^{-8} bp^{-1}g^{-1}. It does not completely explain 8x10^{-7} sbp^{-1}g^{-1} mutation frequency witnessed in DH5α^{GFP}. To alleviate the issues resulting from potential cross-contamination and higher basal mutation rate of *E. coli* in stationary phase, the continuous evolution experiment should be repeated in a turbidostat[108,140]. This continuous culture device would enable running multiple cell culture samples in tandem, with minimal human intervention. The cell cultures can be constantly maintained in log phase to avoid cells switching to a stationary phase physiology.

Overall, from the current analysis, the mutator modules are shown to generate mutations in the target gene at the rate of 0.8x10^{-5} – 2x10^{-6} sbp^{-1}g^{-1}, which is 4 or 5-fold higher than the natural mutation rate of *E. coli*, similar to the OrthoRep system.

Table 5.6: Overall Mutation Frequency of GFP-mut3b resulting from mutator and control modules. Sequencing Reads Filtered at Alignment Score > 1197

	MUT-9 **Strong**	**MUT-22** **Medium**	**MUT-25-1** **Weak**	**MUT-25-2** **Weak**	**ATF + UGI 1**	**ATF + UGI 2**	**T7pol**	**DH5α^{GFP}**
Total mutations called	2881	5783	4694	4678	992	822	980	1447
Total Reads	1099	6303	6895	5955	300	238	5292	5002
Total bases sequenced	1320998	7576206	8287790	7157910	360600	286076	6360984	6012404
Mutation Frequency (sbp^{-1}*) x10^{-3}	2.18	0.76	0.57	0.65	2.75	2.87	0.15	0.24
Mutation Frequency (sbp^{-1}g^{-1}) x10^{-5}	0.76	0.27	0.20	0.23	0.96	1.00	0.05	0.08

*** - per sbp refers to per sequenced base pair**

5.3.4 Analysing the Mutation Spread Generated by the Mutator Modules

Having elucidated the relative mutagenic strength of the strong, medium and weak mutator modules, and characterised the diversity of mutations they can generate; the next step was to analyse the nucleotide positions within GFP-mut3b where the mutations occurred. This analysis of the mutation spread provided clarity on whether mutations are concentrated primarily in the target gene's ORF; whether the upstream promoter and RBS sequences were subjected to mutations; and if the terminator sequence effectively prevented RNA polymerase read-through into downstream DNA. The mutation spread was also used to corroborate the inferences from the gain-of-function experiments discussed in Chapter 4, where mutator modules with the 5'-3'Pol-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex performed mutations not only at C:G, but A:T base-pairs as well. In the gain-of-function experiments, the mutator modules reverted a premature TAA stop codon back to TTA and generated functional β-lactamase. Performing this TAA → TTA mutation required patch repair, where the single nucleotide gap generated by Exo-III at a U:G mismatch needed to be extended in the 5'-3' direction by the 5'-3' exonuclease domain or in the 3'-5' direction by Exo-III to create ssDNA template at the TAA codon. This larger gap is then filled by Pol-IV$^{\Delta12}$ with a chance of an A → T transversion occurring. Identifying mutations at A:T base-pairs flanked by G:C within the sequenced reads would confirm this inference.

Mutations in the Promoter and RBS Sites

An effective continuous evolution system requires repeated cycles of generating mutations, and subsequently selecting from the variant library for functional activity. This process is dependent on the ability to transcribe the mutant gene and translate the mRNA to produce protein molecules for assessment by the selection system. If mutations occur in the DNA sequence of the promoter or RBS, this can potentially inactivate gene expression. Functional protein variants that are not expressed will be lost in the selection process. Therefore, it would be a useful quality if the mutation frequencies displayed by the mutator modules at the promoter and RBS sites is lower, compared to the mutation frequency at the target gene ORF.

A dual promoter system was utilised in the expression cassette for the GFP-mut3b gene. The constitutive promoter, J23116, was placed upstream of the T7-promoter. This was done to ensure J23116 is insulated from the mutation process and can express the downstream gene even when the mutator system is switched off. The T7-promoter was placed downstream, where once the transcription initiation complex is formed, DNA is unwound to provide AID with ssDNA template to perform deamination events. As a result, mutations can be spotted in the P_{T7} DNA sequence (Index Positions 143-165, Figure 5.5), generated by each of the three mutator modules. A mutation frequency of ~ 0.4x10^{-6} sbp^{-1} was displayed across the complete promoter sequence until the TATA-box. Roughly 50-fold higher mutation frequencies were displayed in the GC base-pairs just downstream of the TATA-box. This finding can be explained by the activity of RNA polymerases in the transcription initiation complex. Transferring from initiation to processive elongation requires the RNA polymerase to undergo a favourable conformational change. Until this conformation is achieved, the polymerase generates multiple short RNA fragments because of abortive transcription[194]. This means the AID-T7pol fusion possibly spends a relatively long duration of time near the TATA-box, resulting in the higher mutation rate at the downstream GC base-pairs.

Surprisingly, the mutator modules generated mutations in the J23116 promoters upstream (Index positions 71-105, Figure 5.5). This means AID is able to deaminate deoxycytidine residues up to a 100 base-pairs upstream of the T7-RNA-Pol binding site. We hypothesise two possible explanations for these findings. First, this could be the result of simultaneous binding of T7-RNA-Pol to P_{T7} and a bacterial RNA polymerase binding to J23116. The ability to perform multiple simultaneous transcription cycles of a gene via tandem promoters has been documented[261,262]. Therefore, if the AID molecule is in proximity of the transcription bubble created by bacterial RNA polymerase, it could possibly deaminate deoxycytidine on the nearby ssDNA.

The second hypothesis involves promiscuous DNA binding by T7-RNA-Pol. It has been documented that T7 RNA polymerase can bind to non-promoter DNA sequences and perform abortive transcription, generating RNA dinucleotides[194]. Only under two circumstances does T7-RNA-Pol generate longer RNA transcripts: A stable promoter-RNA-polymerase interaction with the T7-promoter; and if the RNA polymerase binds to a poly-dC site. If the RNA polymerase binds to a poly-dC site comprising atleast two dC residues, slippage of the RNA transcript can occur, resulting in elongation of the nascent RNA chain without polymerase translocation to produce RNA fragment longer than 8-nucleotides[194]. This means AID could have been localised to GC-rich regions upstream of the P_{T7} (Index positions 36-40, 43-49, 103-114 and 135-140), generating U:G lesions in these locations.

Along with promoter sequences, mutations occurred in the RBS sequence as well. These mutations were concentrated between 6 base-pairs of the RBS comprising G:C (Index 211-224, Figure 5.5). The A:T base-pair regions upstream and downstream displayed minimal levels of mutations. The medium and weak mutator modules displayed a mutation frequency of $0.6x10^{-6}$ sbp^{-1} and $0.3x10^{-6}$ sbp^{-1} in the 6-bp region. AID-T7pol+UGI and the strong mutator displayed 5-fold and 2-fold higher mutation frequencies at the dG nucleotides in the RBS, respectively.

Overall, the mutator modules displayed significant mutagenic activity across the two promoters and the RBS sequence. As stated earlier, these mutations can possibly result in inactivation of transcription or translation and potential functional phenotypes and cannot be detected by the selection process.

However, if selection pressure is applied during the continuous evolution experiment, it can result in improved promoter and RBS strength — as shown by Ravikumar et al with the OrthoRep system[109].

Mutations in the GFP-mut3b Open Reading Frame

The mutator modules generated mutations across the complete open reading frame of GFP-mut3b. In the 717 base-pair ORF, mutations occurred at all nucleotide positions except 83 of them (Figure 5.5). This means the mutator modules had a sequence coverage of ~ 88%. Of the 83 unchanged nucleotide positions, 82% were either dA or dT. Only 3 dC and 12 dG residues remained unchanged. GFP-mut3b has a 39% GC-content. This means that the mutator module was able to generate mutations in 370 A:T base-pairs from deamination events produced in 264 G:C nucleotide positions within the ORF. The most frequently mutated dA or dT bases were located 1-5 positions upstream or downstream of G:C. This indicates that the 5'-3' and 3'-5' exonuclease activities of the 5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex is limited to a short distance around the U:G lesion generated by AID-T7pol. By analysing the complete mutational spread, characteristic properties of the mutator modules were identified:

1. Poly-dC and GC-rich regions displayed high mutation frequencies. These results were expected as the mutator system damages DNA via AID's deamination of dC. The BER pathway is initiated and Exo-III generates a single nucleotide gap at the U:G lesion AP-site. This gap may or may not be extended by the exonuclease activity of 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$, making error-prone DNA repair at A:T base-pairs less frequent than G:C by design.

2. Poly-dA, Poly-dT and A:T rich regions that are three nucleotides or longer displayed the lowest frequency of mutations. Also, dA or dT bases located in the centre of such A:T rich regions were not mutated (Index positions 240-252, 305-311, 360-371, 477-480, 780-785; Figure 5.5). Within these nucleotide positions, if an A:T base-pair was positioned more than 3-4 nucleotides away from G:C, mutations did not seem to occur. This suggests that the exonuclease activity of the EP-DNA-repair complex is limited to a short range of nucleotides upstream or downstream of gaps generated by Exo-III.

3. Regularly interspersed AT and GC base-pairs resulted in the longest sequence coverage by the mutator modules for generating mutations. The ability to perform patch repair over short stretches surrounding GC base-pairs means the longest segments of DNA with frequently mutated nucleotides were found in regions such as index positions 370-524 in Figure 5.5. Over 96% of the nucleotide positions in this stretch were mutated by the strong, medium and weak mutator modules.

4. WRC and GYW motifs are hot-spot regions for mutations (where W = A/T, R = Purine, Y = Pyrimidine). It is well documented in literature that these tri-nucleotide motifs are hotspots for somatic hypermutations resulting from AID activity[263,264]. This trend was witnessed with the mutator modules and the AID-T7pol+UGI control. The mutational frequency of dC or dG at these sites was 10- to 100-fold higher than the other nucleotide positions in the 1202 sequenced bases. This identified hotspot motifs are listed in Table 5.7.

Table 5.7: Hotspot Regions for Mutations on the GFP-mut3b Expression Cassette

WRC/GYW Motif	Index Positions (Figure 5.5)	Mutation Frequency (sbp^{-1})
GTT	49, 755, 1172	2.5x10^{-6}
AAC	34, 182, 384, 465, 665	1.3x10^{-5}
AGT	204, 600	1.4x10^{-5}
GTA	234, 677	3.8x10^{-5}
TGC	442, 492	5.0x10^{-5}
AGC	468	3.0x10^{-5}
AGC	752	1.0x10^{-4}

Transcriptional Termination of AID-T7pol

A transcription-based mutator system requires the transcription to be efficiently aborted at terminator sites to avoid generating damage in the DNA sequence downstream of the target gene. The efficiency of transcriptional termination of AID-T7pol activity was assessed by calculating the average mutation frequency for all nucleotide positions in the GFP ORF, the double T7-terminator and comparing it to the average mutation frequency for the 5 nucleotides sequenced downstream of the T7 terminator.

The average mutation frequency per nucleotide for the GFP ORF + T7-terminator was ~ 1.46x10^{-6} sbp^{-1} for the mutator modules and AID-T7pol+UGI. For the five nucleotides downstream of the double terminator, the mutation frequency was 9.7x10^{-8} sbp^{-1}. This indicates a 15-fold reduction in the occurrence of mutations downstream of the terminators (Figure 5.5). It should be noted that most of the mutations counted beyond the terminator sequence resulted from the weak mutator module, MUT-25. The strong mutator and AID-T7pol+UGI did not mutate the downstream DNA sequence. Comparing the mutation spread across the GFP ORF to the T7-terminator on the heatmap indicates a clear reduction in the spread of mutations after AID-T7pol polymerises past GFP-mut3b. Mutations occurred in only 50% of the nucleotides forming the terminator sequence. This finding can be explained by the fact that terminator sequences are naturally rich in AT base-pairs with many poly-dA and poly-dT regions. As described earlier, the mutator modules generate mutations at low frequencies in these regions. Overall, the 15-fold reduction in mutation count downstream of the T7-terminators indicates that a degree of transcriptional insulation exists, but further experimentation by sequencing longer stretches of DNA downstream of the GFP-mut3b expression cassette should provide a better understanding of the insulation provided by the double T7-terminator.

Insertions and Deletions

The python script used for calling mutations from alignments was coded such that, if the alignment score was greater than 1197, then base mismatches would be called from the alignment[xli]. In the current version of the script, only the total mutation count at each index position is returned. The different types of mutations that occurred at each index position is unknown. This meant that insertions and deletions were identified, but the exact index positions where each indel event occurred is currently not elucidated. Creating a revised version of the script in the future will enable accurate identification of indel positions.

However, the frequency and type of insertions could be crudely assessed based on the mutations identified at index positions 1202-1210 in Figure 5.5. As insertions increase the length of the target DNA sequence, mutations counted at index positions 1 - 5 and beyond 1201 indicate that an insertion has occurred. Mutation frequencies at indexes 0 → 4 and 1202 → 1206 indicate potential insertion events of 1 to 5 nucleotides, respectively. Majority of the insertions seem to be +1 or +2. If these insertion events occurred in the GFP-mut3b ORF, then a frameshift would have occurred. This would cause a complete loss of phenotype unless a subsequent deletion event occurs to return the ORF to a functional frame. The mutator modules generated +3 insertions in some instances, which could potentially add a codon into the coding sequence of GFP, allowing the gene to retain a functional ORF. Insertions of more than 3 nucleotides were less frequent. As stated in Section 5.3.1, PacBio platforms can generate insertion errors at a frequency of 0.1 per sequenced base[227]. This could also explain the high frequency of insertions seen in the reads sequenced by PacBio

Overall, the mutational frequency data from Table 5.4 indicates that the mutator modules may possess a high rate of creating insertions. This combined with Figure 5.5, indicates that not all insertions are potentially deleterious, with some +3 and +6 insertions occurring after a 144-hour continuous evolution experiment. Successive cell cycles create the possibility for cumulative insertions overtime. Even if a deleterious +1 frameshift occurs in one generation, knocking out GFP expression, the functional phenotype can still be restored in successive generations it is followed by a -1 deletion or a +2 insertion. Functional indels are a powerful source for generating variation in the protein library for directed evolution[265,266]. Consequently, the insertion and deletion characteristics of the mutator modules needs to be explored further using appropriate *in silico* strategies.

[xli] Appendix 9.5 for Python script

Index Position:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
GFP Sequence:	**A**	**T**	**G**	**A**	**T**	**G**	**T**	**G**	**C**	**T**	**A**	**C**	**A**	**T**	**C**	**T**	**G**	**T**	**C**	**C**	**A**	**C**	**A**	**A**	**T**	**T**	**T**	**T**	**C**	**G**	**A**	**A**	**A**	**A**	**A**	**A**	**C**	**C**	**C**	**G**	**C**	**T**	**T**	**C**	**G**	**G**	**C**	**G**	**G**	**G**	**T**	**T**
Mutation Count								**Prefix Index**																																												
Strong Mutator:	22	15.9	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0	0.76	0.76	0	0.76	0.76	0.76	0	1.51	0	0	0	0.76	0.76	0.76	0	0	0	0	0	4.54	0	0	2.27	0.76	0.76	0	0.76	0.76	0	0.76	0.76	0	0	2.27	0
Medium Mutator:	6.47	7	2.64	0.66	0.53	0.53	0.26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.13	0.13	1.32	0.13	0	0	0	0	13.1	0.92	0.13	0.13	0.13	0.13	0	0.13	0.13	0	0.26	0.13	0	0	3.17	0.13
Weak Mutator-1:	1.45	5.07	1.21	0.72	0.48	0.48	0.24	0.24	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0	0.12	0.24	0.12	0	0.24	0.12	0	0.12	0.12	0.36	1.33	0.12	0	0.12	0	0	12.1	0.24	0	0.36	0.24	0.24	0	0.24	0.24	0.36	0.36	0.36	0	0.24	2.53	0.24
Weak Mutator-2:	9.08	4.75	1.96	0.98	0.42	0.42	0.28	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0	0.14	0.14	0.14	0	0.14	0.14	0	0.14	0.42	0.14	1.12	0.14	0	0	0	0	15.8	0.14	0.14	0.42	0.42	0.42	0	0.42	0.56	0.14	0.56	0.7	0	0.14	2.79	0
AID-T7pol UGI 1:	5.55	2.77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T7pol:	4.24	2.83	1.1	0.79	0.31	0.16	0.16	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0.63	1.41	0.31	0	0	0	0	9.28	0.31	0	0	0	0	0	0	0	0	0	0	0	0	1.73	0
DH5αGFP	34.8	20.8	13.1	0.83	0.33	0.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.17	0	0.17	0.83	0.17	0	0	0	0.17	14	0.17	0	0	0	0	0	0	0	0.17	0.33	0.17	0	0	2.16	0
																																					J23116 Promoter															
Index Position:	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103
GFP Sequence:	**T**	**T**	**T**	**T**	**T**	**A**	**T**	**A**	**G**	**C**	**T**	**A**	**A**	**A**	**A**	**G**	**A**	**T**	**T**	**T**	**G**	**A**	**C**	**A**	**G**	**C**	**T**	**A**	**G**	**C**	**T**	**C**	**A**	**G**	**T**	**C**	**C**	**T**	**A**	**G**	**G**	**G**	**A**	**T**	**T**	**G**	**T**	**G**	**C**	**T**	**A**	**G**
Mutation Count																																																				
Strong Mutator:	0	0	0	0	0	5.3	0.76	0.76	0.76	2.27	0.76	0.76	0	0	0	0.76	0.76	0.76	0	0	0.76	0.76	0.76	0.76	0.76	1.51	0.76	0.76	0.76	0.76	0.76	0.76	1.51	0.76	0.76	0.76	0	0.76	0.76	0.76	0	0	0.76	0.76	0	0.76	0.76	0.76	0.76	0.76	0.76	0.76
Medium Mutator:	0	0.13	0	0	0.13	5.54	2.9	0.79	0.4	0.4	0.4	0.66	0.13	0	0.13	0.79	0.4	0.4	0.13	0.13	0.4	0.66	0.53	0.53	0.53	0.4	0.53	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.26	0.26	0.4	0.66	0.53	0.13	0.26	0.4	0.26	0.26	0.4	0.4	0.53	1.72	0.66	0.4	0.53
Weak Mutator-1:	0.12	0	0	0	0	5.43	2.65	0.48	0.36	0.36	0.36	0.24	0	0	0	0.48	0.24	0.36	0	0.12	0.36	0.36	0.48	0.72	0.36	0.36	0.36	0.24	0.36	0.24	0.24	0.24	0.36	0.36	0.24	0.24	0.12	0.36	0.24	0.36	0.12	0	0.36	0.24	0	0.6	0.36	0.48	1.33	0.24	0.24	0.36
Weak Mutator-2:	0	0	0	0	0.14	5.73	2.37	0.28	0.28	0.28	0.28	0.14	0	0	0.14	0.56	0.42	0.28	0	0.28	0.42	0.7	0.28	0.28	0.28	0.42	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.42	0.42	0.14	0.14	0.42	0.28	0.14	0	0.14	0.28	0.14	0.28	0.28	0.56	0.42	0.84	0.56	0.42	0.56
AID-T7pol UGI 1:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0
T7pol:	0	0	0	0	0.47	4.72	1.73	0	0	0	0	0	0	0	0	0.16	0	0	0	0	0.31	0.16	0	0.31	0	0	0	0	0.16	0.47	0.16	0	0.16	0.47	0.16	0	0	0	0	0	0	0	0	0.47	0.31	0.31	0	0	0.94	0	0	0.16
DH5αGFP	0	0	0	0	0.17	3.49	3.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.17	0.17	0.17	0	0	0	0	0.17	0.17	0.17	0	0	0.17	0	0	0	0	0	0	0.17	0.17	0	0	0	0	0.33	1.33	0	0	0.33
		J23116																																											**T7 Promoter**							
Index Position:	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155
GFP Sequence:	**C**	**G**	**C**	**G**	**T**	**C**	**C**	**G**	**G**	**C**	**G**	**T**	**A**	**G**	**A**	**G**	**G**	**A**	**T**	**C**	**G**	**A**	**G**	**A**	**T**	**C**	**T**	**C**	**G**	**A**	**T**	**C**	**C**	**C**	**G**	**C**	**G**	**A**	**A**	**A**	**T**	**T**	**A**	**A**	**T**	**A**	**C**	**G**	**A**	**C**	**T**	**C**
Mutation Count																																																			
Strong Mutator:	0.76	0.76	0.76	0.76	0.76	0.76	0	0.76	0.76	0.76	1.51	0.76	0.76	1.51	0.76	0.76	0	0.76	0.76	0.76	1.51	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0	0	0.76	1.51	0.76	0.76	0	0	0.76	0	0.76	0	0.76	0.76	2.27	0.76	0.76	0.76	0.76	0.76
Medium Mutator:	0.66	0.66	0.53	0.53	0.66	0.53	0.13	0.26	0.79	0.66	0.53	0.4	0.53	0.53	0.4	0.4	0.13	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.66	0.92	0.53	0.4	0	0.26	0.79	0.79	0.53	0.26	0	0.13	0.4	0.26	0.53	0.26	0.53	0.66	0.66	0.53	0.53	0.4	0.53	0.53
Weak Mutator-1:	0.72	0.6	0.24	0.48	0.36	0.24	0	0.48	0.6	0.6	0.24	0.24	0.24	0.24	0.36	0.36	0.12	0.24	0.24	0.24	0.24	0.24	0.24	0.48	0.36	0.24	0.36	0.24	0.48	0.36	0.24	0.36	0.24	0.24	0.84	0.36	0.24	0.24	0	0.12	0.48	0.12	0.48	0.24	0.48	0.48	1.09	0.48	0.36	0.24	0.24	0.36
Weak Mutator-2:	0.56	0.56	0.42	0.56	0.42	0.42	0.28	0.28	0.42	0.56	0.42	0.42	0.42	0.42	0.28	0.56	0.14	0.28	0.28	0.28	0.28	0.42	0.28	0.42	0.42	0.42	0.28	0.28	0.28	0.42	0.42	0.42	0	0.28	1.12	0.56	0.56	0.14	0	0.14	0.28	0.14	0.28	0.14	0.28	0.28	0.84	0.28	0.42	0.28	0.28	0.28
AID-T7pol UGI 1:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0	2.77	0	0
AID-T7pol UGI 2:	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T7pol:	0.63	0	0	0	0	0.16	0.16	0.31	0.31	0.16	0.31	0	0	0	0	0	0	0	0	0	0	0.16	0.16	0	0	0	0.16	0	0	0	0	0	0.16	0	1.26	0	0	0.16	0.16	0.16	0	0	0.31	0	0	0.16	0.47	0	0	0	0	0
DH5αGFP	0.5	0.33	0	0	0	0	0.17	0	0.17	0.33	0.33	0	0.17	0.33	0.33	0.17	0.17	0.17	0	0	0	0	0.33	0.17	0	0.17	0	0.17	0	0	0	0.17	0	0.17	0.67	0	0	0	0	0	0	0	0.17	0	0	0.17	0.17	0.17	0.17	0.17	0	0
					T7 Promoter																																															
Index Position:	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
GFP Sequence:	**A**	**C**	**T**	**A**	**T**	**A**	**G**	**G**	**G**	**T**	**A**	**C**	**T**	**A**	**G**	**A**	**G**	**G**	**G**	**C**	**T**	**C**	**G**	**T**	**T**	**G**	**A**	**A**	**C**	**A**	**C**	**C**	**G**	**T**	**C**	**T**	**C**	**A**	**G**	**G**	**T**	**A**	**A**	**G**	**T**	**A**	**T**	**C**	**A**	**G**	**T**	**T**
Mutation Count																																																			
Strong Mutator:	0.76	0.76	0.76	0.76	0.76	0.76	23.5	1.51	12.1	3.03	0.76	11.4	1.51	1.51	0.76	0.76	1.51	0	0.76	3.79	0.76	0.76	10.6	1.51	0	1.51	0.76	0	12.9	6.81	9.84	3.79	4.54	0.76	1.51	0.76	1.51	0.76	1.51	4.54	1.51	0.76	0	4.54	1.51	0.76	0.76	1.51	0.76	8.33	0.76	0
Medium Mutator:	0.4	0.4	0.53	0.4	0.53	0.53	0.53	0	0.13	0.4	0.53	0.53	0.4	0.53	0.4	0.53	0.4	0	0.26	0.53	0.66	0.53	0.66	0.53	0.4	0.79	0.53	0.26	0.79	0.79	0.79	0.13	0.66	0.66	0.53	0.66	0.66	1.06	0.53	0.13	0.92	0.53	0.13	0.79	0.66	0.66	0.66	0.66	0.79	0.66	0.92	0.13
Weak Mutator-1:	0.24	0.48	0.36	0.24	0.36	0.24	0.24	0	0	0.48	0.36	0.36	0.24	0.36	0.24	0.24	0.36	0.12	0	0.36	0.24	0.24	0.36	0.36	0	0.24	0.24	0	0.24	0.36	0.36	0	0.48	0.48	0.24	0.36	0.24	0.24	0.36	0.12	0.6	0.24	0	0.24	0.36	0.36	0.36	0.24	0.24	0.36	0.6	0
Weak Mutator-2:	0.28	0.42	0.28	0.28	0.28	0.28	0.28	0.14	0.28	0.7	0.42	0.84	0.42	0.42	0.42	0.56	0.28	0	0.28	0.56	0.42	0.42	0.42	0.42	0.28	0.98	0.28	0.42	0.7	0.56	0.7	0.14	0.56	0.7	0.56	0.56	0.42	0.42	0.42	0.14	0.98	0.28	0.14	0.42	0.56	0.56	0.56	0.56	0.98	0.84	0.98	0.42
AID-T7pol UGI 1:	0	0	0	0	0	0	55.5	2.77	36.1	5.55	0	11.1	0	0	0	0	0	0	8.32	8.32	0	0	16.6	0	0	0	0	0	27.7	8.32	16.6	0	8.32	0	0	0	2.77	0	2.77	2.77	0	0	0	5.55	0	0	0	0	0	13.9	0	0
AID-T7pol UGI 2:	0	0	0	0	0	0	24.5	6.99	38.5	10.5	0	17.5	0	0	0	0	0	0	0	10.5	0	0	17.5	0	0	3.5	0	0	35	6.99	28	0	3.5	0	0	0	0	0	0	3.5	6.99	0	0	6.99	0	3.5	0	0	0	14	0	0
T7pol:	0	0.16	0	0	0	0	0	0	0	0.16	0.31	0.16	0	0	0	0	0	0	0	0.31	0	0	0.16	0	0	0.16	0	0.16	0.16	0	0.16	0	0	0.31	0.16	0	0	0.31	0	0	0	0.16	0.16	0	0	0.16	0	0	0.31	0.16	0	0
DH5αGFP	0	0	0	0	0	0	0	0.17	0.17	0.33	0	0.33	0.17	0	0	0.17	0.17	0.17	0.17	0.33	0	0	0.17	0.33	0	0	0	0	0.5	0.17	0.17	0	0	0.17	0	0.33	0	0	0.33	0	0	0	0	0	0.33	0	0	0	0.17	0	0.17	0.17
										RBS Sequence																																									
Index Position:	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259
GFP Sequence:	**G**	**T**	**A**	**A**	**A**	**A**	**A**	**G**	**A**	**G**	**G**	**A**	**G**	**A**	**A**	**A**	**T**	**A**	**G**	**T**	**C**	**C**	**A**	**T**	**G**	**C**	**G**	**T**	**A**	**A**	**A**	**G**	**G**	**A**	**G**	**A**	**A**	**G**	**A**	**A**	**C**	**T**	**T**	**T**	**T**	**C**	**A**	**C**	**T**	**G**	**G**	**A**
Mutation Count																																																			
Strong Mutator:	34.1	4.54	1.51	0	0	0	0	0.76	0.76	0.76	0	0.76	1.51	0.76	0	0	0.76	0.76	1.51	0.76	6.06	0.76	0.76	0.76	9.84	15.1	26.5	6.06	0.76	0	0	1.51	0	0.76	0.76	0.76	0	1.51	0.76	0	2.27	0.76	0	0	0	2.27	0.76	6.06	1.51	6.81	1.51	0.76
Medium Mutator:	0.79	0.66	1.06	0.13	0	0	0.13	1.06	0.66	0.53	0.26	0.92	0.79	0.66	0	0.26	1.06	0.79	0.79	0.66	0.53	0.13	0.79	0.66	0.92	0.79	2.11	0.79	0.79	0	0.13	0.66	0.13	0.79	0.92	0.79	0.26	0.92	0.79	0.26	1.06	0.92	0.13	0.13	0.26	1.19	0.79	0.92	0.79	0.79	0.13	0.79
Weak Mutator-1:	0.36	0.24	0.48	0.12	0	0	0	0.48	0.24	0.24	0	0.24	0.24	0.24	0	0	0.24	0.24	0.24	0.36	0.48	0.12	0.24	0.36	0.48	0.48	0.48	0.36	0.24	0	0	0.24	0	0.24	0.36	0.24	0.24	0.48	0.24	0	0.36	0.24	0	0.12	0	0.48	0.36	0.6	0.24	0.24	0	0.24
Weak Mutator-2:	0.56	0.7	1.12	0.28	0.14	0.14	0.28	0.84	0.7	0.28	0.14	0.42	0.42	0.42	0	0.14	0.56	0.42	0.42	0.42	0.42	0.14	0.56	0.56	0.56	0.56	0.7	0.56	0.28	0	0.14	0.28	0.14	0.42	0.56	0.28	0.28	0.42	0.28	0.14	0.42	0.28	0.14	0	0.14	0.42	0.56	0.7	0.56	0.28	0.14	0.42
AID-T7pol UGI 1:	61	2.77	0	0	0	0	0	0	0	2.77	0	0	5.55	0	0	0	0	0	0	0	2.77	0	2.77	0	0	41.6	38.8	0	0	0	0	5.55	0	0	5.55	0	0	0	0	0	0	0	0	0	0	16.6	0	25	0	5.55	0	0
AID-T7pol UGI 2:	24.5	0	0	0	0	0	0	0	0	3.5	0	0	3.5	0	0	0	0	0	0	0	0	0	3.5	0	3.5	41.9	38.5	3.5	0	0	0	0	0	0	0	0	0	0	0	0	6.99	0	0	0	0	6.99	0	17.5	0	0	6.99	0
T7pol:	0	0	0.31	0	0	0	0	0.47	0	0	0	0	0	0.16	0	0	0	0	0	0.31	0	0	0.16	0	0.16	0	0.16	0.16	0	0	0	0	0	0.16	0.31	0.16	0.31	0.31	0.47	0.31	0	0	0	0.16	0	0	0	0	0.16	0.16	0	0
DH5αGFP	0.33	0.33	0.33	0.17	0	0	0	0.5	0.17	0	0	0	0	0	0	0	0.17	0.17	0	0.17	0	0	0	0.17	0.33	0.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0.17	0.17	0	0	0

Figure 5.5

Index Position:	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311
GFP Sequence:	G	T	T	G	T	C	C	C	A	A	T	T	C	T	T	G	T	T	G	A	A	T	T	A	G	A	T	G	G	T	G	A	T	G	T	T	A	A	T	G	G	G	C	A	C	A	A	A	T	T	T	T
Mutation Count																																																				
Strong Mutator:	7.57	1.51	0	0.76	0.76	3.79	3.79	1.51	3.79	0	0.76	0	1.51	0.76	0	19.7	2.27	0	11.4	0.76	0	0.76	0.76	0.76	3.79	0.76	0.76	0.76	6.81	3.79	3.03	0.76	0.76	22.7	4.54	0.76	0.76	0	0.76	0.76	0	3.79	0.76	0.76	1.51	2.27	0.76	0	0.76	0	0	0
Medium Mutator:	0.79	0.79	0.26	1.06	0.79	1.58	0	0.13	0.66	0.13	0.66	0.13	0.92	0.66	0.26	0.79	0.66	0.13	1.06	0.66	0.26	0.92	0.13	0.92	0.79	0.92	0.79	0.79	0.26	1.06	0.92	0.92	0.92	0.92	0.92	0.13	0.79	0.13	0.92	0.92	0	1.72	1.32	0.92	1.06	0.79	0	0.4	0.92	0	0	0.13
Weak Mutator-1:	0.24	0.24	0	0.36	0.36	0.84	0.12	0	0.36	0	0.36	0	0.36	0.36	0.12	0.36	0.36	0	0.36	0.6	0.24	0.6	0	0.48	0.24	0.24	0.24	0.24	0.12	0.24	0.24	0.24	0.24	0.24	0.24	0	0.36	0.12	0.24	0.48	0.24	2.65	0.6	0.24	0.36	0.24	0	0.12	0.36	0	0	0
Weak Mutator-2:	0.42	0.42	0.14	0.42	0.7	1.54	0.28	0.14	0.28	0.14	0.56	0.14	0.56	0.42	0.14	0.56	0.7	0.14	0.56	0.42	0.14	0.56	0.28	0.56	0.7	0.7	0.56	0.42	0.14	0.56	0.56	0.56	0.7	0.56	0.42	0.14	0.56	0.14	0.56	0.84	0	3.35	1.12	0.7	0.7	0.56	0	0.28	0.56	0.14	0	0.28
AID-T7pol UGI 1:	0	0	0	0	0	0	13.9	2.77	0	0	0	0	0	0	0	25	2.77	0	27.7	0	0	0	0	0	0	0	0	2.77	8.32	5.55	2.77	2.77	0	33.3	0	0	0	0	0	0	0	0	2.77	2.77	0	0	0	0	0	0	0	0
AID-T7pol UGI 2:	3.5	3.5	0	3.5	0	6.99	3.5	3.5	0	0	0	0	10.5	0	0	28	0	0	24.5	3.5	0	0	0	0	14	0	0	3.5	6.99	0	3.5	0	0	17.5	0	0	0	0	0	0	0	0	0	0	6.99	0	0	0	0	0	0	0
T7pol:	0.16	0	0	0.16	0	0.16	0	0.16	0.16	0	0	0.16	0	0	0	0	0	0.16	0.16	0	0	0	0	0	0	0	0	0	0.16	0	0.47	0	0	0.16	0	0	0.16	0.16	0	0	0.16	3.3	0.16	0	0	0	0	0	0.16	0	0	0
DH5αGFP	0.33	0.17	0.17	0	0.17	0.83	0.33	0.17	0.17	0	0.33	0	0.17	0	0	0.17	0.17	0.17	0	0.17	0.17	0.33	0	0	0.17	0	0	0	0	0.17	0	0	0	0	0.17	0.17	0.17	0.33	0	0.17	0.17	3.66	0.67	0	0	0.33	0.17	0.17	0.33	0.17	0.17	0

GFP-mut3b

Index Position:	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363
GFP Sequence:	C	T	G	T	C	A	G	T	G	G	A	G	A	G	G	G	T	G	A	A	G	G	T	G	A	T	G	C	A	A	C	A	T	A	C	G	G	A	A	A	A	C	T	T	A	C	C	C	T	T	A	A
Mutation Count																																																				
Strong Mutator:	2.27	0.76	0.76	1.51	0.76	0.76	1.51	0.76	3.03	0	0.76	2.27	0.76	1.51	0.76	1.51	1.51	5.3	0.76	0.76	2.27	2.27	0.76	1.51	0.76	0.76	2.27	3.03	0.76	0	3.03	1.51	0.76	0.76	12.9	9.84	1.51	0.76	0	0	0	4.54	0.76	0	0.76	4.54	9.84	2.27	1.51	0	0.76	0
Medium Mutator:	1.06	1.06	0.92	1.06	0.92	0.92	0.92	0.92	0.79	0.26	1.19	0.92	0.92	1.19	0.4	0.26	1.45	0.92	0.79	0.13	0.92	0.13	1.32	1.19	0.92	1.06	0.92	1.06	0.79	0.53	1.19	0.92	0.92	1.19	1.06	1.19	0.92	0.79	0	0	0.26	1.19	0.66	0.26	1.06	0.66	0.13	0.26	0.66	0.26	0.79	0
Weak Mutator-1:	0.48	0.48	0.48	0.24	0.24	0.24	0.24	0.48	0.36	0	0.48	0.24	0.24	0.24	0.24	0.12	0.24	0.6	0.48	0	0.36	0	0.24	0.36	0.24	0.24	0.48	0.84	0.72	0.12	0.24	0.24	0.24	0.24	0.24	0.72	0.12	0.6	0	0	0	0.24	0.24	0	0.36	0.24	0	0	0.36	0	0.24	0
Weak Mutator-2:	0.7	0.56	0.56	0.7	0.56	0.56	0.7	0.7	0.42	0.7	0.98	0.56	0.56	0.56	0.14	0.28	0.84	0.7	0.98	0.42	0.56	0.28	0.56	0.7	0.56	0.84	0.56	1.12	0.7	0.28	0.56	0.56	0.56	0.84	0.56	1.12	0.42	0.98	0.28	0	0.14	0.84	0.56	0.14	0.7	0.7	0	0.14	0.7	0.14	0.56	0
AID-T7pol UGI 1:	0	0	2.77	0	0	0	5.55	0	2.77	0	0	0	0	0	0	2.77	2.77	0	0	0	0	0	0	2.77	0	0	2.77	11.1	2.77	0	5.55	0	0	0	22.2	0	5.55	2.77	0	0	0	13.9	0	0	0	5.55	11.1	0	0	0	0	0
AID-T7pol UGI 2:	3.5	0	3.5	3.5	0	0	10.5	0	6.99	3.5	0	0	0	0	0	0	0	3.5	0	0	0	0	3.5	0	0	0	3.5	6.99	6.99	3.5	0	3.5	0	0	10.5	10.5	0	0	0	0	0	0	0	0	0	14	10.5	3.5	0	0	0	0
T7pol:	0	0	0	0	0	0	0	0	0.31	0	0.16	0.16	0	0	0	0.16	0	0.16	0	0	0.16	0	0	0.16	0	0	0	0.16	0	0	0	0	0.16	0	0	0	0.16	0.16	0.16	0.16	0	0	0	0	0	0.16	0	0	0	0	0.31	0
DH5αGFP	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0.17	0.17	0.33	0.17	0	0	0	0.17	0	0	0	0.33	0.33	0.33	0	0.33	0	0	0	0.17	0.5	0.17	0.17	0	0	0	0.17	0	0	0	0.33	0	0	0	0	0	0

GFP-mut3b

Index Position:	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415
GFP Sequence:	A	T	T	T	A	T	T	T	G	C	A	C	T	A	C	T	G	G	A	A	A	A	C	T	A	C	C	T	G	T	T	C	C	A	T	G	G	C	C	A	A	C	A	C	T	T	G	T	C	A	C	T
Mutation Count																																																				
Strong Mutator:	0	0.76	0	0	0.76	0.76	0	0	3.79	2.27	5.3	9.08	0.76	1.51	9.08	1.51	2.27	0.76	0.76	0	0	0	20.4	3.79	0.76	3.79	2.27	1.51	6.06	2.27	0	23.5	5.3	3.03	0.76	0.76	0	5.3	0	0.76	0.76	5.3	1.51	3.79	0.76	0	3.03	0.76	0.76	0.76	3.79	0.76
Medium Mutator:	0.26	0.79	0.13	0.26	0.92	0.79	0.13	0.26	1.06	1.85	1.19	0.92	0.92	0.92	0.92	1.19	0.66	0.26	0.66	0	0	0.4	1.45	1.06	1.19	0.92	0.26	0.92	0.92	0.66	0.4	0.92	0.26	1.06	1.06	0.79	1.45	1.85	0.4	0.79	0.26	0.92	0.92	0.92	0.79	0.53	0.92	1.19	0.92	0.92	1.32	1.06
Weak Mutator-1:	0	0.24	0	0	0.24	0.24	0	0.12	0.36	0.84	1.09	0.84	0.36	0.36	0.36	0.84	0.6	0	0.72	0.24	0	0	0.97	0.36	0.48	0.48	0.12	0.36	0.36	0.36	0	0.48	0	0.48	0.6	0.36	0.72	0.72	0.36	0.36	0	0.36	0.36	0.48	0.36	0	0.48	0.36	0.48	0.36	0.36	0.48
Weak Mutator-2:	0.14	0.7	0.14	0.28	0.98	0.56	0	0.42	0.84	1.54	0.84	1.12	0.7	0.7	0.84	0.98	0.7	0.14	0.7	0	0.14	0.14	0.84	0.84	0.7	0.56	0.28	0.7	0.84	0.56	0.14	0.7	0.28	0.84	0.7	0.56	1.12	1.4	0.28	0.7	0.14	0.84	0.84	0.7	0.56	0.14	0.84	0.84	0.84	0.7	0.7	0.84
AID-T7pol UGI 1:	0	0	0	0	0	0	0	0	5.55	2.77	2.77	5.55	0	0	11.1	0	0	5.55	0	0	0	0	44.4	2.77	0	2.77	2.77	0	5.55	0	0	63.8	2.77	0	0	0	0	5.55	2.77	0	0	8.32	2.77	0	0	0	0	0	0	0	5.55	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	0	6.99	14	3.5	3.5	0	0	10.5	0	3.5	0	0	0	0	0	45.4	0	0	0	3.5	0	3.5	0	0	55.9	0	3.5	0	0	0	6.99	3.5	0	0	0	0	0	0	0	0	0	0	0	14	0
T7pol:	0	0	0	0	0.16	0	0.16	0.16	0.16	0.79	0.16	0.31	0	0	0	0.16	0.16	0	0.31	0.16	0.16	0.16	0.16	0	0	0.16	0	0	0.16	0	0	0.31	0	0.16	0	0	0.31	0.31	0	0	0	0	0.16	0	0.16	0.16	0.16	0	0	0.16	0.16	0.16
DH5αGFP	0	0.17	0	0.17	0.17	0.17	0.17	0	0.17	1.66	0.33	0	0	0	0.17	0.33	0.5	0	0	0	0.17	0	0	0	0.17	0	0	0.17	0	0	0	0.17	0	0.17	0.17	0.17	0.83	1.33	0	0	0	0.17	0	0.17	0.17	0	0.17	0.17	0.33	0.17	0	0

GFP-mut3b

Index Position:	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467
GFP Sequence:	A	C	T	T	T	C	G	G	T	T	A	T	G	G	T	G	T	T	C	A	A	T	G	C	T	T	T	G	C	G	A	G	A	T	A	C	C	C	A	G	A	T	C	A	T	A	T	G	A	A	A	C
Mutation Count																																																				
Strong Mutator:	1.51	1.51	0.76	0	0	1.51	1.51	8.33	3.03	0.76	0.76	0.76	1.51	6.06	1.51	16.7	3.03	0	0.76	0.76	0	1.51	3.79	6.81	0.76	0	0	3.79	27.3	12.1	1.51	2.27	1.51	0.76	0.76	2.27	6.81	0.76	0.76	3.03	2.27	0.76	3.79	0.76	0.76	0.76	0.76	2.27	0.76	0	0	25
Medium Mutator:	0.92	0.92	0.66	0	0.26	1.06	0.79	0.53	0.66	0.26	0.92	0.92	0.79	0.26	1.06	1.06	0.66	0.4	1.06	0.79	0.26	1.32	1.06	0.92	0.66	0	0.26	1.06	1.45	1.32	1.06	1.19	1.19	1.19	0.92	0.92	0.13	0.4	0.92	0.92	1.06	0.92	1.06	1.19	0.92	0.92	1.19	1.06	0.66	0	0.26	0.92
Weak Mutator-1:	0.48	0.48	0.36	0	0	0.48	0.48	0	0.6	0.12	0.48	0.48	0.36	0.24	0.48	0.48	0.36	0	0.6	0.48	0	0.36	0.36	0.48	0.48	0	0	0.48	1.09	1.09	0.6	0.48	0.48	0.48	0.6	0.72	0.12	0	0.84	0.6	0.48	0.48	0.48	0.48	0.48	0.48	0.48	1.09	0.72	0.24	0.36	0.84
Weak Mutator-2:	0.7	0.7	0.56	0	0.14	0.84	0.56	0.28	0.7	0.28	0.98	0.84	0.84	0.14	0.7	0.7	0.84	0.28	0.84	0.7	0.42	0.84	0.7	0.7	0.7	0	0.42	0.7	0.84	1.12	0.7	0.7	0.7	0.7	0.7	0.7	0.14	0.14	1.26	0.84	0.7	0.98	0.84	0.7	0.84	0.98	0.7	0.84	0.56	0	0.28	0.84
AID-T7pol UGI 1:	0	0	0	0	0	5.55	5.55	13.9	8.32	0	0	0	0	11.1	0	22.2	0	0	0	0	0	0	2.77	8.32	0	0	0	2.77	49.9	8.32	0	2.77	0	0	0	13.9	2.77	5.55	2.77	2.77	0	0	0	0	0	0	0	0	0	0	0	33.3
AID-T7pol UGI 2:	0	10.5	0	0	0	10.5	0	28	0	0	0	0	0	14	0	17.5	3.5	0	0	0	0	0	17.5	3.5	0	0	0	3.5	66.4	6.99	6.99	0	0	0	0	10.5	3.5	6.99	0	0	0	0	6.99	0	0	0	0	3.5	0	0	0	52.4
T7pol:	0	0	0.16	0.16	0	0.16	0.16	0	0	0.16	0	0	0.16	0.16	0.16	0.16	0	0	0.16	0.16	0	0.16	0	0	0	0	0.16	0.47	0.79	0.31	0	0	0	0	0.16	0.16	0.16	0	0	0	0.16	0	0	0.16	0.16	0.16	0	0.16	0	0	0	0.16
DH5αGFP	0	0.33	0	0	0	0.17	0	0	0.17	0	0	0.17	0.17	0	0	0	0.17	0	0	0.17	0.17	0	0	0	0.17	0	0	0.17	0.5	0.17	0	0	0	0	0	0.67	0	0.17	0	0	0	0	0	0	0.17	0	0	0	0.33	0.17	0.17	0.17

GFP-mut3b

Index Position:	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519
GFP Sequence:	A	G	C	A	T	G	A	C	T	T	T	T	T	C	A	A	G	A	G	T	G	C	C	A	T	G	C	C	C	G	A	A	G	G	T	T	A	T	G	T	A	C	A	G	G	A	A	A	G	A	A	C
Mutation Count																																																				
Strong Mutator:	7.57	6.06	5.3	2.27	1.51	3.03	1.51	0.76	0.76	0	0	0	0	1.51	0.76	0.76	1.51	1.51	2.27	0.76	2.27	9.08	0.76	0.76	0.76	0.76	22	5.3	0.76	1.51	1.51	0	6.81	2.27	0.76	0.76	0.76	0.76	1.51	0.76	1.51	2.27	2.27	5.3	0.76	0.76	0	0	0.76	0.76	0	6.81
Medium Mutator:	1.06	1.19	1.19	0.92	1.06	0.92	0.92	1.06	0.79	0	0.13	0	0.26	1.06	0.66	0.92	1.19	0.92	1.06	1.06	1.45	2.38	0.4	0.92	0.92	1.06	1.85	0.4	0.26	1.06	0.66	0.26	0.66	0.26	0.79	0.26	1.06	0.92	0.92	0.92	0.92	0.92	0.92	0.79	0.26	0.66	0	0.4	1.19	0.79	0.26	0.92
Weak Mutator-1:	0.6	0.6	1.21	0.48	0.48	0.48	0.48	0.72	0.6	0	0	0.24	0.12	0.84	0.97	0.12	0.6	0.48	0.48	0.48	1.33	1.81	0.24	0.48	0.6	0.6	3.38	0	0.12	0.84	0.6	0.12	0.6	0.12	0.6	0	0.72	0.6	0.72	0.48	0.6	0.84	0.6	0.48	0	0.48	0	0	0.48	0.48	0	0.48
Weak Mutator-2:	0.84	0.98	0.7	0.84	0.84	0.84	0.7	0.7	1.12	0.28	0.14	0.14	0.14	0.84	0.56	0.28	0.84	0.7	0.84	0.84	1.26	1.68	0.28	0.84	0.84	0.7	1.96	0	0.42	0.7	0.56	0.14	0.56	0.14	0.56	0.14	0.84	0.7	0.7	0.7	0.84	0.7	0.7	0.56	0.14	0.84	0.28	0.28	0.7	0.56	0.14	0.84
AID-T7pol UGI 1:	8.32	5.55	25	2.77	0	0	2.77	2.77	0	2.77	0	0	0	0	2.77	0	0	0	0	0	0	8.32	2.77	0	0	0	69.3	11.1	5.55	8.32	0	0	5.55	8.32	2.77	5.55	0	0	11.1	0	0	25	8.32	2.77	0	0	0	0	0	0	0	19.4
AID-T7pol UGI 2:	3.5	3.5	31.5	10.5	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	90.9	6.99	0	3.5	0	0	6.99	14	6.99	0	0	0	3.5	0	0	17.5	3.5	3.5	0	0	0	0	0	0	0	0
T7pol:	0.16	0	0	0	0	0	0	0.16	0.31	0.16	0.16	0	0.16	0.31	0.16	0.16	0	0	0	0	0.94	1.73	0.47	0.31	0	0	1.57	0.16	0	0	0	0	0	0	0.16	0.16	0.31	0	0	0	0	0	0.16	0	0	0	0	0	0	0	0	0.47
DH5αGFP	0	0	0.17	0	0	0.33	0.17	0	0.17	0	0	0	0	0.33	0.5	0	0.17	0	0	0	0.5	2.33	0	0.17	0.17	0.17	1.5	0	0.17	0.17	0	0	0.33	0	0	0	0	0.17	0	0	0	0.17	0	0	0	0	0	0	0	0	0	0

Figure 5.5

Index Position:	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571
GFP Sequence:	T	A	T	A	T	T	T	T	T	C	A	A	A	G	A	T	G	A	C	G	G	G	A	A	C	T	A	C	A	A	G	A	C	A	C	G	T	G	C	T	G	A	A	G	T	C	A	A	G	T	T	T
Mutation Count																																																				
Strong Mutator:	0.76	2.27	0.76	0.76	0.76	0	0	0	0	0.76	0.76	0	0.76	0.76	0.76	0.76	0.76	0.76	1.51	3.03	0.76	0.76	0.76	0	1.51	1.51	0.76	0.76	4.54	0	0.76	0.76	0.76	1.51	4.54	3.03	1.51	1.51	2.27	1.51	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0	5.3	2.27	0.76	0
Medium Mutator:	1.06	1.19	0.92	0.92	0.79	0.13	0	0	0.26	1.06	0.79	0	0.4	0.92	0.92	0.92	0.92	0.92	0.92	0.66	0	0.26	0.66	0.26	1.06	0.92	1.06	1.19	1.06	0.79	1.06	0.92	0.92	0.92	1.19	0.92	1.06	1.58	1.32	0.92	0.92	0.66	0.53	1.06	1.06	1.06	0.79	0.4	1.06	0.79	0.13	0.4
Weak Mutator-1:	0.6	0.6	0.48	0.48	0.97	0.12	0.12	0.12	0	0.6	0.97	0.12	0.12	0.84	0.72	0.48	0.48	0.48	0.6	0.84	0	0.12	0.6	0	0.6	0.6	0.6	0.72	0.6	0.24	0.6	0.6	0.72	0.6	0.6	0.97	0.6	0.84	0.84	0.84	0.6	0.84	0.12	0.72	0.6	0.6	0.6	0.12	0.72	0.72	0	0
Weak Mutator-2:	0.84	0.84	0.98	0.7	0.84	0.28	0.14	0.28	0.28	0.84	0.7	0.14	0.28	0.98	0.98	0.98	0.84	0.7	0.7	0.84	0	0.14	0.56	0.28	1.12	0.84	0.7	0.7	0.56	0.14	1.12	0.84	0.84	0.84	0.7	0.7	0.7	0.84	0.84	0.84	0.84	0.56	0.14	0.84	0.7	0.7	0.7	0.28	0.84	0.56	0	0.14
AID-T7pol UGI 1:	2.77	0	0	0	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	5.55	2.77	0	2.77	0	0	8.32	2.77	0	0	0	5.55	0	0	0	0	11.1	5.55	2.77	0	11.1	0	2.77	0	0	0	0	0	0	0	19.4	0	0	0
AID-T7pol UGI 2:	3.5	0	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	3.5	0	10.5	0	0	14	0	0	6.99	0	0	0	0	0	0	10.5	10.5	14	3.5	10.5	0	0	0	0	0	0	0	0	0	14	0	0	0
T7pol:	0	0	0	0	0	0	0	0	0	0.16	0.16	0.16	0	0.31	0	0	0	0.16	0.16	0.31	0	0	0.31	0	0	0	0	0	0	0.16	0.47	0.16	0	0	0.16	0	0	0.16	0.31	0	0.16	0	0	0	0	0.16	0	0.16	0.47	0	0	0
DH5αGFP	0.17	0	0	0	0	0	0	0	0	0.17	0.33	0.67	0.17	0	0	0	0	0.17	0.67	0.33	0	0	0.17	0.17	0	0.17	0	0.17	0.33	0	0.67	0.17	0	0	0.17	0	0.17	0.67	0.33	0	0	0	0	0.17	0	0	0.17	0.33	0.17	0	0	0

GFP-mut3b

Index Position:	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623
GFP Sequence:	G	A	A	G	G	T	G	A	T	A	C	C	C	T	T	G	T	T	A	A	T	A	G	A	A	T	C	G	A	G	T	T	A	A	A	A	G	G	T	A	T	T	G	A	T	T	T	T	A	A	A	G
Mutation Count																																																				
Strong Mutator:	1.51	0.76	0.76	1.51	0.76	0.76	1.51	0.76	0.76	0.76	9.08	3.79	4.54	0.76	0	14.4	3.79	0	0.76	0	0.76	1.51	2.27	1.51	0	2.27	14.4	1.51	2.27	18.9	4.54	0.76	1.51	0	0	0	1.51	3.03	2.27	0.76	1.51	0	3.03	2.27	2.27	0	0	0.76	0.76	0	0	1.51
Medium Mutator:	1.06	0.66	0.4	0.92	0.26	1.45	0.92	0.92	0.92	0.92	0.66	0	0.26	0.66	0.26	0.92	0.66	0.26	0.92	0.26	1.06	1.06	1.32	0.92	0.26	1.06	1.32	1.19	0.92	0.92	0.79	0.26	0.79	0.13	0.13	0.53	0.92	0.53	1.06	0.92	0.66	0.4	0.92	0.92	0.79	0	0	0.26	0.66	0	0.13	1.32
Weak Mutator-1:	0.72	0.6	0.12	0.84	0	1.09	0.6	0.6	0.72	0.6	0.84	0	0	0.72	0	0.72	0.6	0	0.84	0.12	0.72	0.6	0.72	0.72	0.12	0.84	1.09	0.72	0.84	0.84	0.84	0	0.6	0	0	0	0.84	0	0.6	0.6	0.6	0	0.6	0.72	0.84	0.12	0.12	0.24	0.6	0	0	0.84
Weak Mutator-2:	0.98	0.56	0.14	0.7	0.14	0.84	0.7	0.7	0.7	0.98	0.98	0	0.28	0.84	0.14	0.84	0.7	0.14	0.7	0.14	0.84	0.84	0.7	0.56	0.14	0.7	0.98	0.84	0.7	1.12	0.98	0.28	0.7	0.14	0	0.28	0.56	0.14	0.7	0.7	0.56	0.28	0.7	0.7	0.84	0	0	0.14	0.56	0	0.14	0.98
AID-T7pol UGI 1:	0	0	0	2.77	0	0	0	0	0	0	16.6	0	0	0	0	30.5	11.1	0	0	0	0	0	0	0	0	0	19.4	5.55	0	41.6	5.55	0	0	0	0	0	0	11.1	5.55	2.77	0	0	16.6	0	0	0	0	0	2.77	0	0	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	0	0	0	21	3.5	6.99	0	0	35	3.5	0	0	0	0	0	0	0	0	0	28	3.5	0	48.9	10.5	6.99	0	0	0	0	0	6.99	0	0	0	0	10.5	0	0	0	0	0	0	0	0	0
T7pol:	0	0.16	0	0.31	0	0.31	0.16	0	0	0	0.16	0.16	0	0	0	0.31	0.16	0	0	0	0	0	0	0	0	0.16	0	0.16	0.16	0	0.31	0	0.16	0	0	0	0	0	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0.31
DH5αGFP	0.17	0	0.17	0.33	0	0.33	0	0	0.17	0	0.33	0	0	0	0	0.17	0	0	0	0	0	0	0	0.17	0	0.17	0.17	0.17	0.33	0.17	0	0	0	0	0	0	0.17	0	0	0	0.17	0	0	0	0	0.17	0	0.17	0	0	0.17	0

GFP-mut3b

Index Position:	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675
GFP Sequence:	A	A	G	A	T	G	G	A	A	A	C	A	T	T	C	T	T	G	G	A	C	A	C	A	A	A	T	T	G	G	A	A	T	A	C	A	A	C	T	A	T	A	A	C	T	C	A	C	A	C	A	A
Mutation Count																																																				
Strong Mutator:	0.76	0	0.76	0.76	0.76	0.76	0.76	0.76	0	0	3.03	0.76	0.76	0	0.76	0.76	0	3.79	0	0.76	1.51	3.79	1.51	1.51	0.76	0	0.76	0	0.76	1.51	0.76	0	0.76	0.76	12.1	3.79	0	4.54	1.51	1.51	0.76	0.76	0	25	4.54	3.79	0.76	0.76	3.03	6.06	3.03	1.51
Medium Mutator:	0.66	0.13	0.92	0.79	0.79	0.66	0.13	0.66	0	0.13	0.79	0.79	0.66	0.13	0.79	0.66	0.13	0.66	0.13	0.79	0.79	0.79	0.92	1.06	0	0.13	0.79	0.13	0.66	0.13	0.66	0.13	0.92	0.79	0.79	0.66	0.13	0.79	0.79	0.79	0.92	0.66	0.13	0.79	0.79	0.79	0.79	0.79	1.06	0.79	0.79	0.13
Weak Mutator-1:	0.72	0	0.6	0.6	0.6	0.6	0	0.72	0	0	0.72	0.6	0.6	0	0.6	0.6	0	0.6	0.12	0.6	0.6	0.6	0.6	0.6	0	0	0.97	0.12	0.6	0	0.6	0	0.6	0.72	0.72	0.6	0	0.72	0.72	0.72	0.6	0.6	0	0.72	0.6	0.72	0.72	0.72	0.97	0.97	0.84	0.12
Weak Mutator-2:	0.7	0.28	0.98	0.98	0.98	0.84	0.14	0.84	0	0.28	0.98	0.98	0.84	0.14	0.98	0.56	0.14	0.56	0.14	0.98	0.84	0.7	0.84	0.7	0.14	0.14	0.56	0.14	0.56	0.14	0.56	0.14	0.56	0.56	0.56	0.42	0.14	0.7	0.7	0.56	0.56	0.42	0.14	0.56	0.7	0.7	0.56	0.56	0.7	0.7	0.56	0.28
AID-T7pol UGI 1:	0	0	2.77	0	0	0	2.77	0	0	0	11.1	2.77	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	0	0	0	30.5	0	0	8.32	0	0	0	0	0	58.2	8.32	0	2.77	5.55	0	8.32	0	2.77
AID-T7pol UGI 2:	0	0	6.99	0	0	0	3.5	0	0	0	6.99	6.99	0	3.5	0	0	0	0	0	0	3.5	0	3.5	3.5	0	0	0	0	0	3.5	0	0	0	0	28	3.5	0	6.99	0	0	0	0	0	90.9	3.5	6.99	3.5	0	0	10.5	3.5	0
T7pol:	0.31	0	0	0	0.16	0.31	0.16	0	0	0	0.16	0	0	0.16	0	0	0	0	0	0	0.31	0.16	0.16	0	0	0	0	0.16	0.16	0	0	0	0	0.16	0	0.16	0.16	0.31	0	0	0	0	0	0.16	0	0	0	0	0.16	0.16	0	0
DH5αGFP	0	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0.17	0	0	0	0.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0	0	0.17	0.17	0

GFP-mut3b

Index Position:	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727
GFP Sequence:	T	G	T	A	T	A	C	A	T	C	A	T	G	G	C	A	G	A	C	A	A	A	C	A	A	A	A	G	A	A	T	G	G	A	A	T	C	A	A	A	G	T	T	A	A	C	T	T	C	A	A	A
Mutation Count																																																				
Strong Mutator:	0.76	24.2	1.51	3.79	0.76	0.76	5.3	1.51	1.51	1.51	0.76	0.76	0.76	1.51	2.27	3.79	1.51	0.76	1.51	0.76	0	0	5.3	3.03	1.51	0	0	0.76	0.76	0	0.76	2.27	3.79	0	0	0	0.76	0.76	0	0	0	0	0	0	0	5.3	0	0	1.51	0.76	0.76	0
Medium Mutator:	0.92	0.79	0.79	0.79	0.79	0.92	0.79	0.92	0.92	0.92	1.06	0.79	0.66	1.32	0.92	0.79	0.79	1.06	0.79	0.92	0	0.26	0.79	0.79	0.13	0	0.13	0.79	0.66	0.13	0.92	0.66	0.13	0.66	0.13	0.79	0.79	0.66	0	0.26	1.06	0.79	0.26	0.66	0.13	0.79	0.66	0.13	0.79	1.06	0.26	0
Weak Mutator-1:	0.6	0.84	0.72	0.72	0.6	0.6	0.72	0.72	0.84	0.6	0.72	0.6	0.72	0.72	0.6	0.84	0.6	0.84	0.72	0.72	0	0	0.72	0.97	0.12	0.12	0.24	0.97	0.72	0	0.72	0.72	0	0.72	0	0.72	0.72	0.84	0	0	0.72	0.84	0	0.84	0	0.72	0.84	0.12	0.97	1.09	0.12	0
Weak Mutator-2:	0.7	0.7	0.56	0.7	0.7	0.7	0.56	0.56	0.84	0.56	0.84	0.84	0.56	0.56	0.98	0.56	0.7	0.7	0.7	0.42	0.14	0.42	0.84	0.42	0	0	0.14	0.42	0.28	0.14	0.42	0.42	0.14	0.28	0.14	0.42	0.42	0.28	0	0.14	0.56	0.28	0.14	0.28	0.14	0.42	0.42	0.14	0.42	0.28	0	0
AID-T7pol UGI 1:	0	13.9	5.55	0	0	0	22.2	2.77	0	5.55	0	0	0	2.77	0	0	2.77	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	2.77	2.77	0	0	0	2.77	0	0	0	8.32	0	0	0	0	13.9	0	0	5.55	0	0	0
AID-T7pol UGI 2:	0	21	3.5	0	0	0	24.5	6.99	3.5	3.5	0	0	0	0	0	0	0	0	0	0	0	0	6.99	0	0	0	0	6.99	0	0	0	0	3.5	0	0	0	3.5	0	0	0	0	0	0	0	0	10.5	3.5	0	0	0	0	0
T7pol:	0	0	0	0.31	0	0	0.16	0	0	0	0	0.16	0.16	0.47	0.16	0.16	0	0	0	0	0	0	0	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0	0.16	0	0
DH5αGFP	0	0.33	0	0	0	0.17	0.33	0	0	0.17	0	0	0.17	0.5	0.17	0.33	0.17	0	0	0.17	0.17	0	0	0	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0	0	0	0.17	0	0

GFP-mut3b

Index Position:	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779
GFP Sequence:	A	T	T	A	G	A	C	A	C	A	A	C	A	T	T	G	A	A	G	A	T	G	G	A	A	G	C	G	T	T	C	A	A	C	T	A	G	C	A	G	A	C	C	A	T	T	A	T	C	A	A	C
Mutation Count																																																				
Strong Mutator:	0	0	0	0	0	0	0.76	0	0	0.76	0	3.79	3.03	0	0	0.76	0	0	0	0	0	0	0	0	0.76	2.27	104	31.8	10.6	1.51	0.76	0.76	0.76	3.03	0	0.76	0	6.06	1.51	0.76	0	0.76	0.76	2.27	0	0	0	0	0.76	0	0	2.27
Medium Mutator:	0.13	0.66	0.13	0.79	0.79	1.06	0.92	0.79	0.92	0.66	0.13	0.79	0.79	0.66	0.13	0.79	0.92	0.13	0.92	0.79	0.79	0.66	0.13	0.92	0.26	0.92	0.79	0.92	0.92	0.26	0.92	0.79	0.13	0.79	0.92	0.79	1.06	0.92	0.79	0.79	1.06	0.79	0.13	0.92	0.79	0.26	0.79	0.79	0.79	0.79	0.13	0.79
Weak Mutator-1:	0	0.72	0	0.84	0.72	0.84	0.72	0.72	0.72	0.84	0.12	0.84	0.72	0.84	0	0.84	0.97	0.12	0.72	0.84	0.84	0.72	0	0.97	0.24	0.97	1.09	0.97	0.84	0	0.72	0.97	0.24	0.72	0.72	0.97	1.21	0.97	0.72	0.84	0.72	0.97	0	0.72	0.97	0.24	0.84	0.72	0.84	0.84	0	0.97
Weak Mutator-2:	0.14	0.28	0.14	0.42	0.42	0.7	0.42	0.7	0.42	0.28	0.14	0.56	0.42	0.28	0.14	0.42	0.28	0.14	0.42	0.42	0.42	0.28	0.28	0.42	0.28	0.56	1.12	0.84	0.56	0.14	0.7	0.56	0.28	0.56	0.84	0.56	0.98	1.26	0.84	0.84	0.98	0.56	0.14	0.56	0.42	0.14	0.7	0.56	0.56	0.42	0.14	0.56
AID-T7pol UGI 1:	0	0	0	0	0	2.77	0	0	5.55	0	0	13.9	2.77	0	0	2.77	0	0	0	0	0	0	0	0	0	0	191	44.4	11.1	2.77	0	0	0	2.77	0	0	2.77	16.6	0	0	0	0	0	0	0	0	0	0	5.55	0	0	5.55
AID-T7pol UGI 2:	0	0	0	0	0	0	3.5	0	0	0	0	3.5	0	0	0	3.5	0	0	0	0	0	3.5	0	0	0	0	157	38.5	14	3.5	3.5	0	0	0	0	0	3.5	21	10.5	0	0	0	10.5	6.99	0	0	0	0	3.5	0	0	3.5
T7pol:	0	0.16	0	0	0	0	0	0.16	0	0	0.16	0.47	0	0	0	0.31	0	0.16	0	0	0	0	0	0	0	0.16	0	0.16	0.16	0.31	0	0.31	0.31	0.47	0.16	0.16	0.16	0	0	0.16	0.16	0.16	0.16	0.31	0.16	0	0.16	0.16	0	0	0.16	0.16
DH5αGFP	0	0	0	0	0.17	0.5	0.17	0	0.17	0	0.17	0	0	0.17	0	0.17	0.33	0.17	0.5	0	0	0	0.17	0.33	0.17	0.17	0	0	0.17	0	0	0	0	0.17	0	0	0.83	0	0.17	0	0	0	0	0	0.33	0	0	0	0	0	0	0.17

Figure 5.5

Index Position:	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831
GFP Sequence:	A	A	A	A	T	A	C	T	C	C	A	A	T	T	G	G	C	G	A	T	G	G	C	C	C	T	G	T	C	C	T	T	T	T	A	C	C	A	G	A	C	A	A	C	C	A	T	T	A	C	C	T
Mutation Count																																																				
Strong Mutator:	2.27	1.51	0.76	0	0	0	9.08	0.76	0.76	0	0	0	0	0	0	0	4.54	0.76	0	0	0	0	0.76	0.76	0	0	0	0	0	0	0	0	0	0	0	20.4	14.4	2.27	3.03	0	3.03	1.51	1.51	1.51	1.51	0.76	0	0	0	4.54	2.27	0
Medium Mutator:	0.92	0.26	0	0.26	0.92	0.92	0.92	0.92	0.79	0.26	0.92	0.13	0.66	0.13	0.66	0.13	0.79	0.79	0.79	1.06	0.79	0.26	2.24	0.4	0.26	0.92	1.06	1.06	0.92	0.13	1.32	0.13	0	0.13	0.92	0.92	0.13	1.06	1.19	1.19	0.92	0.79	0.13	0.92	0.13	1.06	1.06	0.26	1.06	0.92	0.13	0.92
Weak Mutator-1:	0.84	0.12	0	0	0.84	0.72	0.72	0.84	0.84	0.12	0.72	0	0.84	0.12	0.84	0	1.33	1.33	0.72	0.72	0.72	0	2.78	0.24	0.6	1.09	0.72	0.72	0.72	0	0.97	0	0.12	0.12	0.84	0.97	0	0.84	0.72	0.72	0.72	0.72	0	0.97	0.12	0.72	0.72	0	0.72	0.72	0.12	0.72
Weak Mutator-2:	0.7	0	0	0.14	0.98	0.7	0.7	0.56	0.42	0.14	0.42	0.28	0.42	0.14	0.42	0.14	1.26	1.12	0.7	0.7	0.7	0.28	1.4	0.28	0.7	0.7	0.7	0.56	0.42	0.28	0.56	0	0	0.14	0.7	0.56	0.28	0.56	0.56	0.84	0.7	0.56	0.14	0.56	0.42	0.56	0.56	0.14	0.56	0.56	0	0.56
AID-T7pol UGI 1:	0	0	0	0	0	0	8.32	0	0	0	0	0	0	0	0	0	5.55	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0	0	0	38.8	16.6	2.77	0	0	8.32	0	0	2.77	0	0	0	0	0	0	2.77	2.77
AID-T7pol UGI 2:	6.99	0	0	0	0	0	3.5	3.5	3.5	3.5	0	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	3.5	0	0	0	0	0	28	6.99	0	0	0	6.99	0	0	6.99	3.5	0	0	0	0	14	3.5	0
T7pol:	0.16	0	0	0	0	0	0.16	0.16	0.16	0	0	0	0.16	0	0.16	0	0.47	0	0.16	0	0.31	0.47	1.89	0	0.47	0.31	0.63	0	0	0.16	0.16	0	0	0	0.16	0.16	0	0	0.16	0	0	0	0	0.16	0	0	0	0.16	0.16	0.16	0.31	0
DH5αGFP	0	0	0	0	0	0	0	0	0.33	0.17	0	0	0	0	0	0.17	0.83	0.33	0	0.33	0	0.17	2	0.33	0.33	0.5	0	0	0	0.17	0	0	0.17	0.17	0.33	0.17	0.33	0.33	0.17	0.17	0	0	0	0	0	0	0	0	0	0	0	0

																										GFP-mut3b																										
Index Position:	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883
GFP Sequence:	G	T	C	C	A	C	A	C	A	A	T	C	T	G	C	C	C	T	T	T	C	G	A	A	A	G	A	T	C	C	C	A	A	C	G	A	A	A	A	G	A	G	A	G	A	T	C	A	C	A	T	G
Mutation Count																																																				
Strong Mutator:	0.76	0	0	0.76	0.76	2.27	0	2.27	0.76	0	0	4.54	0	0.76	0.76	0	0.76	0	0	0	0.76	0	0	0.76	0	0	0	0	0	25	8.33	6.06	0	12.1	6.06	2.27	1.51	0	0	0	0	0	0	0.76	0	0	0.76	0	3.03	3.03	0	0
Medium Mutator:	0.92	0.92	0.79	0.13	1.06	1.19	0.92	1.19	0.79	0.13	1.06	1.06	0.92	1.19	1.98	0.26	0.13	0.92	0	0.26	1.06	1.19	0.92	0.13	0.26	1.06	0.92	0.92	1.19	0.13	0.13	0.79	0.13	1.19	1.06	1.06	0.13	0	0.13	1.06	0.92	1.06	0.92	0.92	0.92	0.92	1.06	0.92	1.06	1.06	0.92	0.92
Weak Mutator-1:	0.72	0.84	0.84	0.24	0.97	1.09	0.84	0.72	0.84	0	0.84	0.72	0.84	0.97	1.57	0.24	0.12	0.97	0.24	0.24	0.84	0.72	0.97	0	0	0.84	0.84	0.84	1.33	0	0	0.84	0.12	0.84	0.84	0.97	0	0	0	0.84	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.84	0.84	0.72	0.72
Weak Mutator-2:	0.7	0.7	0.56	0	0.7	0.56	0.7	0.7	0.56	0.14	0.7	0.56	0.56	0.7	1.4	0	0.14	0.56	0	0.14	0.98	0.98	1.26	0.28	0.42	0.84	0.7	0.7	0.84	0.28	0	0.7	0	0.7	0.7	1.12	0.14	0	0.14	0.98	0.7	0.84	0.7	0.7	0.7	0.7	0.7	0.7	0.84	0.7	0.7	0.84
AID-T7pol UGI 1:	0	0	0	8.32	0	0	0	2.77	0	0	0	5.55	0	0	2.77	0	0	0	0	0	0	2.77	0	0	0	0	0	0	2.77	33.3	13.9	0	0	30.5	13.9	2.77	0	0	0	0	0	0	0	0	0	0	0	0	5.55	0	2.77	0
AID-T7pol UGI 2:	0	0	0	17.5	0	0	0	0	0	3.5	0	28	0	0	0	3.5	0	0	0	0	3.5	0	0	0	0	0	0	0	0	24.5	6.99	0	0	35	3.5	0	3.5	0	0	0	0	0	0	6.99	0	0	0	0	0	0	0	0
T7pol:	0	0	0	0.16	0.31	0	0	0	0	0	0	0	0	0.31	0.94	0	0	0.16	0	0.16	0.16	0	0	0	0	0	0	0	0.16	0.16	0	0	0	0.31	0	0	0	0	0	0	0	0	0	0.16	0	0	0	0	0.16	0.16	0.16	0
DH5αGFP	0	0	0	0	0.17	0	0.17	0	0	0	0	0.17	0.17	0.5	1.66	0.33	0	0.17	0.17	0.17	0	0	0.17	0.33	0	0.17	0	0	0.17	0.17	0.33	0	0	0.17	0	0	0	0	0	0	0	0.17	0	0.33	0	0.17	0	0	0.33	0	0	0

																										GFP-mut3b																										
Index Position:	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935
GFP Sequence:	G	T	C	C	T	T	C	T	T	G	A	G	T	T	T	G	T	A	A	C	A	G	C	T	G	C	T	G	G	G	A	T	T	A	C	A	C	A	T	G	G	C	A	T	G	G	A	T	G	A	A	C
Mutation Count																																																				
Strong Mutator:	0	0	0.76	0	0	0	0	0	0	0	0	1.51	2.27	0	0	1.51	0.76	0	0	10.6	0.76	0.76	9.08	0.76	0.76	1.51	0	0	5.3	2.27	0.76	0	0.76	0	0.76	0	0.76	0.76	0	0	0	2.27	0.76	0	0	0	0	0	0.76	0	0	0.76
Medium Mutator:	0.13	0.92	0.79	0.13	0.92	0.13	0.92	0.79	0.26	0.92	0.92	0.92	0.79	0.13	0.13	1.19	1.06	0.92	0.13	0.92	0.92	1.06	1.06	0.92	0.92	1.06	1.19	1.06	0.13	0.13	0.92	0.79	0.13	1.06	1.06	1.06	0.92	1.19	0.92	0.79	0.79	1.06	1.06	0.92	0.79	0.13	0.92	0.92	1.19	0.92	0.13	1.32
Weak Mutator-1:	0	0.72	0.84	0	0.84	0	0.97	0.84	0	0.97	0.84	0.97	0.97	0.12	0.36	1.57	1.09	1.21	0.24	0.97	0.97	0.84	0.84	0.84	0.97	0.84	1.21	0.97	0.12	0.12	0.97	0.84	0	0.97	0.97	0.84	1.09	0.84	0.84	1.09	1.21	0.97	1.09	0.84	0.84	0.12	0.84	0.97	0.97	1.09	0	1.33
Weak Mutator-2:	0	0.98	0.84	0	0.98	0.14	0.98	0.84	0	0.7	0.7	0.84	0.7	0	0	0.98	0.84	0.84	0.28	1.12	0.7	0.7	0.7	0.7	0.98	0.7	0.7	0.7	0	0.14	0.7	0.84	0.14	1.12	0.84	0.84	0.7	0.7	0.7	0.7	1.26	0.84	0.7	0.7	0.84	0	0.84	0.7	0.7	0.7	0	0.7
AID-T7pol UGI 1:	0	0	0	0	0	0	2.77	0	0	0	0	16.6	0	0	0	2.77	0	0	0	25	0	0	8.32	2.77	0	0	2.77	0	2.77	2.77	0	0	0	0	2.77	0	0	2.77	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0
AID-T7pol UGI 2:	3.5	0	0	0	0	0	0	0	0	0	0	10.5	0	0	0	3.5	0	0	0	10.5	10.5	3.5	6.99	0	0	3.5	0	0	0	0	3.5	0	0	0	3.5	0	6.99	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0
T7pol:	0	0.16	0	0	0.16	0.47	0.47	0	0.16	0.16	0.16	0.31	0	0	0	0.16	0	0	0.31	0	0.16	0.16	0.16	0	0.16	0	0.47	0.16	0	0.31	0.16	0	0	0.16	0	0	0.31	0.16	0.16	0.16	0.31	0.16	0.16	0	0	0	0	0	0.16	0	0	0
DH5αGFP	0	0	0	0	0	0.17	0.17	0.17	0	0.17	0	0	0.17	0.17	0	0.17	0.33	0.17	0	0	0	0	0	0	0.17	0	0.17	0.33	0	0	0	0.17	0	0	0.17	0.33	0.17	0.17	0	0	1.5	0.67	0	0.17	0.17	0.17	0.17	0	0.17	0	0	0.33

						GFP-mut3b																																													
Index Position:	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987
GFP Sequence:	T	A	T	A	C	A	A	A	T	A	A	G	G	C	T	C	G	A	T	C	G	G	T	G	T	G	A	A	A	A	G	T	C	A	G	T	A	T	C	C	A	G	T	C	G	T	G	T	A	G	T	T
Mutation Count																																																				
Strong Mutator:	0	0	0	0	69.6	9.08	3.79	2.27	0	0	0	0	0	1.51	0	0	0	0	0	0.76	0	0.76	0	1.51	1.51	1.51	0	0	0	0	0	0	0.76	0	0	0	0	0	0	2.27	0.76	0	0	0	0	0.76	3.03	0.76	0	3.79	2.27	0.76
Medium Mutator:	0.92	1.06	0.92	1.06	0.92	0.79	0	0.13	1.19	0.79	0.13	0.92	0.4	1.06	0.92	0.92	0.92	0.92	0.92	1.19	0.79	0.26	1.06	1.19	0.92	0.92	0.92	0	0	0.13	1.32	0.92	0.92	0.92	1.06	1.19	1.06	0.92	0.79	0.4	1.06	0.92	0.92	0.92	1.06	1.06	1.06	0.92	0.92	1.06	0.79	0.13
Weak Mutator-1:	0.84	0.97	0.97	1.09	1.09	1.09	0.24	0.24	1.09	1.09	0	0.97	0	1.33	0.97	1.09	1.09	1.09	1.09	1.33	0.97	0	1.33	1.21	1.21	1.09	1.21	0.24	0	0	1.09	1.09	0.97	1.09	0.97	1.09	0.97	1.09	0.97	0.12	1.09	1.09	0.97	0.97	0.97	0.97	0.97	1.09	1.09	0.97	0.97	0
Weak Mutator-2:	0.7	0.7	0.84	0.7	0.84	0.84	0.14	0.14	0.98	0.7	0.14	0.98	0.14	0.7	0.98	0.84	0.7	0.98	0.7	0.84	0.98	0.14	1.54	0.98	0.98	0.84	1.26	0.28	0	0	0.7	0.7	0.84	0.7	0.7	0.7	0.84	0.84	0.7	0.14	0.84	0.84	0.84	0.7	0.84	0.7	0.7	0.84	0.7	0.84	0.84	0
AID-T7pol UGI 1:	0	0	0	0	108	25	5.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	0	5.55	0	0	0	0	0	2.77	0	0	0	0	0	5.55	0	0	0	5.55	0
AID-T7pol UGI 2:	0	0	0	0	83.9	24.5	10.5	0	0	0	0	0	0	0	0	0	3.5	3.5	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	6.99	3.5	0	0	0	0	0	0	0	0	0	0	6.99	3.5	0	6.99	0	0
T7pol:	0	0	0	0.31	0	0	0	0	0	0	0.16	0	0	0	0	0	0.16	0.16	0.16	0	0	0	0	0.16	0.16	0	0	0	0	0	0	0	0.16	0.16	0	0	0	0	0.31	0.31	0.16	0	0	0	0	0	0.16	0	0	0.16	0	0
DH5αGFP	0	0.17	0	0	0	0	0	0	0	0.17	0	0.17	0	0.67	0.33	0	0	0.33	0.17	0	0.33	0.17	0	0.17	0.17	0	0.17	0	0	0.17	0.17	0	0	0	0	0.33	0	0	0	0	0	0	0	0	0	0.17	0.17	0	0	0	0	0

																																												T7 Terminator								
Index Position:	988	989	990	991	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023	1024	1025	1026	1027	1028	1029	1030	1031	1032	1033	1034	1035	1036	1037	1038	1039
GFP Sequence:	C	T	T	A	T	T	A	C	C	T	G	T	C	C	C	C	T	A	G	C	A	T	A	A	C	C	C	C	G	C	G	G	G	G	C	C	T	C	T	T	C	G	G	G	G	G	A	C	T	C	G	C
Mutation Count																																																				
Strong Mutator:	0	0	0	0	0	0	0	2.27	1.51	0.76	0.76	0	0	3.03	0	0	0.76	1.51	0.76	3.79	0.76	1.51	1.51	0	9.08	6.06	3.03	0.76	1.51	1.51	2.27	0	0	0	0.76	0	0.76	0.76	0.76	0	6.81	7.57	0	0	0	0	0	0	0	0	0	2.27
Medium Mutator:	0.92	0.79	0.13	1.19	0.79	0.4	0.92	0.92	0.13	1.19	0.92	0.92	0.79	0	0	0.13	1.19	1.06	1.45	1.19	1.19	1.06	0.79	0.13	0.92	0	0	0.13	1.45	0.92	1.32	0.13	0	0.26	1.06	0.4	0.79	0.79	0.92	0	0.66	9.77	0	0.13	0	0	0.26	0.13	0.13	0.13	0.26	0.66
Weak Mutator-1:	1.09	1.21	0.24	1.09	1.09	0	1.21	1.21	0	0.97	0.97	1.09	1.33	0.12	0	0	1.21	0.97	1.57	1.09	1.21	1.09	0.97	0	1.21	0.12	0.12	0	1.57	0.97	1.57	0.12	0	0.6	1.21	0	1.09	1.09	1.21	0	0.97	8.33	0.12	0.24	0	0	0.24	0.12	0.12	0.12	0.36	0.36
Weak Mutator-2:	1.12	1.12	0.28	0.98	0.84	0.14	0.98	1.26	0.14	0.98	1.12	1.12	0.98	0	0.14	0	0.84	0.84	1.54	0.98	0.84	0.84	0.84	0	1.26	0.28	0	0	1.26	0.84	1.4	0	0	0.14	0.98	0	0.98	0.84	0.84	0.14	0.98	9.78	0.28	0.42	0.28	0.56	0.84	0.56	0.56	0.56	0.84	0.7
AID-T7pol UGI 1:	0	0	0	0	0	0	0	2.77	5.55	0	0	0	0	8.32	0	0	0	0	0	8.32	0	0	0	0	8.32	2.77	0	0	2.77	0	0	0	0	0	2.77	0	0	0	0	0	5.55	2.77	0	0	0	0	0	0	0	0	0	2.77
AID-T7pol UGI 2:	0	0	0	0	0	0	0	10.5	3.5	0	0	0	0	10.5	0	0	0	0	6.99	3.5	0	0	0	0	17.5	0	10.5	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	3.5
T7pol:	0.16	0	0.16	0.16	0	0	0.16	0.16	0.31	0.16	0	0	0	0	0	0.16	0.47	0.31	0.31	0	0.31	0.16	0	0.16	0.63	0.16	0	0	0.16	0	0.16	0	0	0.16	0	0	0	0.16	0	0	0	10.1	0.31	0.16	0	0	0.31	0.31	0	0	0.16	0.16
DH5αGFP	0	0	0	0.17	0.17	0.17	0.5	0.5	0	0	0.17	0	0.5	0	0	0	0	0	0.5	0.17	0	0.33	0	0	0.67	0.33	0	0	0	0.33	0.33	0	0	0.17	0.83	0	0.17	0	0	0	0.17	7.15	0.33	0	0.17	0.17	0.17	0	0	0	0.17	0.33

Figure 5.5

Index Position:	1040	1041	1042	1043	1044	1045	1046	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056	1057	1058	1059	1060	1061	1062	1063	1064	1065	1066	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076	1077	1078	1079	1080	1081	1082	1083	1084	1085	1086	1087	1088	1089	1090	1091
GFP Sequence:	G	G	G	G	T	T	T	T	T	T	G	C	T	G	A	A	A	G	A	A	T	T	A	T	C	A	A	A	T	A	A	A	A	C	G	A	A	A	G	G	C	T	C	A	G	T	C	G	A	A	A	G
Mutation Count																																																				
Strong Mutator:	0	0.76	0	1.51	1.51	0	0.76	0	0	0	0.76	5.3	3.79	0	0	0	0	0	0	0	0	0	0	0	0.76	0.76	0	0	0	0	0	0	0	3.03	1.51	0	0	0	0	1.51	0.76	0	0	0.76	0.76	0	0	0	0	0	0	0
Medium Mutator:	0.53	0.13	0.13	0	0.26	0.13	0.13	0	0.13	0	1.32	1.06	0.66	0	0	0	0	0	0.13	0.13	0.13	0.13	0	0	0	0.13	0	0.13	0.13	0.13	0	0	0.13	0.13	0	0.13	0.4	0	0	0.13	0.13	0.26	0.13	0.13	0.26	0.4	0	0	0.26	0.13	0.26	0.66
Weak Mutator-1:	0.24	0.12	0	0	1.93	0	0	0.12	0.12	0	0.97	0.72	0.12	0.24	0.12	0	0	0	0	0	0.24	0.12	0.12	0.12	0	0.12	0	0.12	0.12	0.12	0	0	0	0.24	0	0	0	0	0.12	0.12	0.12	0.12	0.12	0.12	0	0.12	0	0	0	0	0.12	0.6
Weak Mutator-2:	0.84	0.14	0	0.14	2.1	0.28	0.14	0	0	0.14	1.26	1.68	0.56	0.42	0.28	0	0.14	0.42	0.56	0.14	0.28	0.14	0.56	0.42	0.42	0.7	0.14	0.28	0.7	0.28	0	0	0.14	0.56	0.56	0.28	0.14	0.42	0.28	0.28	0.42	0.7	0.56	0.7	0.56	0.42	0.42	0.42	0.28	0.14	0.14	0.56
AID-T7pol UGI 1:	0	0	0	5.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	5.55	2.77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	6.99	6.99	0	0	0	0	3.5	0	0	0	0	3.5	0	0	0	0	0	0	0
T7pol:	0	0.16	0.16	0	1.26	0.16	0	0	0	0	0.94	0.47	0.16	0	0	0	0	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0	0	0	0	0	0.16	0	0	0	0	0	0	0	0	0	0	0	0.31
DH5αGFP	0	0	0	0	1.5	0.17	0	0	0	0	1.5	0.33	0	0	0	0	0	0.17	0.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0	0	0	0.33	0	0.5	0.17	0	0.17	0	0.17	0.17	0.17	0.17	0	0	0	0.33

T7 Terminator

Index Position:	1092	1093	1094	1095	1096	1097	1098	1099	1100	1101	1102	1103	1104	1105	1106	1107	1108	1109	1110	1111	1112	1113	1114	1115	1116	1117	1118	1119	1120	1121	1122	1123	1124	1125	1126	1127	1128	1129	1130	1131	1132	1133	1134	1135	1136	1137	1138	1139	1140	1141	1142	1143
GFP Sequence:	A	C	T	G	G	G	C	C	T	T	T	C	G	T	T	T	T	A	T	C	T	G	T	T	G	T	T	T	G	T	C	G	C	T	G	C	A	T	T	A	C	T	A	G	C	A	T	A	A	C	C	C
Mutation Count																																																				
Strong Mutator:	0	0	0	0	0	0	0.76	0	0	0	0	4.54	5.3	3.03	0	0	0	0	0	0	1.51	0.76	0	0	0.76	0	0.76	0	0	0	0	0	0.76	0	3.03	0	0	0	0	0	0.76	0	0.76	1.51	0	0	0	0	0	8.33	4.54	0.76
Medium Mutator:	0	0	0.13	0.13	0	0.26	0	0.13	0	0	0	0	0	0.66	0.13	0	0	0.13	0	0	0.26	0	0.13	0.4	0.4	0.13	0	0	0	0	0	0.13	0	0.79	1.06	0	0.4	0.13	0	0	0	0.13	0.26	0.53	0.13	0.13	0.13	0.13	0	0	0	0
Weak Mutator-1:	0.12	0.24	0.36	0	0.12	0.24	0.12	0	0	0	0	0	0.24	0	0	0	0	0	0.36	0	0.12	0	0	0	0.12	0	0	0	0.12	0	0	0	0	0.36	0.48	0	0.12	0	0.12	0	0	0	0.24	0.24	0.12	0	0	0	0	0.12	0	0
Weak Mutator-2:	0.56	0.42	0.56	0.42	0	0.28	0.42	0.14	0.42	0	0.14	0.7	0.56	0.42	0	0	0.14	0.28	0.28	0.42	0.56	0.42	0.14	0.14	0.28	0.14	0	0.28	0.56	0.56	0.28	0.28	0.42	0.98	0.84	0.42	0.28	0.14	0.14	0.56	0.28	0.56	0.28	1.4	0.28	0.56	0.42	0.14	0.14	0.28	0	0
AID-T7pol UGI 1:	0	0	0	0	0	0	0	0	0	0	0	0	13.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.77	0	0	0	2.77	0	0	0	0	0	0	0	0	0	0	0	2.77	0	2.77	0	0	0	11.1	5.55	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	3.5	0	0	0	0	17.5	0	0	0	0	0	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10.5	0	0
T7pol:	0	0	0	0.16	0	0.63	0.16	0	0	0	0	0.16	0	0	0	0	0	0.16	0.16	0.16	0	0	0	0	0	0	0	0.16	0.16	0.16	0	0.16	0	0	0.47	0	0	0	0	0	0	0	0	0.79	0.16	0	0.16	0	0	0.16	0	0
DH5αGFP	0	0	0	0	0.17	0	0.17	0	0	0	0	0	0	0.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0.17	0.17	0.67	0	0.5	0.17	0.17	0	0.17	0	0	0.17	0.33	0	0.83	0.17	0	0	0	0.17	0.17	0	0

T7 terminator **Suffix Index**

Index Position:	1144	1145	1146	1147	1148	1149	1150	1151	1152	1153	1154	1155	1156	1157	1158	1159	1160	1161	1162	1163	1164	1165	1166	1167	1168	1169	1170	1171	1172	1173	1174	1175	1176	1177	1178	1179	1180	1181	1182	1183	1184	1185	1186	1187	1188	1189	1190	1191	1192	1193	1194	1195
GFP Sequence:	C	T	T	G	G	G	G	C	C	T	C	T	A	A	A	C	G	G	G	T	C	T	T	G	A	G	G	G	G	T	T	T	T	T	T	G	G	G	C	T	C	G	G	A	G	A	T	C	G	A	C	A
Mutation Count																																																				
Strong Mutator:	0	0.76	0	0	0	0	0	0	0	0	0	0	0	0.76	0	15.9	4.54	2.27	0	0.76	0	0	0	0	0	0	0	0.76	2.27	4.54	0.76	0	0	0.76	0	0	1.51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Medium Mutator:	0	0.4	0.13	0.4	0.26	0	0.26	0.26	0.13	0	0.26	0	0	0	0	0.13	0.26	0.26	0.26	0.4	0	0.13	0.13	0.13	0.13	0	0	0.13	0	3.17	0	0.13	0	0.13	0.4	1.19	1.06	0.13	0.4	0	0	0.26	0	0	0	0	0	0	0	0	0.66	0.66
Weak Mutator-1:	0	0.36	0	0.48	0.24	0.12	0	0	0	0.12	0	0	0.12	0	0.24	0.12	0.36	0.24	0.36	0.12	0.12	0.12	0	0.12	0	0	0	0.12	0	2.05	0	0	0	0	0.24	1.57	0.48	0.12	0.24	0	0.12	0.84	0	0.12	0	0	0	0	0	0	0	0
Weak Mutator-2:	0.14	0.14	0.14	0.28	0	0.14	0.14	0.42	0.14	0.42	0.42	0.28	0.28	0	0.28	0.42	0.42	0	0.28	0.28	0.28	0.14	0.14	0.28	0.42	0.28	0	0	0.28	2.1	0	0	0	0	0.42	0.98	0.42	0.42	0.56	0.42	0.56	0.84	0.42	0.42	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28
AID-T7pol UGI 1:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19.4	2.77	2.77	2.77	0	0	0	0	0	0	0	0	0	5.55	0	0	0	0	0	0	2.77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AID-T7pol UGI 2:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24.5	10.5	0	0	0	0	0	0	0	0	0	0	3.5	3.5	0	0	0	0	0	0	3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T7pol:	0	0.16	0	0.16	0	0	0	0	0.31	0.31	0.16	0	0	0	0	0	0.16	0.16	0	0	0	0	0	0	0	0.16	0	0	0	2.52	0	0	0	0	0	1.57	0.16	0.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH5αGFP	0.17	0	0	0.5	0	0.17	0.17	0.33	0.17	0.33	0	0	0.17	0.17	0	0	0	0	0	0	0.17	0.67	0	0	0	0.33	0	0.17	0.17	2.99	0	0	0	0.17	0	1.16	1.16	0	0	0	0	0.33	0	0	0	0	0	0	0	0	0	0

Index Position:	1196	1197	1198	1199	1200	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210
GFP Sequence:	G	T	C	T	C	G									
Mutation Count															
Strong Mutator:	19.7	22.7	3.79	0.76	0.76	0.76	8.33	5.3	0.76	0.76	0.76	0	0	0	0
Medium Mutator:	0.66	0.66	0	0	0	0.4	5.28	1.45	1.58	0.79	0.4	0.26	0.26	0.13	0
Weak Mutator-1:	0	0	0	0.72	0.72	0	4.71	2.53	1.69	1.33	1.09	0.72	0.6	0.48	0.12
Weak Mutator-2:	0.28	0.28	0.28	0.28	0.7	1.4	3.63	1.68	0.7	0.84	0.42	0.42	0.14	0	0
AID-T7pol UGI 1:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AID-T7pol UGI 2:	0	0	0	0	0	0	3.5	3.5	3.5	0	0	0	0	0	0
T7pol:	0	0	0	0	0	0.31	4.72	1.57	1.26	0.94	0.47	0.47	0.31	0	0
DH5αGFP	0	0	0	0	0.17	0.17	4.99	1.83	1.16	1.16	1	0.5	0.33	0.17	0

Figure 5.5*: The mutation spread across the GFP-mut3b expression cassette produced from PacBio sequencing reads. The frequency of a nucleotide substitution or an indel occurring at each index position (x10^{-6} per sequenced base-pair) in the sequenced reads is represented in this heat map (Colour code: Red = Highest, Green = Lowest). The frequency was calculated by dividing the total identified mutations at an index position by the total bases sequenced. The mutator modules with AID-T7pol and the EP-DNA-repair complex generated mutations across the complete GFP-mut3b ORF. AID-T7pol+UGI controls only generated mutations in GC regions, which indicates the EP-DNA-repair complex is essential for performing mutations at A:T base-pairs in the ORF. Mutations generally occurred in a 3-5 nucleotide range surrounding a GC base-pair. This suggests that the 5'-3' exonuclease cleaves a very short patch of nucleotides upstream of the nicked gap generated by an AP-endonuclease.*

5.4 Analysing the Sequencing Data Obtained from Illumina iSeq100

PacBio Sequel is a small-molecule real time sequencing technique where the addition of fluorescently tagged nucleotides to the elongating DNA template is detected, leading to the sequencing of the DNA molecule[225]. This method has its limitations: the error-rate of individual reads is in the range of 10-15% and the most common error introduced into the reads tends to be insertions or deletions[253,267]. Conversely, a sequence-by-synthesis sequencing platform like Illumina generates more substitution errors than indel errors[235]. This meant that for accurately assessing the mutation characteristics of the mutator modules, different sequencing platforms had to be used to compensate for the error-biases generated by each platform.

5.4.1 Issues with Error-Correction in the Context of the Mutagenesis Experiment

In literature, one of the ways for improving the accuracy of detecting SNPs is by having a greater sequence coverage of the genome or target genes. If 4x or higher sequence coverage is achieved, then the error-removal process involves aligning the individual reads to the reference genome, using a database of known SNPs, and using individual read overlaps to accurately detect mutations[232]. If a read possesses erroneous mutations that are not represented in the population, the read is removed or trimmed to improve alignment. Such an error-removal process cannot be applied in our case, where mutations that are randomly generated across the DNA sequence need to be detected. Aligning to a GFP-mut3b reference for error-removal would discard any random mutations that were not called in most of the overlapped reads.

Alternately, k-mer scanning error-correction techniques look to correct for errors by scanning the individual reads for short segments of nucleotides (called k-mers) present within the reads. The reads where certain k-mers are not identified are rejected. Subsequently, a statistically averaged k-mer sequence is generated from scanning all the reads. This averaged k-mer replaces the original k-mer sequence in all the reads, to eliminate incorrect base calls[232,268]. Such techniques are only limited to correcting substitution errors and do not work for indels. Like error-removal, this method of error-correction would also result in loss of random mutations generated by the mutator modules, as the likelihood of the same mutation being found in most of the aligned k-mers is low.

As a result, the best strategy to validate the sequencing data obtained from PacBio Sequel was to utilise another next generation sequencing technique. If the same pattern of mutations is witnessed in sequenced reads across different sequencing platforms, it would provide a strong validation for the observed mutation characteristics of the mutator modules. Therefore, the experimental samples from the 144-hr continuous evolution experiment were also sequenced using the Illumina iSeq100 platform.

5.4.2 Quality of the Illumina Reads

Illumina is a sequence-by-synthesis platform that can only process short DNA fragments of ≤ 300 bases. Pair-end sequencing utilises barcoded forward-reverse primer pairs. Each primer sequences ~150-nt from the 5'-end of the sense and antisense strands to provide the full coverage of 300-nt per primer pair. For the GFP-mut3b mutant library, a 275-bp section (Index 245-289, Figure 5.5) from the 5'-end of the GFP-mut3b ORF was sequenced using pair-end primers and Illumina iSeq100. To ensure there are no gaps in the sequencing coverage, the target DNA sequence was amplified with the barcoded primers such that:

length-of-read-per-primer ≤ length-of-DNA-target ≤ 2x(length-of-read-per-primer).

Generating overlaps in the pair-end reads helps significantly with the downstream error-correction process, but greater overlaps mean significantly shorter DNA fragments are sequenced per pair-end primer-pair. The forward (R1) and reverse (R2) reads of the sequenced GFP-mut3b library were ~ 135-nt in length, with an overlap of 13 nucleotides. The forward and reverse reads were trimmed, filtered based on Phred score using the AfterQC package[234]. The R1 and R2 subreads were combined using FLASH[244] to generate complete reads of the 245-bp DNA insert.

The confidence score of base-calling for Illumina was significantly lower than that of PacBio CCS reads. With multiple read-throughs of the same DNA molecule, PacBio called bases with an average Phred score of 60 (an error every 10^{-6} bases). With Illumina, the bases were called with a Phred score of 35 (error every $0.5\text{x}10^{-3}$ bases). Roughly $1\text{x}10^{8}$ nucleotide bases were sequenced per barcoded sample in the Illumina dataset. At QC_{35}, this equates to ~ 20,000 incorrectly called substitutions and indels due to sequencing error[240,241].

For a fair comparison between the mutations called with Illumina and PacBio methods, both data sets were error-corrected by subtracting the predicted number of incorrectly called mutations from the total identified mutations after alignment to the reference sequence (Table 5.8 and 5.9). Even after extensive data processing, Illumina and PacBio sequencing platforms have been shown to incorrectly call 0.1% and 1.3% of identified SNPs, respectively[227,269]. The data was corrected to account for this general platform-specific error-rate as well. The low-quality score of bases called by Illumina resulted in removal of 14%-68% percent of the mutations counted across the different barcoded samples. For PacBio, this was significantly lower, with only 0%-1% of the mutations being discarded.

$$Seq\ Error\ Mismatches = \left(\frac{Total\ Sequenced\ Bases}{10^{-\frac{Q}{10}}}\right) + (Total\ Mismatches\ x\ Platform\ Error\ Rate)$$

Where Q = Phred Score

Table 5.8: Adjusting the total mutations called to account for base-calling error and platform-specific error of Illumina iSeq100

Illumina iSeq100	MUT-9 Strong	MUT-22 Medium	MUT-25-1 Weak	MUT-25-2 Weak	ATF + UGI 1	ATF + UGI 2	T7pol
Total Reads	363106	559151	483648	434269	329753	431655	394093
Total Mismatches Called	434748	105996	94971	95455	479748	36513	36161
Total Bases Sequenced	88960970	136991995	118493760	106395905	80789485	105755475	96552785
Predicted Error Calls at QC35	17792	27398	23699	21279	16158	21151	19311
Inherent Illumina error of 0.1%	43475	10600	9497	9546	47975	3651	3616
Total Error calls	61267	37998	33196	30825	64133	24802	22927
% Correctly Called Mutations Above Sequencing Error Rate	85.9	64.2	65.0	67.7	86.6	32.1	36.6

Table 5.9: Adjusting the total mutations called to account for base-calling error and platform-specific error of PacBio Sequel

PacBio Sequel	MUT-9 Strong	MUT-22 Medium	MUT-25-1 Weak	MUT-25-2 Weak	ATF + UGI 1	ATF + UGI 2	T7pol	DH5α^{GFP}
Total Reads	1099	6303	6895	5955	300	238	5292	5002
Total Mismatches Called	2864	5749	4659	4645	989	819	958	1414
Total Bases Sequenced	1320998	7576206	8287790	7157910	360600	286076	6360984	6012404
Predicted Error Calls at QC60	1.3	7.6	8.3	7.2	0.4	0.3	6.4	6.0
Inherent PacBio Error of 1.3%	28.6	57.5	46.6	46.5	9.9	8.2	9.6	14.1
Total Error calls	30.0	65.1	54.9	53.6	10.3	8.5	15.9	20.2
% Correctly Called Mutations Above Sequencing Error Rate	99.0	98.9	98.8	98.8	99.0	99.0	98.3	98.6

5.4.3 Alignment Condition for Reads Generated by Illumina

Similar to the alignment of PacBio CCS reads, an alignment condition was required to reduce noise in the called mutations. For comparability to PacBio reads, the condition of alignment score greater than 240 was used (where, Match = +1, Mismatch = 0, Gap creation = -1, Gap extension = -0.1) to allow for up to 5 mismatches per aligned read. Also, no mutations were recoded between index positions 375 and 386. Regardless of the alignment condition used, roughly 240,000 mismatches were called at index positions 380 and 381 for all samples. This mismatch count was 100- to 1000-fold higher than the mismatches recorded at other index positions, which resulted in skewed mutation frequencies being calculated (Figure 5.6). These index positions were sequenced using the R2 (reverse primer) and were located near the 3'-end of the barcoded sample. Due to the imperfect chemistry and imaging system of the Illumina platform, base calling errors increase significantly closer to the 3'-end during sequencing[227,235,269].

Illumina platforms involve simultaneous sequencing of large clusters of identical ssDNA templates using fluorescently labelled reversible terminator-bound dNTPs. Only one labelled dNTP (either dA, dC, dT or dG) is added per sequencing cycle and when bound to the template, further polymerisation is prevented due to the blocked 3'-end and the fluorescence signal is detected from each DNA molecule in the cluster by random addition[235,270]. The process is repeated with all 4 labelled nucleotides. While sequencing, a template strand that fails to incorporate a base in a given cycle will continue to lag behind, a phenomenon referred to as phasing. Conversely, if multiple bases are added in a single cycle, this is called pre-phasing. Phasing, pre-phasing and the decay of signal intensity from one cycle to another results in an increase of base-calling errors towards the end of reads. Therefore, index positions 375-386 were trimmed to allow for more accurate calculation of mutation frequency generated by the mutator modules and the controls.

Lack of Overlap Between R1 and R2 Sub-Reads Significantly Limited the Accuracy of Mutation Calling

A possible explanation for the high mismatch count seen at some index positions and generally across single reads could be the high error-rate of Illumina sequencing platforms. The substitution error-rate for Illumina technologies has been documented in the range of 0.1 – 0.001 base^{-1}. The sample preparation can sometimes determine the degree of errors that can result from the sequencing process[235]. Schirmer and colleagues showed that the three established sample preparation techniques, Fusion Golay, Nextera XT and Universal Tail Tag, each result in completely different error-profiles. Fusion Golay sample prep results in heavily biased mutation rates at dG (3 to 10-fold higher than mutation rates at other nucleotides); Universal Tail Tag has a uniform error-rate of ~ 0.002 base^{-}

^{1} across all nucleotides; and Nextera XT displays a bias for mutations at dT and dC on the R1 (forward primer) subread and bias for dT and dG on the R2 (reverse primer) subread[235]. The error-rate at dT and dG is documented at 0.005 base^{-1}. As the Nextera XT sample preparation protocol was used in GFP-mut3b sequencing, it explains why the sequenced reads were limited to a Phred score of 35. Overall, the limitations of the Illumina sequencing platform and the sample preparation technique resulted in single reads with error-rates of ~ 0.005 base^{-1}.

The high single-molecule error-rate can be offset with downstream error-correction, which involves overlapping the R1 and R2 subreads to generate error-free contigs. This error-correction is dependent on the degree of overlap between R1 and R2 subreads. It has been documented that with an overlap of ≥ 80% between pair-end reads, the sequencing error mismatches can be reduced by ~90%, using correction software such as PEAR[231], PANDAseq[230], FLASH[244] and BayesHammer[235]. In our experiment, the pair-end overlap was only 5%. Consequently, for the 232 non-overlapped nucleotides, no error-correction could be performed. This made it difficult to distinguish substitution events caused by the mutator modules from the mismatch error introduced by Illumina. This also resulted in higher mutation frequencies being witness in the T7-RNA-Pol and DH5α^{GFP} controls for the Illumina data[xlii], compared to PacBio. As a result, a comprehensive analysis for assessing the mutation diversity and mutational spread generated by the mutator modules could not be performed using Illumina reads. However, after adjusting the Illumina dataset to the specified alignment score condition of > 240, certain mutational trends could be identified, which helped to validate the analysis performed on the PacBio sequencing dataset.

5.4.4 Mutation Frequencies Calculated from the Illumina Dataset

The total mutations called by each mutator and control module was corrected for the QC$_{35}$ base-call error-rate (Table 5.8) and the remaining mutation count witnessed in each sample was assumed to be generated by either mutator activity or due to the background cell mutation rate in the case of T7pol and DH5a^{GFP} controls. Roughly 80 to 100 million bases of GFP-mut3b were sequenced for each of the samples. The DNA itself accumulated mutations for 144-hours, which equates to ~288 bacterial cell cycles. Using these values, the mutation frequency of GFP-mut3b resulting from the different mutators was calculated (Table 5.10).

The mutation frequencies calculated using Illumina base-calling were generally comparable to the PacBio data. The weak and medium strength mutators displayed a mutation rate of 0.2x10^{-5} sbp^{-1}g^{-1},

[xlii] Refer to Table 5.12

which was an order of magnitude higher than T7-RNA-Pol and DH5α^{GFP}. Sequencing error and possible cross-contamination between samples during the continuous evolution experiment could be the likely reasons for the high mutation rate seen in GFP-mut3b reads from DH5α^{GFP} control cells.

The mutation frequency with the strong mutator and AID-T7pol+UGI was two-fold higher in the Illumina data compared to PacBio (Table 5.10). This difference probably resulted from more reads passing through the alignment score > 240 condition with Illumina, compared to number of reads qualifying for alignment score > 1197 in the PacBio data. This stringent alignment condition resulted in 69% of the reads for the strong mutator being discarded in the PacBio dataset, while only 28% of the reads were discarded in the Illumina dataset (Tables 5.1 & 5.11). The Sanger sequencing data presented in Section 3.4 showed AID-T7pol+UGI to generate ~ 6-10 mutations per read after 144-hrs. With the alignment score > 1197 condition, such mutants with >5 mismatches would have been discarded, resulting in a lower estimate of the mutation frequency with the strong mutator and AID-T7pol+UGI. Overall, the strong mutator was shown to generate mutations at a rate of ~ 1x10^{-5} sbp^{-1}g^{-1}, which is on par with the established OrthoRep system[109].

Table 5.10: Mutation Frequency Calculated from the Illumina Reads. Sequencing Reads Filtered at Alignment Score > 240 and Corrected for Sequencing Error

	MUT-9 <u>Strong</u>	**MUT-22 <u>Medium</u>**	**MUT-25-1 <u>Weak</u>**	**MUT-25-2 <u>Weak</u>**	**ATF + UGI 1**	**ATF + UGI 2**	**T7pol**
Total mutations called	373481	67998	61775	64630	415615	11711	13234
Total Reads	363106	559151	483648	434269	329753	431655	394093
Total base pairs sequenced	88960970	136991995	118493760	106395905	80789485	105755475	96552785
Mutation Frequency (per sbp*) x10^{-3}	4.20	0.50	0.52	0.61	5.14	0.11	0.14
Mutation Frequency (per sbp^{-1}g^{-1}) x10^{-5}	1.46	0.17	0.18	0.21	1.79	0.04	0.05

*** - per sbp refers to per sequenced base pair**

Table 5.11: Percentage of Illumina Sequencing Reads Discarded at Alignment Score > 240							
	MUT-9 Strong	MUT-22 Medium	MUT-25-1 Weak	MUT-25-2 Weak	ATF + UGI 1	ATF + UGI 2	T7pol
Total Reads Before Conditional Filtering	505381	609402	529306	473766	456361	542969	509785
Total Reads Passing Alignment Score Condition	363106	559151	483648	434269	329753	431655	394093
% Reads Discarded	28.2	8.2	8.6	8.3	27.7	20.5	22.7

5.4.5 Diversity of Mutations Identified Using Illumina

As discussed in Section 5.4.3, nucleotide substitutions are the most common error created by Illumina sequencing platforms. While the substitution error-rate is generally 0.001 base^{-1}, its increased 4-fold if the Nextera XT sample preparation protocol is used. This issue was compounded by the limited ability to perform error-correction as the sequence overlap between the R1 and R2 reads was only 5% in our experiments. As a result, the frequency of each substitution event could not be accurately predicted.

As a qualitative analysis, some common trends could be identified between the two datasets. C → T and G → A were the most frequently occurring substitutions for all the mutators and AID-T7pol+UGI. The frequency of these substitutions was 2-fold greater within Illumina reads (Table 5.12). Except for A → C, T → G, G → T and G → C, all other substitutions displayed comparable frequencies across the two sequencing platforms. A → G, T → A and T → C substitutions occurred at a high frequency across all seven samples. This is most likely explained by the error-bias from the Nextera XT sample preparation[235]. When DNA is prepared using this protocol, the pair-end reads display an error-rate of 0.02–0.1 base^{-1} at these nucleotides, which probably resulted in a high mutation count in the control samples. Future sequencing experiments to elucidate the characteristics of the mutator modules should be performed with greater overlap between the pair-end reads and using different sample preparation protocols, which would enable efficient error-correction of the reads and identify mutations generated by the mutator modules with improved accuracy.

Indel Frequency Significantly Lower with Illumina

The frequency of generating indels errors is known to be 10- to a 100-fold lower (~ 4x10^{-5} base^{-1}) than substitution errors with Illumina[235,271,227]. On the other hand, insertions are the most common error introduced by PacBio[227], at frequencies of ~ 0.1 base^{-1} [227]. This difference in error-rates can be clearly seen with a high frequency of insertion predicted in the PacBio dataset for all the samples including controls. With Illumina, the insertion frequency was 10- to 100-fold less. This indicates that most of the insertions witnessed in the PacBio dataset are most likely an artifact of the sequencing process. Illumina data predicts the frequency of indels to be 10 to 1000-fold lower than the frequency of generating unique nucleotide substitutions with the different mutator modules.

The rate of deletions, however, was higher in the Illumina data, with dG and dC being deleted 4-time more frequently than dA or dT. These findings loosely correlate with the deletion characteristics of DNA polymerase IV[252]. Kobayashi and colleagues have shown DNA Pol-IV to preferentially cause -1 frameshifts by deleting dC and dT at frequencies of $2x10^{-4}$ $base^{-1}$. dC and dG (possibly deletion of dC in the antisense strand) were the most frequently deleted nucleotides by the mutator modules, but at a 10-fold lower rate than published reports. This 10-fold reduction could be the result of using the truncated Pol-$IV^{\Delta 12}$, which has a reduced affinity to bind to replication forks via the β-clamp[252].

Overall, the Illumina dataset is potentially 100-fold more accurate than PacBio for calling indels, while the PacBio dataset is more accurate for calling substitution as the technique involves sequencing the same DNA molecule multiple times and generating a CCS. A coverage of ~ 20x was achieved for most of the individual reads in the PacBio library (Figure 5.3). By overlapping the ~ 20 subreads, a consensus sequence is generated, where each base is called to a Phred Score of 60 ($1x10^{-6}$ chance for error). As a result, to gain an accurate depiction of the mutagenic characteristics of the mutator modules, the advantageous qualities of PacBio and Illumina sequencing platforms were combined. PacBio sequencing, with its highly quality CCS reads provides an accurate insight into the nucleotide substitutions generated by the mutator modules, while Illumina platforms provide a clearer insight into the indels generated.

Table 5.12: Mutation Frequency per bp per read at Alignment Score > 240 (x10^{-6}) (QC35 Error Corrected and Adjusted for Illumina Error-Rate of 0.1%)

	Mutation	Strong Mutator	Medium Mutator	Weak Mutator 1	Weak Mutator 2	ATF + UGI	T7pol	DH5α^{GFP}
Substitutions	A --> T	11.4	11.0	11.3	11.7	8.2	2.4	2.6
	A --> C	4.6	4.9	4.8	4.5	2.0	0.5	0.8
	A --> G	39.5	34.5	35.5	35.1	38.3	13.7	15.8
	T --> A	40.8	45.3	46.8	46.7	35.2	7.1	9.6
	T --> C	45.3	40.0	41.0	42.1	43.9	14.4	17.5
	T --> G	4.7	5.5	5.5	5.3	1.8	0.4	0.5
	C --> T	2435.0	183.1	193.2	259.2	3157.6	25.9	32.1
	C --> A	14.4	15.9	17.2	17.3	10.9	1.7	2.3
	C --> G	22.8	23.6	24.9	24.0	22.6	6.0	7.7
	G --> T	7.7	7.3	7.1	7.8	4.0	1.1	1.4
	G --> C	3.4	3.6	3.6	3.7	2.8	0.7	1.0
	G --> A	1573.1	120.6	130.1	152.6	1839.0	36.5	47.4
Deletions	T --> -	6.4	7.1	7.0	7.6	5.7	1.9	2.4
	A --> -	3.2	3.9	4.2	4.1	2.7	0.8	1.0
	G --> -	22.9	25.9	26.4	26.6	19.2	7.4	8.2
	C --> -	20.3	21.8	23.2	24.2	17.7	6.6	7.9
Insertions	- --> T	1.8	1.6	2.0	1.8	1.4	0.7	0.7
	- --> A	11.8	10.7	11.2	12.0	11.3	4.0	4.6
	- --> G	1.4	1.3	1.4	1.5	1.4	0.5	0.6
	- --> C	5.9	4.5	5.0	5.5	5.4	2.0	2.3

Table 5.13: Spread of Substitutions and Indels in the Reads from Illumina

Sequences with additions	1966	3407	3109	2856	1623	2203	1984
Sequences with deletions	5419	12328	10863	9601	4187	5528	5102
Sequences with substitutions	307591	329967	290869	265008	300395	231029	213726
Total unchanged sequences	48130	213449	178807	156804	23548	192895	173281
Total Sequences Aligned	363106	559151	483648	434269	329753	431655	394093

5.4.6 Mutational Spread Across 245-bases of GFP-mut3b

The mutation frequency spread across the GFP-mut3b ORF displayed similar patterns for reads from both sequencing platforms. dC and dG displayed the highest mutation frequencies, while mutation frequencies at AT regions was more uniform for the strong, medium and weak mutator modules (Figure 5.7). In AT-rich regions (Index positions 246-252, 268-277, 294-298, 305-311 and 350-370), AID-T7pol+UGI generated little to no mismatches, while the mutation frequency at these sites with the mutator modules was 2 to 5-fold higher. This provides evidence that the error-prone DNA polymerase, DNA Pol-IV$^{\Delta12}$ and 5'-3' or 3'-5' exonuclease activity of the error-prone DNA repair complex is needed to reliably introduce mutations at AT-sites. The mutation frequency for AID-T7pol+UGI in AT-regions was in the range of 0.0 – 1x10^{-7}, which is lower than the average mutation rate found in similar regions for the T7-RNA-Pol and DH5α^{GFP} controls. The mutation frequencies for the mutator modules was ~ 5-fold higher at AT sites compared to AID-T7pol+UGI.

The negative controls displayed significantly higher mutation frequencies at AC (CA) and GT (TG) motifs. This is likely due to bias in the Illumina chemistry, which causes the platform to generate sequence-specific errors[270–272]. GGC is the most common motif, where Illumina generates significantly more errors[271]. The GGC motif in Figure 5.7 (Index position 399) displayed high error-rate for all the samples. Another limitation of Illumina is the higher base-calling error at AC and GT, which results from crosstalk between the emission spectra of the four fluorophores, with highest crosstalk between the fluorophores of A-C and G-T, respectively[273]. In our data, this error-rate seems to become worse when AC and GT are found in tandem. At index ranges 371-374 and 391-394, the presence of these nucleotide pairs in tandem resulted in 10-fold to 100-fold higher mutation frequencies being recorded for the T7-RNA-Pol and DH5α^{GFP} controls.

Overall, the general trend that could be identified from the Illumina data is that frequency of mutations is higher at AT-sites with the mutator modules compared to the AID-T7pol+UGI control. This give an indication that the of 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex is functional and is able to diversify the range of the mutator modules to AT-sites. But, due to the high substitution error-rate of Illumina platforms and the lack of overlap-based error-correction means a more comprehensive analysis could not be performed with this dataset. The common trends identified in the mutation spread from both datasets helped validate the PacBio sequencing data, where the base substitutions were called to a 1000-fold higher quality score. The PacBio data provides an accurate look into the substitution profile generated by the mutator modules, while Illumina provides a more accurate indel profile.

Index Position:	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395
GFP Sequence:	**A**	**A**	**C**	**T**	**T**	**A**	**C**	**C**	**C**	**T**	**T**	**A**	**A**	**A**	**T**	**T**	**T**	**A**	**T**	**T**	**T**	**G**	**C**	**A**	**C**	**T**	**A**	**C**	**T**	**G**	**G**	**A**	**A**	**A**	**A**	**C**	**T**	**A**	**C**	**C**	**T**	**G**	**T**	**T**	**C**
Number of Mutations						**GFP-Mut3b**																																							
Strong Mutator:	124	244	13163	273	165	778	8747	21545	4067	249	14592	198	111	152	748	87	167	181	266	257	918	32041	17187	693	71026	1198	332	97730	2784	247761	196243	1160	224	180	297	100995	1428	429	14832	13517	1264	18795	1510	434	74653
Medium Mutator:	172	320	362	216	200	912	371	304	323	250	17048	218	177	161	796	109	219	195	310	300	1207	27657	8936	775	45495	1209	381	85849	3322	295830	235874	1413	299	221	367	46894	1295	445	10448	3789	1576	11720	1679	539	2542
Weak Mutator 1:	102	202	288	191	162	747	344	251	262	198	14551	174	112	121	664	124	185	160	282	250	1081	25154	8217	700	40361	1092	344	77730	3017	260164	206420	1291	247	203	332	43084	1198	391	9945	3734	1367	10652	1534	419	2486
Weak Mutator 2:	125	264	311	182	182	725	434	277	272	234	12348	160	105	121	541	84	171	153	233	240	954	21810	10247	612	34450	889	299	67127	2559	233162	184858	1120	216	162	248	36888	939	352	8149	2905	1131	9064	1254	349	2247
AID-T7pol UGI:	89	195	17776	200	157	543	12247	20914	3179	240	11623	152	92	132	733	90	181	143	236	243	928	30155	18330	687	46744	1047	347	84875	2560	218527	171614	988	230	183	274	91983	1278	405	12216	8031	1181	17714	1459	434	74783
T7pol:	61	69	39	83	96	92	107	115	75	79	1212	58	80	60	51	48	67	47	73	112	181	6472	2058	140	4254	119	83	9665	216	157149	114507	309	90	84	71	4222	109	111	920	182	105	568	149	118	279
T86:	54	57	37	80	62	94	80	83	61	76	1357	53	77	42	36	48	54	64	56	113	188	7130	2131	159	4211	112	86	10485	210	143401	103133	279	88	65	73	4771	107	109	931	237	106	619	136	106	281
	A	A	C	T	T	A	C	C	C	T	T	A	A	A	T	T	T	A	T	T	T	G	C																						
											A	A	A	A	T	T	T	A	T	T	T	G	C	A	C	T	A	C	T	G	G	A	A	A	A	C	T	A	C	C	T	G	T	T	C
	Forward Primer Read										Overlap															Reverse Primer Read																			

***Figure 5.6: Substitution errors generated by Illumina**. Illumina sequencing platforms are prone to generating substitution errors at a frequency of 0.1 – 0.005 base^{-1}. A high level of mutation was seen at index positions 380 and 381 across all the barcoded samples, including native DH5α cells. Near the tail end of the sequencing reads, to base-calling error increases due to phasing or pre-phasing. Error-correction using different in silico tools can be performed to rectify the issue, if the forward R1 reads and reverse R2 reads have significant overlap. In this instance, there was only a 13-base overlap for a 245 base-pair long read. Error-correction could not be performed as a result. To reduce noisy mutation calls due to the platform-specific error-rates, mutations were not counted from index position 375-386.*

Index Position:	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295
GFP Sequence:	G	A	A	C	T	T	T	T	C	A	C	T	G	G	A	G	T	T	G	T	C	C	C	A	A	T	T	C	T	T	G	T	T	G	A	A	T	T	A	G	A	T	G	G	T	G	A	T	G	T	T
Number of Mutations																																																			
Strong Mutator:	1.66	0.30	0.13	0.12	0.11	0.12	0.12	0.12	0.55	0.11	3.05	0.15	1.23	1.60	0.18	5.80	0.07	0.15	0.41	0.17	1.51	2.55	2.61	0.11	0.04	0.07	0.22	1.76	0.05	0.16	12.27	0.07	0.12	16.94	0.08	0.15	0.15	0.10	0.10	1.88	0.07	0.10	0.63	5.78	0.08	2.52	0.08	0.14	18.28	0.05	0.11
Medium Mutator:	3.21	0.55	0.29	0.29	0.33	0.22	0.27	0.17	0.63	0.22	0.20	0.37	0.10	0.13	0.37	0.10	0.16	0.28	0.18	0.47	0.48	0.25	0.26	0.28	0.15	0.13	0.41	0.11	0.24	0.25	0.13	0.28	0.28	0.22	0.33	0.20	0.14	0.21	0.21	0.09	0.22	0.22	0.21	0.13	0.23	0.23	0.18	0.21	0.12	0.31	0.31
Weak Mutator 1:	3.12	0.54	0.29	0.32	0.26	0.21	0.20	0.26	0.71	0.15	0.17	0.42	0.15	0.20	0.39	0.08	0.18	0.36	0.10	0.39	0.63	0.25	0.29	0.19	0.20	0.17	0.55	0.09	0.19	0.35	0.16	0.18	0.22	0.31	0.28	0.24	0.20	0.24	0.21	0.10	0.21	0.36	0.21	0.12	0.25	0.15	0.21	0.22	0.18	0.22	0.30
Weak Mutator 2:	3.36	0.51	0.30	0.31	0.27	0.29	0.27	0.23	0.81	0.15	0.40	0.40	0.19	0.21	0.31	0.10	0.24	0.34	0.17	0.56	0.50	0.27	0.22	0.15	0.15	0.14	0.48	0.08	0.27	0.29	0.24	0.19	0.34	1.20	0.31	0.20	0.17	0.19	0.24	0.08	0.20	0.32	0.16	0.22	0.24	0.28	0.22	0.26	0.17	0.22	0.27
AID-T7pol UGI:	0.73	0.11	0.04	0.02	0.09	0.05	0.06	0.14	0.40	0.06	4.00	0.10	1.88	2.02	0.12	7.27	0.01	0.09	0.51	0.12	0.55	4.15	0.83	0.05	0.02	0.02	0.10	2.91	0.03	0.07	15.19	0.01	0.06	9.54	0.07	0.07	0.03	0.06	0.05	3.12	0.01	0.02	0.94	3.06	0.09	3.60	0.04	0.06	17.90	0.04	0.07
T7pol:	1.79	0.34	0.21	0.26	0.14	0.14	0.11	0.15	0.43	0.12	0.10	0.23	0.10	0.11	0.24	0.03	0.11	0.16	0.10	0.28	0.34	0.16	0.14	0.17	0.10	0.09	0.28	0.07	0.10	0.17	0.07	0.16	0.12	0.19	0.18	0.11	0.09	0.10	0.14	0.06	0.12	0.13	0.12	0.08	0.17	0.16	0.10	0.13	0.09	0.14	0.18
DH5αGFP	2.15	0.36	0.23	0.20	0.19	0.12	0.15	0.16	0.39	0.11	0.15	0.30	0.11	0.09	0.27	0.07	0.14	0.22	0.17	0.26	0.37	0.16	0.12	0.20	0.10	0.13	0.34	0.03	0.12	0.20	0.13	0.17	0.23	0.20	0.16	0.13	0.10	0.19	0.18	0.07	0.14	0.22	0.13	0.09	0.17	0.17	0.11	0.13	0.13	0.13	0.22

Index Position:	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346
GFP Sequence:	A	A	T	G	G	G	C	A	C	A	A	A	T	T	T	T	C	T	G	T	C	A	G	T	G	G	A	G	A	G	G	G	T	G	A	A	G	G	T	G	A	T	G	C	A	A	C	A	T	A	C
Number of Mutations																																																			
Strong Mutator:	0.10	0.10	0.11	0.18	0.11	2.16	0.62	0.15	2.46	0.09	0.07	0.09	0.08	0.10	0.07	0.07	0.84	0.15	0.85	0.12	0.67	0.22	1.16	0.14	0.95	0.33	0.19	1.12	0.17	0.39	1.07	1.38	0.11	2.59	0.17	0.22	0.85	1.06	0.09	1.09	0.12	0.13	1.65	9.80	0.18	0.21	1.62	0.06	0.09	0.14	17.36
Medium Mutator:	0.20	0.14	0.19	0.21	0.11	3.11	0.21	0.25	0.11	0.27	0.22	0.21	0.12	0.19	0.19	0.19	0.13	0.25	0.12	0.24	0.31	0.28	0.09	0.26	0.15	0.12	0.46	0.07	0.21	0.16	0.22	0.21	0.24	0.22	0.26	0.38	0.45	0.31	0.22	0.36	0.26	0.34	0.30	0.57	0.35	0.31	0.11	0.15	0.17	0.25	0.22
Weak Mutator 1:	0.19	0.16	0.15	0.21	0.19	3.06	0.20	0.27	0.14	0.25	0.26	0.20	0.14	0.18	0.21	0.22	0.15	0.19	0.19	0.20	0.29	0.39	0.09	0.27	0.16	0.14	0.53	0.08	0.25	0.19	0.16	0.20	0.30	0.17	0.30	0.43	0.38	0.32	0.22	0.36	0.21	0.28	0.29	0.49	0.32	0.30	0.09	0.14	0.10	0.24	0.44
Weak Mutator 2:	0.23	0.15	0.20	0.42	0.19	2.86	0.22	0.27	0.15	0.21	0.20	0.13	0.22	0.17	0.17	0.21	0.17	0.17	0.15	0.22	0.27	0.42	0.14	0.24	0.13	0.13	0.40	0.15	0.21	0.15	0.52	0.39	0.26	0.26	0.42	0.33	0.59	0.38	0.27	0.39	0.31	0.40	0.32	1.49	0.27	0.35	0.13	0.20	0.15	0.22	0.34
AID-T7pol UGI:	0.03	0.10	0.02	0.16	0.05	2.67	0.95	0.06	2.34	0.03	0.07	0.02	0.03	0.07	0.00	0.04	1.21	0.06	1.46	0.05	0.49	0.07	2.05	0.11	1.62	0.54	0.07	0.43	0.07	0.37	0.95	1.52	0.04	2.59	0.06	0.12	0.99	1.16	0.05	1.67	0.04	0.11	2.94	5.37	0.07	0.07	2.87	0.04	0.04	0.10	20.15
T7pol:	0.13	0.09	0.10	0.10	0.11	1.91	0.12	0.25	0.13	0.15	0.09	0.11	0.08	0.12	0.11	0.14	0.07	0.16	0.06	0.15	0.21	0.21	0.07	0.18	0.07	0.06	0.24	0.04	0.14	0.08	0.08	0.10	0.15	0.12	0.17	0.24	0.14	0.15	0.13	0.21	0.18	0.13	0.12	0.32	0.21	0.20	0.05	0.06	0.09	0.16	0.19
DH5αGFP	0.14	0.13	0.11	0.20	0.07	2.07	0.13	0.17	0.07	0.21	0.14	0.10	0.14	0.08	0.19	0.12	0.06	0.17	0.08	0.13	0.20	0.21	0.05	0.22	0.07	0.08	0.32	0.06	0.15	0.09	0.08	0.11	0.17	0.13	0.23	0.30	0.22	0.17	0.18	0.25	0.15	0.18	0.22	0.32	0.21	0.25	0.08	0.09	0.08	0.19	0.25

Index Position:	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397
GFP Sequence:	G	G	A	A	A	A	C	T	T	A	C	C	C	T	T	A	A	A	T	T	T	A	T	T	T	G	C	A	C	T	A	C	T	G	G	A	A	A	A	C	T	A	C	C	T	G	T	T	C	C	A
Number of Mutations																													GFP-Mut3b																						
Strong Mutator:	2.05	1.02	0.11	0.15	0.08	0.10	5.22	0.09	0.16	0.07	3.44	8.01	1.02	0.08	0.73	0.10	0.10	0.03	0.04	0.06	0.09	0.07	0.02	0.20	0.14	8.80	4.69	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.15	2.98	4.12	0.11	2.57	0.19	0.18	26.06	0.72	0.16
Medium Mutator:	0.55	0.64	0.28	0.29	0.26	0.29	0.11	0.23	0.30	0.24	0.18	0.22	0.15	0.22	1.46	0.17	0.28	0.15	0.20	0.17	0.16	0.17	0.27	0.37	0.55	12.28	3.93	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.23	1.56	0.32	0.31	0.97	0.35	0.39	0.43	0.21	0.23
Weak Mutator 1:	0.66	0.74	0.35	0.38	0.18	0.14	0.08	0.25	0.25	0.22	0.24	0.20	0.18	0.20	1.46	0.14	0.24	0.16	0.15	0.19	0.15	0.16	0.13	0.33	0.58	13.22	4.45	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.26	1.66	0.46	0.35	1.01	0.38	0.35	0.52	0.19	0.32
Weak Mutator 2:	1.25	1.06	0.33	0.24	0.20	0.26	0.10	0.25	0.33	0.23	0.27	0.26	0.18	0.24	1.43	0.15	0.23	0.08	0.14	0.17	0.18	0.19	0.23	0.45	0.52	12.93	11.96	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.26	1.66	0.37	0.35	1.27	0.38	0.38	0.97	0.23	0.27
AID-T7pol UGI:	4.11	1.58	0.12	0.03	0.05	0.04	8.47	0.05	0.06	0.09	5.95	10.38	1.23	0.10	0.44	0.05	0.02	0.05	0.03	0.03	0.05	0.07	0.05	0.12	0.18	7.97	6.81	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.09	1.89	2.14	0.04	4.56	0.10	0.04	41.55	1.61	0.18
T7pol:	0.30	0.42	0.14	0.16	0.11	0.11	0.06	0.13	0.17	0.15	0.17	0.15	0.09	0.13	0.76	0.09	0.13	0.11	0.08	0.07	0.13	0.09	0.11	0.20	0.27	8.09	2.67	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.20	0.97	0.20	0.14	0.57	0.26	0.21	0.31	0.11	0.15
DH5αGFP	0.39	0.41	0.17	0.28	0.11	0.11	0.05	0.17	0.15	0.18	0.15	0.15	0.11	0.17	0.98	0.10	0.18	0.10	0.07	0.08	0.14	0.15	0.12	0.25	0.39	11.63	3.44	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.20	1.21	0.32	0.21	0.84	0.25	0.23	0.44	0.10	0.16

Index Position:	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448
GFP Sequence:	T	G	G	C	C	A	A	C	A	C	T	T	G	T	C	A	C	T	A	C	T	T	T	C	G	G	T	T	A	T	G	G	T	G	T	T	C	A	A	T	G	C	T	T	T	G	C	G	A	G	A
Number of Mutations																																																			
Strong Mutator:	0.07	0.23	0.79	1.97	0.69	0.12	0.12	5.04	0.09	1.85	0.15	0.15	0.64	0.17	0.30	0.16	3.63	0.09	0.11	2.20	0.10	0.13	0.08	1.99	0.84	14.47	0.07	0.07	0.17	0.07	0.53	5.66	0.11	10.14	0.15	0.13	0.66	0.17	0.16	0.10	4.41	5.89	0.05	0.10	0.26	0.75	34.60	0.23	0.16	1.25	0.05
Medium Mutator:	0.30	0.55	1.46	0.52	0.99	0.22	0.21	2.81	0.14	0.52	0.31	0.33	0.96	0.33	0.35	0.29	0.14	0.17	0.28	0.07	0.27	0.33	0.21	0.40	0.21	0.16	0.18	0.12	0.25	0.20	0.21	0.17	0.25	0.18	0.28	0.35	0.43	0.34	0.25	0.19	0.16	0.19	0.19	0.25	0.28	0.14	1.64	0.32	0.26	0.09	0.15
Weak Mutator 1:	0.27	0.49	1.50	0.49	0.95	0.34	0.19	3.16	0.21	0.47	0.29	0.31	1.03	0.39	0.26	0.31	0.13	0.24	0.23	0.04	0.22	0.35	0.16	0.50	0.18	0.15	0.21	0.11	0.27	0.21	0.23	0.15	0.22	0.19	0.26	0.35	0.41	0.35	0.29	0.18	0.21	0.24	0.13	0.31	0.31	0.17	1.57	0.27	0.29	0.10	0.16
Weak Mutator 2:	0.31	0.42	1.53	0.50	1.11	0.25	0.22	2.86	0.20	0.54	0.17	0.27	1.53	0.33	0.32	0.36	0.12	0.24	0.22	0.11	0.19	0.33	0.20	0.50	1.19	0.18	0.20	0.18	0.22	0.26	0.17	0.16	0.20	0.95	0.21	0.24	0.47	0.32	0.27	0.45	0.21	0.35	0.10	0.26	0.36	0.15	1.51	0.26	0.20	0.07	0.21
AID-T7pol UGI:	0.11	0.11	0.56	2.15	0.63	0.07	0.07	4.45	0.07	4.01	0.06	0.10	0.72	0.10	0.39	0.09	6.55	0.04	0.05	5.12	0.05	0.04	0.01	4.03	1.72	10.62	0.02	0.02	0.02	0.02	1.14	4.16	0.06	9.63	0.06	0.07	1.12	0.06	0.12	0.02	4.41	8.79	0.05	0.06	0.09	0.93	48.74	0.36	0.06	2.30	0.03
T7pol:	0.21	0.30	0.87	0.36	0.54	0.18	0.10	1.78	0.12	0.29	0.16	0.18	0.54	0.19	0.19	0.13	0.08	0.10	0.13	0.04	0.12	0.12	0.13	0.25	0.12	0.09	0.10	0.08	0.15	0.17	0.08	0.08	0.12	0.08	0.13	0.15	0.20	0.19	0.17	0.09	0.06	0.08	0.08	0.14	0.17	0.04	1.02	0.16	0.16	0.05	0.10
DH5αGFP	0.21	0.28	0.92	0.34	0.64	0.22	0.14	2.43	0.15	0.44	0.14	0.18	0.74	0.27	0.24	0.16	0.11	0.16	0.19	0.05	0.20	0.20	0.15	0.28	0.11	0.13	0.15	0.10	0.16	0.15	0.13	0.11	0.16	0.13	0.20	0.24	0.25	0.20	0.25	0.10	0.08	0.10	0.10	0.13	0.19	0.10	0.92	0.20	0.17	0.07	0.10

Index Position:	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489
GFP Sequence:	T	A	C	C	C	A	G	A	T	C	A	T	A	T	G	A	A	A	C	A	G	C	A	T	G	A	C	T	T	T	T	T	C	A	A	G	A	G	T	G	C
Number of Mutations																																									
Strong Mutator:	0.11	0.05	4.18	2.00	1.43	0.15	1.68	0.07	0.10	1.21	0.09	0.10	0.09	0.14	0.67	0.12	0.16	0.15	24.68	0.08	2.39	6.77	0.11	0.08	0.69	0.15	0.62	0.12	0.06	0.12	0.14	0.10	1.43	0.22	0.10	0.17	0.08	0.06	0.07	0.13	0.21
Medium Mutator:	0.11	0.22	0.40	0.28	0.17	0.28	0.12	0.16	0.15	0.18	0.23	0.25	0.23	0.24	0.27	0.28	0.31	0.21	0.20	0.22	0.26	0.31	0.25	0.25	0.23	0.20	0.07	0.23	0.15	0.20	0.30	0.27	0.34	0.56	0.24	0.07	0.19	0.07	0.19	0.30	0.50
Weak Mutator 1:	0.15	0.19	0.37	0.45	0.21	0.22	0.10	0.18	0.17	0.21	0.20	0.23	0.25	0.28	0.24	0.31	0.26	0.22	0.29	0.17	0.48	0.38	0.26	0.18	0.14	0.24	0.06	0.20	0.09	0.27	0.19	0.30	0.40	0.60	0.32	0.09	0.19	0.13	0.21	0.26	0.37
Weak Mutator 2:	0.17	0.14	0.25	0.40	0.22	0.20	0.10	0.19	0.20	1.05	0.20	0.21	0.23	0.31	0.21	0.24	0.26	0.28	1.88	0.26	0.52	0.82	0.29	0.24	1.40	0.22	0.04	0.24	0.19	0.22	0.30	0.34	0.39	0.54	0.25	0.09	0.23	0.12	0.22	0.30	0.35
AID-T7pol UGI:	0.04	0.01	5.50	3.32	2.24	0.06	1.78	0.04	0.05	2.19	0.04	0.04	0.06	0.11	1.10	0.10	0.03	0.06	31.51	0.02	4.73	7.49	0.07	0.02	0.34	0.11	0.51	0.04	0.03	0.05	0.06	0.02	1.67	0.15	0.06	0.27	0.03	0.00	0.02	0.02	0.13
T7pol:	0.09	0.12	0.20	0.18	0.14	0.19	0.07	0.08	0.14	0.09	0.11	0.15	0.13	0.14	0.16	0.17	0.18	0.14	0.17	0.13	0.11	0.14	0.17	0.10	0.08	0.15	0.03	0.13	0.10	0.11	0.15	0.14	0.17	0.38	0.19	0.05	0.13	0.06	0.13	0.14	0.25
DH5αGFP	0.11	0.19	0.26	0.26	0.13	0.17	0.08	0.11	0.11	0.13	0.14	0.14	0.15	0.16	0.17	0.18	0.21	0.18	0.16	0.21	0.11	0.16	0.16	0.15	0.10	0.20	0.04	0.19	0.08	0.13	0.21	0.22	0.20	0.41	0.19	0.07	0.15	0.07	0.17	0.22	0.32

***Figure 5.7**: **Mutation spread analysed from the Illumina reads**. The figure shows a heat map representing the frequency of mutations (x10^{-6} sbp^{-1})occurring at each index position across a 245-bp region from the 5'-end of GFP-mut3b (Colour code: Red = High, Green = Low). The alignment result showed AID-T7pol+UGI to primarily perform C → T and G → A mutations. The frequency of mutations occurring at GC-regions was 20- to 100-fold higher than the frequency at AT-regions for AID-T7pol+UGI. The weak and medium mutator modules displayed a fairly uniform mutation frequency across the ORF. In AT-rich regions, the mutation frequency found more than 5-nt upstream or downstream of a GC base-pair was 10- to 100-fold lower.*

5.5 Summary of the Findings from Next Generation Sequencing of the Mutant GFP-mut3b Library

The aim of this chapter was to comprehensively characterise the mutagenic properties of the strong (MUT-9), medium (MUT-22) and weak (MUT-25) mutator modules identified in Chapter 4. To do this, a continuous evolution experiment was performed on a GFP-mut3b target integrated into the genome of the DH5α strain of *E. coli*. The experiment was performed for 144-hours to generate a diverse library of GFP-mut3b variants and subsequently assess the diversity, frequency and spread of mutations across the GFP ORF. The library of variants was sequenced using two next-generation sequencing platforms, PacBio Sequel and Illumina iSeq100. One technique performs real-time sequencing of one DNA molecule at a time, while the other employs a clustered sequence-by-synthesis approach using fluorescently labelled reversible terminator-bound dNTPs.

Due to differences in the sequencing chemistry, the two platforms produce different types of errors in the sequenced reads. Illumina generates substitution errors at frequencies of 0.1 – 0.001 per base, while the error-rate for indels is 10-fold to 100-fold lower. Conversely, PacBio methods generate indels with an error-rate of 0.1 – 0.01 per base, while the substitution error-rate can be as low as $1x10^{-6}$ if consensus reads can be generated. Therefore, assessing indel frequencies from the Illumina reads and substitution frequencies from the PacBio dataset provide a comprehensive strategy to study the mutational characteristics of the mutator modules.

The PacBio dataset showed that all mutator modules can perform a complete range of nucleotide substitutions. The average frequency of each substitution was ~ $4x10^{-5}$ sbp^{-1} for the medium mutator and $3.9x10^{-5}$ sbp^{-1} for the weak mutator. The strong mutator displayed similar frequencies for generating all substitutions except C → T and G → A. The frequency was 30-fold higher for these substitutions, which indicates that AID-T7pol may be generating many U:G lesions in the GFP ORF that are not being converted to AP-sites by uracil-N-glycosylase. Consequently, these mismatches cannot be repaired by either the cell's native DNA repair pathways or the EP-DNA-repair complex of the mutator modules. This possibly results in C → T and G → A mutations being passed on to daughter cells. This hypothesis should be tested by expressing UNG in the mutator modules. AID-T7pol+UGI only generated C → T and G → A substitutions, which provided validation that other types of nucleotide substitutions required the activity of the 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex. With mutations being spotted at A:T base-pairs up to five bases downstream of G:C, it provided evidence that the 5'-3' exonuclease domain in the 3-protein fusion of the EP-DNA-repair complex is functional and can perform error-prone short-patch repair.

The Illumina dataset could not be efficiently error-corrected due a lack of significant overlap between the R1 and R2 sub-reads. With the data being filtered at a Phred score = 35, it meant an incorrect mismatch is likely called every 5000 bases in the sequenced reads. As a result, a comprehensive substitution profile for the mutator modules could not be generated from this dataset. However, comparable frequencies to the PacBio data were seen for all, except AT, AC, GT and GC substitutions. The mutation frequencies were 5-fold to 10-fold higher for the reads from the mutator modules, compared to the ones from T7-RNA-Pol and DH5α^{GFP} controls in the Illumina data. This higher frequency can be inferred to result from the activity of the mutator modules. The main validation provided by the Illumina dataset was regarding indels. The Illumina platform has a 1000-fold lower error-rate for indels compared to PacBio. Illumina-sequenced libraries displayed a 10-fold to 100-fold lower insertion frequency than PacBio reads, at ~ 5x10^{-6} sbp^{-1} for each sample. The deletion frequency was higher at 2x10^{-5} sbp^{-1}, which most likely results from DNA polymerase IV activity. This deletion rate was 10-fold lower than what is documented in literature for Pol-IV. This reduction may have resulted from the 12-amino acid truncation in Pol-IV$^{\Delta 12}$ to prevent it from binding to the β-clamp during DNA replication.

The overall mutation rate per bacterial generation was calculated to be ~ 1x10^{-5} sbp^{-1}g^{-1} for the strong mutator and AID-T7pol+UGI. The medium and weak mutator modules displayed a 10-fold lower mutation rate. This places the strong mutator module on par with the established OrthoRep continuous evolution system and at a 100-fold higher mutation frequency than PACE[274,275]. The mutation frequencies displayed by these established systems is sufficient to generate a functional library of mutants in each cycle of library generation and consequently, achieved protein variants possessing the desired phenotype in a matter of days or few weeks using appropriate selection methods. With the diverse range of mutations that can generated using the AID-T7pol + 5'-3'Pol-I-Exo—Exo-III—Pol-IV$^{\Delta 12}$ mutator system, and mutation frequency of 1x10^{-5} sbp^{-1}g^{-1}, the possibility of these mutator modules to be applied in a continuous evolution system for library generation seems to be a promising endeavour.

Chapter 6: Conclusion and Future Work

6.1 Introduction

The Biotechnology field is centred around developing biological techniques for producing commercial products in the food, petrochemical, consumer, and pharmaceutical industries. This endeavour typically involves engineering synthetic gene circuits and artificial metabolic pathways that enable the production of desired chemicals and products inside model organisms, like bacteria, yeast and algae[30,31]. Engineering complex metabolic pathways and gene circuits requires identifying the appropriate enzymes where the substrates or products of one enzymatic reaction can be linked to another and perform a cascade of reactions with the desired chemicals as the final product[29–31]. In some instances, the chosen enzymes may not function properly in the model organism, which requires modifying the enzyme to have improved physical and chemical properties[31,276]. In other cases, the required enzymatic reaction may not exist in nature and must be artificially synthetised from existing proteins[104,133,266]. Protein engineering enables one to overcome these challenges. It is the process of altering the amino acid composition of a polypeptide sequence to obtain protein variants with desired enzymatic qualities. There are two main approaches to protein engineering: rational design and directed evolution.

Rational design approaches can be applied when there is vast knowledge about the structure and function of a protein. Various *in silico* tools can be applied to model protein characteristics, such as thermostability, protein folding, substrate interaction, and the effect of precise amino acid substitutions on enzyme behaviour can be monitored[35,36,47]. If the desired enzyme properties are achieved *in silico*, the protein is resynthesised with the new peptide sequence. The rational protein engineering approach cannot be applied when there is limited structural information about the protein. In such cases, directed evolution is the preferred technique.

Directed evolution involves using mutagenic techniques to alter the DNA sequence of the gene encoding the target protein, producing a library of variants in the process[92,246]. This library is subsequently scanned using appropriate screening or selection strategies to identify variants possessing the desired qualities. Multiple rounds of library generation and screening eventually result in desired protein variants. With library generation, the greatest challenge is the ability to introduce nucleotide substitutions reliably and efficiently into the target gene sequence. In the early years of research, this was a difficult task to perform *in vivo*, due to a lack of expertise on how to target mutations to the gene of interest, while maintaining genomic fidelity of the host cell. Chemical mutagens and mutator strains of bacteria were used, but these methods suffer from mutation biases and a low mutation rate, which makes it difficult to generate diverse protein libraries[95,157,277].

Researchers naturally began adopting *in vitro* techniques for more precise manipulation of the gene-of-interest[278]. Error-prone PCR with Taq polymerase became a widely applied technique for library creation. This method can suffer from mutation biases and a large portion of the library being non-functional if the mutation rate is very high[140,101]. DNA shuffling techniques can be coupled with EP-PCR to generate greater diversity in the library, but the overall library sizes are much smaller as 60-70% sequence homology is needed for successful DNA shuffling[279,280]. Aside from the precise, labour-intensive experimental protocols for such techniques, a major limitation to *in vitro* library generation is the upper limit on the library size that can be screening per directed evolution cycle. Screening or selection is generally performed *in vivo*, and the limit on library size is based on the transformation efficiency of the organism (~10^9 variants per cycle)[51,281]. Repeated cycles of library generation, transformation and screening can become tedious, and researchers resort to limiting the evolutionary search-space for protein evolution.

The evolutionary search-space is imagined to be a fitness landscape of mountains and valleys, where each favourable mutation allows the evolving protein to climb higher on the mountain, while unfavourable mutations move it further down[100,101] (Section 1.4, Figure 1.3). Evolving novel or promiscuous protein function involves traversing from one mountain to another by accumulating a series of unfavourable and subsequent favourable mutations. Such complex evolutionary pathways cannot be explored efficiently with the restricted library sizes generated by *in vitro* mutagenic techniques. An ideal directed evolution system for this endeavour would require library generation and selection to simultaneously occur *in vivo*, such that each replicating cell in a batch culture expresses a protein variant, and over multiple cell cycles, the protein can continuously evolve until the desired phenotype is achieved[282]. Such a scalable system would address library-size-related bottlenecks. As stated earlier, the challenge with this approach has been the need to develop an *in vivo* mutator system to specifically introduce mutations in the target gene, while performing minimal off-target mutations. This challenge was recently tackled by mutator systems, like, PACE[72], CRISPR-Cas9[111], and OrthoRep[109].

The PACE evolution system creates a link between the evolving protein and the virulence of M13 bacteriophage. If the desired protein function is achieved, virulent phage would be produced that infect other cells in the batch culture, leading to enrichment of the functional phenotype[275,283]. The limitation of this technique is the lack of targeting by the mutator system. DNA damaging enzymes and error-prone DNA polymerases are expressed in *E. coli* cells, which randomly mutate the target gene or the bacterial genome. The reduced genomic fidelity over multiple cell cycles affects cell viability for analysing long evolutionary pathways. Also, the mutation-rate is only 200-fold higher than the natural error-rate of cells, which can restrict genetic diversity in the protein library.

CRISPR-Cas systems provided a solution to the targeted mutagenesis challenge by fusing nucleotide deaminases to Cas9 or dCas9[133,284]. Using specific guide RNA, deamination events could be targeted to complimentary DNA sequence. The precise targeting, however, is limited to a small window of nucleotide bases surrounding the PAM site[117]. Performing continuous evolution of a gene with such a mutator system would require multiple gRNA to cover the complete ORF of the gene, and new gRNA would need to be synthesised to match the evolving DNA sequence, which is not feasible. This system is more applicable for semi-rational approaches where mutations need to be introduced in precise locations across the ORF.

OrthoRep is currently the only published mutator system that can generate targeted mutations across the complete open reading frame of the gene-of-interest, while also shielding the host cell's genome. The system utilises an orthogonal plasmid-DNA-polymerase pair with an error-prone mutant of TP DNA polymerase I, where the target gene is placed in a linear plasmid flanked by p1 terminal proteins. This system generates mutations at the rate of 1x10^{-5} substitutions per base-pair, which is 5-fold higher than the mutation frequency of the yeast genome in cells expressing the OrthoRep system. As a result, this mutator system enables long continuous evolution experiments lasting 100-500 hours[108,109]. However, OrthoRep's application is limited to yeast. Therefore, the aim of this doctoral thesis was to engineer a targeted mutator system for performing continuous evolution of genes in *E. coli*. Such a system would enable better engineering of proteins whose downstream application is in prokaryotic expression systems[285,286].

6.2 Summary of Research Findings

6.2.1 Biological Mechanism of the Mutator Module

The mutator system was developed to emulate the process of somatic hypermutations in *E. coli*. This mechanism enables maturing immunoglobulin cells to rapidly generate mutations in the genes encoding the variable region of antibodies and create an antibody library that is screened for its ability to bind to pathogenic antigens[125,287]. This mechanism utilises activation induced cytidine deaminase (AID) to cleave an amide group from deoxycytidine, and generate a U:G lesion in the DNA. Such hydrolytic damage activates the base excision repair pathway (BER), and in this instance, uracil-N-glycosylase (UNG) is expressed which recognises the U:G mismatch. UNG excises the deoxyuridine to generate an abasic site (AP-site). The AP-site is nicked by an AP-endonuclease, creating a gap. This gap would normally be repaired by a proof-reading polymerase, but during somatic hypermutation, this activity is performed by an error-prone polymerase (DNA polymerase η)[125,132,288]. This can potentially generate nucleotide substitutions at G:C base-pairs. If an exonuclease (exonuclease-1) is recruited at the nicked AP-site, the gap is extended downstream of the damaged site. This larger template allows DNA polymerase η to perform error-prone patch repair, with a chance of mutations occurring at A:T base-pairs[125].

To emulate this mechanism in bacteria, the mutator module required a nucleotide deaminase, an AP-endonuclease, a 5’-3’ or 3’-5’ exonuclease and an error-prone DNA polymerase. During somatic hypermutation, the exact mechanism by which AID activity is localised to the loci encoding the variable regions of antibodies is unknown. Outside of its natural context, previous research showed that AID’s activity can be localised to the target gene by placing it downstream of a T7 promoter and fusing AID to T7 RNA polymerase[121,120]. This orthogonal promoter-RNA-polymerase system has been shown to introduce C → T and G → A mutations in the gene-of-interest. Consequently, the AID-T7pol fusion protein was utilised as the DNA damage device in the mutator module.

The next challenge was to focus the activity of a 5’-3’ exonuclease and an error-prone DNA polymerase to the damaged DNA site, generated by AID-T7pol. An essential protein in the BER pathway is the AP-endonuclease, which generates a gap at AP-sites. By creating a 3-protein fusion comprising the 5’-3’-exonuclease, AP-endonuclease and the EP-DNA-polymerase (5’-3’Exo—AP-Endo—EP-DNA-Polymerase), it was hypothesised that the activity of all these proteins could be localised to the AP-sites generated by UNG because of AID-T7pol-mediated deamination. This 3-protein error-prone DNA repair complex with the AID-T7pol DNA damage device would form the mutator module designed to emulate somatic hypermutations in *E. coli*.

The research goal in this doctoral thesis was three-fold: (1) optimise the expression of the AID-T7pol DNA damage device to reduce its toxicity and enable long time-course mutator experiments; (2) build, test and validate 5’-3’Exo—AP-Endo—EP-DNA-Polymerase fusion proteins assembled with different biological parts for their ability to perform targeted error-prone DNA repair of damaged DNA; and (3) Perform a long time-course mutagenesis experiment with shortlisted mutator modules on a gene-of-interest, sequence the mutant gene library using NGS platforms and analyse the diversity of mutations that the mutator modules can generate.

6.2.2 Optimising the Expression of AID-T7pol in *E. coli*

Previously published expression cassettes for AID-T7pol performed targeted deamination events, but was severely toxic, which prevented *E. coli* cells from growing past ~100 cell cycles[121]. The exact reason for the toxicity was not investigated, but efforts were made to optimise the expression system for AID-

T7pol. The optimisation was achieved using a library of degenerative RBS sequences to optimise the expression of T7-RNA-pol and AID-T7pol at the translational level. Characterisation experiments with sfGFP fused to the two proteins were performed, which revealed RBS sequences that enabled high-, medium- and low-level expression of T7-RNA-pol and AID-T7pol, while being significantly less toxic to the host cell (Section 3.3). The impact of having an optimised AID-T7pol expression cassette was tested by performing a 144-hour mutator experiment where genome-integrated GFP-mut3b was targeted for mutations by AID-T7pol with UGI (Section 3.4). UGI (uracil glycosylase inhibitor) was added to this mutator module to prevent the cell's native BER pathway from repairing the U:G lesions generated by AID. With the optimised expression cassettes active, cells were able to grow at comparable rates to the wildtype DH5α control. The AID-T7pol protein was able to continually accumulate mutations in the target gene overtime. At 24-hours, only 1-2 mutations were spotted in the GFP ORF; but after 144-hours, 6-10 mutations could be seen in each GFP ORF sequenced (Figures 3.7 and 3.8). Thus, optimising the expression of AID-T7pol enabled the DNA damage device to be used in the final mutator module for *in vivo* library generation.

AID-T7pol should only be able to generate C → T or G → A mutations. This was experimentally tested using the gain-of-function experiment with inactive β-lactamase[xliii]. The gene was inactivated by substitution each nucleotide of the ATG start codon with deoxycytidine. Cells were co-transformed with plasmids containing inactive β-lactamase genes and plasmid expressing AID-T7pol+UGI. Only in the instance of A(C)G → A(T)G mutations were carbenicillin-resistant colonies spotted on LB-agar plates. After 72-hours of mutagenesis, roughly 320,000 cells out of 640 million had acquired a functional C → T mutation ($5x10^{-4}$ $cell^{-1}$). To assess the rate of targeted vs off-target mutations generated by AID-T7pol, the rifampicin reversion experiment was performed[185,186]. Bacterial cells are not naturally resistant to rifampicin, but a precise T(C)G → T(T)G mutation in the RpoB gene can provide this adaptation to cells[124]. Consequently, AID-T7pol activity generating rifampicin-resistant cells would provide a qualitative indication of its global mutagenic activity (Section 3.6). After 72-hours of mutagenesis with AID-T7pol+UGI, 343 rifampicin-resistant cells were generated out of ~800 million cells ($4x10^{-7}$ $cell^{-1}$). Compared to the β-lactamase gain-of-function experiment, the off-target mutation frequency was 1000-fold lower.

Overall, the experiments performed in this thesis were shown to optimise the expression of AID-T7pol for long time-course mutagenic experiments and illustrated the targeted activity of AID-T7pol being 1000-fold higher than its off-target mutagenic activity. All these factors should enable the mutator

[xliii] Section 2.5.2 for overview of methodology

module to maintain the host cell's fitness and genomic fidelity for performing continuous evolution of a gene over numerous cell cycles.

6.2.3 Engineering the Error-prone DNA Repair Complex

With the AID-T7pol DNA damage device optimised for use in the final mutator module, the next step was to engineer an EP-DNA-repair complex comprising a 5'-3' exonuclease, an AP-endonuclease and an EP-DNA-polymerase (Chapter 4). The different protein candidates chosen to perform these biological functions were first characterised for their expression in DH5α cells using a fixed library of 15 RBS sequences (Section 4.3). This endeavour provided a comprehensive understanding of the expression level that can be achieved, and the fitness burden imposed on cells by each RBS-Biopart combination. Once complete mutator modules are assembled with AID-T7pol and an EP-DNA-repair complex, the characterisation toolbox could be applied for optimising the expression rate of EP-DNA-repair complexes to achieve a balance between the desired mutation frequency on the gene-of-interest, and the fitness burden on cells. Such optimisation would enable replication of *E. coli* cells at an optimal rate and therefore, result in larger and more diverse protein libraries (each cell potentially expresses a unique library member).

The next step was to assemble the characterised bioparts into the 2-protien and 3-protein fusions of the EP-DNA-repair complex. The three proteins were fused such that the 5'-3' exonuclease was the N-terminal domain, the AP-endonuclease in the middle, and the EP-DNA-polymerase as the C-terminal domain. Such a fusion emulated the conserved structure of single-gene polymerases like DNA polymerase I[210]. Two EP-DNA-polymerases were tested in this study, an engineered DNA polymerase-IV (Pol-IV$^{\Delta12}$, Section 4.4) and an error-prone DNA polymerase-I (EP-Pol-I) engineered by Loeb and colleagues[138,198]. Using different combinations of 5'-3'-Exonucleases and AP-endonucleases, 11 different EP-DNA-repair complexes were assembled with Pol-IV$^{\Delta12}$, which were combined with the optimised AID-T7pol expression cassette to generate 11 mutator modules (Mut-1 – Mut-11, Section 4.5). Similarly, 8 different EP-DNA-repair complexes with the polymerase domain of EP-Pol-I were combined with AID-T7pol to generate Mut-12 – Mut-19. The mutagenic capabilities of Mut-1 – Mut-19 was tested using the loss-of-function assay with GFP-mut3b and gain-of-function assays with inactive β-lactamase genes.

The gain-of-function experiments to revert inactive β-lactamase genes using the EP-Pol-I-based Mut-12 – Mut-19 displayed a strong bias for C → T transitions. C → G and C → A transitions were rarely generated, and at a 100-fold lower frequency than C → T (Section 4.8). As mutation biases and a lack of diversity in nucleotide substitutions is not ideal for protein library generation, the EP-Pol-I-based mutator modules were discarded.

The Pol-IV$^{\Delta 12}$-based Mut-1 – Mut-11 were first tested for their mutagenic strength based on the percentage of cell populations that lost GFP fluorescence using the loss-of-function experiment. The general trend identified was that mutator modules utilising Exo-III as the AP-endonuclease caused a larger population of cells to lose fluorescence, compared to mutators utilising NAPE. This trend carried over to the gain-of-function experiments as well, where Mut-4 was the only NAPE-based mutator module that generated nucleotide substitutions in β-lactamase to generate carbenicillin-resistance (Section 4.7.1). Due to the stronger mutagenic activity and the ability to generate all 4 nucleotide substitutions being investigated (C → A, T, G and A → T), the Exo-III-based mutator modules were shortlisted for further characterisation. The key quality used to differentiate the 5'-3'-Exo-Exo-III—Pol-IV$^{\Delta 12}$-based mutators was the ability to perform patch repair, where the 5'-3'-Exo domain cleaves nucleotides upstream of U:G mismatches, creating a patch of ssDNA template to be filled by the error-prone DNA Pol-IV$^{\Delta 12}$. This mechanism is what enables mutations to be incorporated at A:T base-pairs in the gene's ORF. Only Mut-7 and Mut-9, utilising DNA Pol-I's 5'-3' exonuclease domain, were able to reliably perform patch repair to revert a premature TAA stop codon to create functional β-lactamase (Section 4.7.2). Mut-9 displayed a 10-fold higher frequency for performing patch repair than Mut-7 (Figure 4.21). Consequently, Mut-9 (EP-DNA-repair complex = 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$) and its low, medium expression variants (Mut-25 and Mut-22, respectively) were shortlisted as the final versions of the mutator module.

Using rifampicin-resistance assays, the rate of targeted mutation was shown to be 100- to 1000-fold higher than off-target mutations by Mut-9 (Figures 4.21 and 4.32). Also, the β-lactamase gain-of-function experiment was performed with appropriate control modules, which showed that targeted nucleotide substitutions and patch repair can only occur when both AID-T7pol and the 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ components are expressed in the mutator module (Section 4.10).

All the experiments performed in this chapter allowed for building and testing numerous versions of the 5'-3'Exo—AP-Endo—EP-DNA-Polymerase 3-protein fusions for their ability to perform error-prone DNA repair of the target gene. The gain-of-function experiment with β-lactamase enabled identifying mutators with EP-DNA-repair complexes that reliably generate different nucleotide substitutions and can perform patch repair. Mutator modules expressing AID-T7pol with 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta 12}$ were found to generate a variety of mutations at high frequencies, perform patch repair, and possess a targeted mutation frequency that is ~ 100-, 1000-fold greater than the off-target mutation rate. The next step was to thoroughly analyse the mutation profile that can be generated using Mut-9, Mut-22 and Mut-25. This goal was achieved using next generation sequencing.

6.2.4 Utilising NGS to Elucidate the Mutation Characteristics of the Finalised Mutator Modules

Having developed and tested the *in vivo* library generation technique comprising AID-T7pol and the 5’-3’Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex, the final aim in this doctoral thesis was to investigate the mutagenic characteristics of these mutator modules, like the diversity of mutations they can generate, the mutation frequency and the spread of mutations across a gene’s ORF. For this endeavour, a 144-hour mutator experiment was performed where the strong (Mut-9), medium (Mut-22) and weak (Mut-25) mutator modules were used in a loss-of-function experiment to generate a library of GFP-mut3b mutant DNA sequences. This mutant library was isolated from cellular genomes via PCR with the high-fidelity Phusion polymerase. This amplified DNA library was subsequently prepped and barcoded for next generation sequencing using two platforms, PacBio Sequel and Illumina iSeq100.

Two different NGS techniques were utilised to offset the unique limitations of each platform and achieve a clearer understanding of the mutator modules’ characteristics. Illumina sequencing platforms suffer from a high substitution sequencing error-rate (~ 0.1 – 0.005 per base), but a low rate of creating indel errors (10^{-4} – 10^{-5} base^{-1})[227,235,269,271]. Conversely, PacBio platforms are prone to generating insertions (rate of ~0.1 base^{-1})[227] but can have a low nucleotide substitution error-rate of ~ 1x10^{-6} base^{-1} by overlapping multiple subreads of the same DNA molecule to generate consensus sequences (CCS)[225,289]. Therefore, the alignments from Illumina reads were used to assess the mutator modules frequency of generating indels, while PacBio reads were used to accurately assess the substitution diversity and frequency of occurrence of each nucleotide substitution.

Aligned reads from both platforms showed the strong mutator possessed a mutation frequency of ~ 1x10^{-5} per sequenced base-pair per generation (sbp^{-1}g^{-1}), while the mutation frequency of the medium and weak mutators was 10-fold lower. The medium and weak mutators displayed comparable frequencies for generating each of the 12 nucleotide substitutions (~ 5x10^{-6} sbp^{-1}), while the strong mutator displayed a 10-, 30-fold higher frequency for generating C → T and G → A mutations (Table 5.4). The AID-T7pol+UGI control only performed C → T and G → A mutations; and at the A:T-base-pairs in the GFP-mut3b ORF, mutations were only generated by the mutator modules (Figure 5.5). This provided confirmation that the 5’-3’Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex is necessary for extending the gap from the U:G lesions to upstream A:T bases, creating more ssDNA template to be filled by the error-prone DNA Pol-IV$^{\Delta12}$. Therefore, with the throughput of NGS platforms, it was shown that the mutator modules comprising AID-T7pol to damage DNA and 5’-3’Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ to perform error-prone DNA repair were capable of generating a wide diversity of mutations across the complete ORF is the gene-of-interest, with a targeted mutation frequency of ~ $2x10^{-4}$ – $1x10^{-5}$ $sbp^{-1}g^{-1}$. This frequency is 100-fold higher than what is achieved by the PACE system, and on par with the mutation rate of the OrthoRep *in vivo* library generation technique.

6.3 Recommendation for Future Work

Incorporating Uracil-N-Glycosylase (UNG) in the Mutator Module

Analysis of the mutation profile generated by the strong mutator module (Mut-9) revealed a 30- to 100-fold higher frequency for generating C → T mutations compared to other nucleotide substitutions. One hypothesis for this finding is the possible bottlenecking in the BER pathway when AID-T7pol creates dC → dU deaminations. The DNA repair pathway for this lesion involves expression of uracil-N-glycosylase to recognise the deoxyuridine and cleave the pyrimidine base to generate an AP-site. If the number of U:G lesions generated by AID-T7pol are greater than what can be process by UNG per cell cycle, the mismatch would be passed onto daughter cells, causing an accumulation of C → T and G → A mutations. To test this hypothesis and potentially eliminate the bottleneck, UNG should be expressed constitutively in the mutator module. Higher number of UNG molecules per cell should result in more U:G mismatches being successfully converted to AP-sites, which can be recognised by the 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex to potentially generate mutations.

Using an Adenine and a Cytidine Deaminase in the Mutator Module

The NGS data for the strong mutator module from PacBio and all the mutators from Illumina sequencing showed a bias for C → T and G → A mutations. This result is not surprising as a cytidine deaminase was used to generate U:G mismatches, and when the AP-endonuclease generates a gap during BER, the error-rate of DNA Pol-IV is what determines if a nucleotide substitution will occur or cytosine will be readded to the damaged site. This higher frequency C → T and G → A mutations also suggest that AID-T7pol might be significantly more active in the cell, compared to the EP-DNA repair complex. If the U:G mismatches are not repaired before mitosis, one of the daughter cells would incorporate the mutated strand, resulting in an accumulation of C → T and G → A mutations over successive generations. To counteract this bias, an adenine deaminase could be added to the DNA damaging device of the mutator module. Recent research resulted in a dual base editor fused to cas9 to perform CRISPR-mediated concurrent A → G and C → T substitutions[290]. This dual base editor should be fused to T7 RNA polymerase and tested for its ability to deaminate adenine and cytidine bases.

Using Engineered T7 RNA Polymerases to Increase the Activity Rate of Deaminases

T7 RNA polymerase is a highly processive enzyme. It can process 240 nucleotides per second, which potentially provides very little time for the AID deaminase to interact with the ssDNA being released from the transcription bubble and deaminate cytidine[291]. One potential solution to increasing the number of nucleotides deaminated per transcription cycle by T7 RNA pol is to slow down its processivity. Makarova et al showed that certain mutations in the polymerase domain of T7 RNA pol can reduce is processivity by 2-to-6-fold compared to wildtype and transcribe at the rate of 40 nucleotides per second. T7 RNA Pol Mutants F644A, Q649S and G645A are ideal candidates for potentially testing this hypothesis as they have been shown to have slower processivity than the wildtype polymerase and produce stable mRNA molecules, leading to an increased yield of the protein being evolved[291].

Plugging the Mutator Module with a Selection Module to Perform Continuous Evolution of a Target Gene *In Vivo*

The goal of this doctoral thesis was to develop an mutator system capable of generating a diverse protein library inside *E. coli* cells. The experiments performed throughout this thesis resulted in a mutator module that can reliably introduce a range of nucleotide substitutions in the gene-of-interest, at a mutation frequency of ~ $1x10^{-5}$ $sbp^{-1}g^{-1}$, which is ~1000-fold higher than the off-target mutation rate. These qualities enabled performing a mutagenesis assay for 144-hours where DH5α cells stably replicated for ~ 288 generations, with mutations continually accumulating in the GFP-mut3b gene to generate a diverse genetic library. The next logical step is to combine this library creation technique with an *in vivo* selection module to test the system's ability to evolve a protein towards the desired phenotypic properties.

As a proof-of-concept, the mutator module will be tested for its ability to evolve the LuxR transcriptional activator. When the LuxR protein binds to 3-oxo-hexanoyl-homoserine lactone (C6-HSL), it undergoes a conformational change, enabling the protein to bind to the orthogonal pLux promoter and activate the expression of downstream genes[292]. Francis Arnold and colleagues utilised *in vitro* techniques like DNA shuffling and site-directed mutagenesis to generate a library of LuxR variants and evolve the protein to recognise butanoyl-homoserine lactone (C4-HSL) as the activating ligand[276]. Their research revealed three key amino-acid substitutions required to convert LuxR's ligand specificity from C6-HSL to C4-HSL.

To recreate this experiment with the *in vivo* continuous evolution system, the designed selection module would need to provide feedback to the mutator module once a C4-HSL responsive LuxR variant is achieved. The feedback mechanism would express regulatory proteins that would shut off the mutator module and prevent further mutation of the genotype linked to the desired phenotype. If C4-HSL-sensitive LuxR variants are evolved using this system, it would provide strong validation for the applicability of the AID-T7pol + 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ mutator system as a library generation technique for performing *in vivo* continuous evolution of target genes in *E. coli*.

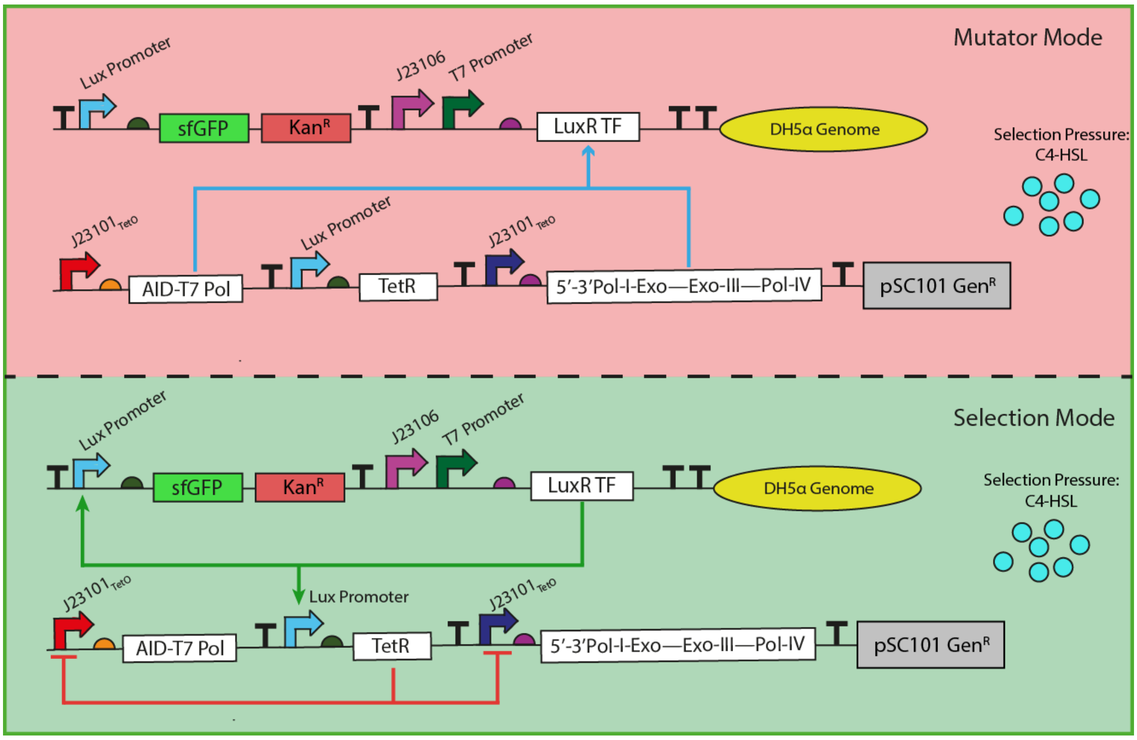


***Figure 6.1**: **SBOL schematic of the continuous evolution system to evolve LuxR** – The mutator system would operate in two modes, mutator and selection modes. The mutator mode is active when LuxR variants cannot bind to C4-HSL. AID-T7pol and the EP-DNA-repair complex are expressed which continually mutate the ORF of the LuxR transcription factor. Once a LuxR variant is achieved that can bind to C4-HSL and become activated, the circuit will go into selection mode. Mutant LuxR would express TetR to shut off the mutator module and express a fluorescent marker and an antibiotic to provide a visual signal to the user that desired protein variants have been achieved.*

6.4 Closing Remarks

In vivo continuous evolution is emerging to be a powerful technique in protein engineering that enables the user to explore vast possibilities in the evolutionary search-space to acquire proteins with desired properties. As both library generation and selection occur within a model organism, this technique is analogous to natural evolution, where mutations providing a selective advantage in a fixed environment are passed on to future generations. Such a system can run for weeks or months without human intervention, providing great scalability to the process of generating desired proteins and enzymes. PACE and OrthoRep are currently the only published continuous evolution techniques capable of generating protein libraries via random mutations across the target gene's open reading frame. PACE has a low mutation rate and a non-targeted mutator system, while OrthoRep's application is limited to yeast. This created scope for developing a targeted *in vivo* mutator system for generating protein libraries in *E. coli*.

The work presented in this thesis was aimed at developing a mutator system in *E. coli*, designed to generate protein libraries by emulating the process of somatic hypermutations, found in immunoglobulin cells for generating antibody diversity. To emulate this process, a DNA damage device comprising AID fused to T7 RNA polymerase was optimised for expression in *E. coli*, such that extensive protein libraries can be generated from mutating the gene-of-interest over multiple days, with minimal mutagenic activity on the host cell genome. The targeted DNA damage device was combined with an engineered error-prone DNA repair complex comprising a 3-protein fusion designed to hijack the base excision repair pathway, and using an error-prone DNA polymerase, incorporate mutations at and around the site of damage. Using β-lactamase as the gene-of-interest and RpoB as the off-target gene, the mutator system was shown to generate mutations in the gene-of-interest at a 1000-fold higher frequency than off-target genes. Using next generation sequencing platforms, a library of GFP-mut3b variants generated by the mutator module were analysed to elucidate its mutational characteristics. The mutator module comprising AID-T7pol and the 5'-3'Pol-I-Exo(s)—Exo-III—Pol-IV$^{\Delta12}$ EP-DNA-repair complex was shown to generate each of the 12 possible nucleotide substitutions, at an overall frequency of 1x10^{-5} sbp^{-1}g^{-1}, which is comparable to the mutation frequency of the OrthoRep system.

Therefore, the findings in this doctoral thesis demonstrate a functional *in vivo* mutator system that can preferentially mutate the gene-of-interest over 100s of continuous bacterial cell cycles and generate a diverse library of protein variants resulting from a diverse range of nucleotide substitutions and functional indels. This mutator system now needs to be tested for its applicability in a continuous evolution system by evolving a protein towards desired physical and chemical properties.

Chapter 7: Materials and Methods

7.1 Materials Used in Experiments

7.1.1 Bacterial Strains

Four different strains of *E. coli* were used: DH5α, BL21(DE3), MG1655 and GM31. DH5α was primarily used throughout this report; being used for the storage of the BASIC DNA parts as glycerol stocks, for the RBS characterisation assays and for the testing of the mutator modules in the gain-of-function workflow. BL21 was used to test the expression of the inactivated β-lactamase gene placed downstream of the T7 promoter in the target plasmids. Once the gene variants were confirmed to be inactive, the gain-of-function assays in Chapter 3 and Chapter 4 were performed in the DH5α strain. MG1655 was used as a template for cloning parts like RecJ, RecE and 5'-3'-DNA-Polymerase-I-Exonuclease from the genome via colony PCR. GM31 strain was used for the loss-of-function assays for testing the mutagenic capability of the mutator modules constructed assembled in Chapter 4. A P_{T7}—GFP-mut3b—T7-terminator cassette was integrated into the GM31 genome using the CRIM genome integration method[293]. This strain is deficient in the Dcm methylase[294].

Table 7.1: List of bacterial strains used in this study

Strain	Genotype	Antibiotic Resistance	Source
DH5α	(K-12 strain, F– *endA1 glnV44 thi1 recA1 relA1 gyrA96 deoR nupG purB20* ϕ80d*lacZ*ΔM15 (*lacZYA-argF*)U169, hsdR17($r_K^- m_K^+$), λ^-)	None	NEB, #C2987H
MG1655	(K-12 strain, F– λ– *ilvG– rfb-50 rph-1*)	None	Laboratory stock
BL21(DE3)	(B strain, F−*omp*T *hsd*SB (rB–, mB–) *gal dcm* *(*DE3))	None	NEB, C2527H
GM31	F- thr-1 araC14 leuB6(Am) fhuA31 lacY1 tsx-78 glnX44(AS) galK2(Oc) galT22 λ- dcm-6 hisG4(Oc) rpsL136 xylA5 mtl-1 thiE1	Streptomycin	Laboratory stock

7.1.2 Media for Bacterial Growth

All media for bacterial growth were autoclaved at 121°C. Media and buffers containing salts with multiple oxidative states were sterilised using 0.22 µm pore filter paper prior to use.

Table 7.2: List of bacterial growth media used in experiments

Medium	Composition	Source
LB (Luria-Bertani Broth)	25 g/L LB powder	ForMedium, #LMM0104
LB Agar	25 g/L LB powder, 10 g/L agar	Formedium, #LMM0204
SOC	31.5 g/L SOC powder	ForMedium, #SOC0202

7.1.3 Antibiotics used in Experiments

All antibiotics were dissolved in specific solution at the following stock concentration and filter sterilized (0.22 µm) prior to use. The stock antibiotics were kept in the -20°C. To achieve working concentration, the antibiotics were diluted 1000- to 2000-fold in fresh media.

Table 7.3: Antibiotics utilised in the experiments

Antibiotic	Solvent	Stock Concentration	Working Concentration
Carbenicillin	dH_2O	100 mg/ml	50 µg/ml
Kanamycin	dH_2O	50 mg/ml	50 µg/ml
Chloramphenicol	Ethanol	25 mg/ml	25 µg/ml
Gentamycin	dH_2O	15 mg/ml	15 µg/ml
Rifampicin	DMSO	50 mg/ml	50 µg/ml

7.1.4 Dyes, Enzymes and Reagents

Table 7.4: List of reagents and enzymes used in experiments

Reagents and Enzymes	Commercial Supplier and Catalogue Number	Experimental Purpose
SYBR Green	Invitrogen, #S7563	Agarose gel electrophoresis
6x loading buffer	NEB, #B7024S	Agarose gel electrophoresis
1 kb DNA ladder	NEB, #N3232L	Agarose gel electrophoresis
100 bp DNA ladder	NEB, #N3231S	Agarose gel electrophoresis
Magnetic beads (AMPureXP)	Beckman Coulter, #A63881	BASIC DNA Assembly
BsaI-HFv2	NEB, #R3733L	Restriction digests
T4 Ligase	Promega, #M1804	Ligation reactions
Pfu polymerase and buffer	Produced in-house	Colony PCR, Diagnostic PCR
Phusion polymerase and buffer	NEB, #M0530S	Site-directed mutagenesis, Illumina and PacBio RSII NGS sample preparation
Taq polymerase and buffer	NEB, #M0267S	Error-prone PCR
dNTP molecules	Sigma Aldrich dATP, #11934511001 dCTP, #11934520001 dGTP, #11934538001 dTTP, #11934546001	PCR
Qubit dsDNA High Sensitivity Assay Kit	ThermoFisher, #Q32854	DNA quantitation
PreCR DNA Repair Mix	NEB, #M0309S	Repair DNA samples intended for next generation sequencing
Exonuclease III	NEB, #M0206S	Removing primer and linear DNA from PCR mix
Illumina iSeq 100 i1 Reagent Kit (iSeq cartridge, i100 flow cell, and reagents)	Illumina, 20021533	Next Generation Sequencing
PhiX Control v3	Illumina, FC-110-3001	Next Generation Sequencing
Nextera XT DNA Library Preparation Kit (24 samples)	Illumina, FC-131-1024	Next Generation Sequencing

7.1.5 Chemical Inducers for Inducible Promoters

Stock solutions of chemical inducers were sterilised using 0.22 µm filter paper and stored at -20°C. The stock concentrations and solvents used can be found in Table 7.4.

Table 7.5: List of chemical inducers used in experiments

Inducible Promoter	Inducer Molecule	Solvent	Stock Concentration
P_{Tet}	aTc	DMSO	100 µg/ml
P_{Lac}	IPTG	dH_2O	1 M
P_{Lux}	3OC6-HSL	Dimethyl formamide (DMF)	10 mM

7.1.6 Competent Cell Preparation

Depending on the level of competency needed in terms of colony formation units, two different protocols were used to make the strains chemically competent for transformation.

1. **Inoue Chemo-competent Cells**

Bacterial transformation was done via heat-shock treatment. The DH5α competent cells used are prepared in-house using the Inoue Chemo-competent Cells protocol[295]. This protocol utilises a nutrient-rich media (SOC) for the cell growth and they are grown at 18°C and shaken at 220 RPM. This ensures that more of the cells that are captured will be in the log phase and in the same stage of the cell cycle. Along with SOC media, this protocol requires the Inoue transformation buffer (ITB) and PIPES. ITB and PIPES are prepared as follows:

ITB Buffer Prep		**1 l**	**300ml**
$MnCl_2 4H_2O$	55 mM	10.88 g	3.27g
$CaCl_2$ $2H_2O$	15mM	2.2g	0.66g
KCl	10mM	18.65g	5.7g
H_2O		**980 ml**	**294 ml**

ITB can be stored at 4° C for upto 3 months.

PIPES pH 6.7:	**Concentration**	**Quantity**
	0.5M	100ml
PIPES		15.1 g
H_20		80 g

- Adjust pH to 6.7 with KOH for solving it

H_2O		Adjust to 100ml

Methodology: Vials of the bacterial strains were purchased from New England Biosciences (NEB). Under sterile conditions, the cells were streaked on an agar plate without antibiotics. After 20-hours, an individual colony was picked into 300 ml of SOC media under sterile conditions. The inoculated SOC media was incubated at 18°C and shaken at 220 RPM until the cell culture reached an OD_{600} between 0.3 and 0.4. The cell cultures were kept on ice for the remainder of the protocol. The cells were spun down using a centrifuge and resuspended in ITB buffer with PIPES. The prepped cells were subsequently divided into 200 ul aliquots in microcentrifuge tubes and stored in -80°C for future use. Using this protocol, cells achieve a transformation efficiency of ~ $1x10^9$ $CFUml^{-1}$.

2. **Calcium Chloride Competent cells**

This protocol utilises 0.1M $CaCl_2$ and 0.1M $CaCl_2$ + 15% glycerol. On the first day, a 5 mL overnight culture is setup by picking one colony from a LB-agar plate under sterile conditions. On the second day, 25 mL of fresh LB is inoculated with 250 µl of the overnight culture, incubated at 37°C and shaken at 220 RPM. Once the cultures reached an OD600 of 0.4, the cells were spun down and resuspended in 0.1M $CaCl_2$. The cells are spun down again and resuspended in 0.1M $CaCl_2$ + 15% glycerol for storage in -80°C. Using this protocol, cells achieve a transformation efficiency of ~ $1x10^7$ $CFUml^{-1}$.

7.2 Commercial Kits for Cell Culture Prep and Agarose Gels

7.2.1 Purifying plasmids and genomic DNA from Bacterial cell cultures

Bioparts were stored in pUC-Ampicillin^R high copy number plasmid vectors as glycerol stocks in DH5α cells. Isolating these bioparts for assembling the mutator modules and expression plasmids required use of bacterial cell culture prep kits. The E.Z.N.A Plasmid isolation kits were purchased from Omega BIO-TEK.

Glycerol stocks of Bioparts and assembled plasmids were streaked on LB-agar plates containing the appropriate antibiotic. A colony is inoculated into the recommended volume of LB media with the appropriate antibiotic and incubated at 37°C overnight. The liquid cultures were treated with Solution I, II and III according to the protocol provided by Omega BIO-TEK. Depending on the elution volume, the prepped plasmid was isolated within a concentration range of 100-500 $ng\mu l^{-1}$.

7.2.2 Imaging and Purification of DNA from Agarose Gels Post Electrophoresis

DNA products resulting from restriction digests or PCR reactions were separated using 1% agarose gel prepared in 1x TBE buffer. 5 µl of sample was combined with 2 µl purple loading dye (6x, NEB) and 1 µl SYBR Green (10x stock, Invitrogen), and then loaded into the wells. The agarose gels were run at 100V for 90 minutes. An appropriate DNA ladder (100bp or 1 kb, NEB) was used for size determination of the DNA products. SYBR Green dye was used for DNA visualisation and images of the gels were captured using a Fuji LAS-3000 imaging system.

DNA fragments were extracted from agarose gel pieces using the Gel Extraction Kit developed by Omega BIO-TEK. The protocol utilising XPS buffer was used as advised by the manufacturer.

7.2.3 DNA Quantitation for Sequencing and Downstream Applications

DNA quantification of miniprepped plasmids for sequencing and general laboratory use was performed using the NanoDrop OneC (Thermo Fisher Scientific), in accordance with the manufacturer's instruction.

For more delicate downstream applications, which required highly accurate quantification of the DNA samples, the Qubit fluorometric quantitation method was used[296]. The high sensitivity dsDNA kit was utilised for quantifying the mutant DNA library to be sequenced using Illumina iSeq100 and PacBio Sequel next generation sequencing platforms. The appropriate reagents and instrument settings were used, based on the manufacturer's instructions.

7.3 Plasmid Construction – BASIC DNA Assembly

7.3.1 Preparation of the BASIC Linkers

Lyophilised BASIC DNA linkers were synthesised by IDT and Biolegio. The BASIC DNA linkers synthesised through IDT consist of 2 separate fragments for each linker section: an adaptor (smaller fragment) and linker (longer fragment). The lyophilised BASIC DNA linkers were then eluted into TE buffer as the recommended stock solution. The BASIC DNA linkers synthesised from Biolegio consist of a combined linker (small and long parts) for each linker section. The lyophilised BASIC DNA linkers from Biolegio were then eluted into 200 ul of linker annealing buffer provided as working solution. To facilitate the annealing process, both working solutions obtained from IDT and Biolegio then heated up at 95°C and after 5 min allow them cool down to room temperature. The stock and working solution were stored at -20°C until next use for up to 3 months.

7.3.2 Three-Step BASIC Assembly Reaction

All DNA assemblies were performed using the Biopart Assembly Standard for Idempotent Cloning (BASIC), developed in the Baldwin lab. The assembly method involves the use of orthogonal linkers that produce a 21 base-pair (bp) overhang[146]. These linkers can also be designed to be functional. Some of the linkers in our library have RBS or BsaI restriction sites coded into them. Currently there are 36 different linkers in our library (7 orthogonal linkers, 3 fusion linkers, 45 RBS-encoded and 2 BsaI restriction-site encoded linkers), allowing for a large multi-part assembly[xliv].

The assembly method is divided into three steps, linker ligation, magnetic bead DNA purification and parts assembly.

1. Linker ligation: This experimental step involves ligating appropriate prefix and suffix linkers to each Biopart that is to be assembled into a genetic circuit. This reaction utilises the BsaI restriction enzyme (ordered from NEB) to cut and release bioparts from the storage plasmid and the T4 DNA ligase (ordered from Promega Corporation) to attach the BASIC prefix and suffix linkers to the Biopart via the complimentary sticky ends. This reaction utilises the T4 ligase buffer (30mM Tris-HCl, 10mM $MgCl_2$, 10mM DTT and 1mM ATP at pH 7.8). 30 µl reactions are setup for each Biopart with the following composition:

Table 7.6: BASIC Linker Ligation Reaction Composition

Reagent	Volume
dH_20	22.5 µl
Promega T4 buffer (10x)	3 µl
Prefix Linker	1 µl
Suffix Linker	1 µl
BASIC Biopart (at 200 ngul^{-1})	1 µl (or 50 ng per 1kb PCR product)
NEB BsaI-HF v2 enzyme (R3733) 20 U/µl	1 µl
Promega T4 ligase (M1801) 1-3 U/µl	0.5 µl
	Mix by pipetting up and down

The reactions are setup in PCR tubes and placed in the thermal cycler with the following conditions:

Temperature	Time	
37°C	2 min	**X 25 cycles**
20°C	1 min	
55°C	5 min	
4°C	store	

[xliv] Appendix 9.1 for Linker sequences and description of function

2. Magnetic Bead Purification of Linker Ligated Bioparts

SPRI beads were used for the purification of linker ligated bioparts from the reaction mixture. The magnetic bead solution was mixed with the bioparts in a 1.6:1 ratio. Therefore, each 30 µl sample of linker-ligated biopart solution was mixed with 48 µl of the bead solution. The clean-up reaction requires a U-bottom 96-well plate and 96-well magnetic-ring stand.

The SPRI beads are paramagnetic and form a ring in the U-bottom 96-well plate once its placed on the magnetic-ring stand. The presence of polyethylene glycol (PEG) enables the SPRI beads to reversibly bind to the DNA molecules[297,298]. The following steps were utilised for the protocol:

Prepare fresh 70% EtOH (0.5 ml per BASIC reaction) and bring magnetic beads (AmpureXP or Ampliclean) stored at 4°C back into homogeneous mix by shaking thoroughly.

We recommend using a 96-well U-bottom 96-well plate in combination with an Amgen magnetic-ring stand for quick magbead immobilisation and easy pipetting access.

1. Add 48 µl of magnetic beads into 96-well U-bottom plate (one well per linker-ligated biopart) and add the 30 µl linker ligation solution from the PCR machine step; mix by pipetting 10 times.
2. Wait 5min to allow DNA molecules to bind to the SPRI beads.
3. Place 96-well plate on magnetic-ring stand and wait for rings for the beads to form and the solution to become clear.
4. Remove the supernatant from the centre of each well, to prevent disrupting the ring.
5. Add 150 µl of 70% EtOH to each well and wait 30 seconds.
6. Remove the ethanol from each well.
7. Add 150 µl of 70% EtOH to each well and wait 30 seconds.
8. Remove the ethanol from each well.
9. Leave the plate to dry for 1-2 min. This is to ensure all the ethanol evaporates, as residual ethanol can impede with downstream applications.
10. Remove 96-well plate from the magnetic stand and resuspend the DNA-bound SPRI beads in 30 µl of dH_20.
11. Wait 1 minute for DNA to unbind from the beads.
12. Place 96-well plate back on magnetic stand and allow the ring to form and the solution to become clear.
13. Pipette 30 µl of H_20 with eluted DNA into fresh 1.5 ml microcentrifuge tubes for direct use in DNA assembly or storage at -20°C for up to 1 month.

3. Biopart Assembly Step: In this step, each of the bioparts are mixed in an equal ratio in a buffer containing potassium and magnesium salts. These salts ensure efficient annealing of DNA molecules via the complimentary BASIC linker overhangs. We used the NEB Cutsmart buffer (50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate and 100 µg/ml BSA at pH 7.9) for performing our Biopart assembly reactions.

The reaction volumes are as follows:

Table 7.7: BASIC Parts Annealing Reaction

Component	Amount (ul)
DNA part	1 for each part
NEB Cutsmart	1
dH2O	Upto 7
Total	10
	Mix by pipetting

The reaction is subsequently incubated at 50°C for 45 minutes to enable the annealing process. The assembled plasmid is then transformed into the bacterial strain of choice for amplification and storage.

7.3.3 Genome Integration with BASIC and One Step Integration Plasmids (pOSIP)

GFP-mut3b was integrated into the DH5a genome using the pOSIP method[299]. The integration method uses a phage λ integrase, which interacts with its bespoke attachment sites — attB in bacteria and attP in phage. The phage-specific attP site can be cloned into a bacterial plasmid along with the integrase expression system, enabling cargo DNA on the plasmid to be inserted into the bacterial genome at the attB site. These OSIP vector plasmids have two functional modules, the integration/propagation module, and a cargo DNA module. The integration/propagation module contains the site-specific recombinase gene, under the control of the λ CI repressor, and the attP attachment sequence. The integration module is flanked by FRT sites, which bind to the yeast FLP recombinase[300]. Therefore, once the cargo DNA and the integration module are integrated into the bacterial genome at the attB site, FLP is expressed, which removes the integration module from the genome by circularising the DNA contained within the FRT sites.

The pOSIP integration method was optimised for use with the BASIC assembly protocol. The DNA sequences corresponding to attP, λ integrase, FRT sites and FLP integrase were synthesised as BASIC-ready bioparts — with BsaI recognition and cutting sites flanking the 5' and 3' ends. The bioparts were assembled into functional OSIP vector plasmids with methylated BASIC linkers flanking the integration module. These methylated linkers enable downstream modular assembly of the OSIP vector with the cargo DNA and an antibiotic marker in a two-part DNA assembly reaction using the protocol described in 7.3.2. Successful integration of the cargo DNA in the bacterial genome was verified by high-fidelity PCR and Sanger sequencing of the PCR products.

7.4 Sequence Verification of Bioparts and Assembled Plasmids

Bioparts and assembled genetic circuits were verified via Sanger Sequencing, performed by Source BioScience. The company required the samples to be delivered at a concentration of 100 ngul^{-1} along with a 5 µl sample of each primer for each sequencing reaction to be performed.

The sequencing reads generated by Source BioScience were 600-1000 nucleotides long. Bioparts that are smaller than 1000-nucleotides in length were verified using a standard forward and a standard reverse primer, designed to bind to the Prefix and Suffix sites of BASIC assembly linkers, flanking the bioparts. Bioparts longer than a thousand nucleotides were verified using the standard forward and reverse primers, and unique primers designed to bind within the ORF of the gene-of-interest.

7.4.1 Generating Sequence Alignments

The sequencing files were uploaded onto the open source web-tool Benchling and its native sequence alignment tool was used to align the sequencing reads to the bioparts and assembled plasmids. This ensured genetic fidelity of the bioparts used in our project. If DNA sequencing revealed the Biopart to contain an SNP, the complete gene fragment was reordered using the gBlock Gene Fragments service from Integrated DNA Technologies (IDT), Europe or amplified again from purified bacterial genome via high fidelity PCR with Phusion polymerase.

7.4.2 Primers for Sequencing and PCR

Primers for Sanger Sequencing and for performing PCR protocols were ordered from Integrated DNA Technologies (IDT), Europe. Primers to be used for Sanger Sequencing and general PCR experiments (like troubleshooting and ORF amplification) were ordered with standard desalting for purification, which ensures that ~70% of the molecules are the correct length with no nucleotide truncation at the 5’-end.

Primers for EP-PCR, colony PCR and site-directed mutagenesis were ordered with HPLC purification. This purification method ensures ~90% of the nucleotide molecules are the correct length with no 5’ truncation[301]. This is crucial for experiments that are dependent on accurate primer design to achieve the desired DNA product[xlv].

7.4.3 *In Silico* Assembly of Plasmids Constructed for Experiments

The BASIC bioparts used in our work can be found in the Baldwin Laboratory Benchling inventory; access to which is available upon request. The sequence files for bioparts (.xdna) were generated from Serial Cloner. To generate a full sequence of the assembled plasmids, *in silico* plasmid assembly were done from individual '.xdna' file using a custom Python script (Python 2.7) written by Dr Marko Storch[xlvi].

7.4.4 Verification of Plasmids and DNA Assemblies Using BsaI Digests

Bioparts in the BASIC format are stored in plasmids with a pUC-Amp^R backbone with LMP and LMS methylated linkers flanking the biopart open reading frame. Similarly, assembled expression cassettes for AID-T7pol, Tet repressor, UGI, the error-prone DNA repair complexes and the inactive B-lactamase genes were flanked by the methylated linkers containing functional BsaI restriction enzyme sites. This enabled quick verification of successful BASIC assembly reactions using a BsaI digest, which would produce two bands in the gel — one for the ORI-antibiotic backbone and another for the biopart/assembled expression system. Comparison of the experimental digest to *in silico* digests performed on Benchling enabled quick verification of successful DNA assembly. The plasmids were then further verified for genetic fidelity via Sanger sequencing.

75 – 100 ng of the target plasmid was added to a 10 µl reaction mixture with 1X NEB Cutsmart buffer and 1 µl of BsaI enzyme. The mixture was incubated at 37°C for 1-hour. The samples were subsequently loaded onto a 1% agarose gel with 2 µl of 6X loading dye, 1 µl of 10X SybrGreen and 100ng of either 1 kbp or 100 kbp DNA ladder. The electrophoresis reaction was performed for 90-minutes at 1X TBE buffer and at static 100 Volts. The gels were then imaged using the FujiFilm LAS-3000 luminescent analyser. Comparison between the gel image and the *in silico* digest performed on Benchling enabled assessment of correctly sized bands.

[xlv] Appendix 9.3 for full list of primers used in this study

[xlvi] Appendix 9.4 for Python script

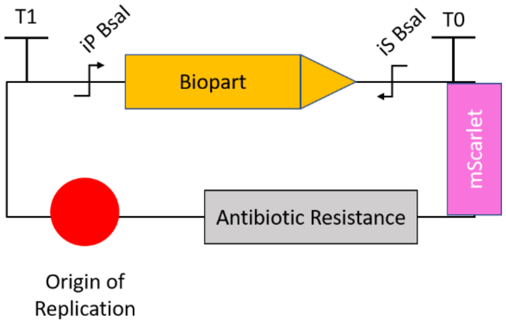
7.5 Bacterial Transformations

A standard protocol was applied for performing bacterial transformations[302]. 40 ul of competent cells are mixed with 5 ul of the assembled plasmid. The mixture is incubated on ice for 20 minutes. This is followed by heatshock treatment in a waterbath for 45 seconds at 42°C, then back on ice for 2 minutes. 400 ul of autoclaved SOC media is added to the transformed cells and the samples are incubated at 37°C for one hour. The samples are then streaked on aa LB-agar plate containing the required antibiotic.

7.6 Applying the SEVA Architecture for Plasmid Construction

The SEVA (Standard European Vector Architecture)[303–305] library contains an array of antibiotic resistance genes and origins of replication that enable standardised assembly of plasmids with unique 'cargo' DNA, ie, gene sequences and genetic circuits. The antibiotic resistance and origin of replication bioparts contain a terminator sequence upstream and downstream of their coding regions respectively. The plasmid is assembled such that these transcription terminator sites (T0 and T1) flank the cargo DNA. This ensures minimal leaky expression of the cargo DNA when the RNA polymerase expresses the antibiotic resistance gene and prevents the need to manually add terminator sequences into the genetic circuit being assembled.

All the SEVA bioparts were optimised for use in BASIC assembly by flanking the DNA sequences with iP and iS sites containing BsaI recognition and cut sites. The bioparts were stored in high copy number Ampicillin-pUC plasmids to achieve high concentration of plasmid DNA from the plasmid DNA prep kit mentioned in Section 6.2. Incorporating the mScarlet chromaprotein into the Amp-pUC backbone of the storage plasmids allows for easy selection of correctly assembled plasmids over false positives from an agar plate. The false positive colonies would display a pink colour on the plates[306], while colonies with the correct plasmid appear colourless.



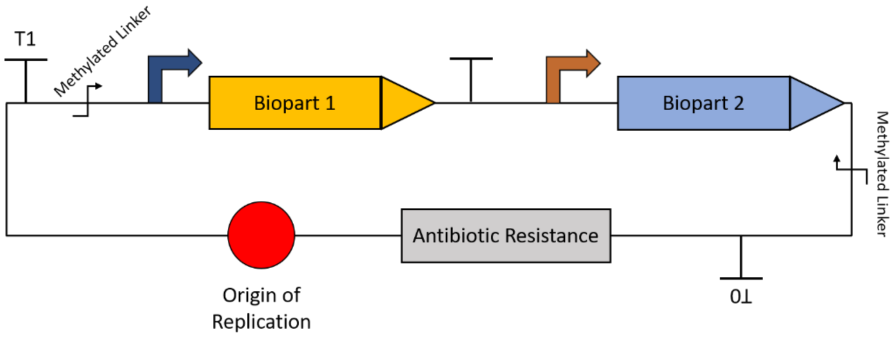


Figure 7.1: SBOL schematic highlight the Standard European Vector Architecture. *The SEVA arcitecture contains a library of antibiotic and ORI parts that are assembled with the cardo DNA such that terminators T0 and T1 flank the cargo.*

7.6.1 Storing PCR amplicons and in silico synthesised bioparts in pUC-Ampicillin

The primers for PCR amplifying parts from various sources were designed to add the iP sequence upstream and iS sequence downstream of the Biopart. PCR was done using either the Phusion or PFU polymerase. These parts were then cloned into the SEVA architecture pUC-Ampicillin^{R}-mScarlet storage plasmid. iP and iS contain BsaI restriction sites, which allows the parts to be released from the storage plasmid during the assembly protocol and ligate to the BASIC linkers.

iP: TCTGGTGGGTCTCTGTCC iS: GGCTCGGGAGACCTATCG

7.7 Plate Reader Experimental Conditions for Measuring Cell Culture Growth and Fluorescence

The time-course experiments for measuring the cell culture growth and the corresponding fluorescence of said cultures were performed in 96-well plates with biological triplicates for each construct of the samples. *E. coli* DH5α cells carrying an empty backbone (containing simply an origin of replication and an antibiotic resistance gene) were used as a negative control. For the characterisation experiments, after assembling the genetic circuits and transforming them into *E. coli*, three colonies were picked for each genetic construct into a 96-well plate containing LB media and the appropriate antibiotics. The colonies were grown for 18-hours at 30°C and 600 rpm in a benchtop plate shaker (Mikura shaker). The cultures were then diluted 200x (10 µl into 190 µl then 10 µl into 90 µl) in LB media using an automated liquid handling robot (CyBio Felix) or manually via a multi-channel pipette. The LB media contained the appropriate antibiotics and inducer molecules.

The 96-well plate was then incubated in a microplate reader (Clariostar, BMG Labtech) with continuous shaking at 30°C and 600 rpm shaking for 6-8 hours. Absorbance measurements for cell culture confluency and fluorescence measurements were programmed to be collected every 15-minutes. Absorbance values were recorded at 600 nm with orbital shaking of the plate. Appropriate chromatic filters were utilised for measure the emission spectra of the various fluorescent proteins used in the experiments: GFP fluorescence (F:482-16/F:530-40), RFP fluorescence (F:540-20/F:590-20).

7.7.1 Data Processing for the Plate Reader Experiments

The absorbance and fluorescence data from plate reader experiments were processed using a custom MS Excel sheet designed by Dr Geoff Baldwin. The absorbance and fluorescence values recorded at 15-minute intervals for 25-30 cycles are copied into the custom Excel sheet. The absorbance values were corrected by subtracting the optical density of blank wells containing only growth media, while the fluorescence values were corrected by subtracting the auto-fluorescence of the negative control (bacterial cells expressing an antibiotic-ORI backbone and no fluorescent protein). The Excel sheet automatically combined the data from biological triplicates for each sample to give an average absorbance and fluoscence value at each time point. The relative fluorescence (FL/OD) for each sample was then calculated.

7.8 Flow cytometry Settings and Data Analysis

During this project, the Attune NxT was used for flow cytometry measurement of cell samples. The flow cytometer settings for the Attune NxT were: 500 volts for forward scattered (FSC), 420 volts for side scattered (SSC), 550 volts for yellow light (YL), and a threshold of 0.3x1000. In total 10,000-50,000 events were collected for each sample depending on the optical density of cell cultures before diluting into PBS. The data was stored as FCS 3.0 datafiles and analysis was done using FlowJo V10.

Single cell population gating was performed after plotting FSC-H against SSC-H and histograms of channel of fluorescences. The outputs from the Attune NxT were using BL1 and YL2. The chosen gating covered 90-99% of the total cell population. The fluorescence intensity of the sample was calculated by subtracting the geometric mean of the sample with auto-fluorescence from the negative control (cells carrying backbone-only plasmid).

7.9 PCR Reactions

7.9.1 Colony PCR

Colony PCR was performed to isolate the bioparts RecJ, RecE, DNA polymerase I, exonuclease III and DNA polymerase IV from the genome of MG1655 strain of *E. coli*. The T7 RNA polymerase was isolated from the BL21(DE3) strain. The DNA sequences for the bioparts were identified from the Uniprot database and appropriate forward and reverse primers were designed to amplify the bioparts from the bacterial genome. The primers were designed to with a melting temperature of 60°C and the forward primers contained the iP sequence (TCTGGTGGGTCTCTGTCC) in the overhang, while the reverse primers were designed with iS (GGCTCGGGAGACCTATCG) in the overhang. The amplified bioparts could subsequently be cloned into the pUC-Amp backbone via BASIC DNA assembly and used as bioparts. For the PCR reaction, PFU polymerase stocks prepared in house were used.

Experimental steps:

Bacterial cells from a glycerol stock were streaked onto a LB-agar plate under sterile conditions. Bacterial cultures can also be spread on plates after serial dilution of the liquid culture. The LB-agar plates were incubated at 37°C for 18-24 hours. Individual colonies were picked and diluted in 20 ul of deionised water (dH$_2$O). The following reaction mixture was setup with the diluted cells for colony PCR:

Table 7.8: Colony PCR with PFU Polymerase

Component	**Reaction Volume**		
	25µl	**50µl**	**100x25µl=96well**
dH$_2$0	19	40	1900
PFU Buffer	2.5	5	250
FOR Primer 10µM	0.25	0.5	25
REV Primer 10µM	0.25	0.5	25
dNTPs (10 mM)	0.5	1	50
Diluted colony mixture	2	2	2
PFU Polymerase	0.5	1	50
Total	**25**	**50**	**2500**

The samples are then placed in the thermal cycler with the following settings:

Temperature	**Time**	
98°	10 min	
98°	30s	**30 cycles**
Primer specific Tm (~ 60°C)	30s	
72°	2min / 1kb PCR product	
72°	10 min last elongation step	
4°	Forever	

After the PCR cycle, 10 ul of the sample is mixed with 2 ul of 6x loading dye and 1 ul of 10x SybrGreen. This mixture is loaded onto a 1% agarose gel and ran in 1% TBE buffer for 90 minutes at 100V to check for the correct size bands. The gels were imaged using the Fujifilm Las-3000 luminescent image analyser. If the bands corresponding to the target DNA sequence were of the correct size, the DNA sequences were purified using the gel extraction protocol described in 7.2.2. The isolated parts were subsequently cloned into pUC-Amp^R plasmids for storage and verification of genotype fidelity via Sanger sequencing.

7.9.2 High Fidelity PCR with Phusion Polymerase

Site-Diected Mutagenesis

Phusion polymerase is a high-fidelity polymerase with an error-rate of 4.2x10^{-7} per bp amplified[307]. This polymerase was used in site-directed mutagenesis PCR to generate the error-prone variants of DNA polymerase I, EP-Pol-I^{46}, EP-Pol-I^{150} and EP-Pol-I^{1100}. The primers for site-directed mutagenesis were designed to bind upstream and downstream of the site of mutation, with the mutation being present in the overhang of both primers. Primers were phosphorylated at the 5' ends to promote circularisation of the amplicons after treatment with T4 ligase. Primers were HPLC purified to ensure over 90% of the synthesised primer population was the correct length. Oligonucleotides are synthesised by adding single nucleotides in the 3' → 5' direction via solid-phase synthesis cycles, where coupling efficiency greatly affects the length and yield of primers[308,309]. Truncated primers may not contain the desired mutation in the overhang region, therefore HPLC is necessary to ensure the oligonucleotides are of the correct length. After the mutagenesis PCR, DpN1 was added to the reaction mixtures and incubated at 37°C for 1-hour. DpN1 degrades methylated DNA, ensuring the template plasmids are removed from the reaction mixture, leaving only the mutated PCR amplicons.

Amplification of the mutant GFP library from DH5α genome

The loss-of-function experiments with the mutator modules to target mutations to the genome-integrated GFP expression cassette generated a library of GFP variants with mutations. The bacterial genomes were isolated from cells using the genomic DNA prep kit from OmegaBiotek. High-fidelity PCR was performed to isolate the mutant GFP expression cassettes from the prepped pool of bacterial genomes. The amplified sequences were purified using the gel extraction kit and subsequently prepped for Illumina and PacBio next generation sequencing by ligating appropriate barcoded adapters for each sequencing method.

Experimental method

For both site-directed mutagenesis and amplification of mutant DNA library, a standard reaction mixture was created.

Table 7.9: Phusion Polymerase PCR Reaction

Reaction Component	**Total Volume (µl)**	
	1 x 50 µl	8 x 12 µl (Gradient Melting Temperature)
H20	31.5	63
5x HF buffer	10	20
10mM dNTPs	1	2
Primer A 10uM	2.5	5
Primer B 10uM	2.5	5
Template Plasmid or Genomic DNA (500pg/ul)	2	4
Phusion (2U/ul)	0.5	1
Total	50	96

The reaction mixtures were incubated in a thermal cycler with the following conditions:

Temperature	**Time**	
98°	3 min	
98°	10s	25 – 30 Cycles
Prime-specific Tm (~ 60°C)	30s	
72°	15-30s / 1kb plasmid	
72°	10 min last elongation step	
4°	forever	

For verification of the amplicon size, they were loaded on agarose gels and electrophoresis was performed with identical conditions to colony PCR (7.9.1). The gels were imaged using the Fujifilm Las-3000 luminescent image analyser. If the amplicon DNA were of the correct size, DNA from the remaining reaction mixture was purified using SPRI magnetic bead purification (as described in 7.3.2).

For site-directed mutagenesis performed on plasmid DNA, the purified amplicons were incubated in Promega T4 ligase buffer with 0.5 µl of the ligase at room temperature for 30 minutes. The circularised plasmids were subsequently transformed into DH5α cells.

For the mutant GFP library, the purified linear DNA was stored at -20°C, until downstream steps to prepare the samples for next generation sequencing.

7.10 Loss of Function Assays with GFP

Screening the Mutagenic Capability of Mutator Modules

Two different bacterial strains were integrated with an expression cassette for GFP. The GM31 strain was developed in previous research, where P_{T7}—GFP-mut3b—T7-terminator was integrated into the genome using the CRIM integration method[121,293]. This strain was used in the experiments performed in Chapter 3 and Chapter 4 for testing and screening of the AID-T7pol DNA damage device and the library of error-prone DNA repair complexes. The $GM31^{GFP}$ strain of cells were made competent using the $CaCl_2$ protocol and stored in -80°C for use. The mutator plasmids were transformed into $GM31^{GFP}$ cells and plated on LB-agar containing the appropriate antibiotic. After 24-hours, individual colonies were suspended in LB-media in a flat-bottom 96-well plate; three colonies were picked for each mutator module to average the fluorescence output from biological triplicates. After 18-hours of overnight growth, the cells were resuspended in fresh LB media with appropriate antibiotics to a starting OD_{600} or 0.05. 20 $ng\mu l^{-1}$ anhydrotetracycline (aTc) was added to the medium to activate the expression of the mutator modules. The 96-well plate was covered in aluminium foil (aTc is light-sensitive) and incubated in a plate shaker at 31°C and 600 rpm for 24-hours. Subsequently, cell cultures were diluted to an OD_{600} of 1.0 and 2 µl (15 million cells) of the diluted culture was resuspended in 198 µl of 10 mM phosphate buffer saline. These samples were passed through the Attune NxT flow cytometer to analyse the loss of fluorescence within the cell populations. $GM31^{GFP}$ cells expressing T7 RNA polymerase were used as the positive control for fluorescence, while $GM31^{GFP}$ cells expressing empty ORI-antibiotic backbones were the negative control. The fluorescence data was stored in FCS 3.0 files and analysed using FlowJo v10.

Long Time-course Mutator Assay

A double promoter expression cassette, J23116-pT7—GFP-mut3b—double-T7-terminator, was integrated into the DH5α genome using the pOSIP λ-integrase protocol[299]. The $DH5\alpha^{GFP}$ strain was used to perform the 144-hour long mutator assay described in Chapter 5. $DH5\alpha^{GFP}$ cells were made chemically competent using the $CaCl_2$ protocol and stored in -80°C for future use. These cells were transformed with the three mutator and three control modules and subsequently spread on LB-agar plates with 25 $\mu g ml^{-1}$ gentamycin. After 24-hours, colonies picked for each of the 6 samples were suspended in 200 µl of LB with gentamycin for overnight growth at 31°C. After 24-hours, the cells were resuspended in 900 µl of fresh LB medium containing gentamycin and 20 $ng\mu l^{-1}$ anhydrotetracycline (aTc) to activate expression of the mutator proteins. The samples in 2 ml deep-well plates were wrapped in foil to prevent photodegradation of aTc and incubated in a plate shaker at 31°C with shaking at 600 rpm.

Every 24-hours, the optical density of the cell cultures was measured using a nanodrop spectrophotometer. A portion of the cells were resuspended in fresh LB media with gentamycin and aTc to keep the mutagenic assay active. Cells were resuspended to a starting OD_{600} of 0.05 (~ 36 million cells transferred to fresh growth medium). 2 µl of the cell cultures were resuspended in 198 µl of 10 mM PBS for flow cytometry measurements every 24-hours, to track the loss of fluorescence in the cell populations. This cycle was repeated every 24-hours for a total time of 144-hours. At the end of the time course, the cell cultures for the 6 mutator and control samples were resuspended in 5 ml of LB with gentamycin and grown overnight. This was done to achieve an appropriate volume of cells for using the genomic DNA prep kit. Genomic DNA was isolated for all the samples; their concentration measured using a nanodrop; and stored in -20°C for downstream applications.

The GFP expression cassette from the genomic library was isolated via Phusion polymerase PCR, using primers designed to bind upstream and downstream of the J23116-P_{T7}—GFP-mut3b—double-T7-terminator expression system. For PacBio sequencing, GFP cassettes were amplified using primers that added iP (GTCC overhang post BsaI treatment) and iS sequences (CGAG overhang), flanking the amplicon DNA sequence. The PacBio prefix and suffix SMRTbells were designed to produce complimentary overhangs to iP and iS respectively[310]. Therefore, sample preparation for PacBio sequencing could be performed using the BASIC assembly workflow.

For Illumina sequencing, the GFP library was amplified using primers that added TAG sequences flanking the amplicons. The tagged amplicons could then be amplified in a second round of PCR with Illumina Nextera Index primers[311].

7.11 Gain of Function Assay with β-lactamase (Amp^R)

These experiments were performed in the DH5α strain using a two-plasmid system. The target plasmid, containing an expression cassette for the active or inactive amp^R genes assembled with a p15A-Kanamycin backbone. The mutator modules were assembled in the pSC101-Gentamycin^R backbone. Both plasmids were co-transformed into competent DH5α cells and plated on Kan + Gen containing LB-agar plates. After 24-hours of incubation at 37°C, individual colonies for the different target and mutator plasmid combinations were suspended in 200 µl of LB media with 50 µgml^{-1} kanamycin and 25 µgml^{-1} gentamycin in a 96-well plate. The samples were grown overnight in a plate shaker at 31°C and 600 rpm shaking. The optical density of the overnight culture samples was measured using a nanodrop spectrophotometer and the samples were diluted to an OD_{600} of 0.05 in 900 µl of fresh LB with kanamycin, gentamycin and 20 ngµl^{-1} of aTc to activate the mutator assay. The 2-ml deep-well plate was wrapped in foil to prevent photodegradation of aTc and incubated at 31°C with 600 rpm shaking in a plate shaker.

Every 24-hours, the optical density of the samples was measured. A portion of the cells were diluted in fresh LB media with the antibiotics and aTc to keep the mutator assay active for another 24-hour cycle. 10^1–10^5 fold serial dilution was performed for the remaining cell culture and the appropriate dilution factor was plated on LB-agar plates with 50 µgml^{-1} carbenicillin and 25 µgml^{-1} gentamycin. After 24-hours of incubating the plate at 37°C, the number of colonies on each agar plate were counted and multiplied by the dilution factor to obtain the total carbenicillin-resistant colony formation units per millilitre (CFU/ml) of the bacterial culture at an OD_{600} of 1.0. This enabled a qualitative comparison of the rate of occurrence of each nucleotide substitution using the mutator modules MUT-1 – MUT-19, MUT-20, MUT-22, MUT-24 and MUT-25, based on the number of bacterial colonies that regained the amp^R phenotype. The gain of function mutator assays were performed for time periods of up to 96-hours.

Mutator Assays with selection pressure were performed in identical fashion, with the addition of 5 µgml^{-1} carbenicillin in the LB media with kanamycin, gentamycin and aTc.

7.12 Rifampicin Selection Assays

The mutator and control modules in Chapter 4 were transformed in DH5a competent cells and plated on LB-agar containing 25 µgml^{-1} gentamycin. After 24-hours of growth at 37°C, individual colonies for cells containing each control and mutator module were suspended in a 96-well plate containing LB with gentamycin and incubated in a plate shaker at 31°C and shaking at 600 rpm overnight. The overnight cultures were resuspended in 900 µl of LB media with gentamycin and 20 ngµl^{-1} aTc to activate the expression of the mutator and control modules. The deep-well plate was wrapped in foil and incubated in a plate shaker at 31°C with shaking at 600 rpm for 24-hours. After this period, cell cultures were diluted to an OD_{600} of 1.0. 800 µl of the diluted culture (640 million cells) were plated on LB-agar containing 50 µgml^{-1} rifampicin and incubated at 37°C overnight. The number of rifampicin resistant colonies per 640 million cells resulting from the activity of the different mutator and control modules were counted. Colony counts divided by 640 million provided the rate of occurrence of rifampicin^R phenotype per cell plated for each of the samples, which were compared to the occurrence of rifampicin^R in wildtype DH5α. This helped to assess the off-target mutagenic activity of the mutator modules.

7.13 Sample Preparation for PacBio NGS

BASIC PacBio protocol

HPLC purified, 5'-phosphorylated barcoded PacBio SMRTbell adapters were synthesised from Biolegio with 4 bp overhangs complementary to the BASIC prefix (iP) and suffix (iS) sites that flank bioparts. This enables PacBio sequencing libraries to be generated with DNA sequences in the BASIC format (plasmids or PCR products).

Strategy:

Using a standard BASIC protocol outlined in Section 7.3, PacBio SMRTbell-BASIC linkers were ligated to both ends of the target DNA sequence in the BASIC format. A total of 8 prefix PacBio linkers (PX-A – PX-H) and 12 suffix linkers (SX-1 – SX-12) were designed in the BASIC format, enabling a library of 96 uniquely barcoded DNA samples to be created for sequencing. Template DNA can be supplied as a plasmid or PCR product in concentrations needed for the BASIC protocol (50ng/1kb template DNA). 8 prefix linkers and 8 suffix linkers were used to generate library of 8 uniquely barcoded DNA samples from the long time-course loss-of-function assay described in 7.10.

Once the PacBio linkers were ligated to the library of mutant GFP, the next step is to repair any DNA damage that may have been introduced during the PCR to amplify the target DNA or during assembly of the target to the SMRTbell-BASIC linkers. DNA damage must be repaired, while free linkers and severely damaged DNA must be removed, as they impede with the sequencing process and generate errors[312]. DNA was repaired with the PreCR mix from NEB (M0309S) and subsequent degradation of single stranded and nicked DNA and linkers was achieved by treatment with Exonuclease III (NEB M0206S). The DNA samples were purified from the mixture after each enzymatic step using Ampure XP SPRI bead solution and the magnetic bead purification protocol described in 7.3.2.

Table 7.10: List of SMRTbell PacBio linkers adapted for use with the BASIC assembly protocol

SMRTbell Name	Sequence (5' → 3')
Pbsmart_PX_A	GGACATGACGCATCGTCTGAATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATTCAGACGATGCGTCAT
Pbsmart_PX_B	GGACGCAGAGTCATGTATAGATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATCTATACATGACTCTGC
Pbsmart_PX_C	GGACGAGTGCTACTCTAGTAATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATTACTAGAGTAGCACTC
Pbsmart_PX_D	GGACCATGTACTGATACACAATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATTGTGTATCAGTACATG

Pbsmart_PX_E	GGACAGTGTGTCATGCGTGTATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATACACGCATGACACACT
Pbsmart_PX_F	GGACGCATATAGTAGAGATCATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATGATCTCTACTATATGC
Pbsmart_PX_G	GGACCAGCAGTATAGACTGTATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATACAGTCTATACTGCTG
Pbsmart_PX_H	GGACAGATGTAGCACATCATATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATATGATGTGCTACATCT
Pbsmart_SX_1	CTCGCATAGCGACTATCGTGATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATCACGATAGTCGCTATG
Pbsmart_SX_2	CTCGCATCACTACGCTAGATATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATATCTAGCGTAGTGATG
Pbsmart_SX_3	CTCGCGCATCTGTGCATGCAATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATTGCATGCACAGATGCG
Pbsmart_SX_4	CTCGTATGTGATCGTCTCTCATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATGAGAGACGATCACATA
Pbsmart_SX_5	CTCGGTACACGCTGTGACTAATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATTAGTCACAGCGTGTAC
Pbsmart_SX_6	CTCGCGTGTCGCGCATATCTATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATAGATATGCGCGACACG
Pbsmart_SX_7	CTCGATATCAGTCATGCATAATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATTATGCATGACTGATAT
Pbsmart_SX_8	CTCGGAGATCGACAGTCTCGATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATCGAGACTGTCGATCTC
Pbsmart_SX_9	CTCGCACGCACACACGCGCGATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATCGCGCGTGTGTGCGTG
Pbsmart_SX_10	CTCGCGAGCACGCGCGTGTGATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATCACACGCGCGTGCTCG
Pbsmart_SX_11	CTCGGTAGTCTCGCACAGATATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATATCTGTGCGAGACTAC
Pbsmart_SX_12	CTCGGAGACTCTGTGCGCGTATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATACGCGCACAGAGTCTC

Sample Preparation Protocol:

The first step was PCR amplification of target GFP sequences flanked by methylated BASIC linkers with 2 standard primers. The primers are designed to bind to genomic DNA upstream of the LMP and LMS linker sequences. This ensures the BsaI recognition and cutting sites are preserved to use amplicons as a BASIC part. The purified genome libraries from 7.10 were used to setup a high-fidelity Phusion polymerase reaction.

LM-PX_for:	CTATTATCTGGTGGGTCTCT	Tm 56
LM-SX_rev:	TTACCGATAGGTCTCCCG	Tm 57

PCR was performed as described in 7.9.2 for each of the 8 samples to be barcoded. The amplicons were subsequently purified from the reaction mixture using the magnetic bead purification steps described in section 7.3.2. The purified samples were stored at -20°C for downstream applications.

1. PacBio SMRTbell Linker Preparation

The lyophilised Biolegio PacBio linkers PBsmartA-H and PBsmart1-12 are stored in 10mM Tris pH 8 at a concentration of 100 µM. To achieve the working concentration, 0.5 µl of the 100 µM stock solution was diluted in 200 µl of BASIC linker annealing buffer (10mM TRIS-HCl buffer pH7.9, 100mM NaCl, 10mM MgCl$_2$). Once added to the annealing buffer, linkers are heated to 95°C in a heat block for 5 minutes and and immediately transferred to ice. This sudden change in temperature enables favourable intramolecular interactions to form the stem-and-loop structure of the Pacbio Smartbells with the 4-bp prefix or suffix overhang of BASIC.

2. SMRTbell + Target DNA Ligation Step

For each BASIC ligation reaction, 30 µl reactions were setup in 200 µl microcentrifuge tubes:

Table 7.11: SMRTbell Ligation Reaction Mixture

Reagent	Volume
dH$_2$0	17 µl
Promega T4 buffer (10x)	3 µl
PBsmrt_PX	2 µl
PBsmrt_SX	2 µl
BASIC biopart	0.5-6 µl (50 ng per 1kb PCR product)
dH$_2$0	add dH$_2$0 to reach 28.5 µl volume
NEB BsaI-HF v2 enzyme (R3733) 20 U/µl	1 µl
Promega T4 ligase (M1801) 1-3 U/µl	0.5 µl
	Mix by pipetting up and down

After mixing, tubes were placed in a PCR machine running the following programme:

Temperature	Time	
37°C	2 min	**X 25 cycles**
20°C	1 min	
55°C	5 min	
4°C	store	

3. Magnetic Bead Purification: SPRI bead purification was performed as described in 7.3.2 and the SMRTbell-ligated DNA samples were eluted in 30 µl of 10 mM Tris-HCl pH 8.0.

4. DNA repair reaction with PreCR Kit

Table 7.12: PreCR DNA Repair Reaction Mixture

Reagent	Volume
SMRTbell-ligated DNA sample	23 µl
PrePCR buffer 10x (from kit)	3 µl
NAD (from kit)	0.5 µl

dNTP 10mM	3 µl
PrePCR enzyme mix (from kit)	0.5 µl
	Mix by pipetting up and down

The reaction mixture was incubated in a thermal cycler for 1-hour at 37°C to facilitate repair of any DNA damage that occurred during the Phusion PCR step or SMRTbell BASIC assembly reaction.

5. Second Magnetic Bead Purification Step: The repaired SMRTbell-ligated DNA samples were purified in 30 µl of 10 mM Tris-HCl pH 8.0.

6. Exonuclease-III sample clean-up: SMRTbell-ligated DNA sequences are circular and protected from Exo-III activity, an exonuclease that cleaves linear and nicked DNA in the 3'-5' direction. This clean-up reaction degrades unbound SMRTbell linkers and damaged linear DNA sequences from the reaction mixture. The clean-up reaction is setup as follows:

Table 7.13: Exonuclease-III Clean-up Reaction

Reagent	Volume
PreCR purified DNA sample	26.8 µl
10x Buffer-1 (from Exo-III kit)	3 µl
Exo-III (in 1x buffer-1)	0.2 µl
	Mix by pipetting up and down

Temperature	Time
37°C	15 min
4°C	store

After 15 minutes of incubation at 37°C, the remaining DNA is immediately purified form this reaction mixture, as longer incubation with Exo-III will degrade most DNA.

7. Third Magnetic Bead Purification Step: Purified SMRT-bell-ligated DNA samples are eluted in 20 µl of 10 mM Tris-HCl pH 8.0.

8. Quantitation of SMRTbell-ligated samples using Qubit

- The first step to pooling the samples together is to calculate the concentration of the samples in nM units.
- This requires knowing the nucleotide length of the DNA samples plus the SMRTbell prefix and suffix sequences. The total length of all these sequences combined was 1400 bp.
- The total length of the barcoded sequence was applied in the following formula to convert concentration from ngul^{-1} to nM:

$$\frac{Concentration\ of\ sample\ in\ ngul^{-1}}{(660\ gmol^{-1})\ x\ average\ library\ size}\ x\ 10^{6} = Concentration\ in\ nM$$

- The individual samples were diluted to a concentration of 0.45 nM in 10 mM Tris pH 8.5.
- 5 ul of each diluted sample was pooled together in one microcentrifuge tube.
- The pooled library maintains a concentration of 0.45 nM.

9. Sample Pooling

Each of the 8 barcoded samples were combined in equimolar amounts to a final pooled volume of 25 µl. These samples were placed on dry ice and delivered to the PacBio sequencing facility at the Earlham Institute for processing.

7.14 Sample Preparation for Illumina NGS

This protocol is for isolating gene targets from the genome of *E. coli* cells and preparing the target for Illumina next generation sequencing. The protocol involves two PCR steps: Amplicon PCR and Index PCR, to amplify the gene target and attach the barcoded i7 and i5 adapters to the target.

I. Bacterial genome isolation

- Setup a 5 ml overnight culture for each strain of cells whose genome needs to be isolated.
- Measure OD of the cells the next day. Only 1x10^8 cells can be prepped using the bacterial genome prep kit.
- Follow the instructions for genome prep given in the kit by Omega Biotek.
- Measure the concentration of each sample using the nanodrop or Qubit fluorometric assay. Generally, a concentration of 100-1000 ngul^{-1} would be achieved.
- Dilute a small portion of the purified genome to a concentration of 20 ngul^{-1} in 10 mM Tris pH 8.5. 20 ul of diluted samples should be enough.

II. Amplicon PCR

- This PCR step involves amplifying the target gene from the bacterial genome using target-specific primers.
- For Illumina sequencing, the target DNA fragment cannot be larger than 275 bp. If the target gene is larger than this, select a region of the gene to sequence or create 275 bp fragments of the entire gene.
- The primers needed for this PCR must be designed to bind the target in the genome but also contain overhangs (called probes) that will anneal to the i7 and i5 indexes during the Index PCR.
- The index and probe sequences can be found on the Illumina website.

- Once the required primers and DNA template are in place, the following Phusion polymerase PCR should be set up:

Table 7.14: Illumina Amplicon PCR Reaction

Components	Volume (ul)	Gradient PCR 8 x 12 ul
PCR grade water (MilliQ)	14.5	65
5x High Fidelity Buffer	5	20
Forward Primer (10 mM)	1.25	3
Reverse primer (10 mM)	1.25	3
dNTPs (10 mM)	1	2
DNA (20 ngul^{-1})	1.5	6
Phusion polymerase	0.5	1
Total (ul)	**25**	**100**

Temperature	Time	
98°C	3 min	
98°C	10 sec	**25 cycles**
55°C-65°C gradient	30 sec	
72°C	10 sec (15-30 sec per kb)	
72°C	10 min	
4°C	Store indefinitely	

- Once the ideal melting temperature is known, setup a larger 25 ul reaction for all the different samples.

III. Magbead Purification I

IV. Index PCR

- This step is to attach the index primers to the amplicons to be sequenced.
- These indices can hybridize with the flow cell during sequencing, each index possessing a unique barcode sequence allowing for easy recognition of the downstream DNA sample being sequenced.
- The prefix primer is known as an i7 index and found in orange capped tubes in the Nextera XT kit.
- The suffix primer is known as an i5 index and found in white capped tubes in the Nextera XT kit.
- Choose a unique i7 and i5 index combination for each unique sample being sequenced.
- Once the index combinations have been identified, setup the following Phusion PCR protocol to attach the indices to the samples:
-

Table 7.15: Illumina Index PCR Reaction	
Components	**Volume (ul)**
PCR grade water (MilliQ)	23.5
5x High Fidelity Buffer	10
I7 Index (Orange cap)	5
I5 index (White cap)	5
dNTPs (10 mM)	1
DNA (From amplicon PCR)	5
Phusion polymerase	0.5
Total (ul)	**50**

- Setup the thermal cycler with the following temperature conditions:

Temperature	Time	
98°C	3 min	
98°C	10 sec	**8 cycles**
65°C	30 sec	
72°C	10 sec (15-30 sec per kb)	
72°C	5 min	
4°C	Store indefinitely	

V. Magbead Purification II

VI. DNA Quantification using Qubit

- The dsDNA high sensitivity kit should ideally be used for this quantification.
- Prepare the high sensitivity dsDNA reagent by diluting 200x in HS dsDNA buffer.
- 190 ul of the working solution of the reagent is needed per sample, so prepare a master-mix accordingly.
- Add 10 ul of each sample with the working solution of the dsDNA HS reagent to a final volume of 200ul in 0.5ml PCR tubes.
- Vortex the samples and incubate at room temperature for a minimum of 3 minutes.
- Normalise the Qubit device for HS dsDNA quantitation using the standards provided in the kit.
- Place each tube in the machine, close the lid and measure the concentration. Qubit measures concentrations in ugml^{-1} (ie, ngul^{-1}).
- Multiply the readings by 20 (samples 20x diluted in HS reagent buffer) to obtain the true concentration of the samples in ngul^{-1}.

VII. Sample pooling

- The first step to pooling the samples together is to calculate the concentration of the samples in nM units.
- This requires knowing the nucleotide length of your samples. Add the length of i7 and i5 indices to the length of the insert DNA. I7 indices are 66 bp long, while i5 indices are 70 bp long.

- Once the length of the indexed samples is known, apply the following formula to convert concentration from ngul^{-1} to nM:

$$\frac{Concentration\ of\ sample\ in\ ngul^{-1}}{(660\ gmol^{-1})\ x\ average\ library\ size}\ x\ 10^6 = Concentration\ in\ nM$$

- Dilute each sample to a concentration of 4 nM in 10 mM Tris pH 8.5.
- Take 5 ul of each diluted sample and pool together.
- The pooled library maintains a concentration of 4 nM.

VIII. Loading pooled sampled into flowcell for iSeq

The pooled samples were loaded into the flowcell as described in the manufacturer's instruction manual.

7.15 NGS Data Processing

The R1 and R2 subreads were first filtered for subreads with bases at a Phred score > 35 using the AfterQC software[234]. The filtered R1 and R2 subreads were then combined into one read using the 13-nucleotide overlap between them. FLASH[244] was used to combine the subreads and obtain full-length reads of the 245-bp insert. The filtered and combined reads were aligned to the reference sequence using the python script in Appendix 9.5. The 'Alignment_Condition' variable was set to 240 for the Illumina reads to filter and count mutations from alignments that achieved an alignment score of > 240. This means that up to 5 mismatches were allowed in each alignment. The called mutations and index position of said mutations were collected in a CSV file and analysed in MS Excel.

The PacBio Sequel sequencing was performed by the Earlham Institute in Norwich. The data processing to convert subreads of a single DNA molecule into CCS reads were done by the institute. The data was provided filtered to a Phred Score of 60. The CCS reads for all barcoded samples were provided in a single fastq file. These were reorganised into individual files based on barcodes and aligned to a reference gene using the python script in Appendix 9.5 to call for mutations.

7.16 Automation Using Liquid Handling Robots

CYBi-FeliX: Conducting an RBS characterisation assay with 15 different RBS and 7 different parts required constructing 107 different plasmids including controls. Such a large number of constructs required working in a 96-well plate format and using robots to achieve precise dispensing of cells in media to obtain a normalised starting optical density (OD). The robot was also used to dispense cells from the gain of function mutator assay onto rectangular agar plates.

Clariostar Microplate Reader: The microplate reader was used to measure GFP fluorescence in RBS characterisation assays. OD measurements were made at 600 nm to evaluate the fitness of cells in characterisation and mutator experiments. OD measurements of overnight cultures also allowed for programming FeliX to dispense cells accurately to obtain a starting OD of 0.05 in each well of the microplate.

Chapter 8 Bibliography

1. Simply... A History Of Biotechnology | New Internationalist. Available at: https://newint.org/features/1991/03/05/simply. (Accessed: 4th September 2020)

2. Manchester, K. L. Louis Pasteur (1822-1895) - chance and the prepared mind. *Trends Biotechnol.* **13**, 511–515 (1995).

3. Barnett, J. A. & Lichtenthaler, F. W. A history of research on yeasts 3: Emil Fischer, Eduard Buchner and their contemporaries, 1880-1900. *Yeast* **18**, 363–388 (2001).

4. Ratzsch, G. K. Hellriegel's work at Bernburg. *Philos. Trans. R. Soc. London. B, Biol. Sci.* **317**, 107–109 (1987).

5. *Bacteria as multicellular organisms*. *Bacteria as multicellular organisms* (Oxford University Press, 1997). doi:10.1038/scientificamerican0688-82

6. Cobb, M. Oswald Avery, DNA, and the transformation of biology. *Current Biology* **24**, R55–R60 (2014).

7. Watson, J. D. & Crick, F. H. C. *A Structure for Deoxyribose Nucleic Acid*. *Nature* **171**, (1953).

8. Cotterill, F. P. D. Crick, F. H. C. (1958) On protein synthesis.

9. CHAPEVILLE, F. *et al.* On the role of soluble ribonucleic acid in coding for amino acids. *Proc. Natl. Acad. Sci. U. S. A.* **48**, 1086–1092 (1962).

10. Nirenberg, M. *et al.* The RNA code and protein synthesis. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 11–24 (1966).

11. Nirenberg, M. & Leder, P. RNA codewords and protein synthesis. *Science (80-.).* **145**, 1399–1407 (1964).

12. NIRENBERG, M. W. & MATTHAEI, J. H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 1588–1602 (1961).

13. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

14. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).

15. Greene, P. J. *et al.* Restriction and modification of a self-complementary octanucleotide containing the EcoRI substrate. *J. Mol. Biol.* **99**, 237–261 (1975).

16. Dugaiczyk, A., Boyer, H. W. & Goodman, H. M. Ligation of EcoRI endonuclease-generated DNA fragments into linear and circular structures. *J. Mol. Biol.* **96**, 171–184 (1975).

17. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3240–3244 (1973).

18. Bud, R. History of Biotechnology. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2003). doi:10.1038/npg.els.0003086

19. Williams, D. C., Van Frank, R. M., Muth, W. L. & Burnett, J. P. Cytoplasmic inclusion bodies in Escherichia coli producing biosynthetic human insulin proteins. *Science (80-.).* **215**, 687–689 (1982).

20. Choi, J. H., Keum, K. C. & Lee, S. Y. Production of recombinant proteins by high cell density culture of Escherichia coli. *Chemical Engineering Science* **61**, 876–885 (2006).

21. Sugrue, R. J., Cui, T., Xu, Q., Fu, J. & Cheong Chan, Y. The production of recombinant dengue virus E protein using Escherichia coli and Pichia pastoris. *J. Virol. Methods* **69**, 159–169 (1997).

22. Mazur, B. J. & Falco, S. C. *THE DEVELOPMENT OF HERBICIDE RESIST ANT CROPS*. *Annu. Rev. Plant Physiol. Plant Mol. Bioi* **40**, (1989).

23. Easwar Rao, D., Divya, K., Prathyusha, I. V. S. N., Rama Krishna, C. & Chaitanya, K. V. Insect-Resistant Plants. in *Current Developments in Biotechnology and Bioengineering: Crop Modification, Nutrition, and Food Production* 47–74 (Elsevier Inc., 2017). doi:10.1016/B978-0-444-63661-4.00003-7

24. Temme, K., Zhao, D. & Voigt, C. A. Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc. Natl. Acad. Sci.* **109**, 7085–7090 (2012).

25. Ryu, M.-H. *et al.* Control of nitrogen fixation in bacteria that associate with cereals. doi:10.1038/s41564-019-0631-2

26. Keasling, J. D. Manufacturing molecules through metabolic engineering. *Science* **330**, 1355–1358 (2010).

27. Hempel, F. *et al.* Microalgae as bioreactors for bioplastic production. *Microb. Cell Fact.* **10**, 81 (2011).

28. Mehrotra, P. Biosensors and their applications - A review. *Journal of Oral Biology and Craniofacial Research* **6**, 153–159 (2016).

29. Stephanopoulos, G. Challenges in engineering microbes for biofuels production. *Science* **315**, 801–804 (2007).

30. Banerjee, C., Dubey, K. K. & Shukla, P. Metabolic Engineering of Microalgal Based Biofuel Production: Prospects and Challenges. *Front. Microbiol.* **7**, 432 (2016).

31. Bastian, S. *et al.* Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli. *Metab. Eng.* **13**, 345–352 (2011).

32. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science (80-.).* **329**, 309–313 (2010).

33. Gross, M. *Evolving new types of enzymes*. *CURBIO* **23**, (2013).

34. Protein Engineering - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/neuroscience/protein-engineering. (Accessed: 4th September 2020)

35. Son, H. F. *et al.* Rational Protein Engineering of Thermo-Stable PETase from Ideonella sakaiensis for Highly Efficient PET Degradation. *ACS Catal.* **9**, 3519–3526 (2019).

36. Korendovych, I. V. Rational and semirational protein design. in *Methods in Molecular Biology* **1685**, 15–23 (Humana Press Inc., 2018).

37. Lanio, T., Jeltsch, A. & Pingoud, A. On the possibilities and limitations of rational protein design to expand the specificity of restriction enzymes: A case study employing EcoRV as the target. *Protein Eng.* **13**, 275–281 (2000).

38. Chovancova, E. *et al.* CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **8**, e1002708 (2012).

39. Reetz, M. T., Soni, P., Fernández, L., Gumulya, Y. & Carballeira, J. D. Increasing the stability of an enzyme toward hostile organic solvents by directed evolution based on iterative saturation mutagenesis using the B-FIT method. *Chem. Commun.* **46**, 8657–8658 (2010).

40. Wijma, H. J., Fürst, M. J. L. J. & Janssen, D. B. A Computational Library Design Protocol for Rapid Improvement of Protein Stability: FRESCO. in *Methods in Molecular Biology* **1685**, 69–85 (Humana Press Inc., 2018).

41. Oroz-Guinea, I., Zorn, K. & Brundiek, H. Protein engineering of enzymes involved in lipid modification. in *Lipid Modification by Enzymes and Engineered Microbes* 11–43 (Elsevier, 2018). doi:10.1016/B978-0-

12-813167-1.00002-5

42. Bommarius, A. S., Blum, J. K. & Abrahamson, M. J. Status of protein engineering for biocatalysts: How to design an industrially useful biocatalyst. *Current Opinion in Chemical Biology* **15**, 194–200 (2011).

43. Zhang, K., Sawaya, M. R., Eisenberg, D. S. & Liao, J. C. Expanding metabolism for biosynthesis of nonnatural alcohols. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20653–20658 (2008).

44. Rotticci, D., Rotticci-Mulder, J. C., Denman, S., Norin, T. & Hult, K. Improved enantioselectivity of a lipase by rational protein engineering. *ChemBioChem* **2**, 766–770 (2001).

45. Selzer, T., Albeck, S. & Schreiber, G. Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.* **7**, 537–541 (2000).

46. Bartsch, S. *et al.* Redesign of a Phenylalanine Aminomutase into a Phenylalanine Ammonia Lyase. *ChemCatChem* **5**, 1797–1802 (2013).

47. Cheng, Y. S. *et al.* Improving the catalytic performance of a GH11 xylanase by rational protein engineering. *Appl. Microbiol. Biotechnol.* **99**, 9503–9510 (2015).

48. Alvizo, O., Allen, B. D. & Mayo, S. L. Computational protein design promises to revolutionize protein engineering. *Biotechniques* **42**, 31–39 (2007).

49. Wang, J., Cao, H., Zhang, J. Z. H. & Qi, Y. Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* **8**, 1–9 (2018).

50. Koonin, E. V. Darwinian evolution in the light of genomics. *Nucleic Acids Res.* **37**, 1011–1034 (2009).

51. Xiao, H., Bao, Z. & Zhao, H. High Throughput Screening and Selection Methods for Directed Enzyme Evolution. *Ind. Eng. Chem. Res.* **54**, 4011–4020 (2015).

52. Herzenberg, L. A., Sweet, R. G. & Herzenberg, L. A. Fluorescence-activated Cell Sorting. **234**, 108–118 (1976).

53. Bhagat, A. A. S. *et al.* Microfluidics for cell separation. *Medical and Biological Engineering and Computing* **48**, 999–1014 (2010).

54. Wittrup, K. D. Protein engineering by cell-surface display. *Current Opinion in Biotechnology* **12**, 395–399 (2001).

55. Lagus, T. P. & Edd, J. F. A review of the theory, methods and recent applications of high-throughput single-cell droplet microfluidics. *Journal of Physics D: Applied Physics* **46**, 114005 (2013).

56. Osborne, G. W. *Recent advances in flow cytometric cell sorting*. *Methods in Cell Biology* **102**, (Methods Cell Biol, 2011).

57. Becker, S., Schmoldt, H. U., Adams, T. M., Wilhelm, S. & Kolmar, H. Ultra-high-throughput screening based on cell-surface display and fluorescence-activated cell sorting for the identification of novel biocatalysts. *Current Opinion in Biotechnology* **15**, 323–329 (2004).

58. Becker, S. *et al.* Single-Cell High-Throughput Screening To Identify Enantioselective Hydrolytic Enzymes. *Angew. Chemie Int. Ed.* **47**, 5085–5088 (2008).

59. Speight, R. E., Hart, D. J., Sutherland, J. D. & Blackburn, J. M. A new plasmid display technology for the in vitro selection of functional phenotype-genotype linked proteins. *Chem. Biol.* **8**, 951–965 (2001).

60. Takahashi, T. T., Austin, R. J. & Roberts, R. W. mRNA display: Ligand discovery, interaction analysis and beyond. *Trends in Biochemical Sciences* **28**, 159–165 (2003).

61. Zahnd, C., Amstutz, P. & Plückthun, A. Ribosome display: Selecting and evolving proteins in vitro that specifically bind to a target. *Nat. Methods* **4**, 269–279 (2007).

62. Hanes, J. & Plückthun, A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4937–4942 (1997).

63. Li, R., Kang, G., Hu, M. & Huang, H. Ribosome Display: A Potent Display Technology used for Selecting and Evolving Specific Binders with Desired Properties. *Molecular Biotechnology* **61**, 60–71 (2019).

64. Shembekar, N., Chaipan, C., Utharala, R. & Merten, C. A. Droplet-based microfluidics in drug discovery, transcriptomics and high-throughput molecular genetics. *Lab on a Chip* **16**, 1314–1331 (2016).

65. Lee, S. Y., Choi, J. H. & Xu, Z. Microbial cell-surface display. *Trends in Biotechnology* **21**, 45–52 (2003).

66. Van der Vaart, J. M. *et al.* Comparison of cell wall proteins of Saccharomyces cerevisiae as anchors for cell surface expression of heterologous proteins. *Appl. Environ. Microbiol.* **63**, (1997).

67. Lee, S. M., Jellison, T. & Alper, H. S. Directed evolution of xylose isomerase for improved xylose catabolism and fermentation in the yeast Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **78**, 5708–5716 (2012).

68. Reetz, M. T., Höbenreich, H., Soni, P. & Fernández, L. A genetic selection system for evolving enantioselectivity of enzymes. *Chem. Commun.* **0**, 5502–5504 (2008).

69. Osterwalder, T., Yoon, K. S., White, B. H. & Keshishian, H. A conditional tissue-specific transgene expression system using inducible GAL4. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12596–12601 (2001).

70. Bongaerts, R. J. M., Hautefort, I., Sidebotham, J. M. & Hinton, J. C. D. Green fluorescent protein as a marker for conditional gene expression in bacterial cells. *Methods Enzymol.* **358**, 43–66 (2002).

71. Brödel, A. K. & Isalan, M. Engineering of biomolecules by bacteriophage directed evolution. *Curr. Opin. Biotechnol.* **51**, 32–38 (2018).

72. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).

73. Brödel, A. K., Jaramillo, A. & Isalan, M. Engineering orthogonal dual transcription factors for multi-input synthetic promoters. *Nat. Commun.* **7**, 13858 (2016).

74. Dickinson, B. C., Packer, M. S., Badran, A. H. & Liu, D. R. A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nat. Commun.* **5**, 1–8 (2014).

75. Meyer, A. J. & Ellington, A. D. Molecular evolution picks up the PACE. *Nat. Biotechnol.* **29**, 502–503 (2011).

76. Rasila, T. S., Pajunen, M. I. & Savilahti, H. Critical evaluation of random mutagenesis by error-prone polymerase chain reaction protocols, Escherichia coli mutator strain, and hydroxylamine treatment. *Anal. Biochem.* **388**, 71–80 (2009).

77. Burney, S., Caulfield, J. L., Niles, J. C., Wishnok, J. S. & Tannenbaum, S. R. The chemistry of DNA damage from nitric oxide and peroxynitrite. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **424**, 37–49 (1999).

78. Zimmermann, F. K. Genetic effects of nitrous acid. *Mutation Research/Reviews in Genetic Toxicology* **39**, 127–148 (1977).

79. Pfeifer, G. P., You, Y. H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **571**, 19–31 (2005).

80. Goosen, N. & Moolenaar, G. F. Repair of UV damage in bacteria. *DNA Repair* **7**, 353–379 (2008).

81. Denamur, E. & Matic, I. Evolution of mutation rates in bacteria. *Molecular Microbiology* **60**, 820–827 (2006).

82. Alexander, D. L. *et al.* Random mutagenesis by error-prone pol plasmid replication in escherichia coli. *Methods Mol. Biol.* **1179**, 31–44 (2014).

83. Wilson, D. S. & Keefe, A. D. Random Mutagenesis by PCR. in *Current Protocols in Molecular Biology* **Chapter 8**, 8.3.1-8.3.9 (John Wiley & Sons, Inc., 2001).

84. Copp, J. N., Hanson-Manful, P., Ackerley, D. F. & Patrick, W. M. Error-prone PCR and effective

generation of gene variant libraries for directed evolution. *Methods Mol. Biol.* **1179**, 3–22 (2014).

85. Lin-Goerke, J. L., Robbins, D. J. & Burczak, J. D. PCRr-based random mutagenesis using manganese and reduced DNTP concentration. *Biotechniques* **23**, 409–412 (1997).

86. Rasila, T. S., Pajunen, M. I. & Savilahti, H. Critical evaluation of random mutagenesis by error-prone polymerase chain reaction protocols, Escherichia coli mutator strain, and hydroxylamine treatment. *Anal. Biochem.* **388**, 71–80 (2009).

87. Wong, T. S., Tee, K. L., Hauer, B. & Schwaneberg, U. Sequence saturation mutagenesis (SeSaM): a novel method for directed evolution. *Nucleic Acids Res.* **32**, e26–e26 (2004).

88. Acevedo-Rocha, C. G., Hoebenreich, S. & Reetz, M. T. Iterative saturation mutagenesis: A powerful approach to engineer proteins by systematically simulating darwinian evolution. *Methods Mol. Biol.* **1179**, 103–128 (2014).

89. Williams, E. M., Copp, J. N. & Ackerley, D. F. Site-saturation mutagenesis by overlap extension PCR. *Methods Mol. Biol.* **1179**, 83–101 (2014).

90. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).

91. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320**, 85–95 (2002).

92. Gumulya, Y., Sanchis, J. & Reetz, M. T. Many Pathways in Laboratory Evolution Can Lead to Improved Enzymes: How to Escape from Local Minima. *ChemBioChem* **13**, 1060–1066 (2012).

93. Behrendorff, J. B. Y. H., Johnston, W. A. & Gillam, E. M. J. Restriction enzyme-mediated DNA family shuffling. *Methods in Molecular Biology* **1179**, 175–187 (2014).

94. Gibbs, M. D., Nevalainen, K. M. H. & Bergquist, P. L. Degenerate oligonucleotide gene shuffling (DOGS): A method for enhancing the frequency of recombination with family shuffling. *Gene* **271**, 13–20 (2001).

95. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

96. Arnold, F. H., Georgiou, G., Udit, A. K., Silberg, J. J. & Sieber, V. Sequence Homology-Independent Protein Recombination (SHIPREC). in *Directed Evolution Library Creation* 153–164 (Humana Press, 2003). doi:10.1385/1-59259-395-x:153

97. Patrick, W. M. & Gerth, M. L. ITCHY: Incremental truncation for the creation of hybrid enzymes. *Methods Mol. Biol.* **1179**, 225–244 (2014).

98. Kawarasaki, Y. *et al.* Enhanced crossover SCRATCHY: construction and high-throughput screening of a combinatorial library containing multiple non-homologous crossovers. *Nucleic Acids Res.* **31**, e126–e126 (2003).

99. Bao, Z., Cobb, R. E. & Zhao, H. Accelerated genome engineering through multiplexing. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **8**, 5–21 (2016).

100. MAYNARD SMITH, J. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 563–564 (1970).

101. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **10**, 866–876 (2009).

102. Böttcher, D. & Bornscheuer, U. T. Protein engineering of microbial enzymes. *Current Opinion in Microbiology* **13**, 274–282 (2010).

103. Lutz, S. Beyond directed evolution-semi-rational protein engineering and design. *Current Opinion in Biotechnology* **21**, 734–743 (2010).

104. Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73–76 (2005).

105. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* **11**, 572–582 (2010).

106. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9205–9210 (2004).

107. Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2**, (2007).

108. Zhong, Z. *et al.* Automated continuous evolution of proteins in vivo. *ACS Synth. Biol.* 2020.02.21.960328 (2020). doi:10.1101/2020.02.21.960328

109. Ravikumar, A. *et al.* Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds Resource Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. *Cell* **175**, 1946-1957.e13 (2018).

110. Butt, H. *et al.* CRISPR directed evolution of the spliceosome for resistance to splicing inhibitors. *Genome Biol.* **20**, 73 (2019).

111. Butt, H., Zaidi, S. S. e. A., Hassan, N. & Mahfouz, M. CRISPR-Based Directed Evolution for Crop Improvement. *Trends in Biotechnology* **38**, 236–240 (2020).

112. Malmborg, A. C., Söderlind, E., Frost, L. & Borrebaeck, C. A. K. Selective phage infection mediated by epitope expression on F pilus. *J. Mol. Biol.* **273**, 544–551 (1997).

113. Packer, M. S., Rees, H. A. & Liu, D. R. Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nat. Commun.* **8**, 1–11 (2017).

114. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

115. Nishida, K. *et al.* Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science (80-.).* **353**, (2016).

116. Hess, G. T. *et al.* Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).

117. Ma, Y. *et al.* Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat. Methods* **13**, 1029–1035 (2016).

118. Sadanand, S. EvolvR-ing to targeted mutagenesis. *Nature Biotechnology* **36**, 819 (2018).

119. Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* **560**, 248–252 (2018).

120. Moore, C. L., Papa, L. J. & Shoulders, M. D. A Processive Protein Chimera Introduces Mutations across Defined DNA Regions in Vivo. *J. Am. Chem. Soc.* **140**, 11560–11564 (2018).

121. Mackrow, B. P. A novel biological circuit to control directed evolution in vivo. *Spiral Imperial College* (2016). Available at: https://spiral.imperial.ac.uk:8443/handle/10044/1/78426. (Accessed: 19th August 2020)

122. Ravikumar, A., Arrieta, A. & Liu, C. C. An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).

123. Tan, Z. L. *et al.* In vivo continuous evolution of metabolic pathways for chemical production. *Microbial Cell Factories* **18**, 1–19 (2019).

124. Goldstein, B. P. Resistance to rifampicin: A review. *Journal of Antibiotics* **67**, 625–630 (2014).

125. Di Noia, J. M. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).

126. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103–107 (2003).

127. Smith, B. T. & Walker, G. C. *Mutagenesis and More: umuDC and the Escherichia coli SOS Response*. (1998).

128. Martin, A. *et al.* Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. *Nature* **415**, 802–806 (2002).

129. Stavnezer, J. *et al.* Differential expression of APE1 and APE2 in germinal centers promotes error-prone repair and A:T mutations during somatic hypermutation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9217–9222 (2014).

130. Broyde, S., Wang, L., Rechkoblit, O., Geacintov, N. E. & Patel, D. J. Lesion processing: high-fidelity versus lesion-bypass DNA polymerases. *Trends Biochem. Sci.* **33**, 209–219 (2008).

131. Rajewsky, K. & Cumano, A. *Evolutionary and Somatic Selection of the Antibody Repertoire in the Mouse Downloaded from*.

132. Maul, R. W. & Gearhart, P. J. AID and somatic hypermutation. in *Advances in Immunology* **105**, 159–191 (Academic Press Inc., 2010).

133. Gehrke, J. M. *et al.* An apobec3a-cas9 base editor with minimized bystander and off-target activities. *Nat. Biotechnol.* **36**, 977 (2018).

134. Lv, J. *et al.* The length of guide RNA and target DNA heteroduplex effects on CRISPR/Cas9 mediated genome editing efficiency in porcine cells. *J. Vet. Sci.* **20**, (2019).

135. Tahlrov, T. H. *et al.* Structure of a T7 RNA polymerase elongation complex at 2.9 Å resolution. *Nature* **420**, 43–50 (2002).

136. Dickerson, S. K., Market, E., Besmer, E. & Papavasiliou, F. N. AID mediates hypermutation by deaminating single stranded DNA. *J. Exp. Med.* **197**, 1291–1296 (2003).

137. Badran, A. H. & Liu, D. R. In vivo continuous directed evolution. *Current Opinion in Chemical Biology* **24**, 1–10 (2015).

138. Camps, M., Naukkarinen, J., Johnson, B. P. & Loeb, L. A. Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9727–9732 (2003).

139. Bryson, D. I. *et al.* Continuous directed evolution of aminoacyl-tRNA synthetases. *Nat. Chem. Biol.* **13**, 1253–1260 (2017).

140. Jang, S., Kim, M., Hwang, J. & Jung, G. Y. Tools and systems for evolutionary engineering of biomolecules and microorganisms. *J. Ind. Microbiol. Biotechnol.* **46**, 1313–1326 (2019).

141. Loh, E., Salk, J. J. & Loeb, L. A. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1154–9 (2010).

142. Wagner, J. & Nohmi, T. Escherichia coli DNA polymerase IV mutator activity: genetic requirements and mutational specificity. *J. Bacteriol.* **182**, 4587–95 (2000).

143. Bunting, K. A., Roe, S. M. & Pearl, L. H. Structural basis for recruitment of translesion DNA polymerase Pol IV/DinB to the beta-clamp. *EMBO J.* **22**, 5883–92 (2003).

144. Borkowski, O., Ceroni, F., Stan, G.-B. & Ellis, T. Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. *Curr. Opin. Microbiol.* **33**, 123–130 (2016).

145. Ceroni, F. *et al.* Burden-driven feedback control of gene expression. *Nat. Methods* **15**, 387–393 (2018).

146. Storch, M. *et al.* BASIC: A New Biopart Assembly Standard for Idempotent Cloning Provides Accurate, Single-Tier DNA Assembly for Synthetic Biology. *ACS Synth. Biol.* **4**, 781–787 (2015).

147. Ehrt, S. *et al.* Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor.

Nucleic Acids Res. **33**, e21 (2005).

148. Storch, M., Dwijayanti, A., Mallick, H., Haines, M. C. & Baldwin, G. S. BASIC: A Simple and Accurate Modular DNA Assembly Method. in 239–253 (Humana, New York, NY, 2020). doi:10.1007/978-1-0716-0908-8_14

149. Madsen, C. *et al.* Synthetic Biology Open Language (SBOL) Version 2.3. *J. Integr. Bioinform.* **16**, (2019).

150. Hecht, A. *et al.* Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* **45**, 3615–3626 (2017).

151. Nakao, R. *et al.* Illumina iSeq 100 and MiSeq exhibit similar performance in freshwater fish 1 environmental DNA metabarcoding 2 Running title: Fish eDNA metabarcoding using iSeq and MiSeq Correspondence 1 4. *bioRxiv* 2020.08.04.228080 (2020). doi:10.1101/2020.08.04.228080

152. Singh, N. *et al.* IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform. *Meta Gene* **7**, 70–75 (2016).

153. Lang, D. *et al.* Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacbio Sequel II system and ultralong reads of Oxford Nanopore. *bioRxiv* 2020.02.13.948489 (2020). doi:10.1101/2020.02.13.948489

154. Foster, P. L. In vivo mutagenesis. *Methods Enzymol.* **204**, 114–125 (1991).

155. Badran, A. H. & Liu, D. R. Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nat. Commun.* **6**, 1–10 (2015).

156. Muteeb, G. & Sen, R. Random mutagenesis using a mutator strain. *Methods Mol. Biol.* **634**, 411–419 (2010).

157. Arnold, F. H., Georgiou, G., Nguyen, A. W. & Daugherty, P. S. Production of Randomly Mutated Plasmid Libraries Using Mutator Strains. in *Directed Evolution Library Creation* 39–44 (Humana Press, 2003). doi:10.1385/1-59259-395-x:39

158. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature Reviews Genetics* **19**, 770–788 (2018).

159. Zheng, K. *et al.* Highly efficient base editing in bacteria using a Cas9-cytidine deaminase fusion. *Commun. Biol.* **1**, 1–6 (2018).

160. Li, C. *et al.* Expanded base editing in rice and wheat using a Cas9-adenosine deaminase fusion. *Genome Biol.* **19**, 59 (2018).

161. Álvarez, B., Mencía, M., de Lorenzo, V. & Fernández, L. Á. In vivo diversification of target genomic sites using processive T7 RNA polymerase-base deaminase fusions blocked by RNA-guided dCas9. *bioRxiv* 850974 (2019). doi:10.1101/850974

162. Chamberlin, M. & Ryan, T. 4 Bacteriophage DNA-Dependent RNA Polymerases. *Enzymes* **15**, 87–108 (1982).

163. Martin, C. T., Muller, D. K. & Coleman, J. E. Processivity in Early Stages of Transcription by T7 RNA Polymerase. *Biochemistry* **27**, 3966–3974 (1988).

164. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nature Reviews Microbiology* **12**, 381–390 (2014).

165. Smanski, M. J. *et al.* Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* **32**, 1241 (2014).

166. Zelcbuch, L. *et al.* Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Res.* **41**, e98–e98 (2013).

167. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).

168. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).

169. Jeschek, M., Gerngross, D. & Panke, S. Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat. Commun.* **7**, 1–10 (2016).

170. Farasat, I. *et al.* Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Mol. Syst. Biol.* **10**, 731 (2014).

171. Kelly, J. R. *et al.* Measuring the activity of BioBrick promoters using an in vivo reference standard. *J. Biol. Eng.* **3**, 4 (2009).

172. Dennis, P. P. & Bremer, H. Regulation of ribonucleic acid synthesis in Escherichia coli B r: An analysis of a shift-up. 1. Ribosomal RNA chain growth rates. *J. Mol. Biol.* **75**, 145–159 (1973).

173. Erickson, D. W. *et al.* A global resource allocation strategy governs growth transition kinetics of Escherichia coli. *Nature* **551**, 119–123 (2017).

174. Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).

175. Looman, A. C., Bodlaender, J., de Gruyter, M., Vogelaar, A. & van Knippenberg, P. H. Secondary structure as primary determinant of the efficiency of ribosomal binding sites in Escherichia coli. *Nucleic Acids Res.* **14**, 5481–5497 (1986).

176. Jin, L., Nawab, S., Xia, M., Ma, X. & Huo, Y. Context-dependency of synthetic minimal promoters in driving gene expression: a case study. *Microb. Biotechnol.* **12**, 1476–1486 (2019).

177. Jin, L., Nawab, S., Xia, M., Ma, X. & Huo, Y. Context-dependency of synthetic minimal promoters in driving gene expression: a case study. *Microb. Biotechnol.* **12**, 1476–1486 (2019).

178. Serra, M. C. & Haccou, P. Dynamics of escape mutants. *Theor. Popul. Biol.* **72**, 167–178 (2007).

179. d'Oelsnitz, S. & Ellington, A. Continuous directed evolution for strain and protein engineering. *Curr. Opin. Biotechnol.* **53**, 158–163 (2018).

180. Weinreich, D. M., Watson, R. A. & Chao, L. PERSPECTIVE: SIGN EPISTASIS AND GENETIC COSTRAINT ON EVOLUTIONARY TRAJECTORIES. *Evolution (N. Y).* **59**, 1165–1174 (2005).

181. Boucher, J. I. *et al.* Viewing protein fitness landscapes through a next-gen lens. *Genetics* **198**, 461–471 (2014).

182. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. Mutant types. (2000).

183. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2774–E2783 (2012).

184. Sprouffske, K., Aguílar-Rodríguez, J., Sniegowski, P. & Wagner, A. High mutation rates limit evolutionary adaptation in Escherichia coli. *PLoS Genet.* **14**, (2018).

185. Lin, Y. H., Tai, C. H., Li, C. R., Lin, C. F. & Shi, Z. Y. Resistance profiles and rpoB gene mutations of Mycobacterium tuberculosis isolates in Taiwan. *J. Microbiol. Immunol. Infect.* **46**, 266–270 (2013).

186. Zhu, J. H. *et al.* Rifampicin can induce antibiotic tolerance in mycobacteria via paradoxical changes in rpoB transcription. *Nat. Commun.* **9**, 1–13 (2018).

187. Ovchinnikov, Y. A. *et al.* RNA polymerase rifampicin resistance mutations in Escherichia coli: Sequence changes and dominance. *MGG Mol. Gen. Genet.* **190**, 344–348 (1983).

188. Goodman, M. F. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annual Review of Biochemistry* **71**, 17–50 (2002).

189. Foster, P. L. Stress responses and genetic variation in bacteria. *Mutation Research - Fundamental and*

Molecular Mechanisms of Mutagenesis **569**, 3–11 (2005).

190. Rogers, S. G. & Weiss, B. [26] Exonuclease III of Escherichia Coli K-12, an AP Endonuclease. *Methods Enzymol.* **65**, 201–211 (1980).

191. Kim, Y.-J. & Wilson, D. M. Overview of base excision repair biochemistry. *Curr. Mol. Pharmacol.* **5**, 3–13 (2012).

192. Lovett, S. T. The DNA Exonucleases of Escherichia coli. *EcoSal Plus* **4**, (2011).

193. Tabor, S. & Tabor, S. Expression Using the T7 RNA Polymerase/Promoter System. in *Current Protocols in Molecular Biology* 16.2.1-16.2.11 (John Wiley & Sons, Inc., 2001). doi:10.1002/0471142727.mb1602s11

194. Sousa, R., Patra, D. & Lafer, E. M. Model for the mechanism of bacteriophage T7 RNAP transcription initiation and termination. *J. Mol. Biol.* **224**, 319–334 (1992).

195. Carpenter, E. P. *et al.* AP endonuclease paralogues with distinct activities in DNA repair and bacterial pathogenesis. *EMBO J.* **26**, 1363–72 (2007).

196. Wagner, J. *et al.* The dinB gene encodes a novel E. coli DNA polymerase, DNA pol IV, involved in mutagenesis. *Mol. Cell* **4**, 281–286 (1999).

197. Wagner, J., Fujii, S., Gruz, P., Nohmi, T. & Fuchs, R. P. P. The β clamp targets DNA polymerase IV to DNA and strongly increases its processivity. *EMBO Rep.* **1**, 484–488 (2000).

198. Loh, E., Salk, J. J. & Loeb, L. A. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc. Natl. Acad. Sci.* **107**, 1154–1159 (2010).

199. Han, E. S. *et al.* RecJ exonuclease: Substrates, products and interaction with SSB. *Nucleic Acids Res.* **34**, 1084–1091 (2006).

200. Currin, A., Swainston, N., Day, P. J. & Kell, D. B. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.* **44**, 1172–1239 (2015).

201. Nagata, Y., Mashimo, K., Kawata, M. & Yamamoto, K. *The Roles of Klenow Processing and Flap Processing Activities of DNA Polymerase I in Chromosome Instability in Escherichia coli K12 Strains*.

202. Thiel, K. *et al.* Translation efficiency of heterologous proteins is significantly affected by the genetic context of RBS sequences in engineered cyanobacterium Synechocystis sp. PCC 6803. *Microb. Cell Fact.* **17**, 1–12 (2018).

203. Foglieni, C. *et al.* Split GFP technologies to structurally characterize and quantify functional biomolecular interactions of FTD-related proteins. *Sci. Rep.* **7**, 1–15 (2017).

204. Kaddoum, L., Magdeleine, E., Waldo, G. S., Joly, E. & Cabantous, S. One-step split GFP staining for sensitive protein detection and localization in mammalian cells. *Biotechniques* **49**, 727–736 (2010).

205. Yamagata, A., Kakuta, Y., Masui, R. & Fukuyama, K. The crystal structure of exonuclease RecJ bound to Mn2+ ion suggests how its characteristic motifs are involved in exonuclease activity. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 5908–5912 (2002).

206. Kuban, W., Banach-Orlowska, M., Schaaper, R. M., Jonczyk, P. & Fijalkowska, I. J. Role of DNA polymerase IV in Escherichia coli SOS mutator activity. *J. Bacteriol.* **188**, 7977–80 (2006).

207. Wagner, J. & Nohmi, T. Escherichia coli DNA polymerase IV mutator activity: genetic requirements and mutational specificity. *J. Bacteriol.* **182**, 4587–95 (2000).

208. Kuban, W. *et al.* Role of Escherichia coli DNA polymerase IV in in vivo replication fidelity. *J. Bacteriol.* **186**, 4802–4807 (2004).

209. Morlock, G. P., Plikaytis, B. B. & Crawford, J. T. Characterization of spontaneous, In vitro-selected, rifampin-resistant mutants of Mycobacterium tuberculosis strain H37Rv. *Antimicrob. Agents Chemother.* **44**, 3298–301 (2000).

210. Simon, C., Herath, J., Rockstroh, S. & Daniel, R. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl. Environ. Microbiol.* **75**, 2964–8 (2009).

211. Huang, Y.-P., Downie, J. A. & Ito, J. *Primary Structure of the DNA Polymerase I Gene of an-Proteobacterium, Rhizobium leguminosarum, and Comparison with Other Family A DNA Polymerases*. *J. Ito CURRENT MICROBIOLOGY* **38**, (Springer-Verlag New York Inc, 1999).

212. Bebeneks, K., Joycei, C. M., Fitzgerald+, M. P. & Kunkelsn, T. A. *The Fidelity of DNA Synthesis Catalyzed by Derivatives of Escherichia coli DNA Polymerase I**. **265**, (1990).

213. Engstrom, M. D. & Pfleger, B. F. Transcription control engineering and applications in synthetic biology. *Synthetic and Systems Biotechnology* **2**, 176–191 (2017).

214. Strom, S. P. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biology and Medicine* **13**, 3–11 (2016).

215. Nagy, P. L. & Worman, H. J. Next-Generation Sequencing and Mutational Analysis: Implications for Genes Encoding LINC Complex Proteins. in *Methods in Molecular Biology* **1840**, 321–336 (Humana Press Inc., 2018).

216. Mazouzi, A. *et al.* Insight into mechanisms of 3'-5' exonuclease activity and removal of bulky 8,5'-cyclopurine adducts by apurinic/apyrimidinic endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3071–E3080 (2013).

217. Golan, G., Ishchenko, A. A., Khassenov, B., Shoham, G. & Saparbaev, M. K. Coupling of the nucleotide incision and 3' → 5' exonuclease activities in Escherichia coli endonuclease IV: Structural and genetic evidences. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **685**, 70–79 (2010).

218. Wang, T. W. *et al.* Mutant library construction in directed molecular evolution: Casting a wider net. *Molecular Biotechnology* **34**, 55–68 (2006).

219. Rozkov, A. *et al.* Characterization of the metabolic burden onEscherichia coli DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. *Biotechnol. Bioeng.* **88**, 909–915 (2004).

220. Liu, Q., Schumacher, J., Wan, X., Lou, C. & Wang, B. Orthogonality and Burdens of Heterologous and Gate Gene Circuits in E. coli. *ACS Synth. Biol.* **7**, 553–564 (2018).

221. Henrikus, S. S. *et al.* DNA polymerase IV primarily operates outside of DNA replication forks in Escherichia coli. *PLoS Genet.* **14**, (2018).

222. Duret, L. Neutral Theory: The Null Hypothesis of Molecular Evolution. *Nat. Educ.* **1**, 803–806 (2008).

223. Fraebel, D. T. *et al.* Environment determines evolutionary trajectory in a constrained phenotypic space. *Elife* **6**, (2017).

224. Hanawalt, P. C. Density matters: The semiconservative replication of DNA. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 17889–17894 (2004).

225. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* **13**, 278–289 (2015).

226. Levene, H. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (80-.).* **299**, 682–686 (2003).

227. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

228. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Research* **46**, 2159–2168 (2018).

229. Jiao, X. A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the

PacBio RS. *J. Data Mining Genomics Proteomics* **04**, (2013).

230. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).

231. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

232. Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y. & Fostier, J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* **18**, 1–13 (2017).

233. Nikolenko, S. I., Korobeynikov, A. I. & Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14**, S7 (2013).

234. Chen, S. *et al.* AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* **18**, 80 (2017).

235. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**, 37 (2015).

236. Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved Protocols for Illumina Sequencing. *Curr. Protoc. Hum. Genet.* **79**, 18.2.1-18.2.42 (2013).

237. Parikh, H. I. *et al. MeFiT: Merging and Filtering Tool for Illumina Paired-End Reads for 16S rRNA Amplicon Sequencing*.

238. Wang, A., Wang, Z., Li, Z. & Li, L. M. BAUM: Improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics* **34**, 2019–2028 (2018).

239. Liu, Q. *et al.* Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* **13 Suppl 8**, 1–8 (2012).

240. Zhang, S., Wang, B., Wan, L. & Li, L. M. Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling. *BMC Bioinformatics* **18**, 335 (2017).

241. Prosdocimi[1], F., Peixoto[2], C. & Ortega[3], J. M. *DNA SEQUENCES BASE CALLING BY PHRED: ERROR PATTERN ANALYSIS*.

242. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**, 704–714 (2016).

243. Trombetta, J. J. *et al.* Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4.22.1-4.22.17 (2014).

244. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

245. Zhao, J., Kardashliev, T., Joëlle Ruff, A., Bocola, M. & Schwaneberg, U. Lessons from diversity of directed evolution experiments by an analysis of 3,000 mutations. *Biotechnol. Bioeng.* **111**, 2380–2389 (2014).

246. Neylon, C. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: Library construction methods for directed evolution. *Nucleic Acids Research* **32**, 1448–1459 (2004).

247. Jee, J. *et al.* Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* **534**, 693–696 (2016).

248. Roa, S., Fei, L. K. & Scharff, M. D. Does antisense make sense of AID targeting? *Proceedings of the National Academy of Sciences of the United States of America* **105**, 3661–3662 (2008).

249. Bowers, P. M. *et al.* Nucleotide insertions and deletions complement point mutations to massively expand the diversity created by somatic hypermutation of antibodies. *J. Biol. Chem.* **289**, 33557–33567 (2014).

250. Shimatani, Z. *et al.* Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion. *Nat. Biotechnol.* **35**, 441–443 (2017).

251. Wong, T. S., Roccatano, D. & Schwaneberg, U. Are transversion mutations better? A Mutagenesis Assistant Program analysis on P450 BM-3 heme domain. *Biotechnol. J.* **2**, 133–142 (2007).

252. Kobayashi, S., Valentine, M. R., Pham, P., O'Donnell, M. & Goodman, M. F. Fidelity of Escherichia coli DNA polymerase IV. Preferential generation of small deletion mutations by dNTP-stabilized misalignment. *J. Biol. Chem.* **277**, 34198–34207 (2002).

253. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, 1–20 (2013).

254. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

255. Foster, P. L. Escherichia coli strains with multiple DNA repair defects are hyperinduced for the SOS response. *J. Bacteriol.* **172**, 4719–4720 (1990).

256. Arnold, F. H., Georgiou, G., Cirino, P. C., Mayer, K. M. & Umeno, D. Generating Mutant Libraries Using Error-Prone PCR. in *Directed Evolution Library Creation* **231**, 3–10 (Humana Press, 2003).

257. Foster, P. L., Lee, H., Popodi, E., Townes, J. P. & Tang, H. Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5990–E5999 (2015).

258. Williams, A. B. Spontaneous mutation rates come into focus in Escherichia coli. *DNA Repair* **24**, 73–79 (2014).

259. Nyström, T. Stationary-phase physiology. *Annual Review of Microbiology* **58**, 161–181 (2004).

260. Loewe, L., Textor, V. & Scherer, S. High Deleterious Genomic Mutation Rate in Stationary Phase of Escherichia coli. *Science (80-.).* **302**, 1558–1560 (2003).

261. Sneppen, K. *et al.* A mathematical model for transcriptional interference by RNA polymerase traffic in Escherichia coli. *J. Mol. Biol.* **346**, 399–409 (2005).

262. Li, M. *et al.* A strategy of gene overexpression based on tandem repetitive promoters in Escherichia coli. *Microb. Cell Fact.* **11**, 1–10 (2012).

263. Besmer, E., Market, E. & Papavasiliou, F. N. The Transcription Elongation Complex Directs Activation-Induced Cytidine Deaminase-Mediated DNA Deamination † Downloaded from. *Mol. Cell. Biol.* **26**, 4378–4385 (2006).

264. Shen, H. M. & Storb, U. *Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled*. *PNAS August* **31**, (2004).

265. Tóth-Petróczy, Á. & Tawfik, D. S. Hopeful (protein InDel) monsters? *Structure* **22**, 803–804 (2014).

266. Emond, S. *et al.* Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nat. Commun.* **11**, 1–14 (2020).

267. Wagner, J. *et al.* Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol.* **16**, 1–17 (2016).

268. Sheikhizadeh, S. & De Ridder, D. ACE: Accurate correction of errors using K-mer tries. *Bioinformatics* **31**, 3216–3218 (2015).

269. Manley, L. J., Ma, D. & Levine, S. S. Monitoring error rates in Illumina sequencing. *J. Biomol. Tech.* **27**, 125–128 (2016).

270. Ledergerber, C. & Dessimoz, C. Base-calling for next-generation sequencing platforms. *Brief. Bioinform.* **12**, 489–497 (2011).

271. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90–

e90 (2011).

272. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).

273. Wang, B., Wan, L., Wang, A. & Li, L. M. An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters. *Sci. Rep.* **7**, 1–11 (2017).

274. Leconte, A. M. *et al.* A population-based experimental model for protein evolution: Effects of mutation rate and selection stringency on evolutionary outcomes. *Biochemistry* **52**, 1490–1499 (2013).

275. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).

276. Hawkins, A. C., Arnold, F. H., Stuermer, R., Hauer, B. & Leadbetter, J. R. Directed evolution of Vibrio fischeri LuxR for improved response to butanoyl-homoserine lactone. *Appl. Environ. Microbiol.* **73**, 5775–81 (2007).

277. Reetz, M. T. *Introduction to Directed Evolution 1.1 General Definition and Purpose of Directed Evolution of Enzymes*. (2017).

278. Gillam, E. M. J., Copp, J. N. & Ackerley Editors, D. F. *Directed Evolution Library Creation Methods and Protocols Second Edition Methods in Molecular Biology 1179*.

279. Rockah-Shmuel, L., Tawfik, D. S. & Goldsmith, M. Generating targeted libraries by the combinatorial incorporation of synthetic oligonucleotides during gene shuffling (ISOR). *Methods Mol. Biol.* **1179**, 129–137 (2014).

280. Wang, J. D., Herman, C., Tipton, K. A., Gross, C. A. & Weissman, J. S. Directed evolution of substrate-optimized GroEL/S chaperonins. *Cell* **111**, 1027–39 (2002).

281. Leemhuis, H., Kelly, R. M. & Dijkhuizen, L. Directed evolution of enzymes: Library screening strategies. *IUBMB Life* **61**, 222–228 (2009).

282. Jang, S., Kim, M., Hwang, J. & Jung, G. Y. Tools and systems for evolutionary engineering of biomolecules and microorganisms. *J. Ind. Microbiol. Biotechnol.* **46**, 1313–1326 (2019).

283. Popa, S. C., Inamoto, I., Thuronyi, B. W. & Shin, J. A. *Phage Assisted Continuous Evolution (PACE): a How-to Guide for Directed Evolution*. (2019).

284. Hsu, C. T. *et al.* Application of Cas12a and nCas9-activation-induced cytidine deaminase for genome editing and as a non-sexual strategy to generate homozygous/multiplex edited plants in the allotetraploid genome of tobacco. *Plant Mol. Biol.* **101**, 355–371 (2019).

285. Yin, J., Li, G., Ren, X. & Herrler, G. Select what you need: A comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. *Journal of Biotechnology* **127**, 335–347 (2007).

286. Verma, R., Boleti, E. & George, A. J. T. Antibody engineering: Comparison of bacterial, yeast, insect and mammalian expression systems. *J. Immunol. Methods* **216**, 165–181 (1998).

287. Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nature Reviews Immunology* **6**, 573–583 (2006).

288. Pérez-Durán, P. *et al.* UNG shapes the specificity of AID-induced somatic hypermutation. *J. Exp. Med.* **209**, 1379–1389 (2012).

289. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics. Proteomics Bioinformatics* **13**, 278–289 (2015).

290. Grünewald, J. *et al.* A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. *Nat. Biotechnol.* **38**, 861–864 (2020).

291. Makarova, O. V., Makarov, E. M., Sousa, R. & Dreyfus, M. Transcribing of Escherichia coli genes with

mutant T7 RNA polymerases: Stability of lacZ mRNA inversely correlates with polymerase speed. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 12250–12254 (1995).

292. Fuqua, W. C., Winans, S. C. & Greenberg, E. P. Quorum sensing in bacteria: The LuxR-LuxI family of cell density- responsive transcriptional regulators. *Journal of Bacteriology* **176**, 269–275 (1994).

293. Haldimann, A. & Wanner, B. L. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *J. Bacteriol.* **183**, 6384–93 (2001).

294. Marinus, M. G. & Morris, N. R. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J. Bacteriol.* **114**, 1143–50 (1973).

295. Inoue, H., Nojima, H. & Okayama, H. High efficiency transformation of Escherichia coli with plasmids. *Gene* **96**, 23–28 (1990).

296. Brunker, K. DNA quantification using the Qubit fluorometer. (2020). doi:10.17504/protocols.io.bc6vize6

297. CoreGenomics: How do SPRI beads work? Available at: http://core-genomics.blogspot.com/2012/04/how-do-spri-beads-work.html. (Accessed: 17th August 2020)

298. Stortchevoi, A., Kamelamela, N. & Levine, S. S. SPRI beads-based size selection in the range of 2-10kb. *J. Biomol. Tech.* **31**, 7–10 (2020).

299. Cui, L. & Shearwin, K. E. Clonetegration using OSIP plasmids: One-step DNA assembly and site-specific genomic integration in bacteria. in *Methods in Molecular Biology* **1472**, 139–155 (Humana Press Inc., 2017).

300. Pan, G., Luetke, K. & Sadowski, P. D. Mechanism of cleavage and ligation by FLP recombinase: classification of mutations in FLP protein by in vitro complementation analysis. *Mol. Cell. Biol.* **13**, 3167–3175 (1993).

301. Oligo synthesis: Why IDT leads the oligo industry. Available at: https://eu.idtdna.com/pages/education/decoded/article/oligo-synthesis-why-idt-leads-the-oligo-industry. (Accessed: 16th August 2020)

302. Addgene: Protocol - Bacterial Transformation. Available at: https://www.addgene.org/protocols/bacterial-transformation/. (Accessed: 17th August 2020)

303. Silva-Rocha, R. *et al.* The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res.* **41**, D666-75 (2013).

304. Martínez-Garćía, E., Aparicio, T., Goñi-Moreno, A., Fraile, S. & De Lorenzo, V. SEVA 2.0: An update of the Standard European Vector Architecture for de-/re-construction of bacterial functionalities. *Nucleic Acids Res.* **43**, D1183–D1189 (2015).

305. Martínez-García, E. *et al.* SEVA 3.0: An update of the Standard European Vector Architecture for enabling portability of genetic constructs among diverse bacterial hosts. *Nucleic Acids Res.* **48**, D1164–D1170 (2020).

306. Bindels, D. S. *et al.* MScarlet: A bright monomeric red fluorescent protein for cellular imaging. *Nat. Methods* **14**, 53–56 (2016).

307. McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol. Biol. Int.* **2014**, 1–8 (2014).

308. Iyer, R. P. & Beaucage, S. L. Oligonucleotide Synthesis. in *Comprehensive Natural Products Chemistry* 105–152 (Elsevier, 1999). doi:10.1016/b978-0-08-091283-7.00126-0

309. DNA Oligonucleotide Synthesis | Sigma-Aldrich. Available at: https://www.sigmaaldrich.com/technical-documents/articles/biology/dna-oligonucleotide-synthesis.html. (Accessed: 3rd September 2020)

310. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format

for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).

311. Lamble, S. *et al.* Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol.* **13**, 1–10 (2013).

312. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (80-.).* **355**, 752–756 (2017).

Chapter 9: Appendix

9.1 Chapters 3 and 4 – List of BASIC Linker Sequences

Table 9.1: List of BASIC Neutral Linkers

Function: The neutral linkers are the simplest of linker pairs in the BASIC library. Their only function is to join DNA components together and do not server a genetic function.

Linker Name	Abbreviation	Linker Sequence (5’ → 3’)
Linker 1	L1	ctcgttacttacgacaCTCCGAGACAGTCAGAGGGTAtttattgaactagtcc
Linker 2	L2	ctcgatcggtgtgaaaAGTCAGTATCCAGTCGTGTAGttcttattacctgtcc
Linker 3	L3	ctcgatcacggcactaCACTCGTTGCTTTATCGGTATtgttattacagagtcc
Linker 4	L4	ctcgagaagtagtgccACAGACAGTATTGCTTACGAGttgatttatcctgtcc
Linker 5	L5	ctcggtattgtaaagcACGAAACCTACGATAAGAGTGtcagttctccttgtcc
Linker 6	L6	ctcgaacttttacgggTGCCGACTCACTATTACAGACttactacaatctgtcc

Table 9.2: List of BASIC Methlated Linkers

The 21-bp overhangs of methylated linkers contain an intact BsaI recognition site. This allows these linkers to serve a very important function in BASIC assembly, the ability to cut an assembled plasmid with BsaI. This is particularly important when assembling genetic circuits containing 10 or more components. Assembling so many parts simultaneously reduces the efficiency of the process. Instead, the process can be split into assembling two or more sets of five DNA parts and sequentially combining the sets into the final circuit with significantly greater efficiency.

Since these linkers contain an active restriction enzyme recognition site, they are methylated by Dam methylase to prevent the host cell from digesting the plasmid in vivo. Both methylated linkers (LMP and LMS) must be used to successfully cut the desired DNA components from an assembled plasmid. LMP is used at the 5’ end of the full circuit to be assembled and LMS is used at the 3’ end, allowing the entire circuit to be flanked with BsaI recognition and cut sites.

Linker Sequence	Abbreviation	Linker Sequence (5’ → 3’)
Methylated Linker A	LMP	ctcgggtaagaactcgCACTTCGTGGAAACACTATTAtctggtgggtctctgtcc
Methylated Linker B	LMS	ctcgggagacctatcgGTAATAACAGTCCAATCTGGTGTaacttcggaatcgtcc

Table 9.3: List of Degerative RBS Linkers

RBS sequences interact with the ribosome during translation initiation and is an essential genetic component in a circuit. RBS sequences are only 3-9 bp long, which makes them a tricky Biopart to configure for a DNA assembly reaction. Instead, the RBS are encoded into the BASIC linkers between the 21-bp overhang and the BsaI cut site to simplify the assembly process.

There are currently 15 unique RBS sequences encoded into three different linker families. Within a linker family, the 21-bp overhangs are identical with unique RBS sequences upstream.

Three RBS linkers are designed to possess a degenerative RBS library. If the 15 defined RBS sequences to not provide stable expression of the target gene, the degenerative RBS linkers can be screened experimentally to identify a suitable RBS. The three degenerative linkers encode 12, 24 and 36 randomised RBS linkers.

Linker Name	Abbreviation	Linker Sequence (5’ → 3’)
UTR1 Denerative RBS Library	U1-DegRBS12	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatcYVRGGAGGtagtcc
	U1-DegRBS24	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatcYVRGGRGGtagtcc
	U1-DegRBS36	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatcBVRGGRGGtagtcc
UTR2 Denerative RBS Library	U2-DegRBS12	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatcYVRGGAGGtagtcc
	U2-DegRBS24	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatcYVRGGRGGtagtcc
	U2-DegRBS36	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatcBVRGGRGGtagtcc
UTR3 Denerative RBS Library	U3-DegRBS12	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatcYVRGGAGGtagtcc
	U3-DegRBS24	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatcYVRGGRGGtagtcc
	U3-DegRBS36	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatcBVRGGRGGtagtcc

Table 9.4: List of Fusion Linkers

The fusion linkers are designed to create fusion proteins. They have been designed to encode complete codons to ensure no frameshifts are introduced to the coding sequence of the downstream protein in the chimera. Three of the fusion linkers (FL2, FL3, FL4) are designed to be flexible, containing glycine-serine repeats. FL6 is an α-helical rigid linker.

Linker Name	Abbreviation	Linker Sequence (5’ → 3’)	Linker Feature
Fusion Linker 1	FL1	CTCGGGCTCGGGCTCCGAAAACTTGTACTTCCAGGGATCGGGCTCCGGGTCC	TEV site
Fusion Linker 2	FL2	CTCGGGCTCGGGCTCCGGATCTGGTTCAGGTTCAGGATCGGGCTCCGGGTCC	18-aa GS flexible
Fusion Linker 3	FL3	CTCGGGCTCGGGCTCCGGATCAGGATCTGGTTCAGGTTCAGGATCGGGCTCCGGGTCC	20-aa GS flexible
Fusion Linker 4	FL4	CTCGGGCTCGGGCTCCGGATCAGGATCTGGTTCAGGTTCAGGATCAGGATCGGGCTCCGGGTCC	22-aa GS flexible
Fusion Linker 5	FL4	CTCGGGCTCGGGCTCCCTGGAAGTTCTGTTTCAAGGTCCATCGGGCTCCGGGTCC	3C site
Fusion Linker 6	FL6	CTCGGCCGAAGCGGCTGCTAAAGAAGCAGCTGCTAAAGAGGCGGCCGCCAAGGCAGGGTCC	21-aa rigid

Table 9.5: List of RBS Containg BASIC linkers

Linker Name	Abbreviation	Linker Sequence (5’ → 3’)
UTR-1	U1-RBS1	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatccaaggaggtagtcc
	U1-RBS2	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatccagggaggtagtcc
	U1-RBS3	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatcccaggaggtagtcc
	U1-RBS4	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatcccgggaggtagtcc
	U1-RBS5	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatccgaggaggtagtcc
	U1-RBS6	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatccggggaggtagtcc
	U1-RBS7	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatctaaggaggtagtcc
	U1-RBS8	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatctagggaggtagtcc
	U1-RBS9	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatctcaggaggtagtcc
	U1-RBS10	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatctcgggaggtagtcc
	U1-RBS11	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatctgaggaggtagtcc
	U1-RBS12	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatctggggaggtagtcc
	U1-RBS13	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAatcacacaggactagtcc
	U1-RBS14	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAaaagaggggaaatagtcc
	U1-RBS15	ctcgttgaacaccgtcTCAGGTAAGTATCAGTTGTAAaaagaggagaaatagtcc
UTR-2	U2-RBS1	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatccaaggaggtagtcc
	U2-RBS2	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatccagggaggtagtcc
	U2-RBS3	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatcccaggaggtagtcc
	U2-RBS4	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatcccgggaggtagtcc
	U2-RBS5	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatccgaggaggtagtcc
	U2-RBS6	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatccggggaggtagtcc
	U2-RBS7	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatctaaggaggtagtcc
	U2-RBS8	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatctagggaggtagtcc
	U2-RBS9	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatctcaggaggtagtcc
	U2-RBS10	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatctcgggaggtagtcc
	U2-RBS11	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatctgaggaggtagtcc
	U2-RBS12	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatctggggaggtagtcc
	U2-RBS13	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAatcacacaggactagtcc
	U2-RBS14	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAaaagaggggaaatagtcc
	U2-RBS15	ctcgtgttactattggCTGAGATAAGGGTAGCAGAAAaaagaggagaaatagtcc
UTR-3	U3-RBS1	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatccaaggaggtagtcc
	U3-RBS2	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatccagggaggtagtcc
	U3-RBS3	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatcccaggaggtagtcc
	U3-RBS4	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatcccgggaggtagtcc
	U3-RBS5	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatccgaggaggtagtcc
	U3-RBS6	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatccggggaggtagtcc
	U3-RBS7	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatctaaggaggtagtcc
	U3-RBS8	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatctagggaggtagtcc
	U3-RBS9	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatctcaggaggtagtcc

	U3-RBS10	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatctcgggaggtagtcc
	U3-RBS11	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatctgaggaggtagtcc
	U3-RBS12	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatctggggaggtagtcc
	U3-RBS13	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAatcacacaggactagtcc
	U3-RBS14	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAaaagaggggaaatagtcc
	U3-RBS15	ctcggtatctcgtggtCTGACGGTAAAATCTATTGTAaaagaggagaaatagtcc

9.2 Chapter 3 – IUPAC Code for Degenerative Nucleotide Bases Symbols

List of IUPAC Degenerate Base Symbols							
Description	**Symbol**	**Nucleotide Base Represented by Symbol**					**Complimentary Base**
		No. of Bases Represented	**A**	**C**	**G**	**T**	
Adenine	**A**	1	A				T
Cytosine	**C**			C			G
Guanine	**G**				G		C
Thymine	**T**					T	A
Uracil	**U**					U	A
Weak	**W**	2	A			T	W
Strong	**S**			C	G		S
A*m*ino	**M**		A	C			K
Keto	**K**				G	T	M
Pu*r*ine	**R**		A		G		Y
P*y*rimidine	**Y**			C		T	R
Not A	**B**	3		C	G	T	V
Not C	**D**		A		G	T	H
Not G	**H**		A	C		T	D
Not T	**V**		A	C	G		B
An*y* one base	**N**	4	A	C	G	T	N
Zero	**Z**	0					Z

9.3 Chapters 3 and 4 – List of DNA parts and Primers Used in this Study

Table 9.6: List of Promoter and Terminator Bioparts

Part Name	Type	DNA Sequence (5' --> 3')
J23101TetO	Inducible	TCTGGTGGGTCTCTGTCCATCTACCTCAGCTTTACAGCTAGCTCAGTCCTAGGTATTATGCTAGCACTCTATCATTGATAGAGTGGACACGTGGCTCCCTATCAGTGATAGAGATTCACTGCGTAAGGCTCGGGAGACCTATCG
J23106	Constitutive	tctggtgggtctctgtccCTCGGTACCAAATTCCAGAAAAGAGGCCTCCCGAAAGGGGGGCCTTTTTTCGTTTTGGTCCGTGCCTACTCTGGAAAATCTTTTACGGCTAGCTCAGTCCTAGGTATAGTGCTAGCAGCTGTCACCGGATGTGCTTTCCGGTCTGATGAGTCCGTGAGGACGAAACAGCCTCTACAAATAATTTTGTTTAAggctcgggagacctatcg
J23108	Constitutive	tctggtgggtctctgtccCTCGGTACCAAATTCCAGAAAAGAGACGCTTTCGAGCGTCTTTTTTCGTTTTGGTCCGTGCCTACTCTGGAAAATCTCTGACAGCTAGCTCAGTCCTAGGTATAATGCTAGCAGCGCTCAACGGGTGTGCTTCCCGTTCTGATGAGTCCGTGAGGACGAAAGCGCCTCTACAAATAATTTTGTTTAAggctcgggagacctatcg
J23116	Constitutive	tctggtgggtctctgtccCTCGGTACCAAATTCCAGAAAAGAGGCCTCCCGAAAGGGGGGCCTTTTTTCGTTTTGGTCCGTGCCTACTCTGGAAAATCTTTGACAGCTAGCTCAGTCCTAGGGACTATGCTAGCAGCTGTCACCGGATGTGCTTTCCGGTCTGATGAGTCCGTGAGGACGAAACAGCCTCTACAAATAATTTTGTTTAAggctcgggagacctatcg
T7 Promoter	Orthogonal to T7 Polymerase	TCTGGTGGGTCTCTGTCCCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGTACTAGAGGGCTCGGGAGACCTATCG
J23116 + pT7	Double Promoter	TCTGGTGGGTCTCTGTCCacaattttcgaaaaaacccgcttcggcgggtttttttatagctaaaagatttgacagctagctcagtcctagggattgtgctagcgcgtccggcgtagaggatcgagatctcgatcccgcgaaattaatacgactcactatagggTACTAGAGGGCTCGGGAGACCTATCG
B15 Terminator	Bacterial RNA Polymerase Terminator	TCTGGTGGGTCTCTGTCCccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctctactagagtcacactggctcaccttcgggtgggcctttctgcgtttataGGTGGGCCTTTCTGCGTTTATAGGCTCGGGAGACCTATCG
T7 Double Terminator	Phage RNA Polymerase Terminator	TCTGGTGGGTCTCTGTCCCCTAGCATAACCCCGCGGGGCCTCTTCGGGGGaCTCGCGGGGTTTTTTGCTGAAAGAattaTCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGCTGCattaCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGGGCTCGGGAGACCTATCG

Table 9.7: List of Bioparts Used in Project

Part Name	Type	DNA Sequence (5' --> 3')
sfGFP	Fluorescent Protein	TCTGGTGGGTCTCTGTCCatgcgtaaaggcgaagaactgttcacgggcgtagttccgattctggtcgagctggacggcgatgtgaacggtcataagtttagcgttcgcggtgaaggtgagggcgacgcgaccaacggcaaactgaccctgaagttcatctgcaccaccggtaaactgccggtgccttggccgaccttggtgacgacgttgacgtatggcgtgcagtgttttgcgcgttatccggaccacatgaaacaacacgatttcttcaaatctgcgatgccggagggttacgtccaggagcgtaccatttccttcaaggatgatggcacttacaaaactcgcgcagaggttaagtttgaaggtgacacgctggtcaatcgtatcgaattgaagggtatcgactttaaagaggatggtaacattctgggccataaactggagtataacttcaacagccataatgtttacattacg

		gcagacaagcaaaagaacggcatcaaggccaatttcaagattcgccacaatgttgaggacggtagcgtccaactggccgaccattaccagcagaacaccccaattggtgacggtccggttttgctgccggataatcactatctgagcacccaaagcgtgctgagcaaagatccgaacgaaaaacgtgatcacatggtcctgctggaatttgtgaccgctgcgggcatcacccacggtatggacgagctgtataaGGCTCGGGAGACCTATCG
GFP-mut3b	Fluorescent Protein	tctggtgggtctctGTCCATGCGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGTCACTACTTTCGGTTATGGTGTTCAATGCTTTGCGAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATATTTTTCAAAGATGACGGGAACTACAAGACACGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATAGAATCGAGTTAAAAGGTATTGATTTTAAAGAAGATGGAAACATTCTTGGACACAAATTGGAATACAACTATAACTCACACAATGTATACATCATGGCAGACAAACAAAAGAATGGAATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGAAGCGTTCAACTAGCAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCCACACAATCTGCCCTTTCGAAAGATCCCAACGAAAAGAGAGATCACATGGTCCTTCTTGAGTTTGTAACAGCTGCTGGGATTACACATGGCATGGATGAACTATACAAATAAGGCTCGggagacctatcg
Split GFP	Fluorescent Protein	TCTGGTGGGTCTCTGTCCACAAGTCGTGACCACATGGTCCTTCATGAGTACGTAAATGCTGCTGGGATTACATAATAATACTAGAGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATATAATACTAGAGTTTACAGCTAGCTCAGTCCTAGGTATTATGCTAGCTACTAGAGAAAGAGGAGAAATACTAGATGAGCAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGAGGAGAGGGTGAAGGTGATGCTACAATCGGAAAACTCACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGTCACTACTCTGACCTATGGTGTTCAATGCTTTTCCCGTTATCCGGATCACATGAAAAGGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAACGCACTATATCTTTCAAAGATGACGGGAAATACAAGACGCGTGCTGTAGTCAAGTTTGAAGGTGATACCCTTGTTAATCGTATCGAGTTAAAGGGTACTGATTTTAAAGAAGATGGAAACATTCTCGGACACAAACTCGAGTACAACTTTAACTCACACAATGTATACATCACGGCAGACAAACAAAAGAATGGAATCAAAGCTAACTTCACAGTTCGCCACAACGTTGAAGATGGTTCCGTTCAACTAGCAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCGACACAAACTGTCCTTTCGAAAGATCCCAACGAAAAGTAATAATACTAGAGTccggcaattaaaaaagcggctaaccacgccgctttttttacgtctgcaGGCTCGGGAGACCTATCG
mCherry	Fluorescent Protein	TCTGGTGGGTCTCTGTCCATGGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCTTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAGACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGTAAGGCTCGGGAGACCTATCG
AID	Cytidine Deaminase	tctggtgggtctctgtccatggatagcctgctgatgaatcgtcgcaaatttctgtatcagtttaaaaatgtgcgttgggccaaaggtcgtcgtgaaacctatctgtgctatgttgtgaaacgtcgtgatagcgcaaccagctttagcctggattttggttatctgcgcaataaaaatggttgtcatgtggagctgctgtttctgcgttatattagcgattgggatctggatccgggtcgttgttatcgtgttacctggtttaccagctggtcaccgtgttatgactgtgcacgtcatgttgcagattttctgcgtggtaatccgaatctgagcctgcgtatttttaccgcacgtctgtatttttgcgaagatcgtaaagcagaaccggaaggtctgcgtcgtctgcatcgtgcaggtgttcagattgccattatgacctttaaagattatttttattgctggaatacctttgtggaaaatcatgaacgcacctttaaagcatgggaaggcctgcatgaaaatagcgttcgtctgtctcgccagctgcgtcgtattctgctgccgctgtatgaagttgatgatctgcgtgatgcctttcgtaccctgggtttaaccggtcaggaagttgcatgcaccgccggcggctcgggagacctatcg
T7 RNA Polymerase	Polymerase orthogonal to T7 promoter	TCTGGTGGGTCTCTGTCCatgaacacgattaacatcgctaagaacgacttctctgacatcgaactggctgctatcccgttcaacactctggctgaccattacggtgagcgtttagctcgcgaacagttggcccttgagcatgagtcttacgagatgggtgaagcacgcttccgcaagatgtttgagcgtcaacttaaagctggtgaggttgcggataacgctgccgccaagcctctcatcactaccctactccctaagatgattgcacgcatcaacgactggtttgaggaagtgaaagctaagcgcggcaagcgcccgacagccttccagttcctgcaagaaatcaagccggaagccgtagcgtacatcaccattaagaccactctggcttgcctaaccagtgctgacaatacaaccgttcaggctgtagcaagcgcaatcggtcgggccattgaggacgaggctcgcttcggtcgtatccgtgaccttgaagctaagcacttcaagaaaaacgttgaggaacaactcaacaagcgcgtagggcacgtctacaag

		aaagcatttatgcaagttgtcgaggctgacatgctctctaagggtctactcggtggcgaggcgtggtcttcgtggcataaggaagactctattcatgtaggagtacgctgcatcgagatgctcattgagtcaaccggaatggttagcttacaccgccaaaatgctggcgtagtaggtcaagactctgagactatcgaactcgcacctgaatacgctgaggctatcgcaacccgtgcaggtgcgctggctggcatctctccgatgttccaaccttgcgtagttcctcctaagccgtggactggcattactggtggtggctattgggctaacggtcgtcgtcctctggcgctggtgcgtactcacagtaagaaagcactgatgcgctacgaagacgtttacatgcctgaggtgtacaaagcgattaacattgcgcaaaacaccgcatggaaaatcaacaagaaagtcctagcggtcgccaacgtaatcaccaagtggaagcattgtccggtcgaggacatccctgcgattgagcgtgaagaactcccgatgaaaccggaagacatcgacatgaatcctgaggctctcaccgcgtggaaacgtgctgccgctgctgtgtaccgcaaggacagggctcgcaagtctcgccgtatcagccttgagttcatgcttgagcaagccaataagtttgctaaccataaggccatctggttcccttacaacatggactggcgcggtcgtgtttacgccgtgtcaatgttcaacccgcaaggtaacgatatgaccaaaggactgcttacgctggcgaaaggtaaaccaatcggtaaggaaggttactactggctgaaaatccacggtgcaaactgtgcgggtgtcgataaggttccgttccctgagcgcatcaagttcattgaggaaaaccacgagaacatcatggcttgcgctaagtctccactggagaacacttggtgggctgagcaagattctccgttctgcttccttgcgttctgctttgagtacgctggggtacagcaccacggcctgagctataactgctcccttccgctggcgtttgacgggtcttgctctggcatccagcacttctccgcgatgctccgagatgaggtaggtggtcgcgcggttaacttgcttcctagtgagaccgttcaggacatctacgggattgttgctaagaaagtcaacgagattctacaagcagacgcaatcaatgggaccgataacgaagtagttaccgtgaccgatgagaacactggtgaaatctctgagaaagtcaagctgggcactaaggcactggctggtcaatggctggctcacggtgttactcgcagtgtgactaagcgttcagtcatgacgctggcttacgggtccaaagagttcggcttccgtcaacaagtgctggaagataccattcagccagctattgattccggcaagggtccgatgttcactcagccgaatcaggctgctggatacatggctaagctgatttgggaatctgtgagcgtgacggtggtagctgcggttgaagcaatgaactggcttaagtctgctgctaagctgctggctgctgaggtcaaagataagaagactggagagattcttcgcaagcgttgcgctgtgcattgggtaactcctgatggtttccctgtgtggcaggaatacaagaagcctattcagacgcgcttgaacctgatgttcctcggtcagttccgcttacagcctaccattaacaccaacaaagatagcgagattgatgcacacaaacaggagtctggtatcgctcctaactttgtacacagccaagacggtagccaccttcgtaagactgtagtgtgggcacacgagaagtacggaatcgaatcttttgcactgattcacgactccttcggtaccattccggctgacgctgcgaacctgttcaaagcagtgcgcgaaactatggttgacacatatgagtcttgtgatgtactggctgatttctacgaccagttcgctgaccagttgcacgagtctcaattggacaaaatgccagcacttccggctaaaggtaacttgaacctccgtgacatcttagagtcggacttcgcgttcgcgtaaGGCTCGGGAGACCTATCG
AID-T7pol	DNA Damage Device; AID fused to T7pol	tctggtgggtctctgtccATGGATAGCCTGCTGATGAATCGTCGCAAATTTCTGTATCAGTTTAAAAATGTGCGTTGGGCCAAAGGTCGTCGTGAAACCTATCTGTGCTATGTTGTGAAACGTCGTGATAGCGCAACCAGCTTTAGCCTGGATTTTGGTTATCTGCGCAATAAAAATGGTTGTCATGTGGAGCTGCTGTTTCTGCGTTATATTAGCGATTGGGATCTGGATCCGGGTCGTTGTTATCGTGTTACCTGGTTTACCAGCTGGTCACCGTGTTATGACTGTGCACGTCATGTTGCAGATTTTCTGCGTGGTAATCCGAATCTGAGCCTGCGTATTTTTACCGCACGTCTGTATTTTTGCGAAGATCGTAAAGCAGAACCGGAAGGTCTGCGTCGTCTGCATCGTGCAGGTGTTCAGATTGCCATTATGACCTTTAAAGATTATTTTTATTGCTGGAATACCTTTGTGGAAAATCATGAACGCACCTTTAAAGCATGGGAAGGCCTGCATGAAAATAGCGTTCGTCTGTCTCGCCAGCTGCGTCGTATTCTGCTGCCGCTGTATGAAGTTGATGATCTGCGTGATGCCTTTCGTACCCTGGGTTTAACCGGTCAGGAAGTTGCATGCACCGCCGGCAACACGATTAACATCGCTAAGAACGACTTCTCTGACATCGAACTGGCTGCTATCCCGTTCAACACTCTGGCTGACCATTACGGTGAGCGTTTAGCTCGCGAACAGTTGGCCCTTGAGCATGAGTCTTACGAGATGGGTGAAGCACGCTTCCGCAAGATGTTTGAGCGTCAACTTAAAGCTGGTGAGGTTGCGGATAACGCTGCCGCCAAGCCTCTCATCACTACCCTACTCCCTAAGATGATTGCACGCATCAACGACTGGTTTGAGGAAGTGAAAGCTAAGCGCGGCAAGCGCCCGACAGCCTTCCAGTTCCTGCAAGAAATCAAGCCGGAAGCCGTAGCGTACATCACCATTAAGACCACTCTGGCTTGCCTAACCAGTGCTGACAATACAACCGTTCAGGCTGTAGCAAGCGCAATCGGTCGGGCCATTGAGGACGAGGCTCGCTTCGGTCGTATCCGTGACCTTGAAGCTAAGCACTTCAAGAAAAACGTTGAGGAACAACTCAACAAGCGCGTAGGGCACGTCTACAAGAAAGCATTTATGCAAGTTGTCGAGGCTGACATGCTCTCTAAGGGTCTACTCGGTGGCGAGGCGTGGTCTTCGTGGCATAAGGAAGACTCTATTCATGTAGGAGTACGCTGCATCGAGATGCTCATTGAGTCAACCGGAATGGTTAGCTTACACCGCCAAAATGCTGGCGTAGTAGGTCAAGACTCTGAGACTATCGAACTCGCACCTGAATACGCTGAGGCTATCGCAACCCGTGCAGGTGCGCTGGCTGGCATCTCTCCGATGTTCCAACCTTGCGTAGTTCCTCCTAAGCCGTGGACTGGCATTACTGGTGGTGGCTATTGGGCTAACGGTCGTCGTCCTCTGGCGCTGGTGCGTACTCACAGTAAGAAAGCACTGATGCGCTACGAAGACGTTTACATGCCTGAGGTGTACAAAGCGATTAACATTGCGCAAAACACCGCATGGAAAATCAACAAGAAAGTCCTAGCGGTCGCCAACGTAATCACCAAGTGGAAGCATTGTCCGGTCGAGGACATCCCTGCGATTGAGCGTGAAGAACTCCCGATGAAACCGGAAGACATCGACATGAATCCTGAGGCTCTCACCGCGTGGAAACGTGCTGCCGCTGCTGTGTACCGCAAGGACAGGGCTCGCAAGTCTCGCCGTATCAGCCTTGAGTTCATGCTTGAGCAAGCCAATAAGTTTGCTAACCATAAGGCCATCTGGTTCCCTTACAACATGGACTGGCGCGGTCGTGTTTACGCCGTGTCAATGTTCAACCCGCAAGGTAACGATATGACCAAAGGACTGCTTACGCTGGCGAAAGGTAAACCAATCGGTAAGGAAGGTTACTACTGGCTGAAAATCCACGGTGCAAACTGTGCGGGTGTCGATAAGGTTCCGTTCCCTGAGCGCATCAAGTTCATTGAGGAAAACCACGAGAACATCATGGCTTGCGCTAAGTCTCCACTGGAGAACACTTGGTGGGCTGAGCAAGATTCTCCGTTCTGCTTCCTTGCGTTCTGCTTTGAGTACGCTGGGGTACAGCACCACGGCCTGAGCTATA

		ACTGCTCCCTTCCGCTGGCGTTTGACGGGTCTTGCTCTGGCATCCAGCACTTCTCCGCGATGCTCCGAGATGAGGTAGGTGGTCGCGCGGTTAACTTGCTTCCTAGTGAAACCGTTCAGGACATCTACGGGATTGTTGCTAAGAAAGTCAACGAGATTCTACAAGCAGACGCAATCAATGGGACCGATAACGAAGTAGTTACCGTGACCGATGAGAACACTGGTGAAATCTCTGAGAAAGTCAAGCTGGGCACTAAGGCACTGGCTGGTCAATGGCTGGCTTACGGTGTTACTCGCAGTGTGACTAAGCGTTCAGTCATGACGCTGGCTTACGGGTCCAAAGAGTTCGGCTTCCGTCAACAAGTGCTGGAAGATACCATTCAGCCAGCTATTGATTCCGGCAAGGGTCTGATGTTCACTCAGCCGAATCAGGCTGCTGGATACATGGCTAAGCTGATTTGGGAATCTGTGAGCGTGACGGTGGTAGCTGCGGTTGAAGCAATGAACTGGCTTAAGTCTGCTGCTAAGCTGCTGGCTGCTGAGGTCAAAGATAAGAAGACTGGAGAGATTCTTCGCAAGCGTTGCGCTGTGCATTGGGTAACTCCTGATGGTTTCCCTGTGTGGCAGGAATACAAGAAGCCTATTCAGACGCGCTTGAACCTGATGTTCCTCGGTCAGTTCCGCTTACAGCCTACCATTAACACCAACAAAGATAGCGAGATTGATGCACACAAACAGGAGTCTGGTATCGCTCCTAACTTTGTACACAGCCAAGACGGTAGCCACCTTCGTAAGACTGTAGTGTGGGCACACGAGAAGTACGGAATCGAATCTTTTGCACTGATTCACGACTCCTTCGGTACCATTCCGGCTGACGCTGCGAACCTGTTCAAAGCAGTGCGCGAAACTATGGTTGACACATATGAGTCTTGTGATGTACTGGCTGATTTCTACGACCAGTTCGCTGACCAGTTGCACGAGTCTCAATTGGACAAAATGCCAGCACTTCCGGCTAAAGGTAACTTGAACCTCCGTGACATCTTAGAGTCGGACTTCGCGTTCGCGTAATAAggctcgggagacctatcg
UGI Expression Cassette with J23115	Stops repair of U:G mismatch by blocking UNG	TCTGGTGGGTCTCTGTCCTttatagctagctcagcccttggtacaatgctagcGcattatttgacctccaatgcgaacaaatcacacaggaaagccctttATGACGAATTTATCGGACATTATCGAAAAGGAGACGGGAAAACAATTAGTCATTCAGGAGTCAATCCTTATGTTGCCCGAGGAAGTCGAGGAAGTTATCGGGAACAAACCGGAGAGTGACATTCTGGTACACACTGCATATGACGAATCAACTGACGAGAATGTAATGTTGTTAACTTCGGACGCGCCGGAGTACAAACCATGGGCCTTGGTGATTCAAGACTCAAACGGGGAAAATAAAATTAAAATGCTGTAAtgtacacgagccattatttctttcctaaggttgaaaataaaaacggcgctaaaaagcgccgtttttttgacggtggtaGGCTCGGGAGACCTATCG
Tet Repressor Expression Cassette with J23116	Blocks transcription via promoters with tet operators unless anhydrotetracycline is present in the medium	TCTGGTGGGTCTCTGTCCttgacagctagctcagtcctagggactatgctagctgagcgctcacaattCTATGGACTATGTTTTCACACAGGAAAGGCCTCGATGtccagattagataaaagtaaagtgattaacagcgcattagagctgcttaatgaggtcggaatcgaaggtttaacaacccgtaaactcgcccagaagctaggtgtagagcagcctacattgtattggcatgtaaaaaataagcgggctttgctcgacgccttagccattgagatgttagataggcaccatactcacttttgccctttagaaggggaaagctggcaagattttttacgtaataacgctaaaagttttagatgtgctttactaagtcatcgcgatggagcaaaagtacatttaggtacacggcctacagaaaaacagtatgaaactctcgaaaatcaattagcctttttatgccaacaaggtttttcactagagaatgcattatatgcactcagcgctgtggggcattttactttaggttgcgtattggaagatcaagagcatcaagtcgctaaagaagaaagggaaacacctactactgatagtatgccgccattattacgacaagctatcgaattatttgatcaccaaggtgcagagccagccttcttattcggccttgaattgatcatatgcggattagaaaaacaacttaaatgtgaaagtgggtcctaaTAACTAGGGCCCATACCCtccggcaattaaaaaagcggctaaccacgccgctttttttacgtctgcaGGCTCGGGAGACCTATCG
β-lactamase WT	Confers resistance to ampicillin/carbenicillin; Internal BsaI site removed	TCTGGTGGGTCTCTGTCCATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGATCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGGCTCGGGAGACCTATCG
ATC β-lactamase with B15 terminator	Start codon changed to ATC	tctggtgggtctctgtccATCAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGATCTCGCGGTAT

		CATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGGCTCGttacttacgacactccgagacagtcagagggtatttattgaactaGTCCCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGGTGGGCCTTTCTGCGTTTATAGGCTCGggagacctatcg
ACG β-lactamase with B15 terminator	Start codon changed to ACG	tctggtgggtctctgtccACGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGATCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGGCTCGttacttacgacactccgagacagtcagagggtatttattgaactaGTCCCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGGTGGGCCTTTCTGCGTTTATAGGCTCGggagacctatcg
CTG β-lactamase with B15 terminator	Start codon changed to CTG	tctggtgggtctctgtccCTGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGATCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGGCTCGttacttacgacactccgagacagtcagagggtatttattgaactaGTCCCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGGTGGGCCTTTCTGCGTTTATAGGCTCGggagacctatcg
β-lactamase with premature TAA stop and B15 terminator	TAA codon highlighted in green	TCTGGTGGGTCTCTGTCCATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATAAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGATCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAAGGCTCGttacttacgacactccgagacagtcagagggtatttattgaactaGTCCCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGGTGGGCCTTTCTGCGTTTATAGGCTCGggagacctatcg

Exonuclease-III	No Stop codon for making fusion protein	TCTGGTGGGTCTCTGTCCATGAAATTTGTCTCTTTTAATATCAACGGCCTGCGCGCCAGACCTCACCAGCTTGAAGCCATCGTCGAAAAGCACCAACCGGATGTGATTGGCCTGCAGGAGACAAAAGTTCATGACGATATGTTTCCGCTCGAAGAGGTGGCGAAGCTCGGCTACAACGTGTTTTATCACGGGCAGAAAGGCCATTATGGCGTGGCGCTGCTGACCAAAGAGACGCCGATTGCCGTGCGTCGCGGCTTTCCCGGTGACGACGAAGAGGCGCAGCGGCGGATTATTATGGCGGAAATCCCCTCACTGCTGGGTAATGTCACCGTGATCAACGGTTACTTCCCGCAGGGTGAAAGCCGCGACCATCCGATAAAATTCCCGGCAAAAGCGCAGTTTTATCAGAATCTGCAAAACTACCTGGAAACCGAACTCAAACGTGATAATCCGGTACTGATTATGGGCGATATGAATATCAGCCCTACAGATCTGGATATCGGCATTGGCGAAGAAAACCGTAAGCGCTGGCTGCGTACCGGTAAATGCTCTTTCCTGCCGGAAGAGCGCGAATGGATGGACAGGCTGATGAGCTGGGGGTTGGTCGATACCTTCCGCCATGCGAATCCGCAAACAGCAGATCGTTTCTCATGGTTTGATTACCGCTCAAAAGGTTTTGACGATAACCGTGGTCTGCGCATCGACCTGCTGCTCGCCAGCCAACCGCTGGCAGAATGTTGCGTAGAAACCGGCATCGACTATGAAATCCGCAGCATGGAAAAACCGTCCGATCACGCCCCCGTCTGGGCGACCTTCCGCCGCGGCTCGGGAGACCTATCG
N. Meninjitidis AP Endonuclease	Codon optimised for expression in K-12 bacterial strains No Stop codon for making fusion protein	TCTGGTGGGTCTCTGTCCATGTTAAAGATCATCTCGGCTAATGTCAACGGAATTCGCTCAGCCTACAAAAAAAGGCTTCTACGAGTACATTGCGGCTTCTGGAGCCGACATTGTGTGTGTTCAGGAATTAAAAGCACAAGAGGCCGATTTATCTGCTGACATGAAGAACCCACACGGTATGCATGGACATTGGCACTGCGCCGAAAAGCGTGGCTACTCTGGAGTGGCTGTCTATAGTAAACGCAAACCTGATAATGTGCAAATTGGGATGGGCATTGAGGAGTTTGACCGTGAGGGTCGTTTTGTTCGTTGCGATTTTGGCCGCTTATCTGTTATTTCGCTGTATCTGCCGTCCGGGAGTAGCGCGGAAGAACGTCAACAGGTAAAGTATCGTTTCCTTGACGCTTTTTACCCCATGTTAGAAGCAATGAAAAATGAGGGCCGCGACATCGTCGTGTGTGGGGACTGGAACATTGCACACCAGAACATCGATCTTAAGAACTGGAAAGGCAATCAGAAGAATTCAGGTTTTCTTCCAGAGGAGCGTGAATGGATCGGGAAGGTGATTCACAAATTAGGTTGGACGGACATGTGGCGTACATTATACCCGGATGTTCCTGGCTATACGTGGTGGAGCAATCGTGGCCAGGCTTATGCAAAGGACGTTGGCTGGCGCATTGATTATCAGATGGTTACGCCTGAGTTAGCTGCAAAGGCCGTTTCTGCCCACGTGTATAAGGACGAGAAATTTTCTGATCATGCACCGCTGGTCGTCGAATATGATTACGCCGCAGAGGGCTCGGGAGACCTATCG
RecJ (from MG1655)	5'-3' Exonuclease No Stop codon	TCTGGTGGGTCTCTGTCCATGAAACAACAGATACAACTTCGTCGCCGTGAAGTCGATGAAACGGCAGACTTGCCCGCTGAATTGCCTCCCTTGCTGCGCCGTTTATACGCCAGCCGGGGAGTACGCAGTGCGCAAGAACTGGAACGCAGTGTTAAAGGTATGCTGCCCTGGCAGCAACTGAGCGGCGTCGAAAAGGCCGTTGAGATCCTTTACAACGCTTTTCGCGAAGGAACGCGGATTATTGTGGTCGGTGATTTCGACGCCGACGGCGCGACCAGCACGGCTCTAAGCGTGCTGGCGATGCGCTCGCTTGGTTGCAGCAATATCGACTACCTGGTACCAAACCGTTTCGAAGACGGTTACGGCTTAAGCCCGGAAGTGGTCGATCAGGCCCATGCCCGTGGCGCGCAGTTAATTGTCACGGTGGATAACGGTATTTCCTCCCATGCGGGGGTTGAGCACGCTCGCTCGTTGGGCATCCCGGTTATTGTTACCGATCACCATTTGCCAGGCGACACATTACCCGCAGCGGAAGCGATCATTAACCCTAACTTGCGCGACTGTAATTTCCCGTCGAAATCACTGGCAGGCGTGGGTGTGGCGTTTTATCTGATGCTGGCGCTGCGCACCTTTTTGCGCGATCAGGGCTGGTTTGATGAGCGTAACATCGCAATTCCTAACCTGGCAGAACTGCTGGATCTGGTCGCGCTGGGGACAGTGGCGGACGTCGTGCCGCTGGACGCTAATAATCGCATTCTGACCTGGCAGGGGATGAGTCGCATCCGAGCCGGAAAGTGCCGTCCGGGGATTAAAGCGCTGCTTGAAGTGGCAAACCGTGATGCACAAAAACTCGCCGCCAGCGATTTAGGTTTTGCGCTGGGGCCACGTCTCAATGCTGCCGGACGACTGGACGATATGTCCGTCGGTGTGGCGCTGTTGTTGTGCGACAACATCGGCGAAGCGCGCGTGCTGGCAAATGAACTCGATGCGCTAAACCAGACGCGAAAAGAGATCGAACAAGGAATGCAAATTGAAGCCCTGACCCTGTGCGAGAAACTGGAGCGCAGCCGTGACACGCTACCCGGCGGGCTGGCAATGTATCACCCCGAATGGCATCAGGGCGTTGTCGGTATTCTGGCTTCGCGCATCAAAGAGCGTTTTCACCGTCCGGTTATCGCGTTTGCGCCAGCAGGTGACGGTACGCTGAAAGGTTCCGGTCGCTCCATTCAGGGGCTGCATATGCGTGATGCGCTGGAGCGATTAGACACACTCTACCCTGGCATGATGCTGAAGTTTGGCGGTCATGCGATGGCGGCGGGTTTGTCGCTGGAAGAGGATAAATTCAAACTCTTTCAACAACGGTTTGGCGAACTGGTTACTGAGTGGCTGGACCCTTCGCTATTGCAAGGCGAAGTGGTATCAGACGGTCCGTTAAGCCCGGCCGAAATGACCATGGAAGTGGCGCAGCTGCTGCGCGATGCTGGCCCGTGGGGGCAGATGTTCCCGGAGCCGCTGTTTGACGGTCATTTCCGTCTGCTGCAACAGCGGCTGGTGGGCGAACGTCATTTGAAGGTGATGGTCGAACCGGTCGGCGGCGGTCCACTGCTGGATGGTATTGCTTTTAATGTCGATACCGCCCTCTGGCCGGATAACGGCGTGCGCGAAGTGCAACTGGCTTATAAGCTCGATATCAACGAGTTTCGCGGCAACCGCAGCCTGCAAATTATCATCGACAATATCTGGCCAATTggctcgggagacctatcg

RecE (from MG1655)	5'-3' Exonuclease No Stop codon	TCTGGTGGGTCTCTGTCCatgagcacaaaaccactcttcctgttacggaaagcgaaaaaatcatccggtgaacctgacgtcgtcctgtgggcaagcaacgattttgaatcgacctgtgccactctggactacctgatcgttaagtcaggtaaaaaactgagcagctattttaaagctgttgccacgaattttcctgtcgttaatgacctgcccgctgaaggtgagatcgattttacctggagtgaacgctatcaactcagcaaagactccatgacatgggaactaaaaccgggagcagcaccagacaacgctcactatcaaggcaataccaacgtcaacggcgaagacatgactgagattgaggagaatatgctactcccaatttctggccaggaactgcccattcgttggcttgctcaacacggcagcgaaaaaccggtaacgcacgtttcacgcgacggactccaggcattacacattgctcgggctgaagaactaccggctgttactgccctggctgtttcccacaaaaccagcctgctcgacccgctggaaattcgcgaactccacaaactggttcgtgacactgacaaagttttccctaatcctggtaattcaaacctgggactgataactgctttttcgaagcatacctgaacgctgactacaccgatcgaggactgctgacaaaagagtggatgaagggtaatcgtgtttcacacatcactcgcacggcttccggtgctaatgctggcggcggaaacctcaccgatcgcggcgaaggtttcgtacacgatctgacgtcactggcgcgcgacgtagccactggcgtactggcccgttcaatggatctggacatctataaccttcatccggcacacgctaaacgcattgaggaaattatcgctgaaaataaaccgcccttttctgttttccgcgacaaattcatcaccatgcctggcgggctggattattcccgcgccatcgtggttgcgtccgtaaaagaagcaccaattgggatcgaggtcatccccgcgcacgtcactgaatatctgaacaaagtactgactgaaaccgatcatgccaaccctgatccggaaatcgtggatattgcctgcggtcgctcctctgccccgatgccgcagcgagtaacagaagaaggaaaacaggatgatgaagaaaaaccgcaaccatctggaacaacggcagttgaacagggagaggctgaaacaatggaaccggacgcaactgaacatcatcaggacacgcagccgctggatgctcagtcacaggtaaattctgttgatgcgaaatatcaggaactgcgggcagaactccatgaagcccggaaaaacattccatcaaaaaatcctgtcgatgacgataaattgcttgctgcatcacgtggtgaatttgttgacggaattagcgacccgaacgatccgaaatgggtaaaggggatccagactcgcgattgtgtgtaccagaaccagccagaaacggaaaaaaccagcccagatatgaatcaacctgagccagtagtgcaacaggaaccggaaatagcctgcaatgcctgcggccagactggcggggataactgccctgactgtggtgcggtgatgggcgacgcaacataccaggaaacattcgatgaagagagtcaggttgaagctaaggaaaatgatccggaggaaatggaaggcgctgaacatccgcacaatgagaatgctggcagcgatccgcatcgcgattgcagtgatgaaactggcgaagtcgcagatcccgtaatcgtagaagacatagagccaggtatttattacggaatttcgaatgagaattaccacgcgggtcccggtatcagtaagtctcagctcgatgacattgctgatactccggcactatatttgtggcgtaaaaatgcccccgtggacaccacaaagacaaaaacgctcgatttaggaactgctttccactgccgggtacttgaaccggaagaattcGGCTCGGGAGACCTATCG
5'-3' Exonuclease Domain of DNA Polymerase I	5'-3' Exonuclease No Stop codon	TCTGGTGGGTCTCTGTCCATGGTTCAGATCCCCCAAAATCCACTTATCCTTGTAGATGGTTCATCTTATCTTTATCGCGCATATCACGCGTTTCCCCCGCTGACTAACAGCGCAGGCGAGCCGACCGGTGCGATGTATGGTGTCCTCAACATGCTGCGCAGTCTGATCATGCAATATAAACCGACGCATGCAGCGGTGGTCTTTGACGCCAAGGGAAAAACCTTTCGTGATGAACTGTTTGAACATTACAAATCACATCGCCCGCCAATGCCGGACGATCTGCGTGCACAAATCGAACCCTTGCACGCGATGGTTAAAGCGATGGGACTGCCGCTGCTGGCGGTTTCTGGCGTAGAAGCGGACGACGTTATCGGTACTCTGGCGCGCGAAGCCGAAAAAGCCGGGCGTCCGGTGCTGATCAGCACTGGCGATAAAGATATGGCGCAGCTGGTGACGCCAAATATTACGCTTATCAATACCATGACGAATACCATCCTCGGACCGGAAGAGGTGGTGAATAAGTACGGCGTGCCGCCAGAACTGATCATCGATTTCCTGGCGCTGATGGGTGACTCCTCTGATAACATTCCTGGCGTACCGGGCGTCGGTGAAAAAACCGCGCAGGCATTGCTGCAAGGTCTTGGCGGACTGGATACGCTGTATGCCGAGCCAGAAAAAATTGCTGGGTTGAGCTTCCGTGGCGCGAAAACAATGGCAGCGAAGCTCGAGCAAAACAAAGAAGTTGCTTATCTCTCATACCAGCTGGCGACGATTAAAACCGACGTTGAACTGGAGCTGACCTGTGAACAACTGGAAGTGCAGCAACCGGCAGCGGAAGAGTTGTTGGGGCTGTTCAAAAAGTATGAGTTCAAACGCTGGACTGCTGATGTCGAAGCGGGCAAATGGTTACAGGCCAAAGGGGCAAAACCAGCCGCGAAGCCACAGGAAACCAGTGTTGCAGACGAAGCACCAGAAGTGACGGCAACGGGCTCGGGAGACCTATCG
Shorter version: 5'-3' Exonuclease Domain of DNA Polymerase I	5'-3' Exonuclease No Stop codon	TCTGGTGGGTCTCTGTCCATGGTTCAGATCCCCCAAAATCCACTTATCCTTGTAGATGGTTCATCTTATCTTTATCGCGCATATCACGCGTTTCCCCCGCTGACTAACAGCGCAGGCGAGCCGACCGGTGCGATGTATGGTGTCCTCAACATGCTGCGCAGTCTGATCATGCAATATAAACCGACGCATGCAGCGGTGGTCTTTGACGCCAAGGGAAAAACCTTTCGTGATGAACTGTTTGAACATTACAAATCACATCGCCCGCCAATGCCGGACGATCTGCGTGCACAAATCGAACCCTTGCACGCGATGGTTAAAGCGATGGGACTGCCGCTGCTGGCGGTTTCTGGCGTAGAAGCGGACGACGTTATCGGTACTCTGGCGCGCGAAGCCGAAAAAGCCGGGCGTCCGGTGCTGATCAGCACTGGCGATAAAGATATGGCGCAGCTGGTGACGCCAAATATTACGCTTATCAATACCATGACGAATACCATCCTCGGACCGGAAGAGGTGGTGAATAAGTACGGCGTGCCGCCAGAACTGATCATCGATTTCCTGGCGCTGATGGGTGACTCCTCTGATAACATTCCTGGCGTACCGGGCGTCGGTGAAAAAACCGCGCAGGCATTGCTGCAAGGTCTTGGCGGACTGGATACGCTGTATGCCGAGCCAGAAAAAATTGCTGGGTTGAGCTTCCGTGGCGCGAAAACAATGGCAGCGAAGCTCGAGCAAAACAAAGAAGTTGCTTATCTCTCATACCAGCTGGCGACGATTAAAGGCTCGGGAGACCTATCG
DNA Polymerase IV wildtype	Error-prone Polymerase	TCTGGTGGGTCTCTGTCCATGCGTAAAATCATTCATGTGGATATGGACTGCTTTTTCGCCGCAGTGGAGATGCGCGACAATCCCGCCCTGCGCGATATCCCTATTGCTATTGGCGGCAGCCGCGAACGTCGGGGGGTGATCAGCACCGCCAATTATCCCGCGCGTAAATTTGGCGTACGTAGCGCTATGCCGACAGGGATGGCGCTCAAATTATGCCCACATCTCACCTTGCTTCCGGGGCGCTTTGACGCCTACAAAGAAGCCTCAAATCATATCCGTGAAATCTTCTCGCGCTACACCTCGCGCATTGAACCGTTGTCACTGGATGAGGCTTATCTCGATGTCACCGATAGCGTCCATTGCCACGGTTCTGCGACCCTCATCGCCCAGGAAATCCGCCAGACAATCTTCAACGAGCTGCAACTGACGGCGT

		CTGCGGGCGTGGCACCAGTAAAGTTTCTCGCCAAAATCGCCTCCGACATGAATAAACCCAACGGCCAGTTTGTGATTACGCCGGCAGAAGTTCCGGCATTTTTACAAACCTTACCGCTGGCAAAAATCCCCGGCGTCGGCAAAGTCTCAGCGGCAAAACTGGAAGCGATGGGGCTGCGGACCTGCGGTGATGTACAAAAGTGTGATCTGGTGATGCTGCTTAAACGCTTTGGCAAATTTGGCCGCATTTTGTGGGAGCGTAGTCAGGGGATTGACGAACGCGATGTTAACAGCGAACGGTTGCGAAAATCCGTCGGCGTGGAACGCACGATGGCGGAAGATATTCATCACTGGTCTGAATGTGAAGCGATTATCGAGCGGCTGTATCCGGAACTTGAACGCCGTCTGGCAAAGGTAAAACCTGATTTACTGATTGCTCGCCAGGGGGTGAAATTAAAGTTCGACGATTTTCAGCAAACCACCCAGGAGCACGTCTGGCCGCGGCTGAATAAAGCTGATCTAATCGCCACCGCGCGTAAAACCTGGGATGAACGCCGCGGCGGGCGCGGTGTGCGTCTGGTGGGGCTGCATGTGACGTTGCTTGACCCGCAAATGGAAAGACAACTGGTGCTGGGATTATAAGGCTCGGGAGACCTATCG
DNA Polymerase IV Δ5	Error-prone Polymerase	TCTGGTGGGTCTCTGTCCATGCGTAAAATCATTCATGTGGATATGGACTGCTTTTTCGCCGCAGTGGAGATGCGCGACAATCCCGCCCTGCGCGATATCCCTATTGCTATTGGCGGCAGCCGCGAACGTCGGGGGGTGATCAGCACCGCCAATTATCCCGCGCGTAAATTTGGCGTACGTAGCGCTATGCCGACAGGGATGGCGCTCAAATTATGCCCACATCTCACCTTGCTTCCGGGGCGCTTTGACGCCTACAAAGAAGCCTCAAATCATATCCGTGAAATCTTCTCGCGCTACACCTCGCGCATTGAACCGTTGTCACTGGATGAGGCTTATCTCGATGTCACCGATAGCGTCCATTGCCACGGTTCTGCGACCCTCATCGCCCAGGAAATCCGCCAGACAATCTTCAACGAGCTGCAACTGACGGCGTCTGCGGGCGTGGCACCAGTAAAGTTTCTCGCCAAAATCGCCTCCGACATGAATAAACCCAACGGCCAGTTTGTGATTACGCCGGCAGAAGTTCCGGCATTTTTACAAACCTTACCGCTGGCAAAAATCCCCGGCGTCGGCAAAGTCTCAGCGGCAAAACTGGAAGCGATGGGGCTGCGGACCTGCGGTGATGTACAAAAGTGTGATCTGGTGATGCTGCTTAAACGCTTTGGCAAATTTGGCCGCATTTTGTGGGAGCGTAGTCAGGGGATTGACGAACGCGATGTTAACAGCGAACGGTTGCGAAAATCCGTCGGCGTGGAACGCACGATGGCGGAAGATATTCATCACTGGTCTGAATGTGAAGCGATTATCGAGCGGCTGTATCCGGAACTTGAACGCCGTCTGGCAAAGGTAAAACCTGATTTACTGATTGCTCGCCAGGGGGTGAAATTAAAGTTCGACGATTTTCAGCAAACCACCCAGGAGCACGTCTGGCCGCGGCTGAATAAAGCTGATCTAATCGCCACCGCGCGTAAAACCTGGGATGAACGCCGCGGCGGGCGCGGTGTGCGTCTGGTGGGGCTGCATGTGACGTTGCTTGACCCGCAAATGGAAAGATAAGGCTCGGGAGACCTATCG
DNA Polymerase IV Δ12	Error-prone Polymerase	TCTGGTGGGTCTCTGTCCATGCGTAAAATCATTCATGTGGATATGGACTGCTTTTTCGCCGCAGTGGAGATGCGCGACAATCCCGCCCTGCGCGATATCCCTATTGCTATTGGCGGCAGCCGCGAACGTCGGGGGGTGATCAGCACCGCCAATTATCCCGCGCGTAAATTTGGCGTACGTAGCGCTATGCCGACAGGGATGGCGCTCAAATTATGCCCACATCTCACCTTGCTTCCGGGGCGCTTTGACGCCTACAAAGAAGCCTCAAATCATATCCGTGAAATCTTCTCGCGCTACACCTCGCGCATTGAACCGTTGTCACTGGATGAGGCTTATCTCGATGTCACCGATAGCGTCCATTGCCACGGTTCTGCGACCCTCATCGCCCAGGAAATCCGCCAGACAATCTTCAACGAGCTGCAACTGACGGCGTCTGCGGGCGTGGCACCAGTAAAGTTTCTCGCCAAAATCGCCTCCGACATGAATAAACCCAACGGCCAGTTTGTGATTACGCCGGCAGAAGTTCCGGCATTTTTACAAACCTTACCGCTGGCAAAAATCCCCGGCGTCGGCAAAGTCTCAGCGGCAAAACTGGAAGCGATGGGGCTGCGGACCTGCGGTGATGTACAAAAGTGTGATCTGGTGATGCTGCTTAAACGCTTTGGCAAATTTGGCCGCATTTTGTGGGAGCGTAGTCAGGGGATTGACGAACGCGATGTTAACAGCGAACGGTTGCGAAAATCCGTCGGCGTGGAACGCACGATGGCGGAAGATATTCATCACTGGTCTGAATGTGAAGCGATTATCGAGCGGCTGTATCCGGAACTTGAACGCCGTCTGGCAAAGGTAAAACCTGATTTACTGATTGCTCGCCAGGGGGTGAAATTAAAGTTCGACGATTTTCAGCAAACCACCCAGGAGCACGTCTGGCCGCGGCTGAATAAAGCTGATCTAATCGCCACCGCGCGTAAAACCTGGGATGAACGCCGCGGCGGGCGCGGTGTGCGTCTGGTGGGGCTGCATGTGACGTTGCTTTAAGGCTCGGGAGACCTATCG
Wildtype Polymerase Domain of DNA Polymerase I		TCTGGTGGGTCTCTGTCCATGGGGCCGTTGAACGTCTTCGAGAATATCGAAATGCCGCTGGTGCCGGTGCTTTCACGCATTGAACGTAACGGTGTGAAGATCGATCCGAAAGTGCTGCACAATCATTCTGAAGAGCTCACCCTTCGTCTGGCTGAGCTGGAAAAGAAAGCGCATGAAATTGCAGGTGAGGAATTTAACCTTTCTTCCACCAAGCAGTTACAAACCATTCTCTTTGAAAAACAGGGCATTAAACCGCTGAAGAAAACGCCGGGTGGCGCGCCGTCAACGTCGGAAGAGGTACTGGAAGAACTGGCGCTGGACTATCCGTTGCCAAAAGTGATTCTGGAGTATCGTGGTCTGGCGAAGCTGAAATCGACCTACACCGACAAGCTGCCGCTGATGATCAACCCGAAAACCGGGCGTGTGCATACCTCTTATCACCAGGCAGTAACTGCAACGGGACGTTTATCGTCAACCGATCCTAACCTGCAAAACATTCCGGTGCGTAACGAAGAAGGTCGTCGTATCCGCCAGGCGTTTATTGCGCCAGAGGATTATGTGATTGTCTCAGCGGACTACTCGCAGATTGAACTGCGCATTATGGCGCATCTTTCGCGTGACAAAGGCTTGCTGACCGCATTCGCGGAAGGAAAAGATATCCACCGGGCAACGGCGGCAGAAGTGTTTGGTTTGCCACTGGAAACCGTCACCAGCGAGCAACGCCGTAGCGCGAAAGCGATCAACTTTGGTCTGATTTATGGCATGAGTGCTTTCGGTCTGGCGCGGCAATTGAACATTCCACGTAAAGAAGCGCAGAAGTACATGGACCTTTACTTCGAACGCTACCCTGGCGTGCTGGAGTATATGGAACGCACCCGTGCTCAGGCGAAAGAGCAGGGCTACGTTGAAACGCTGGACGGACGCCGTCTGTATCTGCCGGATATCAAATCCAGCAATGGTGCTCGTCGTGCAGCGGCTG

		AACGTGCAGCCATTAACGCGCCAATGCAGGGAACCGCCGCCGACATTATCAAACGGGCGATGATTGCCGTTGATGCGTGGTTACAGGCTGAGCAACCGCGTGTACGTATGATCATGCAGGTACACGATGAACTGGTATTTGAAGTTCATAAAGATGATGTTGATGCCGTCGCGAAGCAGATTCATCAACTGATGGAAAACTGTACCCGTCTGGATGTGCCGTTGCTGGTGGAAGTGGGGAGTGGCGAAAACTGGGATCAGGCGCACTAAGGCTCGGGAGACCTATCG
DNA Polymerase I	From MG1655	tctggtgggtctctgtccATGGTTCAGATCCCCCAAAATCCACTTATCCTTGTAGATGGTTCATCTTATCTTTATCGCGCATATCACGCGTTTCCCCCGCTGACTAACAGCGCAGGCGAGCCGACCGGTGCGATGTATGGTGTCCTCAACATGCTGCGCAGTCTGATCATGCAATATAAACCGACGCATGCAGCGGTGGTCTTTGACGCCAAGGGAAAAACCTTTCGTGATGAACTGTTTGAACATTACAAATCACATCGCCCGCCAATGCCGGACGATCTGCGTGCACAAATCGAACCCTTGCACGCGATGGTTAAAGCGATGGGACTGCCGCTGCTGGCGGTTTCTGGCGTAGAAGCGGACGACGTTATCGGTACTCTGGCGCGCGAAGCCGAAAAAGCCGGGCGTCCGGTGCTGATCAGCACTGGCGATAAAGATATGGCGCAGCTGGTGACGCCAAATATTACGCTTATCAATACCATGACGAATACCATCCTCGGACCGGAAGAGGTGGTGAATAAGTACGGCGTGCCGCCAGAACTGATCATCGATTTCCTGGCGCTGATGGGTGACTCCTCTGATAACATTCCTGGCGTACCGGGCGTCGGTGAAAAAACCGCGCAGGCATTGCTGCAAGGTCTTGGCGGACTGGATACGCTGTATGCCGAGCCAGAAAAAATTGCTGGGTTGAGCTTCCGTGGCGCGAAAACAATGGCAGCGAAGCTCGAGCAAAACAAAGAAGTTGCTTATCTCTCATACCAGCTGGCGACGATTAAAACCGACGTTGAACTGGAGCTGACCTGTGAACAACTGGAAGTGCAGCAACCGGCAGCGGAAGAGTTGTTGGGGCTGTTCAAAAAGTATGAGTTCAAACGCTGGACTGCTGATGTCGAAGCGGGCAAATGGTTACAGGCCAAAGGGGCAAAACCAGCCGCGAAGCCACAGGAAACCAGTGTTGCAGACGAAGCACCAGAAGTGACGGCAACGGTGATTTCTTATGACAACTACGTCACCATCCTTGATGAAGAAACACTGAAAGCGTGGATTGCGAAGCTGGAAAAAGCGCCGGTATTTGCATTTGATACCGAAACCGACAGCCTTGATAACATCTCTGCTAACCTGGTCGGGCTTTCTTTTGCTATCGAGCCAGGCGTAGCGGCATATATTCCGGTTGCTCATGATTATCTTGATGCGCCCGATCAAATCTCTCGCGAGCGTGCACTCGAGTTGCTAAAACCGCTGCTGGAAGATGAAAAGGCGCTGAAGGTCGGGCAAAACCTGAAATACGATCGCGGTATTCTGGCGAACTACGGCATTGAACTGCGTGGGATTGCGTTTGATACCATGCTGGAGTCCTACATTCTCAATAGCGTTGCCGGGCGTCACGATATGGACAGCCTCGCGGAACGTTGGTTGAAGCACAAAACCATCACTTTTGAAGAGATTGCTGGTAAAGGCAAAAATCAACTGACCTTTAACCAGATTGCCCTCGAAGAAGCCGGACGTTACGCCGCCGAAGATGCAGATGTCACCTTGCAGTTGCATCTGAAAATGTGGCCGGATCTGCAAAAACACAAAGGGCCGTTGAACGTCTTCGAGAATATCGAAATGCCGCTGGTGCCGGTGCTTTCACGCATTGAACGTAACGGTGTGAAGATCGATCCGAAAGTGCTGCACAATCATTCTGAAGAGCTCACCCTTCGTCTGGCTGAGCTGGAAAAGAAAGCGCATGAAATTGCAGGTGAGGAATTTAACCTTTCTTCCACCAAGCAGTTACAAACCATTCTCTTTGAAAAACAGGGCATTAAACCGCTGAAGAAAACGCCGGGTGGCGCGCCGTCAACGTCGGAAGAGGTACTGGAAGAACTGGCGCTGGACTATCCGTTGCCAAAAGTGATTCTGGAGTATCGTGGTCTGGCGAAGCTGAAATCGACCTACACCGACAAGCTGCCGCTGATGATCAACCCGAAAACCGGGCGTGTGCATACCTCTTATCACCAGGCAGTAACTGCAACGGGACGTTTATCGTCAACCGATCCTAACCTGCAAAACATTCCGGTGCGTAACGAAGAAGGTCGTCGTATCCGCCAGGCGTTTATTGCGCCAGAGGATTATGTGATTGTCTCAGCGGACTACTCGCAGATTGAACTGCGCATTATGGCGCATCTTTCGCGTGACAAAGGCTTGCTGACCGCATTCGCGGAAGGAAAAGATATCCACCGGGCAACGGCGGCAGAAGTGTTTGGTTTGCCACTGGAAACCGTCACCAGCGAGCAACGCCGTAGCGCGAAAGCGATCAACTTTGGTCTGATTTATGGCATGAGTGCTTTCGGTCTGGCGCGGCAATTGAACATTCCACGTAAAGAAGCGCAGAAGTACATGGACCTTTACTTCGAACGCTACCCTGGCGTGCTGGAGTATATGGAACGCACCCGTGCTCAGGCGAAAGAGCAGGGCTACGTTGAAACGCTGGACGGACGCCGTCTGTATCTGCCGGATATCAAATCCAGCAATGGTGCTCGTCGTGCAGCGGCTGAACGTGCAGCCATTAACGCGCCAATGCAGGGAACCGCCGCCGACATTATCAAACGGGCGATGATTGCCGTTGATGCGTGGTTACAGGCTGAGCAACCGCGTGTACGTATGATCATGCAGGTACACGATGAACTGGTATTTGAAGTTCATAAAGATGATGTTGATGCCGTCGCGAAGCAGATTCATCAACTGATGGAAAACTGTACCCGTCTGGATGTGCCGTTGCTGGTGGAAGTGGGGAGTGGCGAAAACTGGGATCAGGCGCACTAAggctcgggagacctatcg
Biopart containing J23101TetO-RBSK-AID		TCTGGTGGGTCTCTGTCCATCTACCTCAGCTTTACAGCTAGCTCAGTCCTAGGTATTATGCTAGCACTCTATCATTGATAGAGTGGACACGTGGCTCCCTATCAGTGATAGAGATTCACTGCGTAAGGCTCGttgaacaccgtctcaggtaagtatcagttgtaaaTCCCGGGGGGGtaGTCCATGGATAGCCTGCTGATGAATCGTCGCAAATTTCTGTATCAGTTTAAAAATGTGCGTTGGGCCAAAGGTCGTCGTGAAACCTATCTGTGCTATGTTGTGAAACGTCGTGATAGCGCAACCAGCTTTAGCCTGGATTTTGGTTATCTGCGCAATAAAAATGGTTGTCATGTGGAGCTGCTGTTTCTGCGTTATATTAGCGATTGGGATCTGGATCCGGGTCGTTGTTATCGTGTTACCTGGTTTACCAGCTGGTCACCGTGTTATGACTGTGCACGTCATGTTGCAGATTTTCTGCGTGGTAATCCGAATCTGAGCCTGCGTATTTTTACCGCACGTCTGTATTTTTGCGAAGATCGTAAAGCAGAACCGGAAGGTCTGCGTCGTCTGCATCGTGCAGGTGTTCAGATTGCCATTATGACCTTTAAAGATTATTTTTATTGCTGGAATA

		CCTTTGTGGAAAATCATGAACGCACCTTTAAAGCATGGGAAGGCCTGCATGAAAATAGCGTTCGTCTGTCTCGCCAGCTGCGTCGTATTCTGCTGCCGCTGTATGAAGTTGATGATCTGCGTGATGCCTTTCGTACCCTGGGTTTAACCGGTCAGGAAGTTGCATGCACCGCCGGCTAAGGCTCGGGAGACCTATCG

List of ORI-Antibiotic Backbones used in Project

1. pSC101-Gentamycin^R

tctggtgggtctctGTCCACTAGTCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTT
GCCGCCGGGCGTTTTTTATTGGTGAGAATCCAGGGGTCCCCAATAATTACGATTTAAATTTGACATAAGCCTGTTCGGTTCGTAAACTGT
AATGCAAGTAGCGTATGCGCTCACGCAACTGGTCCAGAACCTTGACCGAACGCAGCGGTGGTAACGGCGCAGTGGCGGTTTTCATGGC
TTGTTATGACTGTTTTTTTGTACAGCCTATGCCTCGGGCATCCAAGCAGCAAGCGCGTTACGCCGTGGGTCGATGTTTGATGTTATGGAG
CAGCAACGATGTTACGCAGCAGCAACGATGTTACGCAGCAGGGCAGTCGCCCTAAAACAAAGTTAGGTGGCTCAAGTATGGGCATCAT
TCGCACATGTAGGCTCGGCCCTGACCAAGTCAAATCCATGCGGGCTGCTCTTGATCTTTTCGGTCGTGAGTTCGGAGACGTAGCCACCTA
CTCCCAACATCAGCCGGACTCCGATTACCTCGGGAACTTGCTCCGTAGTAAGACATTCATCGCGCTTGCTGCCTTCGACCAAGAAGCGGT
TGTTGGCGCTCTCGCGGCTTACGTTCTGCCCAAGTTTGAGCAGCCGCGTAGTGAGATCTATATCTATGATCTCGCAGTCTCCGGAGAGCA
CCGGAGGCAGGGCATTGCCACCGCGCTCATCAATCTCCTCAAGCATGAGGCCAACGCGCTTGGTGCTTATGTGATCTACGTGCAAGCAG
ATTACGGTGACGATCCCGCAGTGGCTCTCTATACAAAGTTGGGCATACGGGAAGAAGTGATGCACTTTGATATCGACCCAAGTACCGCC
ACCTAACAATTCGTTCAAGCCGAGATCGGCTTCCCGGCCGCGGAGTTGTTCGGTAAATTGGACAACGGTCGGCTCGttacttacgacactccg
agacagtcagagggtatttattgaactaGTCCGGCCGGCCTCAGATCCTTCCGTATTTAGCCAGTATGTTCTCTAGTGTGGTTCGTTGTTTTTGCGT
GAGCCATGAGAACGAACCATTGAGATCATACTTACTTTGCATGTCACTCAAAAATTTTGCCTCAAAACTGGTGAGCTGAATTTTTGCAGT
TAAAGCATCGTGTAGTGTTTTTCTTAGTCCGTTATGTAGGTAGGAATCTGATGTAATGGTTGTTGGTATTTTGTCACCATTCATTTTTATCT
GGTTGTTCTCAAGTTCGGTTACGAGATCCATTTGTCTATCTAGTTCAACTTGGAAAATCAACGTATCAGTCGGGCGGCCTCGCTTATCAAC
CACCAATTTCATATTGCTGTAAGTGTTTAAATCTTTACTTATTGGTTTCAAAACCCATTGGTTAAGCCTTTTAAACTCATGGTAGTTATTTTC
AAGCATTAACATGAACTTAAATTCATCAAGGCTAATCTCTATATTTGCCTTGTGAGTTTTCTTTTGTGTTAGTTCTTTTAATAACCACTCATA
AATCCTCATAGAGTATTTGTTTTCAAAAGACTTAACATGTTCCAGATTATATTTTATGAATTTTTTTAACTGGAAAAGATAAGGCAATATCT
CTTCACTAAAAACTAATTCTAATTTTTCGCTTGAGAACTTGGCATAGTTTGTCCACTGGAAAATCTCAAAGCCTTTAACCAAAGGATTCCT
GATTTCCACAGTTCTCGTCATCAGCTCTCTGGTTGCTTTAGCTAATACACCATAAGCATTTTCCCTACTGATGTTCATCATCTGAGCGTATT
GGTTATAAGTGAACGATACCGTCCGTTCTTTCCTTGTAGGGTTTTCAATCGTGGGGTTGAGTAGTGCCACACAGCATAAAATTAGCTTGG
TTTCATGCTCCGTTAAGTCATAGCGACTAATCGCTAGTTCATTTGCTTTGAAAACAACTAATTCAGACATACATCTCAATTGGTCTAGGTG
ATTTTAATCACTATACCAATTGAGATGGGCTAGTCAATGATAATTACATGTCCTTTTCCTTTGAGTTGTGGGTATCTGTAAATTCTGCTAG
ACCTTTGCTGGAAAACTTGTAAATTCTGCTAGACCCTCTGTAAATTCCGCTAGACCTTTGTGTGTTTTTTTTGTTTATATTCAAGTGGTTAT
AATTTATAGAATAAAGAAAGAATAAAAAAAGATAAAAAGAATAGATCCCAGCCCTGTGTATAACTCACTACTTTAGTCAGTTCCGCAGTA
TTACAAAAGGATGTCGCAAACGCTGTTTGCTCCTCTACAAAACAGACCTTAAAACCCTAAAGGCTTAAGTAGCACCCTCGCAAGCTCGGG
CAAATCGCTGAATATTCCTTTTGTCTCCGACCATCAGGCACCTGAGTCGCTGTCTTTTTCGTGACATTCAGTTCGCTGCGCTCACGGCTCT
GGCAGTGAATGGGGGTAAATGGCACTACAGGCGCGGCGCGCCCAGCTGTCTAGGGCGGCGGATTTGTCCTACTCAGGAGAGCGTTCA
CCGACAAACAACAGATAAAACGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTGATGCCTTTAATTAAGGCTCGggagacctatcg

2. pSC101-Chloramphenicol^R

tctggtgggtctctGTCCACTAGTCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTT
GCCGCCGGGCGTTTTTTATTGGTGAGAATCCAGGGGTCCCCAATAATTACGATTTAAATTGGCGAAAATGAGACGTTGATCGGCACGTA
AGAGGTTCCAACTTTCACCATAATGAAATAAGATCACTACCGGGCGTATTTTTTGAGTTATCGAGATTTTCAGGAGCTAAGGAAGCTAAA
ATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAA
TGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCACAAGTTTTATCCGGCCTTTATTC
ACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTT
GTTACACCGTTTTCCATGAGCAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTC
GCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGT
GAGTTTCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATATTATACGCAAGGCGA
CAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTA
CTGCGATGAGTGGCAGGGCGGGGCGTAATTTGACTTTTGTCGGCTCGttacttacgacactccgagacagtcagagggtatttattgaactaGTCCGGC
CGGCCTCAGATCCTTCCGTATTTAGCCAGTATGTTCTCTAGTGTGGTTCGTTGTTTTTGCGTGAGCCATGAGAACGAACCATTGAGATCAT
ACTTACTTTGCATGTCACTCAAAAATTTTGCCTCAAAACTGGTGAGCTGAATTTTTGCAGTTAAAGCATCGTGTAGTGTTTTTCTTAGTCC
GTTATGTAGGTAGGAATCTGATGTAATGGTTGTTGGTATTTTGTCACCATTCATTTTTATCTGGTTGTTCTCAAGTTCGGTTACGAGATCC

ATTTGTCTATCTAGTTCAACTTGGAAAATCAACGTATCAGTCGGGCGGCCTCGCTTATCAACCACCAATTTCATATTGCTGTAAGTGTTTA
AATCTTTACTTATTGGTTTCAAAACCCATTGGTTAAGCCTTTTAAACTCATGGTAGTTATTTTCAAGCATTAACATGAACTTAAATTCATCA
AGGCTAATCTCTATATTTGCCTTGTGAGTTTTCTTTTGTGTTAGTTCTTTTAATAACCACTCATAAATCCTCATAGAGTATTTGTTTTCAAAA
GACTTAACATGTTCCAGATTATATTTTATGAATTTTTTTAACTGGAAAAGATAAGGCAATATCTCTTCACTAAAAACTAATTCTAATTTTTC
GCTTGAGAACTTGGCATAGTTTGTCCACTGGAAAATCTCAAAGCCTTTAACCAAAGGATTCCTGATTTCCACAGTTCTCGTCATCAGCTCT
CTGGTTGCTTTAGCTAATACACCATAAGCATTTTCCCTACTGATGTTCATCATCTGAGCGTATTGGTTATAAGTGAACGATACCGTCCGTT
CTTTCCTTGTAGGGTTTTCAATCGTGGGGTTGAGTAGTGCCACACAGCATAAAATTAGCTTGGTTTCATGCTCCGTTAAGTCATAGCGAC
TAATCGCTAGTTCATTTGCTTTGAAAACAACTAATTCAGACATACATCTCAATTGGTCTAGGTGATTTTAATCACTATACCAATTGAGATG
GGCTAGTCAATGATAATTACATGTCCTTTTCCTTTGAGTTGTGGGTATCTGTAAATTCTGCTAGACCTTTGCTGGAAAACTTGTAAATTCT
GCTAGACCCTCTGTAAATTCCGCTAGACCTTTGTGTGTTTTTTTTGTTTATATTCAAGTGGTTATAATTTATAGAATAAAGAAAGAATAAA
AAAAGATAAAAAGAATAGATCCCAGCCCTGTGTATAACTCACTACTTTAGTCAGTTCCGCAGTATTACAAAAGGATGTCGCAAACGCTGT
TTGCTCCTCTACAAAACAGACCTTAAAACCCTAAAGGCTTAAGTAGCACCCTCGCAAGCTCGGGCAAATCGCTGAATATTCCTTTTGTCTC
CGACCATCAGGCACCTGAGTCGCTGTCTTTTTCGTGACATTCAGTTCGCTGCGCTCACGGCTCTGGCAGTGAATGGGGGTAAATGGCAC
TACAGGCGCGGCGCGCCCAGCTGTCTAGGGCGGCGGATTTGTCCTACTCAGGAGAGCGTTCACCGACAAACAACAGATAAAACGAAAG
GCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTGATGCCTTTAATTAAGGCTCGggagacctatcg

3. p15A-Kanamycin^{R}

tctggtgggtctctGTCCACTAGTCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTT
GCCGCCGGGCGTTTTTTATTGGTGAGAATCCAGGGGTCCCCAATAATTACGATTTAAATTTGTGTCTCAAAATCTCTGATGTTACATTGCA
CAAGATAAAAATATATCATCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAGCGTGAA
ACGAGCTGTAGCCGTCCGCGTCTGAACAGCAACATGGATGCGGATCTGTATGGCTATAAATGGGCGCGTGATAACGTGGGTCAGAGCG
GCGCGACCATTTATCGTCTGTATGGCAAACCGGATGCGCCGGAACTGTTTCTGAAACATGGCAAAGGCAGCGTGGCGAACGATGTGAC
CGATGAAATGGTGCGTCTGAACTGGCTGACCGAATTTATGCCGCTGCCGACCATTAAACATTTTATTCGCACCCCGGATGATGCGTGGCT
GCTGACCACCGCGATTCCGGGCAAAACCGCGTTTCAGGTGCTGGAAGAATATCCGGATAGCGGCGAAAACATTGTGGATGCGCTGGCC
GTGTTTCTGCGTCGTCTGCATAGCATTCCGGTGTGCAACTGCCCGTTTAACAGCGATCGTGTGTTTCGTCTGGCCCAGGCGCAGAGCCGT
ATGAACAACGGCCTGGTGGATGCGAGCGATTTTGATGATGAACGTAACGGCTGGCCGGTGGAACAGGTGTGGAAAGAAATGCATAAA
CTGCTGCCGTTTAGCCCGGATAGCGTGGTGACCCACGGCGATTTTAGCCTGGATAACCTGATTTTCGATGAAGGCAAACTGATTGGCTG
CATTGATGTGGGCCGTGTGGGCATTGCGGATCGTTATCAGGATCTGGCCATTCTGTGGAACTGCCTGGGCGAATTTAGCCCGAGCCTGC
AAAAACGTCTGTTTCAGAAATATGGCATTGATAATCCGGATATGAACAAACTGCAATTTCATCTGATGCTGGATGAATTTTTCTAATAATT
AATTGGACCGCGGTCGGCTCGttacttacgacactccgagacagtcagagggtatttattgaactaGTCCGGCCGGCCCTAGAAATATTTTATCTGATTA
ATAAGATGATCTTCTTGAGATCGTTTTGGTCTGCGCGTAATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTTTCGAAG
GTTCTCTGAGCTACCAACTCTTTGAACCGAGGTAACTGGCTTGGAGGAGCGCAGTCACCAAAACTTGTCCTTTCAGTTTAGCCTTAACCG
GCGCATGACTTCAAGACTAACTCCTCTAAATCAATTACCAGTGGCTGCTGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTGGACTCAAG
ACGATAGTTACCGGATAAGGCGCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGCTTGGAGCGAACTGCCTACCCGGAA
CTGAGTGTCAGGCGTGGAATGAGACAAACGCGGCCATAACAGCGGAATGACACCGGTAAACCGAAAGGCAGGAACAGGAGAGCGCA
CGAGGGAGCCGCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCACTGATTTGAGCGTCAGATTTCGTGATGC
TTGTCAGGGGGGCGGAGCCTATGGAAAAACGGCTTTTGCCGCGGCCCTCTCACTTCCCTGTTAAGTATCTTCCTGGCATCTTCCAGGAAA
TCTCCGCCCCGTTCGTAAGCCATTTCCGCTCGCCGCAGTCGAACGACCGAGCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCCG
GCGCGCCCAGCTGTCTAGGGCGGCGGATTTGTCCTACTCAGGAGAGCGTTCACCGACAAACAACAGATAAAACGAAAGGCCCAGTCTT
TCGACTGAGCCTTTCGTTTTATTTGATGCCTTTAATTAAGGCTCGggagacctatcg

4. pUC-Ampicillin^{R}

TCTGGTGGGTCTCTGTCCgtaataacagtccaatctggtgtaacttcggaatCGTCCACTAGTCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCA
GAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAGGGGTCCCCAATAATTACGATTTA
AATTAGTAGCCCGCCTAATGAGCGGGCTTTTTTTTAATTCCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGA
CAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGCATTCAGCATTTTCGTGTGGCGCTGATTCCGTTTTTTGCG
GCGTTTTGCCTGCCGGTGTTTGCGCATCCGGAAACCCTGGTGAAAGTGAAAGATGCGGAAGATCAACTGGGTGCGCGCGTGGGCTATA
TTGAACTGGATCTGAACAGCGGCAAAATTCTGGAATCTTTTCGTCCGGAAGAACGTTTTCCGATGATGAGCACCTTTAAAGTGCTGCTGT
GCGGTGCGGTTCTGAGCCGTGTGGATGCGGGCCAGGAACAACTGGGCCGTCGTATTCATTATAGCCAGAACGATCTGGTGGAATATAG
CCCGGTGACCGAAAAACATCTGACCGATGGCATGACCGTGCGTGAACTGTGCAGCGCGGCGATTACCATGAGCGATAACACCGCGGCG
AACCTGCTGCTGACGACCATTGGCGGTCCGAAAGAACTGACCGCGTTTCTGCATAACATGGGCGATCATGTGACCCGTCTGGATCGTTG
GGAACCGGAACTGAACGAAGCGATTCCGAACGATGAACGTGATACCACCATGCCGGCAGCAATGGCGACCACCCTGCGTAAACTGCTG
ACGGGTGAGCTGCTGACCCTGGCAAGCCGCCAGCAACTGATTGATTGGATGGAAGCGGATAAAGTGGCGGGTCCGCTGCTGCGTAGC

GCGCTGCCGGCTGGCTGGTTTATTGCGGATAAAAGCGGTGCGGGCGAACGTGGCAGCCGTGGCATTATTGCGGCGCTGGGCCCGGAT
GGTAAACCGAGCCGTATTGTGGTGATTTATACCACCGGCAGCCAGGCGACGATGGATGAACGTAACCGTCAGATTGCGGAAATTGGCG
CGAGCCTGATTAAACATTGGTAAACCGATACAATTAAAGGCTCCTTTTGGAGCCTTTTTTTTTGGACGACCCTTGTCGGCTCGACCCACGA
CTATTGACTGCTCTGAGAAAGTTGATTGTTACGATTAGTCCGGCCGGCCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTC
TGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGA
AGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCAC
CGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGAT
AGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGA
GATACCTACAGCGTGAGCTTTGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAA
CAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGA
TTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTT
GCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAA
CGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTA
ATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCAGGC
GCGCCCAGCTGTCTAGGGCGGCGGATTTGTCCTACTCAGGAGAGCGTTCACCGACAAACAACAGATAAAACGAAAGGCCCAGTCTTTC
GACTGAGCCTTTCGTTTTATTTGATGCCTTTAATTAAGGCTCGggtaagaactcgcacttcgtggaaacactattaGGCTCGGGAGACCTATCG

5. mScarlet Chromoprotein

Chromoprotein gives distinct pink colour to bacterial colonies. It is released from backbone for successful BASIC DNA assembly.

TCTGGTGGGTCTCTGTCCATGGTTAGCAAAGGCGAGGCGGTTATCAAGGAGTTTATGCGTTTTAAGGTTCACATGGAGGGTAGCATGAA
TGGTCACGAGTTCGAGATCGAGGGTGAAGGCGAGGGTCGTCCGTACGAAGGCACCCAGACCGCGAAGCTGAAAGTGACCAAGGGTG
GCCCGCTGCCGTTCAGCTGGGACATCCTGAGCCCGCAGTTCATGTATGGCAGCCGTGCGTTTACCAAACACCCGGCGGACATTCCGGAT
TACTATAAGCAAAGCTTCCCGGAAGGTTTTAAATGGGAGCGTGTTATGAACTTCGAAGATGGTGGCGCGGTGACCGTTACCCAGGACAC
CAGCCTGGAGGATGGCACCCTGATTTACAAGGTGAAACTGCGTGGCACCAACTTTCCGCCGGATGGTCCGGTTATGCAGAAGAAAACG
ATGGGTTGGGAAGCGAGCACCGAGCGTCTGTATCCGGAAGATGGCGTGCTGAAGGGTGATATCAAAATGGCGCTGCGTCTGAAGGAC
GGTGGCCGTTACCTGGCGGATTTTAAGACCACCTATAAAGCGAAGAAACCGGTGCAAATGCCGGGTGCGTACAACGTTGACCGTAAAC
TGGATATTACCAGCCACAACGAGGATTATACCGTGGTTGAGCAATATGAGCGTAGCGAGGGTCGCCACAGCACCGGCGGCATGGACGA
ACTGTATAAGGGATCCtaaGGCTCGGGAGACCTATCG

6. Purple Chromoprotein

Chromoprotein gives a purple colour to bacterial colonies

TCTGGTGGGTCTCTGTCCttgacagctagctcagtcctaggtattgtgctagctactagtgaaagaggagaaatactagTTTTACAGCTAGCTCAGTCCTAGGT
ATTATGCTAGCTACTAGAGAAAGAGGAGAAATACTAGATGGCATCTCTGGTTAAAAAAGATATGTGTGTTAAGATGACAATGGAAGGA
ACAGTGAATGGATACCATTTCAAATGTGTCGGAGAGGGCGAAGGCAAGCCTTTTGAGGGAACCCAGAATATGCGCATCCGCGTCACTG
AGGGCGGGCCGTTACCCTTTGCATTTGACATTCTTGCTCCATGTTGCATGTACGGCTCGAAAACCTTCATCAAGCATGTCTCAGGGATTC
CGGACTATTTCAAAGAATCCTTTCCCGAAGGCTTCACATGGGAACGTACCCAAATCTTTGAGGACGGGGGCGTCTTGACGGCTCATCAA
GATACTTCTCTGGAAGGTAACTGCTTGATTTATAAAGTAAAGGTTCTGGGCACGAATTTTCCGGCTAATGGACCAGTGATGCAGAAGAA
AACCGCTGGTTGGGAGCCTTGCGTAGAAATGTTATACCCGCGTGATGGAGTGCTGTGTGGTCAGAGCCTGATGGCATTAAAATGCACG
GATGGTAACCACCTGACCTCACACCTGCGTACAACATATCGCAGCCGTAAACCCAGCAACGCGGTGAATATGCCTGAGTTTCACTTTGGC
GACCATCGCATTGAGATTCTGAAAGCTGAACAAGGAAAGTTTTATGAGCAGTACGAATCAGCCGTTGCTCGTTATAGCGACGTACCGGA
AAAAGCAACTTGATAATACTAGAGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGT
CGGTGAACGCTCTCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGGCTCGGGAGACCTATCG

List of Primers Used in this Thesis

ID	Name	Sequence 5'-3'	Application
P1	T7 RNA Pol 800-For	TTCCAACCTTGCGTAGTTCCTCCTAAG	Sequencing
P2	T7 RNA Pol 1500-For	GCTAAGTCTCCACTGGAGAACACTTGGT	Sequencing
P3	T7 pol 2700-Reverse	TCCGACTCTAAGATGTCACGGAGGTTC	Sequencing
P4	T7 pol 800-Reverse	TAGGAGGAACTACGCAAGGTTGGAACA	Sequencing
P5	RecJ_PCR_for + iP	TCTGGTGGGTCTCTGTCCATGAAACAACAGATACAACTTCGTCGCCG	PCR
P6	RecJ_PCR_rev NOSTOP + iS	CGATAGGTCTCCCGAGCCAATTGGCCAGATATTGTCGATGATAATTTGCAGG	PCR
P7	UGI-Forward + iP	TCTGGTGGGTCTCTGTCCATGACGAATTTATCGGACATTATCGAAAAGGAGAC	PCR
P8	UGI+Terminator-Reverse + iS	CGATAGGTCTCCCGAGCCtaccaccgtcaaaaaaaacggcgc	PCR
P9	AMP-WT_forward	TCTGGTGGGTCTCTGTCCATGAGTATTCAACATTTCCGTGTCGCCCT	PCR
P10	AMP-ACG_forward	TCTGGTGGGTCTCTGTCCACGAGTATTCAACATTTCCGTGTCGCCCTTATTC	Site-directed Mutation
P11	AMP-ATC_forward	TCTGGTGGGTCTCTGTCCATCAGTATTCAACATTTCCGTGTCGCCCTTATTC	Site-directed Mutation
P12	AMP-CTG_forward	TCTGGTGGGTCTCTGTCCCTGAGTATTCAACATTTCCGTGTCGCCCTTATTC	Site-directed Mutation
P13	AMP-Reverse primer	CGATAGGTCTCCCGAGCCTTACCAATGCTTAATCAGTGAGGCACCTATC	PCR/Site-directed Mutation
P14	Amp-TAA primer for	AAATAGACTGGATGGAGGCGGATAAAGTTGC	Site-directed Mutation
P15	Amp-TAA primer rev	ATTGTTGCCGGGAAGCTAGAGTAAGTAGT	Site-directed Mutation
P16	ExoIII_ATG_for	TCTGGTGGGTCTCTGTCCATGAAATTTGTCTCTTTTAATATCAACGGCCTGC	PCR
P17	ExoIII_no stop_rev	CGATAGGTCTCCCGAGCCGCGGCGGAAGGTCGCC	PCR
P18	PolIV_WT_for	TCTGGTGGGTCTCTGTCCATGCGTAAAATCATTCATGTGGATATGGACTGC	PCR
P19	PolIV_WT_stop_rev	CGATAGGTCTCCCGAGCCTTATAATCCCAGCACCAGTTGTCTTTCCATT	PCR
P20	PolIV_-5_nostop_rev	CGATAGGTCTCCCGAGCCTCTTTCCATTTGCGGGTCAAGCAACG	PCR
P21	PolIV_-12_STOP_rev	CGATAGGTCTCCCGAGCCTTAAAGCAACGTCACATGCAGCCC	PCR
P22	Pol I Exo - Forward	TCTGGTGGGTCTCTGTCCATGGTTCAGATCCCCCAAAATCCACTTATCC	PCR
P23	Pol I Exo - No Stop reverse	CGATAGGTCTCCCGAGCCCGTTGCCGTCACTTCTGGTGCT	PCR
P24	Pol I Exo Short - No Stop Rev	CGATAGGTCTCCCGAGCCTTTAATCGTCGCCAGCTGGTATGAGAGATA	PCR
P25	NAPE no stop for	TCTGGTGGGTCTCTGTCCATGTTAAAGATCATCTCGGCTAATGTCAACGGA	PCR
P26	NAPE No Stop Rev	CGATAGGTCTCCCGAGCCCTCTGCGGCGTAATCATATTCGACGAC	PCR
P27	HP2-RBSK-AID_For	TCTGGTGGGTCTCTGTCCATCTACCTCAGCTTTACAGCTAGCTCAGTCC	PCR
P28	HP2-RBSK-AID(s)_Rev	CGATAGGTCTCCCGAGCCTTAGCCGGCGGTGCATGCAAC	PCR
P29	RpoB_iP_for	TCTGGTGGGTCTCTGTCCATGGTTTACTCCTATACCGAGAAAAAACGTATTCGT	PCR
P30	RpoB_iS_rev	CGATAGGTCTCCCGAGCCTTACTCGTCTTCCAGTTCGATGTTGATACCCA	PCR
P31	2mut1_F742Y_for	ATGGTTTGCCACTGGAAACCGTCA	Site-directed Mutation
P32	2mut1_F742Y_rev	ACACTTCTGCCGCCGTTGCC	Site-directed Mutation

P33	2mut1_P796H_for	ATGGCGTGCTGGAGTATATGGAACGC	Site-directed Mutation
P34	2mut1_P796H_rev	GGTAGCGTTCGAAGTAAAGGTCCATGTACTTC	Site-directed Mutation
P35	2mut2_D548A_for	CTCCGAAAGTGCTGCACAATCATTCTGAAGAG	Site-directed Mutation
P36	2mut2_D548A_rev	CGATCTTCACACCGTTACGTTCAATGCG	Site-directed Mutation
P37	2mut3_I709N_for	ATGAACTGCGCATTATGGCGCATC	Site-directed Mutation
P38	2mut3_I709N_rev	TCTGCGAGTAGTCCGCTGAGACAAT	Site-directed Mutation
P39	2mut3_A759R_for	CGGATCAACTTTGGTCTGATTTATGGCATGAGT	Site-directed Mutation
P40	2mut3_A759R_rev	TTTCGCGCTACGGCGTTGCT	Site-directed Mutation
P41	2mut2_P549L_for	TGAAAGTGCTGCACAATCATTCTGAAGAGC	Site-directed Mutation
P42	2mut2_P549L_rev	GATCGATCTTCACACCGTTACGTTCAATGC	Site-directed Mutation
P43	GFP (post mutation PCR) For	CCAGTGCCAAGCTTGCATGC	PCR
P44	GFP (post mutation PCR) Rev	CCTCGCTTTGTAACGGAGTAGAGACG	PCR
P45	GFP (PM PCR) outer For	GCGAGAGTAGGGAACTGCCAGGC	PCR
P46	GFP (PM PCR) outer Rev	GCAAGGCTATGTGCCATCTCGATACTCG	PCR
P47	Amp target plasmid_for	CACGATGCGTCCGGCGTAGA	Sequencing
P48	Amp target plasmid_rev	ccCGAGCCTATAAACGCAGAAAGGC	Sequencing
P49	Methylated Linker A - For	CTATTATCTGGTGGGTCTCT	PCR/Sequencing
P50	Methylated Linker B - Rev	TTACCGATAGGTCTCCCG	PCR/Sequencing

9.4 Python script to concatenate biopart sequences to generate assembled expression cassettes and plasmids

```
__author__ = 'Marko Storch'

"""Edited by Haris Mallick"""

#coding: utf8

import glob,os,string,shutil,math


prefix='TCTGGTGGGTCTCTGTCC'

suffix='GGCTCGGGAGACCTATCG'


#linker library, define all the linker you are using without the 'GGCTCG' and 'GTCC' as strings
```

```
#Define neutral Linkers:
linker01='ttacttacgacactccgagacagtcagagggtatttattgaacta'
linker02='atcggtgtgaaaagtcagtatccagtcgtgtagttcttattacct'
linker03='atcacggcactacactcgttgctttatcggtattgttattacaga'
linker04='acccacgactattgactgctctgagaaagttgattgttacgatta'
linker05='agaagtagtgccacagacagtattgcttacgagttgatttatcct'
linker06='gtattgtaaagcacgaaacctacgataagagtgtcagttctcctt'
linker07='aacttttacgggtgccgactcactattacagacttactacaatct'

#Define Methylated Linkers:
linkerMA='ggtaagaactcgcacttcgtggaaacactattatctggtgggtctct'
linkerMB='ggagacctatcggtaataacagtccaatctggtgtaacttcggaatc'

#Define Fusion Linkers
linkerFU1='gccgaagcggctgctaaagaagcagctgctaaagaggcggccgccaaggc'
linkerFU2='gggtcgggctccggatctggttcaggttcaggatcgggctccgg'
linkerFU3='ctgcttgagagccctaaagcattagaagaagcaccttggcctccaccagaggg'

#Define RBS Linkers
linker1RBS1='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcCAAGGAGGTA'
linker1RBS2='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcCAGGGAGGTA'
linker1RBS3='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcCCAGGAGGTA'
linker1RBS4='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcCCGGGAGGTA'
linker1RBS5='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcCGAGGAGGTA'
linker1RBS6='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcCGGGGAGGTA'
linker1RBS7='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcTAAGGAGGTA'
linker1RBS8='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcTAGGGAGGTA'
linker1RBS9='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcTCAGGAGGTA'
linker1RBS10='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcTCGGGAGGTA'
linker1RBS11='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcTGAGGAGGTA'
linker1RBS12='TTGAACACCGTCTCAGGTAAGTATCAGTTGTAAatcTGGGGAGGTA'
linker1RBS13='ttgaacaccgtctcaggtaagtatcagttgtaaatcacacaggacta'
linker1RBS14='ttgaacaccgtctcaggtaagtatcagttgtaaaaagaggggaaata'
linker1RBS15='ttgaacaccgtctcaggtaagtatcagttgtaaaaagaggagaaata'

linker2RBS13='tgttactattggctgagataagggtagcagaaaatcacacaggacta'
linker2RBS14='tgttactattggctgagataagggtagcagaaaaaagaggggaaata'
```

```
linker2RBS15='tgttactattggctgagataagggtagcagaaaaaagaggagaaata'

#number of parts: Specify number of parts being used
partsnumber=7

#Name you Assembled plasmid here:
NAME='Name'
assembly=''
#linker assignment: assign linker[1] to linker[partsnumber] to linkers from above
linker=['1','1','1','1','1','1','1','1','1','1']
linker[1]=linkerMA
linker[2]=linkerFU2
linker[3]=linkerFU3
linker[4]=linkerMB
linker[5]=linker01
linker[6]=linker05
linker[7]=linker01
linker[8]=linker07

#define all the part files in .xdna you are using in the order they will be assembled
part=['1','1','1','1','1','1','1','1','1','1','1']
part[1]='Promoter'
part[2]='Biopart'
part[3]='Terminator'
part[4]='Promoter'
part[5]='Biopart 2'
part[6]='Antibiotic'
part[7]='ORI'
part[8]=''

def partextract(xdnafile):
    with open(xdnafile+'.xdna', 'r') as content_file:
        sequence = content_file.read()
    s=(sequence[(sequence.find('ÿ')+1):]).upper()
    #The DNA sequence within the BASIC prefix and suffix is identified with the find command
    part=s[(s.find(prefix)+len(prefix)-4):(s.find(suffix)+6)]
    part=part.upper()
    print(len(part))
    return(part)

for i in range(partsnumber):
    assembly=assembly+linker[i+1]+partextract(part[i+1])
print(assembly)
#The assebmled DNA is extracted into a new xdna file in the same folder destination
xdna_file = open(NAME+'.xdna', 'w')
xdna_file.write(assembly)
xdna_file.close()
```

9.5 Chapter 5 – Python Script to Process NGS FastQ Files and Identify Mutations

```
"""
@author: Haris & Rohan
"""
from Bio import SeqIO
from Bio import pairwise2
import numpy as np
import os

#List of PacBio Prefix Linkers:
Prefix_A='TCAGACGATGCGTCAT'
Prefix_B='CTATACATGACTCTGC'
Prefix_C='TACTAGAGTAGCACTC'
Prefix_D='TGTGTATCAGTACATG'
Prefix_E='ACACGCATGACACACT'
Prefix_F='GATCTCTACTATATGC'
Prefix_G='ACAGTCTATACTGCTG'
Prefix_H='ATGATGTGCTACATCT'

#List of PacBio Suffix Linkers:
Suffix_1='CATAGCGACTATCGTG'
Suffix_2='CATCACTACGCTAGAT'
Suffix_3='CGCATCTGTGCATGCA'
Suffix_4='TATGTGATCGTCTCTC'
Suffix_5='GTACACGCTGTGACTA'
Suffix_6='CGTGTCGCGCATATCT'
```

```
Suffix_7='ATATCAGTCATGCATA'

Suffix_8='GAGATCGACAGTCTCG'

Suffix_9='CACGCACACACGCGCG'

Suffix_10='CGAGCACGCGCGTGTG'

Suffix_11='GTAGTCTCGCACAGAT'

Suffix_12='GAGACTCTGTGCGCGT'


#The Target DNA sequenced via PacBio

TargetDNA='GTCCACAATTTTCGAAAAAACCCGCTTCGGCGGGTTTTTTTATAGCTAAAAGATTTGACAGCTAGCTCAGTCCTAGGGATT
GTGCTAGCGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGTACTAGAGGGCTCGTTGAAC
ACCGTCTCAGGTAAGTATCAGTTGTAAAAAGAGGAGAAATAGTCCATGCGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCT
TGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTA
AATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGTCACTACTTTCGGTTATGGTGTTCAATGCTTTGCGAGATACCC
AGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATATTTTTCAAAGATGACGGGA
ACTACAAGACACGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATAGAATCGAGTTAAAAGGTATTGATTTTAAAGAAGATGGA
AACATTCTTGGACACAAATTGGAATACAACTATAACTCACACAATGTATACATCATGGCAGACAAACAAAAGAATGGAATCAAAGTTAA
CTTCAAAATTAGACACAACATTGAAGATGGAAGCGTTCAACTAGCAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCT
TTTACCAGACAACCATTACCTGTCCACACAATCTGCCCTTTCGAAAGATCCCAACGAAAAGAGAGATCACATGGTCCTTCTTGAGTTTGTA
ACAGCTGCTGGGATTACACATGGCATGGATGAACTATACAAATAAGGCTCGATCGGTGTGAAAAGTCAGTATCCAGTCGTGTAGTTCTT
ATTACCTGTCCCCTAGCATAACCCCGCGGGGCCTCTTCGGGGGACTCGCGGGGTTTTTTGCTGAAAGAATTATCAAATAAAACGAAAGG
CTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGCTGCATTACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGG
GGTTTTTTGGGCTCG'

print(len(TargetDNA))

#Create barcoded reference sequences for downstream alignment by combining Prefix+TargetDNA+Suffix:

Ref_1= str(Prefix_A)+str(TargetDNA)+str(Suffix_1)

#print(Ref_1)

Ref_2= str(Prefix_B)+str(TargetDNA)+str(Suffix_2)

#print(Ref_2)

Ref_3= str(Prefix_C)+str(TargetDNA)+str(Suffix_3)

#print(Ref_3)

Ref_4= str(Prefix_D)+str(TargetDNA)+str(Suffix_4)

#print((Ref_4))

Ref_5= str(Prefix_E)+str(TargetDNA)+str(Suffix_5)

#print((Ref_5))

Ref_6= str(Prefix_F)+str(TargetDNA)+str(Suffix_6)

#print(Ref_6)

Ref_7= str(Prefix_G)+str(TargetDNA)+str(Suffix_7)

#print(Ref_7)

Ref_8= str(Prefix_H)+str(TargetDNA)+str(Suffix_8)

#print((Ref_8))


#reference for PacBio alignment (len is 1202):
```

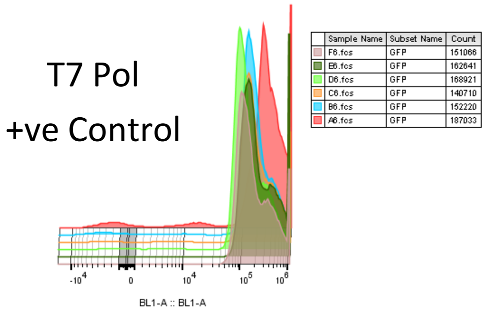
```
reference='GATCTCTACTATATGCGTCCACAATTTTCGAAAAAACCCGCTTCGGCGGGTTTTTTTATAGCTAAAAGATTTGACAGCTAGC
TCAGTCCTAGGGATTGTGCTAGCGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGTACTAG
AGGGCTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAAAAAGAGGAGAAATAGTCCATGCGTAAAGGAGAAGAACTTTTCACTGG
AGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACG
GAAAACTTACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGTCACTACTTTCGGTTATGGTGTTCAATGC
TTTGCGAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTATATTTTTC
AAAGATGACGGGAACTACAAGACACGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATAGAATCGAGTTAAAAGGTATTGATTT
TAAAGAAGATGGAAACATTCTTGGACACAAATTGGAATACAACTATAACTCACACAATGTATACATCATGGCAGACAAACAAAAGAATG
GAATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGAAGCGTTCAACTAGCAGACCATTATCAACAAAATACTCCAATTGGC
GATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCCACACAATCTGCCCTTTCGAAAGATCCCAACGAAAAGAGAGATCACATGGTC
CTTCTTGAGTTTGTAACAGCTGCTGGGATTACACATGGCATGGATGAACTATACAAATAAGGCTCGATCGGTGTGAAAAGTCAGTATCC
AGTCGTGTAGTTCTTATTACCTGTCCCCTAGCATAACCCCGCGGGGCCTCTTCGGGGGACTCGCGGGGTTTTTTGCTGAAAGAATTATCA
AATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGCTGCATTACTAGCATAACCCCTTGGGGCCTCT
AAACGGGTCTTGAGGGGTTTTTTGGGCTCGCGTGTCGCGCATATCT'

nonsense_sequence_1 = 'TTTTTTTTTTTTTTTTTTTTTT'

nonsense_sequence_2 = 'NNNNNNNNNN'

max_position_for_mutation_heatmap = 1400

alignment_condition = 1197


def id_mutation(reference_mut,sequence_mut):

   mutationlist = []

   positionlist = [0] * max_position_for_mutation_heatmap

   for base in range(0,len(reference_mut)):

      if reference_mut[base]!=sequence_mut[base]:

         #print(reference[base],'->',sequence_mut[base])

         mutationlist.append(reference_mut[base]+sequence_mut[base])

         #print(base)

         positionlist[base] +=1

   return mutationlist, positionlist


# creates countlists for each mutation type and iterate through sequences in selected fastqfile counting encountered mutations


def countmut(fastqfile):

   startlist=[]

   positionlist = [0] * max_position_for_mutation_heatmap

   countlist={}

   countlist['AT'] = 0

   countlist['AC'] = 0

   countlist['AG'] = 0

   countlist['TA'] = 0

   countlist['TC'] = 0

   countlist['TG'] = 0
```

```
    countlist['CT'] = 0
    countlist['CA'] = 0
    countlist['CG'] = 0
    countlist['GT'] = 0
    countlist['GC'] = 0
    countlist['GA'] = 0
    countlist['T-'] = 0
    countlist['A-'] = 0
    countlist['G-'] = 0
    countlist['C-'] = 0
    countlist['-T'] = 0
    countlist['-A'] = 0
    countlist['-G'] = 0
    countlist['-C'] = 0
    sequences_with_additions = 0
    sequences_with_deletions = 0
    sequences_with_substitutions = 0
    nonsense_sequence_count = 0
    unchanged_sequence_count = 0
    total_number_of_mutations_in_file = 0
    number_of_sequences_in_file = 0

    for record in SeqIO.parse(fastqfile, "fastq"):
        startlist.append(str(record.seq))

    #test =
'ATGCGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAATGGGCACAAATCGTCTGTCAG
TGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCCTTAAA'
    print('Total sequences in file are ',len(startlist))
    for sequence in range(0,len(startlist)):
        #nonsense sequence test: We try to look for a substring consisting  of 20 Ts or Ns
        #find() function returns -1 if not found
        if (startlist[sequence].find(nonsense_sequence_1) != -1 or startlist[sequence].find(nonsense_sequence_2) != -1) or
(len(startlist[sequence]) > 1210) or (len(startlist[sequence]) < 1195):
            nonsense_sequence_count += 1
            continue

        number_of_sequences_in_file += 1
        if (len(startlist[sequence]) == len(reference)):
            alignment_score = int(pairwise2.align.globalxs(reference, startlist[sequence], -1.0, -0.1, score_only = True))
    #      print(alignment_score)
            if(alignment_score != len(reference) and alignment_score > alignment_condition):
                #print('Sequence is ', sequence)
                sequences_with_substitutions += 1
                mutationlist, temppositionlist = id_mutation(reference, startlist[sequence])
                positionlist = np.add(positionlist,temppositionlist)
                for index in range(0,len(mutationlist)):
                    countlist[mutationlist[index]] +=1
            elif (alignment_score == len(reference)):
                unchanged_sequence_count += 1

        elif (len(startlist[sequence]) < len(reference)):
            alignment_score2 = int(pairwise2.align.globalxs(reference, startlist[sequence], -1.0, -0.1, score_only = True))
    #      print(alignment_score)
            if (alignment_score2 != len(reference) and alignment_score2 > alignment_condition):

                sequences_with_deletions += 1
                alignment_result = pairwise2.align.globalxs(reference, startlist[sequence], -1.0, -0.1, one_alignment_only=True)
                #print(alignment_result)

                first_alignment = alignment_result.pop()
                new_reference = first_alignment[0]
                new_sequence = first_alignment[1]
                #print('New Reference is ',new_reference)
                #print('New Sequence is ',new_sequence)
                mutationlist, temppositionlist = id_mutation(new_reference,new_sequence)
                positionlist = np.add(positionlist,temppositionlist)
                for index in range(0,len(mutationlist)):
                    countlist[mutationlist[index]] +=1

        elif (len(startlist[sequence]) > len(reference)) :
            alignment_score3 = int(pairwise2.align.globalxs(reference, startlist[sequence], -1.0, -0.1, score_only = True))
    #      print(alignment_score)
            if (alignment_score3 != len(reference) and alignment_score3 > alignment_condition):
                sequences_with_additions += 1
                alignment_result = pairwise2.align.globalxs(reference, startlist[sequence], -1.0, -0.1, one_alignment_only=True)
                #print(alignment_result)


                first_alignment = alignment_result.pop()
                new_reference = first_alignment[0]
                new_sequence = first_alignment[1]
                # print('New Reference is ',new_reference)
                # print('New Sequence is ',new_sequence)
                mutationlist, temppositionlist = id_mutation(new_reference,new_sequence)
                positionlist = np.add(positionlist,temppositionlist)
                for index in range(0,len(mutationlist)):
                    countlist[mutationlist[index]] +=1


    total_number_of_mutations_in_file = np.sum(positionlist)
    avg_mutation_per_sequence = total_number_of_mutations_in_file / number_of_sequences_in_file
    print('Additions in ',sequences_with_additions,' sequences, deletions in ',sequences_with_deletions,' sequences &
substitutions in ',sequences_with_substitutions)
    print('Number of nonsense sequences in file are ',nonsense_sequence_count)
    print('Number of unchanged sequences in file are ', unchanged_sequence_count)
    print('Number of sequences in file are ', number_of_sequences_in_file)
    print('Avg no. of mutations per sequence are ', avg_mutation_per_sequence)


    return(countlist, positionlist, sequences_with_additions, sequences_with_deletions, sequences_with_substitutions,
unchanged_sequence_count, number_of_sequences_in_file, avg_mutation_per_sequence)


for file in os.listdir('input_mutation'):
    if file.endswith(".fastq"):
        print()
        print(file)
        score_distribution, mutationpositions, number_of_additions, number_of_deletions, number_of_substitutions,
unchanged_sequences, Total_sequences, avg_mutations_per_seq = (countmut('input_mutation/'+file))
        print(score_distribution)
        with open('output_mutation/'+file[:-5]+'csv', 'w') as f:
            for item in score_distribution:
                f.write("%s\n" % str(item+','+str(score_distribution[item])))
            f.write("%s\n" % str('Number of sequences with additions,'+ str(number_of_additions)))
            f.write("%s\n" % str('Number of sequences with deletions,'+ str(number_of_deletions)))
```

```
f.write("%s\n" % str('Number of sequences with substitutions,'+ str(number_of_substitutions)))
f.write("%s\n" % str('Number of unchanged sequences,'+ str(unchanged_sequences)))
f.write("%s\n" % str('Number of Total sequences,'+ str(Total_sequences)))
f.write("%s\n" % str('Avg. no. of mutations per sequence,'+ str(avg_mutations_per_seq)))
for position in range(0, len(mutationpositions)):
f.write("%s\n" % str(str(position) + ',' + str(mutationpositions[position])))
```

9.6 Chapter 5 – Flow Cytometry Analysis of the 144-Hour Mutagenesis of GFP-mut3b

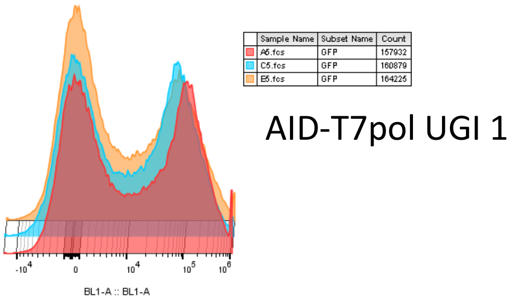
Flow cytometry analysis shown from three different time-points: 96-hrs (RED), 120-hrs (CYAN) and 144-hrs (Orange)



T7 Pol

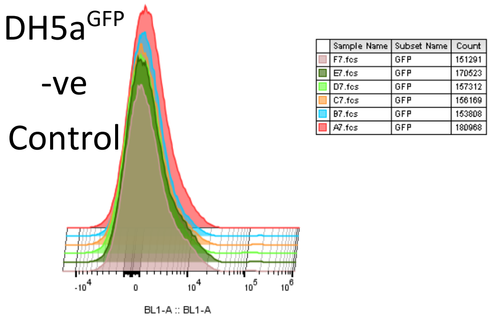
+ve Control

	Sample Name	Subset Name	Count
■	F6.fcs	GFP	151066
■	B6.fcs	GFP	162641
■	D6.fcs	GFP	168921
■	C6.fcs	GFP	140710
■	B6.fcs	GFP	152220
■	A6.fcs	GFP	187033



AID-T7pol UGI 1

	Sample Name	Subset Name	Count
🟥	A5.fcs	GFP	157932
■	C5.fcs	GFP	160879
■	E5.fcs	GFP	164225

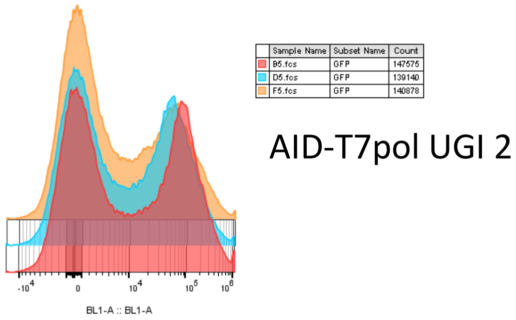


DH5a^{GFP}

-ve

Control

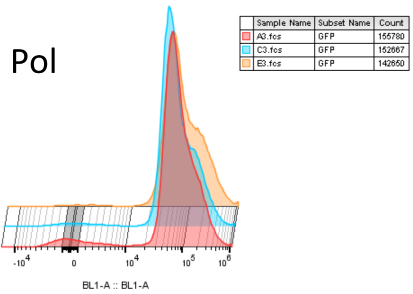
	Sample Name	Subset Name	Count
■	F7.fcs	GFP	151291
■	E7.fcs	GFP	170523
■	D7.fcs	GFP	157312
■	C7.fcs	GFP	156169
■	B7.fcs	GFP	153808
🟥	A7.fcs	GFP	180968



AID-T7pol UGI 2

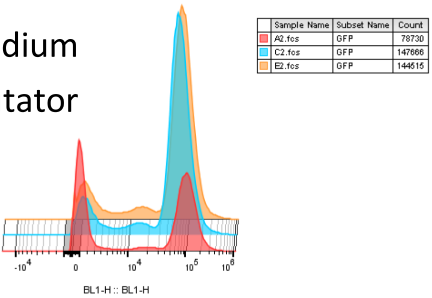
	Sample Name	Subset Name	Count
🟥	B5.fcs	GFP	147575
■	D5.fcs	GFP	139140
■	F5.fcs	GFP	140878

T7 Pol



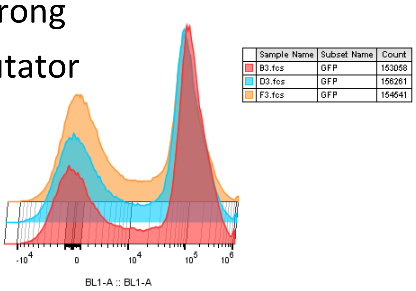
	Sample Name	Subset Name	Count
🟥	A3.fcs	GFP	155780
🟦	C3.fcs	GFP	152667
🟧	E3.fcs	GFP	142650

Medium Mutator



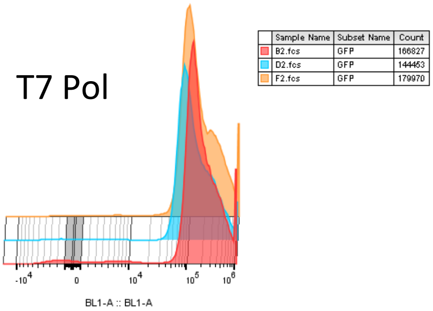
	Sample Name	Subset Name	Count
🟥	A2.fcs	GFP	78730
🟦	C2.fcs	GFP	147666
🟧	E2.fcs	GFP	144515

Strong Mutator



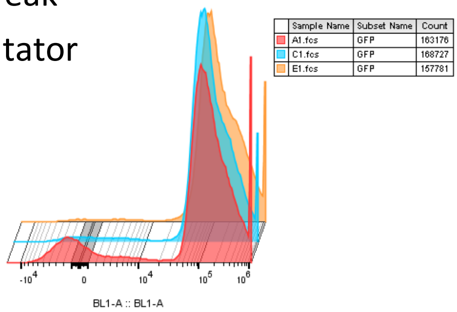
	Sample Name	Subset Name	Count
🟥	B3.fcs	GFP	153058
🟦	D3.fcs	GFP	156261
🟧	F3.fcs	GFP	154541

T7 Pol



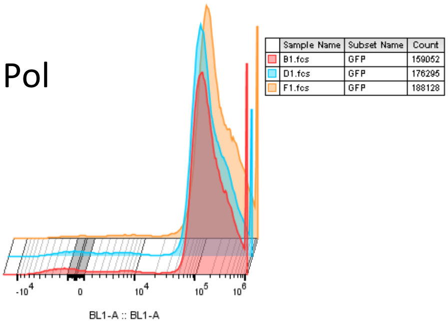
	Sample Name	Subset Name	Count
🟥	B2.fcs	GFP	166827
🟦	D2.fcs	GFP	144453
🟧	F2.fcs	GFP	179970

Weak Mutator



	Sample Name	Subset Name	Count
🟥	A1.fcs	GFP	163176
🟦	C1.fcs	GFP	168727
🟧	E1.fcs	GFP	157781

T7 Pol



	Sample Name	Subset Name	Count
🟥	B1.fcs	GFP	159052
🟦	D1.fcs	GFP	176295
🟧	F1.fcs	GFP	188128