A Novel Biological Circuit to Control Directed Evolution *In Vivo*

Ben Paul Mackrow

Imperial College London,

Department of Life Sciences

DPhil in Synthetic Biology

September 2016

i) Statement of Originality

I declare that this thesis, which I submit for examination in consideration for the award of DPhil in Synthetic Biology, to be awarded by Imperial College London, is my own. The experimental work presented here is a complete and true reflection of the research carried out during three years of laboratory time during my doctoral studies. I hereby state that all content is original and a result of my own personal findings. Where there has been physical or intellectual input or data received from other individuals, groups or taken from previously published research, I have acknowledged all persons with complete transparency.

ii) Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

iii) With Thanks

Firstly, I would like to thank my supervisor Dr Geoff Baldwin for his guidance and support throughout my time as part of his research group. It goes without saying that I would not have had this opportunity without him and he has provided me with ample opportunity to improve my scientific expertise, technical background and business acumen. I would not have been able to complete this body of work without his ideas and advice and for that I am hugely appreciative.

I would also like to thank the rest of Baldwin Group for all the help and suggestions needed during day to day life in the lab. I personally want to thank Marko S, James F and Matt H for their professional support and close friendship throughout. Marko, in particular, has been a constant source of advice and has worked alongside with me other the 3 years. It has been an absolute pleasure to work alongside you all. I must also remember the very many BSc and MRes students who I have had the responsibility of guiding through smaller projects. I am particularly thankful to Richard, Harry and Will for their hard work and dedication during data collection and analysis. I am also indebted to fellow members of CSynBi, specifically Dr Chris Hurst and Catherine Ainsworth who have helped with robotic automation and data analysis, respectively.

Most importantly, I want to say a huge thank you to my parents Clive and Sherri, and my sister Chelsey, for their everlasting support and total belief in me. They have provided me with the confidence and the platform to extend postgraduate study and understand that I'm doing something that I enjoy and feel extremely passionately about.

Without all of you I wouldn't be where I am now and for that, I am eternally grateful.

iv) Thesis Statement / Abstract

Work outlined in this report describes the development a biological device which can precisely control the type, speed, and spread of mutations directed towards a target gene of interest. The intention was to continuously search through a vast DNA sequence landscape by constructing a self-controlling system *in vivo*, which can direct the evolution of the target protein using a non-rational approach, in an effort to identify new or improved functionality, whilst circumnavigating the need for complex and unreliable rational protein design.

I describe the creation of a novel fusion protein, consisting of the mutagenic protagonist, Activation Induced Deaminase (AID), tethered to a highly targetable transcription factor (T7 RNA Polymerase). By specifically mutating only the intended target gene this system is able to increase the number of interesting mutations i.e. those which have a direct consequence on protein function. Furthermore, results show that reducing background mutations helps maintain a viable host cell. Initial results suggest that when this mutator protein is expressed in an *E.coli* strain with a limited base excision repair (BER) pathway, there is a ~1500 fold increase in the frequency of mutation events. Although early iterations of the system primarily showed transition mutations (C-T 75.5%, G-A 24.5%) were by far the most common, later improvements suggest a broader spectrum of mutations can be achieved by the adaptation of error-prone repair mechanisms.

I also present a novel DNA assembly method named BASIC, which was designed with automation in mind, in order to streamline construction and reiteration of expression plasmids, reporter devices and the numerous fusion proteins required throughout the project. BASIC has shown exceptional performance, achieving ~90% accuracy for assemblies containing 7 DNA fragments of varying lengths.

With principle mutator and construction methods in place, further work highlighted here describes the creation and characterisation of non-natural biological parts in the form of hybrid promoters/operators responsive to the inducer Tetracycline. I provide a set of variants which achieve a range of expression strengths with tight transition between uninduced (off) and induced (on) states. Estimations of relative promoter strength and critical inducer concentrations were incorporated into a deterministic model of the overall mutator circuit using a number of ordinary differential equations, including both my own parameterised kinetic functions and those taken from the literature. Although simplified, such computational models quickly revealed limitations of early circuit designs. This included the overly engineered solution to premature stoppage of the mutator device in the form of an AND gate.

In its current form, the mutator system will be used to select for novel forms of the LasR molecule which can bind or associate with a number of small AHL molecules required for quorum sensing activity and activation of the downstream transcriptional activation. More specifically this could be used to

screen for species specific variants, and future inclusion in biosensor circuits. From a wider perspective, the mutator system could be introduced alongside alternative display screening methods depending on the target protein and its functional activity to accelerate evolution through multiple rounds of mutation and selection. Furthermore, BASIC assembly has already been widely cited by the synthetic biology community and has been integrated into the work flow of industry collaborators. It is expected that the move towards complete automation of molecular biology lab work, and the increasing integration of liquid handling platforms across biotech, the approach will become increasingly popular.

Table of Contents

1	**Introduction**	**18**
1.1	Biological Diversity	18
1.2	Genetic Engineering	18
1.3	Synthetic Biology	21
1.4	Directed Evolution	22
1.4.1	*In Vitro* Approaches	24
1.4.2	Mixed *In Vitro* / *In Vivo* strategies for mutagenesis and selection	29
1.4.3	*In Vivo* mutagenesis	33
1.4.4	Simultaneous Positive and Negative Selection	40
1.5	Key Considerations	42
2	**Creating a Mutator Device**	**44**
2.1	Background	44
2.1.1	Activation Induced Deaminase - AID	44
2.1.2	*In vitro* studies of AID activity	46
2.1.3	*In vivo* studies of AID activity	47
2.1.4	AID target sequences	48
2.1.5	Models of AID activity and multimeric dimerisation	48
2.2	Testing AID Mutational Activity	50
2.3	Methodology	52
2.3.1	Plasmid Construction	52
2.3.2	Genome Integration	53
2.3.3	Host Strain Preparation	54
2.3.4	High-throughput Repeat Assay	55
2.4	Results and Discussion	55
2.4.1	Assessment of Strain Fitness and Viability	55
2.4.2	Targeted Mutation Assay	57
2.4.3	Multiple Rounds of Targeted Mutation	61
2.4.4	Untargeted mutation assay	63
2.4.5	Sequencing of GFP target	65
2.4.6	Sequencing off-target mutations in T7RNApol	70
2.5	Conclusions	73
3	**BASIC Assembly**	**77**
3.1	Background	77
3.1.1	DNA Assembly Standards	78
3.2	Materials and Method	80
3.2.1	Overview	80
3.2.2	Considerations of BASIC Linker Design	83

3.2.3	Standard Linkers	83
3.2.4	Hierarchical Assembly – Methylated Linkers	84
3.2.5	Creating Fusion Proteins- Fusion Linkers	85
3.2.6	Combinatorial Pathway Construction - RBS inclusive linkers	86
3.2.7	BASIC assembly Protocol	88
3.2.8	Automation on Liquid Handling Platform	90
3.3	Results and Discussion	90
3.3.1	General Assembly	90
3.3.2	Optimising the Protocol	91
3.3.3	Efficiency and accuracy	95
3.3.4	Hierarchical Assembly	97
3.3.5	Proof of Principle	99
3.3.6	Automated Pathway Construction	103
3.4	BASIC Compatible Genome Integration	104
3.4.1	Introduction	104
3.4.2	Methodology	105
3.4.3	Proof of Principle Experiment	107
3.5	Conclusions	110
4	**Building a Controllable Gene Circuit**	**111**
4.1	Background	112
4.1.1	Circuit Design	112
4.1.2	Adapting the Tetracycline Resistance Operon	114
4.1.3	LasR Target Protein	116
4.2	Methodology	117
4.2.1	Creating a Library of Tetracycline Repressible Promoters	117
4.2.2	pJ23101.TetOp Library Construction	120
4.2.3	Building the Tet Repressor and Output Module	121
4.2.4	Final Construction of Test Device	122
4.2.5	Manual Induction and Characterisation Assays	123
4.2.6	Automated Characterisation on a Liquid Handling Platform	123
4.2.7	Sequencing	125
4.3	Results and Discussion	125
4.3.1	Qualitative Assessment of p101.Tet Library	125
4.3.2	Quantitative Analysis of Strong ‘On’ State in P101.Tet HP Library	126
4.3.3	Qualitative Assessment of Tight ‘Off’ State in P101.Tet HP Library	135
4.4	Conclusions	136
5	**Circuit modelling**	**139**
5.1	Justification and Proposal	139

5.2	Methodology	140
5.3	Results and Discussion	141
5.3.1	Preliminary Modelling Efforts	141
5.3.2	Inclusion of p101Tet HP Parameters	142
5.4	Conclusions	143
5.4.1	Strength and Accuracy of Model	143
5.4.2	Requirement for Dual Selection in Directed Evolution	144
5.4.3	Directed Evolution of Complete Gene Circuits	145
6	**Improving the AID-T7 Mutator Device**	**146**
6.1	Introduction	146
6.2	Tuning expression of the AID-T7 mutator	147
6.2.1	Background	147
6.2.2	Justification and Proposal	147
6.2.3	Methodology	148
6.2.4	Results and Discussion	150
6.2.5	Conclusions	151
6.3	Improving the enzymatic relationship between AID and T7 RNA Pol	152
6.3.1	Background	152
6.3.2	Justification and Proposal	152
6.3.3	T7RNApol Activity	155
6.3.4	Methodology	160
6.3.5	Results and Discussion	161
6.3.6	Conclusions	164
7	**Increasing Mutagenic Spectrum and Diversity**	**165**
7.1	Background	166
7.1.1	Reactivating DNA repair	167
7.1.2	TLS DNA Polymerases and the SOS response	168
7.1.3	Understanding the link between AID Deamination and DNA Repair	170
7.2	Justification and Proposal	172
7.2.1	Polymerase IV	172
7.2.2	Pol V	175
7.2.3	Exonuclease III	177
7.2.4	DNA Pol I	177
7.3	Proposal and Experimental Strategy	178
7.4	Methodology	180
7.5	Results and Discussion	188
7.6	Conclusions	193
8	**Final Conclusions and Future Work**	**195**

8.1 Recent Advancements in the Field of Directed Evolution 195
8.2 Alternative Forms of the Mutator Device 198
8.3 Planned Improvements to the Mutator Device 199
8.4 Suggested Improvements to the Complete *In Vivo* Circuit 202
8.5 Closing Remarks 204
9 General Methods 206
9.1 DNA Purification 206
9.1.1 Phenol/Chlorophorm Extraction of Genomic DNA 206
9.1.2 PCR Purification 206
9.1.3 Magnetic Bead DNA purification 207
9.1.4 Gel Extraction 207
9.1.5 Plasmid DNA preparation 208
9.2 DNA Amplification 209
9.2.1 Polymerase Chain Reaction - Standard Conditions 209
9.2.2 Colony PCR 211
9.2.3 Genomic PCR 211
9.3 DNA modification 211
9.3.1 Site-Directed Mutagenesis 211
9.3.2 Klenow Extension 212
9.4 DNA Analysis 213
9.4.1 DNA size separation by Gel Electrophoresis 213
9.4.2 DNA Quantification by Spectrophotometry 213
9.5 DNA Assembly Methods 214
9.5.1 Gibson One-Step Isothermal Assembly 214
9.5.2 BASIC Assembly 217
9.5.3 Restriction Digest 217
9.5.4 Ligation 218
9.5.5 pJET 1.2 218
9.6 DNA Sequencing 219
9.7 DNA Synthesis 219
9.8 Protein Expression 219
9.8.1 *E.coli* expression 219
9.8.2 Cell Lysis 220
9.9 Protein Analysis 221
9.9.1 Protein Size Separation by SDS-PAGE 221
9.9.2 Coomassie Staining 222
9.9.3 Membrane Transfer 223
9.9.4 Western Blotting 224

9.10 Cell work 225
9.10.1 Preparing Competent Cells 225
9.10.2 Transformation 226
9.10.3 Antibiotic Selection 226
9.10.4 Growth Media 226
9.11 Other 227
9.11.1 OSIP Genome Integration 227
10 Bibliography 229
11 Appendices 244
11.1 Relating to Chapter 2 - Creating a Mutator Device 244
11.2 Relating to Chapter 3 - BASIC Assembly 251
11.3 Relating to Chapter 4 - Building a Controllable Gene Circuit 260
11.4 Relating to Chapter 6 - Improving the AID-T7 Mutator Device 264
11.5 Relating to Chapter 7 - Increasing Mutagenic Spectrum and Diversity 269
11.6 Relating to Chapter 8– Final Conclusions and Future Work 276
11.7 Other 279

List of Figures

Figure 1.1 – Compartmentalisation Self Replication – an *in vitro* method for cell free transcription and translation 26
Figure 1.2 – MAGE summary. 34
Figure 1.3 – PACE summary 35
Figure 1.4 – PACE chemostat control 36
Figure 1.5 – Variations of PACE. 38
Figure 2.1 – Proposed AID mode of action for deamination of cytosine nucleotide 45
Figure 2.2 – Inhibition of Base Excision Repair pathway encourages accumulation of C-T transition mutations. 49
Figure 2.3 – Proposed model for activity of the AID-T7 fusion protein during active transcription. 51
Figure 2.4 – My initial design for an *in vivo* system for directed evolution. 52
Figure 2.5 - Growth curves demonstrated the viability of expressing the mutator devices within the target E.coli host cell. 55
Figure 2.6 - Growth curves demonstrated the viability of expressing the mutator devices within the target E.coli GM31 host strain when co expressed with the repair inhibitor UGI. 56
Figure 2.7 – GM31 UGI+ strains were transformed with variants of the mutator device. 58
Figure 2.8 - Average percentage of GFP + cells after 5 rounds of mutation without variations of the mutator device 59
Figure 2.9 – Mutagenic effects of expressing the ATF mutator device with the GM31 UGI strain. 59
Figure 2.10 - Changing percentages of GFP+ cells within each culture across 5-day of sequential outgrowth. 62
Figure 2.11 - Multiple rounds of mutation were useful in highlighting a sudden drop off of GFP+ populations when expressing the A+T UN unfused mutator construct. 63
Figure 2.12 – Rifampicin reversion assays were used to show the number of reversion colonies within samples that had undergone 5 days outgrowth. 64
Figure 2.13 - A detailed mutation map provides a cumulative summary of the mutations identified within the genome integrated GFP target gene and flanking 5'UTR and 3'UTR regions.. 68
Figure 2.14 – Example sequence trace showing evidence of C-T transition mutations.. 69
Figure 2.15 – The average number of mutations in any single clone shows fairly wide deviation. 69
Figure 2.16 – Comparison of results from both targeted and untargeted mutation assays 75
Figure 3.1 - Prefix and Suffix sequences selected for use within the BASIC assembly format. 80
Figure 3.2 – General steps required for BASIC assembly 81
Figure 3.3 - Three major steps for 'linker based' BASIC assembly 82
Figure 3.4 –Standard Linkers during BASIC assembly. 84
Figure 3.5 – Methylated linkers used to regenerate Prefix and Suffix BsaI recognition sites ensuring that hierarchical idempotent cloning is possible in BASIC assembly. 85
Figure 3.6 - Fusion linkers used to add functionality to the BASIC linker region 86
Figure 3.7 – Fusion protein linker sequences which required *de novo* design strategy 86
Figure 3.8 – Functionalised RBS linkers which can be used for the combinatorial assembly of pathways 87

Figure 3.9 – A simple test construct assembled using the BASIC protocol which provided an immediate assessment of accuracy and efficiency 91
Figure 3.10 – Optimisation of the BASIC assembly protocol. 94
Figure 3.11 - DNA part digestion carried out during BASIC assembly 94
Figure 3.12 – Test constructs used to assess BASIC assembly efficiency and accuracy 95
Figure 3.13 – Test constructs used to assess BASIC assembly efficiency and accuracy including a second resistance marker 95
Figure 3.14 – Multi part assembly using the BASIC assembly standard 96
Figure 3.15 – Various fluorescent profiles seen in the initial test for BASIC assembly accuracy 97
Figure 3.16 – Two stage hierarchical assembly was used to demonstrate both efficiency and accuracy of using BASIC methylated linkers. 98
Figure 3.17 – Hierarchical assembly using the BASIC standard requires the application of methylated linkers 98
Figure 3.18 – BASIC Diagnostic Step 1 101
Figure 3.19 – BASIC Diagnostic Step 2 101
Figure 3.20 – BASIC Diagnostic Step 3 102
Figure 3.21 – SDS PAGE was used to show constitutive protein expression from a range of construct created using BASIC assembly, including a number of fusion proteins.. 103
Figure 3.22 - Preparation, assembly and integration steps for the BASIC pOSIP protocol 106
Figure 3.23 - Verification steps via PCR for BASIC pOSIP protocol 107
Figure 3.24 – Diagnostic colony PCR was used to determine whether selection colonies showed successful genome integration 108
Figure 3.25 – Histogram traces for single gene cassette in which a GFP gene was integrated into the Lambda locus in MG1665 *E.coli* strain. 109
Figure 3.26 - Histogram traces for double gene cassette in which both a GFP and RFP gene were assembled together and integrated into the Lambda locus in MG1665 simultaneously. 109
Figure 4.1 – Overall circuit to control directed evolution of a target protein 113
Figure 4.2 – A simple test construct used to characterise the library of hybrid pJ23101.TetOp promoters 119
Figure 4.3 – Auto fluorescence of blank *E.coli* strain MG1665 in Neinhardt MOPS media for samples grown over a 6 hour induction assay 128
Figure 4.4 – FL and OD curves for sample P101.TetHP1A across a 6 hour outgrowth at differing concentrations. 130
Figure 4.5 – Normalised FL values for p101.Tet.HP1A induced with 20ng/ml aTc and grown over a 6 hour time course. 131
Figure 4.6 – RPU for each P101.Tet library variant, plotted against increasing concentration of the aTc inducer.. 132
Figure 4.7 – Dose response curves for the complete set of P101.Tet.HP hybrid promoters. 134
Figure 4.8 – Mean levels of expression identified in un-induced in variants of P101Tet hybrid promoter library using flow cytometry 136
Figure 5.1 – Cell Designer manifestation of the complete circuit. Numbers correlate to individual kinetic laws applied to the deterministic model. 140

Figure 5.2 – Initial Cell Designer simulation excluding true P101TetHP hybrid promote parameters. 141
Figure 5.3 – Changeable levels of ATF protein as a direct result of different Hill coefficients (n) of P101TetHP..... 142
Figure 5.4 – Model simulation including parameters of the p101TetHP hybrid promoter set showing changing levels of the ATF mutator protein..... 143
Figure 6.1 – General arrangement of constructs use to test the toxicity of overexpression of the mutator protein 148
Figure 6.2 – Final 6-hour end point measurements for both OD (blue bars, left Y axis) and RFU (red bars, right Y axis) are shown constructs used to explore toxicity effects of overexpressing ATF mutator device. The data shown here corresponds to an average measurement across 3 repeats of the assays. Standard error bars are shown. The data presented corresponds to strongly induced cultures (250 ng/ul). Sample names reflect the constructs named in Table 6.1..... 150
Figure 6.3 – The differing fluorescent profiles of GM31 UGI+ strain expressing T7RNApol mutants 161
Figure 6.4 - The differing fluorescent profiles of GM31 UGI+ strain expressing ATF mutants..... 162
Figure 7.1 – Diagnostic Gel for Exo III – Pol IV constructs after immediate amplification of transformant colonies.. 183
Figure 7.2 – Diagnostic gel for 5’ 3’ Exo – Pol IV constructs after immediate amplification of transformant colonies.. 183
Figure 7.3 - Western Blot of ExoIII – Pol IV fusion proteins..... 184
Figure 7.4 – Hierarchical assembly of final library constructs using BASIC standard. 186
Figure 7.5 – The initial loss of function assay in which ATF fusion protein was co expressed with individual Exo III and Pol IV protein variants. 189
Figure 7.6 – Loss of function assay in which ATF fusion protein was co expressed with Exo III- Pol IV variants on a single plasmid 190
Figure 7.7 - Loss of function assay in which ATF fusion protein was co expressed with Exo III – Pol IV variants on a single plasmid. Selected variants of the Exo III Pol IV fusion protein contain a flexible linker. 191
Figure 7.8 - Loss of function assay in which ATF fusion protein was co expressed with Exo III – Pol IV variants on a single plasmid. Selected variants of the Exo III Pol IV fusion protein contain a rigid linker. 191
Figure 7.9 – Evidence of a frame shift mutation within the third codon of the GFP cds which creates in a false stop codon 192
Figure 8.1 – Mutagenesis plasmid created for general in vivo mutagenesis with *E.coli* host strains..... 196
Figure 8.2 – Proposed system for improved gain of function assay to enable selection of variants of the mutator capable of different types of mutations..... 200
Figure 11.1 – Examples test plates from Rifampicin reversion assays. Images of each plate were analysed using colony counting software..... 246
Figure 11.2 – Example gel for colony PCR used to amplified variants of the target gene – post mutation..... 247
Figure 11.3 - Complete sequence map of T7RNApolymerase gene, highlighting off target mutations incurred within the AID + T7 unfused mutator device. 248

Figure 11.4 – Assembly of single and double gene cassettes using the standard BASIC assembly protocol into the BASIC.pOSIP KL vector 257
Figure 11.5 – sfGFP expression from pJ101.Tet characterisation constructs 260
Figure 11.6 – RFU expression profiles for HP1 variants of the J101.TetOp hybrid promoter across a 4 hour preliminary induction assay 260
Figure 11.7 - RFU expression profiles for HP2 variants of the J101.TetOp hybrid promoter across a 4 hour preliminary induction assay 261
Figure 11.8 - RFU expression profiles for HP3 variants of the J101.TetOp hybrid promoter across a 4 hour preliminary induction assay 261
Figure 11.9 – Absorbance measurements taken at 15 min intervals across 6 hours outgrowth for each sample were used to validate whether or not there was expected logistic growth of the MG1665 strain (lack of sigmoidal shape). 262
Figure 11.10 – Diagnostic PCR products to confirm the correct assembly of constructs used to test the toxicity of the AIDT7 mutator protein 265
Figure 11.11 – Plate layout used for toxicity induction assay 265
Figure 11.12 – Expressing T7RNApol mutants in GM31 UGI results in varying % GFP+ cell populations after 3 days successive outgrowth 267
Figure 11.13 – Expressing ATF mutants in GM31 UGI results in varying % GFP+ cell populations after 3 days successive outgrowth. 267
Figure 11.14 – Variation in fluorescent profiles of T7RNApol and ATF variants, when expressed in GM31 UGI+ strain across 3 consecutive days, during the loss of function assay. 268
Figure 11.15 – Recovery of antibiotic resistance in a preliminary gain of function assay used to explore the effect of expression error-prone ExoIII-Pol IV variants alongside the ATF mutator device 276
Figure 11.16 – Automated G/C optimisation of target genes to increase number of sites for deamination as points of entry for the ATF mutator 276
Figure 11.17 – Tuning protein expression using modular RBS parts using the BASIC assembly standard 277
Figure 11.18 - Tuning protein expression using functional RBS linkers using the BASIC assembly standard 278
Figure 11.19 – Final plasmid maps of AID expression vectors engineered to improve downstream purification. .. 280

List of Tables

Table 1.1 – Directed Protein Evolution Strategies 24

Table 2.1 – Investigations into AID activity 46

Table 2.2 –Plasmid constructs used to test the activity of the proposed mutator device 53

Table 2.3 – Description of residue changes identified with target GFP gene 70

Table 2.4 – Identified mutations in T7RNApolymerase and corresponding residue changes in the T7RNApol gene in the AID+T7 unfused device. Silent mutations do not affect T7 transcriptional activity. All other mutations may potentially cause negative affects to transcriptional efficiency. Starred mutations were seen in all GFP negative substitutions. 72

Table 3.1- Variables tested during the development of the BASIC assembly standard as exists today. Each independent variable was changed in isolation and efficiency and accuracy were assessed using the standard 4-part assembly mentioned previously (RFP-GFP-Cam-pMB1KAN). 93

Table 3.2 - Composite parts required for the parallel assembly of the 6 constructs variants with the additional linkers required. Note that the PX and SX linkers change depending on the final intended construct and thus the number of total parts required for this proof of principle assembly quickly increases. 100

Table 3.3 – Analysis of PCR product banding patterns used for BASIC – OSIP diagnostic 107

Table 5.1 – Summary of the parameters included in the latest iteration of the Cell Designer model 140

Table 6.1 – Constructs built in order to test whether high expression of AID T7RNApolymerase can increase the frequency of mutation activity or whether it leads to cell toxicity. 149

Table 6.2 – Functional Domains of T7RNApol 156

Table 6.3 – Single position T7RNApol mutants - T7RNApol mutants and single position mutation affects are described. Many of these mutations have actually been identified in a similar in vivo direct evolution system in which T7RNAPol was the target protein (Esvelt, 2011). ^ mutation lost in highly stringent conditions, # include mutations identified in Temme and Voigt, 2012. 157

Table 6.4 – Activity of T7RNA polymerase mutants measured as b-galactosidase yield generated from expression of lacZ under as synthesised by various mutant T7RNApol. In vitro speed and processivity of each variant were published in the work by Bonner et al, 1994. Adapted from Makarova et al, 1995. 158

Table 6.5 - Primers used to construct a number of T7RNApol mutants by Site-Directed Mutagenesis 160

Table 7.1 – Description of the B and Y family of DNA polymerases particularly relevant to repair in Eukaryotes, with comparative polymerases in prokaryotes (E.coli) provided. 170

Table 7.2 – Comparison of the Pol IV and Pol V and their applicability to this work 180

Table 7.3 – Summary of the constructs built to test the activity of variants of the Exo – Pol fusion proteins when co-expressed alongside the known ATF mutator fusion protein on a single expression vector, within E.coli GM31 host 187

Table 11.1 – Complete data set from 'loss of function' GFP knockout mutator assay. 244

Table 11.2 – Average data sets across a 5 day loss of function mutator assay. Flow cytometry was used to count at least 5000 true E.coli events and gated into GFP+ and GFP- populations. Standard deviation and standard errors are shown. Note that AID T7 unfused constructs shows almost a completely non GFP population by day 5. This is actually in response to critical T7RNApol mutations that create a selection bias, and thus propagate as a consequence of reduced burden on the cell. 245

Table 11.3- All possible C-T mutations and corresponding residue changes. This shows a limited spectrum of diversity. Complete diversity would equate to 400 potential amino acid changes (20x20) not including STOP codons. In this system, only 41 of 400 are possible. The probability that a single deamination event will result in an active mutation is 70.69% a in an active mutation. 249

Table 11.4 – Synthesised oligonucleotides which form the standard linkers used in BASIC assembly. GC content and predicted linker annealing melting temperatures are given. 251

Table 11.5 - Sequences of synthesised oligonucleotides for the creation of methylated linkers used in BASIC assembly for idempotent cloning. GC content and predicted linker annealing TM is given (calculated by IDT Oligo Analyzer). Key: Adapter sites, Adapter Oligonucleotides containing Methylated Cytosine base, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region. 252

Table 11.6 - Sequences of synthesised oligonucleotides required for fusion linkers used to created fusion proteins via the BASIC assembly method. GC content and predicted linker annealing TM is given (calculated by IDT Oligo Analyzer). Key: Adapter sites, Adapter Oligonucleotides, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region. 252

Table 11.7 - Sequences of synthesised oligonucleotides required for generation of RBS linkers. 3 linker overlap regions were generated, with the specific RBS of interest placed within the Adapter Oligo site. Key: Adapter sites, Adapter Oligonucleotides, RBS containing Oligos, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region. 253

Table 11.8 - Primers were designed so that internal linker sequences acted as priming sites for both Forward and Reverse primers. Considerations were made to ensure that all primers had a similar melting temperature (Tm) any thus theoretically any two could be chosen to amplify across a functional part. PCR product bands could then be compared to predicted band size for that specific construct. 254

Table 11.9 – Raw colony counts from BASIC assembly tests for accuracy and efficiency 255

Table 11.10 – Calculation of the average correct and incorrect colonies across 4 repeats of BASIC assembly test for efficiency 256

Table 11.11 - Calculation of the average correct and incorrect colonies across 4 repeats of BASIC assembly test for efficiency 256

Table 11.12 - Raw colony counts from BASIC assembly tests for accuracy and efficiency when using methylated linkers in hierarchical assembly 256

Table 11.13 - Calculation of the average correct and incorrect colonies across 3 repeats of BASIC assembly test for efficiency and accuracy of hierarchical assembly. 257
Table 11.14 – Improved PCR primers for colony PCR diagnostics check of genome integrates produced using the BASIC-OSIP method described in Chapter 3.4 258
Table 11.15 – Description of target insert positions, the loci are governed by known phage attachment sites with the E.coli MG1665 genome. 258
Table 11.16 - A description of the BASIC compatible parts required for BASIC assembly of genome integration vectors within a proof of principle for the BASIC-OSIP standard 258
Table 11.17 – A description of the constructs used in the proof of principle experiment to determine efficiency and usability of the BASIC-OSIP standard 259
Table 11.18 – Final RPU values generated for p101.Tet Hybrid Promoter Characterisation assays and varying concentrations of inducer in order to formulate a does response curve for each variant within the library. Data collection required an automated characterisation platform and analysis tools. 4 repeats of the assay were carried out, shown here is the mean average of the run along with standard deviation and standard errors. 263
Table 11.19- Part required for build a range of test constructs via BASIC assembly to test toxicity of the AIDT7 mutator protein. 264
Table 11.20 – Description of each test construct used to test toxicity of the AIDT7 mutator protein 264
Table 11.21 – Summary of final OD 600nm readings from toxicity induction assay. Samples are normalised against blank media control with standard errors across 3 repeats (6x independent samples for each). 265
Table 11.22 - Summary of final RFU readings from toxicity induction assay. Note that RFP outputs within operon – like constructs did not produce a fluorescent output. 266
Table 11.23 – Raw data collected generated from gating changing fluorescent profiles of various T7RNApol and ATF mutants during loss of function assay. Data shown reflects the final data point collected after 3 days of subsequent outgrowth. 266
Table 11.24 - Primers used to amplify target genes of interest from the wildtype MG1665 E.coli strain 269
Table 11.25 -Primers used to create BASIC compatible parts BASIC Prefix, BASIC Suffix, RBS34, Annealing region 269
Table 11.26 - Primer sequences required for addition of FLAG tag sequence to both ExoIII and 3-5’ Exo BASIC parts. FLAG Tag sequences are shown in red and primer annealing sequences are underlined. 270
Table 11.27 – Initially plan for combinatorial assembly of ExoIII - Pol IV variants using BASIC assembly, detailing Linker addition and size of the each individual part 270
Table 11.28 - Initially plan for combinatorial assembly of 5’ 3’ Exo - Pol IV variants using BASIC assembly, detailing Linker addition and size of the each individual part 271
Table 11.29 – Summary of Stage 1 constructs used to determine the effects of expressing Exo – Pol fusion proteins. 272

Table 11.30 – Summary of the Stage 2 Constructs used to explore the error-prone repair activity when co expressing AID-T7 and Exo-Pol IV fusion proteins 273

Timeline

Research Focus	Year 1				Year 2				Year 3			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Creating a Mutator Device (Chapter 2)												
BASIC Assembly (Chapter 3)												
Building a Controllable Gene Circuit (Chapter 4)												
Improving the AID-T7 mutator Device (Chapter 6)												
Increasing Mutational Spectrum/Diversity (Chapter 7)												

List of Abbreviations

AID	Activation Induced Deaminase
APOBEC	Apolipoprotein B editing catalytic
CDS	Coding Sequence
CPR	Compartmentalized partnered replication
CSR	Compartmentalisation self-replication
CSR	Class Switch Recombination
ePCR	Error-prone Polymerase Chain Reaction
GFP	Green Fluorescence Protein
HP	Hybrid Promoter
HP	Hybrid Promoter
ITS	Initially Transcribed Sequences
MAGE	Multiplex Genome Evolution
Op	Operator
PACE	Phage Assisted Continuous Evolution
RPR	Random Priming Recombination
SELEX	Systematic evolution of ligands by exponential enrichment
sfGFP	Super Folder Green Fluorescence Protein
SynBio	Synthetic Biology
T7RNApol	T7 RNA Polymerase
Tet	Tetracycline
TLS	Trans Lesion Synthesis
TSSM	Targeted site directed mutagenesis

Word Count – 73,014

1 Introduction

1.1 Biological Diversity

The vast and overwhelming level of genetic diversity that exists throughout the natural world is a direct consequence of Darwinian evolution. Selection pressures exist not only at the species level but down at the cellular and molecular levels. Even seemingly insignificant nuances can constitute an advantageous genotype within the 'fittest' organism, with the smallest external selection pressure influencing evolutionary change.

At its most fundamental, evolution is driven by mutagenic changes of nucleotide sequences. Whilst common usage of the term might suggest that a mutation is undesirable or detrimental to fitness, given that it suggests a movement away from the norm, such changes can very often be beneficial. Unquestionably, the short-term survival of the individual organism requires strategies to limit or repair mutations prior to DNA replication. However, the small chance for favourable change is actually to the benefit of the species as a whole, enabling progressive steps of adaptation and optimisation. In highly pressured conditions, change is accelerated through multiple rounds of replication under strong selection, which ultimately dictates which variants work best or indeed survive.

In mammals we often consider evolution to be a very slow process, extended over hundreds of thousands of years and multiple generations, given long periods of species replication. However, if we consider a unicellular species, we begin to observe the frequent mutagenic change which is allowed to propagate quickly as a consequence of short replication time. In prokaryotes, cell doubling takes just minutes, meaning that evolution is highly accelerated and the concept of directed (or molecular) evolution is conceivable. Here the underlying aim is to use the fast-paced replication of prokaryotes to quickly identify advantageous mutations, which have resulted in a desirable phenotypic change. In its most potent form, directed evolution seeks to actively drive, and yet control, mutations within a target organism or localised to a specific gene of interest, in an effort to rapidly reach the most favourable outcome.

1.2 Genetic Engineering

From a traditional stand point, genetic engineering attempts to apply the incredibly vast array of evolved genes found within nature and apply them to alternative contexts or within alternative environments. This may include wholesale changes to an organism via the introduction of exogenous genes, or by more subtle manipulation of the target gene sequence itself. Along with advances in molecular biology techniques and the emergence of central concepts underpinning synthetic and

systems biology, there is now a realisation that this wealth of naturally occurring sequence data can be exploited and a drive to create “made-to-fit” genetic elements that perform very specific functions.

It is first important to underline the differences between classical genetic engineering and natural evolution. Many current engineering techniques aim to create vast libraries of variants of a particular target protein, influenced in design by either the structural or phylogenetic information available, before performing wide reaching screens in order to identify the most useful mutant. One such example is the continuing development of SCOPE, Structure Based Combinatorial Protein Engineering, which aims to create gene libraries of defined composition with the design based on the structural and probabilistic constraints of the encoded proteins (*O’Maille et al, 2002*).

Darwinian evolution does not necessarily create libraries with the same level of diversity, yet by expanding the search over very long periods of time, it ensures that ‘fit for purpose’ mutant variants are allowed to propagate due to the presence of an overarching selection pressure.

Protein evolution can follow both formats i.e. desirable changes can be guided with or without prior knowledge. Although the final goal of this project is to construct a system for accelerated molecular evolution, it is not completely necessary to apply a selection pressure in order to reach the desirable outcome, if the system has the capacity to very quickly screen a library of extended diversity. Therefore, the ideas of directing and accelerating evolution of the target protein are often used interchangeably within this work.

Rational protein engineering remains a long and laborious task. Attempts to carefully engineer proteins using ‘informed’ and predictive models are increasingly met with failure. Whether guided by homologous alignments, prior structural understanding, interaction models or complex *in silco* build algorithms, attempts to predict the consequence of single residue changes are very difficult. In many cases, large assumptions about structure function relationships or specific and non-specific interactions must be made, whilst these approaches require extensive computing power (*Simons et al. 1999; Xu et al. 2003; Soding et al. 2005*).

Furthermore, highly informed protein engineering, regardless of which mutational techniques are employed, often only introduce changes to very well understood regions of the target protein (i.e. functional domains, active sites) and will rely on low frequency, stepwise substitutions at well-defined locations. Thus, the engineer is often limited to making precise and highly specific changes to domains of interest, or small target areas often described as ‘local fitness peaks’, because they are more likely to produce a viable, active and selectable phenotype. Changes centred at these positions will be less likely to result in major phenotypic differences. This is a fundamental weakness to using a top-down approach to protein engineering, in which a ‘close to desirable’ target protein is a designated starting point upon which to layer subsequent changes. However, it is very unlikely that mutant variants will be able to

escape the tight constraints placed upon the wild-type protein, given that it will itself be a highly evolved variant. It will have already undergone numerous changes driven by strong selection pressures in order to work or function to the optimal level required within the normal environmental conditions. Undoubtedly, any pre-evolved protein consists of core structural foundations essential for function, correct folding, activity, expression, solubility and stability. Therefore, there is a miniscule probability that an accumulation of mutations will be tolerated, with the majority of changes resulting in reduced function under the same or pre-existing environmental selection pressures. Nevertheless, the ability to disconnect target protein function from cellular environmental context could provide the opportunity to independently evolve new variants in a beneficial direction.

To truly understand this idea, we can consider that at a very early stage in the evolutionary pathway of a natural protein, there is a high probability that an external pressure will cause a divergent or fundamental change, meaning all future variants possess that particular fitness advantage. If the same selection pressures exist, subsequent reversion of this beneficial change is very doubtful. This dramatically limits the ability for protein redesign. Alternatively, early changes could result in two distinct proteins which shift toward totally unrelated functions, offering two different solutions and general fitness advantage. This is reflected in the divergence seen within evolutionary trees of associated organisms that are quite distinct but still relate back to a common ancestor. It is equally feasible that the early protein framework, in place before this major shift, could form a fixed basis of protein function, thus resulting in two orthogonal proteins which have similar, yet separate, functions.

This design problem is confounded in many current protein engineering approaches, which often employ positive selection and have a limited tolerance for periods of regression, where suboptimal protein function is not permitted. Detrimental changes to fundamental sequence cannot be selected for without a means for negative selection, and therefore we are unlikely to see a significant shift away from fitness peaks i.e. the system does not have the opportunity to move from functional protein, to non-functional and back to a functional one once again. Given that a more complete search through broad and diverse sequence space may result in large scale structural changes to the protein backbone, permitting negative changes would be helpful to identify truly novel and distinct protein functionality. In order to achieve this, the system outlined here will allow for continual changes, which not be solely driven by positive selection, but will be allowed to progress through highly divergent periods of mutation and undergo continual screening until a desirable protein function is identified and the endpoint is reached.

Another limitation placed upon widely used protein engineering methods is the overall screening capacity of the selection assay. Given that the challenge of understanding protein structure and function relationships is as infinitely complex as the size of the nucleotide and amino acids sequence

space which defines them, small subsets of changes are unlikely to improve function. If the screening capacity is small, then a library of variants containing more focused and rational changes is essential, but this would again require some form of prior knowledge. By carrying out the mutation and selection phases within a completely self-contained *in vivo* device, my system should overcome this limitation by screening variants in a continuous manner, only limited by cell growth, doubling rate and maximal culture density.

Another major weakness of existing protein engineering methods is the lack of consideration for how the target protein will act within a true cellular context, as many favour *in vitro* screening methods given their larger screening capacity. This means that selected proteins are often evaluated in complete isolation from any external factor. In reality, the intracellular environment has a huge bearing on protein expression, folding and function and therefore it is very difficult to expect selected protein variants to work as first intended. Importantly, my system will be a completely *in vivo* device, ensuring selection of viable protein variants under biologically relevant conditions e.g. pH, temperature, interference from metabolic by-products.

Finally, if we consider for a moment, the amount of sequence variation possible in just a 3 kilobase nucleotide sequence; 3000 base pairs with 4 nucleotide possibilities at every position leading to 4^{3000} DNA variants. This results in an explosion of variants at the amino acid level, meaning that the corresponding 1000 amino acid polypeptide (with 20 amino acids possible at every position) means there is a possible 20^{1000} protein variants. Unsurprisingly, this complete sequence space can never be fully explored due to the sheer size of the task involved, but the key point is that a directed evolution system, which has the ability to introduce any and all nucleotide changes at defined positions of the target gene (resulting in any amino acid change) is essential for true exploration of sequence space.

For all of these reasons, rational protein engineering is often not fit for purpose, as it only allows for a very limited and narrow search at defined fitness peaks.

Accelerating evolution within an *in vivo* device would carry clear and major advantages; proteins with poorly understood function can be directed away from current local fitness peaks in a completely random nature, a far larger number of variants can be screened directly, and the selection of non-viable false positives can be avoided when evolution is carried out in a true cellular environment.

1.3 Synthetic Biology

Synthetic biology (SynBio) can be considered a sub-division within the wider field of genetic engineering, fuelled by extensive DNA sequencing efforts in recent decades and now driven onwards by the falling cost of DNA synthesis (*Carlson, 2009*). Its exact definition remains a topic of great debate, given that it is an amalgamation of concepts, skills and techniques from cell biology, molecular biology,

systems biology, structural protein engineering, bioinformatics and mathematics. However, in my opinion it should be considered as an attempt to reprogram biology and build toward completely novel and non-natural functions, many of which have been detailed in countless reviews from the field (*Marguet et al, 2007., Church et al, 2014*).

SynBio attempts to move past the limitations and constraints of the natural world but in a way that attempts to apply rational standardisation to biology. Extensive characterisation of individual components is performed in preparation for creating complex multi-part devices, which can then be brought together to build highly intelligent systems. By introducing engineering principles such as the concept of logic gates constructed using a modular parts-based approach, the belief is that the researcher will be able to quickly re-form, carefully control and reliably predict the outcomes of complex biological systems (*Singh, 2014*, *Andrianantoandro et al 2006*).

Advances in fundamental molecular biology have pushed the field way beyond classical methods of genetic modification (selective breeding, guided gene transfer) and even beyond traditional restriction digestion/ ligation cloning, given our improved capability to write and assemble DNA. This has helped in the creation of a whole range of exciting systems including those which; contain an expanded version of the genetic code (to incorporate non-natural amino acids) (*Wang et al 2001, Chin et al 2003*), complex logic gates, programmable switches in bacteria (*Collins et al, 2001*), yeast (*Becksei et al, 2001*) and mammalian systems (*Kramer and Fusenegger, 2005*) or are in fact completely synthetic organisms (*Gibson et al, 2010*) or minimal viable cell chassis (*Hutchinson et al, 2016*). It is believed that such unprecedented control over biology will allow us to rewire organisms to produce valuable end products in the form of renewable fuels, greener chemical alternatives, advanced biomaterials, foodstuffs and pharmaceuticals or therapies.

The SynBio community continues to expand the toolkit of characterised biological parts, which include genetic elements (stretches of DNA sequence) in the form of promoters, ribosomal binding sites, transcriptional enhancers or repressors as well as manipulation of gene coding sequences (CDS). Often these sequences are well understood as independent entities and with the opportunity to link genotype with phenotypic function.

It is these principles that I have sought to employ throughout my work as I believe they provide the best opportunity to build an entirely programmable, predictable, robust and autonomous biological system capable of carrying out directed evolution.

1.4 Directed Evolution

The concept of directed evolution actually moves somewhat counter to the standardised parts-based approach engrained in the ideologies of synthetic biology. The complexity and stochasticity seen

in biology means that even well-defined components are unlikely to function as intended in many scenarios, particularly when introduced into larger systems or transferred between different host organisms. The community's current list of characterised genetic elements remains small and has been extended out from a limited subset of parts chosen on the basis of their original function. Therefore, it is highly unlikely that any single one of these will perfectly meet all desirable characteristics.

In the context of this report, directed evolution is considered as the mutations incurred within a protein coding sequence i.e. gene, which result in the evolution of novel protein function, rather than changes to a non-protein coding stretches of DNA which control transcriptional expression and translation or indeed the amount of said protein produced within the system. Numerous industrially relevant proteins, such as therapeutics (*Jespers et al, 1994*), biochemical enzymes (*Coelho et al, 2013*) or reagents (*McIssac et al, 2014*) have been evolved in this way, for a wide range of end applications e.g. the production of medications, fine chemicals and fragrances. This serves to highlight the importance of directed evolution as a technique which is widely used across biotechnology, healthcare, agriculture and manufacturing.

Unsurprisingly, there are many proposed systems for directed evolution, developed to work either *in vivo*, *in vitro* or by combining a mixture of both approaches. However, none of them offer the perfect solution or fulfil all the necessary requirements needed for optimal protein selection. Only the most recent advances demonstrate our ability to construct intelligent and self-controlling systems. For example, recent work shows that it was possible to control the rate of mutation using a circuit capable of feedback regulated evolution of phenotype in the form of increase lycopene production (*Chou & Keasling, 2013*). However, no biological circuit is without limitation and will require careful adaptation and fine tuning depending on the particular target molecule.

In brief, *in vitro* systems have a large screening capacity, but cannot select for the desirable protein structure/function within a cellular environment. Alternative *in vivo* systems employ widespread mutator phenotypes and introduce untargeted random mutations which may weaken the cell or obscure true positive selection.

Although the application of directed evolution as a powerful engineering tool has been outlined in a number of recent review articles (*Zhao & Zha, 2004., Dougherty & Arnold 2009, Blagodatski & Katanaev 2010, Cobb et al., 2012, Cobb et al., 2013*) my initial discussions attempt to contextualise my own work in relation to published methods, whilst offering a detailed critique of current gold standards (***Table 1.1***).

I also suggest potential ways to overcome existing limitations and improve the accessibility and usability of these approaches. The potential benefits of a complete *in vivo* system outlined herein are

significant, as it could provide the opportunity to engineer a wide range of target proteins without having deep understanding of protein structure or function.

Table 1.1 *– Directed Protein Evolution Strategies*

Method	Context	Diversity	Control	Screening Capacity	Cost	Adaptability	Source
Self-replicating RNA	*In Vitro*	High	Low	High	Low	Low	Mills et al, 1967
SELEX - Aptamers	*In Vitro*	High	Low	High	Medium	Low	Beaudry & Joyce, 1992
Oligo Primer Randomisation	*In Vitro*	Low	Medium	High	Low	High	Shao et al, 1998
CSR	*In Vitro*	High	Low	High	Medium	Medium	Tawfik & Griffiths 1998 Miller et al, 2006 Cohen et al, 2004 Ghadessy et al, 2001
CPR	*In Vitro*	High	Low	High	Medium		Ellefson et al, 2014
epPCR - Error-prone PCR	*In Vitro* / Mixed	High	Medium	Low	Low	High	Eckert & Kunkel, 1991 Cadwell & Joyce, 1992
TSSM – Targeted site-directed mutagenesis	*In Vitro* / Mixed	Medium	High	Low	Low	High	Nair & Zhao, 2008
Overlap Extension PCR	*In Vitro* / Mixed	Medium	High	Low	Low	High	Nair & Zhao, 2008
DNA Shuffling	*In Vitro* / Mixed	Low	Low	Low	Low	High	Stemmer et al, 1994 Crameri et al, 1996 Zhao & Arnold 1997
Synthetic Libraries	*In Vitro* / Mixed	High	High	Low	High	High	Gen9 2016 LabGenius 2016
Mutagenic Strains	*In Vivo*	Medium	Low	High	Low	High	Greener et al, 1997
MAGE	*In Vivo*	Medium	Medium	High	High	Medium	Wang et al, 2009
PACE	*In Vivo*	Medium	Medium	High	Medium	Medium	Esvelt et al, 2011 Carlson et al, 2014 Dickinson et al, 2014
This Work	*In Vivo*	?	Medium	High	Low	?	Mackrow, 2016

1.4.1 *In Vitro* Approaches

True *in vitro* systems attempt to alter target DNA in complete isolation from any cellular context. The main benefits are that potent mutagens can be used to introduce a high frequency of mutations without risking off-target changes that could be to the detriment to the host cell. This means that vast libraries can be generated quickly and efficiently before immediate screening.

In vitro methods are able to introduce mutations in a variety of ways, such as using chemical mutagenesis (*Botstein and Shortle, 1985*), incorporating mutagenic nucleotides (*Wells et al, 1985*) or replication of DNA using low fidelity polymerases (*Zhou et al, 1991*). It should be noted that *in vitro* methods offer a more useful solution when targeting non-protein coding target DNA/RNA sequences which possess protein like functions i.e. those which have the capacity to affect other biomolecules. This is due to a clear associated link between genotype and phenotype meaning that as the evolutionary process continues, subsequent rounds of replication allow for the propagation of desirable

variants, particularly if a selection pressure is applied. Both DNA and RNA can have enzyme-like capabilities, without the need for protein translation, as they are able to form complex folded structures which can bind, interact and then affect their own starting nucleotide sequence e.g. cleavage, ligation, peptide bond formation. Consider for a moment the RNA world hypothesis (*Bernhardt, 2012*). As such, the very first example of directed evolution at a molecular level was carried out on a self-replicating RNA, where RNA specific RNA polymerase replicated RNA viruses and results in a loss of non-essential sequence (*Mills et al, 1967*). Subsequently, this led the way for research into the evolution of functional nucleotides such as ribozymes, oligonucleotide aptamers and riboswitches. Extensive libraries can be screened for binding activity to a chosen target, using a completely *in vitro* method entitled the systematic evolution of ligands by exponential enrichment or SELEX (*Tuerk and Gold, 1990., Gold et al, 1995., Hager et al, 1996*). Furthermore, there can be multiple rounds of mutation and selection using error-prone reverse transcription.

In vitro screening of nucleic acid variants is often linked to binding affinity for a pre-selected target ligand, before subsequent retrieval of successful binders (*Ellington and Szostak, 1990*), or indeed the ability to carry a desired catalytic reaction (*Robertson and Joyce, 1990*). One such example describes the directed evolution of an RNA enzyme to cleave DNA under desirable conditions (*Beaudry & Joyce, 1992*). The authors were able to carry out a cyclic amplification of RNA templates by using a reverse transcriptase to convert back to cDNA, before using T7 RNA polymerase to transcribe RNA. Only in the presence of a catalytically active RNA ribozyme could transcription occur, given that cleavage of a phosphodiester bond within an ssDNA oligonucleotide substrate was needed to produce a specific primer required for reverse transcription. Therefore, during multiple rounds of the two-step process of converting RNA to cDNA and back to RNA, there is a propagation of more active variants of the ribozyme.

Given that the proposal here is to use directed evolution to create protein variants, a truly cell free system would require active transcription and translation machinery in order to convert template DNA into its corresponding and functional protein. One such 'true' *in vitro* approach, developed by Tawfik and colleagues, creates a cell-free environment, named compartmentalisation self-replication (CSR). This utilises cell-like aqueous droplets which hold a cell extract containing all the necessary molecular machinery (*Tawfik & Griffiths, 1998*).

Each droplet maintains individuality, is self-contained and becomes a completely unique 'pot' for selection with the freedom to move along its own evolutionary path. Firstly, a large random mutagenesis library is produced and transferred into the cell-like compartments. Genes are then transcribed and translated within their respective compartments. In this case, the screening identified variants of a DNA HaeIII methyltransferase produced inside the compartment which were able to

methylate the DNA contained within it. The compartments are then broken with the genes linked to production of the enzyme easily selected for, as they were now resistance to a subsequent round of digestion by the HaeIII endonuclease. Useful variants were then amplified and sequenced to reveal genotype (***Figure 1.1***)

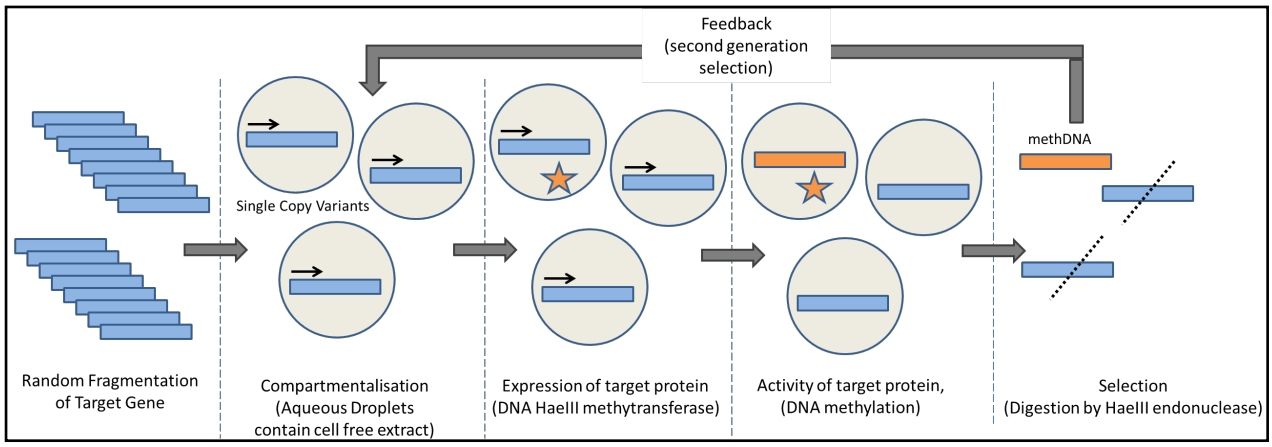


Figure 1.1 *– Compartmentalisation Self Replication – an in vitro method for cell free transcription and translation. In this example, only genes which produce viable, which are able to then methylate the starting DNA template sequence, can avoid digested by endonuclease and are then transferred to the next round. Separating single variants into individual compartments enables a direct link between genotype and phenotype.*

This approach has been extended to use microbeads encapsulated by water-oil emulsions. In this instance, the protein is then exposed to a soluble substrate and catalytic turnover rate is recorded. This enabled the development of a bacterial phosphotriesterase capable of very fast turnover and enzymatic activity (*Griffiths & Tawfik, 2003*). Interestingly, this enzyme is thought to have a very a recent evolutionary history that reflects the timeline of widespread use of organophosphate pesticides, which may go some way explain why a dramatically improved version of the enzyme (63 times increase in turnover rate), could quickly be identified within accelerated and directed evolutionary conditions.

Similar approaches have also now been used to select for new function in restriction endonucleases and DNA polymerases (*Miller et al, 2006., Cohen et al, 2004., Doi et al, 2004., Ghadessy et al, 2001*). As mentioned, the major advantage of CSR is that selection is not just based on binding affinity with the target but can also select for improved catalytic or regulatory activity. Alternative applications have been comprehensively reviewed by the group leading in this area (*Ghadessy et al, 2007*).

A very recent improvement to the CSR protocol claims to avoid unintended selective pressures which have the potential to push the evolutionary trajectory in the wrong direction. Compartmentalized partnered replication (CPR) links genetic parts or complete synthetic circuits with the production of Taq DNA polymerase. In doing so, improved variants of Taq DNA pol gene can easily be identified because protein variants which result in a high rate of replication produce multiple copies of the starting gene sequence. This means that they can propagate in subsequent generations, as a consequence of

increased efficiency during an *in vitro* PCR step and this provides a clear link between genotype and phenotype (*Ellefson et al, 2014*). The mixture of PCR mutant products is transferred back into individual compartments for further rounds of selection. The ability to select for different protein functions by coupling them to Taq polymerase production provides an advantage over earlier CSR methods. Prior to this, target genes were those inherently linked to transcription i.e. transcription factors.

Interestingly, Ellefson and colleagues share similar goals to my own as their system attempts to create orthogonal promoters by evolving a T7 RNA polymerase variant which could recognise orthologous promoter sequences with high activity and specificity. When purposefully mutating the specificity loop of the standard T7RNApol (positions 746, 747, 748 and 756, 757, 758) and subjecting the gene to rounds of mutation, they saw a reversion to the wildtype sequence after selection. Once again, the promoter sequence was changed and once again the polymerase gene was mutated and although comparable, none of these mutants achieved a greater transcriptional activity than the standard T7 promoter/polymerase pair. This would suggest that the existing T7RNApol has underlying and highly evolved structural features providing maximal transcriptional activity, but these well-defined domains also place constraints on the enzyme and given that the system did not include a means of negative selection or allow for period of fitness regression, a variant with increase processivity remained elusive.

CPR was also used to engineer a tRNA synthetase: tRNA pair from *S.cerevisiae* to overcome the amber stop codon by incorporation of an unnatural amino acid, 5-hydroxy-L-tryptophan (5HTP). The selection method in this example had to be altered, with the Taq pol gene engineered so that it contained an amber stop codon at a fixed position. Only tRNA mutants capable of overcoming this stop codon would produce viable Taq pol and could propagate following the *in vitro* PCR step, which would also be useful for novel amino acid incorporation. It does appear that this method was very effective in generating new and exciting mutants for this niche purpose. The main advantages of CPR are the large screening capacity and the flexibility provided by of interval periods without active evolution to allow the opportunity for expression and identification of useful mutants. Its real usability is questionable and will depend on how easily the *in vivo* circuit can be rewired to link alternative protein function (such as enzymes) to its inbuilt Taq polymerase selection method. Although the author explores the idea of evolving repressors, riboswitches or larger operons and pathways, the examples given are relatively easy to link to the existing circuit because their activity is once again closely related to transcriptional strength.

In reality, the CPR protocol becomes increasingly labour intensive, requiring intermittent rounds of error-prone PCR (ePCR) to generate a mutant library with enough variants to begin the process. Given modest mutation rates, and uncontrolled changes, the limited diversity of the end-library generated in this way will only allow a small fraction of the theoretical sequence space to be sampled.

PCR methods can also introduce biasing given that primer hybridisation will favour certain sequence positions and any mutation in the first round of amplification will be highly represented in all subsequent rounds. In this case, error-prone PCR was carried out prior to selection, yet the process had to be stopped during the evolution rounds to tune up mutation levels by carrying out more error-prone amplification. A completely *in vivo* system would avoid such inefficient, uncontrolled steps and ensure a larger sequence space could be sampled.

The importance of the selection method cannot be underestimated and is a key consideration when developing the overall circuit design described within this work which will be discussed later. Although *in vitro* systems provide an opportunity to create huge numbers of genetic variants, up to the magnitude of 10^15, it is essential that a strong selection method is established in advance, relevant to the desired phenotypic function. This search needs to be compatible with high throughput methodology in order to extract the greatest value from any mutant libraries. Often these types of assays take a significant amount of time and labour to develop and have to be changed to suit new protein targets, before identification of useful variants can actually even begin. Their design is based on known protein activity and therefore the limited *in vitro* search must be concentrated or localised to a well-defined domain in order for selection to be possible. There is also no guarantee that the protein will function as intended when it is once again expressed within a cellular environment.

For example*, in vitro* systems are not useful for proteins which undergo complex folding, conformational changes, post translation modifications or are involved in off target cellular interactions. Structural arrangement of the extracellular form of said protein is likely to be markedly different. If we consider a target gene, coding for an intermediate enzyme which would normally function within the host metabolic pathway, we see that there is no real way to link *in vitro* mutations carried out in isolation, with truly desirable function. Even purification and evaluation of intermediate protein enzyme variants does not necessarily describe the protein structure-function relationship and there is a large risk that many identified variants will be false positives. Alternatively, the target protein could be detrimental or toxic to the host and in such cases, *in vitro* approaches may be most suitable.

It can even be argued that cell free systems fail to accurately recreate a true cellular environment. Protein folding, activity and viability will change based on ionic charge, pH, and water composition of the surrounding environment. It is almost impossible to predict cross reactivity with other intracellular biomolecules such as free cellular nucleic acids, carbohydrates, lipids in the cell and thus they fall short in identifying meaningful protein variants.

Therefore, a truly *in* vivo system is still a preferred choice and will enable selection of more accurate and efficient protein variants, allow flexibility to control selection pressures and could be

particularly useful if the target protein is an enzymatic intermediate within a more complex biosynthesis pathway or genetic circuit.

1.4.2 Mixed *In Vitro* / *In Vivo* strategies for Mutagenesis and Selection

The distinction between *in vitro* and *in vivo* systems becomes somewhat blurred given that both strategies can be combined to overcome many of their independent limitations. The motivation here is that an i*n vitro* approach can be used to create large diverse libraries before transferring into the cell for selection of any interesting and viable protein variants. Whilst this appears to be an improvement on primitive *in vitro* systems, it should be noted that there are still limitations relating to transformation efficiency and screening capacity.

Traditionally mixed approaches will employ an *in vitro* mutagenesis step made possible by error-prone PCR (ePCR) techniques (*Cadwell & Joyce, 1992*). It has long been recognised that certain DNA polymerases have greater accuracy and fidelity i.e. incorporating fewer incorrect bases and possessing added proof reading capabilities which enable them to check and remove erroneous nucleotide bases. Site directed mutagenesis of the thermostable enzyme Taq polymerase (widely used for PCR amplification) aimed to decrease its fidelity, whilst maintaining inherent levels of processivity. PCR products effectively become a library of mutants, due to the introduction of incorrect bases during synthesis. This pool of linear DNA can then be cloned into an expression vector for downstream screening and selection. All nucleotide base substitutions can be achieved, across the complete length of the target DNA sequence, which makes this approach useful for sampling complete sequence space.

The mutation rate in ePCR might be considered somewhat low, with early work suggesting a 0.6% chance of a mutation (per nucleotide position, per PCR reaction) but this can be increased by altering reaction conditions in a number of ways; adding excess manganese, increasing the excess magnesium chloride concentration, changing the pH of the reaction buffer, varying concentrations of individual dNTP's (to introduce a biased predisposition for certain nucleotides over any other), or by using a longer period of strand denaturing at very high temperatures (*Eckert & Kunkel, 1990., Eckert & Kunkel, 1991*).

A similar method uses nucleotide randomisation in which oligonucleotide primers are synthesised and extended with low fidelity polymerase to introduce mutations along the amplified product (*Shao et al, 1998*). Short mutated fragments of DNA could be re-joined to produce a large library of the full gene sequence, each mutant containing distinct differences using a process named 'random-priming recombination' (RPR). The short fragments are misincorporated and undergo mis-priming events but retain enough homology to prime to one another although the major weakness of this approach would be the frequency of destructive frame shift mutations.

An advantage of these approaches is that different error-prone polymerases can be used to provide the user with an extracted level of control, where resulting variants can be subjected to different mutation bias between selection rounds, given that one polymerase mutant may favour certain types of base inclusion. Although ePCR does not allow for precise control over the point of mutation, a synthetic random primer approach can in theory be used to achieve a uniform frequency of rational changes across the full length of the target gene. However, in practice these examples revealed biasing with transition mutations (purine to purine or pyrimidine to pyrimidine) occurring more often than transversion mutations (purine to pyrimidine). Once again, highly detrimental frame shift mutations are also likely to be very common due to the use of lower fidelity polymerases.

DNA recombination has also been used to create genetic variants by 'shuffling' DNA fragments from position to position (*Stemmer, 1994*). Randomised and blunt ended dsDNA fragments, consisting of smaller stretches of the complete target gene sequence are generated by digestion with the endonuclease DNaseI which could create di, tri and multi-base dsDNA fragmens. The fragments are artefacts of the pre-evolved wildtype gene which can be reassembled via ligation. Although there are clear benefits to recombining a collection which contain useful sequence, there are technical difficulties associated with excessive fragmentation and a loss of overlap homology. This greatly reduces the probability that a viable coding sequence can be recovered. Early examples only demonstrate simplistic methods of selection given the gene of interest coded for an antibiotic resistance marker. Therefore, any mutant which showed an increased resistance could be easily identified at higher antibiotic concentrations during a plating assay to sample a huge number of cells.

DNA shuffling was later used to create an improved version of the green fluorescent protein (GFP), which had a 45-fold increase in fluorescence against its wildtype (*Crameri et al, 1996*). It has also been used to fragment genes / proteins showing similar function or strong homology and phylogeny to create chimeric versions which have a higher probability of revealing novel yet functional activity (*Zhao and Arnold, 1997., Crameri et al, 1998., Minshull et al, 1999., Chang et al, 1999*). Furthermore, it can be used to target complete pathways or complete host genomes in the effort to create unique strains (*Crameri et al, 1997*., *Zhang et al, 2002*).

Regardless of selection method, resolving the link between improved function specific DNA/protein sequence changes can be elucidated by DNA sequencing and reveal acquired structural properties. Alignment mapping of chosen positive variants is likely to reveal consensus sequence or highly similar changes.

Mixed *in vivo* mutagenesis / *in vitro* selection techniques have been used for a range of purposes such as: evolving biocatalysts to improve yields (*Yano et al, 1998., Kumamaru et al, 1998., Arnold & Volkov, 1999*), altering protein dimerization properties (*MacBeath et al, 1998*), and even complete

evolution of an existing genetic circuit (*Yokobayashi et al, 2002*). The latter example is particularly interesting to my own work in that it suggests an alternative to the rational design of genetic circuits.

One example of using mixed systems to evolve an industrially relevant target is highlighted by the work to engineer the promiscuous enzyme xylose reductase. The wildtype version can reduce more than one substrate, resulting in a mixed product requiring purification at an increased cost to the manufacturer. The enzyme's activity was directed toward the D-Xylose substrate using rounds of targeted site saturation mutagenesis (TSSM) and ePCR, with variants subjected to multiple rounds of panning under both positive and negative selection which are directed against both the desirable and unwanted sugar substrates (*Nair & Zhao, 2008*). The outcome was an enzyme with reduced overall catalytic activity, but one with increased preference for the D-xylose substrate. Once again this represents an important need for a negative selection within the system to help ensure specificity and a reduction in cross reactivity.

Another example of success can be seen within the work to increase long-chain hydrocarbon production in the 2-keto acid-based pathway within *E.coli*. An exogenous gene for citramalate synthase (*cimA*) was taken from *Methanococcus jannaschii* to enable production of 1-propanol and 1-butanol from glucose. An enzyme with 9-fold and 22-fold higher activity respectively, was created using epPCR and DNA shuffling approaches and identified by growing on selective minimal media which does not contain the amino acid isoleucine (*Atsumi & Liao, 2008*). CimA partially recovers the isoleucine auxotroph, as it is able to circumvent a step in the synthesis pathway and shortens the overall pathway toward hydrocarbon formation from glucose. Specificity and activity were also selected over a wider range of growth temperatures.

Although error-prone PCR and DNA shuffling like methods were traditionally used to create variant libraries, the changes are infrequent, uncontrolled and likely to show biasing. An attractive alternative is to build intelligent synthetic libraries, better suited to the selection assay and the functional protein target. Cost effective dsDNA on miniaturised silicon plates, containing 10000 separate synthesis reactions, can be used to create rational libraries in which every desirable variant is made independently. However, this approach can only ever produce small subsets of the complete theoretical library size (*Twist Bioscience, 2016., Gen 9, 2016*). The other option is to introduce degenerate nucleotide bases during synthesis, to create to vast libraries (up to 10^{13}), target specific positions of interest, or introduce changes at frequencies that are likely to produce a valuable end library (compatible with screening capacity). It is important to note that traditional dsDNA synthesis methods rely on short oligo synthesis and long sequence hybridisation which often introduce errors that are not part of the original design. Multiple rounds of PCR amplification will then be used to fill in the intervening gaps and increase final yield, which significantly reduces end-pool diversity due to biased

and non-uniform amplification. Very recent advancements use 'PCR-free' synthesis methods which dramatically reduce biasing, achieving larger and more diverse libraries well suited to high capacity screening (*LabGenius, 2016*).

Nevertheless, the construction of large wide reaching synthetic libraries *in vitro* is only so useful given host cell transformation limits. Consider a 10 aa saturation library which results in 20^{10} variants, which is far in excess of standard *E.coli* transformation limits even using the most highly competent commercial cells used within electro transformation (~10^9). This is the physical limit of how much DNA can be introduced into the cell by temporarily disrupting the cell membrane. Although large scale plating of cultures and subsequent scrapings can help increase the number of independent library variants that can be screened, essentially the researcher can never reach beyond theses limits. Therefore, it would seem ill advised to create a large unfocused library if it can never be comprehensively screened.

Alternatively, intelligent library design, guided by any prior knowledge may be useful to create a library size which can be screened effectively i.e. below maximum transformation efficiency in the host. A naive first-generation library can be used during early rounds of panning, in which the stringency of selection is gradually increased. Retrieval of interesting variants can reveal favourable sequence information before second generation libraries are designed according to commonly seen mutations.

Common display screening methods such as phage display in *E.coli* and surface protein fusions in yeast successfully merge the benefits of large *in vitro* libraries with meaningful *in vivo* screening. In fact, early examples of directed evolution were carried out on therapeutic proteins which could be identified by high affinity target binding (*Roberts et al, 1992*). These approaches have been widely employed for screening antibodies / therapeutic proteins identified via successful antigen binding but underlying transformation efficiencies remain at 10^9 cfu/µg DNA (*E.coli*) and 10^7 cfu/µg DNA (S.*cerevisiae*). *E.coli* estimates are based on using 50 pg of circular pUC19 plasmid DNA for heat shock transformation into 50 µl of competent cells prior to outgrowth in SOC medium and selection on Ampicillin.) It should also be noted that there has been limited success when attempting to identify biologically active enzymes (*Amstutz et al 2001, Wittrup, 2001*).

In summary, *in vitro* libraries benefit from large diversity but need to be first transferred into a host vector. Whilst there may be concerns about cloning efficiency and the subsequent need to oversample the library in order to overcome this limitation, the major issue is that a circular construct is far greater in size. Therefore a larger pmol concentration of dsDNA is required to fully encapsulate the library, and it is even more likely that the physical amount of DNA needed required goes well above is the transformable limit. Overall this preparation process is inefficient, unstable and unreliable, requiring expensive competent cells in an attempt to maximise final transformation efficiency. In order to avoid

transformation, the proposed *in vivo* display system generates the library within the host, ensuring that complete variance can be screened.

1.4.3 *In Vivo* mutagenesis

In vivo systems are increasingly recognised as the optimal way to engineer proteins using directed evolution, because they lower the probability of selecting 'false' positives. The later examples detailed here are in my opinion the most powerful solutions to this challenging task.

Early efforts to introduce changes to target genes *in vivo* have included the use of chemical mutagens, ultraviolet (UV) light or indeed the creation of mutagenic strains which lack proper replication or repair machinery (*Myers et al, 1985., Greener et al, 1997., Horst et al, 1999., Marinus, 2010*). However, all of these approaches have limitations. Firstly, chemical mutagens often achieve a very limited mutational spectrum yet come at a large expense to cell fitness. UV light causes a wide range of non-specific mutations throughout the host cell genome. The commercially available XL1-Red strain is hypermutatable and achieves very high mutation rates given changes to the mismatch and oxo-GTP repair pathways, as well as the 3'-5' exo proofreading domain of main replicative polymerase DNA Pol III. In practice, this host strain is unpredictable, uncontrollable and particularly difficult to grow or manipulate. Determining functional variants is troublesome, given the difficulty in relating the interesting phenotype back to the initial target gene. It is inevitable that a broad and uncharacterised host genotype will make the job of identifying interesting protein mutants much more complex. This is particularly unfavourable when that specific target protein interacts with an extended pathway of cellular proteins.

A very strong method named Multiplex Automated Genome Engineering (MAGE) is extremely powerful given that it has the capacity to simultaneously target and mutate all genes within a complex pathway (*Wang et al, 2009*). This approach requires prior generation of a diverse oligonucleotide library and as such might also be considered a mixed approach. However, given the fact that these short ssDNA oligos are not viable dsDNA variants themselves, the actual mutant library is formed *in vivo*. ssDNA oligos are directed to the exposed lagging strand during DNA replication and can be incorporated into chromosomal DNA at the replication fork (***Figure 1.2***). This is a consequence of expressing the beta-protein of the lambda phage within the host strain, which is known to bind to ssDNA, facilitating homologous recombination of ssDNA fragments (*Ellis et al, 2001*). This is possible thanks to the clever design of the host *E.coli* strain which expresses the bacteriophage λ-Red single strand DNA binding protein. The system can also be tuned, as the level of homology between the pre-designed oligo library and the chromosomal target will directly influence the likelihood of mismatch, insertions or deletion modifications. Impressively, 24 genetic components of the complete lycopene pathway (an industrially important biomolecule and model system) were optimized using an estimated 4.3 billion combinatorial

genomic variants per day during screening. Whilst this is an exciting leap forward for the field, the method does require the considered and semi-rational design of a vast synthetic oligonucleotide library, meaning that there is a significant labour and cost that must be incurred with any new protein target. However, with the falling costs of DNA synthesis this approach may become increasingly accessible, so it will be interesting to see which other proteins or pathways can be evolved using this approach.

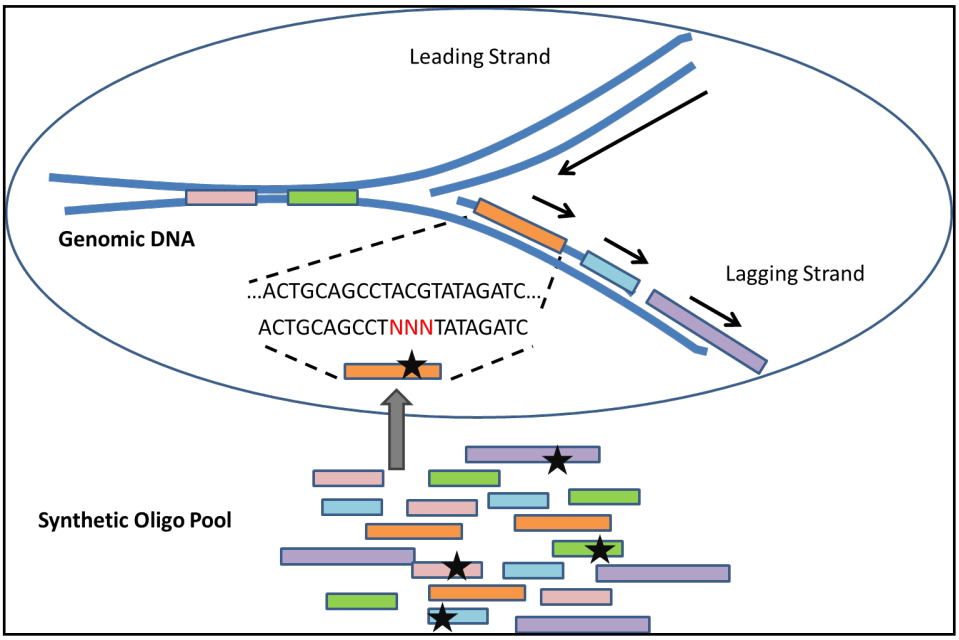


Figure 1.2 *– MAGE requires design and synthesis of a large synthetic pool of oligos which can be targeted for incorporation into the host genome within an engineered strain of E.coli which favours ssDNA recombination. Short oligos with homology to the target genes can be used to introduce substitutions, insertions or deletions.*

Phage assisted continuous evolution (PACE) was recently described and its publication coincided with the commencement of the work outlined within this report (*Esvelt et al, 2011*). It is significantly promising as a method for cycling rounds of mutagenesis of a target protein, linking the short replicative cycle of the M13 filamentous bacteriophage with an *E.coli* host strain. The creation of a positive feedback loop means the target gene/protein can be directed through continuous rounds of mutation and propagation within the host. The process is carried out within a continuous culture system which can be regularly sampled to identify genotypic changes by retrieving sequencing information.

Firstly, an *E.coli* host strain is exposed to the phage DNA containing the target gene of interest. This bacteriophage requires a specific protein (PIII) in order to become infectious to *E.coli*. The *pIII* gene is removed from the phage and inserted it into the host *E.coli* strain. The target gene itself is incorporated in the viral genome. This means a direct link between the virus and bacteria can be created by placing *pIII* gene expression under the control of target gene activity. Therefore, the most favourable target gene variants result in the most infectious phage progeny. The simplest example of

the system targets its corresponding T7 RNA polymerase. By placing the *pIII* gene under the control of a T7 specific promoter, the orthologonal activity of T7RNApol drives *pIII* expression. This means that the fittest variants can be identified i.e. those which are most active and express the greatest amount of PIII protein, will be most infectious and continue to propagate within the system. In summary, phage vectors containing mutants of the target gene infect *E.coli*, express *pIII* and allow that mutant version to propagate in subsequent generations (***Figure 1.3***).

Initial work was successful in creating a T7RNApol variant that showed 6x improved activity for an alternative T3 promoter. After sequencing revealed the full spectrum of mutations, a previously characterised residue change (N748D), known to allow recognition of the T3 promoter base at position -11, was identified (*Raski et al, 1992*).

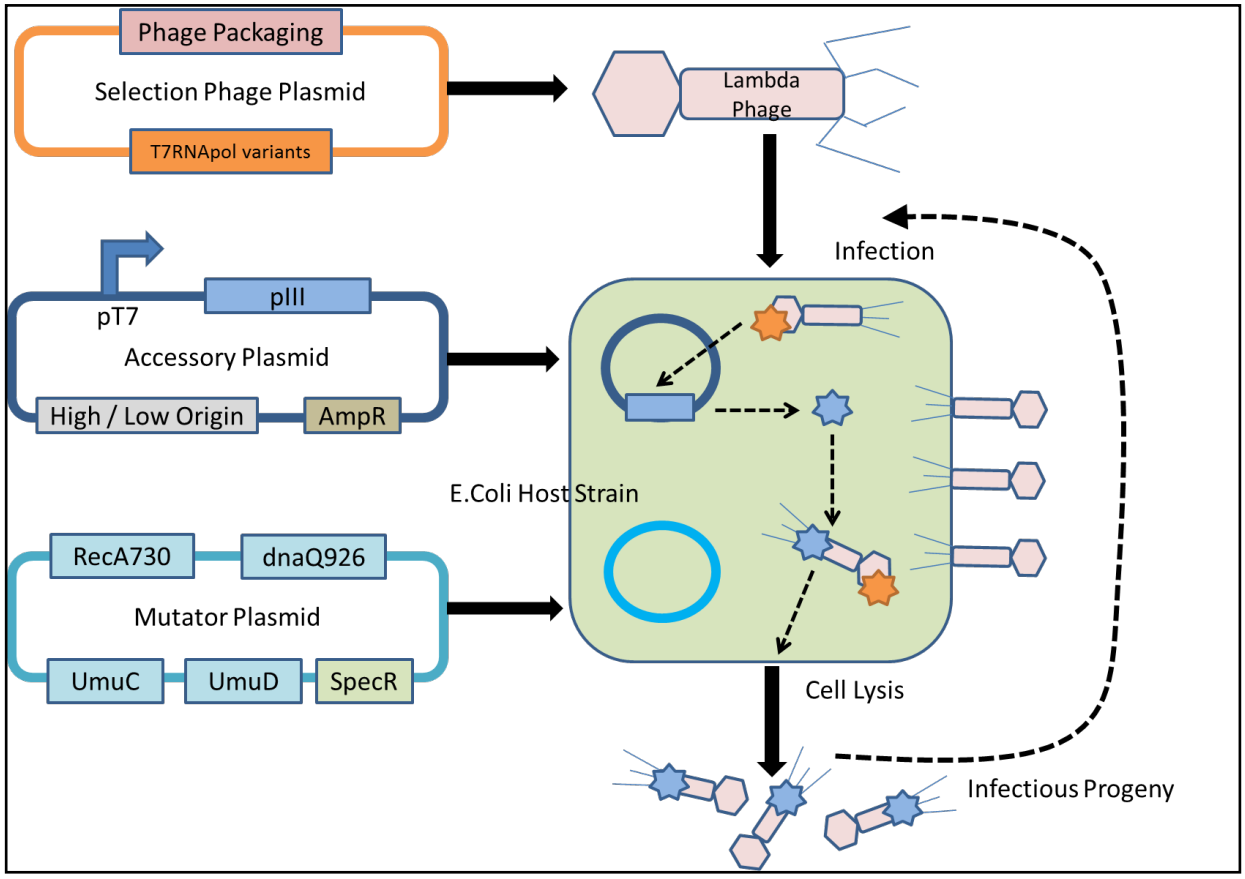


Figure 1.3 *– PACE summary outlining the phage and E.coli host replicative cycle. The example given corresponds to the evolution of T7RNApol variants.*

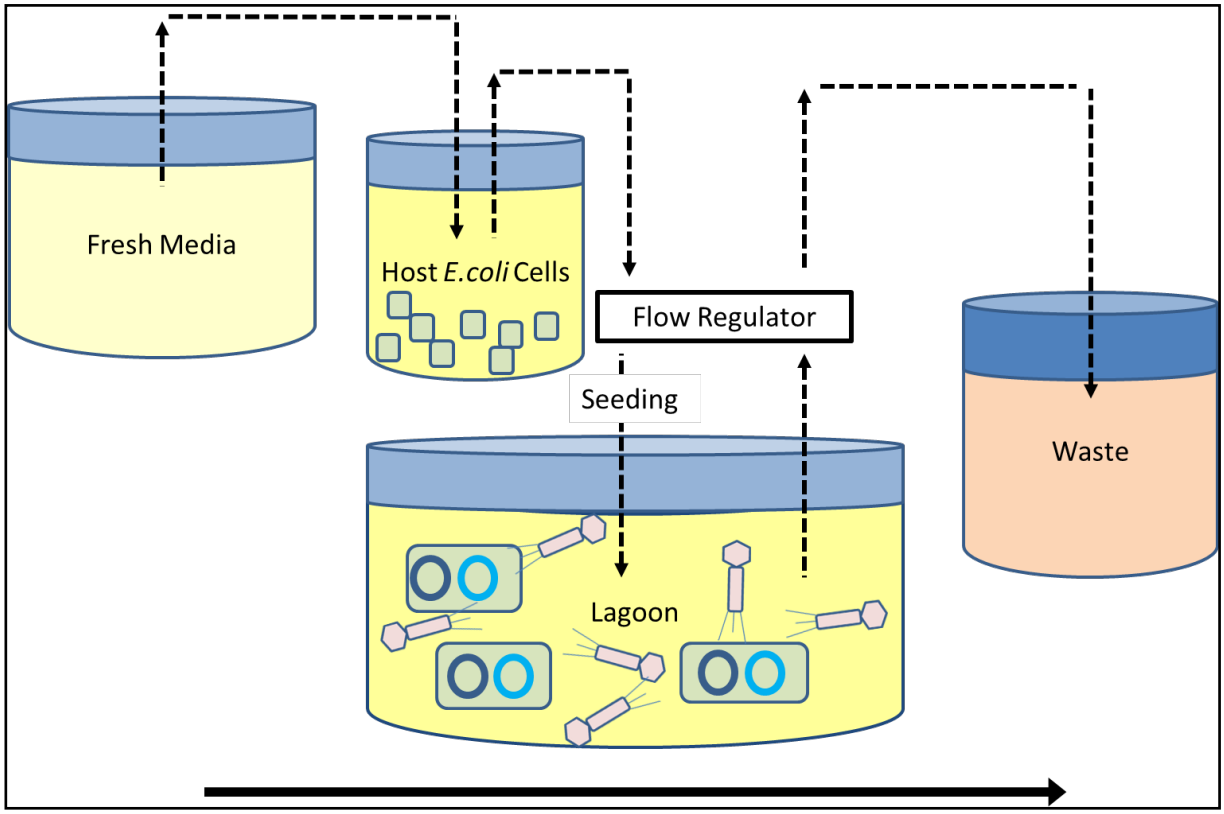


Figure 1.4 *– PACE – Regulation of the mutagenic symbiosis within the lagoon, containing both host E.coli cells (required for protein expression) and the viral bacteriophage (required for rapid gene transfer). Control of flow rate is essential to maintain continuous mutagenesis.*

PACE requires a carefully controlled culture environment and uses a bioreactor to manage flow rates of fresh media and waste culture (***Figure 1.4***). An advantage of this system is the speed of the phage life cycle, far shorter than that of the host *E.coli*, which accelerates the evolutionary process. However, the period of time that host cells are held within the lagoon needs to be managed, to reduce the possibility of potentially disruptive mutations occurring within to the host genome. Given that the phage will cause cell lysis in *E.coli*, host cells do not need to replicate and can be considered temporary hosts which mainly function to allow phage propagation.

This system carries out multiple rounds of mutation and expression of the target gene, translation of target protein variants and selection *in vivo*, avoiding the bottleneck of cell transformation. It eliminates intermediate cloning or purification steps and does not require intervention from the researcher once it is up and running. As such, it has become a very attractive method and the versatility has been demonstrated against a wide variety of target proteins to identify improved enzymatic activity and improve protein binding (*Dickinson et al, 2014., Badran et al, 2015., Hubbard et al, 2015*).

In reference to my own work system, the author outlines the requirement of a mutator plasmid which includes an error-prone polymerase, DNA polymerase V (Pol V), it's subunits named *umuD'* and

umuC. By placing it under the direct control of an arabinose promoter, a mutagenic phenotype can be induced. Over-expression of Pol V encourages random errors during low fidelity DNA replication because the polymerase itself does not possess any proofreading capabilities.

It is important to understand the makeup of the plasmid in the context of my own work, given that even the highly accelerated PACE system showed low rates of mutations without active expression of numerous mutagenic proteins.

Firstly, expression of a RecA mutant (*recA730*), which displays enhanced single stranded binding capacity was required. This mutant does not require the RecBCD enzyme to load the RecA protein onto the ssDNA strand in order to initiate DNA repair mechanisms. Truncation of the RecA mutant, removes negative residues and means the protein carries a less negatively charged overall. This variant can activate Pol V with approximately 2-fold higher efficiency than wildtype protein (*Eggler et al, 2003., Jiang et al, 2009*).

It is important to highlight the use of another mutator protein, *dnaQ926*, a mutant of the *dnaQ* gene, which encodes the proofreading 3' exonuclease subunit of DNA Pol III holoenzyme (*Fijalkowska & Schaaper, 1996*). Unsurprisingly it confers a strong, dominant mutator phenotype lacking all proofreading capabilities. The fact that DNA Pol III is responsible for the majority of DNA synthesis within the cell, could quite conceivably be the overriding cause of error-prone DNA replication.

In a phage-assisted flow system such as this, with regular addition of fresh media and new cells, only short periods of 'healthy' *E.coli* cell growth are required for the phage to replicate and for progeny to be expelled. Overexpression of these mutagenic proteins would normally result in a reduction in fitness and strain viability, but by supplementing that system with fresh cells, destructive changes can be tolerated.

There have been further improvements to the PACE protocol in an effort to tailor the system for engineering enzymes with varied functionalities (*Carlson et al, 2014*). In a later iteration of the protocol, selection stringency could be controlled at greater precision, and allow non-functional target enzymes to proliferate even through periods of inactivity (***Figure 1.5***).

Alterations to the system included placing an additional *pIII* gene under the control of an inducible phage-shock promoter/ Tet operator hybrid region so that the amount of PIII could be regulated by varying concentrations of the inducer anhydrotetracycline (aTc). This made it possible to disconnect a positive feedback between the target gene (on the selection phage) and the expression of *pIII* from the accessory plasmid. Inducing expression of *pIII* at brief intervals allows for a library of inactive variants to propagate during phage replication and pass through the cycle into the next round of mutation/selection. The induced periods can be considered phases of 'evolutionary drift' which appear to result in a more dynamic change between fitness peaks.

The latest work also provided a tool for negative selection under control of the same system, eliminating unwanted or untargeted mutants by introducing a replicative penalty. This is essential to prevent off-target effects and ensure orthogonal function of the target protein. Tuneable and proportional negative selection was possible by linking undesired target activity to the expression of a protein that limits PIII expression itself. Impressive work to highlight a key domain in the *pIII* gene (located at the C-terminus), essential for phage attachment and disassociation, revealed that a non-infective mutant was dominant over the wildtype, and could form the foundation of negative selection. By linking the unwanted target with expression of the PIII negative mutant and placing it under its own inducible control, the unwanted non-specific mutants are 'replication-impaired' and begin to be lost as the complete mixed *E.coli* population grows.

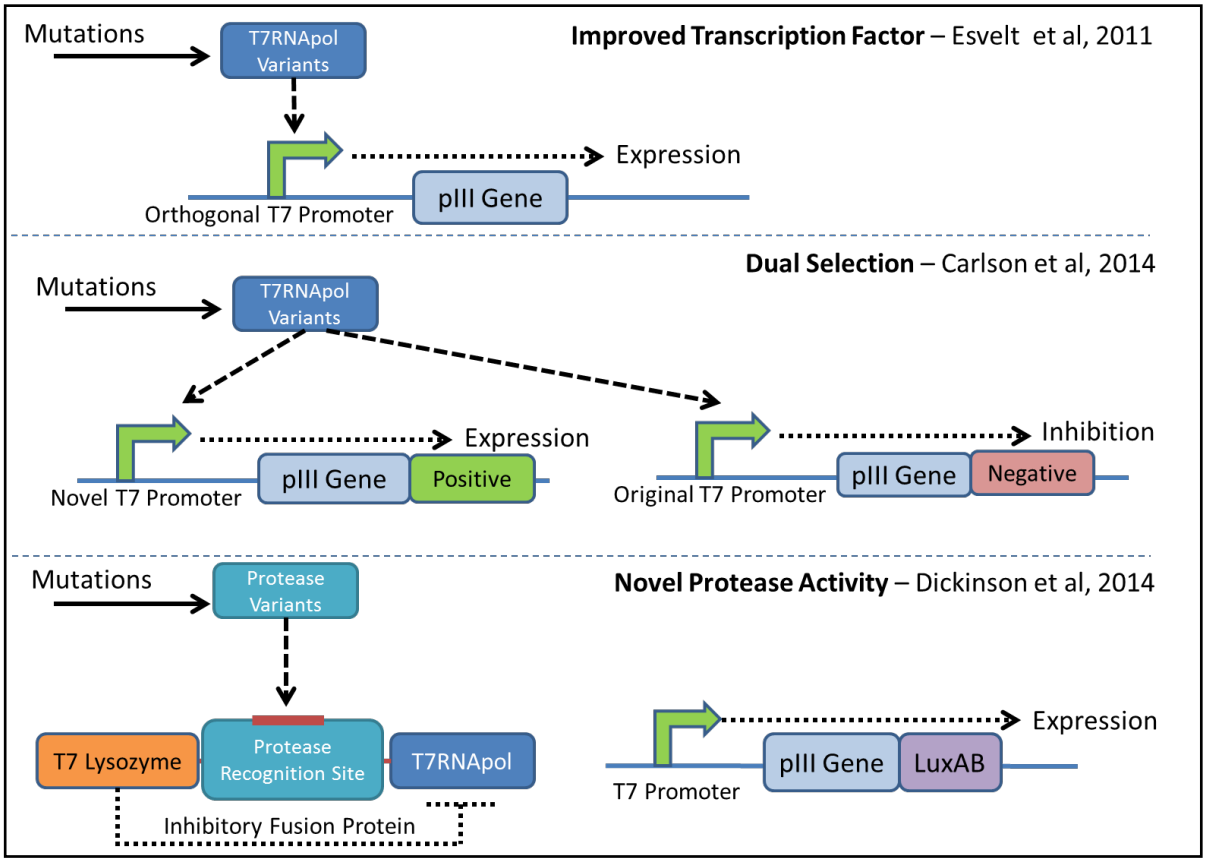


Figure 1.5 *– Variations of PACE providing examples of how the T7RNApol – pIII gene system can be linked to screening for a range of protein functions, including novel binding, improved specificity, and increase processivity.*

Even more recently, a further extension of the PACE protocol describes the directed evolution of proteases, with reference to the clinically relevant hepatitis C virus protease and its ability to mutate readily to overcome therapeutic drug candidates which were designed to inhibit it (*Dickinson et al, 2014*). The key addition to the design of the evolution principle here was the engineering of a protease-activated RNA polymerase (PA-RNAP), so that polypeptide cleavage events could be identified by a transcriptional output (luciferase expression) (***Figure 1.5***). Intelligent structural design of a fusion

protein consisting of the T7 lysozyme located at the N-terminus of T7 RNAP via a carefully constructed linker region (PA-RNAP), prevented T7RNA activity. The flexible linker included a protease specific recognition site that would undergo proteolysis when its cognate protease was expressed and can therefore liberate active T7 RNAP to allow propagation of the phage. In simple terms, the protease gene replaced the *pIII* gene within the selection phage (SP), the *pIII* gene was placed under the control of the T7 promoter on an accessory plasmid (AP) and PA-RNAP was constitutively expressed on a complementary plasmid (CP).

Interestingly, in this latest approach, the author fails to describe the mutagenesis plasmid (MP). Supplementary data suggests that expression of three genes were required to create a mutator phenotype, namely *seqA* and *dam* (DNA adenine methyltransferase) and *dnaQ926*. The *dnaQ* mutant creates an error-prone strain due to the loss of 3' exonuclease proofreading capabilities of replicative Pol III as described previously. However, there is seemingly no use of Pol V subunits or *recA* gene seen in previous PACE methods.

The overexpression of Dam within this system relates to its ability to methylate newly synthesised daughter strands of DNA, resulting in a temporary hemi-methylated state and therefore less opportunity for mismatch repair. Normally during this period, DNA repair via the methyl-directed mismatch repair system is strongly directed at newly synthesised polynucleotide strands, in order to overcome any errors introduced during DNA replication. Previous work has shown that mismatch repair is particular biased toward unmethylated strands during *in vivo* experiments, with infrequent repair observed on fully methylated DNA (*Wagner & Meselson,1976; Pukkila et al, 1983*). Consequently, it does appear that cells with increased levels of Dam would lead to rapid methylation of newly replicated DNA, limit the opportunity for mismatch repair, and become hyper-mutable (*Herman et al, 1981*).

The decision to express *seqA* seems unusual given that it is a negative modulator of replication and binds to hemi-methylated DNA (GATC sequences) localised to the replication fork, sequestering OriC and preventing re-initiation of replication (*Guarné et al, 2000., Waldminghaus and Skarstad, 2009*). It would be logical to think that high frequency replication would only serve to accelerate evolution and be favourable within this system, but perhaps this approach was taken to dampen the number of mutations to a level that will not be totally destructive to the host within the residency timeframe i.e. active period of mutation within the lagoon.

Furthermore, the choice to express both *dam* and *seqA* is of further interest because they are both recognised to cause frame-shift mutations (*Yang et al. 2004)*. This is not beneficial when evolving a target protein in systematic stepwise fashion, such as the target DNA protease, because these changes can be highly destructive, resulting in nonsense polypeptides and therefore dramatically reducing the chance of selecting a functional mutant. It seems to me a logical assumption that unwanted frame shifts

would significantly devalue this approach and stretch the limits of the screen, given that a high number of redundant proteins would mask the small number of interesting variants produced. Nevertheless, the power of the cyclical feedback and continuous selection did lead to the identification of advantageous single amino acid substitutions of the HCV protease (e.g. D168E and D168Y) which resulted in a 30-fold increase in drug resistance.

The study by Yang and colleagues is fairly relevant given my own research goes and the need to identify a potent mutagen. Here *lacZ* reversion assays were used to identify potential mutator genes using a random shotgun library made up of fragments of the complete *E.coli* genome. These were categorised as a function of their mutagenic potency. Interestingly *dinB* (Pol IV) and *dam* (DNA methylase) were identified alongside *seqA* as being highly potent and it was confirmed that the mutagenic effect of overexpression was not as a result of SOS response. A strain possessing a RecA mutant (deficient in both SOS induction and homologous recombination) was used to confirm the mutagenic effects of overexpressing *dinB*, *dam* and *seqA genes*.

As outlined above, power of the PACE protocol has been comprehensively demonstrated. Initially, RNA polymerase variants with 10,000-fold improved promoter specificity, able to accept an orthogonal promoter and reject its innate promoter, have been evolved (*Esvelt et al, 2011., Carlson et al, 2014*). Later work improved the specificity of DNA-binding proteins, recombinases and proteases. However, the functional application once again relates to the evolutionary pathway of a T7 RNA polymerase and linking the evolution of a transcriptional enzyme to its promoter and subsequent expression of an observable or selectable output is inherently simple. Both the examples of reduced stringency and negative selection have also been linked to the same transcriptional capabilities. It remains to be seen if difficult enzymatic protein targets could be coupled with the general circuit machinery and evolved in such a way.

In summary, it appears that PACE is the current gold standard for *in vivo* protein engineering by directed evolution and should serve as a useful reference point when comparing and critiquing my own system.

1.4.4 Simultaneous Positive and Negative Selection

The value of a directed evolution system that can couple positive and negative selection has become increasingly apparent (*Esvelt et al, 2011., Carlson et al, 2014*). Unsurprisingly, a positive selection is required to direct the evolutionary process and ensures that a highly specific protein with the intended function can be identified. However, a negative selection is needed to eliminate any possible cross-reactivity or non-specific off target interactions.

The advantages of such dual '+/- selection' are seen in a method for evolving endonucleases, to ensure mutants have specificity for a designed cleavage site and avoid the chance of cleavage at

unintended sequences (*Doyon et al, 2006*). In this example, the cell death B protein (*ccdB* gene) was placed under the control of the inducible Lac promoter flanked by the intended cleavage sites, meaning cell survival was only possible if the endonuclease mutant could cleave the sites and thus remove the *ccdB* gene, forming the basis of positive selection. Furthermore, by positioning undesirable wildtype recognition sites within an antibiotic resistance marker gene only those mutants that did not cut the wildtype sites could propagate under chloramphenicol selection creating a negative selection. Cell survival was only possible if the desired cleavage sites could be cut and unwanted cleavage sites remained unchanged.

When the activities of new endonuclease mutants were characterized, they demonstrated strong substrate specificity that was actually comparable to the wildtype endonuclease-target, something not achieved in earlier attempts where large levels of non-specific background mutants persisted (*Gruen et al, 2002*). This approach required preparation of plasmid DNA from clones, diversification of the library between steps, and had a very limited screen size as a consequence of re-transformation between the positive and negative selections as the system was not set up to carry out both selections simultaneously.

The importance of positive and negative selection will be discussed in reference to my own system later on.

1.4.5 Determining Mutation Rates

Prokaryotic organisms, such as the ones often employed for the purpose of directed evolution, are favourable given rapid periods of exponentially growth and short replicative cycles. This is advantageous when trying to achieve a high 'frequency' of mutation and when attempting to sample across a broad fitness landscape. Early evolutionary divergence will emerge, widening the search, and desirable variants will emerge during subsequent propagation under a selection pressure. However, this characteristic is also a disadvantage when attempting to define the frequency of mutation, given that in a bacterial host such as *E.coli*, any single mutational event that occurs during early stages of growth, is likely to be highly represented in the final end count. In fact, non-target 'rates' are better reflections of the influence of any underlying mutagenic protagonist, expressed as part of the engineering directed evolution approach. Note that 'frequency' refers to the average number of mutations counted within a small sample of end cultures, whereas 'rate' is the speed and probability that a mutational event will occur, which is far more reproducible and a closer description of true mutational events during any given outgrowth of the sample culture.

Therefore, it has been suggested that a comparison of spontaneous mutations between parallel cultures can nullify the effects of growth, given that this exponential explosion can be widely misleading when counting final mutation frequency. Many publications have described models for calculating the

number of mutational events by observing the distribution of mutations across replicates of the samples.

A summary of suitable methods is provided across a number of comprehensive reviews (*Rosche and Foster, 2000., Foster, 2006., Wong et al, 2006*). The original Luria-Delbruck method for fluctuation analysis outlines quite plainly that whilst there is an incredibly small probability that any one cell will incur a mutation during its existence (and prior to replication), there is still a chance nonetheless (*Luria and Delbruck, 1943*). Due to the fact that the mutated cell may double and produced numerous clones within the final population, with no way to track these events, the whole process is incredibly stochastic. In this original experiment, parallel cultures were grown to saturation and plated on selective media to count the number of mutants. Dilution plates were used to count the total viable number of cells. The initial equation to estimate mutation rate used mean of the distribution, although later work identified that the median provides a more accurate final parameter (*Lea and Coulson, 1949*).

The accuracy of any of these methods is often questionable, but one important consideration is to predict the expected range of the number of mutational events, identified across parallel cultures, so that the upper and lower bounds of distribution can be incorporated. Furthermore, the experimental procedure is particularly important and should be adhered to as closely as possible. In terms of making a comparative assessment of directed evolution tools, often a gain of function assay such as reversion to antibiotic resistance is the preferred choice.

In order to compare my own system with other directed evolution methods, it would be useful to calculate mutation frequency, but stronger comparison could be made by determining mutation rate.

1.5 Key Considerations

My own *in vivo* system for directed evolution is directly compared to the well-established and powerful methods of MAGE and PACE, which offer high throughput cyclic screening and selection, with limited need for external interference once the system has been activated. Although I considered these methods to be the leading techniques within the field upon commencement of the project, they both have certain limitations. MAGE requires a labour intensive and costly design, build and test process. PACE requires constant maintenance of a chemostat / turbidostat and significant fine tuning, yet fails to provide definitive control over the rates or type of mutation. Neither are able to react or respond to interesting phenotypes as there is no form integrated feedback.

My main aims and objectives were to construct a self-controlling circuit which allowed for a significant, yet sensible, rate of highly targeted mutation and I expected to be able to compare my approach with similar methodologies. My belief was that by finding the correct balance between mutation frequency and potency, new and interesting phenotypes would be identifiable amidst periods

of fast cell replication and growth. In order to achieve this, each component of the cell circuit needed to be well understood and in particular, the ability to switch between states of protein repression to expression was required.

The early literature study was extremely beneficial because it helped identify a number of mutagenic candidates which could be suitable for inclusion with my own system. It also highlighted common analytical standards and the value of assessing the final circuit in the same manner.

2 Creating a Mutator Device

My underlying aim was to develop a robust method for directed protein evolution *in vivo* by taking advantage of the rapid bacterial replication and to avoid the requirement to transfer the genetic target between hosts, or between an *in vitro* mutation platform and an *in vivo* selection system, which only serves to reduce the size of the genetic sequence space that can be sampled. The key challenge is to be able to direct mutation and diversification at the genetic level towards only the target gene, whilst eliminating mutations elsewhere in the host genome and maintaining the overall fitness of the cell.

The novel *in vivo* circuit aims to concentrate multiple rounds of directed molecular evolution toward the target gene/protein by linking a mutagenic protein to its transcription. Essential to this system is a protein named activation-induced deaminase (AID), which functions within the immune system of higher order organisms. This protein will be fused to the viral transcription factor T7RNA polymerase (T7RNApol) that has widely used in prokaryotic expression systems and is well understood, achieving very high levels of transcription from its orthogonal promoter, which does not exist within the wildtype *E.coli* genome. Ultimately, the target gene will incur multiple mutations during transcriptional events and propagate through rounds of cell replication.

2.1 Background

2.1.1 Activation Induced Deaminase - AID

AID biochemistry and physiological function has most comprehensively been reviewed in a recent article (*Larijani & Martin, 2012*). Here I outline the main evidence for its mode of activity within the existing literature and consider its potential application as a functional mutator within the exogenous context of a bacterial host.

In higher order organisms, the broad range of genetically diverse antibodies found within early activated B cells at germinal centres, are principally generated by homologous recombination events that create a set of antibodies with wide reaching affinity to numerous target antigens. However, there is a requirement for fine tuning affinity against important antigenic targets, by specifically targeting variable (VH) regions and altering constant regions (Fc). At this stage, antibody diversification relies on somatic hypermutation (SHM) to introduce nucleotide substitutions responsible for the generation of small and distinct changes at complementary determining regions (CDR's) and creates vast libraries during affinity maturation, with preferred variants showing the highest affinity to the target antigen (*Teng & Papavasiliou, 2007*). Class switch recombination (CSR) is able to completely change the constant region of the immunoglobulin and thus defines the mode of immune response. Both SHM and CSR require the activity of the DNA binding enzyme AID (*Muramatsu et al, 2000., Arakawa et al, 2002.,*

Durandy, 2003) which is able to deaminate cytosine to uracil by nucleophile attack, through addition of a zinc bound water molecule and loss of the amide group (*Petersen-Mahrt et al, 2002*) ***(Figure 2.1)***. Therefore, AID can be considered as a master regulator of secondary antibody diversification.

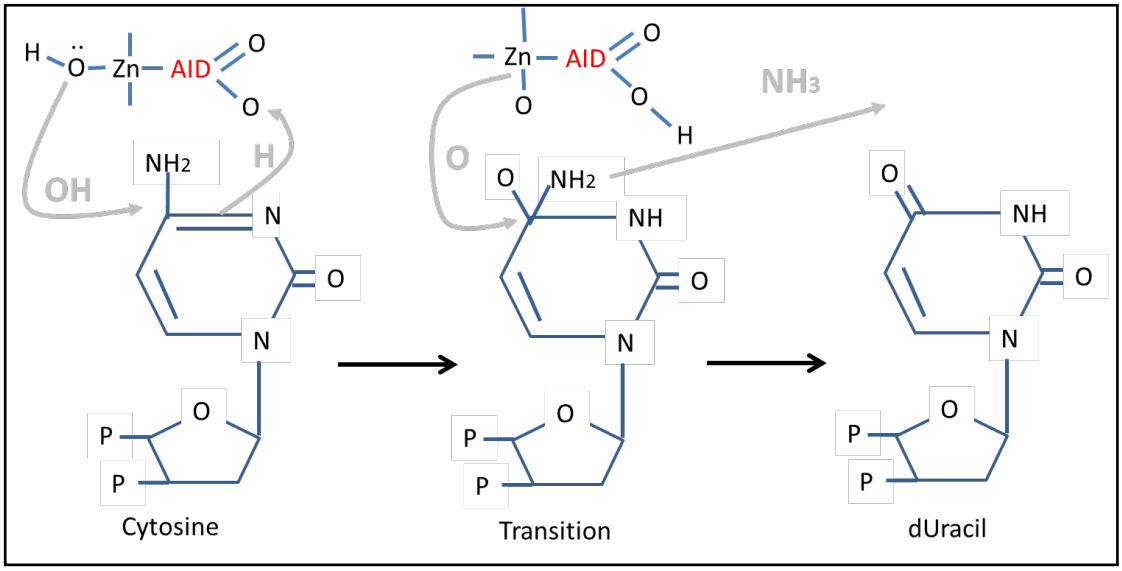


Figure 2.1 *– Proposed AID mode of action for deamination of cytosine nucleotide*

Mutations occur at a rate far above normal background mutation rate when AID is actively expressed within specific B cells. Unsurprisingly, immuno-compromised disease phenotypes can result if the AID gene incurs a disruptive mutation itself, and this has been shown in both mice and human models. For example, an inability to class switch due to AID inactivity is known as Type 2 Hyper IgM syndrome. Weaknesses in the immune system as a direct consequence of AID inactivity or deficiency, range from increased susceptibility to bacterial infection to a prevalence of debilitating autoimmune reactions (*Revy et al, 2000., Muramatsu et al, 2000., Meyers et al, 2011*).

It is clear that AID is essential for creating genetic diversity within the immune system, but the mechanics of AID activity remain a topic of significant debate given that protein crystal structures and protein-DNA interaction modules remain elusive. AID activity has been shown to have different effects on DNA substrates depending on the environmental context of the assay, i.e. the outcomes of expressing the protein *in vivo* or using purified samples for *in vitro* testing, within either eukaryotic hosts and prokaryotic hosts, give very different results.

My exploration into the wider published literature has been useful when considering the application and benefits of AID as an active mutator. I have also attempted to summarise the numerous and contradictory findings below in an effort to justify the choice of AID as a suitable protagonist in my *in vivo* directed evolution system (***Table 2.1***).

***Table* 2.1** – *Investigations into AID activity*

Study	Context	Target DNA	Target Strand	References
Transcription of Synthetic Oligos	In Vitro	ssDNA oligos	Sense	Bransteitter 2003 Pham 2003 Dickerson 2003
Loss of function	In Vitro	Transcribed dsDNA	Mixed	Ramiro 2003
Gain of function	In Vitro	Transcribed dsDNA	Mixed	Chaudhuri 2003 Shen 2005
Active Transcription	In Vivo	Transcribed dsDNA	Mixed	Chaudhuri 2003 Shen 2005
Promoter Positioning	In Vivo	Transcribed dsDNA	Mixed	Besmer 2006 Shen 2005
Protein Dimerisation	In Vitro	-	-	Xie 2004 Conticello 2008 Brar 2008 Basu 2005

2.1.2 *In Vitro* Studies of AID Activity

A number of studies have used *in vitro* assays to analyse AID dependent deamination events on synthetic ssDNA oligonucleotides during active transcription (*Bransteitter et al, 2003. Pham et al, 2003. Dickerson et al, 2003*).

Initial work suggested that there was a requirement for RNase in order to remove small RNA inhibitors from the purified AID protein, but I would suggest this is merely a consequence of the insect expression system used and the assay employed. However, this study was useful in identifying AID preferential association with ssDNA targets (*Bransteitter, 2003*).

The follow up study by the same group suggested AID deamination events occur at high frequency on selected synthetic ssDNA oligos, whereas other target oligos would remain untouched. This is perhaps a result of low AID concentration and high template concentration, meaning there was incomplete saturation of enzyme to substrate during the assay. However, it does provide some evidence of AID's strong binding to ssDNA, given that a weak association might be more likely to result in repeat binding and decoupling events. The author suggests this could be a consequence of a terminal domain region, consisting of mainly basic residues, resulting in a positive charged protein that would readily associate with a negatively charged DNA template (*Pham, 2003*). Interestingly, this study identified mutations at hot spots along the full length of the target sequence and demonstrated AID processivity. This is something that would be critical for my own directed evolution system, with the potential of introducing changes throughout the complete target gene.

Dickerson and colleagues carried out detailed EMSA and TLC binding assays which once again show AID's propensity for single stranded DNA templates. Importantly they were able to distinguish difference between AID and RNA-editing enzyme apolipoprotein B editing catalytic subunit 1 (APOBEC-1) which shows strong binding and the ability to deaminate RNA, whereas AID did not (*Dickerson, 2003*).

All of these studies appear to be in agreement in that they show transcription dependent deamination events and increased mutations in highly transcribed genes, although selection based on loss or gain of function of the target gene itself revealed different preferences for the transcribed or non-transcribed strand (*Ramiro et al, 2003., Chaudhuri et al, 2003., Shen et al, 2005*) or in fact no preference what so ever (*Besmer et al, 2006*).

There is some correlation between the strength of transcription and frequency of mutations which is reduced if the target gene is positioned further from the promoter transcriptional start site (*Besmer et al, 2006; Shen et al, 2005*). This might suggest instability of the transcriptional complex when AID is active at the transcription bubble, perhaps a consequence of following the active RNA polymerase along the gene and could suggest a model for transient, fleeting periods of mutational activity which is something I hope to utilise within my system.

The work by Sohail and colleagues, in which a human AID variant was expressed and purified, is a good example as to why *in vitro* assays are unlikely to provide strong evidence for my own model, given that they do not recreate true cellular environments. In fact, in this example, the non-natural conditions are over exaggerated and harsh, with high concentrations of AID and UDG added into a mix containing synthetic dsDNA strands with a fixed loop region, in an effort to recreate a transcription complex. It was revealed that cytosine bases within the loop region undergo deamination events, UDG removed Uracil bases and the addition of NaOH lead to cleavage at the abasic sites (*Sohail et al, 2003*). In this case, significant secondary structure and a large amount of protein aggregation is likely to result in excessive amounts of AID stalling at the specific target sequence, increasing the chance for mutation.

2.1.3 *In Vivo* Studies of AID Activity

In vivo analysis of AID activity during active transcription does not seem to show a preference for either strand (*Chaudhuri et al, 2003., Shen et al, 2005*).

Further work explored over expression of AID in an *E.coli* plasmid system, in which C – T mutations are identified in a repair deficient strain. This target strain reverts back to being Kanamycin resistant more readily when under periods of active transcription (*Sohail et al, 2003*). This paper is of strong relevance given the gain of function selection used, in which active mutation can be detected by a reversion to resistance, meaning positive events are easy to identify. This provides a means to quantify rates of mutation. Nevertheless, a plasmid based target is not optimal, given that a true number of events cannot be determined accurately given variant plasmid copy numbers. Furthermore, there is evidence of spontaneous reversion mutations.

2.1.4 AID Target Sequences

Some studies have suggested the AID shows preferentially targeting of particular DNA motifs (*Bransteitter et al, 2002*). This argument is fuelled by the regularity of common sequences within the target immunoglobulin genes and the occurrence of the WRCY motif. Therefore, clustered mutations give the impression that AID is acting in a sequence dependent manner, but in my opinion, I believe there are two reasons why this may not be the correct conclusion. Firstly, overloading the assay with purified AID protein is likely to result in protein aggregation and thus reduce the processivity of the enzyme. Secondly, these hotspots may well create looped secondary structures that further encourage stalling at a defined position. Furthermore, there are also far more numerous studies which have shown that AID does not have a propensity for particular cytosine bases, and that the sequence context makes no discernible difference (*Dickerson et al, 2003*). As I will allude to in my own results, I believe the suggestion of target sequence or recognition sites can now be rejected.

An alternative concept of location specific targeting of the AID protein makes reference to the increased activity of AID within CSR regions where DNA – RNA hybrid R-loops are formed, as these are required for recombination events. The proposal is that the loops leaves one strand of DNA exposed, similar to that exposed single strands during active transcription, and thus AID has the opportunity to act upon this stretch of sequence (*Chaudhuri et al, 2003*). This phenomenon may lead to the same outcome as when any target gene is under active transcription. In fact, AID activity is likely to favour extended periods of exposed ssDNA and indeed it is the coding strand which is subjected to deamination events given that the non-coding (transcribed) strand, read in the 3' – 5' direction by the RNA polymerase, forms an immediate and protective DNA – mRNA intermediate hybrid. Given my proposed model of fixing the association of an RNA polymerase with the AID protein itself, it is important to consider these interactions in the context of the complete transcription bubble.

2.1.5 Models of AID Activity and Multimeric Dimerisation

Certain publications have hypothesised that AID exists as a dimer or possibly a high-order oligomer based on structural models generated using homologous cytosine deaminases *(Xie et al. 2004)*. In this example, a crystal structure of the RNA-editing enzyme apolipoprotein B editing catalytic subunit 1 (APOBEC-1) which demonstrates deaminase activity, was used to make a comparative model of AID. However, if we take into account that AID has already been shown to mutate exposed ssDNA, the reliability of this hypothesis is somewhat questioned, and in fact APOBEC-1 and AID only share very limited sequence similarity (*Conticello, 2008*). Importantly, a recent study which uses atomic force microscopy to reveal the oligomerisation state of AID activity *in vitro* shows that it is preferentially directed toward ssDNA exposed within the DNA transcription bubble and exists as a monomer (*Brar et al, 2008*). However, this does not rule out the possibility that post-translational modifications events,

such as phosphorylation by protein kinase A, may create dimer functionality *in vivo* (*Basu et al, 2005*). In light of these conclusions, my investigations into AID activity will be carried out *in vivo* and will be extended to assess functional AID proteins in both free monomeric and constrained dimer form.

Whilst the underlying biochemistry, mode of action and function of AID remains unresolved, the overall consequence of AID activity within mammalian systems is better understood. AID causes a significant increase in deamination events at cytosine (C) – guanine (G) base pairs as cytosine is deaminated into Uracil (U). This creates a U:G mismatch which could potentially introduce incorrect T:A pairs after subsequent rounds of DNA replication as both strands remain viable templates. However, base excision repair (BER) is initiated through activity of Uracil DNA Glycosylase (UNG) which excises the errant uracil and is followed by activity of an AP (Apurinic/Apyridindic) endonuclease (APE1) which creates a nick in the phosphodiester backbone. The resulting single strand breaks can be processed by displacing or non-displacing DNA polymerase during either short or long patch BER e.g. DNA Pol β (Beta) or DNA pol δ to move past the abasic site using the template strand to add in the correct bases in place (***Figure 2.2***).

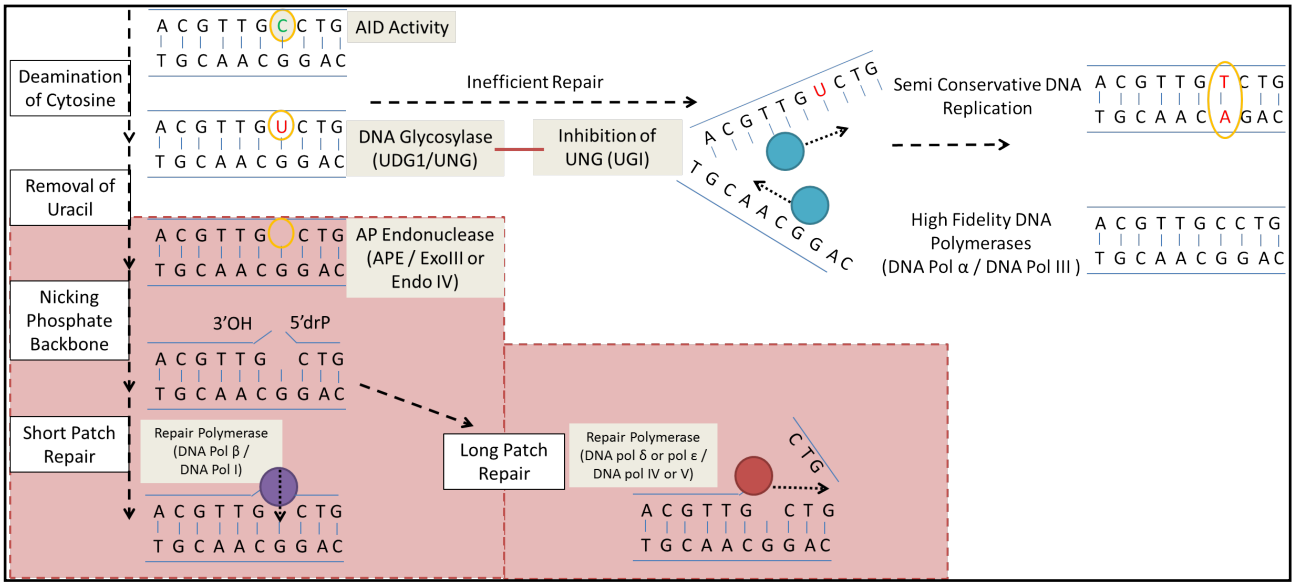


Figure 2.2 – *Inhibition of Base Excision Repair pathway encourages accumulation of C-T transition mutations*

Correct BER will prevent the overall C to T transition, without errors, which is of course desirable in most cells within the body. However, the activity of AID as a mutator is validated by the evidence that UNG acts downstream, and in fact UNG inactivity also results in impaired class switching leading to Type 5 Hyper IgM Syndrome where mutations do not occur (*Imai et al, 2003*). It is therefore important to consider the role of translesion polymerases, such as DNA pol η (eta), in overcoming the abasic site during replication, prior to activity of UNG, but leading to possible inclusion of any of the four nucleotides regardless of the template strand. Furthermore, it should be noted that if two uracil bases

are removed simultaneously on opposite strands, this will equate to a staggered double strand break that is then processed by nucleases, resulting in blunt end fragments permissible for CSR.

It would also seem likely that AID may also initiate a mismatch repair mechanism and acts as a recognisable substrate to the MutS complex (*Green et al, 2011*). There are many published examples which show the recruitment of low fidelity polymerases in response to AID generated abasic sites e.g Pol eta, Pol zeta and Pol theta (*Poltoratsky et al, 2001., Zeng et al, 2001., Zan et al, 2005*).

The reason that incomplete excision and error-prone repair is allowed to happen only when in specific response to AID mutational activity, remains a topic of much debate (*Honjo et al, 2004., Chahwan et al, 2012*). One hypothesis is that UNG activity directly affects low-fidelity error-prone repair in these cells, and that it is influenced by the sequence context in which the mutation is located (*Pérez-Durán et al, 2012*). Hypervariable IgV regions actually show numerous A:T pair mutations and these are common place when expressing AID *in vivo*, yet this spectrum of mutation remains completely unseen in previous *in vitro* studies, suggesting that true cellular AID activity is influenced by another factor.

It is clear that AID has characteristics that are favourable for a directed evolution system, specifically an increased mutational activity across the full length of highly transcribed genes by targeting exposed ssDNA at the transcription bubble. Therefore, I will explore the frequency of AID induced mutations achieved by placing a target gene within the genome of the host, under a strong orthogonal promoter with the general belief that the greater the transcriptional rate, the higher the level of mutation.

Consider for a moment, an *E.coli* cell culture grown overnight for 12 – 16 hours that reaches 1.0 OD600 nm. This will contain in excess of 8 x 10^8cells/ml. Each cell will contain a single copy of the target gene, meaning that there is a huge potential library size created, without the need to interfere directly with the cells. This should mean that interesting variants are able to develop within just a few rounds of growth.

2.2 Testing AID Mutational Activity

In its most basic form, the proposed system will rely on the overexpression of an AID-T7 polymerase fusion protein (*the mutator*) that is able to transcribe a single copy GFP gene integrated into an *E.coli* strain (*the host*) under the control of a highly specific, non-native and orthogonal T7 promoter (*the target*). The fusion protein should allow for specific targeting of AID molecule to the T7 transcription bubble, thus it will only mutate the transcribing gene. This should reduce the possibility of off target mutations within the host genome, providing the foundations for targeted directed evolution *in vivo*. A graphical model for the proposed system is given (***Figure 2.3*** *and* ***Figure 2.4***).

Initial experiments were used to validate my earlier assumptions and provide answers to the following questions:

1. Does expression of AID within *E.coli* host create a mutagenic phenotype?
2. Does AID show a greater level of activity in monomeric or dimeric forms?
3. Can AID be targeted to a specific gene or interest located on the host genome?
4. Can we limit the number of background mutations incurred by the host organism?
5. What types of mutations, if any, can be identify?
6. What is the overall frequency of mutation achieved?

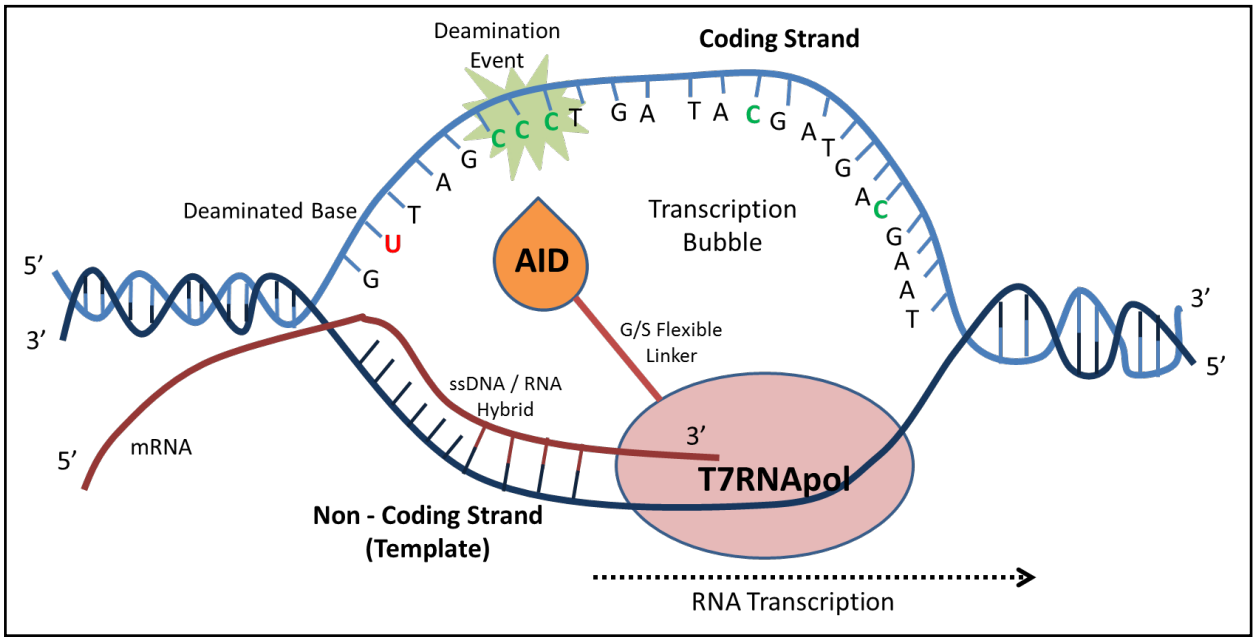


Figure 2.3 *– Proposed model for activity of the AID-T7 fusion protein during active transcription. T7 RNA polymerase is direct to the target gene by association with its highly specific cognate promoter.*

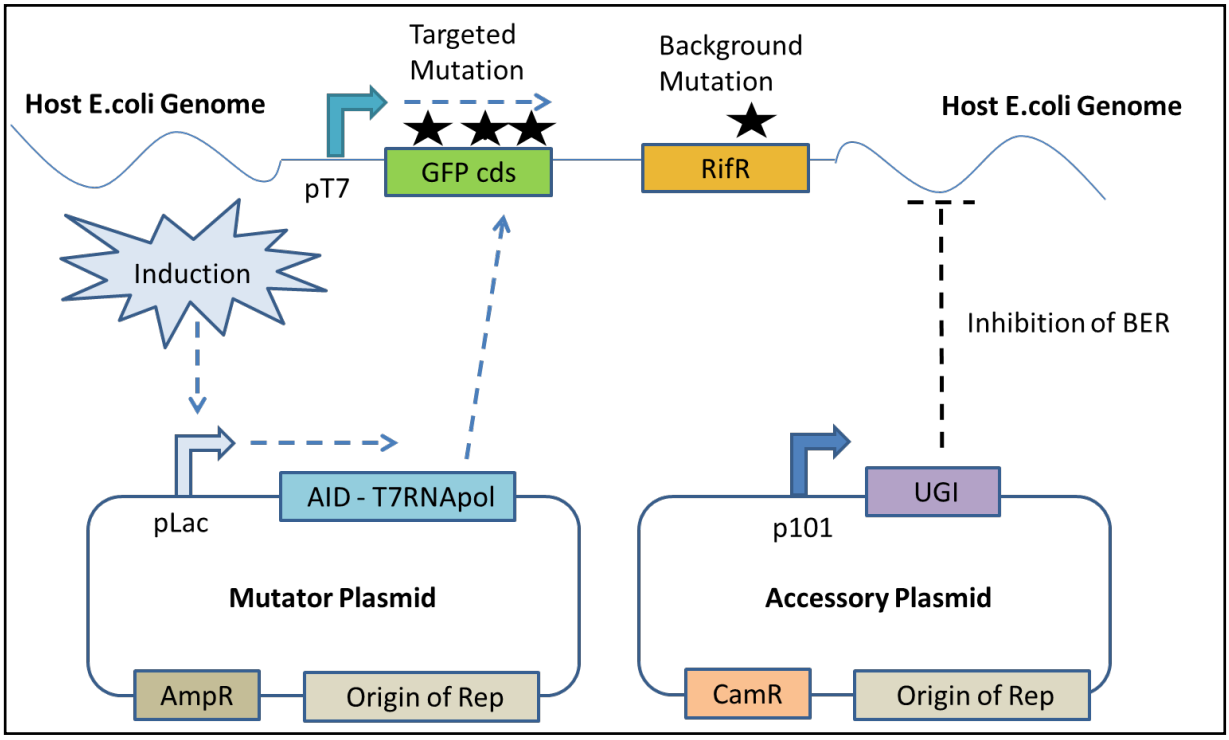


Figure 2.4 *– My initial design for an in vivo system for directed evolution. A GFP target gene is introduced in the host E.coli genome. Targeted mutations are a consequence of expressing the ATF mutator fusion protein and UGI inhibitor which cause deamination events and inhibit BER respectively. The intended result is a loss of GFP fluorescence. Background mutations are identified by resistance reversion to Rifampicin.*

It was believed this approach would result in a very high level of transcription and ultimately significant number of mutation events. The target gene itself encodes a fluorescent protein GFP which has been widely used as a biological marker due to its ease of visualisation. Given that it can be expressed as a large percentage of total cell protein, mutations should be identifiable during a loss of function assay in which fluorescence is lost. Simultaneously, off target background mutations that may occur within the host could be identified using an antibiotic plating 'gain of function' assay to determine if the host shows Rifampicin resistance.

2.3 Methodology

2.3.1 Plasmid Construction

Initially the AID and T7 RNA polymerase genes were sourced from the registry of standard biological parts and cloned separately into the blank pSB1AK3 plasmid containing a low copy origin and Ampicillin resistance marker, using the BioBrick standard assembly method - note this step was conducted in previous work (Adie, 2013). Both genes were placed under the control of the inducible Lac0-1 promoter. Unfused AID+T7 was made by cloning AID cds and T7RNApol CDS into the same vector retaining both Lac0-1 promoters and individual start and stop codons, ensuring individual mRNA production from a single plasmid system. The AID-T7RNApol fusion protein (ATF) was made by cloning

the AID CDS upstream of the T7RNApol cds, eliminating the AID stop codon (TAA) thus creating N'-AID-T7RNApol-C' fusion protein. This arrangement is favourable as T7 C-terminal fusions are known to disrupt the transcriptional activity of the polymerase as revealed by its structure with evidence in the literature (Sousa 1993, Dunn 1988). To explore the possibility of increased AID activity upon dimerization, and AID-AID dimeric fusion was created to incorporate a long flexible Glycine/Serine linker sequence which should permit dimerization upon expression. The AID-AID dimer was incorporated into the existing vectors in both fused and unfused conformations via Gibson Assembly (***Chapter 9.5.1***). Final constructs are detailed in ***Table 2.2.***

Table 2.2 *–Plasmid constructs used to test the activity of the proposed mutator device*

Construct Name	**Description**	**Backbone**	**Antibiotic Resistance**	**Short Hand**
p.T7RNApol Only	T7RNApol expressed individually	pBR3222 Low Copy	Ampicillin	T7
p.AID-T7RNApol	AID fused to at N terminus of T7RNApol and expressed as a single protein	pBR3222 Low Copy	Ampicillin	ATF
p.AID + T7RNApol	AID and T7RNApol expressed individually	pBR3222 Low Copy	Ampicillin	A+T UN
p.AIDAID	Dimeric form of AID	pBR3222 Low Copy	Ampicillin	AA
p.AIDAID-T7RNApol	Dimeric AID fused at the N terminus of T7RNApol	pBR3222 Low Copy	Ampicillin	AATF
p.AIDAID + T7RNApol	Dimeric AID and T7RNApol expressed individually	pBR3222 Low Copy	Ampicillin	AA+T UN
pUGI.Cam	UGI (inhibitor of UDG) expressed individually	pLys Low Copy	Chloramphenicol	UGI+
pGFPmut3b.Gent	Integration plasmid containing GFP mut3b target gene cassette for insertion into phi80 att site in *E.coli*	pAH153 R6K	Gentamycin	GFP.INT
GM31 GFP	Engineered strain positive for GFP under T7 expression	N/A	N/A	GM31 GFP

2.3.2 Genome Integration

The single copy of the target gene was introduced into the genome of the host *E.coli* strain via the CRIM based system genome integration system (*Haldimann and Wanner, 2001*). This was required for clear identification of changes to target protein because the alternative approach, using multi copy plasmid expression, would make identification of new gene/protein variants very difficult if the majority of template DNA remains unchanged. GFPmut3b was chosen as the target gene of interest for simple visualisation *(Cormack 1996)*. Initial attempts to introduce a 'T7promoter-GFPmut3b-terminator' cassette into a repair deficient strain BH156 (UNG-) were unsuccessful. Therefore, it was agreed that *E.coli* strain GM31 would be used as it was known to be more suitable for genome integration. The target gene cassette was cloned into pAH153 and successfully incorporated at the Phi80 phage

attachment site within the GM31 genome. CRIM plasmid integration protocol was used to successfully integrate the target cassette, and selection of transformants was then conducted on Gentamycin (20μg/ml) *(Adie, 2013)*. This method of integration is outlined in ***Chapter 2.***

An alternative approach to generate a repair deficient strain was sought, which involved expression of Uracil Glycosylase Inhibitor (UGI) within the GM31. UGI is a small protein retrieved from a bacteriophage which is able to inhibit UNG by reversible protein binding with 1:1 ratio *(Wang and Mosbaugh, 1988)*. Thus, expression of UGI within our target strain prevents efficient base excision repair by limiting UNG activity and should increase the possibility that AID induced mutations are evident *(Noia and Neuberger, 2002)*. The UGI gene was cloned into the low copy pLys vector, resistant to Chloramphenicol, thus allowing co-transformation with the mutator device plasmid *(Adie, 2013)*.

2.3.3 Host Strain Preparation

Competent GM31 and GM31.GFPmut3b host cell were transformed by heat shock. Co-transformation and double selection of pUGI.Cam (DNA repair inhibitor) and all variants of the mutator device was carried out on rich SOC+ media plates (2% tryptone, 0.5% Yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10mM $MgSO_4$, 2% Glucose) with added antibiotics Carbenicillin (100 μg/ml) and Chloramphenicol (25 μg/ml).

Early problems associated with the inherent leakiness of Lac0-1 promoters, determined that crude LB media was not suitable and permitted early induction of mutation, skewing results downstream (*data not shown*). Therefore, initial selection and outgrowth was carried out on high glucose SOC media was used for catabolite repression.

For each variant mutator device, 8 transformant colonies were picked and grown overnight in SOC+ media to prevent early induction. The use of existing glycerol stocks was avoided as this appeared to result in mutation bias, potentially a consequence of previous periods of outgrowth which allowed propagation of early advantageous 'jackpot' mutations (early preliminary data showed significant divergence in samples expressing the same mutator device. This could only be explained by an underlying pre-mutation that was passed on in all subsequent generations. AI media contains a limited amount of glucose, which is metabolized preferentially during early growth and becomes depleted at mid to late log phase. Therefore, a switch between carbon sources results in lactose being taken up by the cell and activates the pLac promoter, driving expression of the mutator device. This approach allowed for normal growth of culture early on and helps avoid further mutation biasing.

2.3.4 High-throughput Repeat Assay

Repeats of this assay were adapted to favour a 96-plate format to help speed up and streamline the process of cell dilutions, normalisations, OD measurements and FL readings given compatibility with 96-well plate incubator, plate readers and automatic micro-sampling system (AMS) for flow cytometry.

2.4 Results and Discussion

2.4.1 Assessment of Strain Fitness and Viability

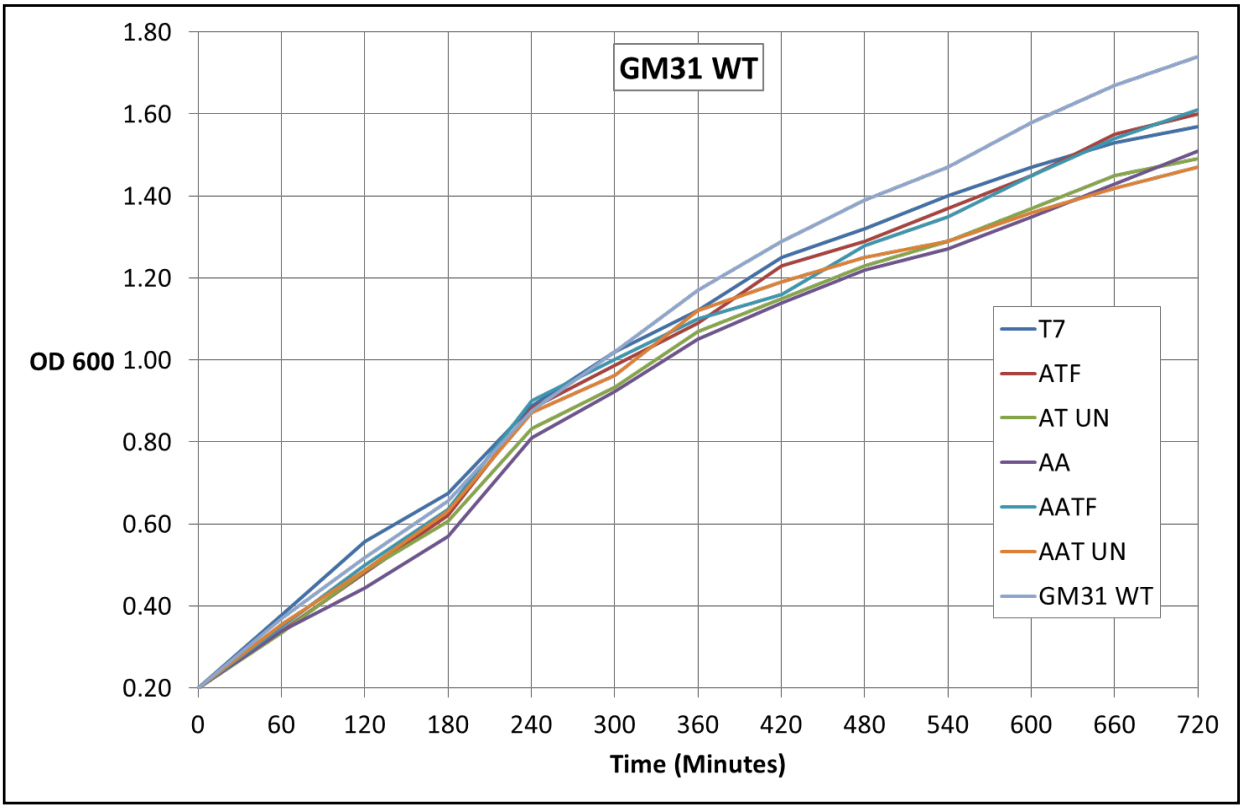


Figure 2.5 *- Growth curves demonstrated the viability of expressing the various mutator devices within the target E.coli GM31 host cell. Cells were normalised to 0.2 and allowed to grow for 12 hours with OD600 measurements taken each hour. 8 biologically independent samples were measured for each mutator device and averages are shown here.*

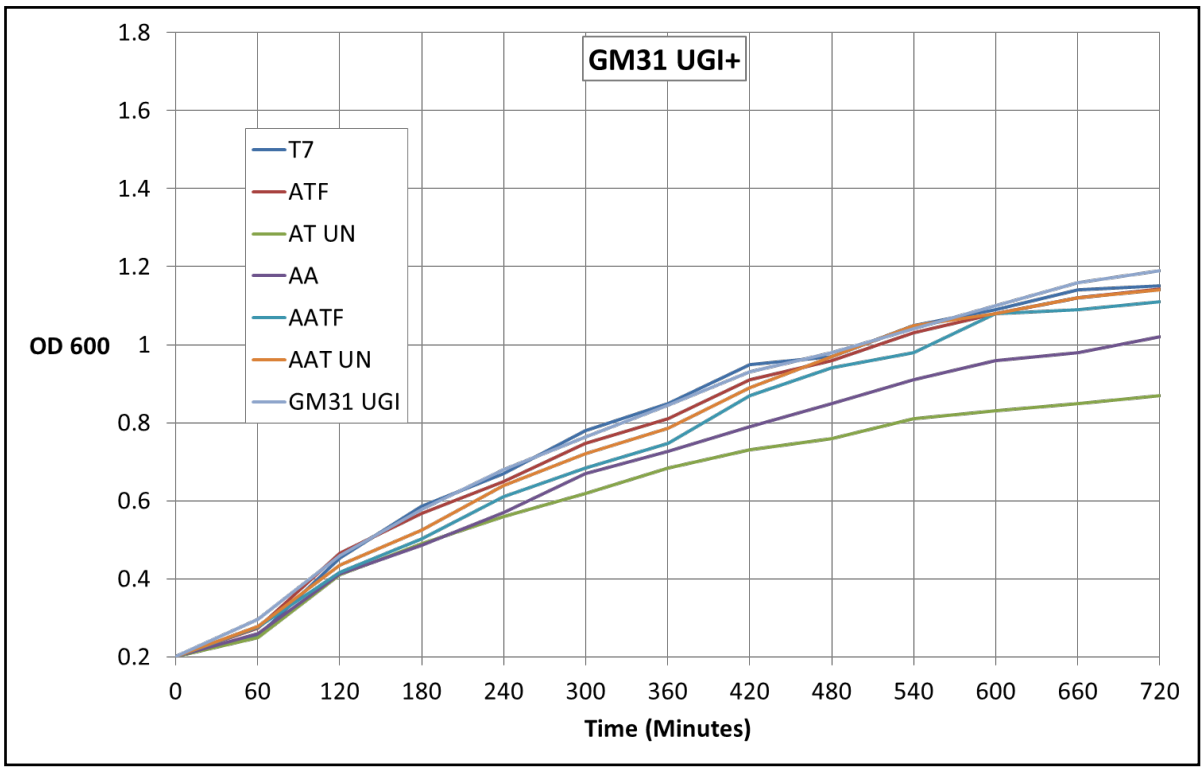


Figure 2.6 *- Growth curves demonstrated the viability of expressing the various mutator devices within the target E.coli GM31 host strain when co expressed with the repair inhibitor UGI. Cells were normalised to 0.2 and allowed to grow for 12 hours with OD600 measurements taken each hour. 8 biologically independent samples were measured for each mutator device and averages are shown here.*

To ensure that the target *E.coli* host can tolerate expression of the various mutator devices, a comparison in cell fitness between the wildtype GM31 and the repair deficient GM31 UGI+ strain was made by growing small volume cultures under inducing conditions. Overnight SOC+ cultures (non-mutated) were sub cultured into 96 well plates containing auto induction medium and allowed to grow for 10 hours at 30°C, with OD600nm measurements taken at hourly intervals on a BMI plate reader.

The lack of clear sigmoidal growth means it is somewhat difficult to calculate doubling time of these cultures, as measurements would normally reference a period of exponential growth. Although loose estimates can be made be based on the known rate of *E.coli* doubling within SOC medium, in which log phase continues until OD 0.6 – 1.0 (600 nm), this only provides qualitative comparison between strains. I have chosen to calculate doubling times based upon fitting a second order polynomial and taking the measurements across a broad time period that would encapsulate logarithmic growth, specifically OD 0.2 and OD 1.0. Estimates suggest that GM31 has a doubling time of 12.7 minutes, and GM31 UGI has a doubling time of 20.9 minutes.

The comparison of final OD's between both GM31 and GM31 UGI+ reveals a small reduction in cell fitness in the latter (***Figure 2.5 and Figure 2.6***) and is likely a consequence of the inhibition of BER

pathways, given that UGI will to some extent block the activity of UDG once it is expressed within the host. This is represented by a 30% decrease in final ODs (after 12 hours outgrowth) for all variants.

Expression of the mutator devices within the both strains does appear to slow growth rate (10 – 15%) compared to wildtype, likely due to high level expression of exogenous proteins from the plasmid constructs. However, I believe this metabolic burden to be minimal and I was still able to achieve a suitable final OD for downstream experiments.

In general, the growth curves do not show normal sigmoidal growth, and are in fact more drawn out with far shallow periods of logarithmic group. As such, it appears that cells are only just entering stationary phase at the 10 to 12-hour mark, slightly longer than expected. Although we would expect the basal level of mutations to increase during stationary phase, because this mutator system requires transcriptional activity it seems that this system will remain viable for overnight periods of outgrowth and I have since transferred strains through 5 successive rounds of growth.

An important observation is the reduction in cell fitness in the AID + T7 unfused device (AT UN) when expressed in the GM31 UGI+ strain, which is expected to show a mild mutator phenotype. Final ODs after day 1 were lower in comparison to the other variations of the mutator device (***Figure 2.6***) and by day 5 this strain was difficult to work with and rather sickly, with many individual replicates failing to grow above 0.5 OD600 (*data not shown*). I believe this to be a direct consequence of the off-target mutational activity and an accumulation of disruptive and undesirable mutations, which although not completely disastrous for cell survival, would clearly weaken the strain. There are numerous housekeeping genes which are essential for normal cell activity and may directly or indirectly effect growth rate and doubling time. One such study explored how every single gene knockout effected growth in both rich and minimal media (*Baba et al, 2006*). It is quite conceivable that a disruptive mutation in any essential gene could lead to a reduction in growth rate, regardless of the orthologous group the gene is placed within i.e. its functional categorisation. It should also be noted that expression of the AID T7 fusion (ATF) in either strains (GM31 or GM31 UGI+), does not appear to have any significant effect on overall growth rates.

2.4.2 Targeted Mutation Assay

To assess activity of various mutator devices on the target GFP gene, single cell expression profiles were analysed using flow cytometry. Samples were grown as above, with growth time extended to 16 hours. 20 µl of each culture was diluted in 200 µl of distilled water (1 in 10 dilution) into 4ml BD Falcon tubes compatible with flow cytometry. Alternatively, samples were suspended in isotonic PBS solution to maintain correct osmotic conditions, although no evidence of cell lysis was apparent when samples were stored for less than 4 hours.

Flow cytometry was carried out on BD FACScan machine operated by CellQuestPro software. Optimal strength of excitation (blue wavelength) was controlled by adjusting the voltage of the laser and was determined by using negative and positive controls prior to data collection. Emission filter was fixed at 510nm.

Data analysis was completed using the freeware 'Flowing Software 2' *(Terho, 2013)* or FlowJo v10.1 software package. The complete dot plot representing all events was used to gate for at least 10,000 true *E.coli* events and these were clearly distinguishable from blank controls (distilled water) to avoid potential background events. The histogram produced from true *E.coli* events revealed distinct subpopulations defined as GFP+ and GFP-. GFP positive controls (T7 construct) gave a strong fluorescent peak. The negative control (a blank GM31 GFPmut3b strain) did not (***Figure 2.7***).

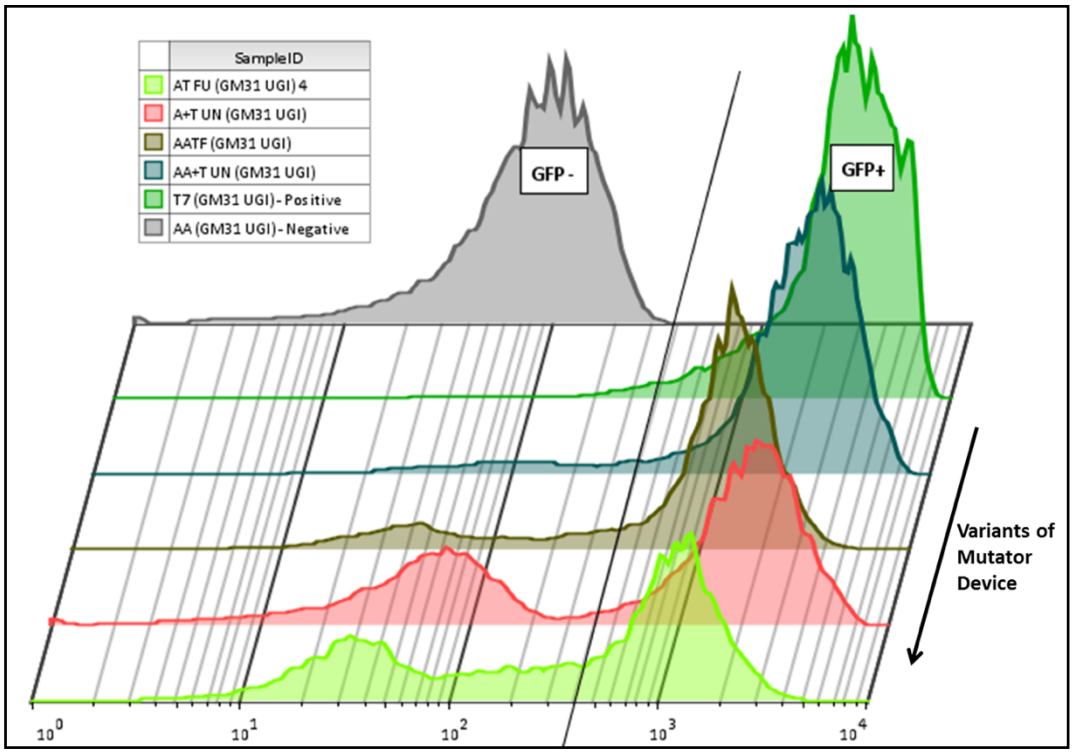


Figure 2.7 *– GM31 UGI+ strains were transformed with variants of the mutator device. Excitation (490nm) and emission (510nm) filters were used to generate fluorescence values plotted along the horizontal axis on a logarithmic scale. Positive control (T7) contains a plasmid that strongly express T7RNApol and leads to large GFP output . Negative control (AA) does not express T7RNApol and therefore no GFP is produced. These were used to define regions for both GFP+ and GFP- populations. Gating was consistent across all samples.*

Initial flow cytometry data appear to show an emergence of a GFP negative population, but it was decided to continue mutation through subsequent growth periods. Therefore, samples were allowed to grow for 16 hours before measurement by flow cytometry and were briefly held at room temperature. Each sample was then used to inoculate fresh AIM media and grown for a further 16 hours. This process continued for a total of 5 days.

As before, total cell counts from flow cytometry were declared either positive or negative events for GFP fluorescence. A summary of this data is provided for a complete 5 day outgrowth is provided (***Appendices Table 11.1 and Table 11.2***) and the percentage of GFP+/GFP- knockouts obtained for each device is shown (***Figure 2.8***).

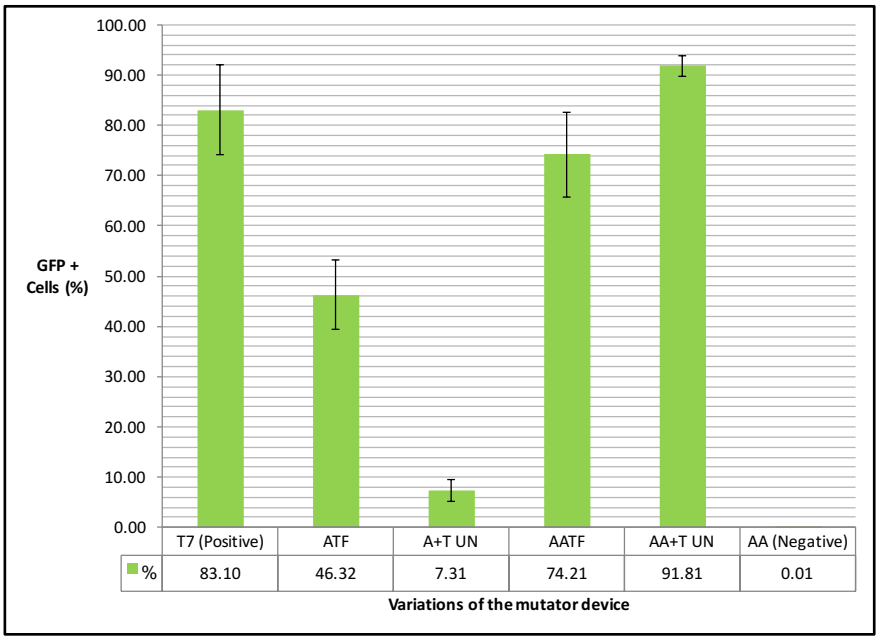


Figure 2.8 *- The average percentage of GFP + cells after 5 rounds of mutation. 8 biologically independent samples cultures were measured for each variant of the mutator device. T7 (positive control) gave high numbers of GFP positive cells as expected. AA (negative control) showed a negligible number of positive GFP cells, likely as a result of stochastic firing. Standard error bars are shown.*

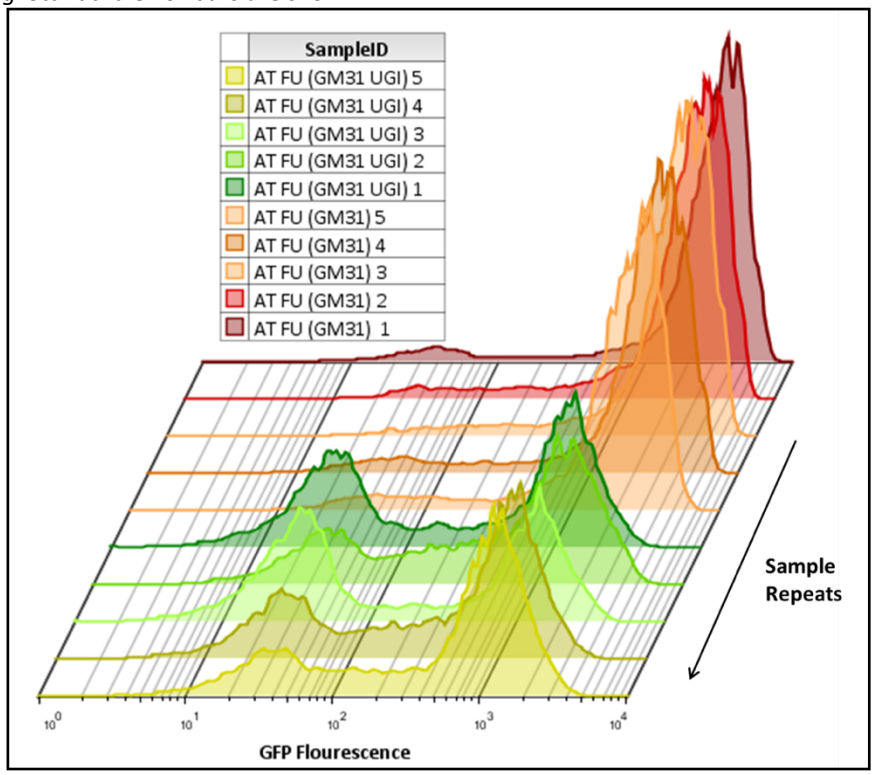


Figure 2.9 *– Mutagenic effects of expressing the ATF mutator device with the GM31 UGI strain. 5 biologically independent samples were analysed by flow cytometry and reveal a clear and distinct non-fluorescent population in the repair deficient GM31 UGI+ strain as reflected by obvious peak within the histogram trace.*

Initial results show clear differing mutational activity between variants of the mutator device. Specifically, the monomeric form of AID when fused or unfused to T7, appears to be produce an active mutator phenotype.

The negative control (AA) does not express T7RNApol and thus does not lead to GFP expression, which is controlled by the orthogonal T7 promoter. Clearly this produces a tight transcriptional response as very limited GFP fluorescence is ever seen in the negative control. There is a very small chance of stochastic firing which could account for very few GFP+ population counts.

The positive control (T7) shows strong GFP expression across the total cell population as expected. Over extended rounds of mutation, the GFP- population does begin to appear in the positive control, which might suggest an accumulation of a small number of background mutations, given that the GM31 UGI+ strain is repair deficient and that during very high levels of T7 guided transcription it has been suggested that mutagenic events are likely (*Beletskii et al, 2000*). Alternatively, it could also represent a residual population of dormant cells which have yet to undergo active transcription of the target GFP gene, possibly as a result of inefficient T7RNApol levels.

Importantly, the inhibitory activity of UGI in the repair deficient GM31 (UNG-) strain is also evident, as without co-expression of UGI, AID activity is virtually absent. This is represented by the similar measurements for T7RNApol only and ATF in the GM31 non-UGI strain, due to efficient base excision repair within the cell in which the majority of deamination events are repaired by UNG (***Figure 2.9***). There is also some resolution in relation to AID monomer and/or AID dimer activity. Whether fused or unfused to T7RNApol, AID monomer mutational activity appears much greater than AID dimer activity, as reflected by a minimal GFP- population (***Figure 2.8***)

This does not categorically state that AID is active as a monomer *in vivo*, given that dimerisation of two monomers of the protein could occur within the context of the cellular environment. Furthermore, because this approach creates a fused dimeric form of AID by placing a fixed linker sequence between both monomers, this may restrict the correct protein conformation and prevent the active dimeric form from forming. Protein dimerisation is perhaps only likely to occur if the individual monomers are able to fold correctly, and indeed an active dimeric form would require them to interact very specifically.

Previous work has described innate activity of AID in mammalian systems as markedly different to *in vitro* activity, once again suggesting that post translational modification may play a role. Furthermore, innate AID activity is known to create multiple cluster mutations with high sequence variance, something that has not been seen in any *in vivo* assay within prokaryotes. Therefore, it remains a strong possibility that post translational modifications may be essentially for this specific capability.

Importantly, I am able to identified a significant reduction in the number of GFP+ cells in the monomeric ATF fusion device (ATF) and unfused AT+UN device and therefore expression of the monomeric form appears to be most suitable for our overall purpose of building an *in vivo* system within an *E.coli* host. Interestingly, on initial review, the unfused device appeared to be the most active of all, with only 7.31% of cells showing GFP fluorescence after 5 rounds of mutation.

2.4.3 Multiple Rounds of Targeted Mutation

By extending the mutation assay over the course of 5 days, a distinct shift toward GFP- populations becomes more pronounced. This can be seen when overlaying histograms for ATF when expressed in the GM31 UGI+ strain (***Figure 2.11***).

This is very encouraging because it means that active mutation can be paused, before subsequent rounds of inducible mutation. Furthermore, there is no obvious effect on strain fitness when the ATF expressed, with cells remaining completely viable throughout, growing to similar cell density during each 16-hour growth period. It appears that both off-target and background mutation are of negligible consequence when this mutator is expressed.

This data collected across the extended time course is merely an extension of the data shown previously. Average populations for 8 biological repeats show that the ATF construct leads to significant GFP- population (54%) appears after 5 rounds. Furthermore, the lack of mutator activity when expressing the AID dimer fusion is also more pronounced after 5 rounds.

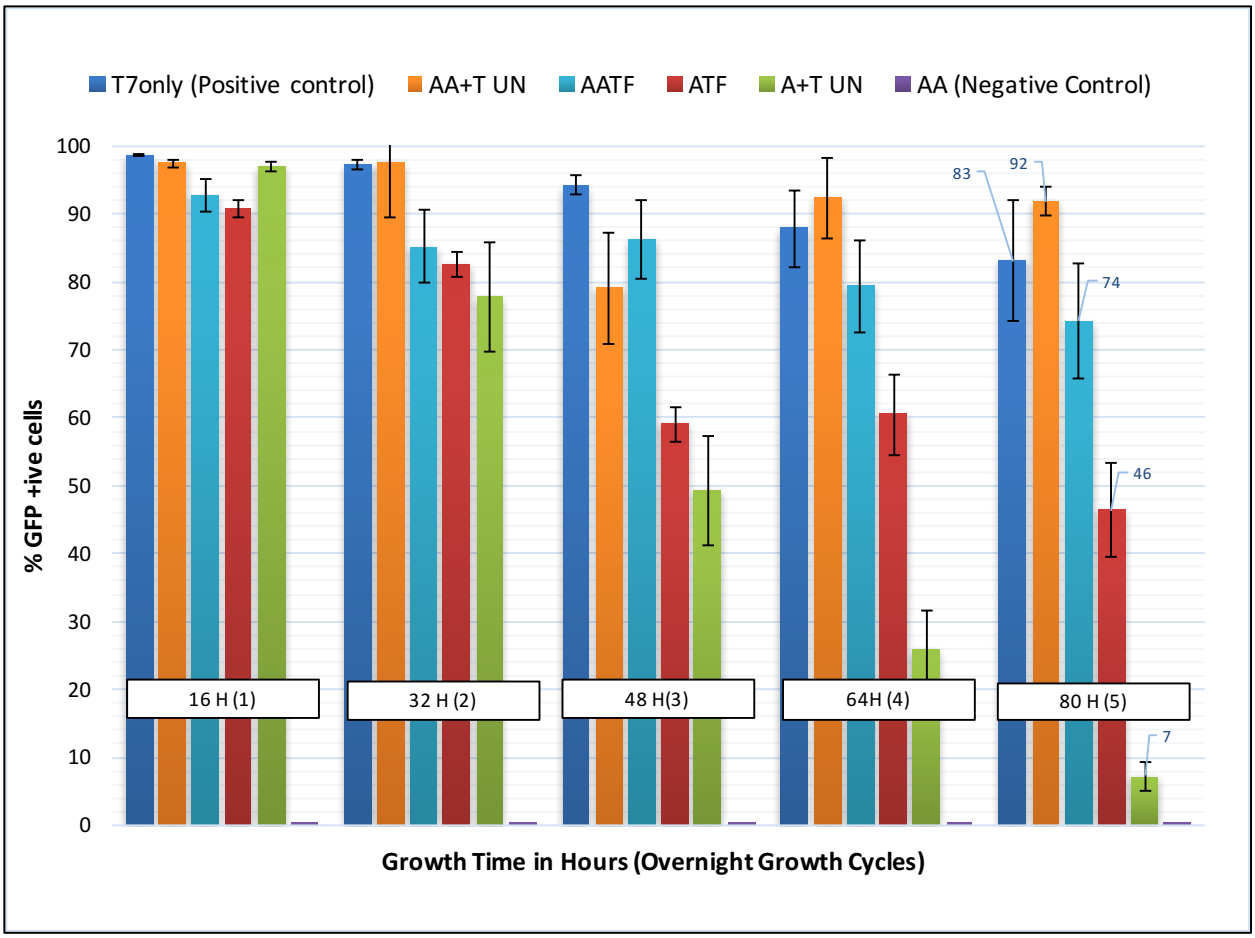


Figure 2.10 *- Changing percentages of GFP+ cells within each culture across 5-day of sequential outgrowth. A progressive decrease in the number of fluorescent cells is seen for all samples, with the most significant decrease seen for the A+T UN sample. Standard error bars are shown*

This extended growth period also served to highlight an interesting characteristic of the A+T unfused device. Certain samples expressing the AT UN mutator lose almost complete GFP fluorescence in a single round of mutation and thus the average results show a far greater standard deviation and error. A gradual decrease in fluorescence across a mixed cell population sample would be expected, given that the most likely scenario is that the gene will slowly accumulate random point mutations (***Figure 2.7***). However, a sudden and complete knockout seems extremely unlikely because GFP fluorescence is critically governed by the formation of a small chromophore region in the centre of the protein, meaning there are very few possible mutations which could affect these residues and result in complete remove fluorescence (*Cormack et al, 1996*). There is such low probability that a single mutation could result in this phenotype, I was alerted to the possibility that the GFP- population was actually a result of incorrect GFP expression, rather than expression of a mutated gene and protein variant that was non-fluorescent.

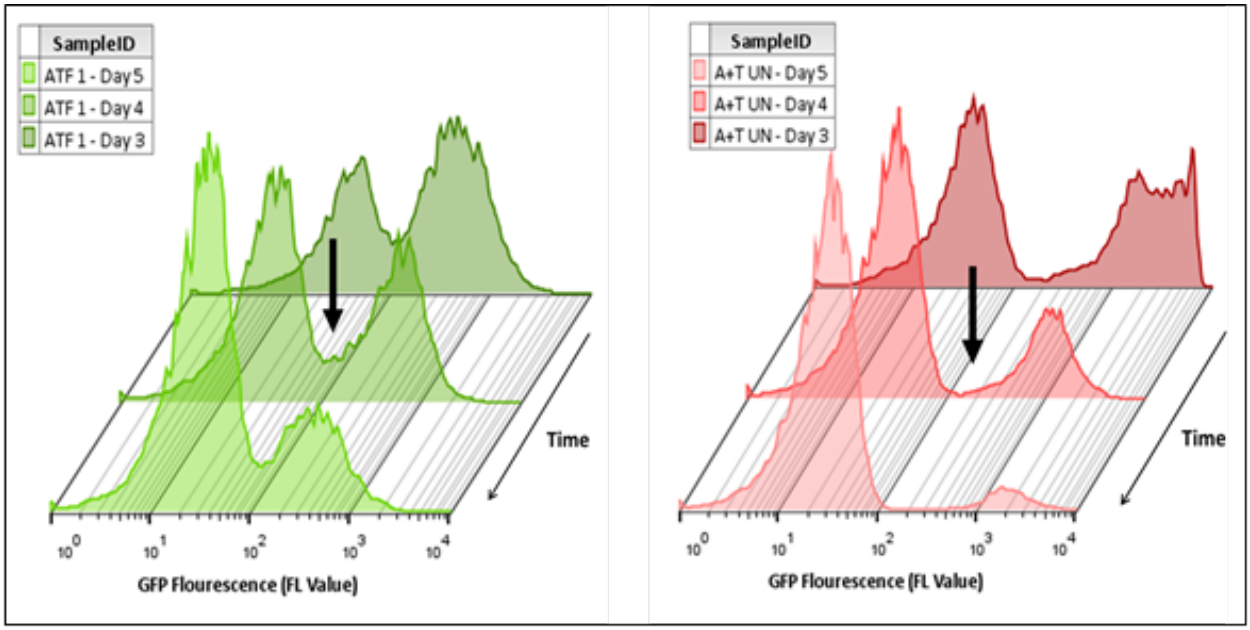


Figure 2.11 *- Multiple rounds of mutation were useful in highlighting a sudden drop off of GFP+ populations when expressing the A+T UN unfused construct (red traces) which persisted from very early on. This is in direct comparison with ATF fusion construct (green traces) which shows a differing population profile which includes a transition through intermediate or 'knockdown' GFP populations (black arrows).*

The most logical explanation was that the transcription of the GFP was interrupted due to mutations in the T7RNApol gene. My initial hypothesis was that the unfused AID variant was actually capable of mutating the T7RNApol gene itself, during relatively high levels of co-expression from the same plasmid, given that active transcription is a prerequisite for AID activity. This phenomenon is not evident when expressing the ATF construct and it was suggested that the unfused AID monomer is allowed to mutate any gene undergoing active transcription. A lac inducible plasmid system such as the one used here, will result in high levels of transcription that frequently expose plasmid ssDNA meaning it is susceptible to AID-induced deamination. Therefore, to elucidate the cause, sequencing of both ATF and AT UN was carried out (***Section 2.4.5***).

Regardless of this unwanted effect, it was possible to extend mutational activity and thus sample a larger sequence space by subsequent re-culturing.

2.4.4 Untargeted Mutation Assay

After the initial screening variants of the mutator device, the flow cytometry experiment was repeated with a smaller subset, namely T7, ATF, A+T UN in both the wildtype control strain (GM31) and the repair deficient strain (GM31 UGI+) across a 5-day period.

8 separate colonies were chosen for each sample. After 5 days of outgrowth, the samples were normalized to 0.3 OD600nm (2.4 x 10^8 cells/ml). 3ml of each sample was spun down to a pellet, the supernatant was removed, and the pellet was re-suspended in 100 µl of SOC+ media. Finally, 50 µl of

each sample was spread onto two separate rifampicin (100 μg/ml) LB plates to allow for clear separation between individual reversion colonies.

Due to its sensitivity to light, rifampicin LB plates were kept in complete darkness at all times and after plating cells, they were placed in 37°C room for 36 hours (rather than 16 hours) to allow reversion colonies to grow up to the extent that they could be visualised on a plate imager. Colony counting was simplified by using freeware software 'OpenCFU', and all plates were counted using the same parameters throughout *(Geissmann, 2013)*. Example plates are shown (***Appendices - Figure 11.1***). Total colony count for each sample plate was averaged across three biological repeats of the assay.

Plating and colony counting proved time consuming and problematic given the slow outgrowth of reversion colonies. However, growing high density cultures and plating a very high number of cells ensures that a meaningful comparison can be made with flow cytometry results, where it is technically possible to count hundreds of thousands of cells. This approach has provided a strong platform for quantifying background mutations.

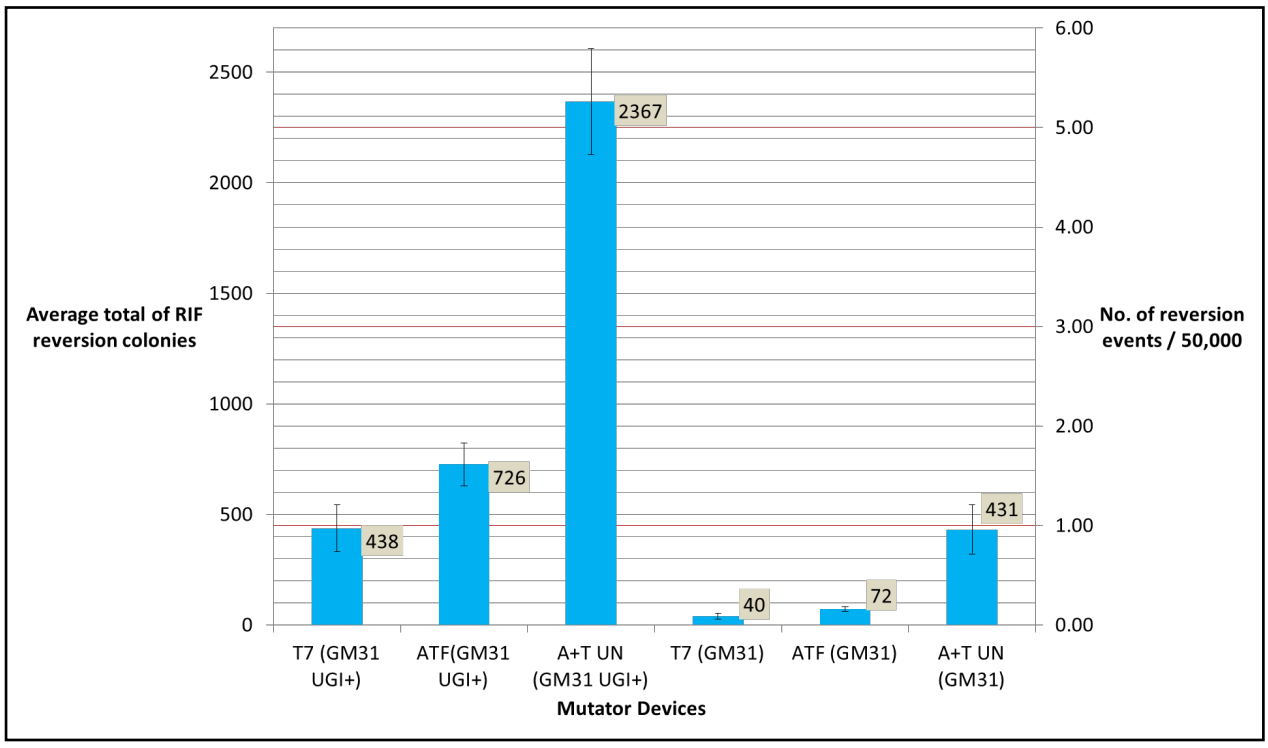


Figure 2.12 *– Rifampicin reversion assays were used to show the number of reversion colonies within samples that had undergone 5 days outgrowth. The number of reversion colonies was markedly high in the repair deficient strain (GM31 UGI) and even higher when expressing AT UN (unfused construct). The secondary axis refers to the number of reversion colonies per 50,000 events (cells counted) taking into account cell density at plating.*

Colony counts showed that the repair deficient strain GM31 UGI gave a higher number of reversion colonies when compared to the wildtype GM31 strain (***Figure 2.12***). This is perhaps as expected given the potential in background mutations above the basal level, or rather it has a weakly

mutagenic phenotype. In comparison, the WT strain shows a relatively low number of background mutations.

Expression of T7 is a useful control in this experiment and has been used to highlight whether it alone can cause a mutagenic phenotype in the GM31 UGI. Very few reversion colonies were identified and given that the wildtype strain showed a similar number of reversion colonies, this can be linked to basal background mutation.

Interestingly, ATF shows a very small increase in mutation rate when compared to the T7 control, but this appears negligible given the somewhat large error bars. Most importantly, expression of the A+T UN causes a significant increase in the total number of reversion colonies as a result of heightened level of background mutation. It is clear that by fusing the ATF it is possible to reduce the number of off target mutations by 70%.

2.4.5 Sequencing of GFP Target

To identify the frequency, type and consequence of mutations within the target GFP gene, a large set of clonal samples needed to be sequenced. After the final flow cytometry run, each sample was kept at 4°C to avoid further cell growth and preventing continued mutation. Both ATF and A+T UN samples, which had already shown to have produced large GFP- populations, were selected. Cultures were diluted 1 in 10,000 and re-streaked onto LB Carbenicillin (100 µg/ml). Fluorescence was not visualised at this point.

Amplification of the target GFP gene sequence from each individual sample was possible using colony PCR (***Chapter 9.2.2***). Two primer sets were designed which flanked the insert cassette; the outside primer pair gave a 1.1kb PCR product and the internal primer pair was then used for sequencing across the complete 1kb GFP cds. An example gel image is provided (***Appendices - Figure 11.2***)

PCR purification was carried out using Agencourt AMPure XP magnetic beads *(Beckman Coulter)* following the manufacturers guidelines, thus multiple reactions in parallel in a 96-well plate format. To ensure high quality reads, PCR conditions were adapted to reduce non-specific products and increase DNA yield. Sequencing was then outsourced to Eurofins MWG Operon.

My initial approach was to pick colonies which did not show fluorescence, however preliminary sequencing data showed significant bias. Given that there are a number of critical residues within the GFP protein target, non-fluorescent colonies are more likely to present with mutations in these positions or around these domains.

In fact, there were two common mutation events which appeared to propagate within the GFP-population these results (*data not shown*). Firstly, uneven weighting of mutations appeared very frequently within the first 30 bp of the sequence and often created a false stop codon. This would mean a short, truncated and non-viable polypeptide would be produced and thus reduce the translation

burden on the cell. If this type of mutation occurs within the first few generations during sigmoidal growth, there is a clear growth advantage and it will therefore propagate in subsequent generations and be highly represented. Secondly, and far less frequently, a small number of mutations occurred within the flurophore domain, which would likely result in incorrect folding of the barrel complex responsible for the excitation and emission at the intended wavelength. In this case, the complete protein mRNA would be produced, so the cells translational machinery would still be burdened by the same work load of producing a complete protein. It seems quite likely that the presence of chromophore mutations within GFP- populations is masked by the early loss of function mutations which can relieve burden of the cell. This early work confirmed that single colonies should be chosen at random, regardless of their fluorescent phenotype.

The potential location of the mutations was still of some concern initially, as it was quite clear that a change within the 5'UTR sequence, specifically within the T7 promoter sequence, could result in low levels of GFP expression, or prevent T7RNApol binding and eliminate transcription entirely, resulting in a large GFP- population. Furthermore, a mutation within the RBS could significantly reduce the level of translated protein and dramatically reduce the amount of protein within the cell. These unwanted mutations seemed increasingly possible in my chosen system given that viral T7 transcriptional behaves rather differently to endogenous RNA polymerases, forming much larger and more frequent transcription complexes. Therefore, if the complete 5'UTR sequence is exposed as ssDNA then it would be susceptible to AID deamination events.

A heat map showing all identified mutations, across all sample reads relating to the ATF fusion construct is shown (***Figure 2.13***). Very few samples showed any mutation outside of the GFP coding sequence and neither of the two mutations that occurred within the 5' UTR sequence were within the promoter or RBS consensus sequences. It was immediately evident that mutations are seen across the full length of the gene. A single sequencing trace, aligned to the original GFP target sequence, provides clear evidence of a T nucleotide base substitution (***Figure 2.14***).

The most significant and positive result, is the presence of multiple C-T or G-A mutations, the frequency and distribution of which are summarised in Figure 2.15.

Initially, one might expect that the reason for loss of fluorescence in a fluorescent protein would be due to mutations within the chromophore or those surrounding sequences that have previously been described as important for chromophore integrity (*Cormack et al, 1996*). A quick analysis of the amino acid changes caused by the C-T point mutations introduce residues with alternative charge or polarity and thus would be likely to cause conformational changes in the resulting GFP polypeptide (***Table 2.3***). This fact, in addition to absence of mutations in the chromophore indicate that GFP

‘knockouts’ are more likely to be a result of poor solubility or issues with protein folding and are in reality ‘knockdown’ events.

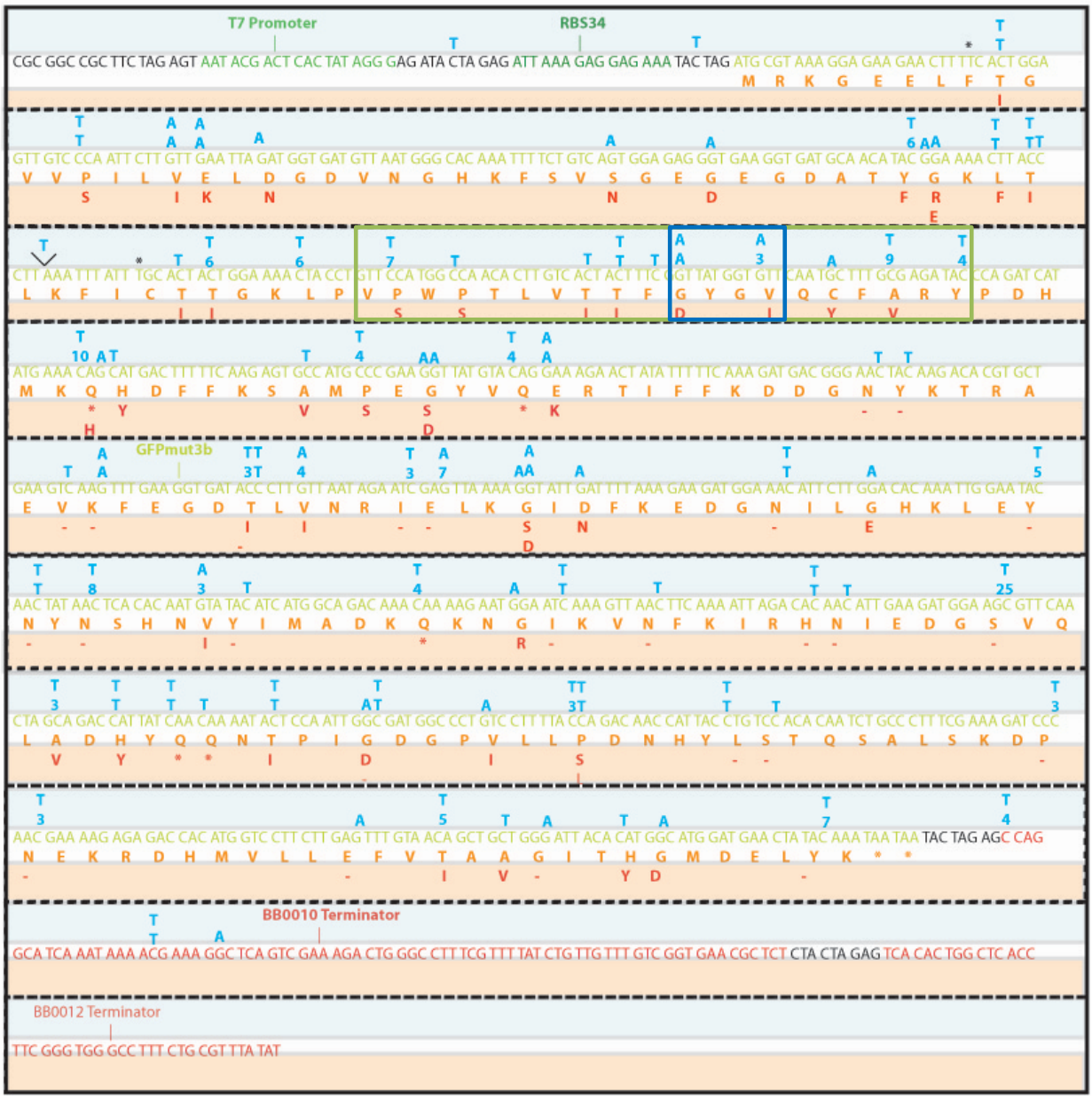


Figure 2.13 *- A detailed mutation map provides a cumulative summary of the mutations identified within the genome integrated GFP target gene and flanking 5'UTR and 3'UTR regions. Purified PCR products generated from a direct colony PCR of 120 samples were used for traditional Sanger end point sequencing. Subscript numbers represent the number of times that a particular mutation event was shown to occur at that specific nucleotide. Sequence features are colour coded; T7 promoter (Turquoise), RBS (Dark Green), GFP cds (Light Green), Terminator (Red). C-T and G-A mutations correspond to deamination events caused by AID activity. Importantly there were no mutations present in the T7 promoter sequence and very few mutations were seen outside of the GFP coding sequence. The GFP chromophore region is boxed (dark blue) and another region considered important for fluorescence of the protein is also boxed (green) and mutations within this region would have an increased chance of fluorescent knockout. Note that only 112 reads were obtained. Failures were likely due to a low PCR product concentration.*

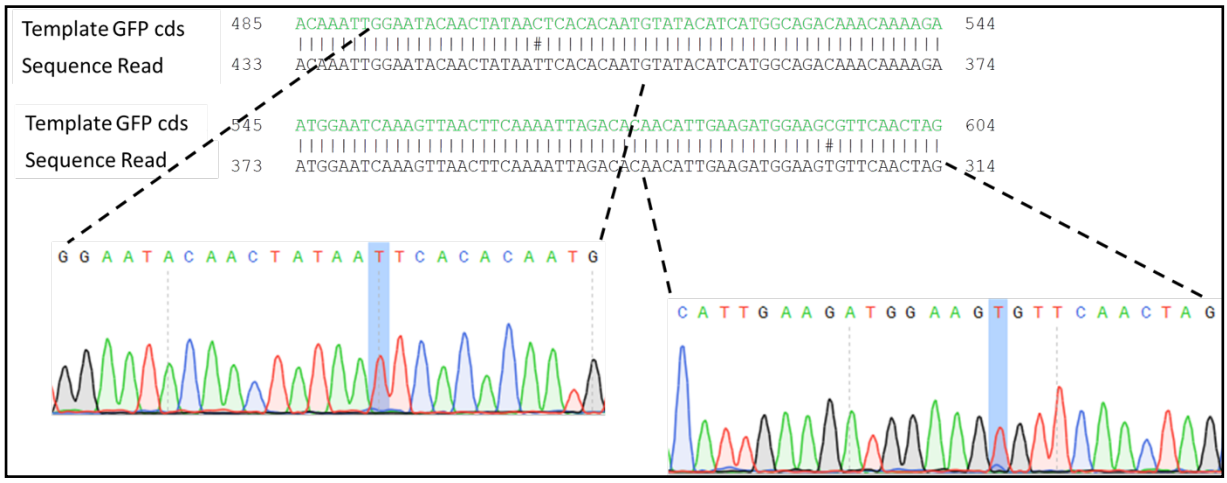


Figure 2.14 *– Example sequence trace showing evidence of C-T transition mutations. The green sequence refers to the template wildtype GFP sequence and the black sequence was a sequence read generated for a single colony PCR.*

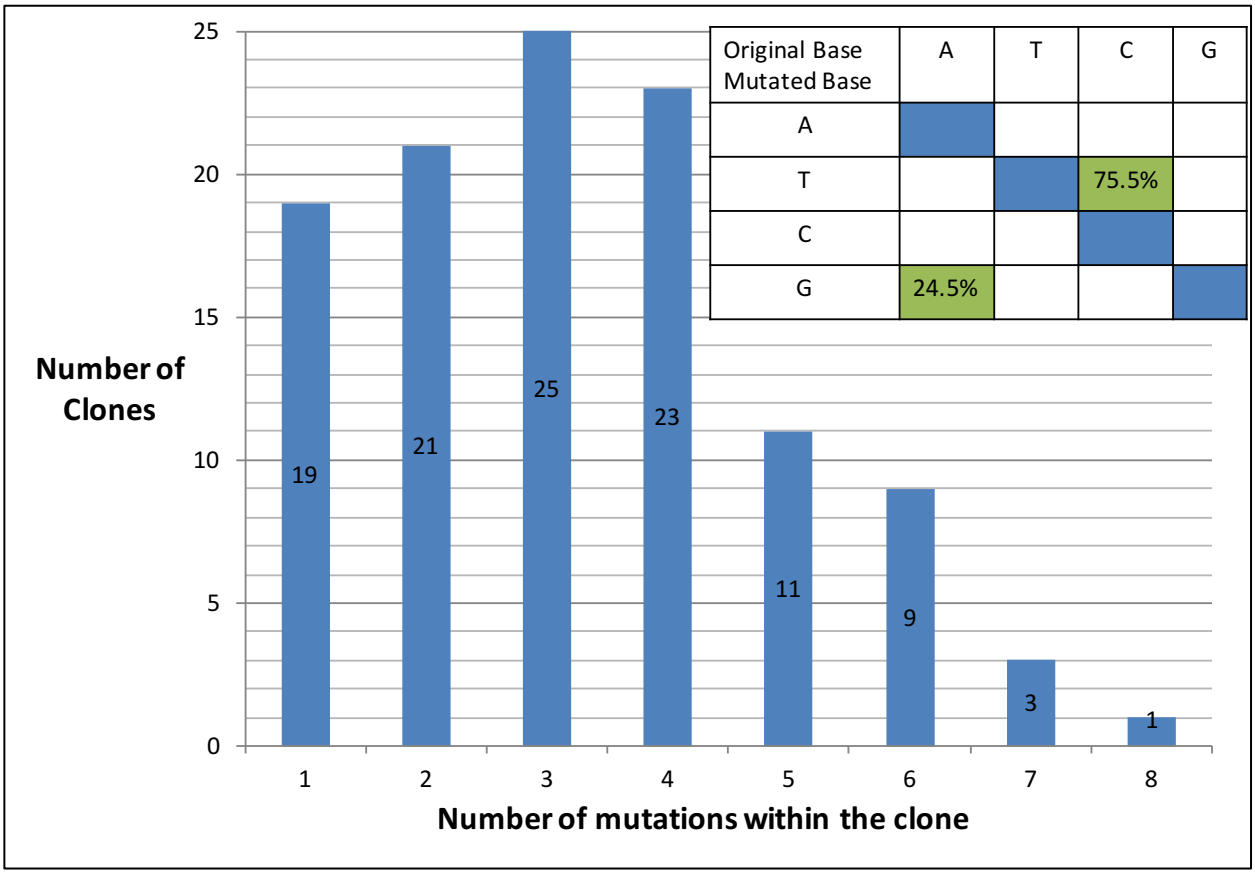


Figure 2.15 *– The average number of mutations in any single clone shows fairly wide deviation. The cumulative total suggests that the mean average number of mutations within any one single sample is 2.3. The inner set box chart describes the general position of mutations and indicates an increased probability that the deamination event occurs on the transcribed strand (192 C-T, 62 G-A). Total mutation = 254, an average of 2.3 mutations per clone.*

Table 2.3 *– Description of residue changes identified with target GFP gene*

Characteristics (Side Chain)	Resulting AA	Characteristic (Side Chain)	Disruptive Change	Frequency
None	S	Polar Neutral	Maybe	13
Non-Polar Aliphatic	I	Non-Polar Aliphatic	No	11
Polar Negative	K	Polar Positive	Yes	4
Polar Negative	N	Polar Neutral	Maybe	2
Polar Neutral	N	Polar Neutral	No	1
None	D	Polar Negative	Maybe	7
Non-Polar Aromatic	F	Non-Polar Aromatic	No	6
None	R	Polar Positive	Maybe	2
None	E	Polar Negative	Maybe	2
Non-Polar Aliphatic	F	Non-Polar Aromatic	Maybe	1
Polar Neutral	I	Non-Polar Aliphatic	Yes	25
None	Y	Non-Polar Aromatic	Maybe	1
Non-Polar Aliphatic	V	Non-Polar Aliphatic	No	14
Polar Neutral	H	Polar Positive	Maybe	1
Polar Positive	Y	Non-Polar Aromatic	Maybe	4
None	S	Polar Neutral	Maybe	1

2.4.6 Sequencing Off-Target Mutations in T7RNApol

As mentioned previously, the mutator relies on high level active transcription carried out by T7RNApol. Therefore, expression of viable T7RNApol is critical to the activity of the system. In order to determine whether the active mutator phenotype, seen during expression of the ATF and A+T UN mutator devices, was indeed the consequence of GFP knockouts, a small number of variants were tested for the integrity of the T7RNApol gene itself.

Firstly, the same cultures analysed in ***Section 2.4.5*** were plated out and GFP- colonies were identified by imaging on a fluorescence plate reader. Given that T7RNApol was expressed from a plasmid, multiple copies of the T7RNApol gene are infact present. Given the considerable size of the T7RNApol gene (3.2kb) multiple overlapping reads were required to achieve full sequence coverage, only 5 clones for each construct were sequenced in their entirety to reduce costs. Non-fluorescent colonies were picked, grown and plasmid DNA was mini-prepped.

Interestingly, multiple mutations were identified in the T7RNApol gene within the A+T UN construct (***Appendices - Figure 11.3***). The underlined CCT codon identifies an unknown codon change already present in the T7RNApol gene and the C-T change (underlined red) was also an original artefact which is in fact a silent mutation.

I believe this is a consequence of the unfused AID having the potential to deaminate any ssDNA target on any actively transcribing gene. The T7RNApol gene which is induced in our assay is naturally undergoing a high rate of transcription on a plasmid.

There are a number of possible reasons for such off-target mutation. Firstly, the relaxed nature of plasmid-based transcription, compared to the compact and highly folded genomic DNA, where the target gene is positioned, might make it more accessible to the AID based mutator. One suggested model suggests plasmid DNA is held in a perpetual phase of replication in which ssDNA is continually exposed and thus free to interact with the AID protein (*del Solar, 1998*). Another explanation is that because there are multiple copies of the plasmid within each cell there is ample opportunity for single for a divergent library of mutants to exist. If a certain mutations provides a greater fitness advantage for that specific cell they will win out and thus propagate and outnumber the less fit variant in subsequent generations. This can also be coupled to the idea that high levels of T7RNApol production may be toxic whilst virulent levels of transcription and protein translation become a significant burden to the cell. Put simply, disruptive mutations in T7RNApol will provide a fitness advantage given that the cell is no longer forced to produce large quantities of GFP, meaning that non-functional T7RNApol mutants are able to propagate through to later generations.

These conclusions are consistent with the actual mutations and subsequent residues changes identified within T7RNApol as some cause alterations to key domains that are responsible for the functional activity of the protein i.e. the active site (***Table 2.4***)

Whilst there is still a spread of mutations in A+T UN unfused T7RNApol gene, there are three mutations in particular that are present in all the GFP knockout samples. Having explored the T7RNApol protein structure in greater detail (***Chapter 6.3***) it appears highly likely that the H.623.Y and P.665.L mutations in combination would completely destroy the primary function of any RNA polymerase since they are residues located proximal to the active site. Therefore, these non-functional mutants could not transcribe the GFP gene and mean there would be no recordable GFP fluorescence. Once again, the prevalence of these mutations corresponds to the idea that they are advantageous to the fitness of the cell. If they occur in early generations, they would be become highly represented throughout the population I subsequent generations.

Importantly, the AT fusion mutator device showed no evidence of disruptive mutations within the T7RNApol gene, and this is likely explained by the fact that the fixed linker, which tethers AID to the larger polymerase protein, and prevents it from accessing ssDNA at anywhere other than the T7RNApol related transcription bubble.

Table 2.4 *– Identified mutations in T7RNApolymerase and corresponding residue changes in the T7RNApol gene in the AID+T7 unfused device. Silent mutations do not affect T7 transcriptional activity. All other mutations may potentially cause negative affects to transcriptional efficiency. Starred mutations were seen in all GFP negative substitutions.*

Identified Mutation	Amino Acid Change	Location	Description of domain
GGC – GAC	Gly – Asp	G.97.D	A-T Recognition Loop
CGC – CGT	Arg – Arg	R.99.R	A-T Recognition Loop
AGG – AAG	Arg – Lys	R.389.K	Thumb
GGT – AGT	Gly – Ser	G.449.S	Palm
CGC – TGC	Arg – Cys	R.478.C	Palm
GAG – GAA	Glu – Glu	E.565.E**	-
CAC – TAC	His – Tyr	H.623.Y**	Finger – residues proximal to active site
CCG – CTG	Pro – Leu	P.665.L**	Finger – residues proximal to active site
AGC – AAC	Ser – Asn	S.789.N	Palm
AGC – AGT	Ser – Ser	S.789.S	Palm
CTG – TTG	Leu – Leu	L.824.L	-
TTC – TTT	Phe – Phe	F.845.F	-

2.4.7 Fluctuation Analysis

When deciding on a method for fluctuation analysis, it is important to first consider the differing number of mutational events counted across all parallel cultures, for all variant samples. This is the most likely cause of inaccuracy when using such presumptive calculations, and therefore a model which fits the anticipated median distribution of the data should be chosen.

An interesting example of this inaccuracy is seen when comparing the studies of AID mutational activity by calculating the number of reversion mutants using a Kanamycin plating assay (*Sohail et al, 2003., Ramiro et al, 2003*). These publications gave respective mutation rate estimations, in direct response to AID expression, of 50-fold and 20-fold above background. The large discrepancy can be linked to a use of a primitive method for fluctuation analysis in *E.coli (Lea and Coulson, 1949)*. This approach makes a huge number of assumptions, specifically that: cells are under exponential growth, previous mutations due not affect subsequent mutation events, the probability of a mutation is constant across the cells lifetime, mutants and non-mutants have similar growth rates, reverse mutations happen so infrequently that are not considered, that every individual mutant is detected and that no further mutations occur once the assay is stopped. Although good experimental procedure can help fulfil the requirements of the model, in practical terms this is a very long and time-consuming process and I can quite confidently say that the chosen assays are not entirely suitable. For example, it is quite clear that off target mutations will affect growth rate and that outgrowth periods are extended to the point where cells enter stationary phase (where mutations become more frequent).

Given that I had not been able to reliably predict the number of expected mutation events (or the anticipate the number of colonies) during the gain of function rifampicin reversion assay, the most

suitable method for fluctuation analysis is considered to be the Ma-Sandri-Sarkar approach (MSS) (*Sarkar et al, 1992*). This is a regressive model that uses the distribution range of a hypothetical number of mutational events, which is then fit against the real observed data.

There are also a number of software packages available that provide a simplified way to analyse final data sets using MSS. For example, the 'FALCOR' package offers a template spread sheet for data input to generate mutation rates based on simple frequency, Lea-Coulson and MSS methods (*Hall et al, 2009*). Furthermore, the 'bz-rates' package implements a generalised version of MSS that considers the changeable growth rate of the mutant and wild type strains (*Hamon and Ycart, 2012*).

Nevertheless, given the focus on identifying global population changes using the loss of fluorescence flow cytometry assay, coupled with the lack of a targeted reversion assay, it was agreed that any fluctuation analysis would be avoided.

2.5 Conclusions

There are a number of conclusions that can be drawn from this initial work which principally validates the mutator concept.

Firstly, the GM31 (UGI+) repair deficient strain demonstrates a slightly lower cell fitness as reflected by slower growth rates. Background mutations which occur in GM31 (UGI+) strain have the potential to be disruptive to overall cellular function, but in reality the mutator strain retains relatively high cellular fitness even over 5 successive growth periods which suggests that it can tolerate the increased mutator activity (***Figure 2.5 and 2.6***).

Secondly, it was immediately clear that overexpression of AID within our designed strain resulted in a heightened mutagenic phenotype. Increased mutational events were visualised as increasingly non-fluorescent sub-populations, likely a consequence of the mutations within the target GFP gene. Targeted mutations occur at a significantly higher frequency than basal level of background mutation (***Figure 2.8***).

Whilst I have not attempted to establish a fixed rate of mutation within this work, the targeted assay confirms AID induced deamination events which are verified by sequencing results (***Figure 2.13***). Multiple rounds of mutation have also been used to emphasise the effect of an active mutator phenotype and reveal that the ATF fusion is indeed the most active of all mutator variants tested (***Figure 2.11***). In an effort to reduce non-specific and off target mutation events, mutational activity can clearly be directed to sites of active transcription. This is reflected by a lower level of off-target mutation in the ATF fusion construct, which remain close to the levels shown in negative controls.

I have noted that previous work to link T7 transcriptional systems with AID activity, did not manage to reproduce the typical hypermutation events seen when using endogenous RNA polymerase

for transcription in an *E.coli* host *(Shen et al 2005., Besmer et al, 2006)*. I believe this to be advantageous, because viral T7RNA polymerase is likely to generate extensive and highly active transcription bubbles, transcribing at rates 20 times faster than *E.coli* RNAP *(Uptain et al, 1997)*. Furthermore, the key targeting feature provided by a highly specific T7 promoter-T7RNA polymerase is essential for specific mutation of the genome integrated target.

Within the context of this initial study, GFP has been a very useful selection output, given that it can be visualised easily, making it possible to measure 10^6 within 5 minutes using flow cytometry. Establishing a high throughput assay which can reliably screen and measure targeted loss of function was essential.

However, in reality, GFP output expression is a fairly poor measure of mutation rate, given that only certain mutations (or changes in gene sequence) are directly linked to the observed phenotypic change. It fails to provide any immediate information about the context or type of mutation that has been incurred. Essentially, mutations within GFP could result in a partial reduction or increase in fluorescence, shift its excitation and emission spectra, change its folding capabilities and solubility. Single mutations that do not occur at its fluorophore centre may still affect the strength of the output signal. Consequently, GFP output selection can never be used to accurately measure mutation rate in the context of a loss of function assay.

Therefore, it would be preferable to construct an assay which has a binary output, i.e. a yes (1) or no (0) response. In terms of a biological assay, this would possible by using a 'survive or die' selection method e.g. antibiotic resistance.

The rifampicin reversion (gain of function) assay requires only a single mutation for the cell to become resistant and is easily identified as a positive event when selecting on an antibiotic. Although expression of AID increases mutation rates above the basal background level, a much more pronounced difference is observed when we consider the actual number of cells that are counted in both assays. When the results are normalised to 50,000 events, given that this was the maximum number of cell counted by flow cytometry and yet hundreds of thousands of cells can be screened in the plating assay, we quickly begin to understand how potent the targeted mutator device can become. A second Y axis provides comparable scales between both the targeted and non-targeted assays and relates to the number of mutations seen per 50,000 cells counted (***Figure 2.16***).

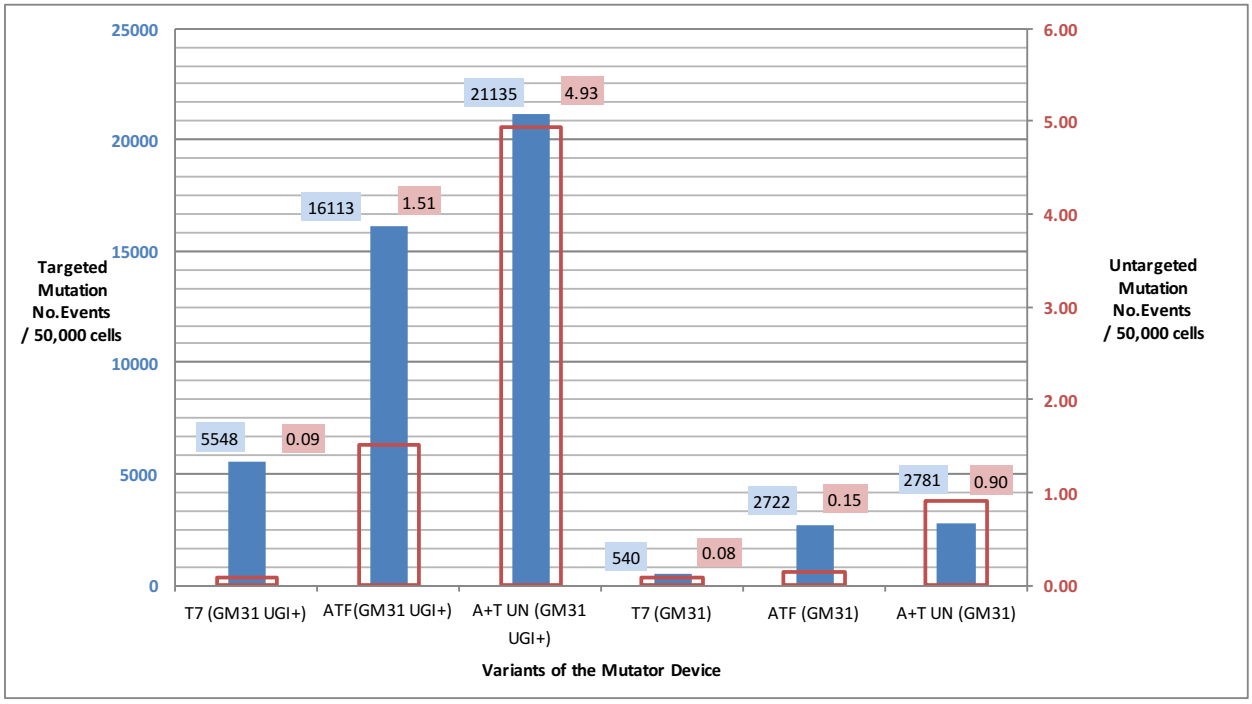


Figure 2.16 *– Comparison of results from both targeted and untargeted mutation assays when testing variants of the mutator device. Targeted loss of function assay (left Y axis) considers the total number of cells counted by flow cytometry. Untargeted gain of function assay (right Y axis) considers the total amount of cells plated on antibiotic. In this experiment, T7 is considering the negative control, given that high transcription would potentially increase mutation rate.*

Throughout this initial work I was unable quantify mutation rate. There are a number of reasons for this, but the major problem was that the targeted mutation assay does not provided a definitive read out of single mutagenic events. Qualitatively speaking, any comparison made between targeted and non-targeted is over-conservative and will actually reflect a lower level of targeted mutation. This is because cells categorising cells into GFP+ or GFP- pots does not consider silent (same amino acid) or non-disruptive (amino acid change which has no bearing on protein) mutation events. This has likely resulted in an underestimation of targeted mutator efficiency and efficacy. Loosely speaking, an underestimation is not of great concern, given that overall mutation rate is low. In fact, the ATF device results in a low rate of mutation which is somewhat useful in the wider context as the intention is to embed the targeted mutator within a system which is able to select for interesting mutants and implement self-control. Therefore, it would be preferable that only a handful of mutations within each cell variant prior to undergoing transcription and translation of the protein. Gradual stepwise changes with a lower mutation rate are actually very favourable to the hypothetical *in vivo* system.

It should be noted that there appears to be a small preference for the coding strand given the prevalence of C-T changes rather than the corresponding G-A changes that would be a consequence of deamination of the template strand (non-coding). Previous work has suggested that there is no clear

preference for either strand during *in vivo* expression of AID (*Chaudhuri et al. 2003; Shen et al. 2005).* However, the engineered ATF fusion protein is likely to restrict AID interaction with exposed ssDNA, with the coding strand exposed for longer periods, given that the template strand will likely forming hybrid interactions with the emerging mRNA. In fact, *in vitro* studies that have attempted to link AID activity to viral T7 transcription have shown a preference for the coding strand (*Pham et al, 2003*).

When considering the viability and usability of the ATF mutator system at this stage, there were a number of reason to reflect positively on its potential application. As mentioned, the relatively low number of mutations should enable the wider system to select favourable mutants. Furthermore, the potential to pause mutation by holding cells within a dormant state may be beneficial for ensuring that all variants within the library are expressed. Currently the lac inducible system appears somewhat leaky, and to truly achieve this switch-like activity, I have attempted to improve promoter/operator and inducer/repressor associations (***Chapter 4***)**.**

3 BASIC Assembly

3.1 Background

Biopart Assembly Standard for Idempotent Cloning (BASIC) first emerged within my research group at the very beginning of my doctoral studies and given my previous experience in DNA assembly, I have been heavily involved in its progression and evolution.

Early discussions within our team quickly established the need for a universal DNA assembly method which was efficient, reliable, robust and automatable, whilst being easy and accessible for new users. Previous work to establish an assembly standard had had very limited success and it was agreed that it would be to the benefit of not only my project but also our whole research group. In fact, the wider Synthetic Biology community as a whole would benefit from a robust method and it was one of the primary research goals of the Centre for Synthetic Biolog (CSynBi) at Imperial.

Synthetic biology aims to characterise composite genetic parts as individual components that can be combined together to form complex yet predictable biological systems. In recent years, the ability to quickly and efficiently read DNA has enabled detailed understanding of nucleotide sequences and there are widely accessible registries which detail genetic elements in order to provide the researcher with qualitative and quantitative analysis of their function i.e. promoters, repressors, enhancers, terminators, protein coding sequences and reporters. The falling cost of DNA synthesis is driving the field forward at an exciting pace with the potential to dramatically reduce manual labour and explore well beyond natural occurring genetic sequence and into the realms of complete *de novo* design.

However, the fact remains that researchers are still required to piece together these modular parts. It would be far more beneficial if they were able to focus on experiment design or data analysis rather than repetitive manual cloning techniques. The ability to quickly navigate through the design, build, test cycle provides further opportunity for iterative improvements and learning from each round.

Therefore, it is essential that foundational and enabling technologies, such as DNA assembly, are improved upon. Although sizeable stretches of the DNA can now be synthesised quickly and cost effectively (1kb dsDNA fragment ~£60), our ability to create larger constructs, such as complete expression plasmids (>5kb), is still limited by the inherently low fidelity of solid phase DNA synthesis. The traditional approach of synthesising small oligos at high fidelity and then combing these together through hybridisation methods is still susceptible to INDEL mutations and errors. Whilst companies have attempted to lower the cost of DNA synthesis by CHIP base synthesis or improve fidelity by using proof reading and error repair enzymes, the need to be able to assemble multiple parts of 500 bp – 1 kb in length remains essential.

We set out to develop a novel standard that was able to overcome many of the limitations of the existing assembly methods e.g. Gibson Assembly (PCR-based) or Golden Gate (restriction-based). Significant efforts were made to ensure that the protocol was as experimentally simple with the low costs, needing only common and accessible laboratory reagents.

This work resulted in the publication of the BASIC standard last year (*Storch, 2015*). This highly efficient and enabling methodology has become an accepted standard across the wider department and is currently in use by a number of close academic collaborators, being trialled by industry partners, taught as part of the undergraduate biochemistry and postgraduate synthetic biology degrees and has very recently been adapted for use within SynBiCite's DNA foundry, who have developed an optimised version of BASIC as the chosen method for DNA assembly for integration onto its state of the art robotic platform.

Although a significant amount of time was spent developing the method it has been a particular rewarding experience and has since become a great advantage when carrying numerous cloning steps required prior to the experiments outlined in this report (***Chapter 4, 5, 6***). This capability is very desirable when constructing complex and multi-faceted system such as the directed evolution circuit proposed within this report.

Here I provide a summary of the steps undertaken in the design, testing and implementation of the protocol. I intend to highlight my personal involvement, the numerous troubleshooting steps and improvements I have been responsible for and real applications applicable to my own research goals.

3.1.1 DNA Assembly Standards

A number of varying approaches for DNA assembly have been proposed and these have been extensively reviewed by my colleagues *(Casini 2015, Ellis 2011)*. I do not intend to reproduce their detailed descriptions here, but the numerous methods can be sub divided into two main categories; 1) restriction-ligation reactions e.g. BioBrick standard or Golden Gate assembly 2) PCR based overlap techniques e.g. CPEC or Gibson assembly.

Restriction enzymes were fundamental to the advancement of molecular biology and enabled the rational and uni-directional cloning of exogenous DNA from one organism to another but are dependent on highly specific recognition sites to cut DNA in correct positions and provide short overhangs for annealing to adjoin sequences. Even though there is now a very large repertoire of restriction enzymes, each with their own well-defined and specific restriction site, it remains highly unlikely that the target DNA position of interest would contain any matching restriction site. PCR preparation of parts would enable the addition of desirable restriction sites, for subsequent cloning, but this is hugely time consuming and is worsened by the potential presence of the same motif appearing as non-target cryptic sites elsewhere in the DNA sequence.

The ability to access a vast catalogue of restriction enzymes is only so helpful, and in fact the ability to share a vast registry of biological parts would actually be more beneficial. It is this mantra that saw the establishment of the BioBrick standard which became widely used during the early stages of what we now recognise as the field of synthetic biology (*Knight, 2003., Shetty et al, 2008*). Its popularity was built on ideas of standardisation and the principle of idempotency i.e. that each part retained the same flanking prefix and suffix restriction recognition sites, meaning that multiple parts could be prepared in the same way and assembled in exactly the same format. Furthermore, parts can be reused and will retain independence even after assembly, providing useful modularity. Parallel assemblies and subsequent rounds of cloning allowed for construction of complex devices in a hierarchical fashion.

However, a multipart construction using this format can be considered labour intensive and time consuming because only two parts can be combined in each reaction. Furthermore, retention of the prefix and suffix sequence themselves introduce 'scar-like' sequences which could potentially have negative implications to context dependent parts and would be unwanted in certain circumstances such as during complete genome construction. The use of four different restriction sites means that these 6bp sequences are then forbidden at all other locations within the various parts and need to be eliminated which requires a significant amount of work.

In more recent years, the Gibson isothermal assembly method has grown in popularity (*Gibson 2009, Gibson 2011*). This approach requires each part to have overlapping dsDNA sequence at its flanks, corresponding to the adjacent parts upstream and downstream and therefore designed changes can be made to the overlap itself. It also eliminates the problems caused by so-called forbidden sequences. Each overlap region should retain strong specificity for its corresponding part, ensuring the accuracy of the assembly is maintained. A single one-step assembly of multiple parts also appears advantageous over the 2-part BioBrick assembly. However, each construction requires bespoke design, with each overlap having to be tailored to suit the intended application. Overlapping sequences can be added using simple PCR extension but there may be concerns over accuracy and fidelity for larger sequences. This concept of overlap assembly has been extended to include a collection of widely used modular parts in the GAMP protocol (*Akama-Garren et al, 2015*). In this example, 5 common overlaps were designed and added to parts in order to dictate their position in the final construct. Therefore, the user can choose from a collection of promoters, genes and vector backbones with reliable overhangs that will anneal with high specificity upon enzymatic processing (in keeping with the Gibson protocol). A likely issue here would be a rather significant 30 bp scar region, which is not functionalised.

Linker based approaches have also been developed to help standardise target annealing sequences such as the MODAL approach (*Casini et al, 2014*). However, the reality is that PCR reaction, which are directly influenced by the presence of non-specific sites or secondary structure formation, are

inherently difficult to standardise and the purity of products cannot always be maintained. Such an approach is not compatible to automation if it requires DNA size separation and purification steps.

3.2 Materials and Method

3.2.1 Overview

BASIC assembly aims to be as widely applicable and as robust as possible, with flexibility and adaptability in mind. It relies on the use of integrated prefix and suffix sequences, containing a Type IIs restriction site which is recognised by the enzyme BsaI. This enzyme cuts outside of its specific target sequence, and therefore the overhang that left is entirely dependent on the sequence downstream of the recognition site (***Figure 3.1***).

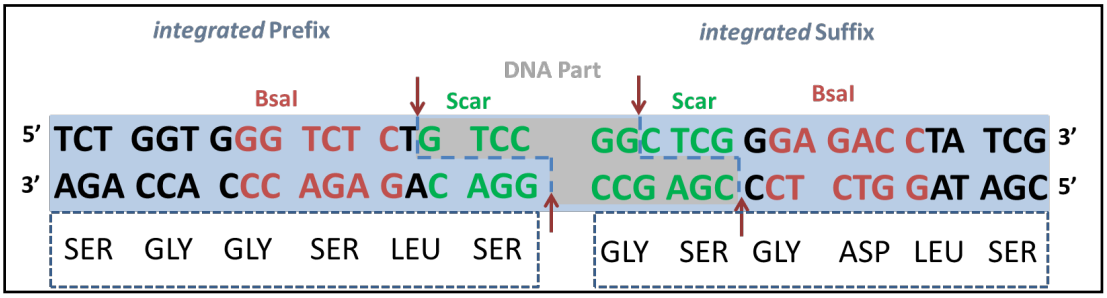


Figure 3.1 *- Prefix and Suffix sequences selected for use within the BASIC assembly format. BsaI recognition sites are positioned at flanking regions but cut outside and internal from this site, thus leaving unique overhangs suitable for prefix and suffix linker addition*

This particular characteristic has been used to carefully design both prefix and suffix sequences, which contain the same BsaI recognition site, but leave unique overhangs different with sufficient sequence independence. The target DNA part is released from a universal storage vector by a single BsaI digestion reaction allowing compatible prefix and suffix linkers to be added. These anneal with high specificity and provide built in directionality in the downstream assembly step (***Figure 3.2***)

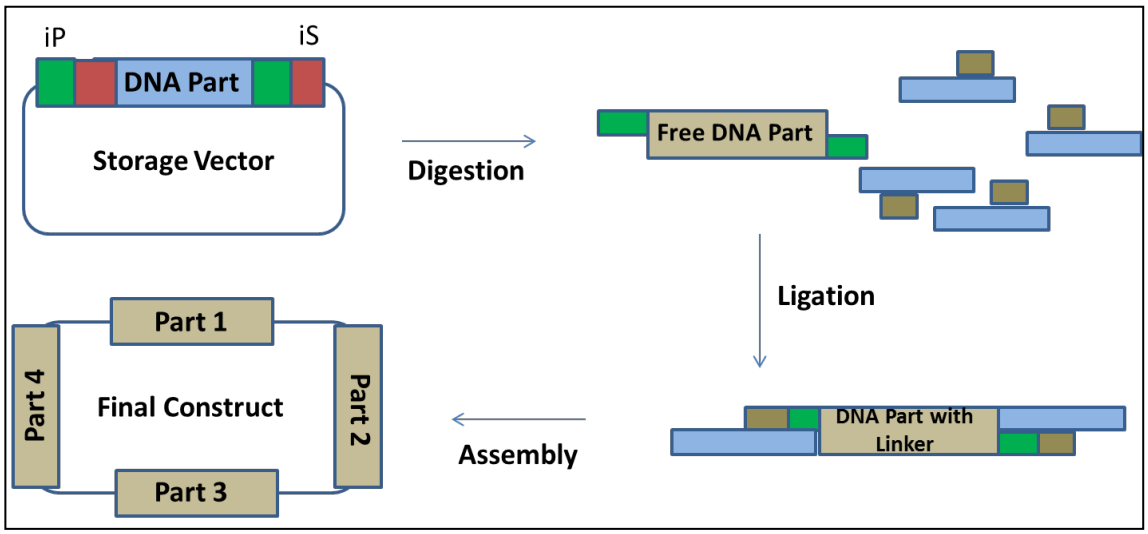


Figure 3.2 *– General steps required for BASIC assembly. Formatted parts are prepared and held in the universal storage vector (pJET 1.2). Digestion by BsaI releases the target part from the vector and creates overhangs required for linker addition. Note that linkers have been pre-annealed to their corresponding adapter oligo to create a partial double stranded linker that readily ligates. Finally, all parts are added together in a final ligation steps that only requires hybridisation between the corresponding long overhangs created by the linkers.*

Considerations were to ensure the short scar sequences which remained contained neutral amino acids residues. This would allow functional DNA parts to be used to create fusion proteins if the downstream application requires. Furthermore, previous work in our group explored the use of specific optimal sequences to direct overlap assembly of multiple parts with high accuracy - MODAL (*Casini et al, 2014*). Therefore, the BASIC standard was a further extension of these principles, utilising biological neutral linker sequences that were favourable as PCR landing pads. We employed a computational design tool, R2oDNA software, which considers GC content, dimerisation, comparison to other linkers within the complete linker set and homology to host organism genome (*Casini et al, 2014*).

Each linker sequence in the set was close to the 40% GC content revealed to be most accurate and efficient within this earlier work (*Casini et al, 2014*). By default, the linker sequences generated are 45bp in length. This includes with a 21bp overlap providing specific homology between adjacent parts, ensuring the accuracy of the overall assembly process. It also includes a 2 x 12bp sequencing flanking the long overlap region on either side. This was required for controlled annealing of adapter oligonucleotides, creating a partial dsDNA sequence which dramatically improves ligation between the linker and DNA part through activity of DNA ligase (***Figure 3.3***). The individual prefix and suffix linkers were generated from these complete linker sequences.

The process continues once individual parts are digested out from the storage vector and include exposed 4bp overhangs. The prepared and partially double stranded oligonucleotides are ligated at both ends, directed by homology to the exposed overhangs. A key advantage is that the digestion and

ligation steps are carried out simultaneously but each individual part and linker is prepared separately. Importantly, the overall final order of assembled parts is governed by the positioning of the compatible linkers sets, driven by high specificity of complementary sequence.

Before the final ligation, purification and removal of excess linker sequence is essential because residual oligonucleotides would be free to bind to complementary linkers, saturating the free overhangs of correctly ligated linkers, and promote non -specific end-ligation if carried through to the final assembly step. This would significantly reduce the chance that adjacent parts could come together and hybridise in the intended fashion. Essentially this would block the final assembly (circularisation) of the complete construct.

The final assembly of all parts requires addition of equimolar concentrations of all parts within an ionic buffer, promoting efficient and accurate hybridisation, guided by long linker homology. A combinatorial strategy can be easily employed at this stage by altering the final mix of composite parts e.g. similar parts of variants of the same part which have been prepared with the same linker set, can be weighted accordingly to build a wide combinatorial library.

Finally, the assembled construct is transformed into a cloning strain, selected on appropriate antibiotic media and individual clones screened using downstream diagnostics namely PCR amplification using the linker sequences as priming sites.

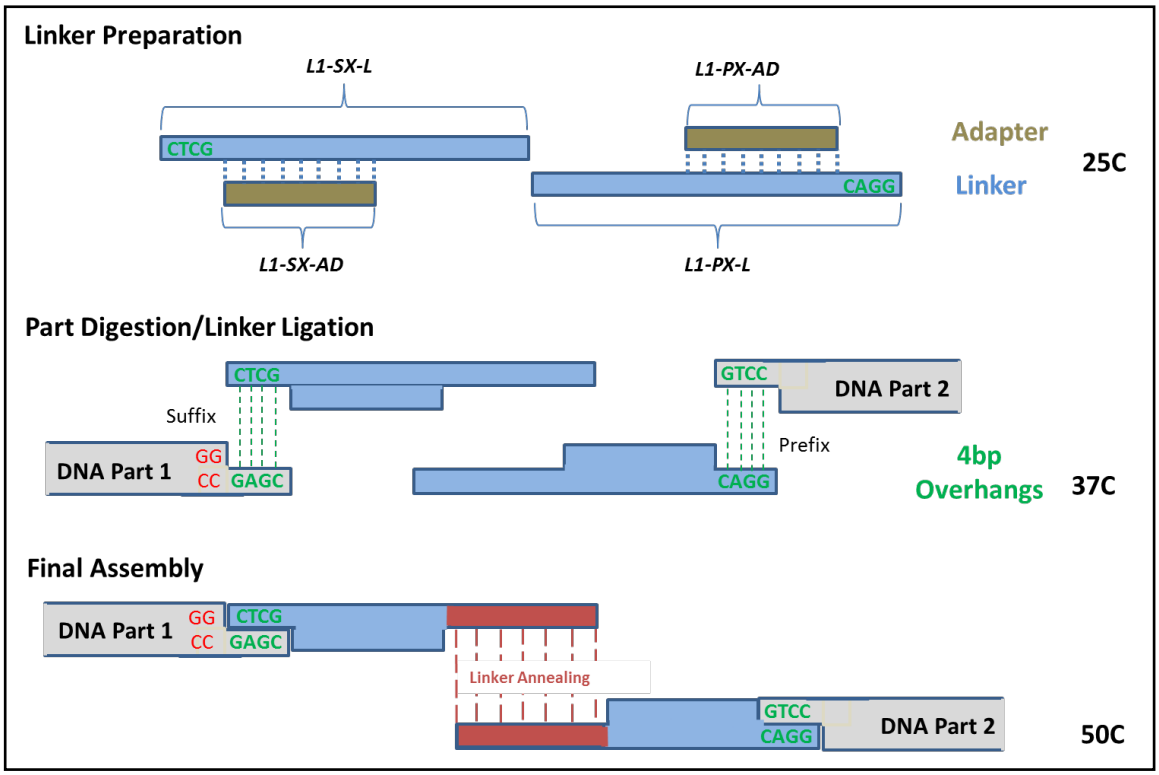


Figure 3.3 *- Three major steps are required for 'linker based' BASIC assembly. 1) Linker preparation requires Linker and Adapter oligos to be annealed, creating partially double stranded DNA. 2) Parts are then digested from the storage vector and the DS linkers are annealed simultaneously. 3) Linker specific homology permits accurate assembly of both prefix and suffix linker halves, completing the join.*

3.2.2 Considerations of BASIC Linker Design

The importance of intelligent linker design could not be underestimated when developing BASIC assembly as the linker sequences themselves ensure highly specific and accurate assembly of parts, and thus a final construct that is extremely likely to be correct. Many of these considerations are listed below:

- Addition of Prefix and Suffix sequences
- Biologically neutral
- Linker sequence specificity
- Internal BsaI forbidden site
- Avoidance of internal false RBS
- GC content and melting temperature
- Orientation of linker and adapter sequences prior to ordering
- Linker sequences are 5'phosphorylated to improve annealing and activity of ligase
- Adapter sequences are not 5'phosohorylated
- HPLC purification to remove errant synthesis products that would otherwise reduce efficiency.

3.2.3 Standard Linkers

A set of 7 standard linkers with sequence that is favourable for general molecular biology or more specifically, their application in cloning methods and *in vivo* protein expression were chosen. For example, the sequence were considered as suitable PCR landing pads, are translated into neutral residues and avoid repeat sequences that might result in secondary structures (*L1-L7*) (***Appendices - Table 11.4***). Each consists of 4 individual oligos that were synthesised from IDT. The nomenclature given to these relates to their positioning within the complete linker sequence. The naming of Prefix and Suffix linkers is in reference to the whether they anneal to the 4bp overhang left by the prefix or suffix sequence after BsaI digestion (e.g. *PX, SX*).

Initially, both the Prefix and Suffix sections of the linker are made up of 2 oligos named the Linker and Adapter (e.g. *L1-PX-L* and *L1-PX-AD*). These are annealed to make a partially double stranded DNA (in order to improve the efficiency of ligation by increasing the activity of the ligase). Once linker and adapter have been annealed, a 4bp overhang remains at the 5'end, which complements the 4bp overhang left after BsaI digestion of the DNA part which in the case of Prefix is 'GTCC' and for the Suffix it is 'CGAG'.

In the final assembly step, the complementary 21bp ssDNA overhangs that remain on both L1-PX and L1-SX are annealed and thus the final connection is referred to as Linker 1 (*L1*) (***Figure 3.4***).

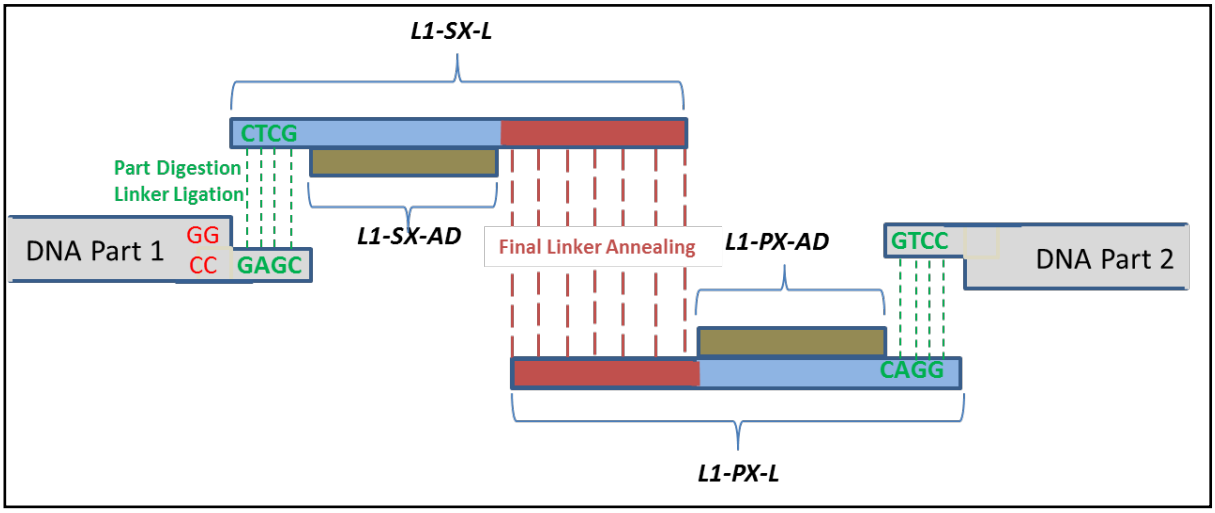


Figure 3.4 *– Standard Linkers during BASIC assembly.*

3.2.4 Hierarchical Assembly – Methylated Linkers

A key characteristic of Type II restriction endonucleases is the ability to cut outside of their recognition site and thus this sequence itself can be lost after successful assembly. This means that if the site can be restored then there is the possibility for hierarchical assembly. If initial assemblies are flaking by the same prefix and suffix sequences, they can become intermediates in larger and more complex construct assemblies.

We explored the use of methylated linkers to restore and protect the BsaI recognition sites (***Figure 3.5***). Essentially, small modules consisting of different DNA parts can be assembled with the methylated linker sequences position at alternative ends. The already assembled 'module of interest' can then be retrieved and released using the same BASIC assembly process.

These linkers are named 'Linker Methylated Prefix and Suffix' (LMP + LMS respectively) contain the standard prefix and suffix BASIC sequences, but have a 5'methylation of a Cytosine (C) residue within the BsaI recognition sites (***Appendices -Table 11.5***). Methylation was carried out during production synthesis. This protects against digestion during the initial assembly and regenerates the prefix and suffix whilst the module of interest is retained as a functional part in its own right. After subsequent transformation into a cloning strain, the module interest can be removed and shuffled into a new double selection vector backbone along with any other parts the user wishes to add. Thus, it becomes particularly easy to construct multi-part pathways or complex circuits in very few steps.

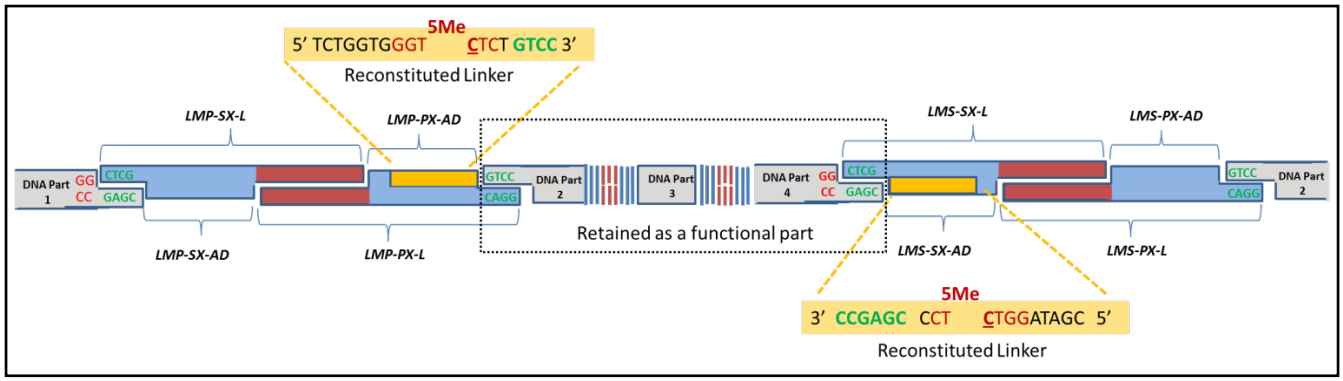


Figure 3.5 *– Methylated linkers used to regenerate Prefix and Suffix BsaI recognition sites ensuring that hierarchical idempotent cloning is possible in BASIC assembly. The internal parts can be retained as a complete functional part in subsequent assemblies.*

3.2.5 Creating Fusion Proteins- Fusion Linkers

During the design of the BASIC standard, considerable care was taken to ensure compatibility with potential downstream applications, broadening its usability. One such application, which was particularly applicable to my own work, was to use BASIC assembly to build libraries of fusion proteins. Therefore, a bespoke linker sequence needed to be designed so that it permits translation of a single multi-domain polypeptide, built from two protein coding sequences (***Figure 3.6***).

When there is limited knowledge of protein domains or indeed separated protein domain structure, it is often favourable to screen variants of the fusion proteins which include different linker sets. For example, these might include flexible, rigid or biologically inspired linker sequences. Therefore, I sought to design three linker sequences which fit into each of these categories, namely an alpha-helical rigid linker (*LFU1*), a flexible Glycine-Serine repeat linker (*LFU2*), and a linker sequence copied from the Taq polymerase / DNA Pol I homolog (*LFU3*). The relevance of LFU3 is detailed later, given the general aim to fuse a 5'-3- Exonuclease domain with a DNA Polymerase domain (***Chapter 7***).

Previous design of the general BASIC linker used the R2oDNA software (*Casini et al, 2014*) but given further linker sequence constraints, I decide upon a manual 'trial and error' design process and utilised the 'reverse mode' function available within the tool, to assess whether the fusion linkers met with the general linker criteria. Importance was placed upon optimising the linker so that the target residues were included. However, the linker sequence needed to retain enough specificity at their overlaps to achieve accurate assembly. GC content, melting temperature, linker length, biological neutrality, secondary structure formation and the possibility of non-specific annealing to alternate linker sets were all considered.

Importantly, the small scar sequences created by the BASIC prefix and suffix could not be changed, however the 6bp suffix scar (GGC TCG) coded for Glycine and Serine residues. The prefix scar (G TCC) was slightly malleable given that the first codon is incomplete, but to ensure consistency, I chose to fix the linker sequence as (GGG TCC) which once again coded for Glycine and Serine residues.

Therefore, the two functional protein domains are 'capped' by two flexible and neutral residues, yet the linker sequence could be rigid or flexible, depending on the added linker sequence. I believe this to be very useful when screening a wide range of fusion proteins variants.

Another minor change was needed to alter the established linker length to ensure that an addition part remains within a single reading frame. Therefore, each fusion linker length was carefully thought-out to ensure the reading frame is upheld (***Appendices - Table 11.6***)

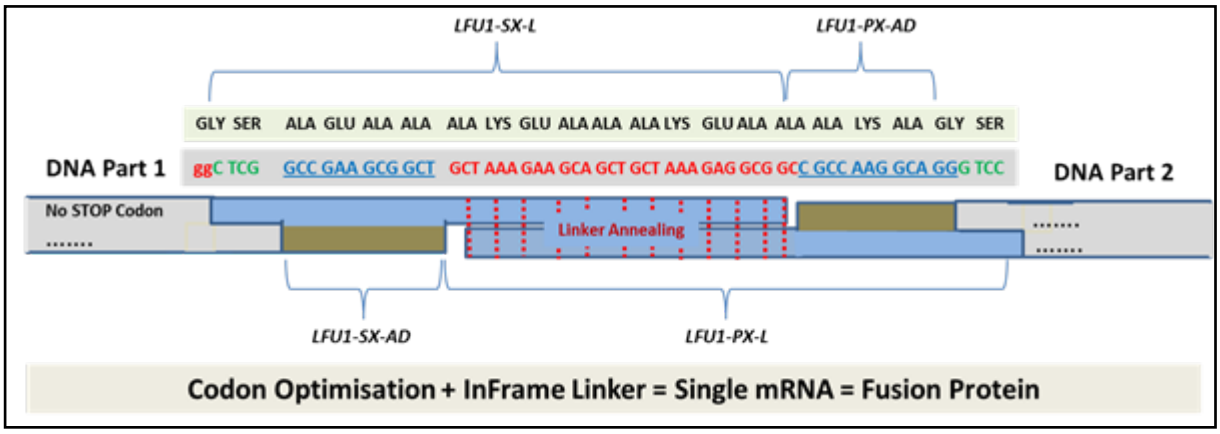


Figure 3.6 *- Fusion linkers used to add functionality to the linker region and ensure that a single mRNA containing both DNA part 1 and DNA part 2 will be read in-frame and translated as a single polypeptide, without compromising the efficiency of BASIC assembly.*

Rigid linker (Alpha-Helical)

L-FU -1 GSAEAAAKEAAAKEAAAKAGS 21aa

ggC TCG GCC GAA GCG GCT **GCT AAA GAA GCA GCT GCT AAA GAG GCG/T GC**C GCC AAG GCA GGG TCC*

Flexible Linker (Glycine-Serine Repeat)

L-FU-2 GSGSGSGSGSGSGSGS 18aa

ggC TCG GGC TCG GGC TCC **GGA TCT GGT TCA GGT TCA GG**A TCG GGC TCC GGG TCC *

Copied Linker (TAQ / DNAPolI homolog)

L-FU-3 GSLLESPKALEEAPWPPPEGS 21aa

ggC TCG CTG CTT GAG AGC **CCT AAA GCA TTA GAA GAA GCA CCT TGG CC**T CCA CCA GA**A** GGG TCC

Figure 3.7 *– Fusion protein linker sequences which required de novo design strategy in order to maintain optimal conditions for adapter annealing and reduce secondary structure formation due to repetitive use of certain codons.*

3.2.6 Combinatorial Pathway Construction using RBS-inclusive linkers

Another extension to the BASIC assembly method involves the introduction of functional sequence within the linker sequences themselves, via the addition of ribosomal binding sequences (RBS). The theory here is to increase linker usability within multi-gene pathway construction, in which complex metabolic pathways can be assembled in a combinatorial fashion from a range of RBS linkers,

promoters and the relevant gene coding sequences. This also increases the underlying modularity of the BASIC standard, because can genes can be retained as individual coding sequences without the need to introduce RBS into the part itself. This strategy should allow for parallel assembly of different versions of the same pathway, and if an accurate and quantitative screening method is in place, a large number of variants can be analysed.

Once more, the R2oDNA software was used to make informed choices for RBS linker design. The RBS sequence itself is placed at the 3'end of the linker sequence, encapsulated by the 14bp adapter oligonucleotide annealing site. This is beneficial because even if RBS sequences are repeated across different linkers, this region is not exposed and thus we can avoid non-specific annealing during the final assembly step.

It is important to note that the original prefix scar of 4bp is extended to 6bp by addition of 'TA', to provide an optimal distance between the RBS position and the ATG start codon and thus the adapter region is also extended from 12bp to 14bp. R2oDNA allows the user to define certain regions of the linker sequence and randomise the rest of the sequence under the same set of stringent criteria mentioned previously.

5 different RBS sequences were chosen and were expected to give a broad range of translational strengths. A precursor sequence was added to the chosen RBS sequences which formed part of the 21 bp overlapping sequence between neighbouring linkers. Low scoring sequences, which were likely to form secondary structures e.g. hairpins and self-dimers, were eliminated. 4 precursor sequences were chosen along with 4 RBS sequences, providing a possible set of 20 different RBS linkers (***Appendices - Table 11.7***).

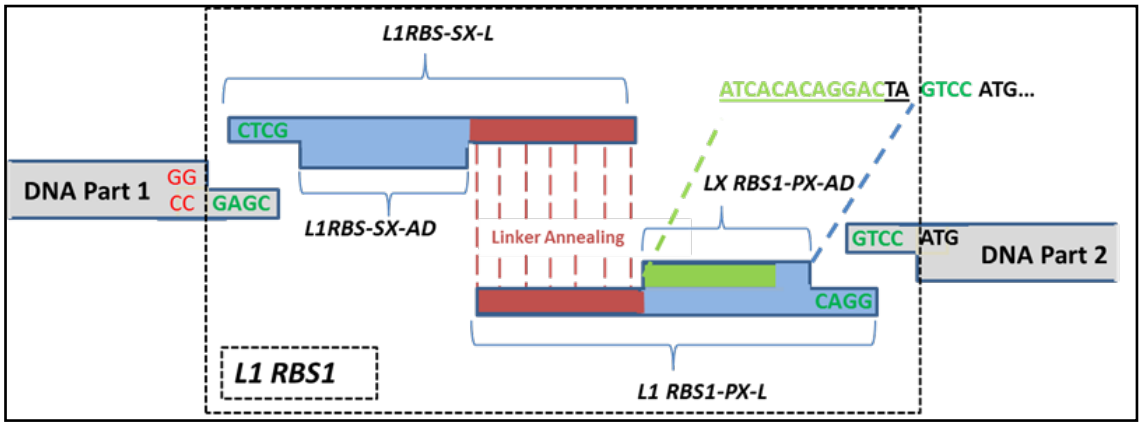


Figure 3.8 *– Functionalised RBS linkers which can be used for the combinatorial assembly of pathways, providing a varied level of expression with multi-gene operons*

3.2.7 BASIC assembly Protocol

I include here a detailed breakdown of the different steps in BASIC assembly. For a simplified working protocol please refer to ***Chapter 9.5.2***.

Construct Design

Firstly, it is extremely important to have a clear design of the final construct and its composite parts. Ultimately the position of the various parts within the final construct is governed by which linkers are joined during the initial ligation step. Furthermore, the choice of fusion, RBS or methylated linkers should be clearly noted so that they are placed in the correct position to provide the desired added function. Therefore, it is of paramount importance that the design is accurate and annotated in a useful way. Given that the ambition is to run multiple assemblies in parallel within a 96 well plate format, with a single plate containing parts and a single plate with linkers, a simple script was built to ensure that each build was correct.

Part Preparation

Initially all functional parts were amplified using standard PCR conditions with primer extension to add the BASIC prefix and suffix. The PCR product was then cloned into a commercial high copy storage vector, pJET1.2 (*Thermo Scientific*) and sequence verified. This enables the preparation of a high yield of plasmid DNA for later use in the assembly reaction whilst ensuring that a re-useable part can be kept in a high copy backbone for later use.

Linker Preparation

Single strand DNA oligonucleotides were ordered with 5' phosphorylated ends to ensure they were favourable for ligation and then HPLC purified to limit the chance of incorrect oligo variants being carried through into the reaction, something which would likely reduce efficiency of assembly. DNA pellets of both linker and adapter oligos required re-suspension in 10 mM TRIS-HCL buffer pH 7.9 to a final concentration of 100 µM. A high ionic strength annealing buffer consisting of 10mM TRIS-HCl, 100 mM NaCl and 10 mM $MgCl_2$ was used to promote annealing of the adapter and linker oligonucleotides during a 30-minute incubation at room temperature. 0.5 µl of adapter and linker oligonucleotides were added to 49 µl of annealing buffer in a final reaction volume of 50 µl. The partially double stranded oligonucleotides that are formed were stored at -20°C and used in subsequent assembly reactions for up to 1 month afterwards.

Digestion/Ligation

Simultaneous digestion of the functional part from the storage vector and ligation of compatible linkers is advantageous in reducing the overall time frame of the complete protocol. A great deal of optimisation was required to ensure that complete digestion of plasmid and release of the functional part was possible. Thus, the concentration of both the DNA and ratio of BsaI restriction enzyme and T4 DNA ligase was evidently essential to promote maximal binding of linker oligos to the exposed overhangs. This was also of vital importance as we needed to reduce the possibility of vector background. This was possible because BsaI cuts outside of its recognition site, meaning the site is then lost and once the linker has been ligated onto the part, there is no possibility of re-digestion. Given that the restriction enzyme and the ligase are both contained within the same reaction mix, this should push the reaction in a forward direction, increasing the success of part/linker ligation. Each part requires the addition of both a prefix and suffix linker and a 30 µl reaction was set up consisting of 3 µl 10mM ATP, 3 µl 10x NEB Buffer 4, 3 µl 10x BSA, 5 µl of prefix linker, 5 µl of suffix linker, 1 µl of DNA part normalised to 50ng/µl per Kb of the total storage plasmid, 1 µl BsaI-High Fidelity, 0.5 µl T4 ligase, 8.5 µl dH20. The incubation cycle consisted of holds at 37°C for 2-4 hours, 20°C for 20 minutes, 65°C for 20 minutes and was carried out in a thermocycler.

Purification of Ligated Products

Removal of excess linker oligonucleotides prior to the final assembly is essential because otherwise there is the potential for 'free' linkers to anneal with their homologous part/linker and in affect 'block' the complete assembly of the intended circular construct. The decision to use the AMPure XP magnetic beads (Agencourt) for DNA purification was based on the long-term goal to automate the assembly reaction on a robotic liquid handling platform. This method was seen as quick, simple and cost-effective solution whilst the 96-plate format ensured that future transfer onto the robot should be possible. Theoretically the small linker oligonucleotides below 50bp should be lost during this purification step. As stated in the manufacturer's guidelines, 54 µl of magnetic bead solution was added to the total 30µlof the previous reaction. The final elution of purified DNA was carried out in 40 µl of dH_20. DNA was stored at 4ºC until required for use (should the part and linker combination allow). The prepared DNA parts were used for up to 2 weeks

Final Assembly

1 µl of each purified DNA part was combined during the final assembly within an ionic buffer consisting of 1 µl 10x BSA, 1 µl 10x NEB4 and dH_20 up to a final volume of 10 µl. Accurate and specific

annealing of homologous linker overlaps was encouraged by incubating the reaction mix at 50°C for 1 hour.

Transformation and Selection

5 µl of the final assembly mix was transformed into 40 µl of competent *E.coli* DH5a cells (high competency >10^9 CFU/ µg pUC19) using the standard heat shock method (***Chapter 9.10.2***). The mix was held on ice for 20 min and then placed into a water bath set at 42°C for 1 min. 950 µl of rich SOC media was added to each sample and placed at 37C for 1hour to allow for recovery. Finally, the samples were spun down to pellet cells, 900 µl of media was discarded and cells were resuspended in 100 µl and plated onto SOC agar plates with the addition of appropriate antibiotics for selection. As standard, Kanamycin (50 µg/ml) and Chloramphenicol (25 µg/ml) were used for double selection plates.

Checking constructs

As with any cloning technique, it was good practice to check that the reaction had worked correctly. Successful colonies were selected, and preparation of DNA was required for downstream diagnostic tests, namely sequencing across the various linker joins, using restriction digests to reveal the complete construct size or colony PCR to provide a banding pattern which can be compared to the expected (***Appendices – Table 11.8***)

3.2.8 Automation on Liquid Handling Platform

A major focus during the development of the BASIC assembly standard was to ensure that preliminary data could be collected using a bench-top protocol, but more importantly, make allowances in design so that each step of the protocol was easily transferable to an automated liquid handling robot. This was underlying reason to move away from PCR based assembly, which is technically very difficult to automate. However, the use of a standard set of linker sequences with predefined characteristics could be completely adaptable to high-throughput combinatorial assembly approach given a technically capable robotic platform. Thus, we chose to adapt our protocol in terms of reagents, volumes and conditions (e.g. the use of a mag bead purification system) so that the entire protocol could be transferred once manual optimisation and the laborious test phase had been completed.

3.3 Results and Discussion

3.3.1 General Assembly

Initial proof of principle experiments consisted of 4-part assembly which included two genes coding for a red and green fluorescent proteins, along with an antibody resistance marker and an origin of replication. Therefore, if correct assembly of all parts was achieved, the target construct could be

selected and visualised (***Figure 3.9***). A total colony count of the transformant plate could be compared to the number of colonies which showed excitation of both fluorophores (i.e. comparing incorrect background and correct assemblies). This allowed for a simple estimation of efficiency and accuracy. Subsequent preparation of 'correct' colonies was used to sequence across each linker join, (although this was not necessarily used in later diagnosis of future constructs due to significant cost). Preliminary experiments showed very high efficiency and accuracy for 4-part assemblies and thus extensive sequencing of multiple clones was not needed. All combinations of standard linkers were used with minimal change in the number of colonies on each transformant plate, and negligible differences in accuracy. We employed a simple colony PCRs for diagnostic purposes, which were fine-tuned to allow the functional part to be amplified out from the final construct each time, with primer sites embedded with the linker sequences.

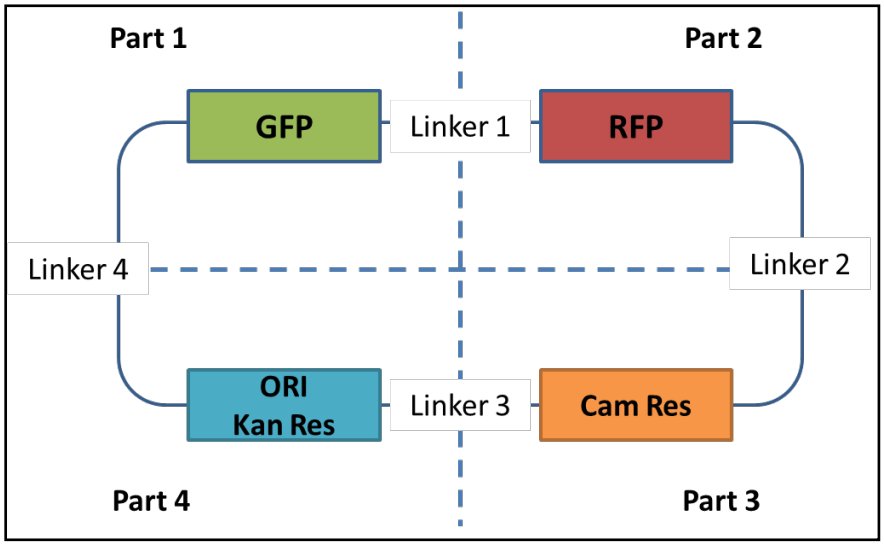


Figure 3.9 *– A simple test construct assembled using the BASIC protocol which provided an immediate assessment of accuracy and efficiency*

3.3.2 Optimising the Protocol

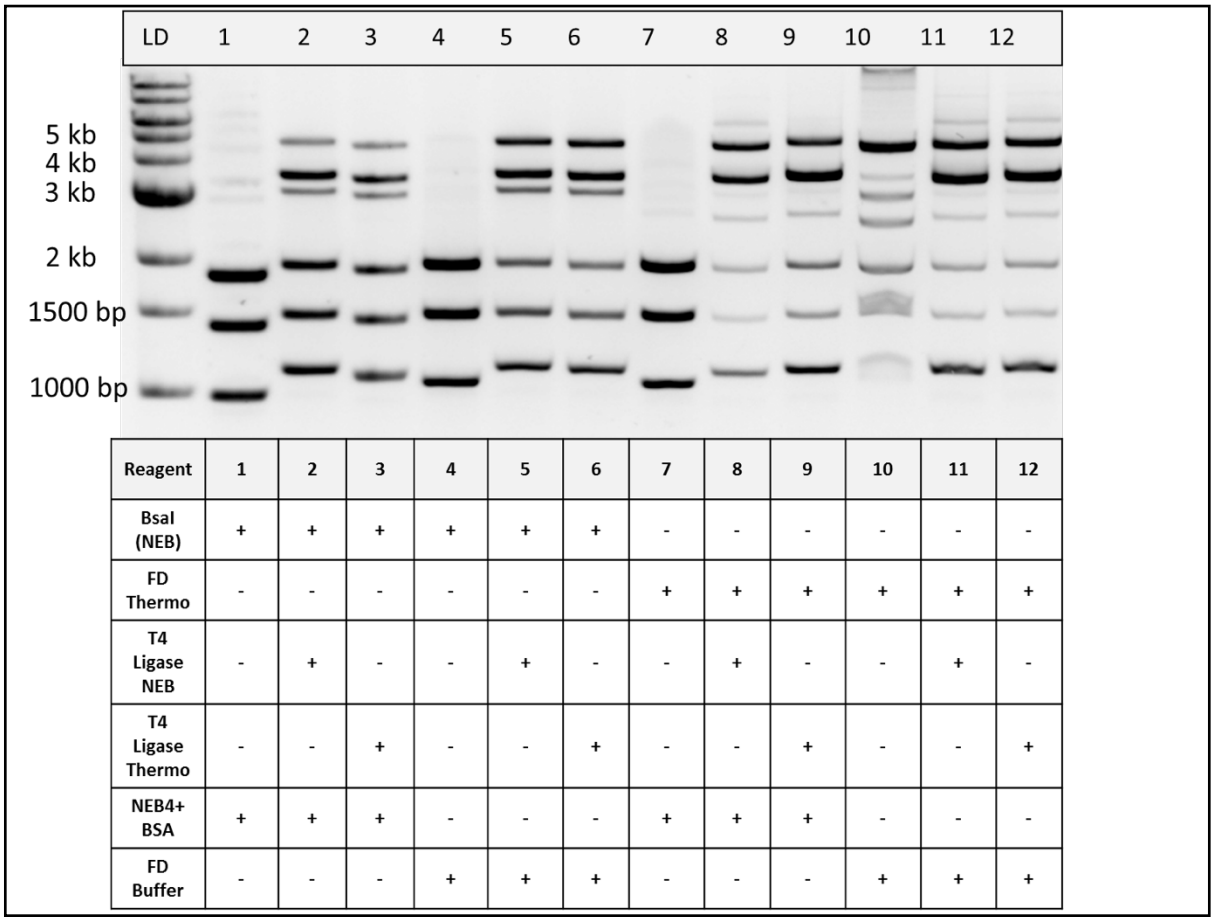
Whilst the protocol above represents the current published BASIC standard, the process of creating a working and reliable protocol took a total of 18 months, given that there were numerous variables to test. We found that certain conditions had a significant impact on the success of the assembly and often a trial and error process was required to identify the more elusive, yet vitally important factors. Major variables and conditions that were changed in order to assess which version of the BASIC standard was most accurate and efficient which are described below (***Table 3.1***).

Primarily this optimisation required improvements to digestion and ligation steps. A common DNA part containing the RFP gene was digested from the storage vector and ligated to L1PX and L2SX oligos with variations of the BsaI restriction enzyme, commercial ligases and differing buffer conditions

through a 1-hour reaction. Varying patterns and intensity of product bands were used to determine optimal conditions (***Figure 3.10***). Namely the smallest band (~1 kb) represents the RFP part which has been completely released from the vector. BsaI (NEB) lanes 1-6 showed improved activity over Fast Digest (Thermo) in lanes 7-12, as demonstrated by stronger bands for all intermediate fragments (below 4.5kb). The comparisons between buffers did not reveal any significant differences as both contained the necessary cofactors required. The second step in optimisation considers the duration of the digestion and ligation step (***Figure 3.11***). Lane 1 included a control reaction in which no ligase was added and therefore digestion was much more efficient. Lanes 2 and 3 represent the same digestion reaction but with the addition of different linker set added to the ligation mix. Incomplete digestion was seen at 1 hour, in which the linearised vector remains (red arrows). Complete digestion could be achieved by 4 hours.

***Table 3.1**- Variables tested during the development of the BASIC assembly standard as exists today. Each independent variable was changed in isolation and efficiency and accuracy were assessed using the standard 4-part assembly mentioned previously (RFP-GFP-Cam-pMB1KAN).*

Stage of Protocol	Consideration	Variables	Consequences
Linker Preparation	Oligo Synthesis	5' Phosphorylated	Phosphorylation improves activity of ligation in subsequent steps
		Non - phosphorylated	
		HPLC Purified	Purification was essential to prevent non-specific oligo sequences that would ultimately reduce assembly efficiency by sequestering correct linker annealing
		Unpurified	
	Annealing Linker + Adapter Buffer	Different Ionic Buffers	A number of ionic buffers were considered with varying salt concentrations
	Annealing of Linker and Adapter – Temperature	Hold at RT	A high ionic buffer at RT was sufficient to anneal linker and adapter but improved overall efficiency was seen when linkers were heated to 90C and allowed to cool during annealing
		Cool from 90°C	
Digestion/Ligation	Restriction Enzyme - Type	BsaI-HF (NEB)	BsaI-HF was the preferred choice given the design of the buffer for the simultaneous digestion and ligation reaction. Fast Digest actually showed reduced activity (***Figure 3.10***)
		Fast Digest (Thermo Scientific)	
	Restriction Enzyme - Buffer	NEB4	Neither buffer had a significant effect on digestion activity (***Figure 3.10***)
		Fast Digest	
	Ratio of BsaI and T4 Ligase		The digestion/ligation reaction moves in a forward direction due to the loss of BsaI in the linker ligated part. However, excess or overly active ligase could allow the part to re-join with the vector (***Figure 3.10***).
	Reaction time	1-4 hours	Complete digestion of the part from the storage plasmid was essential, and thus a longer reaction time showed improved overall efficiency (***Figure 3.11***)
Purification of Ligated Products	Purification Strategy	Spin Column	Initial use of spin column DNA purification was tried. Although smaller linker oligos were expected to be lost through the column, the lower efficiency would suggest excess linkers were not completely removed. The use of mag beads separated parts (>100bp) with excess linkers and had the added of advantage of automation.
		Mag Beads	
Final Assembly	Ligase	T4 Ligase	The use of ligase to aid assembly in the final step was considered, particularly thermostable Taq ligase at 50°C. However, adding ligase led to a large drop in accuracy so was excluded.
		Taq Ligase	
		No Ligase	
	Temperature	50-60°C	It was suggested that a higher temperature would improve the accuracy of complementary binding of linker overlaps, with 50-60°C considered best.
	Reaction Time	1 hour	An extended hold at 50°C led to increased accuracy and efficiency
Transformation	Competent Cells	Transformation Efficiency	The number of colonies on selection plates drops an order of magnitude with each extra part as transformational efficiency becomes limiting factor



Reagent	1	2	3	4	5	6	7	8	9	10	11	12
BsaI (NEB)	+	+	+	+	+	+	-	-	-	-	-	-
FD Thermo	-	-	-	-	-	-	+	+	+	+	+	+
T4 Ligase NEB	-	+	-	-	+	-	-	+	-	-	+	-
T4 Ligase Thermo	-	-	+	-	-	+	-	-	+	-	-	+
NEB4+ BSA	+	+	+	-	-	-	+	+	+	-	-	-
FD Buffer	-	-	-	+	+	+	-	-	-	+	+	+

Figure 3.10 *– Optimisation of the BASIC assembly protocol. Shown here are products from the Digestion/Ligation step with varying concentrations and combinations of reagents. Products were separated on a 1% agarose gel. linearised template DNA contains a total of 3 x BsaI recognition sites (1 in the vector backbone and 2 flanking the RFP part) resulting in a possible 6 different linear products*

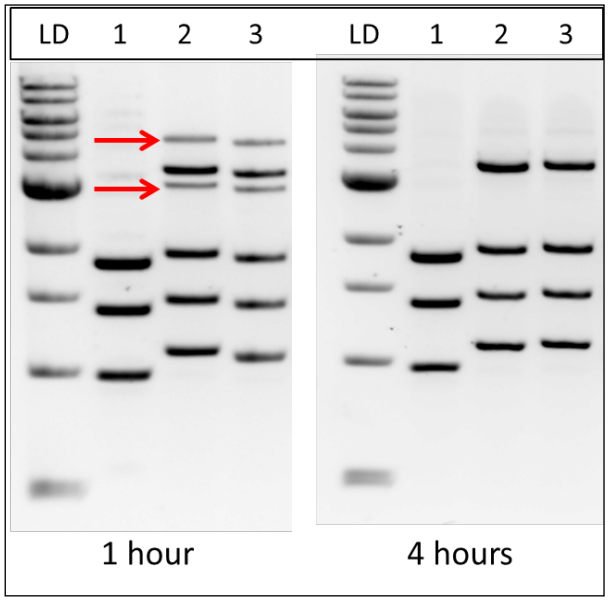


Figure 3.11 *- DNA part digested used to assess BASIC assembly efficiency and accuracy. This digestion was carried out under standard reaction conditions (3 µl 10mM ATP, 3 µl 10x NEB buffer 4, 3 µl 10x BsaI, 5 µl of prefix linker, 5 µl of suffix linker, 1 µl of DNA) for 1 hour and 4 hours. Products were separated by gel electrophoresis. Red arrows indicate incomplete digestion products corresponding to linearised vector*

3.3.3 Efficiency and Accuracy

In order to determine the efficiency and accuracy of the BASIC assembly method, a set of constructs were built using the standard set of linkers. In its simplest form this consisted of a 2-part assembly with a resistance cassette (Kanamycin resistance required for selection on antibiotic plates), and a fluorescent gene cassette (used as a visual output). Constructs were also built with the addition of a second resistance cassette (Chloramphenicol). With the addition of a second fluorescent cassette and building promoters as individual parts a 6 or 7-part assembly could be attempted.

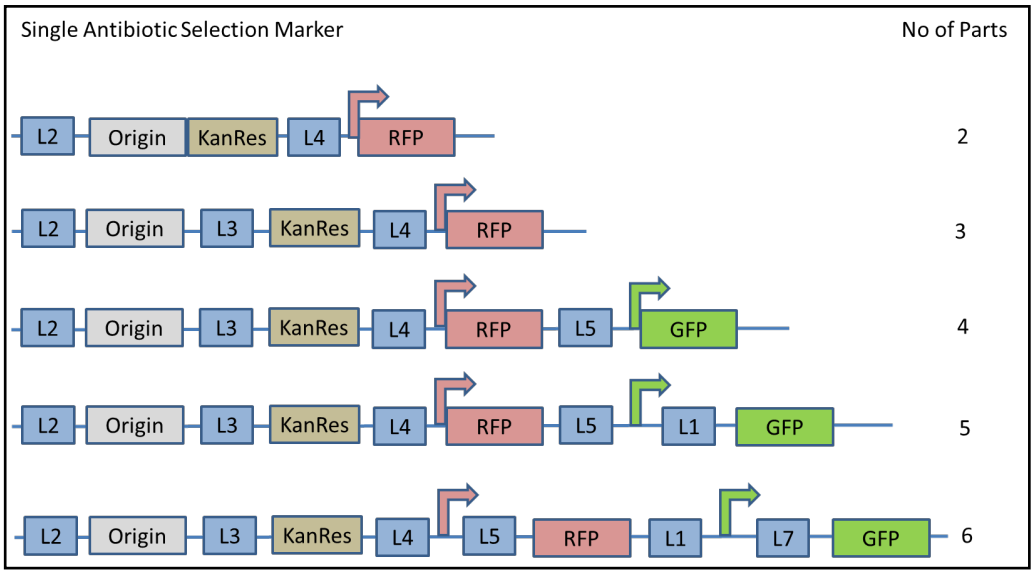


Figure 3.12 *– Test constructs used to assess BASIC assembly efficiency and accuracy. Initially, constructs included a single antibiotic resistance cassette. A visual read out for accuracy could be determined by record fluorescent output containing a single*

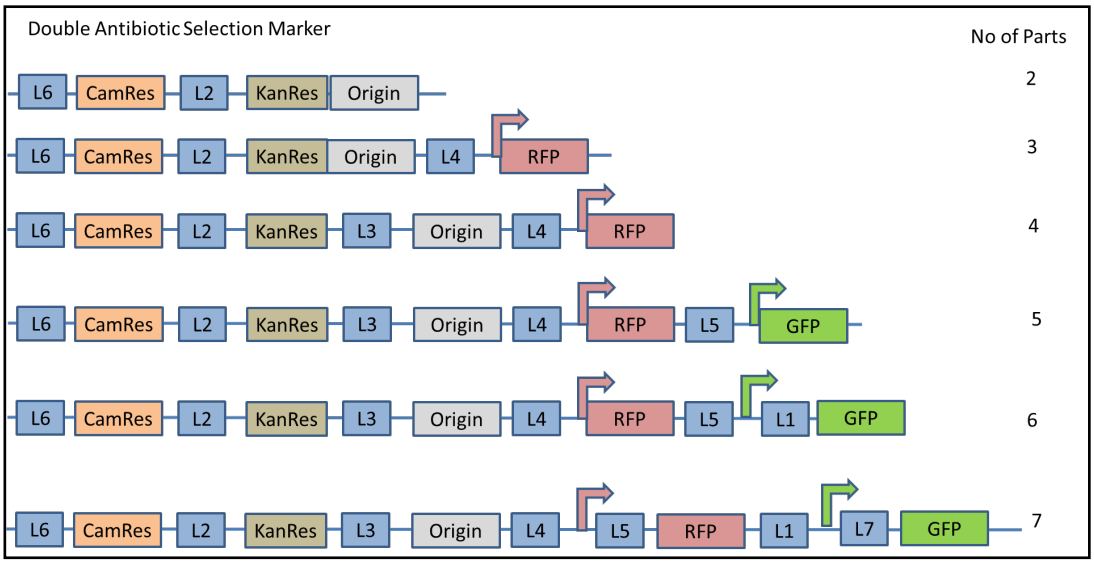


Figure 3.13 *– Test constructs used to assess BASIC assembly efficiency and accuracy which include a second resistance marker*

Assemblies were carried out using the BASIC assembly method previously mentioned and transformed into DH5a *E.coli* cells, plated on LB and relevant antibiotic. Efficiency was determined by

total transformant colony counts (***Appendices - Table 11.9***). By virtue of the fact that the replicative construct was able to survive under selection, this guarantees that both the origin of replication and resistance marker gene must be included in the final assembly. In order to determine accuracy, the transformant plates were visualised on a plate imager, with excitation with blue and green laser for detection by green and red channels, respectively. Where both fluorescent proteins were expressed, images were overlaid, and a yellow filter was applied.

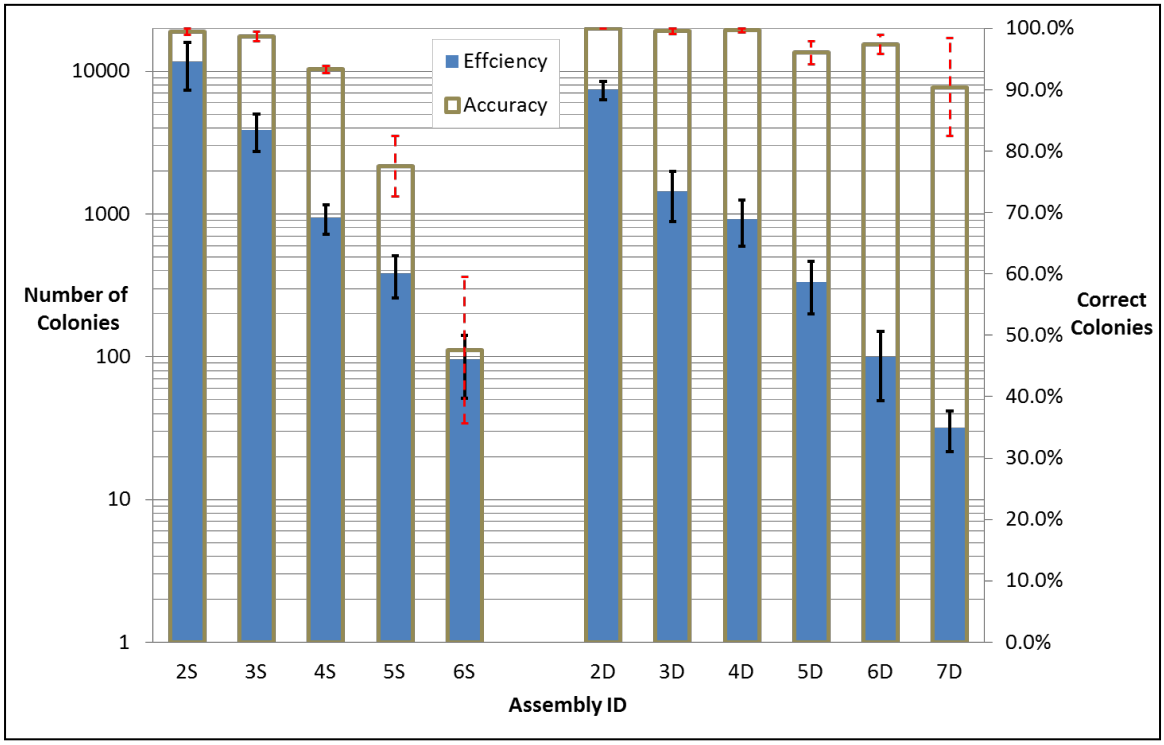


Figure 3.14 *– Multi part assembly using the BASIC assembly standard, comparing one or two resistance markers comparing efficiency and accuracy based on transformant colonies. Averages were taken across 4 repeats of each assembly. Efficiency was determined by the number of colonies on transformation plates and remains similar regardless of the single (S) or double (D) antibiotic selection, falling by half an order of magnitude per additional DNA part. The accuracy seen when two resistance markers were used is remarkable with 7 parts assemblies showing ~90% accuracy (2D – 7D). Standard errors were calculated (**Appendices - Table 11.10 and Table 11.11**) and shown here.*

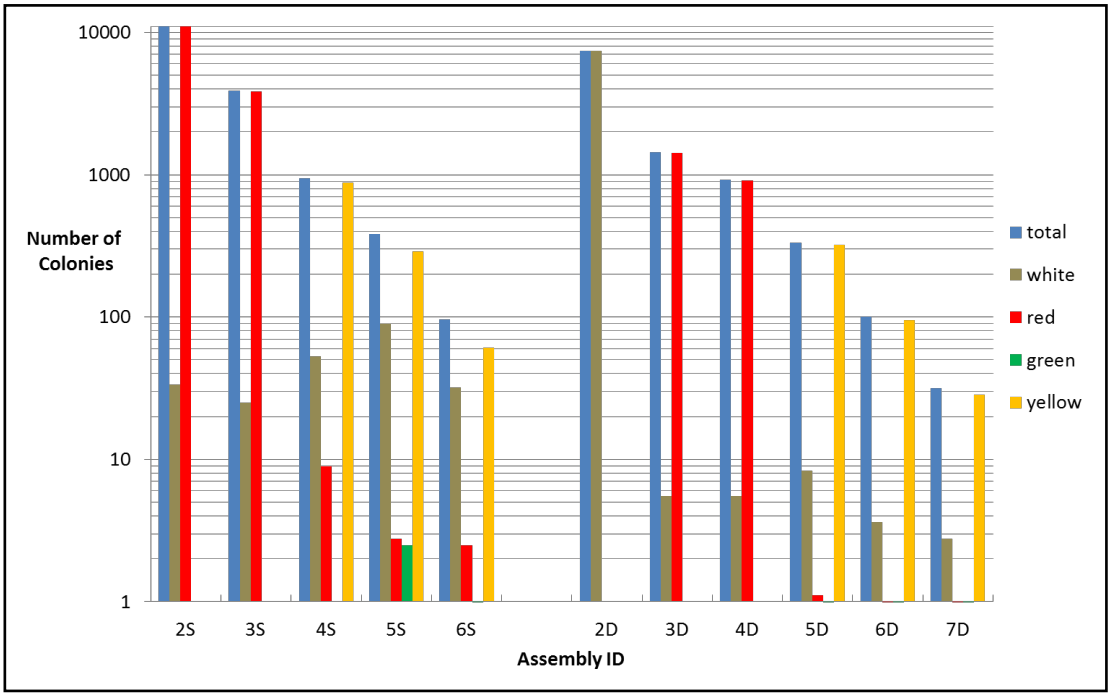


Figure 3.15 *– Various fluorescent profiles seen in the initial test for BASIC assembly accuracy. Blue bars represent the average total number of colonies counted. Where only 1 gene encoding a fluorescent protein is included (2S,3S,2D,3D,4D), red bars represent the correct assembly. Where 2 genes encoding 2 fluorescent proteins are included, yellow bars represent the correct assembly (given that both red and green are overlaid as yellow output). White (no fluorescence) bars represent incorrect assemblies.*

3.3.4 Hierarchical Assembly

In order to demonstrate that the BASIC standard could be used to build constructs in a hierarchical fashion, a test assembly which integrated the pre-designed methylated linkers (described above) was broken into two distinct stages. Efficiency and accuracy could be determined in much the same way as in the initial validation tests already described, given that two fluorescence genes were again used. The key difference being that during Stage 1, both GFP and RFP were separated out into two different circularised constructs.

Stage 1 constructs were simple 4-part assemblies, containing a double antibiotic resistance marker, high copy origin of replication, a constitutive promoter and the gene encoding either GFP or RFP, ensuring easy selection and screening (***Figure 3.15***). The methylated linkers ensure that the BASIC prefix and suffix sequences are retained and mean that the promoter and fluorescent gene can be used a single functional part in any subsequent assembly.

The final circularised construct built in stage 2 required the release of the functional part, and the addition of a second resistance antibiotic marker to avoid carry through of the intermediate vectors from Stage 1. Although the final assembly was in fact a 4-part assembly, it is made up from 6 composite parts, which given that the general efficiency of BASIC tends to drop by and order of magnitude with each additional part, this underlines the benefit of using a hierarchical approach.

Control constructs did not contain the methylated linkers. Only correct Stage 1 assemblies were used in the Stage 2 assembly. Three repeats of the experiment were carried out and total colony counts are provided (***Appendices - Table 11.30***).

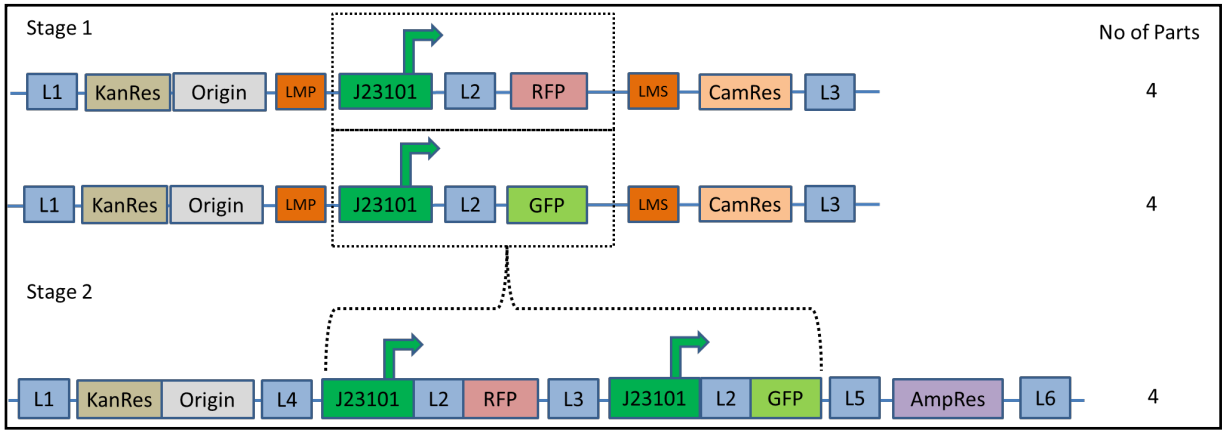


Figure 3.16 *– Two stage hierarchical assembly was used to demonstrate both efficiency and accuracy of using BASIC methylated linkers (LMP and LMS). In this assembly, both RFP and GFP genes were again used to give visual outputs whilst being selected on antibiotic media. Stage 1 involved the construction of a functional output module, combining a promoter and a fluorescent gene, yet these were flanked by methylated linkers. Stage 2 required both output modules to be retained as individual parts, which could be released due to the standard BASIC prefix and suffix within methylated linker sequences. The final assembled construct required a difference secondary resistance marker (AmpRes) in order to avoid any presence of background Stage 1 constructs.*

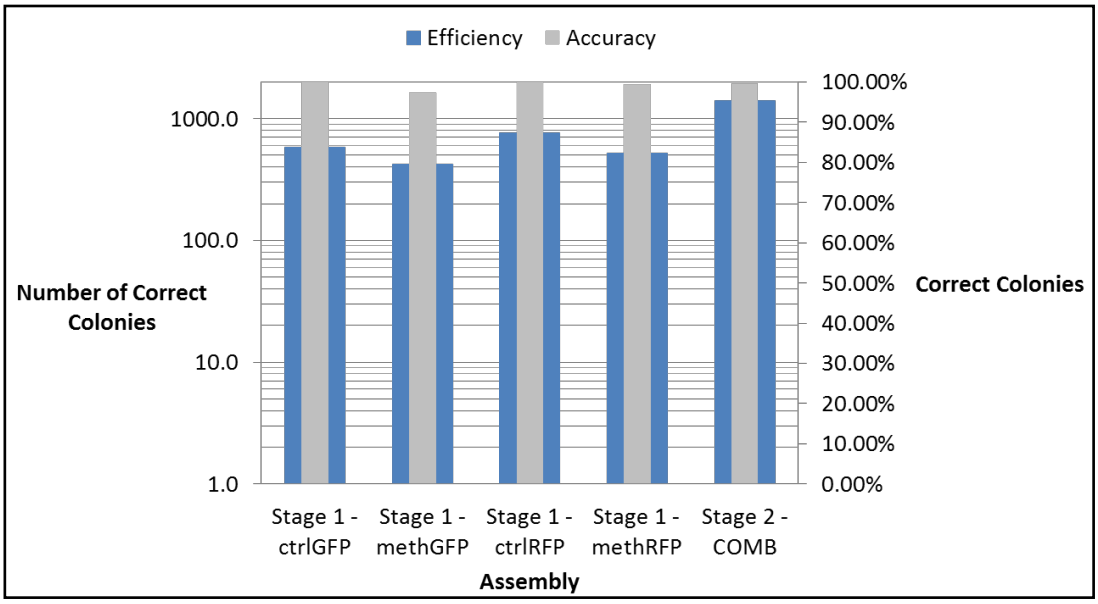


Figure 3.17 *– Hierarchical assembly using the BASIC standard requires the application of methylated linkers. Data presented on the left y axis represents total colony counts on double antibiotic selection plates as an estimate of efficiency. The right Y axis is a measure of accuracy based on the number of correct colonies at both stage 1 and stage 2 of the hierarchical assembly i.e. those which demonstrate the correct fluorescence profile.*

Results from the 2-stage assembly process are very favourable and showed exceptional accuracy of the final construct. A small number of Stage 1 assemblies showed the incorrect fluorescence profile, but only correct constructs were carried through to Stage 2. Interestingly, the efficiency of Stage 2 does

seem to be greater, as represented by a higher number total colony count which would suggest that the methylated linkers used in Stage 1 are potentially limiting in terms of the efficiency of annealing, but when using a lower number of parts I do not anticipate that this will be an issue when scaling more complex assemblies.

3.3.5 Proof-of-Principle

As a demonstration of the usability and versatility of the BASIC assembly method, I chose to assemble a similar 4-part assembly consisting of the same standard parts Ori-Kan, Cam, RFP and sfGFP. However, in an effort to provide proof that standard linkers, RBS linkers and fusion linkers could all be used simultaneously, I made a number of construct variants in parallel. The composite DNA parts required were prepared in accordance to the BASIC standard. It is often beneficial to plan assemblies in tabulated format to ensure that PX and SX linkers are added in the correct positions (***Table 3.2***).

Firstly, the 4-part construct with standard linkers (as shown previously) was prepared and acted as a control for positive expression. To evaluate the use of RBS linker and fusion linkers together, the RBS was removed from the functional GFP part with the RFP part placed downstream, thus creating a simple 2 gene operon under control of a single promoter. It is important to note that the intermediate sequence was evaluated for potential terminator sequences (i.e. AT loops).

Next, the two fluorescent genes were expressed as a complete fusion protein by removing the stop codon from the upstream GFP gene. The used of the aforementioned fusion linkers (LFU1-3) to join both GFP and RFP coding sequences enable translation of the complete fusion protein.

I also prepared a positive control for both GFP and RFP expression under a constitutive promoter. The negative control consisted of a blank vector built in a simplified 2-part BASIC assembly. Efficiency of the assembly was evaluated in much the same way as before, given that double selection plates ensure a reduced level of background and the two distinct fluorophores could be observed if excited at a suitable wavelength.

Table 3.2 *- Composite parts required for the parallel assembly of the 6 constructs variants with the additional linkers required. Note that the PX and SX linkers change depending on the final intended construct and thus the number of total parts required for this proof of principle assembly quickly increases.*

Part Number	Part DNA	Prefix Linker	Suffix Linker	Part Size (cut from pJet storage vector)
1	pMB1.KAN	L1-PX	L2-SX	1665 bp
2	CAM	L2-PX	L3-SX	977 bp
3	J23101.RBS34.GFP	L3-PX	L5-SX	943 bp
4	J23101.RBS34.RFP	L5-PX	L1-SX	930 bp
5	J23102	L3-PX	L1RBSX-SX	171 bp
6	ATG.GFP.NoTerm	L1RBS3-PX	L5-SX	749 bp
7	ATG.GFP.NoTerm	L1RBS3-PX	L2RBSX-SX	749 bp
8	RBS34.ATG.RFP.Term	L5-PX	L1-SX	887 bp
9	ATG.GFP.Term	L2RBS3-PX	L1-SX	899 bp
10	ATG.GFP.NoStop	L1RBS3-PX	LFU1-SX	730 bp
11	ATG.GFP.NoStop	L1RBS3-PX	LFU2-SX	730 bp
12	ATG.GFP.NoStop	L1RBS3-PX	LFU3-SX	730 bp
13	ATG-RFP.Term	LFU1-PX	L1-SX	861 bp
14	ATG-RFP.Term	LFU2-PX	L1-SX	861 bp
15	ATG-RFP.Term	LFU3-PX	L1-SX	861 bp
16	J23101.RBS34.GFP	L3-PX	L1-SX	943 bp
17	J23101.RBS34.RFP	L3-PX	L1-SX	930 bp

Table 3.3 *- Constructs contain standard, RBS and fusion linkers. Diagnostic PCR was used to reveal whether transformants were the correct construct.*

Assembly ID	Description	No. of Parts	Construct	Parts Required	PCR Diagnostic Product Size (L3 For – L1 Rev)
A	RBS Linker	5	1-pMB1Kan-2-CAM-3-J23102-RBSLinker1.3.-ATG.GFP.NoTerm-5-RBS.ATG.RFP	1,2,5,6,8	1807 bp
B	Positive Control	4	1-pMB1Kan-2-CAM-3-J23101.RBS34.GFP-5-J23101-RBS34-RFP	1,2,3,4	1873 bp
C	Fusion Linker 1 (Rigid)	5	1-pMB1Kan-2-CAM-3-J23102-RBSLInker1.3.-ATG.GFP.NoStop-FusionLinker1-ATG.RFP	1,2,5,10,13	1762 bp
D	Fusion Linker 2 (Flexible)	5	1-pMB1Kan-2-CAM-3-J23102-RBSLinker1.3.-ATG.GFP.NoStop-FusionLinker2-ATG.RFP	1,2,5,11,14	1762 bp
E	Fusion Linker 3 (Taq)	5	1-pMB1Kan-2-CAM-3-J23102-RBSLinker1.3.-ATG.GFP.NoStop-FusionLinker3-ATG.RFP	1,2,5,12,15	1762 bp
F	GFP Positive Control	3	1-pMB1Kan-2-CAM-3-J23101.RBS34.GFP	1,2,16	943 bp
G	RFP Positive Control	3	1-pMB1Kan-2-CAM-3-J23101.RBS34.RFP	1,2,17	930 bp
H	Blank Vector / Strain Control	2	1-pMB1Kan-2-CAM	1,2	N/A

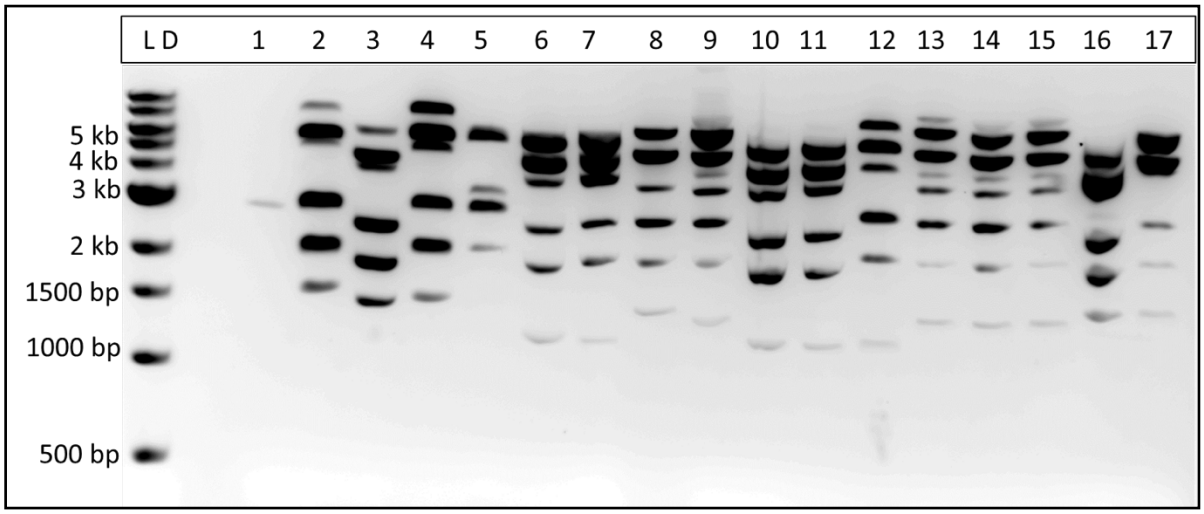


Figure 3.18 *–BASIC Diagnostic Step 1 – 10 µl of each purified DNA part post linker addition and before final assembly were separated by gel electrophoresis run. This was used to confirm that complete digestion of the part from the vector backbone had occurred.*

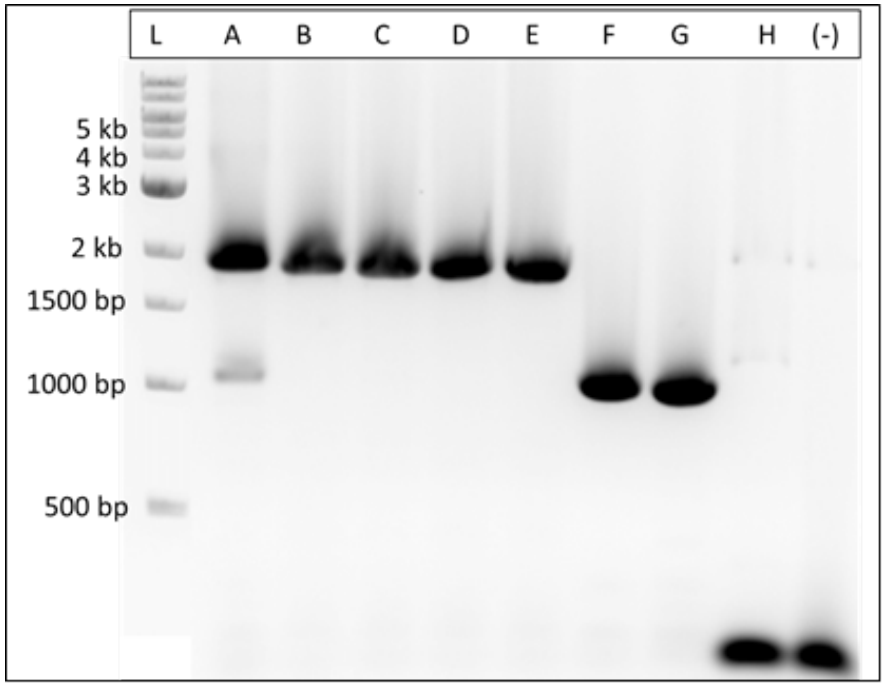


Figure 3.19 *– BASIC Diagnostic Step 2 - Single colonies were picked from transformant selection plates and used directly in colony PCR. Primers used were annealed within Linker 3 and Linker 1, ensuring that the amplified product represented the combined length of functional parts in each construct (i.e. not the pMB1Kan.Cam backbone as antibiotic selection can validate these are both present).*

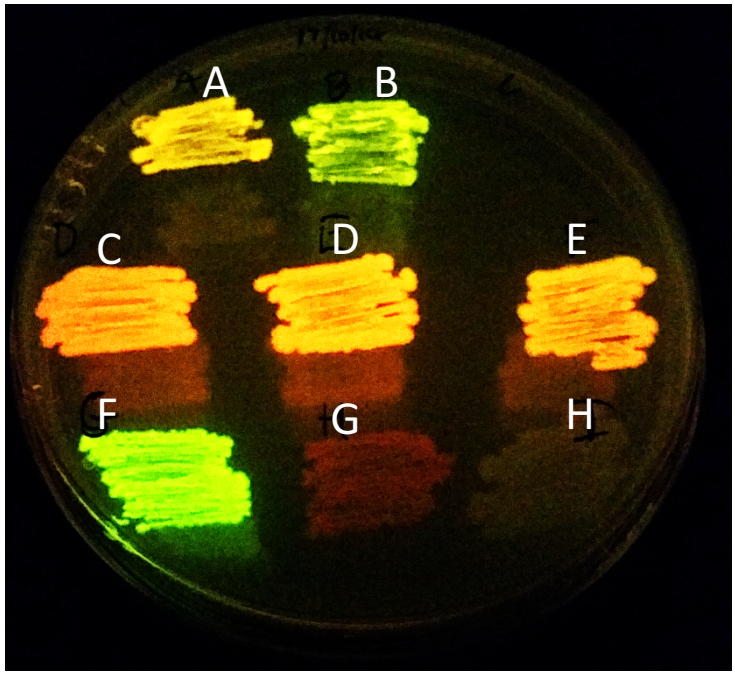


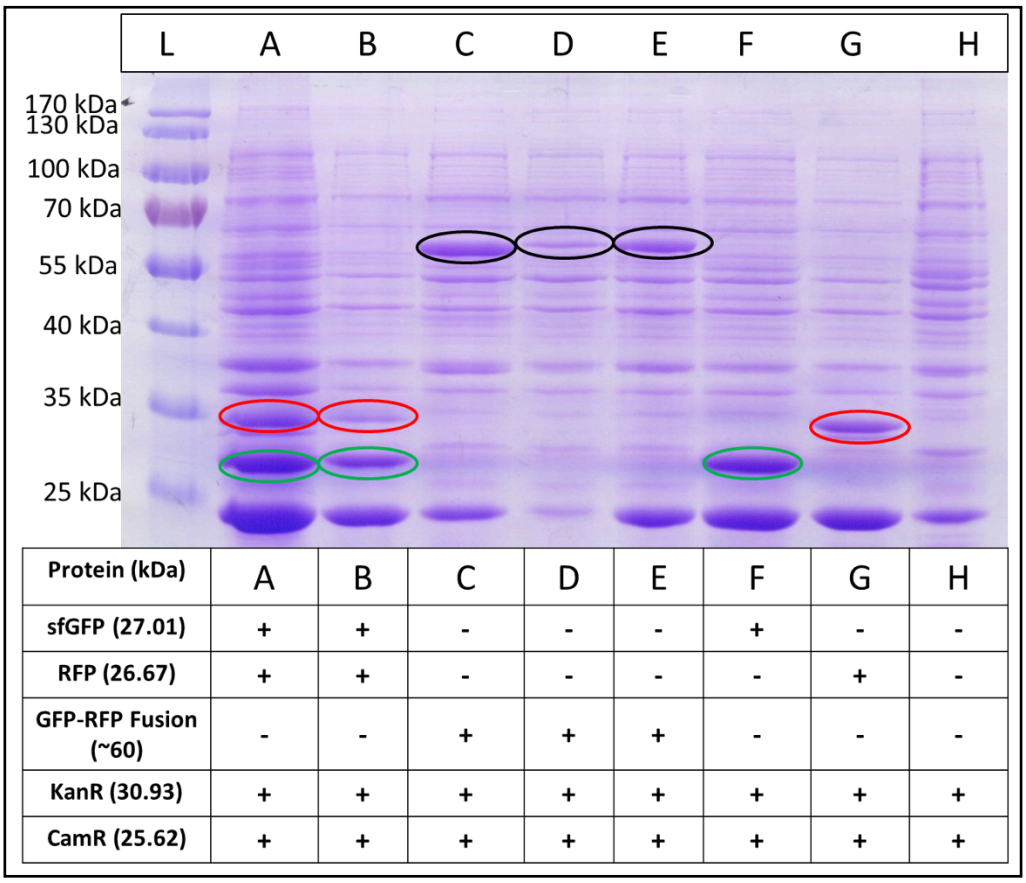
Figure 3.20 *– Diagnostic Step 3 - Expression of two fluorescent proteins as independent single genes and as a combined fusion protein allowed for simple visualisation in white light and the combined fluorescence was also confirmed when streak plates were placed on a blue box.*

Initial results were encouraging given that all sample plates presented with >100 transformant colonies on the double antibiotic selection plates, as this would suggest that the concentration of assembled (circular) plasmid DNA was significant and reflects the efficiency of the reaction. A further troubleshooting step is possible if there are no observable transformants (although it was not necessary here, given that all assemblies appeared to work) in which all purified DNA parts that have undergone digestion/ligation can be visualised on an agarose gel to ensure that correctly prepared DNA and linker were added to the final assembly step (***Figure 3.18***). Individual colonies were picked and used in a colony PCR with diagnostic primers (***Table 11.8***) and shown to give the anticipated PCR product sizes (***Figure 3.19***). Correct colonies were then re-streaked onto a single plate that showed evidence of GFP and RFP protein expression (in relation to construct variant) when observed normally (***Figure 3.20 - A***) and under blue light (***Figure 3.20 - B***).

The chosen colonies were grown and mini prepped in order to sequence across the RBS and fusion linker connected regions to confirm the accuracy of assembly (*data not shown*).

Finally, protein samples were evaluated using denaturing SDS-PAGE (***Chapter 9.9.1***) to reveal distinct protein bands of expected sizes (***Figure 3.21***). In constructs C, D and E that were built using all variations of fusions linkers, the desirable fusion protein (GFP-RFP) could be clearly be observed along with the absence on the individual proteins. A clear comparison can be made between the A-B (single expression of both RFP and GFP), C-E (expression of the fusion protein) and F-G (individual expression of

GFP and RFP respectively. Sample H is a blank control, consisting of the assembled backbone containing both Kan and Cam resistance cassettes.



Protein (kDa)	A	B	C	D	E	F	G	H
sfGFP (27.01)	+	+	-	-	-	+	-	-
RFP (26.67)	+	+	-	-	-	-	+	-
GFP-RFP Fusion (~60)	-	-	+	+	+	-	-	-
KanR (30.93)	+	+	+	+	+	+	+	+
CamR (25.62)	+	+	+	+	+	+	+	+

Figure 3.21 *– SDS PAGE was used to show constitutive protein expression from a range of construct created using BASIC assembly, including a number of fusion proteins. Letters correspond to the same constructs referenced in Table 5.3 and Figures 5.12 – 5.14.*

3.3.6 Automated Pathway Construction

Very recently, the BASIC protocol has become further established by work carried out within the newly opened SynBiCite foundry, which contains a state of the art suite of liquid handling robots. BASIC assembly is perfectly suited to automation and whilst all of the data shown previously has been performed manually at the bench, the transition to a liquid handling platform is enabling the construction of 100's of variants in parallel. This may also lend itself to multi-gene pathway construction using a combinatorial approach. For example, a broad screen could be conducted, in which the expression level of each gene can be controlled.

Given that protein expression is a function of both transcriptional and translational activity, theoretically both promoter and RBS regulatory sequences can be disconnected and retained as individual parts which are then assembled in multiple combinations. The proposed combinatorial strategy is difficult using existing part-based assembly methods, given the fact they are short in length (<30bp). The potential solution is to add the promoter and RBS to the gene CDS prior to assembly, but

this leads to an exponential increase in the number of parts needed whilst being extremely costly in terms of time, money and labour.

Using BASIC, promoters can exist as single parts, and functional RBS linkers can be used to assemble a minimal set of composite parts into maximal combinatorial library of the same pathway. We now have the potential to quickly gain insights about optimal expression levels for each intermediate protein, general metabolic burden and ultimately the ideal combination required to achieve the highest yield of target biomolecules.

The logistics of building such a large library is intimidating given that a 3-gene pathway exploring 3 different promoters and 3 RBS sequences, there are 19,683 combinations. This approach is only possible with a reliable and robust DNA assembly method and a high throughput automated platform.

3.4 BASIC Compatible Genome Integration

3.4.1 Introduction

The underlying focus of my work is to develop a directed evolution system which accurately targets a chosen gene of interest. As detailed in ***Chapter 2,*** a single copy variant of the target gene is embedded within the host *E.coli* genome. This is essential if we are able to select for new and interesting protein function.

It is of critical importance that a reliable and accurate method for genome integration was identified. Unfortunately, and somewhat surprisingly given the previous work to attempt to manipulate *E.coli* strains, a quick and efficient genome integration method remained elusive within our laboratory during the early stages of my research.

Therefore, a portion of time was dedicated to establishing a simple and reproducible protocol that could take advantage of the high efficiency and accuracy of the BASIC assembly standard already described by adapting a recently published integration method.

My own work did not suit a plasmid-based system and there is also a strong argument that the wider synthetic biology community would benefit from characterising individual genetic parts as single copies, in order to eliminate inaccuracies associated with plasmid copy number. Furthermore, novel metabolic pathways are often developed for use in large scale industrial fermentation processes and these would ideally involve an engineered strain rather than use a plasmid based expression system which requires continuous replenishment of antibiotics for selection.

We had previously used the CRIM standard for genome integration that relies of phage attachment sites within the *E.coli* genome (*Haldimann and Wanner, 2001*). However, this approach proved to laborious and unreliable and further attempts to integrate target genes at different loci and within alternative *E.coli* strains of interest where unsuccessful.

The decision was made to adopt a collection of vectors detailed in a recently published method, 'One-Step Cloning and Chromosomal Integration of DNA' or OSIP (*St-Pierre et al, 2013*). By converting pOSIP integration vectors into BASIC format, it was possible to insert the gene of interest (or multiple pathway genes) into the pOSIP plasmid backbone via BASIC assembly, with the final construct then facilitating subsequent genome integration. The ability to achieve integration with just a single cloning step will dramatically reduce workload.

We anticipated that the intrinsically high accuracy and efficiency of both the BASIC assembly and pOSIP integration methods could be combined to produce an extremely powerful approach (**BASIC-OSIP**), capable of moving from single parts to complex integrated multi-gene pathways in 24 – 48 hours. In relevance to the general focus of my work, it could also allow for detailed quantification of mutation rate at different genome loci and their varying susceptibility to target mutations. There is significant evidence to suggest that epigenetic control of gene expression in prokaryotes is a consequence of DNA methylation and directly influences interaction with transcription factors and DNA binding proteins.

3.4.2 Methodology

Preparation of Vectors

Firstly, chemically competent cells MG1655 and Pir+ (BW25141) were prepared using the method outlined previously to achieve high competency (***Chapter 9.10.1***). pOSIP integration vectors were obtained from AddGene and sequence verified. The collection of vectors includes various integrases which targeted different genome loci based on phage attachment sites. The pOSIP-Flipase plasmid (pE-FLP) was also prepared for subsequent curing of Kanamycin resistance from the integrated strain.

The vectors were converted in the BASIC standard by PCR amplification to remove a multiple cloning site whilst adding the BASIC prefix and suffix sequences (***Figure 3.22***). The gene or genes to be integrated were stored in the BASIC assembly format, and thus newly established BASIC part library could be used. To prevent background (survival of colonies which only included the resistance gene) these insert parts were stored on the Amp resistant pJET storage vector.

pOSIP-BASIC vectors contained the R6Ky origin of replication meaning they were only able to replicate in a Pir+ strain, therefore the plasmid would be lost from the target MG1665 strain. This ensures that antibiotic selection was a result of the resistance gene having been directly integrated in the host genome (***Figure 3.22***).

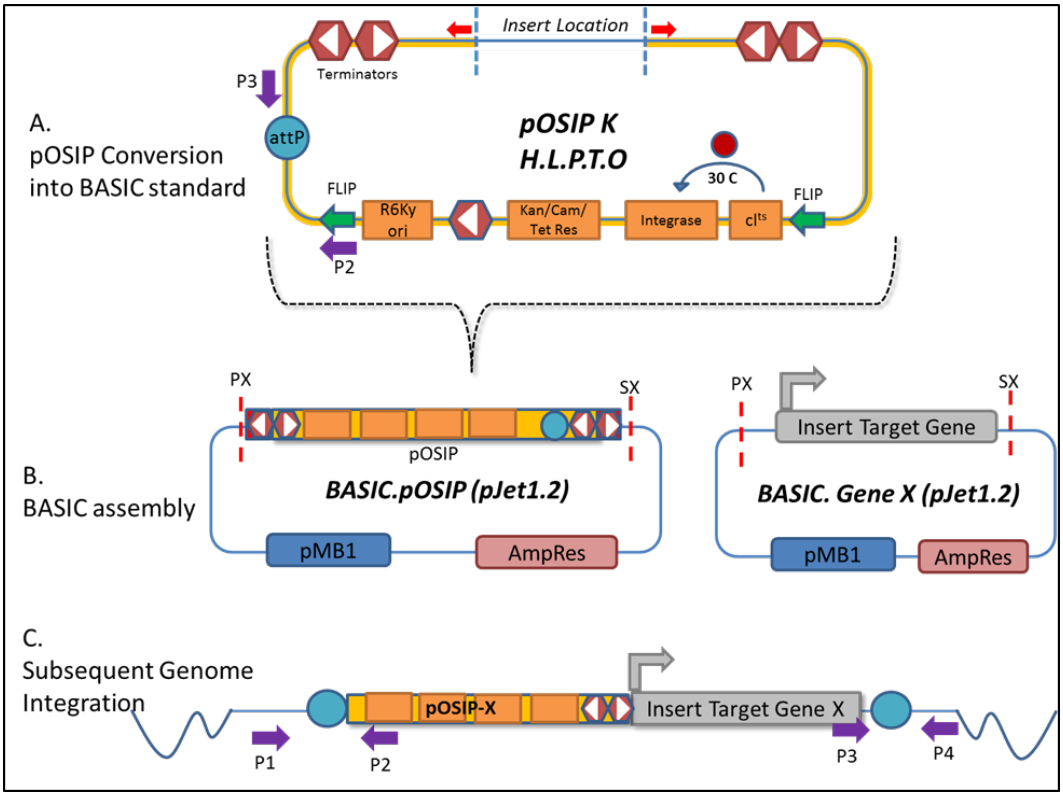


Figure 3.22 *- Preparation, assembly and integration steps for the BASIC pOSIP protocol*

Target Gene Insertion and Vector Assembly

By treating the pOSIP vector and insert genes as standard BASIC parts allowed us to simply follow the BASIC protocol to assembly them together efficiently and accurately (***Chapter 3.2.7***). The assembly reaction mix was transformed into both MG1665 and Pir+ cells simultaneously. 5 µl of the assembly mix was added with 45 µl competent *E.coli* cells. Genome integration should take place in MG1665 and plasmid replication should be possible in Pir+ cells only.

The complete integration and verification methods are described in ***Chapter 9.11.1.*** For the purpose of this work, pOSIP BASIC H, L, P, T, O were used. It should be noted that the ϕC31 integrase attB sequence is not normally present in the MG1665 *E.coli* chromosome. Integration sites and their relative positions are described (***Appendices – Table 11.5)***

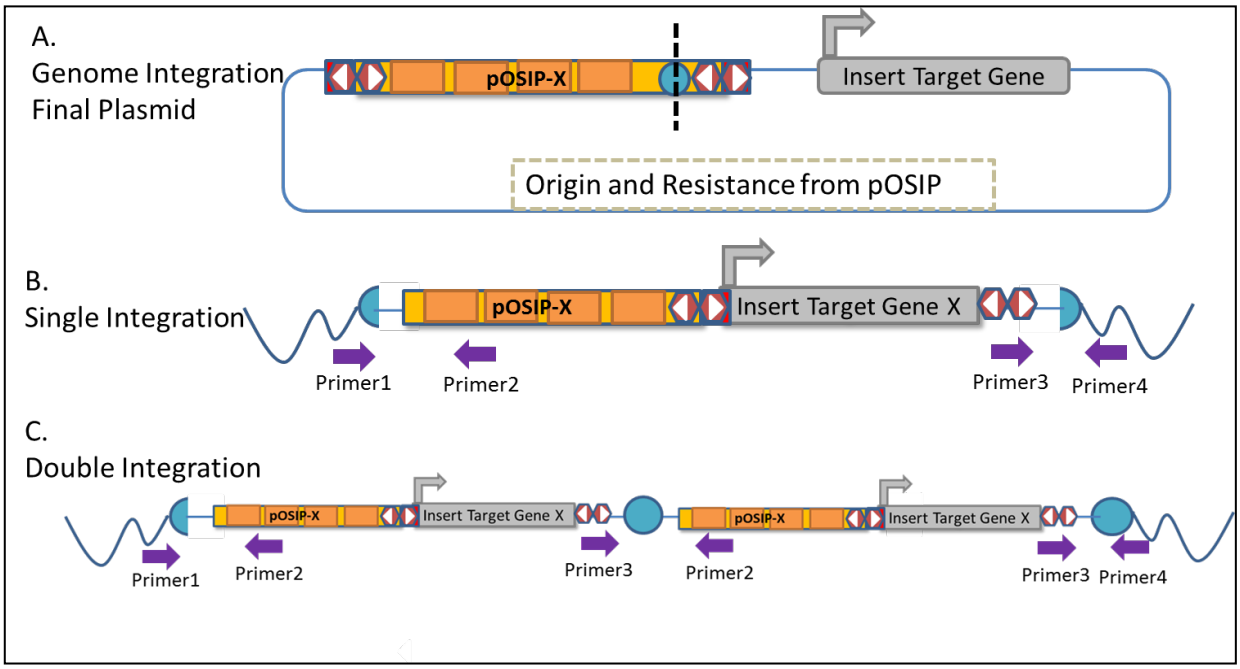


Figure 3.23 *- Verification steps via PCR for BASIC pOSIP protocol*

Table 3.3 *– Analysis of PCR product banding patterns used for BASIC – OSIP diagnostic*

Vector pOSIP	Integration Site	PCR Annealing (C)	No Integration	Single Integration	Double Tandem Integration (additional band)
H	HK022	60	740	343 + 824	427
L	Lambda	60	741	631 + 1156	1046
P	Phi80	55	569	303 + 499	233
T	P21	57	506	622 + 1110	1226
O	186 (1st locus)	60	241	328 + 389	476
O	186 (2nd locus)	60	601	429 + 648	476
C	C31*	60	1150	446 + 1093	389

3.4.3 Proof of Principle Experiment

A proof of principle experiment was conducted to show that the BASIC-OSIP protocol could be used for single and double gene assembly and integration. GFP and RFP expressing cassettes were assembled into the suite of BASIC-OSIP vectors and integrated into the host MG1665 *E.coli* genome, before confirmation of integration and identification of successful expression of the fluorescent proteins.

Firstly, BASIC compatible parts were prepared and a number of different BASIC-OSIP constructs were built (***Appendices - Table 11.16*** *and* ***Table 11.17***). These included both single and double expression cassettes (***Figure 11.4***).

10 colonies were picked for both single GFP integrates and double GFP + RFP integrates from restreaked transformation plates ensuring loss of the initial plasmid. Colony PCR using locus specific primers was used to confirm successful integration as revealed by differing banding patterns (***Figure***

3.24). Correct integration and expression could also be confirmed by flow cytometry to measure fluorescence as an output for expression of the insert genes (***Figure 3.25*** *and* ***Figure 3.26***).

The data presented here refers to genome integration into Lambda attB locus using BASIC-OSIP-KL integration vector, although successful integration was also achieved using the BASIC-OSIP KO and KH vectors, which target the HK022 and 186 att sites respectively (*data not shown*).

This proof-of-principle test used a small sample size (10 independent colonies) but it does appear that there is a reduction in efficiency as the number of insert parts increases. This was initially apparent due to the decreasing number of colonies on transformation plates when complex multi-part assemblies were performed. This is also in keeping with previous builds that use BASIC assembly, potentially a consequence of transformation efficiency.

Importantly, accuracy of assembly remains high for single and double gene inserts and in practice this has meant only 2 or 3 clones need to be checked for each assembly in order to identify a successful clone. The results presented here shows 100% accuracy for single part assembly integration (10/10) and 80% accuracy for double assembly integration (8/10).

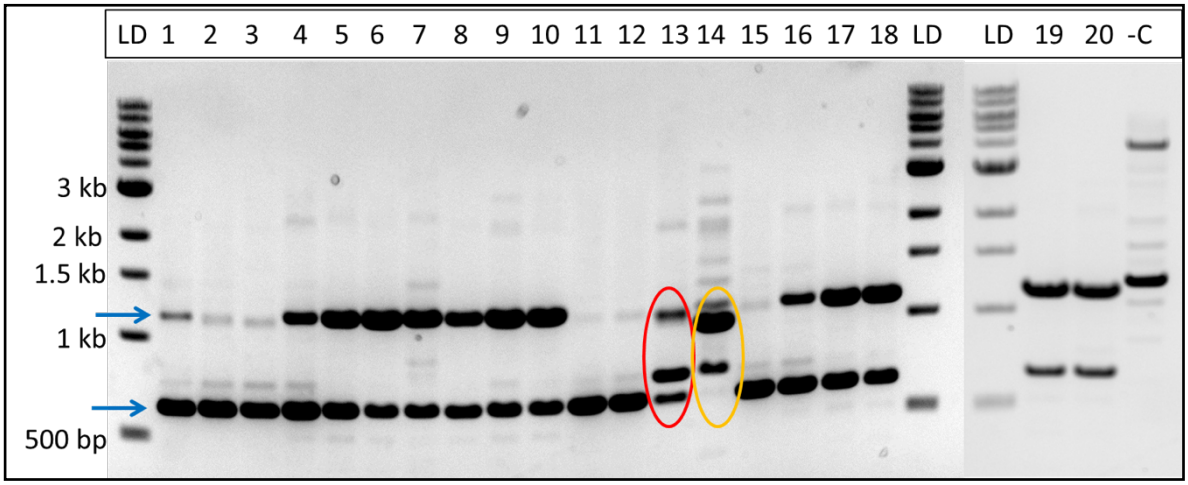


Figure 3.24 *– Diagnostic colony PCR was used to determine whether selection colonies showed successful genome integration. Single (1-10) and double part (11-20) assemblies into the BASIC-OSIP KL vector. Blue arrows correspond to the anticipated bands for correct single integration at Lambda site (631+1156 bp). Wildtype MG1665 strain was used as a negative control (-C) and shows distinctly different banding pattern when no integration has taken place. Double integration 13 and 14 (circled) appear to show incorrect banding patterns as indicated by a band at 741 bp. 1kb ladder (LD) is used as a reference marker during gel electrophoresis separation of PCR products.*

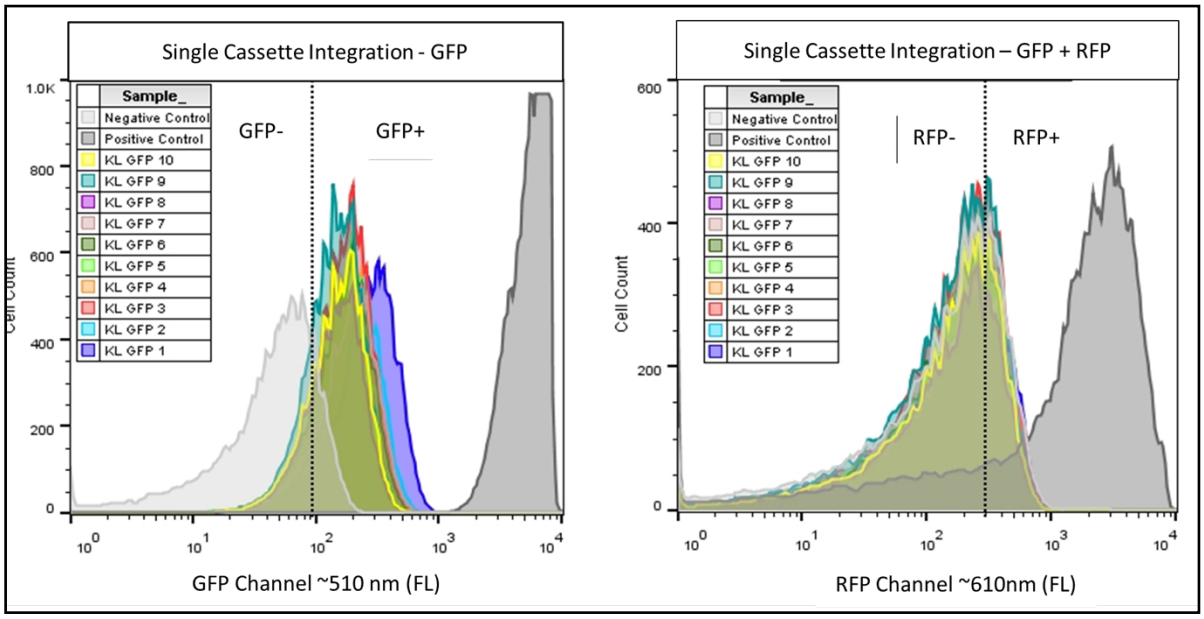


Figure 3.25 *– Histogram traces for single gene cassette in which a GFP gene was integrated into the Lambda locus in MG1665 E.coli strain. Flow cytometry confirms GFP emission for all 10 transformant colonies (emission at ~510nm). Single gene cassettes did not include RFP and thus no emission was seen from and RFP channel (~610nm). Comparisons were made against blank cells (negative control) and GFP/RFP expressing plasmid (positive control). Positive and negative cell populations are separated by a dotted line.*

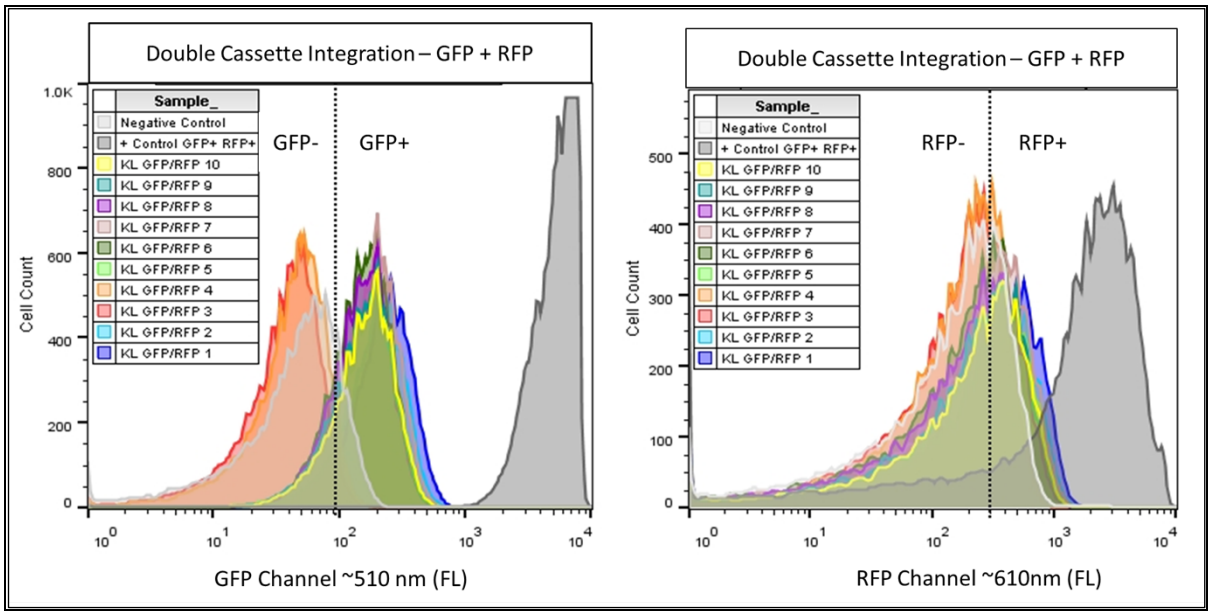


Figure 3.26 *- Histogram traces for double gene cassette in which both a GFP and RFP gene were assembled together and integrated into the Lambda locus in MG1665 simultaneously. Results correspond with diagnostic PCR verification showing that double integrates 13 + 14 (orange and red traces) do not show GFP or RFP expression. All other samples gave strong emission in the GFP channel but weak emission in the RFP channel.*

Unfortunately, due to the fact that both assembly and integration happen within a single step, it is difficult to determine whether incorrect clones are a consequence of inaccurate assembly or inaccurate integration.

Encouragingly, it does appear that high overall accuracy provides a more robust method. The substantial time and effort to improve this capability now makes it possible to insert any target gene into a range host *E.coli* genome loci. In theory, it should now be possible to build a large library of target gene variants and to explore questions of loci specific expression and whether particular positions are more susceptible to mutation by the AT7 mutator device.

I believe this capability extends the value of the BASIC standard, and that the BASIC-OSIP vectors should be included in any future DNA part distribution. Highly accurate and quantitative characterisation of functional DNA parts may favour single gene copy expression platforms if we consider that plasmid-based expression systems cannot distinguish between rates of transcription and translation because plasmid copy number remains an unknown variable. For example, the cell plasmid copy number is likely to change depending on growth phase. Therefore, copy number cannot be uniform within a batch culture and thus quantification is favour total population measurements using plate reader measurements, but these are not conclusive.

Alternatives options include cell free expression systems to control template DNA and more accurately quantify transcription and translation, but these fail to accurately recreate true cellular conditions (*Chappel et al, 2013*). Therefore, single cell measurement using techniques such as flow cytometry would be favourable, and thus an efficient, accurate and robust genome integration method is needed meaning the method outlined here has real value, providing the researcher with a tool to move from prepped DNA parts through to integration of multi-part pathways within just two days.

3.5 Conclusions

BASIC assembly has been very successful in providing a reliable means for modular assembly of quite complex constructs throughout my own research and has shown some uptake in the wider research community. The accuracy and efficiency of multi part assembly compare favourably with the latest improving to existing methods such as Golden Gate and USER assembly (*Marillonet and Werner, 2015., Caveleiro et al, 2015*). Numerous research groups and industry partners have decided that BASIC is an excellent DNA assembly standard, becoming their preferred choice, specific to their application or as inspiration to their own assembly methods with the original paper being citied in numerous works (Mitchell et al, 2015., Iverson et al, 2015., Jin et al, 2016).

4 Building a Controllable Gene Circuit

Synthetic biology is built on the predication that if biological parts, including DNA elements such as promoters, repressors, enhancers, can be highly characterised and understood as individually entities, then a parts-based engineering-like approach can be used throughout design, build and test and dramatically improve our ability to create complex biological systems. Ultimately, it is now widely accepted that these design principles can be used to construct multi-level gene circuits (*Brophy and Voigt, 2014*).

During my time as part of the Centre for Synthetic Biology and Innovation (CSynBi) at Imperial, there have been two main underlying goals.

The first departmental goal was to create a large database, or registry, of professionally characterised genetic parts known as Synthetic Biological Information system (SynBIS). This should include in-depth data sheets, available via open-access to the international Synthetic Biology community. It is critically important that there are fixed characterisation standards to ensure that data is transferable and can be used to build highly accurate models and as such the International Open Facility Advancing Biotechnology (BIOFAB) (*Mutalik et al, 2013*). The number and type of characterised parts remains sparse and therefore it was my intention to abide by the general standards outlined by CSynBi /SynBIS and to provide clear data sets for parts that which are not only beneficial to my own model but could be added into the parts registry.

The second target was to establish a state of the art liquid handling platform which has the capability to automate high-throughput characterisation. The need for a robust method quickly became apparent and I was able to use a newly established protocol to characterise two sets of inducible promoters for inclusion in my model.

This chapter outlines the work carried out to establish a library of inducible promoters based on the Tetracycline operons found within prokaryotic systems.

I intended to create and characterise two sets of repressible promoters by taking advantage of the current infrastructure available within CSynBi. As such I am indebted to the work of Dr Chris Hirst and Catherine Ainsworth for their work to established automated characterisations protocols and data analysis, respectively.

I believe this collection of promoters would provide me with the repertoire to more accurately predict and finely tune the overall directed evolution circuit.

4.1 Background

4.1.1 Circuit Design

Fundamental to the overall project was the design, modelling and implementation of a genetic circuit with the ability to control mutation. As mentioned, the system was to be programmed for directed evolution of the target gene and linked with an output signal that the desirable end point has been reached. When the desired phenotype has been identified, the circuit will respond by switching off any further mutation. In return, it will switch on an output and selectable marker. The specifications of the circuit also require the threshold for selection (or switching) to be tuneable so that evolution can take place at varying levels of stringency. This should be possible with correct engineering of the response module in accordance with small molecule inputs. For example, lowering the concentration or type of small molecule (for which specificity is being engineered), it should be possible to increase stringency of selection.

The complete circuit can be briefly explained as follows (**Figure 4.1**):

1. Activation of the mutator module is identified by expression of RFP
2. Constitutive expression of the AT7 mutator protein (***Chapter* 2**) and an error-prone polymerase (***Chapter 5***) will direct mutations to the target gene of interest as a result of guided deamination and subsequent error-prone repair.
3. Many variants of the target protein will be produced, in this instance, the quorum sensing Las receptor.
4. Addition of substrate small molecules (AHL) will eventually result in binding with compatible LasR variants.
5. The LasR/AHL complex will activate the Output module and cells will begin to show green fluorescence.
6. Initially, expression of a Tet Repressor protein will inhibit any further mutation and expression of the wildtype T7RNA polymerase will encourage high levels of the specific LasR variant of interest (creating a positive feedback).
7. The output module will also lead to expression of an antibiotic resistance (Kanamycin) so that cells of interest can be easily selected and sequenced.

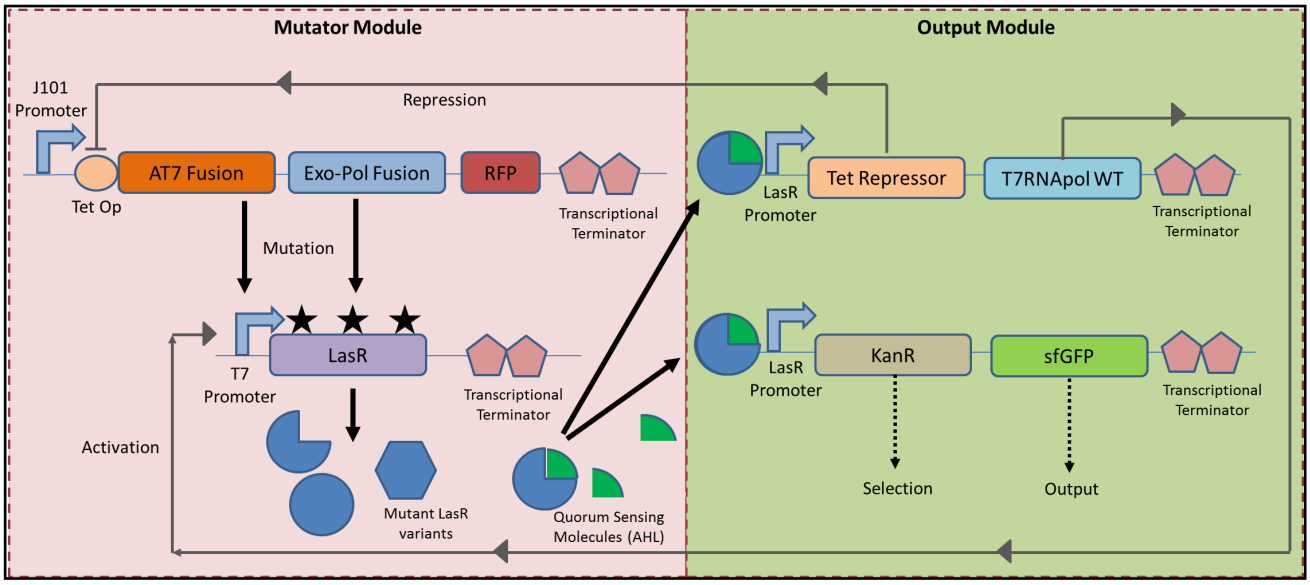


Figure 4.1 *– Overall circuit to control directed evolution of a target protein. Expressing of the mutator protein (AT7 fusion) alongside an error-prone repair protein (Exo-Pol variants) will lead to mutation of the LasR protein. Binding of the LasR to target small molecules (AHL) result in activation of the output module and leads to repression of the mutator module, providing an automated selection of the desirable target protein*

The main focus of this work is to add an extra level of stability to the gene circuit. Transcriptional and translational regulation in biological systems is inherently noisy and even 'simple' prokaryotic hosts are complex multi-dimensional systems, with numerous metabolic pathways and cross-signalling events happening concurrently. The stochastic nature of such an environment is an underlying problem when attempting to predict, model and control the activity of the circuit within the biological host.

However, through the use of well characterised genetic elements such as repressors, activators, operators and enhancers, protein expression can be controlled at a transcriptional level. Firstly, the localisation of an active transcription factor (TF) to the promoter region upstream of the target gene is required to initiate transcription of mRNA and subsequent translation of a functional protein. The TF must interact with the coding strand in some way and will likely contain a DNA binding motif, specific to its target promoter sequence.

The activity of the TF itself can be controlled in two ways. Firstly, a repressor protein that demonstrates high binding affinity for a target operator sequence (positioned in close proximity to the target promoter sequence) can block transcription. Alternatively, the TF may exist as a non-active form until an activator / inducer is introduced into the system.

In this instance, the proposed circuit is dependent on both forms of transcriptional regulation / activation. A robust genetic switch is critically important to this system's ability to change from mutator mode to output mode. Firstly, binding of the optimal LasR variant with its corresponding AHL is required to form an active TF complex which can initiate downstream expression from its cognate promoter.

Furthermore, the decision to link the pLasR promoter with both output selection and the repression of the mutator module means that the activation strength of pLasR needs to be very well characterised. The ability to strongly repress the mutator module is also vital, in order to avoid moving beyond the desirable end point i.e. when a functional protein has been identified. This is dependent of understanding the interaction between TetR and corresponding TetOp binding sequences and also needs to be fully characterised.

4.1.2 Adapting the Tetracycline Resistance Operon

The prokaryotic tetracycline resistance operon, was first identified within a bacterial transposon and provides a phenotypic resistance to the antibiotic tetracycline (*Hillen et al, 1984*). It is well understood as containing operator regions that are capable of repressing gene expression and the organisation of the operon is such that the Tet promoter drives expression of the antibiotic resistance protein (TetA) as well as the repressor protein (TetR). In normal conditions (without the antibiotic), expression of the TetA resistance protein is repressed due to the fact that an operator region (TetOp) located at the 5' UTR will strongly bind to TetR. This blocks the hosts innate transcription factor (RNApol) by steric hindrance of the template DNA strand. However, in the presence of tetracycline, the TetR molecule will preferentially bind to the antibiotic and thus the affinity of the TetR for the TetOp region is greatly reduced. Ultimately, when the antibiotic (or an analogue of Tet e.g. anhydrotetracylcine -aTc) is present in high concentration then the antibiotic/repressor binding outcompetes the repressor/operator binding, resulting in expression of the TetA resistance protein, which is in fact a membrane bound protein able to export the antibiotic from the cell. Once a sufficient level of Tet has been exported, the concentration of TetR increases once again and negatively regulates TetA expression. Therefore, the cell produces the resistance protein needed for survival only when it is required, and this serves to reduce the metabolic burden of having to maintain production of a membrane protein channel.

Further exploration reveals that the 208 aa TetR protein forms a homodimer which is able to interact with the major groove of two palindromic Tet operators (TetO1 and TetO2) sequences with differing binding affinities and varying disassociation constants. Extensive studies into the structural basis of TetR homodimer and TetOp interaction has been revealed by crystal structures of the complex, with binding directly linked to Thr26 and Lys48 of the TetR protein (*Orth et al, 2000., Tiebel et al, 2000*). The conformational changes that are caused when the inducer (Tetracycline or aTc) binds with TetR also reveal the mechanism of induction (*Hinrichs, 1994*). The unwinding and shifting of helices in the tertiary structure of both TetR monomers causes subsequent separation of the DNA binding domains and abolishes binding of the TetR with TetO1/TetO2.

It is important to note that there is some genetic variation amongst different bacterial species which contain Tet resistance operons, across five distinct determinant classes (A-E). The crystal structures revealed in the aforementioned report is based on class D, whereas class B is more commonly found in *Enterobacteria* such as *E.coli*. The author explains that TetR proteins between species share about two thirds homology and therefore the comparison is unlikely to be completely correct. However, there are strong similarities in terms of mode of action, given that the corresponding TetO1 sequence only differs by a single base pair and the TetR monomer contains the same two small N terminal DNA binding domains which corresponding to the major groove of TetOp across all classes.

These specific operator-repressor-antibiotic binding affinities have been widely exploited to construct carefully controlled transcriptional activation systems in the form of Tet-On and Tet-Off approach used in numerous eukaryotic expression assays (*Gossen and Bujard 1992., Goosen et al 1995*). Both are inducible systems that respond to Doxycycline (a stable analogue of Tetracycline) and both turn gene expression on. 'On' refers to the fact that expression is possible when the system is in the presence of Dox and 'Off' refers to a system which expresses when Dox is absent from the system.

The Tet-Off system relies on a fusion protein consisting of the TetR protein to ensure targeting to the 5' TetOp sequence and a viral protein (VP16) activator domain from the herpes virus, which encourages transcription. Therefore, this fusion protein can bind to the TetOp region located upstream of a minimal CMV promoter. In normal conditions, the fusion protein binds to the operator and the promoter leading to increased expression of the intended target gene located downstream. However, when Dox is added to the system, TetR binds with Dox and the TetR/TetO association is lost meaning that transcription is prevented.

Alternatively, the Tet-On system and specifically another variation of the fusion protein, requires binding to Dox to allow for any association with the operator and to permit transcription. Although the Tet Operon is a prokaryotic system, its use in mammalian cells is very affective because it remains biologically inert and orthogonal in many hosts.

Focusing on my current work in prokaryotic *E.coli* host, this would be an excellent opportunity build a library of Tet inducible promoters that would ideally lend themselves to the overall control of the *in vivo* device in question.

I intended to construct a number of inducible promoters that are a hybrid of a well described constitutive sigma70 promoter (J23101) and variations of the repressible Tet operators. These could be screened for favourable characteristics, in terms of their repression strength, cooperativity, optimal inducer concentration and levels of expression post induction. I expected to be able to calculate reliable parameters by generating a number of dose response curves.

Although other promoter/operator parts such as Lac and Ara have also been widely used within the field, I have opted to build a Tet controlled system for the fundamental reason that the small molecule inducer is inert and that the alternative choices are often affected by nutrient stress (*Viswanathan et al, 2016*). Therefore, they would be largely affected by a continuing change within the growth medium (*Skerra 1994., Lutz and Bujard, 1997*). For example, Lac based systems are known to be particularly leaky under nutrient stress which is commonplace during extended periods of growth. Arabinose is itself is a monosaccharide that could be found as a carbon source in many variations of growth media and induction is therefore very difficult to control or predict (*Guzmann et al, 1995*).

Given that early iterations of the mutator system suggest that long periods of changeable logarithmic growth and subsequent stationary phases are needed for any significant levels of mutation, the Tet system is likely to be much more reliable over a longer period. Furthermore, external addition of the aTc enables well defined point of induction with the possibility to tune expression levels by varying the concentration of inducer. aTc shows greatly reduced antibiotic activity and is not detrimental to the cell at the very low concentrations required for successful induction and should not alter cell growth. Therefore, I believe a well characterised set of J23101.TetOp hybrids will be extremely beneficial to the overall circuit design and subsequent modelling of the system.

4.1.3 LasR Target Protein

The proof -of-principal circuit is designed to engineer LasR based transcriptional activators for the purpose of developing detection devices in security and medical applications. LasR variants work in response to acyl homoserine lactone molecules (AHLs) and trigger biofilm formation in many pathogenic organisms, such as *Pseudomonas aeruginosa (Riedel et al, 2001)*. Different organisms typically produce and respond to different AHL signals, although there is some cross-species reactivity. Therefore, orthogonal LasR AHL pairs could be tailored to create highly specific signal and response pairs which are discrete for a specific organism or they may be engineered to respond to a broad spectrum of small molecules specificity. Ultimately, these genetic parts could be tuned for biosensor applications and expanding the synthetic biology toolkit.

A number of publications have outlined work to diversify another quorum sensing system, LuxR, using directed evolution (*Yokobayashi et al, 2003., Collins et al 2005., Hawkins et al, 2007*). However, the experimental approaches are labour intensive and incorporate many of the techniques outlined previously (***Chapter 1.4***). They do not allow for a continuous, streamlined and automated approach. Therefore, engineering the LasR system should also provide a strong proof of principle with clear comparisons to existing directed evolution approaches.

4.2 Methodology

General methods referred to below are explained in further detail in (***Chapter 9 - General Methods)***. Any differences or exceptions from these standards are clearly identified.

4.2.1 Creating a Library of Tetracycline Repressible Promoters

The Tet system had been identified as a suitable candidate system which offers a model for strong transcriptional regulation and repression. TetR binds to the TetOp through interaction of an N-terminal DNA binding domain and therefore binding affinity is likely to be affected by both critical TetR residues and the TetOp sequences themselves.

One might anticipate that the wildtype forms of TetOp and TetR to have maximal binding affinities if their innate biological purpose is to strongly repress expression of a resistance gene. However, this is an incorrect assumption, because natural evolution has ensured that the binding and association parameters of TetOp, TetR and Tet are finely tuned to levels of antibiotic tolerance. Therefore, TetR/Tet binding affinity is very high given that the cell needs to immediately switch on expression of TetA resistance gene as soon as needed. It follows that the TetR/TetOp binding affinity must be sufficiently reduced to the point that TetR conformational changes cause disassociation with TetOp. Therefore, it seems possible that a higher strength repression could be engineered by changing the TetR/Tet binding interface, or the TetOp sequence itself.

All things considered, attempting to engineer a protein/analog interface is likely to be very problematic given that changes to protein structure will inevitably lead to many non-functional variants. However, generation a diverse library of hybrid Tet operators with a range of characteristics should be feasible, given that previous work in the literature has already established differences in binding affinities of both TetO1 and TetO2.

Initial brainstorming identified three ways to generate variable and analogous operator libraries that would demonstrate variable expression profiles based on their ability to repress *E.coli* sigma70 RNA polymerase binding to a J23101 constitutive promoter and prevent expression of downstream output sfGFP:

1. The position and distance of proximal TetOp sites to the J23101 promoter.
2. The choice of TetO1 and TetO2 given differing binding affinities (the later binds TetR with five-fold greater affinity)
3. The intervening sequence and distance between the operator sites, given that individual TetR homodimers bind to target operator sequences with some degree of cooperativity.

Option 3 is particularly interesting given that TetR homodimers bind to the major groove at both sides of the palindromic DNA operator sequence. The major groove at the recognition site is widened whereas the major groove on the opposite side to the binding interface is narrowed. Partial unwinding of the DNA duplex may improve binding of the second TetR homodimer (*Tiebel et al, 2000*). Furthermore, the angle at which the second bound TetR homodimer is positioned may mean that is brought into close proximity with the adjacent TetR already bound to the other operator sequence, creating a strong intermolecular association. Both factors may influence the level of cooperativity between the two TetR molecules, in that the binding of the first repressor aids the binding of the second repressor and actually leads to an increase in overall binding affinity if both are bound at the same time. This can be best quantified by the Hill coefficient, which explores how the binding of a ligand to a macromolecule can be enhanced by prior binding of separate ligands.

I therefore propose that specific nucleotide changes to the intervening operator sequence will likely affect the dsDNA helical structure which would result in a change in cooperative binding and ultimately this could help create a hybrid library that enables me to fine tune repression within the wider genetic circuit.

If we consider that the system will link the constitutive expression of the TetR protein to the inducible expression of a fluorescent output, the assay will be able to measure strength of repression as different concentrations of inducer molecule anhydrotetracycline (aTc) are added. By building a test construct consisting of a constitutive promoter and Tet operator hybrid, placed upstream of the sfGFP output gene, the visualised level of fluorescence directly infers the level of transcription allowed upon induction. The relative strength of a series of novel promoter/operator hybrids can therefore be characterised. A simple depiction of the test circuit and general library is provided (***Figure 4.2***).

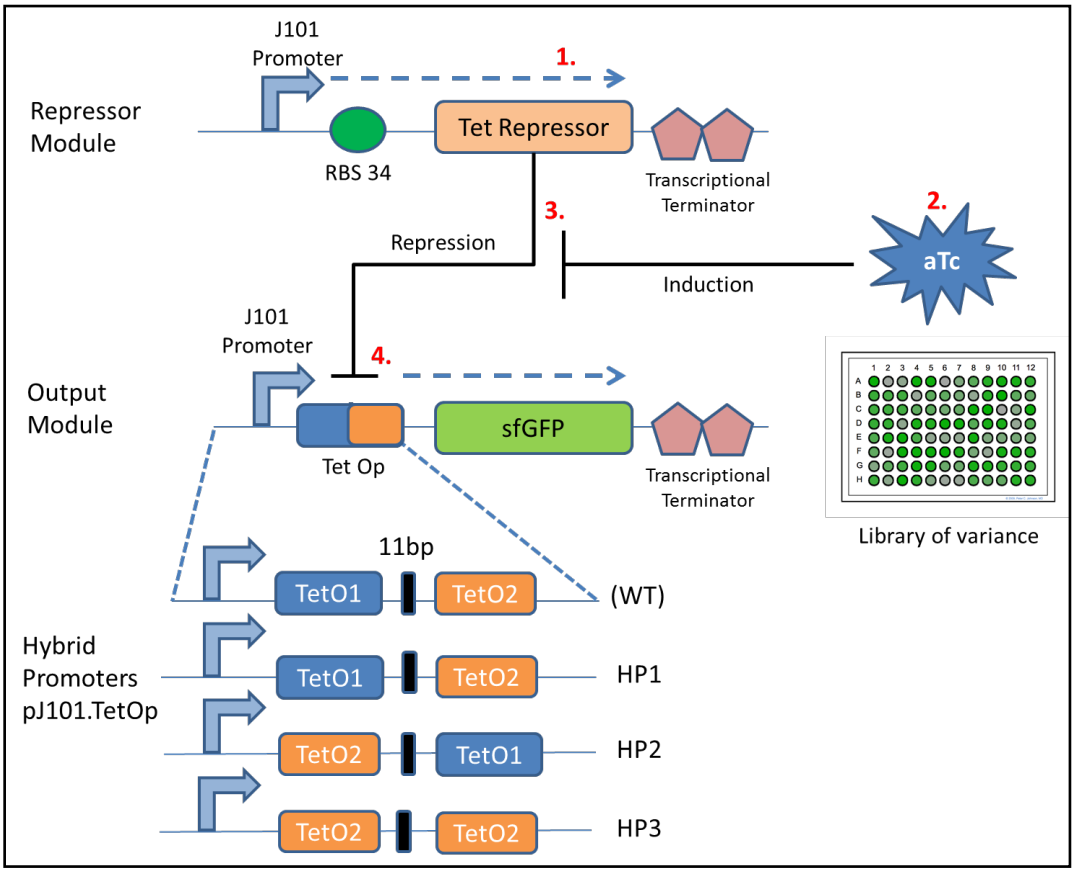


Figure 4.2 *– A simple test construct used to characterise the library of hybrid pJ23101.TetOp promoters*

The level of TetR will be constant throughout and will act as a background control, this can be estimated given that the strength of expression when the J23101 promoter is coupled with RBS34 has already been established. Upon induction with varying concentrations of aTc, TetR/TetOp binding will permit expression of the output sfGFP by measurement using a fluorescence spectrophotometer. The changing levels of sfGFP expression will enable me to determine which operator sequences are most strongly bound to TetR and can one again be directly linked to the strength of the J23101 promoter.

It is likely that low concentrations of the inducer will not allow for transcription to take place and thus there will be no recordable fluorescent output. Furthermore, it should be possible to identify small stochastic events which occur even when the system is not induced. Comparisons with the wildtype promoter/operator may show a reduction in leaky expression of the sfGFP output, if the library created demonstrates improved level of repression.

To avoid the potential effect on repression that the Tet promoter/operator overlap may have, it was decided that a constitutive promoter should be placed upstream of the TetOp sequences as shown (***Figure 4.2***). Therefore, a default level of expression should be consistent and comparable. Furthermore, previous characterisation of this set of sigma70 promoters allows sensible decisions to be made when deciding on an optimal level of output expression.

Another consideration in the design relates to which RBS sequence is chosen because sufficient levels of translation are required to produce the output sfGFP protein. Although it is not necessarily beneficial to produce such large quantities of sfGFP protein because of the excess metabolic load placed on the cell, a medium strength RBS is required to achieve levels of fluorescence within a dynamic range of recordable values and importance is placed on identifying leaky events.

Further considerations were made regarding how best to increase the sensitivity of the assay and how to quickly identify output fluorescence post induction. Therefore a further set of operator libraries was assembled except the sfGFP gene was tagged with the ssrA degradation tag (AANDENYALVA) on the 3' end, creating an extended sequence at the C terminus which marks the protein for identification and degradation by the ClpXp system in *E.coli* and thus greatly reduces the half-life of the protein. Degradation tags have often been employed to minimise the effect of stochastic leaky events such as the events we hope our assay will identify. The rational here was to use the tagged libraries to remove any leakiness that could mask sudden increase in transcription post induction.

4.2.2 pJ23101.TetOp Library Construction

Hybrid (HP) Tet operator libraries were designed. The first consisted of the wildtype TetO1 and TetO2 sequence, named p101.Tet-WT. HP1 introduced random nucleotide variability within the 11bp sequence between TetO1 and TetO2. HP2 included TetO2 and TetO1 in opposite positions compared to the wildtype and also maintained random nucleotide variability across the library. HP3 included two copies of the TetO2 operator (known to show superior binding affinity for TetR) and once again maintained the variable 11bp region.

All test operator libraries were built from synthesised ssDNA oligos with complementary ~20bp of overlaps and included degenerate bases required to make the 11bp variable region. Oligos were annealed and extended using a Klenow Polymerase extension protocol to produce short stretches of double stranded DNA (***Chapter 9.3.2***). The dsDNA blunt end fragments were then ligated into the pJET 1.2 cloning vector (***Chapter 9.5.5***). By avoiding any PCR amplification and using a highly efficiency blunt ligation whilst using a positive selection for accurate inserts, the vast library could be screened more accurately and comprehensively. It should be noted that the transformed cells were not plated or selected as single colonies, rather they were grown up as liquid cultures overnight at 37°C (LB media, Ampicillin). This was to ensure that complete libraries (HP1-4) could be screened. The library cultures were retained as glycerol stocks and stored long term at -80°C until required. A single transformation control for each sample was plated and sequenced verified to confirm that the insert was correct.

4.2.3 Building the Tet Repressor and Output Module

Functional parts, were obtained from the Registry of Standard Biological Parts, and were sequence verified.

Firstly, the TetR module was built. Initially the BBa_P0040_pSB1A2 plasmid was taken from the registry (pSB1A2 refers to the standard BioBrick vector, A denotes Ampicillin resistance), (P0040 refers to the BioBrick part itself). Site directed mutagenesis was used to add the J23101 promoter sequence upstream of RBS34. A gradient PCR was used to amplify the complete vector with 5' phosphorylated primers used to introduce the J23101 sequence at the correct location. Correct PCR products were ligated and transformed into DH5a *E.coli* and plated on LB Ampicillin (100 µg/ml) agar plates for selection. Two colonies were picked, grown overnight, miniprepped and sent for high-throughput sequencing to confirm that the correct changes had been introduced and the TetR module was complete.

The sfGFP output module was constructed using similar steps. BBa_I20260_pSBK3 was acquired from the registry because it already contained the J23101, RBS32 and B0015. GFP-mut3 was not the preferred choice in this assay because of the slower maturation time and the presence of an internal BsaI, which would be detrimental to later assembly using the BASIC assembly standard. Primers with 5'end extensions were used amplify sfGFP and the insert vector under standard PCR conditions and correct bands were isolated by gel extraction. Both parts included long overlaps compatible with my own version of the Gibson Assembly protocol (***Chapter 9.5.1***). Assembled constructs were heat shock transformed into DH5a cells and selection on LB Kanamycin (50 µg/ml) plates. 10 colonies were picked from the plates and grow overnight in LB Kanamycin at 37°C before plasmid DNA preparation. Diagnostic restriction digests were performed, including HindIII (which cut in the vector) and EcoRI (which cut in the insert) to reveal two distinct bands on an agarose gel. 8 out of 10 colonies revealed the correct banding pattern (in agreement with estimated accuracy of a 2-part Gibson Assembly).

As discussed in the design rationale, the sfGFP fluorescent proteins included an LVA degradation tag. Untagged versions were also required, and site-directed mutagenesis was employed to eliminate the tag by inserting a double stop codon immediately upstream in the CDS.

During the time taken to construct the individual modules, the BASIC assembly standard was simultaneously being developed (***Chapter 3 - BASIC Assembly***). It was decided that this would be the most suitable way to construct the final test plasmid and so each module was amplified in its entirety, with the addition of BASIC prefix and suffix sequences.

The final test construct also contained the low copy origin of replication p15A and a Kanamycin resistance cassette in BASIC standard. A low copy number plasmid was believed to be suitable given previous work to establish the downstream characterisation protocol on the CSynBi robot platform, and

it was anticipated that the level of fluorescence would be easily recordable, without producing excessive burden on the cell. The functional modules were added in the opposite orientation to the origin and resistance cassette to minimise transcriptional read-through.

Transformants were selected on Kanamycin LB agar plates, picked, miniprepped and confirmed by restriction digest using HindIII (cut in vector) and ScaI (cut in the sfGFP module) revealing correct banding patterns in a number of the chosen colonies.

4.2.4 Final Construction of Test Device

The strategy for inserting the pJ101.Tet library into the test vector was performed by adding BsaI recognition sites to the vector template by site directed mutagenesis. BsaI is a Type IIs endonuclease and cuts outside of its recognition site, meaning that assembly can be scarless. This was particularly important in order to maintain sequence context within the 5'UTR region (as changes here are likely to have a significant bearing on repression and expression.

This approach required the removal of the J23101 promoter located upstream of the sfGFP output gene by site directed mutagenesis. Inverse PCR primers were designed to amplify the vector backbone, remove the promoter site and add two flanking BsaI recognition sites with compatible overhangs.

Correct products of the site directed mutagenesis step were identified by gel electrophoresis, transformed into DH5a cells and selected on LB agar Kanamycin plates. The loss of the promoter meant that subsequent colonies did not show fluorescence imaging. Sequencing confirmed the presence of the flanking BsaI sites.

Insertion of the library was now possible. Cutting out the final vector and the pJ23101.TetOp library variants with BsaI exposed complementary 4bp overhangs. Linearization of the vector enabled the small library fragment to be ligated in the correct orientation. This results in a loss of the internal BsaI recognition site within the final construct and encourages the digestion/ligation to progress in the forward direction only. The single pot reaction includes both BsaI and T4 ligase and was held at 37°C for 1 hour, 20°C for 1 hour and 65°C for 20 minutes. The final ligation mixture was then transformed into DH5a cells and plated onto Kan (100 µg/ml) and aTc (2000 ng/ml) LB agar plates for selection.

Selection on the inducer aTc revealed fluorescence in those colonies that had correct promoter/operator inserts. Imaging was performed on a FujiFilm FLA-3000 fluorescence image analyser using a blue excitation laser and green channel emission filter. sfGFP expression was used as a marker for correct assembly as it was not feasible to sequence each individual construct generated by such an extensive library. Some colonies did not show GFP output suggesting the accuracy of assembly was not 100%, and most likely due to parental vector background. Final confirmation by restreaking onto AMP (100µg/ml) and KAN (50µg/ml) LB plates simultaneously showed that without the inducer, expression of

TetR was sufficient to completely repress GFP output. Each individual colony was used to prepare a glycerol stock for downstream characterisation assays (***Appendices -Figure 11.5***).

In total, ~20 individual variants from 8 separate libraries were tested i.e. the four different promoter operator arrangements (HP 1-4) placed upstream of 2 different output reporter gene modules (sfGFP and sfGFP.LVA).

4.2.5 Manual Induction and Characterisation Assays

Given the size of the test library, subsequent induction assays were carried out in flat bottom 96-well plates (*Costar*), a 96-well plate incubator (*Mikura Ventura 2000*) and spectrophotometer plate reader (*BioTek Synergy HT*) and KC4 analysis software.

200 µL EZ Rich Defined Medium (MOPS) with kanamycin was added to each well in across a 96-well plate. Addition of 5µL of the glycerol stock from 8 promoter variants from each promoter library were used to inoculate each well, alongside 8 positive control wells (expressing GFP), 8 negative control wells (blank vector) and 8 empty wells (media only), repeated twice across the plate. The plate was then incubated at 30˚C 700 RPM. OD 600nm measurements were recorded using the plate reader and samples were normalised to 0.2 in 200 µL MOPS media. Half of the samples and controls also received 1000 ng ml^{-1} aTc. The samples were incubated at 30˚C at 700 RPM with OD600 and fluorescence was recorded (excitation: 485 nm, emission: 520 nm) were taken every hour for a total of 4 hours. Average media OD600 and fluorescence readings were subtracted from sample readings at each time point to determine a relative fluorescence unit for each promoter. This initial screen demonstrated broad differences across the operator library and helped determine which variants to characterise more thoroughly on the robotic platform.

Flow cytometry was also used to help explore low levels of expression and leaky events, with the assumption that small fluorescent subpopulations could be identifiable. Therefore, 20 µl of each sample was removed from the 4-hour induction plate and added to 180 µl dH_20 in falcon tubes for immediate measurement on the flow cytometer. Raw data was collected from 50,000 events for each sample, with excitation 490 nm laser and emission 510 nm filter used. Flowing Software (*Terho, 2013*) was used to analyse the data, suitable gates were selected, and distinct regions of fluorescent and non-fluorescent populations were identified in comparison with controls. Useful overlay histograms were generated which showed obvious differences in the fluorescent peaks across the libraries.

4.2.6 Automated Characterisation on a Liquid Handling Platform

As mentioned previously, a robotic liquid handling platform had already been established within CSynBi and was capable of running very accurate high-throughput characterisation experiments (*Hirst, 2014*). This was used to prepare plates and media, normalise cell cultures, maintain steady growth, take

OD and FL values, aliquot samples for flow cytometry analysis and output data points as raw data to create accurate growth and induction curves. Therefore, existing protocols were adapted to fit with this experiment and with the help of Dr Chris Hirst a new script was devised which allowed multiple repeats to be measured within the same run. For ease of use, all samples and controls were condensed onto a single 96 well plate per run.

Chosen promoter operator variants were transformed into chemically competent MG1665 *E.coli* and plated on LB Kanamycin plates. Individual colonies were picked with a sterile pipette tip and use to inoculate a starter culture of 1 mL LB kanamycin and grown 37°C for 4 hours. A flat-bottomed 96-well plate was prepared with 200 µL LB kanamycin across all wells. Each well was inoculated with 20 µL of the starter culture. The plate was then covered with a Breathe-Easy® sealing membrane (*Sigma Aldrich*), incubated at 30°C and shaken at 600RPM overnight for around 16 hours.

The following day, the sample plate was placed into the same model of incubator which was preset to 30°C and 600RPM except this was within the footprint of the CSynBI robotic platform. All following steps were automated by the robot, executed under a pre-written script. The robot was able to continue independently for much of the assay, with limited human interaction.

Firstly, 15 µL of each sample was diluted in 282 µL MOPS defined media. The absorbance at 600 nm was then measured using a microplate reader, followed by normalization with MOPS media to an absorbance of 0.014. The plate was then incubated for 80 minutes at 30˚C and shaken at 600RPM. Following this outgrowth, the OD of the samples were measured once again, and the samples normalised to 0.035. 15µL of the samples were then diluted into 85µL of MOPS media (with varying aTc inducer concentrations (0, 5, 10, 20, 100, 250, 500, 1000 ng/µl for Ptet101 induction) in a black microplate (*Greiner*). OD600 and GFP fluorescence (excitation of 485 nm and emission of 520 nm) were recorded every 15 minutes over the 6-hour assay. Raw data was exported into a pre-defined Excel spread sheet which allowed for initial visualisation of results and a qualitative assessment as to the success of that particular run.

After 3 hours, 10 µL of sample was removed from each well into a fresh plate containing 190 µL ddH_2O in each well (1 in 20 dilution). After the assay finished, a further 5 µL of sample was removed from each well and diluted into 195 µL of ddH_2O (1 in 40 dilution). These two plates would be later used for flow cytometry analysis.

The three-hour and six-hour plates from the robot characterisation were analysed using the flow cytometer with all 200 µL of sample run through the cytometer. Forward scatter, side scatter and GFP fluorescence were measured (485 nm excitation, 520 nm emission) with the aim to asses 'leakiness' of the various promoters, by observing single cell stochastic events when samples were in supposedly 'non-induced' conditions (0 ng/µl aTc). An automatic micro-sampling system (AMS) was used to

improve the high-throughput nature of the assay and was coupled with flow cytometer to streamline analysis of the complete 96 well plates. Raw data was once again collected, with 50000 events (individual cells) counted for each sample. Data was initially analysed using Flowing Software.

4.2.7 Sequencing

To reveal the exact sequence of each promoter/operator hybrid, selected samples were miniprepped and sent for sequencing (*Source Bioscience*) using a reverse read primer that confirmed the presence of the correct sfGFP CDS. Sequencing reads also helped rule out the possibility of mutations within the 5'UTR promoter/operator regions that had the potential to alter results.

4.3 Results and Discussion

4.3.1 Qualitative Assessment of p101.Tet Library

Initial crude analysis of the various promoter libraries did indeed show a range of expression profiles as a direct consequence of induction with aTc. A preliminary assessment of these chosen variants is given, with relative fluorescent unit calculated for each sample by normalising FL units against OD values (***Appendices - Figure 11.6, Figure 11.7, Figure 11.8***). However, the crude nature of this initial assay did not appear to reveal any trends between the differing libraries in terms of expression strength or levels of repression. In fact, considering that as a first attempt, a maximal inducer concentration (1000 ng/ml aTc) was used, it is quite likely that repression would be completely bypassed these samples.

Numerous individual variants were chosen for extended characterisation and were sequenced to reveal a significant level of variance across the internal 11 bp variable region (Error! Reference source not found.**4.1**). This was promising as it provided some validation that the earlier library contained a high level of diversity.

Table 4.1 *- Identified sequence from the set of P101.Tet hybrid promoters taken forward for characterisation on the robot platform.*

Promoter name	Binding Region 1	11 bp Variable Region	Binding Region 2
P101.Tet-WT	**5'** ACTCTATCATTGATAGAGT	TATTTTACCAC	TCCCTATCAGTGATAGAGA **3'**
P101.Tet-HP1A	**5'** ACTCTATCATTGATAGAGT	CGGGTGGGCAG	TCCCTATCAGTGATAGAGA **3'**
P101.Tet-HP1B	**5'** ACTCTATCATTGATAGAGT	GGACACGTGGC	TCCCTATCAGTGATAGAGA **3'**
P101.Tet-HP2A	**5'** TCCCTATCAGTGATAGAGA	AAATGATAAGG	ACTCTATCATTGATAGAGT **3'**
P101.Tet-HP2B	**5'** TCCCTATCAGTGATAGAGA	GGAGCTATGGA	ACTCTATCATTGATAGAGT **3'**
P101.Tet-HP3A	**5'** TCCCTATCAGTGATAGAGA	CGTACCAAACA	TCCCTATCAGTGATAGAGA **3'**
P101.Tet-HP3B	**5'** TCCCTATCAGTGATAGAGA	ACCCGGAAGTC	TCCCTATCAGTGATAGAGA **3'**
P101.Tet-HP3C	**5'** TCCCTATCAGTGATAGAGA	CTTTCCAAGGA	TCCCTATCAGTGATAGAGA **3'**
P101.Tet-HP3D	**5'** TCCCTATCAGTGATAGAGA	TCCCTCCCCCC	TCCCTATCAGTGATAGAGA **3'**

4.3.2 Quantitative Analysis of Strong 'On' State in P101.Tet HP Library

As discussed, the automated liquid handling platform made accessible by Centre for Synthetic Biology (CSynBi) was suitable for high throughput data collection and did not require manual pipetting or movement of plates. Each run on the robot included 2 biological repeats for each variant at a range of aTc induction concentrations and thus over multiple runs averages could be taken for each data point. This included a measurement of cell density (600 nm) and fluorescence (510 nm) at 15-minute intervals across a 6-hour time frame.

In practice, the scale of raw data produced was fairly substantial and so a streamline approach to analysis was desirable. It should be made clear that data collection and analysis was very much a collaborative process. Fellow CSynBI members Dr Chris Hirst (built the original robotic platform), Catherine Ainsworth (designed analysis scripts) and Harry Rickerby (co-worker throughout) were hugely influential in helping progress the work.

Initially, all raw data was imported into an Excel template that quickly arranged data points, normalised values and produced growth curves for each sample well (Hirst, 2014). A suite of analysis scripts was already available thanks to parallel work (Ainsworth, Unpublished) invaluable when generating many of the following graphs and curve fittings.

Prior experimentation using the characterisation platform had identified examples of growth failure and can be linked back to general sensitivity of the samples to temperature changes and the external environment with issues of contamination and biofilm formation upon stress. Therefore, it was

reasoned that before any of the data sets were taken forward, that each sample should be assessed for growth fitness. Normal growth of the MG1665 *E.coli* strain within Neidhardt MOPS medium (from here referred to as just MOPS medium) had been well characterised, and could be fitted to a logistic growth. Any sample which did not show sigmoidal growth was excluded from the analysis, an example is provided (***Appendices - Figure 11.9***). In some instances, this phenomenon was perhaps a consequence of an extended period of logarithmic growth and never showing signs of slowing or reaching a lag/stationary phase. Nevertheless, this data was removed at if it did not satisfy this checkpoint.

Another major limitation of the characterisation in the *E.coli* host strain, was an inherent level of cell auto-fluorescence when grown in MOPS, which was chosen because it is a highly defined minimal medium (all nutrient concentrations are controlled). This improved the reproducibility of the results. However, it too results in a level of auto-fluorescence at the desired emission wavelength for detecting sfGFP output. To compound this issue, earlier experimental work had identified the fact that the level of auto-fluorescence did not show a linear increase or match an increasing cell density. The suggestion here is that because both cells and medium showed changeable levels of auto-fluorescence depending on cell density, particularly as the culture reaches logarithmic growth, the overall auto-fluorescence at any one time cannot be presumed.

Therefore, an MG1665 strain without any vector (negative control) was measured during data collection runs. The fact that the FL and OD for blank cells were recorded at the exact same time points as the test samples, meant that normalised measure for auto-fluorescence could be generated. In this case, 8 blank samples were recorded during every repeat, and plotted FL against OD600. A third order polynomial equation was used to fit a curve to this data (***Figure 4.3***). I believe this significant spike in auto-fluorescence at the onset of the assay can be attributed to the MOPS medium itself. Beyond a certain point (0.4 OD600 nm) there is a sharp increase as cells begin to saturate the medium (although not completely).

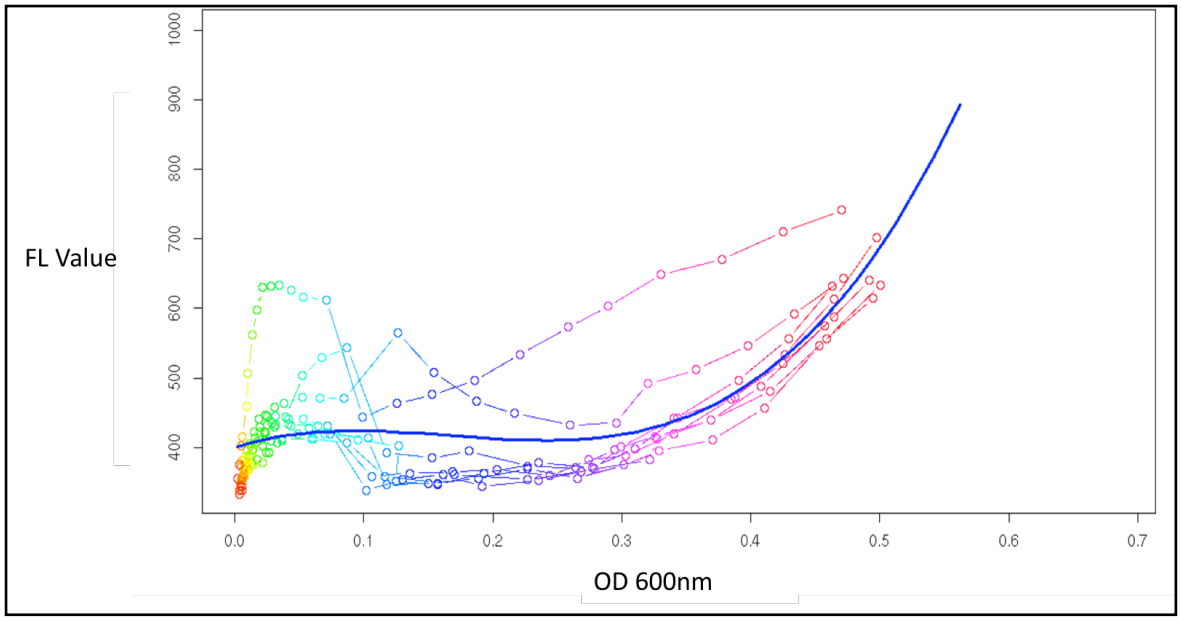


Figure 4.3 *– Auto fluorescence of blank E.coli strain MG1665 in Neinhardt MOPS media for samples grown over a 6 hour induction assay. This example showed a third order polynomial fit against 8 repeats from a single characterisation run*

Having normalised FL against OD to generate RFU values, and with a subsequent adjustment made to take into account auto-fluorescence at each time point, it would be possible to plot each sample promoter variant in terms of total relative expression strength, with some degree of accuracy. However, given the true purpose of this work is to generate a meaningful parameter which can be used to describe the actual level of expression achieved by each promoter under differing levels of induction, it was decided that the data would be fitted to a previously devised model for constitutive promoters, devised by Catherine Ainsworth (*unpublished DPhil submission*).

In simple terms, this model attempts to calculate protein synthesis rate by considering the growth of the individual sample in relation to an initial lag phase, maximal rate of growth, maximum population size maintained within the environmental conditions as well as the normal maturation and degradation rates of the sfGFP protein. Given that this mathematical model was beyond the scope of this project but built as part of the wider CSynBi / SynBis project, it made sense to use it to fit my normalised data.

The results were fairly promising for a number of the samples tested, particularly those which showed a higher level of expression at strong levels of induction. Strong curve fits were achieved which enable the gradient of the curve to be taken as the actual protein synthesis rate (***Figure 4.4***). When a high concentration of aTc (1000ng/ml) is used to induce the culture, P101.TetHP1A shows a particularly high level of expression and achieves a maximum sfGFP expression calculated as 18700 fluorescent absorbance units / hour i.e. the number of fluorescent emitting molecules produced per hour. The

fitted promoter model curve (represented by the green line), is clearly well positioned in relation to the normalised FL values at each time point and carries a high degree of confidence.

However, this model was less successful when a lower concentration of inducer was used, and indeed when lower FL values are achieved. As can be seen for curves associated with 100 ng/ml and 20 ng/ml data sets clearly FL values begin to level off at lower concentrations, most likely a consequence of the inducer become depleted within the system or perhaps degraded within the environment. Considering that the repressor is constitutively expressed, this is perhaps a dampened induction when extended out across the 6-hour growth period. This does not seem to be a consequence of the cell culture reaching stationary phase, as there is no discernible levelling off of the growth curve (represented in blue). Unfortunately, the promoter model curve does not consider a sudden reduction in fluorescence and therefore curve fitting is poor.

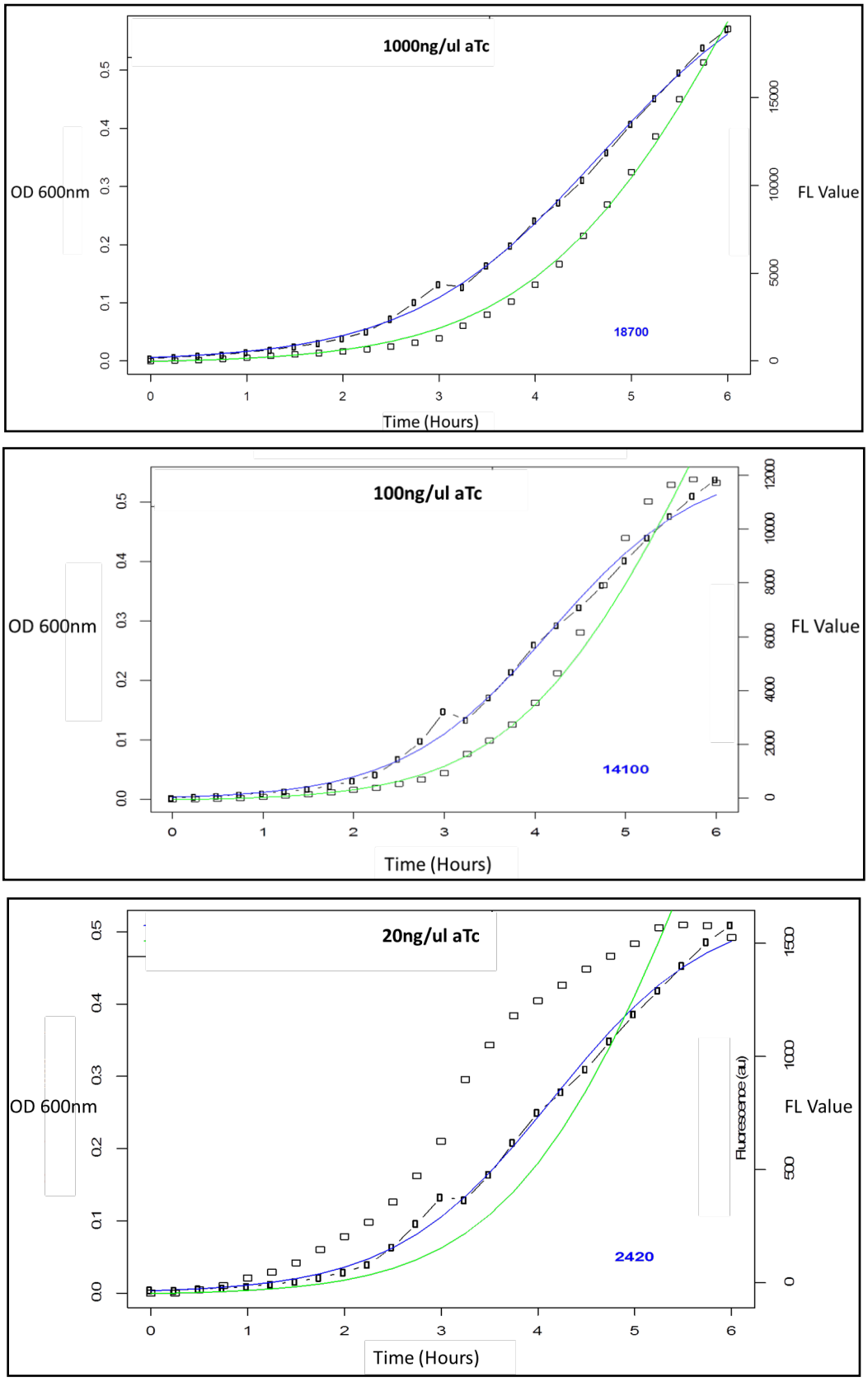


Figure 4.4 *– FL and OD curves for sample P101.TetHP1A across a 6-hour outgrowth at differing concentrations. Normalised OD data points (small circles) are used to fit a logistic curve (blue line) which provides an estimate of maximal growth rate. Normalised FL values (hollow squares) are used to fit a curve (green line) to determine protein synthesis as function of gradient defined by a promoter model.*

The exclusion of the final data points was considered, in order to eliminate the effect of any reduction in FL values, but it was not desirable as this could dramatically skew the results for many of the high expression variants.

The most suitable alternative was to use the promoter model to fit data to intermediate periods of time, in an attempt to more accurate describe protein synthesis rate at a particular point during outgrowth. Therefore, the promoter model was used to fit a curve to 4 data points (4 x 15-minute intervals) encompassing a 1-hour window (***Figure 4.5***).

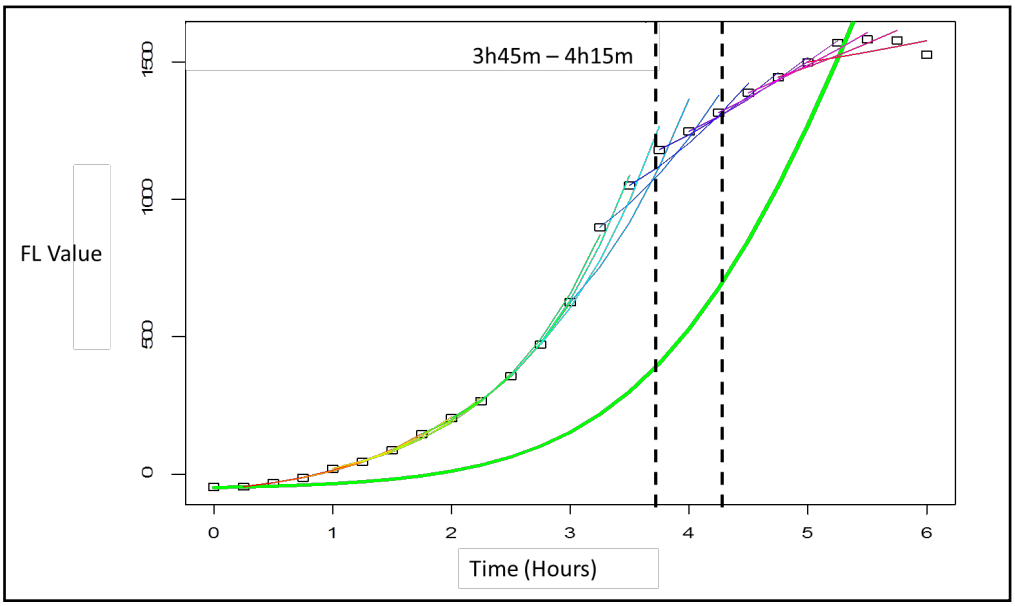


Figure 4.5 *– Normalised FL values (hollow squares) for p101.Tet.HP1A induced with 20ng/ml aTc and grown over a 6 hour time course. A promoter synthesis model for determining protein synthesis rate was used to fit a curve against all data without success (green trace). Subsequently, intermediate 1-hour windows (consisting of 4 data points) were used to fit multiple curves that better reflected the changing protein synthesis rate across the time course. A final window (3h45m - 4h15m) was chosen to avoid errors relating to the removal of a small volume of sample at 3 hours.*

Relative promoter units for each sample were taken as the gradient of the fitted curve between 3 hours 45 minutes and 4 hours 15 minutes during the 6-hour outgrowth. This was seen as the most suitable option as all samples were still within logarithmic growth at that point of the assay, and the effect of removing a small sample volume at 3-hour time point was minimised (note that removal of this sample was required for downstream flow cytometry analysis). Curve fits for all recorded samples during this window were used to generate a relative promoter unit (RPU) in accordance with the gradient of the curve.

A summary of RPU values is provided with calculated average mean, standard deviation and standard errors across 4 biological repeats of the assay (***Appendices - Table 11.18***). The majority of the p101.Tet HP library demonstrates strong levels of induction beyond 20 ng/ml of aTc inducer and in fact increasing the concentration of the inducer above 250 ng/ml seems to have very little effect on overall

protein synthesis rate (***Figure 4.6***). Interestingly, there is quite a distinct difference in the range of RPUs across the library. The constitutive J23101 control is not affected by the inducer but serves as a reference to normalise RPU against this highly characterised constitutive promoter, which is given an RPU of 1. It is clearly obvious that certain HP variants provide an RPU much greater than this when induced (above the critical lower bound of aTc concentration).

In an effort to understand which library offers the best solution (in relation to tight off, strong on), it is important to relate these results back to the original organisation of the TetO1 and TetO2 sequences. It does appear that the HP3 variants, which included TetO2 operator binding sequences in both positions has a lower RPU than the wildtype TetOp configuration (TetO1 followed by TetOp2). This was perhaps as anticipated given that TetOp2 is known to have a five times greater binding affinity for TetR.

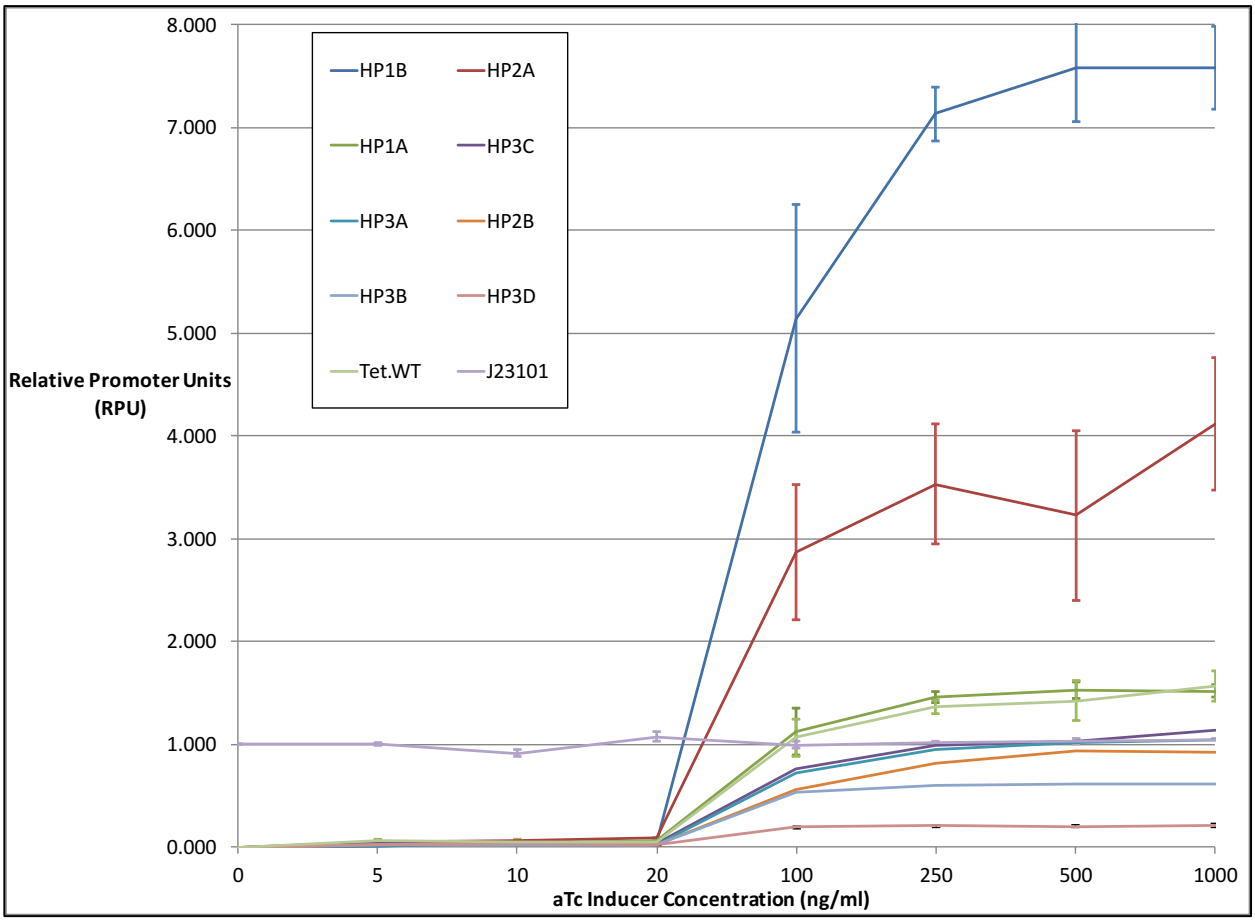


Figure 4.6 *– RPU for each P101.Tet library variant, plotted against increasing concentration of the aTc inducer (as described in Table 4.1). J23101 exists as a positive control for constitutive expression but remains unaffected by inducer concentration. Standard error bars are shown.*

It is important to point out that these measurements are related to maximal expression achieved upon induction and do not provided resolution of binding strength or leakiness of the promoter (this

will be detailed when assessing the flow cytometry data below). However, one hypothesis could be that the low RPU is a consequence of very strong TetOp2 / TetR binding, with improved cooperativity, to the extent that TetR/aTc binding is unable to disrupt repressor/operator association. Although there is option to increase the concentration of inducer to a higher level in an attempt to alleviate this affect, it does not seem feasible given that aTc demonstrates an antibiotic-like effect on the cells beyond 2000 ng/ml (*Gossard and Bujan, 1993*).

In an effort to obtain meaningful parameters for each HP variant, the average RPU data points were plotted as shown, to create does response curves (***Figure 4.6***). It was decided that a Hill function could be fit to this data given the known level of cooperativity between TetR (protein macromolecule) and aTc (ligand) binding. In order to ensure a strong fit to the data, the non-linear least square method was used, given that the dose response does not follow a linear equation i.e. beyond a certain concentration of inducer, there is no further increase in RPU. The least squares method attempts to minimises the sum of the squares of the differences, between observed responses and the predicted logistic Hill function.

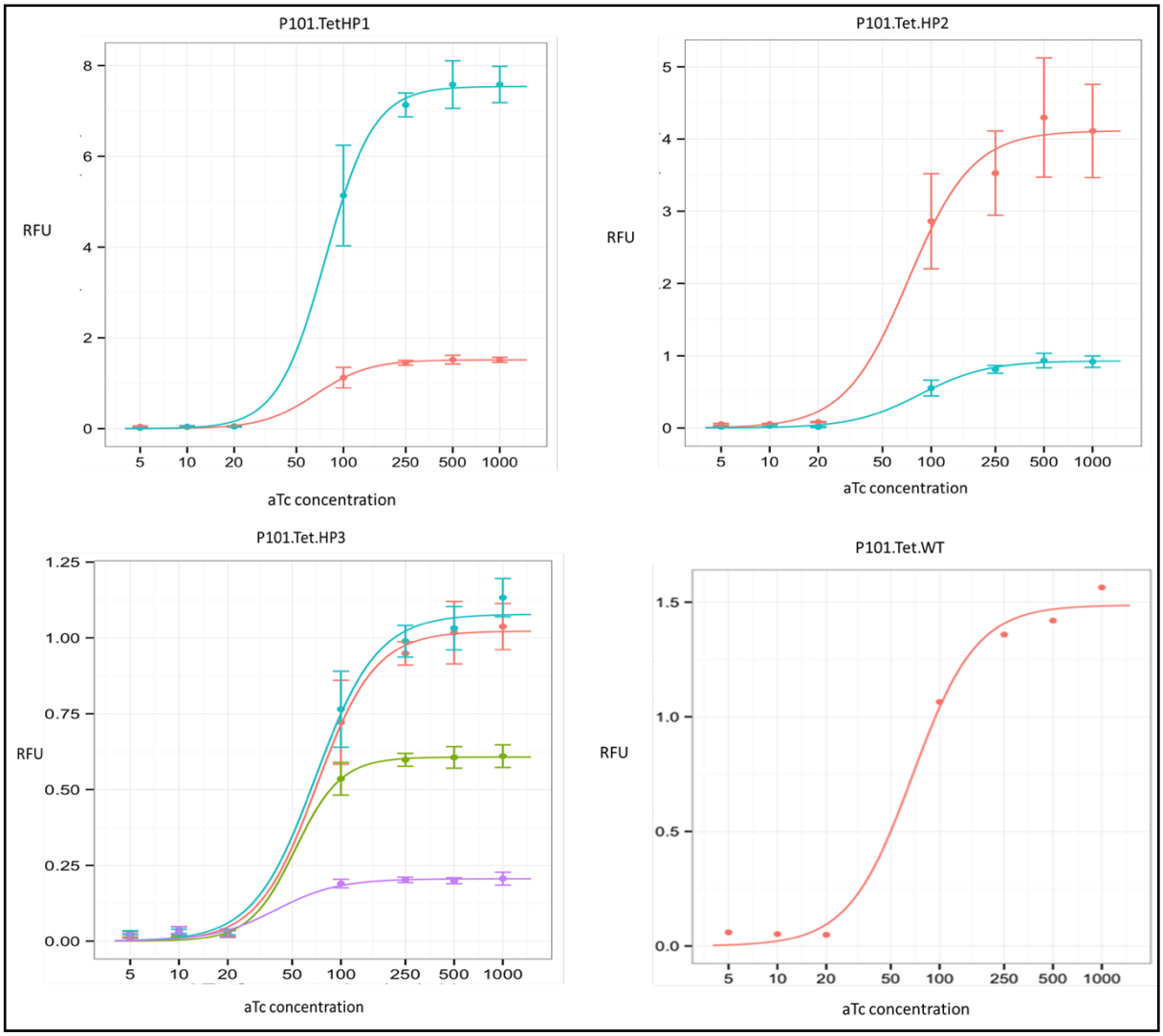


Figure 4.7 *– Dose response curves for the complete set of P101.Tet.HP hybrid promoters (as described in Table 4.1). Curves were generated using a logistic Hill function and fit in accordance to the least squares methods (recreated from Rickerby, 2014). Standard error bars are shown.*

Biologically relevant parameters are able to be generated from these dose response curves given that the Hill function provides a measure for Vmax (maximal transcription/protein synthesis rate) generated by the P101.Tet promoter, upon the basis of its interaction with the number (n) of aTc (activator) molecules within the system (activator), whilst considering the Kd (disassociation constant) of aTc.

A more accurate understanding of the complete genetic circuit proposed within this work highlights the fact that the system will be in a perpetual state of 'on' until the point at which a viable LasR mutant can activated repression of the system by expressing TetR (***Figure 4.1***). Therefore, increasing the level of aTc (as has been shown in this assay) will not directly influence the activity of system. However, given that TetR expression has been controlled by a known promoter (J23101) an

estimation of the number of free TetR molecules within the cell at any one time can be made based on the protein synthesis rate of this promoter. Previous work conducted by colleagues at CSynBi, characterised the expression strength of this promoter, taking into consideration the known binding/dissociation constant of TetR/aTc (*Kamionka et al, 2004*). I made use of this plug and play equation outlined in the PhD thesis submission by Dr Chris Hirst to generate values for the steady state levels of TetR.aTc at the given concentrations of aTc (*Hirst, 2014*). It is then possible to determine the number of free TetR levels within the cell. The final parameters are shown below (***Table 4.2***).

Table 4.2 *– Parameters for the complete P101.Tet.HP hybrid promoter library variants tested within this work (recreated from Rickerby, 2014).*

Promoter	Maximum transcription rate *(Vmax)* (RPUs)	Dissociation Constant (*Kd*) (TetR molecules cell^{-1})	Hill coefficient (*n*)
P101.Tet.WT	1.49	1833	2.2
P101.Tet.HP1A	1.52	1838	2.6
P101.Tet.HP1B	7.54	1798	2.8
P101.Tet.HP2A	4.11	1815	2.2
P101.Tet.HP2B	0.93	1765	2.1
P101.Tet.HP3A	1.02	1822	2.4
P101.Tet.HP3B	0.61	1894	3.1
P101.Tet.HP3C	1.08	1827	2.3
P101.Tet.HP3D	0.21	1944	2.3

4.3.3 Qualitative Assessment of Tight 'Off' State in P101.Tet HP Library

Flow cytometry permits single cell fluorescence to be observed and its therefore ideal for measuring very small levels of leaky expression from the P101.Tet HP library. The theory was to be able to identify weakly repressed hybrid promoter variants by observing a population of GFP fluorescing cells when no inducer was added to the culture. Therefore, 30 µl of the culture was removed at 3 hours and again at 6 hours.

Forward scatter is often used to measure the overall size of the cell, and side scatter relates to the refraction of light based in the internal complexities within the cell. During analysis of the HP library, side scatter was shown to have a better correlation with GFP expression within individual cells of a control sample, ensuring that the gated population were actually cells that were mature enough to be producing GFP.

FL was plotted against side scatter to create dot plots for all samples and analysed using a positive control to define the gate. Corresponding histograms showed a Gaussian distribution meaning that the mean FL values could be interpreted as a marker for overall fluorescence with the sample. Mean averages for all samples were calculated based on a minimum of three biological repeats of this

assay and compared against the TetOp.WT i.e. the true wild type sequences as found within the Tet transposon (positive control) and a blank MG1665 *E.coli* strain (negative control).

From this data set, it was evident that all versions of the P101.TetHP library are strongly repressed and in fact even more strongly than the TetOp.WT variant. No clear difference between subset HP libraries was apparent, likely due to a limited number or repeats within a sizeable standard error about the mean, however it was clear that the hybrid promoters, particular those variants containing the TetOp2 in the first position, achieve strongest repression.

Although this data does not provide a quantitative assessment of repression strength it was used to inform later iterations of the genetic circuit model and decide which functional promoter part could be should be used in future test constructs (***Chapter 5, 6 and 7***).

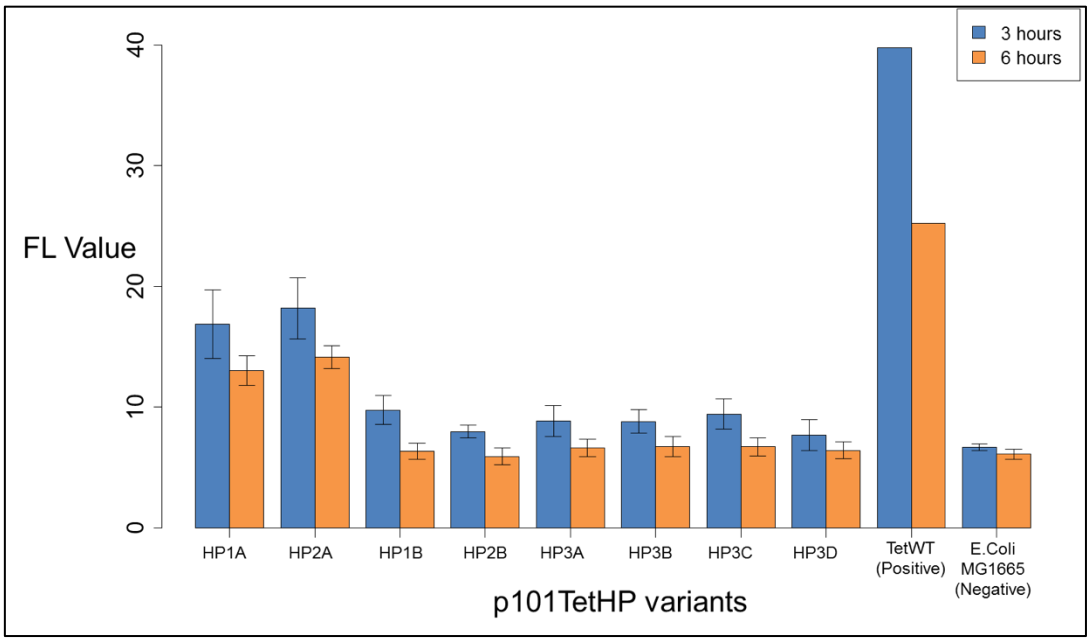


Figure 4.8 *– Mean levels of expression identified in un-induced in variants of P101Tet hybrid promoter library using flow cytometry (as described in Table 4.1). Note that the improvement in repression strength in comparison to the wildtype Tet operon, in many instances this is comparable to the negative control, demonstrated a strong 'off' state. Standard error bars are shown.*

4.4 Conclusions

The general aim to build and test an inducible/repressible promoter library with the aid of an automated characterisation platform was achieved. It should be said that the amount and quality of the data obtained using the robot was excellent. By way of example, 8 biological repeats, across 8 induction concentrations, were measured at 24-time points within a single run. Although there were a few examples of the failed runs (often due to poor growth of the cells during the overnight preparation) in general the adapted protocol was fairly reliable. The consistency and reproducibility of results is reflected by relatively small error bars for many of the test samples.

Nevertheless, there were some very important checkpoints need to eliminate inaccuracies within the final data set. Firstly, a comparison against normal sigmoidal growth was used, given that previous experiments had referenced sporadic contamination, biofilm formation, poor mixing and/or air bubbles within sample wells and extended lag phase growth (likely due to a temperature shock when inoculating into fresh medium. I believe this system would benefit from regular sterilisation, calibration and a more global control of temperature. The reproducibility of the data would likely benefit from the robot being held in a contained workspace environment with the ability to maintain a fixed temperature at all footprint positions. Subsequent work within CSynBI to transfer this characterisation protocol onto a cutting edge robotic platform, provides greater temperature control. It would be extremely interesting to see how comparable and reproducible these results would be on the new system.

A further issue in the experimental protocol was a sudden decrease in expression levels during the final few time points, particular in weaker promoters at low induction concentrations. The working solution was to fit use small intermediate windows at a selected time point, which was successful. In future work, all effort should be made to minimise aTc degradation. This inducer should be kept in blackened tubes to avoid long term exposure to light and stored at -20°C. However, I believe the most likely reason for the sudden reduction in gene expression is a direct consequence of cell saturation in high density cultures. *E.coli* cells have the potential to actively uptake aTc from the surrounding environment as well as passively uptake through lipid-mediated or porin-mediated diffusion channels (*Delcour, 2009*).

This work might be extended to make a comparison between variants of a single subset library given that there seems to be a pronounced difference between expression and repression strengths of those hybrid promoters (HP's) which contain the same operator sequence context. For example, P101Tet.HP1A-B are calculated as having a remarkably different RPU (1.52 – 7.54 respectively). The fact that these variants show no difference in their operator binding sequences suggests that the 11bp variable region between both operators must have an influence upon the tightness of the repressor binding and overall strength of expression that can be achieved.

An initial hypothesis might point to the differing GC content of the intermediate regions in both HP1A (5'-CGGGTGGGCAG-3' = 91% GC) and HP1B (5'- GGACACGTGGC-3' = 72% GC). This will undoubtedly affect the bendability and curvature of B-DNA nucleotide sequences (*Goodsell and Dickerson, 1994., Gabrielian and Pongor, 1996*). Given that cooperativity plays a significant role in the binding of the TetR at adjacent TetOp sequences, binding of the repressor to the first operator will cause conformational changes downstream, encouraging a relaxing or unwinding, and altering the conformity of the second operator site, making it more or less favourable to binding of the second TetR homodimer. Whilst assumptions could be made by designing fixed 11bp regions, a true interaction

module would require a direct binding assay between purified TetR protein and the target promoter/operator sequence within a dsDNA template, way beyond the scope of this project.

It should be noted within the wildtype TetOp, the TetO1 sequence actually overlaps the Tet promoter. Therefore, there are two potential explanations for the leaky characteristic of the wildtype compared to the variants of the HP library. In the first instance, RNA polymerase is never bound to its cognate promoter because TetR is already directed to the operator sequence and completely obscures the promoter site. Alternatively, the operator sequence may be more easily bypassed by RNA polymerase because it is only loosely directed to the promoter site, thus it can begin transcription downstream. As some stochastic expression has been seen in the wildtype, it seems quite possible that the second option holds true, and positioning of the 'weaker' operator TetO1 at this locus is likely to increase the chances of stochastic expression.

Ultimately, this set of hybrid promoters shows examples of strong repression and high expression and would be favourable for the wider circuit. The parameters generated can be used to inform the deterministic model of the genetic circuit through integration as ordinary differential equations (ODE's) (***Chapter 5***).

5 Circuit modelling

As previously stated, a highly characterised set of genetic parts provides the opportunity to simulate a model for the biological circuit with a high degree of confidence. Ideally, parameters for every interaction described within the circuit design should be established in advance and should consider the biological context of the intended host chassis. The later point is particularly relevant given the stochastic nature of biological systems, as the engineered device framed within a cellular environment will inevitably interact with the host.

All things considered, the initial computational design is likely to provide a far more accurate simulation of events. Undoubtedly the integration of an entirely novel system such as this will require numerous iterative rounds of design and modelling before it truly works as intended. Establishing a strong model should simplify and speed up the redesign process and guide changes based on discrepancies.

5.1 Justification and Proposal

Throughout this work, numerous models of the complete genetic circuit have been generated using the software Cell Designer (*Funahashi et al, 2008*). These models were built from a number of ordinary differential equations (ODE's), which are described by a parameterised kinetic function for each event across a certain time period, to create a deterministic response to input/output variables.

Given my relative lack of experience using system biology tools, this software provided a useful design interface which is fairly intuitive and enables the model to be extended by simply adding new parameters such as those generated for both p101Tet Hybrid promoter and pLas sets.

The complete model can also be exported in 'Systems Biology Markup Language' for transfer to more sophisticated software packages that may be require for complex stochastic modelling such as COPASI (*Hoops et al, 2006*).

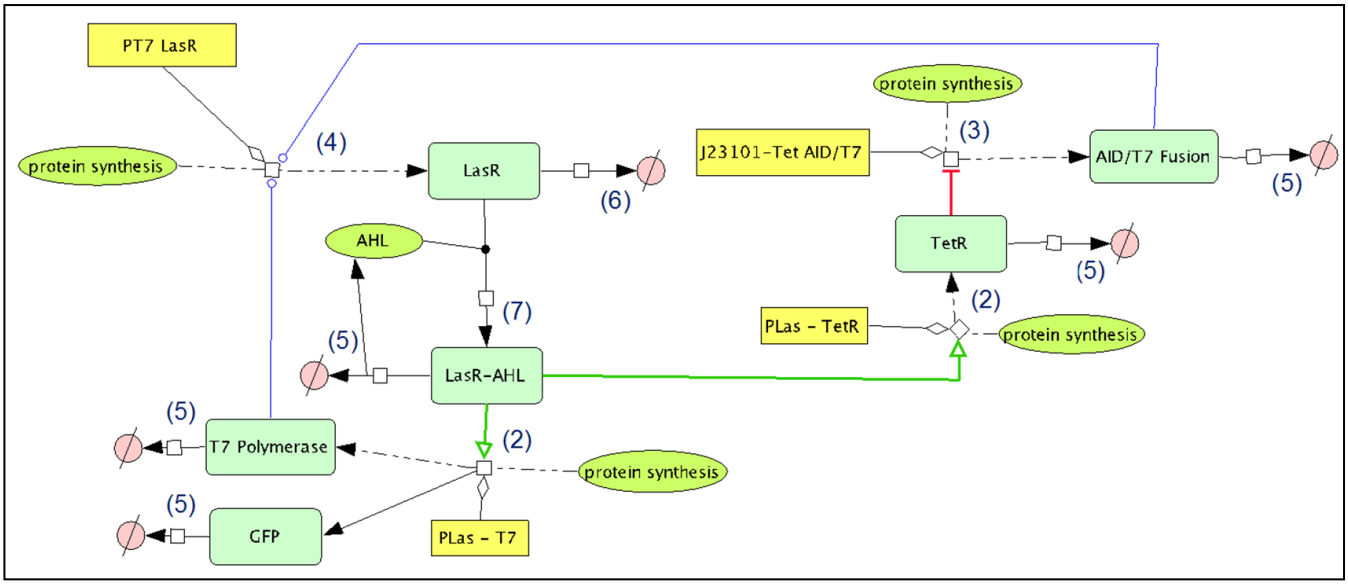


Figure 5.1 *– Cell Designer manifestation of the complete circuit. Numbers correlate to individual kinetic laws applied to the deterministic model. Light green rounded boxes reference the protein or protein complex intermediates. Yellow boxes reference expression constructs. Pink circles refer to unknown protein degradation*

5.2 Methodology

The majority of parameters were either retrieved from the literature or from unpublished data collected as part of the on-going efforts within CSynBi.

Table 5.1 *– Summary of the parameters included in the latest iteration of the Cell Designer model*

Parameter	Value	Units	Source	Notes
E.coli Cell Volume	3.0	Femtolitres	Volkmer and Heinemann, 2011	Based on MG1665 strain grown in LB media
TetR : P_{tet101} *Vmax*	77	Protein molecules $cell^{-1}$ $minute^{-1}$	CSynBI (Unpublished)	Based on J23101 synthesis rate
TetR : P_{tet101} K_d	0.32	Molecule number $cell^{-1}$	Kamionka et al, 2004	Based on wildtype P_{TetR}
P_{tet101} Hill Coefficient	2.0	-	Orth et al, 2000	Estimate based on dimerization of TetR with two TetO binding sites
LasR : P_{lasR} *Vmax*	551	Protein molecules min^{-1} $cell^{-1}$	Chappell et al, 2013	Based on pLasR3 - estimated from supplementary data
P_{lasR} K_d	18.06	Molecule number $cell^{-1}$	Chappell et al, 2013	Based on pLasR3
P_{lasR} Hill Coefficient	1.2	-	Chappell et al, 2013	Based on pLasR3
P_{T7} *Vmax*	98	Protein molecules min^{-1} $cell^{-1}$	Garcia and Molineux, 1995	Very rough estimate based on J23101 PoPS measurement (Kelly et al, 2009)
P_{T7} *Kd*	99.33	Polymerases $cell^{-1}$	Jia et al, 1996	Based on the use of φ10 T7 promoter
Protein degradation* rate constant	0.035	$Minutes^{-1}$	Sezonov et al 2007	Based on doubling time of MG1665 *E.coli* in LB media. Assumed to exceed degradation rate (Maurizi, 1992)
AHL-LasR association rate constant	1000	$Minutes^{-1}$	Schuster et al 2004	Estimate based on strong irreversible binding of AHL to LasR

5.3 Results and Discussion

Cell designer models are output as traces showing the predict number of molecules for all protein or small molecule intermediates cross the chosen time course. Models were generated using the parameters described in ***Table 5.1***. The model assumes that 700 minutes outgrowth of the target MG1665 strain, which is inclusive of all components previously described in the mutator and output modules.

5.3.1 Preliminary Modelling Efforts

Initially, the preliminary simulations included the addition of 1000 molecules of AHL at the 330 minutes and demonstrated that the modelled circuit worked as expected. The fictitious addition of AHL here is used to demonstrate successful binding to a compatible LasR protein, which leads to subsequent expression of the T7RNApol and repression of ATF mutator. However, initial P101Tet parameters were only based on PJ23101 and *in vitro* TetR binding which are likely to be quite far removed from the reality of association within the true circuit, given that they do not account for DNA secondary structure formation (***Figure 5.2***).

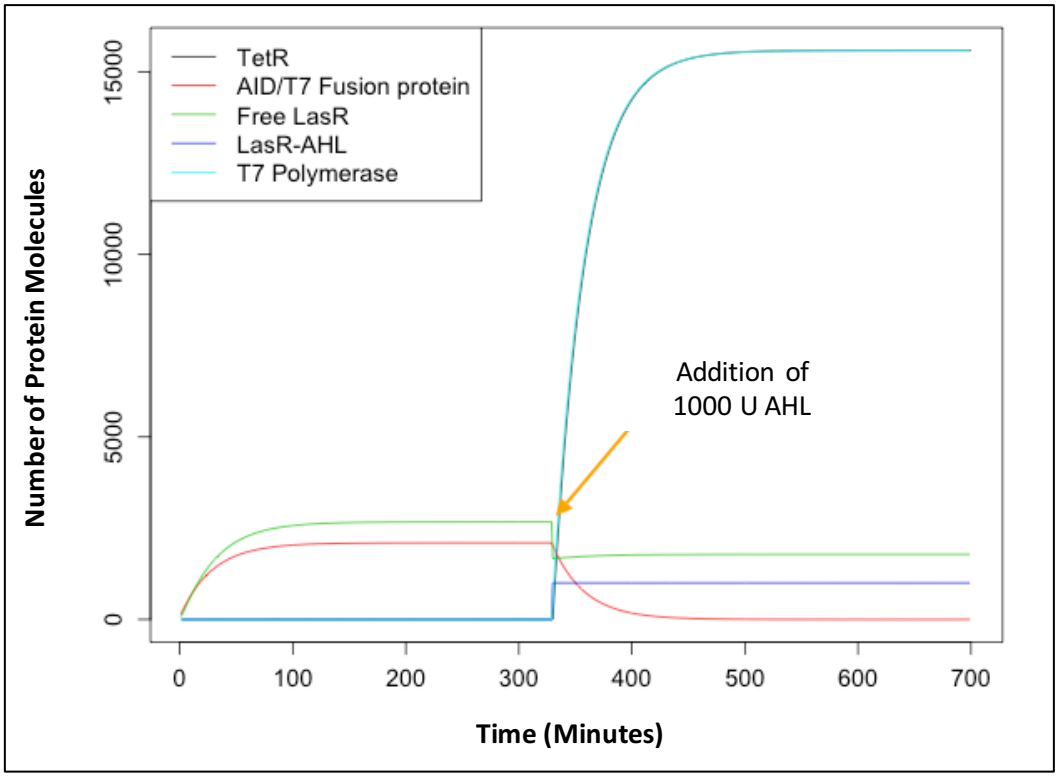


Figure 5.2 *– Initial Cell Designer simulation excluding true P101TetHP hybrid promote parameters. The addition of 1000 AHL molecules at the midpoint is used to replicate successful AHl/LasR binding.*

This initial test highlighted the need for a P101Tet.HP promoter with a high Hill coefficient based on the predicted dissociation constant, in order to achieve the desirable repression of the mutator (***Figure 5.3***)

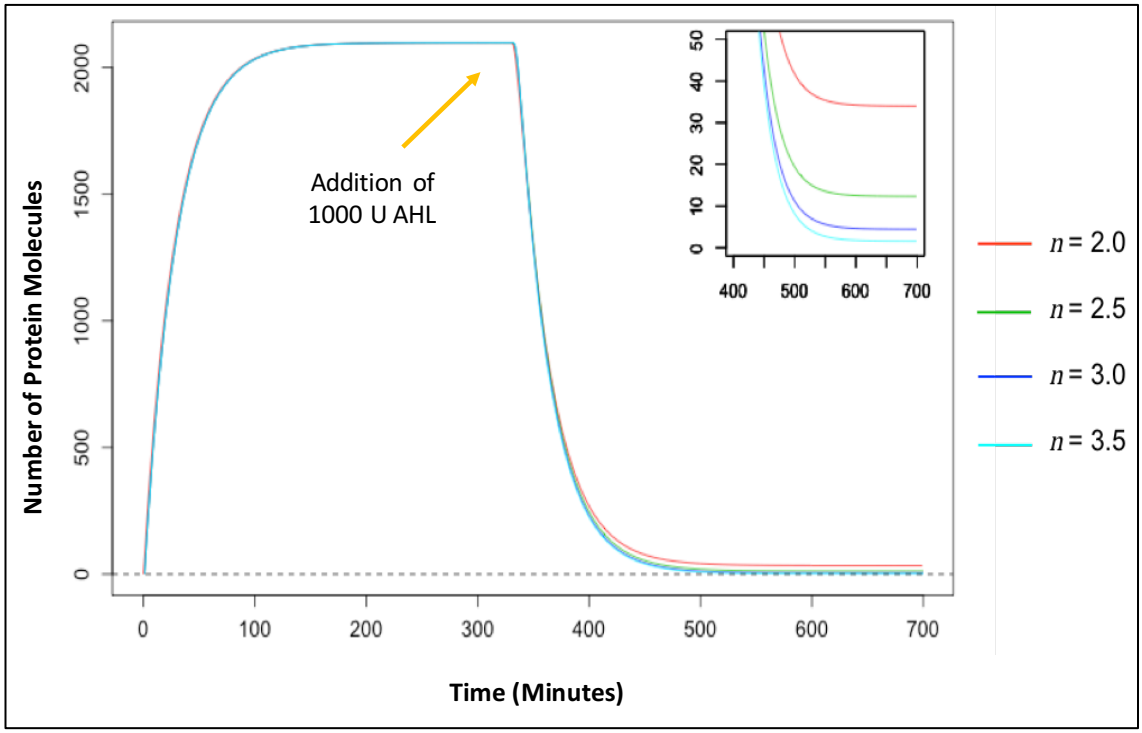


Figure 5.3 *– Changeable levels of ATF protein as a direct result of different Hill coefficients (n) of P101TetHP. Once again, the addition of 1000 molecules of AHL is included at 330 minutes.*

5.3.2 Inclusion of p101Tet HP Parameters

Characterisation of a set of novel P101TetHP hybrid promoters was carried out in ***Chapter 4.*** Parameters were generated for each variant, with dissociation constants now describing free TetR molecules per cell. It should be noted that changes to P101TetHP characteristics will only affect the repression of the ATF mutator and have no bearing on the initial number of molecules within the system. Once again, AHL addition is used to simulate the emergence of a desirable LasR variant during active mutation. All P101TetHP parameters are listed in ***Table 4.1***.

and the most up to date simulation of the model demonstrates very changeable levels of ATF expression and repression. Therefore, it would follow that there would be varying levels of mutation, if ATF concentration is directly linked to frequency of mutation.

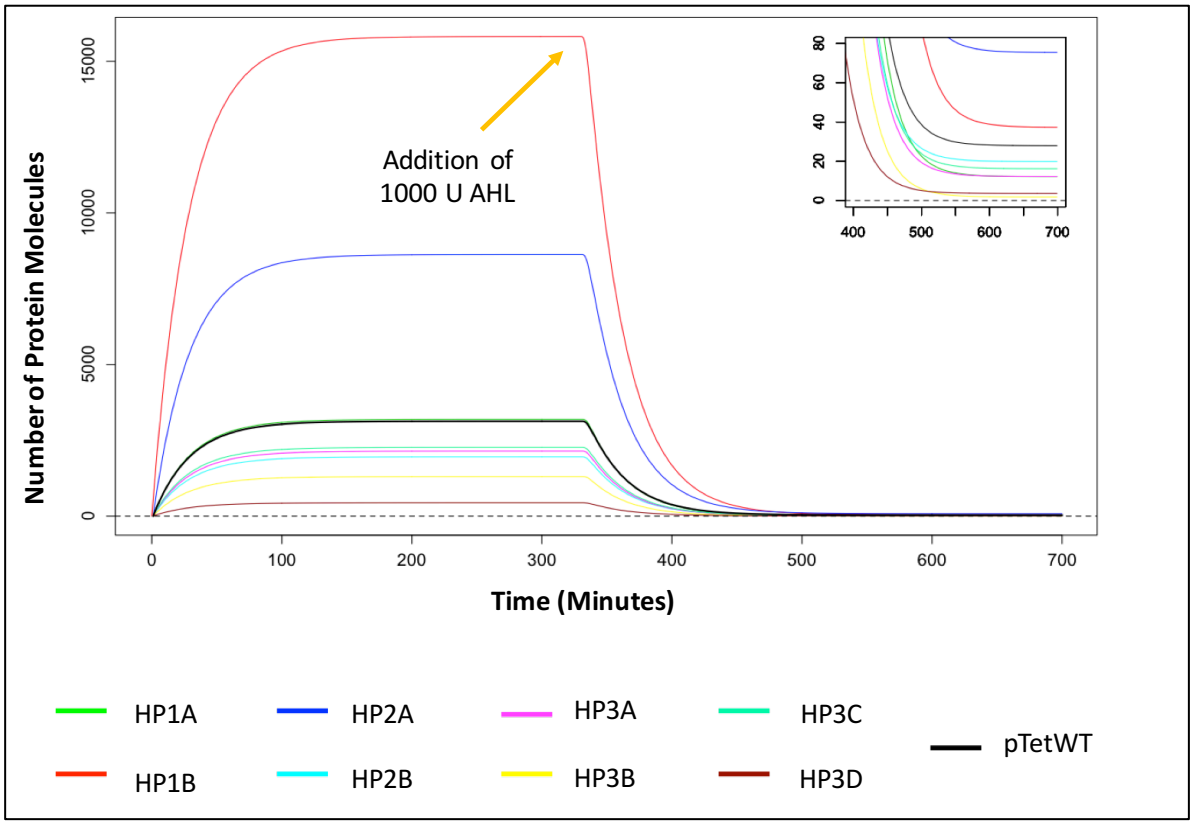


Figure 5.4 *– Model simulation including parameters of the p101TetHP hybrid promoter set (built and characterised in Chapter 4) showing changing levels of the ATF mutator protein. Once again, the addition of 1000 molecules of AHL is included at 330 minute time point*

5.4 Conclusions

5.4.1 Strength and Accuracy of Model

There were some very early benefits from constructing a rough deterministic model of the complete circuit, which were not identified from simple diagrammatic versions. Early designs had included a AND gate, based upon the activation of a sigma 54 hrpL promoter orthogonal to *E.coli* (*Wang et al, 2011*). The idea was to use this logic gate to help reduce non-specific activation of the output module, which could potential cause the mutator system to be repressed far too early (*data not shown*). In my opinion, true AND functionality is questionable, given that both inputs would be a direct response to LasR / AHL binding. Regardless, this added level of circuit complexity was not required, as the circuit would require a significant level of TetR to build up in order for repression of the mutator.

The real success of this work has been the creation of a set of hybrid promoters that provides a level of tuning, which is clearly evident within the model simulations. For example, a final decision about the most suitable P101TetHP will likely be based on the desired level of mutation and whether small differences in repression of ATF expression are likely to result in the desired end point from being obscured.

It should be noted that this basic simulation makes many assumptions, given that many of the parameters are broad estimates pulled from numerous different sources. Many of the assays used generate parameters which are highly specific and not particularly universal (such as *in vitro* expression and target binding), which underlines the importance of a standard approach to part characterisation carried out within the host organism.

Sensitivity analysis was also used to explore how each parameter directly affects the expression of ATF as it is considered the most essential protein within the system. ATF is the mutagenic protagonist that drives the system towards completion i.e. eventually causing it to reach an 'off' state. This approach is useful, given the number of uncertainties within the model, as it seeks to determine which variable will have the biggest influence on a selected species (in this case ATF expression). Cell designer software simplifies the process, by using the parameters entered within the model to highlight the consequence of changing any parameter (which may also itself be an assumed parameter). In this instance, the most important factor effecting the levels of the ATF protein fusion protein is the protein degradation rate which is followed by the strength of PTet101 promoter (data not shown). This may influence decisions and direct the focus of later work to characterise these variables to improve our understanding.

5.4.2 Requirement for Dual Selection in Directed Evolution

Although not a direct comment on the validity of the model, an important consideration moving forward is the need for the selection system to be able to tolerate and make allowances for periods of both positive and negative selection. If we consider the evolution of a LasR protein in our system, in which we aim to evolve an effective response to a novel AHL molecules to produce transcriptional output. Positive selection will more than likely produce a LasR mutant that is capable of forming a complex with the chosen AHL derivative, and it should be possible to increase specificity and affinity of the binding between substrate and analogue by lowering the concentration of AHL. Nevertheless, there is no form of negative selection that can apply a preventative pressure on such a wide range of mutants i.e. those mutants which can complex with a variety of AHL analogues. It would be as equally likely for non-specific mutants to evolve out from the original starting LasR protein. Furthermore, the similar nature of separate species-specific quorum sensing genes suggests that cross reactivity is likely. There is an option to pre-engineer a non- functional LasR target gene, including mutations within the essential domains of the LasR protein, but we might expect a reversion of these mutations and a predominance of the wildtype gene after selection.

In regard to the development of a dual positive-negative selection system, previous work to generate a LuxR variant which could respond to a novel AHL homolog was only able to generate a broad-range variant i.e. it could respond to a range of AHL, including the new and original homologs

(*Collins et al, 2005*). Only when a dual selection was employed did the unwanted response 'evolve away', leaving a highly specific mutant (*Collins et al, 2006*).

5.4.3 Directed Evolution of Complete Gene Circuits

A recent publication has identified the numerous complexities of attempting to build multi-part biological circuits in a robust and predictable manner (*Dougherty and Arnold, 2009*). The difficulty of implementing pathways or devices within living cells is confounded by huge unknowns which are a consequence of stochastic background events caused by intracellular protein or pathway cross reactivity. It is inevitable that a final model which has been assembled by combining models from a small subset of individually characterised components will have limited accuracy. A recent example showed that the combining diversified promoter parts with in depth modelling of feed forward networks could produce a predictable gene circuit in a eukaryotic host, without the need for fine tuning *post hoc*. (*Ellis et al, 2009*).

Accurately predicting how the mutator system will work within any host cellular environment is a daunting task, and whilst the mature model may provide validation and provide the best chance of the circuit working as planned, it is unlikely that our first iteration will work perfectly.

An interesting solution could be to employ a directed evolution approach for the purpose of improving the components (*Farinas et al., 2001*), devices, pathways (*Schmiddt-Dannert et al, 2000., Crameri et al, 1997*), and in fact the complete AIDT7 mutator circuit devised here. In one such example, the author explains that 'circuit engineering in cells can tolerate designs that are bad in some features because we can force the system to evolve' and thus it 'helps compensate for the designer's profound ignorance of sequence-function relationships when programming cell behaviour' *(Yokobayashi et al, 2002)*. Here a gene circuit which functioned as an inverter, was constructed from a library of the genetic elements was generated by combining ePCR and random digest products in a ligation mixture and subjected to further rounds of amplification to create second generation libraries. In actual fact, the circuit did not work on initial testing due to low levels of leaky expression. However, the directed evolution approach found a solution which was revealed as a change in repressor-operator binding affinity. Interestingly, this circuit was previously 'debugged' using a rational approach employing modelling and computer aided simulations, but this was considerably labour intensive (*Weis, 2000., Weiss and Basu, 2002*). Future work in this project may indeed require reiterations by simplistic promoter or RBS tuning.

6 Improving the AID-T7 Mutator Device

6.1 Introduction

This section is a natural progression from the early developmental work outlined in ***Chapter 2*** and has been enabled by foundational method development in ***Chapter 3.***

The results from initial work to validate the mutator device were very encouraging and suggested that the AID T7 fusion (ATF) variant can produce mutations that are both targeted and restricted to a specific genomic target gene, ensuring that number of non-specific background mutations remained negligible and that the host strain remains viable. The fact the mutations occur across the full length of the target gene of interest is also promising.

Having extensively sequenced a large number of clones after mutation, I identified two clear effects of expressing the ATF protein; i) deamination events occur slightly more frequently on the non-template (coding) strand, consistent with previous observations in the literature, ii) the frequency of mutations seems to decline as the length of the transcript increases. Both of these behaviours presumably reflect the nature of the relationship and interaction between T7 RNA polymerase (T7RNApol) and its target DNA sequence.

Firstly, ssDNA is more accessible to AID on the non-template strand given that the larger T7RNApol protein directly interacts with the template strand. Furthermore, immediately following transcription an ssDNA/mRNA hybrid structure is likely to be formed. Therefore, it seems likely that there is only a fleeting period in which AID could interact with the template strand.

Secondly, T7RNApol is known to undergo an abortive phase of transcription in which only short mRNA transcripts are produced. In this instance, T7RNApol does not transcribe the complete length of the gene, dropping off the template strand very quickly. Given that the AID monomer is fused to its N terminus, there is very little opportunity for AID to interact anywhere other than the 5' end of the coding strand. We might therefore expect the number of mutation events to decrease as an inverse relation to the overall gene length and distance from the transcriptional start point.

Unfortunately, in its current form, the AT7 protein generates infrequent mutations across a limited spectrum and does not meet the prerequisite for sampling complete sequence space. At a nucleotide level this means the system cannot introduce all transition and transversion mutations and only a small number of amino acid residue changes are possible. This is sub-optimal for a directed evolution system that seeks to sample complete genetic diversity.

Given the earlier positive results, there seemed enough promise in the mutator device to continue its development and consider ways to optimise control and regulation and attempt to increase the frequency, distribution and variance of mutations that can achieved.

6.2 Tuning expression of the AID-T7 mutator

6.2.1 Background

In this section I refer to my efforts to increase the frequency of mutations by controlling the level of expression of the ATF protein using a range of inducible promoters (transcriptional control) and RBS strength (translation control) with the aim of creating a tuneable level of targeted mutation with tighter control and faster induction response.

Whilst the ATF protein remained unchanged, constructs were built in accordance to the BASIC standard (***Chapter 3***) and although there were minor alterations to non-coding sequences within each construct, I tried to reduce the chance that context dependency would have an effect on results. All variants were screened for targeted mutational activity, toxicity and burden using the protocols and analysis outlined in ***Chapter 2***. Therefore, the data presented here should be directly comparable to previous data sets.

6.2.2 Justification and Proposal

Previously studies have shown that AID mutational activity is directly linked to transcription. This is likely a consequence of exposed ssDNA at the transcription bubble, with AID known to act only on ssDNA targets. It follows therefore, that a higher transcription rate exposes the ssDNA target more frequently and for longer periods, and thus AID has the opportunity to bind ssDNA and introduce mutations more readily.

In order to specifically target AID, the monomer was fused to the N-terminus of the viral T7RNApol. Expression of this viral protein in *E.coli* has been widely explored and its activity is distinct to host transcription factors. When compared to endogenous *E.coli* RNA polymerases, it clearly results in a very high number of mRNA transcripts, yet the conversion of mRNA into polypeptide during translation is still limited, which is likely a consequence of host ribosomal machinery reaching maximal capacity. Unfortunately, there is no way to separate these transcription and translation events *in vivo*, but one might suggest that regardless of final protein yield, a higher transcription rate will increase the frequency of AID mutations.

Previous work to characterise expression strength within our laboratory (*data not shown*) had provided anecdotal evidence that overexpression of the T7RNApol protein itself may actually build to a level which is toxic to the host strain and have significant consequences on cell fitness and viability. It was suggested that expression of the T7RNApol should be driven from a low copy plasmid in order to limit this effect, and therefore comparisons were made between high and low copy variants. It was also decided to investigate whether the addition of a negatively charged LVA degradation tag (which marks the protein for protease degradation and removal by the ATP dependent ClpXP system in *E.coli*) could

help recover this toxicity effect (*Gottesman et al, 1998,. Baker and Sauer, 2012*) by encouraging the host cell to remove residual protein.

Another important consideration was to maintain tight regulation over expression of the ATF mutator protein in order to control active mutagenic state. In relation to the complete circuit, any potential 'leaky' expression could have a pronounced effect on the switching of the circuit and obscure the link between interesting phenotype with genotypic variant. Extensive work to generate a new set of hybrid promoters/operator sequences, inducible by Tetracycline, has been described in ***Chapter 4***. It was decided to use this new and improved set of promoters to control ATF expression given their tightened off states.

6.2.3 Methodology

Plasmid Construction

A number of constructs were built using the BASIC assembly standard, with the parts described verified accordingly (***Appendices - Table 11.19, Table 11.20*** and ***Figure 11.10***). All constructs conform to the general plasmid map shown in ***Figure 6.1***. All test constructs are described in ***Table 6.1***.

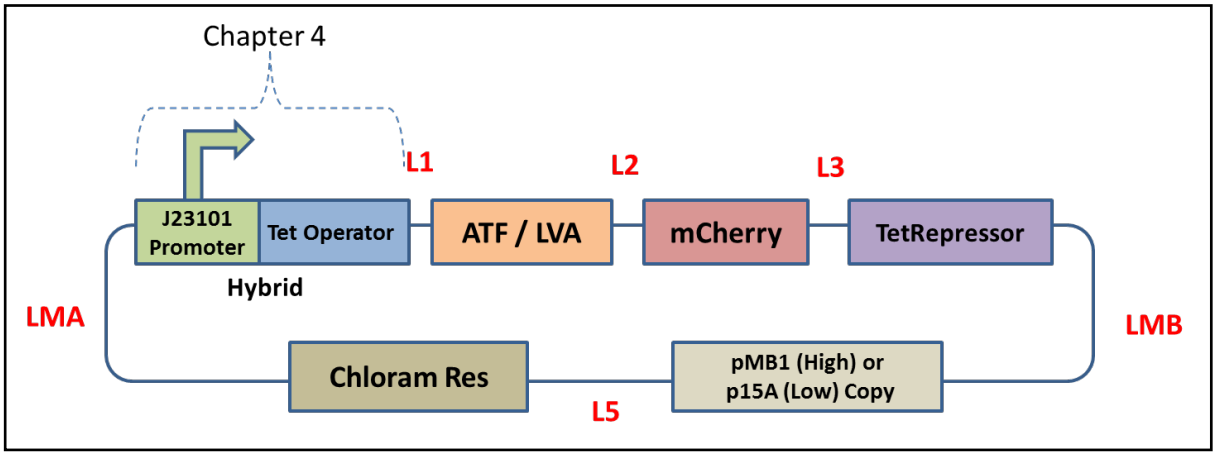


Figure 6.1 *– General arrangement of constructs use to test the toxicity of overexpression of the mutator protein*

Table 6.1 *– Constructs built in order to test whether high expression of AID T7RNApolymerase can increase the frequency of mutation activity or whether it leads to cell toxicity.*

Construct	Description	Promoter	Degradation Tag	Copy Number
Tox A	HP2.ATF(No Tag).mCherry.TetRep.CamR.pMB1-Kan	HP2.A	No	High
Tox B	HP2.ATF(LVA).mCherry.TetRep.CamR.pMB1-Kan	HP2.A	Yes	High
Tox C	HP2.ATF(No Tag).mCherry.TetRep.CamR.p15a-Kan	HP2.A	No	Low
Tox D	HP2.ATF(LVA).mCherry.TetRep.CamR.p15a-Kan	HP2.A	Yes	Low
Tox E	HP3.ATF(NO tag).mCherry.TetRep.CamR.pMB1-Kan	HP3.A	No	High
Tox F	HP3.ATF(LVA).mCherry.TetRep.CamR.pMB1-Kan	HP3.A	Yes	High
Tox G	HP3.ATF(NO tag).mCherry.TetRep.CamR.p15a-Kan	HP3.A	No	Low
Tox H	HP3.ATF(LVA).mCherry.TetRep.CamR.p15a-Kan	HP3.A	Yes	Low
Tox I	HP2.mCherry.TetRep.CamR.pMB1-Kan	HP2.A	-	High
Tox J	HP2.mCherry.TetRep.CamR.p15a-Kan	HP2.A	-	Low
Tox K	HP3.mCherry.TetRep.CamR.pMB1-Kan	HP2.A	-	High
Tox L	HP3.mCherry.TetRep.CamR.p15a-Kan	HP2.A	-	Low
Tox M	CamR.pMB1-Kan	-	-	High
Tox N	CamR.p15a-Kan	-	-	Low
Tox O	HP2.T7RNApol.TetRep.CamR.pMB1-Kan	HP2.A	-	High
Tox P	HP2.T7RNApol.TetRep.CamR.p15a-Kan	HP2.A	-	Low

Induction Assay

Initially, all test constructs were transformed into the repair deficient host *E.coli* strain (GM31 UGI+) described previously and plated onto LB agar plates supplement with Chloramphenicol (100 µg/ml).

Biologically independent colonies were picked from transformant plates and used to directly inoculate into 96-well plates. Plate layout is described in ***Appendix- Figure 11.11*** and contained two individual samples for each variant. Negative controls consisting of blank media and the original GM31 UGI+ strain were also included. Each well contained 200 µl of LB supplemented with Kanamycin (50 µg/ml) and Chloramphenicol (100 µg/ml). The plate was covered with BreatheEasy ™ membrane and grown overnight at 30°C to permit continued aeration and to avoid condensation with the sample wells.

The samples were then measured for cell density at optical density (OD) 600nm using a BMG plate reader. Calculations were automated to ensure accurate normalisation to 0.2 OD600, and correct volumes of each culture were added into fresh MOPS medium (without antibiotic) in a clean 96-well plate (Costar) with clear plastic lid. Samples were then grown for 2 hours at 30°C in a bench top incubator, with shaking set at 200 RPM. This second period of non-induced outgrowth was important to obtain samples that were at similar stages of growth, prior to induction.

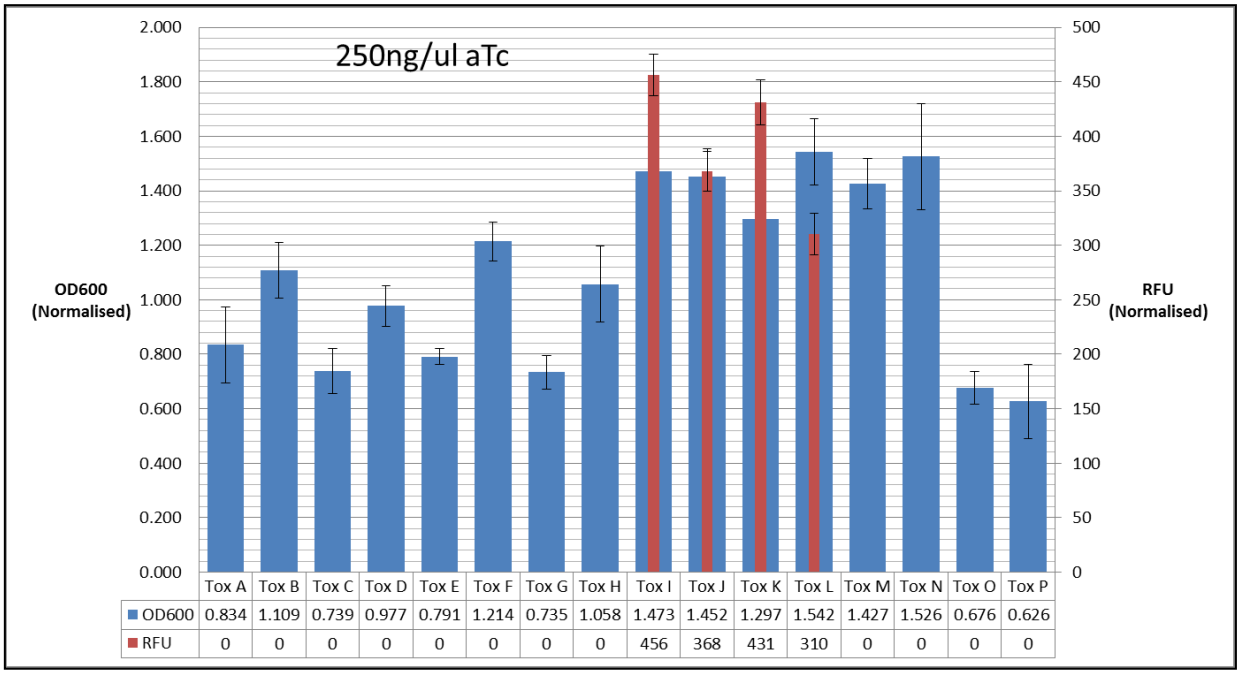
A further normalisation step (0.2 OD600) was performed, with cultures added to fresh MOPS medium and differing concentrations of anhydrotetracycline (aTc) inducer (0 ng/µl, 60 ng/µl, 250 ng/µl) in a black plastic, clear bottom, clear lid, 96-well plate (Greiner) to ensure both OD and fluorescence could be measured during outgrowth. Media plates had been pre-warmed to 30°C to avoid any possibility of cold shock, as previous anecdotal evidence had suggested that this could dramatically affect the rate of outgrowth.

All outgrowth, OD (600 nm) and fluorescence, FL (Ex 587 nm – Em 610 nm) measurements were carried out in the BMG plate reader at 30-minute intervals. Samples were allowed to grow for 6-hours at 30°C.

6.2.4 Results and Discussion

2 independent samples for each variant tested were tested across 3 repeats of the induction assay over successive days. Therefore, each end point reflects the mean average of 6 samples from each variant. OD values were normalised against medium only blank wells and relative fluorescence units (RFU) were then normalised against empty strain cells at comparative OD time points. Summary data is provided in ***Appendices, Table 11.21 and Table 11.22.***

Although every effort was made to ensure that conditions were consistent each day, it was noted that the growth of every sample on Day 1 was somewhat slower across all samples. This was most likely a consequence of temperature shock when sub-culturing into fresh media. To avoid any further possibility of this during later repeats (Day 2 and 3), the final plate was left for 1 hour at 30°C prior to induction.



	Tox A	Tox B	Tox C	Tox D	Tox E	Tox F	Tox G	Tox H	Tox I	Tox J	Tox K	Tox L	Tox M	Tox N	Tox O	Tox P
OD600	0.834	1.109	0.739	0.977	0.791	1.214	0.735	1.058	1.473	1.452	1.297	1.542	1.427	1.526	0.676	0.626
RFU	0	0	0	0	0	0	0	0	456	368	431	310	0	0	0	0

Figure 6.2 *– Final 6-hour end point measurements for both OD (blue bars, left Y axis) and RFU (red bars, right Y axis) are shown for constructs used to explore toxicity effects of overexpressing ATF mutator device. The data shown here corresponds to an average measurement across 3 repeats of the assays. Standard error bars are shown. The data presented corresponds to strongly induced cultures (250 ng/ul). Sample names reflect the constructs named in* ***Table 6.1***

It was clearly evident that constructs which actively express the ATF mutator (A – H) do not reach the same cell density as the empty vectors grown under the same conditions (I – N). I believe this

indicates that overexpression of T7RNApol can be detrimental to the fitness of the host strain and indeed the toxicity effect of overexpressing wildtype T7RNApol is clearly demonstrated here (O – P). In fact, the final end point OD reading suggests overexpression of T7RNApol results in the host strain growing >50% slower. The literature reveals that early T7 expression systems, in which a purified form of T7RNApol was used for induction, often saw a toxicity effect (*Studier, 1991*). As such, systems with the ability to control and inhibit T7RNApol often express the T7 lysozyme. It was also noted that in order to protect cell fitness, tight repression of T7RNApol expression can be maintained under strong inducible promoters such as the pLac (*Dubendorf and Studier, 1991*) or the novel pTet promoter set created within this work (***Chapter 4***).

The expression of ATF mutator appears somewhat advantageous and indeed less toxic to the cell than expression of wildtype T7RNApol. I would suggest that this is because the N terminal fusion (N-AID-T7RNApol-C) is potentially blocking a specific protein-protein interaction that is the causes the underlying toxicity effects in this first instance.

Interestingly, the addition of an LVA degradation tag (B, D, E, H) also appeared to marginally improve cell fitness as reflect by a higher OD in all such variants. However, tagged variants did not recover to the same fitness as those variants that did not include ATF.

Unfortunately, expression of an RFP (mCherry) protein, positioned downstream of the AT7 to create a single mRNA transcript was unsuccessful. The hope was that this would provide a clear read out for the level of ATF transcription, but I did not see any fluorescent output at the anticipated excitation and emission wavelengths for these samples (A-H). Coincidentally, I did not see any evidence of mCherry protein expression in SDS-PAGE or Western Blot analysis of total cell protein lysate (*data not shown*). Positive controls showed mCherry expression from a single gene cassette was possible, revealing a strong emission profile at 610 nm (I-L).

Levels of fluorescence seem to show some correlation with the strength of hybrid promoters used to drive expression. Previous work has identified that hybrid promoter 2 (HP2) is a marginally stronger promoter than hybrid promoter 3 (HP3) when considering samples I/K or J/L, respectively.

6.2.5 Conclusions

The data collected here does not show that expression of the mutator protein on a high or low copy plasmid backbone has any significant consequence on growth rate. However, I anticipate that an excessively high copy number will only serve to worsen the toxicity effect seen in many T7 expression systems. There are two key points to take forward into future work; 1) T7RNA polymerase expression should be regulated to a lower level by using minimum inducer concentration and potentially using a low copy replication of origin 2) An LVA degradation tag might be useful in dampening any toxicity effect and this may also be beneficial in removing the mutator protein from the *in vivo* system.

6.3 Improving the enzymatic relationship between AID and T7 RNA Pol

6.3.1 Background

The ATF fusion protein has already been identified as a capable mutator protein, able to generate mutations within a target gene. However, previously work has shown that numerous rounds of mutational activity are required to introduce pronounced and a meaningful number of changes with very few mutations during a single 16-hour outgrowth (***Chapter 2***).

An extreme mutator phenotype would limit the users control of the overall circuit, given that high frequent changes would bypass the ability to screen or identify a useful protein. However, current mutation rates are relatively low and if we truly intend to sample across a vast sequence space (which may be the required to identify truly novel protein functionality) then it would be valuable to have the capability to increase rate of mutation.

6.3.2 Justification and Proposal

As previously described T7RNApol is from the T7 bacteriophage, and is capable of highly specific transcription from its cognate promoter, achieving a very high rate of transcription within *an E.coli* host. AID has been shown to cause mutations on highly transcribed genes (*Bransteitter et al, 2003., Pham 2003, Dickerson et al, 2003 and this work - Chapter 3.3*). However, many questions regarding AID biochemistry remain i.e. a direct correlation of ssDNA binding affinity and *in vivo* function has yet to be revealed and information regarding AID processivity and substrate specificity remain elusive. Although the requirement of ssDNA for *in vivo* AID activity has been shown in numerous studies, it is hoped that a 3D crystal structure of AID will elucidate these more complex answers.

I quickly became aware that both proteins of interest (T7RNApol and AID) and different rates of activity. T7RNApol is a viral transcription factor that has an exceptionally high elongation speed, whilst AID is a highly regulated mutator protein and much lower rate of activity. It was my belief that by carrying out certain site-directed changes on important functional domains, it would be possible to synchronised the level of activity of both, which would likely mean that the fusion construct can work as intended.

My proposed model of activity allows for two potential outcome scenarios (***Figure 2.2***). Firstly, an increase in elongation speeds may result in a greater number of transcriptional events, meaning a larger number of exposed ssDNA targets. This could increase mutational activity if there is an increased opportunity for AID to bind to ssDNA and induce deamination events. Alternatively, slower elongation speeds may increase mutational frequency by closely matching T7RNApol and AID rate of activity. This could create a larger window of opportunity for AID ssDNA binding and encourage a greater number of deamination events.

I was able identify C-T and G-A transitions (average of 2 – 3 mutations per 5-day outgrowth) above the negative control and this must suggest that current ATF fusion does permit some level of AID induced mutation even with unassociated rates of activity. My initial belief is that T7RNApol tendency to stall during transcription would then provide the opportunity for AID to interact with its target. Furthermore, T7RNApol is known to stack up along the transcription bubble, because host *E.coli* translational machinery will not be in sync with the consequence being that surplus mRNA is marked for removal. Although previous results have shown that ATF has a lower level of transcription than wildtype T7RNApol it is likely above the rates of AID activity.

The principal aim was to connect both functional proteins enzymatically, rather than just physically to ensure both are able to carry out their desired function.

6.3.2.1 AID Activity

AID enzymatic activity has recently been summarised in the excellent review by Larijani and Martin (*Larijani and Martin, 2012*). In brief, it does seem that AID acts completely independently from any polymerase activity and I was able to quickly ascertain that AID does not share the same enzymatic speeds as many other DNA binding enzymes.

One study compared the *in vitro* catalytic and binding kinetics of purified AID on predicted motif target ssDNA, suggested that the K_m (Michaelis constant) for purified GstTag-AID was 10-15nM and at V_{max} (point of maximal velocity) the K_{cat} (turnover) was calculated as $0.0042s^{-1}$, which equates to around 1 deamination event every 4 minutes per molecule of AID (*Larijani, 2007*). This study also revealed that the K_d (dissociation constant) of bound AID to ssDNA ranged between 0.3 and 1.2nM. The binding affinity of AID appears very high against test substrates, calculated as between 0.3 and 2nM, and remains unaffected by the size of transcription bubble or presence/absence of sequence motifs. It also suggests that the AID-bubble complexes are relatively stable and have prolonged half-lives of around 8 minutes. Maintaining a stable complex would be essential for thorough deamination activity given the low catalytic activity.

The author compares these rates to those of other DNA binding proteins. Firstly, the K_d of AID compares favourably to the K_d of single-stranded binding proteins. Furthermore, a K_d value in the nanomolar range (enzymes typically have K_ds in the micromolar range) coupled with a long half-life and sequence independent binding, suggests that binding to ssDNA is not a fleeting or transient interaction, but in fact the AID protein may remain bound for an extended period, possibly even after the deamination event has occurred. This may also validate the noted appearance of multiple deamination events in close 'patches' along the target, potential due to persistent AID binding at a specific target location (*Pham et al, 2011*). In this study, unlinked T7RNApol expression of a lacZ marker gene is used to

assess the frequency of deamination events. Given the tendency of AID to create multiple mutations on certain ssDNA targets whilst alternative clones remain untouched, suggests that AID demonstrates processivity, becoming strongly associated with the ssDNA target once initially bound. When the AID/target incubation periods were extended, there was some level of normalisation between the number of mutations seen on each ssDNA target. The author also hypothesises that the prolonged association of AID at the site of deamination may actually encourage the recruitment of DNA repair factors and therefore slow and inefficient enzymatic activity of AID is actually beneficial for generating genetic diversity.

AID has been shown to associate with other DNA binding proteins, specifically Replication Protein A (RPA) which binds to ssDNA during DNA replication and nucleotide excision repair. This prevents it from refolding or forming secondary structures (*Chaudhuri et al, 2004*). The bacterial equivalent is named single-strand binding protein (SSB). This might suggest AIDs dependency on co-factors and its direct effect on downstream repair post-deamination.

In comparison, the closely linked Uracil DNA glycosylase (UDG) protein, which also scans ssDNA prior to initiating BER, shows a much higher enzymatic efficiency than AID (*Porecha and Stivers, 2008*).

A number of mutational studies have been carried out on AID to elucidate more information regarding structural properties and functional domains. One such study employed random mutagenesis to identify variants that were able to achieve higher levels of mutations *(Wang et al, 2009)*.

Whilst the K_m, K_{cat} and K_d of AID are likely to be tailored specifically towards its *in vivo* function and its native role in somatic hypermutation, these rates do not reflect AID activity within our system given the constraints caused by fixing AID to T7RNApol. Thus, it is proposed that this would force AID to travel along the exposed ssDNA at a high processivity speed defined by the T7RNApol to which it is tethered

A general correlation between the catalytic rate of AID and its efficiency has been shown *in vivo* (*Larijani and Martin, 2012*). However, there is still no direct correlation with AID ssDNA binding affinity and its functional activity *in vivo*. A high transcription rate clearly increases the frequency of ssDNA exposure and opportunity for AID activity *in vivo,* as previously referred to in **Chapter 2**, but a reduction in binding affinity is likely to greatly reduce catalytic efficiency and may prevent any deamination events from occurring. There is currently no information regarding AID processivity *in vivo*.

A key consideration when detailing these *in vitro* studies is that they require purification of the AID protein with various labels and tags, requiring complex purification steps that may result in changes to protein structure and folding. Indeed, the author states that a large fraction of the AID protein was inactive, likely as a result of harsh conditions used to overcome its notoriously poor solubility, which has prohibited purification efforts and the elucidation of a crystal structure. My early intentions were also

to attempt purification of AID and cloning of the CDS was performed but expression of the protein was initially unsuccessful (***Appendices 11.6***).

Another consequence of over expressing AID is inevitably protein aggregation, something which would never occur *in vivo* given that AID is so tightly controlled and lowly expressed. There is the potential of overloading *in vitro* assays with highly concentrated purified AID protein which may actually reduce enzymatic activity.

6.3.3 T7RNApol Activity

The crystal structure of T7RNApol reveals both the catalytic and promoter recognition sites and enabled researchers to orientate the transcribing polymerase protein against the DNA template on a tentative model (*Sousa and Wang, 1993*). A catalytic pocket sits inside a large template binding cleft and the residues at these locations directly influence transcriptional activity.

An important consideration is that the N terminus (the position which the AID protein has been fused within the ATF variant), is not located near the catalytic site, and thus polymerase activity should not be affected, which is somewhat in agreement with results from **Chapter 2.** This is also confirmed by evidence showing residues can be removed from the N-terminus before signal markers are attached, and yet polymerase activity is unaffected (*Dunn et al, 1988*).

In contrast, there is significant importance placed on the C-terminus, which interacts with the catalytic region. C-terminal fusions are extremely detrimental to catalysis and transcription. Therefore, these regions are potential targets for creating a number of T7 polymerase mutants, which may alter the rate of transcription or cause a transcriptional stallage event. The functional domains of T7RNApol are summarised (***Table 6.2***).

Table 6.2 *– Functional Domains of T7RNApol*

Domain	Function	Residues
N-Terminal Region	Promoter recognition and DNA melting	1-324
AT-rich recognition loop	Inherent flexibility allows indirect binding to the DNA sequence in the minor groove at base pair -17 and -13. Side chains of Arg96 and Lys98 interact with the phosphate groups causing distortion to the phosphodiester backbone of minor groove (*Cheetham, 1999*)	93-100
Intercalating b-hairpin loop	Transcription initiation – formation and stabilization of the transcription bubble. Melting of promoter duplex so template strand is accessible to the binding site of T7RNAP. Val237 residue directly separates template and non-template strand (*Cheetham, 1999*)	
Polymerase Domain	All of the below	325-883
Thumb	Stabilize ternary complex during transcription by wrapping around the template. Mutant with shorter thumbs show less processivity due to lower affinity for template DNA (Bonner, 1994)	325-411
Palm	Situated at base of deep cleft, forming b-sheets containing catalytically most critical and conserved amino acids Asp537 and Asp812, directly interact with two Mg(II)ions (Bonner, 1992)	412-565 and 785-883
Finger	Raises above the palm domain, forms a wall on the left side of the active-site. Includes specificity loop, involved in direct interaction with the promoter DNA. Also includes important catalytic residues that are in proximity to active site Tyr639, Lys631 (Sousa, 1993).	566-784
'Active Site'	Located in the deep pocket bounded by the fingers, thumb and palm and N-terminal domain. Contains two aspartate residues that are involved directly in catalysis. Tyr639 discriminates the incoming sugar by forming hydrogen-bond interaction with 2'OH group with the incoming ribonucleotide. dNTPs lack the 2'-OH group and thus would not be incorporated. His784 may also have a role as a sugar discriminator (*Brieba, 2000*)	639
Specificity loop	Specific to T7RNAP, projects into the DNA binding cleft. Direct interactions with the upstream promoter sequence. Residues Asn748 (-10 and -11), Arg746, Arg756 (-9), Gln758 (-8) are important. (interaction with promoter sequence in brackets)	739-770

Many publications have noted that although *E.coli* expression systems which employ T7RNApol achieve very high transcriptional rates, the final increase in the level of translated protein does not seem to correlate. This has been explored further by has expressing various target genes under wildtype RNA polymerase (RNApol) or T7RNApol, which revealed that protein yield was reduced for T7RNApol (*Iost & Dreyfus, 1995*). This appears to be counter intuitive, but to understand this strange effect we must first consider that the mechanics of transcription and translation are not entirely independent events. As a matter of fact, newly synthesised mRNA interacts with the ribosome almost immediately after transcription, and thus the RNA polymerase and ribosome are located in close proximity. RNApol and the ribosomal machinery are well matched in processivity rates which prevents long stretches of exposed mRNA. In the case of T7RNApol transcription, excessive mRNA production, coupled with an 8-fold increase in elongation speed (compared to RNApol), means that large stretches of mRNA are exposed to nucleases within the cell. In this case, the T7RNApol moves along the template DNA strand far ahead of the WT ribosome (*Iost et al, 1992*). Therefore, T7RNApol transcripts will often be degraded before they can be translated into viable polypeptides although evidently this lack of stability doesn't

necessarily mean that T7RNApol produces less protein because of the sheer number of mRNA transcripts it produces.

Table 6.3 *– Single position T7RNApol mutants - T7RNApol mutants and single position mutation affects are described. Many of these mutations have actually been identified in a similar in vivo direct evolution system in which T7RNAPol was the target protein (Esvelt, 2011). ^ mutation lost in highly stringent conditions, # include mutations identified in Temme and Voigt, 2012.*

Point Mutation	Domain Location	Anticipated Affect
Y639F	Active Site	Introduces dNTPs into mRNA – false transcripts
G645A	Finger/Active Site	Reduction in transcription elongation speed and catalytic activity
I810S	Palm	Significantly weakens elongation complex – incomplete mRNA
R627S	Finger	Slow elongation time
F644A	Finger/Active Site	Reduction in transcription elongation speed and catalytic activity
Q649S	Finger/Active Site	Reduction in transcription elongation speed and catalytic activity
P266L	Hinge	Weakens promoter binding, reduces abortive cycling, could increase transcription rate by early clearance through promoter
I4M*, G175R*, E222K*, E63V*, N165S*^, P248S**^	N-Terminus	Improved binding to T3 / T7 promoter hybrid sequences
E484A*, G542V*, E877K**^	Palm	Improved binding to T3 / T7 promoter hybrid sequences
V574A*, R632S#, G675R*, E775K*, L699I*^, I681L**^	Finger	Improved binding to T3 / T7 promoter hybrid sequences
Q238L*^, Q239R**	B-hairpin loop	Improved binding to T3 consensus promoter sequence. Direct interaction with the promoter sequence, separating both strands.
K98R**	AT-rich recognition loop	Improved binding to T3 consensus promoter sequence. Indirect binding to the DNA sequence in the minor groove
N748D**	Specificity loop	Known to allow recognition of T3 base at -11 position of promoter

Extensive work to determine whether or not mRNA transcripts could be stabilised by slowing T7RNApol elongation were useful for establishing a range of T7RNApol mutants (*Bonner et al, 1992., Makarova et al, 1995*). Many of these variants are accurately described and have well-defined effects on polymerase function (***Table 6.4***). In much of this work, the level of β-galactosidase produced under expression from T7 mutant variants was normalised against the activity of wildtype version of the T7 RNA polymerase. Total yield per transcript was dependent on expression strain. The fact that the strains used in this study were actually derived from a GM31 *E.coli* strain so the rates of expression are highly relevant to my proposed expression system.

Interestingly, the mutants with slowest elongation time *in vitro* (R637S and I810S) failed to show complete transcription of the target gene *in vivo*. These specific residue changes are located close to the C-terminus of the protein which is essential for elongation. It is believed that the transcription complex

formed by T7RNApol is rather unstable and thus the polymerase has to have a fast elongation time to compensate. This is vital to ensure that complete mRNA transcripts can be produced before catastrophic failure of the transcription bubble. In order to compensate for this instability, T7RNApol demonstrates sudden and increased processivity of T7RNApol once it reaches 9 bp from the transcriptional start site. However, this characteristic was less likely to be observed seen in the slow R637S and I810S mutants.

Mutants with elongation speeds that are comparable to the wildtype T7RNApol, showed a similar level of protein production per mRNA transcript (R551S and Y639F), but this discrepancy was likely due to miss-incorporation of dNTPs by T7RNApol i.e. the resulting protein was non-viable. Interestingly, the Y639F mutant consists of a single residue change in the catalytic active site of the protein and demonstrates a substrate dependent DNA polymerase activity *in vivo* i.e. the polymerases were able to synthesise RNA and DNA transcripts (*Sousa & Padilla, 1995*). The author hypothesises that this position is essential for sensing incorrect geometry in the structure of the substrate, template or product and is important for identifying mismatches. The inclusion of dNTPs in target gene transcripts may not have an obvious be benefit for my own application given that it will reduce GFP expression, with one publication suggesting Y639F mutant produces only 50% of the transcribed protein compared to wildtype T7RNApol *in vivo* (*Makarova et al, 1995*). However, this mutant could be potentially useful for reducing the level of mRNA produced and avoid overloading the host system. The effects of adding dNTP bases into mRNA transcripts are relatively unknown.

Table 6.4 *– Activity of T7RNA polymerase mutants measured as b-galactosidase yield generated from expression of lacZ under as synthesised by various mutant T7RNApol. In vitro speed and processivity of each variant were published in the work by Bonner et al, 1994. Adapted from Makarova et al, 1995.*

Mutation	***In Vitro* Speed (nt/sec)**	***In Vitro* processivity**	**tRNA (37°C)**	**β-Galactosidase (37°C)**	**37°C rne+ cells (nmol)**	**42°C rne+ cells (nmol)**	**42°C rne-50 cells (nmol)**
T7RNAPol	240	0.94	100	100	100	100	100
Y639F	240	0.93 – 0.94	45 +/- 5	13.5 +/- 1	30 +/- 4	31 +/- 1	46 +/- 10
R551S	240	ND	110 +/ 5	114 +/- 20	105 +/- 15	79 +/- 20	75 +/- 10
F644A	170	0.90 – 0.93	5.5 +/- 1	7.5 +/- 1	136 +/- 30	290 +/- 40	92 +/- 10
Q649S	160	0.88 – 0.91	25 +/- 5	50 +/- 8	200 +/- 10	230 +/- 10	113 +/- 12
G645A	90	0.81 – 0.87	20 +/- 4	68 +/- 7	340 +/- 30	460 +/- 20	100 +/- 12
R627S	80	0.81 – 0.85	<2	<0.5	ND	ND	ND
I810S	40	0.70 – 0.75	<2	<0.5	ND	ND	ND
E.coli RNA pol					1975 +/- 200	1440 +/- 30	9300 +/- 700

Mutants which demonstrate slightly slower elongation time (F644A and Q649S) showed a significant increase in the amount of target protein. Most importantly, the G645A mutant was considered the slowest mutation still able to transcribe the complete gene, and surprisingly the mRNA transcripts were less sensitive to RNase than even WT RNAPol generated transcripts.

A very interesting paper by Canugovi and colleagues investigated the affect that the rate of transcription *in vitro* would have on the activity of the AID (*Canugovi et al, 2009*). They found that the G645A mutation did reduce the rate of polymerase elongation by 3-4 fold, without changing the rate of transcription initiation, suggesting that the stallage effect causes the transcription bubble to be open for a longer period of time, without effecting the likelihood of transcription events occurring. Indeed, the G645A mutant T7RNAP showed an increase propensity for 'clustered' mutations and may result in the AID protein residing at one location for a longer period of time, as it follows the active polymerase with access to ssDNA. The same effect was seen if transcription was stalled by a sudden removal of dNTPs from the reaction. Note here that the dNTP removal was only possible *in vitro* and in fact the group showed no evidence of clustered mutations *in vivo.*

However, I saw still saw value in exploring mutant T7RNApol's within my own *in vivo* system. Clustered mutations may actually be beneficial if mutations can be concentrated to important functional domains as they will reduce the probability that any induced changes are silent. In practice two mutations within the same triplet codon will inevitably cause amino acid residue changes and this is desirable when exploring protein sequence space.

Unfortunately, no variant with increased transcriptional activity has previously been identified literature and given that the proposed system uses a fixed cognate promoter recognition sequence it would be difficult to identify one. Although a considered attempt to engineer T7RNApol using PACE directed evolution was able to engineer a T7RNApol mutant with improved transcriptional activity on a T3 promoter sequence (600% increase in activity over wtT7RNApol), the overall rate of transcription was not increased (*Esvelt et al, 2011*). Although another recent attempt to create T7RNApol variants claims to improve orthogonality and activity of variants from previous studies, they were unable to recover complete activity of wildtype T7RNApol (*Ellefson et al, 2014*). However, it should also be noted that if said T7RNApol variants strongly target their orthogonal promoters with a high transcriptional rate, then it may be useful for later iterations of the complete *in vivo* system.

One further observation is that the P266L T7RNApol variant might reduce the number of small abortive transcripts seen during initial transcriptional complex formation. Abortive cycling may be increased by pyrimidine A or T bases within the first 10bp of transcribed sequence. The P266L should make the T7RNApol more tolerant for a wide range of initially transcribed sequences (ITS). The earlier paper establishes that abortive cycling is more likely if the catalytic turnover of the T7RNApol is

reduced. This agrees with the other stallage mutants previously mentioned, which directly affect the catalytic site. These are far more likely to drop off at the initial transcription complex (*Guillerez et al, 2005*).

Therefore, combining mutagenic changes which create a T7RNApol mutant with reduced processivity (G645A) and with the capability to bypass abortive cycling (P266L) is likely to be well suited for my application, and for use in the ATF fusion mutator device.

6.3.4 Methodology

Firstly, a number of T7RNApol variants taken from the literature were created by site-directed mutagenesis (***Section 9.3.1***). The same changes were also made to the ATF mutator device, previously shown to be an active and targeted mutator. A summary of the phosphorylated primers required to introduce these changes is provided, along with a description of the anticipated characteristics of each variant (***Table 6.5***). It should be noted that in this instance all variants remain under a lac inducible promoter and are activated by a switch in catabolism during logarithmic growth.

As in Chapter 2, the same targeted mutation / loss of function assay was used to observe a loss of fluorescence. GFP+ and GFP- cells were visualised across 3 days of successive outgrowth. It is important to note that a reduction in transcription rate could make it difficult to distinguish between GFP+ or GFP- cells during flow cytometry, as both population peaks will move closer together or begin to blur and overlap. If any variant of the AID mutator is particularly active, it is possible that an intermediate population showing a weakened fluorescence profile will begin to appear, making qualitative analysis much more difficult. Furthermore, a loss of complete transcriptional activity would represent a GFP- population, so in order to distinguish between true knockdown/knockout mutations of GFP and to avoid false positive, T7RNApol controls (which do not cause mutations) are also included within the assay.

Table 6.5 *- Primers used to construct a number of T7RNApol mutants by Site-Directed Mutagenesis*

Mutant	Primers Used for Site-Directed Mutagenesis	Tm	Characteristics
Y.639.F	FOR 5' PHOS-TTTgggtccaaagagttcggc 3'	65C	Similar elongation speed to wtT7RNAP
	REV 5' PHOS-agccagcgtcatgactgaac 3'	66C	
G.645.A	FOR 5' PHOS-GCTttccgtcaacaagtgctgg 3'	65C	Slightly slower than wtT7RNAP
	REV 5' PHOS- gaactctttggacccgtaagc 3'	64C	
I.810.S	FOR 5' PHOS-AGCcacgactccttcggtaccattc 3'	67C	Slowest compared to wtT7RNAP Incomplete transcription
	REV 5' PHOS-cagtgcaaaagattcgattccg 3'	68C	
P.266.L	FOR 5'PHOS – CTGatgttccaaccttgcgtagttcc	67C	Reduction in abortive transcripts Facilitates promoter clearance
	Rev 5'PHOS – agagatgccagccagcgcac	72C	
G.645.A P.266.L	Double Mutant		As Above

6.3.5 Results and Discussion

Initial testing of the various T7 and ATF mutants (outlined in ***Table 6.3, Table 6.4, Table 6.5***) seems to agree with much of the work outlined previously (***Chapter 2***) and confirmed within the literature (*Sousa et al, 1993., Bromer et al, 1994., Cheetham et al, 1999., Brieba et al, 2000*). Summary data is provided along with overlaid histogram traces for each variant. This highlights consistent profiles across all repeats (***Appendices - Figure 11.12, Figure 11.13***). Qualitative analysis of the various population profiles shown in ***Figure 6.3*** and ***Figure 6.4*** seem to show the anticipated effect for each mutant.

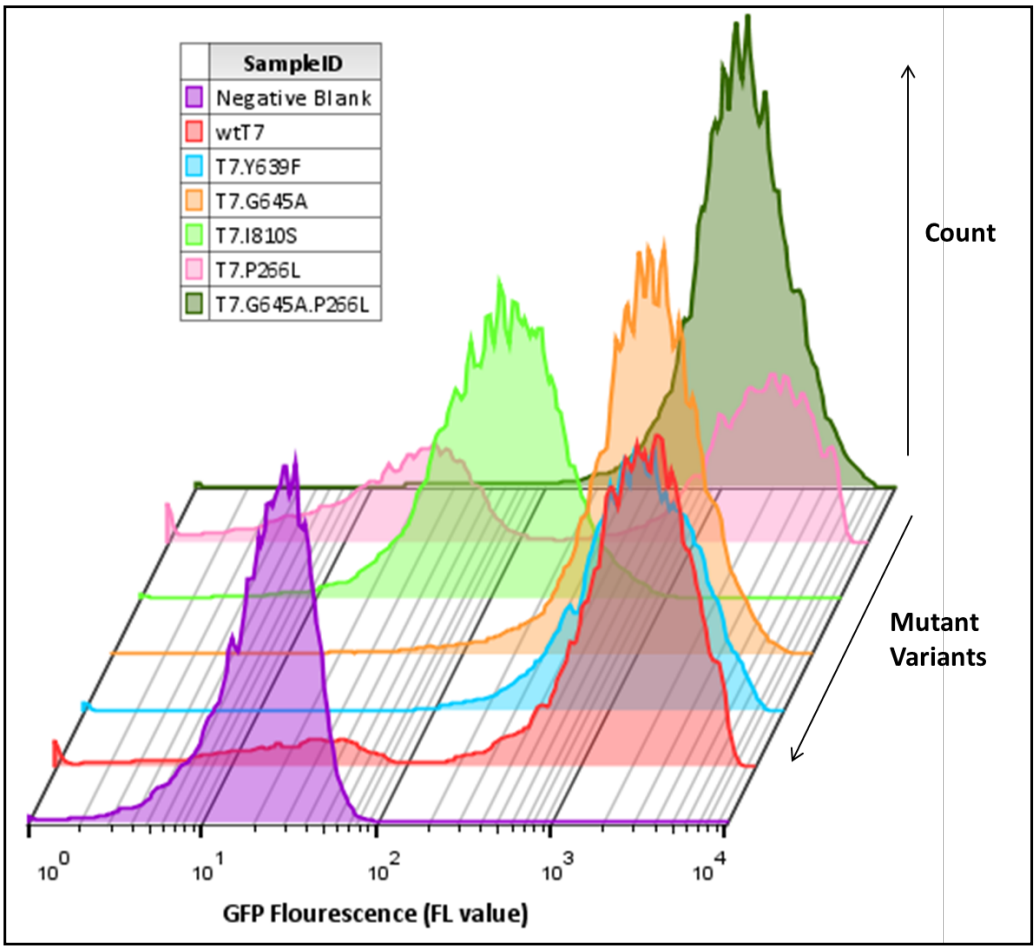


Figure 6.3 – *The differing fluorescent profiles of GM31 UGI+ strain expressing T7RNApol mutants*

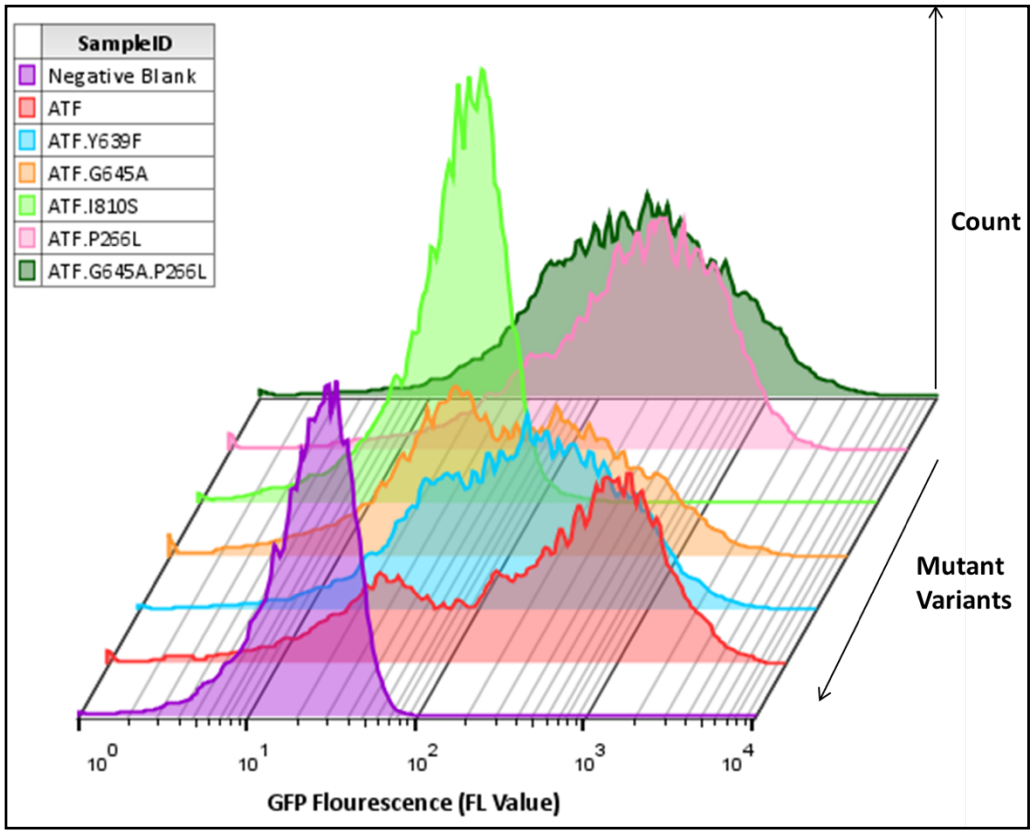


Figure 6.4 *- The differing fluorescent profiles of GM31 UGI+ strain expressing ATF (mutator device) mutants.*

It is important to remember that flow cytometry measures the level of fluorescence for each single event, or in this case, each individual cell. Therefore, peaks are generated when cells with similar fluorescence intensity accumulate during the sampling period. The width of the peak reflects the range of fluorescence intensity across the cells within a sampled culture i.e. the wider peak, the larger the deviation in FL value. When referring to the X axis (log scale), a peak at 10^4 indicates a greater fluorescence than a peak at 10^3.

Therefore, increased transcriptional activity may lead to higher levels of fluorescence within any individual cell if the host translational machinery is able to efficiently convert mRNA into protein. However, the flow cytometry assay does not allow us to disconnect transcription and translational efficiencies and so the assay might be of limited use for this particular purpose, given that I would ideally like to visualise the effect of mutations of T7RNApol transcriptional activity, speed and processivity.

Nevertheless, I am able to make some assumptions given the use of T7RNApol controls. The real benefit of this assay is to quickly compare the fluorescent profiles of a range of mutants and identify obvious differences with the active ATF mutator device.

Firstly, a true negative sample contained blank GM31 UGI+ cells and results in a strong non-fluorescence peak profile (purple trace - ***Figure 6.3*** and ***Figure 6.4***). This was used to gate raw data into either GFP+ or GFP- populations.

The results for both wtT7RNApol and ATF expression reflect my previous analysis which was also carried out in this repair deficient strain. Primarily the wtT7 (red trace - ***Figure 6.3***) shows a large fluorescent population and a small number of non-fluorescent cells which are likely a consequence of infrequent background mutations accumulated over the multiple rounds of outgrowth. The profile for ATF is much as expected, showing that the maximum florescence within each cell is somewhat lower, a consequence of the creation of the N-terminal fusion protein. It also presents with a characterised and gradual slope towards a non-fluorescent population (red trace – ***Figure 6.4***)

The first mutant to be observed, Y.639.F (blue traces), was expected to present with a lower level of fluorescence given that the final level of translated protein is likely to be lower due to its tendency to add dNTPs within transcribed mRNA. A marginal reduction in FL value can be identified in both T7RNApol Y.639.F and ATF Y.639.F variants compared to wildtype.

P.266.L mutants (pink traces) were not expected to change processivity of the transcription factor, merely provide a recovery of abortive cycling, which is a characteristic of viral T7RNApol transcription. The results shown here seem in agreement given that there is strong similarity with both wtT7RNApol and ATF variants (red traces). In fact, the T7RNApol P.266.L trace seems to show the emergence of a non-fluorescent population, similar to the one which begins to appear in wildtype T7RNApol after successive outgrowths. This is interesting observation, given that previously this peak was described as a consequence of background mutations due to the high levels of viral transcription. If the P.266.L mutant is in fact reducing the number of aborted transcripts, we may also expect that the basal level of mutation to be slightly increased by virtue of the fact these background mutations occur at the exposed transcription bubble.

I810S variants (light green traces) are known to severely reduce transcriptional activity to the extent that the mutant polymerase is unable to achieve successful transcription of the target gene. This is reflected by the large non-fluorescent peak which in fact overlays the peak shown by the negative control (***Figure 6.3***). It is likely that the I810S variants produces short abortive transcripts that do not result in a viable protein.

G.645.A mutants (orange traces) are the most interesting variants in relation to my intended application. They are particularly exciting given that they demonstrate a slower elongation speed and processivity *in vitro*, when compared to wildtype T7RNApol. This is reflected by a small but noticeable drop in FL values for GFP+ population for the T7RNApol G.645.A mutant compared to wildtype.

Most importantly, the emergence of a significant GFP- population when expressing the ATF G645A is fairly pronounced. In comparison to ATF, there is a ~40% decrease in GFP+ population. This variant also shows a transitional trace passing through an intermediate knockdown population and this would suggest that the ATF G645A variant is indeed both an active and potential even more potent mutator (***Figure 6.4***)

Interestingly, the double mutant G645A+P266L (dark green traces) was expected to be the most potent mutator by slowing elongation speed and ensuring that promoter clearance (avoidance of abortive cycling). This would mean a larger number of transcripts and an increased opportunity for the ATF mutator to act. However, it appears that the overall affect is much the same as when G645A mutant is expressed on its own, and so I would question the P266L mutation really has on increasing the number of transcripts produced.

6.3.6 Conclusions

Unfortunately, the conclusions made here are not informed by final sequencing data. In order to generate a meaningful heat map and reveal a significance change in mutation frequency, as a consequence of expression any of the ATF variants, I would need to comprehensively sequence 100s of individual colonies post mutation. However, the relatively high associated costs meant that the GFP loss of fluorescence assay was used as an initial read out of mutational activity.

I have also noted the importance of 5'UTR leader sequence, immediately following the promoter, given that wttT7RNApol favours initiation with GTP (*Imburgio et al, 2000., Brieba et al, 2002*). Therefore, it would be preferable to engineer a leader sequence that contains G/C bases to ensure that there is sufficient transcription regardless of the mRNA sequence. The alternative approach could be to trial different variants of T7RNApol, which have already been engineered to tolerate and initiate transcription from a broader range of nucleotide bases (*Esvelt et al, 2011*).

Another similar consideration, in as much as the activity of the mutator is defined by transcriptional activity, is that all potential transcriptional terminator motifs i.e. those that form secondary hairpin structures, should be avoided as this will result in aborted transcripts. A relic of T7RNApol abortive cycling during early initiation of transcription could appear as mutational biasing toward the start of the target gene.

In summary, this investigative tangent suggests that T7RNApol mutants can potentially improve overall activity of the AID-T7 fusion protein. Qualitative work appears to have been successful in identifying a polymerase variant that has a reduced propensity to undergo abortive cycling after transcription initiation. This can be combined with a mutant demonstrated a slow elongation speed in order to create a more potent mutator that achieves sufficient level of target protein expression.

7 Increasing Mutagenic Spectrum and Diversity

The initial ATF mutator device, which fuses AID to T7RNA polymerase, was shown to be an active mutator in the engineered *E.coli* host GM31 strain which also expresses the Uracil DNA glycosylase inhibitor (UGI) and results in a dampened base excision repair (BER) pathway. This interferes with the host's ability to remove deaminated bases caused by AID. Consequently, aberrant base changes propagate through subsequent rounds of DNA replication as the intermediate U/G pair is copied as if it were a C/G position.

Whilst this approach enabled me to test the efficacy of targeted deamination, the reality is that the current system is limited to creating a small number of mutations i.e. only C - T and G – A nucleotide transition can occur (dependent on which strand is targeted).

This section begins to explore ways to increase the spectrum of mutations, with specific focus on hijacking innate *E.coli* repair mechanisms by engineering translesion bypass (TLS) polymerases to encourage error-prone repair.

The general hypothesis is based on the theory that by re-enabling DNA repair in our host strain and ensuring that a very high frequency of deamination events will be targeted to the gene of interest by ATF, then normal high-fidelity repair will be overloaded. Therefore, it is expected that error-prone repair will compensate.

BER pathways rely on removal of the deaminated based by uracil deglycosylase (UNG) which creates abasic sites within the dsDNA target. The presence of concentrated and frequent abasic sites, become target sites for apurinic/apyrimidinic exonuclease (APE) which nick the phosphodiester backbone and leave a break within the dsDNA strand, which is accessible by certain DNA polymerases (*Krokan and Bjoras, 2013*). Understanding which DNA polymerases are involved, and indeed how they are recruited to these specific sites remains somewhat unclear

My proposed model for AID activity incorporates these associated proteins, although within the context of the prokaryotic host. Ultimately, I intend to add a higher level of tunability to this pathway, by controlling the expression for proteins responsible for active mutation and subsequent repair.

The strategy I have chosen is to co-express the low fidelity Pol IV Polymerase, known to be involved in non-targeted mutagenesis in *E. coli* and therefore tightly controlled by the host. Pol IV activity in both *in vitro* and *in vivo* assays has been shown to be highly dependent on its close association with the β-clamp of Pol I during SOS cell response. An interaction of overexpressed exogenous Pol IV with the host's normal DNA repair pathway would likely cause a significant increase in global mutations and result in a non-viable cell strain. I have identified that C-terminal deletions of 5 and 12 amino acids disconnect it from the innate replicative and repair machinery.

In an effort to ensure Pol IV is targeted to the specific lesions caused by AID deamination the intention I fused the replicative Pol IV domain to an exonuclease domain. By overexpressing AID-T7RNApol and Exo-DNAPol fusion proteins with the same strain, they may respectively lead to a high frequency of targeted deamination events and semi targeted removal of abasic sites prior to subsequent error-prone repair at the site of the original lesion. In principal, this semi-targeted strategy should help to maintain a viable host strain.

Given the complexity of fusing two very distinct protein domains, a library of variants has been constructed, which utilises the BASIC assembly standard to build numerous constructs with a modular approach. In particular, 3 different linker variants have been used to improve the possibility of identifying functional fusion proteins. Expression of the fusion variants tightly controlled using a recently characterised set of inducible promoters which respond to minute concentrations of inducer (***Chapter 4***).

7.1 Background

Early work established ATF is an active mutator that can be directed to a highly transcribed target gene. However, the overall goal is to build a device that is capable of sampling across complete sequence space. The current approach is limited to transition mutations, of which <70% are C:T transitions of the non-transcribed coding strand. A truly effective system would be able to achieve both transversion and transition mutations across the full length of the target gene.

It is also important to ensure that any increase in mutagenic diversity remains targeted to the gene of interest. For example, variants of dnaQ gene which encodes the proofreading subunit of DNA polymerase III holoenzyme produces the strongest known mutator phenotype with mutation frequencies of up to 100,000-fold higher than wildtype strains (*Cox et al, 1982*). However, this would be unacceptable within my system device and be extremely detrimental to the overall fitness of the *E.coli* strain.

When going back in to the literature, my attention was drawn to the innate eukaryotic processes which enable the fast-paced evolution of antibodies (Ab). AID is known to be highly active at specific regions within immunoglobulin genes (Ig) and highly expressed in within the context of earlier B cells. However, the level of observed genetic diversity identified far surpasses the limited spectrum of deamination transitions caused by AID. It was immediately obvious that the increased mutational spectrum and overall diversity seen with the hyper variable IG target genes is only possible if error-prone repair is allowed to occur (*Delker et al, 2009*).

Extensive work over recent years has revealed the importance of translesion by-pass (TLS) polymerases within DNA damage response pathways and demonstrates that they are essential to cell

survival. TLS polymerases enable cells to replicate when DNA lesions would cause the normal high-fidelity replicative polymerases to stall, something which is itself a marker for initiation of the SOS-response in prokaryotes by RecA (*Broyde et al, 2008*). The expression of TLS polymerase (Pol V) has recently been exploited as a general mechanism for introducing genetic diversity within a directed evolution system, albeit in an untargeted manner (*Esvelt et al, 2011*). It therefore seems logical to explore the use of error-prone TLS polymerases within my own directed evolution system.

7.1.1 Reactivating DNA repair

Both prokaryotic and eukaryotic systems are regularly exposed to endogenous and exogenous sources of DNA damage including the build-up of reactive oxygen species or UV damage, respectively. Quick and effective repair of changes to template DNA sequence is vital to cell survival and can be achieved in numerous different ways depending on the type of lesion.

Initially, the direct reversal of the mutational change might seem to be the most obvious way to revert any single nucleotide base change and restore the original sequence. Such mechanisms for direct reversal do exist in bacteria, fungi and even higher order mammals, but these approaches are often resource heavy, given that novel enzymes for each type of mutagenic change are needed. One such example is the methylation of guanine nucleotide bases requires expression of the methyl guanine methyl transferase (MGMT) protein. Although this process shows excellent fidelity, the fact that it is a stoichiometric rather than enzymatic, in as much as one molecule of MGMT is required for every methyl guanine target and it can be considered a general burden on the cell (*Lindahl et al, 1982*).

Alternatively, repair mechanisms seek to completely excise the lesion from DNA and are therefore more general response mechanisms which are particularly well suited to correcting single strand damage. These included base excision (BER) pathway which excise the modified base only, nucleotide excision (NER) which removes small sections of ssDNA as short oligonucleotides and mismatch repair (MMR) pathway seen as a responsive to replicative errors rather than removal of spontaneous or induced mutations by way of dsDNA and have been widely reviewed (*Friedberg et al, 1995., Wood, 1996*).

In reference to my own proposal, overexpression of AID leads to deamination of cytosine and results in the presence of incorrect Uracil bases within dsDNA. Therefore, BER is required to remove lesions from the DNA and is first defined by activity of a DNA glycosylase (such as UDG/UNG) which is able to identify the incorrect bases and catalyse the hydrolysis of the N-glycosidic bond between it and the deoxyribose – phosphate backbone. Both the human *udg1* and *E.coli ung* genes share a significant amount of sequence homology and have the ability to angle the target base within dsDNA so that it is positioned at the active site of the enzyme, ready for removal. UDG/UNG are specifically designed to remove uracil although an alternative Mug protein is also known to remove Uracil in response to a U:G

mismatch within *E.coli (Lutsenko and Bhagwat, 1999)*. It should be noted that UNG is strongly expressed in *E.coli* and shows a very high turnover rate which should encourage BER repair within my host strain.

After removal of the deaminated base, an apurinic/ apyrimidic site (AP/Abasic site) remains. Subsequent repair is initiated by an AP endonuclease, which breaks the phosphodiester backbone and leaves 5' end deoxyribose-phosphate and an exposed 3' end hydroxyl. In mammalian systems the endonuclease is APE and in *E.coli* it is in fact either Exo III (*Xth*) or Endo IV (*nfo*) (*Hedge et al, 2008*).

The final stage of repair requires the removal of the 5' end deoxyribose-phosphate residues and DNA synthesis to add the lost bases. This process can occur through short or long patch BER and includes the activity of translesion bypass polymerases. Firstly, a model for short patch repair in human cells requires activity of a deoxyribophodiesterase (dRPase) and is coupled with activity of DNA pol β to create a single nucleotide patch repair although given that no known homolog for dRPase has been identified in *E.coli*. The model for long patch repair in mammalian systems describes the activity of DNA pol δ or DNA pol ε which are capable of intereacting with proliferating cell nuclear antigen (PCNA) DNA clamp to carry out strand displacement in the 5' – 3' direction (*Akbari et al, 2009., Svilar et al., 2011*).

There are conflicting arguments around what is most important for cell survival. In the immediate it would be preferable to correct the mutation with the highest fidelity, but in practice the host repair machinery may be overworked and not have the capacity to do so. Slow and stalling DNA replication is clearly non-optimal, so a sub-set of low fidelity TLS polymerases with the capability to move beyond lesions is desirable. Furthermore, large scale corrective changes may be required, depending on the type of mutation, and the associated change can in fact lead to hugely disruptive dramatic changes to DNA sequence, particularly at double strand breaks. Therefore, in certain instances it may be the lesser of two evils to allow single point mutations to occur and subsequently propagate.

7.1.2 TLS DNA Polymerases and the SOS response

At their most basic, DNA polymerases (DNApol) are the enzymes that drive DNA synthesis through addition of free deoxyribonucleotides to the 3' end of the daughter strand which is newly formed in the 5' – 3' direction, guided by the corresponding Watson-Crick base pair on the adjacent template strain. DNA polymerase itself moves along the template strand in the 3'–5' direction

In practice DNA polymerase activity requires numerous other cofactor enzymes to prepare the DNA prior for replication, namely DNA helicases and topoisomerase to unwind the DNA template, along with formation of the DNA clamp inclusive of loading proteins which encourages association at the replication fork (*Tuteja and Tuteja, 2004*).

The structure of DNA polymerases are fairly well conserved, with a large degree of homology in terms of general quaternary structure and often described as giving the appearance of a right hand with finger, thumb and palm domains. In general, the palm domain is the most highly conserved across the

all families of DNA pol, forming a catalytic centre understood to associate or bid with the template DNA strand. The thumb is important for processivity (this being the speed at which nucleotides are added) (*Garcia-Diaz and Bebenek, 2007*).

Most significant in this context is the finger domain, which is believed to be responsible for binding to the correct nucleotide base and thus important to overall fidelity of the polymerase. Replicative DNA polymerases are by and large very accurate, but they do infrequently introduce mistakes. However, some possess a proofreading capability that enables them to excise incorrect bases through a 3'-5' exonuclease activity but in order for this to occur, the finger domain must envelope the template strand and incoming nucleotide and discriminate between aberrant mismatches. Incorrect base pairing results in a conformational change to the DNA structure and causes the DNA pol + template complex to be repositioned so that the daughter strand is in contact with the exonuclease domain. Unless the induced fit is in perfect confirmation polymerase activity cannot continue. Thus, the 'finger' seems like an obvious starting point when making comparisons between the structure of high and low-fidelity polymerases, something which will be explored in greater detail in this section (*Khare and Eckhert, 2002., da Silva and Reha-Krantz, 2007*).

A particular class of DNA polymerases are known to tolerate errors within dsDNA and can insert nucleotide bases where there is no guide or complementary base (most likely an abasic AP site), or where the guide is incorrect. These are termed translesion by-pass polymerase (TLS) and are categorised as the Y-family polymerases (*described in **Table 5.1***). Many of these proteins have been directly linked to error-prone repair as a directed consequence of the SOS response. This is particularly apparent in the control of umuDC operon which is inhibited by lexA binding to its operator and is therefore controlled by a RecA SOS response. A detailed comparison between the numerous error-prone TLS polymerases in prokaryotes and eukaryotes is provided in the comprehensive review by Goodman and colleagues (*Goodman et al, 2002*). A more in-depth description of the known structural similarities and difference of Y family polymerase across a bacteria, archaea and eukaryotes are reviewed (*Pata, 2010*).

Table 7.1 *– Description of the B and Y family of DNA polymerases particularly relevant to repair in Eukaryotes, with comparative polymerases in prokaryotes (E.coli) provided.*

Family	Eukaryotic Polymerase	Proof Reading	Accuracy	Processivity	Function	Prokaryotic Equivalent
B	Pol α (alpha)	No	Medium	Medium	Replication, Primase	
	Pol δ (delta)	Yes	High	High	Replication	
	Pol ε (epsilon)	Yes	High	High	Replication, Repair	
	Pol ζ (zeta)	No	Medium / Low	Low	TLS, SHM	
Y	Pol η (eta)	No	Low	Low	TLS, SHM	Pol V (UmuCD)
	Pol ι (iota)	No	Low	Low	TLS, SHM	Pol IV (DinB)
	Pol κ (kappa)	No	Low	Low	TLS, SHM	
	Rev1	No	Low	Low	TLS	Pol R1

7.1.3 Understanding the link between AID Deamination and DNA Repair

The relationship between AID activity and the hypermutation of the Ig genes in eukaryotes is of key importance to this project and has been explored in a number of review articles (*Diaz et al, 2005*). AID is able to introduce complete genetic diversity at targeted hypervariable receptor regions and mutated both G:C and A:T pairs. This is striking given that AID enzymatic activity exclusively deaminates cytosine nucleotides. The suggestion is that AID activity must be strongly associated with error-prone DNA synthesis and highlights a link with downstream TLS polymerase activity.

There is a great deal of data which justifies these claims. Firstly, AID deficient B-cells lack both sets of mutations in the Ig hypervariable region (*Muramatsu et al, 2000*). Furthermore, over-expression of AID in non-B cells only leads to mutations at G: C pairs (*Yoshikawa et al, 2002*).

As the expression of AID is exclusive to eukaryotic organisms, its role in somatic hypermutation (SHM) has been explored largely within mammalian cell lines. Work by Diaz and colleagues explored the role of TLS DNA polymerases and identifies the eukaryotic B- and Y- families as being essential for effective SHM due to a number of factors: i) an unattainable mutation frequency and diversity that would not occur otherwise, ii) base substitutions are predominant over deletions or additions, iii) high probability of transition mutation events (A-G or C-T) compared to transversions.

Subsequent studies have been carried out to assess the mutational spectrum seen in Ig regions when various TLS genes have been knocked down. For example, a knockdown of the B-family DNA polymerase ζ (*zeta*), which lacks 3’ – 5’ exonuclease and proofreading activity, resulted in a mouse model with reduced mutation frequency (*Zan et al 2001, Diaz et al 2001*). Another clear link can be seen in human case studies of the genetic disorder Xeroderma pigmentosaum (XP) in which damage caused by ultraviolet light is amplified, due to a lack of efficient DNA repair. One particular variant of the disease (XP Type V) actually relates to a mutation in the gene which codes for the TLS DNA polymerase

η (*eta*), and thus the ability to replicate DNA over the site of the damage is lost. These individuals are known to have a reduced mutation frequency in their Ig V regions in regards to changes which can occur at A:T base pairs, once again suggesting an obvious link to TLS replication (*Zeng et al, 2004*). Another investigation explored the consequence of knocking out DNA polymerase ι (*iota*) in human B cells, and showed that G:C pair mutations were only seen when AID and DNA polymerase ι were present, although this was not replicable in a mouse model (*Faili et al, 2002*)

Mouse model knockouts demonstrating a reduced level of expression for various X-family DNA polymerases μ (*mu*), λ (*lambda*), β (*beta*) and κ (*kappa*) did not show any unusual SHM phenotype or mutation spectrum. Thus, much of the current literature suggests that Y-family TLS DNA polymerases have a beneficial association with SHM because they have a tendency to introduce base substitutions rather than the insertions or deletions that would cause negative frame shift mutations, something which does not favour a gradual evolutionary modification of antibodies. Single substitutions events would be highly desirable in our evolution approach because this would introduce potentially useful amino acid residue changes yet eliminate the chance of catastrophic frame shift changes.

As explained earlier in this report, deamination events as a result of AID activity will inevitably lead to abasic lesions on dsDNA as they will be processed by Uracil DNA glycosylase. There is some evidence to suggest that TLS DNA polymerases show specificity for these abasic sites and are somehow enlisted to bypass the lesion as part of the repair mechanism. A previous study was able to create a mutator phenotype in yeast that showed a high number of abasic sites, by increasing expression of DNA glycosylase above the level of apurinic DNA endonuclease. The author demonstrated that DNA polymerase ζ (*zeta*) is required for extension after a cytosine base had been inserted opposite the abasic site by deoycytidyl transferase (Rev1/TDT) (*Glassner et al, 1998*). By blocking the TDT and DNA polymerase ζ, spontaneous mutations were reduced. The use of TDT in my proposed system may have some value given the fact that it adds dNTPs at any 3' end overhang without the need of a template. This was considered as an alternative approach as the creation of a high number of abasic sites within the target gene may be enough to increase association with TDT which has the potential to introduce single base changes and whilst avoiding strand regression. It is a much smaller protein that could be used to create an exonuclease-TDT fusion protein, which ensures it is recruited to the sites of interest. However, there is no immediate way to disconnect TDT activity from introducing off target mutations elsewhere in the host genome.

A completely alternative model for repair of the U:G mismatch pair, created following AID activity, still exists. Whilst the BER repair pathway may seem the most likely candidate for the repair of mutational events present in non-replicative DNA strands, there are increasing suggestions that mismatch repair (MMR) proteins are preferentially recruited to the site of the lesions and can somehow

interact with TLS polymerases to promote error-prone repair (*Diaz et al, 2005*). Such a model would advocate a direct interaction between the TLS polymerase and MMR proteins which would require a protein co-factor able to encourage recruitment of both proteins to the site of the lesion. AID, DNA polymerase ζ, proliferating cell nuclear antigen (PCNA) or an altogether unknown protein have been hypothesised as potentially fulfilling this role of co-factor.

7.2 Justification and Proposal

My focus was to explore error-prone repair in the prokaryotic *E.coli* host and therefore I intended to express low fidelity error-prone TLS polymerases which can overcome abasic sites and introduce random nucleotide bases whilst replication happens across the specific lesion. Essentially this means that there is a high probability of mutation at the initial site of deamination.

Y-family polymerases are identifiable as low-fidelity by virtue of the fact that they do not contain a specific 3'-5' exonuclease domain pertaining proofreading capabilities. Up to a 100-fold increase in mutation rate has been identified when comparing Y-polymerase activity against the A, B, and X families (*Kunkel, 2009*). In prokaryotes, both Pol IV and Pol V can by-pass a variety of lesions and are induced as part of the SOS response of the cell (*Kuban et al, 2006*).

7.2.1 Polymerase IV

Pol IV is a promising candidate for the proposed error-prone repair system, given its ability to target abasic sites and introduce all nucleotide bases. Therefore, I decided to take a deeper look into the literature surrounding its activity and understand how it could potentially be engineered fit for purpose.

Early work demonstrated that over expression of the *dinB* gene, later revealed to code for Pol IV, increased the likelihood of random -1 frame-shift mutations (*Kim et al, 1997*). Subsequently, the mutator phenotype, seen during the SOS response in *E.coli*, was linked to chromosomal and episomal overexpression of DinB (*Kim et al, 2001*).

However, it was only when the protein was purified and tested *in vitro*, that Pol IV was revealed as an individual polymerase, separate from Pol III and Pol V. This work was then continued, and numerous site saturation mutations were performed to specifically understand the protein domains responsible for this mutator affect when expressed *in vivo* (*Wagner et al, 1999*).

Pol IV is present at high intracellular concentrations (~250 molecules per cell), far more than would be needed at all active replication forks in the cell at any one time (*Kim, 2001*). Chromosomal Pol IV (DinB) expression levels are 6 – 12 times higher than that of other associated SOS-inducible error-prone polymerases (Pol II and Pol V) and is far higher abundance than the main replicative polymerase Pol III (*McHenry and Kornberg, 1977*). It is believed that under stress conditions, Pol IV contributes to

untargeted mutagenesis, affecting DNA which is undamaged and its expression may be up-regulated by other stresses such as those often encountered during late stationary phase, controlled by an SOS response or the stress-induced sigma factor RpoS (*Galhardo et al, 2009.*, *Layton and Foster, 2003*).

Pol IV function has now been investigated in a number of expression studies. In fact, an early investigation into the role that the main replicative DNA polymerase, Pol III, plays in spontaneous mutation attempted to mutate the *dnaE* gene, which expresses the alpha subunit of DNA pol III holoenzyme (*Strauss et al, 2000*). This actually revealed that Pol IV must play a role in spontaneous mutations within *E.coli* as it quickly became apparent that a *dnaE +/dinB+* strain gave a far higher level of substitutions and frame shift mutations compared to a *dnaE+/dinB-* knockout strain.

The association between Pol III and Pol IV activity has been studied further both *in vitro* and *in vivo* and reveals that Pol IV is highly dependent on its association with the β-clamp, the beta subunit of the DNA polymerase III in prokaryotes (*Heltzel et al, 2012*). A five-residue region located at the C-terminus of Pol IV is essential for this interaction, specifically residues 346-QLVLGL-351 that form a consensus β-clamp binding motif (*Lenne-Samuel, 2002*). This interaction tethers the polymerase to the clamp via a C-terminal peptide that forms the little finger (LF) domain of Pol IV (*Bunting, 2003*). Secondary structure predictions suggest that this LF domain is common throughout all Y-polymerases, although specific differences likely represent adaptations for overcoming target lesions (*Ling, et al 2001*). Therefore, the LF domain could become a useful target for creating a mutant polymerase that is able to overcome basic sites, although it is hoped that Pol IV is already able to do so.

Recent understanding of the association between Pol IV and the β-clamp has been widely discussed (*Hetzel et al, 2012*). Most recently, Pol III and Pol IV interaction has been explored by overexpression of DinB under an arabinose promoter. This leads to an immediate stallage of the replication fork at random genetic positions and a decrease in cell fitness, which was identified as a reduction in colony-forming ability (*Uchida et al, 2008*). The author suggests that DinB targets Pol III and acts as a brake at replication folk, which may only operate when the cell is under stress conditions, and in fact this is likely to be beneficial way of stalling Pol III processivity in order to create a checkpoint for subsequent repair, which may in fact be by Pol IV itself (*Mori et al, 2012*). Although overexpression of Pol IV in this work seems to abolish DNA replication almost immediately after induction, with the potential to become lethal to the cell, quantification of Pol IV concentration suggests it was expressed at 72-fold higher levels than in an SOS-induced cell. It is logical to think that beneficial Pol IV activity is entirely concentration dependent, and needless overexpression will cause complete stallage of Pol III, impeding DNA synthesis and normal DNA replication. The author suggests that DNA replication is completely blocked when the level of DinB reaches about 15-fold above SOS-induced cells. Mutants of the β-clamp binding region and polymerase activity domain revealed that the polymerase activity of

DinB has no effect on inhibiting the movement of Pol III but the β-clamp binding domain is partially required for inhibition. It is important to note that even when a single copy of *dinB* gene was incorporated into the chromosome to reduce rate of overproduction, an estimated 1.8-fold excess of Pol IV (compared to normal SOS levels) could still inhibit DNA synthesis. This work serves to highlight the importance of disconnecting Pol IV interaction with the host Pol III replicative machinery when expressing it within my own system and underlines the importance of tight regulatory control.

The specific type of DNA damage bypassed by Pol IV was revealed by extensive repair deficient mutants and mutation specific knockout strains (*Bjedov et al, 2007*). It was hoped that one knockout pathway corresponding to one particular type of mutation, would show a significant peak in background mutations. This suggested that Pol IV is involved in tolerance of cytotoxic alkylating DNA lesions *in vivo* (*alkA* and *tag* gene deficient knockouts). It should be noted that these errors can be checked and repaired by MMR mechanisms as previously described (*Strauss et al, 2000*). Therefore, all strains were MMR deficient (by knocking out *mutS* gene) and high numbers of abasic sites were maintained by inactivating exonuclease III (*xth* gene) and endonuclease IV (*nfo* gene), which would otherwise nick the backbone and remove the abasic site.

The complete crystal structure of Pol IV remains elusive, but some assumptions can be made, by exploring the structure of DinB homolog Dbh, from the archaea *Sulfolobus solfataricus* (*Zhou et al, 2001*). The structure is very similar to the catalytic domain of other polymerase families and suggests minimal contact between the polymerase protein and DNA (PDB file, 1IM4: Fingers 20-75, Palm 1-19 and 76-169, Thumb 170-205). Importantly, the nascent base pair binding pocket is quite accessible which would suggest that the protein is only capable of low fidelity and low processivity. Whilst Dbh only shares 38% sequence similarity with *E.col*i DinB*,* the model does show that the fingers of the protein appear to be in a closed conformation, suggesting the 'induced fit' mechanism that normally ensures high fidelity in DNA polymerases cannot actually be active. Interestingly, the author shows that this specific *Dbh* polymerase is able to by-pass abasic sites.

Further work, published around the same time explored a different DinB ortholog from the same species call *Dpo4* which shared limited sequence similarity (*Ling et al, 2001*). Crystal structures revealed ternary complexes in which the polymerase interacts with the target DNA strand and an incoming nucleotide (PDB files, 1JX4, 1JXL). It also confirmed that *Dpo4* contains a conserved active site and a characteristic hand-like configuration. However, the limited and nonspecific contact with the replicating base pair once again permits a relaxed base selection and low fidelity replication. The author also alludes to the fact that the crystals were captured translocating two nucleotides simultaneously, suggesting a mechanism for error-prone lesion by-pass synthesis which differs from that of DinB.

A later study was able to generate a crystal structure that confirmed the presence of a novel structural domain specific to Y-family DNA polymerases (*Bunting et al, 2003*). Termed the 'little finger' (LF), this domain in *E.coli* Pol IV is shown to associate with the *E.coli* β-clamp processivity factor. Most importantly, this is specifically the *E.coli* polymerase DinB, and can be considered the most useful structural information available. Although the main interaction between the C-terminal domain (residues 340-351) of Pol IV LF and hydrophobic channel on the surface of the β-clamp is identifiable even in non-processive mutants, a secondary and essential protein–protein interface exists. There appears to be an interaction between the surface loops of Pol IV LF core and the edge of two β-clamp ring subunits. In this state, the polymerase is maintained in an inactive orientation, which would imply that there is a mechanism for switching between replicative polymerase and the lesion bypass polymerases. Tight regulation of the lesion bypass polymerase activity would be required because of their inherent low fidelity, and the belief is that the β-clamp is largely the regulator. This model actually shows that when Pol IV is bound to the clamp, it would not have access to the DNA template which actually lies between the clamp, and only disruption of this secondary protein-protein interaction could alter the confirmation sufficiently to allow access. The author hypothesises that the secondary protein-protein interaction provides a stable state at which DNA is unbound, preventing Pol IV from outcompeting high-fidelity polymerases at the replication fork.

It seems very likely that the LF domain does in some way provide a lesion by-pass function since it is only found in the Y-family polymerases. Furthermore, mutations targeted to the same area of the LF domain, in both Pol IV and Pol V seem to impair lesion bypass activity (*Bunting et al, 2003., Woodgate et al, 1994., Boudsocq et al, 2002*).

Considering the substantial evidence to suggest that Pol IV is responsible for introducing multiple mutations as a consequence of its polymerase activity, it appears to be a potential candidate for our system. Pol IV could in theory introduce any of the four nucleotides when overcoming Pol III stallage in response to abasic site formation. Further evaluation of protein structure may also reveal specific domain regions that encourage lesion binding and further reduce fidelity, with a distinct shift from normal high-fidelity repair towards error-prone lesion bypass DNA synthesis.

7.2.2 Pol V

Pol V consists of two subunits which are both required for polymerase activity, umuC and umuD. Importantly, umuD requires the assembly of an active RecA filament which is able to bind to ssDNA at stalled points of replication (*Opperman et al, 1999., Jiang et al, 2009*). The formation of RecA filaments results in the release of the LexA inhibitor from a number of regulatory operator sites including those within the *umuCD* operon. Importantly, the RecA – ssDNA complex has been shown to cause a conformational change to umuD protein domain, resulting in an altered form umuD'. A subsequent

heterodimeric interaction between umuD and umuD' allows for association with umuC. This results in an umuD'$_2$C complex equates to the active form of Pol V. Given the highly mutagenic activity of Pol V, which is also capable of carrying out translesion synthesis, this multi stage regulation helps avoid unwanted error-prone repair. It is also important to note that active umuD' heterodimer is targeted for destruction by ClpXp protease and subsequent removal (*Wood, 1996*).

They are a number of studies which detail UmuC's catalytic activity and attempt to characterise its substrate lesions and subsequent repair effect. A comprehensive study by Tang and colleagues details the speed, processivity and specific capability of the active umuD'$_2$C complex to bypass abasic sites (*Tang et al, 2000*).

In direct reference to my proposed system for directed evolution, I once again refer to the promising method of Phage Assisted Continuous Evolution (PACE) (*Esvelt et al, 2011*). This approach has been shown to be capable of creating vast sequence diversity of a target gene through expression of both Pol V subunits umuD' and umuC under an arabinose promoter along with co-expression of a RecA mutant (recA730). This mutant displayed an enhanced single stranded binding capacity and bypasses the need for RecBCD enzyme normally required to load the RecA protein onto the ssDNA strand. The recA730 mutant, also called RecAΔC17/E38K double mutant (*Eggler et al, 2003*) can activate Pol V with approximately 2-fold higher efficiency than wildtype protein (*Jiang et al, 2009*). This is because the truncated protein has a less negative charge due to a removal of the last 17 negatively charged residues from the C-terminus.

Although the PACE strategy was extremely successful in terms of the sequence diversity that was achievable and would suggest the promise of expressing Pol V within my own directed evolution system, it is important to highlight the use of another mutator protein, DNAQ926. This is in fact a mutant of the dnaQ gene that encodes the proofreading 3' exonuclease subunit of DNA pol III holoenzyme (*Figalkowska et al, 1996*). This mutant confers a strong, dominant mutator phenotype likely due to two reasons. The first being the lack of a proofreading domain and the second being that Pol III is responsibility for the majority of replicative DNA synthesis within the cell and thus the overexpress mutant will likely outcompete the wildtype DNA Pol III. Therefore, this may be the true overriding cause of increased mutations as a consequence of error-prone DNA synthesis. Given that the phage assisted system only requires short periods of *E.coli* cell growth, an extreme mutator phenotype could briefly be tolerated.

In terms of utilising Pol V to increase the mutational spectrum my system can achieve, it may also be worth considering the use of the G.23.P mutant of UmuD which is known to be cleaved more readily and may ensure that overexpressing of UmuC and UmuD remains viable within the host strain. Given that Pol V requires a complex interaction between multiple substrates, the potential for creating a

viable fusion protein seems unlikely. The alternative would be to regulate overexpression of the UmuD mutant and encourage the active Pol V complex to target sites through its interaction with the RecA complex formed at the abasic site in response to stalled replication.

7.2.3 Exonuclease III

Exonuclease III (Exo III), coded for by the *xth* gene in *E.coli*, is recognised as the main AP endonuclease in the BER pathway and is particularly relevant for the removal of abasic sites which are a left as a direct consequence of removing Uracil lesions by activity of UDG.

A previous investigation identified the importance of Exo III in BER by creating mutants of unable to express deoxyuridine triphosphate (dut-), leading to an increase in uracil substitution if thymine cannot be readily sourced during DNA replication (*Taylor & Weiss, 1982*). Subsequent removal of uracil by UNG produced the apyrididinic abasic sites at which Exo III is known to have high levels of endonucleolytic activity. Double *dut/xth* mutants were lethal unless there was a mutation in the *ung* gene, meaning uracil was not removed from DNA in the first instance, and thus repair was likely directed along an alternative pathway. This is a clear indication that when initial glycosylase activity is allowed to happen, Exo III is required downstream and therefore must be recruited to the abasic site for its removal.

The crystal structure of Exo III (*PDB file, 1AKO*) also reveals clear similarities with bacterial endonucleases such as DNA Nuclease I. The large number of positive residues within its active site confirms that phosphate-bond cleavage at AP sites occurs via a nucleophilic attack (*Mol et al, 1995*).

It should also be noted that Endonuclease IV is an AP endonuclease responsible for about 10% activity in *E.coli*. Furthermore, Endonuclease V is also recognised as cleaving Uracil containing DNA. Both of these endonucleases are potential candidates to explore in later iterations.

7.2.4 DNA Pol I

DNA pol I is a responsible for a large majority of DNA repair within prokaryotes. It contains 5’ – 3’ DNA dependent polymerase activity, 3 – 5’ exonuclease activity required for proofreading and a 5’ – 3’ exonuclease domain which is involved in nick translation during DNA repair. It is hypothesised that this domain could be added to alternative polymerase domains (e.g. Pol IV) which are known to show lower fidelity. Therefore, it could possible to create a fusion protein with added nick-translation activity to a low fidelity polymerase which is effective at removing bases ahead of the site of insertion with erroneous base addition.

7.3 Proposal and Experimental Strategy

I intended to directly link the error-prone activity of TLS polymerases to specific abasic sites following targeted AID activity of the ATF mutator device by creating a large fusion protein. These abasic sites have the potential to become 'points of entry' for DNA repair.

As previously mentioned, overexpression of AID during active transcription leads to deamination events and this activity can be targeted to a gene of interest by a fusing AID to an orthogonal T7 RNA polymerase. It transpires that cytosine bases throughout the full length of the target gene may become deaminated to uracil. This modality will be marked for BER via glycosylase activity of UDG and the errant base is removed to leave an abasic site lesion. In its previous form, my system expresses and inhibitor of UDG (UGI) to prevent BER from taking place.

This iteration of my design eliminates the need for expressing UGI and seek to hijack BER mechanisms and encourage error-prone repair. At this point in the BER pathway an AP endonuclease (e.g. Exo III), is able to cleave the DNA backbone leaving a 3' hydroxyl group that can be extended by a DNA polymerase to fill the single nucleotide gap, prior to nick ligation. As outlined above, the activity of translesion bypass polymerases is likely to be essential in order for replication at such non-guided positions. Therefore, over-expression of Pol IV or Pol V may encourage low fidelity synthesis at these specific target sites and ultimately increase the genetic diversity seen within the target gene.

The simplest and most direct method for introducing error-prone repair will be to express the DinB (Pol IV) and UmuCD (Pol V) genes on the same plasmid that encodes our AID-T7 fusion protein, whilst excluding the UGI expression plasmid. In addition, strategies to localise the TLS polymerases specifically to sites of BER have been considered to attempt to avoid off target error-prone repair and to limit global background mutations that could weaken host strain viability.

Firstly, the aim of disassociating error-prone TLS polymerases from the innate repair could be achieved by engineering Pol IV to remove the C-terminus (residues 346-351) and disrupt its interaction with the β-clamp of Pol III (*Lenne-Samuel 2002, Bunting 2003*), or by removing a larger C-terminal region to preventing Pol IV access at 'normal' sites of active replication (*Bunting et al, 2003*). This may provide a convenient mechanism to abrogate background mutations if the truncate protein remains active at directed site of interest whilst remaining unassociated with host *E.coli* repair pathways.

There is further scope for create fusion proteins in an effort to guide TLS polymerase activity to specific abasic lesion sites. Since Exo III activity defines the intermediate step linking strand incision and the recruitment of TLS DNA polymerases, a protein fusion consisting of Exo III and Pol IV may provide the spatial and temporal localisation of both enzymes required for error-prone BER.

A further advantage of using exonucleases that possess 3' - 5' activity is that they may continue regression of the target strand, even removing non-target bases. Subsequent activity of the TLS

polymerase may then create a patch repair effect. Given that many high-fidelity polymerases contain a 3'-5' exonuclease activity that pertains a proofreading capability, I will explore the application of the 3'-5' exonuclease domain of DNA Pol I, fused to Pol IV polymerase.

Furthermore, DNA pol I is linked to nick translation during DNA repair due to its 5'-3' exonuclease activity which may be useful for introducing more pronounced changes. The theory here is that the exonuclease removes the nucleotide ahead of the polymerase as it progresses in the 5' – 3' direction. If said polymerase domain happens to be low fidelity, then the chances of error-prone repair are increased.

DNA pol I will also be used to inform my design, given that this large polymerase displays 5' – 3' polymerase activity. When creating a random linker library for building exonuclease – polymerase (EXO-Pol) fusion proteins, I will replicate the relevant stretch of residues in DNA Pol I, as this is known to tolerate correct folding and activity of both functional domains. In fact, these two distinct enzymatic activities co-exist during high fidelity DNA replication and can remove and add new nucleotides simultaneously. The lifetime of ssDNA gaps is short to avoid prolonged exposure to attack by Exo III which may direct repair along a more error-prone repair pathway (*Setlow et al, 1972*).

The final spectrum of mutations that might be achieved post error-prone repair were questionable, given the possibility of patch-repair i.e. a nucleotide patch consisting of more than one base could be excised and replaced, encouraging significant changes to the protein coding sequence. However, this expands the spectrum of mutants even further and could be advantageous when searching for novel protein function.

Ultimately, it was decided that Pol IV would be first investigated for its mutagenic TLS capabilities within my proposed system with a summary of my choices provided below (***Table 7.2***).

My decision was heavily influenced by the exploration into the literature and this is comprehensively summarised in the excellent comparisons between Pol IV and Pol V made by Tang and colleagues (*Tang et al, 2000*). In this work they consider response speeds, processivity, bypass rates and activity at various different lesions types.

Table 7.2 *– Comparison of the Pol IV and Pol V and their applicability to this work*

Pol IV	**Pol V**
Both Pol IV and Pol V can by-pass abasic sites.	
Both Pol IV and Pol V show low processivity and dissociate after 6 – 8 nucleotides are added.	
Pol IV is 5 – 10 times more accurate on undamaged DNA	
Pol IV is a single domain, with many known mutants, particularly the C-terminal deletion that can disconnect it from normal DNA replication.	Pol V consists of multiple protein domains that are likely unsuitable for creating fusion proteins
Pol IV does require association with Pol III beta-clamp which is useful given that it will stall at the abasic lesion and offers a point of entry.	Pol V requires RecA ssDNA to activate of UmuDC complex so therefore no way to disconnect it from normal repair pathways.

7.4 Methodology

7.4.1 Cloning Strategy

Initially I decided to observe the effect of over-expressing a number of the repair linked proteins previously described, namely Pol IV, Pol V and Exo III. Firstly, Pol II (*dinA*), Pol IV (*dinB*) and Pol V (*umuCD*) gene sequences were identified in the wildtype K12 MG1655 *E.coli* genome using the NCBI reference genome (http://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521) and relevant primers were designed (***Appendices - Table 11.24*** - Primers used to amplify target genes of interest from the wildtype MG1665 *E.coli* strain) to amplify the coding sequences without regulatory 5' UTR regions. PCR amplification required total *E.coli* genomic DNA prepared using phenol/chloroform extraction (***Appendices 9.1.1***). Blunt products were subsequently cloned into pJET 1.2, prior to sequencing. Sequence reads genes were verified using BLAST alignment against the target gene. These intermediate vectors were then retained for use downstream in subsequent PCR steps.

7.4.2 BASIC part preparation

Parallel efforts to establish the novel BASIC assembly technique were on going (***Chapter 3***). Therefore, significant effort was made to standardise all parts into the BASIC format in order to streamline multi-part combinatorial assembly downstream. A range of functional parts were prepared by amplifying the target genes, with addition of BASIC compatible prefix and suffix sequences to PCR

primers (***Appendices*** - ***Table 11.25***). Various modifications were made to the genes of interest through addition of 5' or 3' end flanking sequences added to primer design. This included the addition of characterised RBS sequences or the removal of stop/start codons for building fusion proteins. All new PCR products were cloned back into the pJET storage vector ready for use with BASIC assembly downstream.

As outlined earlier, the C terminus of the Pol IV protein has been heavily associated with the β-clamp of DNA pol III responsible for normal DNA replication. Theoretically, over expression of Pol IV could lead to a significant increase in global mutation. Therefore, I prepared Pol IV variants lacking the last 5 amino acids and the last 12 amino acids within CDS. These truncated variations of the protein were expected to prevent interaction with β-clamp.

7.4.3 Fusion Linker Creation

The general basis for fusion linker was to ensure that both Exo and Pol domains retained their functionality, whilst avoiding any potential interaction between the two. As previously described, a number of bespoke BASIC compatible linkers were designed with the intention of creating fusion proteins. All design constraints were considered; melting temperature, overlap length, codon usage and the avoidance of non-specific assembly by eliminating homologous sequences. Ultimately, they were designed so that any two individual domains could be retained separately as BASIC compatible parts, and then assembled together to form a fusion protein (***Chapter 3.2.5***).

A rigid alpha helical linker was designed, which is likely to be very stable and can form intra segment hydrogen bonds that result in a closely packed backbone. It is believed the rapid nature of this association and rigid spacer that is formed will ensure both protein domains fold correctly and have little interference with each other (*Chen et al, 2013*).

Furthermore, a flexible glycine serine repeat linker was built. These small, polar amino acids are also likely to avoid interaction with neighbouring protein domains because they readily form hydrogen bonds with water, adding to their stability.

Given the target application in question, I also designed a functional linker which was influenced by the structure of DNA Pol I, given the close association of both exonuclease and polymerase domains within this protein. Exploration into the literature did not reveal a complete structure for DNA Pol I but descriptions of the Klenow fragment reveal accurate locations for the functional domains and specific active sites. Proteolysis results in the removal of the 5'-3' Exo domain for DNA Pol I and thus models of the Klenow fragment depict and N-terminal flap 'flap' which provides a clue to the linker sequence between the smaller 5'-3' domain, which is lost, and the rest of the polymerase. This resulted in a 51bp/17aa linker which consists of the residues 'LLESPKALEEAPWPPPE'. This sequence is clearly proline

rich which is interesting given that is distinct for other amino acids in that its cyclic side chain has a very restricted conformation and lacks an amide hydrogen, meaning that it cannot act as hydrogen donor which may limit hydrogen bond formation with other residues (*Williamson 1994., Chen et al, 2013*). Therefore, this is predicted to be a rigid linker that has limited interaction within the neighbouring functional domains and it appears to be biological relevant. It should be noted that many naturally existing multi domain proteins often contain proline residues (*Motram et al, 1989., Turner et al, 1993*).

In reality, the optimal linker will provide enough rigidity to prevent interaction of both Exo and Pol domains yet ensure there is flexibility, so that each domain can each interact with the target DNA strand at the correct angle. The small set of linker variations FL1 (flexible), FL2 (rigid), FL3 (natural) will be screened for viability in this context, but their general application had already been shown in a proof of principle study (***Chapter 3.3.5***).

7.4.4 Construction - Stage 1 – Assembly of Fusion Protein Variants

I intended to explore the effects of over expressing Exo III, 5'-3' Exo and Pol IV proteins within the target GM31 host *E.coli* strain. Therefore, numerous constructs were built using the BASIC parts described previously. A complete list of constructs is provided (***Appendices - Table 11.29***). Multiple variations of the Exo III - Pol IV fusion protein were made, consisting of variations of the fusion linker and expression was controlled under inducible and constitutive promoters. Control constructs included a functional operon which expressed both proteins from a single mRNA transcript, with proteins translated individually.

Interestingly, constructs expressing Pol IV could not be recovered and it seemed likely this known mutagen was lethal to the cell when placed on a multi-copy plasmid. This was immediately obvious in the diagnostic colony PCR (***Figure 7.1*** *and* ***Figure 7.2***). It should be noted that colonies which did form on these plates were later understood to contain frame shift mutations, which removed any possibility of a functioning Pol IV protein being expressed.

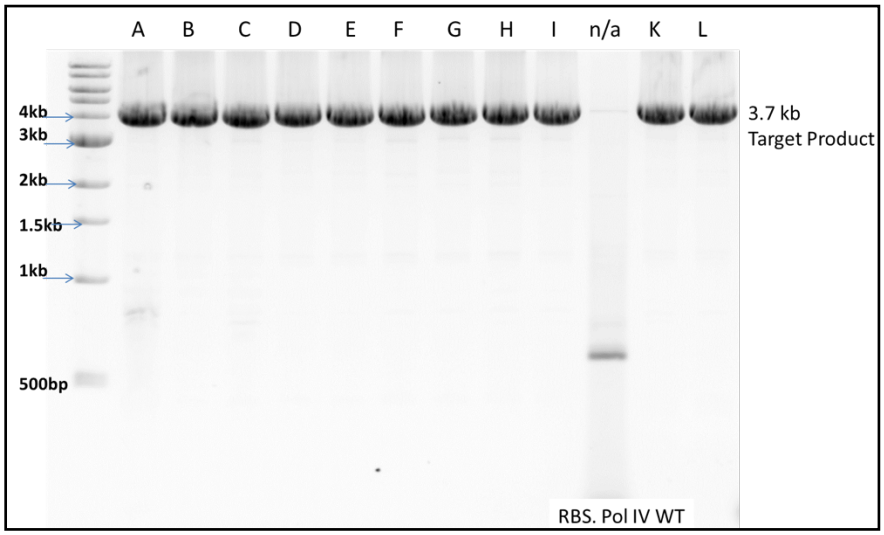


Figure 7.1 *– Diagnostic Gel for Exo III – Pol IV constructs. Colony PCR was carried out on assembled constructs, by picking colonies directly from selection plates. Primers used anneal to sequence which flank the complete fusion protein CDS. Individual expression of Pol IV in its wildtype form was lethal to the cell and is represented here by an empty lane (between lanes I and K). Please refer to* ***Table 11.29*** *for complete descriptions of all constructs.*

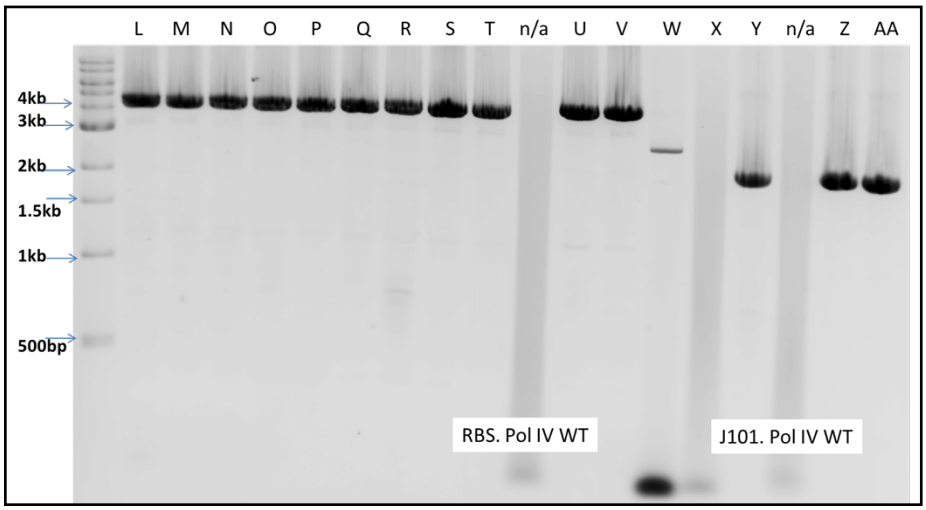


Figure 7.2 *– Diagnostic gel for 5' 3' Exo – Pol IV constructs after immediate amplification of selected colonies. Colony PCR was carried out on assembled constructs, by picking colonies directly from selection plates. Primers used anneal to sequence which flank the complete fusion protein CDS. Individual expression of Pol IV was lethal to the cell and is represented here by to empty lanes (between samples T and U and between Y and Z). Note that constructs W and X are control plasmids that did not contain the required PCR priming sites. Please refer to* ***Table 11.29*** *for complete descriptions of all constructs.*

7.4.5 FLAG Tag addition

To simplify visualisation of the fusion protein and confirm that it is being expressed in the host strains I decided to add a FLAG tag motif. Whole cell lysate was used for protein separation and compatible antibodies were used for subsequent Western blotting.

The FLAG tag sequence ('N-Asp-Tyr-Lys-Asp-Asp-Asp-Asp-Lys-C') which corresponds to DYKDDDDK is considered an optimal choice of peptide tag because it can be fused to either the N or C

terminus of the target proteins with strong downstream antibody targeting. Furthermore, new improved versions of the ANTI-FLAG antibody are also able to recognise internal loci of the target sequence which provides added flexibility of design. Importantly, the FLAG tag sequence contains more hydrophilic residues than other widely used epitope tags, which should reduce the effect on the function of the target protein. Ideally, these variants should retain the functional activity of the intended Exo-Pol IV fusion proteins, even with the additional residues.

The FLAG tag was added to the N-terminus of the Exo III BASIC part by using site-directed mutagenesis with 5' end phosphorylated primers (***Appendices - Table 11.26***). This resulted in N-FLAG-EXO-Pol IV fusion protein after assembly.

7.4.6 Protein Expression

Preparation of whole cell protein lysate from 50ml of each culture was immediately performed, and separation of proteins was performed on a 12% denaturing SDS-PAGE gel as described (***Chapter 9.9***). A Coomassie stain was used to visualise the anticipated band corresponding to the relevant fusion proteins. Even though Exo III and Pol IV have the potential to cause a mutagenic and lethal phenotype controlling expression under the tightly repressible Tet101 promoters did allow for normal growth prior to induction. Unfortunately, Coomassie stain visualisation on small polyacrylamide gels did not allow for observation of the target protein bands, given the relatively poor resolution of separation. However, the FLAG peptide sequence was useful for downstream visualisation by Western blotting using primary Anti-FLAG (1:3000) and secondary Anti-Mouse HRP conjugate (1:50000) antibodies using the chemiluminescent peroxidase assay. The target 72kDa target bands corresponding to the Exo III – Pol IV fusion protein variants were identified (***Figure 7.3***).

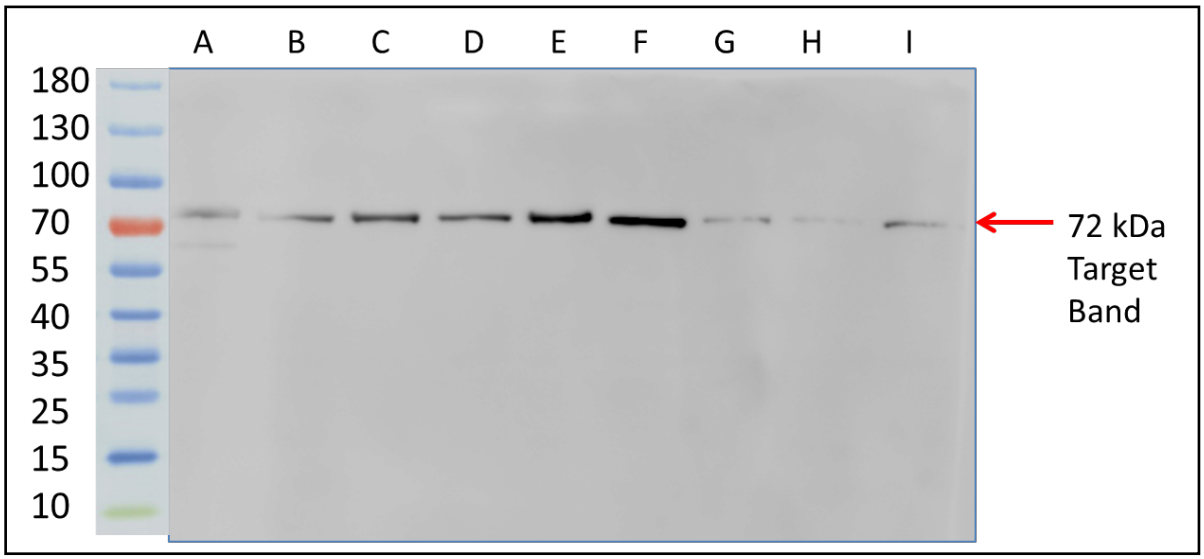


Figure 7.3 *- Western Blot of ExoIII – Pol IV fusion proteins (Primary Anti FLAG 1:3000, Secondary Anti Mouse 1:50000)*

7.4.7 Construction - Stage 2 - Hierarchical Assembly of Final Test Library

As described in my proposal, I wanted to explore the viability and activity of co-expressing EXOIII and Pol IV or indeed the EXO - Pol IV fusion protein alongside the known ATF mutator.

Given the size and complexity of said constructs, the final library was built using a hierarchical BASIC assembly. The initial EXO – Pol IV fusion protein variants were built and tested for expression and viability. The ATF device had previously been used to assess toxicity of over-expressing the mutator protein (***Chapter 6.2***). Both functional parts were retained thanks to the use of methylated linkers during BASIC assembly, which restored the prefix and suffix sequences, and meant that the Exo-Pol IV fusion and the ATF fusion could be used in the subsequent assembly as parts in their own right. The general process is outlined (***Figure 7.4***)

Once again, all constructs were validated using a diagnostic PCR. It is important to point out that given the size of the complete constructs, they were not sequenced. However, given previous results had showed that BASIC assembly has excellent accuracy and efficiency, the fact that each composite part had been sequence validate prior, provided enough confidence to progress immediately to the screening assay.

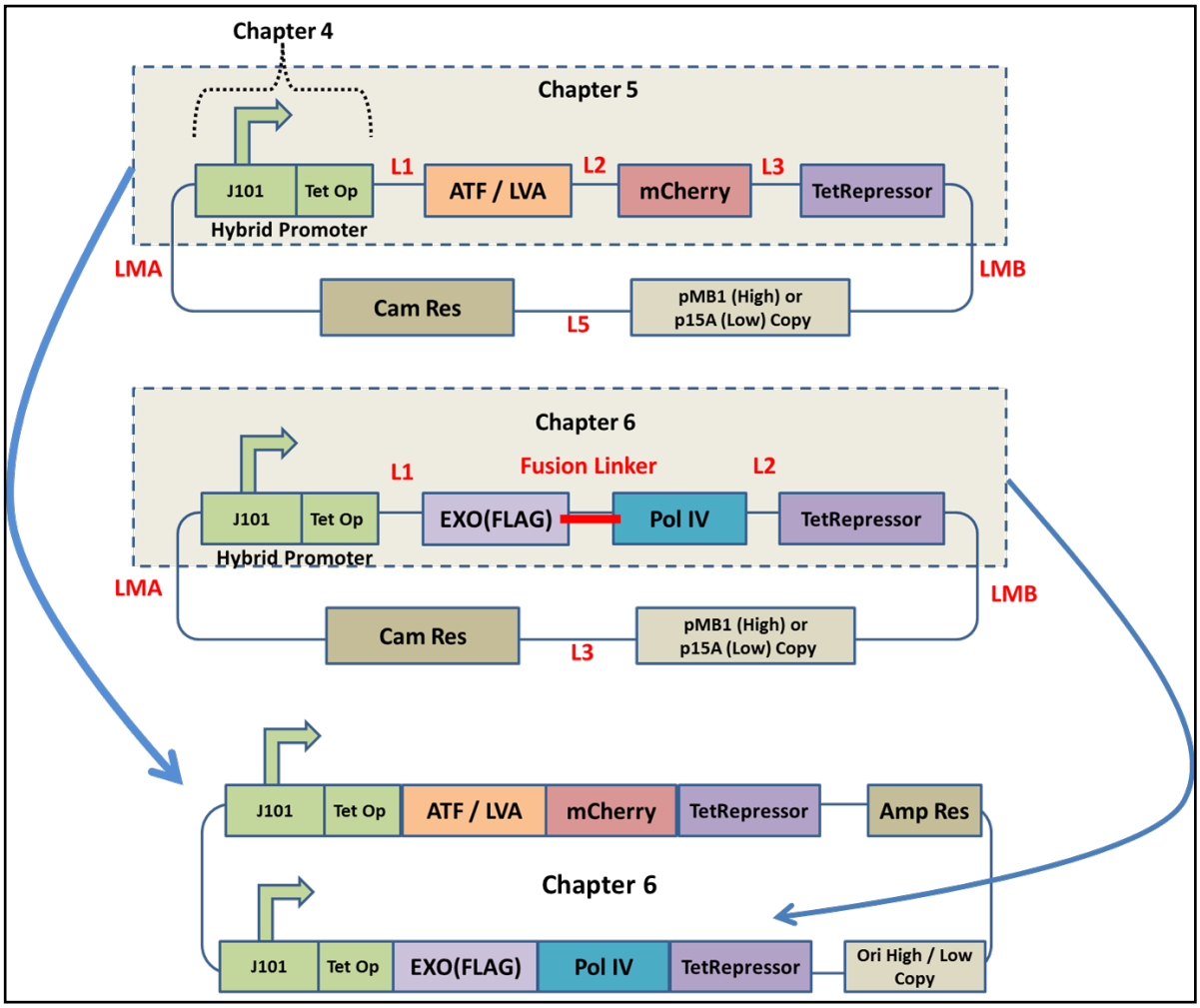


Figure 7.4 *– Hierarchical assembly of final library constructs using BASIC standard. A combination of parts including characterised inducible promoters (**Chapter 4**), known variants of the ATF mutator fusion proteins (**Chapter 5.4.2**) and alternative forms of the EXO-Pol IV fusion protein (**Chapter 7**) were used to build the final test library.*

Table 7.3 *– Summary of the constructs built to test the activity of variants of the Exo – Pol fusion proteins when co-expressed alongside the known ATF mutator fusion protein on a single expression vector, within E.coli GM31 host*

Construct	Exo III	Flexible Linker	Rigid Linker	Natural Linker	No Fusion	Pol IV WT	Pol IV -5	Pol IV -12
A	+	+				+		
B	+	+					+	
C	+	+						+
D	+		+			+		
E	+		+				+	
F	+		+					+
G	+			+		+		
H	+			+			+	
I	+			+				+
J	+				+		+	
K	+				+			+
Construct	**5' – 3' EXO**	**Flexible Linker**	**Rigid Linker**	**Natural Linker**	**No Fusion**	**Pol IV WT**	**Pol IV -5**	**Pol IV -12**
L	+	+				+		
M	+	+					+	
N	+	+						+
O	+		+			+		
P	+		+				+	
Q	+		+					+
R	+			+		+		
S	+			+			+	
T	+			+				+
U	+				+		+	
V	+				+			+
W	Empty Vector							

7.4.8 Cell Outgrowth

Growth conditions have been described previously (***Chapter 2.3.4***). Single transformant colonies were used to inoculate 96 well plates supplemented with SOC+ media and relevant antibiotics before being grown overnight at 30°C. Samples were then normalised to 0.2 OD600 and grown for a further 2 hours under the same conditions. A second normalisation step was carried out and grown for another 2 hours prior to induction. Induction by injection of aTc to working concentration (250 ng/ul) was performed manually. Samples were then allowed to grow for 6 hours post induction at 30°C.

7.4.9 Loss of Function Assay

In order to test the general activity of the newly built Exo – Pol library, I followed the same loss of function assay as previously described within the GM31 T7.GFP host strain (***Chapter 2.4.2***). However, there were some major differences. Firstly, the UGI expressing plasmid was eliminated from the system. Secondly, samples were grown for a single 8-hour period post induction. Once again, 50,000 events were counted for each sample prior to gating. Positive (T7RNApol) and negative (blank strain) controls were used to determine GFP+ and GFP- gates respectively. Therefore, the changing fluorescent populations were an initial insight into whether an improved mutator phenotype was likely.

7.5 Results and Discussion

7.5.1 Cell Fitness

It was important to observe the overall fitness of the cell and indeed the viability of expressing both ATF and Exo-Pol IV variants were co expressed with the host GM31 strain. Samples were grown as described above, with OD600nm measurements taken at 1-hour intervals.

It was immediately clear that all variants expressing the Pol IV WT showed an immediate drop off in growth post induction, again signalling lethality. Importantly, Pol IV -5c and -12c variants did not lead to a significant reduction in growth rate and it can therefore be assumed that the disconnect with the host replicative machinery avoided the destructive mutator phenotype.

7.5.2 Loss of Function Assay

Numerous variants of the Exo III-Pol IV proteins were trialled during co-expression with the ATF mutator fusion protein, within the target GM31 strain. As noted, there was no longer any inhibition of repair within this strain.

Initially, both Exo III and Pol IV proteins were expressed individually alongside ATF (***Figure 7.5***). It was immediately obvious that the positive (T7RNApol) and negative (blank vector) controls showed markedly different fluorescent profiles. Expression of the ATF mutator alone did once again show a reduction in fluorescence strength as a consequence of lower transcriptional strength. However, there was no evidence of a mutator phenotype when the ATF is expressed in the wildtype GM31 strain. The ATF mutator requires co expression of a UDG inhibitor to prevent BER in order to produce a mutator phenotype, as previously shown.

The Exo III -Pol IV variants were of particularly interest. Expression of the Pol IV resulted in a completely non-fluorescent population which was likely a consequence of a lethal phenotype (purple trace – ***Figure 7.5***). Upon expression of Pol IV, I would expect the majority of cells to face catastrophic

mutational events. Given the high count, the cells that do still survive are likely to have incurred a disruptive mutation in the Pol IV gene itself. In this case, removal of a functional Pol IV results ina growth advantage. Therefore, within this particular loss of function assay, sample expressing Pol IV appear to be false positives i.e. they appear as complete GFP- population.

Expression of ATF alongside the Pol IV C-terminal knockout provides some particularly exciting results. The common characteristic of an active mutator and true positive result transitional intermediate state, as identified for active mutator is seen for both Pol IV -5c and Pol IV -12c variants. Although the loss of function assay provides a fairly crude measurement of mutation rate, this qualitative data clearly shows that the GFP target gene has been largely disrupted upon expression of the ExoIII and Pol IV proteins, and way beyond what is achievable when expressing ATF.

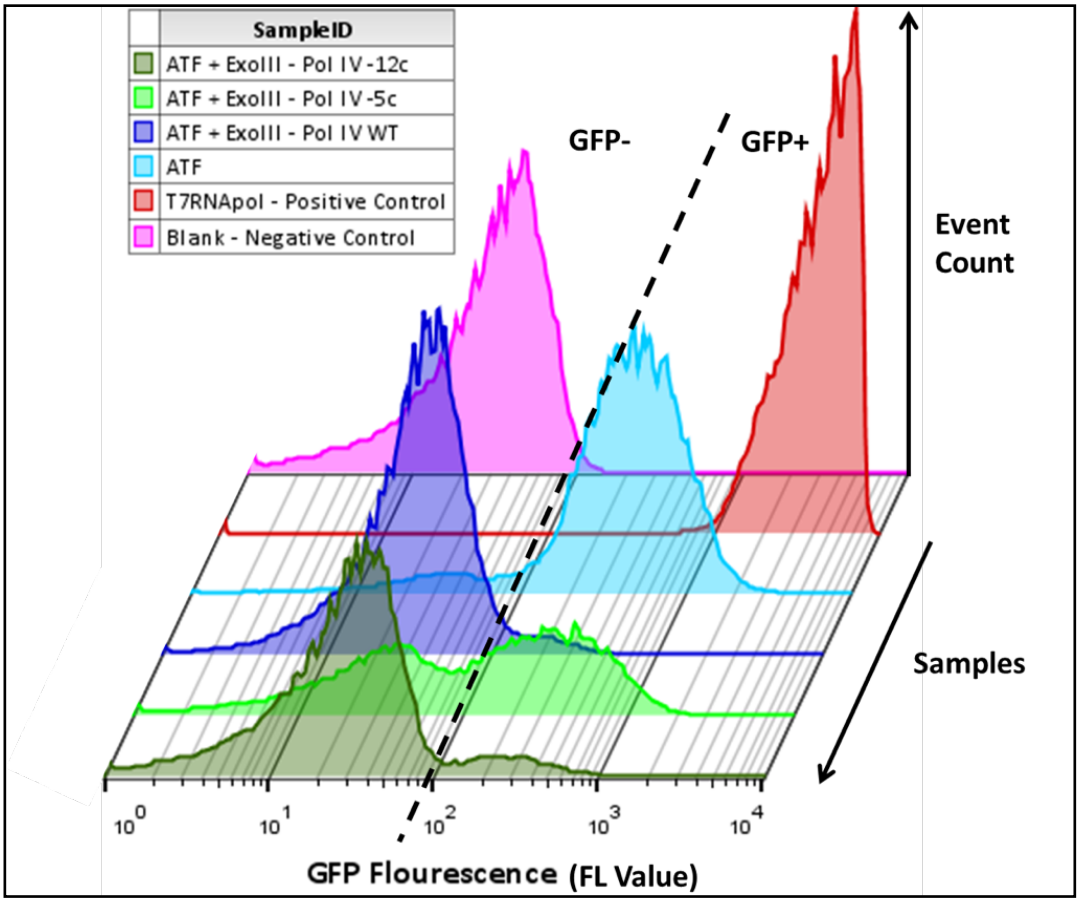


Figure 7.5 *– Histogram traces from initial loss of function assay in which ATF fusion protein was co expressed with individual Exo III and Pol IV protein variants. Selected traces clearly show a dramatic change in fluorescence profiles when Exo III and Pol IV are also expressed in the host GM31 strain (Light and Dark Green traces).*

Nevertheless, there is still the possibility that error-prone repair is happening in a random fashion and may be occurring at the points of ATF induced mutation but by chance, or rather a consequence of the higher frequency of mutations within the target gene. Up until this point, the lack of Exo III – Pol IV fusion means there is no way to know whether error-prone repair is guided to the abasic site or

whether it is an indirect response. For example, overexpression of these novel protein variants might result in the normal high fidelity BER repair proteins being out competed.

Following on from this preliminary test, the complete library of Exo III-Pol IV fusion proteins were co expressed alongside ATF mutator, once again from a single plasmid system. In the interest of clarity, matching version of the fusion linker have been grouped together; Rigid Linker (***Figure 7.6***), Flexible Linker (***Figure 7.7***) and Natural Linker (***Figure 7.8***).

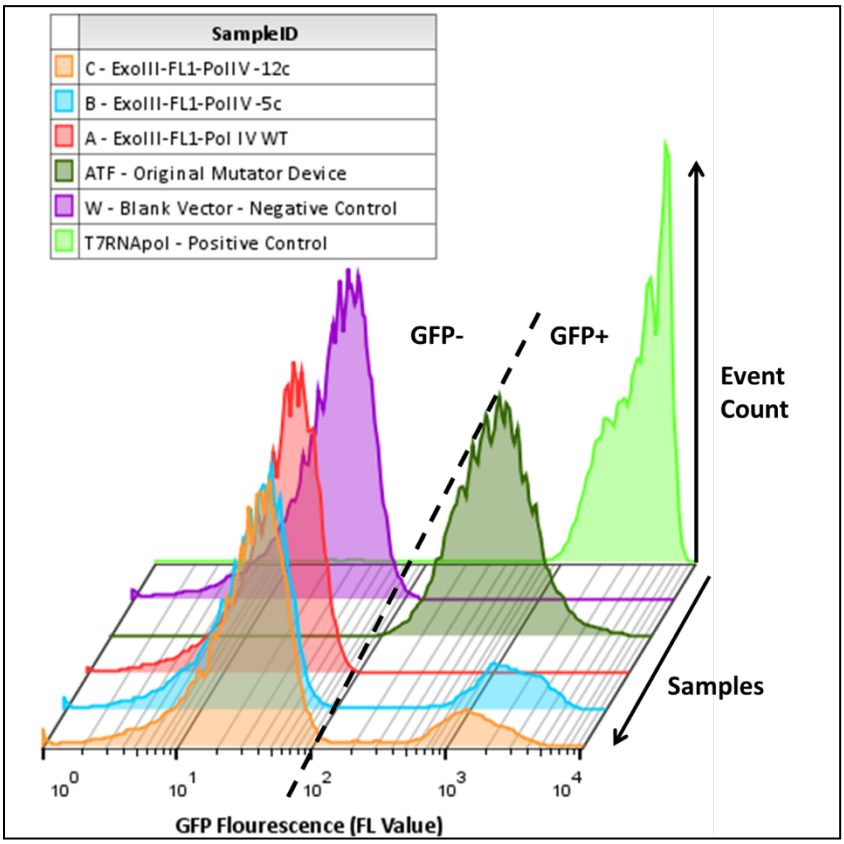


Figure 7.6 *– Histogram traces from loss of function assay in which ATF fusion protein was co expressed with Exo III-Pol IV variants on a single plasmid. Selected variants of the Exo-III Pol IV fusion protein contain a rigid linker.*

Once again, it is fairly noticeable that the constructs expressing the wildtype Pol IV do not show any GFP + population. In fact, given the lethality of over expressing Pol IV, I would expect no growth of these cultures. I suspect that survival of these populations is due to a mutation within the Pol IV wildtype gene. Given that Pol IV is highly transcribed from a plasmid, there is increased opportunity that it will incur background mutations within its coding sequence due to the mutagenic effects of ATF.

Having previously already seen evidence of frame shift mutations within the Pol IV WT gene during initial cloning efforts it seem highly probable that the same thing could be happening here. The destructive nature of the Pol IV protein means that any cell, which has eliminated a functional version gains a fitness advantage and will easily propagate within the growing culture.

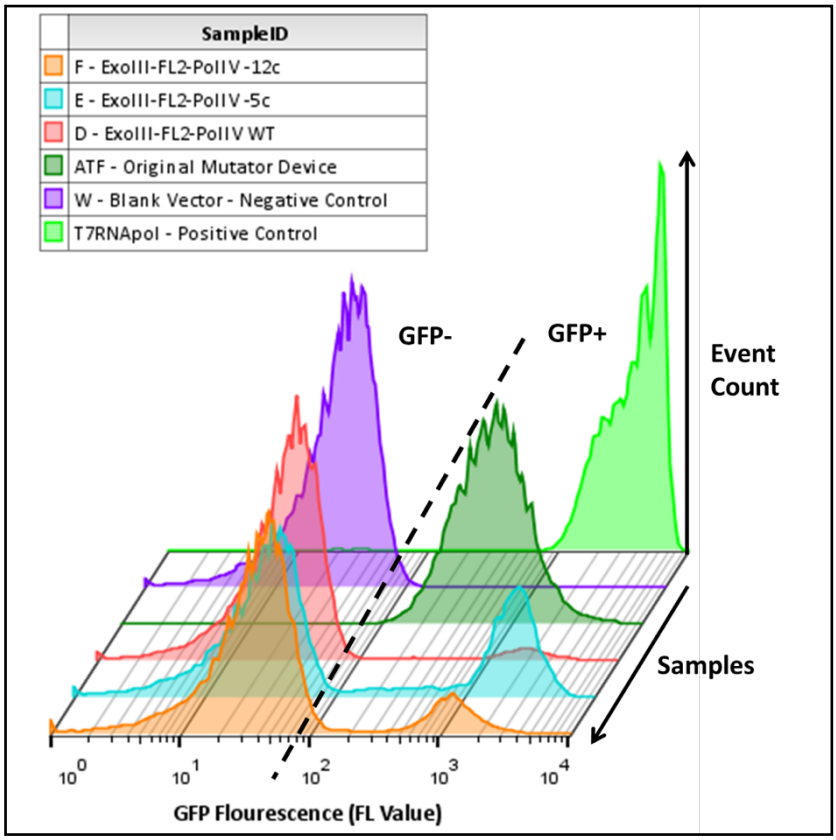


Figure 7.7 *- Histogram traces from loss of function assay in which ATF fusion protein was co expressed with Exo III – Pol IV variants on a single plasmid. Selected variants of the Exo III Pol IV fusion protein contain a flexible linker.*

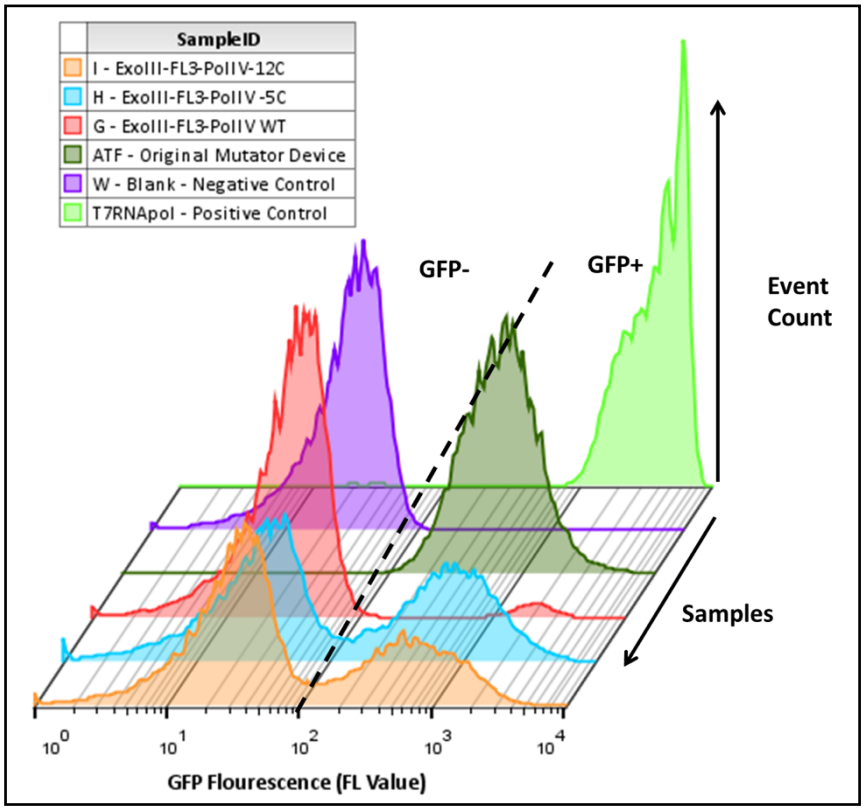


Figure 7.8 *- Histogram traces from loss of function assay in which ATF fusion protein was co expressed with Exo III – Pol IV variants on a single plasmid. Selected variants of the Exo III Pol IV fusion protein contain a rigid linker.*

The most interesting profiles were seen when the Exo III – Pol IV fusion proteins with C-terminal truncations that had been engineered to contain the naturally occurring linker sequence (BASIC-Linker 3). These showed the emergence of the characteristic intermediate transition between the GFP+ and GFP- populations, as clearly seen in blue and orange traces (***Figure 7.8***).

7.5.3 Sequencing

To understand the true cause of the altered fluorescence profiles, and to attempt to identify whether or not the expression of error-prone polymerases within the system had in fact led to an increase mutational spectrum, a large-scale sequencing effort was required. However, due to both time constraints and the significant associated costs, only very few samples were sequenced.

Once again, samples were plated following the flow cytometry analysis, in order to maximise the chance of selecting individual colonies that did actually contain mutations of interest i.e. those populations, which showed a large negative GFP population were chosen. Therefore, samples relating to construct H (ATF.ExoIII-FL3-PolIV-5c) and construct I (ATF.ExoIII-FL3-PolIV-12c) which demonstrated interesting mutation profiles were chosen for sequencing.

Unfortunately, the results were not compelling. The only mutational event identified across 10 sequencing reads was a -1bp deletion within codon 3 of the GFP gene, which would in fact result in highly disruptive frameshift mutation and create an immediate stop codon. Ultimately this would result in a loss of functional GFP protein and provide a fitness advantage to this cell population as they were not required to produce excess quantities of the exogenous GFP protein, allowing them to grow faster, replicate sooner and therefore propagate within the mixed culture.

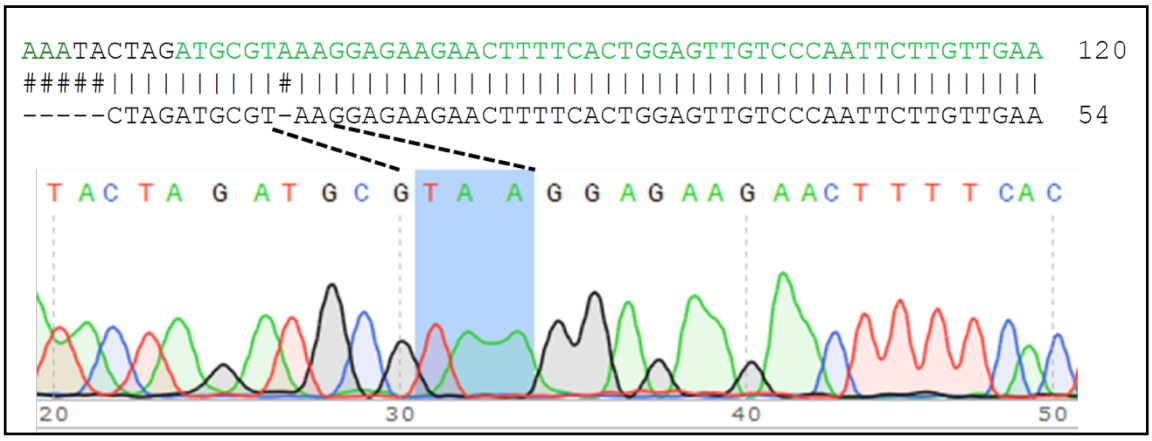


Figure 7.9 *– Evidence of a frame shift mutation within the third codon of the GFP CDS which creates in a false stop codon*

It was also noted that no C-T changes were identified, suggesting that the ATF mutator protein cannot produce the same mutagenic phenotype in the wildtype GM31 strain.

7.6 Conclusions

I was only able to draw very limited conclusions from this phase of work. Primarily, those being that overexpression of Pol IV protein is not tolerated within the target *E.coli* GM31 strain, in fact it results in a complete loss of cell viability. Over-expression of Exo III protein and the truncated version of Pol IV protein was possible, without directly effecting cell viability

I strongly believe this initial work identifies a desperate need for an alternative screening assay which would enable a more definitive and immediate answers when screening such a large library of variants.

The GFP loss of function assay was useful for identifying the loss of a fluorescent population as mutations are accumulated for successive periods of growth, and although a range of alternative histogram profiles were observed, this assay provides no link to the type or indeed the cause of the mutation. There are extensive labour and financial costs associated with sequencing such a vast library and yet this would be unlikely to yield meaningful results given the low mutation rates, along with a high risk of selecting for undesirable mutations (such as those seen in any instance where the cell can gain a fitness advantage). This is also the possibility of off-target mutations occurring in the T7RNApol gene which cannot be easily recorded (as described in ***Chapter 2***).

Alternative, a gain of function assay might be beneficial as this should enable the user to define a particular type or set of mutations needed for production of the functional protein. However, the flow cytometry assay widely used throughout this work is not compatible with a gain of function assay. Due to the fact that the underlying mutation rate is low and the maximum number of cells (events) that can be counted is deceivingly low, the probability that a reversion back to an active GFP protein is unworkably small. The desired change would be so infrequent that any GFP+ population would be miniscule and would remain undetected. Furthermore, alternative mutations elsewhere in the target GFP may be destructive and would in fact mask any improved gain of function, further reducing the probability that any favourable change resulting in the gain of function could ever realistically be identified.

An earlier idea involved engineering the GFP cassette so that it contained a base change within the fluorophore that is critical to fluorescence of the protein. A single disruptive base change, which was permissible by the proposed mutator system (dependent on a C-T change on the coding strand), could not be identified. A further suggestion was to introduce a false stop codon that would result in dramatic loss of function due to an aberrant and truncated form of the protein being produced.

However, even in the later version of the mutator device, which has the potential to cause transversion mutations during error-prone repair, the expected point of entry for mutations remained a deaminated cytosine. Therefore, it was impossible to use a false STOP codon (TAA, TGA, TAG) given these codons do not contain a cytosine base.

As this period of research came to an end, my attention turned to developing an improved screening assay that would provide the opportunity to explore a larger library of fusion proteins with far greater reliability and that would be compatible with large next generation sequencing of mutation events (detailed in ***Chapter 8***).

With a robust assay in place, the library of fusion proteins could actually be expanded further in an effort to find a truly functional variant. In fact, I envision a random linker library between the Exo III - Pol IV domains loosely based on the consensus sequence of naturally occurring linkers found between functional domains of multi domain proteins (polymerases), and those which contain the most viable residues (based on charge and polarity). In this case, phosphorylated oligos that anneal to form the linker sequence could be used. Inclusion of a single blunt cutting restriction site between both domains should enable multiple copies of the linker library to be cloned in without any directional guidance yet still ensure the whole coding sequence remains in frame. This means that both linker length and variance could be explored. Of course, this idea is entirely dependent on first establishing a viable and robust screen and designing a library with a size that is screen able.

Ultimately, the varied library of constructs and indeed the comprehensive list of individual parts will be highly beneficial for future working in upgrading the mutator device.

8 Final Conclusions and Future Work

This body of work has helped move the wider research goal forward, with clear resolution into the activity of an active mutator (ATF) capable of highly target mutagenesis *in vivo*. Nevertheless, there are a number of unanswered questions that remain and will be of vital importance to the overall success of the project.

Firstly, I was unable to accurately measure mutation rate given the inaccuracy of fluctuation analysis (which make large assumptions) and the incompatible experimental procedure. However, many of the alternative methods for molecular evolution employ this form of analysis and calculate mutation rate on the evidence of single mutagenic changes. Therefore, fluctuation analysis could be justified when making direct comparisons. Alternatively, an accurate and quantifiable mutation rate could be achieved *in vitro* but it seems unlikely such estimates would correlate with *in vivo* activity. However, purification of protein variants, such as those used to slow T7 processivity, could be then used to target ssDNA during *in vitro* transcription and provide a strong indication of desirable mutagenic proteins.

Secondly, for this system to be truly successful a wider mutation spectrum is required. Comprehensive screening of the previously built error-prone variants requires improvements in screening. A gain of function assay would provide stronger estimate of mutation rate and can be focused toward the specific type of restorative change.

The *in vivo* system may also favour continuous outgrowth within a bioreactor, ensuring that continuous logarithmic growth can be maintained

Long term focus should also consider alternative methods for selection of protein targets with a wide variety of functions including novel binding affinity and rate of catalytic activity. Numerous screening methods such as two hybrid selection or phage / yeast display could be adapted to provide and immediate output selection for alternative protein targets (*Vidal and Legrain, 1999., Baker et al, 2002*). Depending on the target protein, and indeed the size of the library, the brute force approach of high throughput mass spectrometry could also be used.

8.1 Recent Advancements in the Field of Directed Evolution

During the course of my research, I have been drawn to a number of recent advancements across the field of synthetic biology but with specific focus on molecular tools which could benefit wider efforts in directed evolution. During this time, there have been a number of review articles offering updates and critiques of methods applicable to molecular evolution (*Cobb et al, 2013., Lane and Seelig, 2014*).

A most recent review article by David Liu's group, whom I would consider leaders in the field thanks to their work to develop PACE, provides a strong summary of these improvements (*Packer and Liu, 2015*).

The same group has also published a general set of mutagenesis plasmids for *in vivo* directed evolution that have been shown to produce a broad spectrum of mutations (***Figure 8.1***) (*Badran and Liu, 2015*). This suite of plasmids is an extension of the mutagenesis plasmid (MP) created during the earlier PACE publication, which included a variant of DNA Pol III proofreading domain and Pol V domains and originally achieved substitutions rates of 5.4×10^{-8}. However, this is still substantially lower that the mutagenesis rates that can be achieved by current *in vitro* mutagenesis techniques. Interestingly, a later improvement to the MP demonstrated that error-prone rates could be increased to 9.9×10^{-8} by removing the error-prone repair capabilities all together. Further adaptions of MP to overexpress *dam* methylation protein and *seqA* GATC binding protein encouraged mismatch repair achieving 4.4×10^{-7}. Perhaps an even more interesting improvement, and closely linked to my own strategy, was an MP expressing a cytidine deaminase *CDA1* coupled with the expression of a UDG inhibitor UGI (although the author also notes that p56 binds with UDG by mimicking of structural and electronic features of uracil-containing DNA). Previous investigations have suggested that CDA1 results in stronger mutation rates than AID (*Lade et al, 2011*). Ultimately this MP achieved substitution rates of 2.0×10^{-6}. Further adaptations of note include repression of protein pump proteins to prevent expulsion of mutagenic bases from the cell 6.2×10^{-6}.

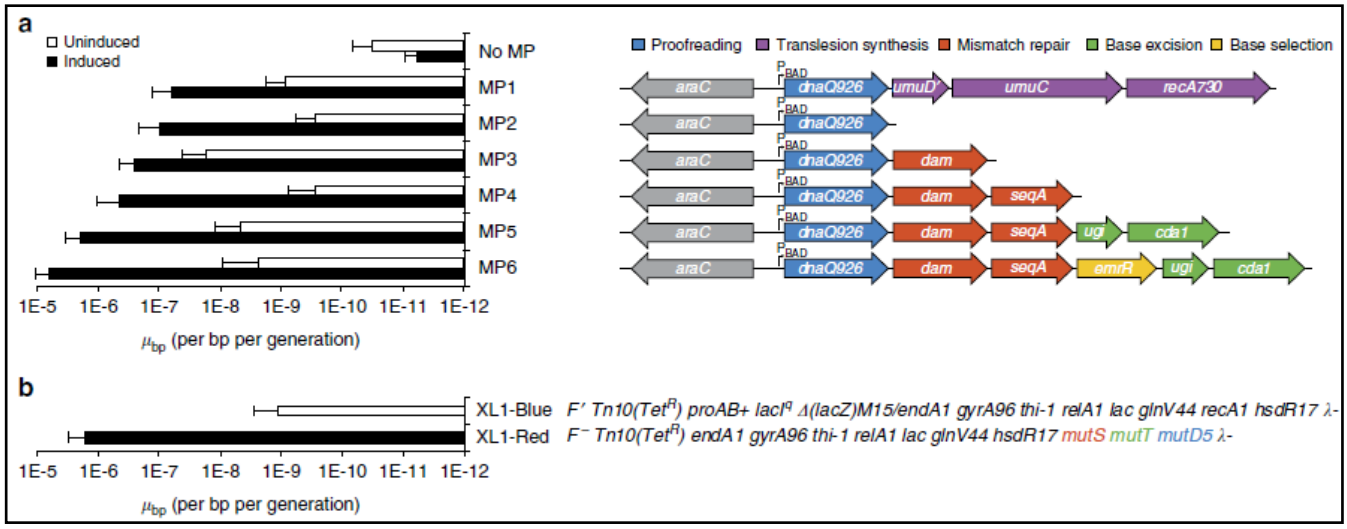


Figure 8.1 *– Mutagenesis plasmid created for general in vivo mutagenesis with E.coli host strains (a). Widely used and commercially available mutagenic strains of E.coli showing detailed genotype (b). Taken from Badran and Liu, 2015.*

This extensive work, which includes the construction and characterisation of over 80 MP, along with various estimations of mutation rates may be particular beneficial for developing future iterations of our own mutator device. The author states that a mutation rate above 4×10^{-7} is likely to result in a loss of host cell viability. This equates to roughly 1.9 per genome generation and appears to be in accordance to the predicted probability of mutagenic changes being incurred by essential genes (10-30% of the genome, 0.2–0.6 mutations per essential gene per generation) (*Mackie et al, 2014*).

Undoubtedly, this level of mutagenic potency could only be reliably increased by truly targeting genes of interest, rather than depending on low level inducible expression of global mutator proteins, which further justifies the work and general MP designs described within my own work.

Another major advance in molecular biology over the past 3 – 4 years has been the discovery and adaptation of the CRISPR (Clustered regularly interspaced short palindromic repeats) system in prokaryotes. A deep level of understanding into this primitive type of learned immunity, which uses guide RNAs to guide the Cas9 endonuclease (CRISPR associated protein) to highly specific restriction sites within the genome, is fuelling interest across biotechnology. The ability to introduce targeted double strand breaks is a substantial upgrade on classic restriction endonucleases, with a wide range of proposed applications across a number of host systems (*Mali et al, 2013*).

Unsurprisingly, this a great deal of crossover within the field of genome editing, given that highly targeted changes can be made *in vivo* with unerring accuracy, way beyond that of ZF's and TALEN's. It has already been used for multiplex mutagenesis in *E.coli* with significant success (*Jiang et al, 2013*). However, these approaches rely on delivering donor DNA at the point of the break and end homology repair mechanisms, rather than a direct enzymatic change to the target site. Recent revelations around the types of DNA repair post Cas9 activity may also be helpful in understanding how best to make use of these double strand breaks as points of entry for error-prone repair (*van Overbeek et al, 2016*).

Excitingly, a number of very recent publications have linked guide sequences with DNA deaminases. One publication from the Church lab at Harvard, details the construction of ZF or TALEN – AID fusions to direct deamination events towards a broken ACG start codon in a GFP reporter gene, meaning that a targeted C-T change would result in a fluorescent output (*Yang et al, 2016*). Editing efficiency was improved by knocking out *mutS* and *ung* genes, key to mismatch and base excision repair of deaminated bases respectively. The author also describes variants of the C-terminal AID fusion and linker length/sequence compositions, highlighting the importance of identifying a viable linker with a clear influence of catalytic and DNA editing frequency. It remains to be seen if these linker sequence lengths, compositions or their positioning could be useful to our own development of AID-T7RNApolymerase fusion proteins.

A parallel study, once again conducted by David Liu's lab, focused on CRISPR-deaminase fusions which can improve editing efficiency by virtue of the fact that CRISPR-Cas9 binding creates the exposed ssDNA loop required for deaminase activity (*Komor et al, 2016*). In fact, the chosen variant was the ratAPOBEC1 fused to the N-terminus of CRISPR-Cas9. This approach was able to be pinpoint cytidine bases to a 5bp window resulting in the corrective change in up to 75% of total cellular DNA. Quite remarkably, the later generations of these highly engineered proteins also included the fusion of UGI.

The strength of this approach has been confirmed by very similar research published at almost the same time (*Nishida et al, 2016*).

I believe that CRISPR guides would be ideally suited for building highly targeted fusion proteins that are perfectly suited for *in vivo* molecular evolution. The wide number of complex nuclease – deaminase fusions also opens up the question of targetable repair, in order to open up the possibility of a much larger mutational spectrum. One could hypothesise a lead 5'UTR sequence which not only contains a high active promoter sequence (to ensure expression of mutated variants) but a CRISPR compatible site with a far broader window of mutation. If this target window could be extended across larger stretches of the target gene or domains of interest, perhaps by extending the intervening linker sequence (larger than the 8-16 residue linkers described) then it would make for a very powerful tool. I believe this quickly expanding field, with many DNA synthesis companies offering bioinformatics tools and fast production of CRISPR guide RNAs and kits (*IDT, Synthego, Twist Biosciences, Desktop Genetics*) numerous variants of said fusion proteins could be made on a target by target basis.

8.2 Alternative Forms of the Mutator Device

The desire for highly targeted mutagenesis *in vivo* was originally tackled using an eloquent system in which a DNA Pol I variant was highly targeted to plasmids with ColE1 origin of replications (*Camps et al, 2003*). This highlight one of the major problems with the system discussed here, given that errors are introduced during transcription and subsequent repair meaning there is a less than direct link between phenotype and genotype. If alternative viral DNA polymerases (rather than an RNA polymerase) are identified, they could be used to target the gene of interest with high specificity and result in low fidelity replication, then clear genotypic changes would make identification of interesting variants far easier, whilst disconnecting the mutagenic from the host replication machinery. I believe an alternative approach may utilise orthogonal replication rather than orthogonal transcription to introduce increase error rates and maintain viability of the host.

In fact, one such system has been developed in yeast, which uses two orthogonal plasmids (pGKL1/2) capable of autonomous protein primed replication when both introduced to the host (*Ravikumar et al, 2014*). This disconnect helps prevent any interaction with the host genome and ensures that the replication polymerase expressed from 1 plasmid can lead to mutations during replication of the second plasmids which houses the target gene. It should be noted that this study did not attempt to calculate the frequency of insertion or deletion (INDEL) mutations, which are not optimal for molecular evolution purposes, although it did suggest that the error-prone polymerase was capable of achieving 10^{-3} per base per replication, which for a 1kb target gene, is ideal for selection

purposes. Further experimentation with this dual plasmid system identified an orthogonal RNA polymerase capability, which makes it even more useful for targeted *in vivo* mutagenesis.

A further musing could potential explore the use of mRNA dependent DNA polymerases such as those than function as reverse transcriptases often found in retroviruses. Given that the current mutator system produces plentiful amounts of mRNA of the target gene, if there lead marker sequences that can directed this class of enzymes to the nascent RNA strand, then once again this offers a more direct link between genotype and phenotype.

8.3 Planned Improvements to the Mutator Device

As previously mentioned, the ability to screen for interesting variants of the mutator device has up until this point been based on a loss of function assay. Unfortunately, given that the target GFP gene can incur numerous mutations without a change in phenotype, it is extremely difficult to link genotypic change to a reliable observation of phenotype. A more optimal solution would be a gain of function assay that could provide an added level of selection and avoid the presentation of false positives.

Numerous solutions were considered, such as introducing a false stop codon with the LacZ gene permissible for blue white screening (*Pham 2003, Bransteitter 2004*). Alternatively, a double fluorescent gene cassette could become the basis for a more comparative 'loss of function' assay. A target sfGFP gene under control of the cognate T7 promoter and a non-target mCherry gene under control of a constitutive J23101 promoter at the same genome locus would avoid any biasing due to greater susceptibility of one locus over another. sfGFP and mCherry show significant sequence homology and flow cytometry could be used to record both emission profiles, with a qualitative assessment between targeted and background mutations. Nevertheless, the output for selection in these cases is still reliant of colorimetric emission readings which will undoubtedly lead to the selection of numerous false positive.

More recently, I have noted suites of rifampicin *rpoB* and *lacZ* mutants covering all 12 transitions and transversion mutations are available and this could be used to comprehensively describe the mutational spectrum and biasing in direct response to expressing future version of mutator device (*Garibyan et al, 2003. Cupples and Miller, 1989*). Both of these target cassettes would help generate a definitive mutation rate via fluctuation analysis which would also be directly comparable to recent high-quality publications *(Badran and Liu, 2015*).

Much of my final few months on this project were spent developing a suitable suite of constructs for a gain of function assay which should produce a binary output i.e. 'I' the desired mutation has been identified or '0' the desired mutation has not been identified (***Figure 8.2***).

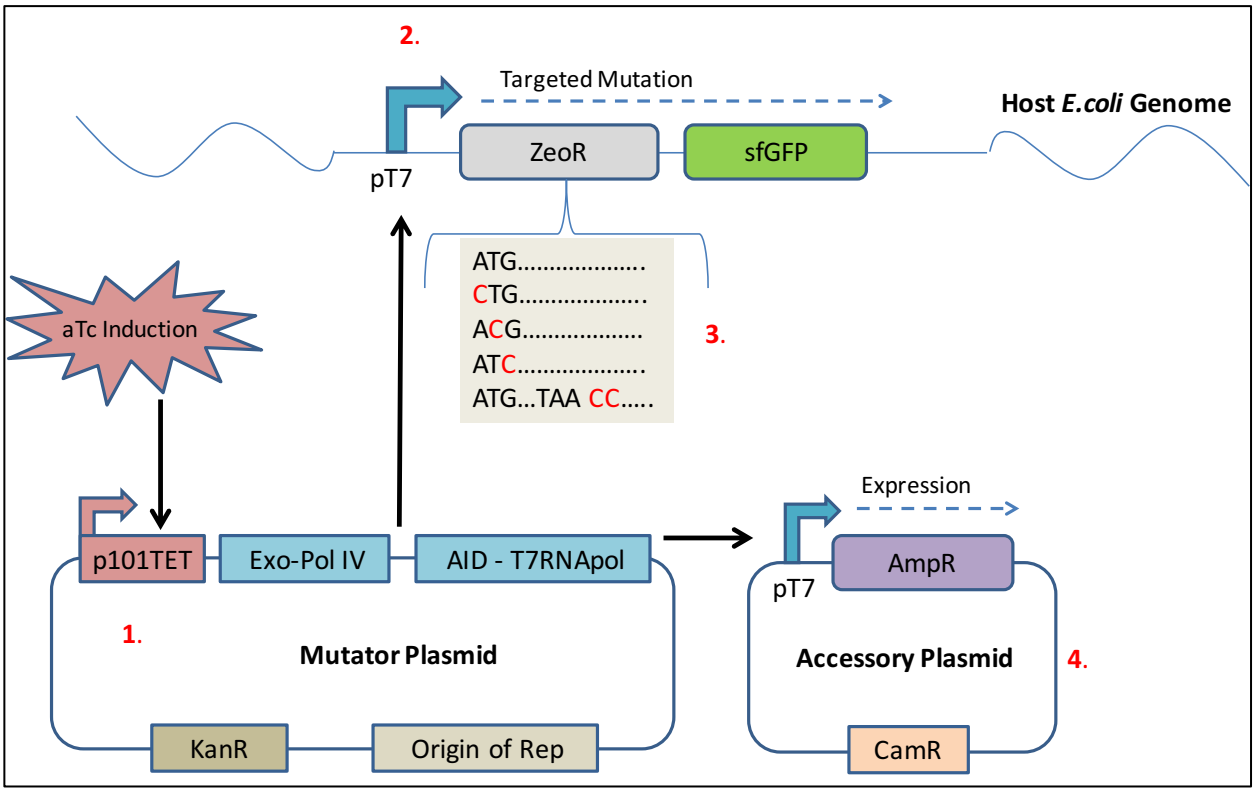


Figure 8.2 *– Proposed system for improved gain of function assay to enable selection of variants of the mutator capable of different types of mutations.*

The new screening approach will consist of a plating assay on varying concentrations and types of antibiotic. Although this may be somewhat labour intensive to set up, it will mean that hundreds of thousands of cells can be selected simultaneously. They key steps are outlined below:

1. The mutator plasmid will remain much the same i.e. both the ATF fusion protein will be expressed under aTc induction, alongside expression of the library of Exo – Pol IV variants
2. Mutation will be directed towards the target gene as previously described by targeting the ATF protein to its cognate T7 promoter. It is anticipated that some deamination events will be repaired erroneously due to the expression of the error-prone repair protein.
3. A newly designed selection cassette will be integrated into the host genome and will consist of an antibiotic resistance marker (Zeomyocin) and sfGFP. Multiple versions of the resistance cassette have been made which alter the start codon of the gene to include a single Cytosine base, thus preventing a functional protein from being translated. Therefore, these cells will be unable to survive on the antibiotic. Given that all three forms of C – T, A, G changes will result in a viable start codon, this will permit and immediate identification of the desirable transversion mutations. Furthermore, an internal stop codon could be used to identify patch repair, as multiple base changes would be required to produce a viable protein.

4. An accessory plasmid containing a further resistance marker under the control of a T7 promoter will ensure that no false positive is selected. Specifically, this will prevent knockouts to T7 if the mutator strain is grown under double antibiotic selection. In effect this creates a AND gate that requires both criteria to be satisfied in order for colonies to grow.

The construction and integration of the parts has already been enabled by BASIC assembly and preliminary data has already been collected which shows recovery of the altered start codons as a consequence of error-prone repair (***Appendices - Figure 11.15***). This new screening effort will be combined with the a recently established suite of robots which include colony pickers, air displacement and acoustic liquid handling robots, plate sealers, thermocyclers, DNA fragment analysers and fluorescence spectrophotometers plate readers, meaning the throughput of the assay will be greatly improved. For example, surviving colonies after large scale plating could be picked for colony PCR verification and PCR master mix could then be prepared and divided in 96 and 384 formats using the acoustic handler. The final PCR products could be sampled to ensure that they contain the target gene fragment prior to sequencing.

To coincide with this, I believe that in order to obtain strong and meaningful sequence information which truly reveals the extent of the mutator phenotype that would enable the construction of comprehensive and detailed heat map, the assay would favour a next generation sequencing effort. An NGS facility has been identified with the capability to run samples upon the 300-400 end runs providing 10,000 reads. The intention is to gel purify and amplify targets using a barcode labelling system followed by final sample purification using DNA bead pull down and dye quantification, which is much in keeping with other published work within the field (*Badran and Liu, 2015).* Data processing would require the development of automated analysis scripts for multi read alignments and comparisons to wildtype sequence. It should be noted that high throughput NGS sample preparation and runs have a particularly high error rate. Therefore, a comparison to a control run will be vital to ascertain correct mutation spectra and a 'true' mutated fraction will need to be carefully calculated.

As previously stated, a robust and high throughput screening capability would enable an even more diverse set of Exo-Pol fusion proteins to be characterised. Given the large assumptions made during the construction of these variants, their activity is not guaranteed. By designing an intelligent library which includes a broader set of changes, the possibility of identifying a functional variant is increased.

Beyond this current work, it remains to be seen how variants of the TLS Pol V behave when co-expressed with the ATF mutator protein. Once again, the level of Pol V expression will need to be carefully controlled to ensure viability of the cell, but this presents a problem given that it is a multi-

subunit complex and requires activation by the RecA master controller. It remains to be seen how expressing RecA, UmuD' or UmuC' mutants will affect the host cell.

A further extension of the error-prone system may include site-directed mutagenesis to engineer even lower fidelity into this range of TLS polymerases. Well establish models point to specific residues which interact with the replicating base pair. If there is indeed an inherent biasing or preference for a particular nucleotide substitution (revealed by extending sequencing efforts), these domains would be interesting candidates for further protein engineering.

Furthermore, there are many of the mutagenic protein candidates which are directly or indirectly linked to error-prone repair. For example, Terminal deoxynucleotidyl Transferase (TdT) is an enzyme involved in antibody diversification with the ability to add random nucleotides to any free DNA 3' end as it is template independent and expected to be active downstream of Exo III (*Seeberg et al, 1995*).

Further improvisations include sequence optimised version of the target gene to create CDS which are heavily weighted to G/C's and thus potential sites of deamination, whilst retaining the same amino acid sequence. A simple python script has been used to produce optimised variants of sfGFP and LasR targets and it remains to be seen how much more susceptible the updated versions of these genes will be to mutation by ATF (***Appendices*** – ***Figure 11.6***). It should be noted that this may result in non-optimal codon optimisation and ultimately lead to a lower level of protein expression and so more mature forms of this script should take into account *E.coli* codon usage and codon pairing.

An interesting proof of principal assay might be to simply target sfGFP with the intention of changing the excitation or emission spectra. Recent access to a fluorescence activated cell sorter (FACS) could automate a high throughput screening assay, selecting interesting single cell variants from a mixed culture which has been subjected to multiple rounds of mutation. The synthetic biological toolkit already contains numerous different fluorescent proteins but there is a need for proteins with far red shifted emission profiles, so this could be an interesting phenotype to search for (*Davidson and Campbell, 2009*).

8.4 Suggested Improvements to the Complete *In Vivo* Circuit

The work to accurately characterise two sets of promoter sets within the widely used MG1665 *E.coli* strain has produced parameters which should be transferrable to alternative models. However, a number of experimental flaws, often as a result of the primitive liquid handling platform, meant that data collection was drawn out and often not as repeatable as was first hoped. Numerous runs were aborted and whole data sets were lost. Furthermore, quality control check points often meant that upon initial first pass analysis it became immediately clear that particular data sets must be removed.

Undoubtedly, automation and standardisation of the characterisation protocol is incredibly desirable and a new state-of-the-art facility within CSynBi is already being used to characterise the original library of promoter variants in an effort to generate even more accurate parameters.

Futhermore, intermediate testing of the various modules should be carried out as true measurements of behaviour and to better predict the complete circuit response. For example, a cognate LasR / AHL (known to show binding within an active complex) could be used to activate pLas and drive expression of the output module, determined by selection on antibiotic and visualised by sfGFP.

The circuit may require a further tuning to maintain the required level of each protein intermediate and numerous points of control exist during the conversion of DNA to protein. For example, transcription is governed by the concentration or binding affinity of the transcription factor (TF) or the copy number of the target gene. Given that the circuit is inherently linked to a selection output governed by the activity of transcription factor, and its association with its cognate promoter, the transcriptional elements do not favour change. However, a more likely approach would involve altering translation strength by RBS tuning. Predictive models based upon interaction thermodynamics between the mRNA transcript and ribosome have been developed, in an attempt to accurately link the strength of translation initiation with protein expression levels, but this has had relatively mixed success (*Sallis et al, 2009*). The more workable solution would be to utilise the modularity offered by the BASIC assembly standard, to change the selected (and characterised) RBS variations through varying linker assembly. Very recent work in our lab has shown that functional RBS linkers can be used successful for combinatorial BASIC assembly (***Appendices - Figure 11.17*** and ***Figure 11.18***).

An alternative way to target translation rates could be to manipulate the mRNA transcript to form hairpin secondary structures which can extend half-life, although the effect on the ability of said transcripts to interact with ribosomal machinery is not well characterised (*Carrier and Keasling, 1999*).

Finally, down regulation of intermediate proteins may be achieved by tuning degradation within the cellular environment by the relatively simple addition of C-terminal tags within mRNA. A major flaw in the current model is that there is no true parameter for protein degradation, merely a dilution of protein levels associated with cell doubling. Work within CSynBi to characterise a system for robust and tuneable protein degradation within *E.coli* are on-going, with attempts to separate transcription and translation and direct proteins for along the innate ClpXP, ClpAP and SspB pathways for removal (*Lies and Maurizi, 2008*). Further efforts within the wider synthetic biology community to establish an orthogonal protease and cognate tag which does interact with the host degradation pathways would enable degradation to be tuned in isolation (*Cameron and Collins, 2014*). These will be important to further inform our model and should provide an added level of control to the circuit.

Construction of a device that is capable of generating an optimal mutation rate *in vivo* against intermediate metabolite protein targets which are not directly linked to growth selection was recently proposed (*Chou and Keasling, 2013*). The author shares a similar goal as to the project described here, with an attempt to overcome the inefficiencies of *in vitro* approaches and the inherent risk to the host when generating the library *in vivo*. Feedback regulated evolution of phenotype (FREP) is comparable to our complete mutator device in that it can regulate mutation rate based on output, except it is designed to identify a changing concentration of the target intermediate which included tyrosine and isoprenoids, with mutations a direct consequence of expressing a DNA Pol III mutant with limited proofreading capability.

Another advantage of the proposed system could be the molecular evolution of a multi gene pathway in complete isolation from the host i.e. targeting more than one gene in parallel. There is nothing to prevent more than one gene being expressed under the control of the specific T7 promoter. Furthermore, by generating variations of the ATF protein, which have differing mutational activity and target very specific promoter sequences (such as those described in *Esvelt 2011*), this could in theory allow for tuneable mutation across the complete pathway (*Temme et al, 2012*).

8.5 Closing Remarks

The majority of the aims outlined at the start of this work have been achieved with some noticeable success.

The first primary aim was to develop a mutator device which has the potential to target mutations to a gene of interest, whilst maintaining host cell viability. This has been established through testing a library of variants in both targeted and untargeted assays, with extensive sequencing revealing the type and speed of mutations. More importantly, it highlighted that earlier conclusions surrounding the validity of particular constructs were incorrect and confirmed that a novel fusion protein was functional.

Another primary aim was to begin to construct a more sophisticated genetic circuit within the host cell and this has been somewhat successful. Early models were useful for reiteration of the circuit design, surpassing and yet simplifying first generation designs. Indeed, much of the work to establish a strong protocol for part characterisation and expand the repertoire of genetic parts available for this project has been completed.

Thirdly, there has been an extended effort to introduce an increased level of mutational diversity into the system by altering DNA repair. This work has been extensively mapped out and the majority of test constructs and target strains required to fully test the updated version of the system are now in place and ready to go. Current efforts are focused on building a reliable screening method.

Lastly, an off shoot of this research has included the design, testing, optimisation, validation and application of a novel DNA assembly method has been very successful. Its continued uptake across the research community has been both an exciting and rewarding experience. Its reliability and usability, along with establishing a well-populated part list which is relevant and compatible genetic parts should help the overall project to accelerate through slower periods that would be otherwise associated with repetitive cloning efforts. I believe over the coming 12 months this will mean that a complete and functioning gene circuit can be tested.

9 General Methods

Introduction

All methods outlined here have been used during completion of my doctoral research and have been utilise throughout the prior chapters. Any variation or deviations from the standard protocols shown here are provided within the method subheading within each specific research chapter.

9.1 DNA Purification

9.1.1 Phenol/Chlorophorm Extraction of Genomic DNA

3ml of cell culture was pelleted at 15,000 RPM for 5 minutes. Cells were resuspended in 500 µl CTAB buffer and then incubated for 1 hour at 65°C. Extraction mix consisted of equal volumes of phenol/chloroform/sample. The sample was then vortexed briefly and centrifuged for 15,000 RPM for 5 minutes. The clear aqueous phase left on the top of sample was transferred to a fresh tube. An additional step extra chloroform step was then performed in which 500 ml chloroform added to the sample and the spun for 5 minutes. Once again, the aqueous phase was removed from the top of the sample and transferred to a fresh tube. 1 ml of 100% ethanol (ice cold) was then added to the sample and spun at 15,000 RPM for 15 minutes. The sample was left on ice for 20 minutes and then the supernatant was poured off carefully to avoid loss of pellet. 1ml of 70% ethanol was added and the sample was spun briefly for 1 minute, before pipetting off as much ethanol as possible and leaving the sample to air dry at room temperature for 30 minutes. Finally, the pellet of slightly cloudy appearance was resuspended in TBE buffer prior to storage at -20°C.

Table 9.1 *– Contents of the CTAB Buffer required for Genomic DNA extraction*

CTAB Buffer	Volume (250 ml stock)
Hexadecyltrimethylammonium Bromide (CTAB)	5 g
Tris (pH 8.0)	25 ml
NaCl	20.45 g
EDTA	1.68 g
2% B-mercaptoethanol	Added just before use

9.1.2 PCR Purification

Purification of PCR products is often required to remove unwanted buffers and retained high quality DNA for downstream applications i.e. restriction digestion, sequencing etc. The Sigma-Aldrich GenElute PCR Clean-Up kit (NA1020) were used for all PCR purifications. The manufacturers guidelines were followed.

9.1.3 Magnetic Bead DNA purification

The Agencourt AMPure XP purification system (*PN: 000387v001*) utilises a solid-phase paramagnetic bead technology for high-throughput purification of dsDNA from samples. This protocol was used when a high number of samples were required for purification, because it can be performed in a 96-well plate using small volumes (under 200 µl). It has also been applied to an automated robotic platform and requires a magnetic base plat which remains fixed on the footprint of the robot, with the 96-well sample plate moved off and on as required. Manufacturers guidelines were followed as summarised here.

The AMPure XP bead solution was kept at 4°C in-between uses, ensuring that the bottle is shaken to resuspended any precipitate before use. 1.8 µl of AMPure XP was added for every 1 µl of PCR reaction sample i.e. for a 50 µl PCR reaction, 90 µl AMPure XP is added. The samples were placed into a 96-well plate with rounded bottom wells, mixed by pipetting up and down a number of times and left to incubate at room temperature for 5 minutes to maximise yield. The plate was then placed on a Magnet plate for 2 minutes to separate beads from solution. At this point, DNA was bound to the beads and remains when the cleared solution was aspirated from each well. 200 µl of 70% ethanol is added to each well and incubated for 30 seconds before removal. The plate was allowed to dry at room temperature for up to 5 minutes to ensure that residue ethanol was lost. The plate is then taken off of the magnetic plate, and DNA was removed from the beads by the addition of 40 µl of MilliQ dH_20 to each well, ensuring that each sample was pipetted numerous times. The plate was then placed back on the magnetic base plate and the beads separate from solution. Finally, the purified DNA sample was eluted off and transferred to a clean Eppendorf tube for long term storage at -20°C. The protocol was improved by using a multichannel pipette or applying a robotic automated liquid handling platform to carry out each step.

9.1.4 Gel Extraction

Gel extraction and subsequent purification of PCR products of a specific size were often required to avoid the inclusion of unwanted products in downstream applications. My laboratory used the Omega Bio-tek E.Z.N.A Gel Extraction Kit (*PN: D2501-01*) using the spin column design for use with a mini bench-top centrifuge. Manufacturers guidelines were followed which are summarised here.

To begin with, PCR products were stained with Invitrogen SYBRGreen (*PN: S-7563*) and loaded onto a 1% Agarose gel with 6 x loading dye. Comparisons were made against the NEB 1kB DNA ladder (*PN: N3232L*) consisting of known molecular weight markers. The correct band was excised from the gel after visualisation of the UV light box. The chosen fragment of gel was weighed and placed in an Eppendorf alongside the corresponding volume of binding buffer as per the manufacturer's guidelines. The gel was melted by placing the Eppendorf in a heatblock at 60°C for 10 minutes. Once completely

melted the sample was loaded on to the provide spin purification column with collection tube and centrifuged at 10000 x g for 1 minute. The filtrate was then placed back into the top of the column and run through three times before being discarded. A further 300 µl of binding buffer was added and centrifuged 10000 x g for 1 minute. 700 µl of wash buffer containing ethanol was loaded onto the column and spun at maximal speed (13,000 x g) to wash the sample. The filtrate was again discarded before the column was spun again at maximal speed to dry the column. A clean collection tube was placed below the spin column and 30 µl of MilliQ dH_20 was added to the column for elution. The column was left to sit for 30 seconds before being spun at maximal speed for 1 minute. The elutant was passed through a total of three times as it was found to greatly increase DNA yield. Purified samples were then transferred to a clean Eppendorf for long term storage at -20°C, and for later use in downstream applications.

9.1.5 Plasmid DNA preparation

Preparation of plasmid DNA is an essential requirement when working with *E.coli*, and the extraction and purification of high quality DNA is essential for use in downstream applications. Standard practice within my laboratory was to use the Omega Bio-tek E.Z.N.A Plasmid DNA Mini Kit II (*product number D6943-01*) which uses a spin column containing a binding matrix which ensures that DNA or RNA bind specifically but reversibly under optimal conditions. Manufacturer's guidelines were followed and are summarised here.

Initially, small bacteria cultures were grown overnight in LB or SOC media. Depending on the plasmid copy number or *E.coli* strain in question, differing volumes of culture were suggested. As standard, 5ml cultures were grown from either single colony or glycerol stocks overnight to obtain dense cultures in exponential phase. 3ml of culture was pelleted in an Eppendorf tube by centrifugation at maximal (13,000 x g) for 5 minutes with removal of the supernatant. 100 µl of 'Resuspension Buffer' with RNaseA was added and used to completely resuspend the cell pellet. 250 µl of 'Lysis Buffer' was added to the tube, followed by gentle inversion of the tube to avoid sheering of the DNA. Complete lysis was allowed to take place for 2 minutes before addition of the 350 µl Neutralization buffer followed by immediate inversion of the tube. The tube was then spun at maximal speed for 10 minutes to pellet cell debris and supernatant was then transferred to the HiBind DNA column. The column was spun at maximal speed before addition of HB buffer and again spun at maximal for 1 minute. A wash step then requires the addition of 700 µl of DNA wash buffer containing ethanol, as per the manufacturer's guidelines. The column is spun twice at maximal speed to dry. Elution of the plasmid DNA requires the addition of 30 µl of MilliQ dH_20 at centrifugation at maximal speed for 2 minutes. Maximal yield of the sample can be obtained by running the elutant back through the column 2 further times. Purified

samples were then transferred to a clean Eppendorf for long term storage at -20°C, and for later use in downstream applications.

9.2 DNA Amplification

9.2.1 Polymerase Chain Reaction - Standard Conditions

The standard Polymerase Chain Reaction (PCR) was widely used throughout many cloning and diagnostic steps within this project. Various types of polymerases were used and a number of reaction conditions have been applied to ensure that the correct PCR product was obtained. This often required a trial-and-error approach when using new primers.

A master mix containing polymerase, buffer and dNTPs was always prepared and kept on ice throughout. Individual aliquots were divided out between 200 µl PCR reaction tubes before addition of relevant primers and DNA template. Primers were designed to have an annealing temperature of between 55-70°C, and considerations of Tm within the buffer used, hairpin/secondary structure/self-dimer/homo-dimer formation were analysed using NEB's Tm Calculator (https://www.neb.com/tools-and-resources/interactive-tools/tm-calculator) and IDT oligoanalyzer (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/) respectively.

Phusion Polymerase PCR Reaction

Phusion polymerase was often used for difficult to amplify targets, or if the size of the product was above 2-3kb, owing to the high-fidelity (3'-5' exonuclease activity) and speed (5'-3' polymerase activity) of the enzyme. It was used when accuracy and speed of amplification was needed.

Table 9.2 – Standard Phusion PCR reaction mix

Reagent	Volume
H20	31.5
5x HF buffer or GC buffer *	10
dNTPs 10mM	1
Primer A 10uM	0.5
Primer B 10uM	0.5
DNA Template	2 (0.1 – 1 µg final conc)
Phusion Polymerase (2U/ul)	0.5
DMSO	(1.5 µl if required)
Total Reaction	50 ul

Table 9.3 *- Phusion Polymerase PCR Cycle*

Step	Temperature	Duration	Cycles
Initial Denaturation	98°	3 min	1
Denaturation	98°	5-10s	25 – 35
Annealing	60° -70° (depending on primers)	10-30s	
Extension	72°	15-30s / 1kb plasmid	
Final Extension	72°	10 min last elongation step	1
Hold	4°	Infinite	-

**HF buffer was used if there was normal GC/AT nucleotide base distribution within the target sequence (between 40 – 60%). High GC content requires GC Buffer and DOCO*

PFU Turbo Polymerase PCR Reaction

Pfu Turbo DNA polymerase was used as the default standard within our laboratory, owing to the fact that members of the group were able to purify their own polymerase protein. The purity and concentration of the protein was checked, and the sample diluted to 5 x working concentration, before long term storage at -20°C. PFU showed good fidelity and processivity that made it useful for amplifying products of up to 3kb.

Table 9.4 *– Standard PFU Polymerase Reaction Mix*

Reagent	**Volume**
H20	31.5
10x Reaction Buffer	5
dNTPs 10mM	1
Primer A 10uM	0.5
Primer B 10uM	0.5
DNA Template	2 (0.1 – 1 µg final conc)
Pfu Turbo Polymerase (2U/ul)	0.5

Table 9.5 *- PFU Turbo Polymerase PCR Cycle*

Step	**Temperature (C)**	**Duration**	**No of Cycles**
Initial Denaturation	95°	2 min	
Denaturation	95°	30s	25 – 30 x
Annealing	55-70 (depending on primers)	30s	
Extension	72°	1 min / 1kb plasmid	
Final Extension	72°	10 min last elongation step	
Hold	4°	Infinite	

Controls:

For all PCR reactions, positive controls included clean template DNA that produces a known fragment size under the intended conditions and has well characterized primers that work at same annealing temperature. Negative controls contained the complete PCR mix without the inclusion of the template DNA and provided evidence that there was no contamination from other DNA sources i.e. cross contamination from colonies.

9.2.2 Colony PCR

Colony PCR was often used for diagnostic purposes i.e. individual *E.coli* colonies were picked directly from the growth media selection plate and diluted into individual PCR reaction tubes. Reaction conditions remained the same as above except the DNA template was replaced with complete whole cells from the colony. The initial denaturation step was extended to 10 minutes to break open cells within the sample, exposing both genomic and plasmid target DNA. Given that there is much broader DNA sequence within the sample, the stringency of the primers was often questioned, so redesign of primers and/or changes to the reaction conditions was needed on some occasion to eliminate non-specific products.

9.2.3 Genomic PCR

Genome DNA targets are sparse in that there is only ever a single copy per cell, and low concentration of target DNA would mean that initial PCR cycles generate much lower concentration of PCR product. Often this was overcome by increasing the number of repeat cycles and increasing the extension time during the reaction.

9.3 DNA modification

9.3.1 Site-Directed Mutagenesis

The protocol relies on the low error rate of Phusion polymerase (Finnzymes) in which the whole plasmid is amplified using an inverse PCR, with any required deletions, insertions, or nucleotide substitutions introduced within the primer design**.**

Firstly, 5’ phosphorylated primers were ordered (IDT) to ensure that the PCR product can circularise by ligation with roughly a 65°C melting temperature, extensions were added to each with up to 20 bp per primer. This approach was used throughout the report for introducing various tags, restriction and multiple cloning sites as well as specific base changes to triplet codons correlating to different amino acid residues. Initially, a high fidelity (HF) reaction buffer was used. In cases where amplification of the desired product failed, specifically in high GC rich regions or where there was a large amount of primer dimer, then a GC buffer and/or DMSO was used to increase PCR efficiency at the cost of a higher error rate.

Table 9.6 *– Standard Site-Directed Mutagenesis Reaction Mix*

Reagent	Volume (µl)
H_20	31.5
5 x HF buffer or GC buffer	10
10mM dNTPs	1
Primer A 10 uM	0.25
Primer B 10 uM	0.25
Template DNA 50 pg/µl	1
Phusion (2U/µl)	0.5
DMSO	(4 µl if required)
Total Reaction Volume	50

Table 9.7 *– Reaction conditions for Site-Directed Mutagenesis*

Step	Temperature (°C)	Duration	Cycles
Initial Denaturation	98	3 min	
Denaturation	98	20 seconds	
Annealing	55-70 (depending on primers)	30 seconds	25 – 35 x
Extension	72	15-30 seconds / 1kb plasmid	
Final Extension	72	10 min last elongation step	
Hold	4	Infinite	

5 µl of the sample was run on an agarose gel to check that the PCR reaction had worked correctly, the presence of the correct sized product and absence of non-specific products. Subsequent gel extraction was required for certain samples where non-specific PCR products could not be avoided. Therefore, 2 x 50 µl PCR reactions for each sample were run on an agarose gel. The correct bands were excised, and DNA as extracted using the gel extraction protocol.

A further clean up step was used to eliminate background template DNA from the reaction by digesting the original parental plasmid DNA. 1 µl DpnI and 1 µl NEB4 was added to PCR product. The sample was then incubated at 37°C for 15 minutes. Heat inactivate at 80°C for 20 minutes.

Positive reactions that showed low level background were ligated. 5 µl of the successful PCR product was mixed with 4 µl H_2O, 1 µl of 10x T4-ligase buffer and 0.5 µl T4 ligase (NEB). The reaction mix was incubated for 1h at room temperature.

The ligation reaction was transformed into 50 µl DH5 chemical competent bacteria by heat-shock at 42°C for 1 minute. 0.9 ml of SOC medium was added and the sample was shaken at 37°C for 1 hour. Finally, 100 µl of the culture was plated onto an appropriate LB antibiotic plate. For diagnostic purposes, I often picked 2 - 4 colonies and grow in appropriate LB medium overnight**.** I then miniprepped each clone and sent each for sequencing (MWG, Cogenics, Source Biosciences).

9.3.2 Klenow Extension

The Klenow extension protocol requires the activity of the Klenow fragment, obtained from DNA polymerase I. This protein retains 5’-3’ polymerase activity and 3’-5’ exonuclease activity required for

adding nucleotides during DNA replication and proofreading respectively. However, it has lost the 5'-3' exonuclease activity of DNA Polymerase I, ensuring that it can be used in such an *in vitro* application as this.

Klenow extension was often used to generate small fragments of DNA (~100bp) for various downstream cloning. Synthesis of two single stranded DNA oligonucleotides was seen as relatively cost effective, when compared to synthesising short dsDNA fragments. Primers were designed to overlap by at least 20 bps in the middle of the total DNA sequence. This allows for the addition of restriction sites if the intended use is to clone the fragment into a target vector by restriction digest. However, blunt ligation was also possible if the primers are 5' phosphorylated. The standard protocol used in this work requires the newly made dsDNA to be cloned into the blunt vector pJET 1.2, which already had phosphorylated ends to ensure efficient ligation.

A 100 µl reaction mix consisted of; 95 µl dH_20, 1 µl 100x BSA, 1 µl of each Oligonucleotide. Addition of 1 µl of dNTP's (0.25mM of each dNTP) and 1 µl Klenow Fragment**.** The reaction mixed was then placed into a thermal cycler incubated at 94°C for 5 minutes then cooled by 0.1°C/s to 5°C below the oligonucleotides Tm (melting point) and held for 5 minutes so that the oligonucleotides may anneal before being cooled by 0.1°C/s to 37°C. Alternatively, the sample can be placed in boiling water and allowed to cool to room temperature. The polymerase extension was subsequently initiated by adding dNTP's and the Klenow fragment itself, incubated at 37°C for a 1 hour to create dsDNA from the annealed ssDNA oligonucleotides**.** After extension, the Klenow polymerase was heat inactivated by incubating at 75°C for 20 minutes and then placed on ice. Subsequent restriction digestion and ligation into target vector followed using standard protocols (***described below***).

9.4 DNA Analysis

9.4.1 DNA Size Separation by Gel Electrophoresis

During this work, 0.5 – 1.5% agarose gels, were frequently used to separate dsDNA fragments based on size. Firstly 0.5 g – 1.5 g agarose was dissolved in TBE buffer and then heated in the microwave. After the sample had cooled briefly, it was poured into small casters that conform to the BioRad gel tank standards. 10 µl – 50 µl of DNA sample were mixed with SYBRGreen and 6 x Purple Loading Dye prior to loading the whole sample. 15 µl of 1 x NEB 1kb or 1 x NEB 50bp DNA ladder were loaded alongside samples and then run for 1 hour at 90 volts. Visualisation was possible using a blue light box.

9.4.2 DNA Quantification by Spectrophotometry

To check the purity and concentration of DNA samples, a Thermo Scientific NanoDrop 1000 spectrophotometer was used. Measurements at 260 nm and 280 nm are taken and the 260/280 sample

absorbance ratio used to assess purity. A sample with a 260/280 ratio of around 1.8 is considered pure. Sample concentration is calculated from the 260nm reading using the Beer-Lambert equation.

The Beer-Lambert equation is modified to use an extinction coefficient with units of ng-cm/ml. Using this extinction coefficient gives a manipulated equation:

$$c = (A * e)/b$$

c is the nucleic acid concentration in ng/microliter, **A** is the absorbance in AU, **e** is the wavelength-dependent extinction coefficient in ng-cm/microliter and **b** is the path length in cm.

The generally accepted extinction coefficients for nucleic acids are:

Double-stranded DNA: 50 ng-cm/ul

Single-stranded DNA: 33 ng-cm/ul, RNA: 40 ng-cm/ul

A blank measurement was made containing only the buffer solution in which the DNA samples was eluted or diluted. 1 µl of sample DNA was then loaded on the NanoDrop stage and measured accordingly.

9.5 DNA Assembly Methods

9.5.1 Gibson One-Step Isothermal Assembly

Original published by Dan Gibson in 2009, the Gibson One-Step Isothermal Assembly technique (referred to as Gibson) has quickly become one of the most widely used approaches to DNA assembly and helped replaced traditional approaches to molecular cloning *(Gibson et al, 2009)*. Gibson requires DNA parts with overlapping sequences at both ends, which are exposed upon exonuclease activity. This creates long overhangs that can be hybridised to complementary base pair overhangs on neighbouring parts. The length of the overhang can be extended to improve specificity and the preparation PCR step, to amplify constitutive parts, is useful for adding in additional sequence. I initially used Gibson assembly during the early cloning steps within this work, and as such adapted the original publication into my own protocol.

Overlapping complement nucleotide sequences are exposed when temporary T5 exonuclease activity causes regression of strands in the 5'-3' direction. When exonuclease activity is stopped, the overhang which remains on part A will bind with the overhang of part B by complementary base pairing. After annealing, the nick in DNA backbone is repaired by DNA ligase and a polymerase is required to fill in any missing nucleotides. The success and subsequent popularity as an assembly technique, relates to the improvements which allow for assembly in one single isothermal step, no longer requiring the thermocycler to control changes in temperature at different stages. This was made possible because the

T5 exonuclease used in this protocol recesses DNA from the 5' end, and consequently will not compete with the polymerase activity in the 5' > 3' direction. Previously, a T4 exonuclease was used, which recesses from the 3' end of DNA, meaning its activity had to be stopped before addition of polymerase and led to a time consuming two step approach. In the one-step approach, the exonuclease and polymerase do not require separation of their activities by changes in temperature. The optimal temperature for this technique is consider as 50°C so therefore a thermostable Taq DNA ligase is used in order to repair the backbone of DNA strands after complementary base pairing has occurred. Phusion polymerase was chosen, as it is now preferred in many cloning techniques due to its proofreading ability. Smaller dsDNA fragments (whether synthesised or amplified) can be joined together to make a much larger sequence, without any 'scar' left by residual restriction sites, by simply ensuring that there are overlapping ends that show complementation to each other. When amplifying individual DNA fragments, it is essential to design primers which have 20-40bp of overlap included as 5'end primer extensions. It is possible to incorporate multiple restriction sites and non-coding sequences into the overlap or make other changes to the DNA sequence.

DNA parts were added in equimolar concentrations, given the differing lengths of individual parts based on the following formula:

pMol x N x 660 pg/pmol x 1 µg/10^6pg = µg

pMol = Defined target DNA concentration, I used 0.025 pMol as standard

N = Number of nucleotides in part sequence

330/pmol = Average molecular weight of single nucleotide (660 pg/pmol for dsDNA)

Initially an Isothermal Reaction Buffer was prepared, ensuring that all reagents are dissolved in solution, before filter sterilisation. The reaction buffer can be stored long term at -20°C. Secondly, T5 exonuclease was added to the buffer and separated into individual reaction run tubes of smaller volumes to avoid continual freeze thawing which is known to degrade the T5 exonuclease. Prior to an assembly reaction, dNTP's, Taq DNA ligase and Phusion DNA polymerase was added to a single reaction run tube, alongside the correct volumes of the various DNA parts in equimolar concentrations. The total volume of the reaction did not exceed 20 µl once the DNA parts have been added. The reaction tube was then placed in a water bath at 55°C for 1 hour before immediate heat shock transformation into 50 µl of competent cells. After an hour outgrowth, the cells were spun down, resuspended in 100 µl of LB media and plated out on relevant LB agar antibiotic plates and grown overnight at 37°C. Successful transformant colonies were screened to ensure they had been correctly assembled by diagnostic PCR and digests.

Table 9.8 *- Isothermal Reaction Buffer*

Reagents	Amount	Final Concentration	Stock Solution Calculations
1M Tris – Cl	500ul	500nM	MW 121.14
1M MgCl2	50ul	50nM	MW 203.30
1M DTT	50ul	50nM	MW 154.25
100mM NAD	500ul	5nM	MW 663.43
PEG-8000	0.25 g		
dH20 (MilliQ)	250ul		
Total Reaction Volume	1ml		

Filter sterilised and stored at -20°C until later use

Table 9.9 *- Reaction Run Tubes*

Reagents	Amount
Isothermal Reaction Buffer	320 µl
T5 Exonuclease (10U/µl)	0.64 µl
dH20 (MilliQ)	679.36 µl
Total Reaction Volume	1ml

Divided into individual 'Run Tubes' (80 x 12.35µl aliquots) to avoid continued freeze thawing

Table 9.10 *- 1 x Final Reaction Mix*

Reagents	Amount
Reaction Run Tube	12.35 µl
dNTP Mix (10mM)	0.4 µl
Taq DNA Ligase	2 µl
Phusion DNA Polymerase	0.25 µl
Total Reaction Volume	15 µl

Added to Run Tubes before carrying out each individual assembly reaction

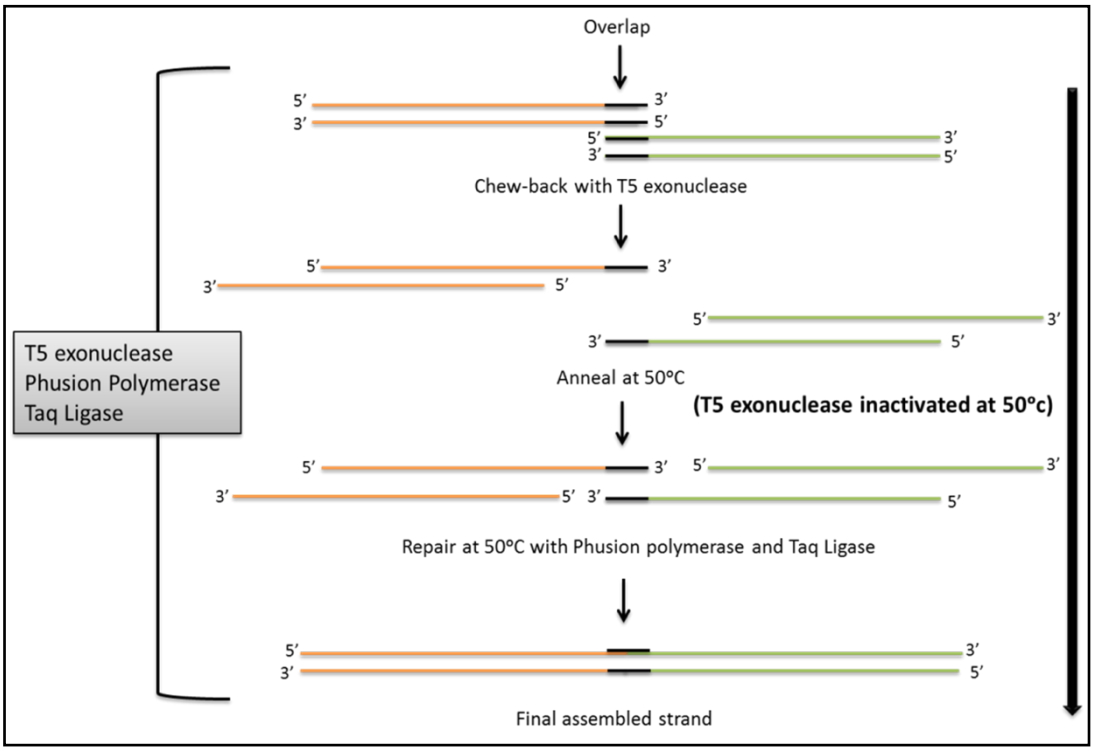


Figure 9.1 *– Overview of Gibson One- Step Isothermal Assembly*

9.5.2 BASIC Assembly

BASIC assembly was developed during my project and the DNA assembly method of choice which can be performed at the bench or using on an automated liquid handling platform. Please refer to **Chapter 3** for detailed description of the BASIC method, reaction mixes are summarised in tabulated format here.

Table 9.11 *– Reaction mix required for BASIC Linker preparation*

Reagent	Volume	Stock concentration	Final concentration
Annealing buffer 10mM TRIS-HCl buffer pH7.9, 100mM NaCl, 10mM MgCl2	49 µl		
Linker oligo	0.5 µl	100 µM	1 µM
Adapter oligo	0.5 µl	100 µM	1 µM
Total Reaction Volume	50 µl		

Table 9.12 *– Reaction mix required for BASIC part digestion and linker ligation*

Reagent	Volume (ul)	Stock concentration	Final concentration
Water	8.5		
ATP	3	10 mM	1 mM
NEB Buffer 4	3	10 x	1 x
BSA	3	10 x	1 x
iP Linker	5	1 µM	166.6 nM
iS Linker	5	1 µM	166.6 nM
DNA Part	1	76 nM	2.5 nM
BsaI- HF	1	20 U/µl	0.66 U/µl
T4 Ligase	0.5	400 U/µl	6.6 U/µl
Total Reaction Volume	30		

Table 9.14 *– Reaction mix required for BASIC part assembly*

Reagent	Volume (µl)	Stock Concentration	Final Concentration
DNA parts (each)	1	1.5 nM	0.15 nM
BSA	1	10 x	1 x
NEB4	1	10 x	1 x
Water	Up to 10		
Total Reaction Volume			

9.5.3 Restriction Digest

Restriction digests were often for diagnostic purposes and validation of constructs. Upscale versions of the digestion mix were used if a specific DNA fragment was required for downstream application given that as significant concentration of DNA may be lost during gel extraction and purification. The standard reaction mix is incubated at 37°C for 1 hour to ensure complete digestion

Table 9.15 *– Reaction mix for standard restriction digests*

Reagent	Volume (μl)
Restriction Enzyme	0.5
10x Buffer	1
BSA	(0.5 if required)
DNA	2
dH20	Up to 10
Total Reaction Volume	10

9.5.4 Ligation

The ligation of DNA fragments based on homology between corresponding overhangs was used in many of the standard cloning procedures. In a simple ligation, in which a smaller insert fragment is inserted into the larger vector DNA it was found to be good practice to calculate the ratio of DNA required, based on the concentration and length of both linear fragments, which by default is 6:1 Insert:Vector.

Table 9.16 *– Reaction mix for standard ligation reactions*

Reagent	Volume (μl)
T4 DNA ligase	1
10x Ligation buffer (including 10uM ATP)	1
Vector DNA (3kb)	50ng
Insert DNA (2kb)	(6 x molar concentration)
dH20	Up to 10
Total Reaction Volume	10

The ligation mix was incubated at 25˚C for 1 hour, or left overnight at 4˚C, before transformation into competent cells, plating onto agar, and checking by colony PCR, further diagnostic digestion or sequencing.

9.5.5 pJET 1.2

pJET 1.2 is a commercially available blunt ended vector that contains a lethal gene so that when both ends are ligated, and the vector is circularised without an intervening sequence at this region, the cell does not survive. In this way, it acts as an alternative to blue-white screening, providing positive selection for the insertion of a dsDNA fragment (Thermo Scientific).

The ligation mix was vortexed briefly and centrifuged for 3-5 minutes. It was then incubated at room temperature (22°C) for 5 minutes and directly used for transformation in competent cells.

Table 9.17 *– Reaction mix for blunt ligation into vector pJET 1.2*

Reagent	Volume (µl)
2X Reaction Buffer	10
PCR Product	1
pJET1.2/blunt Cloning Vector	1
T4 DNA ligase	1
dH20	Up to 20
Total Reaction Volume	20

9.6 DNA Sequencing

Standard Sanger DNA sequencing was outsourced to a number of different companies throughout the course of the research project, namely MWG Eurofins, Beckmann Coulter, Source Bioscience and GATC Biotech.

Although there was variance depending on the company's particular requirements, generally single reaction tubes contained 5 µl of plasmid DNA (100 ng/ul) or purified PCR products (1ng/ µl per 100bp) along with a separate 5 µl of oligonucleotide primer (3.2 pmol/ul) which defined the direction of the sequence read. Common stock primers were established over time.

9.7 DNA Synthesis

DNA synthesis was outsourced to Integrated DNA Technologies (IDT) throughout the duration of my research. This included the frequent production of short oligonucleotides required as PCR primer and infrequent synthesis of small dsDNA regions in the form of 500-1kb G-Blocks™.

9.8 Protein Expression

At numerous points within this work, a range of target proteins were expressed in *E.coli* cells. A crude, small scale approach was used to grow up small batch culture and a number of cell lysis methods were trialled.

9.8.1 *E.coli* expression

Firstly, expression constructs were transformed into the chosen *E.coli* strain (either MG1665, DH5a, BL21 or GM31) and plated on antibiotic (Amp, Kan, Cam) LB or SOC+ plates and grown at 37˚C. Single colonies were then picked and used to inoculate small 5ml starter cultures overnight at 37˚C. 1ml of overnight culture was used to inoculate fresh 50 ml media (1:50) in 250 ml conical flask (suitable for small scale protein preps). Cultures were grown at 25˚C to OD 0.2-0.4 (600nm) measured by spectrophotometry. Where required, IPTG (1 mM) and ATC (250 ng/ml) were added to induce expression of the target gene. 10 ml samples were taken at hourly intervals 0 – 4 hours post induction. Samples were centrifuged, flash frozen in liquid nitrogen and stored at -20°C.

9.8.2 Cell Lysis

Boiling Method

Firstly, 50 µl of 5 x SDS Loading Buffer was added to normalised cell pellet. The sample was then held for 5 minutes at 85-95°C (although it is noted that some proteins will aggregate at higher temp and lead to insolubility). 25 µl of each sample was retained in an effort to visualise both the complete fraction (insoluble and soluble). Insoluble protein samples were treated with 6-8 M urea or Triton X-100 (non-ionic detergent. The remaining sample was spun within pre-chilled microcentrifuge at maximum speed (15,000 rpm) for 10 mins at 4°C. The clear soluble fraction was then removed. Depending on the total amount of expected protein, sample was diluted 1 in 5 if required into 5 x Loading Buffer prior to loading onto SDS PAGE gel.

Table 9.18 *– Contents of 5 x Loading Buffer used in crude cell lysis*

5 x Loading Buffer	Volume (10ml stock)
1 M Tris.HCl pH 6.8	0.6 ml
100% glycerol	2.5 ml
10% (w/v) SDS	2 ml
b-mercaptoethanol	0.5 ml
1% (w/v) bromophenol blue	1 ml
ddH20	3.4 ml

Lysis Method

Firtly, 1 ml of Lysis buffer was added to resuspend 1 g cell pellet. 1 µl of lysozyme (300mg/ml), 10 µl of EDTA (100 mM) was then added and the sample was incubated at 4°C for 4 hours. Afterwards, 1 µl of DNase1 (1 mg/ml) and 5 µl of MgCl2 (1 M) was then added. The sample was incubated at 4°C for a further 30 mins before being centrifuged at 15,000 rpm, maintain a cold temperature of 4°C. The soluble cell lysate was removed and diluted with 5 x loading buffer prior to loading onto SDS-PAGE gel.

Table 9.19 *– Contents of Lysis buffer used for crude cell lysis*

Lysis Buffer	Volume (10 ml stock)
1M Tris-HCl pH 7.5	0.5 ml
1M NaCl	1 ml
1M DTT	10 µl
50% glycerol	1 ml
ddH20	7.5 ml

Freeze Thaw Method

Firstly, cells were normalised to 0.4 OD 600 nm and then centrifuged. The sample was then resuspended in 1ml FreezeThaw buffer and 0.5ml lysozyme (100 µg/ml). Samples were incubated for 20 minutes at 30°C. The sample was then subjected to 3 successive rounds of flash freezing (in liquid nitrogen) and thawing at 37°C. The lysate was then treated with DNase/RNase (5 µg/ml final concentration) for 15 minutes at 30°C. Cell debris was removed by centrifuging at 15,000 rpm for 15 minutes and the soluble fraction within the supernatant was removed and diluted with 5 x loading buffer (*shown above*) prior to loading onto SDS-PAGE gel.

9.9 Protein Analysis

In many instances within this work, proteins with total cell lysate were separated by SDS PAGE and visualised by direct Coomassie staining or via transfer to a membrane for western blotting.

9.9.1 Protein Size Separation by SDS-PAGE

Small SDS PAGE gels were made for use with the BioRad systems electrophoresis system. Stacking gels (5% acrylamide) were used, whereas the choice of resolving gels (7.5%, 12%, 15% acrylamide) was dependent on the size of the target protein band.

Gel preparation first required preparation of 0.75 mm – 1.0 mm glass plates which were washed and secured into holders. Stacking and resolving gel mixes (***Tables 9.20 and 9.21***) were prepared fresh, with the addition of TEMED and AMPS just prior to pouring. The resolving gel was poured and levelled off with 100% ethanol to ensure an even transfer of samples during separation, and then left to set for 30 minutes before the ethanol was removed using filter paper. The stacking gel mix was then added and a compatible multi well comb was immediately put into place, gently lowered down to avoid any bubbles. The gel was then left for further 30 minutes to set.

Table 9.20 *– Contents of the Stacking Gel used for SDS-PAGE*

5% Stacking Gel	Volume
mQ H2O	3.40 mL
30% acrylamide	0.83 mL
1.0M Tris.HCl (pH 6.8)	0.63 mL
10% SDS	50 µL
TEMED	10 µL
10% AMPS	50 µL
TOTAL	5 ml

Table 9.21 *– Contents of the Resolving Gel used for SDS-PAGE*

12% Resolving Gel	Volume
mQ H_2O	3.27 mL
30% acrylamide	4.00 mL
1.5M Tris.HCl (pH 8.8)	2.53 mL
10% SDS	100 µL
TEMED	14.2 µL
10% AMPS	133.3 µL
TOTAL	10 ml

Running Samples

Glass plates were removed from holders and excess gel removed, and then placed into the gel tank. 1 x SDS Running Buffer was then added so that the gels were completely submerged and then a pre-run was carried out at 150 volts for 10 minutes prior to loading any samples. After this each well was flushed through before samples were loaded with a micropipette. 10-12 µl of prestained protein ladder PageRuler (*Thermo Scientific*) was added to the first lane. Equal volumes for each sample were added to the rest of the lanes (most often 25ul). Empty wells were filled with 1 x Loading Buffer to ensure samples run level. The electrophoresis was then carried out at 220 V and the bromophneol dye was used as an indicator of how long to leave the samples running, typically this was around 50 minutes.

Table 9.22 – Contents of the 10 x SDS Running Buffer required for SDS-PAGE

10 x SDS Running Buffer	Amount
TRIS	30 g
Glycine	144 g
SDS	10 g
ddH20	1 litre total volume

9.9.2 Coomassie Staining

After running gels, glass plates were gently removed to avoid any breakages in the gel itself. The outside of the gel was then trimmed to remove rough edges and rinsed in fresh 1 x SDS running buffer before

being placed into 200 ml Coomassie Staining Buffer and left under gentle agitation overnight. The following day, the staining solution was removed, and the gel rinsed with ddH_20. Multiple rounds of destaining were then carried out using the Destaining Buffer described and this was intersected by washes in ddH_20. The final gel was imaged using plate imager under white light exposure.

Table 9.23 *– Contents of the Coomassie Staining Solution required for protein gel analysis*

Coomassie Staining Solution	Amount (1 litre stock)
Coomassie Blue R-250	1 g
100% Methanol	450 mL
Glacial acetic acid	100 mL
ddH20	450 mL

Table 9.24 *– Contents of the Destaining Solution required for protein gel analysis*

Destaining Solution	Amount (1 litre stock)
100% Methanol	100 mL
Glacial acetic acid	100 mL
ddH20	800 mL

9.9.3 Membrane Transfer

Prior to Western Blotting, all protein was transferred onto a nitrocellulose or PDVF membrane. This required the gel to be carefully placed in pre-cooled Transfer Buffer alongside the membrane and filter paper pre-cut to match the size of the gel, for 30 minutes. Gel and membrane were laid on top of one another, ensuring that there was complete contact without bubbles by gently rolling over the top of the gel. Wads of filter paper were then placed either side along with fibrous pads. The sandwich was then placed into a clasped cassette and loading into the blotting module, ensuring that it is in the correct orientation to allow for transfer from with gel position closest to the negative cathode and the membrane closest to the positive anode. The tank was then filled with Transfer Buffer and a cooling ice block to ensure that the tank did not become too hot and lead to warping of the gel. A magnetic stirrer was placed inside and the whole tank positioned on a stirrer plate to ensure even temperature throughout. Electrophoresis was carried for 2 hours at 100v (although was extended for larger proteins).

Table 9.25 *– Contents of the Transfer Buffer used for membrane transfer*

10 x Transfer Buffer	Working Concentration	Amount (1 litre stock solution)
TRIS	25mM	3.02 g
Glycine	190mM	14.26 g
Methanol	20% (w/v)	200 ml
ddH20		Fill to 1 litre

9.9.4 Western Blotting

The membrane was removed from cassette and washed in TBS for 10 minutes. Blocking of the membrane required addition of 5% Milk or 3% BSA in TBS and was left for 1 hour. Membrane was then washed twice in TBST buffer for 10 minutes each time. Primary Antibody was diluted in 5% Milk-TBST (to a range of 1:1000 – 1:3000). Membrane was loaded into a sterilin, and 2 ml of Primary Antibody solution was added. The sterilin was then placed on a rolling incubator to ensure gentle agitation for 2 hours. Numerous washes (>3) were carried out using TBST for 10 minutes each time. Secondary Antibody was diluted antibody in TBST (1: 10,000 – 1: 50,000) for 1 hour under gentle agitation, again using a rolling incubator. Numerous washes (>3) were performed using TBST for 10 minutes each time. A final wash in TBS (no Tween) for 30 mins helped remove excess tween before imaging. HRP conjugated secondary antibodies were visualised using the Pierce ECL method (*Thermo Scientific*). Chemiluminescent substrates were mixed at 1:1 ratio (1 ml of each) and pipetted directly onto the membrane which was then left 2-5mins in a darkened room. Excess liquid was drained from membrane before being placed in a clean plastic wallet ready for imagining. Imagining was performed on the plate imager under normal light exposure with the smallest aperture. Cumulative time course images were taken until the bands became apparent. A normal photograph of the gel membrane was also taken as this clearly showed the ladder position. Both images were overlaid to provide a general guidance for detected band sizes.

Table 9.26 *– Contents of the TRIS-Buffer used for Western Blotting*

TRIS-Buffer (TBS)	Working Concentration	Amount (1 litre stock solution)
Tris	20mM	2.42 g
NaCl	500mM	29.22 g
ddH20		750 ml
HCL		Adjust to pH 7.5
ddH20		Fill to 1 litre

Table 9.27 *– Contents of the TRIS-Buffer Tween used for Western Blotting*

TRIS-Buffer Tween (TBST)	Working Concentration	Amount (1 litre stock solution)
Tris	20mM	2.42 g
NaCl	150mM	8.76 g
ddH20		750 ml
HCL		Adjust to pH 7.5
ddH20		Fill to 990 ml
10% Tween 20	0.1%	10 ml

9.10 Cell work

9.10.1 Preparing Competent Cells

Standard Competent Cells

This technique was more widely used to prepare cells for general cloning purposes. Transformation efficiencies were regularly recorded up to 1 x 10^7 cfu/µg of plasmid DNA.

Firstly, cells were taken from virgin stock and streaked out on LB medium plates (without antibiotic) and grown overnight at 37°C. Single colonies were used to inoculate 5 ml LB in 10 ml sterilin and grown overnight at 37°C at 250 RPM. The following day, 250 µl of culture was used to inoculate 25 ml LB medium in 50 ml falcon and grown to 0.4 (OD600nm) at 37°C. Cultures were kept on ice and then spun at 4,000 RPM for 10 mins at 4°C. The supernatant was poured off and the pellet was resuspended in 2.5ml 50 mM $CaCl_2$ and left for 20 mins. Each sample was once again centrifuged at 4,000 rpm for 10 minutes 4°C. Cell pellets were resuspended in 2 ml 50 mM $CaCl_2$ +15% Glycerol and 200 µl was aliquoted into separate Eppendorf's for long term storage.

High Efficiency Competent Cells

This technique was widely used for applications where maximum transformation efficiency was required, namely library creation and multi part DNA assembly. Transformation efficiencies were regularly recorded up to 1 x 10^9 cfu/µg of plasmid DNA. Adapted from *Inoue et al, 1990*.

Once again cells were taken from virgin stocks and used to inoculate 5ml cultures, and then grown overnight at 37°C. Overnight culture was then used to inoculate 250 ml of SOC media in 2 litre conical flask. Incubation was carried out at 18°C at 220 RPM, until culture reaches 0.3-0.55 OD600 (usually 3-4 hours). The flask was removed and left on ice for 10 minutes before 200 ml of culture was removed and centrifuged at 4500 RPM for 5 min at 4°C. The supernatant was discarded, and the pellet kept on ice. 40 ml of chilled ITB buffer was used to resuspend the pellet, and once again left on ice for 10 minutes before subsequent spin, with the supernatant discarded. Samples were then resuspended in ITB and pooled together before addition of DMSO (7% final concentration), kept on ice for 10 minutes and then divided into 200 µl aliquots prior to storage at -80°C.

Table 9.28 *– Contents of ITB Buffer used in the production of High Efficiency Competent Cells*

ITB	Working Concentration	Amount (1 litre stock)
$MnCl_2$ $4H_2O$	55 mM	10.88 g
$CaCl_2$ $2H_2O$	15 mM	2.2 g
KCl	10 mM	18.65 g
H_2O		980 ml
PIPES	0.5M	20 ml

9.10.2 Transformation

The standard approach used to introduce DNA into competent *E.coli* cells was via a heat shock step, in which a very briefly period heating opens up temporary channels on the outside of the cell membrane, allowing circular plasmid DNA.

Firstly, 0.5 µl of mini prep plasmid DNA was added to 50 µl of chemically competent cells. The mix was then placed in an Eppendorf tube and left for 30 minutes on ice. The Eppendorf was then placed in a 42°C water bath for 60 seconds before being returned to ice. After 30 seconds, 1 ml of growth medium (LB or SOC) was added to the tube. Samples were allowed to recover by incubating at 37°C for 1 hour under gentle agitation.

9.10.3 Antibiotic Selection

Table 9.29 *– Antibiotic stock concentrations and working concentrations for working with E.coli*

Antibiotic	Stock Concentrations	Working Concentration (plasmid expression)
Ampicillin	100 mg/ml	100 µg/ml
Kanamycin	50 mg/ml	50 µg/ml
Chloramphenicol	25 mg/ml	25 µg/ml
Tetracycline	10 mg/ml	10 µg/ml
Gentamycin	10 mg/ml	10 µg/ml
Zeomycin	5 mg/ml	5 µg/ml

9.10.4 Growth Media

Table 9.30 *– LB Medium mixture*

Reagent	Amount (g)
Tryptone	10
Yeast Extract	5
NaCl (10 mM)	10
ddH_20	Made up to 1 litre

Table 9.31 *- SOC Medium mixture*

Reagent	Concentration (mM)	Amount (g)
Tryptone		20 L^{-1}
Yeast Extract		5
NaCl	10	0.58
KCl	2.5	0.19
$MgCl_2$	10	2.0
$MgSO_4$	10	2.46
Glucose	2	0.36
ddH20		Made up to 1 litre

Table 9.32 *- AIM Medium mixture*

Reagent	Amount
Tryptone	20g
Yeast Extract	5g
NaCl (10mM)	5g
60% Glycerol	10ml
10% Glucose	5ml

8% Lactose	25ml
ddH20	Made up to 1 litre

Neidhart MOPS Medium

The MOPS medium recipe was adapted from the established recipe (*Neidhardt et al, 1974*). Firstly, 100 ml 10 x MOPS medium (Teknova, PN: M2106). was mixed with 10 ml K2HPO4 (0.132 M) and 880 ml ddH_20, before adjusting to pH 7.2 with 10 M NaOH. The solution was then filter sterilized and stored at 4C for up to 1 month. Just before the sample was used, 10 ml of glucose was added to make a final concentration of 0.1%.

9.11 Other

9.11.1 OSIP Genome Integration

The OSIP genome integration protocol was adapted from the published method (St-Pierre et al, 2013). In summary, integration required me to grow *E.coli* MG1665 at 37°C in 950 µl SOC medium during the transformation recovery phase for 1 hour, during which period of time an integrase is expressed. The integrase itself is controlled by the cIts heat repressible promoter.

Firstly, MG1665 transformant cell strains were plated on 50 µg/ml Kanamcycin. Pir+ cells achieve low copy plasmid expression of the Kanamycin resistance and were also plated on 50 µg/ml Kanamycin. Both strains were then grown overnight at 28°C to avoid further integrase expression and the possibility of double integrates.

In order to verify that integration had worked, the first step was to identify colonies on the overnight plates (fewer colonies the greater the number of assembled parts). Before picking single colonies, 24-hour growth of the plates was required given the lower temperature. It often appeared that some colonies were significantly bigger than others and the larger colonies were likely to be those that shown successful integration i.e. they have the Kanamycin resistance gene. It was found that smaller colonies were likely to be background and that these cells had not lost the original plasmid. Nevertheless, false positive colonies were infrequent (~1 in 50).

10 colonies for each integration assembly in the MG1665 were restreaked (50 µg/ml Kan) and again grown out at 28°C to ensure loss of the integration plasmid. 2 colonies for each assembly into Pir+ were re-streaked (50 µg/ml Kan) so the plasmid DNA could be retained for later preparation.

For diagnostic purposes, colony PCR was performed to verify correct single integrates. Improved vector specific primers were designed to increase Tm by extending the base pair length. This was in an effort to avoid non-specific background products. 10 colonies from the MG1655 restreak plate were suspended in 25 µl H_2O each and added to the colony PCR mix, which included 4 locus specific primers.

This allowed fast validation due to a varying banding pattern which is dependent on single, double or no integration (***Table 3.3***).

Table 9.35 *- Colony PCR reaction Master Mix used to check OSIP genome integration*

Reagent	Volume (ul)
dH_20	160
PFU polymerase (10x)	20
PFU Buffer	5
Primers 1-4 10uM	2.5 each (10 total)
dNTPs 10uM	5
TOTAL volume	200

Table 9.36 *- Colony PCR reaction conditions used to check OSIP genome integration*

Step	Temperature	Time	Cycles
Initial Denaturation	95°C	2 min	
Denaturation	95°C	20s	30 – 35 x
Annealing	(dependent on primers)	20s	
Extension	72°C	90s	
Final Extension	72°C	10 min last elongation step	
Hold	4°C	Infinite	

Curing of the integration strain was important to remove any trace of antibiotic resistance and allow further downstream cloning. 2 single integrated correct clones were used to inoculate 5 ml SOC medium (50 µg/ml Kan) shaking over night at 28°C at 200 RPM. Overnight cultures were used to make chemically competent using standard protocol outlined previously. 50 µl of competent cells were transformed with 50 ng of pE-FLP via heatshock, recovered at 28°C in SOC medium, plated onto 50 µg/ml Amp and grown overnight at 28°C. A further colony PCR with primers P1-P4 was then used identify clones with successfully excised integration resistance module. 2 colonies showing the correct integration module were re-streaked onto LB agar plates with either no antibiotic, 50 µg/ml Kan and 100 µg/ml Amp plates in parallel to confirm loss of Kan resistance from integration at 28°C. After the loss of Kanamycin resistance is confirmed, removal of the pE-FLP plasmid was performed by re-streaking to single colony without antibiotics and growing at 37°C overnight as it cannot propagate at the higher temperature (without selection the Amp resistance plasmid is lost). Further colony picking and direct plating on 100 µg/ml Ampicillin was once again used to confirm a loss of pE-FLP. This was followed by a further outgrowth in liquid LB medium without antibiotics.

Final sequencing verification was required to confirm correct integration. Therefore, a PCR check using primers P1 + P4 was used to amplify across the complete cassette as these primers bind to genome specific sites only. The PCR product was then checked on a 1% agarose gel and purified using AmpPureX Mag Bead purification. Purified samples were sent for final verification by standard Sanger Sequencing methods using internal sequencing primers. Finally, verified genome integrated strains were kept as glycerol stocks at -80C prior to further downstream assays.

10 Bibliography

1. Adie, T. (2014). Development of a synthetic biology approach to targeted directed evolution of proteins *in vivo*. PhD Thesis. *Imperial College*. Available at: https://spiral.imperial.ac.uk:8443/handle/10044/1/28088
2. Akama-Green, E.H., Joshi, N.S., Tammela, T., Chang, G.P., Wagner, B.L., Lee, D., Rideout, W, M., Papagiannakopoulos, T., Xue, W., Jacks, T. (2016). Modular Assembly Platform for Rapid Generation of DNA Constructs. *Scientific Reports*, 6, p. 16836.
3. Akbari, M., Peña-Diaz, J., Andersen, S., Liabakk, N., Otterlei, M. (2009). Extracts of proliferating and non-proliferating human cells display different base excision pathways and repair fidelity. *DNA Repair*, 8 (7) p. 834-843
4. Amstutz, P., Forrer, P., Zahnd, C. and Pluckthun, A. (2001). *In vitro* display technologies: novel developments and applications. *Current Opinion in Biotechnology*, 12 p.400 - 405.
5. Arakawa, H., Hauschild, J. & Buerstedde, J.M. (2002) Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science,* 295 (5558) p.1301–1306.
6. Arnold, F. & Volkov, A. (1999). Directed evolution of biocatalysts. *Current Opinion in Chemical Biology*, 3 (1) p.54–59.
7. Atsumi, S. & Liao, J.C. (2008). Directed evolution of *Methanococcus jannaschii citramalate* synthase for biosynthesis of 1-propanol and 1-butanol by Escherichia coli. *Applied and Environmental Microbiology*, 74 (24) p.7802–8.
8. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology,* 2.
9. Badran, A.H., Liu, D.R. (2015). Development of potent *in vivo* mutagenesis plasmids with broad mutational spectra. *Nature Communications,* 6 (8425)
10. Baker, K., Bleczinski, C., Lin, H., Salazar-Jimenez, G., Sengupta, D., Krane, S., Cornish, V.W. (2002). Chemical complementation: A reaction-independent genetic assay for enzyme catalysis. *Proceedings of the National Academy of Sciences,* 99 (26) p.16537–16542.
11. Baker, T. A., & Sauer, R.T. (2012). ClpXP, an ATP-powered unfolding and protein-degradation machine. *BBA Molecular Cell Research*, 1823 (1) p15–28.
12. Beaudry, A. & Joyce, G., (1992). Directed evolution of an RNA enzyme. *Science*, 257 (5070) p.635-41
13. Becskei, A., Séraphin, B., Serrano, L., A-Mohammadi, S., Hawkins, R., Bateman, E., She, Z. (2001). Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *The EMBO Journal*, 20 (10) p.2528–2535.
14. Beletskii, A., Grigoriev, A., Joyce, S., & Bhagwat, A. S. (2000). Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *Journal of Molecular Biology*, 300 (5) p. 1057–65.
15. Besmer, E., Market, E. & Papavasiliou, F. (2006). The transcription elongation complex directs activation-induced cytidine deaminase-mediated DNA deamination. *Molecular and cellular biology*, 26 (11) p.4378–4385.
16. Bjedov, I., Dasgupta, C., Slade, D., Le Blastier, S. (2007). Involvement of Escherichia coli DNA polymerase IV in tolerance of cytotoxic alkylating DNA lesions *in vivo*. *Genetics*, 176 (3) p.1431–1440.
17. Blagodatski, A., & Katanaev, V. L. (2011). Technologies of directed protein evolution *in vivo*. *Cellular and Molecular Life Sciences: CMLS*, 68 (7) p.1207–14.
18. Bonner, G., Patra, D., Lafer, E. & Sousa, R. (1992). Mutations in T7 RNA polymerase that support the proposal for a common polymerase active site structure. *The EMBO journal*, 11 (10) p.3767–3775.
19. Botstein, D., & Shortle, D. (1985). Strategies and applications of *in vitro* mutagenesis. *Science,* 229 (4719), p.1193–201.

20. Boudsocq, F., Ling, H., Yang, W. & Woodgate, R. (2002). Structure-based interpretation of missense mutations in Y-family DNA polymerases and their implications for polymerase function and lesion bypass. *DNA repair*, 1 (5) p.343–358.
21. Bransteitter, R., Pham, P., Calabrese, P. & Goodman, M. (2004). Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. The Journal of Biological Chemistry, 279 (49) p.51612–51621.
22. Bransteitter, R., Pham, P., Scharff, M. & Goodman, M. (2003). Activation-induced Cytidine Deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (7) p.4102–4107.
23. Brar, S., Sacho, E., Tessmer, I., Croteau, D. (2008). Activation-induced deaminase, AID, is catalytically active as a monomer on single-stranded DNA. *DNA repair*, 7 (1) p.77–87.
24. Brieba, L. & Sousa, R. (2000). Roles of histidine 784 and tyrosine 639 in ribose discrimination by T7 RNA polymerase. *Biochemistry*. 39 (5) p.919–923.
25. Brieba, L.G, Padilla. R, and Sousa, R. (2002). Role of T7 RNA Polymerase His784 in Start Site Selection and Initial Transcription. *Biochemistry*, 41 (16) p.5144-9.
26. Broyde, S., Wang, L.H., Rechkoblit, O., Geacintov, N.E. and Patel, D.J. (2008). Lesion processing: high-fidelity versus lesion-bypass DNA polymerases. *Trends in Biochemical Sciences*, 33 p.209-219.
27. Brutlag, D., Atkinson, M., Setlow, P. & Kornberg, A. (1969). An active fragment of DNA polymerase produced by proteolytic cleavage. *Biochemical and biophysical research communications*, 37 (6) p.982–989.
28. Bunting,K., Roe,S. and Pearl,L. (2003). Structural basis for recruitment of translesion DNA polymerase Pol IV/DinB to the beta-clamp. *The EMBO journal*, 22 p.5883–92.
29. Cadwell, R.C. & Joyce, G.F. (1992). Randomization of genes by PCR mutagenesis. *PCR Methods and Applications*, 2(1) p.28-33.
30. Cameron, D. E., & Collins, J. J. (2014). Tunable protein degradation in bacteria. *Nature Biotechnology*, *32* (12) p.1276–81.
31. Camps, M., Naukkarinen, J., Johnson, B. P., & Loeb, L. A. (2003). Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (17) p.9727–32.
32. Canugovi, C., Samaranayake, M. & Bhagwat, A. (2009) Transcriptional pausing and stalling causes multiple clustered mutations by human activation-induced deaminase. *FASEB Journal*, 23 (1) p.34–44.
33. Carlson, J., Badran, A., Guggiana-Nilo, D. & Liu, D. (2014) Negative selection and stringency modulation in phage-assisted continuous evolution. *Nature chemical biology*, 10 (3) p.216–222.
34. Casini, A., Christodoulou, G., Freemont, P., Baldwin, G., et al. (2014) R2oDNA Designer: Computational Design of Biologically Neutral Synthetic DNA Sequences. *ACS Synthetic Biology*.
35. Casini, A., Macdonald, J.T., Jonghe, J.D., Christodoulou, G., et al. (2014) One-pot DNA construction for synthetic biology: the Modular Overlap-Directed Assembly with Linkers (MODAL) strategy. *Nucleic acids research*, 42 (1), p.7
36. Cavaleiro, A, M., Kim, S, H., Seppälä, S., Nielsen, M. T., & Nørholm, M. H. H. (2015). Accurate DNA Assembly and Genome Engineering with Optimized Uracil Excision Cloning. *ACS Synthetic Biology*, 4 (9) p.1042–1046.
37. Chahwan, R., Edelmann, W., Scharff, M. & Roa, S. (2012) AIDing antibody diversity by error-prone mismatch repair. *Seminars in immunology*, 24 (4) p.293 - 300.
38. Chang, C. C., Chen, T. T., Cox, B. W., Dawes, G. N., Stemmer, W. P., Punnonen, J., & Patten, P. A. (1999). Evolution of a cytokine using DNA family shuffling. *Nature Biotechnology*, 17 (8), p.793–7.
39. Chappell, J., Jensen, K., & Freemont, P. S. (2013). Validation of an entirely *in vitro* approach for rapid prototyping of DNA regulatory elements for synthetic biology. *Nucleic Acids Research*, 41 (5), p.3471–3481.

40. Chaudhuri, J., Khuong, C. & Alt, F. (2004). Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. *Nature*, 430 (7003) p.992–998.
41. Chaudhuri, J., Tian, M., Khuong, C., Chua, K., (2003). Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature*, 422 (6933) p.726–730.
42. Cheetham, G. & Steitz, T. (1999). Structure of a transcribing T7 RNA polymerase initiation complex. *Science*, 286 (5448) p.2305 - 2309.
43. Chen, X., Zaro, J., & Shen, W.C. (2013). Fusion Protein Linkers: Property, Design and Functionality. Advanced Drug Delivery Reviews, 65 (10), p. 1357–1369.
44. Chin, J. W., Cropp, T. A., Anderson, J. C., Mukherji, M., Zhang, Z., & Schultz, P. G. (2003). An expanded eukaryotic genetic code. *Science*, 301 (5635), p.964–7.
45. Chou H.H & Keasling J.D. (2013). Programming adaptive control to evolve increased metabolite production. *Nature Communications*, 4, p.2595.
46. Cobb, R. E., Si, T., & Zhao, H. (2012). Directed evolution: an evolving and enabling synthetic biology tool. *Current Opinion in Chemical Biology*, 16 (3-4), p.285–91.
47. Cobb, R. E., Sun, N., & Zhao, H. (2013). Directed evolution as a powerful synthetic biology tool. *Methods*, *60*(1), 81–90.
48. Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. (2013). Olefin cyclopropanation via carbine transfer catalyzed by engineered cytochrome P450 enzymes. *Science* 339, p.307–310.
49. Cohen, H., Tawfik, D. & Griffiths, A. (2004) Altering the sequence specificity of HaeIII methyltransferase by directed evolution using *in vitro* compartmentalization. *Protein engineering, design & selection*, 17 (1) p.3–11.
50. Collins, C., Arnold, F. & Leadbetter, J. (2005) Directed evolution of Vibrio fischeri LuxR for increased sensitivity to a broad spectrum of acyl-homoserine lactones. *Molecular microbiology*, 55 (3) p.712–723.
51. Collins, C.H., Leadbetter, J.R. & Arnold, Frances H, (2006). Dual selection enhances the signalling specificity of a variant of the quorum-sensing transcriptional activator LuxR. *Nature biotechnolog*y,24(6), p.708–12.
52. Collins, J. J., Gardner, T. S., & Cantor, C. R. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403 (6767), p.339–342.
53. Cormack, B., Valdivia, R. & Falkow, S. (1996) FACS-optimized mutants of the green fluorescent protein (GFP). Gene. 173 p.33–38.
54. Cox, E. & Horner, D. (1982) Dominant mutators in Escherichia coli. *Genetics*, 100 (1) p.7–18.
55. Crameri, A., Dawes, G., Rodriguez, E., Silver, S., & Stemmer, W. P. (1997). Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nature Biotechnology*, 15 (5) p.436–8.
56. Crameri, A., Raillard, S. A., Bermudez, E., & Stemmer, W. P. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, 391 (6664) p.288–91.
57. Crameri, A., Whitehorn, E., Tate, E. & Stemmer, W. (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature biotechnology*, 14 (3) p.315–319.
58. Cupples, C. G. & Miller, J. H. A. (1989) set of lacZ mutations in Escherichia coli that allow rapid detection of each of the six base substitutions. *Proceedings of the National Academy of Science USA*, 86 p.5345–5349
59. da Silva, E, F., Reha-Krantz, L, J., (2007).; DNA polymerase proofreading: active site switching catalyzed by the bacteriophage T4 DNA polymerase. *Nucleic Acids Research*, 35 (16) p.5452-5463.
60. Davidson, M. W., & Campbell, R. E. (2009). Engineered fluorescent proteins: innovations and applications. *Nature Methods*, *6* (10) p.713–717.
61. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M., & Díaz-Orejas, R. (1998). Replication and control of circular bacterial plasmids. *Microbiology and Molecular Biology Reviews,* 62 (2) p.434–64.

62. Delcour, A. H. (2009). Outer membrane permeability and antibiotic resistance. *Biochimica et Biophysica Acta*, 1794 (5), p.808–16.
63. Delker, R.K., Fugmann, S.D. and Papavasiliou, F.N. (2009). A coming-of-age story: activation-induced cytidine deaminase turns 10. *Nature Immunology*, 10, p.1147-1153.
64. Deutscher, M. & Kornberg, A. (1969). Enzymatic synthesis of deoxyribonucleic acid. XXIX. Hydrolysis of deoxyribonucleic acid from the 5' terminus by an exonuclease function of deoxyribonucleic acid polymerase. *The Journal of Biological Chemistry*. 244 (11) p.3029–3037.
65. Diaz, M. & Lawrence, C. (2005). An update on the role of translesion synthesis DNA polymerases in Ig hypermutation. *Trends in Immunology*. 26 (4) p.215–220.
66. Diaz, M., Verkoczy, L. K., Flajnik, M. F. & Klinman, N. R. (2001). Decreased frequency of somatic hypermutation and impaired affinity maturation but intact germinal center formation in mice expressing antisense RNA to DNA polymerase zeta. *Journal of Immunology*, 167 p.327–335.
67. Dickerson, S., Market, E., Besmer, E. & Papavasiliou, F. (2003). AID mediates hypermutation by deaminating single stranded DNA. *The Journal of Experimental Medicine*. 197 (10) p.1291–1296.
68. Dickinson, B., Packer, M., Badran, A., Liu, D. (2014) A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nature Communications*, 5 p.5352
69. Doi, N., Kumadaki, S., Oishi, Y., Matsumura, N., et al. (2004) *In vitro* selection of restriction endonucleases by *in vitro* compartmentalization. *Nucleic Acids Research*, 32 (12) p.95.
70. Dougherty, M. J., & Arnold, F. H. (2009). Directed evolution: new parts and optimized function. *Current Opinion in Biotechnology*, *20* (4), p.486–91.
71. Doyon, J., Pattanayak, V., Meyer, C. & Liu, D. (2006). Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *Journal of the American Chemical Society*, 128 (7) p.2477–2484.
72. Dubendorff, J. W., & Studier, F. W. (1991). Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor. *Journal of Molecular Biology*, 219 (1) p.45–59.
73. Dunn, J., Krippl, B., Bernstein, K., Westphal, H. (1988). Targeting bacteriophage T7 RNA polymerase to the mammalian cell nucleus. *Gene*, 68 (2) p.259–266.
74. Eckert, K. & Kunkel, T. (1990). High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic acids research*, 18 (13) p.3739–3744.
75. Eckert, K. & Kunkel, T. (1991). DNA polymerase fidelity and the polymerase chain reaction. *PCR methods and applications*, 1 (1) p.17–24.
76. Eggler, A., Lusetti, S. & Cox, M. (2003). The C-terminus of the Escherichia coli RecA protein modulates the DNA binding competition with single-stranded DNA-binding protein. *The Journal of Biological Chemistry*, 278 (18) p.16389–16396.
77. Ellefson, J., Meyer, A., Hughes, R., Cannon, J. (2014). Directed evolution of genetic parts and circuits by compartmentalized partnered replication. *Nature Biotechnology*, 32 (1) p.97–101.
78. Ellington, A. D., & Szostak, J. W. (1990). *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, 346(6287), p.818–22.
79. Ellis, H. M., Yu, D., DiTizio, T., & Court, D. L. (2001). High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (12), p.6742–6
80. Ellis, T., Adie, T. & Baldwin, G.S. (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integrative biology*, 3 (2), p.109–118.

81. Ellis, T., Wang, X., & Collins, J. J. (2009). Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotechnology*, *27* (5), p.465–471.
82. Esvelt, K.M., Carlson, J.C. and Liu, D.R. (2011). A system for the continuous directed evolution of biomolecules. *Nature*, 472, p.499-503.
83. Faili, A. (2002) Induction of somatic hypermutation in immunoglobulin genes is dependent on DNA polymerase I. *Nature*, 419, p.944–947.
84. Farinas, E. T., Bulter, T., & Arnold, F. H. (2001). Directed enzyme evolution. *Current Opinion in Biotechnology*, *12*(6), 545–51.
85. Fijalkowska, I. & Schaaper, R. (1996). Mutants in the Exo I motif of Escherichia coli dnaQ: defective proofreading and inviability due to error catastrophe. *Proceedings of the National Academy of Sciences of the United States of America*, 93 (7) p.2856–2861.
86. Foster, P, L. (2006). Methods for Determining Spontaneous Mutation Rates. *Methods in Enzymology*, 409, 195–213.
87. Friedberg, E, C., Walker, G, C., Siede W. 1995. DNA Repair and Mutagenesis. *ASM Press*
88. Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., Kitano, H. (2008) CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE*, 96 (8) p. 1254-1265
89. Gabrielian, A., & Pongor, S. (1996). Correlation of intrinsic DNA curvature with DNA property periodicity. *FEBS Letters*, 393 (1), p.65–68.
90. Galhardo, R. S., Do, R., Yamada, M., Friedberg, E. C., Hastings, P. J., Nohmi, T., & Rosenberg, S. M. (2009). DinB Upregulation Is the Sole Role of the SOS Response in Stress-Induced Mutagenesis in Escherichia coli. *Genetics*, 182 (1) p.55-68
91. Garcia-Diaz, M., & Bebenek, K. (2007). Multiple functions of DNA polymerases. *Critical Reviews in Plant Sciences*, 26 (2), p.105–122
92. García, L. R., & Molineux, I. J. (1995). Rate of translocation of bacteriophage T7 DNA across the membranes of Escherichia coli. *Journal of Bacteriology*, 177 (14), p.4066.
93. Garibyan, L.. (2003). Use of the rpoB gene to determine the specificity of base substitution mutations on the Escherichia coli chromosome. *DNA Repair*, 2 p.593–608
94. Geissmann, Q., 2013. OpenCFU, a new free and open-source software to count cell colonies and other circular objects. *PloS one*, 8 (2) p.54072.
95. Gen9. (2016) 'Variant Libraries'. Available at: https://www.gen9bio.com/solutions/variant-libraries [Last Accessed 18/08/16].
96. Ghadessy, F. & Holliger, P. (2007). Compartmentalized self-replication: a novel method for the directed evolution of polymerases and other enzymes. *Methods in Molecular Biology*, 352 p.237–248.
97. Ghadessy, F., Ong, J. & Holliger, P. (2001). Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (8) p.4552–4557.
98. Gibson, D. (2011) Enzymatic assembly of overlapping DNA fragments. *Methods in Enzymology*. p.498349–361.
99. Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science (New York, N.Y.)*, *329* (5987), p.52–6.
100. Gibson, D.G., Young, L., Chuang, R.Y. & Venter, J.C. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods,* 6 p.343 - 345
101. Glassner, B., Rasmussen, L., Najarian, M., Posnick, L., (1998). Generation of a strong mutator phenotype in yeast by imbalanced base excision repair. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (17) p.9997–10002.

102. Gold, L., Polisky, B., Uhlenbeck, O., & Yarus, M. (1995). Diversity of oligonucleotide functions. *Annual Review of Biochemistry*, 64 p.763–97.
103. Goodman, M. (2002). Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annual Review of Biochemistry*. 71 p.17–50.
104. Goodman, M.F., Tang, M., Pham, P., Shen, X., Taylor, J, S., O'Donnell, M. and Woodgate, R. (2000). Roles of *E.coli* DNA polymerases IV and V in lesion-targeted and untargeted SOS mutagenesis. *Nature*, 404, p.1014–1018.
105. Goodsell, D, S., & Dickerson, R, E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Research*, 22 (24) p.5497–503.
106. Gossard, F. & Verly, W. (1978). Properties of the main endonuclease specific for apurinic sites of Escherichia coli (endonuclease VI) - Mechanism of apurinic site excision from DNA. *European Journal of Biochemistry*, 82 (2) p.321–332.
107. Gossen M, Freundlieb S, Bender G, Muller G, Hillen W, Bujard H (1995). Transcriptional activation by tetracyclines in mammalian cells. *Science* (268) p.1766−1769
108. Gossen, M. & Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 89 (12), p. 5547–5551.
109. Gossen, M., & Bujard, H. (1993). Anhydrotetracycline, a novel effector for tetracycline controlled gene expression systems in eukaryotic cells. *Nucleic Acids Research*, 21 (18), p.4411–4412.
110. Gottesman, S., Roche, E., Zhou, Y., & Sauer, R. T. (1998). The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes & Development*, 12 (9), p.1338–47.
111. Green B, Belcheva A, Nepal RM, Boulianne B, Martin A. The mismatch repair pathway functions normally at a non-AID target in germinal center B cells. *Blood*, (118), p.3013–8.
112. Greener, A., Callahan, M., & Jerpseth, B. (1997). An efficient random mutagenesis technique using an E. coli mutator strain. *Molecular Biotechnology*, 7 (2), p. 189–95.
113. Griffiths, A.D. & Tawfik, D.S., (2003). Directed evolution of an extremely fast phosphotriesterase by *in vitro* compartmentalization. *The EMBO Journal*, 22 (1), pp.24–35.
114. Gruen, M., Chang, K., Serbanescu, I. & Liu, D. (2002). An *in vivo* selection system for homing endonuclease activity. *Nucleic acids research*, 30 (7) p.29.
115. Guillerez, J., Lopez, P., Proux, F., Launay, H., et al. (2005). A mutation in T7 RNA polymerase that facilitates promoter clearance. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (17) p.5958–5963.
116. Guzman, L. M., Belin, D., Carson, M. J., & Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *Journal of Bacteriology*, *177* (14), p.4121–30.
117. Haldimann, A. & Wanner, B. (2001). Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *Journal of Bacteriology*, 183 (21) p.6384–6393.
118. Hall, B. M., Ma, C.-X., Liang, P., & Singh, K. K. (2009). Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. Bioinformatics (Oxford, England), 25 (12), p1564–5. FALCOR software available at: http://www.mitochondria.org/protocols/FALCOR.html
119. Hamon, A., & Ycart, B. (2012). Statistics for the Luria-Delbrück distribution. *Electronic Journal of Statistics*, *6* (0), p.1251–1272. Bz-rates software package available at: http://www.lcqb.upmc.fr/bzrates
120. Hawkins, A., Arnold, F., Stuermer, R., Hauer, B., et al. (2007) Directed evolution of Vibrio fischeri LuxR for improved response to butanoyl-homoserine lactone. *Applied and Environmental Microbiology*, 73 (18) p.5775–5781.
121. Hegde, M. L., Hazra, T. K., & Mitra, S. (2008). Early Steps in the DNA Base Excision/Single-Strand Interruption Repair Pathway in Mammalian Cells. *Cell Research*, *18* (1), p.27–47.

122. Heltzel, J, M., Maul, R, W., & Wolff, D.W. (2012). Escherichia coli DNA polymerase IV (Pol IV), but not Pol II, dynamically switches with a stalled Pol III* replicase. *Journal of Bacteriology*, 194 (14) p.3589–3600.
123. Hillen, W., Schollmeier, K., & Gatz, C. (1984). Control of expression of the Tn10-encoded tetracycline resistance operon. II. Interaction of RNA polymerase and TET repressor with the tet operon regulatory region. *Journal of Molecular Biology*, *172* (2), p.185–201.
124. Hinrichs, W. (1994). Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance. *Science*, 264 (5157) p.418-20
125. Hirst, C. D. (2014). Automated BioPart Characterisation for Synthetic Biology. PhD Thesis. Available from: https://spiral.imperial.ac.uk/bitstream/10044/1/25961/.../Hirst-CD-2014-PhD-Thesis.
126. Honjo, T., Muramatsu, M. & Fagarasan, S. (2004) AID: how does it aid antibody diversity? *Immunity*, 20 (6) p.659–668.
127. Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Kummer, U. (2006). COPASI—a COmplex PAthway SImulator. *Bioinformatics*, *22*(24) p.3067–3074.
128. Horst, J.P., Wu, T., & Marinus, M. G. (1999). Escherichia coli mutator genes. *Trends in Microbiology*, 7 (1) p.29–36.
129. Hubbard, B.P., Badran, A.H., Zuris, J.A., Guilinger, J.P., Davis, K.M., Chen, L., Tsai, S,Q., Sander, J.D., Joung, K., Liu, D.R. (2015). Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nature Methods,* 12, p.939–942.
130. Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Renaudin, J. (2016). Design and synthesis of a minimal bacterial genome. *Science,* 351 (6280), p.6253.
131. Imburgio, D., Rong, M., Ma, K., & McAllister, W, T. (2000). Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. *Biochemistry*. 39(34), p.10419–30.
132. Inoue, H., Nojima, H., & Okayama, H. (1990). High efficiency transformation of Escherichia coli with plasmids. *Gene*, *96*(1), p.23–8.
133. Iost, I. & Dreyfus, M. (1995). The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation. *The EMBO Journal*. 14 (13) p.3252–3261.
134. Iost, I., Guillerez, J. & Dreyfus, M. (1992). Bacteriophage T7 RNA polymerase travels far ahead of ribosomes *in vivo*. *Journal of Bacteriology*, 174 (2) p.619–622.
135. Ippoliti, P.J., DeLateur, N.A., Jones, K.M. & Beuning, P.J. (2012). Multiple Strategies for Translesion Synthesis in Bacteria. *Cells*, 1 (4) p.799-831
136. Iverson, S.V., Haddock, T.L., Beal, J. and Densmore, D.M., (2015). CIDAR MoClo: improved MoClo assembly standard and new *E. coli* part library enable rapid combinatorial design for synthetic and traditional biology. *ACS Synthetic Biology*, 5(1), p.99-103.
137. Jespers, L. S., Roberts, A., Mahler, S. M., Winter, G. & Hoogenboom, H. R. (1994). Guiding the selection of humanantibodies from phage display repertoires to a single epitope of an antigen. *Biotechnology,* 12 p.899–903.
138. Jia, Y., Kumar, A., & Patel, S. S. (1996). Equilibrium and Stopped-flow Kinetic Studies of Interaction between T7 RNA Polymerase and Its Promoters Measured by Protein and 2-Aminopurine Fluorescence Changes. *Journal of Biological Chemistry*, 271 (48), p.30451–30458.
139. Jiang, Q., Karata, K., Woodgate, R., Cox, M., (2009). The active form of DNA polymerase V is UmuD'(2)C-RecA-ATP. *Nature*, 460 (7253) p.359–363.
140. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotechnolgy, 31 p.233–239.
141. Jin, P., Ding, W., Du, G., Chen, J. and Kang, Z. (2016). DATEL: a scarless and sequence-independent DNA assembly method using thermostable exonucleases and ligase. *ACS Synthetic Biology*, 5 (9), p.1028-1032.

142. Joho, K., Gross, L., McGraw, N., Raskin, C. (1990). Identification of a region of the bacteriophage T3 and T7 RNA polymerases that determines promoter specificity. *Journal of Molecular Biology*, 215 (1) p.31–39.
143. Kamionka, A., Bogdanska-Urbaniak, J., Scholz, O., & Hillen, W. (2004). Two mutations in the tetracycline repressor change the inducer anhydrotetracycline to a corepressor. *Nucleic Acids Research*, 32 (2), p.842–847.
144. Kelly, J. R., Rubin, A. J., Davis, J. H., Ajo-Franklin, C. M., Cumbers, J., Czar, M. J., Alon, U. (2009). Measuring the activity of BioBrick promoters using an *in vivo* reference standard. *Journal of Biological Engineering*, 3 (1), p.4.
145. Khare, V., Eckert, K., (2002). The proofreading 3'-5' exonuclease activity of DNA polymerases: a kinetic barrier to translesion DNA synthesis. *Mutation research*, 510 (1-2) p. 45-54
146. Kim, S., Maenhaut-Michel, G., Yamada, M., Yamamoto, Y., (1997). Multiple pathways for SOS-induced mutagenesis in Escherichia coli: an overexpression of dinB/dinP results in strongly enhancing mutagenesis in the absence of any exogenous treatment to damage DNA. *Proceedings of the National Academy of Sciences of the United States of America*. 94 (25) p.13792–13797.
147. Kim, S.R., Matsui, K., Yamada, M., Gruz, P., et al. (2001) Roles of chromosomal and episomal dinB genes encoding DNA pol IV in targeted and untargeted mutagenesis in Escherichia coli. *Molecular genetics and genomics : MGG*. [Online] 266 (2), 207–215.
148. Klenow, H. & Henningsen, I. (1970). Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. *Proceedings of the National Academy of Sciences of the United States of America*, 65 (1) p.168–175.
149. Klenow, H., Overgaard-Hansen, K. & Patkar, S. (1971). Proteolytic cleavage fo native DNA polymerase into two different catalytic fragments. Influence of assay condtions on the change of exonuclease activity and polymerase activity accompanying cleavage. *European Journal of Biochemistry*, 22 (3) p.371–381.
150. Knight, T. (2003). Idempotent vector design for standard assembly of biobricks. Available from: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA457791.
151. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 533 (7603), p.420–4.
152. Kramer, B. P., & Fussenegger, M. (2005). Hysteresis in a synthetic mammalian gene network. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (27), p.9517–22.
153. Krokan, H., Bjørås, M. (2013). Base excision repair. *Cold Spring Harbor perspectives in biology*, 5 (4).
154. Kuban, W., Banach-Orlowska, M., Schaaper, R. M., Jonczyk, P., & Fijalkowska, I. J. (2006). Role of DNA Polymerase IV in Escherichia coli SOS Mutator Activity. *Journal of Bacteriology*, *188* (22), p.7977–7980.
155. Kumamaru, T., Suenaga, H., Mitsuoka, M., Watanabe, T. (1998). Enhanced degradation of polychlorinated biphenyls by directed evolution of biphenyl dioxygenase. *Nature Biotechnology*, 16 (7) p.663–666.
156. Kunkel, T. A. (2009). Evolving views of DNA replication (in) fidelity. *Cold Spring Harbor Symposia on Quantitative Biology*, *74* p.91–101.
157. LabGenius Ltd. (2016) 'The world's most powerful synthetic DNA libraries, ready to screen'. Available at: http://www.labgeni.us/labgene/about/ [Last Accessed 18/08/2016].
158. Lane, M. D., & Seelig, B. (2014). Advances in the directed evolution of proteins. *Current Opinion in Chemical Biology*, 22 p.129–136.
159. Larijani, M. & Martin, A. (2012). The biochemistry of activation-induced deaminase and its physiological functions. *Seminars inI Immunology*, 24 (4) p.255–263.
160. Larijani, M., Petrov, A., Kolenchenko, O., Berru, M., *et al.* (2007). AID associates with single-stranded DNA with high affinity and a long complex half-life in a sequence-independent manner. *Molecular and Cellular Biology*, 27 (1) p.20–30.

161. Layton, J. C., & Foster, P. L. (2003). Error-prone DNA polymerase IV is controlled by the stress-response sigma factor, RpoS, in Escherichia coli. *Molecular Microbiology*, 50(2), p.549–561.
162. Lea, D.E., & Coulson, C. A. (1949). The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics*, 49(3), p.264–85.
163. Lenne-Samuel, N., Wagner, J., Etienne, H. & Fuchs, R.P. (2002) The processivity factor beta controls DNA polymerase IV traffic during spontaneous mutagenesis and translesion synthesis *in vivo*. *EMBO reports*, 3 (1) p.45–49.
164. Lies, M., & Maurizi, M. R. (2008). Turnover of endogenous SsrA-tagged proteins mediated by ATP-dependent proteases in Escherichia coli. *The Journal of Biological Chemistry*, 283 (34), p.22918–29.
165. Lindahl, T., Demple, B., & Robins, P. (1982). Suicide inactivation of the E. coli O6-methylguanine-DNA methyltransferase. *The EMBO Journal*, 1 (11), p.1359–1363.
166. Ling, H., Boudsocq, F., Woodgate, R. & Yang, W. (2001). Crystal structure of a Y-family DNA polymerase in action: a mechanism for error-prone and lesion-bypass replication. *Cell*, 107 (1) p.91–102.
167. Luria, S. E., & Delbrück, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, 28 (6), p.491–511.
168. Lutsenko, E., Bhagwat, A. (1999). The role of the Escherichia coli mug protein in the removal of uracil and 3, N(4)-ethenocytosine from DNA. *The Journal of Biological Chemistry*, 274 (43) p.31034
169. Lutz, R. & Bujard, H. (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic acids research*, 25 (6), p.1203–1210.
170. MacBeath, G., Kast, P. & Hilvert, D. (1998) Redesigning enzyme topology by directed evolution. *Science*, 279 (5358) p.1958–1961.
171. Mackie, A., Paley, S., Keseler, I., Shearer, A., Paulsen, I. (2014). Addition of Escherichia coli K-12 Growth Observation and Gene Essentiality Data to the EcoCyc Database. *Journal of Bacteriology*, 196 (5) p.982-988
172. Makarova, O., Makarov, E., Sousa, R. & Dreyfus, M. (1995) Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. *Proceedings of the National Academy of Sciences of the USA,* 92 (26) p.12250–12254.
173. Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., … Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science*, 339 (6121) p.823–826.
174. Marillonnet, S., & Werner, S. (2015). Assembly of Multigene Constructs Using Golden Gate Cloning. *Methods in Molecular Biology*, *1321* p.269–84.
175. Marinus, M. G. (2010). DNA methylation and mutator genes in *Escherichia coli* K-12. *Mutation Research/Reviews in Mutation Research*, 705 (2) p.71–76.
176. Maurizi, M. R. (1992). Proteases and protein degradation in *Escherichia coli*. *Experientia*, 48(2), p.178–201.
177. McHenry, C., & Kornberg, A. (1977). DNA polymerase III holoenzyme of Escherichia coli. Purification and resolution into subunits. *Journal of Biological Chemistry*, 252 (18) p.6478–6484.
178. McIsaac, R. S. (2014). Directed evolution of a far-red fluorescent rhodopsin. *Proceedings of the National Academy of Science USA,* 111, p.13034–13039.
179. McMurry, L., & Levy, S. B. (1978). Two transport systems for tetracycline in sensitive Escherichia coli: critical role for an initial rapid uptake system insensitive to energy inhibitors. Antimicrobial Agents and Chemotherapy, 14(2), p.201–9.
180. Meyers, G., Ng, Y.S., Bannock, J., Lavoie, A., (2011). Activation-induced cytidine deaminase (AID) is required for B-cell tolerance in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (28) p.11554–11559.

181. Miller, O., Bernath, K., Agresti, J., Amitai, G. (2006). Directed evolution by *in vitro* compartmentalization. *Nature Methods*, 3 (7) p.561–570.
182. Mills, D. R., Peterson, R. L., & Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences of the United States of America*, *58* (1), p.217–24.
183. Minshull, J., Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Stemmer, W. P. C. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nature Biotechnology*, 17(9), p.893–896.
184. Mitchell, L.A., Chuang, J., Agmon, N., Khunsriraksakul, C., Phillips, N.A., Cai, Y., Truong, D.M., Veerakumar, A., Wang, Y., Mayorga, M. and Blomquist, P., (2015). Versatile genetic assembly system (VEGAS) to assemble pathways for expression in S. cerevisiae. *Nucleic acids research*, 43 (13), p.6620-6630.
185. Mol, C., Kuo, C., Thayer, M., Cunningham, R. (1995) Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature*, 374 (6520) p.381–386.
186. Mottram, J. C., North, M. J., Barry, J. D., & Coombs, G. H. (1989). A cysteine proteinase cDNA from Trypanosoma brucei predicts an enzyme with an unusual C-terminal extension. *FEBS Letters*, 258 (2), p.211–5.
187. Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., et al. (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, 102 (5) p.553–563.
188. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q.A., Endy, D. (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature Methods*, *10*(4), p.354–60. http://doi.org/10.1038/nmeth.2404
189. Myers, R. M., Lerman, L. S., & Maniatis, T. (1985). A general method for saturation mutagenesis of cloned DNA fragments. *Science*, 229 (4710), p.242–7.
190. Nair, N.U. & Zhao, Huimin, (2008). Evolution in reverse: engineering a D-xylose-specific xylose reductase. *Chembiochem : a European journal of chemical biology*, 9(8), p.1213–5.
191. Neidhardt, F,C., Bloch, P, L., & Smith, D, F., (1974). Culture medium for enterobacteria. *Journal of Bacteriology*, 119 (3) p.736-747
192. Nishida, K., Arazoe, T., Yachie, N., Banno, S., Kakimoto, M., Tabata, M., Kondo, A. (2016). Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science, 353* (6305)
193. Noia, J.D. & Neuberger, M.S. (2002) Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature*, 419 (6902) p.43-48
194. O'Maille, P.E., Bakhtina, M., Tsai M.D. (2002) Structure-based combinatorial protein engineering (SCOPE). *Journal of Molecular Biology*, 321 (4) p.677-91.
195. Opperman, T., Murli, S., Smith, B. & Walker, G. (1999). A model for a umuDC-dependent prokaryotic DNA damage checkpoint. *Proceedings of the National Academy of Sciences of the United States of America*, 96 (16) p.9218–9223.
196. Orth, P., Saenger, W., Hinrichs, W., Schnappinger, D., & Hillen, W. (2000). Structural basis of gene regulation by the tetracycline inducible Tetrepressor operator system. *Nature Structural Biology*, 7 (3), p.215–219.
197. Orth, P., Schnappinger, D., Hillen, W., Saenger, W., & Hinrichs, W. (2000). Structural basis of gene regulation by the tetracycline inducible Tet repressor-operator system. *Nature Structural Biology*, *7* (3), p.215–9.
198. Pata, J. D. (2010). Structural Diversity of the Y-family DNA Polymerases. *Biochimica et Biophysica Acta*, *1804*(5), 1124.
199. Pérez-Durán, P., Belver, L., de Yébenes, V., Delgado, P. (2012). UNG shapes the specificity of AID-induced somatic hypermutation. *The Journal of Experimental Medicine*, 209 (7) p.1379–1389.
200. Petersen-Mahrt, S., Harris, R. & Neuberger, M. (2002) AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature*, 418 (6893) p.99–103.

201. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*, 424 (6944) p.103–107.
202. Pham, P., Calabrese, P., Park, S. & Goodman, M. (2011). Analysis of a single-stranded DNA-scanning process in which activation-induced deoxycytidine deaminase (AID) deaminates C to U haphazardly and inefficiently to ensure mutational diversity. *The Journal of Biological Chemistry*, 286 (28) p.24931–24942.
203. Poltoratsky, V., Woo, C, J., Tippin, B., Martin, A., Goodman, M, F., Scharff, M, D. (2001). Expression of error-prone polymerases in BL2 cells activated for Ig somatic hypermutation. *Proceedings of the National Academy of Sciences of the United States of America,* 98:7976–81.
204. Porecha, R, H., & Stivers, J, T. (2008). Uracil DNA glycosylase uses DNA hopping and short-range sliding to trap extrahelical uracils. *Proceedings of the National Academy of Sciences*, *105* (31), p.10791–10796.
205. Rada, C., Williams, G., Nilsen, H., Barnes, D., (2002). Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. *Current Biology*, (20) p.1748–1755.
206. Ramiro, A., Stavropoulos, P., Jankovic, M. & Nussenzweig, M. (2003). Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nature Immunology*. 4 (5) p.452–456.
207. Raskin, C., Diaz, G. & McAllister, W. (1993). T7 RNA polymerase mutants with altered promoter specificities. *Proceedings of the National Academy of Sciences of the United States of America*. 90 (8) p.3147–3151.
208. Ravikumar, A., Arrieta, A., & Liu, C. C. (2014). An orthogonal DNA replication system in yeast. *Nature Chemical Biology*, 10 (3), p.175–7.
209. Revy, P., Muto, T., Levy, Y., Geissmann, F., et al. (2000). Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell*, 102 (5) p.565–575.
210. Riedel, K., Steidle, A., Eberl, L., Wu, H., Geisenberger, O., Molin, S., Huber, B., Givskov, M. (2001). N-Acylhomoserine-lactone-mediated communication between Pseudomonas aeruginosa and Burkholderia cepacia in mixed biofilms. *Microbiology*, *147* (12), p.3249–3262.
211. Robertson, D. L., & Joyce, G. F. (1990). Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, *344* (6265), p.467–8.
212. Rosche, W. A., & Foster, P. L. (2000). Determining mutation rates in bacterial populations. *Methods*, 20 (1), p.4–17.
213. Salis, H. M., Mirsky, E. A., & Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, *27* (10), p.946–50.
214. Sarkar, S., Ma, W. T., & Sandri, G. H. (1992). On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica*, 85 (2), p.173–9.
215. Savva, R. and Pearl, L.H. (1995) Nucleotide mimicry in the crystal-structure of the uracil-dna glycosylase-uracil glycosylase inhibitor protein complex. *Nature Structural Biology*, 2 p.752-757
216. Schmidt-Dannert, C., Umeno, D., & Arnold, F. H. (2000). Molecular breeding of carotenoid biosynthetic pathways. *Nature Biotechnology*, *18* (7) p.750–3.
217. Schuster, M., Urbanowski, M., & Greenberg EP. (2004). Promoter specificity in Pseudomonas aeruginosa quorum sensing revealed by DNA binding of purified LasR.. *Proc National Academic Sciences USA*, 101 (45), p.15833–9.
218. Seeberg, E., Eide, L. and Bjoras, M. (1995). The base excision repair pathway. *Trends in Biochemical Science*, 20 p.391-397.
219. Setlow, P. & Kornberg, A. (1972). Deoxyribonucleic acid polymerase: two distinct enzymes in one polypeptide. II. A proteolytic fragment containing the 5’ leads to 3' exonuclease function. Restoration of intact enzyme functions from the two proteolytic fragments. *The Journal of biological chemistry*, 247 (1) p.232–240
220. Sezonov, G., Joseleau-Petit, D., & D’Ari, R. (2007). Escherichia coli Physiology in Luria-Bertani Broth. *Journal of Bacteriology*, 189 (23), p.8746–8749.

221. Shao, Z., Zhao, H., Giver, L. & Arnold, F. (1998) Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic acids research*, 26 (2) p.681–683.
222. Shen, H., Ratnam, S. & Storb, U. (2005) Targeting of the activation-induced cytosine deaminase is strongly influenced by the sequence and structure of the targeted DNA. *Molecular and cellular biology*, 25 (24) p.10815–10821.
223. Shetty, R.P., Endy, D. & Knight, T.F. (2008) Engineering BioBrick vectors from BioBrick parts. *Journal of Biological Engineering*, 2 (4)
224. Singh, V. (2014). Recent advances and opportunities in synthetic logic gates engineering in living cells. *Systems and Synthetic Biology*, 8 (4), p.271–282.
225. Skerra, A. (1994) Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in Escherichia coli. *Gene* 151, p.131-135.
226. Smyth, D. R., Mrozkiewicz, M. K., McGrath, W. J., Listwan, P., & Kobe, B. (2003). Crystal structures of fusion proteins with large-affinity tags. *Protein Science: A Publication of the Protein Society*, 12 (7), p.1313–22.
227. Sohail, A., Klapacz, J., Samaranayake, M., Ullah, A. and Bhagwat, A.S. (2003). Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. *Nucleic Acids Research*, 31, p.2990-2994
228. Sousa, R. & Padilla, R. (1995). A mutant T7 RNA polymerase as a DNA polymerase. *The EMBO Journal*, 14 (18) p.4609–4621.
229. Sousa, R., Chung, Y., Rose, J. & Wang, B. (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. *Nature*, 364 (6438) p.593–599.
230. St-Pierre, F., Cui, L., Priest, D. G., Endy, D., Dodd, I. B., & Shearwin, K. E. (2013). One-Step Cloning and Chromosomal Integration of DNA. *ACS Synthetic Biology*, 2 (9), 537–541.
231. Stemmer, W. (1994). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, [370 (6488) p.389–391.
232. Storch M., Casini, A., Mackrow B., Trewhitt H., Ellis, T., Baldwin, G., (2014), BASIC: a new Biopart Assembly Standard provides accurate, single-tier DNA assembly for synthetic biology. *ACS Synthetic Biology*, 4 (7) p.781-7.
233. Strauss, B., Roberts, R., Francis, L. & Pouryazdanparast, P. (2000). Role of the *dinB* gene product in spontaneous mutation in Escherichia coli with an impaired replicative polymerase. *Journal of bacteriology*. 182 (23) p.6742–6750.
234. Studier, F. W. (1991) Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *Journal of Molecular Biology*, 219 (1), p37–44.
235. Svilar, D., Goellner, E. M., Almeida, K. H., & Sobol, R. W. (2011). Base Excision Repair and Lesion-Dependent Subpathways for Repair of Oxidative DNA Damage. *Antioxidants & Redox Signaling*, *14* (12), p. 2491–2507.
236. Tang, M., Pham, P., Shen, X., Taylor, J., (2000) Roles of E. coli DNA polymerases IV and V in lesion-targeted and untargeted SOS mutagenesis. *Nature*, 404 (6781) p.1014–1018.
237. Tawfik, D. & Griffiths, A. (1998). Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, 16 (7) p.652–656.
238. Taylor, A. & Weiss, B. (1982) Role of exonuclease III in the base excision repair of uracil-containing DNA. *Journal of Bacteriology*, 151 (1) p.351–357.
239. Temme, K., Hill, R., Segall-Shapiro, T. H., Moser, F., & Voigt, C. A. (2012). Modular control of multiple pathways using engineered orthogonal T7 polymerases. *Nucleic Acids Research*, 40 (17) p.8773-8781
240. Terho, P., 2013. Flowing Software Version 2.5.1. Turku Centre for Biotechnology. Available from http://flowingsoftware.btk.fi/index.php?page=3
241. Tiebel, B., Garke, K. & Hillen, W. (2000). Observing conformational and activity changes of Tet repressor *in vivo*. Nature Structural Biology, 7 (6) p.479-81

242. Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249 (4968), p.505–10.

243. Turner, S. L., Russell, G. C., Williamson, M. P., & Guest, J. R. (1993). Restructuring an interdomain linker in the dihydrolipoamide acetyltransferase component of the pyruvate dehydrogenase complex of Escherichia coli. *Protein Engineering*, 6 (1), p.101–8.

244. Tuteja, N., Tuteja, R. (2004). Prokaryotic and eukaryotic DNA helicases. Essential molecular motor proteins for cellular machinery. *European Journal of Biochemistry*, 271 (10) p. 1835-1848

245. Twist Bioscience. (2016) 'Libraries'. Available at: https://www.twistbioscience.com/products/libraries/ [Last Accessed 18/08/2016].

246. Uchida, K., Furukohri, A., Shinozaki, Y., Mori, T. (2008). Overproduction of Escherichia coli DNA polymerase DinB (Pol IV) inhibits replication fork progression and is lethal. *Molecular microbiology*. 70 (3) p.608–622.

247. Uptain, S., Kane, C. & Chamberlin, M. (1997) Basic mechanisms of transcript elongation and its regulation. *Annual review of biochemistry*, 66 p.117–172.

248. van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Wei, W. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, *63* (4), p.633–646.

249. Vidal, M., & Legrain, P. (1999). Yeast forward and reverse "n"-hybrid systems. *Nucleic Acids Research*, *27*(4), 919–29.

250. Viswanathan, A., Anufrieva, O., Sala, A., Yli-Harja, O., & Kandhavelu, M. (n.d.). Phase-dependent dynamics of the lac promoter under nutrient stress. *Research in Microbiology*, *167* (6), p.451–61.

251. Volkmer, B., & Heinemann, M. (2011). Condition-Dependent Cell Volume and Concentration of Escherichia coli to Facilitate Data Conversion for Systems Biology Modeling. *PLoS ONE*, 6 (7), e23126.

252. Wagner, J, & Meselson, M. (1976). Repair tracts in mismatched DNA heteroduplexes. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(11), p.4135.

253. Wagner, J., Gruz, P., Kim, S., Yamada, M., et al. (1999) The *dinB* gene encodes a novel E. coli DNA polymerase, DNA pol IV, involved in mutagenesis. Molecular cell. [Online]. 4 (2) p.281–286.

254. Wagner,J. and Nohmi,T. (2000) Escherichia coli DNA polymerase IV mutator activity: genetic requirements and mutational specificity. Journal of Bacteriology, 182, p.4587–4595.

255. Wagner,J., Fujii,S., Gruz,P., Nohmi,T. and Fuchs,R. (2000) The β clamp targets DNA polymerase IV to DNA and strongly increases its processivity. *EMBO reports*, 1(6) p.484–488.

256. Waldminghaus, T., & Skarstad, K. (2009). The Escherichia coli SeqA protein. *Plasmid*, *61* (3), p.141–150.

257. Wang, B., Kitney, R.I., Joly, N., Buck, M., (2011). Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nature Communications,* 2, p.508

258. Wang, H., Isaacs, F., Carr, P., Sun, Z., (2009). Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, 460 (7257) p.894–898.

259. Wang, L., Brock, A., Herberich, B., & Schultz, P. G. (2001). Expanding the genetic code of Escherichia coli. *Science,* 292 (5516), p.498–500.

260. Wang, M., Yang, Z., Rada, C., & Neuberger, M. S. (2009). AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. *Nature Structural & Molecular Biology*, 16 (7), p.769–76.

261. Wang, Z. & Mosbaugh, D. (1988). Uracil-DNA glycosylase inhibitor of bacteriophage PBS2: cloning and effects of expression of the inhibitor gene in *Escherichia coli*. *Journal of Bacteriology*, 170 (3) p.1082–1091.

262. Warner, H., Duncan, B., Garrett, C. & Neuhard, J. (1981) Synthesis and metabolism of uracil-containing deoxyribonucleic acid in *Escherichia coli*. *Journal of Bacteriology.* 145 (2) p.687–695.

263. Weiss R (2001) Cellular computation and communications using engineered genetic regulatory networks. Ph.D. thesis, Massachusetts Institute of Technology.
264. Weiss,R., Basu, S. (2002). The device physics of cellular logic gates. In: NSC-1: The First Workshop of Non-Silicon Computing. Boston, MA
265. Wells, J. A., Vasser, M., & Powers, D. B. (1985). Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene*, 34(2-3), p.315–23.
266. Williamson, M. P. (1994). The structure and function of proline-rich regions in proteins. *The Biochemical Journal*, 249–60.
267. Wittrup,K.D. (2001). Protein engineering by cell-surface display. *Current Opinion in Biotechnology*, 12 p.395 - 399.
268. Wong, T. S., Roccatano, D., Zacharias, M., & Schwaneberg, U. (2006). A statistical analysis of random mutagenesis methods used for directed protein evolution. *Journal of Molecular Biology*, 355 (4), 858–71.
269. Wood, R. D. (1996). DNA Repair in Eukaryotes. *Annual Review of Biochemistry*, *65* (1), 135–167.
270. Woodgate, R., Singh, M., Kulaeva, O., Frank, E., et al. (1994) Isolation and characterization of novel plasmid-encoded umuC mutants. *Journal of Bacteriology*, 176 (16) p.5011–5021.
271. Xie, K., Sowden, M., Dance, G., Torelli, A. (2004) The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (21) p.8114–8119.
272. Yang, L., Briggs, A., Chew, W. L., Mali, P., Guell, M., Aach, J., Church, G. (2016). Genome Editing With Targeted Deaminases. *bioRxiv*. Available at: http://biorxiv.org/content/early/2016/07/28/066597
273. Yano, T., Oue, S. & Kagamiyama, H. (1998). Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (10) p.5511–5515.
274. Yokobayashi, Y., Collins, C.H., Leadbetter, J.R., Arnold, F.H., Et Al. (2003) Evolutionary Design Of Genetic Circuits And Cell-Cell Communications. *Advances in Complex Systems*.
275. Yokobayashi, Y., Weiss, R. & Arnold, F. (2002) Directed evolution of a genetic circuit. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (26) p.16587–16591.
276. Yoshikawa, K., Okazaki, I. M., Eto, T., Kinoshita, K. (2002). AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. *Science*, 296 (5575) p.2033–2036.
277. Zan H, Shima N, Xu Z, Al-Qahtani A, Evinger Iii AJ, Zhong Y. (2005). The translesion DNA polymerase theta plays a dominant role in immunoglobulin gene somatic hypermutation. *The EMBO Journal*, 24 p.3757–69.
278. Zan, H., Komori, A., Li, Z., Cerutti, A., (2001). The translesion DNA polymerase zeta plays a major role in Ig and bcl-6 somatic hypermutation. *Immunity*. 14 (5) p.643–653.
279. Zeng, X., Negrete, G. A., Kasmer, C., Yang, W. W., & Gearhart, P. J. (2004). Absence of DNA Polymerase η Reveals Targeting of C Mutations on the Nontranscribed Strand in Immunoglobulin Switch Regions. *Journal of Experimental Medicine*, 199 (7) p.917-924.
280. Zhang, Y.-X., Perry, K., Vinci, V. A., Powell, K., Stemmer, W. P. C., & del Cardayré, S. B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, 415 (6872), p.644–6.
281. Zhao, H. and Zha, W. (2004) Evolutionary Methods for Protein Engineering. *Enzyme functionality*. Markus Deker, New York, p.353.
282. Zhao, H., and F. H. Arnold. (1997). Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Research*, 25 (6) p.1307–1308.
283. Zhou, B., Pata, J. & Steitz, T. (2001) Crystal structure of a DinB lesion bypass DNA polymerase catalytic fragment reveals a classic polymerase catalytic domain. *Molecular cell*, 8 (2) p.427–437.

284. Zhou, Y. H., Zhang, X. P., & Ebright, R. H. (1991). Random mutagenesis of gene-sized DNA molecules by use of PCR with Taq DNA polymerase. *Nucleic Acids Research*, *19* (21) p.6052.

11 Appendices

11.1 Relating to Chapter 2 – Creating a Mutator Device

Table 11.1 *– Complete data set from 'loss of function' GFP knockout mutator assay.*

T7	All events: Events	GFP+: Events	GFP+: % of Vis	GFP-: Events	GFP-: % of Vis
01A0101.001	10698	10557	98.68	141	1.32
01B0101.010	7839	7670	97.84	169	2.16
01C0101.019	9463	9379	99.11	84	0.89
01D0101.028	8823	8751	99.18	72	0.82
01E0101.037	8474	8413	99.28	61	0.72
01F0101.046	5164	5056	97.91	108	2.09
01G0101.055	8291	8222	99.17	69	0.83
01H0101.064	4606	4517	98.07	89	1.93
AT F					
01A0201.002	10718	9991	93.22	727	6.78
01B0201.011	9973	9175	92.00	798	8.00
01C0201.020	11627	10200	87.73	1427	12.27
01D0201.029	9441	8738	92.55	703	7.45
01E0201.038	10982	9973	90.81	1009	9.19
01F0201.047	9161	7714	84.20	1447	15.80
01G0201.056	6731	6367	94.59	364	5.41
01H0201.065	9045	8193	90.58	852	9.42
A+T UN					
01A0301.003	6422	6075	94.60	347	5.40
01B0301.012	4638	4353	93.86	285	6.14
01C0301.021	8851	8637	97.58	214	2.42
01D0301.030	6779	6595	97.29	184	2.71
01E0301.039	11530	11187	97.03	343	2.97
01F0301.048	5143	5103	99.22	40	0.78
01G0301.057	7928	7772	98.03	156	1.97
01H0301.066	9129	9002	98.61	127	1.39
AAT F					
01A0501.005	9869	9472	95.98	397	4.02
01B0501.014	9560	9128	95.48	432	4.52
01C0501.023	8749	8469	96.80	280	3.20
01D0501.032	8511	6876	80.79	1635	19.21
01E0501.041	6683	5580	83.50	1103	16.50
01F0501.050	10676	10252	96.03	424	3.97
01G0501.059	11331	10991	97.00	340	3.00
01H0501.068	7620	7302	95.83	318	4.17
AA+T UN					
01A0601.006	10775	10567	98.07	208	1.93
01B0601.015	9813	9595	97.78	218	2.22
01C0601.024	7858	7762	98.78	96	1.22
01D0601.033	10170	9897	97.32	273	2.68
01E0601.042	8573	8461	98.69	112	1.31
01F0601.051	9177	8699	94.79	478	5.21
01G0601.060	7588	7399	97.51	189	2.49
01H0601.069	6633	6376	96.13	257	3.87
Negative (AA)					
01A0401.004	6052	10	0.17	6042	99.83
01B0401.013	9776	4	0.04	9772	99.96
01C0401.022	10199	2	0.02	10197	99.98
01D0401.031	10517	17	0.16	10500	99.84
01E0401.040	8310	3	0.04	8307	99.96
01F0401.049	10071	17	0.17	10054	99.83

01G0401.058	7290	5	0.07	7285	99.93
01H0401.067	9481	4	0.04	9477	99.96

Table 11.2 *– Average data sets across a 5 day loss of function mutator assay. Flow cytometry was used to count at least 5000 true E.coli events and gated into GFP+ and GFP- populations. Standard deviation and standard errors are shown. Note that AID T7 unfused constructs shows almost a completely non GFP population by day 5. This is actually in response to critical T7RNApol mutations that create a selection bias, and thus propagate as a consequence of reduced burden on the cell.*

Day 1						
Construct	GFP+	GFP+ (%)	GFP-	GFP- (%)	(%) stdev	(%) sterror
T7	7821	99	101	1	0.62	0.22
ATF	8794	91	916	9	3.33	1.18
A+T UN	7341	97	212	3	1.88	0.66
AATF	8509	93	616	7	6.56	2.32
AA+T UN	8595	97	229	3	1.34	0.47
Negative (AA)	8	0	8954	100	6.18	2.19
Day 2						
Construct	GFP+	GFP+ (%)	GFP-	GFP- (%)	(%) stdev	(%) sterror
T7 (Positive)	9553	97	278	3	2.07	0.73
ATF	9718	83	2041	17	4.90	1.73
A+T UN	8544	78	2456	22	22.81	8.07
AATF	8340	85	1397	15	15.42	5.45
AA+T UN	9942	98	255	2	1.18	0.42
AA (Negative)	18	0	10095	100	0.11	0.04
Day 3						
Construct	GFP+	GFP+ (%)	GFP-	GFP- (%)	(%) stdev	(%) sterror
T7 (Positive)	8090	94	522	6	3.98	1.41
ATF	4092	59	2804	41	7.13	2.52
A+T UN	4639	49	5316	51	22.98	8.13
AATF	86	86	86	86	16.46	5.82
AA+T UN	7685	79	812	21	32.08	11.34
AA (Negative)	9	0	9522	100	0.07	0.03
Day 4						
Construct	GFP+	GFP+ (%)	GFP-	GFP- (%)	(%) stdev	(%) sterror
T7 (Positive)	5906	88	860	12	15.77	5.57
ATF	3994	60	2502	40	16.58	5.86
A+T UN	1735	26	4881	74	16.96	6.00
AATF	6333	79	1506	21	19.43	6.87
AA+T UN	5871	92	487	8	3.15	1.11
AA (Negative)	6	0	5804	100	0.08	0.03
Day 5						
Construct	GFP+	GFP+ (%)	GFP-	GFP- (%)	(%) stdev	(%) sterror
T7 (Positive)	7608	83	1694	17	25.28	8.94
ATF	5002	46	5594	54	19.64	6.94
A+T UN	621	7	7616	93	5.99	2.12
AATF	6715	74	2292	26	23.69	8.38
AA+T UN	8951	92	795	8	5.60	1.98
AA (Negative)	1	0	11457	100	0.01	0.00

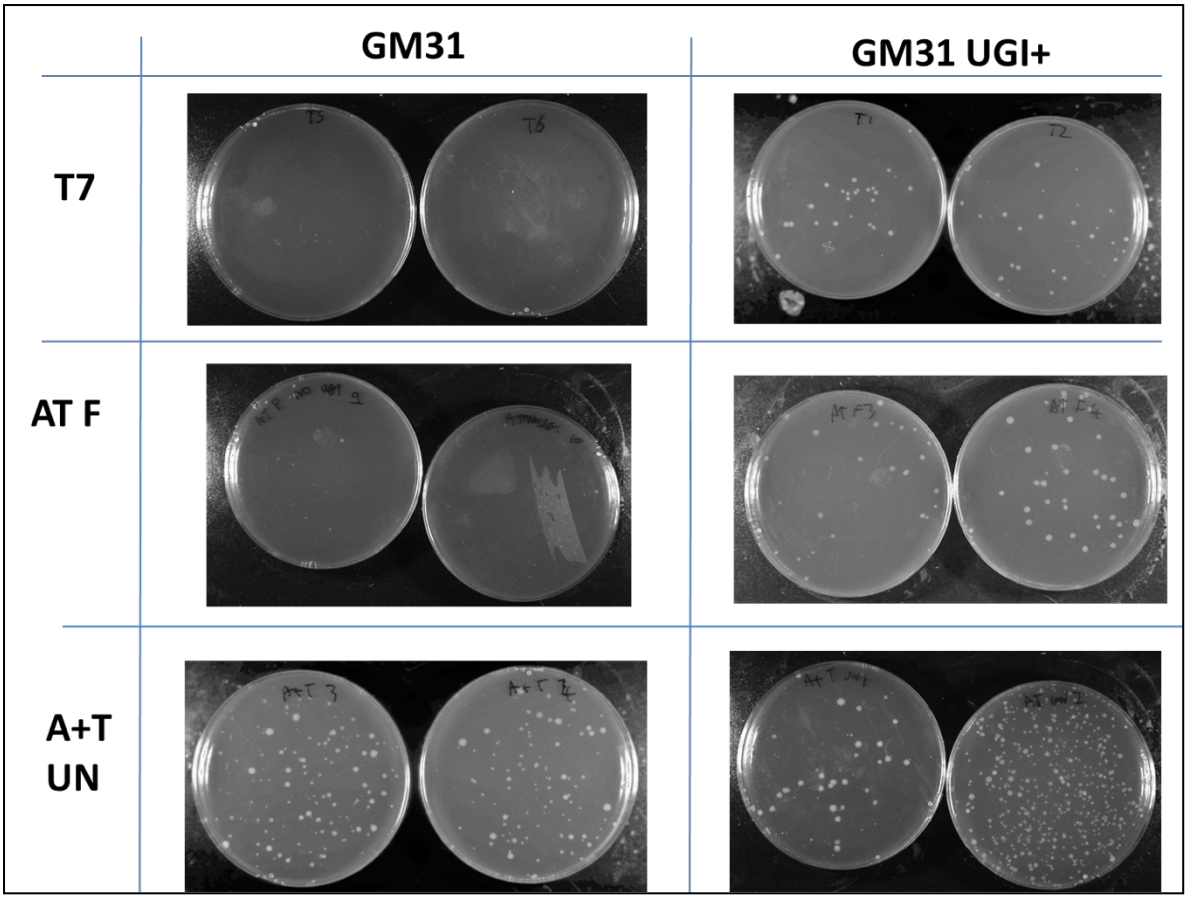


Figure 11.1 *– Examples test plates from Rifampicin reversion assays. Images of each plate were analysed using colony counting software.*

Table 11.3 - Primers used to amplify target GFP gene from the E.coli genome post mutation

Post Mutation – Colony PCR	Direction	Sequence (5' – 3')
Inner	For	CCAGTGCCAAGCTTGCATGC
Inner	Rev	CCTCGCTTTGTAACGGAGTAGAGACG
Outer	For	GCGAGAGTAGGGAACTGCCAGGC
Outer	Rev	GCAAGGCTATGTGCCATCTCGATACTCG

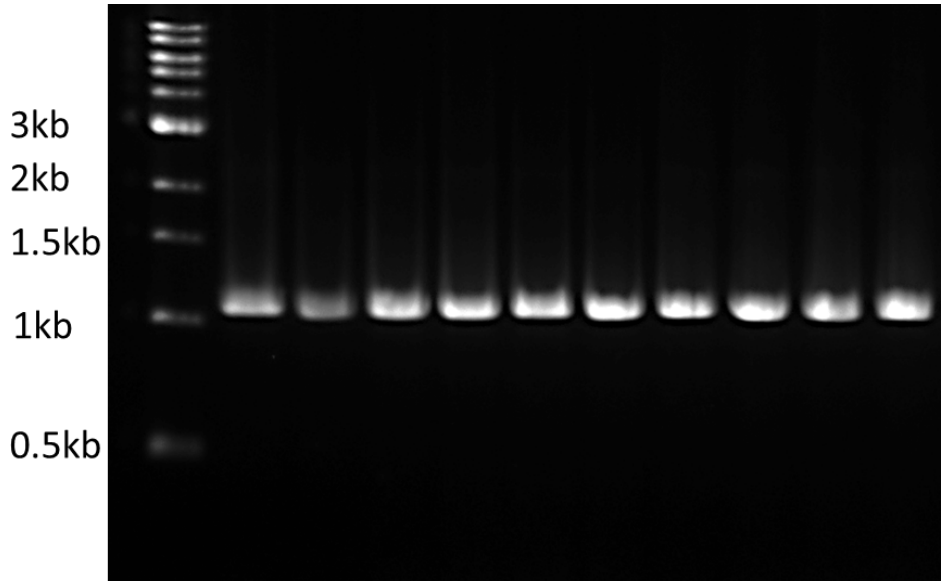


Figure 11.2 *– Example gel for colony PCR used to amplified variants of the target gene – post mutation*

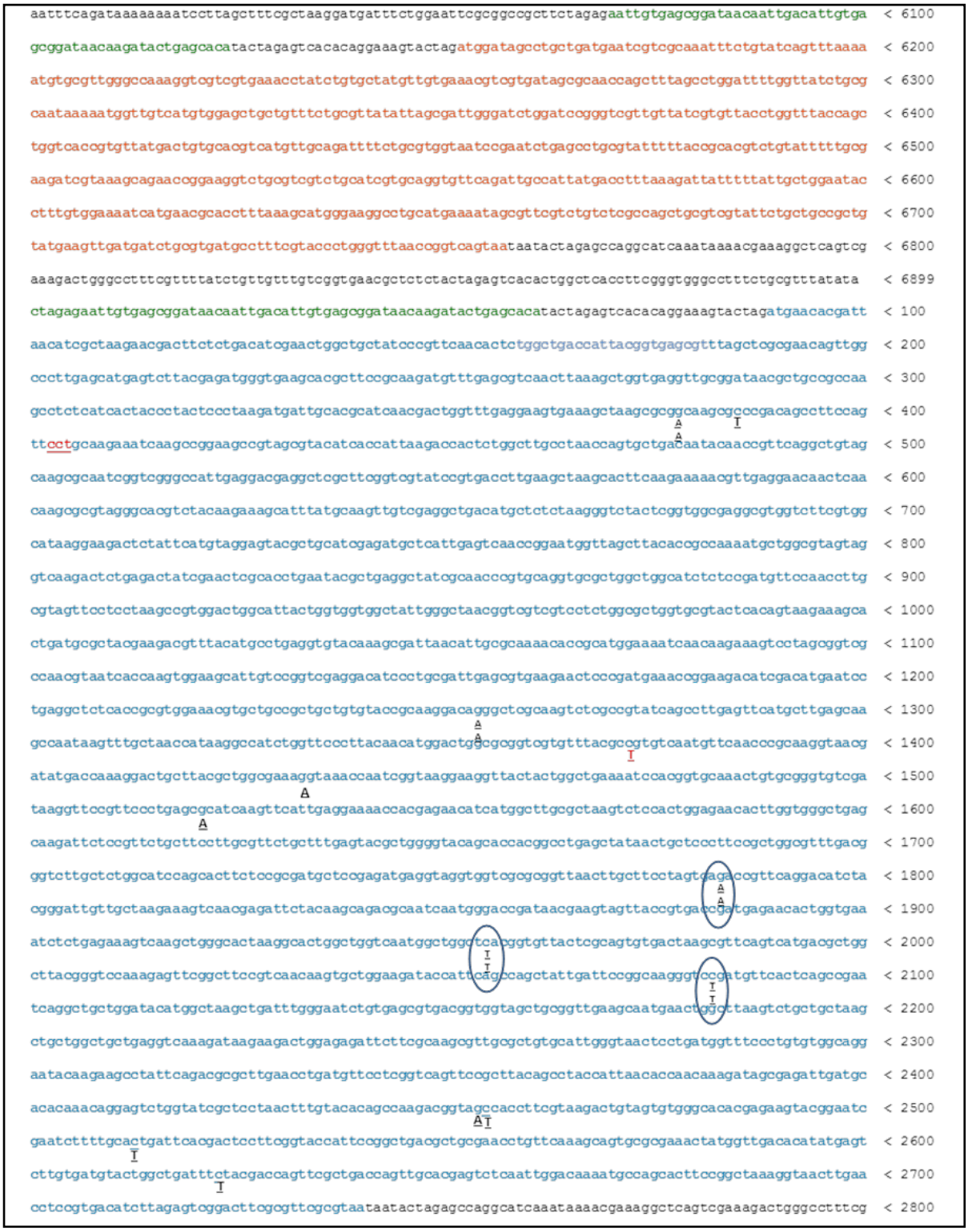


Figure 11.3 *- Complete sequence map of T7RNApolymerase gene (blue), highlighting off target mutations incurred within the AID + T7 unfused mutator device.*

***Table 11.3**- All possible C-T mutations and corresponding residue changes. This shows a limited spectrum of diversity. Complete diversity would equate to 400 potential amino acid changes (20x20) not including STOP codons. In this system, only 41 of 400 are possible. The probability that a single deamination event will result in an active mutation is 70.69% a in an active mutation.*

Original Amino Acid	Codons	Possible C-T Mutations	Amino Acid Change	No. Non-Silent Mutations	No of Silent Mutations	Probability of AA change as a result (%)
Phenylalanine	TTT TTC	 TTT	 No	0	1	0.00
Leucine	TTA TTG CTT CTC CTA CTG	 TTT TTC,TTT,CTT TTA TTG	 Phe Phe,Phe,No No No	3	3	5.17
Isoleucine	ATT ATC ATA	 ATT	 No	0	1	0.00
Methionine (START)	ATG			0	0	0.00
Valine	GTT GTC GTA GTG	 GTT	 No	0	1	0.00
Serine	TCT TCC TCA TCG AGT AGC	TTT TTC,TCT TTA TTG AGT	Phe Phe,No Leu Leu No	4	1	6.90
Proline	CCT CCC CCA CCG	TCT,TTT,CTT TCC,CTC,CCT,TTC,CTT,TTT TCA,CTA,TTA TCG,CTG,TTG	Ser,Phe,Leu Ser,Leu,No,Phe,Leu,Phe Ser,Leu,Leu Ser,Leu,Leu	14	1	24.14
Threonine (Thr)	ACT ACC ACA ACG	ATT ATC,ACT,ATT ATA ATG	Iso Iso,No,Iso Iso Met	5	1	8.62
Alanine	GCT GCC GCA GCG	GTT GTC,GCT,GTT GTA GTG	Val Val,No,Val Val Val	5	1	8.62
Tyrosine (Thy)	TAT TAC	 TAT	 No	0	1	0.00
STOP	TAA TAG TGA			0	0	0.00
Histamine	CAT CAC	TAT TAC,CAT,TAT	Tyr Tyr,No,Tyr	3	1	0.00
Glutamine	CAA CAG	TAA TAG	STOP STOP	2	0	5.17
Asparagine	AAT AAC	 AAT	 No	0	1	3.45
Lysine	AAA AAG			0	0	0.00
Aspartic Acid	GAT GAC	 GAT	 No	0	1	0.00

Original Amino Acid	Codons	Possible C-T Mutations	Amino Acid Change	No. Non-Silent Mutations	No of Silent Mutations	Probability of AA change as a result (%)
Glutamic Acid	GAA GAG			0	0	0.00
Cysteine	TGT TGC	TGT	No	0	1	0.00
Tryptophan (Try)	TGG			0	0	0.00
Arginine	CGT CGC CGA CGG AGA AGG	TGT TGC,CGT,TGT TGA TGG	Cys Cys,No,Cys STOP Trip	5	1	8.62
Glycine	GGT GGC GGA GGG	GGT	No	0	1	0.00
TOTAL				**41**	**17**	

11.2 Relating to Chapter 3 - BASIC Assembly

Table 11.4 *– Synthesised oligonucleotides which form the standard linkers used in BASIC assembly. GC content and predicted linker annealing melting temperatures are given.*
Adapter sites, Adapter Oligonucleotides, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region.

Linker Name	Sequence	Suffix Adapter Tm (GC Content)	Prefix Adapter Tm (GC Content)	Linker Annealing Site Tm (GC content)
Linker 1	ggCTCGTTACTTACGACACTCCGAGACAGTCAGAGGGTATTTATTGAACTAGTCC			66C (57%)
L1-SX-L	CTCGTTACTTACGACACTCCGAGACAGTCAGAGGGTA			
L1-SX-AD	TGTCGTAAGTAA	43C (33%)		
L1-PX-L	GGACTAGTTCAATAAATACCCTCTGACTGTCTCGGAG			
L1-PX-AD	TTTATTGAACTA		36C (18%)	
Linker 2	ggCTCGATCGGTGTGAAAAGTCAGTATCCAGTCGTGTAGTTCTTATTACCTGTCC			63C (48%)
L2-SX-L	CTCGATCGGTGTGAAAAGTCAGTATCCAGTCGTGTAG			
L2-SX-AD	TTTCACACCGAT	49C (41%)		
L2-PX-L	GGACAGGTAATAAGAACTACACGACTGGATACTGACT			
L2-PX-AD	TTCTTATTACCT		38C (25%)	
Linker 3	ggCTCGATCACGGCACTACACTCGTTGCTTTATCGGTATTGTTATTACAGAGTCC			63C (42%)
L3-SX-L	CTCGATCACGGCACTACACTCGTTGCTTTATCGGTAT			
L3-SX-AD	TAGTGCCGTGAT	51C (50%)		
L3-PX-L	GGACTCTGTAATAACAATACCGATAAAGCAACGAGTG			
L3-PX-AD	TGTTATTACAGA		39C (25%)	
Linker 4	ggCTCGACCCACGACTATTGACTGCTCTGAGAAAGTTGATTGTTACGATTAGTCC			64C (43%)
L4-SX-L	CTCGACCCACGACTATTGACTGCTCTGAGAAAGTTGA			
L4-SX-AD	ATAGTCGTGGGT	50C (50%)		
L4-PX-L	GGACTAATCGTAACAATCAACTTTCTCAGAGCAGTCA			
L4-PX-AD	TTGTTACGATTA		41C (25%)	
Linker 5	ggCTCGAGAAGTAGTGCCACAGACAGTATTGCTTACGAGTTGATTTATCCTGTCC	48C (50%)	39C (25%)	63C (43%)
L5-SX-L	CTCGAGAAGTAGTGCCACAGACAGTATTGCTTACGAG			
L5-SX-AD	GGCACTACTTCT			
L5-PX-L	GGACAGGATAAATCAACTCGTAAGCAATACTGTCTGT			
L5-PX-AD	TTGATTTATCCT			
Linker 6	ggCTCGGTATTGTAAAGCACGAAACCTACGATAAGAGTGTCAGTTCTCCTTGTCC	41C (33%)	46C (42%)	63C (43%)
L6-SX-L	CTCGGTATTGTAAAGCACGAAACCTACGATAAGAGTG			
L6-SX-AD	GCTTTACAATAC			
L6-PX-L	GGACAAGGAGAACTGACACTCTTATCGTAGGTTTCGT			
L6-PX-AD	TCAGTTCTCCTT			
Linker 7	ggCTCGAACTTTTACGGGTGCCGACTCACTATTACAGACTTACTACAATCTGTCC	46C (42%)	38C (25%)	64C (48%)
L7-SX-L	CTCGAACTTTTACGGGTGCCGACTCACTATTACAGA			
L7-SX-AD	CCCGTAAAAGTT			
L7-PX-L	GGACAGATTGTAGTAAGTCTGTAATAGTGAGTCGGCA			
L7-PX-AD	TTACTACAATCT			

Table 11.5 *- Sequences of synthesised oligonucleotides for the creation of methylated linkers used in BASIC assembly for idempotent cloning. GC content and predicted linker annealing TM is given (calculated by IDT Oligo Analyzer). Key: Adapter sites, Adapter Oligonucleotides containing Methylated Cytosine base, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region.*

Linker Name	Sequence	Suffix Adapter Tm (GC Content)	Prefix Adapter Tm (GC Content)	Linker Annealing Site Tm (GC content)
Methylated Linker 1	ggCTCGGGTAAGAACTCGCACTTCGTGGAAACACTATTATCTGGTGGGTCTCTGTCC	43C (33%)	36C (18%)	66C (57%)
LMP-SX-L	CTCGGGTAAGAACTCGCACTTCGTGGAAACACTATTA			
LMP-SX-AD	CGAGTTCTTACC			
LMP-PX-L	GGACAGAGACCCACCAGATAATAGTGTTTCCACGAAGTG			
LMP-PX-AD	TCTGGTGGGT/C/TCT			
Methylated Linker 2	ggCTCGGGAGACCTATCGGTAATAACAGTCCAATCTGGTGTAACTTCGGAATCGTCC	49C (41%)	38C (25%)	63C (48%)
LMS-SX-L	CTCGGGAGACCTATCGGTAATAACAGTCCAATCTGGTGT			
LMS-SX-AD	CGATAGGT/C/TCC			
LMS-PX-L	GGACGATTCCGAAGTTACACCAGATTGGACTGTTATTAC			
LMS-PX-AD	AACTTCGGAATC			

Table 11.6 *- Sequences of synthesised oligonucleotides required for fusion linkers used to created fusion proteins via the BASIC assembly method. GC content and predicted linker annealing TM is given (calculated by IDT Oligo Analyzer). Key: Adapter sites, Adapter Oligonucleotides, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region.*

Linker	Sequence	Suffix Adapter Tm (GC)	Prefix Adapter Tm (GC)	Linker Annealing Site Tm (GC)
Fusion Linker 1-Rigid	ggCTCGGCCGAAGCGGCTGCTAAAGAAGCAGCTGCTAAAGAGGCGGCCGCCAAGGCAGGGTCC	76C (75%)	77C (83%)	69C (50%)
LFU1-SX-L	ggCTCGGCCGAAGCGGCTGCTAAAGAAGCAGCTGCTAAAGAGGCGGC			
LFU1-SX-AD	AGCCGCTTCGGC			
LFU1-PX-L	GGACCCTGCCTTGGCGGCAGCCTCTTTAGCAGCTGCTTCTTTAGC			
LFU1-PX-AD	C GCC AAG GCA GG			
Fusion Linker 2-Flexible	ggCTCGGGCTCGGGCTCCGGATCTGGTTCAGGTTCAGGATCGGGCTCCGGGTCC	75C (83%)	75C (75%)	64 (55%)
LFU2-SX-L	CTCGGGCTCGGGCTCCGGATCTGGTTCAGGTTCAGG			
LFU2-SX-AD	GGA GCC CGA GCC			
LFU2-PX-L	GGACCCGGAGCCCGATCCTGAACCTGAACCAGATCC			
LFU2-PX-AD	ATCGGGCTCCGG			
Fusion Linker 3-Taq Homolog	ggCTCGCTGCTTGAGAGCCCTAAAGCATTAGAAGAAGCACCTTGGCCTCCACCAGAAGGGTCC	65C (58%)	69C (68%)	70C (48%)
LFU3-SX-L	CTCGCTGCTTGAGAGCCCTAAAGCATTAGAAGAAGCACCTTGGCC			
LFU3-SX-AD	GCTCTCAAGCAG			
LFU3-PX-L	GGACCCTTCTGGTGGAGGCCAAGGTGCTTCTTCTAATGCTTTAGG			
LFU3-PX-AD	TCCACCAGAAGG			

Table 11.7 *- Sequences of synthesised oligonucleotides required for generation of RBS linkers. 3 linker overlap regions were generated, with the specific RBS of interest placed within the Adapter Oligo site. Key: Adapter sites, Adapter Oligonucleotides, RBS containing Oligos, Prefix and Suffix Scar Sequence, Homologous Linker Annealing Region.*

Linker	Sequence	Suffix Adapter Tm (GC Content)	Prefix Adapter Tm (GC Content)
L1RBS-SX-L	CTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAA		
L1RBS-SX-A	GACGGTGTTCAA	50C (50%)	
L2RBS-SX-L	CTCGTGTTACTATTGGCTGAGATAAGGGTAGCAGAAA		
L2RBS-SX-A	CCAATAGTAACA	40C (33%)	
L3RBS-SX-L	CTCGATTTACCGAGGCTGAATAAGGATTACTTTCCGT		
L3RBS-SX-A	GCCTCGGTAAAT	49C (50%)	
LX RBS1-PX-A (RBS33)	ATCACACAGGACTA		52C (43%)
LX RBS2-PX-A (RBS64)	AAAGAGGGGAAATA		50C (36%)
LX RBS3-PX-A (RBS30)	AAAGAGGAGAAATA		46C (28%)
LX RBS4-PX-A	ATCACAAGGAGGTA		52C (42%)
LX RBS5-PX-A	ATTAGTGGAGGTTA		-
L1 RBS1-PX-L	GGACTAGTCCTGTGTGATTTACAACTGATACTTACCTGA		
L2 RBS1-PX-L	GGACTAGTCCTGTGTGATTTTCTGCTACCCTTATCTCAG		
L3 RBS1-PX-L	GGACTAGTCCTGTGTGATACGGAAAGTAATCCTTATTCA		
L1 RBS2-PX-L	GGACTATTTCCCCTCTTTTTACAACTGATACTTACCTGA		
L2 RBS2-PX-L	GGACTATTTCCCCTCTTTTTTCTGCTACCCTTATCTCAG		
L3 RBS2-PX-L	GGACTATTTCCCCTCTTTACGGAAAGTAATCCTTATTCA		
L1 RBS3-PX-L	GGACTATTTCTCCTCTTTTTACAACTGATACTTACCTGA		
L2 RBS3-PX-L	GGACTATTTCTCCTCTTTTTTCTGCTACCCTTATCTCAG		
L3 RBS3-PX-L	GGACTATTTCTCCTCTTTACGGAAAGTAATCCTTATTCA		
L1 RBS4-PX-L	GGACTACCTCCTTGTGATTTACAACTGATACTTACCTGA		
L2 RBS4-PX-L	GGACTACCTCCTTGTGATTTTCTGCTACCCTTATCTCAG		
L3 RBS4-PX-L	GGACTACCTCCTTGTGATACGGAAAGTAATCCTTATTCA		
L1 RBS5-PX-L	GGACTAACCTCCACTAATTTACAACTGATACTTACCTGA		
L2 RBS5-PX-L	GGACTAACCTCCACTAATTTTCTGCTACCCTTATCTCAG		
L1 RBS1	ggCTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAAATCACACAGGACTAGTCC	59C (33%)	
L1 RBS2	ggCTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAAAAAGAGGGGAAATAGTCC	59C (33%)	
L1 RBS3	ggCTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAAAAAGAGGAGAAATAGTCC	59C (33%)	
L1 RBS4	ggCTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAAATCACAAGGAGGTAGTCC	59C (33%)	
L1 RBS5	ggCTCGTTGAACACCGTCTCAGGTAAGTATCAGTTGTAAATTAGTGGAGGTTAGTCC	59C (33%)	
L2 RBS1	ggCTCGTGTTACTATTGGCTGAGATAAGGGTAGCAGAAAATCACACAGGACTAGTCC	62C (42%)	
L2 RBS2	ggCTCGTGTTACTATTGGCTGAGATAAGGGTAGCAGAAAAAAGAGGGGAAATAGTCC	62C (42%)	
L2 RBS3	ggCTCGTGTTACTATTGGCTGAGATAAGGGTAGCAGAAAAAAGAGGAGAAATAGTCC	62C (42%)	
L2 RBS4	ggCTCGTGTTACTATTGGCTGAGATAAGGGTAGCAGAAAATCACAAGGAGGTAGTCC	62C (42%)	
L2 RBS5	ggCTCGTGTTACTATTGGCTGAGATAAGGGTAGCAGAAAATTAGTGGAGGTTAGTCC	62C (42%)	
L3 RBS1	ggCTCGATTTACCGAGGCTGAATAAGGATTACTTTCCGTATCACACAGGACTAGTCC	60C (33%)	
L3 RBS2	ggCTCGATTTACCGAGGCTGAATAAGGATTACTTTCCGTAAAGAGGGGAAATAGTCC	60C (33%)	
L3 RBS3	ggCTCGATTTACCGAGGCTGAATAAGGATTACTTTCCGTAAAGAGGAGAAATAGTCC	60C (33%)	
L3 RBS4	ggCTCGATTTACCGAGGCTGAATAAGGATTACTTTCCGTATCACAAGGAGGTAGTCC	60C (33%)	

Table 11.8 *- Primers were designed so that internal linker sequences acted as priming sites for both Forward and Reverse primers. Considerations were made to ensure that all primers had a similar melting temperature (Tm) any thus theoretically any two could be chosen to amplify across a functional part. PCR product bands could then be compared to predicted band size for that specific construct.*

Primer Name	Sequence (5' – 3')	PFU Tm	Phusion TM
BASIC L1-FOR	GACACTCCGAGACAGTCAGAGGGTA	61C	69C
BASIC L1-REV	TACCCTCTGACTGTCTCGGAGTGTC	61C	
BASIC L2-FOR	GTGTGAAAAGTCAGTATCCAGTCGTGTAGTTC	59C	70C
BASIC L2-REV	GAACTACACGACTGGATACTGACTTTTCACAC	59C	
BASIC L3-FOR	GCACTACACTCGTTGCTTTATCGGTATTG	59C	70C
BASIC L3-REV	CAATACCGATAAAGCAACGAGTGTAGTGC	59C	
BASIC L4-FOR	CGACTATTGACTGCTCTGAGAAAGTTGATTG	58C	71C
BASIC L4-REV	CAATCAACTTTCTCAGAGCAGTCAATAGTCG	58C	
BASIC L5-FOR	GCCACAGACAGTATTGCTTACGAGTTG	59C	69C
BASIC L5-REV	CAACTCGTAAGCAATACTGTCTGTGGC	59C	
BASIC L6-FOR	GCACGAAACCTACGATAAGAGTGTCAG	58C	68C
BASIC L6-REV	CTGACACTCTTATCGTAGGTTTCGTGC	58C	
BASIC L7-FOR	GGGTGCCGACTCACTATTACAGACTTAC	60C	68C
BASIC L7-REV	GTAAGTCTGTAATAGTGAGTCGGCACCC	60C	
BASIC.FUL1-FOR	GCTAAAGAAGCAGCTGCTAAAGAGGCG	61C	68C
BASIC.FUL1-REV	CGCCTCTTTAGCAGCTGCTTCTTTAGC	61C	
BASIC.FUL2-FOR	CCGGATCTGGTTCAGGTTCAGGA	60C	72C
BASIC.FUL2-REV	TCCTGAACCTGAACCAGATCCGG	60C	
BASIC.FUL3-FOR	CCTAAAGCATTAGAAGAAGCACCTTGGCC	61C	72C
BASIC.FUL3-REV	GGCCAAGGTGCTTCTTCTAATGCTTTAGG	61C	

Table 11.9 *– Raw colony counts from BASIC assembly tests for accuracy and efficiency*

	Assembly ID	Dilution Factor	Total	White (no fluorescence)	Red	Green	Yellow	Percentage Correct	Percentage Incorrect
Repeat 1	2S	9	566	12	554	0	0	97.88%	2.12%
	3S	9	145	5	140	0	0	96.55%	3.45%
	4S	90	840	41	2	0	797	94.88%	4.88%
	5S	90	318	26	5	3	284	89.31%	8.18%
	6S	90	132	35	2	2	93	70.45%	26.52%
	2D	9	664	664	0	0	0	100.00%	0.00%
	3D	9	260	0	260	0	0	100.00%	0.00%
	4D	9	158	2	156	0	0	98.73%	1.27%
	5D	90	599	18	1	1	579	96.66%	3.01%
	6D	90	215	12	2	1	200	93.02%	5.58%
	7D	90	44	0	0	1	43	97.73%	0.00%
Repeat 2	2S	1	230	0	230	0	0	100.00%	0.00%
	3S	9	582	0	582	0	0	100.00%	0.00%
	4S	9	137	6	3		128	93.43%	4.38%
	5S	90	643	213	3	3	424	65.94%	33.13%
	6S	90	24	15	2	0	7	29.17%	62.50%
	2D	1	104	104	0	0	0	100.00%	0.00%
	3D	9	124	0	124	0	0	100.00%	0.00%
	4D	9	100	0	100	0	0	100.00%	0.00%
	5D	90	382	0	1		381	99.74%	0.00%
	6D	90	100	0	1	0	99	99.00%	0.00%
	7D	90	37	0	1	0	36	97.30%	0.00%
Repeat 3	K2	9	355	0	355	0	0	100.00%	0.00%
	K3	9	209	4	205	0	0	98.09%	1.91%
	K4	9	78	6	0	0	72	92.31%	7.69%
	K5	90	315	59	1	1	254	80.63%	18.73%
	K6	90	178	56	5	0	117	65.73%	31.46%
	KC2	9	449	449	0	0	0	100.00%	0.00%
	KC3	9	110	2	108	0	0	98.18%	1.82%
	KC4	9	38	0	38	0	0	100.00%	0.00%
	KC5	90	154	12	2		140	90.91%	7.79%
	KC6	90	39	1	0	0	38	97.44%	2.56%
	KC7	90	30	10	0	0	20	66.67%	33.33%
Repeat 4	K2	1	134	0	134	0	0	100.00%	0.00%
	K3	1	51	0	51	0	0	100.00%	0.00%
	K4	9	41	3	0	0	38	92.68%	7.32%
	K5	90	101	23	1	2	75	74.26%	22.77%
	K6	90	12	9	0	0	3	25.00%	75.00%
	KC2	1	69	69	0	0	0	100.00%	0.00%
	KC3	9	22	0	22	0	0	100.00%	0.00%
	KC4	90	352	0	352	0	0	100.00%	0.00%
	KC5	90	63	0	0	2	61	96.83%	0.00%
	KC6	90	7	0	0	0	7	100.00%	0.00%
	KC7	90	3	0	0	0	3	100.00%	0.00%

Table 11.10 *– Calculation of the average correct and incorrect colonies across 4 repeats of BASIC assembly test for efficiency*

Average Correct	STD DEV	STD ERR	Average Incorrect	STD DEV	STD ERR
99.47%	0.01	0.01	0.53%	0.01	0.01
98.66%	0.02	0.01	0.83%	0.02	0.01
93.33%	0.01	0.01	0.84%	0.02	0.01
77.54%	0.10	0.05	5.16%	0.10	0.05
47.59%	0.24	0.12	11.80%	0.24	0.12
100.00%	0.00	0.00	0.00%	0.00	0.00
99.55%	0.01	0.00	0.45%	0.01	0.00
99.68%	0.01	0.00	0.32%	0.01	0.00
96.03%	0.04	0.02	1.84%	0.04	0.02
97.36%	0.03	0.02	1.33%	0.03	0.01
90.42%	0.16	0.08	8.33%	0.17	0.08

Table 11.11 *- Calculation of the average correct and incorrect colonies across 4 repeats of BASIC assembly test for efficiency*

Assembly ID	Total Colonies	STD ERR	White	STD ERR	Red	STD ERR	Green	STD ERR	yellow	STD ERR
2S	11658	4281.76	33	33.33	11625	4295.80	0	0.00	0	0.00
3S	3875	1145.78	25	14.61	3850	1159.95	0	0.00	0	0.00
4S	944	219.66	53	8.24	9	8.16	0	0.00	883	206.15
5S	383	124.21	89	49.99	3	1.06	3	0.53	288	79.69
6S	96	45.25	32	11.83	3	1.15	1	0.56	61	32.55
2D	7417	1120.37	7417	1120.37	0	0.00	0	0.00	0	0.00
3D	1433	546.18	6	5.56	1428	546.92	0	0.00	0	0.00
4D	920	324.32	6	5.56	914	319.56	0	0.00	0	0.00
5D	333	133.65	8	5.00	1	0.45	1	0.53	323	130.98
6D	100	50.93	4	3.25	1	0.53	0	0.28	96	47.26
7D	32	9.96	3	2.78	0	0.28	0	0.28	28	9.90

Table 11.12 *- Raw colony counts from BASIC assembly tests for accuracy and efficiency when using methylated linkers in hierarchical assembly*

	ID	Dilution Factor	Total	White (no fluorescence)	Red	Green	Yellow	Percentage Correct	Percentage Incorrect
Repeat 1	ctrl GFP	9	53	0	0	53	0	100.00	0.00
	meth GFP	9	25	0	0	25	0	100.00	0.00
	ctrl RFP	9	61	0	61	0	0	100.00	0.00
	meth RFP	9	27	0	27	0	0	100.00	0.00
	COMB	9	191	0	0	0	191	100.00	0.00
Repeat 2	ctrl GFP	90	83	0	0	83	0	100.00	0.00
	meth GFP	90	104	3	0	101	0	97.12	2.88
	ctrl RFP	90	123	0	123	0	0	100.00	0.00
	meth RFP	90	172	2	170	0	0	98.84	1.16
	COMB	9	41	0	0	0	41	100.00	0.00
Repeat 3	ctrl GFP	9	31	0	0	31	0	100.00	0.00
	meth GFP	9	28	2	0	26	0	92.86	7.14
	ctrl RFP	90	567	2	565	0	0	99.65	0.35

	meth RFP	9	34	0	34	0	0	100.00	0.00
	COMB	9	145	1	0	0	144	99.31	0.69

Table 11.13 *- Calculation of the average correct and incorrect colonies across 3 repeats of BASIC assembly test for efficiency and accuracy of hierarchical assembly.*

Average Correct	STD DEV	STD ERR	Average Incorrect	STD DEV	STD ERR
100.00	0.00	0.00	0.00	0.00	0.00
96.66	2.94	1.70	3.34	1.67	0.96
99.88	0.18	0.10	0.12	4.12	2.38
99.61	0.67	0.39	0.39	0.60	0.34
99.77	0.34	0.20	0.23	0.00	0.00

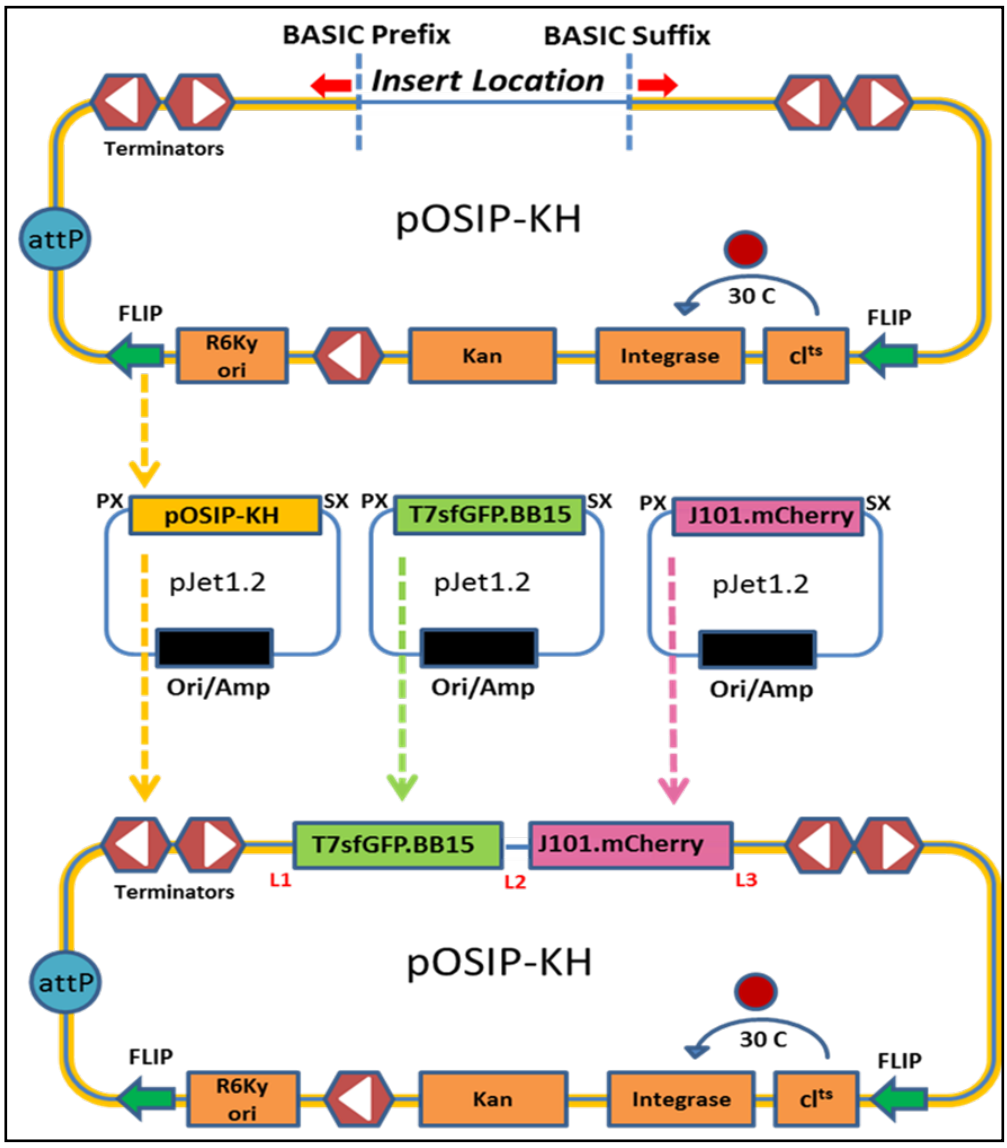


Figure 11.4 *– the assembly of single and double gene cassettes using the standard BASIC assembly protocol into the BASIC.pOSIP KL vector*

Table 11.14 *– Improved PCR primers for colony PCR diagnostics check of genome integrates produced using the BASIC-OSIP method described in* ***Chapter 3.4***

Primer Name (BASIC.OSIP….)	Sequence (5'-3')	Length (bp)	PFU Tm °C
P1.H.Imp	CATGGTAGGAATCAATGCCTGAGTGTTGTC	30bp	60C
P1.L.C.Imp	CTCACCAGGCATCACGGCAATATACTG	27bp	61C
P1.P.Imp	CAAAGGATTTGAGCGAGCAACTGTACC	23bp	59C
P1.T.Imp	GATGTTATCGCCTGTATGAACCTGAACG	27bp	58C
P1.O.1st.Imp	CCCTCATTCGAAACCACCCACCG	25bp	62C
P1.O.2nd.Imp	GAATCCGGAATGCCTGCATTGCGC	24bp	64C
P2.H.L.T.C.O.Imp	CAGGAACACTTAACGGCTGACATGGG	26bp	61C
P2.P.Imp	CACTAGTTCGCAGAGTGTTATGGTT	25bp	56C
P3.H.C.O.Imp	GAACGAGTATCGAGATGGCACATAGCCTTG	30bp	62C
P3.L.T.Imp	GGGAATTAATTCTTGAAGACGAAAGGGCC	26bp	59C
P3.P.Imp	GGGAAAGAAACAGAGAAGGGCAC	23bp	57C
P4.H.Imp	GGTCCGGGCATCAACAGCACATTC	24bp	63C
P4.L.Imp	GGTTTGGCTCTGGTCTGGTAGCAATG	26bp	61C
P4.P.Imp	GCATGGCCTTAACAAAGACATCATC	25bp	55C
P4.T.Imp	CTAGTAGTAGAACTACCACCTGACCGG	27bp	59C
P4.C.Imp	Corresponded loci position does not exist in *E.coli*	-	-
P4.O.1st.Imp	ACATCATATGATCATCATGTTTATTGCGTGGT	29bp	57C
P4.O.2nd.Imp	CCGTAACCCTGGAGCCAAAATATCCTC	27bp	61C

Table 11.15 *– Description of target insert positions, the loci are governed by known phage attachment sites with the E.coli MG1665 genome.*

Integration Site	P1 and P4 Genome position	Neighbouring Genes
HK022*	2638681 - 2639403	Reductase (Periplasmic protein associated with TorRS) Histidine Kinase
Lambda	2887793 - 2888533	YbhB, Kinase Inhibitor Acyl-CoA thioesterase
Phi80	2384324 - 2384892	Between genes. Energy Transducer TonB Unknown hypothetical protein
P21	2498757 – 2499262	Isocitrate dehydrogenase
C31	2888533	GGGTGCCAGGGCGTGCCCTTGGGCTCCCCGGGCGCGTA
186 (Primary locus)	914268 - 914508	No gene – upstream from tRNA-Met
186 (Secondary locus)	484111 – 484711	formamidopyrimidine-DNA glycosylase

Table 11.16 *- A description of the BASIC compatible parts required for BASIC assembly of genome integration vectors within a proof of principle for the BASIC-OSIP standard*

Part Number	BASIC Part	Prefix Linker	Suffix Linker
1	J23101.RBS.sfGFP	L2 PX	L1 SX
2	pT7.RBS.sfGFP	L2 PX	L1 SX
3	pT7.RBS.LasR	L2 PX	L1 SX
4	BASIC – pOSIP KH	L1 PX	L2 SX
5	BASIC – pOSIP KL	L1 PX	L2 SX
6	BASIC – pOSIP KP	L1 PX	L2 SX
7	BASIC – pOSIP KT	L1 PX	L2 SX
8	BASIC – pOSIP KO	L1 PX	L2 SX
9	J23101.RBS.sfGFP	L2 PX	L3 SX
10	J23101.RBS.mcherry	L3 PX	L1 SX
11	pT7.RBS.sfGFP	L2 PX	L3 SX
12	p15A.Kan	L1 Px	L2 Sx
13	pMB1.Kan	L1 Px	L2 Sx

Table 11.17 *– A description of the constructs used in the proof of principle experiment to determine efficiency and usability of the BASIC-OSIP standard*

Assembly	No. of Parts	Construct	Parts Required
A	2	KH - J23101.RBS.sfGFP	1,4
B	2	KL - J23101.RBS.sfGFP	1,5
C	2	KP - J23101.RBS.sfGFP	1,6
D	2	KT -J23101.RBS.sfGFP	1,7
E	2	KO - J23101.RBS.sfGFP	1,8
P	3	KH - J23101.RBS.sfGFP - J23101.RBS.mcherry	4, 9, 10
Q	3	KL - J23101.RBS.sfGFP - J23101.RBS.mcherry	5, 9, 10
R	3	KP - J23101.RBS.sfGFP - J23101.RBS.mcherry	6, 9, 10
S	3	KT - J23101.RBS.sfGFP - J23101.RBS.mcherry	7, 9, 10
T	3	KO - J23101.RBS.sfGFP - J23101.RBS.mcherry	8, 9, 10
U	2	p15aKan – J23101.RBSsfGFP	12, 1
V	2	pMB1Kan – J23101.RBSsfGFP	13, 1

11.3 Relating to Chapter 4 - Building a Controllable Gene Circuit

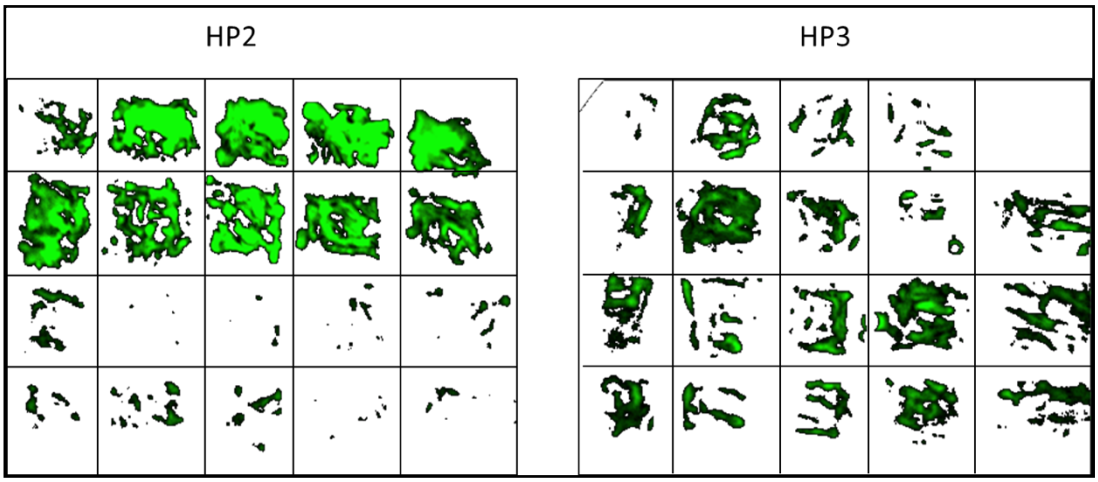


Figure 11.5 *– sfGFP expression from pJ101.Tet characterisation constructs (485nm excitation and 510nm emission)*

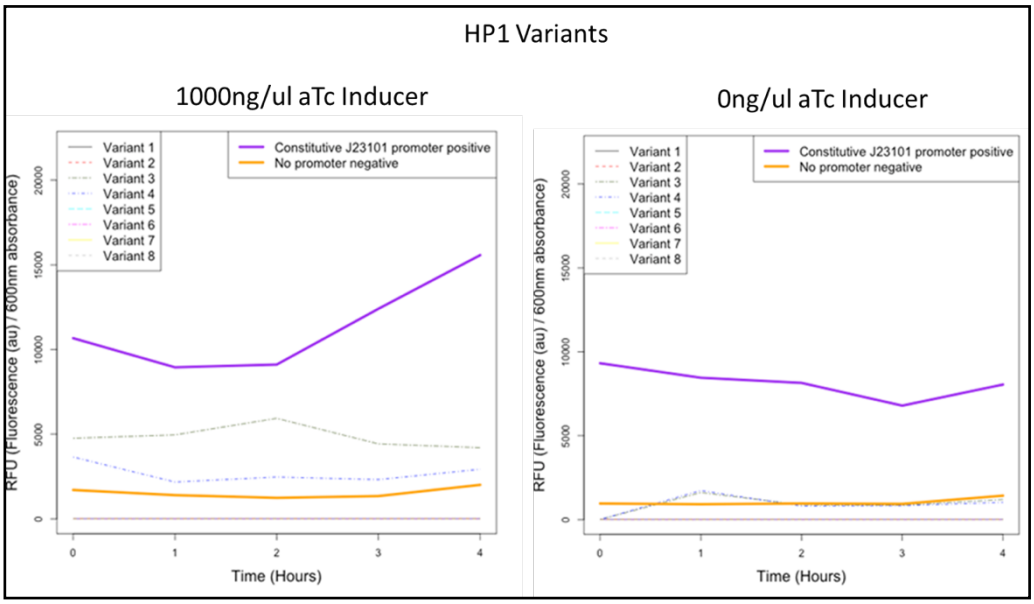


Figure 11.6 *– RFU expression profiles for HP1 variants of the J101.TetOp hybrid promoter across a 4 hour preliminary induction assay*

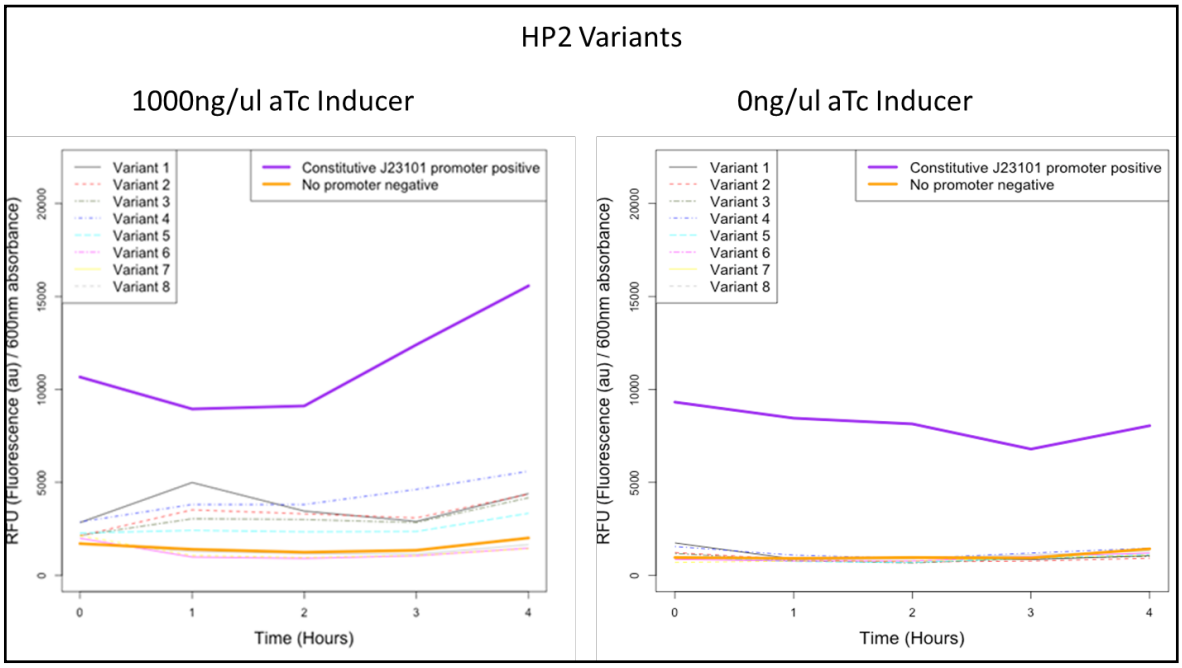


Figure 11.7 *- RFU expression profiles for HP2 variants of the J101.TetOp hybrid promoter across a 4 hour preliminary induction assay*

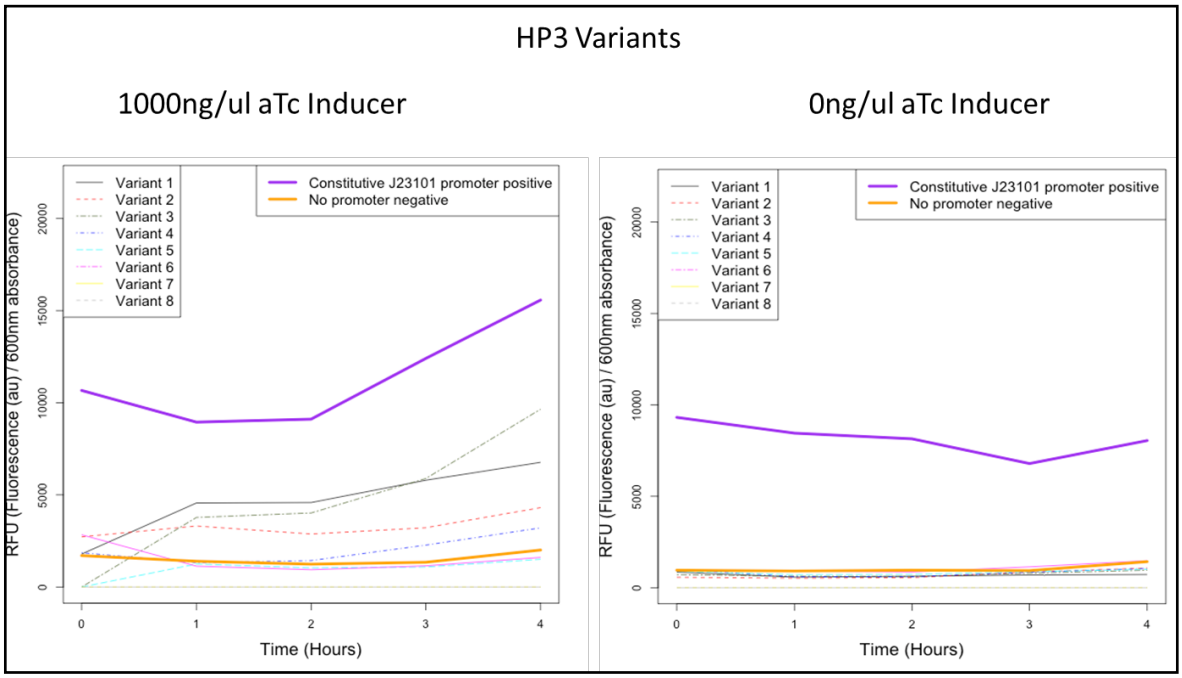


Figure 11.8 *- RFU expression profiles for HP3 variants of the J101.TetOp hybrid promoter across a 4 hour preliminary induction assay*

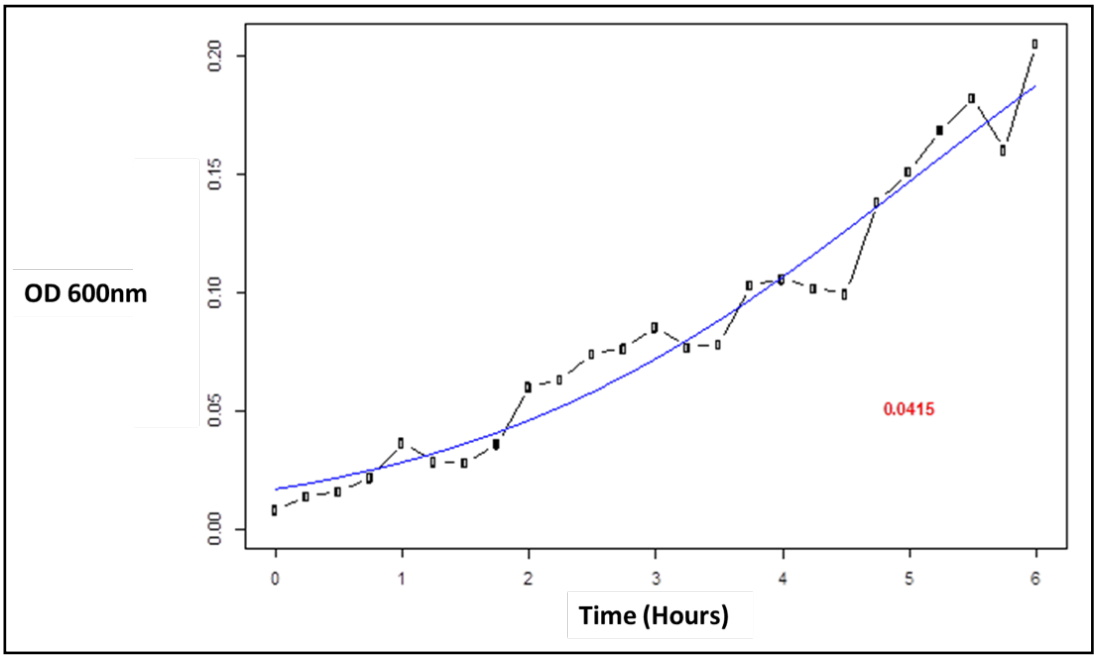


Figure 11.9 *– Absorbance measurements taken at 15 min intervals across 6 hours outgrowth for each sample were used to validate whether or not there was expected logistic growth of the MG1665 strain (lack of sigmoidal shape).*

Table 11.18 *– Final RPU values generated for p101.Tet Hybrid Promoter Characterisation assays and varying concentrations of inducer in order to formulate a does response curve for each variant within the library. Data collection required an automated characterisation platform and analysis tools. 4 repeats of the assay were carried out, shown here is the mean average of the run along with standard deviation and standard errors.*

Average RPU (Mean)										
aTc (ng/ml)	HP1A	HP2A	HP1B	HP2B	HP3A	HP3B	HP3C	HP3D	Tet.WT	J23101
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.043	0.053	0.014	0.017	0.008	0.023	0.028	0.018	0.060	1.006
10	0.034	0.055	0.043	0.031	0.032	0.020	0.030	0.037	0.052	0.912
20	0.054	0.083	0.046	0.015	0.026	0.026	0.031	0.023	0.049	1.072
100	1.123	2.862	5.136	0.553	0.722	0.536	0.765	0.190	1.065	0.992
250	1.452	3.527	7.132	0.812	0.949	0.598	0.989	0.202	1.359	1.019
500	1.526	3.223	7.580	0.934	1.017	0.606	1.032	0.199	1.420	1.023
1000	1.517	4.112	7.581	0.918	1.037	0.610	1.133	0.206	1.564	1.047

StDev										
aTc (ng/ml)	HP1A	HP2A	HP1B	HP2B	HP3A	HP3B	HP3C	HP3D	Tet.WT	J23101
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.028	0.024	0.006	0.010	0.011	0.021	0.013	0.018	0.012	0.027
10	0.029	0.018	0.047	0.024	0.031	0.011	0.018	0.023	0.033	0.077
20	0.017	0.015	0.026	0.014	0.022	0.023	0.019	0.022	0.018	0.102
100	0.453	1.318	2.216	0.219	0.277	0.093	0.251	0.028	0.259	0.064
250	0.108	1.165	0.526	0.107	0.077	0.037	0.104	0.018	0.093	0.027
500	0.169	1.427	1.048	0.202	0.206	0.071	0.143	0.017	0.272	0.058
1000	0.115	1.289	0.803	0.160	0.152	0.075	0.126	0.042	0.200	0.004

StErr										
aTc (ng/ml)	HP1A	HP2A	HP1B	HP2B	HP3A	HP3B	HP3C	HP3D	Tet.WT	J23101
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.016	0.012	0.003	0.005	0.006	0.011	0.007	0.009	0.009	0.014
10	0.015	0.009	0.024	0.012	0.015	0.006	0.009	0.012	0.023	0.039
20	0.010	0.008	0.013	0.007	0.011	0.012	0.009	0.011	0.013	0.051
100	0.226	0.659	1.108	0.110	0.138	0.054	0.125	0.014	0.183	0.032
250	0.054	0.583	0.263	0.054	0.039	0.021	0.052	0.009	0.065	0.014
500	0.085	0.824	0.524	0.101	0.103	0.036	0.071	0.010	0.193	0.029
1000	0.058	0.645	0.402	0.080	0.076	0.037	0.063	0.021	0.142	0.002

11.4 Relating to Chapter 6 - Improving the AID-T7 Mutator Device

***Table 11.19**- Part required for build a range of test constructs via BASIC assembly to test toxicity of the AIDT7 mutator protein.*

Part Number	BASIC Part	Prefix Linker	Suffix Linker
1	HP2	Linker A Meth Px	Linker 1 Sx
2	HP3	Linker A Meth Px	Linker 1 Sx
3	RBS.ATF	Linker 1 PX	Linker 2 Sx
4	RBS.ATF.LVA	Linker 1 PX	Linker 2 Sx
5	RBS.mCHERRY	Linker 2 Px	Linker 3 Sx
6	Tet.REP	Linker 3 Px	Linker B (Meth) Sx
7	CamR	Linker B (Meth) Px	Linker 5 Sx
8	pMB1.Kan	Linker 5 Px	Linker A (Meth) Sx
9	P15.Kan	Linker 5 Px	Linker A (Meth) Sx
10	HP2	Linker A Meth Px	Linker 2 Sx
11	HP3	Linker A Meth Px	Linker 2 Sx
12	T7RNApol	Linker 1 PX	Linker 2 Sx

***Table 11.20** – Description of each test construct used to test toxicity of the AIDT7 mutator protein*

Construct Name	No. of Parts	Construct	Parts Required	PCR Diagnostic – Product Size	Sequencing Primers
A.Tox	6	HP2.ATF(No Tag).mCherry.TetRep.CamR.pMB1-Kan	1,3,5,6,7,8	~4.3	L1 For - L3 Rev
B.Tox	6	HP2.ATF(LVA).mCherry.TetRep.CamR.pMB1-Kan	1,4,5,6,7,8	~4.3	L1 For - L3 Rev
C.Tox	6	HP2.ATF(No Tag).mCherry.TetRep.CamR.p15a-Kan	1,3,5,6,7,9	~4.3	L1 For - L3 Rev
D.Tox	6	HP2.ATF(LVA).mCherry.TetRep.CamR.p15a-Kan	1,4,5,6,7,9	~4.3	L1 For - L3 Rev
E.Tox	6	HP3.ATF(NO tag).mCherry.TetRep.CamR.pMB1-Kan	2,3,5,6,7,8	~4.3	L1 For - L3 Rev
F.Tox	6	HP3.ATF(LVA).mCherry.TetRep.CamR.pMB1-Kan	2,4,5,6,7,8	~4.3	L1 For - L3 Rev
G.Tox	6	HP3.ATF(NO tag).mCherry.TetRep.CamR.p15a-Kan	2,3,5,6,7,9	~4.3	L1 For - L3 Rev
H.Tox	6	HP3.ATF(LVA).mCherry.TetRep.CamR.p15a-Kan	2,4,5,6,7,9	~4.3	L1 For - L3 Rev
I.Tox	5	HP2.mCherry.TetRep.CamR.pMB1Kan	10,5,6,7,8	~1kb	L2 For – L3 Rev
J.Tox	5	HP2.mCherry.TetRep.CamR.p15aKan	10,5,6,7,9	~1kb	L2 For – L3 Rev
K.Tox	5	HP3.mCherry.TetRep.CamR.pMB1.Kan	11,5,6,7,8	~1kb	L2 For – L3 Rev
L.Tox	5	HP3.mCherry.TetRep.CamR.p15aKan	11,5,6,7,9	~1kb	L2 For – L3 Rev
M.Tox	2	CamR.pMB1Kan	6,7,8	Built Previously	-
N.Tox	2	CamR.p15aKan	6,7,9	Built Previously	-
O.Tox	5	HP2.T7RNApol.TetRep.CamR.pMB1	1,12,6,7,8	~3.3 kb	L1 For - L3 Rev
P.Tox	5	HP2.T7RNApol.TetRep.CamR.p15a-Kan	1,12,6,7,8	~3.3 kb	L1 For - L3 Rev

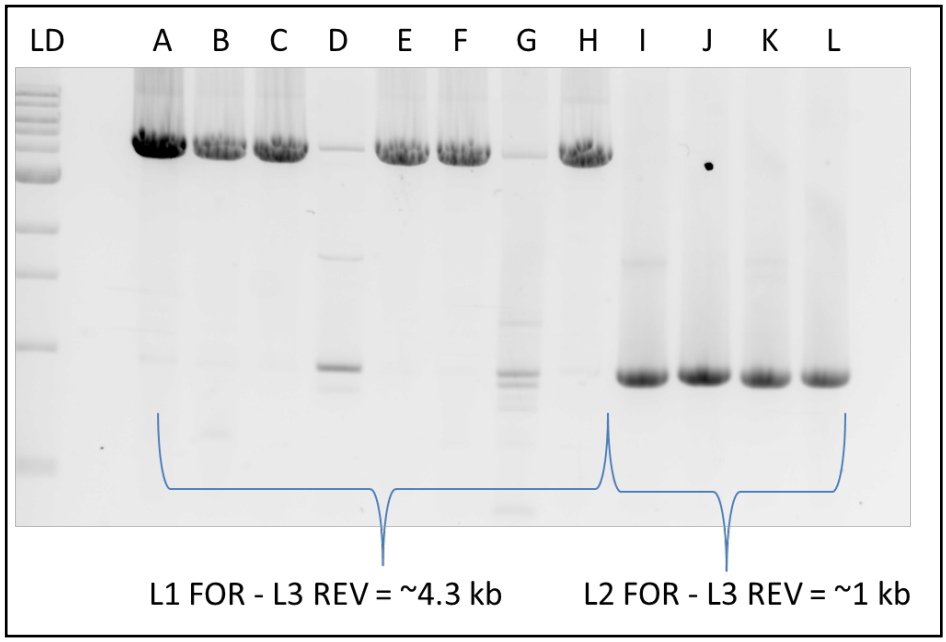


Figure 11.10 *– Diagnostic PCR products to confirm the correct assembly of constructs used to test the toxicity of the AIDT7 mutator protein*

	1	2	3	4	5	6	7	8	9	10	11	12
A	Tox A	Tox H	Tox A	Tox H	Tox A	Tox H	Tox A	Tox H	Tox A	Tox H	Tox A	Tox H
B	Tox B	Tox I	Tox B	Tox I	Tox B	Tox I	Tox B	Tox I	Tox B	Tox I	Tox B	Tox I
C	Tox C	Tox J	Tox C	Tox J	Tox C	Tox J	Tox C	Tox J	Tox C	Tox J	Tox C	Tox J
D	Tox D	Tox K	Tox D	Tox K	Tox D	Tox K	Tox D	Tox K	Tox D	Tox K	Tox D	Tox K
E	Tox E	Tox L	Tox E	Tox L	Tox E	Tox L	Tox E	Tox L	Tox E	Tox L	Tox E	Tox L
F	Tox F	MG1665	Tox F	MG1665	Tox F	MG1665	Tox F	MG1665	Tox F	MG1665	Tox F	MG1665
G	Tox G	MG1665	Tox G	MG1665	Tox G	MG1665	Tox G	MG1665	Tox G	MG1665	Tox G	MG1665
H	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk	Media Blk

0ng/ul atc | 60ng/ul aTc | 250ng/ul aTc

Figure 11.11 *– Plate layout used for toxicity induction assay*

Table 11.21 *– Summary of final OD 600nm readings from toxicity induction assay. Samples are normalised against blank media control with standard errors across 3 repeats (6x independent samples for each).*

Construct	End Point OD 1	End Point OD 2	End Point OD 3	Average End Point OD	Std Deviation	Std Error of the Mean
Tox A	0.676	0.893	0.933	0.834	0.14	0.08
Tox B	1.225	1.064	1.037	1.109	0.10	0.06
Tox C	0.643	0.784	0.789	0.739	0.08	0.05
Tox D	0.895	0.997	1.04	0.977	0.07	0.04
Tox E	0.764	0.788	0.822	0.791	0.03	0.02
Tox F	1.133	1.27	1.24	1.214	0.07	0.04
Tox G	0.665	0.757	0.782	0.735	0.06	0.04
Tox H	0.896	1.134	1.143	1.058	0.14	0.08
Tox I	1.348	1.473	1.597	1.473	0.12	0.07
Tox J	1.355	1.442	1.56	1.452	0.10	0.06
Tox K	1.195	1.243	1.453	1.297	0.14	0.08
Tox L	1.433	1.52	1.673	1.542	0.12	0.07
Tox M	1.322	1.467	1.493	1.427	0.09	0.05
Tox N	1.302	1.654	1.621	1.526	0.19	0.11
Tox O	0.632	0.744	0.653	0.676	0.06	0.03
Tox P	0.693	0.716	0.47	0.626	0.14	0.08

Table 11.22 *- Summary of final RFU readings from toxicity induction assay. Note that RFP outputs within operon – like constructs did not produce a fluorescent output.*

Construct	End Point RFU 1	End Point RFU 2	End Point RFU 3	Average End Point RFU	Std Deviation	Std Error of the Mean
Tox A				0		
Tox B				0		
Tox C				0		
Tox D				0		
Tox E				0		
Tox F				0		
Tox G				0		
Tox H				0		
Tox I	420	486	463	456	33.50	19.34
Tox J	335	398	371	368	31.61	18.25
Tox K	390	456	448	431	36.02	20.80
Tox L	279	345	307	310	33.13	19.13
Tox M				0		
Tox N				0		
Tox O				0		
Tox P				0		

Improving the enzymatic relationship between AID and T7 RNA Pol

Table 11.23 *– Raw data collected generated from gating changing fluorescent profiles of various T7RNApol and ATF mutants during loss of function assay. Data shown reflects the final data point collected after 3 days of subsequent outgrowth.*

Day 3 - EndPoint				
Variant	GFP- %	GFP+%	stdev	sterr
wtT7	8.68	90.84	4.27	1.51
ATF	26.76	70.41	2.50	0.89
T7. Y639F	2.05	97.76	0.62	0.22
ATF.Y639F	35.64	64.36	2.80	1.34
T7. G645A	2.00	97.73	0.66	0.23
ATF.G645A	38.23	56.20	13.61	4.81
T7.I810S	48.46	41.81	9.25	3.27
ATF.I810S	96.35	2.00	0.83	0.29
T7.G645A P266L	1.64	98.11	0.31	0.11
ATF.G645A P266L	35.32	58.66	3.18	1.12
T7 P266L	36.50	62.78	8.52	3.01
ATF P266L	22.71	72.75	3.67	1.30

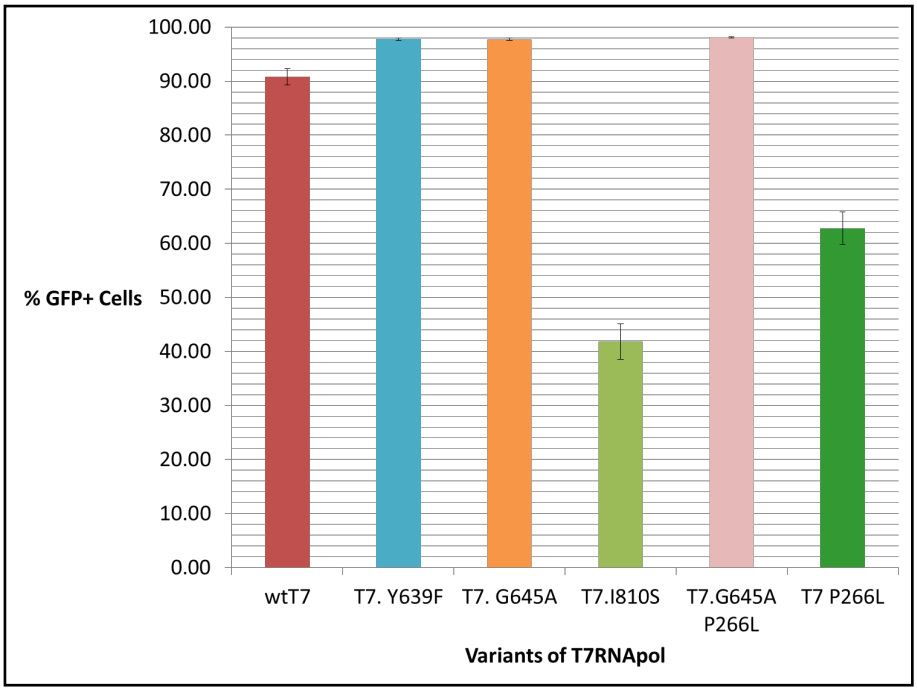


Figure 11.12 *– Expressing T7RNApol mutants in GM31 UGI results in varying % GFP+ cell populations after 3 days successive outgrowth.*

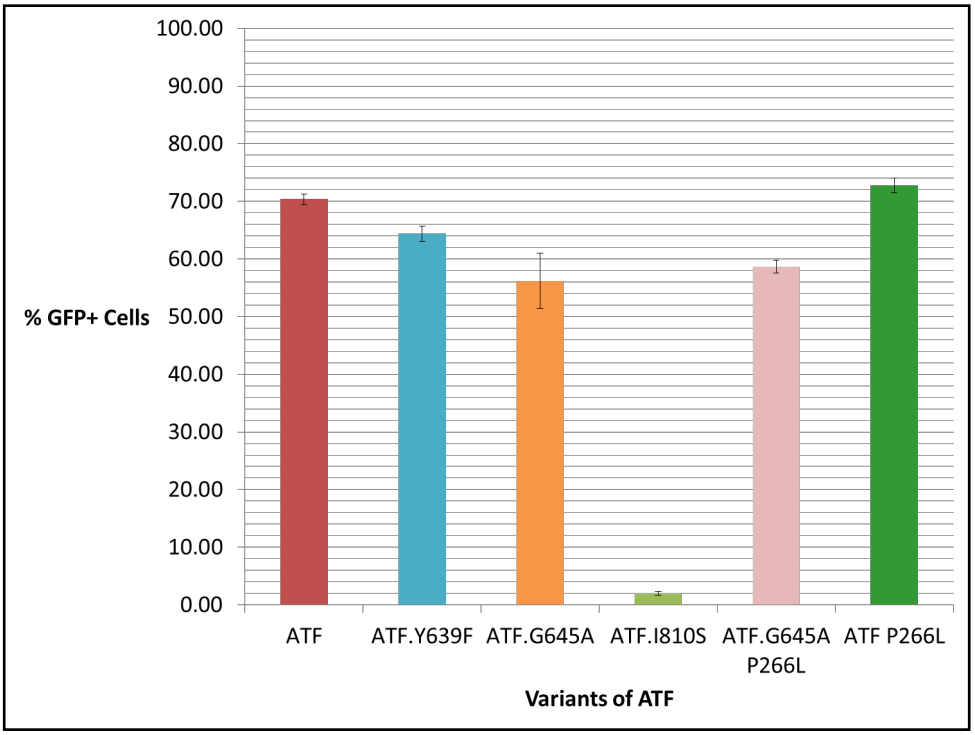


Figure 11.13 *– Expressing ATF mutants in GM31 UGI results in varying % GFP+ cell populations after 3 days consecutive outgrowth. The mutants which reduce T7RNApol elongation speed appear to show an increased mutagenic phenotype as represented by a lower final population of GFP+ cells. Note that ATF.1810S is a false positive i.e. it cannot successfully initiate transcription and thus no GFP is produced.*

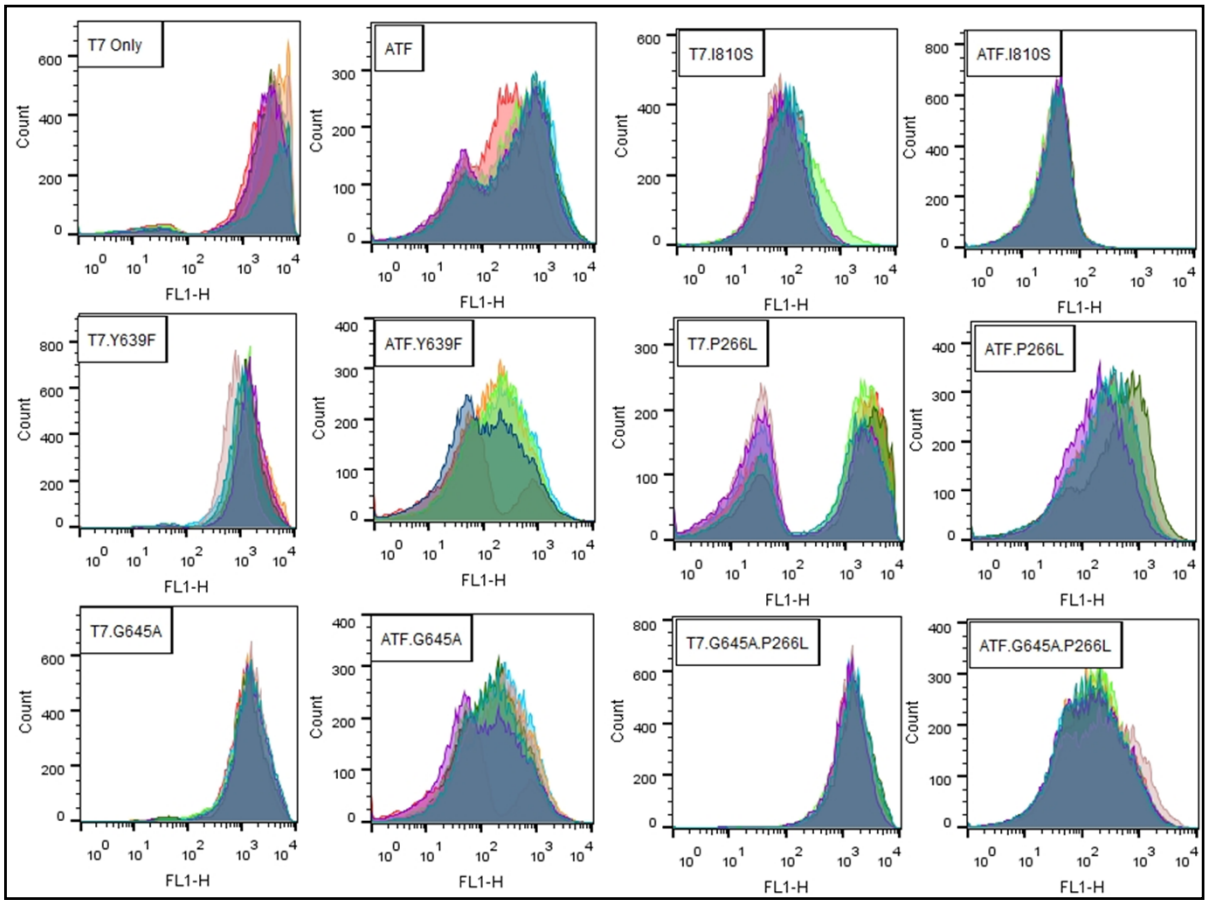


Figure 11.14 *– Variation in fluorescent profiles of T7RNApol and ATF variants, when expressed in GM31 UGI+ strain across 3 successive days, during loss of function assay.*

11.5 Relating to Chapter 7 - Increasing Mutagenic Spectrum and Diversity

Table 11.24 *- Primers used to amplify target genes of interest from the wildtype MG1665 E.coli strain*

Protein / Gene	Primer Sequences (5'-3')	Annealing Temp (PFU)	Product Size (bp)
Pol II / *dinA*	Forward GTG GCG CAG GCA GGT TTT ATC TTA ACC Reverse TCAAAATAGCCCAAGTTGCCCGGTCATAAG	60C	2352
Pol IV / *dinB*	Forward ATG CGT AAA ATC ATT CAT GTG GAT ATG GAC TGC Reverse TCATAATCCCAGCACCAGTTGTCTTTCC	57C	1056
Pol V / *umuCD*	Forward ATG TTG TTT ATC AAG CCT GCG GAT CTC Reverse TTATTTGACCCTCAGTAAATCAGAACTTCGCG	56C	1688
DNA Pol I / *polA*	Forward CCTGTCTGTGGCATCCTTTACCC Reverse CGTGACAGCTTATGTTGCTTACTTACG	55C	2787/ 2888*
ExoIII / *xthA*	Forward CAACAGGCGGTAAGCAACGC Reverse GACCCAGTTTGAGCCAGGAGAGC	59C/70C	807/ 908*
RecA gene (Flanking)	Forward GCTTCAACAGAACATATTGACTATCCGG Reverse CGACCCTTGTGTATCAAACAAGACG	69C	1142
RecA gene (CDS)	Forward ATGGCTATCGACGAAAACAAACAG Reverse TTAAAAATCTTCGTTAGTTTCTGCTACGC	67C	
RecA C-17	Forward ATGGCTATCGACGAAAACAAACAG Reverse TTA CGTTGAGTTCGGGTTGCTCAG	67C	
RecA mutant E.38.K	Forward Phos\AAAACCATCTCTACCGGTTCGC Reverse Phos\CACATCCATGGAACGGTCTTC	66C	

Table 11.25 *-Primers used to create BASIC compatible parts BASIC Prefix, BASIC Suffix, RBS34, Annealing region*

DNA part (BASIC compatible)	Primer Sequences (5'-3')	Annealing Temp (PHU)
PolIV RBS34 CDS	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAATACTAG ATGCGTAAAATCATTCATGTGG Reverse - CGATAGGTCTCCCGAGCC TTATAATCCCAGCACCAGTTG	64C
PolIV NoATG	Forward - TCTGGTGGGTCTCTGTCC CGTAAAATCATTCATGTGGATATGG Reverse - CGATAGGTCTCCCGAGCC TTATAATCCCAGCACCAGTTG	64C
PolIV -5C RBS34 CDS	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAATACTAG ATGCGTAAAATCATTCATGTGG Reverse - CGATAGGTCTCCCGAGCC TTATCTTTCCATTTGCGGGTC	66C
Pol IV -5C NoATG	Forward - TCTGGTGGGTCTCTGTCC CGTAAAATCATTCATGTGGATATGG Reverse - CGATAGGTCTCCCGAGCC TTATCTTTCCATTTGCGGGTC	68C
Pol IV -12C RBS34 CDS	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAATACTAG ATGCGTAAAATCATTCATGTGG Reverse - CGATAGGTCTCCCGAGCC TTAAAGCAACGTCACATGCAGC	
Pol IV -12C NoATG	Forward - TCTGGTGGGTCTCTGTCC CGTAAAATCATTCATGTGGATATGG Reverse - CGATAGGTCTCCCGAGCC TTAAAGCAACGTCACATGCAGC	66C
ExoIII RBS34 CDS	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAA GCGACT ATGAAATTTGTCTCTTTTAATAT Reverse - CGATAGGTCTCCCGAGCC TTAGCGGCGGAAGGTCG	63C
ExoIII RBS34 NoSTOP	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAA GCGACT ATGAAATTTGTCTCTTTTAATAT Reverse - CGATAGGTCTCCCGAGCC GCGGCGGAAGGTCGC	63C
5'-3'Exo RBS34 CDS	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAATACTAGATGGTTCAGATCCCCCAAAATCC Reverse - CGATAGGTCTCCCGAGCC TTACGTTGCCGTCACTTCTGGTG	69C
5'-3'Exo NoSTOP	Forward - TCTGGTGGGTCTCTGTCC AAAGAGGAGAAATACTAGATGGTTCAGATCCCCCAAAATCC Reverse - CGATAGGTCTCCCGAGCC CGTTGCCGTCACTTCTGGTG	69C

Table 11.26 *- Primer sequences required for addition of FLAG tag sequence to both ExoIII and 3-5' Exo BASIC parts. FLAG Tag sequences are shown in red and primer annealing sequences are underlined.*

DNA part (BASIC compatible)	Primer Sequences (5'-3')	Annealing Temp
ExoIII N-FLAG RBS34 NoSTOP	Forward Phos \GATTATAAAGATGACGATGACAAG AAATTTGTCTCTTTTAATATCAACGGCC Reverse Phos\ CATAGTCGCTTTCTCCTCTTTG	66C
3' – 5' EXO N-FLAG RBS34 NoSTOP	Forward Phos\GATTATAAAGATGACGATGACAAG GTTCAGATCCCCCAAAATCCAC Reverse Phos\ CATCTAGTATTTCTCCTCTTTGGACAGAG	67C

Table 11.27 *– Initially plan for combinatorial assembly of ExoIII - Pol IV variants using BASIC assembly, detailing Linker addition and size of each individual part*

Part Number	Part DNA	Prefix Linker	Suffix Linker	Part Size (Cut from Vector – 3kb pJET)
A1	HP2 Promoter	Linker A (Meth) PX	Linker 1 SX	116 bp
A2	HP3 Promoter	Linker A (Meth) PX	Linker 1 SX	116 bp
A3	RBS.Exo III (FLAG) No STOP	Linker 1 Px	Fu.Linker 1 SX	857 bp
A4	RBS.Exo III (FLAG) No STOP	Linker 1 Px	Fu.Linker 2 Sx	857 bp
A5	RBS.Exo III (FLAG) No STOP	Linker 1 Px	Fu.Linker 3 Sx	857 bp
A6	Pol IV WT (NoATG)	Fu.Linker 1 Px	Linker 2 Sx	1061 bp
A7	Pol IV WT (NoATG)	Fu.Linker 2 PX	Linker 2 Sx	1061 bp
A8	Pol IV WT (NoATG)	Fu.Linker 3 Px	Linker 2 Sx	1061 bp
A9	Pol IV -5CTerm	Fu.Linker 1 Px	Linker 2 Sx	1043 bp
A10	Pol IV -5CTerm	Fu.Linker 2 PX	Linker 2 Sx	1043 bp
A11	Pol IV -5CTerm	Fu.Linker 3 Px	Linker 2 Sx	1043 bp
A12	Pol IV -12C	Fu.Linker 1 Px	Linker 2 Sx	1025 bp
A13	Pol IV -12C	Fu.Linker 2 PX	Linker 2 Sx	1025 bp
A14	Pol IV -12C	Fu.Linker 3 Px	Linker 2 Sx	1025 bp
A15	TetRep	Linker 2 Px	Linker B (Meth) Sx	891 bp
A16	CamR	Linker B Px	Linker 3 Sx	967 bp
A17	PMB1 KanR	Linker 3 Px	Linker A (Meth) Px	1655 bp
A18	RBS.EXOIII (FLAG) STOP	Linker 1 PX	Linker 5 Sx	857 bp
A19	RBS.Pol IV WT	Linker 5 PX	Linker 2 Sx	1082 bp
A20	RBS.Pol IV -5C	Linker 5 PX	Linker 2 Sx	1064 bp
A21	RBS.Pol IV -12C	Linker Px	Linker 2 Sx	1046 bp
A22	TetRep	Linker A (Meth) PX	Linker B (Meth) Sx	891 bp
A23	J23101	Linker A (Meth) PX	Linker 1 SX	140 bp
A24	RBS.EXOIII.CDS	Linker 1 PX	Linker B (Meth) Sx	833 bp
A25	RBS.Pol IV.WT CDS	Linker 1 PX	Linker B (Meth) Sx	1082 bp

Table 11.28 *- Initially plan for combinatorial assembly of 5' 3' Exo - Pol IV variants using BASIC assembly, detailing Linker addition and size of each individual part*

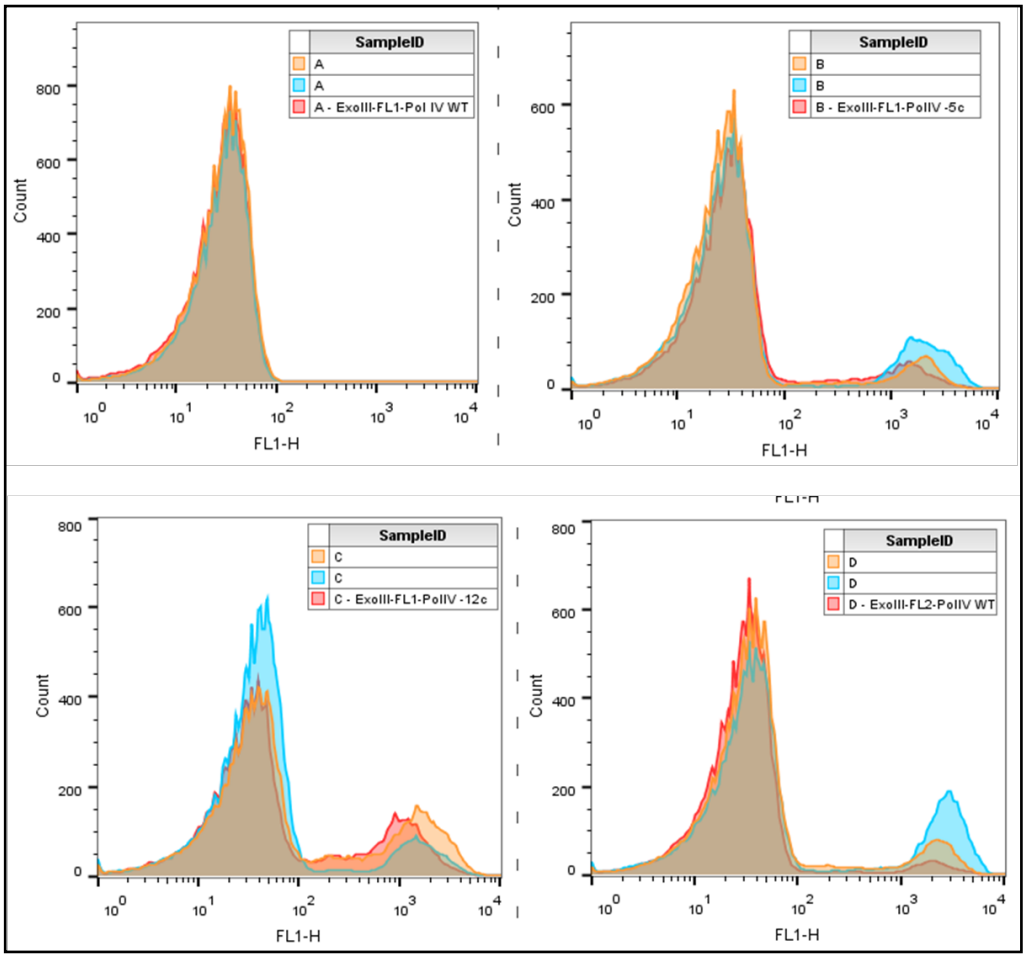
Part Number	Part DNA	Prefix Linker	Suffix Linker	Part Size (Cut from Vector – 3kb pJET)
B1	HP2 Promoter	Linker A (Meth) PX	Linker 1 SX	116 bp
B2	HP3 Promoter	Linker A (Meth) PX	Linker 1 SX	116 bp
B3	RBS.5-3 Exo(FLAG) No STOP	Linker 1 Px	Fu.Linker 1 SX	857 bp
B4	RBS.5-3 Exo(FLAG) No STOP	Linker 1 Px	Fu.Linker 2 Sx	857 bp
B5	RBS.5-3 Exo(FLAG) No STOP	Linker 1 Px	Fu.Linker 3 Sx	857 bp
B6	Pol IV WT (NoRBS)	Fu.Linker 1 Px	Linker 2 Sx	1061 bp
B7	Pol IV WT (NoRBS)	Fu.Linker 2 PX	Linker 2 Sx	1061 bp
B8	Pol IV WT (NoRBS)	Fu.Linker 3 Px	Linker 2 Sx	1061 bp
B9	Pol IV -5CTerm (NoRBS)	Fu.Linker 1 Px	Linker 2 Sx	1043 bp
B10	Pol IV -5CTerm (NoRBS)	Fu.Linker 2 PX	Linker 2 Sx	1043 bp
B11	Pol IV -5CTerm (NoRBS)	Fu.Linker 3 Px	Linker 2 Sx	1043 bp
B12	Pol IV -12C (NoRBS)	Fu.Linker 1 Px	Linker 2 Sx	1025 bp
B13	Pol IV -12C (NoRBS)	Fu.Linker 2 PX	Linker 2 Sx	1025 bp
B14	Pol IV -12C (NoRBS)	Fu.Linker 3 Px	Linker 2 Sx	1025 bp
B15	TetRep	Linker 2 Px	Linker B (Meth) Sx	891 bp
B16	CamR	Linker B Px	Linker 3 Sx	967 bp
B17	PMB1 KanR	Linker 3 Px	Linker A (Meth) Px	1655 bp
B18	RBS.5-3EXO (FLAG) CDS	Linker 1 PX	Linker 5 Sx	857 bp
B19	RBS.Pol IV WT CDS	Linker 5 PX	Linker 2 Sx	1082 bp
B20	RBS.Pol IV -5C CDS	Linker 5 PX	Linker 2 Sx	1064 bp
B21	RBS.Pol IV -12C CDS	Linker Px	Linker 2 Sx	1046 bp
B22	TetRep	Linker A (Meth) PX	Linker B (Meth) Sx	891 bp
B23	J23101	Linker A (Meth) PX	Linker 1 SX	140 bp
B24	RBS.EXOIII.CDS	Linker 1 PX	Linker B (Meth) Sx	833 bp
B25	RBS.Pol IV.WT CDS	Linker 1 PX	Linker B (Meth) Sx	1082 bp
B26	RBS.Pol IV -5C CDS	Linker 1 PX	Linker B (Meth) Sx	1064 bp
B27	RBS.Pol IV -12C CDS	Linker 1 PX	Linker B (Meth) Sx	1046 bp

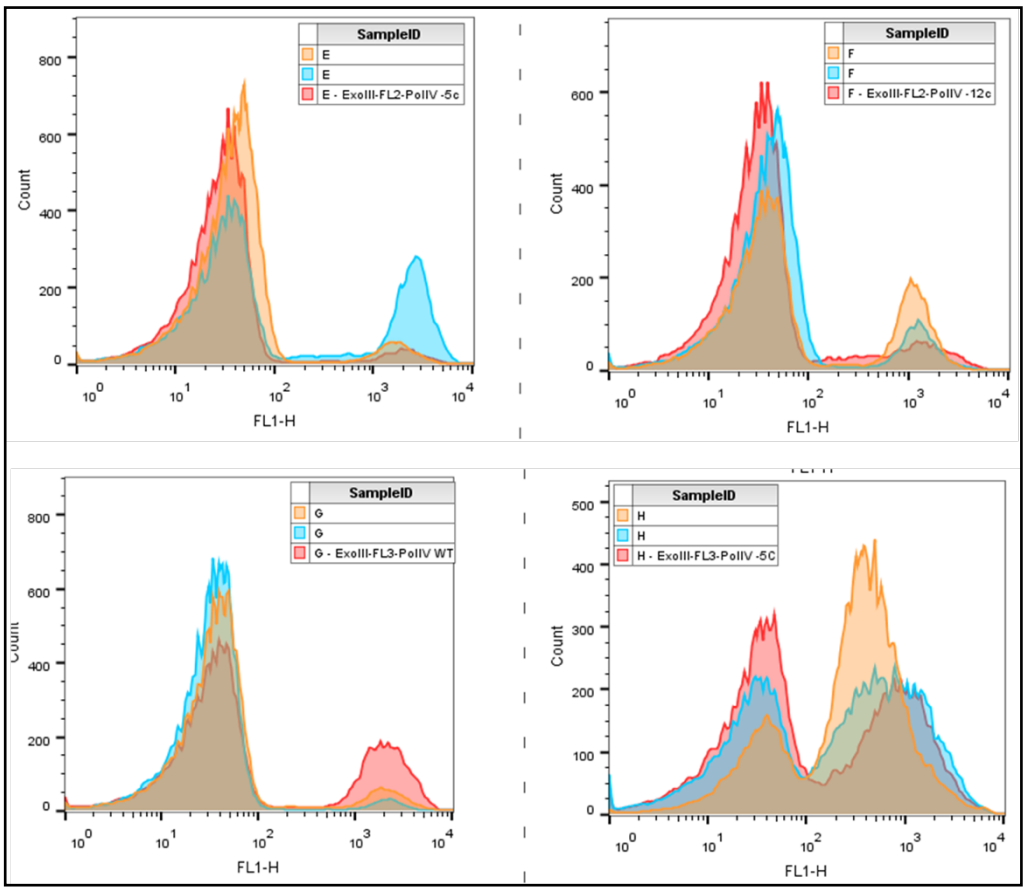
Table 11.29 – Summary of Stage 1 constructs used to determine the effects of expressing Exo – Pol fusion proteins.

Assembly	Construct	Parts Required	Diagnostic PCR – Product Size (PRM L1FOR – L3 REV)
A	HP2 – ExoIII (FLAG) – Fusion Linker 1 – PolIV WT – TetR.CamR.pMB1KanR	A1,3,6,15,16,17	3776 bp
B	HP2 – ExoIII (FLAG) – Fusion Linker 1 – PolIV -5C – TetR.CamR.pMB1KanR	A1,3,9,15,16,17	3758 bp
C	HP2 – ExoIII (FLAG) – Fusion Linker 1 – PolIV12C – TetR.CamR.pMB1KanR	A1,3,12,15,16,17	3740 bp
D	HP2 – ExoIII (FLAG) – Fusion Linker 2 – PolIV WT – TetR.CamR.pMB1KanR	A1,4,7,15,16,17	3776 bp
E	HP2 – ExoIII (FLAG) – Fusion Linker 2 – PolIV-5C – TetR.CamR.pMB1KanR	A1,4,10,15,16,17	3758 bp
F	HP2 – ExoIII (FLAG) – Fusion Linker 2 – PolIV -12C – TetR.CamR.pMB1KanR	A1,4,13,15,16,17	3740 bp
G	HP2 – ExoIII (FLAG) – Fusion Linker 3 – PolIV WT – TetR.CamR.pMB1KanR	A1,5,8,15,16,17	3776 bp
H	HP2 – ExoIII (FLAG) – Fusion Linker 3 – PolIV -5C– TetR.CamR.pMB1KanR	A1,5,11,15,16,17	3758 bp
I	HP2 – ExoIII (FLAG) – Fusion Linker 3 – PolIV -12C – TetR.CamR.pMB1KanR	A1,5,14,15,16,17	3740 bp
N/A	HP2 – ExoIII (FLAG) – Linker 5 – RBS.Pol IV WT - TetR.CamR.pMB1KanR	A1,18,19,15,16,17	Could Not Be Recovered
J	HP2 – ExoIII (FLAG) – Linker 5 – RBS.Pol IV -5C - TetR.CamR.pMB1KanR	A1,18,20,15,16,17	
K	HP2 – ExoIII (FLAG) – Linker 5 – RBS.Pol IV-12C - TetR.CamR.pMB1KanR	A1,18,20,15,16,17	
L	HP2 – 5-3Exo (FLAG) – Fusion Linker 1 – PolIV WT – TetR.CamR.pMB1KanR	B1,3,6,15,16,17	3776 bp
M	HP2 – 5-3Exo (FLAG) – Fusion Linker 1 – PolIV -5C – TetR.CamR.pMB1KanR	B1,3,9,15,16,17	3758 bp
N	HP2 – 5-3Exo (FLAG) – Fusion Linker 1 – PolIV12C – TetR.CamR.pMB1KanR	B1,3,12,15,16,17	3740 bp
O	HP2 – 5-3Exo (FLAG) – Fusion Linker 2 – PolIV WT – TetR.CamR.pMB1KanR	B1,4,7,15,16,17	3776 bp
P	HP2 – 5-3Exo (FLAG) – Fusion Linker 2 – PolIV-5C – TetR.CamR.pMB1KanR	B1,4,10,15,16,17	3758 bp
Q	HP2 – 5-3Exo (FLAG) – Fusion Linker 2 – PolIV -12C –TetR.CamR.pMB1KanR	B1,4,13,15,16,17	3740 bp
R	HP2 – 5-3Exo (FLAG) – Fusion Linker 3 – PolIV WT – TetR.CamR.pMB1KanR	B1,5,8,15,16,17	3776 bp
S	HP2 – 5-3Exo (FLAG) – Fusion Linker 3 – PolIV -5C– TetR.CamR.pMB1KanR	B1,5,11,15,16,17	3758 bp
T	HP2 – 5-3Exo (FLAG) – Fusion Linker 3 – PolIV -12C –TetR.CamR.pMB1KanR	B1,5,14,15,16,17	3740 bp
N/A	HP2 – 5-3Exo (FLAG) – Linker 5 – RBS.Pol IV WT - TetR.CamR.pMB1KanR	B1,18,19,15,16,17	Could Not Be Recovered
U	HP2 – 5-3Exo (FLAG) – Linker 5 – RBS.Pol IV -5C - TetR.CamR.pMB1KanR	B1,18,20,15,16,17	
V	HP2 – 5-3Exo (FLAG) – Linker 5 – RBS.Pol IV-12C - TetR.CamR.pMB1KanR	B1,18,21,15,16,17	
W	TetR.CamR.pMB1KanR (control)	Already Had	
X	pMB1KanR.Cam (control)	Already Had	Already Had
Y	J23101 – RBS.5-3EXO – CamR – p15AKanR	AB 23,24,16,17	**Could Not Be Recovered**
N/A	J23101 – RBS.POLIVWT – CamR –p15KanR	AB 23,25,16,17	**Could Not Be Recovered**
Z	J23101 – RBS.POLIV-5C – CamR –p15KanR	AB 23,26,16,17	2034 bp
AA	J23101 – RBS.POLIV-12C – CamR –p15KanR	AB 23,27,16,17	2019 bp

Table 11.30 *– Summary of the Stage 2 Constructs used to explore the error-prone repair activity when co expressing AID-T7 and Exo-Pol IV fusion proteins*

General Arrangement	pTet.101 – ATF – mCherry - TetRep.- pTet.101 - ExoIII (FLAG) - Fusion Linker – Pol IV - TetRep
Construct	Variant
A	ExoIII (FLAG) Flexible Linker wt PolIV
B	ExoIII (FLAG) .Fusion Linker 1.PolIV -5C.TetRep
C	ExoIII (FLAG).Fusion Linker 1. PolIV-12.TetRep
D	ExoIII (FLAG) Fusion Linker 2.PolIV WT.TetRep
E	ExoIII (FLAG) Fusion Linker 2.PolIV -5C.TetRep
F	ExoIII (FLAG) Fusion Linker 2.PolIV -12C .TetRep
G	ExoIII (FLAG) Fusion Linker 3.PolIV WT.TetRep
H	ExoIII (FLAG) Fusion Linker 3.PolIV -5C .TetRep
I	ExoIII (FLAG) Fusion Linker 3.PolIV -12C .TetRep
J	ExoIII (FLAG).Linker 5.RBS.Pol IV-5
K	ExoIII (FLAG).Linker 5.RBS.Pol IV -12
L	5'3'EXO Fusion Linker 1.PolIV WT.TetRep
M	5'3 EXO Fusion Linker 1.PolIV -5C.TetRep
N	5'3 EXO Fusion Linker 1. PolIV12
O	5'3 EXO Fusion Linker 2.PolIV WT.TetRep
P	5'3 EXO Fusion Linker 2.PolIV -5C.TetRep
Q	5'3 EXO Fusion Linker 2.PolIV -12C .TetRep
R	5'3 EXO Fusion Linker 3.PolIV WT.TetRep
S	5'3 EXO Fusion Linker 3.PolIV -5C .TetRep
T	5'3 EXO Fusion Linker 3.PolIV -12C .TetRep
U	5'3 EXO – Linker 5 – RBS.Pol IV -5
V	5'3 EXO – Linker 5 – RBS.Pol IV -12
W	pMB1Kan-Amp


SampleID
A
A
A - ExoIII-FL1-Pol IV WT
B
B
B - ExoIII-FL1-PolIV -5c
C
C
C - ExoIII-FL1-PolIV -12c
D
D
D - ExoIII-FL2-PolIV WT
Count
FL1-H


SampleID
E
E
E - ExoIII-FL2-PolIV -5c
F
F
F - ExoIII-FL2-PolIV -12c
G
G
G - ExoIII-FL3-PolIV WT
H
H
H - ExoIII-FL3-PolIV -5C
Count
FL1-H

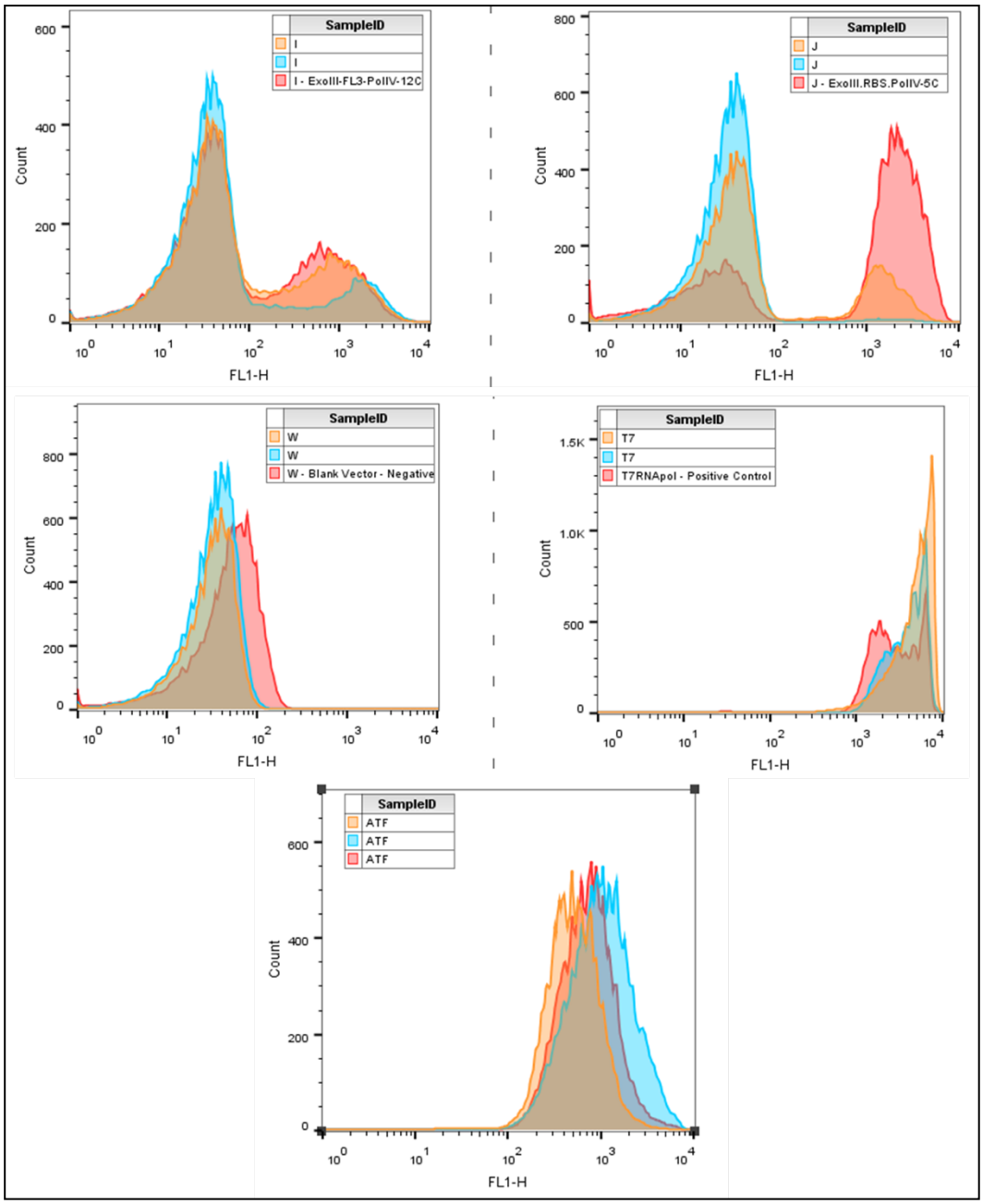


Figure 11.10 *– Differing fluorescence profiles for a range of Exo III – Pol IV fusion constructs expressed within the target GM31 E.coli strain*

11.6 Relating to Chapter 8– Final Conclusions and Future Work

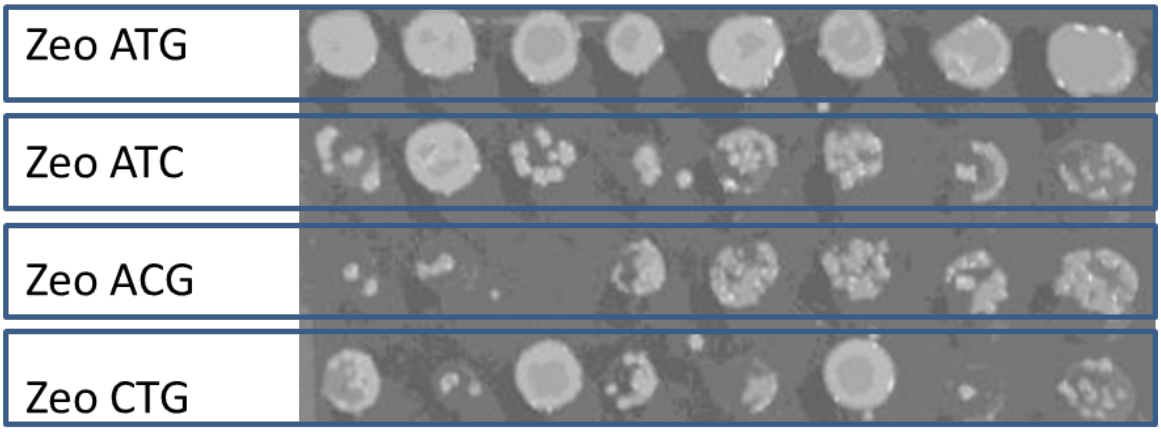
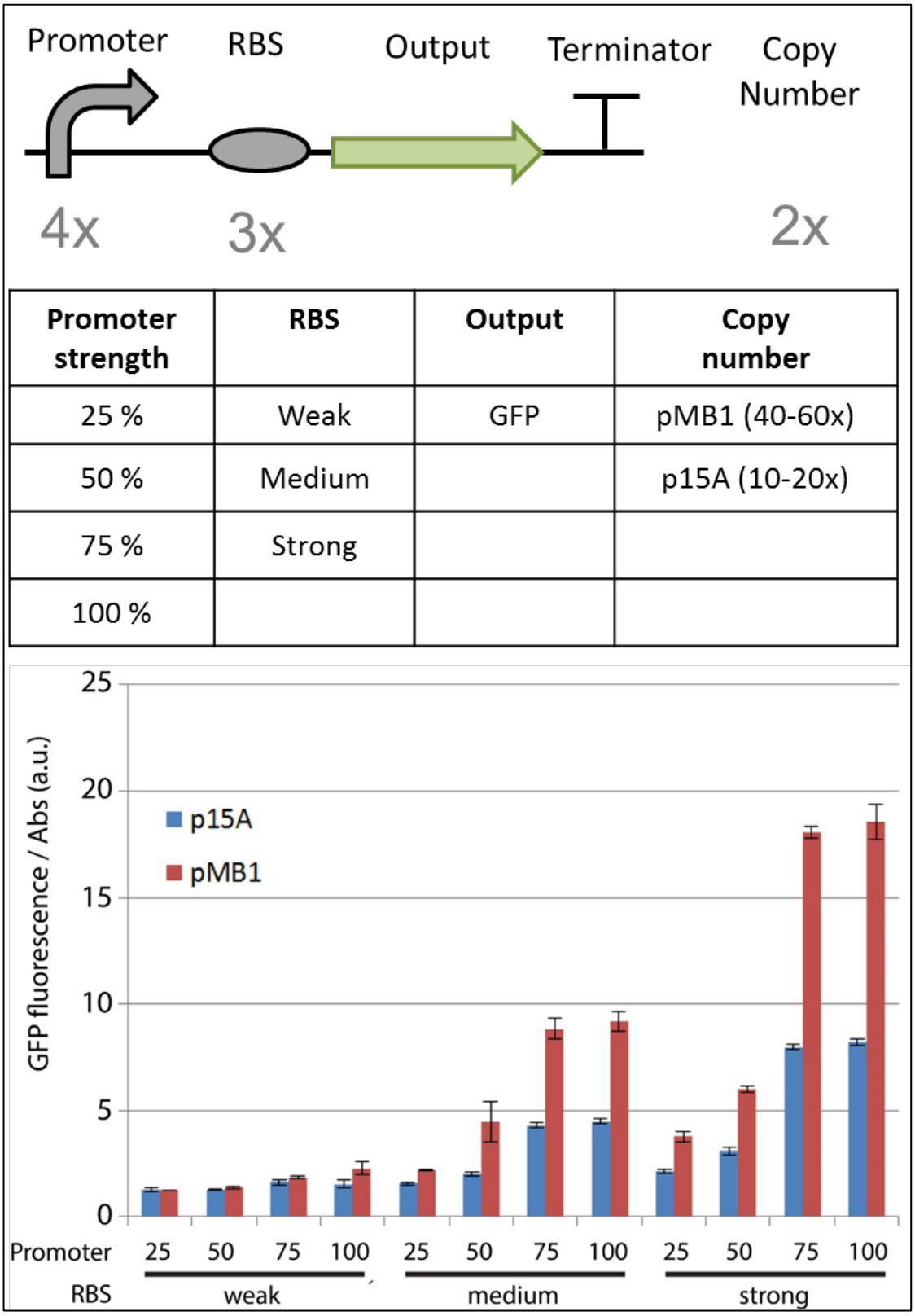


Figure 11.15 *– Recovery of antibiotic resistance in a preliminary gain of function assay used to explore the effect of expression error-prone ExoIII-Pol IV variants alongside the ATF mutator device*

	WT sfGFP Top Line %GC = 51%	sfGFP G/C Bottom Line %GC = 63.8%	
1	ATGCGTAAAGGCGAAGAACTGTTCACGGGCGTAGTTCCGATTCTGGTCGAGCTGGACGGC		60
1	ATGCGCAAGGGCGAGGAGCTCTTCACCGGCGTCGTCCCCATCCTCGTCGAGCTCGACGGC		60
61	GATGTGAACGGTCATAAGTTTAGCGTTCGCGGTGAAGGTGAGGGCGACGCGACCAACGGC		120
61	GACGTCAACGGCCACAAGTTCTCCGTCCGCGGCGAGGGCGAGGGCGACGCCACCAACGGC		120
121	AAACTGACCCTGAAGTTCATCTGCACCACCGGTAAACTGCCGGTGCCTTGGCCGACCTTG		180
121	AAGCTCACCCTCAAGTTCATCTGCACCACCGGCAAGCTCCCCGTCCCCTGGCCCACCCTC		180
181	GTGACGACGTTGACGTATGGCGTGCAGTGTTTTGCGCGTTATCCGGACCACATGAAACAA		240
181	GTCACCACCCTCACCTACGGCGTCCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAG		240
241	CACGATTTCTTCAAATCTGCGATGCCGGAGGGTTACGTCCAGGAGCGTACCATTTCCTTC		300
241	CACGACTTCTTCAAGTCCGCCATGCCCGAGGGCTACGTCCAGGAGCGCACCATCTCCTTC		300

Figure 11.16 *– Automated G/C optimisation of target genes to increase number of sites for deamination as points of entry for the ATF mutator*



Promoter strength	RBS	Output	Copy number
25 %	Weak	GFP	pMB1 (40-60x)
50 %	Medium		p15A (10-20x)
75 %	Strong		
100 %			

Figure 11.17 *– Tuning protein expression using modular RBS parts using the BASIC assembly standard (Storch 2016, unpublished)*

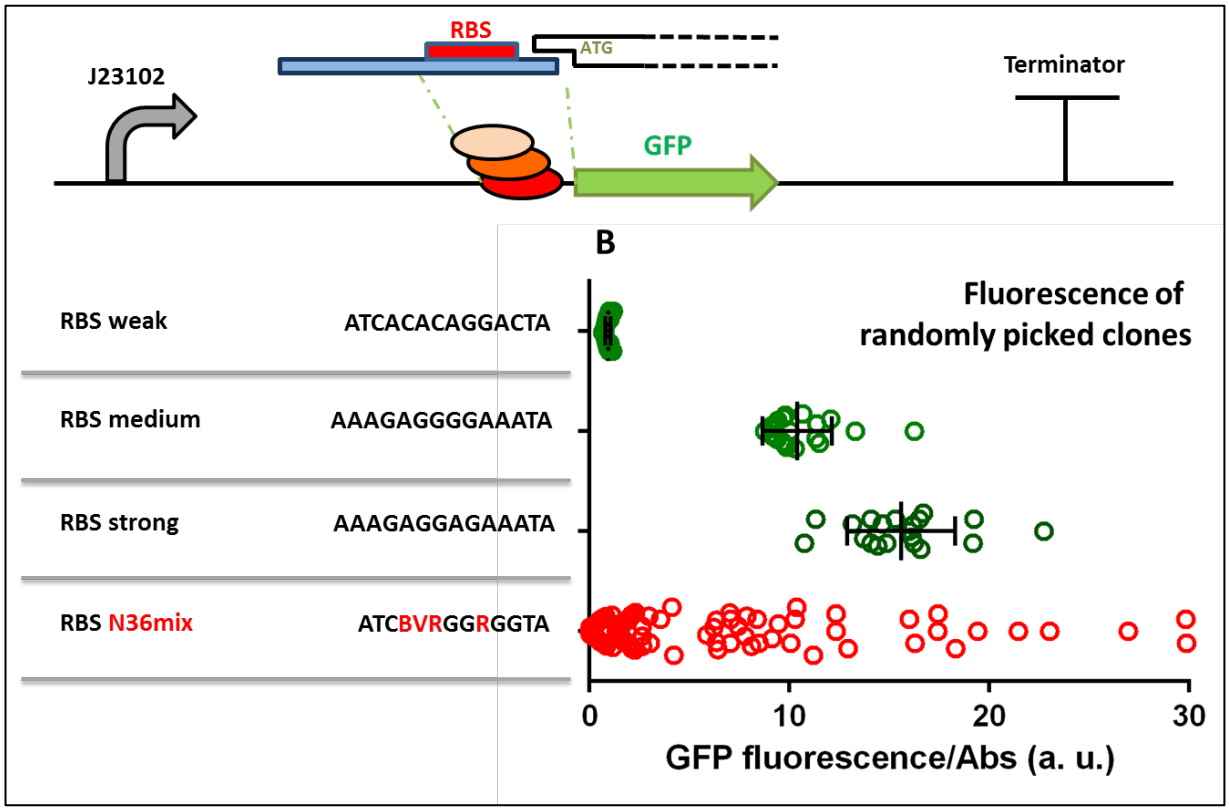


Figure 11.18 *- Tuning protein expression using functional RBS linkers using the BASIC assembly standard (Storch 2016, Unpublished)*

11.6.1 Other

AID protein expression and purification guide

Early intentions were to attempt purification of AID or ATF for *in vitro* characterisation studies although focused changed to assessing *in vivo* activity. Nevertheless, this brief summary describes a set of expression vectors that were created in the hope of achieving improved protein yield should future attempts to generate *in vitro* models, crystals structures, or detailed clarification of dimerisation.

Justification

- Purification of AID protein in sufficient amounts could enable determination of protein crystal structure, ssDNA binding, interaction model, dimerisation and characterisation of activity *in vitro*.
- AID gene was cloned into a protein purification plasmid (MBP.3C) which he had previously been used to for successful purification of troublesome, insoluble proteins (*Storch, unpublished*).
- Large Maltose Binding Protein (MP) tag used to encourage increased expression and enhance solubility and then removed by site-specific proteolysis before crystallisation. There is a 3C cleavage site for removal.
- Smaller His-Tag for purification may give better yield and can remain during crystallization.
- Separation of the protein from the cell lysate on a column, using affinity chromatography i.e. column has affinity for the tag.
- There may be an option to keep the affinity tag attached to protein if the linker between has fixed rigid design, as the 3D structure of MBP has been solved and could be used as search models to solve crystallography phase problems and/or improve crystallisation of the protein of interest (*Smyth et al, 2003*).

Cloning Strategy

- AID monomer was amplified out using primers below, with addition of NcoI prefix and XhoI sites for cloning.
- Two variants were made, including and excluding His-tag
- A variant with NotI site was included to make a construct with Tag RFP downstream

Primer	Sequence	Direction	Tm (C)
AID.For.NcoI	ccgCCATGGCATCGGTACCGGATCCG a**tggatagcctgctgatgaatcg**	Forward	68
AID.Rev.STOP	**cctgggtttaaccggtcagtaa taa**GCCtcgaGCcac	Reverse	66
AID.Rev.HisA	**ccctgggtttaaccggtcag** GCCtcgaGCcac	Reverse	67
AID.Rev.TagRFP	**ccctgggtttaaccggtcag GCG**GCCGCTGA	Reverse	67
AID.Seq.Primer	GTCGTCAGACTGTCGATGAAGC	Forward	n/a

Construct	Expressed Proteins	Protein size
1. MBP.3C.AID.STOP	MBP.AID	63 kDa
2. MBP.3C.AID. HISx6	MBP.AID.HIS	63.4 kDa
3. MBP.3C.AID.TagRFP.HISx6	MBP.AID.TagRFP.HIS	89.88 kDa

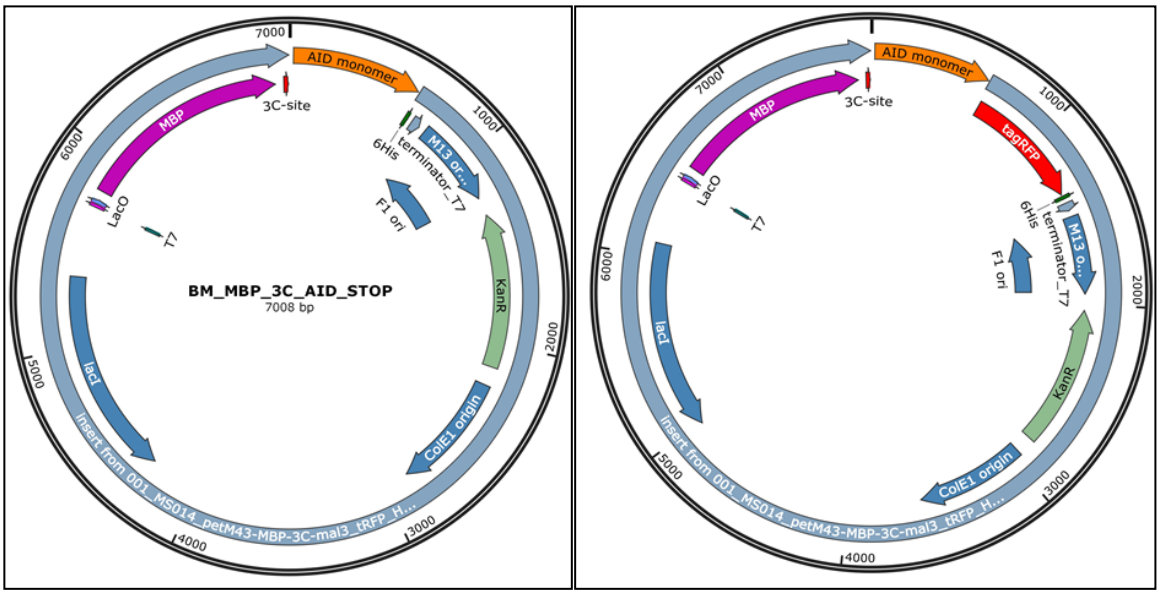


Figure 11.19 *– Final plasmid maps of AID expression vectors engineered to improve downstream purification.*

Proposed Purification Steps

- Use optimal *E.coli* strain (BL21)
- Range of IPTG induction concentrations to be trialled. Inducer added at mid-log phase.
- Lower outgrowth temperature (18°C) to discourage formation of protein inclusion bodies and the loss of insoluble protein.
- Purified AID.TagRFP should be visualised by colour
- AID.HIS can be further purified using HIS column.
- AID.HIS and AID.STOP can also be purified using the HEPARIN column (Heparin has high affinity for DNA binding proteins and has been used for AID purification previously).
- Anti-His Western blot to make sure we have the expected AID protein (~25kDa) using anti-his primary antibody
- Protein concentration measured by Absorbance 280nm based on excitation coefficient (49390 = 1.952 – 1.966 given linear sequence of AID. Denature using 6M guanidinium prior to measurements.
- Protein concentration measured with Bradford Assay – coomassie blue interaction with AID sample and BSA standard should have similar distribution of arginine and aromatic residues for comparison to be highly accurate.
- Protein concentration measured with BCA assay –colometric assay for interactions between peptide backbone, and not the residues that interact with the BCA and therefore doesn't depend on the residues within protein.
- To analyse whether AID exists in monomeric or dimeric forms, analytical ultracentrifuge techniques could be used to check the monodispersity of the sample and potentially obtain information about the shape and molecular mass of the protein. There is a commonly used software package called SedFit that may be useful for analysis. Testing should be carried out using cell free *in-vivo* like conditions using PBS buffers.