

**A novel *in vitro* selection method to aid  
in the development of ribozyme-based  
riboswitches**

Matthew Christopher Haines

Department of Life Sciences

Imperial College London

Submitted in accordance with the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

# Declarations

## Declaration of originality

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

## Copyright declaration

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

# Abstract

Ribozymes are RNA sequences capable of catalysing chemical reactions. For some ribozymes their catalytic rate is influenced by molecular ligands. Those Ligand-responsive ribozymes whose catalytic activity influences gene expression can be referred to as ribozyme-based riboswitches. These sequences have potential applications as therapeutics and in the fields of enzyme and metabolic engineering. However, their application is restricted by the limited variety of detectable ligands. With the aim of addressing this issue, a novel *in vitro* selection method for the enrichment of ribozymes was developed. Compared to analogous methods, this method requires approximately half the time to implement and can be scaled up to 96-well format. Furthermore, the avoidance of size selection steps facilitates larger insertion and deletion mutations, ensuring more diverse regions of the sequence space are explored during selection. In testing and developing this method, new and existing theophylline-activated ribozymes were enriched. This process was assisted by a model, developed to aid the optimisation of selection. As confirmed with NGS data, the model accurately described the dynamics of key sequences for the majority of the selection experiment. Following the enrichment and identification of theophylline-activated ribozymes, several of these sequences were tested for their ability to activate gene expression in *E. coli*. In contrast to the suggestions of previous reports, all tested theophylline-activated ribozymes increased GFP expression in response to ligand concentrations. Although disputed, this result suggests the existence of a relationship between the *in vitro* responses of ribozyme sequences and their ability to function *in vivo*. Though ribozyme-based riboswitches with new ligand-specificities were not characterised during this work, the methods and framework outlined, coupled with the demonstration that selected sequences can regulate gene expression suggests novel ribozyme-based riboswitches should be identified with future work.

# Acknowledgements

Firstly, I wish to thank my three supervisors, Geoff, Guy & Diego for all their support. Geoff, I really appreciated your guidance when alternative avenues of investigation kept coming up during this work and for all your help with the writing. Also thank you for arranging my PIPs placement. I will remember that experience for years to come. Guy, I really valued the fact you made me feel part of your group. I particularly enjoyed all your group's outings and of course the socialising that came with it. Diego, thank you for all your support tackling the various mathematical problems that came up during this work and for your all your general advice. I look forward to seeing all three of you in the future, particularly in relation to the development of a paper based on this work. This, like all of you I hope will follow shortly.

I also wish to acknowledge all the remaining members of Imperial College London who contributed to this work. I must give a special thanks to Dr Marko Storch from the Baldwin lab. Without your persuasion and subsequent work, the last result in this thesis would not exist and the conclusions of this work would be significantly different. I'd also like to thank all the other members of the Baldwin lab and those members who've moved on, such as Ben, Frank & Vasily. Your presence on a day-to-day basis has made this process that much more enjoyable and easier.

I also must acknowledge my friends and family back home, in London and now in Greece. In particular, my Mum and Dad, without whom I'd have probably been living on the streets during the 4<sup>th</sup> year of my PhD. For all of you though, I really valued all the support you have given me over these past 4 years.

Last but certainly not least I must acknowledge Elli. I'm sure you're as relived as I am that this process is, fingers crossed, almost over for me. The only shame is, you've still got a year or two. You'll be pleased to know however that I'm willing to show you the same love and support you've shown me throughout my PhD. I can only hope that all our hard work and devotion will pay off in the end.

# Table of contents

|                                                                                           |    |
|-------------------------------------------------------------------------------------------|----|
| Chapter 1: Introduction .....                                                             | 18 |
| 1.1    Ribozymes .....                                                                    | 18 |
| 1.1.1    Reactions and mechanisms of catalysis .....                                      | 19 |
| 1.2    Natural ribozyme-based riboswitches .....                                          | 21 |
| 1.3    Synthetic ribozymes-based riboswitches.....                                        | 23 |
| 1.3.1    The hammerhead ribozyme sequence and catalysis .....                             | 23 |
| 1.3.2    Ligand-responsive minimal hammerhead ribozymes .....                             | 25 |
| 1.3.3    Hammerhead ribozyme-based riboswitches .....                                     | 26 |
| 1.4    Applications of ribozyme-based riboswitches .....                                  | 29 |
| 1.4.1    Enzyme and metabolic engineering applications .....                              | 29 |
| 1.4.2    Therapeutic applications.....                                                    | 30 |
| 1.5    Current methods for ribozyme-based riboswitch development .....                    | 31 |
| 1.5.1    Identifying a ligand-binding aptamer.....                                        | 32 |
| 1.5.2    Incorporating the aptamer domain into the ribozyme.....                          | 34 |
| 1.6    Ribozyme-based riboswitches with new ligand specificities .....                    | 37 |
| 1.7    Potential and addressing the limitations of in vitro selection methods.....        | 37 |
| 1.8    Summary .....                                                                      | 40 |
| Chapter 2: Developing a novel ribozyme in vitro selection method .....                    | 42 |
| 2.1    Introduction .....                                                                 | 42 |
| 2.2    Objectives.....                                                                    | 46 |
| 2.3    Results.....                                                                       | 47 |
| 2.3.1    Objective 1 – Selection method design .....                                      | 47 |
| 2.3.2    Objective 2 – Previous Trp ribozyme library re-design .....                      | 54 |
| 2.3.3    Objective 3 – Testing the novel cDNA selection steps.....                        | 65 |
| 2.3.4    Objective 4 – Non-specific purification of nucleic acids .....                   | 69 |
| 2.3.5    Objective 5 – PCR primer optimisation and negative selection effectiveness ..... | 73 |

|                                                                                                    |                                                                                  |            |
|----------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|------------|
| 2.3.6                                                                                              | Objective 6 – Effectiveness of positive selection and further optimisation ..... | 78         |
| 2.4                                                                                                | Discussion.....                                                                  | 85         |
| 2.4.1                                                                                              | Selection method .....                                                           | 85         |
| 2.4.2                                                                                              | Library design .....                                                             | 88         |
| 2.4.3                                                                                              | Non-specific purification of nucleic acids .....                                 | 89         |
| 2.4.4                                                                                              | Bait-oligo vs Ligase Mediated Selection.....                                     | 90         |
| 2.5                                                                                                | Summary .....                                                                    | 92         |
| <b>Chapter 3: Enrichment of a theophylline-activated ribozyme .....</b>                            |                                                                                  | <b>93</b>  |
| 3.1                                                                                                | Introduction .....                                                               | 93         |
| 3.2                                                                                                | Objectives.....                                                                  | 94         |
| 3.3                                                                                                | Results .....                                                                    | 94         |
| 3.3.1                                                                                              | Objective 1 – Design of a control library .....                                  | 94         |
| 3.3.2                                                                                              | Objective 2 – Optimisation of selection conditions .....                         | 99         |
| 3.3.3                                                                                              | Objective 3 – Enrichment of a ligand-activated phenotype .....                   | 118        |
| 3.3.4                                                                                              | Objective 4 – Enrichment of a theophylline-activated genotype .....              | 132        |
| 3.4                                                                                                | Discussion.....                                                                  | 137        |
| 3.4.1                                                                                              | Control library design.....                                                      | 137        |
| 3.4.2                                                                                              | Theoretical model .....                                                          | 137        |
| 3.4.3                                                                                              | Selection condition optimisation .....                                           | 139        |
| 3.4.4                                                                                              | Phenotype enriched and selection modifications .....                             | 140        |
| 3.4.5                                                                                              | Genotype enriched .....                                                          | 141        |
| 3.5                                                                                                | Summary .....                                                                    | 142        |
| <b>Chapter 4: Selection for Trp-activated ribozymes &amp; estimating rates of enrichment. ....</b> |                                                                                  | <b>144</b> |
| 4.1                                                                                                | Introduction .....                                                               | 144        |
| 4.2                                                                                                | Objectives.....                                                                  | 145        |
| 4.3                                                                                                | Results.....                                                                     | 145        |

|                                                                                   |                                                                                   |            |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------|
| 4.3.1                                                                             | Objective 1 – Selection for a Trp-activated ribozyme .....                        | 145        |
| 4.3.2                                                                             | Objective 2 – Feasibility of screening larger libraries .....                     | 147        |
| 4.4                                                                               | Discussion.....                                                                   | 157        |
| 4.4.1                                                                             | Selection for a Trp-activated ribozyme .....                                      | 157        |
| 4.4.2                                                                             | Feasibility of screening larger libraries .....                                   | 160        |
| 4.5                                                                               | Summary .....                                                                     | 162        |
| <b>Chapter 5: NGS and characterisation of theophylline-activated riboswitches</b> |                                                                                   | <b>164</b> |
| 5.1                                                                               | Introduction .....                                                                | 164        |
| 5.1.1                                                                             | NGS analysis of the Theophylline Library selection experiment.....                | 164        |
| 5.1.2                                                                             | Characterisation of riboswitch activity for theophylline-activated ribozymes..... | 165        |
| 5.2                                                                               | Objectives.....                                                                   | 166        |
| 5.3                                                                               | Results.....                                                                      | 167        |
| 5.3.1                                                                             | Objective 1 – NGS of the Theophylline Library selection experiment .....          | 167        |
| 5.3.2                                                                             | Objective 2 –NGS data and Sanger sequencing data agree.....                       | 172        |
| 5.3.3                                                                             | Objective 3 – Insertion deletion mutations tolerated by the method .....          | 176        |
| 5.3.4                                                                             | Objective 4 – Accuracy of model simulation.....                                   | 178        |
| 5.3.5                                                                             | Objective 5 – Identification of other theophylline-activated ribozymes.....       | 180        |
| 5.3.6                                                                             | Objective 6 – Characterisation of theophylline-activated riboswitches.....        | 193        |
| 5.4                                                                               | Discussion.....                                                                   | 197        |
| 5.4.1                                                                             | Next-generation sequencing of the selection experiment .....                      | 197        |
| 5.4.2                                                                             | Theophylline Library results.....                                                 | 198        |
| 5.4.3                                                                             | Tolerance of insertion and deletion mutations .....                               | 198        |
| 5.4.4                                                                             | Accuracy of selection simulations.....                                            | 199        |
| 5.4.5                                                                             | Sequence 3 .....                                                                  | 200        |
| 5.4.6                                                                             | Cheaters .....                                                                    | 200        |
| 5.4.7                                                                             | Characterisation of theophylline-activated riboswitches .....                     | 205        |
| 5.5                                                                               | Summary .....                                                                     | 209        |

|                                                                                             |     |
|---------------------------------------------------------------------------------------------|-----|
| Chapter 6: Conclusion .....                                                                 | 211 |
| 6.1 Introduction .....                                                                      | 211 |
| 6.2 Findings and their implications .....                                                   | 211 |
| 6.2.1 Development of an <i>in vitro</i> selection method with the required properties ..... | 212 |
| 6.2.2 Identification of ribozyme-based riboswitches with new ligand specificities .....     | 213 |
| 6.3 Recommendations for future work .....                                                   | 215 |
| 6.4 Closing remarks.....                                                                    | 216 |
| Chapter 7: Materials & methods.....                                                         | 218 |
| 7.1 Computational methods .....                                                             | 218 |
| 7.1.1 Programming languages .....                                                           | 218 |
| 7.1.2 RNA folding .....                                                                     | 218 |
| 7.1.3 Nucleic acid thermodynamic properties and GC contents .....                           | 218 |
| 7.1.4 Densitometry and % cleaved calculations .....                                         | 218 |
| 7.2 Experimental methods.....                                                               | 219 |
| 7.2.1 Preparation of dsDNA templates from synthetic oligonucleotides .....                  | 219 |
| 7.2.2 RNA synthesis.....                                                                    | 219 |
| 7.2.3 Purification of RNA from IVT reactions .....                                          | 219 |
| 7.2.4 cDNA synthesis.....                                                                   | 220 |
| 7.2.5 Non-specific purification of cDNA.....                                                | 221 |
| 7.2.6 DNA ligase mediated purification of full-length or cleaved cDNA.....                  | 222 |
| 7.2.7 Preparation of cDNA for analysis via urea PAGE.....                                   | 225 |
| 7.2.8 General PCR conditions.....                                                           | 225 |
| 7.2.9 Semi-quantitative PCR .....                                                           | 226 |
| 7.2.10 Quantification of nucleic acid in solution .....                                     | 226 |
| 7.2.11 PCR purification.....                                                                | 227 |
| 7.2.12 Electrophoresis .....                                                                | 227 |
| 7.2.13 Staining and imaging of gels .....                                                   | 227 |

|                               |                                                                                           |     |
|-------------------------------|-------------------------------------------------------------------------------------------|-----|
| 7.3                           | Reagents.....                                                                             | 229 |
| 7.3.1                         | Buffers.....                                                                              | 229 |
| 7.3.2                         | Chemically synthesised oligonucleotides.....                                              | 229 |
| 7.4                           | Methods specific to Chapter 2.....                                                        | 232 |
| 7.4.1                         | Bait-oligo mediated full-length cDNA selection.....                                       | 232 |
| 7.4.2                         | Bait-oligo mediated cleaved cDNA selection and adapter ligation .....                     | 232 |
| 7.4.3                         | SPRI-based purification of DNA from PCRs.....                                             | 233 |
| 7.4.4                         | PCR primer sets 1, 2 & 3 evaluation.....                                                  | 233 |
| 7.5                           | Methods specific to Chapter 3.....                                                        | 233 |
| 7.5.1                         | Synthesis of cDNA without RNA purification .....                                          | 233 |
| 7.5.2                         | Sanger sequencing .....                                                                   | 234 |
| 7.6                           | Methods specific to Chapter 5.....                                                        | 234 |
| 7.6.1                         | NGS.....                                                                                  | 234 |
| 7.6.2                         | Characterisation of sequences in <i>E. coli</i> for theophylline-dependent GFP expression | 237 |
| Chapter 8: Bibliography ..... |                                                                                           | 239 |
| Chapter 9: Appendix .....     |                                                                                           | 251 |
| 9.1                           | Functions and parameters for selection simulations.....                                   | 251 |
| 9.2                           | Appendix for Chapter 2.....                                                               | 253 |
| 9.2.1                         | Candidate bait-oligo toehold sequences .....                                              | 253 |
| 9.2.2                         | Candidate RT primer sequences .....                                                       | 253 |
| 9.2.3                         | Timing of selection method .....                                                          | 254 |
| 9.3                           | Appendix for Chapter 4.....                                                               | 254 |
| 9.3.1                         | Maximum number of sequences sampled during each round.....                                | 254 |
| 9.3.2                         | Theo Con rate constants and TRT inactivation time .....                                   | 255 |
| 9.3.3                         | % cleaved and fitness from cleavage rate constant and inactivation/IVT time.....          | 256 |
| 9.3.4                         | Most-fit ligand-unresponsive sequences.....                                               | 256 |

|       |                                                                |     |
|-------|----------------------------------------------------------------|-----|
| 9.4   | Appendix for Chapter 5 .....                                   | 257 |
| 9.4.1 | Algorithm to remove obvious cheaters .....                     | 257 |
| 9.4.2 | Conditions used to measure fitness of selected sequences ..... | 258 |
| 9.4.3 | Flow cytometry histograms .....                                | 258 |

# List of figures

|                                                                                                                                                                                                                                       |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Figure 1-1: Cleavage and ligation by small nucleolytic ribozymes.....                                                                                                                                                                 | 20  |
| Figure 1-2: The Hammerhead ribozyme .....                                                                                                                                                                                             | 24  |
| Figure 1-3: A ligand-responsive ribozyme based on the minimal hammerhead ribozyme .....                                                                                                                                               | 26  |
| Figure 1-4: A common strategy for implementing synthetic ribozyme-based riboswitches in <i>E. coli</i> .                                                                                                                              | 28  |
| Figure 1-5: Process of SELEX for aptamer generation .....                                                                                                                                                                             | 33  |
| Figure 2-1: A previously conceived library <sup>123</sup> , designed with aim of selecting for hammerhead ribozymes activated by the amino acid Trp .....                                                                             | 45  |
| Figure 2-2: Potential phenotypes of sequences during selection.....                                                                                                                                                                   | 48  |
| Figure 2-3: Principle and an alternative scheme for enriching ligand-activated ribozymes .....                                                                                                                                        | 50  |
| Figure 2-4: Bait-oligo mediated selection of cDNA.....                                                                                                                                                                                | 53  |
| Figure 2-5: bait-oligo hybridisation inhibition by stem III formation.....                                                                                                                                                            | 55  |
| Figure 2-6: Effect of removing anti-RBS nucleotides on library stability .....                                                                                                                                                        | 57  |
| Figure 2-7: MFE structure of the ribozyme library containing the shortened stem III and toehold sequences – Roman numerals denote stem numbers, while colours indicate base-pairing probabilities, according to the given scale. .... | 60  |
| Figure 2-8: RT primer landing pads.....                                                                                                                                                                                               | 63  |
| Figure 2-9: Selected library design.....                                                                                                                                                                                              | 65  |
| Figure 2-10: cDNA preparation & bait-oligo full-length cDNA purification .....                                                                                                                                                        | 67  |
| Figure 2-11: bait-oligo mediated cleaved cDNA purification and ligation.....                                                                                                                                                          | 69  |
| Figure 2-12: Optimisation of NaCl and PEG for the purification of cDNA.....                                                                                                                                                           | 71  |
| Figure 2-13: Effect of volume on the efficiency of RNA purification using SPRI.....                                                                                                                                                   | 72  |
| Figure 2-14: DNA purification using SPRI vs silica spin-columns .....                                                                                                                                                                 | 73  |
| Figure 2-15: Tested PCR primer sets.....                                                                                                                                                                                              | 75  |
| Figure 2-16: Products of PCRs using either primer sets 1, 2 or 3. ....                                                                                                                                                                | 76  |
| Figure 2-17: Effectiveness of bait-oligo mediated negative selection.....                                                                                                                                                             | 78  |
| Figure 2-18: Effectiveness of bait-oligo mediated positive selection .....                                                                                                                                                            | 80  |
| Figure 2-19: Ligase mediated cDNA selection .....                                                                                                                                                                                     | 82  |
| Figure 2-20: Ligase Mediated Selection of full-length or cleaved cDNA.....                                                                                                                                                            | 84  |
| Figure 2-21: Effectiveness of bait-oligo and ligase mediated positive selections .....                                                                                                                                                | 85  |
| Figure 3-1: Design of the Theophylline Library .....                                                                                                                                                                                  | 96  |
| Figure 3-2: Theophylline Library constitutively active sequence.....                                                                                                                                                                  | 98  |
| Figure 3-3: Fitness of all sequences .....                                                                                                                                                                                            | 108 |

|                                                                                                                              |     |
|------------------------------------------------------------------------------------------------------------------------------|-----|
| Figure 3-4: Kinetics and fitness of the Theophylline Control during IVT .....                                                | 111 |
| Figure 3-5: Fitness at the RNA vs cDNA level.....                                                                            | 112 |
| Figure 3-6: TRT reaction.....                                                                                                | 114 |
| Figure 3-7: Comparison of methods used to prepare cDNA.....                                                                  | 115 |
| Figure 3-8: TRT reaction produces high yields and a constant response .....                                                  | 116 |
| Figure 3-9: Simulated enrichment of the Theophylline Control from the Theophylline Library .....                             | 118 |
| Figure 3-10: 1 <sup>st</sup> attempt to enrich a Theophylline-activated phenotype from the Theophylline Library .....        | 120 |
| Figure 3-11: Library instability caused by selection.....                                                                    | 122 |
| Figure 3-12: Improved semi-quantitative PCR optimisation .....                                                               | 126 |
| Figure 3-13: Effectiveness of improved semi-quantitative PCR .....                                                           | 127 |
| Figure 3-14: Optimising DNA template removal.....                                                                            | 129 |
| Figure 3-15: 2 <sup>nd</sup> attempt to enrich a Theophylline-activated phenotype from the Theophylline Library .....        | 131 |
| Figure 3-16: Response and stability of the pool during selection .....                                                       | 132 |
| Figure 3-17: Enrichment of a theophylline-activated genotype.....                                                            | 136 |
| Figure 4-1: Attempt to enrich for a Trp-activated phenotype from the Trp Library .....                                       | 147 |
| Figure 4-2: Enrichment of the Theophylline Control from more complex libraries under the previous selection conditions. .... | 150 |
| Figure 4-3: General effect of delaying ribozyme inactivation during negative selection. ....                                 | 153 |
| Figure 4-4: Optimal negative selection conditions for Theophylline Control enrichment. ....                                  | 156 |
| Figure 4-5: The Trp library may lack elements required for Trp binding.....                                                  | 159 |
| Figure 5-1: Relatively low yield from sequencing .....                                                                       | 169 |
| Figure 5-2: Bioanalyzer trace suggests the barcoded selection experiment was of good quality .....                           | 170 |
| Figure 5-3: Quality and quantity of sequences corresponding with the selection experiment .....                              | 172 |
| Figure 5-4: Enrichment of expected sequences from the Theophylline Library.....                                              | 173 |
| Figure 5-5: Dynamics of Theophylline Library sequences & suspected ligand-unresponsive sequences .....                       | 175 |
| Figure 5-6: Distribution of sequence lengths.....                                                                            | 178 |
| Figure 5-7: Accuracy of simulations describing selection dynamics.....                                                       | 180 |
| Figure 5-8: Multiple sequence alignment of high fold change sequences in final selection cycle.....                          | 181 |
| Figure 5-9: Estimated dynamics of obvious cheaters during selection .....                                                    | 183 |
| Figure 5-10: Selection dynamics of sequences selected for analysis .....                                                     | 185 |
| Figure 5-11: Properties of identified sequences.....                                                                         | 187 |

|                                                                                                      |     |
|------------------------------------------------------------------------------------------------------|-----|
| Figure 5-12: Performance and recalculated fitness for selected sequences .....                       | 190 |
| Figure 5-13: Properties of full-length (F) RNA for sequences 3, 4 & the Theophylline Control .....   | 192 |
| Figure 5-14: Ability of theophylline-activated ribozymes to regulate gene expression in E. coli..... | 196 |
| Figure 9-1: Flow cytometry histograms relating to the data in Figure 5-14 .....                      | 259 |

## List of tables

|                                                                                                                                               |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Table 3-1: Frequency and genotypes of sequences identified following selection.....                                                           | 134 |
| Table 4-1: Modification to negative selection stringencies during Trp ribozyme selection.....                                                 | 146 |
| Table 7-1: Components of T7 DNA Ligase reactions.....                                                                                         | 223 |
| Table 7-2: Components of Tth DNA Ligase reactions .....                                                                                       | 224 |
| Table 7-3: Components of PFU polymerase PCRs.....                                                                                             | 225 |
| Table 7-4: PCR conditions for those assembled as described in Table 7-3.....                                                                  | 226 |
| Table 7-5: Buffers.....                                                                                                                       | 229 |
| Table 7-6: Chemically synthesised oligonucleotides acquired from Integrated DNA Technologies, Inc.<br>.....                                   | 231 |
| Table 7-7: Chemically synthesised oligonucleotides acquired from Twist Bioscience. ....                                                       | 231 |
| Table 7-8: Components added to unpurified IVT reaction to synthesise cDNA .....                                                               | 234 |
| Table 7-9: Amplicon PCR components for NGS sample preparation.....                                                                            | 235 |
| Table 7-10: Amplicon PCR conditions for NGS sample preparation .....                                                                          | 235 |
| Table 7-11: Index PCR components for NGS sample preparation .....                                                                             | 236 |
| Table 7-12: Index PCR conditions for NGS sample preparation .....                                                                             | 236 |
| Table 9-1: Parameters used to generate selection simulations in Figure 3-9 & Figure 5-7.....                                                  | 251 |
| Table 9-2: Parameters used to generate selection simulations in Figure 4-2 .....                                                              | 252 |
| Table 9-3: Parameters used to generate the simulation illustrated in Figure 4-4C when inactivation<br>delayed during negative selection ..... | 252 |
| Table 9-4: Candidate bait-oligo toehold sequences .....                                                                                       | 253 |
| Table 9-5: Candidate RT primer sequences .....                                                                                                | 253 |
| Table 9-6: Time required to implement the selection method described in this thesis .....                                                     | 254 |
| Table 9-7: Parameters required to remove obvious cheaters using algorithm.....                                                                | 257 |

## List of equations

|              |     |
|--------------|-----|
| (3-1) .....  | 100 |
| (3-2) .....  | 100 |
| (3-3) .....  | 100 |
| (3-4) .....  | 101 |
| (3-5) .....  | 101 |
| (3-6) .....  | 102 |
| (3-7) .....  | 102 |
| (3-8) .....  | 102 |
| (3-9) .....  | 103 |
| (3-10) ..... | 104 |
| (3-11) ..... | 104 |
| (3-12) ..... | 105 |
| (3-13) ..... | 106 |
| (3-14) ..... | 106 |
| (3-15) ..... | 124 |
| (3-16) ..... | 124 |
| (5-1) .....  | 188 |
| (5-2) .....  | 188 |
| (7-1) .....  | 219 |
| (9-1) .....  | 255 |
| (9-2) .....  | 256 |
| (9-3) .....  | 256 |
| (9-4) .....  | 256 |

# Acronyms

## Nucleic acid sequences

IUPAC nucleotide code<sup>1</sup> was adopted, enabling acronyms for all nucleotides and their combinations.

## General terms

|                |                                                           |
|----------------|-----------------------------------------------------------|
| +ve            | Positive                                                  |
| -ve            | Negative                                                  |
| bp             | Base pair                                                 |
| dNTP           | Deoxynucleotide triphosphate                              |
| DTT            | Dithiothreitol                                            |
| FACS           | Fluorescence-activated cell sorting                       |
| FRET           | Fluorescence resonance energy transfer                    |
| GFP            | Green fluorescent protein                                 |
| IVT            | <i>In vitro</i> transcription                             |
| LOD            | Limit of detection                                        |
| MFE            | Minimal free energy structure                             |
| MW             | Molecular weight                                          |
| ORF            | Open reading frame                                        |
| PEG            | Polyethylene glycol                                       |
| PPP            | Triphosphate                                              |
| RBS            | Ribosome binding site                                     |
| RNAP           | RNA polymerase                                            |
| RT             | Reverse transcription                                     |
| RTase          | Reverse transcriptase                                     |
| SD             | Standard deviation                                        |
| SELEX          | Systematic evolution of ligands by exponential enrichment |
| SPRI           | Solid Phase Reversible Immobilization (beads)             |
| T <sub>m</sub> | Melting temperature                                       |
| Trp            | Tryptophan                                                |
| TRT            | IVT & RT reactions conducted simultaneously               |
| UTR            | Untranslated region                                       |

### *Species*

*B. subtilis*

*Bacillus subtilis*

*E. coli*

*Escherichia coli*

*C. difficile*

*Clostridium difficile*

*Tth*

*Thermus thermophilus*

### Nucleic acids, nucleotides and their derivatives

ATP

Adenosine triphosphate

c-di-GMP

cyclic di-guanosyl-5'-monophosphate

DNA

Deoxyribonucleic acid

HHR

Hammerhead ribozyme

mRNA

Messenger RNA

RNA

Ribonucleic acid

tRNA

Transfer RNA

# Chapter 1: Introduction

## 1.1 Ribozymes

For all known living organisms, the processes of storing genetic information and catalysing biological reactions are primarily mediated by DNA and protein-based enzymes respectively. However, it has been proposed that prior to the emergence of DNA and protein, RNA functioned independently of both these biomolecules storing genetic information and catalysing the reactions necessary to support life<sup>2</sup>. This hypothesis was supported by the identification of viral bacteriophage genomes encoded in RNA, demonstrating that RNA was capable of storing genetic information<sup>3</sup>. 15 to 20 years after this hypothesis, group I introns were discovered<sup>4</sup>, proving that RNA has the necessary chemistry to catalyse chemical reactions. The term “ribozyme” has since been used to describe all RNA molecules which possess the ability to catalyse chemical reactions.

Following the identification of group I introns, several additional ribozymes and ribozyme classes have been discovered. One notable discovery was the presence of a ribozyme within the ribosome core<sup>5,6</sup>. In this case, a key adenine residue is directly involved in catalysing the peptidyl-transferase reaction leading to polypeptide and hence protein synthesis. In this respect, life as it currently functions is entirely dependent on the functionality of ribozymes.

It must be noted however that this peptidyl-transferase ribozyme is unique in that all other known natural ribozymes catalyse phosphoryl transfer reactions<sup>7</sup>. Like the ribosome, some of these ribozymes exist as part of large ribonucleoprotein complexes. These include RNase P<sup>8</sup> and the spliceosome<sup>9</sup> which catalyse the processing of tRNA and pre-mRNA splicing, respectively. However, simpler ribozymes, such as small nucleolytic ribozymes and the previously mentioned group I introns, form the basis of known ribozyme-based riboswitches.

### 1.1.1 Reactions and mechanisms of catalysis

To understand the functioning of ribozyme-based riboswitches, it is important to understand the reactions that the underlying ribozymes catalyse and the mechanisms underpinning their catalysis. Both group I introns and small nucleolytic ribozymes catalyse transesterification reactions leading to cleavage and/or ligation of RNA<sup>10,11</sup>.

In the case of cleavage by small nucleolytic ribozymes, the scissile bond (red dotted-line – Figure 1-1A), is attacked by the 2'-O group from the preceding nucleotide residue<sup>7,10,12-14</sup> (Figure 1-1A). Cleavage results in the breaking of the scissile bond and the formation of two separate RNA products; one containing a 2'-3'-cyclic phosphate and another containing a 5'-OH group (Figure 1-1C). It has been shown that several ribozymes can also catalyse the reverse, ligation reaction if the two cleavage products in Figure 1-1C are brought into proximity of one another<sup>15-17</sup>. This is true even under physiological conditions, although depending on the ribozyme the exact mechanism of ligation may differ from that of cleavage<sup>17</sup>. Therefore, the favourability of ligation and cleavage reactions varies between ribozymes within this class.

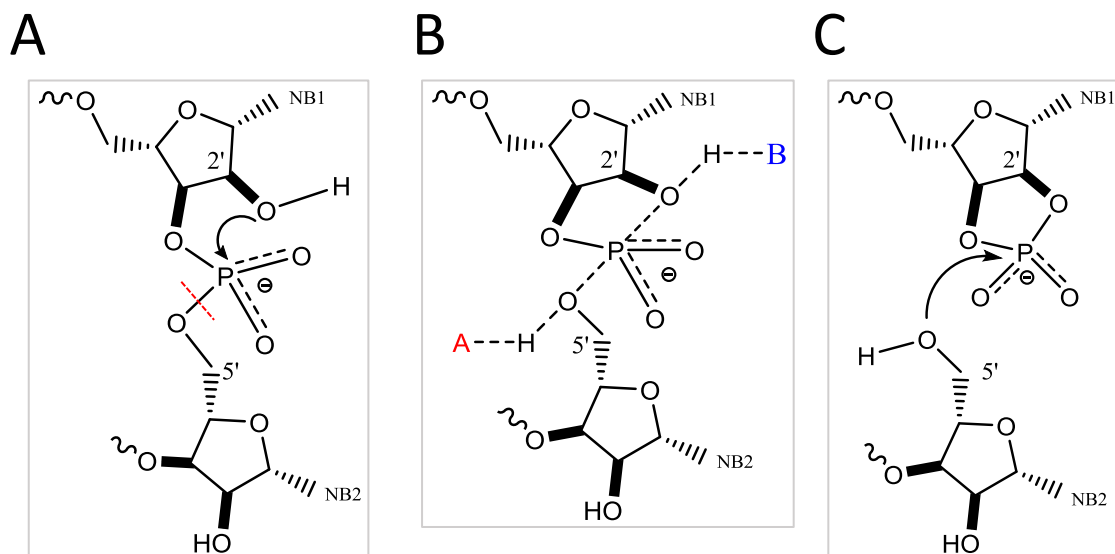


*Figure 1-1: Cleavage and ligation by small nucleolytic ribozymes – (A) Cleavage is initiated by the 2'-O which acts as a nucleophile during this reaction. This nucleophilic attack results in the breaking of the scissile bond (red dotted-line) and the production of the species in (C). (B) In-line orientation of the 2'-O and scissile bond along with general acid-base catalysis often enhances the rate of cleavage. A and B denote acid and base, respectively, where their exact chemical identity depends on the type of the ribozyme. (C) Products of the cleavage reaction in (A). A reverse ligation reaction can also be catalysed by nucleolytic ribozymes (refer to text) whereby the 5'-O attacks the phosphorus atom. NB1/2 denotes nitrogenous bases.*

The exact mechanisms small nucleolytic ribozymes utilise to catalyse transesterification reactions appear to vary subtly from one member to the next<sup>18–20</sup>. However, based on the crystal structures of six of these species<sup>18,19,21–24</sup>, two catalytic themes are often involved<sup>14</sup> (Figure 1B). One of these themes involves the positioning of nucleotides residues before and after the scissile bond in a specific catalytic orientation. This orientation aligns the 2'-O in-line with the scissile bond, leading to the required conformation for nucleophilic attack. General acid-base catalysis is a second commonly found mechanism amongst small nucleolytic ribozymes. In the case of cleavage, the general base increases the rate of the reaction by deprotonating the 2'-O, increasing its electronegativity and as such

encouraging attack. The general acid meanwhile protonates the 5'-O producing a more stable OH leaving group.

In contrast to small nucleolytic ribozymes, group I introns catalyse several transesterification reactions<sup>11,25</sup>. The first of these reactions leads to cleavage between the group I intron and the 5' exon of the mRNA. This reaction is similar to that illustrated in Figure 1-1A except that the 3'-OH group of an exogenous guanosine or guanine ribonucleotide acts as nucleophile instead of the 2'-OH of the preceding nucleotide. The liberated 3'-OH of the 5' exon is now free to attack the phosphoester linkage between the 3' exon and intron splicing the two exons together and releasing the intron. Depending on the group I intron a further transesterification reaction may occur within the released intron.

A further difference between group I introns and small nucleolytic ribozymes is the mechanisms that these ribozymes use to catalyse transesterification reactions. For instance, metal ions appear to play a much larger role in catalysing the transesterification reactions catalysed by group I introns than small nucleolytic ribozymes<sup>7</sup>. Specifically, metal ions have been identified in the group I ribozyme active sites, stabilising the charges on attacking and leaving groups as well as playing roles in orientating the nucleophile<sup>26</sup>.

## 1.2 Natural ribozyme-based riboswitches

Riboswitches are cis-acting, RNA-based elements which regulate the expression of genes in response to the concentration of ligand molecules<sup>27,28</sup>. Ligands modulating the activity of natural riboswitches include amino acids<sup>29,30</sup>, metal ions<sup>31,32</sup>, signalling molecules<sup>33,34</sup> and coenzymes<sup>35</sup>. Several possible processes are exploited by riboswitches to modulate the expression of genes. These include the processes of transcription<sup>36</sup>, translation<sup>34,37</sup> or mRNA degradation<sup>38</sup>. Changes in the rate of one or more of these processes ultimately effects the expression of the regulated gene. To couple the concentration of ligands to the rate of gene expression, riboswitches use a variety of molecular mechanisms including the activity of ribozymes<sup>27</sup>.

Two notable examples of natural ribozyme-based riboswitches are the *glmS* ribozyme<sup>39</sup> and the c-di-GMP dependent group I intron<sup>34</sup>. The *glmS* ribozyme is a small nucleolytic ribozyme which requires a glucosamine-6-phosphate cofactor for efficient cleavage<sup>19,39</sup>. One of the roles glucosamine-6-phosphate plays during catalysis is as a general acid<sup>40</sup>, protonating the 5'-O leaving group (refer to Figure 1-1B). As such, the rate of cleavage of this small nucleolytic ribozyme is heavily dependent on the concentration of this ligand. This ribozyme is located within the 5' UTR of the *glmS* mRNA. Cleavage therefore results in the generation of a transcript with a 5'-OH group (Figure 1-1C). Within host organisms such as *B. subtilis*, this functional group is recognised by native RNases leading to an increase in the rate of mRNA degradation and a reduction in the expression of the *glmS* gene<sup>41</sup>. Importantly, the product of the *glmS* gene directly synthesises the *glmS* ribozyme cofactor, glucosamine-6-phosphate. Therefore, this ribozyme functions to implement a negative feedback loop ensuring that concentrations of glucosamine-6-phosphate are maintained at homeostatic levels within the cell.

The c-di-GMP dependent group I intron has a considerably different architecture compared to the *glmS* ribozyme. Rather than the ribozyme domain directly interacting with the ligand, a separate c-di-GMP aptamer domain senses the ligand and communicates to the group I intron ribozyme<sup>34</sup>. In the presence of low concentrations of the ligand, the c-di-GMP aptamer of the riboswitch is unbound and the self-splicing intron splices out the RBS from the mRNA, inhibiting translation. In the presence of high concentrations of c-di-GMP, the c-di-GMP aptamer binds the ligand causing a conformational change within the riboswitch. This leads to alternative splicing with the formation of a functional RBS and an increase in the expression of virulence genes in the host organism, *C. difficile*<sup>42,43</sup>.

These natural ribozyme-based riboswitches demonstrate the ability of ribozymes to correlate gene expression with ligand concentrations, mediating important cellular functions. However, the complex mechanisms involved in detecting and coupling ligand concentrations to gene expression have meant

that other, simpler and easier to engineer ribozymes have been considered when attempting to generate novel ribozyme-based riboswitches.

### 1.3 Synthetic ribozymes-based riboswitches

With the exception of the first synthetic ribozyme-based riboswitch<sup>44</sup>, the majority of synthetic ribozyme-based riboswitches have been generated using the hammerhead ribozyme<sup>45–56</sup>.

#### 1.3.1 The hammerhead ribozyme sequence and catalysis

The hammerhead ribozyme is one of the smallest and most well-studied ribozymes<sup>57,58</sup>. It is characterised by a three-way junction with stems I - III extending from a core region<sup>59</sup> (Figure 1-2A). Each of the stems can terminate in a hairpin loop or 5'-PO<sub>4</sub> and 3'-OH termini. The core region of the ribozyme is composed of 17 nucleotides numbered 1 through 17 with the nucleotide residue 3' of the cleavage site designated as the first nucleotide (1.1)<sup>60</sup>. Nucleotides involved in stem base-pairings are also given a decimal value to indicate their position in the stem. For instance, A15.1 is the 15<sup>th</sup> nucleotide of the core and is involved in the 1<sup>st</sup> base pair in stem III. Of these 17 core nucleotides, 13 are highly conserved and almost always required for efficient catalysis<sup>59,61,62</sup>. Three of the remaining nucleotides including the N1.1 residue which is immediately downstream of the scissile bond (red dotted-line – Figure 1-2A) are variable. However, base pairing is required between N1.1 and N2.1 for the correct orientation of the active site and hence only N7 is truly degenerate. Finally, a guanine residue is not permitted at position 17, presumably to enable interactions between this residue and A13 or to prevent the disturbance of other critical interactions.

During catalysis the hammerhead ribozyme assumes an active Y-conformation<sup>23</sup> (Figure 1-2B). Cleavage of the ribozyme proceeds with the 2'-O of H17 attacking the scissile bond in a manner analogous to that illustrated in Figure 1-1A. The rate of cleavage is enhanced by general acid-base catalysis similar to that depicted in Figure 1-1B<sup>13,23</sup>. Specifically, evidence suggests that the N1 of G12 acts as a general base, deprotonating the 2'-OH of H17. Meanwhile, the 2'-OH group of G8 stabilised by a divalent metal ion acts as a general acid, stabilising the 5'-O leaving group on N1.1.



### 1.3.2 Ligand-responsive minimal hammerhead ribozymes

Prior to the generation of ribozyme-based riboswitches using the hammerhead ribozyme, ligand-responsive ribozymes which functioned *in vitro* were generated using minimal hammerhead ribozyme sequences<sup>65</sup>. In this pioneering work the authors used a minimal hammerhead ribozyme as a basis from which they engineered in ligand-binding and allosteric transitions. To achieve these two properties in the context of the minimal hammerhead ribozyme, stem II was replaced with an adenosine-binding sequence, referred to as an aptamer, and a communication domain in a similar manner to that illustrated in Figure 1-3. Upon binding of the ligand adenosine or ATP to the aptamer, a conformational change would occur leading to inhibition of substrate cleavage as characterised by a change in the rate of the reaction.

Several ligand-responsive ribozymes similar to the one illustrated in Figure 1-3 were developed using a variety of methods<sup>66–69</sup>. These ligand-responsive ribozymes were engineered to recognise ligands including flavin mononucleotide, cyclic mononucleotides, theophylline and antibiotics, demonstrating flexibility in the types of aptamers that can be incorporated into the design. Given all these ribozymes were based on the minimal hammerhead, they all required higher than physiological concentrations of metal ions for their functionality. As such they could not function as ribozyme-based riboswitches *in vivo*.

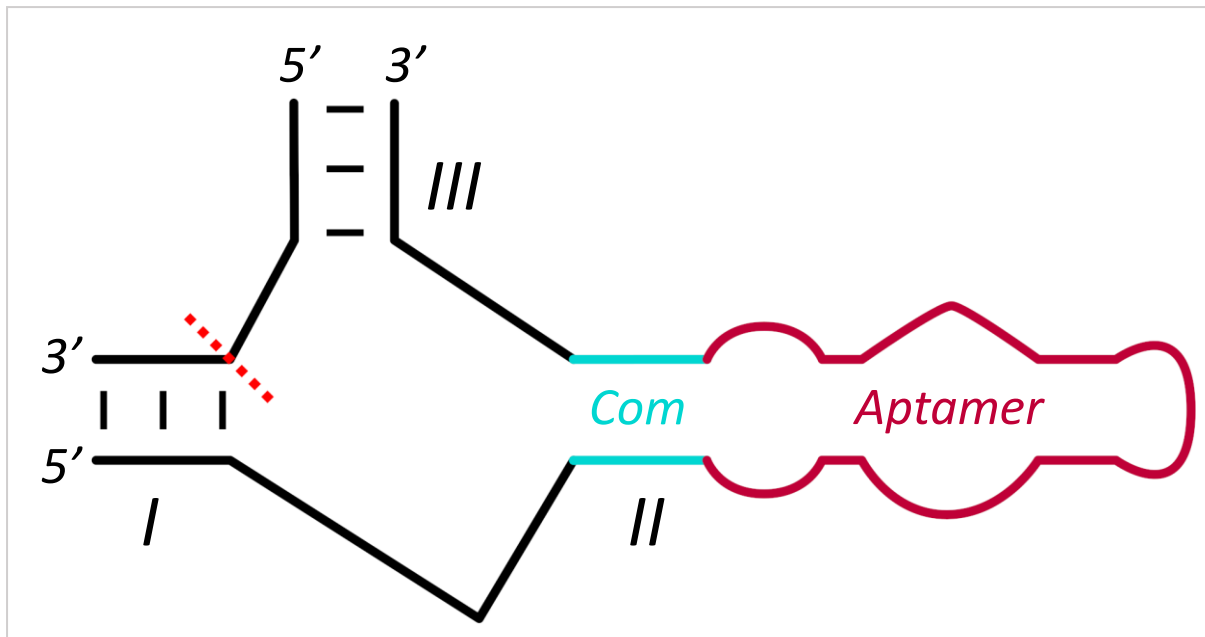


Figure 1-3: A ligand-responsive ribozyme based on the minimal hammerhead ribozyme – Stem II is replaced by a communication domain (Com) fused to an aptamer which binds the ligand of interest. Binding of the ligand to the aptamer domain influences the conformation of the ribozyme, leading to an increase or decrease in the rate of cleavage. The scissile bond is indicated by a red dotted-line while stem numbers are indicated by roman numerals.

### 1.3.3 Hammerhead ribozyme-based riboswitches

Full-length hammerhead ribozymes differ from minimal hammerhead ribozymes such as the one illustrated in Figure 1-3, by the presence of sequences which mediate 3° contacts between loops I and II<sup>23</sup> (Figure 1-4). These 3° contacts are required for cleavage under physiological concentrations of metal ions<sup>23,63</sup>, enabling cleavage *in vivo*. Many of these riboswitches were engineered to respond to theophylline through incorporation of the theophylline aptamer at various loci. However, riboswitches responsive to other small molecules such as thymine pyrophosphate<sup>52</sup>, tetracycline<sup>47,55</sup> and neomycin<sup>49</sup>, in addition to riboswitches responsive to protein motifs<sup>56</sup> have been developed by fusing the relevant aptamers to full-length hammerhead ribozymes.

The cleavage activity of full-length hammerhead ribozymes has been coupled to gene expression using various strategies. To couple cleavage to gene expression in *E. coli*, ribozymes are often incorporated into the 5' UTR of mRNA. This enables regulating of translational initiation<sup>52,53,70</sup> and RNA degradation<sup>71</sup> as exemplified in Figure 1-4. In this case a full-length hammerhead ribozyme has been fused to an aptamer and communication domain via loop I in a similar manner to that conducted previously<sup>47,55,72</sup>. To regulate translational initiation, stem III of the hammerhead ribozyme is replaced with an anti-RBS:RBS stem. This sequesters the RBS from the ribosome, inhibiting translational initiation. Following cleavage, the ribosome can more easily displace the anti-RBS. Potentially, this may involve the action of ribonucleases which can degrade the cleaved anti-RBS from the 2', 3' cyclic phosphate terminus<sup>73</sup>. Secondly, cleavage leads to the formation of a transcript with a 5'-OH terminus instead of a 5' triphosphate (PPP). In contrast to the previously mentioned *glmS* ribozyme present in *B. subtilis*, *E. coli* degrades such transcripts at a slower rate<sup>74</sup>. This ensures that the cleaved HHR persists for longer and as such undergoes translation with higher frequency<sup>71</sup>.

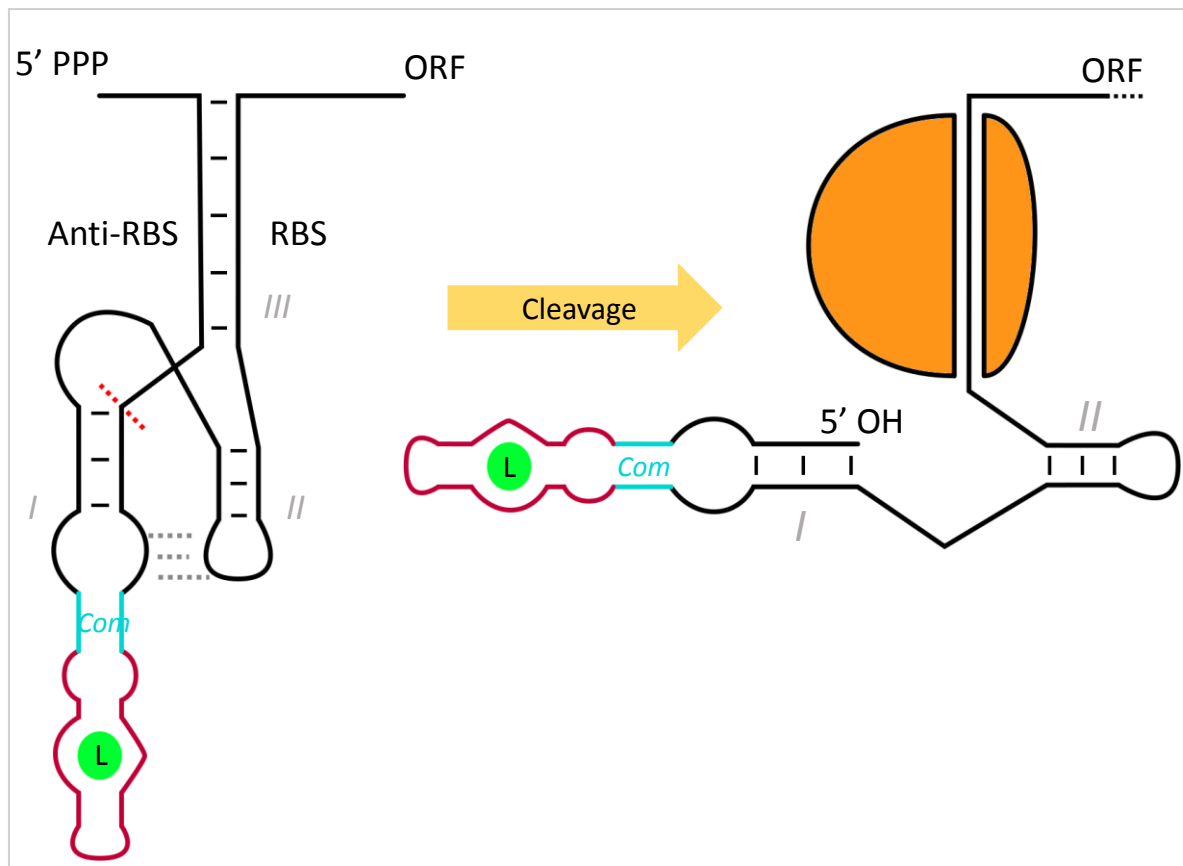


Figure 1-4: A common strategy for implementing synthetic ribozyme-based riboswitches in *E. coli* – Loop I of a full-length hammerhead ribozyme has been fused to an aptamer (maroon) via a communication domain (Com) and inserted into the 5' UTR of a mRNA molecule to generate a synthetic ribozyme-based riboswitch. Ligand (L) binding to the aptamer domain stabilises the active Y-conformation which requires 3° contacts between loops I & II (grey dotted-lines). Cleavage about the scissile bond (red dotted-line) proceeds encouraging displacement of the anti-RBS in stem III, enhancing translational initiation. Cleavage additionally produces a transcript with a 5'-OH instead of a 5' PPP, increasing mRNA half-life.

In addition to functioning as riboswitches in bacteria, different strategies have been used to generate ribozyme-based riboswitches which function in eukaryotes. In this case the ribozyme-based riboswitch is inserted upstream of the 3' polyadenylated tail<sup>46,49–51,54,55</sup> and/or downstream of the 5' cap<sup>47</sup>. In these scenarios, cleavage results in the removal of the polyadenylated tail and/or 5' cap increasing the rate of RNA degradation and hence lowering gene expression.

## 1.4 Applications of ribozyme-based riboswitches

There are several applications of gene regulators responsive to ligand concentrations including those classed as ribozyme-based riboswitches. These are discussed below with an emphasis towards applications previously demonstrated or suggested to be feasible using ribozyme-based riboswitches.

### 1.4.1 Enzyme and metabolic engineering applications

Several ribozyme-based riboswitches are responsive to potential metabolites<sup>47,52,54</sup>. The ability to couple the intracellular concentrations of these molecules to the expression of a gene means these riboswitches are suited to several metabolic and enzyme engineering applications.

One application of these riboswitches is as sensors in ultrahigh-throughput screens or selection experiments conducted to identify functional variants from large libraries<sup>75,76</sup>. In the case of ultrahigh-throughput screens, a reporter gene such as GFP can be placed under the regulation of the riboswitch. In this way the concentration of the metabolite is coupled to a fluorescent signal. A FACS-based screen can then be conducted to sample  $10^7$ - $10^9$  variants from libraries per week<sup>76</sup> and in doing so identify optimal designs from amongst a large number of tested variants. This level of throughput is orders of magnitude larger than what is achieved using traditional approaches e.g. mass spectrometry<sup>75</sup>. As such the probability that optimal designs are identified is increased.

An example of this approach was demonstrated using a theophylline-responsive ribozyme regulating the expression of GFP<sup>77</sup>. Using this sensor, a library of caffeine demethylase mutants generated by error-prone PCR was screened. Following several initial rounds of screening in 96-well plates, a FACS-based screen identified a cell containing an enzyme variant with a further 3-fold increase in theophylline production. Characterisation of the enzyme variant showed it had a lower apparent  $k_M$  and higher product selectivity compared to the wild type, illustrating the potential for such approaches.

In addition to aiding selection and screening experiments, it has been suggested that these ribozyme-based riboswitches could function to dynamically regulate metabolic pathways, improving the production of metabolites<sup>78</sup>. Dynamic regulation is ubiquitous in natural metabolic pathways<sup>79,80</sup>. In contrast to the static control often implemented in synthetic metabolic pathways<sup>81</sup>, dynamically regulated metabolic pathways are regulated according to the environment at any given point in time. To achieve dynamic metabolic pathway regulation, ligand-responsive gene regulators or allosteric enzymes sense metabolites and in doing so, adjust metabolic fluxes through changes in gene expression or turnover rate. Using synthetic biology-based approaches, ligand-responsive gene regulators can be reprogrammed to achieve artificial dynamic control of metabolic pathways<sup>82–86</sup>. This better optimises the use of cellular resources which in turn can improve the yields and titres of commercially relevant compounds.

An example of this approach is the improvement of lysine production in a lysine producing bacterial strain<sup>87</sup>. During this study an activated variant of a natural lysine-inhibited riboswitch was developed. This riboswitch was used to regulate the expression of a lysine transporter protein and as such control transport of lysine to the extracellular medium. The use of this riboswitch in this configuration meant that when the intracellular concentration of lysine was high, lysine was exported to the extracellular medium. Conversely, when the intracellular concentration was low, the opposite scenario occurred. Given that lysine is an essential nutrient, cellular burden issues were likely avoided while lysine export was maximised. The result of this was a 15 % increase in lysine yield<sup>87,88</sup>.

#### 1.4.2 Therapeutic applications

The small size and cis-acting nature of ribozyme-based riboswitches makes them well suited to therapeutic applications as immunogenicity, cross-reactivity and stoichiometry issues are lower compared to other gene regulators<sup>89,90</sup>. These desirable features coupled with the ability to regulate gene expression in response to specific cues has been utilised in several proof-of-concept studies relating to therapeutics<sup>89,91–94</sup>.

For instance, a theophylline-responsive hammerhead ribozyme has previously been incorporated into the 5' and/or 3' UTRs of genes required for the proliferation of oncolytic viruses<sup>94</sup>. Oncolytic viruses are viral vectors that preferentially infect cancer cells and are thus being exploited as cancer therapeutics<sup>95</sup>. The incorporation of a theophylline-responsive riboswitch in this context enabled theophylline-mediated inhibition of viral replication. This approach has several potential applications. One application is as a back-up safety mechanism to prevent non-specific viral infections. Otherwise they may be used to curb the virulence of oncolytic viruses which are currently too toxic for therapeutic applications<sup>96</sup>. A further, potential application is as a tool to inhibit viral activity in carrier cells thereby improving delivery of viral particles to the tumour site<sup>97</sup>.

In addition to responding to external chemical cues, ribozyme-based riboswitches have been engineered to respond to internal cues once transcribed in the host cell<sup>92</sup>. In this example, an aptamer responsive to a hepatitis C viral protein was incorporated into the hammerhead ribozyme. This ribozyme-based riboswitch was then used to regulate the release of a microRNA known to suppress hepatitis C infections. The result of this was that microRNA-mediated gene knockdown was limited in healthy cells. As such, any potential side-effects associated with this microRNA treatment would be reduced.

## 1.5 Current methods for ribozyme-based riboswitch development

The development of ribozyme-based riboswitches has to date been conducted in a two-step process<sup>44–56</sup>. The first of these steps involves the identification of aptamer domains which bind the ligand of interest. Following the identification and characterisation of this domain, a second separate process is conducted to couple the aptamer to a ribozyme and in doing so generate a functional riboswitch. These two processes are discussed below.

## 1.5.1 Identifying a ligand-binding aptamer

### 1.5.1.1 Natural aptamers

One option is to use sequences from natural riboswitches as aptamer domains in synthetic ribozyme-based riboswitches. An example of this approach was highlighted by the generation of a thymine pyrophosphate-responsive ribozyme by Hartig and co-workers<sup>52</sup>. Here, the aptamer domain from a natural thymine pyrophosphate-responsive riboswitch was coupled to a hammerhead ribozyme to generate a synthetic ribozyme-based riboswitch. A potential advantage of this approach is that riboswitches from bacteria could be repurposed to regulate gene expression in eukaryotes as the hammerhead ribozyme can function in both domains. However, the obvious limitation of this approach is that the repertoire of detectable ligands is limited to those recognised by natural riboswitches.

### 1.5.1.2 SELEX-generated aptamers

To overcome this limitation, aptamer domains against bespoke ligands can be generated using SELEX<sup>98,99</sup>. While there are several variations of SELEX<sup>100,101</sup>, the essential process remains the same (Figure 1-5). During this process, an initial pool of up to  $10^{16}$  oligonucleotide sequences is subjected to several rounds involving multiple steps. Briefly, the pool is incubated in the presence of ligand and oligonucleotides which bind the ligand are selected from those that remain unbound. This enables their amplification using RT-PCR or PCR in the case of RNA or DNA, respectively, generating material for subsequent rounds. Once the pool displays a phenotype indicative of ligand-binding, individual sequences can be identified through sequencing and this information used in the generation of functional aptamers.

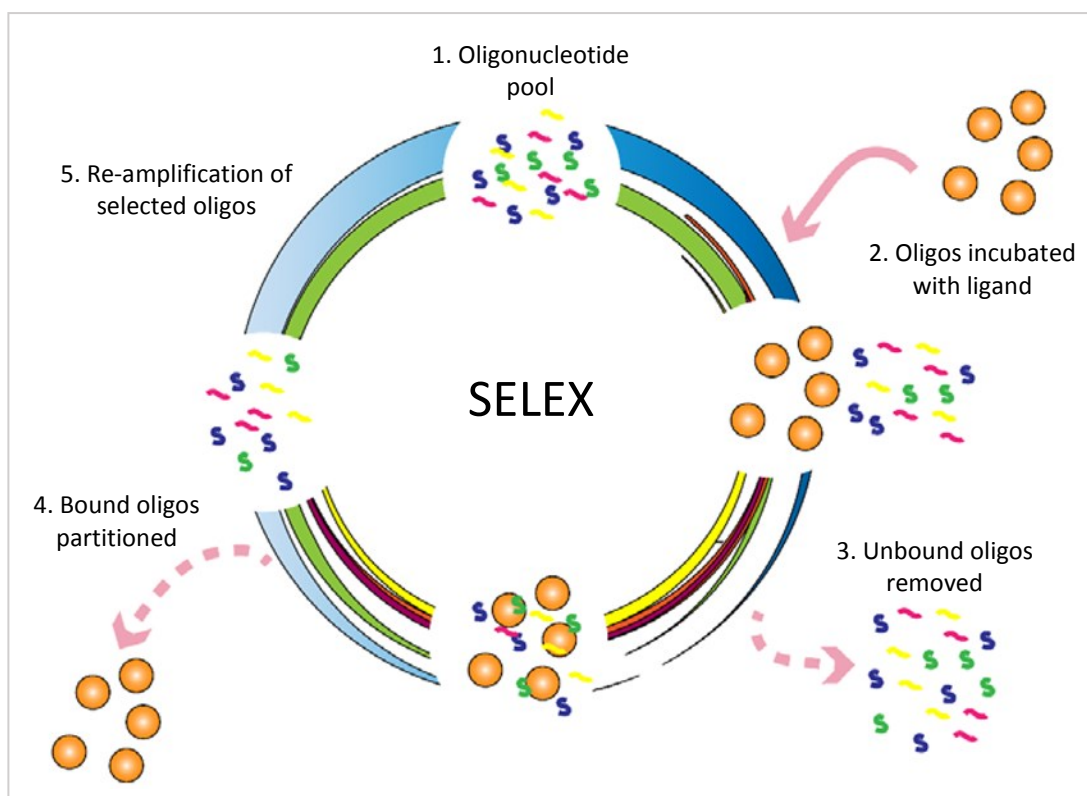


*Figure 1-5: Process of SELEX for aptamer generation – An initial pool of oligonucleotide sequences is generated (step 1). The pool is incubated with the ligand (step 2). Unbound oligonucleotides are removed from the pool (step 3). Bound oligos are then portioned through purification (step 4). The purified oligos are amplified generating a pool for subsequent rounds of selection (step 5). This process is repeated for several cycles until the pool displays a desirable phenotype. Figure adapted from Zhang and co-workers<sup>102</sup>.*

Using this process hundreds of aptamers have been generated against many diverse ligands<sup>103</sup>. Furthermore, the specificity of these aptamers can be high as demonstrated by the theophylline aptamer which effectively discriminates against caffeine, a molecule which differs by a single methyl substitution<sup>104</sup>. These results provide an explanation as to why a great deal of attention has given been given to generating synthetic riboswitches using these aptamers<sup>55,105–110</sup>, as opposed to generating equivalent protein-based gene regulators through the alteration of transcription factor specificity<sup>111</sup>. This later approach is considered more challenging given the complexity of protein-based allosteric mechanisms<sup>112</sup>. However, it should be noted that only a handful of SELEX-generated aptamers have

subsequently been incorporated into ribozyme-based riboswitches, with only three that sense small-molecules ligands<sup>47,49,54,113</sup>.

Two previous studies have shed light on why only a few SELEX-generated aptamers have been used in this context<sup>110,114</sup>. The first of these studies screened a pool of sequences previously generated using SELEX<sup>115</sup> for the ability of each sequence to function as a riboswitch. While several riboswitch-compatible aptamers were identified within this pool, aptamers which could not function as riboswitches were found to be over-represented.

During the subsequent study, two aptamers compatible with riboswitch generation were investigated along with two which were unable to fulfil this function. Analysing these aptamers using NMR, the riboswitch-compatible aptamers underwent significant conformational changes upon ligand binding. In contrast, the two riboswitch-incompatible aptamers were preformed prior to ligand binding, not undergoing conformational changes upon ligand-binding.

These two studies suggest that aptamers compatible with riboswitches can be overlooked by SELEX. This is because other non-compatible aptamers which have similar or improved ligand affinities but are unable to undergo conformational changes are enriched. This limitation with SELEX appears to be a major stumbling block to the identification of compatible aptamers<sup>116,117</sup> and hence the identification of ribozyme-based riboswitches with new ligand specificities.

### 1.5.2 Incorporating the aptamer domain into the ribozyme

Following the identification of a compatible aptamer domain, the aptamer must be coupled to a ribozyme regulating gene expression to generate a ribozyme-based riboswitch such as that illustrated in Figure 1-4. This process usually incorporates an additional communication domain between the ribozyme and aptamer to enhance performance. The exact sequence of the communication domain and resulting ribozyme-based riboswitch has previously been arrived at rationally or with the use of screening or selection methods which are conducted *in vivo* or *in vitro*. These three approaches for

generating ribozyme-based riboswitches by coupling aptamers to ribozymes are discussed in the following sections.

#### 1.5.2.1 Rational methods

The rational incorporation of an aptamer domain is defined by the absence of oligonucleotide libraries which undergo screening or selection. The first ribozyme-based riboswitch was generated in this manner by inserting the theophylline aptamer into a variety of loci present in a group I self-splicing intron<sup>44</sup>. An additional example was illustrated by the insertion of the tetracycline aptamer into a full-length hammerhead ribozyme<sup>47</sup>. During this work, existing structural and biochemical data was used to inform the design of several candidate communication domains, leading to the identification of a functional design.

A more recent example of rational riboswitch engineering involved the development of a thermodynamic model describing riboswitch gene regulation<sup>118</sup>. Having characterised several aptamer sequences in detail, this model was used in conjunction with an algorithm that randomly inserted mutations into the pre- and post-aptamer regions. The model and algorithm were then used to generate several riboswitches not containing ribozyme domains. A theophylline-responsive riboswitch developed by this process could activate gene expression 383-fold. Perhaps a similar approach could be used to generate ribozyme-based riboswitches. This approach could be particularly effective if compatible aptamers are readily available and there continues to be improvements in our understanding of RNA structure and its prediction.

#### 1.5.2.2 In vivo methods

Of the methods discussed for incorporating aptamers into ribozyme-based riboswitches, *in vivo* screening and selection methods have been the most widely used<sup>46,48–54,70</sup>. The first step for both these processes is the generation of a library of sequences containing degenerate residues at key loci such as within the communication domain. Screening of libraries can then be mediated using FACS<sup>46,50</sup> or microtiter plate assays<sup>48,52–54,70</sup> by coupling the activity of each variant to expression of a reporter gene

such as GFP. Otherwise selection of desirable phenotypes can be achieved by coupling riboswitch gene expression to positive and negative selection markers that regulate cell survival<sup>49</sup>. Given libraries are screened or selected under *in vivo* conditions the risk of selected constructs failing to function as characterised is much lower than it is for other methods<sup>72,113</sup>.

However, one limitation associated with the use of *in vivo* methods is the range of ligands that can be used to regulate riboswitch activity. For instance, the ligands used should be non-cytotoxic and able to cross the plasma membrane<sup>117,118</sup>.

### 1.5.2.3 In vitro methods

Compared to *in vivo* methods, *in vitro* methods for ribozyme-based riboswitch development can accommodate a wider spectrum of ligands and enable greater control over ligand concentrations. This is because ligands do not need to cross plasma membranes to influence ribozyme activity during selection. Furthermore, *in vitro* methods can sample a larger number of sequences compared to *in vivo* methods. For instance, ribozyme *in vitro* selection methods sample approximately  $10^{14}$  sequences<sup>72,119</sup>. This is significantly higher than the  $10^4$  -  $10^9$  sequences sampled using *in vivo* methods<sup>49,120</sup>. This ability to sample a larger number of species increases the probability that designs which function under the selection conditions used are identified.

However, one key limitation of *in vitro* methods is the unreliability of selected sequences *in vivo*. While functional ribozyme-based riboswitches have been identified *in vitro*<sup>55,121</sup>, there have been instances where selected *in vitro*-functional ribozymes failed to perform in an *in vivo* environment<sup>55,72,122</sup>. Importantly, the subsequent use of *in vivo* or rational methods for aptamer incorporation successfully incorporated identical aptamer sequences into ribozyme-based riboswitches which functioned in the cellular environment<sup>47,48</sup>. This suggests that current *in vitro* selection methods can enrich for sequences which lack the necessary properties to function as riboswitches.

In addition to the failure of selected sequences to function in an *in vivo* environment, it has been suggested that a relatively large investment of time and resources are required to select ribozymes *in*

*vitro*<sup>121</sup>. This large investment of time and resources partly derives from the requirement to PAGE purify selected sequences multiple times during the process and that each PAGE purification is often implemented with an overnight incubation step<sup>55,72,119</sup>. Specifically, at least 33 PAGE purifications were previously conducted to enrich a tetracycline-responsive ribozyme-based riboswitch to detectable levels<sup>55</sup>, suggesting this endeavour required months of work.

## 1.6 Ribozyme-based riboswitches with new ligand specificities

In the previous section the current process for generating ribozyme-based riboswitches was analysed. While there are a variety of methods to incorporate compatible aptamers into functional ribozyme-based riboswitches, methodologies for the initial identification of compatible aptamers are limited. This has led to a situation where only a handful of ligands are currently detectable by available ribozyme-based riboswitches.

Developing ribozyme-based riboswitches with new ligand specificities could enable additional applications. For example, the development of ribozyme-based riboswitches responsive to other more commercially relevant metabolites, would enable their production to be optimised using metabolic and enzyme engineering methods. Furthermore, the ribozyme-based riboswitches suggested for clinical applications currently require cytotoxic concentrations of ligand for maximum activation<sup>90</sup>. As such, ribozyme-based riboswitches which respond to less toxic ligands might improve the effectiveness of these therapeutics by enabling higher concentrations of ligand to be administered during treatment.

## 1.7 Potential and addressing the limitations of *in vitro* selection methods

In developing ribozyme-based riboswitches with new ligand specificities, four studies suggest the ribozyme *in vitro* selection methods briefly mentioned in Section 1.5.2.3 are of potential use<sup>66,68,69,123</sup>. In three of these studies, ligand-responsive minimal hammerhead ribozymes similar to that illustrated in Figure 1-3, were identified without previously characterised aptamer sequences<sup>66,68,69</sup>. Instead

functional aptamers communicating with the ribozyme domain were identified from randomised regions containing varying numbers of degenerate nucleotides. These studies illustrate that the large numbers of sequences analysed using these methods is sufficient to not only incorporate compatible aptamers but to identify ligand-responsive ribozymes without prior knowledge of the aptamer or communication domains. A further study also used these *in vitro* methods to assay a pool of mutagenized SELEX-generated aptamers, subsequently generating minimal hammerhead ribozymes responsive to caffeine and aspartame<sup>123</sup>. These four studies illustrate that *in vitro* methods for ribozyme selection can improve the ability to identify ligand-responsive ribozymes by either avoiding SELEX altogether or by assaying large numbers of SELEX-generated aptamers in parallel.

However, as previously mentioned there are two key limitations associated with the use of these *in vitro* selection methods for the application of generating ribozyme-based riboswitches. These limitations are the reliability of selected sequences in the *in vivo* environment and the time and resources required in implementing selection.

With regards to the first of these limitations, it was reasoned that given some ribozyme-based riboswitches had previously been developed *in vitro*<sup>55,121</sup>, there is some probability that selected ligand-responsive ribozymes could function as riboswitches. Furthermore, even if selected ligand-responsive ribozymes did not function as riboswitches, subsequent characterisation of these sequences *in vitro* and *in vivo* might reveal properties required to achieve this function. Such information may then be used to update and then iterate selection, ensuring that this time only those sequences containing desirable properties are enriched and subsequently tested.

It was felt that the second limitation associated with ribozyme *in vitro* selection methods could be addressed in this work. Reducing the time and resources required to select ligand-responsive ribozymes *in vitro* would be beneficial for two reasons. Firstly, the method could be improved more rapidly given that less time is required to generate data which can then be acted upon. Secondly, more selection experiments could be conducted given that fewer resources are required per experiment.

As such, larger and more diverse regions of the sequence space could be explored increasing the probability that rare sequences such as those with new ligand specificities are identified<sup>124–126</sup>. Otherwise, more ligands could be investigated for their ability to influence the activity of candidate ribozyme sequences. This is likely to be important given that different ligands have different potentials to bind to RNA and hence influence ribozyme activity<sup>126</sup>. These efforts together should increase the probability that functional sequences are identified.

To reduce the time and resources required for *in vitro* ribozyme selection, a novel method was conceived in the 2<sup>nd</sup> Chapter of this thesis. Using this method 96 independent ribozyme *in vitro* selection experiments can be conducted in parallel, achieving greater economies of scale compared to previous methods. Furthermore, the developed method requires half the time to implement, potentially saving weeks or months per experiment. Given also that all the steps involved have been previously automated using liquid-handling robots, there is scope to further reduce the time and resources required.

To optimise and test this method during the 3<sup>rd</sup> Chapter of this thesis, a control experiment was conducted. This control experiment involved enriching for a derivative of a known theophylline-responsive ribozyme present within a control library. To optimise the selection conditions used to enrich this sequence, a model describing the process of selection was developed. This model enabled the identification of selection conditions which ensured the theophylline-responsive ribozyme was enriched in a practical amount time.

In the 4<sup>th</sup> Chapter, a library of sequences was screened for the presence of ribozymes responsive to the amino acid Tryptophan (Trp). Given to the best of our knowledge a Trp-responsive ribozyme has not previously been characterised, the identification of such a species would aid in achieving the aim of characterising a ribozyme-based riboswitch with new ligand specificities. In addition to conducting this selection experiment, the model of selection developed during the previous chapter was used to determine the feasibility of more challenging selection experiments by estimating the number of

selection cycles required in these cases. Furthermore, this analysis suggested the existence of optimal selection conditions depending on the ribozyme sequence being selected. This latter finding should function to improve future selection experiments.

In the 5<sup>th</sup> Chapter, two aims were addressed. The first of these aims was to analyse the previously conducted control experiment in depth using NGS technology. Following the initial acquisition of an NGS data set describing the control experiment, several individual objectives related to this aim were subsequently set. One of these objectives was to investigate sequences not corresponding with the initial library, with an emphasis towards identifying new theophylline-responsive ribozymes. The second aim of this chapter was to determine the ability of *in vitro*-selected, theophylline-responsive ribozymes to function as riboswitches in *E. coli*. Achieving the latter aim would yield additional information regarding the unreliability of *in vitro*-selected sequences to function as ribozyme-based riboswitches, particularly in this host organism.

## 1.8 Summary

Ribozymes are RNA sequences which catalyse chemical reactions. For example, small nucleolytic ribozymes and group I introns catalyse phosphoryl transfer reactions leading to RNA cleavage and/or ligation. Riboswitches containing these ribozymes have been discovered in nature. The catalytic activity of the ribozyme within these ribozyme-based riboswitches is influenced by the concentration of the ligand they detect. In response to ligand concentrations, downstream gene expression is regulated for example through altered interactions with the ribosome or changes in mRNA stability.

In addition to these natural ribozyme-based riboswitches, synthetic ribozyme-based riboswitches have been developed. The full-length hammerhead ribozyme sequence is often used when developing these constructs due to its small size and well-understood nature. Several studies have suggested that these synthetic constructs could act as therapeutics or fulfil applications in the fields of metabolic and enzyme engineering.

The development of ribozyme-based riboswitches responsive to new ligands could improve their applicability. For instance, the detection of more commercially relevant metabolites would enable the production of these compounds to be optimised using the relevant metabolic and enzyme engineering methods. Otherwise, the detection of less toxic ligands could lead to the development of more effective therapeutics. Unfortunately, the development of ribozyme-based riboswitches with new ligand specificities has proved challenging. It was hypothesised that improvements to the selection of ribozymes *in vitro* would overcome this issue. To this end the overall aims of this thesis were:

1. Develop a ribozyme *in vitro* selection method which is faster to implement and requires fewer resources.
2. Use this method to identify a ribozyme-based riboswitch with new ligand-specificities.

# Chapter 2: Developing a novel ribozyme *in vitro* selection method

## 2.1 Introduction

In the previous chapter, several examples of natural and synthetic ribozyme-based riboswitches were presented. These elements couple the expression of downstream genes to the concentration of ligands through alterations in the rate of catalysis. This ability to detect ligand concentrations *in vivo* means ribozyme-based riboswitches have the potential to fulfil several applications in the clinic and in the fields of enzyme and metabolic engineering<sup>\*1</sup>. The effectiveness and impact of ribozyme-based riboswitches is likely to be improved with the development of other variants responsive to new ligands. However, as reviewed in the previous chapter this have proved challenging.

It was hypothesised that an improved ribozyme *in vitro* selection method which requires less time and fewer resources to implement could alleviate this issue. To develop such a method, it was necessary to understand the factors which currently limit existing methods. Previously, there have been several instances where the direct selection of ribozyme-based riboswitches has been attempted *in vitro*<sup>55,72,127</sup>. The methods employed during these attempts are all roughly based on an *in vitro* selection method originally pioneered by Breaker and co-workers<sup>67,119</sup>. To select and purify nucleic acids using this method, PAGE purification steps are employed on several occasions.

Except for these PAGE purification steps, the remaining steps of this method have previously been implemented in microtiter plate format<sup>128–130</sup>. As such, replacing the PAGE purification steps with alternative microtiter plate-compatible steps would enable the entire method to be conducted within this setting. Such a method was hypothesised to be implementable at a larger scale compared to existing methods and as such would benefit from economies of scale. This is because the majority of methods compatible with microtiter plates can be scaled in at least 96-well format. To the best of our

---

<sup>\*1</sup>Refer to the Introduction chapter, Section 1.4 for a description of these applications.

knowledge commercially available platforms which achieve PAGE purification at this scale do not currently exist. Furthermore, by working in microtiter plate format, components such as multi-channel pipettes and liquid-handling robots can be exploited making the process of working at scale easier.

In addition to generating a methodology which requires less resources, developing a method which avoided PAGE purification was hypothesised to reduce the time required to implement selection. This is because these steps were often implemented with overnight incubation steps meaning they must be conducted on separate days<sup>55,72</sup>.

In designing an *in vitro* selection method not reliant on the PAGE purification of nucleic acids, the general process of selection was first considered. Specifically, the phenotypes of desirable and undesirable sequences were considered along with how previous methods enrich desirable sequences from a pool of sequences displaying different phenotypes. Based on the general requirements of selection, an alternative scheme was conceived which was subsequently implemented using novel selection steps, compatible with microtiter plates.

During the development of this process, ribozyme sequences were required to test the method ensuring it functions as designed. Previously, a library of sequences had been designed with the aim of selecting for ribozymes activated by the amino acid Trp<sup>131</sup> (Figure 2-1A). This library was generated by incorporating the CYA motif (Figure 2-1B) into stem-loop I of a full-length hammerhead ribozyme with the use of several degenerate residues. The CYA motif has previously been identified in 81 % of Trp-binding aptamers characterised in a previous study<sup>132</sup>. This motif was also shown to be the site of Trp binding within these aptamers. As such, ribozyme sequences containing this motif were predicted to have a higher probability to respond to Trp. In addition to containing the CYA motif, stem III of this ribozyme sequence has been replaced with RBS and anti-RBS sequences. It was hypothesised that by including these sequences in the design of this library, selected sequences would have a higher probability of functioning as ribozyme-based riboswitches in *E. coli*; similar to that illustrated in Figure 1-4.

An advantage of working with a library such as this is that to the best of our knowledge a Trp-activated ribozyme or ribozyme-based riboswitch has not yet been characterised. Therefore, the enrichment and identification of such a ribozyme would be impactful and may enable several applications. Potential applications include the optimisation of Trp concentrations in *E. coli* for the production of this metabolite and others that depend on its concentration<sup>105,133</sup>. To better ensure that this library is compatible with the selection method developed during this thesis, it was modified within this chapter.

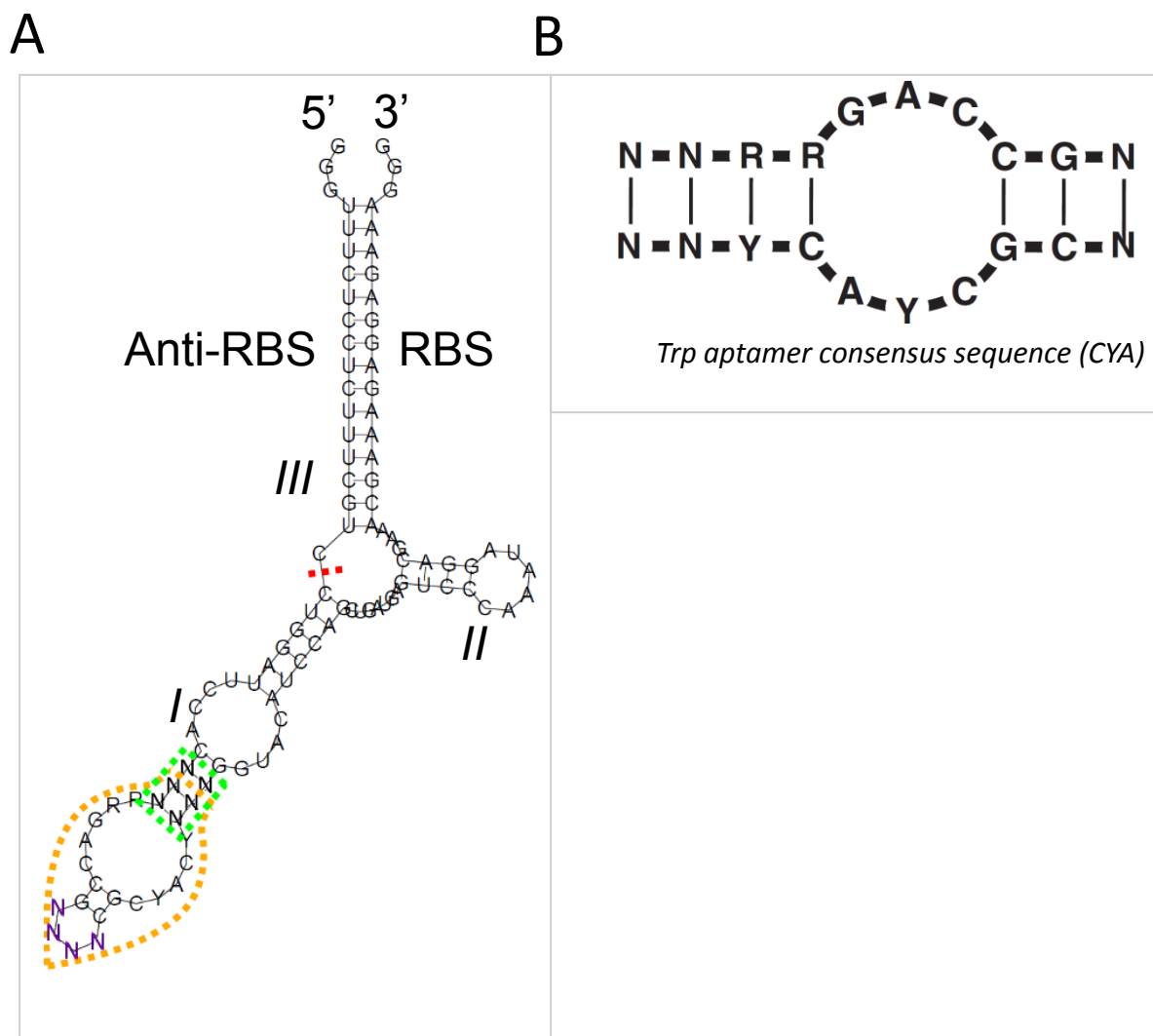


Figure 2-1: A previously conceived library<sup>131</sup>, designed with aim of selecting for hammerhead ribozymes activated by the amino acid Trp – (A) stem III of the hammerhead ribozyme library has been replaced with an anti-RBS/RBS stem such that selected Trp-activated ribozymes may regulate gene expression as illustrated previously in Figure 1-4. The scissile bond is indicated with a red dotted-line while stem/stem-loop numbers are indicated with roman numerals. The CYA element (orange dotted-line) is inserted into stem-loop I of the library with additional degenerate nucleotides (purple and green dotted-line). (B) The sequence of the CYA element, identified as the consensus sequence for a collection of Trp-binding aptamers<sup>132</sup>.

Following the design of a library similar to that illustrated in Figure 2-1A, the novel selection steps required to implement selection were tested for their ability to enrich the expected phenotypes. The

remaining purification steps required to implement this method were then optimised, with the aim of ensuring the method can be easily implemented in microtiter plate format. With the subsequent development of PCR primers, the selection method was partially implemented on the Trp Library designed in this chapter. This enabled the process of selection to be tested and optimised.

## 2.2 Objectives

### *Objective 1:*

Design an effective *in vitro* ribozyme selection method which doesn't require the PAGE purification of nucleic acids and can be implemented in microtiter plates.

### *Objective 2:*

Modify the previously designed Trp Library<sup>131</sup> so that it is compatible with the proposed selection method.

### *Objective 3:*

Test the proposed selection method to determine whether the novel process of selection functions as required.

### *Objective 4:*

Develop and optimise the remaining nucleic acid purification steps, improving the ease at which the method can be scaled.

### *Objective 5:*

Test the functionality of negative selection by conducting multiple rounds of negative selection on the library.

### *Objective 6:*

Test the functionality of positive selection manually by conducting single rounds of positive selection on the product of the previously conducted negative selections.

## 2.3 Results

### 2.3.1 Objective 1 – Selection method design

The first objective in this chapter was to design an *in vitro* selection method to enrich ligand-responsive ribozymes which was faster than current methods and could be scaled in microtiter plates.

#### 2.3.1.1 An alternative scheme for selection

In developing a selection method which fulfilled the above criteria, the phenotypes of several possible sequences likely to be present during selection were considered (Figure 2-2). Two of these phenotypes are ligand-responsive, corresponding with sequences which bind and interact with the ligand. Specifically, ligand-activated and ligand-inhibited sequences have faster and slower cleavage rates in the presence of ligand, producing predominantly more cleaved and full-length transcripts, respectively. The remaining two phenotypes are ligand-unresponsive. Regardless of the ligand concentration, constitutively active and inactive sequences yield predominantly more cleaved or full-length transcripts, respectively.

| Phenotype of sequence | Response -ligand (●)                                                               | Response +ligand (●)                                                                |
|-----------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Ligand-activated      |   |  |
| Ligand-inhibited      |   |  |
| Constitutively active |  |                                                                                     |
| Inactive              |  |                                                                                     |

Figure 2-2: Potential phenotypes of sequences during selection – Four possible phenotypes and their responses in the absence (-) and presence (+) of ligand are given. The red cartoon indicates the majority of transcripts are full-length while the blue cartoon indicates the majority are cleaved.

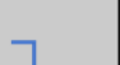
A successful *in vitro* selection method would enrich ligand-responsive ribozymes from a pool of sequences, enabling their subsequent identification. Previous *in vitro* selection methods have achieved this result using positive and negative selection steps. Figure 2-3A illustrates the functionality of these selection steps for enriching ligand-activated ribozymes while discriminating against the three remaining phenotypes. During negative and positive selection, sequences are incubated in the absence and presence of ligand, respectively. Following this incubation, only those sequences which remain full-length or cleave are selected. In conducting both these processes, ligand-inhibited, constitutively active and inactive sequences are selected against, meaning ligand-activated sequences are gradually enriched. By amplifying and iterating negative and positive selection steps, ligand-activated sequences are enriched to a degree where they dominate the pool and can therefore be identified.

Previous *in vitro* selection methods selected full-length or cleaved transcripts with the use of PAGE purification. To avoid these PAGE purification steps, it was found necessary to conduct selection at the cDNA level rather than at the RNA level. Specifically, following *in vitro* transcription of the oligonucleotide pool (step 1 - Figure 2-3B), all transcripts are purified and reverse transcribed (step 2

- Figure 2-3B). Full-length or cleaved cDNA is then selected, enabling negative and positive selection, respectively (step 3 - Figure 2-3B). This discrimination at the cDNA level aimed to enrich ligand-activated ribozymes by ensuring only those cDNA molecules encoding ligand-activated sequences are selected. Following selection of the desired cDNA molecules, DNA templates are regenerated and amplified creating material for subsequent rounds (step 4 - Figure 2-3B).

Figure 2-3B illustrates that a consequence of selecting at the cDNA level is that positive and negative selections are decoupled. Specifically, each time the pool is transcribed, it is subjected to either positive or negative selection but not both. This contrasts with previous *in vitro* selection methods where both positive and negative selection are conducted prior to regeneration and amplification of the DNA template. To acknowledge this, individual positive or negative selections are termed a “round of positive/negative selection”. The two together are referred to as a “cycle of selection”.

A

| Phenotype             | -ve                                                                               | +ve (●)                                                                           |
|-----------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Ligand-activated      |  |  |
| Ligand-inhibited      |  |  |
| Constitutively active |  |  |
| Inactive              |  |  |

B

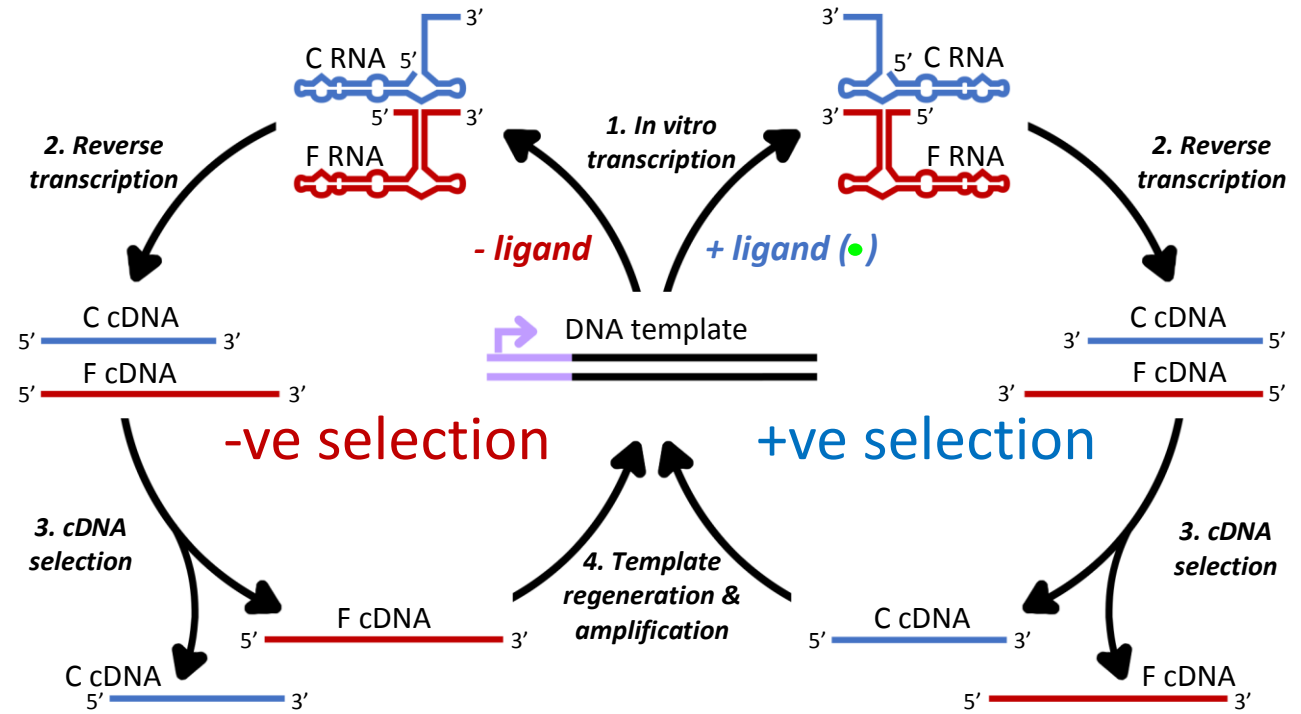


Figure 2-3: Principle and an alternative scheme for enriching ligand-activated ribozymes – (A) Table illustrating how positive & negative selection together can enrich for ligand-activated ribozymes, while discriminating against ligand inhibited, inactive and constitutively active sequences. (B) General scheme for selection. During positive and negative selection, the library is transcribed with and without ligand, respectively, yielding a mixture of cleaved (C) & full-length (F) RNA (step 1). RNA is purified and reverse transcribed yielding C & F cDNA (step 2). Either C or F cDNA is selected depending on whether positive or negative selection is conducted (step 3). Using this selected cDNA, the DNA template is regenerated and amplified yielding material for the next round of selection.

### 2.3.1.2 Bait-oligo mediated selection and template regeneration

To implement the selection method in Figure 2-3B, a procedure for selecting full-length and cleaved cDNA is required (step 3 - Figure 2-3B). In line with the objective of this chapter this method must be compatible with microtiter plates and as such cannot involve PAGE purification. To achieve this, a method using a bait-oligo was conceived (Figure 2-4).

As described in the previous section, to enrich ligand-activated ribozymes full-length cDNA must be selected during negative selection. To achieve this, cDNA generated during selection is first annealed to a bait-oligo which selectively hybridises with full-length cDNA (step 1 - Figure 2-4). This selective hybridisation is achieved from the fact cleaved cDNA is missing a region at the 3' end of the molecule compared to the full-length cDNA. This region has been cleaved off at the RNA-level during ribozyme catalysis. Following hybridisation to full-length cDNA, the full-length cDNA/bait-oligo complex is immobilised on streptavidin-coated paramagnetic beads (step -B – Figure 2-4). The immobilisation of this complex is facilitated by the presence of a 5' biotin-modification on the bait-oligo. Having immobilised the complex, it can be pulled down enabling the purification and selection of full-length cDNA.

In the case of positive selection, cleaved cDNA must be selected. Retaining the supernatant from the negative selection pull-down results in the removal of some full-length cDNA. However, this would not be 100 % effective. This is because oligo hybridisation reactions exist in a chemical equilibrium meaning a fraction of full-length cDNA will be unbound by the bait-oligo<sup>25</sup>. To remove sufficient quantities of full-length cDNA from the reaction, multiple pull-downs are conducted. This is achieved by adding additional bait to the supernatant of each pull-down and repeating the process (repeating step A & +B – Figure 2-4). In this way full-length cDNA is diluted to the desired concentration, purifying and selecting for cleaved cDNA.

Following selection of the appropriate cDNA molecule, the DNA template must be regenerated and amplified for further rounds of selection (step 4 - Figure 2-3B). To regenerate the DNA template from

full-length cDNA during negative selection, overhang PCR is used. Specifically, this involves the use of a sense primer which contains the missing T7 promoter region (steps -3 & 4 - Figure 2-4). To prevent the formation of truncated PCR products during amplification, the bait-oligo used during the pull-down step contains a 3' blocking group. This prevents its extension by the DNA polymerase used during PCR.

To regenerate the DNA template from cleaved cDNA during positive selection, a longer sequence must be incorporated into this molecule. This sequence includes the region cleaved off during RNA catalysis and the T7 promoter. It was suspected that incorporating this sequence using overhang PCR would generate non-specific products. Such sequences are undesirable as they effectively dilute the DNA library. As such, regeneration of the DNA template is achieved in this instance using a ligation reaction. Specifically, the cleaved cDNA molecule is first annealed to a splint in the presence of an adapter encoding the missing sequences. Subsequent ligation using a DNA ligase generates a species which contains both the cleaved-off region and T7 promoter sequence (step +3 - Figure 2-4). This species can then be amplified during PCR to generate material for subsequent rounds of selection (step 4 - Figure 2-4).

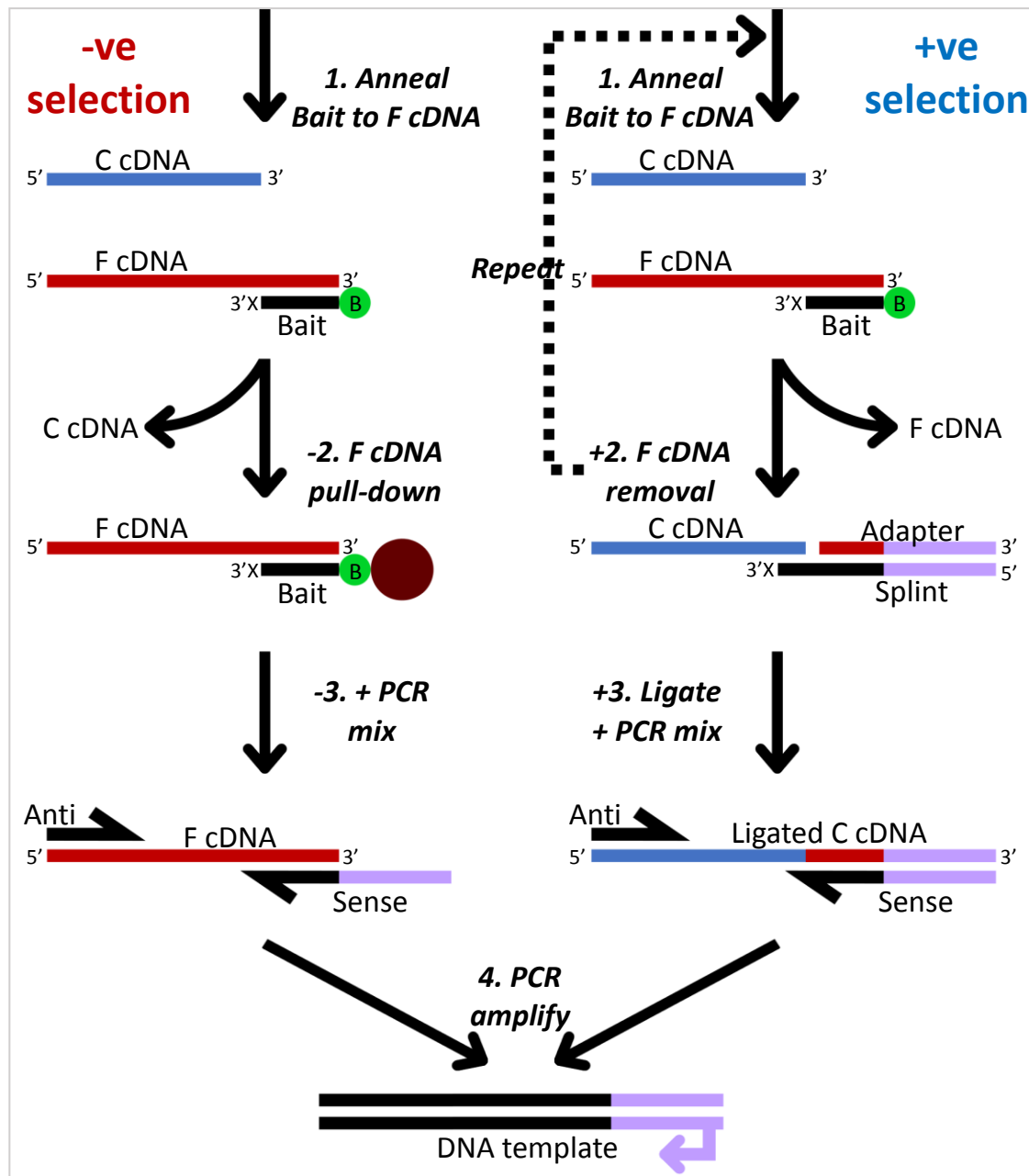


Figure 2-4: Bait-oligo mediated selection of cDNA – For both negative & positive selection, full-length cDNA is annealed to a 3' blocked (X), 5' biotinylated bait-oligo (step 1). During negative selection, bait-oligo/full-length cDNA complex is immobilised on streptavidin-coated paramagnetic beads, enabling a pull-down reaction (step -2). The product is PCR amplified to regenerate & amplify the DNA template (step -3 & 4). During positive selection, the pull-down supernatant is retained (step +2). Steps 1 and +2 are repeated diluting full-length cDNA. The remaining cleaved cDNA is ligated to an Adapter using a Splint (step +3). This regenerates the DNA template enabling PCR amplification (step 4).

### 2.3.2 Objective 2 – Previous Trp ribozyme library re-design

Having outlined the selection method, ribozymes sequences were required to test the method ensuring it was functional. To this end, the previously designed Trp Library illustrated in Figure 2-1A was adopted. However, it was hypothesised that this library would not be compatible with the conceived selection method. To remedy this, the library was modified.

While modifying the previously designed Trp Library, the following requirements were formulated to help encourage the selection of a Trp-activated ribozyme which could function as a riboswitch:

- I. Transcription of the DNA encoding the library should be efficient and specific to full-length and cleaved RNA limiting the production of non-specific transcripts.
- II. Reverse transcription of full-length & cleaved RNA should be efficient, generating sufficient quantities of full-length & cleaved cDNA.
- III. The bait-oligo should selectively and efficiently anneal to full-length cDNA so that the selection method outlined in Figure 2-4 can be implemented.
- IV. Variants in the library should contain all the nucleotides required for cleavage under physiological conditions. Furthermore, without considering degenerate residues, the 2° structure of the library should mimic that of the consensus hammerhead ribozyme sequence (Figure 1-2A). This should maximise the chance of selecting for ribozymes which have wild-type cleavage rates in the presence of ligand.

Based on Requirement IV and the desire to identify a Trp-activated ribozyme, it was reasoned that only stem III and sequences distal to this stem should be considered when modifying the previously designed Trp Library. This is because modifying sequences in stem-loops I and II or within the core of the ribozyme run the risk that all variants are unable to cleave under physiological conditions. The following sections optimise these regions according the above requirements.

### 2.3.2.1 Stem III

As per Requirement III, the bait-oligo should selectively anneal to the full-length cDNA. This requires that the bait-oligo anneals to the reverse transcribed anti-RBS. This portion is cleaved off during ribozyme catalysis and is located at the 3' end of full-length cDNA (Figure 2-5).

Figure 2-5 illustrates that the efficiency of bait-oligo selection is dependent on the design of stem III. This is because bait-oligo hybridisation is inhibited if stem III is a stable element in full-length cDNA. The reason for this is because under these circumstances bait-oligo annealing is in direct competition with stem III hybridisation. Given the complementary stem III strands are part of the same molecule, the effective concentration of stem III will always be larger than the concentration of bait-oligo, making the  $\Delta G$  for bait-oligo annealing less favourable<sup>134</sup>.

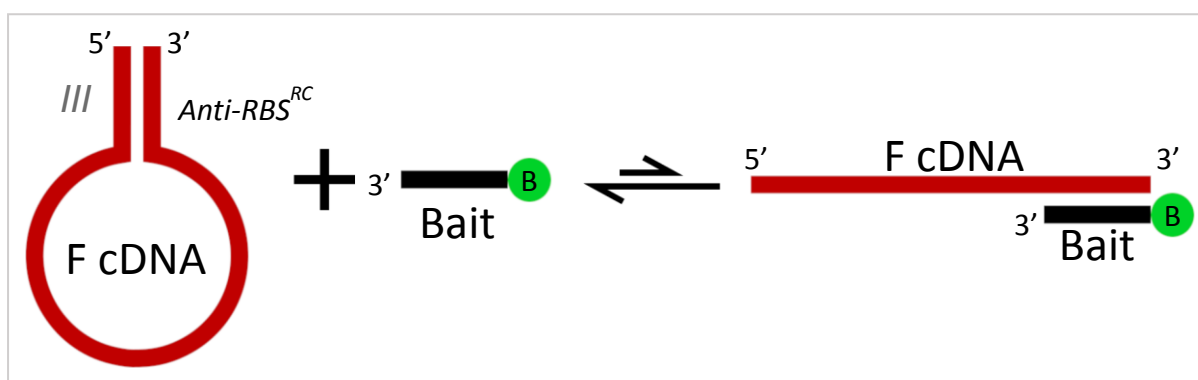


*Figure 2-5: bait-oligo hybridisation inhibition by stem III formation – During bait-oligo hybridisation with the full-length cDNA, Stem III must be displaced for the bait-oligo to anneal to the reverse complement of the anti-RBS (anti-RBS<sup>RC</sup>). In the above scheme stem III is too stable and hence the equilibrium lies to the left of the reaction.*

Therefore, the bait-oligo interaction must involve more nucleotides than those present in stem III. To achieve this, two optimisations are conducted. The first is to shorten stem III as much as possible. The second optimisation involves designing a single stranded region outside of stem III that anneals to the bait. This second optimisation is covered in the subsequent section.

To shorten stem III, fewer anti-RBS nucleotides can be used in the design of the library. This ensures the library contains the RBS sequence and other nucleotides in stem III required for catalysis or important for structural stability. Intuitively, shortening stem III beyond a certain point will destabilise the ribozyme 2° structure. This will significantly reduce the cleavage rates for most of the library members, contradicting Requirement IV above.

To determine the optimal length of the anti-RBS a computational approach was taken. By using a computational approach many candidate designs could be investigated with little effort before specific designs were settled upon and subsequently assayed *in vitro* for their ability to conform to the above requirements. The Vienna RNA-fold software, used in this case calculates MFE 2° structures, base-pairing probabilities, and information about the ensemble of structures for a given sequence<sup>135</sup>. It was hypothesised that a correctly folded MFE structure which was more frequent in the ensemble would indicate a library design which better adheres to Requirement IV. This is because a library design with this property should on average contain more sequences which adopt the consensus hammerhead ribozyme 2° structure more frequently.

With this rational the % frequency of the MFE structure in the ensemble for varying numbers of anti-RBS nucleotides was calculated (Figure 2-6A). As expected when there are too few anti-RBS nucleotides in the design, the stability of the MFE structure drops significantly. Comparing the MFE 2° structures of library designs containing 1 and 7 anti-RBS nucleotides (Figure 2-6B & D), this loss of stability is caused by a reduction in the propensity for the ribozyme stems to base pair. Surprisingly, MFE stability is lower when there is > 9 anti-RBS nucleotides. Comparing the MFE 2° structures of library designs containing 7 and 12 anti-RBS nucleotides (Figure 2-6D & C), this effect appears to be a consequence of U-A base pairs in the distal region of stem III which fray. This creates multiple structures where these bases are paired or unpaired, effectively diluting the MFE in the ensemble. As such, 7 anti-RBS nucleotides were chosen for the library design (Figure 2-6D). This produced the most stable MFE structure and shortened stem III.

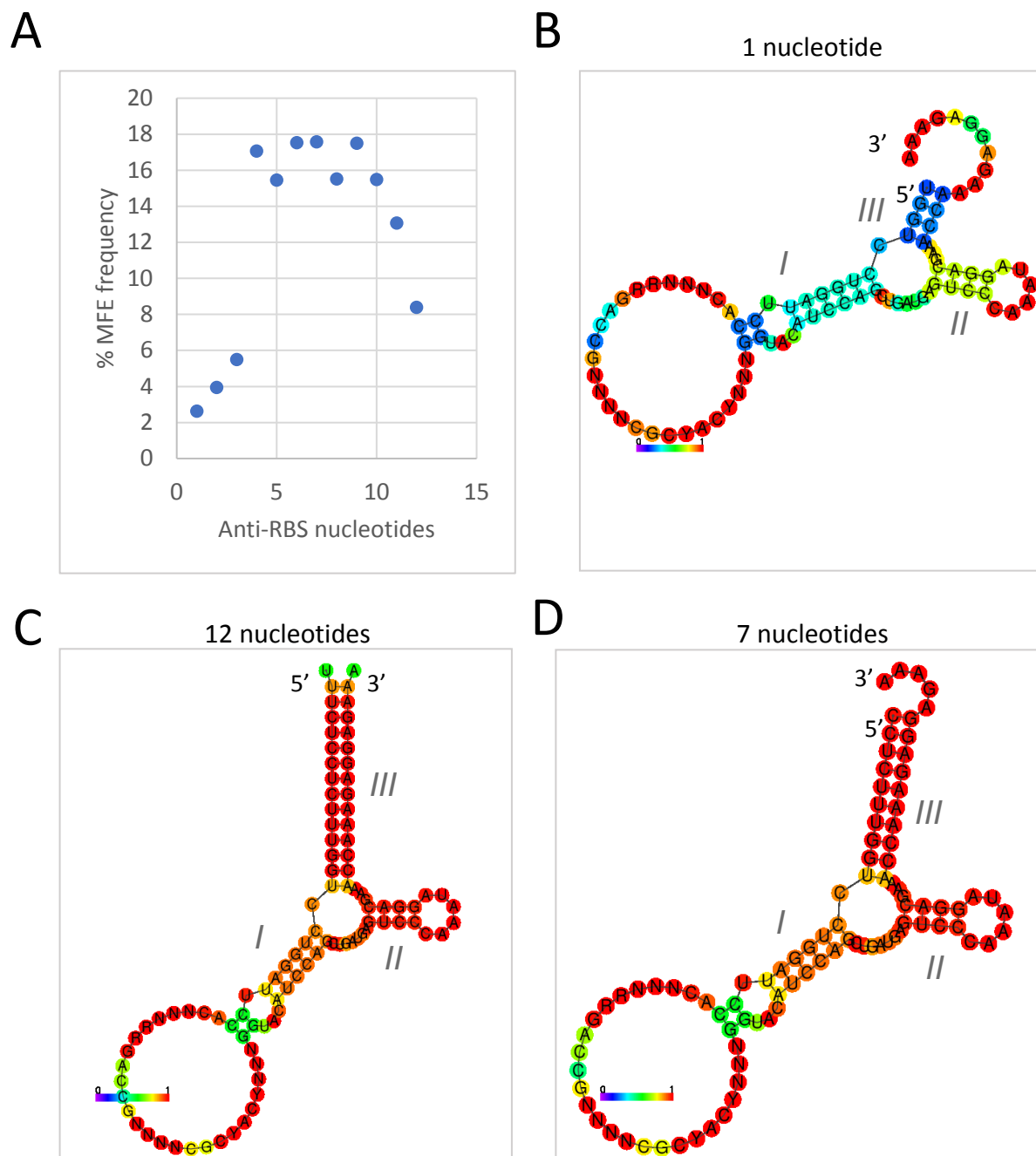


Figure 2-6: Effect of removing anti-RBS nucleotides on library stability – (A) The % frequency of the MFE structure in the ensemble as a function of the number of anti-RBS nucleotides. (B – D) MFE structures for libraries containing 1, 12 & 7 anti-RBS nucleotides, respectively. Roman numerals denote stem numbers, while colours indicate base-pairing probabilities, according to the given scale.

### 2.3.2.2 Bait-oligo toehold

As covered in the previous section, the interaction between bait-oligo & full-length cDNA must include more nucleotides than stem III for this interaction to be stable. To achieve this, a short single-stranded region, also known as a toehold sequence was introduced into the library design, upstream of stem III. It was hypothesised that a good toehold sequence should shift the equilibrium of the reaction illustrated in Figure 2-5 to the right, aiding Requirement III.

The length of the toehold was first considered. From previous studies a toehold should be at least 5 nucleotides in length<sup>136</sup>. Furthermore, strand displacement kinetics tend to saturate for toeholds which are greater than 7 nucleotides. Therefore, it was reasoned that the toehold sequence should be between 5 and 7 bps.

To generate a general sequence for the toehold, the functionality of the residues in the toehold were first considered. For instance, the 5' nucleotide of the toehold is also the +1 position of the transcription product. Varying the +1 position from a G to another nucleotide can drastically reduce the yield of RNA transcripts. For instance, a C or A residue at this position will result in a 10 or 5 fold reduction in RNA transcripts, respectively<sup>137</sup>. Guanine and uracil nucleotides in the remaining positions of the toehold were however avoided. This is because these nucleotides can interact with each other and their Watson-Crick complements<sup>25</sup>. This would encourage non-specific interactions, destabilising library variants which would impact Requirement IV above. To help maintain the interaction between the bait and the full-length cDNA once it had bound, a GC base pairing was introduced in the 3' position of the toehold. With these considerations the general sequence for the toehold is:

$$5' - G(M)^n C$$

where  $3 \leq n \leq 5$ , meaning there are 56 sequences which satisfy the above general sequence.

To identify an exact sequence, the thermodynamic properties & GC content of the toehold were considered. According to the previous study, for maximal strand displacement kinetics,  $\Delta G <$

$-8 \text{ kcal/mol}^{136}$ . Only 30 of the 56 sequences had this property. The GC content of these 30 sequences ranged from 57 to 100 %. Using sequences with a high GC content would encourage the bait to non-specifically anneal to variants with a similarly high GC content in their variable regions<sup>138</sup>. As such our choice of toehold sequence was limited to the 10 sequences with a GC content of 57 %. The first of these sequences contained a cluster of cytosine residues (Appendix, Section 9.2.1). To more evenly distribute these residues the 2<sup>nd</sup> identified sequence was selected.

To guard against the selected toehold sequence destabilising the other constant regions of the ribozyme library, the MFE structure (Figure 2-7) was calculated. The MFE structure in Figure 2-7 illustrates that the inclusion of the toehold does not inhibit the formation of stems I, II or III. Furthermore, comparing the structures in Figure 2-6D and Figure 2-7, the base-pairing probabilities are similar suggesting that the presence of this toehold is well tolerated.

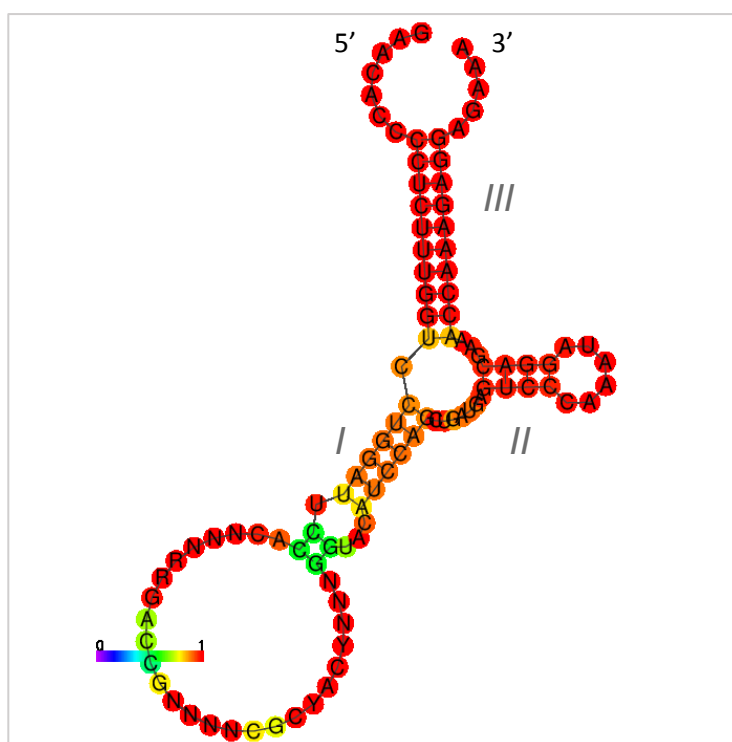


Figure 2-7: MFE structure of the ribozyme library containing the shortened stem III and toehold sequences – Roman numerals denote stem numbers, while colours indicate base-pairing probabilities, according to the given scale.

### 2.3.2.3 RT primer landing pad

According to Requirement II, the library design should enable efficient reverse transcription of full-length and cleaved cDNA. Without any adjustments to the current library design (Figure 2-7), the RT primer would need to be 5 nucleotides in length to ensure this. If the RT primer was longer it would need to anneal within stem III. This would reduce the RT efficiency for the full-length cDNA<sup>139</sup>. It was hypothesised that using a 5 nucleotide RT primer would reduce the specificity of the reverse transcription reaction. This is because there is the potential for the primer to anneal to other regions of the molecule, yielding truncated products in an analogous manner to those produced when random hexamers are used as RT primers<sup>140</sup>. As such the single-stranded region downstream of stem III was extended to enable a longer RT primer to be used without the displacement of this stem. This region is referred to as the “RT primer landing pad”.

In determining a general sequence for the RT primer landing pad, G & U nucleotides were similarly avoided as was reasoned for the previous toehold design. To avoid the production of non-specific transcripts, the two terminal 3' residues were assigned cytosine identities. It has been documented that the T7 RNAP is capable of extending RNA transcripts if the 3' end of the transcript has homology to itself or other nucleotides in solution<sup>141,142</sup>. The use of two terminal cytosine residues as a mechanism to combat this was based on data from previous designs (data not shown) and that the fact cytosine is the most discriminative residue<sup>143</sup>. With these considerations the general sequence for the RT primer landing pad was:

$$AGAAA(M)^nCC - 3'$$

Given there is no upper limit for n, there are an infinite number of sequences which conform to the above general sequence. To limit the number sequences which were evaluated, the effect of the RT primer T<sub>m</sub> was considered. If the T<sub>m</sub> is too high, there is a propensity for the primer to miss-anneal<sup>140</sup>. Furthermore, an RT primer with a high T<sub>m</sub> increases the likelihood that the 3' landing pad will disrupt the library. This is because the propensity for this region to intramolecularly hybridise with non-variable ribozyme domains increases. As mentioned for small RT primers, using an RT primer which has a T<sub>m</sub> which is too low would reduce the specificity of the RT reaction. As such, only sequences which had melting temperatures similar to the 42 °C RT reaction temperature under the reaction conditions were considered. Of the potential RT primers, only 14 fulfilled the criteria of 42 °C ≤ T<sub>m</sub> ≤ 47 °C (Appendix, Section 9.2.2).

Any chosen 3' landing pad should not disrupt the library constant regions in line with Requirement IV. As such, the effect of the potential landing pads on MFE frequency was calculated (Figure 2-8A). Figure 2-8A shows that for each library construction, the frequency of the MFE structure in the ensemble is similar, differing by at most up to ~0.4 %. However, sequence 14 produces a library with the highest value. Sequence 5 was also noted as yielding a library with a particularly frequent MFE structure given its length. Figure 2-8B & C illustrate that these landing pads do not disrupt the formation of stems I, II

and III. Furthermore, the base pairing probabilities in these sequences are similar to the sequences which lacks these elements (Figure 2-7). This suggests they are well tolerated.

A further 3' landing pad sequence, sequence 2 was also considered. While this sequence yielded a library with a less frequent MFE structure in its ensemble (Figure 2-8A), its sequence composition was different to sequences 5 and 14 which are similar. As such had these two library designs failed, it was reasoned sequence 2 would be likely to give a different, potentially beneficial response.

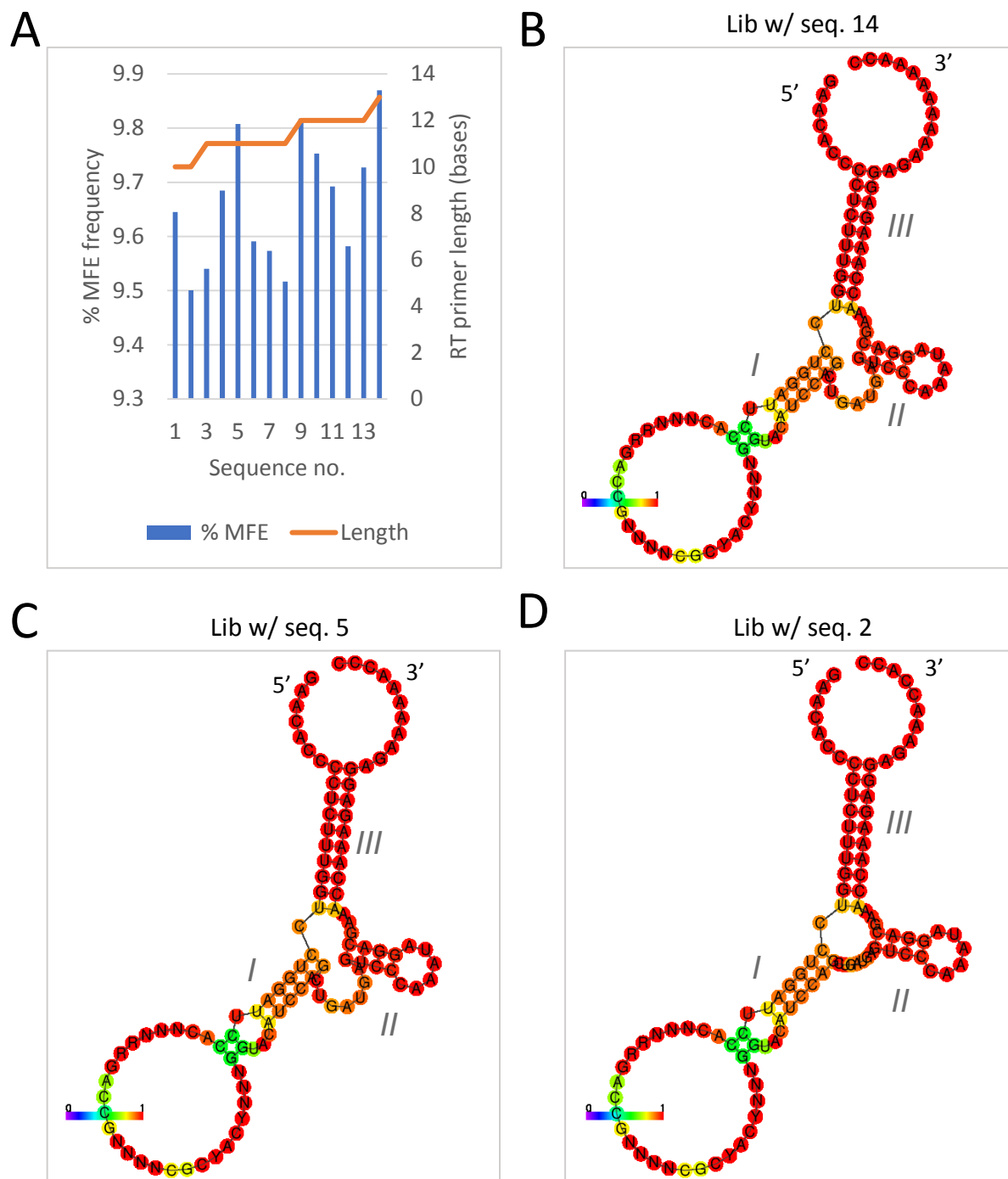


Figure 2-8: RT primer landing pads – (A) The % frequency of the MFE structures in the ensemble for the 14 potential RT primer landing pads (blue bars). The length of each RT primer is given by the orange line. (B – D) MFE structures for libraries containing selected RT primer landing pads. Roman numerals denote stem numbers, while colours indicate base-pairing probabilities, according to the given scale.

#### 2.3.2.4 Selected library designs

Having identified three potential RT primers, the resulting libraries were ordered, assembled and *in vitro* transcribed to determine whether the desired full-length and cleaved RNA species would be produced. Figure 2-9 illustrates the purified RNA product from each IVT reaction in addition to the assembled DNA templates. From Figure 2-9, all the DNA templates are correctly assembled as characterised by bands of the appropriate size. For instance, the DNA template corresponding with a library containing the 14<sup>th</sup> RT primer landing pad was expected to be 128 bps. While all the DNA templates appear to be correctly assembled, libraries using sequences 2 and 5 produce more non-specific IVT products compared to the library using sequence 14. For instance, the library containing sequence 5 produces a clear non-specific product at ~100 bp. As such the library containing sequence 14 was selected. This library is henceforth referred to as the Trp Library and this design forms the basis for all libraries assayed throughout this thesis.

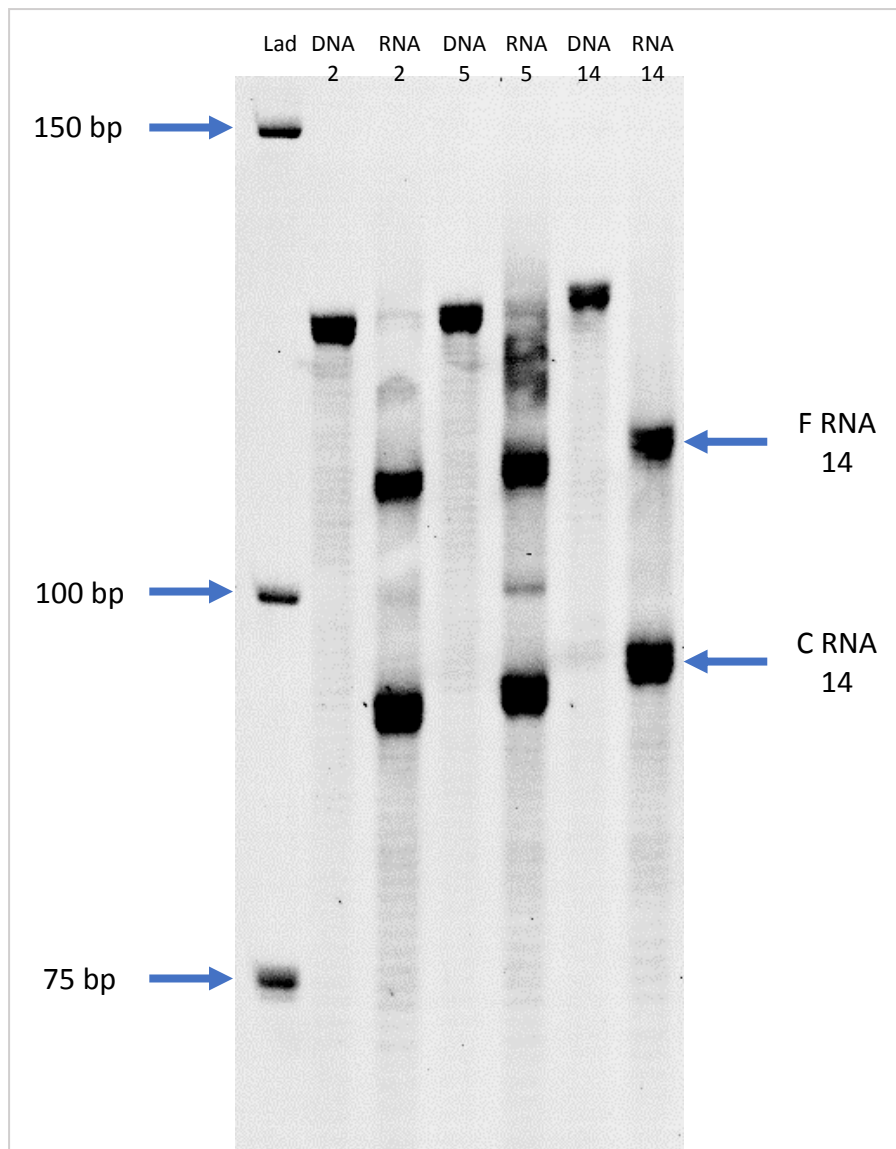


*Figure 2-9: Selected library design – Urea PAGE gel with lanes containing: 2  $\mu$ L NEB LMW ladder (Lad), 0.5  $\mu$ g DNA and 1  $\mu$ g purified RNA for libraries containing RT primer landing pad sequences 2, 5 & 14. full-length (F) and cleaved (C) RNA is indicated for sequence 14.*

### 2.3.3 Objective 3 – Testing the novel cDNA selection steps

Having designed a library, potentially compatible with the selection method, the novel bait-oligo mediated cDNA selection steps were tested for their ability to function.

### 2.3.3.1 Full-length cDNA selection

To better visualise changes in cDNA quantities caused by the above selection methods, the RT primer was labelled with a HEX fluorophore. The HEX-labelled RT primer was shown to have a lower detection limit than that achieved using EtBr staining under the conditions tested (data not shown). Furthermore, the intensity of bands should directly correlate with the molarity of cDNA present, given each RNA molecule should be reverse transcribed by a single RT primer. One limitation of this method however is that the quantity of cDNA cannot be directly compared to the quantity of non-HEX labelled nucleic acids such as RNA or DNA. Furthermore, the presence of the HEX group is likely to affect the migration rate of cDNA due to the additional MW.

To test the functionality of bait-oligo negative selection, full-length cDNA was purified using the methodology illustrated in Figure 2-4 and described in the Materials & methods chapter (Section 7.4.1). The DNA template, RNA, cDNA and pull-down samples were subsequently analysed using urea PAGE (Figure 2-10). From Figure 2-10, the expected bands corresponding with cleaved and full-length cDNA molecules are present following analysis of the purified reverse transcription reaction. Conducting a bait-oligo pull-down on the purified cDNA leads to the disappearance of cleaved cDNA, meaning only the full-length cDNA is present. This suggests the bait-oligo pull-down functions as required.

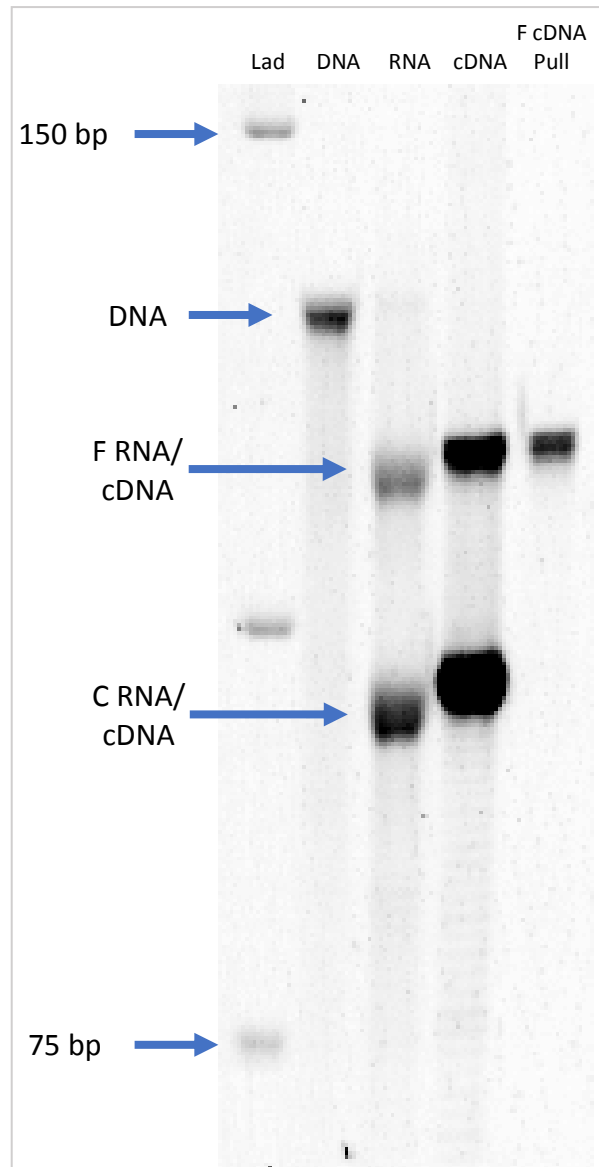


Figure 2-10: cDNA preparation & bait-oligo full-length cDNA purification – (A) Urea PAGE gel with Lanes containing: 4  $\mu$ L NEB LMW ladder (Lad), 0.6  $\mu$ g DNA, 4  $\mu$ g purified RNA & purified cDNA from the RT reaction in addition to the product of the bait-oligo pull-down (F cDNA pull). Full-length (F) and cleaved (C) cDNA/RNA are indicated. Gel was stained with EtBr and imaged for both EtBr and HEX-mediated cDNA fluorescence.

### 2.3.3.2 Cleaved cDNA selection & template regeneration

During the proposed bait-oligo positive selection, cleaved cDNA is first purified by removing sufficient quantities of full-length cDNA. To achieve this, successive bait-oligo pull-downs were conducted and

the supernatant retained as depicted in Figure 2-4 and described in the Materials & methods chapter (Section 7.4.2). Figure 2-11 illustrates this purification step was successful as characterised by the disappearance of the full-length cDNA band from the cDNA sample.

Following cleaved cDNA purification, an Adapter is ligated to cleaved cDNA to re-generate the DNA template. The ligation reaction product should therefore be the same length as the DNA template. Figure 2-11 illustrates the appearance of a band slightly larger than the DNA template following ligation. The additional MW of the ligation product compared to the DNA template can be explained by the presence of a 5' HEX group which is introduced by the RT primer. As such it was concluded that the key steps of bait-oligo positive selection are functional.

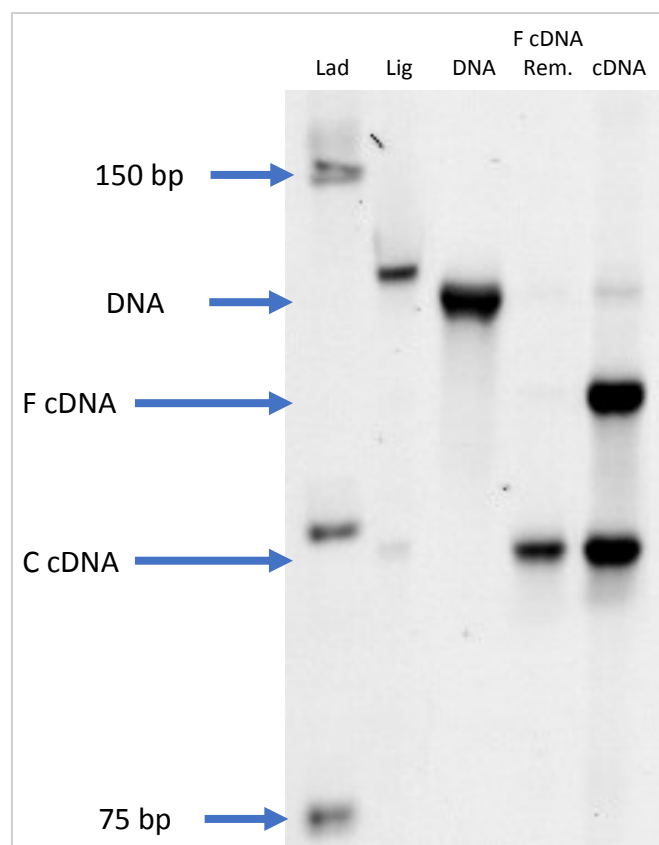


*Figure 2-11: bait-oligo mediated cleaved cDNA purification and ligation - Urea PAGE gel with Lanes containing: 4  $\mu$ L NEB LMW ladder (Lad), Ligated cleaved cDNA (Lig), 0.8  $\mu$ g DNA, cDNA after full-length cDNA removal (F cDNA rem.) and cDNA from the RT reaction. Full-length (F) and cleaved (C) cDNA are indicated. The Gel was stained with EtBr and imaged for both EtBr and HEX-mediated cDNA fluorescence.*

#### 2.3.4 Objective 4 – Non-specific purification of nucleic acids

As explained in the following sub-sections, there are several instances during selection where all the DNA, RNA or cDNA requires purification and concentration. In developing this selection process, purification of total cDNA, RNA and DNA was attempted using SPRI beads<sup>144</sup>. SPRI beads are carboxylated paramagnetic beads. These beads interact with nucleic acids in buffers of high ionic strength containing molecular crowding agents. SPRI bead technology is well suited to protocols which function in microtiter plates<sup>145</sup> as minimal manipulation is required compared to other techniques such as ethanol precipitation or silica spin-columns. The latter had been used up until this point to

implement the required purification steps (refer to the Materials & methods chapter, Sections 7.2.3 & 7.2.5).

#### 2.3.4.1 Optimisation of SPRI for cDNA purification

As part of the bait-oligo selection method, purification of free cDNA in solution is required at least once during each round of selection. Specifically, purification is required following the RT reaction to enable the complete hydrolysis of RNA. This is achieved using RNase A, whose activity on double-stranded RNA and RNA in a DNA hybrid would otherwise be effected under the conditions of the RT reaction<sup>138</sup>. Without complete full-length RNA digestion, annealing the bait-oligo to the full-length cDNA would be inhibited, reducing the effectiveness of selection.

The purification of cDNA and most nucleic acids is feasible using SPRI technology<sup>144,146</sup>. However, the majority of commercially available SPRI kits are designed for single-stranded DNA oligonucleotides larger than 200 bases. The Trp Library full-length and cleaved cDNA molecules which together must be purified using this method are 111 and 93 bases, respectively. As such, using commercially available SPRI kits would lead to a significant loss of material. Fortunately oligonucleotides as small as 50 bp can be purified using SPRI by varying the [PEG] and [NaCl]<sup>145</sup>.

To achieve optimal yields of full-length and cleaved cDNA using this technique, the [PEG] and [NaCl] used in the SPRI buffer was optimised (Figure 2-12). Figure 2-12A suggests a higher [PEG] improves the yield of cDNA recovered using this method. This effect is characterised by an increase in the % yield from 55 to 72 % when using 39 compared 30 % PEG. This result is in line with previous reports which concluded that using a higher [PEG] increases SPRI efficiency for smaller oligonucleotides<sup>145</sup>. Furthermore, Figure 2-12B suggests that increasing the [NaCl] above 1 M reduces the yield directly or indirectly by restricting the volume of PEG. Overall, more than 70 % of cDNA purified using spin-columns could be recovered using SPRI purification under 1 M NaCl and 39 % w/v PEG (Figure 2-12A). Given the advantages of SPRI over spin-columns in terms of ease of use and scaling, this purification strategy was adopted.

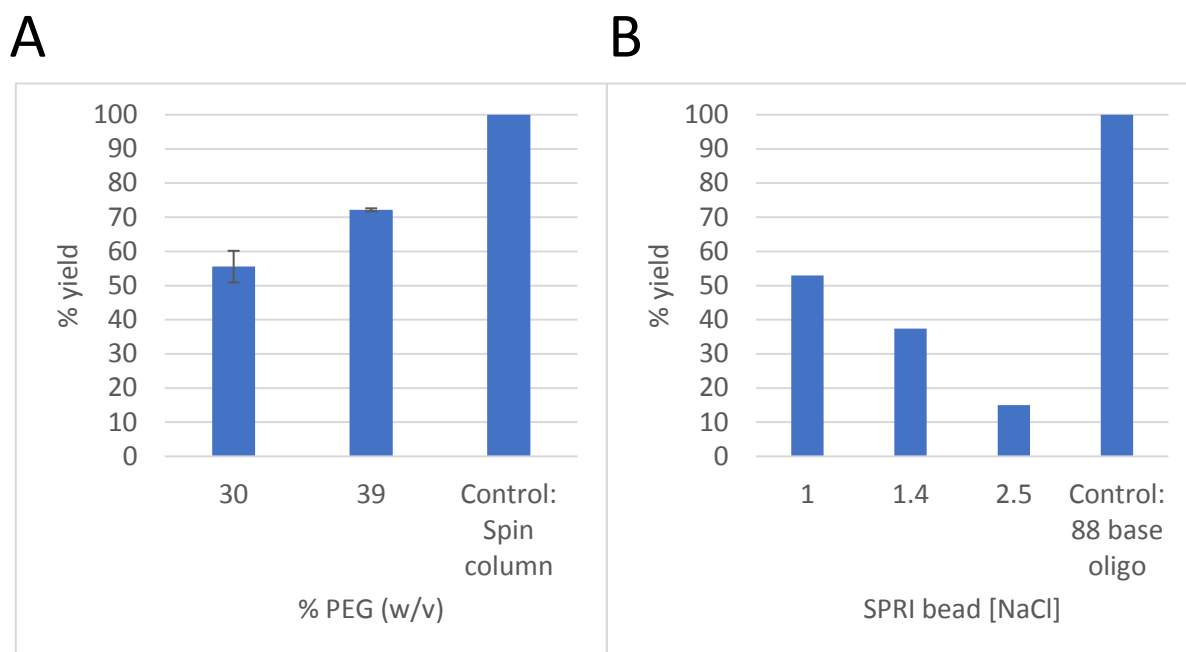


*Figure 2-12: Optimisation of NaCl and PEG for the purification of cDNA – All yields were quantified using densitometry (A) Trp Library cDNA was purified with spin-columns or SPRI under 30 or 39 % PEG. The % yield of full-length & cleaved cDNA compared to the spin-column is calculated. Error bars represent standard deviations calculated from 3 repeats. (B) 1 µg of an 88 base PAGE-purified synthetic oligonucleotide was purified using SPRI. Buffers contained 2.5, 1.4 or 1 M NaCl. Buffers containing 1.4 or 1 M NaCl contained 30 % w/v PEG while buffer containing 2.5 M NaCl contained the maximum 24 % PEG.*

#### 2.3.4.2 Effect of IVT volume on SPRI-based RNA purification

Given selection occurs at the cDNA level, achieving high yields of cDNA is important for maximum throughput. Achieving a high yield of cDNA from the RT reaction requires the purification and concentration of the RNA. Specifically, according to the manufacturer's instructions, purified RNA at a concentration > 100 ng/µL is required to ensure optimal RT yields.

Having identified optimal conditions for the purification of cDNA using SPRI, RNA was purified in a similar manner as described in the Materials & methods chapter (Section 7.2.3.1). To increase the chance of yielding the required titre of 100 ng/µL, the volume of IVT used was optimised. By using

larger volumes, it was hypothesised that the titre of purified RNA could be increased as more material would be captured and eluted in the same volume. Figure 2-13 illustrates that the required 100 ng/ $\mu$ L RNA titre was achieved using SPRI. However, surprisingly increasing the volume of IVT from 23 to 30 or 36  $\mu$ L reduces the titre from 275 to 114 or 87 ng/ $\mu$ L. This result was contradictory with the expectations of the above statement.

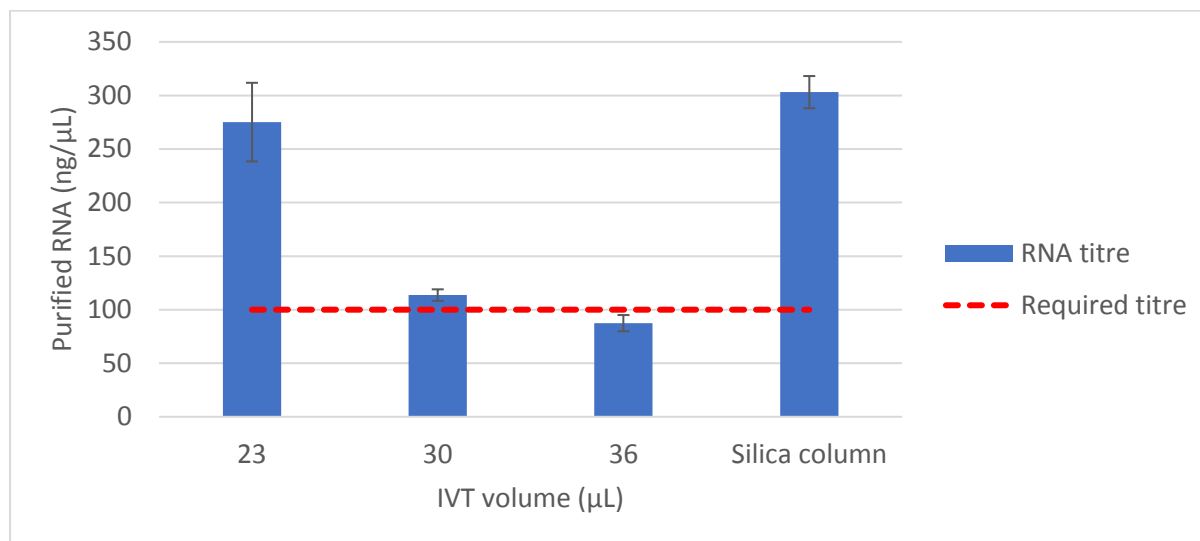


*Figure 2-13: Effect of volume on the efficiency of RNA purification using SPRI – IVT volumes between 23 and 36  $\mu$ L were digested with DNase I and purified using either SPRI or silica spin-columns. The red line indicates the titre of RNA required to maximise cDNA yield according to the manufacturers instructions. Error bars represent standard deviations calculated from 3 repeats*

#### 2.3.4.3 Purification of DNA from PCRs

At the end of each round, the selected cDNA is amplified to provide material for the next round of selection and to determine the response of the library in the presence and absence of ligand. For each standard 20  $\mu$ L transcription reaction, 0.5  $\mu$ g DNA is required (Materials & methods, Section 7.2.2). As such a minimum of 1.5  $\mu$ g DNA is required following each round of selection, enabling the 3 transcription reactions required.

To attempt to generate this material, PCRs were conducted and subsequently purified using either silica spin-columns or using SPRI beads as described in the Materials & methods chapter (Section

7.4.3). In contrast to the purification of cDNA and RNA where size selection is not required, it was reasoned that DNA should be purified from remaining PCR primers. This would ensure consistent selection conditions<sup>119</sup>. As such the [PEG] was not increased when attempting to purify the DNA template as this could encourage the co-purification of the smaller PCR primers. Figure 2-14 illustrates that under these conditions the required yield of 1.5 µg was not achieved using commercially-available SPRI beads. In contrast, this yield was surpassed using silica spin-columns. As such, silica spin-columns and not SPRI was adapted for the purification of PCRs for the remainder of this thesis. This decision is reviewed in Section 2.4.3.

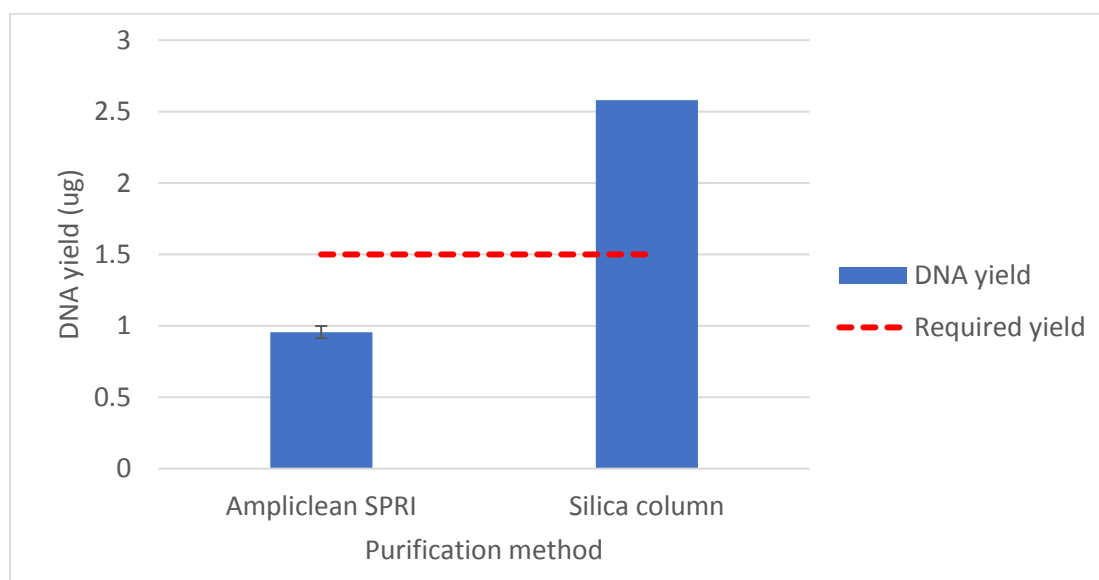


*Figure 2-14: DNA purification using SPRI vs silica spin-columns – PCRs were purified using Ampliclean SPRI beads or silica spin-columns according to the manufactures instructions and the absolute yield of DNA quantified. The red line indicates the required yield of DNA. Where present, error bars represent standard deviations calculated from 3 repeats.*

### 2.3.5 Objective 5 – PCR primer optimisation and negative selection effectiveness

The purpose of *Objective 5* is to ensure that the entire round of negative selection is functional. Prior to checking the functionality of negative selection, PCR primers were designed to facilitate complete

rounds. Following their design, two sets of PCR primers were tested in parallel to determine which would give rise to the most effective negative selection.

#### 2.3.5.1 PCR primer sets used to implement negative selections

In developing the PCR step required to implement multiple negative selections, 3 different sets of PCR primers were considered.

The 1<sup>st</sup> set of PCR primers uses a sense primer analogous to that previously used by Breaker and co-workers<sup>72</sup>. These researchers conducted selection on a similar library design. Figure 2-15 illustrates that the sense primer (Sense1) is the maximum possible length as a longer sense primer would involve annealing to degenerate residues within the library. The corresponding anti-sense primer (Anti1) was designed to have a similar T<sub>m</sub>.

A 2<sup>nd</sup> set of PCR primers was designed to enable variability in stem-loop II. This was achieved by having a shorter anti-sense primer (Anti2 - Figure 2-15). Using this anti-sense primer, residues in stem-loop II are free to vary without any reduction in PCR efficiency or with mutagenesis upon primer annealing. It was hypothesised that variability in stem-loop II would be beneficial for several reasons. The first is that residues in stem-loop II could evolve potentially improving the performance of selected ribozymes as has previously been demonstrated<sup>46</sup>. Secondly it may be beneficial to insert ligand sensing domains into stem-loop II rather than stem-loop I in future library designs (refer to discussion).

A final 3<sup>rd</sup> set of PCR primers was designed on the realisation that the 3' region of library is A-rich. As such it was hypothesised that an anti-sense primer containing a 3' A residue would prevent mis-annealing (Anti3 - Figure 2-15). To complete this set of primers, a sense primer was designed to have a similar T<sub>m</sub> (Sense3).

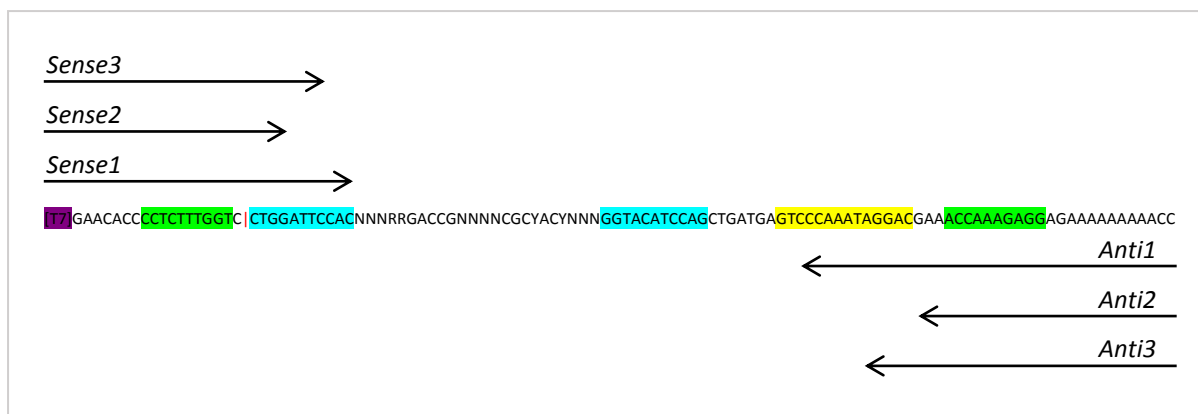


*Figure 2-15: Tested PCR primer sets – The sequence of the Trp Library investigated during this thesis. Regions are colour-coded for clarity: Purple = T7 promoter; green = stem III; red line = scissile bond; light blue = stem-loop I; yellow = stem-loop II. Arrows indicate the length of the sense and anti-sense primers used for each set.*

The three PCR primer sets were assayed for their ability to amplify the initial library without the generation of non-specific products<sup>\*2</sup>. Non-specific products dilute the original library and complicate the analysis of the library at the RNA level through the production of additional bands. As illustrated by Figure 2-16, using the 2<sup>nd</sup> set of primers leads to the generation of a larger non-specific band and as such only primer sets 1 & 3 were used to test multiple negative selections.

<sup>\*2</sup>Refer to the protocol given in the Materials & methods chapter, Section 7.4.4.

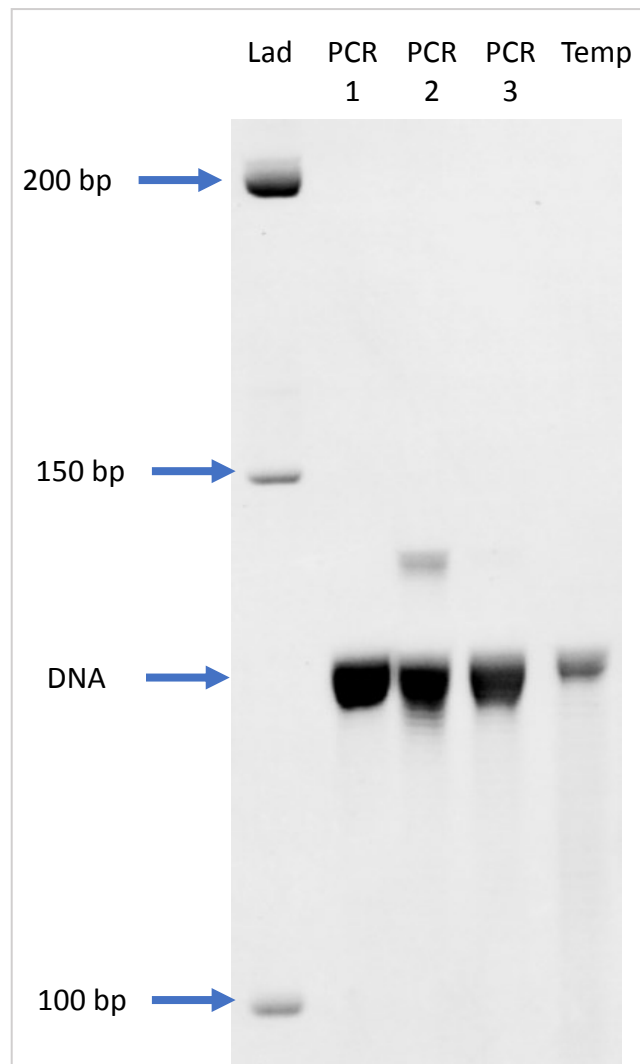


*Figure 2-16: Products of PCRs using either primer sets 1, 2 or 3 - Urea PAGE gel with lanes containing: NEB LMW ladder, products of PCRs using either primer sets 1, 2 or 3 and the PCR template (Temp); equivalent to Trp Library DNA. The desired size of the PCR product is indicated (DNA).*

### 2.3.5.2 Multiple bait-oligo negative selections

The effectiveness of the bait-oligo negative selection was tested by implementing multiple, negative selections using either primer sets 1 or 3. As previously illustrated in Figure 2-3A, if negative selection is functional, there should be an enrichment of ligand-activated and inactive phenotypes. The enrichment of these phenotypes will be characterised by a reduction in cleaved and an increase in full-length RNA transcripts.

In determining if sequences with the expected phenotype could be enriched by negative selection, the Trp library was subjected to multiple rounds of negative selection. This was achieved via transcription and purification prior to reverse transcription. The resulting cDNA was purified and full-length cDNA purified using bait-oligo selection (Materials & methods, Section 7.4.1). Selected cDNA was PCR amplified to regenerate DNA templates for further rounds of negative selection. To measure changes caused by selection, the DNA templates generated following each round were transcribed and the RNA transcripts generated, purified and separated via electrophoresis. Following electrophoresis, % cleaved values were calculated as described in Section 7.1.4 of the Materials & methods chapter. The enrichment of the expected phenotype would be characterised by a decrease in the % cleaved value, as this is a hallmark of fewer cleaved and more full-length transcripts, respectively. As illustrated by Figure 2-17A both sets of primers produce this response suggesting that negative selection is functional.

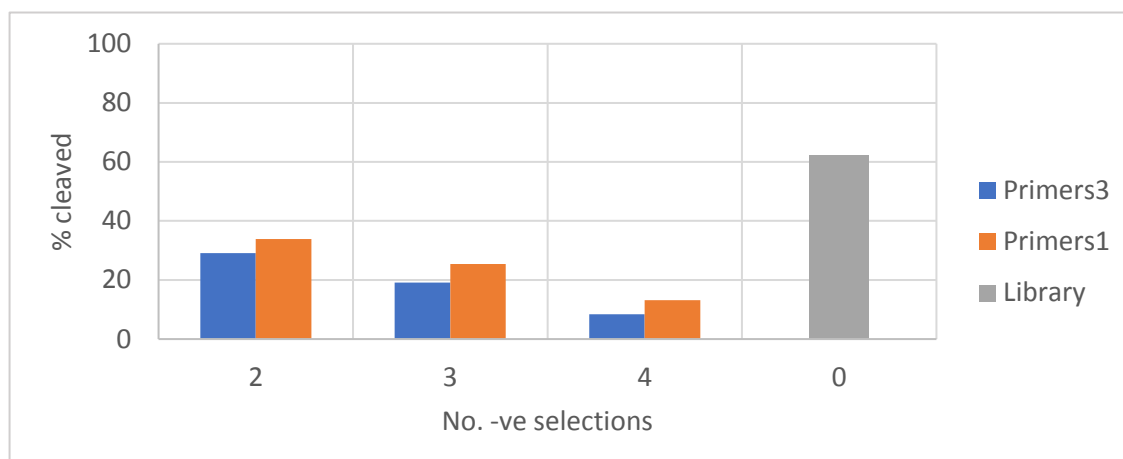
Figure 2-17A also illustrates that when primer set 3 is used the magnitude of this reduction is greater, indicating improved negative selection effectiveness. Specifically, the % cleaved value after 2, 3 & 4 consecutive rounds of negative selection was calculated at 33, 25 & 13 % for primer set 1 and 29, 19 & 8. % for primer set 3, respectively. It is suspected that this effect is caused by the higher propensity for sense primer 1 to amplify any remaining cleaved cDNA. This higher propensity is likely a result of sense primer 1 being longer<sup>\*3</sup>, meaning there is a greater region of complementarity to any cleaved cDNA that escapes purification.

The overall effectiveness of bait-oligo negative selection using primer set 3 is further illustrated in Figure 2-17B as following 4 or 5 consecutive negative selections, cleaved RNA is almost undetectable.

---

<sup>\*3</sup>Refer to Figure 2-15.

A



B

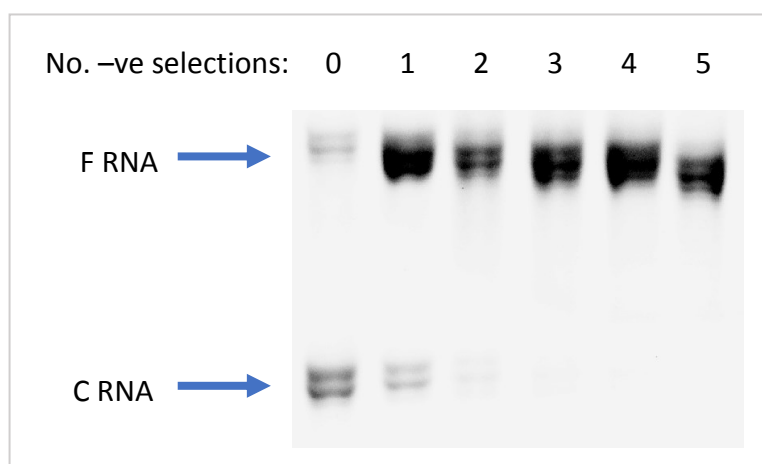


Figure 2-17: Effectiveness of bait-oligo mediated negative selection – (A) Variable numbers of consecutive negative selections were conducted using either primer sets 1 or 3. The resulting DNA templates were transcribed and % cleaved values for the RNA generated calculated. (B) Following each negative (-ve) selection, DNA templates were transcribed and 1  $\mu$ g purified RNA analysed. All negative selections used primer set 3. full-length (F) & cleaved (C) RNA are indicated.

## 2.3.6 Objective 6 – Effectiveness of positive selection and further optimisation

### 2.3.6.1 Bait-oligo mediated positive selection

To measure the effectiveness of positive selection, a single round was conducted using DNA templates generated previously after 0, 1, 2 or 5 rounds of consecutive negative selection (Figure 2-18). If

positive selection is functional, there should be an enrichment of ligand-activated and constitutively active phenotypes according to Figure 2-3A. The enrichment of these phenotypes will be characterised by a reduction in full-length and an increase in cleaved RNA transcripts. Similar to negative selection, this change in phenotype can be quantified by calculating % cleaved values following selection. In contrast with negative selection, an effective positive selection should result in a measurable increase in the % cleaved value.

The data in Figure 2-18 suggests the overall effectiveness of bait-oligo mediated positive selection depends on the number of previous negative selections. For instance, conducting a round of positive selection following 2 or 5 rounds of negative selection results in significant increase in % cleaved values. Specifically, the % cleaved value increases from 19.0 & 1.8 % to 35.7 & 14.3 %, respectively. This suggests that in these cases the expected phenotype was enriched. However, when positive selection is conducted following 1 round of negative selection or on the initial library (0 negative selections), the magnitude of the change in is not as significant. Indeed, when positive selection is conducted on the initial library, a reduction in the % cleaved value from 52.8 to 43.5 % is observed. This suggests that the expected phenotype was not enriched.

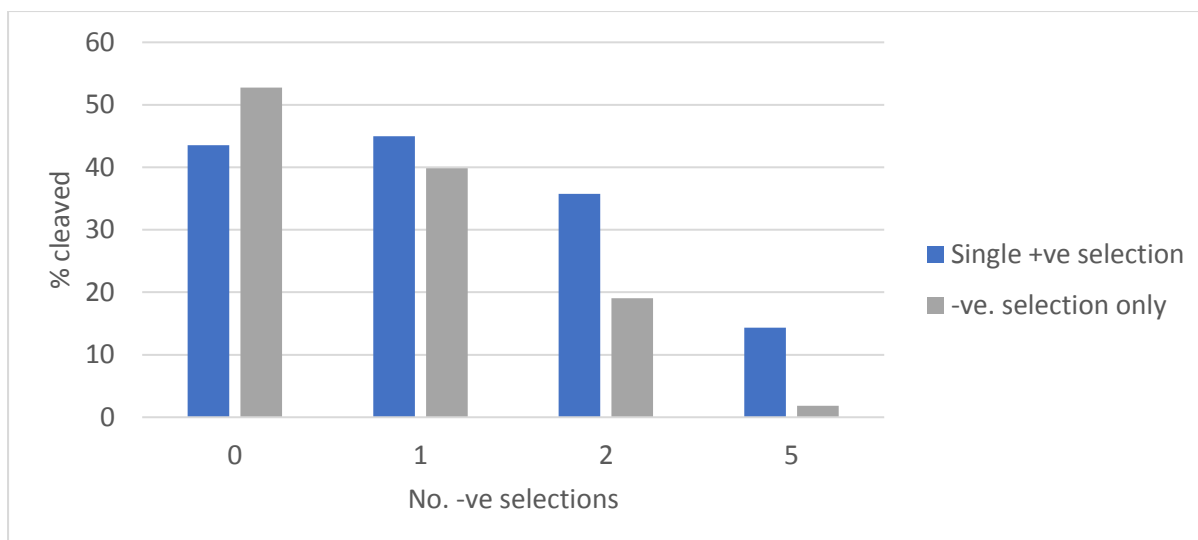


*Figure 2-18: Effectiveness of bait-oligo mediated positive selection – positive selections were conducted after 0, 1, 2 or 5 consecutive negative selections. % cleaved values of the purified RNA transcribed after selection (blue) is shown along with values of the purified RNA transcribed before selection (grey).*

#### 2.3.6.2 Ligase mediated selection of cDNA

To attempt to improve the effectiveness of positive selection, another method for selecting cDNA was conceived. This method relies on the discrimination of DNA ligases to select cleaved or full-length cDNA and hence is referred to as ligase-mediated selection (Figure 2-19). Comparing this method to bait-oligo mediated selection (Figure 2-4), cleaved cDNA molecules are directly selected rather than being enriched through the removal of full-length cDNA molecules. Given the removal of full-length cDNA molecules will not be 100 % efficient<sup>25</sup>, it was hypothesised that ligase mediated positive selection would be more effective than that achieved using the equivalent bait-oligo method.

The first step of the DNA ligase-based method involves selectively ligating the desired cDNA molecule to an adapter. To achieve this both cDNA molecules are annealed to a Splint in the presence of an adapter (step 1 - Figure 2-19). Different adapters are used depending on the cDNA molecule being purified. Specifically, Adapter- & Adapter+ are used for full-length & cleaved cDNA purification, respectively. Both adapters are 5'-phosphorylated enabling ligation to the 3'-OH group of the relevant

cDNA molecule. During positive selection, the Splint/Adapter+ duplex anneals to full-length cDNA forming a 3' flap. This configuration is not recognised by the T7 DNA Ligase<sup>147</sup>, which was the ligase used to implement selection. As such, only cleaved cDNA is ligated. During negative selection, the Splint/Adapter- duplex anneals to cleaved cDNA molecules creating a gap. In an analogous manner to the flap, this substrate is not ligated by the T7 DNA ligase<sup>147</sup> and as such only full-length cDNA is ligated.

Following the ligation reaction, the resulting ligation product is purified. To facilitate this both adapters are 3' biotinylated enabling the immobilisation of the ligation product on streptavidin-coated paramagnetic beads. After immobilisation, the product is washed with NaOH to remove sequences which interact with the product or solid support non-specifically (step 2 - Figure 2-19).

One desirable feature of this protocol is that the DNA template is regenerated in parallel with selection. Specifically, the T7 promoter is added to full-length cDNA while the promoter and cleaved off region are added to cleaved cDNA, respectively. This enables direct amplification of the purified ligation product using PCR, avoiding cDNA purification in the case of positive selection (step 3 – Figure 2-19).

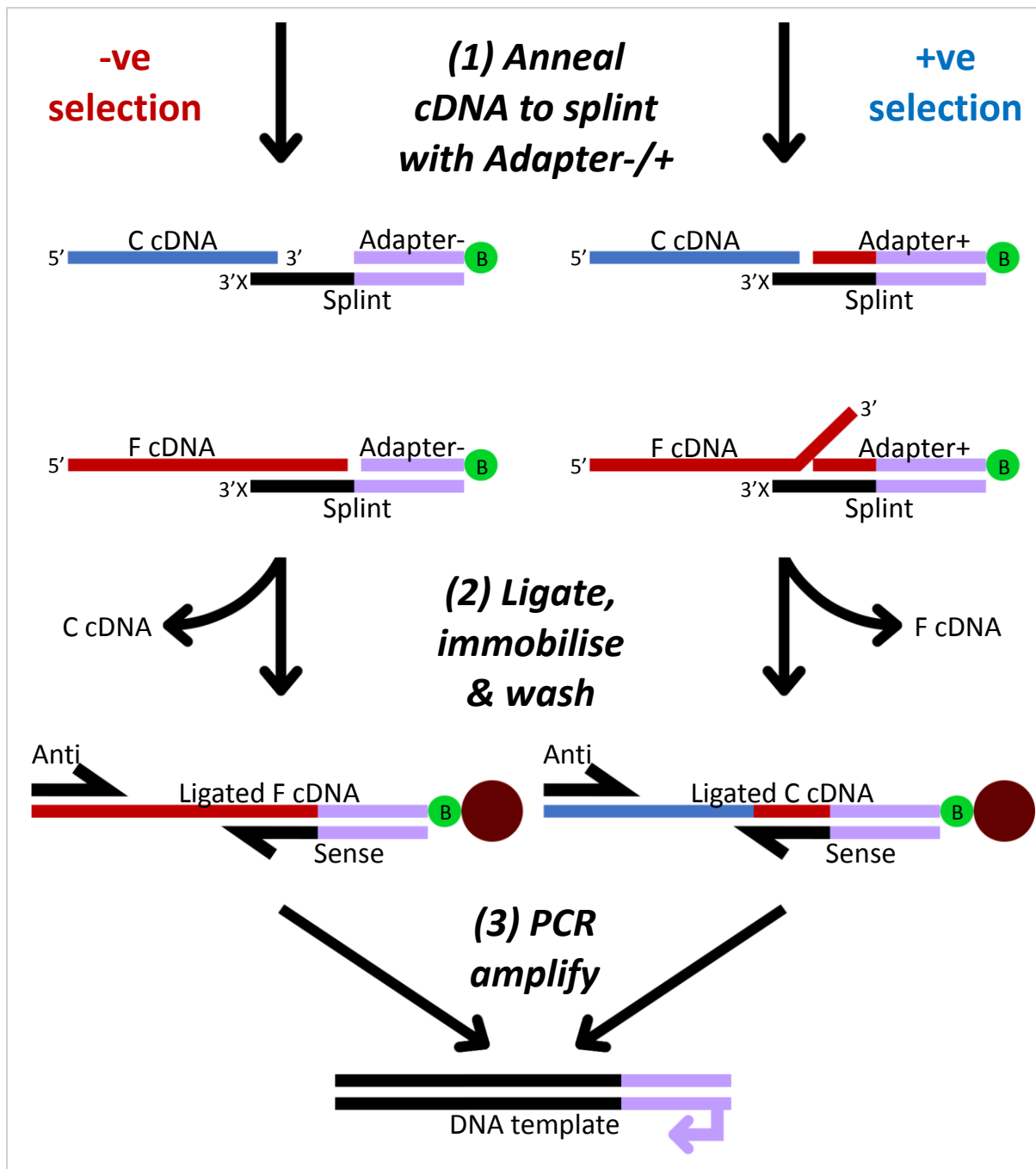


Figure 2-19: Ligase mediated cDNA selection – During negative & positive selection the cDNA is annealed to a 3' blocked (X) Splint with either Adapter- or Adapter+, respectively (step 1). Only the desired cDNA forms a T7 DNA ligase substrate<sup>147</sup>. The selected cDNA is ligated to its respective adapter. Each adapter is 3' biotinylated allowing the ligation product to be immobilised on streptavidin-coated paramagnetic beads and washed with NaOH (step 2). The now purified ligation product is PCR amplified regenerating the DNA template (step 3).

The ligase mediated selection method was analysed for its ability to selectively ligate full-length and cleaved cDNA as this was key to the mechanism of selection. Figure 2-20, illustrates that following both ligation reactions a band with a slightly larger MW than the DNA template is produced. Like the bait-oligo ligation reaction tested earlier (Figure 2-11), this corresponds with the ligation reaction product. Furthermore, following positive selection ligation, the intensity of the band corresponding to cleaved cDNA decreases. This suggests that cleaved cDNA is acting as the substrate during this reaction. Similarly, the amount of full-length cDNA decreases following negative selection ligation reaction, suggesting full-length cDNA acts as the substrate in this case. Therefore, it was concluded that the ligase method is capable of selectively ligating full-length or cleaved cDNA enabling the subsequent purification and amplification of these species.

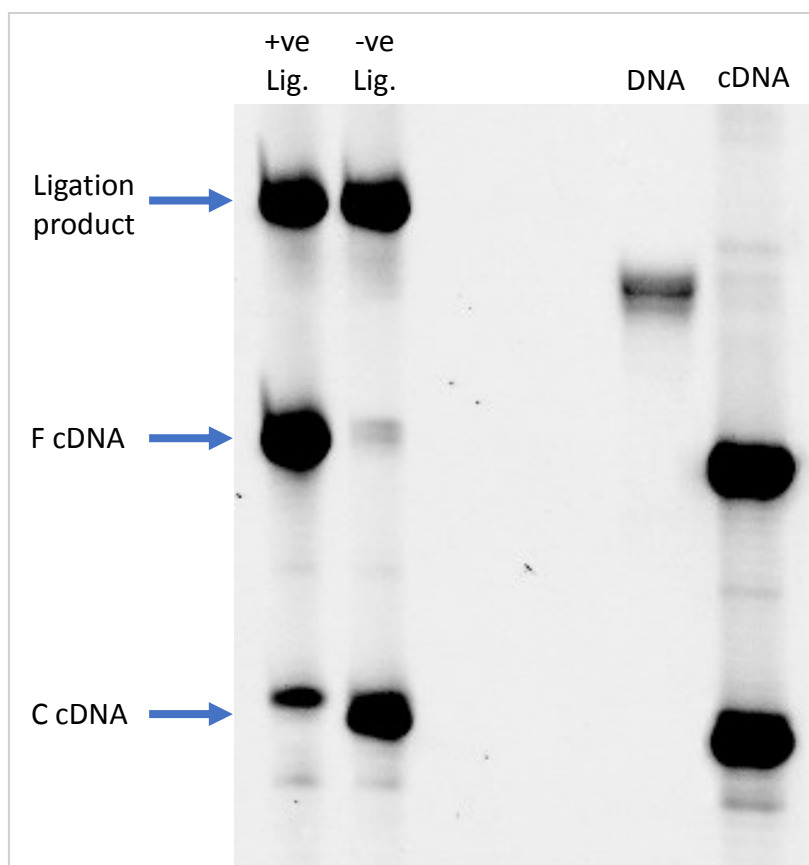


*Figure 2-20: Ligase Mediated Selection of full-length or cleaved cDNA - Urea PAGE gel with Lanes containing: ligation reaction products for positive (+ve) & negative (-ve) selection, the DNA template and the cDNA substrate. Full-length (F) and cleaved (C) cDNA is indicated. Gel is stained using EtBr and imaged for both EtBr and HEX-mediated cDNA fluorescence.*

Having shown that ligase mediated selection method was ligating the correct substrate, the effectiveness of positive selection was measured in a similar manner to that conducted previously for the bait-oligo method. The analysis of bait-oligo and ligase mediated positive selection in Figure 2-21 suggests ligase mediated positive selection is more effective. This is characterised by higher % cleaved values following ligase mediated selection in all cases. Furthermore, the difference between the two selection methods is particularly noticeable when more negative selections have been conducted prior to the positive selection. For instance, after 5 consecutive rounds of negative selection, a % cleaved value of 24.5 % was achieved using ligase mediated selection compared to 14.3 % when using the bait-oligo method. Although triplicates for each of the conditions were not conducted, the fact

that the ligase mediated selection leads to higher % cleaved values under four different conditions supports this interpretation.

While Ligase mediated positive selection appears to be more effective than the equivalent bait-oligo selection method. Conducting Ligase mediated positive selection on the initial Trp Library did not lead to a measurable increase in the % cleaved value. This suggests that in contrast to our expectations, neither constitutively active nor Trp-activated sequences were enriched in this case. This result is further reviewed in the Discussion (Section 2.4.4).

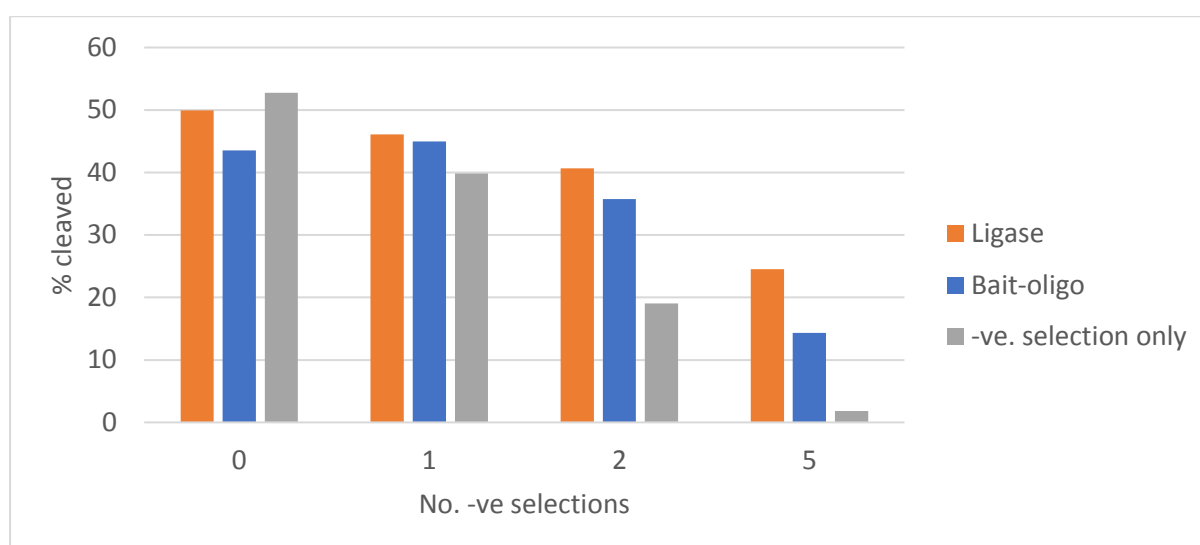


*Figure 2-21: Effectiveness of bait-oligo and ligase mediated positive selections – positive selections were conducted after 0, 1, 2 or 5 consecutive negative selections. % cleaved values of the purified RNA following bait-oligo and Ligase Mediated Selection (blue & orange) are shown with values before selection (grey).*

## 2.4 Discussion

### 2.4.1 Selection method

In line with the objectives of this chapter, an alternative method for the *in vitro* selection of ligand-responsive ribozymes was outlined. This scheme differs from previous examples of *in vitro* ribozyme selection<sup>55,119,127,148,149</sup> and as such there are several advantages and limitations to this method.

#### 2.4.1.1 Advantages of the method

One advantage of this method is the reduction in time achieved compared to previous methods. Based on the selection experiment implemented in the subsequent chapter, rounds of selection can be completed within 8 hours (Appendix, Section 9.2.3) and as such can be accommodated within a working day. Previous methods selecting for ribozyme-based riboswitches *in vitro* often required 3 overnight incubation steps for every cycle of selection<sup>55,72,119</sup>. This implies the equivalent 2 rounds would require at least four 4 days to complete. The method described within this thesis therefore represents a 50 % reduction in the time required. Given that 10 – 15 cycles of selection are often implemented prior to the detection of functional sequences<sup>55,72</sup>, this method has the potential to save months of time per experiment. This time saving should increase the overall throughput of the method and enable data to be collected at a faster rate.

Furthermore, with additional work it should be possible to automate parts or all the selection method. This is because all of the processes required to implement this method have previously been automated using liquid-handling robots<sup>128,129,146,150</sup>. This could further reduce the time and resources required while improving the reliability of the method<sup>151</sup>.

Although some of the nucleic acid purification steps tested didn't meet the specified requirements, as described in Section 2.4.3, this method should still be scalable in 96-well format. This implies that compared to equivalent previous methods<sup>119,149</sup>, more libraries and/or ligands could be screened in parallel, reducing the resources required per experiment. The ability to screen several libraries in parallel would also enable more diverse regions of the sequence space can be explored. For instance, libraries containing randomised regions<sup>66,68,69</sup>, structured regions<sup>152</sup> or SELEX-enriched aptamers<sup>55,72</sup> could all be screened during a single experiment. The ability to screen for responsive ribozymes using many diverse ligands will improve the probability that one of these ligands influences the activity of sequences present during selection. This is expected to prove useful considering that different ligands have different potentials to interact with RNA<sup>126</sup> and hence influence the activity of ribozymes. These

beneficial aspects of scaling selection should improve the probability that novel ligand-responsive ribozymes and ultimately riboswitches are identified.

Another potential advantage of this selection method results from the avoidance of size selection steps. This should enable the accumulation of any number of insertion and/or deletion mutations. These mutations are actively selected against using previous methods employing size selection as the size of the library is corrected following gel purification. Again, this capability to explore a greater region of the sequence space with the accumulation of larger insertion and deletion mutations should improve the propensity for identifying rarer sequences with improved functionality.

In contrast to previous methods, the method outlined in this chapter could select ligand-responsive ribozymes which function through co-transcriptional folding. Co-transcriptional folding in this context means the ligand influences the folding and hence activity of the resulting ribozyme during transcription. This mechanism is often seen in natural riboswitches<sup>153,154</sup> and improves the functionality by increasing the energy required to switch between “Off” and “On” conformations after transcription has occurred<sup>155</sup>. Previous *in vitro* selection methods require the purification of ribozymes between positive & negative selection meaning “On” & “Off” conformations must interconvert without the process of transcription. In contrast our method involves transcription prior to both selections.

#### 2.4.1.2 Limitations of the method

Comparing this selection method to previous methods, there is likely to be less control over the stringency of selection. Using previous *in vitro* methods, the stringency of selection can be varied by changing the time the library is incubated for during selection<sup>119</sup>. For instance, during positive selection to increase the stringency, the library is incubated for a shorter period prior to quenching ribozyme catalysis. This ensures variants with slower ligand-dependent cleavage rates are selected against. Using the method presented, full-length RNA is not purified and as such selection starts immediately following transcription. While positive selection stringency can be increased by

increasing the length of the transcription reaction, this variable would also reduce the throughput of selection by synthesising less RNA. Mechanisms to overcome low RNA yields with short transcription reactions are considered in the subsequent chapter.

A further limitation of this method is that there are more steps and hence the propensity for selection bias is greater. For example, compared to previous methods twice as many reverse transcriptions are required. Potentially, the high error rate and mutational bias of the RTase could bias the selection of sequences with specific properties<sup>156</sup>. Furthermore, the library is amplified by PCR following both positive and negative selection increasing the propensity for PCR biases<sup>157</sup>. With that said, the use of Next-Generation Sequencing and other modifications to SELEX<sup>100,158</sup>, have enabled such biases to be minimised to a degree where desirable sequences can be identified. Similar methodologies could also be applied in this case to limit these biases.

#### 2.4.2 Library design

In this chapter, several domains of a previously designed Trp Library<sup>131</sup> were modified to improve the compatibility of the library with the method. From the data gathered about the performance of the redesigned library, all these requirements were achieved. Specifically, the library design enabled specific transcription of both full-length and cleaved RNA. Both these species were efficiently reverse transcribed yielding cDNA for selection. It was also possible to specifically anneal a bait-oligo to the full-length cDNA species facilitating selection. Finally, the base-pairing probabilities of the MFE 2° structure were close to that of the consensus hammerhead ribozyme sequence suggesting that the redesigned regions are unlikely to inhibit the performance of individual variants.

One specification which was not considered during the library design was the ease at which primers for the resulting library could be designed. A consequence of this was that anti-sense primers which annealed outside of stem-loop II could not be used. As such, variability in this region is restricted for libraries based on this general design. This is disadvantageous as evolving residues in stem-loop II could lead to the selection of ribozymes which function through alternative 3° interactions<sup>46</sup>,

potentially improving functionality. Furthermore, the ability to evolve ligand binding sequences in stem-loop II could also be desirable as this strategy was implemented for the generation of previous ribozyme-based riboswitches<sup>54</sup>.

To enable the use of anti-sense primers which anneal outside of stem-loop II, inspiration can be taken from a recent example which inserted SHAPE-seq cassettes into a stem equivalent to Stem III<sup>152</sup>. The use of these SHAPE-seq cassettes meant that stems equivalent to stem I and II in the library design did not contain primer binding sites. Therefore, including these cassettes or a similar motif in future designs should facilitate the use of an anti-sense primer which does not include stem-loop II sequences.

#### 2.4.3 Non-specific purification of nucleic acids

Methods for the *in vitro* selection of ribozyme sequences, including this method require the purification of all nucleic acids in the pool multiple times during selection<sup>119,127,149</sup>. Previous methods achieved these purifications steps using silica spin-columns and/or ethanol precipitation steps. However, substituting these steps for SPRI-based steps, would improve the ease at which this method could be scaled in microtiter plates and subsequently automated<sup>144,145</sup>.

By optimising the [PEG] and [NaCl] good yields of purified, total cDNA were achieved using SPRI. Using a similar buffer composition, total RNA could also be purified using SPRI to the required concentrations. However, surprisingly when the volume of IVT reaction was increased and the same elution volume used, the titre and hence yield of total RNA was lower. It is suspected that the reduction of yield with a larger IVT volume is due to the high concentration of PEG used. As mentioned, the majority of SPRI applications utilise a lower [PEG]. With a higher [PEG], the viscosity of the solution is increased and hence the force resisting the migration of SPRI beads within the solution is larger. Given as part of purification, SPRI beads must be pelleted, it is plausible that the increased [PEG] prevents complete pelleting for larger volumes, reducing overall yield. This result highlights the importance of optimising novel SPRI steps experimentally before their implementation.

In contrast to total RNA and cDNA purification, the required yield of the purified DNA template from the PCR step was not achieved using SPRI. Given no optimisation of this step was implemented, the required yield might be achieved with further work tuning this strategy. This could include determining the optimal [PEG] which maximises DNA yield without co-purification of PCR primers. On the other hand, the current silica spin-column strategy could be scaled and if necessary automated<sup>159,160</sup>, meaning the aim of developing an *in vitro* selection method which operates at a larger scale is still achieved without SPRI. However, one disadvantage of this approach is that more equipment would be required making the process more expensive and complicated.

#### 2.4.4 Bait-oligo vs Ligase Mediated Selection

In this chapter, two methods were conceived which enabled the selective purification of cDNA molecules, driving selection. It was shown experimentally that both methods could purify the desired cDNA molecules. Furthermore, the functionality of bait-oligo negative selection was demonstrated by enriching for ligand-activated and inactive variants to the point where constitutively-active sequences were no longer detectable in the pool.

However, ligase mediated positive selection was shown to have higher efficacy than the equivalent bait-oligo method. This result likely arises from the direct isolation of cleaved cDNA molecules, rather than indirectly isolating this species with the removal of full-length cDNA molecules based on oligonucleotide hybridisations. Given these hybridisations exist in a chemical equilibrium<sup>25</sup>, some full-length cDNA will always be left in the supernatant following bait-oligo mediated positive selection. This point of view is supported by a prior study which aimed to remove sequences with high levels of background activation using bait-oligo hybridisations and pull-downs<sup>161</sup>. While successful, a significant number of selected sequences had this undesirable phenotype. In contrast, given DNA ligases completely discriminate against incorrect substrates under certain conditions<sup>162</sup>, this purification step has the potential to have 100 % efficacy.

In addition to having improved positive selection efficiency, DNA ligase mediated selection should be easier to automate and is more novel than the bait-oligo method. This latter point arises from the fact that to the best of our knowledge, DNA ligases have not yet been used to mediate directed evolution. The increased ease of automation can be reasoned by comparing the workflow of the two selection methods. Firstly, positive and negative selection workflows for the ligase DNA selection method are identical except for the adapters used in the ligation step. This contrasts with bait-oligo selection meaning more individual steps must be automated. Furthermore, the Ligase Mediated Selection can be conducted in a shorter period. This conclusion comes from the fact that 2 rather than 5 pull-downs are required, along with one less purification step during positive selection. These factors represented a compelling case for the use of ligase mediated selection in subsequent experiments.

Unfortunately, neither method could enrich for the expected phenotype when conducting positive selection on the Trp Library. Given that the expected phenotype was enriched after more negative selections, this effect is unlikely to be caused by full-length cDNA contamination. If this was the case, the magnitude of this effect should be greater when there is more full-length cDNA present in the pool as is the case after more negative selections.

Perhaps the time taken to prepare cDNA inhibits the effectiveness of positive selection. There is some discrepancy between the time taken to prepare cDNA using the method outlined in this chapter and previous positive selection incubation times. For instance, previous *in vitro* selection methods use positive selection incubation times ranging from 30 minutes to 10 seconds<sup>55,72</sup>. In contrast it takes several hours to prepare cDNA using the selection schemes formulated in this chapter. However, other reasons for the observed results such as biases for inactive variants cannot be ruled out. Therefore, to address this issue a control experiment should be formulated to ensure the selection method is functional. The work conducted in the subsequent results chapter develops and implements this control experiment.

## 2.5 Summary

The aim of this results chapter was to conceive an *in vitro* selection method for the enrichment of ligand-responsive ribozymes which could be implemented in shorter times and could achieve a greater scale compared to previous methods. It was hypothesised that the development of such a method would improve the likelihood of identifying novel ribozyme-based riboswitches. To this end, an alternative scheme of selection was developed. Two specific selection strategies which could be used to implement this scheme were formulated during this chapter. To test the functionality of these strategies, a previously conceived Trp Library<sup>131</sup> was modified to increase its compatibility with the method. Using this library, the novel steps of selection and remaining purification steps were tested and optimised. Comparing the two selection strategies formulated during this chapter, ligase mediated selection has several advantages including a more effective positive selection and easier implementation and future automation. Although expected phenotypes were enriched in several cases using this method, this was not the case when positive selection was conducted on the initial library. To ensure the selection method is functional it was suggested to conduct a control experiment, enriching for a known ligand-responsive ribozyme.

## Chapter 3: Enrichment of a theophylline-activated ribozyme

### 3.1 Introduction

In the previous chapter, a novel *in vitro* selection method which could select for ligand-responsive ribozymes was shown to be feasible and capable of selecting for the desired full-length or cleaved cDNA molecules. However, conducting positive selection on the initial Trp Library reduced the % cleaved value of the pool. In contrast to what was predicted, this result suggested ligand-activated and constitutively active sequences were not enriched under these conditions.

To ensure the selection method functions as required, a control experiment which involves the enrichment of a known ligand-activated ribozyme will be designed and implemented in this chapter. Achieving a positive result from this control experiment would demonstrate that the method functions as predicted. This should bring the overall aim of using this methodology to identify a ribozyme-based riboswitch with new ligand specificities closer to realisation.

To fully test the selection method a control library must be designed that contains undesirable sequences as well as the known ligand-activated ribozyme. This is because an effective selection method should discriminate against these undesirable sequences while enriching for ligand-responsive ones (refer to Figure 2-3A). Note that the use of a completely degenerate library would not guarantee that sequences with the required undesirable phenotypes are present during selection. In such cases a non-functional selection method could still deliver a positive result.

Having designed a control library, steps will be taken to ensure that the known ligand-activated sequence present within the library can be enriched in practical amount of time. Following this, the selection method will be implemented to enrich for a ligand-activated phenotype corresponding to that of the known sequence. Enrichment of the ligand-activated sequence or sequences with similar functionality will then be confirmed, demonstrating that the selection method worked as expected.

## 3.2 Objectives

### *Objective 1:*

Design a control library containing a previously characterised ligand-activated ribozyme. In addition to this ribozyme, the library should contain inactive and constitutively active ribozymes. This will facilitate the testing of both positive and negative selections which should discriminate against these sequences, respectively.

### *Objective 2:*

Optimise selection conditions to enable the enrichment of the ligand-activated ribozyme from the control library in a practical amount of time.

### *Objective 3:*

Implement selection, enriching for a ligand-activated phenotype from the control library.

### *Objective 4:*

Confirm the enrichment of the ligand-activated sequence or sequences with similar functionality.

## 3.3 Results

### 3.3.1 Objective 1 – Design of a control library

In line with objective 1, a library will be designed which:

- I. Contains a previously characterised ligand-activated ribozyme.
- II. Is compatible with the proposed selection method enabling the enrichment of the ligand-activated sequence from this library.
- III. Contains both inactive and constitutively active ribozymes. These sequences permit both positive & negative selection to be scrutinised as if either process was not functional, the ligand-activated ribozyme would not be enriched (refer to Figure 2-3A).

So that the library satisfies the first of the above requirements, a previously selected ligand-activated ribozyme must be identified. Furthermore, this ligand-activated ribozyme should only contain residues conferring ligand sensing properties in regions other than those required to implement selection. For instance, ligand sensing nucleotides are prohibited in stem III as this domain contains residues required for reverse transcription, selection and PCR. From the literature, previously selected theophylline and tetracycline-activated ribozymes meet these specifications<sup>55,72</sup>. Given the extensive characterisation and use of the theophylline aptamer<sup>46,104,118,163–167</sup>, theophylline was chosen over tetracycline. Of the theophylline-activated ribozymes identified in the above study, the VI-1 variant (Figure 3-1A) had the highest dynamic range reported. Specifically, transcribing this variant in the presence of 3.16 mM theophylline increases the rate of cleavage 285-fold compared to transcribing in its absence<sup>72</sup>. The VI-1 variant was therefore chosen as the basis for constructing the control library.

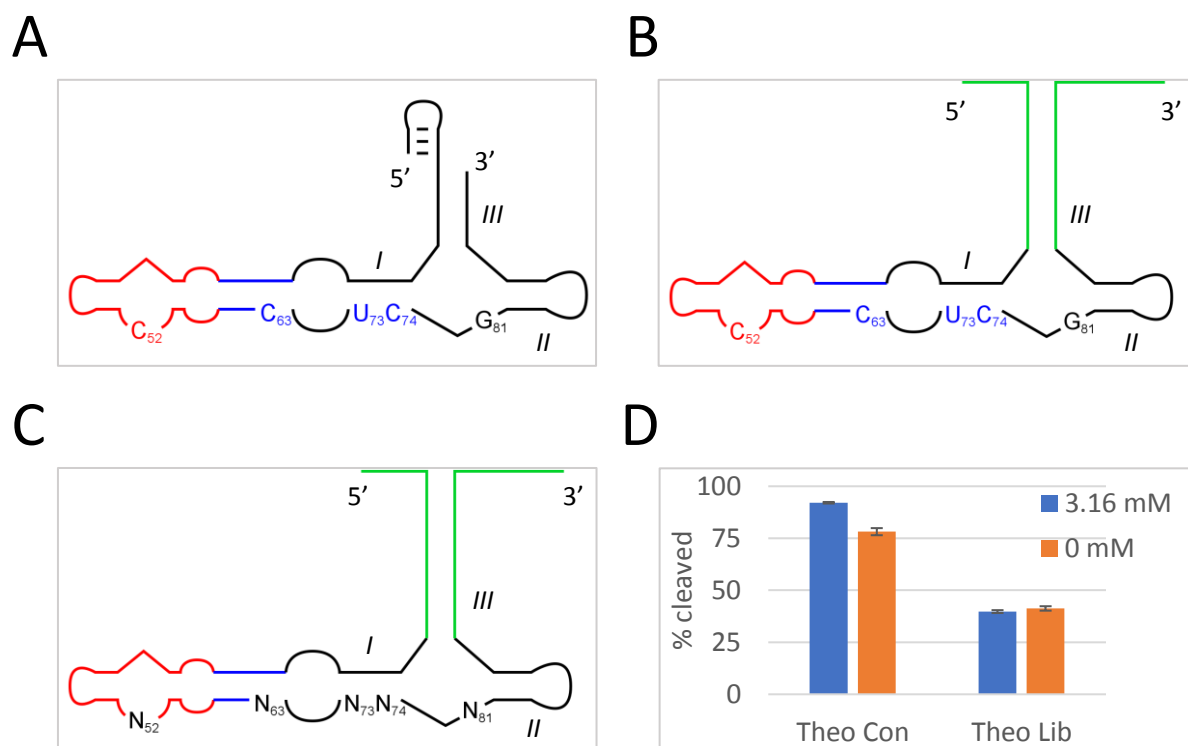


Figure 3-1: Design of the Theophylline Library – (A – C) illustrates 2° structure cartoons for ribozymes. Roman numerals denote stem-loop numbers. The position of nucleotides is indicated by subscripts. Regions randomised in the study by Breaker & co-workers<sup>72</sup>(blue) or ligand binding regions (red) are indicated along with Stem III from the Trp Library (green). (A) VI-1 variant as described previously<sup>72</sup>. (B) Theophylline Control. (C) Theophylline Library. (D) DNA templates for the Theophylline Control (Theo Con) and the Theophylline Library (Theo Lib) were transcribed and the purified RNA % cleaved values determined via electrophoresis. Error bars denote SDs from 3 repeats.

Without any modifications the VI-1 variant is not compatible with the selection method. This is because it does not contain residues in stem III which enable the selection process (refer to Section 2.3.2). To remedy this, Stem III from the Trp Library used in the previous chapter is substituted into the VI-1 variant<sup>72</sup> (Figure 3-1B). This sequence generated from this substitution is henceforth referred to as the Theophylline Control and functions as a positive control during this experiment.

Along with the Theophylline Control, variants with inactive phenotypes should also be present in the control library to scrutinise negative selection. Mutations to key residues involved in ribozyme

catalysis will lead to the formation of inactive ribozymes satisfying this requirement. The 2'-OH of G81 from the parent VI-1 sequence is equivalent to G8 in Figure 1-2 and as such functions as general acid during catalysis<sup>12</sup> (Figure 1-1B). While this mechanism does not directly involve the guanine nucleobase of this residue, biochemical and crystal structure analysis suggests a Watson-Crick base pair is required between G81 and C76 for the correct positioning of the 2'-OH group on this residue<sup>23,168</sup>. As such a G81N substitution will mean  $\frac{3}{4}$  of the starting library will have significantly lower cleavage rates, equivalent to an inactive phenotype.

Constitutively active variants are defined by fast cleavage rates, which do not vary depending on the ligand concentration. To introduce these sequences into the control library a two-step approach was taken. The first step involved inserting one or more degenerate residues at loci shown to interact with theophylline. Such sequences would not be able to interact with theophylline ensuring that their cleavage rates are identical in its presence and absence. Given C52 in the Theophylline Control makes three hydrogen bonds directly with theophylline<sup>163</sup>, introducing a degenerate base at this loci achieves this goal.

The second step to introduce constitutively active variants into the control library involved inserting degenerate residues to ensure that some variants within the library could form the active 2° structure in the absence of ligand. Together with the above substitution, such sequences are likely to be constitutively active. To identify these residues, the regions randomised in the previous library used to select the VI-1 variant were analysed (Figure 3-1A, blue). Given wild type ribozymes contain perfect base-pairing within these regions<sup>169</sup>, bases which lead to imperfect base-pairing were hypothesised to perturb the active conformation until ligand binding. C63, U73 & C74 were all found to prevent perfect base-pairing within these regions and as such degenerate bases were inserted at these loci.

To determine whether the above approach introduced constitutively active sequences into the control library, the MFE 2° structure of a sequence which was hypothesised to be constitutively active was analysed (Figure 3-2A). As illustrated by Figure 3-2A, this variant contains correctly base paired stems

I, II & III due to U63, C73 & A74 identities. This 2° structure corresponds with the active conformation (refer to Figure 1-2) suggesting cleavage occurs in the absence of theophylline. This contrasts with the Theophylline Control which requires theophylline-binding to stabilise the formation of these stems and hence the active conformation (Figure 3-2B). Furthermore, the sequence in Figure 3-2A contains an A52 identity and as such will have reduced affinity for theophylline. Therefore, its rate of cleavage will not be influenced by this ligand, satisfying the above definition.

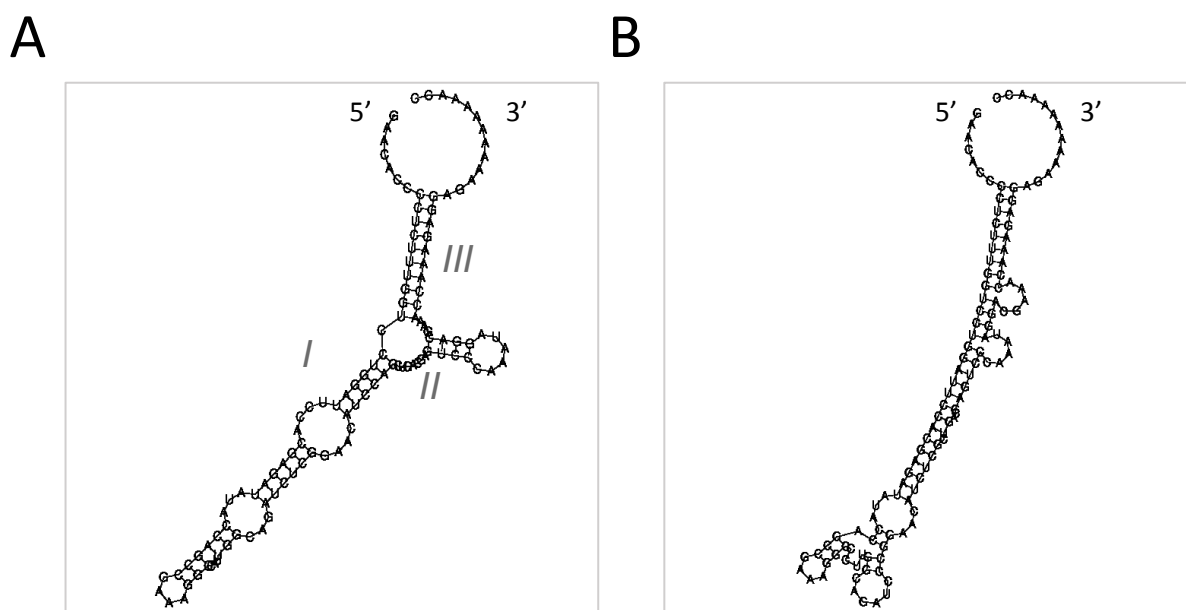


*Figure 3-2: Theophylline Library constitutively active sequence – Roman numerals denote correctly base-paired stems/stem-loops. (A) MFE 2° structure of a variant in the Theophylline Library with A52, U63, C73, A74 & G81 identities. (B) MFE 2° structure of the Theophylline Control.*

Making the above five degenerate base substitutions produces a library henceforth referred to as the Theophylline Library (Figure 3-1C). This library contains 1024 sequences, of which only one is the Theophylline Control sequence. The responses of the Theophylline Library and the control to theophylline were measured under the selection conditions used in the previous chapter. Figure 3-1D illustrates the response of these two species. As expected, the library shows no response to theophylline having identical % cleaved values in the presence and absence of ligand. This suggests that the library is dominated by sequences that are unresponsive to the ligand theophylline.

### 3.3.2 Objective 2 – Optimisation of selection conditions

While the Theophylline Control does display some response to theophylline under the conditions used (Figure 3-1D), the response is marginal given the large change in rate constants previously reported<sup>72</sup>. It was therefore questionable as to how many cycles of selection would be required to enrich the Theophylline Control to a detectable frequency? A further question, was how alterations to the selection conditions might increase the rate of enrichment and to what magnitude? To help answer these questions, a model describing the selection process was first developed and elaborated below.

#### 3.3.2.1 Derivations - Number of cycles required for detection?

To detect a sequence following selection, it is possible to pick random sequences from the pool at the end of selection and determine their sequence identity using sequencing technologies<sup>55,66,68,69,72,170</sup>. The number of sequences sampled will determine the frequency required to guarantee that on average, a sequence is observed. For example, if 10 sequences are randomly selected, a sequence must be present at a frequency of 10 % to ensure it is observed on average.

To determine the number of cycles required to enrich a sequence ( $i$ ) to a given frequency, the dynamics of this sequence during selection can be used. The dynamics of sequence  $i$  during selection can be summarised by the frequency of this sequence as a function of the number of rounds of selection,  $n$ . To determine the number of rounds and hence cycles required to enrich a sequence to frequency, it is simply a matter of finding the value of  $n$  from the dynamics for which the frequency exceeds a given threshold<sup>\*4</sup>.

To determine the frequency of sequence  $i$  at the  $n^{th}$  round of selection, a previously proposed model for an evolving system was considered<sup>171</sup>. This model states that the number of DNA templates of sequence  $i$  ( $x_i$ ) after the  $n^{th}$  round of selection ( $x_i^{(n+1)}$ ) is given by:

---

<sup>\*4</sup>Refer to Figure 3-9 for an example.

$$x_i^{(n+1)} = x_i^{(n)}(k_i^{(n)} + 1) \quad (3-1)$$

where  $k_i^{(n)}$  is the growth rate of sequence  $i$  evaluated at the  $n^{th}$  round of selection. Note that Equation (3-1) can be converted to frequencies ( $\frac{x_i^{(n)}}{X}$ ) by dividing both sides by the total number of DNA templates in the pool ( $X \equiv \sum_{i=1}^I x_i$ ). It should also be noted that  $X$  is assumed to be constant given that a fixed concentration of DNA is used in each IVT reaction. It is also important to consider that while  $n$  is given as a superscript, rather than a subscript, it denotes a series rather than an exponent. To calculate the dynamics of sequence  $i$  using Equation (3-1), it is necessary to know the initial frequency of sequence  $i$  ( $\frac{x_i^{(0)}}{X}$ ) and all the growth rate constants at each round of selection ( $k_i^{(0)}, k_i^{(1)} \dots k_i^{(N)}$ ).

Assuming all sequences are present at the same initial frequency,  $\frac{x_i^{(0)}}{X}$  can be calculated from the number of sequences in the initial library ( $I$ ):

$$\frac{x_i^{(0)}}{X} = \frac{1}{I} \quad (3-2)$$

Although biases in DNA synthesis will affect this assumption<sup>172</sup>, this approximation is often used when gauging library coverage and as such is likely to be relatively accurate.

While a value for  $\frac{x_i^{(0)}}{X}$  is relatively easy to come by, determining the values of the growth rate constants ( $k_i^{(n)}$ ) is not as obvious. To determine these values, the mechanics of the selection process was considered (refer to Chapter 2, Figure 2-3B). In the ideal case, positive selection should select all cleaved cDNA molecules produced during selection ( $C_T^{(n+)}$ ). From this set of molecules some will belong to sequence  $i$  ( $C_i^{(n+)}$ ). Given the number of DNA templates in the system is constant ( $X$ ):

$$x_i^{(n+1)} = \frac{C_i^{(n+)}}{C_T^{(n+)}} X \quad (3-3)$$

where  $x_i^{(n^++1)}$  is the number of molecules of  $x_i$  after the  $n^{th}$  round of selection, given that the  $n^{th}$  round is a positive round of selection ( $n^+$ ). This distinction is important considering positive and negative selections differ in at least the concentration of ligand used and the cDNA molecules which are selected for. As such the dynamics of sequence  $i$  will vary depending on whether positive or negative selection is conducted. Similarly, during a round of negative selection all full-length cDNA molecules ( $F_T^{(n^-)}$ ) should be selected. Again, from this set of molecules some will belong to sequence  $i$  ( $F_i^{(n^-)}$ ), meaning:

$$x_i^{(n^-+1)} = \frac{F_i^{(n^-)}}{F_T^{(n^-)}} X \quad (3-4)$$

where  $x_i^{(n^-+1)}$  is the number of molecules of  $x_i$  after the  $n^{th}$  round of selection, given the  $n^{th}$  round is a negative round of selection ( $n^-$ ).

To derive an expression for the growth rate constants during positive ( $k_i^{(n^+)}$ ) and negative ( $k_i^{(n^-)}$ ) selection, Equations (3-3) & (3-4) are substituted into (3-1), yielding:

$$k_i^{(n^+)} = \frac{X}{C_T^{(n^+)}} \times \frac{C_i^{(n^+)}}{x_i^{(n)}} - 1, \quad k_i^{(n^-)} = \frac{X}{F_T^{(n^-)}} \times \frac{F_i^{(n^-)}}{x_i^{(n)}} - 1 \quad (3-5)$$

To simplify Equation (3-5) further, two assumptions were made. Firstly, it is assumed that the IVT reaction is a pseudo-first-order chemical reaction with a rate directly proportional to  $x_i^{(n)}$ . Second, it is assumed that all RNA is reverse transcribed. This second assumption is supported by previous experiments showing that even with relatively high concentrations of RNA and random primers, yields of 80 – 90 % can be achieved<sup>173</sup>. Note the RT reaction used during this selection method used the recommended concentration of RNA along with a primer which specifically annealed at the 3' end of the RNA molecule, suggesting that high % yields would be achieved. With these assumptions  $F_i^{(n)} + C_i^{(n)}$  and  $F_T^{(n)} + C_T^{(n)}$  can be written as follows:

$$F_i^{(n)} + C_i^{(n)} = k_T t x_i^{(n)}, \quad F_T^{(n)} + C_T^{(n)} = k_T t X$$

Where  $t$  is the length of time the transcription reaction is incubated for and  $k_T$  is the transcription reaction rate constant, respectively. Rearranging these expressions and solving for  $x_i^{(n)}$  and  $X$  yields:

$$x_i^{(n)} = \frac{F_i^{(n)} + C_i^{(n)}}{k_T t}, \quad X = \frac{F_T^{(n)} + C_T^{(n)}}{k_T t} \quad (3-6)$$

Using the result in (3-6) and some algebraic manipulation, Equation (3-5) can be rewritten as:

$$k_i^{(n+)} = \frac{\frac{C_i^{(n+)}}{F_i^{(n+)} + C_i^{(n+)}}}{\frac{C_T^{(n+)}}{F_T^{(n+)} + C_T^{(n+)}}} - 1, \quad k_i^{(n-)} = \frac{1 - \frac{C_i^{(n-)}}{F_i^{(n-)} + C_i^{(n-)}}}{1 - \frac{C_T^{(n-)}}{F_T^{(n-)} + C_T^{(n-)}}} - 1$$

The parameters  $\frac{C_i^{(n+)}}{F_i^{(n+)} + C_i^{(n+)}}$  and  $\frac{C_i^{(n-)}}{F_i^{(n-)} + C_i^{(n-)}}$  are equivalent to the fraction of sequence  $i$  which is cleaved during positive and negative selection, respectively. If identical positive and negative selection conditions are used throughout selection, these parameters will not vary<sup>148</sup> and as such are independent of  $n$ . With this realisation the above expression is rewritten as:

$$k_i^{(n+)} = \frac{\%C_i^{(+)}}{\%C_T^{(+)}} - 1, \quad k_i^{(n-)} = \frac{1 - \%C_i^{(-)}}{1 - \%C_T^{(-)}} - 1 \quad (3-7)$$

where  $\%C_i^{(+)} = \frac{C_i^{(+)}}{F_i^{(+)} + C_i^{(+)}}$  and  $\%C_i^{(-)} = \frac{C_i^{(-)}}{F_i^{(-)} + C_i^{(-)}}$  and are both independent of  $n$ . The identities of the variables  $\%C_T^{(n+)}$  and  $\%C_T^{(n-)}$  are given in Equation (3-8) below:

$$\%C_T^{(n+/-)} = \frac{C_T^{(n+/-)}}{C_T^{(n+/-)} + F_T^{(n+/-)}} \equiv \sum_i \left( \frac{x_i^{(n)}}{X} \%C_i^{(+/-)} \right) \quad (3-8)$$

The “+/-” notation in Equation (3-8) indicates that each variable or parameter could take either “+” or “-” depending on whether positive or negative selection is conducted, respectively. The result of Equations (3-7) is that the growth rate constants for sequence  $i$  throughout selection can be calculated ( $k_i^{(0)}, k_i^{(1)} \dots k_i^{(N)}$ ) if two parameters and one variable are known. Note that  $\%C_T^{(n+/-)}$  is

considered as a single variable given that a selection experiment can only take one course, meaning the response of the pool at a given round of selection can only be one value.

The parameters are the cleavage responses of sequence  $i$  under positive ( $\%C_i^{(+)}$ ) and negative ( $\%C_i^{(-)}$ ) selection conditions, respectively. These parameters can be empirically determined prior to selection as illustrated for the Theophylline Control in Figure 3-1D. In this example, these responses were determined for the current selection conditions.

Equation (3-8), suggests that the variable  $\%C_T^{(n+/-)}$  can be calculated using one of two possible strategies. Either the cleavage response of the pool for each round of selection is calculated empirically\*<sup>5</sup> or the cleavage responses of all the sequences in the pool along with their frequencies must be known\*<sup>6</sup>. Unfortunately, the first strategy cannot be conducted prior to selection and the second strategy is impractical given the large number of sequences often present during *in vitro* selection experiments.

To overcome this limitation, it is assumed in the remainder of this thesis that there are only two sequences present during selection, sequence  $i$  and sequence  $j$ . The accuracy of this assumption is addressed at each point this model is used to estimate the dynamics of sequences during selection.

With this assumption,  $\%C_T^{(n+/-)}$  can be calculated as follows:

$$\%C_T^{(n+/-)} = \frac{x_i^{(n)}}{X} \%C_i^{(+/-)} + \left(1 - \frac{x_i^{(n)}}{X}\right) \%C_j^{(+/-)} \equiv \sum_i \left(\frac{x_i^{(n)}}{X} \%C_i^{(+/-)}\right) \quad (3-9)$$

Rearranging Equation (3-9) to solve for  $\%C_j^{(+/-)}$  yields:

---

\*<sup>5</sup>Equivalent to  $\frac{C_T^{(n+/-)}}{C_T^{(n+/-)} + F_T^{(n+/-)}}$

\*<sup>6</sup>Equivalent to  $\sum_i \left(\frac{x_i^{(n)}}{X} \%C_i^{(+/-)}\right)$

$$\%C_j^{(+/-)} = \frac{\%C_T^{(n+/-)} - \frac{x_i^{(n)}}{X} \%C_i^{(+/-)}}{1 - \frac{x_i^{(n)}}{X}}$$

To yield a specific value for  $\%C_j^{(+/-)}$ , the cleavage response of the initial pool under positive ( $\%C_T^{(0+)}$ ) and negative ( $\%C_T^{(0-)}$ ) selection conditions can be used along with the initial frequency of sequence  $i$  ( $\frac{x_i^{(0)}}{X}$ ). Note the responses of the pool can be acquired prior to selection as illustrated for the Theophylline Library under the current selection conditions (Figure 3-1D). Using these values and the result from Equation (3-2), the above expression can be written as:

$$\%C_j^{(+/-)} = \frac{I\%C_T^{(0+/-)} - \%C_i^{(+/-)}}{I - 1} \quad (3-10)$$

Using the expressions in Equations (3-10), (3-9), (3-7) & (3-1) the frequency of sequence  $i$  after positive ( $\frac{x_i^{(n^++1)}}{X}$ ) and negative ( $\frac{x_i^{(n^-+1)}}{X}$ ) selection can be calculated as follows:

$$\frac{x_i^{(n^++1)}}{X} = p\left(\frac{x_i^{(n)}}{X}, \%C_i^{(+)}, I, \%C_T^{(0+)}\right), \quad \frac{x_i^{(n^-+1)}}{X} = n\left(\frac{x_i^{(n)}}{X}, \%C_i^{(-)}, I, \%C_T^{(0-)}\right) \quad (3-11)$$

where  $p\left(\frac{x_i^{(n)}}{X}, \%C_i^{(+)}, I, \%C_T^{(0+)}\right)$  and  $n\left(\frac{x_i^{(n)}}{X}, \%C_i^{(-)}, I, \%C_T^{(0-)}\right)$  are functions as listed in the Appendix (refer to Section 9.1). To calculate the value of these two functions, five parameters and one variable are required. The five parameters are  $\%C_i^{(+)}$ ,  $\%C_i^{(-)}$ ,  $\%C_T^{(0+)}$ ,  $\%C_T^{(0-)}$  &  $I$ . As mentioned throughout this section, it is possible to calculate all five of these parameters prior to conducting an experiment. Furthermore, so long as the order of positive and negative selections is known, it is possible to calculate the variable,  $\frac{x_i^{(n)}}{X}$  given  $\frac{x_i^{(0)}}{X}$  and the expressions in Equation (3-11). Given that  $\frac{x_i^{(0)}}{X}$  can be calculated from  $I$ ,<sup>\*7</sup> only the five parameters previously listed and the order of positive and

---

\*7Refer to Equation (3-2).

negative selections needs to be known to estimate the dynamics of a sequence during selection and hence the number of cycles required to detect it.

### 3.3.2.2 Derivations - How enrichment varies with modifications to selection conditions?

To identify optimal selection conditions, it is in principle possible to estimate the dynamics of this sequence for each condition using the methods described in the previous section. With this data, the number of cycles required to enrich a sequence can be compared for each of the conditions being tested. However, to obtain this data the cleavage responses of all sequences in the pool must be known or approximations about this variable must be made (refer to Equation (3-9)). In this section this limitation is addressed by deriving an expression which indicates the propensity for a sequence to be enriched without having to consider the response of the entire pool.

To derive this expression, the mechanics of a cycle of selection are first considered. Selection cycles using the method described in this thesis are characterised by two rounds of selection. From Equation (3-1) the frequency of sequence  $i$  ( $x_i^{(n+2)}$ ) after selection is therefore:

$$x_i^{(n+2)} = x_i^{(n)}(k_i^{(n)}k_i^{(n+1)} + k_i^{(n)} + k_i^{(n+1)} + 1)$$

Given for each cycle, one round of positive and one round of negative selection is conducted,  $k_i^{(n)}$  and  $k_i^{(n+1)}$  from the above expression can be substituted for  $k_i^{(n+)}$  and  $k_i^{(n-)}$ , yielding:

$$x_i^{(n+2)} = x_i^{(n)}(k_i^{(n+)}k_i^{(n-)} + k_i^{(n+)} + k_i^{(n-)} + 1) \quad (3-12)$$

Note in this instance it is assumed that for each cycle, a round of positive selection was conducted prior to negative selection. However, the final result in Equation (3-14) would not change if the reverse was considered. Substituting the values from Equation (3-7) into Equation (3-12),  $x_i^{(n+2)}$  can be evaluated as:

$$x_i^{(n+2)} = x_i^{(n)} \frac{\%C_i^{(+)}(1 - \%C_i^{(-)})}{\%C_T^{(n+)}(1 - \%C_T^{(n-)})} \quad (3-13)$$

Equation (3-13) indicates that  $x_i^{(n+2)}$  is dependent on  $x_i^{(n)}$  and a growth term. The denominator of the growth term is dependent on the cleavage responses of all the sequences in the pool and their frequencies (refer to Equation (3-9)). The numerator in contrast is sequence specific and provides the basis to estimate the propensity for sequence  $i$  to be enriched, henceforth referred to as its fitness:

$$fitness_i = \%C_i^{(+)}(1 - \%C_i^{(-)}) \equiv K_{(n)} \frac{x_i^{(n+2)}}{x_i^{(n)}} \quad (3-14)$$

where  $K_{(n)} = \%C_T^{(n+)}(1 - \%C_T^{(n-)})$

Analysing Equation (3-14), several important conclusions are realised. Firstly, the fitness of a sequence is the same irrespective of other sequences in the population. This implies the fitness of sequence can be determined empirically for a given set of selection conditions. Secondly, fitness is directly proportional to the fold change of a sequence following a cycle of selection ( $\frac{x_i^{(n+2)}}{x_i^{(n)}}$ ). Therefore, those sequences with the highest fitness values are enriched over other sequences in the pool given enough cycles of selection. Furthermore, if the fitness of two or more sequences are the same, then the fold-change of these sequences will also be the same and as such, relative to one another there would be no change following selection.

### 3.3.2.3 Fitness of all possible sequences

To quantify the fitness for all possible sequences during selection, fitness values were calculated for all combinations of cleavage responses during positive & negative selection according to Equation (3-14). These responses are illustrated in Figure 3-3.

To aid in identifying phenotypes, positive and negative selection conditions were assumed to be identical except for a difference in the concentration of ligand. Under these conditions, ligand-unresponsive sequences have identical % cleaved responses (blue dotted-line – Figure 3-3A). This is

because these sequences have identical cleavage rates regardless of the concentration of ligand. As reviewed in the discussion and proceeding chapters this scenario is not always the case.

Ligand-activated sequences meanwhile are located to the right of this blue dotted-line (% cleaved (+ve) > % cleaved (-ve)), while ligand-inhibited ribozymes are located to the left (% cleaved (-ve) > % cleaved (+ve)). Figure 3-3A illustrates that of all the sequences, those which are maximally activated by ligand have the highest fitness values and as such will be enriched over the remaining sequences (bottom-right quadrant). This is encouraging as the selection method was designed to select for ligand-activated sequences, while discriminating against sequences with alternative phenotypes (refer to Figure 2-3A).

Figure 3-3B further investigates the fitness of the ligand-unresponsive sequences in Figure 3-3A. Several results are illustrated by this data. Firstly, the inactive and constitutively active ribozymes which have low and high % cleaved values, respectively have comparatively low fitness values amongst these sequences. This is presumably because these phenotypes are strongly selected against during positive and negative selection, respectively. Figure 3-3B illustrates that for these unresponsive sequences, a maximum fitness value is obtained when 50 % cleavage occurs in the presence and absence of ligand. This yields a fitness of 0.25. These ligand-unresponsive sequences are henceforth referred to as semi-active sequences to indicate they are neither constitutively active nor inactive but have median cleavage rates.

Figure 3-3A also suggests some ligand-activated sequences are unlikely to be enriched. Specifically, those located in the upper right and lower left quadrants of Figure 3-3A. These sequences have fitness values lower than 0.25 and hence would not be enriched over semi-active sequences. In conclusion, to guarantee the enrichment of a ligand-activated ribozyme under these conditions, it must have a fitness greater than 0.25.

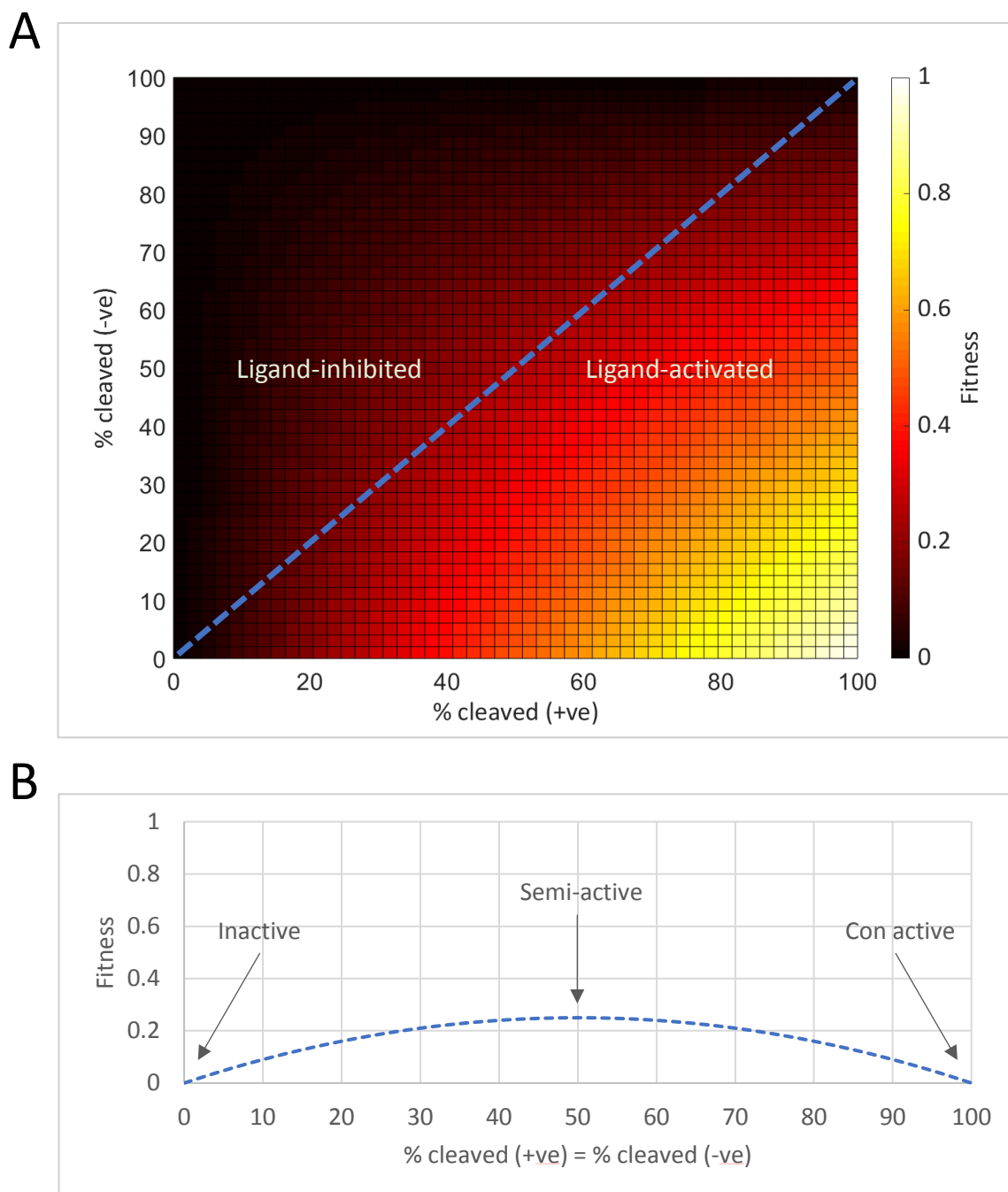


Figure 3-3: Fitness of all sequences – (A) Fitness calculated for all combinations of % cleaved values under positive (% cleaved (+ve)  $\equiv 100\%C_i^{(+)}$ ) and negative (% cleaved (-ve)  $\equiv 100\%C_i^{(-)}$ ) selection conditions according to Equation (3-14). Colour scale indicates fitness. With assumptions in the text, ligand-unresponsive ribozymes lie on the dotted blue-line, while ligand-activated and inhibited sequences are present to the right and left, respectively. (B) Blue dotted-line from (A). Inactive, semi-active and constitutively (Con) active sequences are indicated.

### 3.3.2.4 Optimising RNA preparation

Following the results of the model regarding fitness, selection conditions were optimised to maximise the fitness of the Theophylline Control under positive & negative selection conditions that differed only in the concentration of ligand. This would ensure this sequence is enriched over the other undesirable sequences in the library in as short a time as possible.

To identify variables that would modify fitness, previous studies were consulted. From these examples, a common strategy to vary selection stringency is to alter the library incubation time prior to ribozyme inactivation<sup>55,119,127,148,149</sup>. Taking inspiration from this, the IVT time was varied and the cleavage response of the Theophylline Control measured (Figure 3-4).

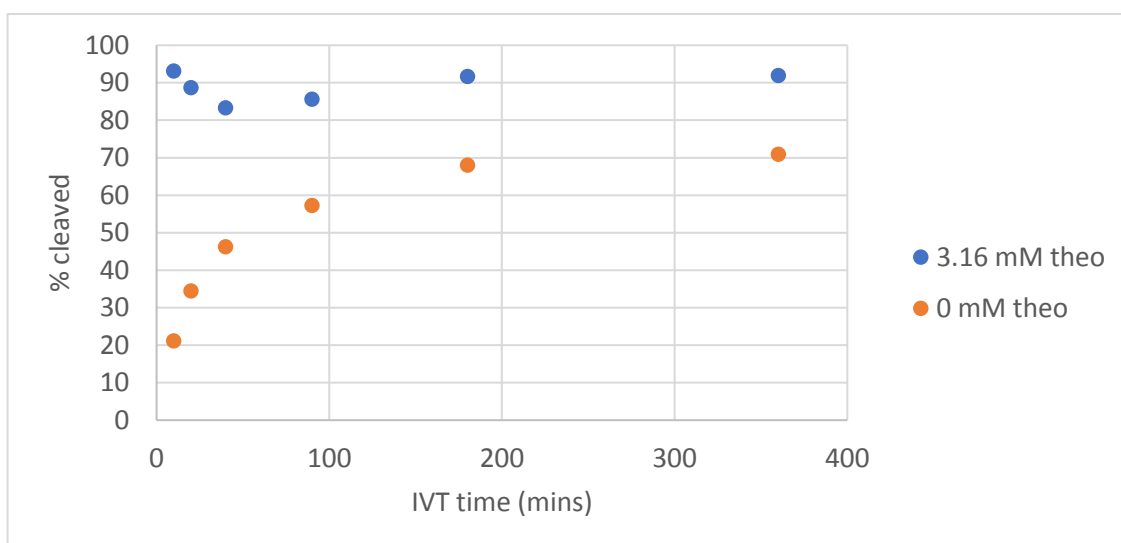
Figure 3-4A illustrates the cleavage kinetics of the Theophylline Control during transcription. Under 0 mM theophylline cleavage increases logarithmically with the IVT time. The response of the control in the presence of 3.16 mM theophylline is more complicated and has previously been described as biphasic<sup>72</sup>, initially decreasing from higher % cleaved values before switching to a logarithmically increasing phase at longer IVT times. This response is further commented on in the discussion (Section 3.4.3).

To further understand the implications of this response, this data was examined as a function of fitness using Equation (3-14) (Figure 3-4B). In line with the conclusions of the previous section, those sequences with higher fitness values will be enriched over sequences with lower fitness values and vice versa. Figure 3-4B illustrates how the fitness of the Theophylline Control and semi-active sequences at the RNA level varies with IVT time. Note, as specified in the previous sections, fitness should be evaluated at the cDNA level, as this is the point at which selection occurs. The analysis of fitness at the cDNA level is conducted in the subsequent section.

Figure 3-4B illustrates that with longer IVT incubations, the fitness of the Control decreases exponentially. Specifically, with an incubation of 180 minutes or more, the control has approximately the same fitness as semi-active sequences previously described (~0.25). This result suggests that under

these conditions the control would not be significantly enriched over these sequences during selection. As such should these sequences be present or evolve from the Theophylline Library, the selection experiment would fail. In conclusion, this data suggests that to guarantee the enrichment of the Theophylline Control, a short IVT time must be used.

A



B

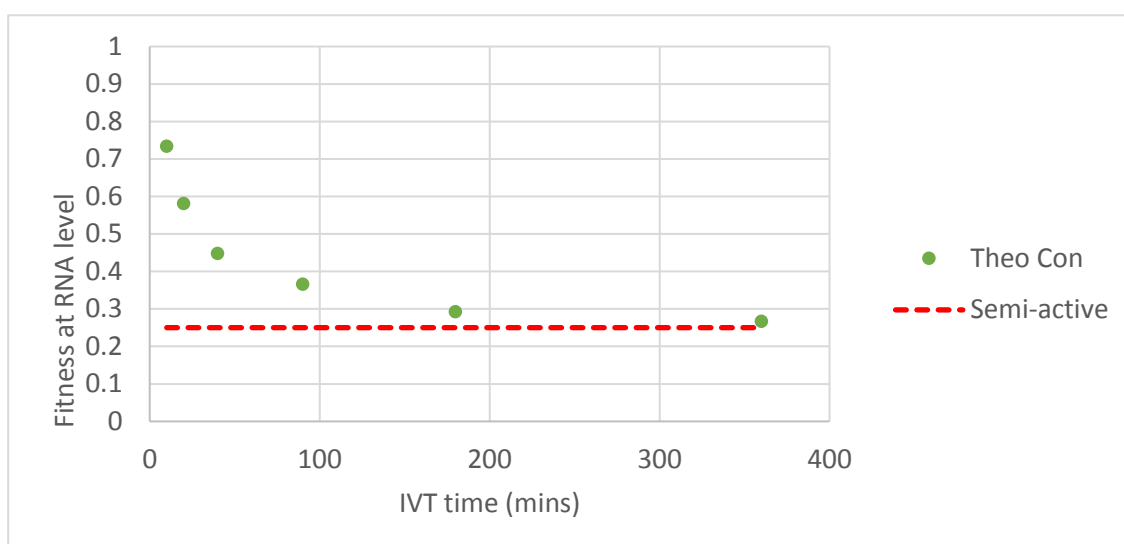


Figure 3-4: Kinetics and fitness of the Theophylline Control during IVT – (A) % cleaved values for the Theophylline Control were calculated for variable IVT times in presence of 3.16 or 0 mM Theophylline. IVT reactions were not purified prior to electrophoresis. (B) Fitness at the RNA level as a function of IVT time was calculated according to Equation (3-14) using the values in (A). Red dotted-line indicates the fitness of ligand-unresponsive, semi-active sequences.

### 3.3.2.5 Optimising cDNA preparation

It is important to monitor the fitness of the Theophylline Control at the cDNA level. This is because selection occurs following reverse transcription and as such changes in the fitness of a sequence may occur between the end of transcription and the beginning of this stage.

Following the conclusion that a short IVT time should be used, the Theophylline Control was *in vitro* transcribed for 10 minutes rather than 180 minutes. The resulting RNA was purified and reverse transcribed, generating cDNA for analysis (Figure 3-5). Figure 3-5 illustrates that the Theophylline Control cDNA has a significantly higher % cleaved value than that of the RNA; particularly under 0 mM theophylline where this value increases from 39.8 to 85.2 %. This result is likely attributed to cleavage during RNA purification and reverse transcription steps. A consequence of this cleavage is that the fitness of the Theophylline Control is reduced below 0.25, suggesting under these conditions the control is unlikely to be enriched. It can therefore be concluded that in addition to shortening IVT times, modifications to cDNA preparation are also required.

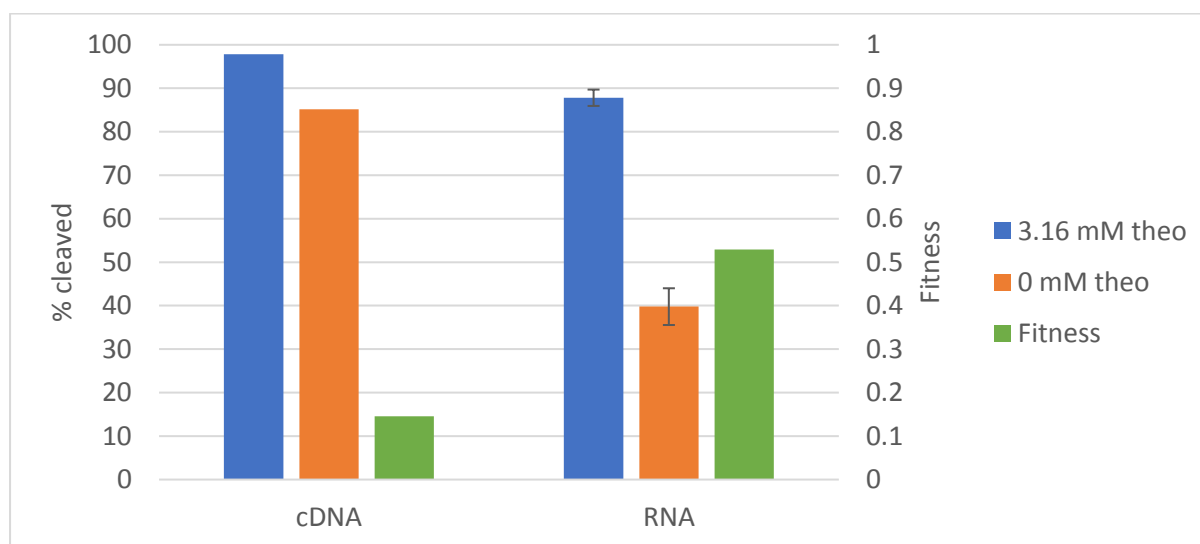


*Figure 3-5: Fitness at the RNA vs cDNA level – Theophylline Control % cleaved and fitness values are calculated from unpurified IVT reactions incubated for 10 minutes (RNA), or subsequently following cDNA generation (cDNA). Both RNA & cDNA samples were generated in the presence of either 3.16 or 0 mM theophylline.*

Two alternative methods for cDNA synthesis were conceived to improve the fitness of the Theophylline Control. The first method involves avoiding RNA purification and is described in detail in the Materials & methods chapter (Section 7.5.1). By avoiding RNA purification, there is less time for the Theophylline Control to cleave. Using this method, it was hypothesised that the proportion of cleaved cDNA would be similar to that present at the RNA-level, where fitness was much higher.

The second alternative method for cDNA synthesis involves conducting the IVT and RT reactions simultaneously (TRT reaction). Figure 3-6 illustrates the principle of the TRT reaction. During the TRT reaction, full-length RNA is inactivated by reverse transcription. Inactivation results from the interaction between full-length RNA and its cDNA complement which prevents the formation of the active conformation by trapping the ribozyme in an RNA:cDNA duplex. This results in the quenching of ribozyme catalysis in the time it takes to synthesise full-length cDNA. Given RT reaction times are relatively fast<sup>59</sup>, the TRT reaction was hypothesised to improve the fitness of the Theophylline Control, in line with the result in Figure 3-4B.

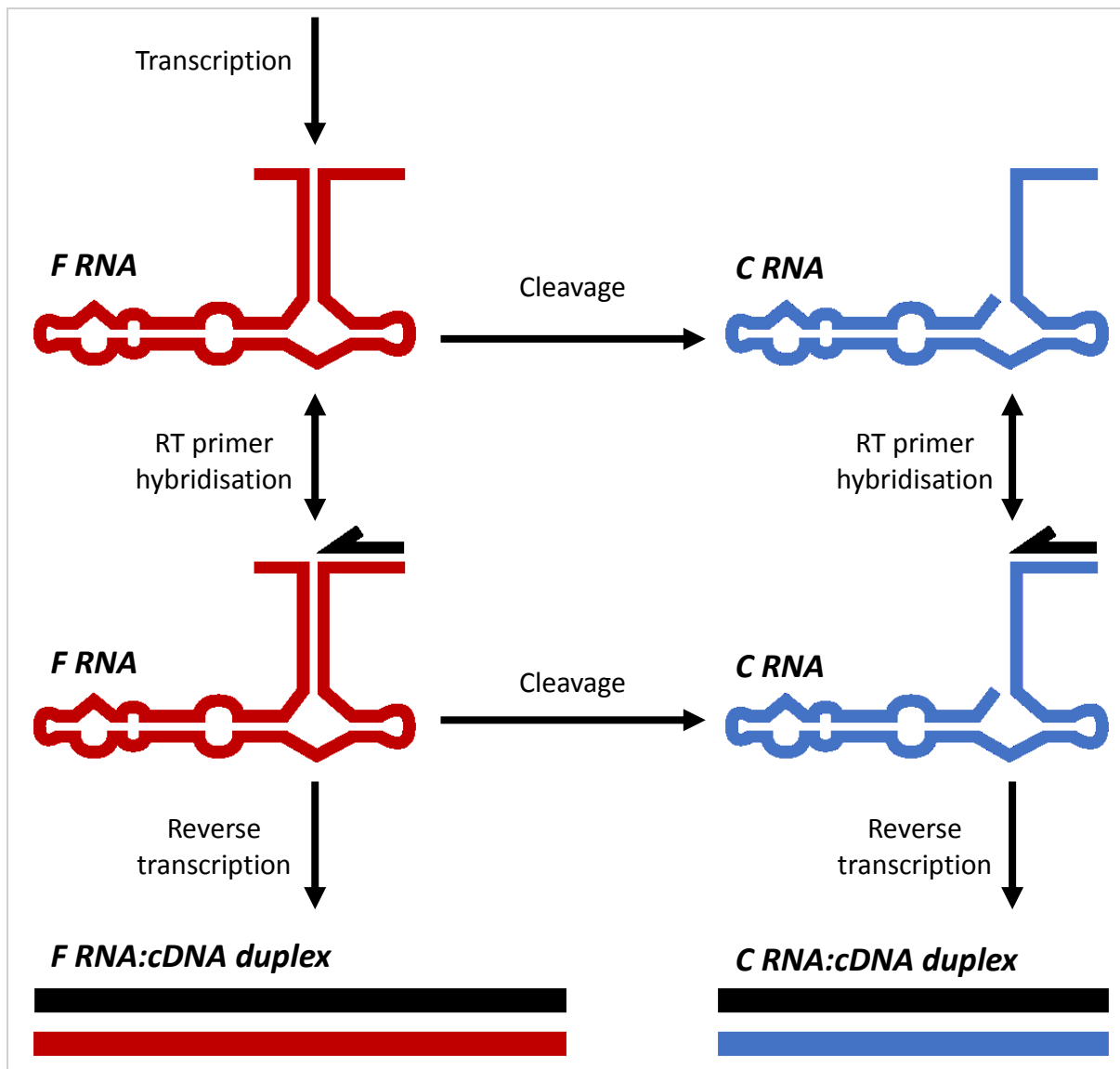


Figure 3-6: TRT reaction – Following transcription of full-length (F) RNA, F RNA is free to cleave producing cleaved (C) RNA or reversibly hybridise with the RT primer. Following RT primer hybridisation, reverse transcription can proceed inactivating F RNA in an RNA:cDNA duplex. Reverse transcription of C RNA also produces an RNA:cDNA duplex.

From Figure 3-7, the fitness of the Theophylline Control is significantly improved using the alternative methods of cDNA preparation. Of the two alternative methods for cDNA synthesis, the TRT reaction yields the highest improvement in fitness, 0.77 vs 0.44. As per the model of selection, those sequences with the largest fitness values are enriched over all other sequences. As such, use of the TRT reaction

instead of simply avoiding RNA purification would maximise the probability that the Theophylline Control is enriched and subsequently identified, achieving the main aim of this chapter.

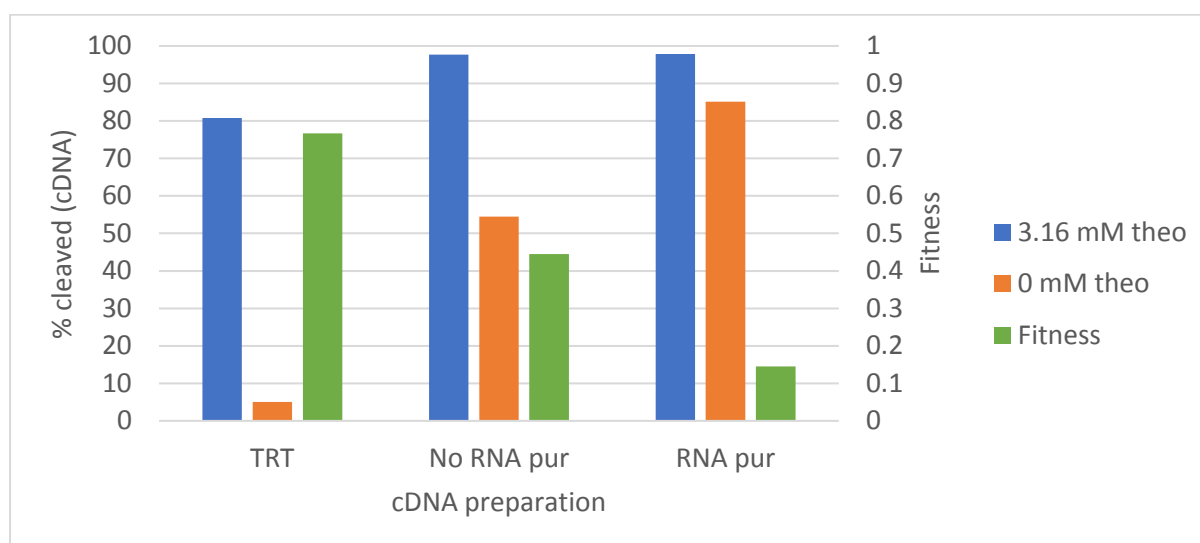
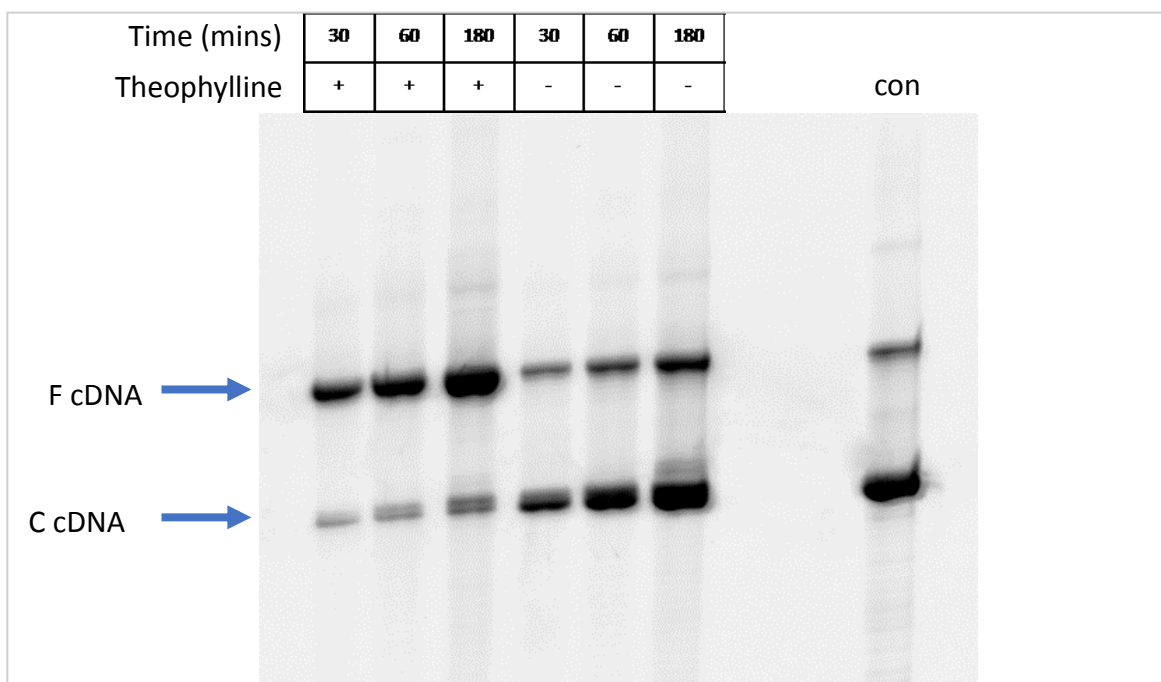


*Figure 3-7: Comparison of methods used to prepare cDNA – Theophylline Control cDNA was prepared in the presence of 0 or 3.16 mM theophylline using the TRT reaction, without RNA purification (No RNA pur) or with the previously used method (RNA pur). % cleaved and fitness values were calculated.*

Figure 3-8A illustrates an additional advantage of the TRT reaction is that high yields of cDNA can be synthesised using this method without compromising fitness. Specifically, similar or greater quantities of cDNA than those achieved using the manufacturer's recommended conditions were yielded by incubating the TRT reaction for 180 mins. As previously mentioned, achieving good yields of cDNA is important for maximising throughput as selection is conducted at the cDNA level. Figure 3-8B illustrates that fitness is maintained, even under long incubation times as the % cleaved values remain approximately constant.

A



B

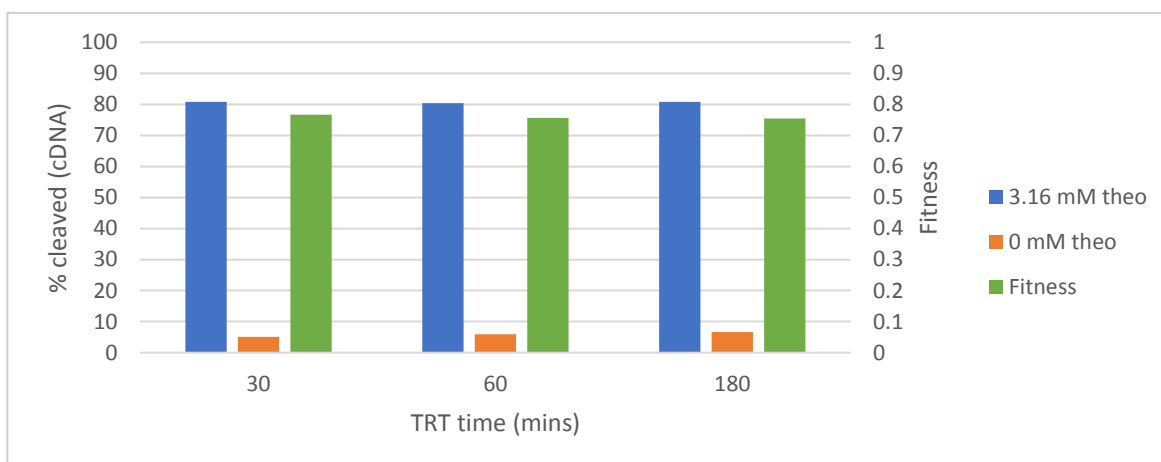


Figure 3-8: TRT reaction produces high yields and a constant response – (A) Urea PAGE gel containing TRT reactions incubated for 30, 60 or 180 mins under 0 (-) or 3.16 mM (+) theophylline. The product of a RT reaction containing 100 ng/ $\mu$ L purified RNA is illustrated for comparison (con). Full-length (F) & cleaved (C) cDNA species are indicated. Gel imaged for HEX-cDNA fluorescence (B). % cleaved and fitness values are calculated for each of the time points.

### 3.3.2.6 Number of selection cycles

Prior to implementing selection, one final question of interest was exactly how many cycles of selection would be required to enrich the Theophylline Control to detectable levels? This question was important given the selection experiment should be completed in a practical period of time. Ultimately, to detect the Theophylline Control, its sequence would need to be identified within the pool. If the Theophylline Control was enriched to 50 % of the pool, on average, half of all randomly picked sequences would be this Control sequence. As such, upon reaching a frequency of 50 %, it should be possible to readily detect the Theophylline Control by picking only a handful of sequences. To estimate the number of cycles required to reach a frequency of 50 %, the model developed in Section 3.3.2.1 can be used if several parameters relating to the Theophylline Control and Theophylline Library are calculated (refer to Appendix, Section 9.1, Table 9-1 for parameters).

As outlined previously, simulations generated by this model are based on several assumptions. One of the key assumptions is that all the remaining sequences during selection except the Theophylline Control have the same phenotype. On the face of it, this approximation seems incorrect given that the phenotypes of inactive and constitutively active sequences are significantly different. Note the Theophylline Library was designed specifically so that these sequences were present. That said, if all these sequences are diluted out at the same rate during selection, the average of their phenotype will remain the same. In this way they can all be assumed to have the same average phenotype, calculated in Equation (3-10). Given inactive and constitutively active sequences have equally low fitness values, they should be diluted out during selection at approximately the same rate<sup>\*8</sup>. As such, it was anticipated that this assumption should still yield a relatively accurate account of the Theophylline Control dynamics during selection.

Figure 3-9 illustrates the estimated dynamics of the Theophylline Control during selection. This data suggests that the Theophylline Control can be enriched to > 50 % of the library in 4.5 cycles. Given a

---

<sup>\*8</sup>Refer to Figure 3-3B for an illustration of the fitness' of these undesirable sequences.

cycle of selection can be implemented manually every 2 days, this represented a feasible timescale to test the functionality of the method.

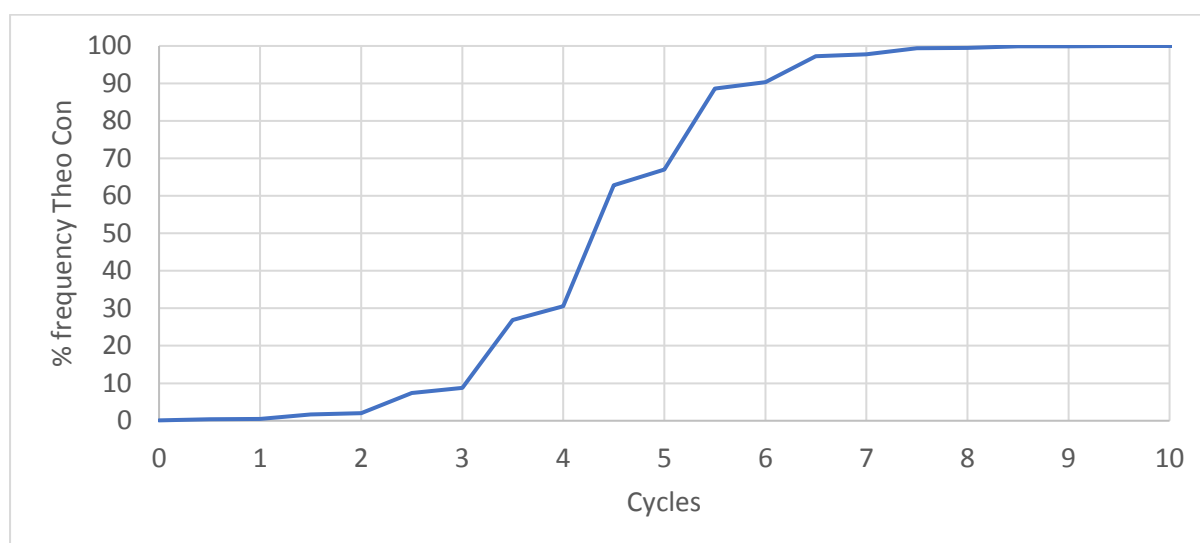


*Figure 3-9: Simulated enrichment of the Theophylline Control from the Theophylline Library – The % frequency of the Theophylline Control (Theo Con) in the Theophylline Library is plotted as a function of selection Cycles. This simulation was generated as described in 3.3.2.1 using the parameters given in the Appendix (Section 9.1, Table 9-1).*

### 3.3.3 Objective 3 – Enrichment of a ligand-activated phenotype

#### 3.3.3.1 1st attempt

To enrich for the Theophylline Control and hence a ligand-activated phenotype, the Theophylline Library was subjected to 4.5 cycles of selection. Each cycle involved 1 round of positive selection followed by 1 round of negative selection. During each round, cDNA was generated using the TRT reaction and full-length or cleaved cDNA selected using DNA-ligase mediated selection.

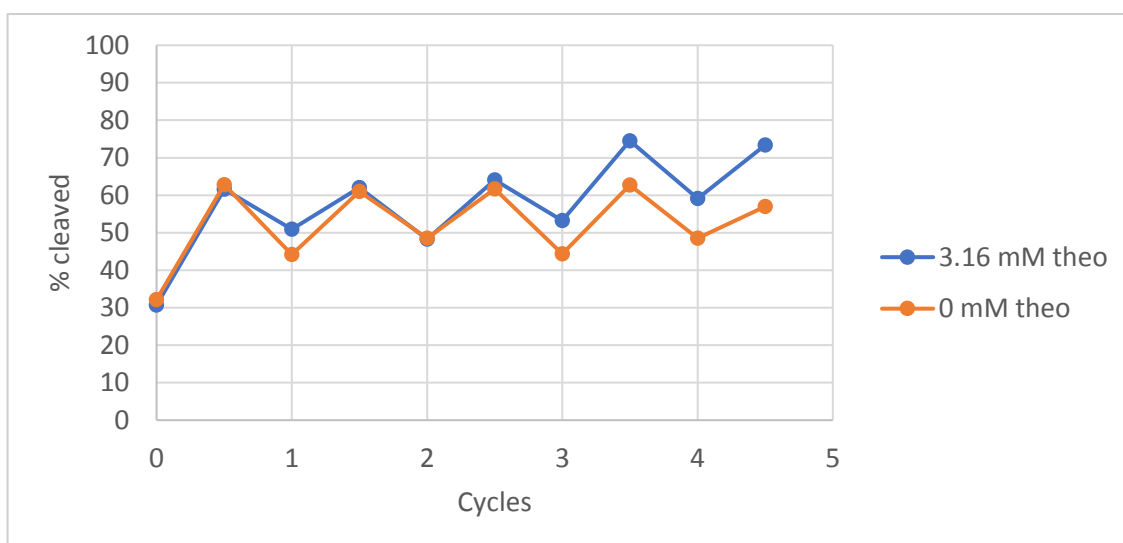
To monitor changes in the phenotype of the pool, DNA following each round of selection was *in vitro* transcribed in the presence of 0 and 3.16 mM theophylline for 10 minutes. The % cleaved values of the resulting RNA were calculated (Figure 3-10A).

Figure 3-10A suggests that both positive & negative selection were functional. This is characterised by an increase in the pool % cleaved value after the first round of a cycle which corresponds to positive selection, followed by a decrease in % cleaved during the subsequent negative selection. For instance, following the positive selection round of the 1<sup>st</sup> cycle, the % cleaved of the population increased from approximately 30 % to 60 % before decreasing again during negative selection.

Additionally, the general trend from Figure 3-10A is that increased cycles of selection increase the response of the pool to Theophylline suggesting that a Theophylline-activated phenotype was enriched. This is characterised by an increasing deviation in the % cleaved values of the pool in the presence and absence of theophylline. For instance, after 4.5 cycles the % cleaved of the pool in the presence of 0 and 3.16 mM theophylline was 57.0 % and 73.4 %, respectively compared to no difference prior to selection.

Furthermore, following 4.5 cycles of selection, the cleavage of the pool in 3.16 mM theophylline had increased to a value similar to the Theophylline Control. Specifically, the control cleaves to 87.8 % under these conditions compared to the Theophylline Library which cleaves to 30.7 %. However, the cleavage of the pool after 4.5 cycles in the absence of theophylline has increased beyond that of the Theophylline Control. This suggests that in addition to enriching for the Control, other sequences with high background cleavage rates have also been enriched.

A



B

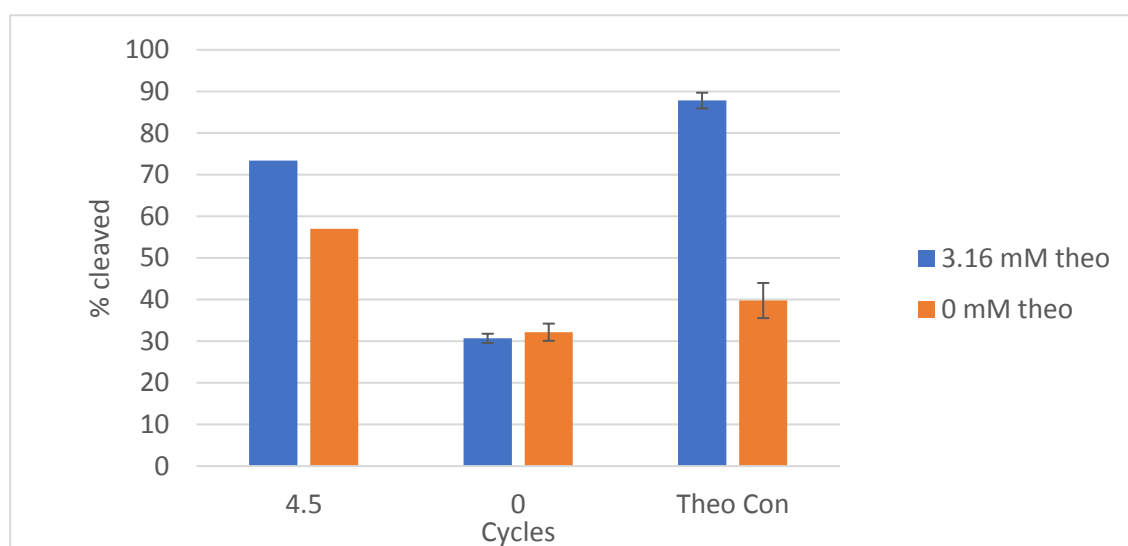


Figure 3-10: 1<sup>st</sup> attempt to enrich a Theophylline-activated phenotype from the Theophylline Library – (A) After each round of selection, the pool was in vitro transcribed for 10 minutes in the presence of 0 or 3.16 mM theophylline and the resulting RNA % cleaved values calculated. (B) % cleaved of the pool after 4.5 cycles is compared with the Theophylline Control & Theophylline Library (0 cycles). Where present, error bars indicate SDs from 4 repeats.

Figure 3-11 illustrates a further complication caused by the process of selection. Specifically, following the 2<sup>nd</sup> cycle of selection, there is a significant increase in the number of non-specific, high MW bands.

At a first approximation, none of these bands appeared to show any response to theophylline (data not shown). Therefore, mutational biases<sup>174</sup> caused by the selection method are the most likely reason for their appearance. Such dominant mutational biases are unlikely to be desirable as sequences are rapidly diluted out of the pool, limiting the number of cycles sequences they are exposed to. This would function to limit overall enrichment.

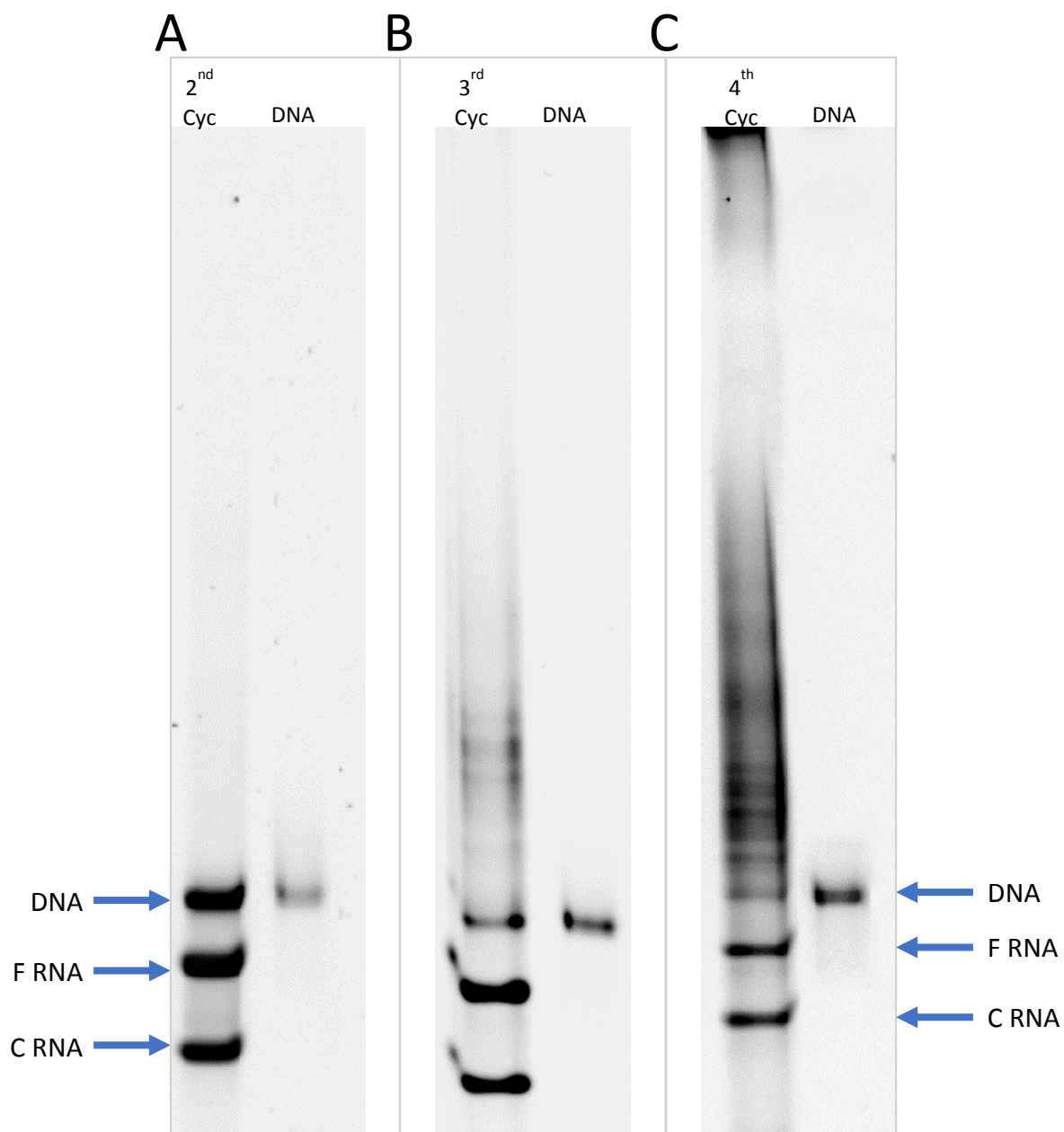


*Figure 3-11: Library instability caused by selection – Urea PAGE gels loaded with the response of the pool in the presence of 3.16 mM theophylline following (A) 2, (B) 3 or (C) 4 cycles of selection. The DNA template encoding the Theophylline Library is loaded for comparison. Full-length (F) & cleaved (C) RNA species are indicated.*

### 3.3.3.2 Improving library stability and selection efficiency

To address the issue of library stability and to generally improve selection efficiency, three modifications to the method were made.

#### 3.3.3.2.1 *Tth* DNA ligase

The first modification to the method aimed at avoiding concatemers during cDNA ligation. Concatemerization during this step would produce longer products and as such could be responsible for the observed library instability. While in principle the T7 DNA ligase shouldn't catalyse the formation of concatemers under the reaction conditions, its specificity is effected by the [PEG]<sup>147</sup>. As such variations in this factor or perhaps background activity could be responsible for the non-specific products observed. To avoid concatemers, T7 DNA ligase was substituted with the *Tth* DNA ligase. In contrast to the T7 DNA ligase, the *Tth* DNA ligase is capable of discriminating against single nucleotide mismatches at the ligation junction under a range of conditions<sup>175</sup>. This makes it a more robust DNA ligase. Furthermore, the *Tth* DNA ligase is a thermostable enzyme meaning the ligation reaction is conducted at a higher temperature inhibiting non-specific annealing.

#### 3.3.3.2.2 Semi-quantitative PCR

Another possible cause of the observed library instability is overamplification during the PCR step. Overamplification of variable oligonucleotide libraries has been shown to encourage the production of product-product hybrids<sup>176</sup>. Given the PCR step was implemented for a fixed number of 20 cycles throughout the previous experiment, it is feasible that product-product hybridisation occurred during this step.

To prevent overamplification and the production of product-product hybrids, PCR should be stopped prior to any significant depletion in PCR primer concentration<sup>46</sup>. However, given the dynamics of PCR, if too few PCR cycles are conducted the yield of this step is severely affected. This reduces the throughput & makes analysis difficult due to the lack of material. Therefore, an optimal number of PCR cycles should be implemented after each round of selection.

To implement the optimum number of PCR cycles, inspiration was taken from Klussmann and co-workers<sup>129</sup>. In their study semi-quantitative PCR was used to prevent overamplification. This involved

pausing PCR after a given number of cycles and quantifying the yield of DNA. If the yield had passed a given threshold, PCR was stopped and the reaction purified.

While the previously proposed semi-quantitative PCR was successful in its purpose, DNA concentrations had to be measured up to four times during thermocycling. Furthermore, the actual number of cycles could differ from the optimal number by up to two. This is because measurements were taken every 3 cycles. It was reasoned that by considering the kinetics of the PCR<sup>177</sup>, a more accurate method could be developed where fewer measurements were required. To this extent the following algorithm was developed to implement such a process:

- I. Conduct a fixed number of cycles ensuring the DNA yield is significantly below the optimal yield.
- II. If the DNA yield is quantifiable, calculate and implement the remaining number of cycles to achieve the maximum DNA yield without overamplification.
- III. Else if the DNA yield is below the LOD, conduct as many cycles as possible without exceeding the optimal DNA yield. Take a further measurement and repeat this process from step II.

To implement the above algorithm several parameters are required. The first of these is the initial number of cycles conducted during step I. The LOD and the optimal DNA yield are also required in addition to the PCR efficiency ( $Eff$ ) which relates the yield ( $DNA$ ) to the number of cycles ( $C$ ) through the following equation<sup>177</sup>:

$$DNA = DNA_0 Eff^C \quad (3-15)$$

Where  $DNA_0$  is the initial [DNA]. Rearrangement of Equation (3-15) yields:

$$\frac{\log(DNA)}{C} = \frac{\log(DNA_0)}{C} + \log(Eff) \quad (3-16)$$

By varying  $DNA_0$  and measuring the yield of  $DNA$  after a fixed number of cycles key parameters of the PCR step can be determined (Figure 3-12).

The optimal DNA yield is indicated in Figure 3-12. Yields higher than the optimal [DNA] are beyond the linear relationship between  $\frac{\log(DNA)}{c}$  and  $\frac{\log(DNA_0)}{c}$ . Including these data points when determining this linear relationship would result in a reduction in the calculated PCR efficiency, characterised by a lower y-axis intercept (Equation (3-16)). A reduction in PCR efficiency indicates incomplete template:primer hybridisation during these cycles<sup>46</sup>. As such, conducting PCR cycles beyond the optimal [DNA] runs the risk of product-product hybridisation and hence the formation of non-specific products.

The number of cycles required to implement step II above can be calculated If the [DNA] is above the LOD. This is because amplification proceeds linearly up to the optimal [DNA] in this region (Figure 3-12). The exact number of cycles required to implement this step of the algorithm can be calculated by rearranging Equation (3-16). The number of cycles required during step III can be calculated assuming the [DNA] = LOD. Using this number of cycles means the yield of DNA would not exceed the optimum. In this way the above algorithm was implemented as summarised in the Materials & methods Chapter, Section 7.2.9.

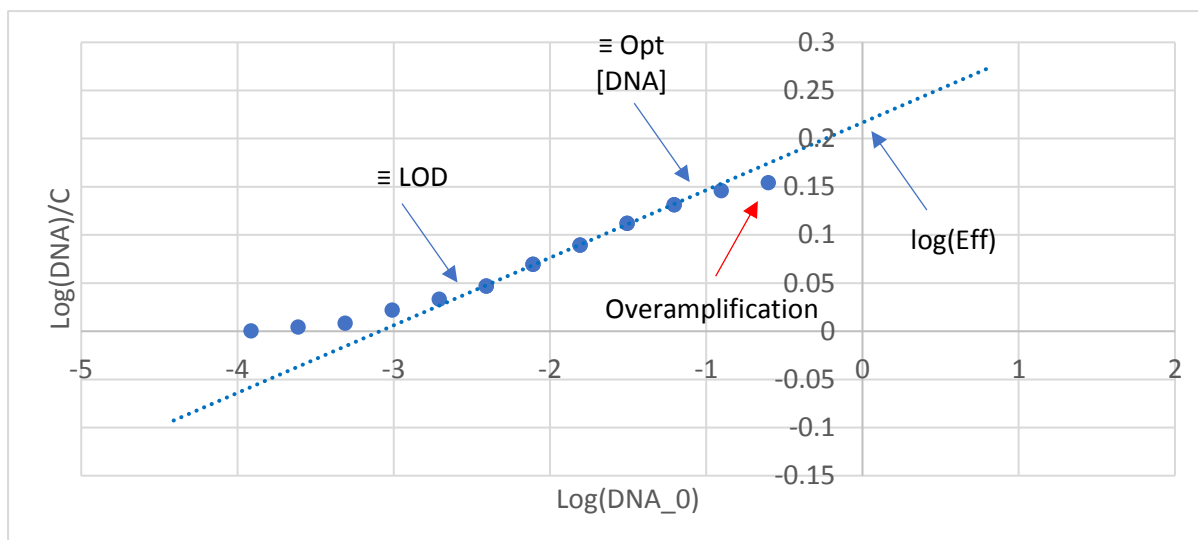


Figure 3-12: Improved semi-quantitative PCR optimisation – Varying concentrations of Theophylline Library DNA were amplified under selection conditions and the resulting [DNA] plotted according to Equation (3-16). A line of best fit was fitted to the linear portion of the results ( $R^2 = 0.9993$ ). The equivalent ( $\equiv$ ) LOD and optimal (opt) [DNA], in addition to the logarithm of PCR efficiency (Eff) are indicated by blue arrows. A data point corresponding to overamplification is indicated by a red arrow.

To determine the functionality of the improved semi-quantitative PCR, the yield of all 52 semi-quantitative PCRs conducted during this work was analysed (Figure 3-13). The optimal [DNA] value calculated from Figure 3-12 was 25.1 ng/ $\mu$ L. Based on the above algorithm, the yield of DNA should be as close to this value without exceeding it. Figure 3-13 illustrates that this was the case as the distribution of yields are centred around 19 – 22 ng/ $\mu$ L.

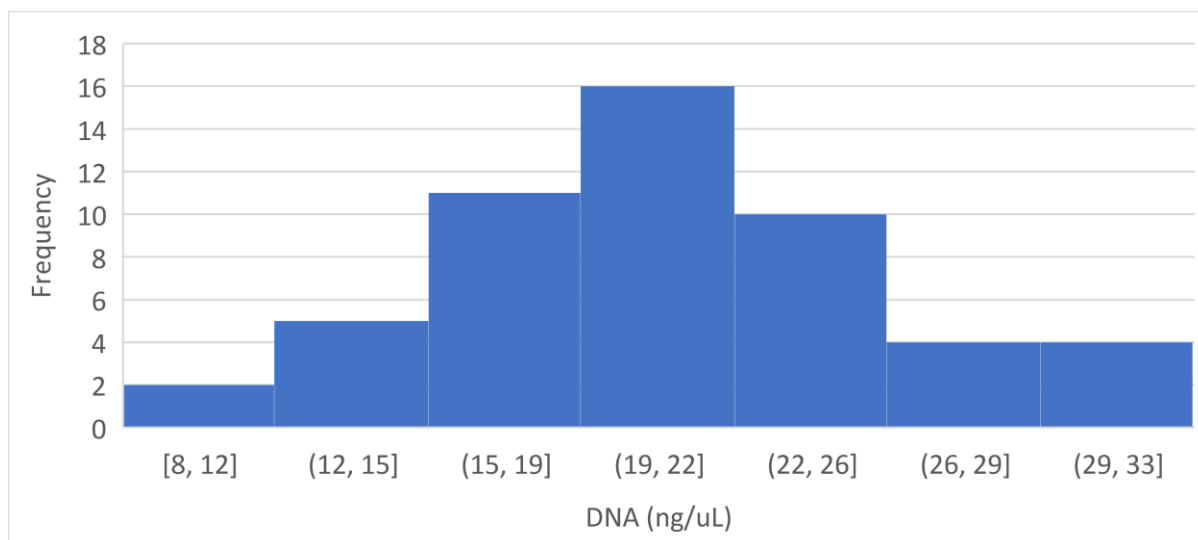


*Figure 3-13: Effectiveness of improved semi-quantitative PCR – Histogram summarising the yields from 52 PCR reactions.*

#### 3.3.3.2.3 Optimising DNA template removal

The above measures primarily aimed at preventing non-specific product formation improving library stability. The last modification primarily aimed at improving selection efficiency. With improved selection efficiency fewer cycles would be required to enrich the control to a desired level. This would also have a positive effect on library stability as fewer steps which introduce mutational biases would be conducted.

At the end of the TRT reaction, RNA, cDNA and the DNA template are present in the reaction mix. By removing the DNA template prior to cDNA selection, the efficiency of the method was improved (data not shown). This was presumably because the remaining DNA template was co-purified to some degree during cDNA selection. To remove the DNA template a pull-down reaction was used. To enable the pull-down, the primers used during PCR were 5' biotinylated. This allows the immobilisation of the biotinylated template on streptavidin-coated paramagnetic beads and its subsequent removal following the TRT reaction. While this improved selection efficiency, any DNA template which remains is biotinylated meaning it could be co-purified with the biotinylated cDNA ligation product during selection.

A pertinent question therefore, was could more DNA template be removed, improving selection efficiency. It was hypothesised that more template could be removed by using more pull-downs. to determine if this was the case, DNA template was incubated under TRT reaction conditions. The template was then removed as previously described using 1 pull-down and 10  $\mu\text{L}$  of paramagnetic beads or 3 pull-downs each using 5  $\mu\text{L}$  of beads. To determine the success of each methodology, any remaining DNA template was first immobilised on 10  $\mu\text{L}$  streptavidin-coated paramagnetic beads and washed as described in the Materials & methods chapter (latter part of Section 7.2.6.3). The DNA on these beads was then amplified for 19 cycles during PCR and the resulting yield determined (Figure 3-14). This process is equivalent to how carryover would occur, enabling it to be quantified.

Figure 3-14 illustrates that by using 3 pull-downs each with 5  $\mu\text{L}$  of beads, the [DNA] from the PCR was equivalent to the control which contained water. In contrast using the previous method, statistically higher yields of DNA were achieved compared to the control. This is indicative of carryover. Therefore using 3 pull-downs instead of 1 would reduce carryover improving selection efficiency.

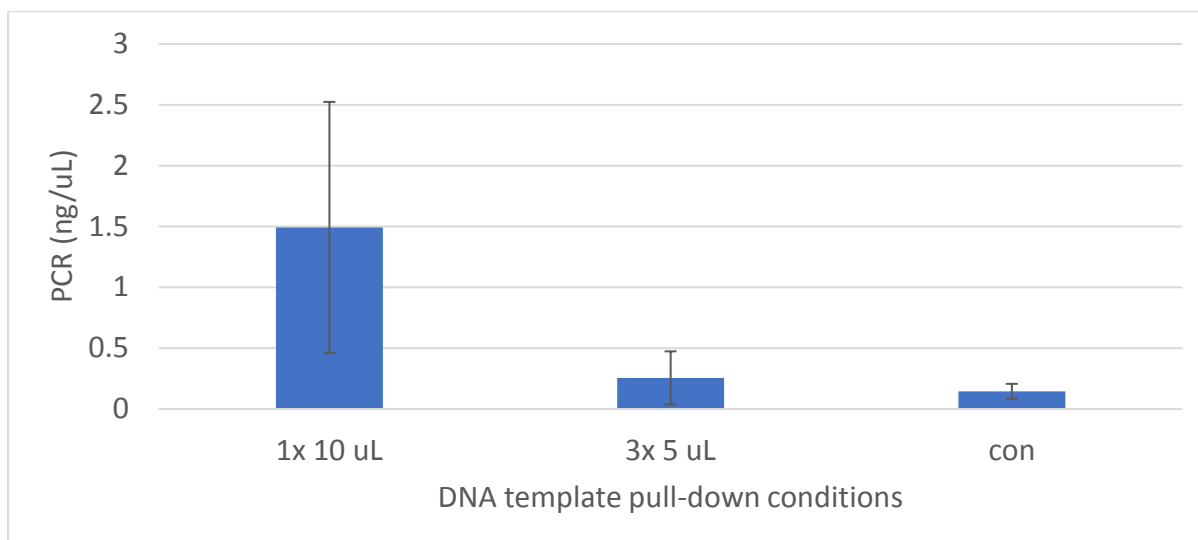


*Figure 3-14: Optimising DNA template removal – Biotinylated DNA template was incubated under TRT reaction conditions. The template was removed using either 1 pull-down and 10  $\mu$ L of strep-coated paramagnetic beads or 3 pull-downs each using 5  $\mu$ L of beads. Any remaining DNA template was immobilised, washed and amplified for 19 cycles. The resulting yield was calculated and compared to the case where the DNA template was replaced with water (con). Error bars indicate the SD of 2 repeats.*

### 3.3.3.3 2nd attempt

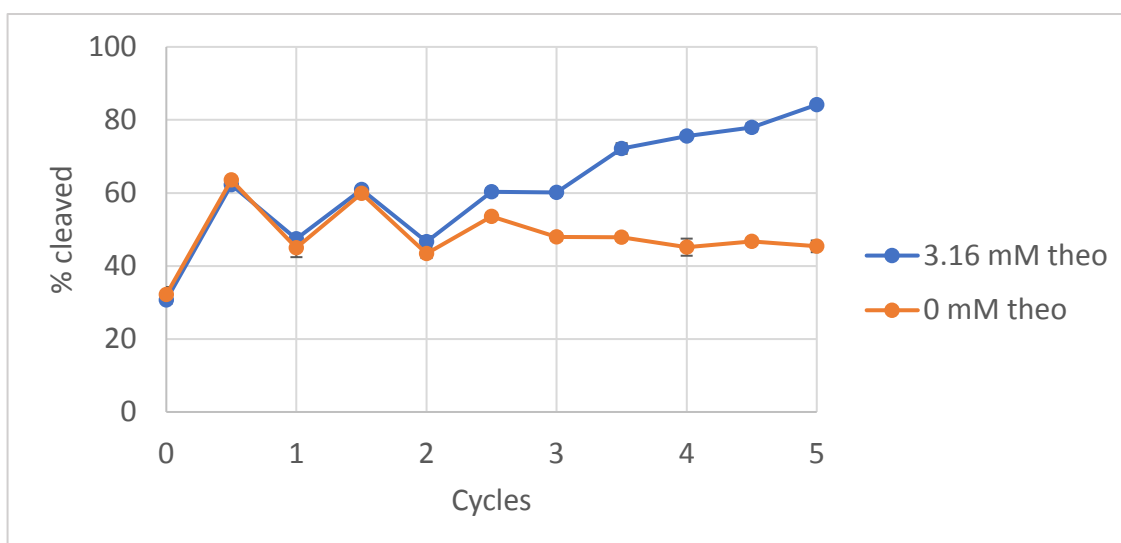
The Theophylline Library was again subjected to 5 cycles of selection. This time, to improve selection efficiency and library stability the 3 above modifications were implemented. Enrichment of a Theophylline-activated phenotype was monitored as described for the 1<sup>st</sup> attempt in Figure 3-10.

Similar to the 1<sup>st</sup> attempt, both positive & negative selection rounds appear functional and a theophylline-activated phenotype is enriched (Figure 3-15A). However, the magnitude of the change is greater during this 2<sup>nd</sup> attempt. For instance, after 4.5 cycles the response of the pool is 78 and 45.4 % in the presence and absence of theophylline, respectively compared to 73.4 and 57 % during the 1<sup>st</sup> attempt. This suggests the additional modifications to the selection method improved its efficiency.

Comparing the response of the pool after 5 cycles to that of the control the response is almost identical, differing by only a few percentage points (Figure 3-15B). For instance, the response of the

pool after 5 cycles is 84.2 and 45.4 % in the presence of 3.16 & 0 mM theophylline, respectively compared to 87.8 and 39.8 % for the Theophylline Control. This suggests that the control or a sequence with similar functionality has been significantly enriched during selection.

A



B

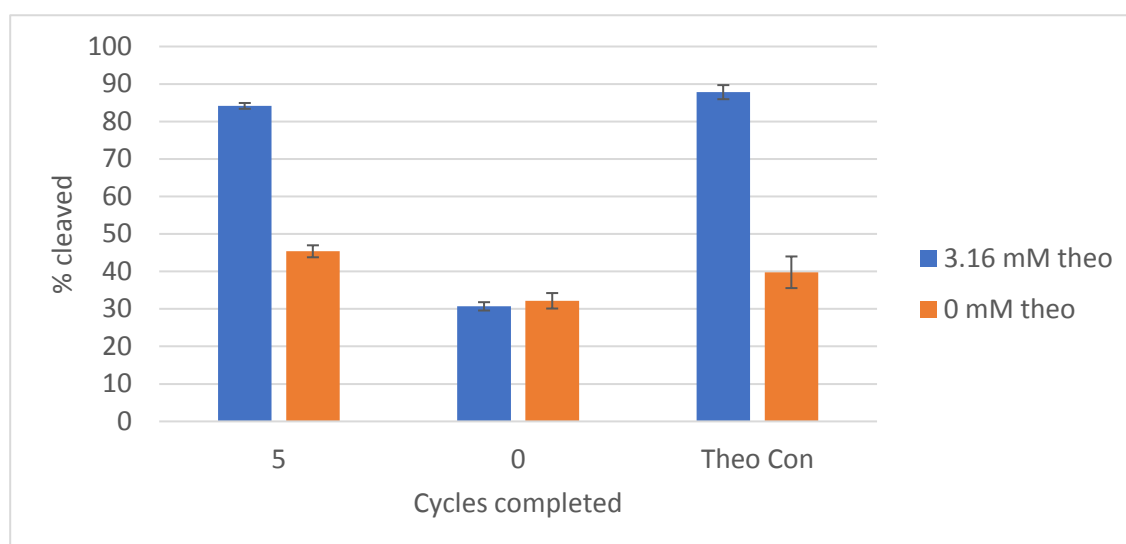


Figure 3-15: 2<sup>nd</sup> attempt to enrich a Theophylline-activated phenotype from the Theophylline Library – (A) After each round of selection, the pool was *in vitro* transcribed for 10 minutes in the presence of 0 or 3.16 mM theophylline and the resulting RNA % cleaved values calculated. (B) % cleaved of the pool after 0 and 5 cycles is illustrated along with % cleaved values for the Theophylline Control. For both (A) & (B) error bars indicate SDs from 4 repeats.

To determine if the modifications to the selection method improved library stability, DNA following every round of selection was *in vitro* transcribed in the presence 0 & 3.16 mM theophylline and the

reaction analysed using electrophoresis (Figure 3-16). In contrast with Figure 3-11, all IVT reactions analysed in Figure 3-16 have 3 distinct bands corresponding to the DNA template, full-length RNA and cleaved RNA. Therefore, it can be concluded that the modifications to the method improved library stability by reducing the formation of non-specific bands.

Furthermore, from Figure 3-16 there is a clear similarity in appearance of the IVT following the 5<sup>th</sup> cycle when compared to that of the Theophylline Control. This supports the view that the control or a sequence with similar functionality has been significantly enriched during selection.

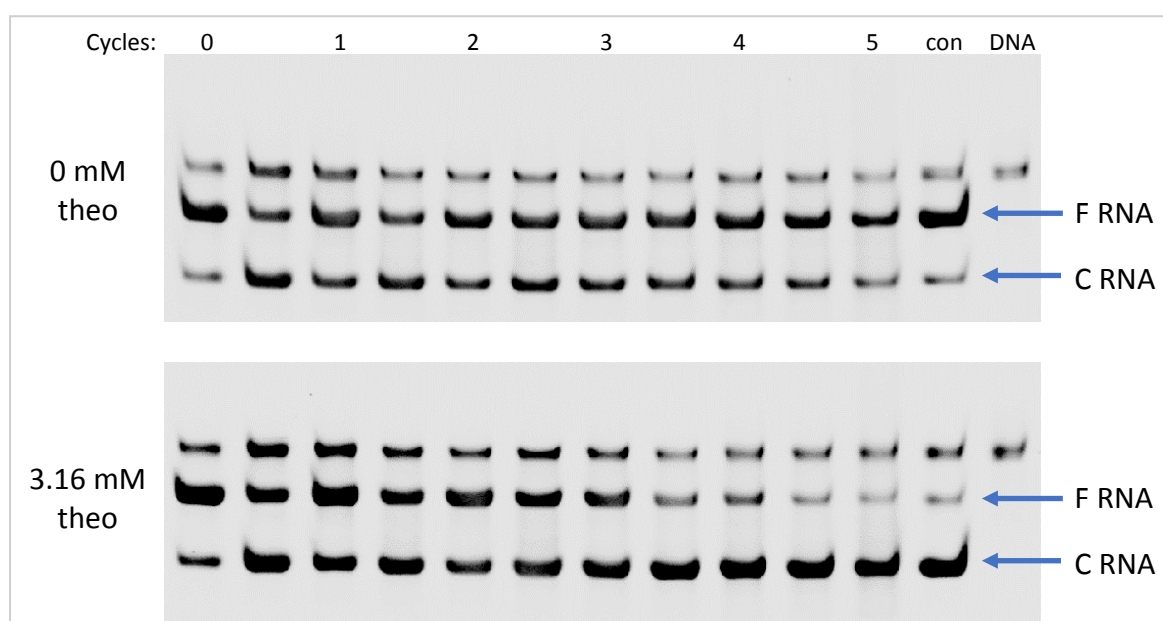


*Figure 3-16: Response and stability of the pool during selection – DNA following each round of selection was in vitro transcribed in 0 (top) and 3.16 (bottom) mM theophylline. The IVT reactions were analysed using urea PAGE. The Theophylline Control (Con) and Theophylline Library DNA (DNA) are loaded for comparison. Full-length (F) and cleaved (C) RNA species are indicated.*

### 3.3.4 Objective 4 – Enrichment of a theophylline-activated genotype

The change in the phenotypic response of the pool strongly suggests that the Theophylline Control or similar sequences were enriched from the Theophylline Library. To validate this, variants from the

pool before and after selection were sequenced\*<sup>9</sup>. To maximise the chances of observing Theophylline Library sequences, the pool before and after selection was gel purified. Gel purification would lower the chances of observing excessively short or long sequences that based on the result in Figure 3-16 should have appeared by chance. This was important considering the identity of only 10 sequences following selection would be determined. In addition to picking 10 sequences at random from the end of the selection experiment, 10 sequences at random were picked from the initial Theophylline Library. This was to ensure biases in DNA synthesis or other processes did not severely affect this selection experiment.

Based on the predicted dynamics of the Theophylline Control in Figure 3-9, it was expected that > 50 % of the sequences assayed from the pool would correspond with the Theophylline Control. Table 3-1 illustrates that 30 % of the sequences picked were the Theophylline Control, which is close to the predicted value.

In addition to the Theophylline Control several other sequences were identified. Some of these sequences were similar to the Theophylline Control. For instance, 20 % of clones sequenced contained a Theophylline Library variant which differed from the Theophylline Control by a single base substitution at position 63. Specifically, this sequence contained an C63A substitution and as such is referred to as the C63A variant of the Theophylline Control.

While some sequences were similar, others contain mutations at significant loci. For instance, G10d contains a substitution mutation at position 73 which is otherwise conserved. This sequence also has a 3-base deletion in a region involved in making 3° interactions which stabilise the active conformation of the ribozyme<sup>63,72</sup>. Another example is G10e which is the only clone to have a substitution mutation at position 52 compared to the control. As described and mentioned in Figure 3-17F, the C residue at position 52 makes 3 direct hydrogen bonds with theophylline. This sequence also contains a A46G substitution in the theophylline-binding domain. Therefore, if G10d and G10e are theophylline-

---

\*<sup>9</sup>Refer to the Materials & methods chapter, Section 7.5.2 for a description of the methodology.

responsive ribozymes, they must function using alternative communication and ligand binding mechanisms compared to the Theophylline Control.

| ID           | % Frequency | Identity at randomised pos. |    |    |    |    | Additional mutations |
|--------------|-------------|-----------------------------|----|----|----|----|----------------------|
|              |             | 52                          | 63 | 73 | 74 | 81 |                      |
| Theo Con     | 30          | C                           | C  | U  | C  | G  |                      |
| C63A variant | 20          | C                           | A  | U  | C  | G  |                      |
| G10a         | 10          | C                           | A  | U  | C  | G  | G30A                 |
| G10b         | 10          | C                           | A  | U  | C  | G  | U34G, G66A           |
| G10c         | 10          | C                           | G  | U  | U  | G  | 65del                |
| G10d         | 10          | C                           | C  | A  | A  | G  | 67_69del             |
| G10e         | 10          | A                           | G  | U  | G  | G  | A46G                 |

*Table 3-1: Frequency and genotypes of sequences identified following selection – For each sequence it's % frequency amongst the 10 clones sequenced is given along with the identity at the randomised positions (pos.) and any additional mutations acquired outside these regions.*

To further analyse these sequences, the nucleotide frequency at each of the initially randomised positions was analysed from the clones before and after selection with respect to the Theophylline Control (Figure 3-17A - E). Figure 3-17 A - E illustrates that for each of these positions, except for position 63 there is convergence towards the identity of Theophylline Control. For instance, at position 81 towards the identity of the Theophylline Control (G81) increases from 10 to 100 %, demonstrating convergence.

To help explain the finding at position 63, the result of the study which identified the Theophylline Control was re-examined. This study identified a sequence, "VI-2", which had identical communication

residues to VI-1, except at position 63 where an A63 identity was present<sup>72</sup>. Significantly, VI-2 was shown to have a theophylline-activated phenotype similar to VI-1. Given A63 and C63 identities comprised 80 % of the sequence variation at this position, the data in Figure 3-17 suggests at each degenerate residue, convergence towards a known theophylline-activated genotype occurred.

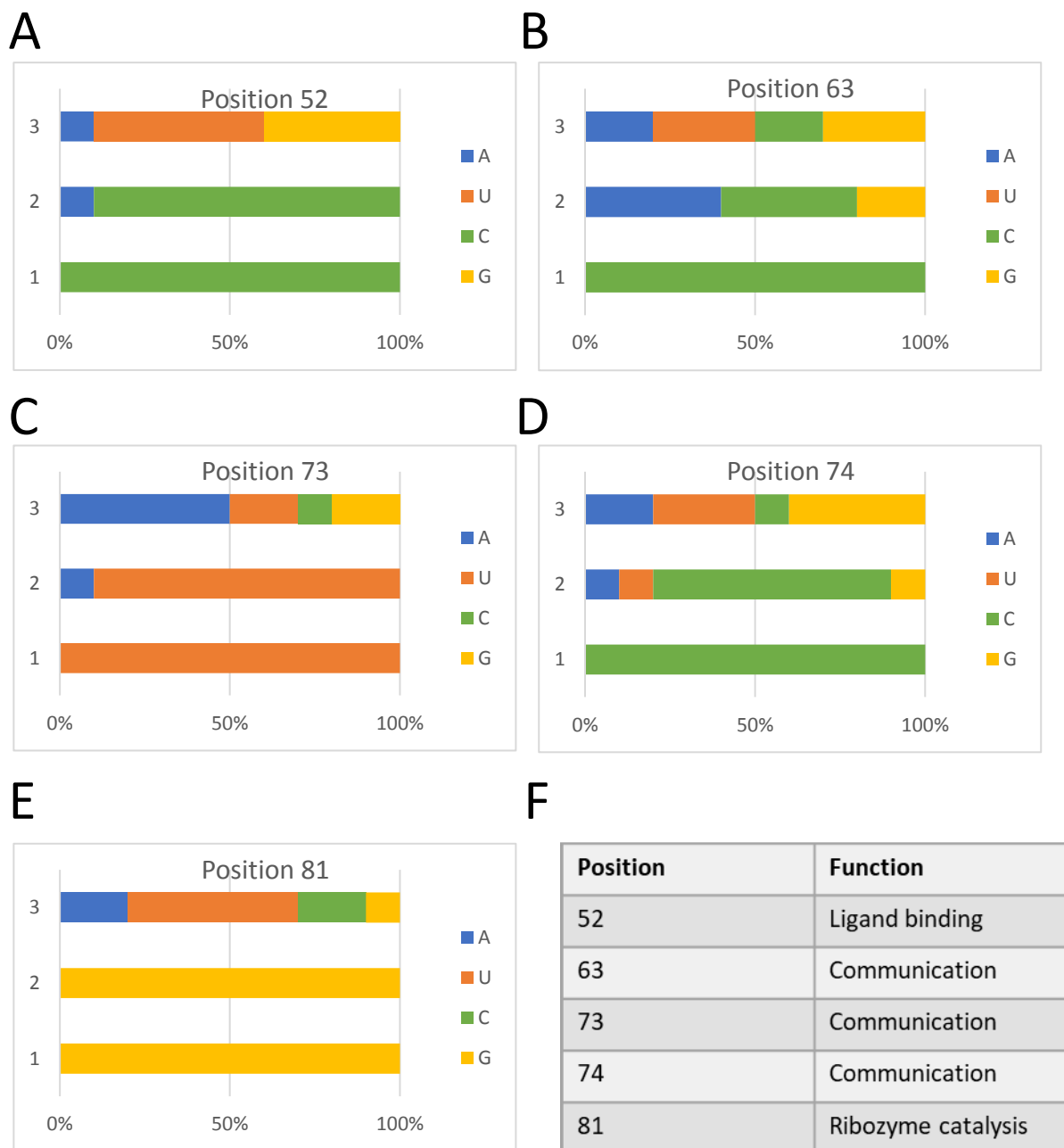


Figure 3-17: Enrichment of a theophylline-activated genotype – (A) – (E) The frequency of each nucleotide residue was determined at the randomised positions in the Theophylline Library for sequences isolated from the pool before and after selection. The identity of the Theophylline Control (Theo Con) is shown for comparison. (F) The function of each residue in the Theophylline Control.

## 3.4 Discussion

### 3.4.1 Control library design

During this work a control experiment to test the functionality of the selection method was implemented. In implementing this experiment, the Theophylline Library was developed. This library met all three prescribed requirements. Firstly, it contained a ligand-activated ribozyme in the form of the Theophylline Control. Secondly, it was compatible with the outlined selection method and thirdly it contained both inactive and constitutively active phenotypes, enabling both positive and negative selection to be scrutinised. Evidence for these sequences was presented in the form of previous biochemical data<sup>12,23,168</sup>, *in silico* folding of MFE structures and the dynamics of the library during selection. As such, enrichment of the Theophylline Control or similar theophylline-activated sequences from the Theophylline Library would demonstrate the functionality of the method at a small but significant scale.

### 3.4.2 Theoretical model

To the optimise selection conditions and estimate the amount of time required to enrich the Theophylline Control, a theoretical model describing the selection process was developed.

To optimise selection conditions, an expression for the fitness of a sequence was derived. This expression contains only two parameters which can be calculated empirically for given sequences based on the selection conditions. The use of this expression facilitated the optimisation of selection conditions to enrich for the Theophylline Control in a practical amount of time.

Furthermore, by considering the fitness of all possible sequences during selection, several results were realised. Firstly, as expected, sequences with maximum fitness are those which are maximally activated by the ligand. Secondly, of all the possible undesirable sequences under the selection conditions considered, a set of ligand-unresponsive sequences referred to as semi-active had the highest fitness value attainable. Interestingly, previous researchers conducting *in vitro* selections for ligand-responsive ribozymes have complained about the difficulty in removing “mis-folded”

sequences<sup>66</sup>. These sequences were suggested to partition into cleaved and full-length transcripts and as such are likely to be equivalent to the semi-active sequences described here. A third result from this analysis was that certain poorly ligand-activated sequences would not be enriched by selection. This result suggests that selection conditions should be carefully chosen to ensure ribozymes with the desired properties do not fall within this region during selection.

With several assumptions, it was possible to estimate the dynamics of the Theophylline Control during selection. Based on these dynamics, it was predicted that approximately 50 % of the sequences in the pool after 5 cycles of selection would be the Theophylline Control. This is relatively close to the 30 % reported by Sanger sequencing. However, it is noted that the pool prior to sequencing was gel purified and that a sample size of 10 is relatively small. To better scrutinise this model, a larger data set describing the dynamics of the Theophylline Control during selection could be acquired using next-generation-sequencing methods. This avenue is explored further in Chapter 5.

Another future application for this model would be to estimate the time required to conduct other selection experiments. An example would be the time required to enrich ligand-responsive ribozymes from more complex libraries. Furthermore, it is important to note that during the selection experiments in this chapter, positive selection conditions were assumed to be identical to negative selection conditions except for a difference in ligand concentration. Significantly, in all other *in vitro* selection experiments, the time at which ribozymes are inactivated during positive selection is different to the inactivation time used during negative selection<sup>55,66–69,72,178,179</sup>. The effect of this parameter on the enrichment of desirable sequences and the time taken to enrich functional sequences from more complex libraries is considered in the subsequent results chapter.

A parameter which was not included in the above model was the affinity of the ribozyme for the ligand. While not included, this parameter is accounted for in the % cleaved parameter calculated when a ribozyme sequence is exposed to ligand. A desirable feature of a previous model was that optimal ligand concentrations to enrich ligand-responsive ribozymes with bespoke  $K_D$  values could be

estimated<sup>69</sup>. With work, it should be possible to make similar estimations, given both this model and the one reported are based on similar principles of ribozyme cleavage.

### 3.4.3 Selection condition optimisation

In the process of optimising the selection conditions the kinetics of the Theophylline Control was characterised. In the presence of theophylline, the response of the Theophylline Control was biphasic. Significantly, the response of the VI-1 from the previous study was also characterised as biphasic<sup>72</sup> and as such these sequences are likely to have similar phenotypes. A potential reason for this biphasic response could be related to the ability of the hammerhead ribozyme to catalyse the reverse ligation reaction from cleavage reaction products<sup>17</sup> (Figure 1-1). Following the initial stages of transcription there is likely to be a build-up of cleavage reaction products leading to an increase in the rate of the backward, ligation reaction. This property of hammerhead ribozymes may explain why the % cleaved value increases rapidly before decreasing and levelling out. However, additional work is required to prove this explicitly.

In the absence of theophylline, Theophylline Control fitness decreased exponentially to the point where it was identical to that of ligand-unresponsive sequences. This limited the use of long IVT times to enrich for this sequence. To overcome this limitation, the TRT reaction was developed. During the TRT reaction the ribozyme is inactivated by the RTase which synthesises a complementary cDNA molecule, inhibiting the formation of an active conformation. To the best of our knowledge such a reaction has not previously been reported. As such demonstrating the functionality and application of the TRT reaction in this context represents a significant contribution by this work to the field. Significantly, under the conditions of the TRT reaction, the Theophylline Control's fitness improved such that its fold change over undesirable sequences was twice as high. Another desirable feature of this reaction was the constant response generated by the Theophylline Control for extended incubation times. This enables a similar amount of cDNA to be generated compared to that achieved under standard RT reaction conditions.

One consequence of using the TRT reaction for both rounds of selection is that positive & negative selection conditions are identical apart from the [ligand]. However, it has often been the case that in previous *in vitro* ribozyme selections, inactivation times differ between positive and negative selections<sup>55,119,127,148,149</sup>. Modifying the TRT reaction to easily accommodate delays is considered in the subsequent results chapter.

#### 3.4.4 Phenotype enriched and selection modifications

Two attempts were reported during this Chapter to enrich the Theophylline Control from the Theophylline Library. While enrichment of the Theophylline Control may have occurred relative to other Theophylline Library sequences during the 1<sup>st</sup> attempt, many non-specific bands arose during the latter rounds of selection.

This result was not noticed during the 2<sup>nd</sup> attempt suggesting that modifications to PCR or the DNA ligase used, restricted the production of these non-specific bands. Following 5 cycles of selection, a phenotype approximately equivalent to the Theophylline Control was enriched from the Theophylline Library. Although as mentioned modifications to the selection method were made, none of these modifications compromised the method's scalability and all are amenable to automation e.g. semi-quantitative PCR<sup>129</sup>. This result is therefore a milestone in demonstrating that the enrichment of ligand-responsive ribozymes is feasible using the *in vitro* selection method developed during this work.

In modifying the selection method, an alternative form of semi-quantitative PCR was developed. This method is more accurate and requires fewer measurements than the previously outlined method<sup>129</sup>. Furthermore, the success of this method was empirically demonstrated by the distribution of yields achieved. To improve this method further, PCRs could be conducted in single tubes rather than in strips. This would enable the optimal number of cycles to be conducted for each PCR rather than having to use an average value for several; achieving a tighter distribution than that illustrated in Figure 3-13.

Although successful, there are alternatives to semi-quantitative PCR which could achieve a similar or better result. One possibility is the use of emulsion PCR. This method limits product-product hybrid formation & other biases while allowing for higher yields than those achieved using semi-quantitative methods<sup>100,157,176</sup>. However, the implementation of emulsion PCR is suggested to be more difficult than traditional solution PCR<sup>157</sup>. Furthermore, a reduction in throughput may occur as only  $10^{10}$  templates can be amplified while maximally limiting the formation of non-specific products<sup>180,181</sup>. A further alternative would be to quantify a portion of the selected cDNA using qPCR<sup>138</sup> and then to implement the required number of cycles on the remaining sample. This method would however require a real-time thermocycler making it a more expensive and less accessible.

There are several other improvements that could be made to the selection method. One of which is the optimisation of the DNA ligase buffer for specificity. It was recently shown that modifications to the pH and salt content of this component has the potential to improve the specificity of the ligase<sup>175</sup>. Considering the importance of this step, implementing modifications to improve its robustness are desirable. Another potential improvement is to use a double-strand specific DNase<sup>182</sup> to degrade the DNA template following the TRT reaction instead of removing it using a pull-down reaction. This step may function to further reduce DNA template carryover.

### 3.4.5 Genotype enriched

By determining the identity of individual sequences before and after selection, it was shown that convergence towards a known theophylline-activated genotype occurred at all randomised positions. Of the initially randomised positions, two produce noteworthy convergences. The first was the 100 % convergence seen at position 81 towards the identity of the control. The requirement of the G81 identity for general acid-base catalysis by the hammerhead ribozyme<sup>23,168</sup>, suggests positive selection was effective at removing inactive sequences. A further noteworthy convergence was seen at position 63. At this position, the method suggested that an alternative A63 identity has similar a functionality to that of the Control. Significantly, this result is supported by a previous study<sup>72</sup> which noted that

both the control and a similar but not identical sequence with this modification have strong theophylline-activated phenotypes<sup>\*10</sup>.

Of the clones sequenced following selection, two had significantly alternative genotypes compared to the Theophylline Control. It is possible that these sequences function through novel mechanisms not previously identified. However, each of these sequences was only identified once. Therefore, these sequences may have been recorded only by chance. Obtaining the identity of more sequences present in the pool after selection should enable more accurate deductions regarding phenotype. This could be achieved with additional Sanger Sequencing or with the use of next-generation sequencing technologies<sup>183</sup>, as is described in Chapter 5.

### 3.5 Summary

The aim of this chapter was to test the function of the selection method using a controlled experiment. To achieve this, a library containing 1024 different ribozymes sequences was designed where only one sequence corresponded to a known theophylline-activated ribozyme. The library was designed so that some of the remaining sequences were inactive or constitutively active, allowing both positive & negative selection to be tested. To optimise the selection method for this task, a theoretical model describing selection was developed. The results of this model and subsequent work led to the development of the TRT reaction for the novel preparation of cDNA encoding ribozyme sequences. Conveniently, this model could also be used to estimate the time required to enrich the Theophylline Control to detectable levels, meaning the resources required to implement this experiment could be estimated beforehand. Incorporating the TRT reaction and several modifications to the selection method led to a phenotype closely approximating the theophylline-activated sequence being enriched in 5 cycles. Enrichment of a theophylline-activated genotype at each of the randomised positions in the library was subsequently confirmed with Sanger sequencing. Furthermore, sequences with mutations in other loci were also characterised suggesting novel theophylline-activated sequences

---

<sup>\*10</sup>Refer to the Conclusion Chapter, Section 6.2.2 for a more detailed description and analysis.

may have evolved during selection. In conclusion, these results highlight the functionality of this method with respect to the control experiment designed and implemented in this Chapter.

## Chapter 4: Selection for Trp-activated ribozymes & estimating rates of enrichment

### 4.1 Introduction

In the previous chapter, a control library was designed which contained over 1000 sequences, only one of which was known to be activated by the drug theophylline, prior to selection. The library was designed such that the majority of sequences were unresponsive to theophylline and had either constitutively active or inactive phenotypes. Subjecting this library to several cycles of *in vitro* selection using the method developed, resulted in the enrichment of a theophylline-activated phenotype and convergence to a theophylline-activated genotype at degenerate positions. This result demonstrated the potential for enriching functional ligand-responsive ribozymes, using the developed method.

While the result of this control experiment was encouraging, further work is required to achieve the aim of characterising a ribozyme-based riboswitch with new ligand specificities. To address this aim in this chapter, the Trp library designed in Chapter 2 will be subjected to several cycles of selection to attempt to enrich a Trp-activated ribozyme. Given to the best of our knowledge a Trp-activated ribozyme has yet to be identified, identifying such a sequence may achieve the remaining aim of this thesis, if selected Trp-activated sequences are functional *in vivo*.

In addition to selecting for Trp-activated ribozymes, the number of selection cycles required to enrich sequences from challenging libraries was investigated using the model of selection. The information generated from this investigation should prove useful when considering the feasibility of such experiments in the future. During this analysis the phenotype of the Theophylline Control was used to approximate the phenotype of desirable sequences. In all the simulations conducted, the enrichment of the Theophylline Control was estimated from challenging libraries, containing undesirable sequences with the maximum fitness possible. In the first analysis, the selection conditions used in the previous chapter were considered. Following this analysis, variations to the ribozyme inactivation time during negative selection are investigated to attempt to speed up the rate of enrichment. A

significant result from this part of the investigation was the suggestion of an optimal delay during negative selection dependent on the parameters of the desired ribozyme.

## 4.2 Objectives

### *Objective 1:*

Perform cycles of selection on the Trp library, attempting to enrich for a Trp-activated phenotype.

### *Objective 2:*

Using the model of selection, investigate the number of selection cycles required to enrich functional sequences to detectable levels from large libraries.

## 4.3 Results

### 4.3.1 Objective 1 – Selection for a Trp-activated ribozyme

To enrich for a Trp-activated ribozyme, the Trp library was subjected to 5 cycles of selection. As before, each cycle involved 1 round of positive selection followed by 1 round of negative selection.

To attempt to enrich Trp-activated sequences at a faster rate, negative selection conditions were modified. The decision to modify negative selection conditions was based on the fact the Trp Library contains more than 10,000 times the number of sequences present compared to the Theophylline Library. Therefore, if functional ligand-responsive ribozymes are present, they are likely to be present at much lower frequencies and as such will require more cycles of selection for their detection. However, due to time constraints, only five cycles of selection could be implemented. This meant that to stand a good chance of identifying these species the rate of enrichment would need to increase.

To attempt to increase the rate of enrichment, the time prior to ribozyme inactivation during negative selection was delayed. This follows reports that implementing longer delays during negative selection selects against ligand-unresponsive ribozymes previously described as semi-active in the prior chapter<sup>119</sup>. It is also noted that all prior *in vitro* selection experiments often do not use the same

ribozyme inactivation time during both positive and negative selection<sup>55,66–69,72,178,179</sup>. However, a more detailed explanation of why this strategy improves the rate of enrichment can be found in the subsequent section of this chapter.

To delay ribozyme inactivation time during negative selection, the reaction scheme of the TRT reaction was consulted<sup>\*11</sup>. During this reaction, the action of the RTase functions to inactivate the ribozyme. It was hypothesised that by delaying the addition of this component the ribozyme inactivation time would be delayed. Depending on the cycle number, the RTase was withheld for various periods of time (Table 4-1). This gradual increase in withholding times followed advice that delaying inactivation for too long in the early stages of selection can be counterproductive<sup>119</sup>.

| Cycle | Negative selection<br>RTase delay (min) |
|-------|-----------------------------------------|
| 1     | 0                                       |
| 2     | 0                                       |
| 3     | 15                                      |
| 4     | 30                                      |
| 5     | 60                                      |

*Table 4-1: Modification to negative selection stringencies during Trp ribozyme selection – List of time delays for the addition of the RTase during negative selection cDNA preparation.*

To monitor changes in the phenotype of the pool, DNA generated between cycles 0 – 1 and 3 – 5 inclusive, were in vitro transcribed in the presence of 0 and 4 mM Trp for 10 minutes. The % cleaved value of the resulting RNA were calculated (Figure 4-1A). Figure 4-1A illustrates that there is some variation in the phenotype during selection. For instance, the response of the pool reduced from

---

<sup>\*11</sup>Reaction scheme as described in Figure 3-6

approximately 55 % to 50 % following the final negative selection in the 5<sup>th</sup> cycle. A similar change in phenotype is also observed in the 1<sup>st</sup> cycle of selection. Together this suggests that selection was altering the composition of the pool.

However, there appears to be no increase in the response of the pool to Trp during selection. This is summarised in Figure 4-1B which shows that in a similar manner to the pool before selection, the pool after selection shows no difference in the % cleaved response in the presence of 0 or 4 mM Trp. Therefore, it is concluded that enrichment of a detectable Trp-activated phenotype did not occur.

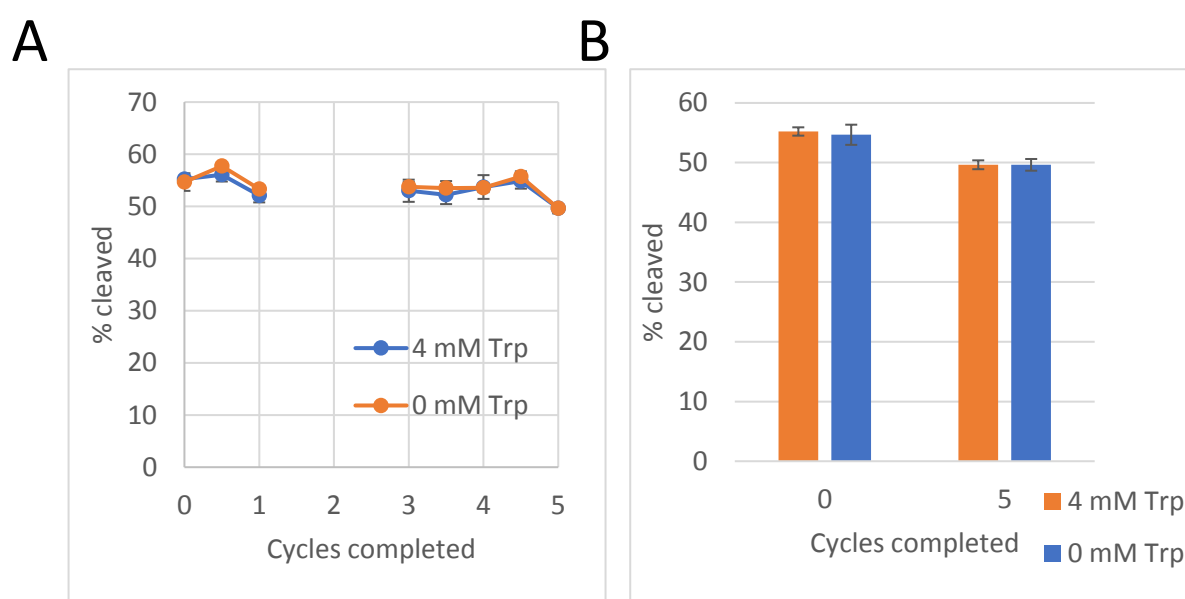


*Figure 4-1: Attempt to enrich for a Trp-activated phenotype from the Trp Library – (A) DNA templates generated following 0 – 1 and 3 – 5 cycles of selection were in vitro transcribed for 10 minutes in the presence of 4 or 0 mM Trp. RNA % cleaved values were subsequently calculated. (B) % cleaved of the pool after 0 and 5 cycles. For both (A) & (B) error bars indicate SDs from 4 repeats.*

#### 4.3.2 Objective 2 – Feasibility of screening larger libraries

Following the results of the previous section, it was of interest as to how many cycles would be required to screen libraries equivalent to or larger than the Trp Library? To provide an answer to this question, the model of selection, formulated in the previous chapter was used. In Section 4.3.2.1

below, selections are assumed to function under the conditions used to screen the Theophylline Library in the previous chapter. In the subsequent section, modification to the ribozyme inactivation time during negative selection are considered to reduce this number.

#### 4.3.2.1 Enrichment under the current conditions

As demonstrated in the previous chapter, the number of cycles required to enrich a sequence to detectable levels can be inferred from the dynamics of that sequence during selection<sup>\*12</sup>. To estimate a sequence's dynamics, the expressions in Equation (3-11) can be used if several parameters are known. These parameters are the cleavage responses of the sequence under the selection conditions, the number of sequences in the library and the cleavage responses of the pool at each round of selection. The latter can be calculated if all other sequences in the library are assumed to have the same phenotype and the order of positive and negative selections are known. The values of these parameters are considered in the following paragraphs.

To assign a cleavage response to the sequence being enriched, the previously determined cleavage response of the Theophylline Control under positive and negative selection conditions was used. However, as reviewed in the discussion it is possible that ribozymes responsive to other ligands will have weaker phenotypes affecting this assignment.

To ensure all potential library sizes were accounted for, the maximum number of variants which can be generated during a round of selection was considered. This value should limit the number of sequences which can be sampled and hence represents a maximum library size. To calculate this value, the maximum number of cDNA molecules which could be synthesised during selection was estimated (refer to the Appendix, Section 9.3.1). This was calculated at approximately  $10^{14}$  cDNA molecules for each round of selection. As such, the following simulations did not consider initial pools with more than this number of sequences.

---

<sup>\*12</sup>Refer to Chapter 3, Section 3.3.2.6 and Figure 3-9.

To account for fitter undesirable sequences being present in these more complex libraries, sequences other than the Theophylline Control were assigned a fitness value of 0.25. Under the assumptions of the model, this is the maximum value attainable by an undesirable sequence (refer to Chapter 3, Section 3.3.2.3, Figure 3-3B). As such, if the Theophylline Control is the only functional sequence in the library, the number of cycles calculated under these conditions should represent a maximum.

Based on the values of these parameters, simulations were conducted for initial libraries containing between  $10^6 - 10^{14}$  sequences, as described in the Appendix, Section 9.1 and Table 9-2. To calculate the number of cycles required to enrich the Theophylline Control to detectable levels, the number of cycles required to enrich this species to 50 % of the pool was calculated. At this point half the sequences randomly picked from the pool should be the Theophylline Control, facilitating easy detection. Figure 4-2 illustrates that when the Theophylline Control is diluted 100-fold, it requires an additional 4 cycles to reach this level of enrichment. For instance, 25 vs 29 cycles of selection are required to enrich the Theophylline Control to 50 % of the library when present in a library containing  $10^{12}$  vs  $10^{14}$  undesirable sequences.

It was also considered whether it might be possible to detect the Theophylline Control at lower frequencies with the use of more advanced technologies. Previously, NGS technologies have been employed to detect ribozymes with specific properties from amongst large numbers of candidate sequences<sup>121,122,184</sup>. During one of these studies,  $10^8$  reads were generated for a single library of ribozymes<sup>122</sup>. From this data, the authors of this study suggested that with 1000 reads corresponding to a given sequence, its phenotype could be accurately determined. Based on these values a % frequency of  $10^{-3}$  % should enable detection with the use of NGS technologies. Figure 4-2 suggests that the Theophylline Control would reach this frequency with 10 fewer cycles of selection. As such, the Theophylline Control could be detected within 19 cycles of selection rather than 29.

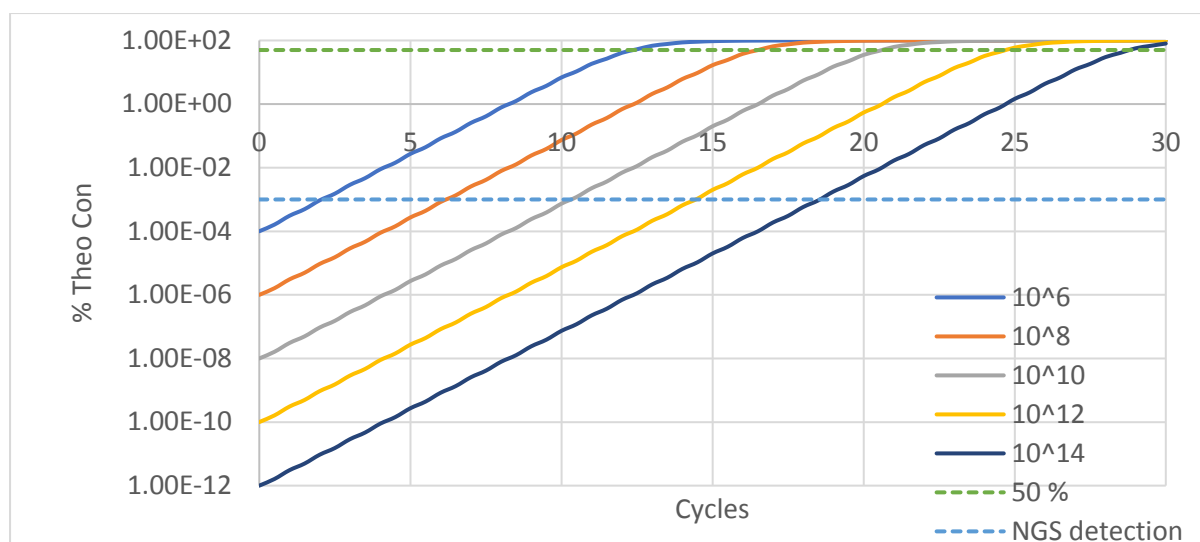


Figure 4-2: Enrichment of the Theophylline Control from more complex libraries under the previous selection conditions – The dynamics of the Theophylline Control as a percentage of the total pool were simulated when the Theophylline Control was present at varying initial frequencies. The remaining species present during selection are undesirable sequences with a maximum fitness value of 0.25 (refer to Appendix, Section 9.1, Table 9-2 for all the parameters used). 50 % of the pool and the % frequency required for detection by NGS<sup>122</sup> methods are indicated by green and blue dotted-lines, respectively.

#### 4.3.2.2 Enrichment with modifications to negative selection

It was hypothesised that delaying ribozyme inactivation during negative selection would increase the rate at which functional sequences are enriched. To determine if this was the case, it was necessary to calculate the current inactivation time under TRT reaction conditions. Once the value of this parameter is known, it can be varied, and its effect determined.

To calculate this time, it was assumed that the Theophylline Control's kinetics in the absence of theophylline follows that of a previous model<sup>148</sup>. With this assumption, the Theophylline Control cleavage rate constant and inactivation time could be calculated as described in the Appendix, Section 9.3.2. In the presence of 0 mM theophylline the Theophylline Control cleavage rate constant was calculated at  $0.1117 \text{ min}^{-1}$  and from this, that ribozymes are inactivated after 57 seconds during the TRT reaction. To determine the accuracy of this value, it was used to back-calculate the cleavage rate

constant for the Theophylline Control in the presence of 3.16 mM theophylline using empirical data (refer to the Appendix, Section 9.3.2). From this calculation, the cleavage rate constant for the Theophylline Control was calculated at  $5.5 \text{ min}^{-1}$  under these conditions. This is close to the  $5.7 \text{ min}^{-1}$  value previously reported for the similar VI-1 sequence<sup>72</sup>, suggesting the values calculated for these parameters were accurate.

By delaying ribozyme inactivation during negative selection, the % cleaved values for all sequences except inactive sequences will change. Therefore, assigning cleavage rate constants instead of % cleaved values to each sequence will simplify any analysis. This is because the fitness of a sequence with a specific set of cleavage rate constants during positive ( $k_{+c}$ ) & negative ( $k_{-c}$ ) selection can be analysed for selection conditions where ribozyme inactivation occurs at different times (refer to the Appendix, Section 9.3.3, Equation (9-4)).

Following a conversion to cleavage rate constants from % cleaved values, it was of interest as to whether previous results would be recovered. Recovering such results would demonstrate the accuracy when converting between units. One result from the previous chapter suggested that of the possible ligand-unresponsive sequences, a maximum fitness of 0.25 was achieved by semi-active sequences. To determine if this was still the case, the fitness of sequences with identical  $k_{+c}$  and  $k_{-c}$  values between 0 and  $6 \text{ min}^{-1}$  were calculated with no delay during negative selection (Figure 4-3D – 0 mins). Figure 4-3D supports the previous conclusion as a maximal fitness of 0.25 occurs for a ribozyme that is unresponsive to ligand and has a cleavage rate constant of  $1.7 \text{ min}^{-1}$ .

To understand the effect of delaying inactivation during negative selection, the fitness of sequences with cleavage rates during positive ( $k_{+c}$ ) & negative ( $k_{-c}$ ) selection between 0 and  $6 \text{ min}^{-1}$  was analysed for 0, 5 and 60 minute delays in ribozyme inactivation (Figure 4-3A – C). Several results are illustrated by this data. For all cases, the sequence which is maximally activated by ligand is still the fittest of all sequences considered. This is characterised by the maximum fitness values observed for sequences which have  $k_{+c}$  and  $k_{-c}$  values of 6 and  $0 \text{ min}^{-1}$ , respectively. Secondly, unless a sequence

has no background cleavage (i.e.  $k_{-c} = 0$ ), a decrease in fitness will occur when inactivation is delayed. Furthermore, the magnitude of this decrease is larger the longer the delay. For instance, Figure 4-3C illustrates that for the majority of sequences considered, their fitness approaches 0 when ribozyme inactivation is delayed by 60 minutes. Intuitively, this general reduction in fitness for active ribozymes is because longer delays leads to more cleavage, meaning fewer full-length cDNA molecules are available at the point of selection.

To more clearly analyse the effect of delays on unresponsive ribozymes, the fitness of these sequences was calculated for 0 and 5-minute delays (Figure 4-3D). This data shows two results. Firstly, with the longer 5-minute delay, the most-fit, unresponsive ribozyme has a lower cleavage rate constant. Specifically, after a 5-minute delay the most-fit, unresponsive ribozyme has a cleavage rate constant of  $0.08 \text{ min}^{-1}$ , significantly less than that obtained without a delay ( $1.7 \text{ min}^{-1}$ ). Secondly, there is a significant drop in the fitness of unresponsive ribozymes when implementing a 5-minute delay. For instance, the fitness of the most-fit, unresponsive ribozyme with a 5-minute delay has a fitness of  $< 0.01$  compared to 0.25 without a delay.

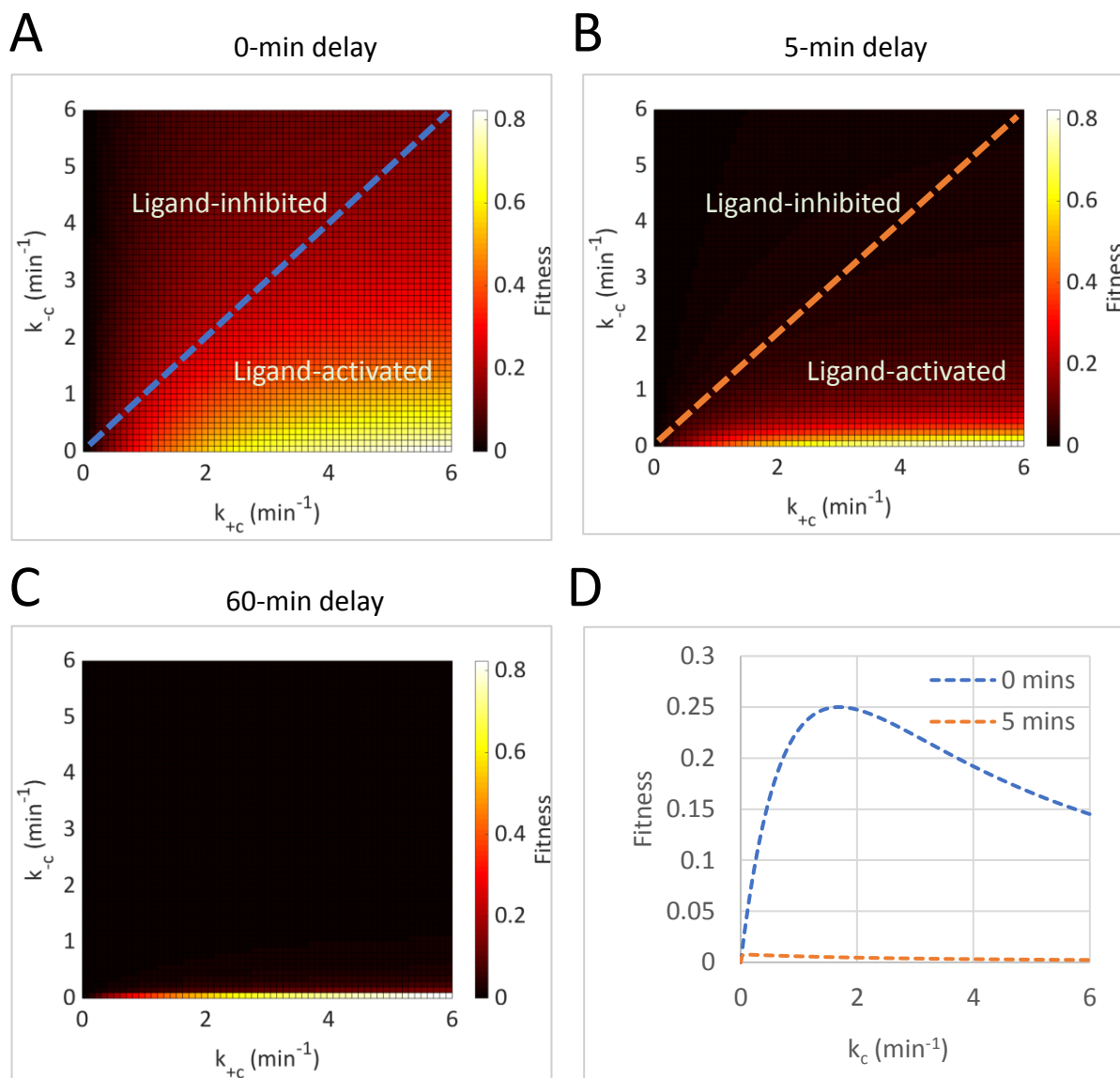


Figure 4-3: General effect of delaying ribozyme inactivation during negative selection – Fitness of sequences with cleavage rate constants during positive ( $k_{+c}$ ) & negative ( $k_{-c}$ ) selection which have values between 0 and 6  $\text{min}^{-1}$ . These fitness values were calculated for a (A) 0, (B) 5 or (C) 60-minute delay in ribozyme inactivation during negative selection. The colour scale indicates the fitness for each combination of  $k_{+c}$  &  $k_{-c}$  values. Where present dotted-lines indicates  $k_{+c} = k_{-c}$  with sequences to the left being ligand-inhibited and those to the right being ligand-activated. (D) Fitness of unresponsive ribozymes (i.e.  $k_{+c} = k_{-c}$ ) when either a 0 or 5-minute delay is used.

Previously, the fitness of a sequence was shown to be proportional to its fold-change following a cycle of selection (refer to Equation (3-14)). As such, dividing the fitness of one sequence by another will yield the fold-change of one sequence relative to another. Maximising this relative fold-change will identify the conditions which maximise the rate at which the more desirable sequence is enriched over the other. Based on the results in Figure 4-3, it appears that for delays in ribozyme inactivation during negative selection, the fitness of undesirable sequences decreases at a faster rate than some ligand-activated ones. If this is the case, these ligand-activated sequences would have higher relative fold-changes and as such would be enriched at a faster rate.

To determine if this was the case, the fitness of the Theophylline Control and the most-fit ligand-unresponsive sequence was calculated for delays between 0 and 180 minutes<sup>\*13</sup> (Figure 4-4A). Figure 4-4A illustrates that increasing the delay during negative selection causes an exponential decrease in the fitness of both sequences. This is equivalent to the results from Figure 4-3A – C. However, the fitness of the unresponsive sequence bottoms out faster, meaning the relative fold-change for these two sequences increases before reaching a maximum of 16 at 45 minutes. This suggests that the maximum rate of enrichment of the Theophylline Control over unresponsive sequences occurs when ribozyme inactivation is delayed by 45 minutes during negative selection.

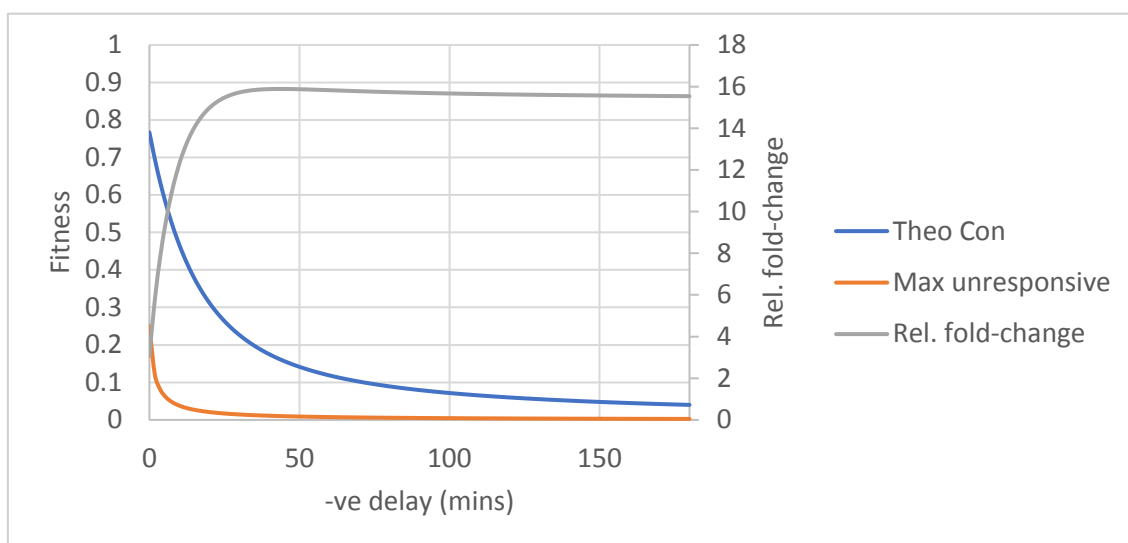
To understand why the relative fold-change reaches a maximum following a 45-minute delay, the % cleaved values of the Theophylline Control with varying delays were compared to those produced by the most-fit, unresponsive ribozyme (Figure 4-4B). Figure 4-4B shows that with a 45-minute delay, the % cleaved values are equal. Under these negative selection conditions, the Theophylline Control neither increases nor decreases in frequency compared to the most-fit, unresponsive sequence. This enables the Control to grow as fast as possible during positive selection without decreasing in frequency during negative selection, relative to this undesirable sequence.

---

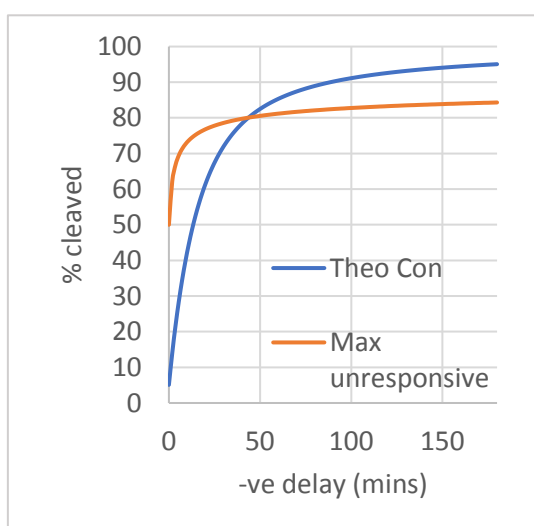
<sup>\*13</sup>Fitness' of sequences calculated according to Equation (9-4) in the Appendix. Cleavage rate constant for the most-fit, ligand-unresponsive sequence calculated as described in the Appendix, Section 9.3.4.

Figure 4-4C illustrates the effectiveness of this approach. Specifically, the simulation illustrated in this figure suggests that by implementing a 45-minute delay during negative selection, the Theophylline Control can be enriched to > 50 % of the pool within 12 cycles, even when present amongst  $10^{14}$  unresponsive sequences. This compares with the 29 cycles that were required under the previous conditions.

A



B



C

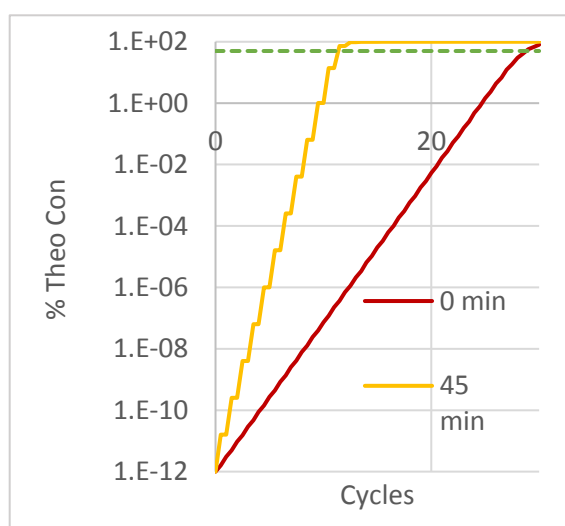


Figure 4-4: Optimal negative selection conditions for Theophylline Control enrichment – (A) Fitness of the Theophylline Control for varying delays in ribozyme inactivation during negative selection along with the maximum fitness for an unresponsive sequence. The relative (Rel.) fold-change is calculated by dividing the two fitness values at each time point. (B) % cleaved values for the sequences in (A) as a function of negative selection delay. (C) Theophylline Control selection dynamics under previous conditions (dark red) or with a 45-minute delay in inactivation during negative selection (yellow). Library contains  $10^{14}$  undesirable sequences with maximum fitness. 50 % of the pool is indicated by a green dotted-line. Refer to the Appendix (Section 9.1, Table 9-3) for the parameters used.

## 4.4 Discussion

### 4.4.1 Selection for a Trp-activated ribozyme

To attempt to enrich ligand-responsive ribozymes with new ligand specificities, the Trp library was subjected to several cycles of selection. With the aim of increasing the rate of enrichment, negative selection was modified to delay ribozyme inactivation. To achieve this, the RTase was withheld for bespoke periods of time. While changes in the overall phenotype of the pool were observed, the pool did not display a Trp-activated phenotype at the end of selection.

There are several potential reasons why a Trp-activated phenotype was not observed following selection. It could be that more cycles of selection were required to enrich ligand-activated sequences to a detectable level. This point of view is supported by the theoretical results described in this chapter suggesting up to 15 cycles of selection could be required to enrich a Theophylline Control-like sequence from a library containing a similar number of sequences ( $10^7$  - refer to Figure 4-2). This of course depends on the selection conditions and the phenotypes of other sequences in the pool.

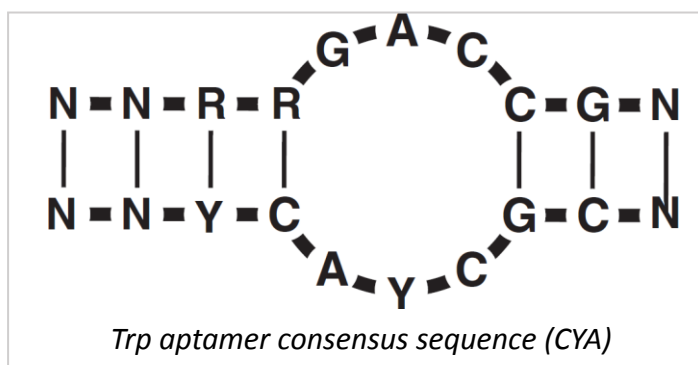
A further possibility is that the modifications to the selection method negatively affected selection. Specifically, the withholding of the RTase during negative selection could have led to the enrichment of sequences that respond to unintended stimuli. Given similar results have previously been observed, this effect cannot be ruled out. For instance, the authors of a previous study unintentionally enriched for acid-sensitive ribozymes<sup>66</sup>. This was caused by the ligand used during positive selection unintentionally altering the pH of the system. Whether withholding the RTase caused similar issues remains to be determined.

It is also conceivable that the Trp Library did not contain any functional sequences. This point of view is supported by Figure 4-5 which shows both the sequence and 2° structure for the CYA motif (Figure 4-5A) and the minimal 70-727 Trp aptamer (Figure 4-5B). While the CYA motif was shown to be the site at which Trp binds in the aptamer (green-dotted box, Figure 4-5B), this aptamer contains additional bulges and hairpin loops which are required for its functionality<sup>132</sup>. Perhaps these

sequences are required to stabilise the formation of the Trp binding pocket as has been noted in the case of other aptamers<sup>152</sup>. Given that the Trp library only contained the CYA motif, it is possible that none of the sequences in the initial pool could bind this ligand. In this scenario enrichment of Trp-responsive ribozymes would have required additional cycles to first evolve functional sequences prior to their enrichment.

However, there is evidence to suggest that a Trp-activated ribozyme might be feasible using the full aptamer sequence (Figure 4-5B). For instance, this sequence has been used on two previous occasions to generate riboswitches that increase gene expression 3 – 3.5 fold in bacteria when Trp is added to the media<sup>105,185</sup>. Furthermore, the riboswitch generated from one of these studies was subsequently used in a high-throughput screen to identify an *E. coli* strain with a 45.4 % improvement in Trp production relative to the parent<sup>186</sup>.

A



B

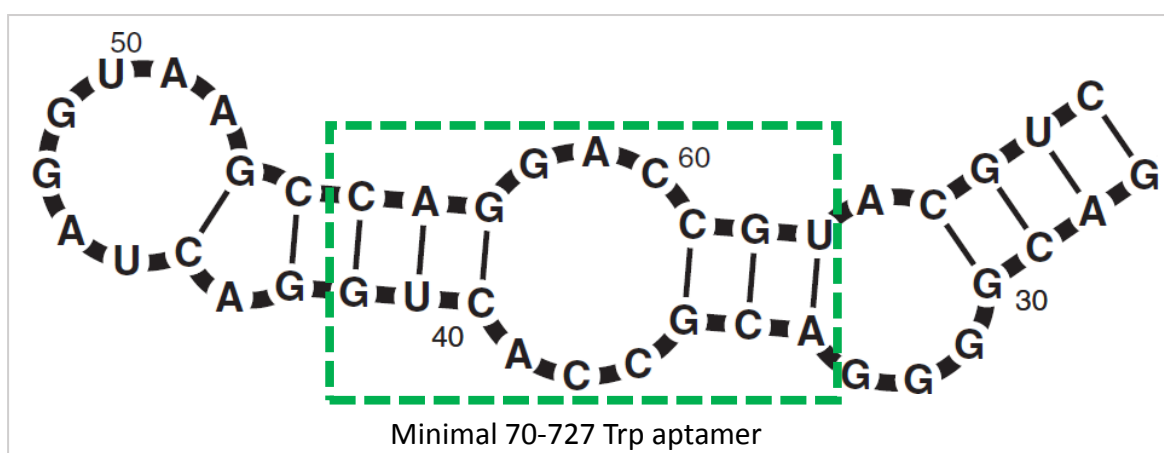


Figure 4-5: The Trp library may lack elements required for Trp binding – (A) The CYA motif present in 81 % of Trp binding aptamer sequences identified in a previous study<sup>132</sup>. (B) The minimal 70-727 Trp aptamer. The site of the CYA consensus sequence is illustrated by a green-dotted box. Figures adapted from Yarus and co-workers<sup>132</sup>.

However, prior to implementing additional work to generate a Trp-activated ribozyme, it is worth considering the value of such a species. This is particularly true considering for the purposes of monitoring Trp in bacteria, it is possible to re-engineer natural sensors. As an example, transcriptional attenuation seen in the Trp operon<sup>187</sup> was recently modified, generating a Trp biosensor<sup>133</sup>. This biosensor had improved dynamic range compared to the above synthetic Trp riboswitches<sup>105,185</sup> and was successfully applied in a high-throughput screen. As such it seems unlikely that a synthetic Trp-activated ribozyme could surpass the functionality of sensors derived from these natural elements,

particularly in the host species. Perhaps it is worth considering whether a Trp-activated ribozyme would be of value in other systems such as in mammalian systems for which ribozyme-mediated gene regulation is feasible<sup>56,188,189</sup>. Otherwise, generating ligand-responsive ribozymes against other ligands should be considered to better demonstrate the application of this method.

#### 4.4.2 Feasibility of screening larger libraries

The number of cycles required to enrich the Theophylline Control to detectable levels was estimated under the current selection conditions and with modifications to negative selection. Using the current selection conditions, the Theophylline Control could be detected with a maximum of 29 cycles of selection, assuming all other sequences were unresponsive to ligand. In addition to being achievable in a realistic period of time, a similar number of cycles has previously been conducted using other *in vitro* selection methods to enrich for ligand-responsive ribozymes based on the minimal hammerhead sequence<sup>66</sup>. However, given more transcription and RT steps are required to implement this method, it remains to be determined whether other factors such as selection bias would prevent this number of cycles from being obtained realistically.

It was also shown that currently available NGS methods<sup>122</sup> could be used to detect functional ligand-responsive ribozymes with fewer cycles of selection. It is worth noting a caveat of this approach is that without a phenotypic change, there is no way of knowing whether selection has been successful or not prior to NGS analysis.

To evaluate the effect of delaying ribozyme inactivation during negative selection, the current inactivation time during the TRT reaction was first calculated. From these calculations, the value of this parameter was approximately 1 minute. Significantly, several previous *in vitro* selection experiments have used no less than a 1 minute inactivation time when enriching functional sequences<sup>55,68,69,178</sup>. This suggests the TRT reaction and this method is not limited in implementing a short enough delay, during positive selection.

To analyse the effect of delaying ribozyme inactivation during negative selection, the fitness of all sequences with cleavage rate constants between 0 and 6 min<sup>-1</sup> were analysed for varying delays. Based on this analysis it was shown that the fitness of all sequences with non-zero cleavage rates decreases with increasing delays. However, compared to the fitness of undesirable sequences, sequences such as the Theophylline Control experience a relatively lower reduction in fitness. As such, these sequences have a higher relative fold-change and are therefore enriched at a faster rate.

By analysing this relative fold-change, an optimal delay was identified which maximised the rate of enrichment. The use of this strategy meant the number of cycles required to enrich the Theophylline Control was halved. Specifically, the maximum number of cycles was reduced from 29 to 12 cycles. Importantly this number of cycles or more is often conducted during previous attempts<sup>55,66,68,69,72,123</sup>. Furthermore, this number is only slightly more than double the number of cycles previously conducted to successfully enrich the Theophylline Control, suggesting it is likely to be achievable.

It is of interest as to whether this relationship holds for all ligand-activated sequences? If this is the case such knowledge would prove useful in establishing strategies for future selection experiments. For instance, sequences with bespoke background cleavage rates could be enriched in as short a time as possible by calculating the optimal negative selection delay.

Several limitations associated with this analysis are noted. Firstly, the number of cycles required to detect other ligand-activated ribozymes may be significantly larger than that for the Theophylline Control sequence. This is because ribozymes responsive to other molecules may have poorer performances relative to this sequence. Evidence to support this point of view has been presented by a previous study suggesting that given its size and complexity, theophylline is an ideal ligand for RNA detection<sup>184</sup>.

Secondly, in conducting these simulations, all the remaining sequences were assumed to be ligand-unresponsive and have maximum fitness values. However, the presence of other ligand-activated ribozymes would likely slow the enrichment of more functional sequences. This is because these

sequences would likely compete with one another. As such, the accuracy of these simulations is likely to be improved if more knowledge regarding the phenotypes of other sequences during selection is known.

Thirdly biases, such as different PCR amplification rates or biases in transcription or reverse transcription rates were not considered. As mentioned, these factors could function to improve the fitness of undesirable sequences. Finally all sequences were assumed to follow simple cleavage kinetics outlined by a prior model<sup>148</sup>. However, the kinetics of the Theophylline Control in the presence of ligand<sup>\*14</sup> suggest more complex kinetics can exist depending on the conditions used.

## 4.5 Summary

The first objective of this chapter was to enrich a Trp-activated ribozyme from the Trp library designed previously in this thesis. To achieve this, the Trp library was subjected to 5 cycles of selection. To speed up the rate of enrichment, the ribozyme inactivation time during negative selection was delayed. While changes in the phenotype of the pool during selection were observed, a Trp-activated phenotype was not detected after 5 cycles. Following this result, work was conducted to achieve the second objective of this chapter. This objective aimed to estimate the feasibility of more challenging selection experiments using the model of selection. Under the selection conditions used in the previous chapter, it could be possible to enrich sequences similar to the Theophylline Control from large libraries. This is particularly true if the pool is analysed prior to a change in phenotype using NGS methods. However, by delaying ribozyme inactivation during negative selection, the number of selection cycles required should be reduced significantly. Furthermore, the results outlined in this chapter suggest that depending on the background cleavage rates of desired ribozymes, an optimal delay during negative selection can be calculated, ensuring that bespoke ligand-responsive ribozymes are enriched in as short a time as possible. Such information should prove useful when implementing

---

<sup>\*14</sup>Refer to Chapter 3, Figure 3-4A

future selections. Further work is however required to demonstrate that this is a general relationship which holds for all ribozymes.

# Chapter 5: NGS and characterisation of theophylline-activated riboswitches

## 5.1 Introduction

Within this results chapter, two aims were addressed. The first of these aims was to analyse the previously conducted control experiment in depth using NGS technology. The second aim of this chapter was to determine the ability of several *in vitro*-selected, theophylline-activated ribozymes to function as riboswitches in *E. coli*. These two aims are introduced separately within this Section.

### 5.1.1 NGS analysis of the Theophylline Library selection experiment

NGS technologies were released a little over a decade ago and since their introduction have had profound effects on the fields of genomics, transcriptomics, epigenetics and protein-nucleic acid interactions<sup>190,191</sup>. More recently NGS technologies have been applied to SELEX<sup>158,192–194</sup> experiments, generating large data sets which describe the enrichment of library variants during selection. This information is used to identify consensus sequences or functional sequences earlier in selection potentially saving time and resources while reducing selection biases.

Previously, in Chapter 3 a selection experiment was conducted to enrich a theophylline-activated ribozyme (Theophylline Control) from a control library (Theophylline Library). Analysing this selection experiment using NGS technologies would be useful for several reasons.

Firstly, it should be possible to analyse the accuracy of the simulation generated by the model of selection which estimated the dynamics of the Theophylline Control (Figure 3-9). If this simulation is shown to be accurate, this would provide evidence to support the accuracy of the simulations in the previous chapter which were based on similar assumptions. If, however inaccuracies are identified, they might give clues as to how the model can be improved to better achieve the objective of optimising future selection experiments.

Using NGS technologies would also enable other predictions and hypotheses to be formally tested. For instance, it was predicted that the selection method outlined in this thesis would be able to sample a larger region of the sequence space than previous methods. This was due to a lack of gel purifications which would otherwise select against functional sequences which contained more or fewer nucleotides than the initial library. To investigate the ability of the selection method to tolerate such sequences, the lengths of all sequences formed during selection are investigated.

It was also previously hypothesised that novel theophylline-activated sequences not originally present in the Theophylline Library evolved during the previous selection experiment. Identifying such ribozymes would demonstrate that the method is capable of evolving functional sequences through directed evolution. Furthermore, some of the sequences identified by Sanger sequencing differed in key ways when compared to the Theophylline Control. Perhaps these ribozymes have unique properties which are either desirable or may improve our understanding of the factors which affect ribozyme-based riboswitch functionality.

To achieve the objective of identifying theophylline-activated ribozymes which were absent from the Theophylline Library, NGS data generated in this chapter was used to make better predictions regarding the phenotypes of sequences present at the end of selection. The phenotypes of several of these sequences are then empirically determined under the selection conditions to see if these sequences match our expectations.

### 5.1.2 Characterisation of riboswitch activity for theophylline-activated ribozymes

In the Introductory Chapter of this thesis, the key limitations of *in vitro* methods for selecting ribozyme-based riboswitches were discussed (Section 1.5.2.3). One of the key limitations discussed was the unreliability of selected sequences to function as ribozyme-based riboswitches *in vivo*. To support this statement, a previous study<sup>72</sup> which failed to generate theophylline-responsive riboswitches in mammalian cells is often quoted<sup>46,113</sup>. Importantly, one of the theophylline-activated

ribozymes tested during this study was the VI-1 sequence. This sequence formed the basis for designing the Theophylline Control (Chapter 3, Section 3.3.1). Specifically, the VI-1 sequence and the Theophylline Control sequence differ only in the stem III region. If the Theophylline Control or any of the Theophylline-activated ribozymes which are derived from this sequence are shown to act as functional riboswitches in *E. coli*, this statement would need to be reconsidered, particularly with regards to generating riboswitches for this host organism.

Following the analysis of the previous experiment using NGS, the ability of *in vitro*-functional theophylline-activated ribozymes, including the Theophylline Control to function as riboswitches is investigated. This is achieved by incorporating these sequences into the 5' UTR region of GFP expression cassettes in a similar manner to that previously described<sup>53</sup> and illustrated in the Introductory Chapter of this thesis (Figure 1-4). The ability of these constructs to regulate GFP expression in *E. coli* in response to theophylline was subsequently quantified with the use of flow cytometry.

## 5.2 Objectives

### *Objective 1:*

Analyse the Theophylline Library selection experiment using NGS.

### *Objective 2:*

Investigate whether the NGS data supports several expected and previously demonstrated results.

These include showing:

1. The Theophylline Control and equivalent sequence with a C63A substitution (C63A variant) were enriched from the members of the Theophylline Library.
2. That inactive sequences in the Theophylline Library were selected against during positive selection.

3. That constitutively active sequences in the Theophylline Library were selected against during negative selection.
4. That non-Theophylline Library sequences increased in frequency during selection.

*Objective 3:*

Investigate whether significant insertion and deletion mutations occurred during selection. Such mutations should be tolerated by the method due to the absence of gel purifications during selection.

*Objective 4:*

Investigate the accuracy of the simulations describing the dynamics of the Theophylline Control during selection.

*Objective 5:*

Identify sequences which evolved during selection and are likely to be theophylline-activated ribozymes. Following their identification, synthesise and evaluate the phenotype and fitness of each sequence.

*Objective 6:*

Characterise the functionality of theophylline-activated ribozymes for their ability to regulate gene expression in *E. coli*.

## 5.3 Results

### 5.3.1 Objective 1 – NGS of the Theophylline Library selection experiment

To facilitate the other objectives in this results chapter, the DNA templates generated following each round of the previous selection experiment, were prepared to facilitate subsequent sequencing using Illumina® NGS technology (Materials and methods, Section 7.6.1.1).

The most common application for Illumina NGS platforms is in genomics research. In this case, there is a large amount of diversity in the nucleotide composition between each read sequenced. In

contrast, the sequences following each round of the selection experiment are expected to be similar, differing only at the degenerate loci in the Theophylline Library or where mutations have occurred during selection. Such low diversity libraries are not well tolerated by Illumina platforms<sup>195</sup>. To overcome this problem, a high diversity sample known as PhiX was spiked into the sample to increase read diversity. Based on previous reports<sup>121,122</sup> and recommendations, a PhiX spike-in of 20 % was requested.

Based on the requested PhiX spike-in and the run conditions (Materials & methods, Section 7.6.1.2), it was expected that 3 – 4 million clusters would be sequenced of which 80 % should correspond with the barcoded experiment. The total number of clusters analysed during sequencing was 3,443,852. This is within the expected range suggesting that DNA quantification prior to sequencing was correct. However, only 6 % of reads generated from these clusters correspond with the barcoded library (Figure 5-1A). This was significantly less than the 80 % expected. 43 % of the remaining clusters produced undetermined reads, meaning that they did not contain a barcode. Figure 5-1B illustrates that almost all these undetermined reads align against PhiX sequences, suggesting a 43 % PhiX spike-in might have occurred instead. While this was higher than the 20 % expected, its effect on the data set was not as detrimental as that caused by the number of failed clusters. Specifically, 51 % of the clusters generated failed to produce reads of sufficient quality (Figure 5-1A).

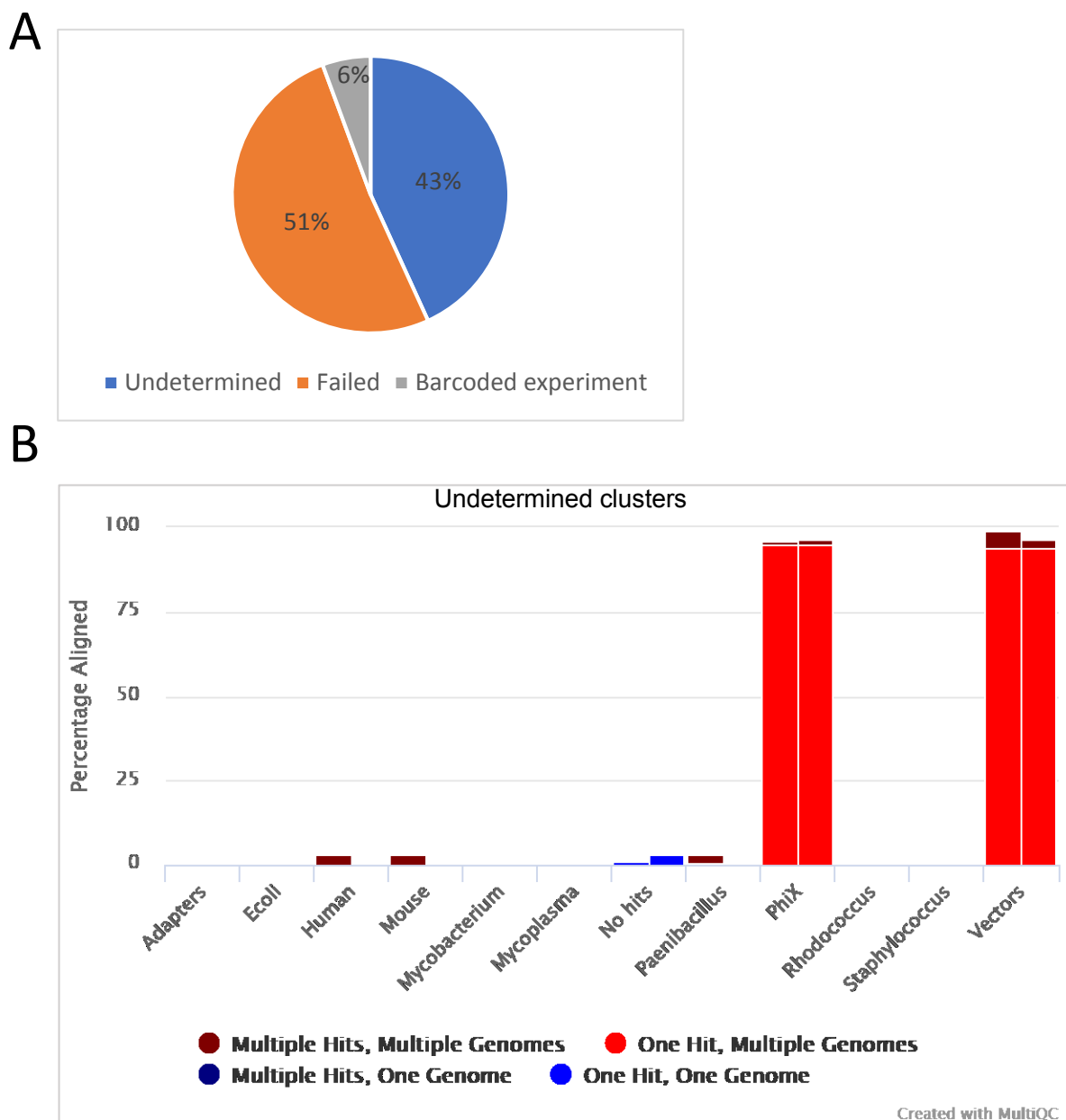


Figure 5-1: Relatively low yield from sequencing – (A) Pie chart illustrating % of failed and undetermined reads, in addition to those which correspond with the barcoded experiment. (B) Alignment of undetermined reads against various sources.

Based on Illumina troubleshooting guidelines<sup>196</sup>, there are two likely reasons which could explain the high number of failed clusters. First, it is possible that the quality of the sample was too low. This would be indicated by the presence of contaminating shorter or longer sequences such as primer

dimers. Such sequences would form clusters which could then not be sequenced. The second reason was that the nucleotide diversity was too low, even with the high percentage of PhiX used.

To rule out library quality as the cause for the low yield, the pooled and barcoded experiment was analysed on a bioanalyzer (Figure 5-2). Based on the length of the Theophylline Library and the size of the sequences added during sample preparation<sup>197</sup>, the size of the combined DNA templates should be approximately 275 bp. As illustrated by Figure 5-2, most of sequences in the pooled sample have a size very close to this, ~287 bp, with only a relatively small peak at 209 bp being observed. This suggests most sequences generated prior to sequencing were the expected size. As such, the low nucleotide diversity of the library appears the most likely cause for the high number of failed clusters.

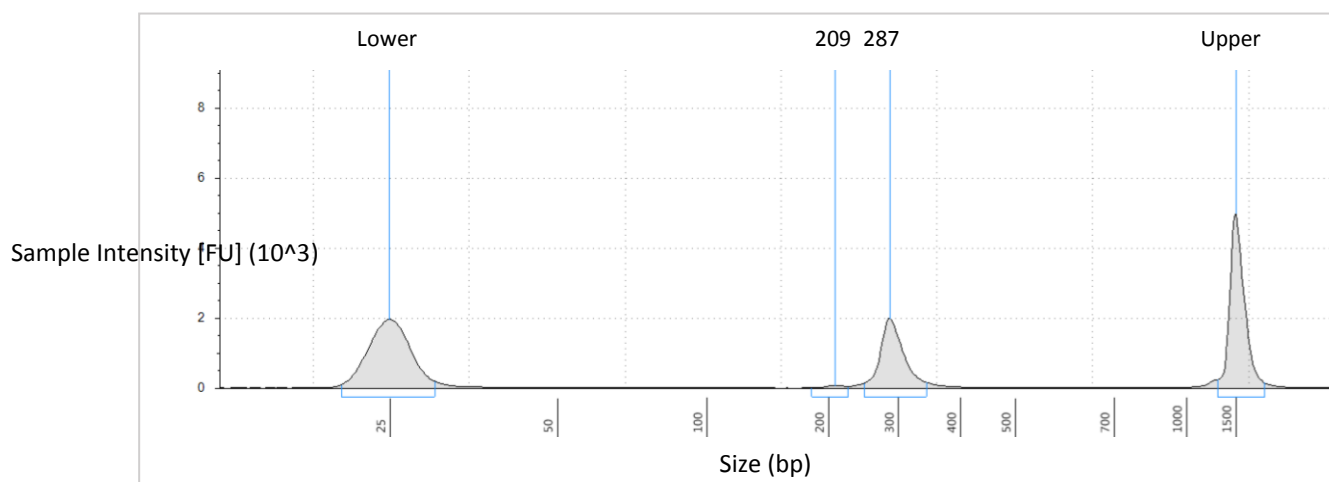


*Figure 5-2: Bioanalyzer trace suggests the barcoded selection experiment was of good quality – Prior to PhiX spiking, the pooled and barcoded selection experiment was analysed using a bioanalyzer to confirm the quality of the library. The size of the peaks generated from this analysis are indicated by blue lines.*

While the number of sequenced clusters was lower than expected, a key question was whether the data generated was sufficient to fulfil the remaining objectives set out in this chapter? To answer this question the demultiplexed reads corresponding with each round of the selection experiment were

processed<sup>\*15</sup> and analysed for their quality and quantity (Figure 5-3). Figure 5-3A illustrates the quality of each read as expressed as a Phred value. A Phred score of 30 indicates that an error in the identity of the base occurs at a frequency of 1:1000<sup>198</sup>. Given the length of each read was ~150 bases, a Phred score of 30 would indicate that the majority of reads contained no errors. As illustrated by Figure 5-3A, a Phred score > 30 was present at each position in all the reads, implying the data set generated was of good quality.

To determine if the quantity of data generated was sufficient, the number of sequences recorded for each round of selection was calculated (Figure 5-3B). Figure 5-3B indicates that for each round of selection almost 20,000 sequences were generated. More precisely, 17,595 sequences were generated on average for each round of selection. Given the Theophylline Library contained 1024 variants, this yields a coverage of more than 17-fold prior to selection. During selection, enrichment of sequences such as the Theophylline Control is expected, implying coverage for important sequences would increase when analysing the later cycles of selection. As such, the quantity of data generated was likely sufficient to enable the majority or all the objectives for this chapter to be met.

---

<sup>\*15</sup>Refer to the methods described in the Materials & methods chapter, Sections 7.6.1.2 & 7.6.1.3.

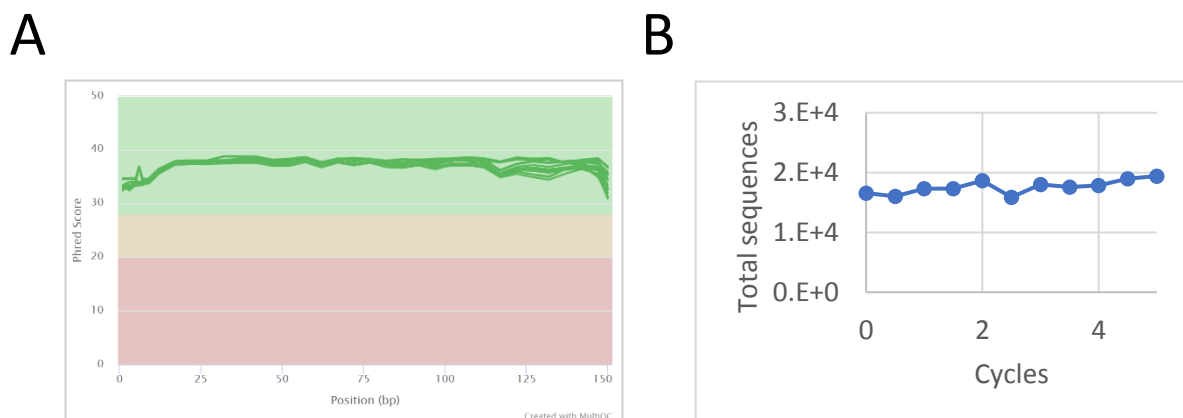


*Figure 5-3: Quality and quantity of sequences corresponding with the selection experiment – (A) Phred score as a function of nucleotide position for each read corresponding with the selection experiment<sup>199</sup>. (B) The total number of sequences obtained for each round during selection. Each round is equivalent to half a cycle. Total number of sequences given in scientific format e.g.  $1.E + 4 \equiv 10,000$ .*

### 5.3.2 Objective 2 –NGS data and Sanger sequencing data agree

Having generated a large and potentially useful data set, it was important to substantiate its accuracy. This was particularly important given the large proportion of clusters which failed during sequencing. For instance, it may have been that the failed clusters corresponded with a specific set of sequences biasing the data. To substantiate the accuracy of this data, several results previously determined from the Sanger sequencing data were investigated.

#### 5.3.2.1 Enrichment of sequences from the Theophylline Library

One of these results was that the Theophylline Control and an equivalent sequence containing a C63A substitution (C63A variant), were enriched from the Theophylline Library (refer to Chapter 3, Section 3.3.4). To determine whether the NGS data supported this result, the dynamics of these sequences as a % of recorded Theophylline Library sequences was calculated (Figure 5-4). Specifically, processed sequences which did not align with any of the 1024 Theophylline Library variants were disregarded in this calculation. The origin of these non-library sequences is considered in the final paragraph of this Section and subsequently in Section 5.3.3. Figure 5-4 illustrates that the NGS data supports this

previous result as the Theophylline Control and C63A variant were enriched from < 0.1 % each, to 35.6 & 54.1 % after 5 cycles of selection, respectively. As such, of the Theophylline Library sequences present at the end of selection, approximately 90 % were these two sequences.

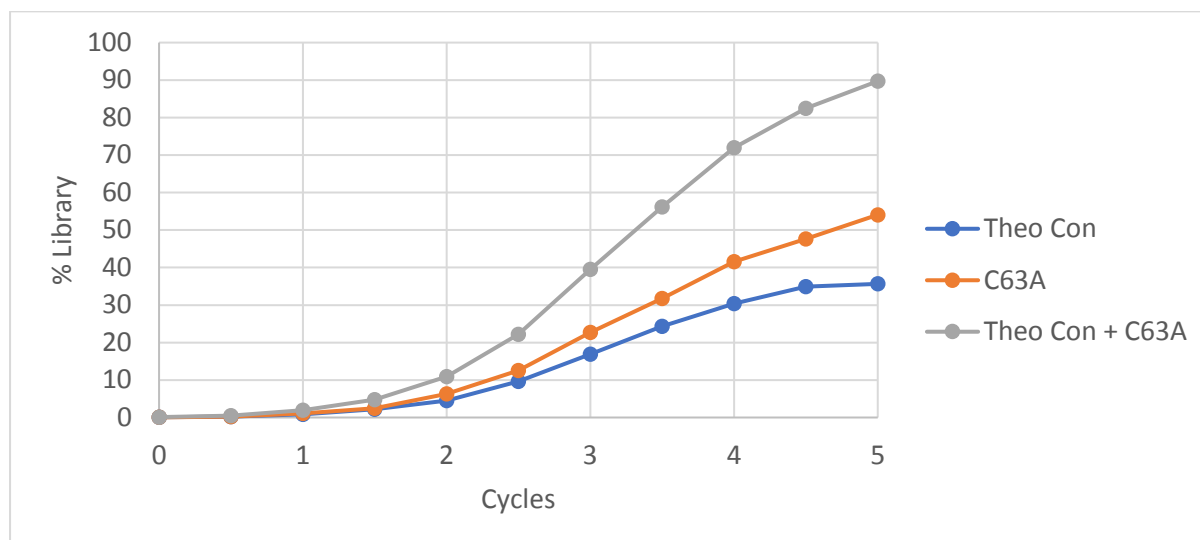


*Figure 5-4: Enrichment of expected sequences from the Theophylline Library – The % of Theophylline Control and C63A variant within the remaining Theophylline Library is illustrated as a function of selection. The sum of these two percentages has also been calculated (Theo con + C63A).*

#### 5.3.2.2 Dynamics of ligand-unresponsive and Theophylline Library sequences

Having shown that the NGS data supported the conclusion that the Theophylline Control and C63A variant were enriched from the initial library, two other previously established conclusions were investigated.

The first of these conclusions was that inactive, non-cleaving sequences were strongly selected against during selection. This result was highlighted by the fact every single clone analysed by Sanger sequencing after selection contained a G81 identity which is required for catalysis<sup>12</sup>. To determine whether the NGS data supported this conclusion, the dynamics of all Theophylline Library sequences lacking a G81 identity were analysed using the entire NGS data set (green-line, Figure 5-5A). These suspected inactive sequences comprise three quarters of the Theophylline Library and account for 768 sequences. Figure 5-5A shows that following 5 cycles of selection there is a significant decrease in

the frequency of processed sequences aligning with these variants. Specifically, the average frequency of these sequences reduces roughly a 1000-fold, from 1:1000 before selection to 1:1 million following selection. This suggests that these sequences were strongly selected against.

Figure 5-5A also illustrates that there is agreement between the dynamics and expected phenotype of these Theophylline Library inactive sequences. For instance, during the 1<sup>st</sup> cycle's positive selection, these sequences decreased in frequency from 0.08 to 0.01 %. This agrees with the expectations of positive selection which should select against inactive sequences which do not cleave (refer to Chapter 2, Figure 2-3A). During the subsequent negative selection, the frequency of these sequences increases from 0.01 to 0.04 %. Again, this is in line with the expectations of negative selection which should enrich for inactive sequences which produce relatively more full-length cDNA.

The dynamics of the three suspected constitutively active sequences from the Theophylline Library were also calculated using the NGS data. These sequences were designed so as not able to bind theophylline and to form the parental Hammerhead ribozyme MFE 2° structure in the absence of this ligand (refer to Chapter 3, Figure 3-2A). Therefore, they were suspected of being theophylline-insensitive and having fast cleavage rates.

Figure 5-5A shows that the 1<sup>st</sup> round of positive selection increased the frequency of these sequences at a faster rate than the Theophylline Control. This is characterised by a steeper gradient in the line describing their selection dynamics for this round (4.8 vs 4.0). This result is in line with the expectation that these sequences were cleaving at a faster rate, producing more cleaved RNA & hence cleaved cDNA for selection. Their frequency decreased significantly following the subsequent round of negative selection, suggesting that these sequences were theophylline-unresponsive. This contrasts with the Theophylline Control which increased in frequency. This data is therefore in agreement with the view that these sequences were constitutively active and that positive and negative selection were functioning as expected.

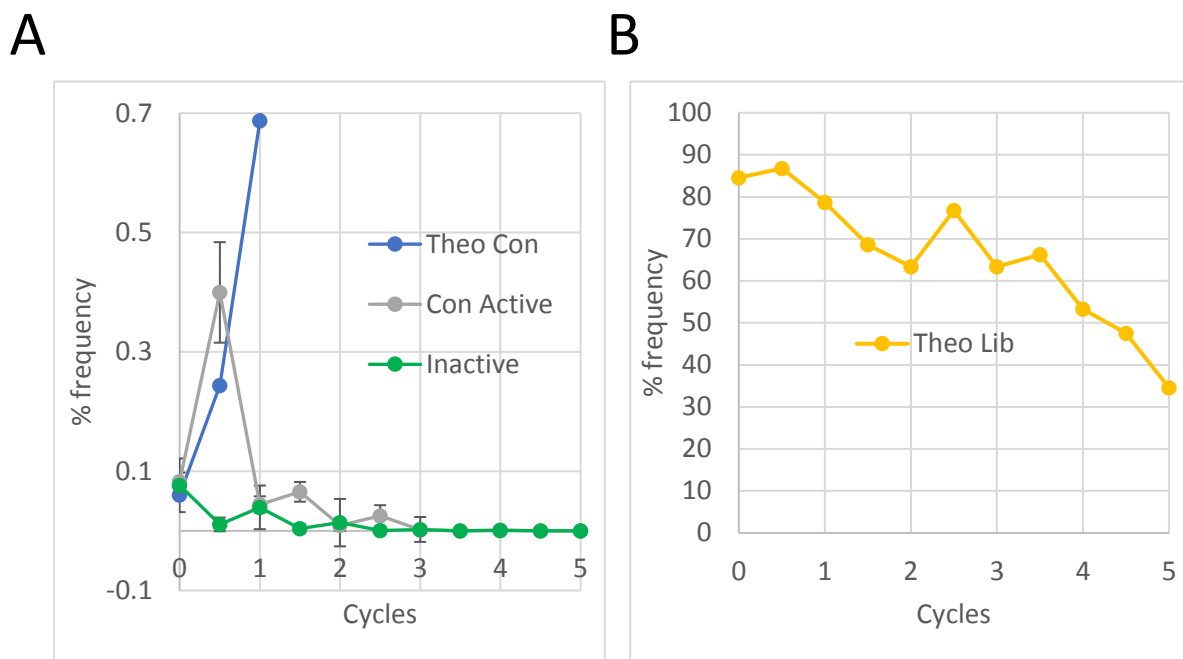


*Figure 5-5: Dynamics of Theophylline Library sequences & suspected ligand-unresponsive sequences – The entire NGS data set was used to determine either (A) the % frequency of Theophylline Library inactive and constitutively active (Con Active) sequences or (B) the % frequency of all Theophylline Library sequences as functions of selection. Error bars indicate SDs. The dynamics of the Theophylline Control during the 1<sup>st</sup> cycle is shown for comparison in (A). Note, selection involved cycles of alternating rounds of positive followed by negative selection.*

Previously generated Sanger sequencing results relating to this selection experiment also suggested that during selection Theophylline Library sequences decreased in frequency. Specifically, this data showed that 70 % of clones before selection aligned against sequences from the Theophylline Library (data not shown), compared to 50 % after selection<sup>\*16</sup>.

To evaluate whether the NGS data supported this conclusion, the % frequency of Theophylline Library sequences during selection was calculated (Figure 5-5B). Figure 5-5B shows that 84.5 % of sequences recorded from the pool prior to selection aligned with one of the 1024 Theophylline Library variants. The remaining 15.5 % of sequences can be accounted for by errors during DNA synthesis and

<sup>\*16</sup>Refer to Chapter 3, Table 3-1.

mutations which occurred during NGS library preparation. Figure 5-5B also illustrates that the frequency of sequences aligning against Theophylline Library variants decreases to 34.6 % after 5 cycles. This is similar to the value previously recorded from the Sanger sequencing results.

The reduction of Theophylline Library sequences during selection either suggests that sequences in the original pool not aligning with the Theophylline Library were preferentially selected, or that there are new emergent sequences. As is illustrated by the results in the following section, the latter case is likely to be true.

### 5.3.3 Objective 3 – Insertion deletion mutations tolerated by the method

One potential benefit of avoiding gel purifications during selection is that insertion and deletion mutations are not selected against. This is because, functional sequences which are not the expected length are not discriminated against. During the process of selecting from the Theophylline Library and preparing it for NGS no gel purifications were conducted. As such the data generated from NGS should demonstrate the ability of the method to tolerate variants with insertion and deletion mutations.

Following the occurrence of an insertion and deletion mutation, the length of the sequence should change. Therefore, to analyse the method's tolerance for these mutations the lengths of all sequences and those recorded prior to selection were analysed (Figure 5-6).

Analysing the distribution of sequence lengths before selection (Figure 5-6A), there is a clear centring of the distribution around the Theophylline Library length (139 bp). This is expected as prior to selection many of the sequences originated from this library. It is worth noting that there are instances of smaller or longer sequences. For instance, there is a collection of sequences around 100 bp and several instances of sequences around 280 bp. As mentioned, these insertion and deletion mutations presumably occurred during either DNA synthesis or during the preparation steps required for NGS e.g. barcoding PCR.

Comparing Figure 5-6A with Figure 5-6B, the distribution of sequence lengths recorded during selection is more diverse than those recorded prior to it. For instance, 52.8 % of sequences recorded prior to selection had the same length as the Theophylline Library compared to 39.9 % for all sequences recorded. Therefore, insertion and deletion mutations must have occurred during selection and these mutations must have been tolerated by the method.

Furthermore, the range of sequence lengths for all sequences recorded almost spans the realistic range. Specifically, only sequences less than 300 bp can be entirely captured with the Illumina kit used in this instance. Furthermore, sequences shorter than 64 bp would have less complementarity to the primers used to prepare the library<sup>\*17</sup>. Given the largest and smallest sequence recorded was 284 and 50 bp, respectively these maximum and minimum lengths were almost observed. This suggests that the range of insertion and deletion mutations achievable by this method is not limited by the method itself.

Observing the distribution of lengths in Figure 5-6B, there appears to be two peaks in the frequency of sequence lengths. One of these peaks occurs around the length of the Theophylline Library (139 bp), while another is observed at 67 bp. Considering the length of barcoding and adapter sequences, the length of these sequences prior to sequencing are equivalent to 203 and 275 bp, respectively. This correlates well with the distribution of sequence lengths observed from the bioanalyzer trace taken prior to sequencing (Figure 5-2). As such, this data supports the view that the NGS data accurately captures properties of the sample sequenced.

---

<sup>\*17</sup>The Theophylline Library shares overlaps of 32 bps with both MH\_IllumSense & MH\_IllumAnti primers used during the amplicon PCR to prepare the NGS sample. As such sequences with < 64 bps would have reduced complementarity.

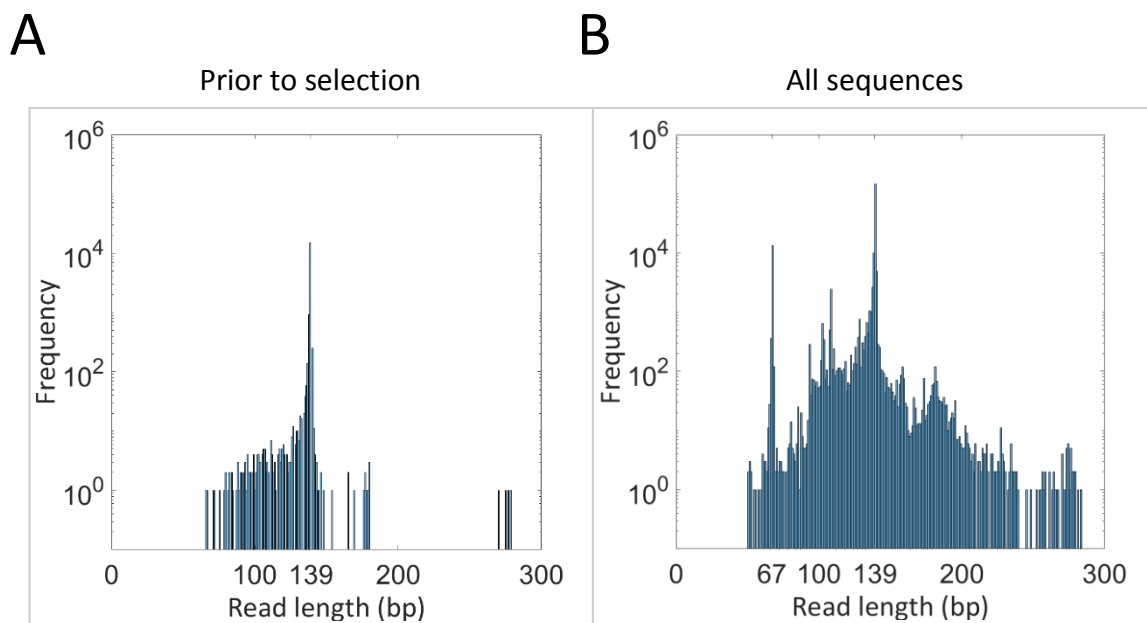


*Figure 5-6: Distribution of sequence lengths – Histograms indicating the length of sequences (A) recorded prior to selection or (B) all sequences detected. The frequency for each bin is given on a log scale.*

#### 5.3.4 Objective 4 – Accuracy of model simulation

In the previous results chapter, simulations of the selection method were used to estimate the time and resources required to enrich desirable sequences under different sets of selection conditions. Determining the accuracy of these predictions is important considering these predictions underpin the feasibility of using the selection method to screen larger libraries.

To estimate the accuracy of these predictions, the accuracy of the simulation describing the enrichment of the Theophylline Control from the Theophylline Library was investigated. Demonstrating that this simulation is accurate would provide evidence to suggest similar simulations based on the same model and assumptions are also accurate. To achieve this objective, the dynamics of the Theophylline Control calculated from NGS data is compared to its predicted dynamics from the simulation conducted previously in Chapter 3 (Figure 3-9). Figure 5-7 illustrates that the dynamics of the Theophylline Control calculated from NGS data are accurately predicted up to and including the

3<sup>rd</sup> cycle ( $R^2 = 0.97$ ). During the 4<sup>th</sup> and 5<sup>th</sup> cycles, there is divergence between the expected and actual % frequencies.

It was thought that some of this divergence could be accounted for by the presence of the C63A variant. The enrichment of this sequence is not accounted for by the model which assumed that all sequences apart from the Theophylline Control, decreased in frequency at the same rate during selection\*<sup>18</sup>. In contrast to this assumption, the C63A variant was enriched along with the Theophylline Control during selection (Figure 5-4). To account for this effect, the enrichment of the Theophylline Control and C63A variant together was simulated using Equation (3-11) and the parameters given in the Appendix (Section 9.1, Table 9-1). This simulation assumed that the Theophylline Control and this mutant have identical phenotypes and were present at the same initial frequency. With this and the previous assumptions, the model accurately predicts the enrichment of these two species up to and including the 4<sup>th</sup> cycle ( $R^2 = 0.94$ ). However, during the 5<sup>th</sup> cycle there was divergence from the predicted and actual enrichment of these two species.

This discrepancy between the simulated enrichment of these two sequences & their actual dynamics could be explained by the presence of other functional sequences which were not accounted for by the model. The presence of such sequences is supported by Figure 5-5A and Figure 5-6. Both these figures suggest sequences not originally present in the pool prior to selection were produced during this experiment. Potentially some of these sequences may have similar or improved fitness' compared to the Theophylline Control, meaning they were enriched during the 5<sup>th</sup> cycle.

---

\*<sup>18</sup> Refer to Chapter 3, Section 3.3.2.6.

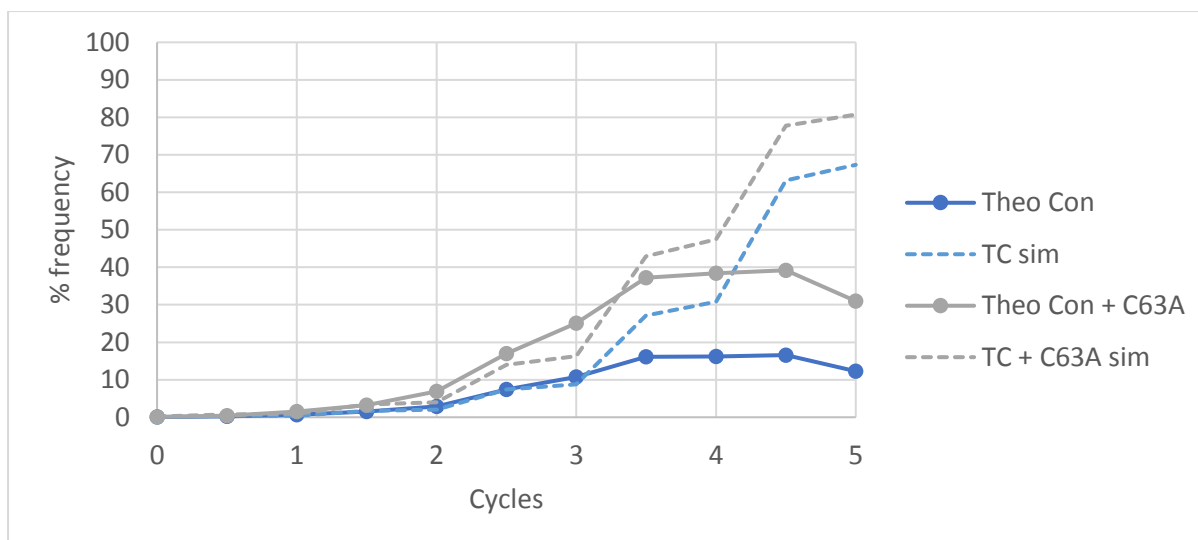


*Figure 5-7: Accuracy of simulations describing selection dynamics – The simulation describing the enrichment of the Theophylline Control from the Theophylline library (TC sim), is compared to its dynamics calculated from NGS data. A similar comparison is made between the simulated enrichment of both the Theophylline Control & C63A variant together (TC + C63A sim) and NGS data describing their dynamics.*

### 5.3.5 Objective 5 – Identification of other theophylline-activated ribozymes

#### 5.3.5.1 Identification of obvious cheaters & sequences for analysis

To identify sequences with desirable phenotypes, the fold change of all sequences from the 4<sup>th</sup> to the 5<sup>th</sup> cycle was calculated. From Equation (3-14), those sequences with the highest fold changes are predicted to have large fitness values and hence the most desirable phenotypes. To this end, only sequences which increased in frequency by  $\geq 1.5$ -fold from the 4<sup>th</sup> to the 5<sup>th</sup> cycle were considered for further analysis. Furthermore, to ensure sequences considered for further analysis did not appear by chance, only sequences recorded at least 25 times in the final pool after selection were considered.

From this analysis, 28 sequences were identified. To analyse the composition of these sequences, they were compared against the Theophylline Library through a multiple sequence alignment conducted in MATLAB (Figure 5-8). From this alignment, it was observed that several sequences have large deletion mutations. For instance, sequences 1 – 7 have deletions in the alignment from bases 27 up to 75 when

compared against the Theophylline Library. Importantly, part of this region aligns with the theophylline aptamer and communication domain. Given these domains are required for a theophylline-activated phenotype, it was suspected that these sequences were not theophylline-activated ribozymes and had instead developed mechanisms to cheat the selection process.

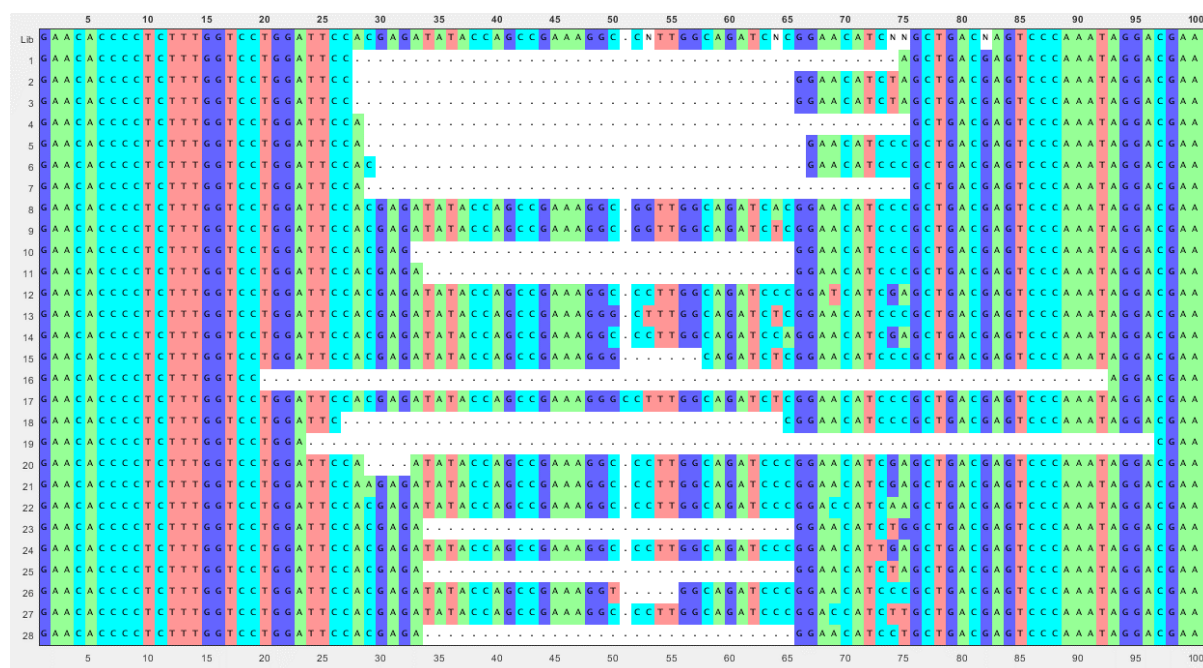


*Figure 5-8: Multiple sequence alignment of high fold change sequences in final selection cycle – A multiple sequence alignment was conducted for the Theophylline Library (Lib) and the 28 sequences identified which had a fold change ( $> 1.5$ ) and were recorded at a frequency greater than 25 after selection. The 28 sequences are ordered in descending fold change. The first 100 bases of this alignment are shown. “.” indicates a gap in the alignment.*

Following the identification of these obvious cheaters, their dynamics during selection was estimated (Figure 5-9). This was conducted to determine how the enrichment of these sequences varied during selection. Such information would give an indication to the extent of these sequences and hence whether such sequences are a real concern to the success of future selection experiments.

To calculate the dynamics of the obvious cheaters in Figure 5-8, they must first be identified from within the NGS data set. To achieve this, common properties of these sequences were considered. It

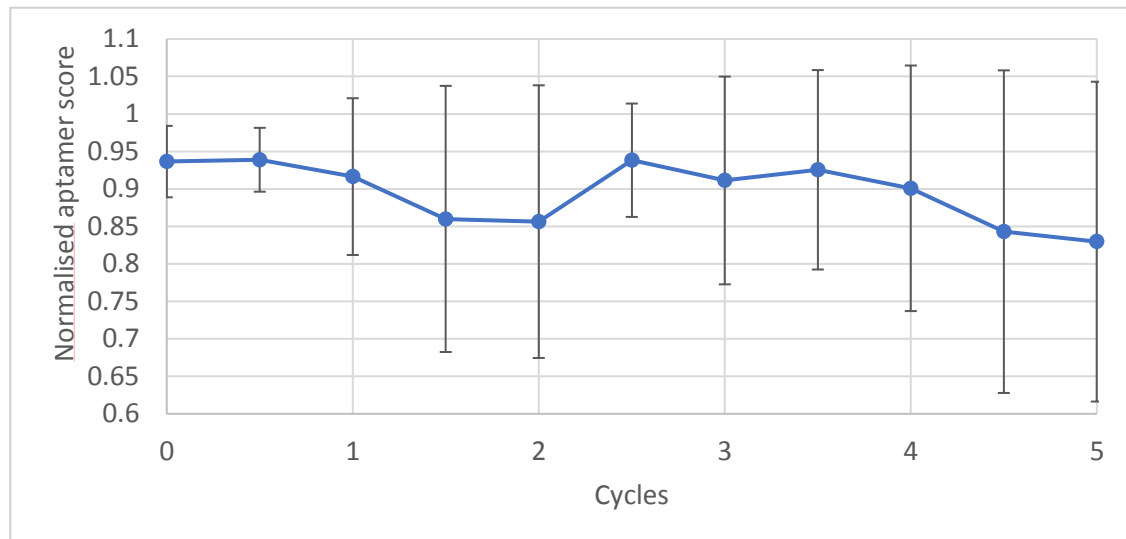
was suspected that each of these sequences would have a low local alignment score when aligned against the theophylline aptamer domain. This is based on observations from Figure 5-8, which suggested these sequences were missing this domain. Following this realisation all processed sequences were locally aligned against the theophylline aptamer. Averages and standard deviations of these scores normalised to the maximum possible score were plotted for each round of selection (Figure 5-9A).

Figure 5-9A illustrates that many sequences present prior to selection align well against the theophylline aptamer, suggesting initially the majority of sequences contained the theophylline aptamer. This is characterised by a small standard deviation and an average normalised alignment score close to 1. Over the course of the first two cycles this value drops, suggesting sequences not containing the theophylline aptamer evolved and were enriched. However, this reverses in the first positive selection of the third selection cycle where the normalised aptamer score increases from 0.86 to 0.94. Following this oscillation there is again a steady decrease in the normalised aptamer alignment score.

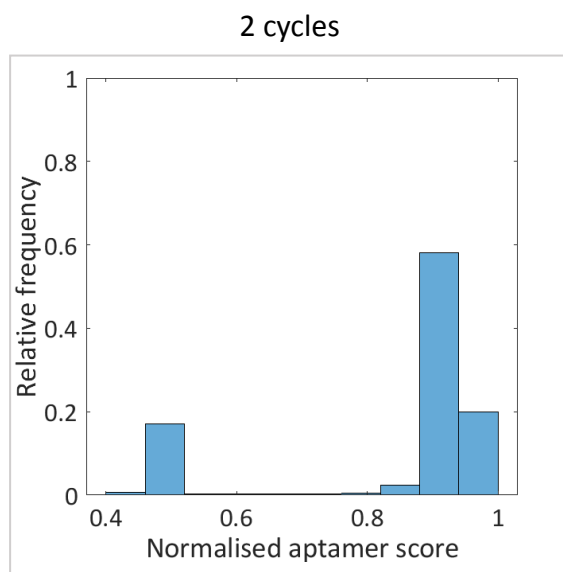
A further observation from Figure 5-9A is the relatively large standard deviations present in some of the normalised aptamer alignment scores. It was thought that this might be due to the presence of a bimodal distribution. Figure 5-9B & C illustrates that this was the case. Specifically, there appears to be two groups of sequences in the pool following the 2<sup>nd</sup> and 5<sup>th</sup> cycles. One group has a low normalised score of around 0.5 and as such is likely to represent obvious cheaters which lack this domain. The other group meanwhile has a score closer to 1, suggesting these sequences contained the theophylline aptamer domain. Comparing Figure 5-9B & C, there is not much difference in distributions recorded following the 2<sup>nd</sup> & 5<sup>th</sup> cycles of selection. Specifically, approximately 70 – 80 % of sequences appear to have the theophylline aptamer after both 2 and 5 cycles of selection. This suggests after an initial enrichment of these sequences, they were not able to break through this

barrier and dominate the pool. Whether this would have occurred with further cycles of selection remains to be determined.

**A**



**B**



**C**

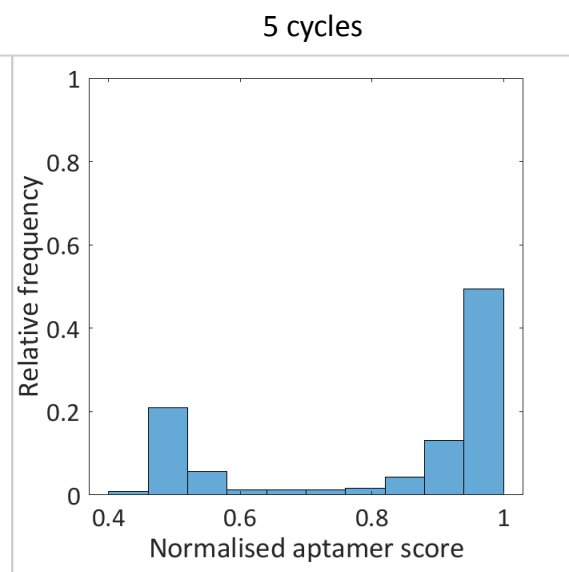


*Figure 5-9: Estimated dynamics of obvious cheaters during selection – (A) For each round of selection, all processed sequences were locally aligned against the theophylline aptamer. Each alignment was normalised to the maximum score. Averages and standard deviations were calculated from this data. The distribution of normalised aptamer alignment scores was calculated after (B) 2 & (C) 5 cycles and plotted as histograms.*

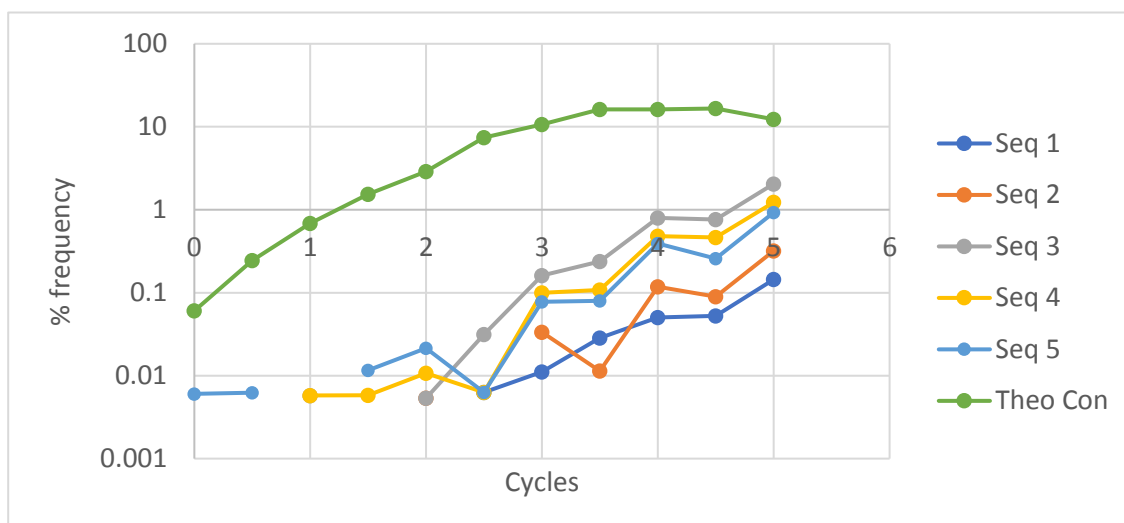
To avoid including these obvious cheaters in any further analysis, sequences unlikely to contain theophylline binding and communication domains were disregarded using the algorithm given in the Appendix (Section 9.4.1). After removing these obvious cheaters, 13 sequences remained. Given the available resources, it was feasible to empirically analyse the phenotype of 5 of these sequences.

While sequences with the highest fold change were suggested to have strongest phenotypes, it was suspected that there might be other factors, such as detection limit issues or mutational bias<sup>174</sup> which could influence the fold change during the final cycle of selection. Therefore, in addition to selecting the two sequences with the highest fold change, the three remaining sequences with the highest frequency after selection were also selected for further analysis. Given higher absolute frequencies, it was reasoned that any predictions regarding these sequences would be more accurate.

To understand the enrichment of these sequences during selection, their dynamics was compared to that of the Theophylline Control (Figure 5-10A). Figure 5-10A, shows that all the selected sequences were either not recorded prior to selection (0 cycles), or in the case of Sequence 5 were present at a lower initial frequency. As such, their relatively lower frequency at the end of selection could be accounted for by their lower initial frequency and not because of a weaker fitness.

To estimate the phenotypes of these sequences relative to the Theophylline Control, the fold-changes of these sequences during the 5<sup>th</sup> cycle of selection were analysed (Figure 5-10B). Figure 5-10B illustrates that all the sequences selected for further analysis had much higher fold-changes during this cycle of selection. Therefore, based on the model of selection they should have larger fitness values and more desirable phenotypes.

A



B

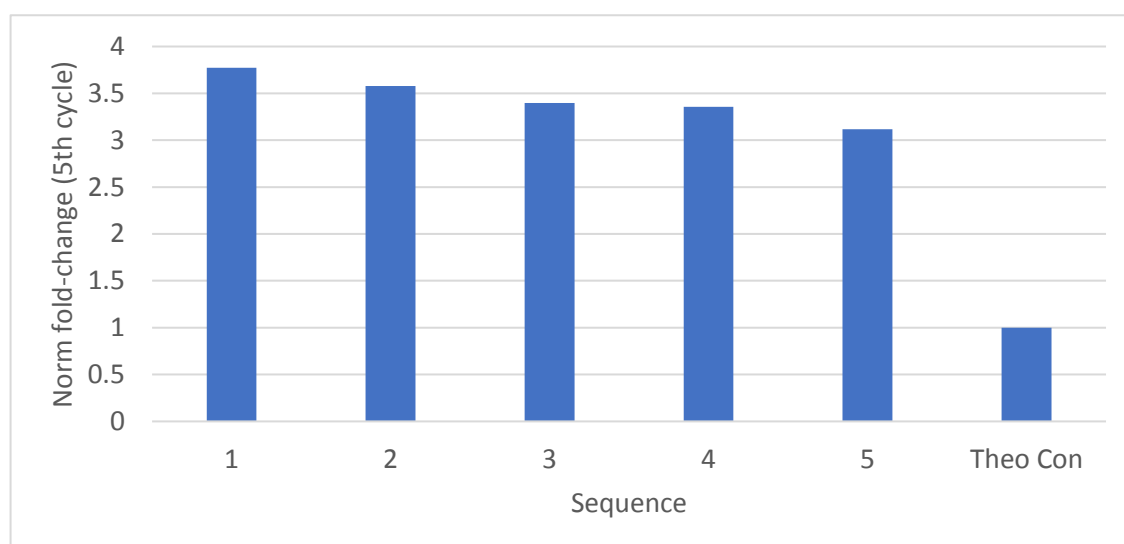


Figure 5-10: Selection dynamics of sequences selected for analysis – (A) The % frequency of the five sequences selected for further analysis as a function of selection cycles is compared to that of the Theophylline Control (Theo Con). (B) The fold change of each sequence during the 5<sup>th</sup> cycle selection is normalised to that of the Theophylline Control.

#### 5.3.5.2 % cleaved & fitness of selected sequences

To determine the accuracy of the predictions made in the previous section, each of the identified sequences was chemically synthesised. dsDNA templates for each sequence were subsequently

generated and incubated under positive and negative selection conditions equivalent to those used during selection<sup>\*19</sup>. The resulting % cleaved vales for each sequence were calculated based on the cDNA generated and these values used to determine the fitness of each sequence according to Equation (3-14).

Figure 5-11 illustrates that in contrast to the predictions made from Figure 5-10B, none of the identified sequences had improved fitness values relative to the Theophylline Control ( $p < 0.01$ ). This is because none of these sequences had larger cleavage responses during positive selection and while sequences 3 & 4 had similar responses during negative selection (9.6 and 11.7 % vs 6.7 %); neither these nor any of the other sequences had statistically lower background cleavage rates ( $p < 0.01$ ). Furthermore, sequences 1, 2 & 5 had similar responses during positive & negative selection, suggesting they were not theophylline-activated ribozymes.

---

<sup>\*19</sup>Refer to the Appendix, Section 9.4.2 for an explanation of the 4.9 mM theophylline used under positive selection conditions.

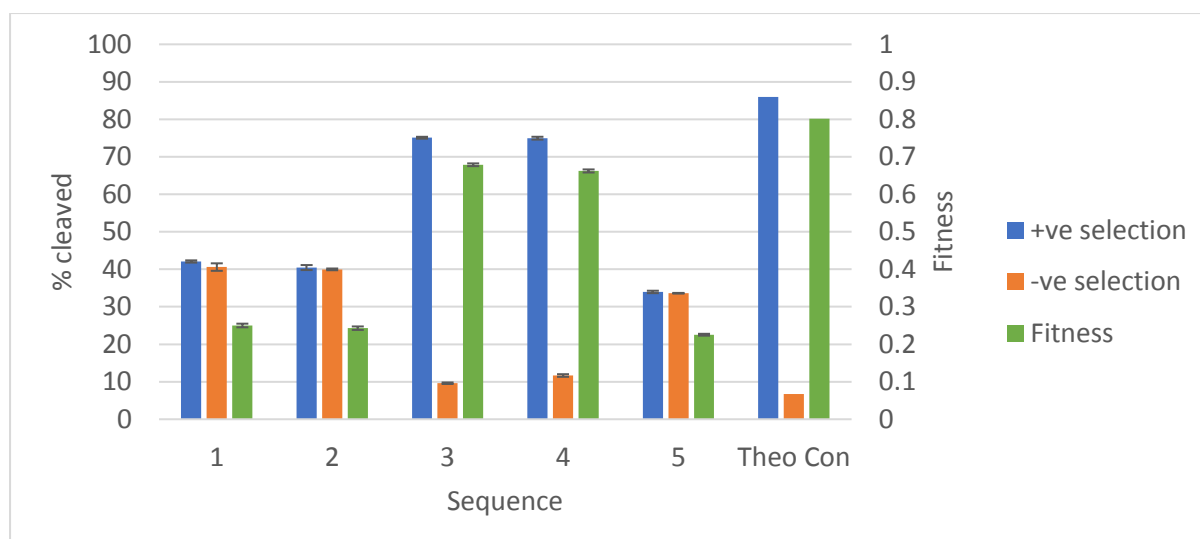


*Figure 5-11: Properties of identified sequences – Sequences identified from NGS data were incubated under positive & negative selection conditions used during Theophylline Library selection (Appendix, Section 9.4.2). % cleaved (blue and orange) and fitness values (green) were calculated from the resulting cDNA and compared to those of the Theophylline Control (Theo Con). Error bars calculated from 3 repeats.*

### 5.3.5.3 Recalculation of fitness with fewer assumptions

To attempt to understand why the predictions made in Section 5.3.5.1 were incorrect, the assumptions of the model were re-examined. Two key assumptions of the model were that all DNA templates were transcribed with equal efficiency and that following transcription, all RNA was reverse transcribed. The validity of the second assumption was based on data generated under standard RT reaction conditions<sup>173</sup> where the efficiency of the RTase is optimal. However, the conditions of the TRT reaction are different and hence are unlikely to be optimal for reverse transcription. For instance, the TRT reaction was incubated at 37 °C which is lower than the optimal RT temperature<sup>140</sup>. Furthermore,  $\text{PPi}^{4-}$  released from the hydrolysis of NTPs forms a precipitate with  $\text{Mg}^{2+}$  in the form of  $\text{Mg}_2\text{PPi}$ . This leads to a lowering of the concentration of free  $\text{Mg}^{2+}$  during the reaction<sup>200</sup>. Given  $\text{Mg}^{2+}$  is a required nucleotide polymerase co-factor<sup>138</sup>, the likelihood of RNA synthesised towards the end of the TRT

reaction being reverse transcribed is less. Therefore, in the specific case of the TRT reaction, the assumption of complete reverse transcription for all RNA is likely to be inaccurate.

To account for this, the fitness of a sequence was derived without assumptions in the rates of RNA & cDNA synthesis by substituting the result from Equation (3-5) into Equation (3-12), followed by algebraic manipulation:

$$fitness_i = \frac{C_i^{(n+)}}{x_i^{(n)}} \times \frac{F_i^{(n-)}}{x_i^{(n)}} \equiv K_{(n)} \frac{x_{i(n+2)}}{x_{i(n)}} \quad (5-1)$$

$$\text{Where } K_{(n)} = \frac{C_i^{(n+)} F_i^{(n-)}}{X^2}$$

Equation (5-1) suggests that as before, fitness is directly proportional to the fold-change of a sequence ( $\frac{x_{i(n+2)}}{x_{i(n)}}$ ) following a cycle of selection. However, the fitness of a sequence is now dependent on two

different terms. The first term ( $\frac{C_i^{(n+)}}{x_i^{(n)}}$ ), is equivalent to the number of cleaved cDNA molecules each

DNA template produces under positive selection conditions. The value of this parameter should be constant under given positive selection conditions assuming sequences do not interact with one another. As such this parameter is considered independent of the round of selection ( $n$ ). The second

term ( $\frac{F_i^{(n-)}}{x_i^{(n)}}$ ), is equivalent to the number of full-length cDNA molecules each DNA template produces

under negative selection conditions. As was the case for  $\frac{C_i^{(n+)}}{x_i^{(n)}}$ , the value of this parameter should also

be independent of  $n$  if the same assumptions are true. With this realisation the fitness of sequence  $i$  is written:

$$fitness_i = \frac{C_i^{(+)}}{x_i} \times \frac{F_i^{(-)}}{x_i} \equiv K_{(n)} \frac{x_{i(n+2)}}{x_{i(n)}} \quad (5-2)$$

To determine whether RT efficiencies or transcription rates improved the fitness of any of the identified sequences, the values of  $\frac{C_i^{(+)}}{x_i}$  &  $\frac{F_i^{(-)}}{x_i}$  were estimated by the intensities of the cleaved and full-length cDNA bands produced by each sequence under the relevant selection conditions. Using these values, the fitness of each sequence was recalculated according to Equation (5-2) (Figure 5-12). If RT efficiencies or transcription rates played a role in the enrichment of a sequence, the fitness for that sequence should be relatively higher compared to the values calculated in Figure 5-11. Figure 5-12 shows that sequences 1, 2, 4 & 5 still all had lower fitness values compared to the Theophylline Control. However, in contrast to the fitness previously calculated for Sequence 3, the recalculated fitness for this sequence was similar to that determined for the Theophylline Control.

Analysing the amount of full-length and cleaved cDNA produced by this sequence under the relevant selection conditions, this improved fitness appears to be the result of more full-length cDNA being produced under negative selection conditions. Specifically, the amount of full-length cDNA quantified for this sequence was  $1358 \pm 33.9 AU$  compared to the Theophylline Control's 1149.9. This result suggests that the full-length Sequence 3 species was transcribed or reverse transcribed with greater efficiency than the Theophylline Control during negative selection.

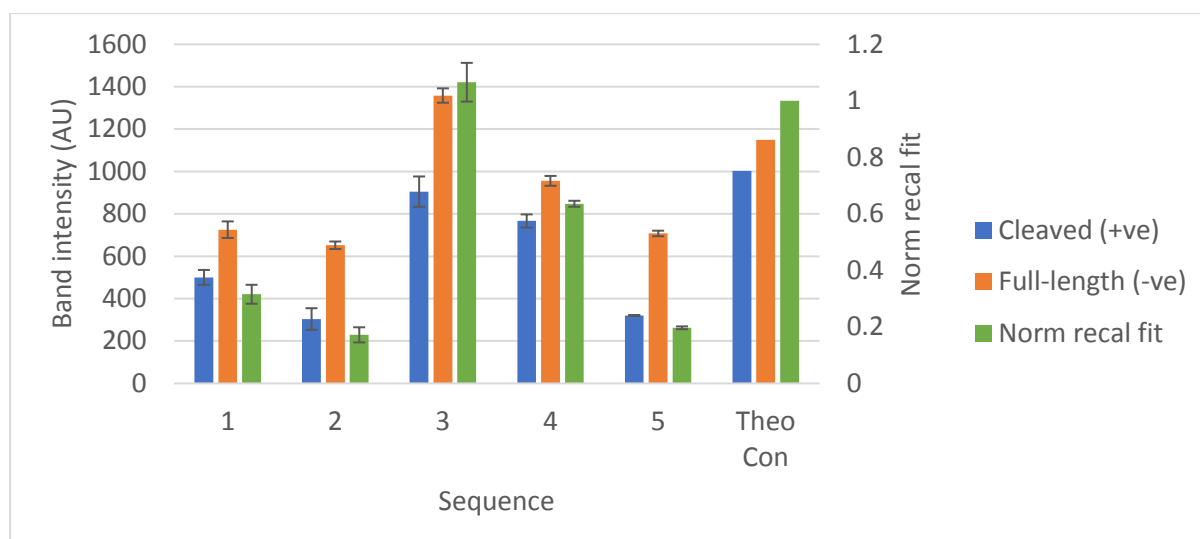


*Figure 5-12: Performance and recalculated fitness for selected sequences – The quantity of cleaved (blue) and full-length (orange) cDNA generated under positive & negative selection conditions was determined using densitometry to yield band intensities. The multiplication product of these two values was used to recalculate fitness (refer to Equation (5-2)). The resulting fitness values were subsequently normalised to that of the Theophylline Control (green). Error bars calculated from 3 repeats.*

To explain the result in Figure 5-12, factors that affect the rate of transcription were first considered. One factor affecting the rate of transcription is the length of a sequence. This affects the rate of transcription as shorter sequences have shorter elongation times and hence can be synthesised at faster rates. Another factor which affects the rate of transcription is the nucleotide composition of a sequence. As an example, the T7 RNAP has been shown to pause when encountering the “ATCTGTT” motif during transcription<sup>201</sup>.

Sequences 3, 4 & the Theophylline Control were analysed in the context of these two factors (Figure 5-13A). Sequence 4 was included as compared to Sequence 3 it has similar cleavage rates (Figure 5-11) but a weaker fitness (Figure 5-12). Therefore, differences between this sequence and Sequence 3 could help explain whether Sequence 3 was transcribed with higher efficiency. Figure 5-13A, shows that all the sequences considered are the same length. Investigating the composition of these sequences, none of the sequences contained the “ATCTGTT” motif (data not shown) and there were

only 4 positions where different nucleotide identities were noted. These were positions 64, 68, 73 & 74. At each of these positions, at least two of these sequences have identical identities. For instance, at position 64 both Sequence 3 and the Theophylline Control contain a cytosine identity. Given similar lengths and compositions, it is unlikely that the transcriptions rates for these sequences were significantly different from one another.

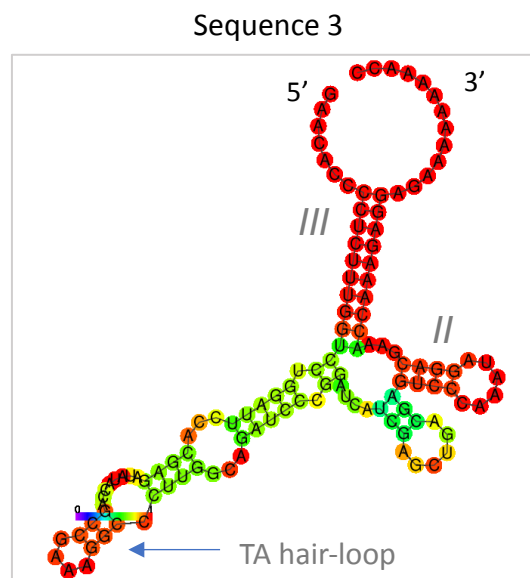
The fact that transcription rates for these sequences are likely to be similar suggests that the full-length RNA from Sequence 3 was reverse transcribed with greater efficiency. To identify a potential molecular explanation for this phenomenon, the MFE 2° structure of the full-length RNA for each of these sequences was analysed (Figure 5-13B – D). From these structures, the full-length RNA corresponding to Sequence 3 contains nucleotides in stem-loops I and the core ribozyme domain which are less likely to base pair than those in the equivalent Theophylline Control or Sequence 4 molecules. This suggests that the 2° structure of Sequence 3 is weaker. Given the lower reaction temperature used during the TRT reaction, it is plausible that this factor improved the rate of reverse transcription for Sequence 3<sup>140</sup>, explaining its improved fitness.

From Figure 5-13B & D, it is also worth noting the differences in calculated MFE 2° structures for Sequence 3 & the Theophylline Control (Figure 5-13B & D). For instance, Sequence 3 contains a correctly base-paired stem-loop II and stem III which are absent in the Theophylline Control. The terminal hair-pin loop of the theophylline aptamer is also correctly base-paired in the case of Sequence 3. As such, it is possible that this sequence will have different properties relative to the Theophylline Control when analysed *in vivo*.

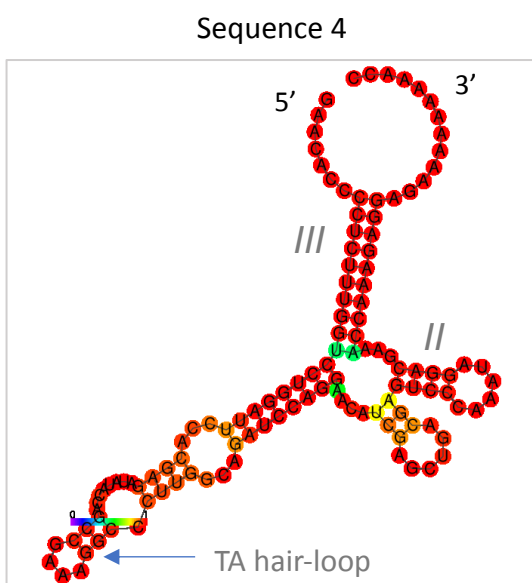
A

| Seq         | F RNA<br>(bases) | Identity at position |    |    |    |
|-------------|------------------|----------------------|----|----|----|
|             |                  | 64                   | 68 | 73 | 74 |
| 3           | 122              | C                    | U  | G  | A  |
| 4           | 122              | A                    | A  | G  | A  |
| Theo<br>Con | 122              | C                    | A  | U  | C  |

B



C



D

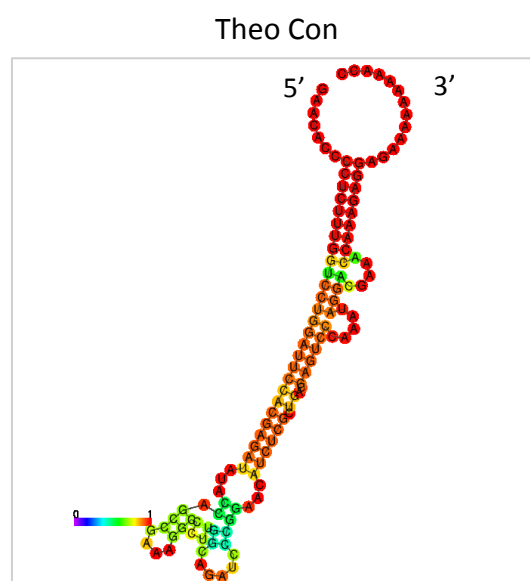


Figure 5-13: Properties of full-length (F) RNA for sequences 3, 4 & the Theophylline Control – (A) The length of each F RNA sequence and the identities of each sequence at positions 64, 68, 73 & 74. (B - D) MFE 2° structures for each full-length RNA sequence. Colours indicate base-pairing probabilities, according to the given scale. Roman numerals denote correctly formed stems/stem-loops, while a correctly formed hairpin-loop in the theophylline aptamer (TA hair-loop) is indicated.

### 5.3.6 Objective 6 – Characterisation of theophylline-activated riboswitches

Having characterised new theophylline-activated ribozymes, the ability of these and the Theophylline Control to function as riboswitches in *E. coli* was determined. Conducting this work was important for two reasons. Firstly, the reliability of *in vitro* methods to generate functional riboswitches has been actively questioned in recent years<sup>46,113</sup>. If any of the identified theophylline-activated ribozymes are shown to regulate gene expression in *E. coli*, this point of view would need to be reconsidered; particularly with regards to this host organism. Secondly, characterising similar theophylline-activated ribozymes which nevertheless have different properties may allude to variables which influence the functionality of ribozyme-based riboswitches. Given the Theophylline Control and Sequence 3 sequences have similar compositions but different 2° structures, the characterisation of these sequences may allude to the effect of this parameter on the functionality of ribozyme-based riboswitches.

In addition to characterising several theophylline-activated ribozymes, the functionality of constitutively active<sup>\*20</sup> and inactive<sup>\*21</sup> sequences from the Theophylline Library was also determined. These sequences are suspected of being theophylline-unresponsive given previous biochemical and structural data, in addition to data describing their dynamics during selection (Figure 5-5B). Characterising these sequences will provide a reference point from which to analyse the theophylline-activated sequences.

The ability of each of the above sequences to regulate gene expression in response to theophylline was to be assessed by expressing each in the 5' UTR of a GFP reporter gene. Specifically, it was envisaged that each sequence would regulate reporter gene expression as previously described<sup>53</sup> and illustrated in the Introduction chapter (Figure 1-4). To achieve this configuration, modifications to these sequences were first required. As illustrated in Figure 5-14A & B, the bait-oligo toehold & RT

---

<sup>\*20</sup>Constitutively active sequence is equivalent to that displayed in Chapter 3, Figure 3-2A.

<sup>\*21</sup>Inactive sequence contains a G81A mutation, inhibiting catalysis<sup>12</sup>.

primer landing pad sequences from the Theophylline Library were removed and replaced with a complete anti-RBS:RBS stem. It was hypothesised that in this configuration, the RBS would be more effectively masked, ensuring that gene expression was more strongly inhibited until cleavage and dissociation of the anti-RBS occurred. In addition to this modification, a 5'-GG, distal of the RBS:anti-RBS stem was included to ensure high yields when expressed in *E. coli* BL21(DE3) via the T7 promoter. This modification was specifically based on the required consensus sequence of the T7 promoter<sup>137</sup>.

As described in the Materials & methods chapter<sup>\*22</sup>, sequences equivalent to those in Figure 5-14B with a T7 promoter were cloned into a high copy number plasmid, upstream of a GFP reporter using BASIC DNA assembly<sup>202</sup>. Transformed cells containing the constructs were then incubated under theophylline concentrations spanning from 0 to 2.5 mM. GFP expression from individual cells was subsequently quantified using flow cytometry. From this data, unimodal populations across all samples were observed (Appendix, Section 9.4.3, Figure 9-1), implying that geometric means from this data could be used to compare responses. For each theophylline concentration, we normalised geometric mean values to those calculated from the constitutively active sequence to account for non-specific affects caused by varying theophylline concentrations (Figure 5-14C).

Relative to cells harbouring the constitutively active sequence, cells harbouring Theophylline Control, C63A variant and Sequence 3 sequences all increase GFP expression in response to increasing concentrations of theophylline. Overall, absolute GFP gene expression increases 5.8, 5.7 & 3.0-fold when incubated in the presence of 2.5 mM theophylline, respectively. It is also worth noting the similarity of the responses by Theophylline Control and C63A variant riboswitches. This suggests that a C63A substitution to the Theophylline Control does not affect its riboswitching activity.

In contrast to cells harbouring theophylline-activated ribozymes, those harbouring inactive sequences display low GFP expression levels regardless of the theophylline concentration. This result is in line with the expectation that this inactive ribozyme has a much lower cleavage rate if any, limiting gene

---

<sup>\*22</sup>refer to Section 7.6.2.2 for a detailed description of the protocol.

expression. Conversely, cells harbouring constitutively active ribozymes display much higher GFP expression levels. For all the theophylline concentrations tested, GFP expression is higher than that achieved by cells harbouring the Theophylline Control which differs from this sequence by a single G81A substitution. Similarly, this result is in line with our expectation that this ribozyme has faster cleavage rates than the Theophylline Control and as such produces higher levels of GFP expression.

One final observation from Figure 5-14C is the relatively high level of GFP expression produced by Sequence 3. For instance, the level of GFP expression for this sequence at 0 mM theophylline is 3.5-fold higher than it is for the Theophylline Control. However, based on the data from Figure 5-11, both these sequences appear to have similar cleavage rates *in vitro* under 0 mM theophylline. Furthermore, at theophylline concentrations above 0.63 mM, GFP expression from cells harbouring Sequence 3 exceeds that from all the others, including those harbouring the constitutively active ribozyme. This result is curious given the *in vitro* data from Figure 5-11 suggests Sequence 3 should have a slower cleavage rate compared to the Theophylline Control under these conditions. Potential underlying reasons to explain this phenomenon are reviewed in the following Discussion Section.

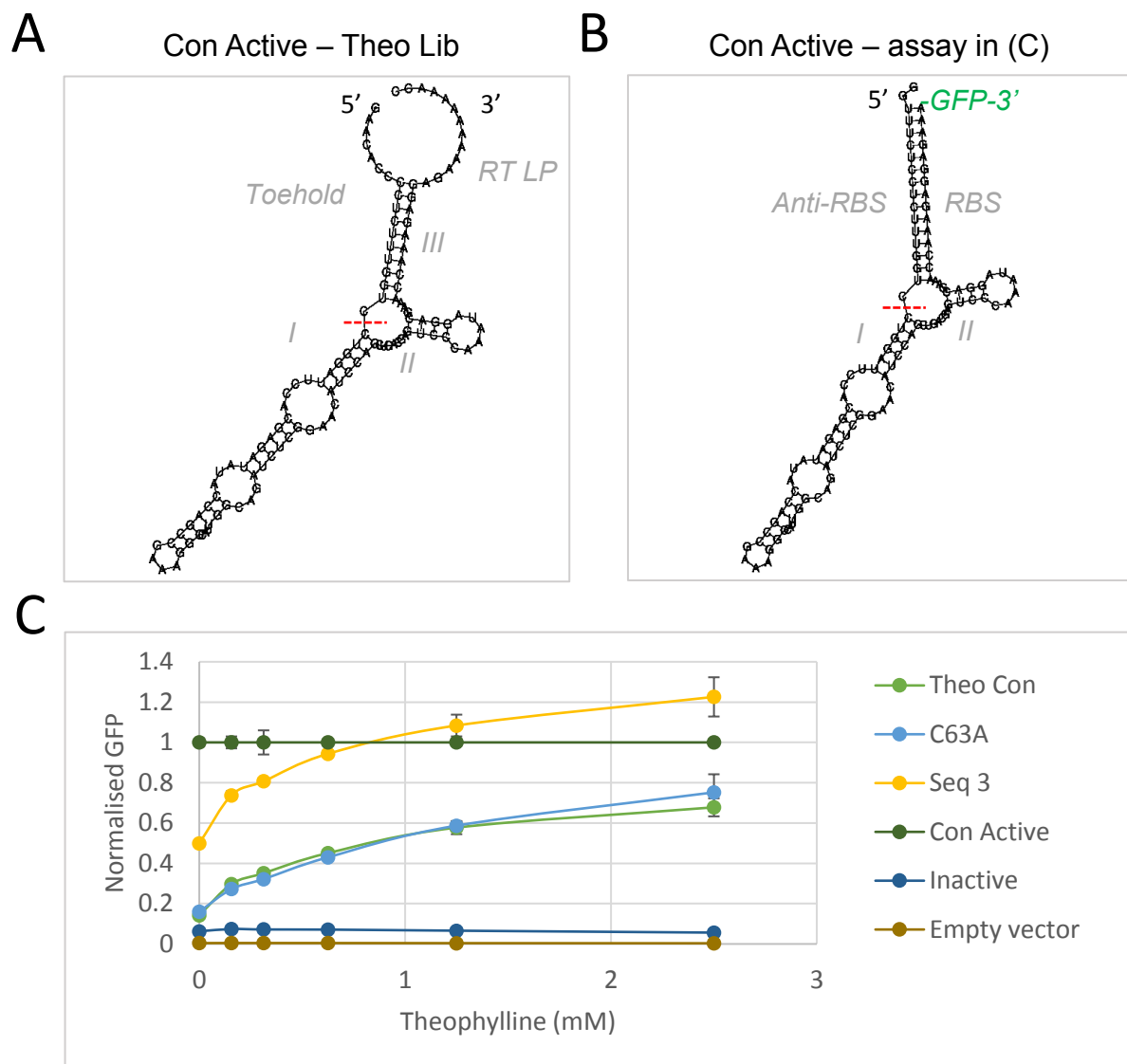


**Figure 5-14: Ability of theophylline-activated ribozymes to regulate gene expression in *E. coli* – RNA sequence & 2° structure of a constitutively (Con) active ribozyme from (A) Theophylline Library and (B) equivalent to that assayed in (C). Roman numerals denote correctly formed stems. The scissile bond (red-dotted line), Anti-RBS, RBS, bait-oligo toehold, RT primer landing pad (RT LP) and GFP loci are indicated. (C) Cells harbouring an empty vector or expression cassettes containing Theophylline Control (Theo Con), C63A variant, Sequence 3, constitutively active or inactive sequences were analysed under variable theophylline (0 – 2.5 mM). Geometric mean fluorescence analysed via flow cytometry was normalised to that determined from the constitutively active sequence. Error bars indicate standard deviations from 2 repeats.**

## 5.4 Discussion

### 5.4.1 Next-generation sequencing of the selection experiment

To achieve the remaining objectives set out in this chapter, DNA templates produced during the selection of theophylline-activated sequences from the Theophylline Library were sequenced using Illumina NGS technology. While data of sufficient quality and quantity was generated to fulfil the remaining objectives, a large proportion of the clusters produced during sequencing failed. This meant the quantity of data produced was roughly an order of magnitude lower than what was expected. Based on the accurate quantification of the NGS sample and the predominant absence of contaminating shorter and longer sequences, it was suggested that the reason for this low yield was the high homogeneity of the pool. However, other reasons relating to the equipment cannot be ruled out.

While sufficient data for this experiment was generated, had the analysed selection experiment involved a larger library, the functionality of sequences identified may have suffered. This is because less data would have been available to characterise dynamics and frequencies of key sequences. Given sample homogeneity was suggested as being an issue, strategies to address this include increasing the percentage of PhiX spike-in and lowering the density of clusters<sup>196</sup>. However, both these strategies come at the expense of the number of reads generated which correspond with the sample.

Therefore, should nucleotide diversity be identified as the issue affecting data yield further strategies might be considered. One possible solution would be to include a degenerate region in the sense primer used to prepare the sample. Including a heterogeneous 25 base region in the sense primer prior to the amplicon should reduce the number of failed clusters as the 1<sup>st</sup> 25 cycles of read 1 are used to determine whether a cluster fails or passes built-in filters<sup>196</sup>. Although this would reduce the maximum length by 25 bases, the impact of this effect is likely to be minimal given the general size of ribozymes and the current ability to sequence up to 600 bases using Illumina technology<sup>191</sup>.

### 5.4.2 Theophylline Library results

To determine the accuracy of the NGS data, three conclusions previously arrived at from Sanger sequencing data were investigated. One of these conclusions was that the Theophylline Control and C63A variant were enriched from the Theophylline Library. A second conclusion was that inactive sequences from the Theophylline Library were selected against. Finally, it was predicted from Sanger sequencing data that the overall frequency of Theophylline Library sequences reduced during selection. In all three instances, the results from the NGS and Sanger sequencing data agreed, suggesting the selection experiment was accurately summarised by this data.

The dynamics of sequences deemed inactive or constitutively active in the Theophylline Library were further investigated using this data. It was shown that both these sets of sequences had selection dynamics which matched their predicted phenotypes. It was also observed that the constitutively active sequences were enriched at a faster rate during the first positive selection compared to the Theophylline Control. This suggested that these sequences had faster cleavage rates during positive selection compared to the Theophylline Control.

### 5.4.3 Tolerance of insertion and deletion mutations

Measuring the lengths of all sequences and those recorded prior to selection, it was shown that the lengths of sequences in the pool became more diverse. This suggested that insertion and deletion mutations occurred during selection and were tolerated by the method. The distribution of sequence lengths also matched that predicted by the bioanalyzer trace. This provided further support that the NGS data accurately captured information about the sample sequenced.

Selecting ligand-responsive ribozymes on the basis of gel purification limits the number of insertion and deletion mutations that can occur by at most a couple of base pairs<sup>66,68,69,72,178</sup>. In contrast the distribution of lengths recorded following selection almost spanned the maximum range detectable by the Illumina NGS platform used. This suggests that the method can sample a much larger region of the sequence space than previous methods. Such a property could be considered advantageous in the

development of novel ligand-responsive ribozymes; particularly when large regions of the sequence space must be covered as is likely to be the case for identifying novel ribozyme-based riboswitches with new ligand specificities<sup>124–126</sup>. However, it is also noted that the ability to sample larger regions of the sequences space is likely to increase the propensity for sequences which can cheat selection to evolve.

#### 5.4.4 Accuracy of selection simulations

The accuracy of the simulation describing the dynamics of the Theophylline Control during selection was also investigated. It was shown that this simulation was accurate up to and including the 3<sup>rd</sup> cycle of selection.

Assuming the C63A variant had an identical phenotype to the Theophylline Control, an accurate simulation describing the dynamics of both these species together up to the 4th cycle was also generated. However, the dynamics of these sequences during the 5<sup>th</sup> cycle could not be accurately modelled. The presence of other sequences recorded towards the end of selection which had improved selective advantages compared to these sequences provides one explanation for this observation. This was because such sequences were not accounted for by the model of selection. In conclusion, this data suggests that accurate simulations using the model of selection are possible so long as the majority of sequences are accounted for. This was largely the case for the first 3 to 4 cycles of selection.

In the previous chapter selection simulations describing the enrichment of the Theophylline Control from libraries of variable sizes assumed all other variants were ligand-unresponsive and had the highest fitness values attainable. Based on the data collected in this chapter, the accuracy of these simulations will depend on how accurate the assumption is that all the remaining sequences present during selection can be summarised by this one phenotype. This is further complicated as based on the results of this chapter and previous studies<sup>203</sup>, sequences will eventually evolve which develop mechanisms to cheat selection.

With that said, by conducting more challenging control selection experiments, it should be possible to determine the frequency with which cheaters such as these arise and their respective fitness'. This knowledge can then be built into the model such that these sequences are controlled for during simulations. Furthermore, modifications to the selection method which reduce the propensity for the enrichment of cheaters will improve the accuracy of simulations based under the current assumptions. Such modifications are considered in Section 5.4.6.2 of this discussion.

### 5.4.5 Sequence 3

By analysing the NGS data generated, 5 sequences were selected from the pool at the end of selection. Based on their dynamics during the 5<sup>th</sup> cycle of selection and their recorded frequencies, these sequences were suspected of having strong theophylline-activated phenotypes. Of these 5 sequences, data was presented to suggest Sequence 3 has a similar fitness to the Theophylline Control under the selection conditions. Based on the properties of Sequence 3 compared to similar sequences, it was suggested that this effect was in part due to weaker 2° structure in the full-length RNA species which was more easily reverse transcribed during selection. Sequence 3 has not been previously characterised and nor was it present in the Theophylline Library<sup>72</sup>. As such, its identification demonstrates the ability of the selection method to evolve sequences with similar or improved fitness's under the selection conditions through directed evolution.

### 5.4.6 Cheaters

During this chapter, several sequences were identified which were either shown to cheat selection or were strongly implicated in doing so. The types of cheaters identified along with their impact in this chapter are first reviewed. Subsequently, potential mechanisms these sequences use to cheat selection are discussed along with potential solutions to inhibit these mechanisms. Finally, an alternative and potentially better methodology for estimating the fitness of sequences in the pool at the end of selection is suggested.

#### 5.4.6.1 Types of cheaters and their implications

For the purposes of this analysis, cheaters are defined as sequences which achieve higher levels of enrichment during selection based on mechanisms not related to their cleavage responses. Some of these sequences could be recognised from their nucleotide composition, reported through NGS analysis and as such were considered as obvious cheaters. Several other sequences which appeared to cheat selection were not identified until their phenotypes were empirically measured. These two types of cheaters are discussed separately in the remainder of this section.

The dynamics of obvious cheaters during selection was estimated based on the frequency of the theophylline aptamer in the pool. From this data it was shown that following an initial increase in the frequency of these sequences, their frequency did not increase beyond approximately 25 % of the pool for the remaining portion of the experiment. As such, these cheaters did not come to dominate the pool, inhibiting the enrichment and identification of functional sequences such as the Theophylline Control. However, further work is required to determine whether these sequences would prevent the enrichment of functional sequences when more diverse libraries or cycles of selection are implemented. That said, so long as these sequences do not come to dominate the pool, the algorithm implemented to remove these sequences from further analysis should limit the impact of these sequences. Given also that this algorithm does not require information regarding the aptamer or communication domain<sup>\*23</sup>, this approach should function as a general method, irrespective of the ligand-responsive ribozyme being selected.

Unfortunately, this algorithm did not identify four sequences which were subsequently shown to cheat the 5<sup>th</sup> cycle of selection. Based on the empirically analysed phenotypes of these sequences, they were able to cheat the 5<sup>th</sup> cycle of selection, being enriched to a higher degree than they should otherwise have been. The inability to in the first instance prevent the enrichment of these sequences

---

<sup>\*23</sup>Refer to the Appendix, Section 9.4.1 for the algorithm and parameters required.

and then in the second instance to prevent their being selected for further characterisation was undesirable given the time and resources invested.

It is worth first noting that given to the best of our knowledge previous ribozyme *in vitro* selection methods have not been analysed in the same depth with the use of NGS, it is difficult to make direct comparisons between this method and those implemented previously. Indeed, previous selection methods characterised enriched sequences with the use of Sanger sequencing methods. Based on this metric, it is likely that the majority of sequences identified by this method would have been functional<sup>72</sup>. That said, given this methodology explores wider regions of the sequence space and relies on the enzymatic rather than physical identification of sequences using PAGE, it is conceivable that the propensity for the formation of cheaters is higher compared to previous methods. Nevertheless, with a better understanding of the mechanisms cheaters exploit, it may be possible to prevent their enrichment.

#### 5.4.6.2 Potential mechanisms and solutions to cheating

As suggested in previous studies<sup>121</sup>, biases in the numerous process that are required during selection may aid the enrichment of cheaters. One form of biases are mutational biases which may arise from the properties of several enzymes used during the process. For instance, the RTase used during selection has been shown to favour G to A mutations and resist G to C and T to G mutations<sup>156</sup>. Furthermore, PCR steps have been shown to favour the introduction of pyrimidines over purines<sup>157</sup>. It is possible that mutational biases such as these have favoured the production of cheaters from other relatively more frequent sequences such as the Theophylline Control. The propensity for mutational biases would be particularly high towards the end of selection due to the high frequency of enriched sequences in the pool.

In addition to general mutational bias, several of the processes of selection are likely to favour sequences with specific characteristics. For instance, the reverse transcription step of the TRT reaction was shown to favour sequences with weaker RNA 2° structures such as Sequence 3. This is because

such sequences were more efficiently reverse transcribed. The ligation step of the selection method may also favour sequences based on 2° structure. This is because ligation efficiencies can be inhibited by 2° structure elements which inhibit hybridisation with splint molecules during ligation<sup>204</sup>.

It is also conceivable that cheaters are utilising mechanisms which bypass the process of selection altogether. One hypothetical strategy to avoid the process of selection would involve the same RNA molecule yielding different cDNA molecules depending on the selection conditions. In this configuration the majority of cDNA produced would be ligated to regenerate the starting template. One phenomenon sequences may exploit to cheat in this manner is the template switching property of the RTase<sup>205</sup>. This property enables the elongating cDNA strand to switch to other RNA or even DNA templates. Perhaps the obvious cheaters identified during this chapter are efficient at cheating in this manner. It is plausible that during the reverse transcription of these shorter sequences, they can template switch to other longer sequences. Cheaters using this phenomenon could rely on the functionality of other sequences in the pool by copying their responses.

As a solution to this problem, the TRT reaction could be conducted in an emulsion with individual templates separated in droplets. This would prevent sequences present within the pool from influencing the functionality of one another. Given both transcription<sup>206</sup> and reverse transcription<sup>156,207</sup> reactions have been compartmentalised in this manner, it should be feasible to do the same for the TRT reaction. A reduction in throughput as a result of compartmentalisation<sup>180,181</sup> might be avoided by using many templates per droplet during the initial cycles of selection. This would maximise throughput for the first few cycles when the library has high integrity and each variant is at a relatively low copy number. The number of templates per droplet could then be gradually reduced to inhibit processes such as template switching, increasing selection robustness as the pool converges.

It might also be possible to address issues associated with the ligation reaction. For example, the ligation reaction buffer could be optimised to improve the specificity of this step<sup>175</sup>. This would function to reduce the chance of ligating incorrect sequences. Another modification would be to pre-

anneal the splint and adapter prior to ligation and avoid denaturing these species during the reaction. This would prevent the adapter annealing in other configurations. Such configurations might be conducive to the ligation of cheaters.

Furthermore, it might be possible to increase the ligation reaction temperature by changing the splint or adapter sequences. Higher temperatures would inhibit the formation of competing 2° structure elements which as mentioned could affect ligation reaction efficiencies. To increase the ligation reaction temperature during negative selection, sequences upstream of the T7 promoter could be included in the library design enabling a longer splint and Adapter- interaction and hence  $T_m^{*24}$ . To increase the reaction temperature during positive selection, the ligand-binding domain could be moved from stem-loop I to stem-loop II facilitating a longer splint and cleaved cDNA interaction.

#### 5.4.6.3 Improving the estimation of fitness at the end of selection

Unfortunately, even implementing procedures such as those described above, it is likely to be impossible to completely avoid biases in some of the fundamental processes of selection. As such cheaters are likely to occur at some frequency during this and other similar directed evolution experiments<sup>203</sup>. In this regard it is of the utmost importance that such sequences can be readily identified following selection so that time and resources are not wasted further characterising their phenotypes.

In this chapter, the fold-change of sequences caused by the 5<sup>th</sup> cycle of selection in addition to the frequency of each sequence recorded at the end of selection was used to estimate fitness. However, a more accurate estimation of fitness might be obtained by acquiring more NGS data and considering the expression in Equation (5-2). Achieving a more accurate estimation of fitness should better ensure only functional sequences are selected for further analysis.

---

<sup>\*24</sup>Refer to Chapter 2, Section 2.3.6.2, Figure 2-19 for an illustration of the relevant oligonucleotides.

To calculate the fitness for a sequence according to Equation (5-2), three parameters are required. The first is the frequency of the sequence ( $x_i$ ). This parameter can be calculated for sequences at the end of selection using the NGS methods and downstream analysis described in this chapter. The other two parameters are the number of cleaved cDNA molecules yielded under positive selection conditions ( $C_i^{(+)}$ ) and the number of full-length cDNA molecules yielded under negative selection conditions ( $F_i^{(-)}$ ).

To calculate these two additional parameters for many sequences enriched following selection, the relevant cDNA molecules must also be sequenced using NGS methods. This could be achieved by first generating those cDNA molecules under the relevant selection conditions. Subsequently, the desired cDNA molecule would be ligated to a biotinylated NGS adapter using a splint and adapter combination that anneals in a homologous manner to that illustrated for Ligase mediated selection (Figure 2-19). Purifying the ligation reaction product as described during selection would allow the relevant cDNA molecules to be isolated prior to their specific amplification and sequencing. Subsequently quantifying the relevant sequences from the NGS data set would yield all three parameters and hence the fitness of many sequences enriched during the selection process according to Equation (5-2).

In addition to enabling calculations of fitness, this data may also indicate whether incorrect sequences are being ligated during selection. Furthermore, sequences which cheat through other mechanisms should be identifiable by comparing their fitness' to their fold-changes during selection. Such data is therefore likely to be useful to identify large numbers of cheaters, enabling better hypotheses regarding the mechanisms they use to cheat selection and hence developing more effective solutions to limit their enrichment in future experiments.

#### 5.4.7 Characterisation of theophylline-activated riboswitches

To achieve the final objective of this chapter, the ability of several theophylline-activated ribozymes characterised throughout this work to function as ribozyme-based riboswitches in *E. coli* was tested. The results of this experiment showed that all sequences activated by theophylline *in vitro*, were able

to activate GFP expression in response to increasing concentrations of this ligand in *E. coli*, demonstrating that each sequence was able to act as a ribozyme-based riboswitch in *this organism*. This result and others<sup>55,121</sup> suggests that there is some relationship between the functionality of sequences under bespoke selection conditions and their functionality *in vivo*.

The results from this experiment also showed that the C63A variant was able to regulate gene expression in an almost identical manner to the Theophylline Control sequence. This validates the methods ability to identify equivalently functional ribozyme-based riboswitches from the Theophylline Library given that both these sequences were enriched to similar degrees during selection.

Both the Theophylline Control and C63A variant sequences were able to activate gene expression approximately 6-fold in response to 2.5 mM theophylline. This is close to the 10-fold activation under 4 mM theophylline previously reported for a similar theophylline-activated ribozyme-based riboswitch<sup>53</sup>. This implies that with respect to fold-changes in gene expression, these riboswitches are unlikely to be significantly improved with minor changes to sequence composition. This in turn suggests that close-to optimal ribozyme-based riboswitches which function in *E. coli* can be generated using only *in vitro* methods.

As has been noted previously in this Chapter<sup>\*25</sup>, a sequence similar to the Theophylline Control was previously reported as being unable to regulate gene expression in mammalian cells<sup>72</sup>. Specifically, this sequence and the Theophylline Control only differ in the identity of residues within stem III. While the influence stem III sequences cannot be ruled out, there are fundamental differences between mammalian and bacterial gene expression that could explain why the Theophylline Control was able to regulate gene expression in *E. coli* but may not be able to in mammalian cells.

---

<sup>\*25</sup>Refer to Section 5.1.2.

One difference which could explain this result is the differences in mRNA degradation rates between these two host organisms. Specifically, mRNA half-lives in mammalian cells are much longer, averaging approximately 3 - 4 hours compared to approximately 5 minutes in *E. coli*<sup>208,209</sup>. Significantly, the kinetics of the Theophylline Control\*<sup>26</sup> suggest that the longer this ribozyme persist without inactivation, the closer the cleavage responses will be in the presence and absence of theophylline. As such, the longer half-life of this ribozyme in mammalian cells should negatively influence its functionality.

Another difference between mammalian and *E. coli* host cells which could explain this phenomenon is differences in protein actuators involved in mediating gene expression. Specifically, there are many more protein actuators involved in manipulating mRNA in mammalian cells compared to *E. coli*<sup>210</sup>. These actuators could influence the fold and functionality of the ribozyme domain. In comparison, processes such as mRNA transport and splicing that occur in mammalian gene expression are absent in *E. coli*. Furthermore, expression using the T7 RNAP should improve the probability that the ribozyme functions more similarly to the way it does *in vitro*. This is because the processes of transcription by the T7 RNAP and translation by the ribosome are uncoupled<sup>211</sup>. As such, the ribozyme domain should have more time to fold prior to perturbations from the ribosome or other mRNA-binding proteins.

With that said, it was found that relative to Sequence 3, the Theophylline Control produced a lower level of GFP expression in all cases even though *in vitro* data suggested that the Theophylline Control cleaved at a faster rate at high theophylline concentrations. This result is in contradiction to the point of view that gene expression completely correlates with *in vitro* cleavage rates.

There are several potential reasons which could explain this phenomenon. One potential reason is that the cleavage rates for Sequence 3 are different relative to the Theophylline Control when analysed *in vivo*. Other potential reasons could be related to differences in RNA degradation rates for any of the RNA species involved. As an example, the cleaved RNA species for Sequence 3 could be

---

\*<sup>26</sup>Refer to Chapter 3, Section 3.3.2.4, Figure 3-4A.

degraded at a slower rate or perhaps the cleaved anti-RBS is more easily degraded, increasing the rate of translational initiation following cleavage. A further possibility is that full-length RNA translational initiation rates are higher for Sequence 3 than the Theophylline Control. Perhaps the weaker 2° structure of the full-length RNA sequence from Sequence 3 is more easily displaced by the small ribosomal subunit, leading to increased background translation rates.

To help identify the underlying reason for this phenomenon, it should be possible to determine whether this reason is intrinsic to the process of catalysis or extrinsic. To this end, an inactive variant of Sequence 3 can be generated in a similar manner to the Theophylline Control through a G81A substitution. Comparing the response of this inactive Sequence 3 to the inactive Theophylline Control should determine whether the underlying reason is intrinsic to catalysis or extrinsic to this process. If the underlying process is extrinsic to catalysis, the inactive Sequence 3 should yield a higher level of GFP expression compared to the inactive Theophylline Control. This is because the underlying process should affect both the inactive and active versions of these sequences to the same extents.

In addition to characterising an inactive version of Sequence 3, Sequence 4 could also be characterised in *E. coli*. Evidence was presented in this chapter to suggest this sequence had similar rates of cleavage *in vitro* but a more rigid 2° structure. As such, comparing the functionality of this sequence with Sequence 3 may inform on the influence of this variable.

Determining the influence of variables such as 2° structure on the functionality of these ribozyme-based riboswitches should enable better predictions regarding the functionality of sequences selected *in vitro* for their ability to function as ribozyme-based riboswitches. It may also be possible given this information to propose modifications to selected sequences to improve performances or to perhaps identify residues which could be targeted for mutagenesis, enabling further screening. This should function to improve the propensity with which *in vitro* selected sequences are translated into optimal ribozyme-based riboswitches.

## 5.5 Summary

The first aim of this chapter was to analyse the previously conducted control experiment using NGS technology so that several objectives could be met. The quantity and quality of data generated using this technology was sufficient to achieve all the objectives set out. Furthermore, recommendations were made which could improve the quantity of data yielded in future experiments. The accuracy of this data was validated by proving several previously shown and expected results. Following this, the ability of the selection method to explore more diverse regions of the sequence space compared to other methods was demonstrated. In accordance with the fourth objective of this chapter, the accuracy of simulations describing the dynamics of theophylline-activated sequences during selection was investigated. The high accuracy of these simulations for most of the selection experiment suggest that the model developed during this thesis might yield useful predictions for future experiments. The usefulness of this model is likely to increase significantly if the types of sequences which are present during selection are better understood. In accordance with the fifth objective of this chapter, two not-previously characterised theophylline-activated sequences which evolved during selection were identified. One of these sequences had an equivalent fitness to the Theophylline Control under the selection conditions. In addition to this sequence, sequences empirically shown to cheat selection or strongly implicated in doing so were identified. Some of these sequences could be readily identified and did not appear to influence the enrichment or characterisation of functional sequences. The remaining cheaters were not identified until after their phenotypes had been empirically measured and as such were more insidious. Potential methods to better identify and limit the enrichment of cheaters in future experiments were discussed. To meet the last objective of this Chapter, two theophylline-activated ribozymes identified during this work along with the Theophylline control were analysed for their ability to regulate gene expression in *E. coli*. All three of these sequences were empirically shown to function as ribozyme-based riboswitches in this host organism. While all these sequences were functional, relative differences in gene expression could not be completely accounted for by relative *in vitro* cleavage rates. Further characterisation of these and similar sequences is

expected to reveal the variables which account for these differences, aiding the translation of *in vitro* selected sequences into functional ribozyme-based riboswitches.

## Chapter 6: Conclusion

### 6.1 Introduction

The property of ribozyme-based riboswitches to regulate the activity of downstream genes in response to ligand concentrations means these components have several potential applications. These include applications in the clinic such as regulating the activity of oncolytic viruses<sup>90,94,95</sup> as well as in the fields of metabolic and enzyme engineering by optimising the production of commercially relevant molecules<sup>75,77</sup>.

The development of functional ribozyme-based riboswitches with new ligand specificities could lead to more examples of these applications while also improving the potential to achieve existing applications. For instance, detecting other commercially relevant metabolites *in vivo* using these sensors would enable the production of these metabolites to be optimised. Furthermore, the ribozyme-based riboswitches suggested for clinical applications currently require cytotoxic concentrations of ligand for maximum activation<sup>90</sup>. As such, ribozyme-based riboswitches which respond to less toxic ligands should enable higher concentrations of ligand to be administered during treatment, potentially yielding a more effective therapeutic.

To date, the development of ribozyme-based riboswitches with new ligand specificities has proved challenging. This has mainly been due to prior characterisation of aptamers using SELEX and the limitations associated with this strategy. The development of a ribozyme *in vitro* selection method which requires fewer resources and can be implemented in shorter times was hypothesised to address this issue, leading to the characterisation of ribozyme-based riboswitches with new ligand specificities.

### 6.2 Findings and their implications

In testing this hypothesis, it was first necessary to develop an *in vitro* selection method which had the required specifications. Subsequently this methodology could be implemented to attempt to identify

novel ribozyme-based riboswitches with new ligand specificities. Within this section, the findings reported throughout this thesis are discussed in the context of these two objectives.

### 6.2.1 Development of an *in vitro* selection method with the required properties

To develop a ribozyme *in vitro* selection method which can be implemented in shorter times with fewer resources, an alternative scheme of selection was formulated. The benefit of using this scheme of selection is that it could be implemented without the requirement of PAGE purification steps. This enabled all steps of the method to be implemented in 96-well format, enabling greater economies of scale and hence potentially fewer resources.

Given also that these PAGE purification steps were previously conducted with overnight incubations<sup>55,72</sup>, the removal of these steps led to a 50 % reduction in the time required to implement rounds of selection. It was also noted that the techniques utilised during selection are compatible with liquid-handling robots<sup>128,129,146,150</sup>. As such, further time savings and easier scaling in future experiments might be achieved by automating parts of or all of the methodology.

In addition to requiring less time and fewer resources, further examination of the method revealed it could have several additional advantages. One of these advantages again derives from the avoidance of PAGE purification steps. Given the avoidance of PAGE purification steps, large insertion and deletion mutations which otherwise are selected against are instead tolerated. The tolerance of these mutations enables more diverse mutations to occur during selection and hence a wider region of the sequence space to be explored. This property is likely to be desirable when attempting to characterise ribozyme-based riboswitches with new ligand specificities given these sequences are likely to be rare within the sequence space<sup>124–126</sup>.

While the method demonstrated several desirable properties, it was also found that several ligand-unresponsive sequences were able to cheat selection, being enriched during the latter rounds. This property is undesirable for two reasons. Firstly, these sequences take up a portion of the pool, effectively lowering the number of functional sequences that can be examined. Secondly, as

demonstrated, it can complicate the identification of functional sequences from those that have been enriched. However, given to the best of our knowledge previous methods have not been analysed in the same depth with the use of NGS, it is difficult to make direct comparisons between this method and those implemented previously. Indeed, the sequence most commonly identified using previously used Sanger sequencing methods was functional and the second most frequent sequence identified was also highly unlikely to be a cheater<sup>72</sup>. That said, given this methodology explores wider regions of the sequence space and relies on the enzymatic rather than the physical identification of sequences using PAGE, it is conceivable that the propensity for the formation of cheaters is higher compared to previous methods.

In addition to the direct development of this methodology, a model describing the underlying process of selection was formulated during this work. As demonstrated in this thesis, it is possible to use this model to estimate optimal selection conditions and the dynamics of sequences during selection. This knowledge is expected to be desirable when planning and implementing future experiments given that sequences could be enriched using fewer cycles and the resources required for enrichment estimated beforehand.

### 6.2.2 Identification of ribozyme-based riboswitches with new ligand specificities

To attempt to identify a ribozyme-based riboswitch with new ligand specificities, the method developed in this thesis was used to screen a library for the presence of ribozymes responsive to the amino acid Trp. To the best of our knowledge, a Trp-responsive ribozyme has not yet been characterised, implying that the identification of such a sequence could have led to the fabrication of a ribozyme-based riboswitch with new ligand specificities. Unfortunately, following five cycles of selection the pool did not display a Trp-responsive phenotype, suggesting Trp-responsive ribozymes

were not enriched to detectable levels. While additional cycles may have identified such a sequence, as discussed previously<sup>\*27</sup> there could be multiple reasons for this result.

While ribozymes responsive to new ligands were not presented during this thesis, several theophylline-activated ribozymes not previously reported were identified. These include Sequence 3 & 4 and a variant of the Theophylline Control containing a C63A substitution.

In addition to identifying three new theophylline-activated ribozymes, two of these sequences along with the Theophylline Control were tested for their ability to regulate gene expression in *E. coli*. All three of these sequences were able to function as ribozyme-based riboswitches in *E. coli*, activating GFP expression in response to theophylline. This result is significant given that a sequence analogous to the Theophylline Control had previously been reported as being non-functional when tested in mammalian cells<sup>72</sup>. This study is often cited as an example of why engineering ribozyme-based riboswitches *in vitro* is unreliable<sup>46,113</sup>. In contrast, the result presented in this thesis and by others<sup>55,121</sup> suggests that there is likely to be some relationship between the cleavage responses of sequences during selection and their cleavage responses *in vivo*. It should however be noted that the data presented suggested that *in vitro* cleavage responses do not completely define the functionality of a sequence *in vivo*. This was highlighted by comparing the functionality of Sequence 3 & the Theophylline Control.

However, it is expected that with a better understanding of the variables which effect gene expression, it should be possible to bridge the gap between *in vitro* cleavage responses and *in vivo* functionality. Bridging this gap should make the challenge of identifying ribozyme-based riboswitches with new ligand-specificities easier. This is because *in vitro* selection methods such as that described in this thesis, are less reliant on the prior characterisation of compatible aptamer sequences<sup>66,68,69,123</sup>. The

---

<sup>\*27</sup>Refer to Chapter 4, Section 4.4.1.

identification of such sequences appears to be the key challenge in developing riboswitches with new ligand specificities<sup>116,117</sup>.

### 6.3 Recommendations for future work

One important lesson from this work is that selection experiments can be made significantly easier if you know what you're selecting for and how long selection takes. This is highlighted during this thesis by the work done prior to enriching several theophylline-activated sequences from the Theophylline Library. Without this prior work, it is conceivable that these sequences would not have been enriched. Based on this conclusion, before undertaking a future selection experiment, thought should be given to the parameters which would yield a functional ribozyme-based riboswitch. Parameters which should be considered include the concentration of ligand sensed by the resulting riboswitch and the rates of cleavage under selection conditions. With these parameters it should be possible to estimate optimal selection conditions and the time it takes to enrich sequences based on the model described in this work.

It is noted that estimations of desirable cleavage rates could prove difficult. This is because data presented in this work and by others<sup>121</sup> suggests *in vitro* cleavage rates do not directly translate into levels of gene expression, perhaps due to the influence of other variables e.g. RNA 2° structure<sup>212</sup>. That said, with a better understanding of all the variables which influence gene expression, it should be possible to provide a range of target cleavage rates to aim for. Furthermore, as discussed in Chapter 5, a better understanding of the other variables which influence gene expression might be obtained by investigating several of the sequences characterised in this work in greater detail.

In addition to understanding what needs to be selected, there are several improvements to the methodology itself which could improve its future implementation. Many of these improvements have been suggested throughout the individual discussion sections of the results chapters. For example, automation of this method was suggested as a future improvement in Chapter 2 to further increase the scalability of the method while reducing the resources required. In addition to these previously

mentioned improvements, the use of more “*in vivo-like*” selection conditions should also be considered. An example would be the use of macro-molecular crowding agents in selection buffers. These components affect RNA folding and structure and have been used to better approximate the *in vivo* environment<sup>213,214</sup>.

Future selection experiments should also take full advantage of the scalability of this method. In addition to screening many candidate ligands, the screening of many libraries in parallel could also be considered. Perhaps candidate aptamer sequences could be inserted into stem-loop II instead of stem-loop I or perhaps in both stem-loops. Furthermore, rather than using motifs previously characterised by SELEX to guide the design of the aptamer domain, random or structured sequences<sup>152</sup> could be used to explore other regions of the sequence space in parallel. A previously described hybrid approach<sup>123</sup> could also be used to generate libraries for screening. The advantage of this approach is that aptamers which bind the ligand in the context of the ribozyme sequence can first be identified using SELEX.

Other more ambitious library designs could also be considered. For instance, libraries aimed at modifying the ligand specificity of the *glmS* ribozyme<sup>\*28</sup> might be generated. The benefit of characterising a ligand-responsive ribozyme which functions in a similar manner to this sequence is that it might have improved properties compared to current synthetic ribozymes<sup>215</sup>. This is because in contrast to the ligand-responsive hammerhead ribozymes mentioned throughout this work, the ligand is directly involved in catalysis, acting as a cofactor.

## 6.4 Closing remarks

To achieve a key objective of this thesis an *in vitro* selection method was developed which selects ligand-responsive ribozymes in 96-well format and can be implemented in shorter periods of time.

---

<sup>\*28</sup>Refer to the Introduction chapter, Section 1.2 for a description of the *glmS* ribozyme.

Furthermore, the avoidance of size selection steps enables larger insertion and deletion mutations during selection, resulting in a larger exploration of the sequence space.

The functionality of this methodology was demonstrated through the enrichment of theophylline-activated ribozymes from a library containing a large proportion of undesirable sequences. To achieve this result, a model describing the process of selection was developed. This model enabled novel selection conditions to be identified which resulted in the enrichment of functional sequences in practical periods of time. The model also accurately predicted the dynamics of key sequences for most of this experiment.

Analysing the result of this experiment using NGS meant several theophylline-activated ribozymes were reported which have not been previously characterised. Two of these sequences evolved during selection, empirically demonstrating the ability of the method to evolve and enrich functional sequences through directed evolution. While potential issues were identified from this analysis, as discussed, the techniques used throughout this thesis provide a framework to further characterise and limit the impact of these issues.

Two of the theophylline-activated ribozymes identified by this work along with a sequence similar to a previous theophylline-activated ribozyme were tested for their ability to regulate gene expression in *E. coli*. In contrast to the expectations of previous studies, all three of these sequences were able to regulate gene expression in this organism, suggesting there is some relationship between the *in vitro* performances and *in vivo* functionalities of sequences for this organism. Further work is expected to elucidate a better description of this relationship and in doing so bridge the gap between *in vitro* selections and ribozyme-based riboswitch development. It is anticipated that achieving this aim will lead to the generation of novel ribozyme-based riboswitches, enabling a range of applications.

## Chapter 7: Materials & methods

### 7.1 Computational methods

#### 7.1.1 Programming languages

MATLAB scripts were run in R2018a or earlier versions. Bash scripts and programmes requiring a Linux environment were run in ubuntu 16.04 LTS.

#### 7.1.2 RNA folding

Individual MFE 2° structures for RNA sequences were generated using the RNAfold WebSever available at <http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>. Where parameters from large numbers of sequences were required, the Vienna RNA v2.4.4 programme was run within ubuntu 16.04 LTS.

#### 7.1.3 Nucleic acid thermodynamic properties and GC contents

Using the MATLAB function “oligoprop”, the GC content of sequences in addition to the  $\Delta G$  upon hybridisation were calculated. In the latter case, the parameters required for this calculation were calculated by averaging those from several previous reports<sup>134,216,217</sup>.

Duplex T<sub>m</sub> values were calculated according to the model proposed by Honda and co-workers<sup>134</sup>. To correct T<sub>m</sub> values for divalent cations the model developed by Walder and co-workers was used<sup>218</sup>.

#### 7.1.4 Densitometry and % cleaved calculations

To measure the quantity of nucleic acids separated by electrophoresis, the intensity of the corresponding bands was measured using TOTALLAB CLIQS 1D Gel Image Analysis software. Background subtraction was performed using the rolling ball method.

When quantifying molar amounts of HEX-labelled nucleic acids, the band intensity was directly used. Otherwise band intensity was divided by the MW of the relevant nucleic acid.

% cleaved values for RNA or cDNA were calculated from this data as follows:

$$\% \text{ cleaved} = \frac{100C}{C + F} \quad (7-1)$$

where  $C$  and  $F$  are the molar amounts of cleaved and full-length species, respectively.

## 7.2 Experimental methods

### 7.2.1 Preparation of dsDNA templates from synthetic oligonucleotides

Apart from the Trp Libraries described in the following paragraph, all dsDNA templates were prepared from the relevant synthetic oligonucleotides given in Table 7-6 using PCR, as described in Section 7.2.8 or 7.2.9.

Trp Libraries containing RT primer landing pads 2, 5 & 14 were prepared using a protocol similar to that previously described<sup>131</sup>. Briefly HHRv2.4.1\_Large and either MH\_v2.4.4D, E or F were annealed and extended using the Klenow fragment<sup>219</sup>, yielding libraries containing RT primer landing pads 2, 5 or 14, respectively. The resulting dsDNA was purified using GenElute™ PCR Clean-Up kit spin-columns and resuspended in TE buffer. Note, the DNA library containing RT primer landing pad 14 is equivalent to the Trp Library used throughout this thesis.

### 7.2.2 RNA synthesis

IVT reactions were assembled under the following conditions: 40 mM Tris-HCl (pH 7.9), 2 mM spermidine, 8.4 mM MgCl<sub>2</sub>, 2 mM each rNTP, 25 ng/μL DNA template, 1 U/μL NEB RNase Inhibitor Murine; 5 mM DTT & 5 U/μL NEB T7 RNA Polymerase.

Where Trp or theophylline were present, these ligands were used at the specified concentrations. All IVT reactions were incubated at 37 °C. To generate the results prior to Figure 3-4, all IVT reactions were incubated for 3 hours. Following this figure, all IVT reactions were incubated for 10 minutes.

### 7.2.3 Purification of RNA from IVT reactions

To purify RNA from IVT reactions, each reaction was resuspended in 1/10<sup>th</sup> volume 10X NEB DNase I buffer and 1/10<sup>th</sup> volume NEB DNase I. The reaction was incubated on ice for 15 minutes followed by

a 5-minute incubation at 37 °C. 0.1 mM. EDTA pH 7.5 was subsequently added to a final concentration of 17.6 mM and the reaction snap-cooled by heating at 75°C for 10 minutes followed by a 4-minute incubation on ice. Note that RNA separated via urea PAGE following Chapter 3, Section 3.3.1 was not purified. Instead the unpurified IVT reaction was analysed.

To generate results prior to Chapter 2, Section 2.3.4, the remaining RNA was purified using ZYMO RESEARCH RNA Clean & Concentrator™ -25 spin-columns. Following this section SPRI beads were used as described in Section 7.2.3.1 to purify RNA from IVT reactions.

#### 7.2.3.1 SPRI-bead based purification of RNA

To generate SPRI beads capable of purifying the relevant RNA molecules a similar protocol to that used to generate SPRI beads for cDNA purification (Section 7.2.5) was used with the following modifications:

- Pelleted beads were washed with 1 mM Sodium citrate (pH 6.4), 0.05 % Tween 20 instead of 1x TET buffer.
- 10 mM Tris-HCl (pH 8.0) was substituted with 1 mM Sodium citrate, pH 6.4 (M) in the solution used to resuspend the beads following washing.

The resulting beads were then used as described in Section 7.2.5.

### 7.2.4 cDNA synthesis

#### 7.2.4.1 Reverse transcription of purified RNA

Reverse transcription was performed using Protoscript® II RTase from NEB in volumes of 20 µL. The recommended reaction conditions were used along with 5 µM RT primer and up to 2 ug RNA. RT reactions were incubated at 42 °C for 30 minutes and the RTase inactivated at 65 °C for 20 minutes.

#### 7.2.4.2 TRT reaction and biotinylated DNA template removal

All TRT reactions were assembled in volumes of 20 µL under the following conditions: 40 mM Tris-HCl (pH 7.9), 2 mM spermidine, 10.4 mM MgCl<sub>2</sub>, 75 mM KCl, 2 mM each rNTP, 0.5 mM each dNTP, 25

ng/ $\mu$ L DNA template, 2.5  $\mu$ M RT primer or RT primer(HEX), 1 U/ $\mu$ L NEB RNase Inhibitor Murine; 5 mM DTT, 5 U/ $\mu$ L NEB T7 RNA Polymerase, 10 U/ $\mu$ L NEB Protoscript<sup>®</sup> II RTase.

Where Trp or theophylline were present, these ligands were used at the specified concentrations. All TRT reactions were incubated at 37 °C. Unless otherwise specified, TRT reactions were incubated for 30 minutes before inactivation at 65 °C for 20 minutes.

When implementing the selection method, the biotinylated DNA template was removed. This was achieved by adding each reaction to Dynabeads<sup>™</sup> MyOne<sup>™</sup> Streptavidin C1 washed according to the manufacturer's instructions and resuspended in 20  $\mu$ L 2x B&W buffer. For the results in Chapter 3, Section 3.3.3.1, 10  $\mu$ L of dynabeads was used. For the remainder of this thesis, 5  $\mu$ L of dynabeads were used. The biotinylated DNA was immobilised by incubating at room temperature for 15 minutes on a rocker set at 40 rpm. The dynabeads were separated using a magnet for 3 minutes and the supernatant retained. Where 5 instead of 10  $\mu$ L dynabeads were added initially, the process was repeated twice for a total of three pull-downs.

#### 7.2.5 Non-specific purification of cDNA

For the results generated prior to Chapter 2, Section 2.3.4, cDNA was desalted using BIO-RAD Micro Bio-Spin<sup>™</sup> P-6 Gel Columns in Tris Buffer. For the remainder of this thesis, cDNA was purified using SPRI beads as described in the following paragraphs. Unless otherwise stated and where required, any remaining RNA was digested by adding RNAase A to a final concentration of 0.1 mg/mL and incubating at 37 °C for 30 minutes.

To prepare SPRI beads for the non-specific purification of cDNA, commercially available SPRI beads (e.g. Agencourt AMPure XP) were resuspended in an equivalent volume of 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 1 M NaCl, 39 % (w/v) PEG 8000 and 0.05 % Tween 20. To achieve this the commercially available beads were first pelleted and washed twice with 1.5x the volume of 1x TET buffer before being resuspended in a buffer devoid of PEG 8000 & Tween 20. These two latter components were subsequently added with vigorous mixing and/or vortexing.

To purify cDNA using the above beads a similar protocol to the manufacturers was used. Briefly, 1.8x the volume of the above SPRI beads was added and the reaction mixed. The reaction was then incubated for 5 minutes and the beads pelleted using a magnet. The supernatant was aspirated and the beads washed twice with 70 % EtOH. The beads were then dried for up to 5 minutes and resuspended in aqueous solution. Following re-pelleting of beads, the purified cDNA in the supernatant was transferred to a clean tube.

## 7.2.6 DNA ligase mediated purification of full-length or cleaved cDNA

DNA ligase mediated purification was conducted on cDNA samples purified non-specifically as described in Section 7.2.5. For the results presented prior to Chapter 3, Section 3.3.3.2, T7 DNA Ligase was used as outlined in Section 7.2.6.1 below. For the remaining results of thesis, *Tth* DNA ligase was used as outlined in the following Section (7.2.6.2). In both cases, the ligation reaction product was purified as described in Section 7.2.6.3.

### 7.2.6.1 Ligation of selected cDNA using T7 DNA Ligase

To ligate full-length or cleaved cDNA to Adapter- or Adapter+, respectively, the cDNA was first annealed to the Splint in the presence of the relevant adapter. This was achieved by combining the following components from Table 7-1 and conducting an annealing reaction: Adapter(+/-), Splint, Purified cDNA, T7 DNA Ligase Annealing Buffer and H<sub>2</sub>O.

To initiate the ligation reaction, the remaining components of the ligation reaction (Table 7-1) were added and the reaction incubated at room temperature for 30 minutes.

| Component                      | Final concentration/volume  |
|--------------------------------|-----------------------------|
| Adapter(+/-)                   | 2.5 $\mu$ M                 |
| Splint                         | 2.5 $\mu$ M                 |
| Purified cDNA                  | 12 $\mu$ L                  |
| T7 DNA Ligase Annealing Buffer | 1x                          |
| H <sub>2</sub> O               | Variable                    |
| ATP                            | 1 mM                        |
| PEG 6000                       | 7.5%                        |
| NEB T7 DNA Ligase              | 150 U/ $\mu$ L              |
| <b>Total volume</b>            | <b>20 <math>\mu</math>L</b> |

Table 7-1: Components of T7 DNA Ligase reactions

#### 7.2.6.2 Ligation of selected cDNA using Tth DNA Ligase

A similar protocol to that described in the previous section (7.2.6.1) was used to anneal full-length or cleaved cDNA to Adapter- or Adapter+. Specifically, the following reagents from Table 7-2 were combined prior to annealing: Adapter(+/-), Splint, Purified cDNA, NEB *Taq/Tth* DNA Ligase Reaction Buffer and H<sub>2</sub>O specified.

To initiate the ligation reaction, the NEB *Taq/Tth* DNA Ligase (refer to Table 7-2) was added on ice and the reaction incubated at 51.7 °C or 46.4 °C for 30 minutes. The reaction was then transferred to ice and purified immediately as described in Section 7.2.6.3.

| Component                                     | Final concentration/volume  |
|-----------------------------------------------|-----------------------------|
| Adapter(+/-)                                  | 1 $\mu$ M                   |
| Splint                                        | 1 $\mu$ M                   |
| Purified cDNA                                 | 12 $\mu$ L                  |
| NEB <i>Taq/Tth</i> DNA Ligase Reaction Buffer | 1x                          |
| NF-H <sub>2</sub> O                           | Variable                    |
| NEB <i>Taq/Tth</i> DNA Ligase                 | 1.6 U/ $\mu$ L              |
| <b>Total volume</b>                           | <b>20 <math>\mu</math>L</b> |

Table 7-2: Components of *Tth* DNA Ligase reactions – N.B. The NEB *Taq* DNA Ligase is referred to as NEB *Taq/Tth* DNA Ligase to indicate that it was originally characterised from *Thermus thermophilus* strain HB8<sup>220</sup> and to align with the rest of this document.

#### 7.2.6.3 Pull-down and purification of ligation reaction products

Ligation reactions generated in Sections 7.2.6.1 or 7.2.6.2 were added to Dynabeads™ MyOne™ Streptavidin C1 washed according to the manufacturer's instructions and resuspended in 20  $\mu$ L 2x B&W buffer. Where NEB *Taq/Tth* DNA Ligase was used, this buffer was supplemented with an additional 20 mM EDTA to inhibit the ligase during immobilisation. The following formulae was used to calculate the volume of Dynabeads™ MyOne™ Streptavidin C1 ( $V_D$ ) used during immobilisation:

$$V_D = \frac{A}{5} V_L$$

where  $A$  and  $V_L$  are the concentration of the adapter and the volume of the ligation reaction, respectively.

The immobilisation reaction was incubated at room temperature for 15 minutes on a rocker set at 40 rpm. The Dynabeads were separated using a magnet for 3 minutes and the supernatant aspirated.

To remove non-specifically bound oligonucleotides the following protocol was used: The Dynabeads were washed once with 100  $\mu$ L 1x SSC buffer, resuspended in 40  $\mu$ L 0.15 M NaOH and incubate at room temperature for 10 minutes. Following this, the beads were washed once with 80  $\mu$ L 0.1 M NaOH; once with 100  $\mu$ L 1x B&W buffer and once with 100  $\mu$ L TE buffer. The now purified ligation

reaction product could be resuspended in TE buffer and amplified via PCR as described in Section 7.2.8 or 7.2.9.

### 7.2.7 Preparation of cDNA for analysis via urea PAGE

To prepare free cDNA in solution for analysis via urea PAGE, all samples were non-specifically purified according to Section 7.2.5. Following purification, samples were dried using a Speed Vac® Plus AR and resuspended in 1x RNA loading buffer.

To prepare cDNA immobilised on streptavidin-coated beads, beads were resuspended in 95% formamide, 10 mM EDTA and heated at 65 °C for 5 mins. An equal volume of 0.025% (w/v) bromophenol blue, 0.025% xylene cyanol, 0.025% (w/v) SDS was added, effectively resuspending the sample in 1x RNA loading buffer for subsequent analysis via urea PAGE.

### 7.2.8 General PCR conditions

PCRs implemented using PFU polymerase were assembled as described in Table 7-3 and incubated in a thermocycler as described in Table 7-4.

| Component          | Final concentration |
|--------------------|---------------------|
| PFU buffer         | 1-fold              |
| dNTPs              | 200 µM each         |
| Sense3(bio)*       | 0.5 µM              |
| Anti3(bio)*        | 0.5 µM              |
| Template           | Variable            |
| PFU DNA polymerase | 2.8 µg/mL           |

*Table 7-3: Components of PFU polymerase PCRs – \*Apart from the PFU polymerase PCRs conducted in Chapter 2, Section 2.3.5, Sense3(bio) and Anti3(bio) primers were used.*

| Step                 | Temp (°C) | Time (mins) | Number of cycles |
|----------------------|-----------|-------------|------------------|
| Initial denaturation | 98        | 1           | 1                |
| Denaturation         | 98        | 0.5         | Variable**       |
| Annealing            | 65*       | 0.5         |                  |
| Extension            | 72        | 1           |                  |
| Final Extension      | 72        | 5           | 1                |

*Table 7-4: PCR conditions for those assembled as described in Table 7-3 – \*All primers were annealed at 65 °C except for PCR primer set 2 (Sense2 & Anti2), which were annealed at 57.7 °C. \*\*10 – 20 cycles were often conducted to yield sufficient material.*

### 7.2.9 Semi-quantitative PCR

PCRs were first assembled as described in Table 7-3 and then incubated for 10 cycles as described in Table 7-4 with a final extension of 2 minutes. PCRs were then held at 50 °C and a 20 µL sample removed and the concentration of DNA quantified. If the concentration of DNA exceeded 2.24 ng/µL<sup>\*29</sup>, the remaining number of cycles required to achieve a yield of 25.1 ng/µL was calculated using the expression below:

$$\log\left(\frac{25.1}{DNA_{(10)}}\right) / \log(1.65)$$

where  $DNA_{(10)}$  is the concentration of DNA after 10 cycles in ng/µL. Otherwise if the concentration of DNA was below this value, a further 5 cycles was conducted and the concentration of DNA remeasured. This was repeated until the concentration of DNA exceeded 2.24 ng/µL enabling the remaining number of cycles to be determined.

### 7.2.10 Quantification of nucleic acid in solution

dsDNA was quantified using the Qubit™ dsDNA BR Assay Kit or a NanoDrop 1000 Spectrophotometer. All remaining nucleic acids were quantified using a NanoDrop 1000 Spectrophotometer.

---

<sup>\*29</sup>(*LOD* – refer to Chapter 3, Section 3.3.3.2.2)

### 7.2.11 PCR purification

Apart from the results presented in Figure 2-14, all PCRs using PFU DNA polymerase were purified using GenElute™ PCR Clean-Up spin-columns according to the manufacturer's instructions.

### 7.2.12 Electrophoresis

#### 7.2.12.1 Urea PAGE

Gels were run using either the Sequi-Gen® GT Nucleic Acid Electrophoresis Cell or the BIO-RAD Mini-PROTEAN Tetra Cell. 8 % polyacrylamide, 8M urea gels were pre-warmed to approximately 50 °C and samples were denatured prior to loading by resuspending in 1x RNA loading buffer and heating at 70 °C for 10 minutes.

#### 7.2.12.2 Agarose gel electrophoresis

dsDNA templates were separated using 2 % agarose in 1x TBE buffer.

### 7.2.13 Staining and imaging of gels

For the results generated following Figure 3-4, Chapter 3, all RNA in urea PAGE gels was stained using SYBR™ Green II. Prior to this result, all RNA was stained using EtBr. dsDNA in agarose gels was stained using SYBR™ Green I.

#### 7.2.13.1 HEX fluorescence

The fluorescence from HEX-labelled cDNA was measured using a 532 nm laser and Cy3 (/570DF20) filter on a Fujifilm FLA 5000 imager.

#### 7.2.13.2 EtBr staining and imaging

Where gels were only stained using EtBr, staining involved a 30-minute incubation in 0.5 µg/mL EtBr in 1x TBE buffer followed by a 15 minutes incubation in H<sub>2</sub>O. Where detection of HEX-labelled nucleic acids was required prior to EtBr staining, gels were imaged first for HEX fluorescence before being subsequently fixed for 45 minutes in 10 % V/V acetic acid and stained using EtBr as above.

Imaging of EtBr fluorescence was conducted in an identical manner to that described for HEX fluorescence.

#### 7.2.13.3 SYBR™ Green II Staining and imaging

Gels were stained with SYBR™ Green II according to the manufacturer's recommendations. Following staining gels were imaged using the relevant settings on a Fujifilm LSA-3000 imager.

#### 7.2.13.4 SYBR™ Green I staining and imaging

dsDNA in agarose gels was stained with SYBR™ Green I according to the manufacturer's recommendations. Following staining agarose gels were imaged as described in the previous section for SYBR™ Green II.

## 7.3 Reagents

### 7.3.1 Buffers

| Buffer                            | Components                                                                                                                                                      |
|-----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| TE buffer                         | 10mM Tris-HCl (pH 8.0), 0.1 mM EDTA                                                                                                                             |
| 1x RNA loading buffer             | 50 % deionized formamide, 0.0125 % bromophenol blue, 0.0125% xylene cyanol, 0.0125 % SDS, 5 mM EDTA (pH 8.0), 50 mM NaCl                                        |
| 1x Annealing buffer               | 100 mM NaCl, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA.                                                                                                                |
| 2x B&W buffer                     | 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl.                                                                                                                   |
| 1x T7 DNA Ligase annealing buffer | 66 mM Tris:HCl (pH 7.5), 10 mM MgCl <sub>2</sub> , 1 mM DTT                                                                                                     |
| 1x TET buffer                     | 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.05 % Tween 20                                                                                                             |
| 10x PFU buffer                    | 200mM Tris-HCl (pH 8.8 at 25°C), 100mM KCl, 100mM (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 20mM MgSO <sub>4</sub> , 1.0% Triton® X-100 and 1mg/ml BSA. |

Table 7-5: Buffers

### 7.3.2 Chemically synthesised oligonucleotides

Chemically synthesised oligonucleotides were ordered from either Integrated DNA Technologies, Inc. or from Twist Bioscience as described in Table 7-6 & Table 7-7, respectively.

| ID      | Sequence                                        | IDT purification code |
|---------|-------------------------------------------------|-----------------------|
| Adapter | /5Phos/GACCAAAGAGGGGTGTTCTATAGTGAGTCGTATTA/3C6/ | HPLC                  |

|                        |                                                                                                                                                    |      |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|------|
| Adapter-               | /5Phos/TATAGTGAGTCGTATTA/3Bio/                                                                                                                     | HPLC |
| Adapter+               | /5Phos/GACCAAAGAGGGGTGTTCTATAGTGAGTCGTATTA/3Bio/                                                                                                   | HPLC |
| Anti1                  | GGTTTTTTTTCTCCTCTTTGGTTTCGTCTATTTGGG                                                                                                               | STD  |
| Anti2                  | GGTTTTTTTTCTCCTCTTTGGTTTC                                                                                                                          | STD  |
| Anti3                  | GGTTTTTTTTCTCCTCTTTGGTTTCGTCTA                                                                                                                     | STD  |
| Anti3bio               | /5Biosg/GGTTTTTTTTCTCCTCTTTGGTTTCGTCTA                                                                                                             | STD  |
| Bait-oligo             | /5Biosg/GAACACCCCTCTTTGGTC/3C6/                                                                                                                    | PAGE |
| HHRv2.4.<br>1_5T_Large | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACNNRRGACCGNNNCGCYACYNNN<br>GGTACATCCAGCTGATGAGT                                                       | STD  |
| MH_Illum<br>Anti       | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTTTTTTTTCTCCTCTTTGGTTTCGTCTA                                                                                   | STD  |
| MH_Illum<br>Sense      | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAATACGACTCACTATAGAACACCCCTCTTTG                                                                                  | STD  |
| MH_v2.4.<br>4D_small   | GGTGGTTTCTCCTCTTTGGTTTCGTCTATTTGGGACTCATCAGCTGGATGTACC                                                                                             | STD  |
| MH_v2.4.<br>4E_small   | GGGTTTTTCTCCTCTTTGGTTTCGTCTATTTGGGACTCATCAGCTGGATGTACC                                                                                             | STD  |
| MH_v2.4.<br>4F_small   | GGTTTTTTTTCTCCTCTTTGGTTTCGTCTATTTGGGACTCATCAGCTGGATGTACC                                                                                           | STD  |
| RT primer              | GGTTTTTTTTCT                                                                                                                                       | STD  |
| RT<br>primer(HE<br>X)  | /5HEX/GGTTTTTTTTCT                                                                                                                                 | HPLC |
| Sense1                 | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCAC                                                                                                     | STD  |
| Sense2                 | TAATACGACTCACTATAGAACACCCCTCTTTGGTCC                                                                                                               | STD  |
| Sense3                 | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTC                                                                                                        | STD  |
| Sense3bio              | /5Biosg/TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTC                                                                                                | STD  |
| Sequence<br>1          | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGCCGGT<br>TGGCAGATCACGGAACATCCCGTGACGAGTCCCAAATAGGACGAAACCAAAGAGGAGAAAAAAAAC<br>C | STD  |
| Sequence<br>2          | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGCCGGT<br>TGGCAGATCTCGGAACATCCCGTGACGAGTCCCAAATAGGACGAAACCAAAGAGGAGAAAAAAAAC<br>C | STD  |
| Sequence<br>3          | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGCCCTT<br>GGCAGATCCCGGATCATCGAGCTGACGAGTCCCAAATAGGACGAAACCAAAGAGGAGAAAAAAAACCC    | STD  |
| Sequence<br>4          | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGCCCTT<br>GGCAGATCCAGGAACATCGAGCTGACGAGTCCCAAATAGGACGAAACCAAAGAGGAGAAAAAAAACCC    | STD  |

|                      |                                                                                                                                                    |      |
|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|------|
| Sequence 5           | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGGCCTT<br>TGGCAGATCTCGGAACATCCCGCTGACGAGTCCCAAATAGGACGAAACCAAGAGGAGAAAAAAAAC<br>C | STD  |
| SP74                 | CAAGAAAACCCACGCCACCTAC                                                                                                                             | STD  |
| Splint               | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCAC/3C6/                                                                                                | HPLC |
| Theophylline Control | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGGCCTT<br>GGCAGATCCCGGAACATCTCGCTGACGAGTCCCAAATAGGACGAAACCAAGAGGAGAAAAAAAAC       | STD  |
| Theophylline Library | TAATACGACTCACTATAGAACACCCCTCTTTGGTCCTGGATTCCACGAGATATACCAGCCGAAAGGCCNTT<br>GGCAGATCNCGGAACATCNGGCTGACNAGTCCCAAATAGGACGAAACCAAGAGGAGAAAAAAAAC       | STD  |
| SEVA_T0_rev          | GGACCCCTGGATTCTCACC                                                                                                                                | STD  |
| SEVA_T1_for          | GGCGGCGGATTTGTCCTAC                                                                                                                                | STD  |

**Table 7-6: Chemically synthesised oligonucleotides acquired from Integrated DNA Technologies, Inc. –**

The IDT purification code for each sequence is given. “/5Phos/”, “/5HEX/”, “/5Biosg/” & “/3C6/” denote the presence of a 5' Phosphate group, 5' HEX, 5' standard biotin or 3' Hexanediol group, respectively.

| ID              | Sequence                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Theo Con part   | TCTGGTGGGTCTCTGTCCGTGCCTACTCTGGAAAATCTTAATACGACTCACTATAGGTTTCTCCTCTTTGGTCCTGGATT<br>CCACGAGATATACCAGCCGAAAGGCCCTTGGCAGATCCCGGAACATCTCGCTGACGAGTCCCAAATAGGACGAAACCA<br>AAGAGGAGAAATAGTGGAGACCTATCG                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| C63A part       | TCTGGTGGGTCTCTGTCCGTGCCTACTCTGGAAAATCTTAATACGACTCACTATAGGTTTCTCCTCTTTGGTCCTGGATT<br>CCACGAGATATACCAGCCGAAAGGCCCTTGGCAGATCACGGAACATCTCGCTGACGAGTCCCAAATAGGACGAAACCA<br>AAAGAGGAGAAATAGTGGAGACCTATCG                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Seq3 part       | TCTGGTGGGTCTCTGTCCGTGCCTACTCTGGAAAATCTTAATACGACTCACTATAGGTTTCTCCTCTTTGGTCCTGGATT<br>CCACGAGATATACCAGCCGAAAGGCCCTTGGCAGATCCCGGATCATCGAGCTGACGAGTCCCAAATAGGACGAAACCA<br>AAAGAGGAGAAATAGTGGAGACCTATCG                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Con Active part | TCTGGTGGGTCTCTGTCCGTGCCTACTCTGGAAAATCTTAATACGACTCACTATAGGTTTCTCCTCTTTGGTCCTGGATT<br>CCACGAGATATACCAGCCGAAAGGCCATTGGCAGATCTCGGAACATCCAGCTGACGAGTCCCAAATAGGACGAAACCA<br>AAAGAGGAGAAATAGTGGAGACCTATCG                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Inactive part   | TCTGGTGGGTCTCTGTCCGTGCCTACTCTGGAAAATCTTAATACGACTCACTATAGGTTTCTCCTCTTTGGTCCTGGATT<br>CCACGAGATATACCAGCCGAAAGGCCCTTGGCAGATCCCGGAACATCTCGCTGACGAGTCCCAAATAGGACAAAACCA<br>AAGAGGAGAAATAGTGGAGACCTATCG                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| sfGFP part      | TCTGGTGGGTCTCTTAGTCCATGCGTAAAGGCGAAGAAGTGTTCACGGGCGTAGTTCGATTCTGGTCGAGCTGGACG<br>GCGATGTGAACGGTCATAAGTTTAGCGTTTCGCGGTGAAGGTGAGGGCGACGCGACCAACGGCAAAGTACCTGAAG<br>TTCATCTGCACCACCGGTAAAGTCCCGGTGCCTTGGCCGACCTTGGTGACGACGTTGACGTATGGCGTGACGTGTTTTG<br>CGCGTTATCCGGACCACATGAAACAACACGATTCTTCAAATCTGCGATGCCGGAGGGTTACGTCCAGGAGCGTACCA<br>TTTCCTTCAAGGATGATGGCACTTACAAAACCTCGCGCAGAGGTTAAGTTTGAAGGTGACACGCTGGTCAATCGATCG<br>AATTGAAGGGTATCGACTTAAAGAGGATGGTAACATTCTGGGCCATAAACTGGAGTATAACTCAACAGCCATAATG<br>TTTACATTACGGCAGACAAGCAAAAGAACGGCATCAAGGCCAATTTCAAGATTGCCACAATGTTGAGGACGGTAGCG<br>TCCAAGTGGCCGACCATTACCAGCAGAACACCCCAATTGGTGACGGTCCGGTTTTGCTGCCGGATAATCACTATCTGAG<br>CACCCAAAGCGTGCTGAGCAAAGATCCGAACGAAAAACGTGATCACATGGTCCTGCTGGAATTTGTGACCGCTGCGG<br>GCATACCCACGGTATGGACGAGCTGTATAAGCGTCCGTAATAAGGCTCGGAGACCTATCG |

**Table 7-7: Chemically synthesised oligonucleotides acquired from Twist Bioscience.**

## 7.4 Methods specific to Chapter 2

### 7.4.1 Bait-oligo mediated full-length cDNA selection

Synthesised cDNA was non-specifically purified, in a manner similar to that described in Section 7.2.5, however RNase A was added to a final concentration of 30 µg/mL.

Bait-oligo pull-down reactions were conducted by first diluting purified cDNA into 1x annealing buffer with 2.5 µM Bait-oligo. The following thermocycler programme was then conducted to anneal Bait-oligo to full-length cDNA:

- Incubate 95 °C For 2 mins.
- Cool reaction by 0.1 °C/second to 56.5 °C.
- Hold temperature for 5 mins.
- Store at 4 °C.

The reaction was then added to Dynabeads™ MyOne™ Streptavidin C1 which were washed according to the manufacturer's instructions and resuspended in an equal volume 2x B&W buffer. The annealed sample was added to the resuspended Dynabeads and the Bait-oligo immobilised by incubating at room temperature for 15 minutes. Dynabeads were then pelleted with a magnet for 3 minutes and the supernatant aspirated. The beads were washed 3x with 1x annealing buffer prior to further manipulation.

### 7.4.2 Bait-oligo mediated cleaved cDNA selection and adapter ligation

Synthesised cDNA was non-specifically purified, in a manner similar to that described in Section 7.2.5, however RNase A was added to a final concentration of 10 µg/mL.

To select cleaved cDNA from purified cDNA, a similar pull-down to that described in the previous section (7.4.1) was conducted. However, rather than aspirating the supernatant it was retained. To this supernatant an additional volume of Bait-oligo was added and the process repeated using an

annealing temperature of 65.2 °C rather than 56.5 °C to hybridise Bait-oligo to full-length cDNA. This process was repeated until four pull-down reactions were completed.

To ligate cleaved cDNA to the Adapter oligonucleotide, the protocol outlined in Section 7.2.6.1 was used except the Adapter oligonucleotide was used instead of Adapter+. Following ligation, the reaction was purified as described in Section 7.2.5 and the sample incubated at 65 °C for 10 minutes to ensure complete inactivation of the T7 DNA Ligase.

#### 7.4.3 SPRI-based purification of DNA from PCRs

PCRs were conducted by amplifying 18 ng of Trp Library DNA in 90 µL for 10 cycles using Sense3 & Anti3 primers<sup>\*30</sup>. Reactions were purified using AmpliClean™ Magnetic Bead-based PCR Cleanup beads according to the manufacturers protocol. Yields were calculated and compared to those achieved using GenElute™ PCR Clean-Up spin-columns.

#### 7.4.4 PCR primer sets 1, 2 & 3 evaluation

For each primer set, 60 ng of Trp Library DNA was amplified in a volume of 200 µL according to the protocol given in Section 7.2.87.2.6. The resulting PCRs were purified using GenElute™ PCR Clean-Up spin-columns and 600 ng of DNA analysed via urea PAGE.

### 7.5 Methods specific to Chapter 3

#### 7.5.1 Synthesis of cDNA without RNA purification

IVT reactions were first assembled and incubated for 10 minutes as described previously in Section 7.2.2. Following this, EDTA was added to a final concentration of 1.1 mM and the reaction incubated at 70 °C for 1 minute to inactivate the T7 RNAP. The reaction was subsequently transferred to ice and the components listed in Table 7-8 added. cDNA was synthesised at 42 °C for 30 minutes and the RTase inactivated at 65 °C for 20 minutes.

---

<sup>\*30</sup>PCRs were conducted under the conditions specified in Section 7.2.8

| Component                  | Final concentration |
|----------------------------|---------------------|
| Tris-HCl (pH 8.3)          | 16.7 mM             |
| KCl                        | 75 mM               |
| MgCl <sub>2</sub>          | 3.75 mM*            |
| RT primer(HEX)             | 2.5 µM              |
| DTT                        | 5 mM                |
| dNTPs                      | 0.5 mM each         |
| NEB Murine RNase inhibitor | 1 U/µL              |
| NEB Protoscript® II RTase  | 10 U/µL             |

*Table 7-8: Components added to unpurified IVT reaction to synthesise cDNA – \*Excludes the MgCl<sub>2</sub> present in the IVT reaction.*

## 7.5.2 Sanger sequencing

Biotinylated DNA templates corresponding with the Theophylline Library and the pool following 5 cycles of selection were separated via agarose gel electrophoresis and purified using the Omega Bio-tek E.Z.N.A.® Gel Extraction Kit according to the manufacturer's instructions. The purified DNA was blunt cloned into the pJET1.2/blunt Cloning Vector from Thermo Scientific and the resulting plasmid transformed into *E. coli* Dh5α cells. Transformed cells were incubated on carbenicillin agar plates and individual colonies picked and cultured overnight. Plasmid DNA from each culture was purified using the Omega Bio-tek E.Z.N.A.® Plasmid Mini Kit I. Samples were then sequenced by Source BioScience using the SP74 oligonucleotide as a sequencing primer.

## 7.6 Methods specific to Chapter 5

### 7.6.1 NGS

#### 7.6.1.1 Sample preparation

DNA templates following rounds of selection were prepared for NGS in a similar manner to that described previously<sup>197</sup>. Specifically, the amplicon PCR stage of this method was assembled as

illustrated in Table 7-9 and amplified as described in Table 7-10. Amplicon stage PCRs were then purified using 1.5x the volume of Agencourt AMPure XP beads. Index PCRs were then assembled as illustrated in Table 7-11 and amplified as described in Table 7-12. The resulting PCR was purified using 1.1x the volume of Agencourt AMPure XP beads. The resulting purified index PCR products were analysed via agarose gel electrophoresis and compared to the original DNA templates to confirm incorporation of adapter sequences (data not shown). To pool the indexed samples prior to sequencing, they were quantified using the Qubit™ dsDNA BR Assay Kit.

| Component                  | Final concentration |
|----------------------------|---------------------|
| NEB Phusion HF buffer      | 1-fold              |
| dNTPs                      | 200 µM each         |
| MH_IllumSense              | 0.2 µM              |
| MH_IllumAnti               | 0.2 µM              |
| DNA template               | 500 fg/µL           |
| NEB Phusion DNA polymerase | 0.04 U/µL           |

*Table 7-9: Amplicon PCR components for NGS sample preparation.*

| Step                 | Temp (°C) | Time (seconds) | Number of cycles |
|----------------------|-----------|----------------|------------------|
| Initial denaturation | 98        | 30             | 1                |
| Denaturation         | 98        | 10             | 20               |
| Extension            | 72        | 15             |                  |
| Final Extension      | 72        | 300            | 1                |

*Table 7-10: Amplicon PCR conditions for NGS sample preparation.*

| Component                  | Final concentration/volume |
|----------------------------|----------------------------|
| NEB Phusion HF buffer      | 1-fold                     |
| dNTPs                      | 200 µM each                |
| i7 Adapter*                | 5 µL                       |
| i5 Adapter*                | 5 µL                       |
| Purified Amplicon PCR      | 5 µL                       |
| NEB Phusion DNA polymerase | 0.04 U/µL                  |
| H <sub>2</sub> O           | To 50 µL                   |

Table 7-11: Index PCR components for NGS sample preparation – \*Index primers were used from Nextera XT Index Kit (FC-131-1001).

| Step                 | Temp (°C) | Time (seconds) | Number of cycles |
|----------------------|-----------|----------------|------------------|
| Initial denaturation | 98        | 30             | 1                |
| Denaturation         | 98        | 10             | 20               |
| Annealing            | 65        | 15             |                  |
| Extension            | 72        | 15             |                  |
| Final Extension      | 72        | 300            | 1                |

Table 7-12: Index PCR conditions for NGS sample preparation.

#### 7.6.1.2 Pooled library analysis and Illumina sequencing

The pooled library sample generated in the previous section was kindly analysed and then sequenced by individuals at the NIHR Imperial BRC Genomics Facility. Specifically, the sample was quantified via qPCR and the quality assessed using a Bioanalyzer. The sample was then sequenced using an Illumina MiSeq v2 150 PE micro run with a PhiX spike-in of 20 %, as requested.

#### 7.6.1.3 NGS data processing

Indexed reads identified from the sequencing run in the previous section were kindly demultiplexed and provided as “fastq.gz” files by individuals at the NIHR Imperial BRC Genomics Facility. Furthermore, a MultiQC report<sup>199</sup> of the sequencing run and a Flowcell Summary were kindly provided. For all reads, adapter sequences were removed using cutadapt version 1.16 and overlapping paired-end reads merged using PEAR v0.9.11. The remaining analysis was conducted in MATLAB including

sequence alignments where required. Briefly, unique sequences from the entire data set were identified. A matrix describing the dynamics of each unique sequence was subsequently generated by counting frequencies from amongst sequences recorded following each round of selection. This matrix was then used in calculations for % frequencies, fold-changes etc.

### 7.6.2 Characterisation of sequences in *E. coli* for theophylline-dependent GFP expression

Designing of the T7 promoter-ribozyme-RBS sequences was achieved with the aid of MATLAB. Except for advice relating to theophylline concentrations, all the remaining steps of this experiment were designed and implemented by Dr Marko Storch.

#### 7.6.2.1 Construction of expression cassettes

For each of the sequences assayed, an expression cassette containing the sfGFP part from Table 7-7 and one of the remaining parts was generated using BASIC DNA assembly<sup>202</sup>. Briefly, two parts from Table 7-7, including the sfGFP part were incubated under BASIC linker ligation reaction conditions in the presence of LMP and LMS linkers. Following insert assembly, this BASIC part was cloned into the AmpR-pUC19 backbone (also linker ligated) and transformed into *E. coli* DH5 $\alpha$  via heatshock. Picked clones were grown in LB+50  $\mu$ g/ml carbenicillin and incubated at 37 °C overnight. Plasmid DNA from these cultures was subsequently purified using Omega Bio-tek E.Z.N.A.<sup>®</sup> Plasmid Mini Kit I and insert identities confirmed by Source BioScience using SEVA\_T0\_rev & SEVA\_T1\_for oligonucleotides as sequencing primers.

#### 7.6.2.2 Assay and flow cytometry analysis

Expression cassettes were transformed into BL21(DE3) via heat shock and successful transformants selected on agar plates supplemented with 50  $\mu$ g/mL carbenicillin. 3 colonies were picked for each construct and incubated in 96 well plates, growing in 200  $\mu$ L LB medium supplemented with 50  $\mu$ g/mL carbenicillin shaking overnight at 30°C. Overnight cultures were diluted 200 times into the assay plate with 100  $\mu$ L LB supplemented with 50  $\mu$ g/mL carbenicillin, 250  $\mu$ M IPTG and 0 - 2.5 mM theophylline.

Cultures were grown shaking at 30 °C to mid-log phase (6 hours) and 2 µl were off-sampled into 200µl Phosphate Buffer Saline supplemented with 2 mg/ml kanamycin to arrest protein synthesis.

Samples were directly analysed for single cell GFP fluorescence using a BD Fortessa flow cytometer. Cell populations were gated by the same FSC/SSC setting across all samples. At least 10,000 cells were analysed for each sample, the data was exported in FCS 3.0 file standard and analysed using FlowJo\_V10.

## Chapter 8: Bibliography

1. Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* **13**, 3021–3030 (1985).
2. Rich, A. *Horizons in Biochemistry*. (Academic Press Inc, 1962).
3. LOEB, T. & ZINDER, N. D. A bacteriophage containing RNA. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 282–9 (1961).
4. Kruger, K. *et al.* Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **31**, 147–157 (1982).
5. Nissen, P., Nissen, P., Hansen, J., Ban, N. & Moore, P. B. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. *Sciences (New. York)*. **289**, 920–930 (2000).
6. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å Resolution. *Science (80-. )*. **289**, 905–920 (2000).
7. Wilson, T. J., Liu, Y. & Lilley, D. M. J. Ribozymes and the mechanisms that underlie RNA catalysis. *Front. Chem. Sci. Eng.* (2016). doi:10.1007/s11705-016-1558-2
8. Altman, S. Ribonuclease P. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366**, 2936–41 (2011).
9. Shi, Y. The Spliceosome: A Protein-Directed Metalloribozyme. *J. Mol. Biol.* **429**, 2640–2653 (2017).
10. Jimenez, R. M., Polanco, J. A. & Lupták, A. Chemistry and Biology of Self-Cleaving Ribozymes. *Trends Biochem. Sci.* **40**, 648–661 (2015).
11. Cech, T. R. *SELF-SPLICING OF GROUP I INTRONS*. *Annu. Rev. Biochem* **59**, (1990).
12. Scott, W. G., Horan, L. H. & Martick, M. The hammerhead ribozyme: structure, catalysis, and gene regulation. *Prog. Mol. Biol. Transl. Sci.* **120**, 1–23 (2013).
13. Thomas, J. M. & Perrin, D. M. Probing general acid catalysis in the hammerhead ribozyme. *J. Am. Chem. Soc.* **131**, 1135–43 (2009).
14. Lilley, D. M. J. How RNA acts as a nuclease: some mechanistic comparisons in the nucleolytic ribozymes. *Biochem. Soc. Trans.* **45**, 683–691 (2017).
15. McLeod, A. C. & Lilley\*, D. M. J. Efficient, pH-Dependent RNA Ligation by the VS Ribozyme in Trans<sup>†</sup>. (2003). doi:10.1021/BI035790E
16. Nahas, M. K. *et al.* Observation of internal cleavage and ligation reactions of a ribozyme. *Nat. Struct. Mol. Biol.* **11**, 1107–1113 (2004).
17. Marella D. Canny, Fiona M. Jucker, A. & Pardi\*, A. Efficient Ligation of the Schistosoma Hammerhead Ribozyme<sup>†</sup>. (2007). doi:10.1021/BI062077R
18. Rupert, P. B. & Ferré-D'Amaré, A. R. Crystal structure of a hairpin ribozyme–inhibitor complex with implications for catalysis. *Nature* **410**, 780–786 (2001).
19. Klein, D. J. & Ferré-D'Amaré, A. R. Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science* **313**, 1752–6 (2006).
20. Liu, Y., Wilson, T. J., McPhee, S. A. & Lilley, D. M. J. Crystal structure and mechanistic investigation of the twister ribozyme. *Nat. Chem. Biol.* **10**, 739–744 (2014).

21. Ren, A. *et al.* Pistol ribozyme adopts a pseudoknot fold facilitating site-specific in-line cleavage. *Nat. Chem. Biol.* **12**, 702–708 (2016).
22. Ke, A., Zhou, K., Ding, F., Cate, J. H. D. & Doudna, J. A. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature* **429**, 201–205 (2004).
23. Martick, M. & Scott, W. G. Tertiary Contacts Distant from the Active Site Prime a Ribozyme for Catalysis. *Cell* **126**, 309–320 (2006).
24. Suslov, N. B. *et al.* Crystal structure of the Varkud satellite ribozyme. *Nat. Chem. Biol.* **11**, 840–846 (2015).
25. Voet, D. & Voet, J. Biochemistry. Hoboken. *John Wiley Sons* (2004).
26. Stahley, M. R. & Strobel, S. A. Structural evidence for a two-metal-ion mechanism of group I intron splicing. *Science* **309**, 1587–90 (2005).
27. Breaker, R. R. Riboswitches and Translation Control. *Cold Spring Harb. Perspect. Biol.* a032797 (2018). doi:10.1101/cshperspect.a032797
28. Smith, A. M., Fuchs, R. T., Grundy, F. J. & Henkin, T. M. Riboswitch RNAs: regulation of gene expression by direct monitoring of a physiological signal. *RNA Biol.* **7**, 104–10 (2010).
29. Mandal, M. *et al.* A Glycine-Dependent Riboswitch That Uses Cooperative Binding to Control Gene Expression. *Science (80-. )*. **306**, 275–279 (2004).
30. Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S. & Breaker, R. R. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev.* **17**, 2688–2697 (2003).
31. Dann, C. E. *et al.* Structure and Mechanism of a Metal-Sensing Regulatory RNA. *Cell* **130**, 878–892 (2007).
32. Furukawa, K. *et al.* Bacterial riboswitches cooperatively bind Ni(2+) or Co(2+) ions and control expression of heavy metal transporters. *Mol. Cell* **57**, 1088–1098 (2015).
33. Nelson, J. W. *et al.* Riboswitches in eubacteria sense the second messenger c-di-AMP. *Nat. Chem. Biol.* **9**, 834–839 (2013).
34. Lee, E. R., Baker, J. L., Weinberg, Z., Sudarsan, N. & Breaker, R. R. An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science* **329**, 845–848 (2010).
35. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–56 (2002).
36. Winkler, W. C., Cohen-Chalamish, S. & Breaker, R. R. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15908–13 (2002).
37. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).
38. Caron, M.-P. *et al.* Dual-acting riboswitch control of translation initiation and mRNA decay. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E3444-53 (2012).
39. Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–6 (2004).
40. Bingaman, J. L. *et al.* The GlcN6P cofactor plays multiple catalytic roles in the glmS ribozyme. *Nat. Chem. Biol.* (2017). doi:10.1038/nchembio.2300
41. Collins, J. A., Irnov, I., Baker, S. & Winkler, W. C. Mechanism of mRNA destabilization by the

- glmS ribozyme. *Genes Dev.* **21**, 3356–68 (2007).
42. Bordeleau, E. *et al.* Cyclic di-GMP riboswitch-regulated type IV pili contribute to aggregation of *Clostridium difficile*. *J. Bacteriol.* **197**, 819–32 (2015).
  43. Purcell, E. B., McKee, R. W., Bordeleau, E., Burrus, V. & Tamayo, R. Regulation of Type IV Pili Contributes to Surface Behaviors of Historical and Epidemic Strains of *Clostridium difficile*. *J. Bacteriol.* **198**, 565–77 (2016).
  44. Thompson, K. M., Syrett, H. A., Knudsen, S. M. & Ellington, A. D. Group I aptazymes as genetic regulatory switches. *BMC Biotechnol.* **2**, 21 (2002).
  45. Klauser, B. *et al.* Post-transcriptional Boolean computation by combining aptazymes controlling mRNA translation initiation and tRNA activation. *Mol. Biosyst.* **8**, 2242–8 (2012).
  46. Townshend, B., Kennedy, A. B., Xiang, J. S. & Smolke, C. D. High-throughput cellular RNA device engineering. *Nat. Methods* **12**, 989–994 (2015).
  47. Beilstein, K., Wittmann, A., Grez, M. & Suess, B. Conditional control of mammalian gene expression by tetracycline- dependent hammerhead ribozymes. *ACS Synth. Biol.* (2014). doi:10.1021/sb500270h
  48. Ausländer, S., Ketzer, P. & Hartig, J. S. A ligand-dependent hammerhead ribozyme switch for controlling mammalian gene expression. *Mol. Biosyst.* **6**, 807 (2010).
  49. Klauser, B., Atanasov, J., Siewert, L. K. & Hartig, J. S. Ribozyme-Based Aminoglycoside Switches of Gene Expression Engineered by Genetic Selection in *S. cerevisiae*. *ACS Synth. Biol.* **4**, 516–25 (2015).
  50. Liang, J. C., Chang, A. L., Kennedy, A. B. & Smolke, C. D. A high-throughput, quantitative cell-based screen for efficient tailoring of RNA device activity. *Nucleic Acids Res.* **40**, e154 (2012).
  51. Chen, X., Denison, L., Levy, M. & Ellington, A. D. Direct selection for ribozyme cleavage activity in cells. *RNA* **15**, 2035–45 (2009).
  52. Wieland, M., Benz, A., Klauser, B. & Hartig, J. S. Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angew. Chem. Int. Ed. Engl.* **48**, 2715–8 (2009).
  53. Wieland, M. & Hartig, J. S. Improved aptazyme design and in vivo screening enable riboswitching in bacteria. *Angew. Chem. Int. Ed. Engl.* **47**, 2604–7 (2008).
  54. Win, M. N. & Smolke, C. D. A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14283–8 (2007).
  55. Wittmann, A. & Suess, B. Selection of tetracycline inducible self-cleaving ribozymes as synthetic devices for gene regulation in yeast. *Mol. Biosyst.* **7**, 2419–27 (2011).
  56. Ausländer, S. *et al.* A general design strategy for protein-responsive riboswitches in mammalian cells. *Nat. Methods* **advance on**, (2014).
  57. PENEDO, J. C., Wilson, T. J., Jayasena, S. D., Khvorova, A. & Lilley, D. M. J. Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *RNA* **10**, 880–888 (2004).
  58. de la Peña, M. *et al.* The Hammerhead Ribozyme: A Long History for a Short RNA. *Molecules* **22**, 78 (2017).
  59. Perreault, J. *et al.* Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput. Biol.* **7**, e1002031 (2011).

60. Hertel, K. J. *et al.* Numbering system for the hammerhead. *Nucleic Acids Res.* **20**, 3252 (1992).
61. Salehi-Ashtiani, K. & Szostak, J. W. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature* **414**, 82–4 (2001).
62. Nelson, J. A. & Uhlenbeck, O. C. Hammerhead redux: Does the new structure fit the old biochemical data? (2008). doi:10.1261/rna.912608
63. Khvorova, A., Lescoute, A., Westhof, E. & Jayasena, S. D. Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nat. Struct. Biol.* **10**, 708–12 (2003).
64. Nelson, J. A. & Uhlenbeck, O. C. Minimal and extended hammerheads utilize a similar dynamic reaction mechanism for catalysis. *RNA* **14**, 43–54 (2008).
65. Tang, J. & Breaker, R. R. Rational design of allosteric ribozymes. *Chem. Biol.* **4**, 453–459 (1997).
66. Koizumi, M., Soukup, G. A., Kerr, J. N. & Breaker, R. R. Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. *Nat. Struct. Biol.* **6**, 1062–71 (1999).
67. Soukup, G. A. & Breaker, R. R. Engineering precision RNA molecular switches. *Proc. Natl. Acad. Sci.* **96**, 3584–3589 (1999).
68. Piganeau, N., Jenne, A., Thuillier, V. & Famulok, M. An Allosteric Ribozyme Regulated by Doxycycline. *Angew. Chemie* **39**, 4369–4373 (2000).
69. Piganeau, N., Thuillier, V. & Famulok, M. In vitro selection of allosteric ribozymes: theory and experimental validation. *J. Mol. Biol.* **312**, 1177–1190 (2001).
70. Saragliadis, A., Krajewski, S. S., Rehm, C., Narberhaus, F. & Hartig, J. S. Thermozymes: Synthetic RNA thermometers based on ribozyme activity. *RNA Biol.* **10**, 1010–6 (2013).
71. Carothers, J. M., Goler, J. A., Juminaga, D. & Keasling, J. D. Model-driven engineering of RNA devices to quantitatively program gene expression. *Science* **334**, 1716–9 (2011).
72. Link, K. H. *et al.* Engineering high-speed allosteric hammerhead ribozymes. *Biol. Chem.* **388**, 779–86 (2007).
73. Cannistraro, V. J. & Kennell, D. The Processive Reaction Mechanism of Ribonuclease II. *J. Mol. Biol.* **243**, 930–943 (1994).
74. Deana, A., Celesnik, H. & Belasco, J. G. The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature* **451**, 355–358 (2008).
75. Lim, H. G., Jang, S., Jang, S., Seo, S. W. & Jung, G. Y. Design and optimization of genetically encoded biosensors for high-throughput screening of chemicals. *Curr. Opin. Biotechnol.* **54**, 18–25 (2018).
76. Lin, J.-L., Wagner, J. M. & Alper, H. S. Enabling tools for high-throughput detection of metabolites: Metabolic engineering and directed evolution applications. *Biotechnol. Adv.* **35**, 950–970 (2017).
77. Michener, J. K. & Smolke, C. D. High-throughput enzyme evolution in *Saccharomyces cerevisiae* using a synthetic RNA switch. *Metab. Eng.* **14**, 306–316 (2012).
78. Stevens, J. T. & Carothers, J. M. Designing RNA-Based Genetic Control Systems for Efficient Production from Engineered Metabolic Pathways. *ACS Synth. Biol.* **4**, 107–115 (2015).
79. Bhartiya, S., Rawool, S. & Venkatesh, K. V. Dynamic model of *Escherichia coli* tryptophan

- operon shows an optimal structural design. *Eur. J. Biochem.* **270**, 2644–2651 (2003).
80. Weaver, R. *Molecular Biology*. (McGraw-Hill Higher Education, 2011).
  81. Venayak, N., Anesiadis, N., Cluett, W. R. & Mahadevan, R. Engineering metabolism through dynamic control. *Curr. Opin. Biotechnol.* **34C**, 142–152 (2015).
  82. Farmer, W. & Liao, J. Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nat. Biotechnol.* **18**, 533–537 (2000).
  83. Oyarzún, D. A. & Stan, G.-B. V. Synthetic gene circuits for metabolic control: design trade-offs and constraints. *J. R. Soc. Interface* **10**, 20120671- (2012).
  84. Zhang, F., Carothers, J. M. J. & Keasling, J. J. D. J. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat. Biotechnol.* **30**, 354–9 (2012).
  85. Dahl, R. H. *et al.* Engineering dynamic pathway regulation using stress-response promoters. *Nat. Biotechnol.* **advance on**, (2013).
  86. Xu, P., Li, L., Zhang, F., Stephanopoulos, G. & Koffas, M. Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proc. Natl. Acad. Sci.* 1406401111- (2014). doi:10.1073/pnas.1406401111
  87. Zhou, L.-B. & Zeng, A.-P. Engineering a Lysine-ON Riboswitch for Metabolic Control of Lysine Production in *Corynebacterium glutamicum*. *ACS Synth. Biol.* (2015). doi:10.1021/acssynbio.5b00075
  88. Zhou, L.-B. & Zeng, A.-P. Exploring lysine riboswitch for metabolic flux control and improvement of L-lysine synthesis in *Corynebacterium glutamicum*. *ACS Synth. Biol.* 150109140058000 (2015). doi:10.1021/sb500332c
  89. Ketzer, P., Haas, S. F., Engelhardt, S., Hartig, J. S. & Nettelbeck, D. M. Synthetic riboswitches for external regulation of genes transferred by replication-deficient and oncolytic adenoviruses. *Nucleic Acids Res.* **40**, e167 (2012).
  90. Ho Lee, C., Ryul Han, S. & Lee, S.-W. Therapeutic Applications of Aptamer-Based Riboswitches. *Nucleic Acid Ther.* **26**, 44–51 (2016).
  91. Bell, C. L. *et al.* Control of alphavirus-based gene expression using engineered riboswitches. *Virology* **483**, 302–311 (2015).
  92. Lee, C. H., Kim, J. H., Kim, H. W., Myung, H. & Lee, S.-W. Hepatitis C Virus Replication-Specific Inhibition of MicroRNA Activity with Self-Cleavable Allosteric Ribozyme. *Nucleic Acid Ther. (Formerly Oligonucleotides)* **22**, 120104060718000 (2012).
  93. Chen, Y. Y., Jensen, M. C. & Smolke, C. D. Genetic control of mammalian T-cell proliferation with synthetic RNA regulatory systems. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8531–6 (2010).
  94. Ketzer, P. *et al.* Artificial riboswitches for gene expression and replication control of DNA and RNA viruses. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E554-62 (2014).
  95. European Medicines Agency. First oncolytic immunotherapy medicine recommended for approval. (2015). Available at: [http://www.ema.europa.eu/ema/index.jsp?curl=pages/news\\_and\\_events/news/2015/10/news\\_detail\\_002421.jsp&mid=WC0b01ac058004d5c1](http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2015/10/news_detail_002421.jsp&mid=WC0b01ac058004d5c1). (Accessed: 4th August 2018)
  96. Kelly, E. J., Hadac, E. M., Greiner, S. & Russell, S. J. Engineering microRNA responsiveness to

- decrease virus pathogenicity. *Nat. Med.* **14**, 1278–1283 (2008).
97. Willmon, C. *et al.* Cell Carriers for Oncolytic Viruses: Fed Ex for Cancer Therapy. *Mol. Ther.* **17**, 1667–1676 (2009).
  98. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* (80-. ). **249**, 505–510 (1990).
  99. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–22 (1990).
  100. Ouellet, E., Foley, J. H., Conway, E. M. & Haynes, C. Hi-Fi SELEX: A high-fidelity digital-PCR based therapeutic aptamer discovery platform. *Biotechnol. Bioeng.* **112**, 1506–1522 (2015).
  101. Nutiu, R. & Li, Y. In vitro selection of structure-switching signaling aptamers. *Angew. Chem. Int. Ed. Engl.* **44**, 1061–5 (2005).
  102. Zhuo, Z. *et al.* Recent Advances in SELEX Technology and Aptamer Applications in Biomedicine. *Int. J. Mol. Sci.* **18**, 2142 (2017).
  103. Lee, J. F., Hesselberth, J. R., Meyers, L. A. & Ellington, A. D. Aptamer database. *Nucleic Acids Res.* **32**, D95–100 (2004).
  104. Jenison, R. D., Gill, S. C., Pardi, A. & Polisky, B. High-resolution molecular discrimination by RNA. *Science* **263**, 1425–9 (1994).
  105. Yang, J. *et al.* Synthetic RNA devices to expedite the evolution of metabolite-producing microbes. *Nat. Commun.* **4**, 1413 (2013).
  106. Groher, F. *et al.* Riboswitching with ciprofloxacin—development and characterization of a novel RNA regulator. *Nucleic Acids Res.* **46**, 1–12 (2018).
  107. Sinha, J., Reyes, S. J. & Gallivan, J. P. Reprogramming bacteria to seek and destroy an herbicide. *Nat. Chem. Biol.* **6**, 464–70 (2010).
  108. Lynch, S. A. & Gallivan, J. P. A flow cytometry-based screen for synthetic riboswitches. *Nucleic Acids Res.* **37**, 184–92 (2009).
  109. Davidson, M. E., Harbaugh, S. V, Chushak, Y. G., Stone, M. O. & Kelley-Loughnane, N. Development of a 2,4-dinitrotoluene-responsive synthetic riboswitch in *E. coli* cells. *ACS Chem. Biol.* **8**, 234–41 (2013).
  110. Weigand, J. E. *et al.* Screening for engineered neomycin riboswitches that control translation initiation. *RNA* **14**, 89–97 (2008).
  111. Taylor, N. D. *et al.* Engineering an allosteric transcription factor to respond to new ligands. *Nat. Methods* **advance on**, (2015).
  112. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331–9 (2014).
  113. Felletti, M. & Hartig, J. S. Ligand-dependent ribozymes. *Wiley Interdiscip. Rev. RNA* **8**, e1395 (2017).
  114. Weigand, J. E. *et al.* Mechanistic insights into an engineered riboswitch: a switching element which confers riboswitch activity. *Nucleic Acids Res.* **39**, 3363–72 (2011).
  115. Wallis, M. G., von Ahsen, U., Schroeder, R. & Famulok, M. A novel RNA motif for neomycin recognition. *Chem. Biol.* **2**, 543–52 (1995).

116. Schneider, C. & Suess, B. Identification of RNA aptamers with riboswitching properties. *Methods* **97**, 44–50 (2016).
117. Berens, C. & Suess, B. Riboswitch engineering — making the all-important second and third steps. *Curr. Opin. Biotechnol.* **31**, 10–15 (2015).
118. Espah Borujeni, A., Mishler, D. M., Wang, J., Huso, W. & Salis, H. M. Automated physics-based design of synthetic riboswitches from diverse RNA aptamers. *Nucleic Acids Res.* gkv1289- (2015). doi:10.1093/nar/gkv1289
119. Roth, A. & Breaker, R. R. in *Ribozymes and siRNA protocols* 145–164 (Humana Press, 2004). doi:10.1385/1-59259-746-7:145
120. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).
121. Kobori, S., Nomura, Y., Miu, A. & Yokobayashi, Y. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Res.* gkv265- (2015). doi:10.1093/nar/gkv265
122. Kobori, S., Takahashi, K. & Yokobayashi, Y. Deep Sequencing Analysis of Aptazyme Variants Based on a Pistol Ribozyme. *ACS Synth. Biol.* acssynbio.7b00057 (2017). doi:10.1021/acssynbio.7b00057
123. Ferguson, A. *et al.* A novel strategy for selection of allosteric ribozymes yields RiboReporter sensors for caffeine and aspartame. *Nucleic Acids Res.* **32**, 1756–66 (2004).
124. Vinkenborg, J. L., Karnowski, N. & Famulok, M. Aptamers for allosteric regulation. *Nat. Chem. Biol.* **7**, 519–527 (2011).
125. Carothers, J. M., Oestreich, S. C., Davis, J. H. & Szostak, J. W. Informational Complexity and Functional Activity of RNA Structures. *J. Am. Chem. Soc.* **126**, 5130–5137 (2004).
126. Carothers, J. M., Goler, J. A., Kapoor, Y., Lara, L. & Keasling, J. D. Selecting RNA aptamers for synthetic biology: investigating magnesium dependence and predicting binding affinity. *Nucleic Acids Res.* **38**, 2736–47 (2010).
127. Goler, J. A., Carothers, J. M. & Keasling, J. D. Dual-selection for evolution of in vivo functional aptazymes as riboswitch parts. *Methods Mol. Biol.* **1111**, 221–35 (2014).
128. Cox, J. C., Rudolph, P. & Ellington, A. D. Automated RNA Selection. *Biotechnol. Prog.* **14**, 845–850 (1998).
129. Eulberg, D., Buchner, K., Maasch, C. & Klussmann, S. Development of an automated in vitro selection protocol to obtain RNA-based aptamers: identification of a biostable substance P antagonist. *Nucleic Acids Res.* **33**, e45 (2005).
130. OpenWetWare contributors. Ethanol precipitation of nucleic acids. *OpenWetWare* (2012). Available at: [https://openwetware.org/wiki/Ethanol\\_precipitation\\_of\\_nucleic\\_acids](https://openwetware.org/wiki/Ethanol_precipitation_of_nucleic_acids). (Accessed: 4th September 2018)
131. Haines, M. The application of de novo RNA-based sensor controllers to metabolic engineering. (Imperial College London, 2014).
132. Majerfeld, I. & Yarus, M. A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res.* **33**, 5482–93 (2005).
133. Fang, M. *et al.* Intermediate-sensor assisted push–pull strategy and its application in

- heterologous deoxyviolacein production in *Escherichia coli*. *Metab. Eng.* **33**, 41–51 (2016).
134. Sugimoto, N., Nakano, S. -i., Yoneyama, M. & Honda, K. -i. Improved Thermodynamic Parameters and Helix Initiation Factor to Predict Stability of DNA Duplexes. *Nucleic Acids Res.* **24**, 4501–4505 (1996).
  135. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36**, W70–4 (2008).
  136. Zhang, D. Y. & Winfree, E. Control of DNA strand displacement kinetics using toehold exchange. *J. Am. Chem. Soc.* **131**, 17303–14 (2009).
  137. Milligan, J. F. & Uhlenbeck, O. C. *RNA Processing Part A: General Methods. Methods in Enzymology* **180**, (Elsevier, 1989).
  138. Sambrook, J. & Russell, D. *Molecular cloning A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2001).
  139. Isel, C., Ehresmann, C. & Marquet, R. Initiation of HIV Reverse Transcription. *Viruses* **2**, 213–43 (2010).
  140. O’Connell, J. *RT-PCR protocols*. (Humana Press, 2002).
  141. Cazenave, C. & Uhlenbeck, O. C. RNA template-directed RNA synthesis by T7 RNA polymerase. *Proc. Natl. Acad. Sci.* **91**, 6972–6976 (1994).
  142. Nacheva, G. A. & Berzal-Herranz, A. Preventing undesired RNA-primed RNA extension catalyzed by T7 RNA polymerase. *Eur. J. Biochem.* **270**, 1458–65 (2003).
  143. Khodakov, D. A., Khodakova, A. S., Huang, D. M., Linacre, A. & Ellis, A. V. Protected DNA strand displacement for enhanced single nucleotide discrimination in double-stranded DNA. *Sci. Rep.* **5**, 8721 (2015).
  144. Hawkins, T. L., O’Connor-Morin, T., Roy, A. & Santillan, C. DNA purification and isolation using a solid-phase. *Nucleic Acids Res.* **22**, 4543–4 (1994).
  145. Clarke, A. C. *et al.* From cheek swabs to consensus sequences: an A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. *BMC Genomics* **15**, 68 (2014).
  146. Lennon, N. J. *et al.* A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol.* **11**, R15 (2010).
  147. Ashford, S. R. Bacteriophage T7 DNA Ligase. *J. Biol. Chem.* **271**, 11083–11089 (1996).
  148. Long, D. M. & Uhlenbeck, O. C. Kinetic characterization of intramolecular and intermolecular hammerhead RNAs with stem II deletions. **91**, 6977–6981 (1994).
  149. Piganeau, N. in 317–328 (2012). doi:10.1007/978-1-61779-545-9\_19
  150. Jurinke, C., van den Boom, D., Cantor, C. R. & Köster, H. in *PCR Mutation Detection Protocols* 179–192 (Humana Press, 2002). doi:10.1385/1-59259-273-2:179
  151. Tozzoli, R., D’Aurizio, F., Villalta, D. & Bizzaro, N. Automation, consolidation, and integration in autoimmune diagnostics. *Auto-Immun. highlights* **6**, 1–6 (2015).
  152. Porter, E. B., Polaski, J. T., Morck, M. M. & Batey, R. T. Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat. Chem. Biol.* **13**, 295–301 (2017).
  153. Uhm, H., Kang, W., Ha, K. S., Kang, C. & Hohng, S. Single-molecule FRET studies on the

- cotranscriptional folding of a thiamine pyrophosphate riboswitch. *Proc. Natl. Acad. Sci. U. S. A.* 201712983 (2017). doi:10.1073/pnas.1712983115
154. Watters, K. E., Strobel, E. J., Yu, A. M., Lis, J. T. & Lucks, J. B. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nat. Struct. Mol. Biol.* **23**, 1124–1131 (2016).
  155. Garst, A. D., Edwards, A. L. & Batey, R. T. Riboswitches: structures and mechanisms. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).
  156. Ellefson, J. W. *et al.* Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* **352**, 1590–3 (2016).
  157. Takahashi, M. *et al.* High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency. *Sci. Rep.* **6**, 33697 (2016).
  158. Reinholt, S. J., Ozer, A., Lis, J. T. & Craighead, H. G. Highly Multiplexed RNA Aptamer Selection using a Microplate-based Microcolumn Device. *Sci. Rep.* **6**, 29771 (2016).
  159. Clarke, S. C. & Diggle, M. A. Automated PCR/Sequence Template Purification. *Mol. Biotechnol.* **21**, 221–224 (2002).
  160. IVANOVA, N. V., DEWAARD, J. R. & HEBERT, P. D. N. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol. Ecol. Notes* **6**, 998–1002 (2006).
  161. Martini, L. *et al.* In Vitro Selection for Small-Molecule-Triggered Strand Displacement and Riboswitch Activity. *ACS Synth. Biol.* **4**, 1144–50 (2015).
  162. Wiedmann, M. *et al.* Ligase chain reaction (LCR)--overview and applications. *Genome Res.* **3**, S51–S64 (1994).
  163. Zimmermann, G. R., Jenison, R. D., Wick, C. L., Simorre, J.-P. & Pardi, A. Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. *Nat. Struct. Biol.* **4**, 644–649 (1997).
  164. Dwidar, M. & Yokobayashi, Y. Controlling *Bdellovibrio bacteriovorus* Gene Expression and Predation Using Synthetic Riboswitches. *ACS Synth. Biol.* **6**, 2035–2041 (2017).
  165. Westbrook, A. M. & Lucks, J. B. Achieving large dynamic range control of gene expression with a compact RNA transcription–translation regulator. *Nucleic Acids Res.* **45**, 5614–5624 (2017).
  166. Eckdahl, T. T. *et al.* Programmed evolution for optimization of orthogonal metabolic output in bacteria. *PLoS One* **10**, e0118322 (2015).
  167. Bloom, R. J., Winkler, S. M. & Smolke, C. D. Synthetic feedback control using an RNAi-based gene-regulatory device. *J. Biol. Eng.* **9**, 5 (2015).
  168. Przybilski, R. & Hammann, C. The tolerance to exchanges of the Watson Crick base pair in the hammerhead ribozyme core is determined by surrounding elements. *RNA* **13**, 1625–30 (2007).
  169. Dufour, D., de la Peña, M., Gago, S., Flores, R. & Gallego, J. Structure-function analysis of the ribozymes of chrysanthemum chlorotic mottle viroid: a loop-loop interaction motif conserved in most natural hammerheads. *Nucleic Acids Res.* **37**, 368–81 (2009).
  170. Ryckelynck, M. *et al.* Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. *RNA* **21**, 458–69 (2015).
  171. Taylor, P. D. & Jonker, L. B. Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156 (1978).

172. Li, A. *et al.* Beating Bias in the Directed Evolution of Proteins: Combining High-Fidelity on-Chip Solid-Phase Gene Synthesis with Efficient Gene Assembly for Combinatorial Library Construction. *ChemBioChem* **19**, 221–228 (2018).
173. Stangegaard, M., Høgh Dufva, I. & Dufva, M. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. *Biotechniques* **40**, 649–657 (2006).
174. Hershberg, R. & Petrov, D. A. Selection on Codon Bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).
175. Lohman, G. J. S. *et al.* A high-throughput assay for the comprehensive profiling of DNA ligase fidelity. *Nucleic Acids Res.* **44**, e14 (2016).
176. Yufa, R. *et al.* Emulsion PCR Significantly Improves Nonequilibrium Capillary Electrophoresis of Equilibrium Mixtures-Based Aptamer Selection: Allowing for Efficient and Rapid Selection of Aptamer to Unmodified ABH2 Protein. *Anal. Chem.* **87**, 1411–1419 (2015).
177. Ramakers, C., Ruijter, J. M., Deprez, R. H. L. & Moorman, A. F. . Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* **339**, 62–66 (2003).
178. Gu, H., Furukawa, K. & Breaker, R. R. Engineered allosteric ribozymes that sense the bacterial second messenger cyclic diguanosyl 5'-monophosphate. *Anal. Chem.* **84**, 4935–41 (2012).
179. Hartig, J. S. & Famulok, M. Reporter Ribozymes for Real-Time Analysis of Domain-Specific Interactions in Biomolecules: HIV-1 Reverse Transcriptase and the Primer–Template Complex. *Angew. Chemie Int. Ed.* **41**, 4263–4266 (2002).
180. Witt, M. *et al.* Comparing two conventional methods of emulsion PCR and optimizing of Tegosoftware-based emulsion PCR. *Eng. Life Sci.* **17**, 953–958 (2017).
181. O'Hare, H. M. & Johnsson, K. The Laboratory in a Droplet. *Chem. Biol.* **12**, 1255–1257 (2005).
182. Nilsen, I. W. *et al.* The Enzyme and the cDNA Sequence of a Thermolabile and Double-Strand Specific DNase from Northern Shrimps (*Pandalus borealis*). *PLoS One* **5**, e10295 (2010).
183. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
184. Kobori, S. & Yokobayashi, Y. Analyzing and Tuning Ribozyme Activity by Deep Sequencing to Modulate Gene Expression Level in Mammalian Cells. *ACS Synth. Biol.* **7**, 371–376 (2018).
185. Jang, S. & Jung, G. Y. Systematic Optimization of L -Tryptophan Riboswitches for Efficient Monitoring of the Metabolite in Escherichia coli. *Biotechnol. Bioeng.* (2017). doi:10.1002/bit.26448
186. Jang, S. *et al.* On-chip analysis, indexing and screening for chemical producing bacteria in a microfluidic static droplet array. *Lab Chip* **16**, 1909–1916 (2016).
187. Yanofsky, C. Transcription attenuation: once viewed as a novel regulatory strategy. *J. Bacteriol.* **182**, 1–8 (2000).
188. Al-Rubeai, M., Kuystermans, D., Wieland, M., Ausländer, D. & Fussenegger, M. Engineering of ribozyme-based riboswitches for mammalian cells. *Methods* **56**, 351–357 (2012).
189. Young, D. D., Garner, R. A., Yoder, J. A. & Deiters, A. Light-activation of gene function in mammalian cells via ribozymes. *Chem. Commun. (Camb.)* 568–70 (2009). doi:10.1039/b819375d
190. Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nat.*

- Rev. Genet.* **11**, 476–86 (2010).
191. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
  192. Cheng, C., Chen, Y. H., Lennox, K. A., Behlke, M. A. & Davidson, B. L. In vivo SELEX for Identification of Brain-penetrating Aptamers. *Mol. Ther. - Nucleic Acids* **2**, (2013).
  193. Schütze, T. *et al.* Probing the SELEX Process with Next-Generation Sequencing. *PLoS One* **6**, e29604 (2011).
  194. Szeto, K. *et al.* RAPID-SELEX for RNA Aptamers. *PLoS One* **8**, e82667 (2013).
  195. Illumina. Low-Diversity Sequencing on the Illumina. **22**, 5–7 (2013).
  196. Illumina. Optimizing Cluster Density on Illumina Sequencing Systems Table of Contents. (2016). doi:10.1016/j.ydbio.2007.02.036
  197. Illumina. 16S Metagenomic Sequencing Library Preparation. 1–28 (2013). Available at: [http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry\\_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf](http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf).
  198. Illumina. *Quality Scores for Next-Generation Sequencing*. (2011).
  199. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
  200. Kern, J. A. & Davis, R. H. Application of solution equilibrium analysis to in vitro RNA transcription. *Biotechnol. Prog.* **13**, 747–56 (1997).
  201. Lyakhov, D. L. *et al.* Pausing and termination by bacteriophage T7 RNA polymerase. *J. Mol. Biol.* **280**, 201–213 (1998).
  202. Storch, M. *et al.* BASIC: A New Biopart Assembly Standard for Idempotent Cloning Provides Accurate, Single-Tier DNA Assembly for Synthetic Biology. *ACS Synth. Biol.* **4**, 781–787 (2015).
  203. Gallagher, R. R. The Great Escape: How to Avoid Cheaters in Biological Containment, Directed Evolution, and Genome Engineering. (2016).
  204. Gao, Y., Wolf, L. K. & Georgiadis, R. M. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Res.* **34**, 3370–7 (2006).
  205. Luo, G. X. & Taylor, J. Template switching by reverse transcriptase during DNA synthesis. *J. Virol.* **64**, 4321–8 (1990).
  206. Griffiths, A. D. & Tawfik, D. S. Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol.* **24**, 395–402 (2006).
  207. Nakano, M. *et al.* Single-molecule reverse transcription polymerase chain reaction using water-in-oil emulsion. *J. Biosci. Bioeng.* **99**, 293–295 (2005).
  208. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–56 (2012).
  209. Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. & Rosenow, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **13**, 216–23 (2003).
  210. Carmody, S. R. & Wenthe, S. R. mRNA nuclear export at a glance. *J. Cell Sci.* **122**, 1933–1937

- (2009).
211. Iost, I., Guillerez, J. & Dreyfus, M. Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo. *J. Bacteriol.* **174**, 619–22 (1992).
  212. Espah Borujeni, A., Channarasappa, A. S. & Salis, H. M. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.* **42**, 2646–2659 (2014).
  213. Paudel, B. P. & Rueda, D. Molecular Crowding Accelerates Ribozyme Docking and Catalysis. *J. Am. Chem. Soc.* **136**, 16700–16703 (2014).
  214. Leamy, K. A., Assmann, S. M., Mathews, D. H. & Bevilacqua, P. C. Bridging the gap between in vitro and in vivo RNA folding. *Q. Rev. Biophys.* **49**, e10 (2016).
  215. de Silva, C. & Walter, N. G. Leakage and slow allostery limit performance of single drug-sensing aptazyme molecules based on the hammerhead ribozyme. *RNA* **15**, 76–84 (2009).
  216. SantaLucia, J., Allawi, H. T. & Seneviratne, P. A. Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability <sup>†</sup>. *Biochemistry* **35**, 3555–3562 (1996).
  217. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad.* (1998).
  218. Owczarzy, R., Moreira, B. G., You, Y., Behlke, M. A. & Walder, J. A. Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry* **47**, 5336–53 (2008).
  219. Stemmer, W. P. C., Cramer, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).
  220. OSHIMA, T. & IMAHORI, K. Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a Nonsporulating Thermophilic Bacterium from a Japanese Thermal Spa. *Int. J. Syst. Bacteriol.* **24**, 102–112 (1974).
  221. Popa, N., Novac, O., Profire, L., Hritcu, D. & Popa, M. I. Inclusion and release of theophylline from chitosan based microparticles. *Turkish J. Chem.* **34**, 255–262 (2010).

## Chapter 9: Appendix

### 9.1 Functions and parameters for selection simulations

Refer to the derivations in Chapter 3, Section 3.3.2.1 for a description of the following functions:

$$p\left(\frac{x_i^{(n)}}{X}, \%C_i^{(+)}, I, \%C_T^{(0+)}\right) = \frac{\frac{x_i^{(n)}}{X} \%C_i^{(+)}}{\frac{x_i^{(n)}}{X} \%C_i^{(+)} + \frac{I \%C_T^{(0+)} - \%C_i^{(+)}}{I - 1} \left(1 - \frac{x_i^{(n)}}{X}\right)}$$

$$n\left(\frac{x_i^{(n)}}{X}, \%C_i^{(-)}, I, \%C_T^{(0-)}\right) = \frac{\frac{x_i^{(n)}}{X} (1 - \%C_i^{(-)})}{1 - \left(\frac{x_i^{(n)}}{X} \%C_i^{(-)} + \frac{I \%C_T^{(0-)} - \%C_i^{(-)}}{I - 1} \left(1 - \frac{x_i^{(n)}}{X}\right)\right)}$$

Individual  $\%C_T^{(0+)}$ ,  $\%C_T^{(0-)}$ ,  $\%C_{i(+)}$  and  $\%C_{i(-)}$  parameters for each selection simulation are given in Table 9-1 - Table 9-3 below. All simulations involved alternating rounds of positive and negative selection with the first round of selection being positive.

| Parameter      | Value      | Reasoning                               |
|----------------|------------|-----------------------------------------|
| $\%C_T^{(0+)}$ | 0.2102543  | fraction cleaved Theo Lib, 3.16 mM theo |
| $\%C_T^{(0-)}$ | 0.2083738  | fraction cleaved Theo Lib, 0 mM theo    |
| $\%C_{i(+)}$   | 0.8079652  | fraction cleaved Theo Con, 3.16 mM theo |
| $\%C_{i(-)}$   | 0.05085209 | fraction cleaved Theo Con, 0 mM theo    |
| $I$            | 1024*      | Theo Lib size                           |

Table 9-1: Parameters used to generate selection simulations in Figure 3-9 & Figure 5-7 – All Theophylline Control (Theo Con) & Library (Theo Lib) parameters empirically measured under TRT reaction conditions under the stated theophylline concentration. \*Where the Theophylline Control & C63A variant were simulated together in Figure 5-7, the value of  $I$  was halved to 512.

| Parameter      | Value      | Reasoning                               |
|----------------|------------|-----------------------------------------|
| $\%C_T^{(0+)}$ | 0.5        | Response of semi-active sequences       |
| $\%C_T^{(0-)}$ | 0.5        |                                         |
| $\%C_{i(+)}$   | 0.8079652  | fraction cleaved Theo Con, 3.16 mM theo |
| $\%C_{i(-)}$   | 0.05085209 | fraction cleaved Theo Con, 0 mM theo    |
| $I$            | variable   | Refer to figure                         |

Table 9-2: Parameters used to generate selection simulations in Figure 4-2 – All Theophylline Control

(Theo Con) parameters empirically measured under TRT reaction conditions under the stated theophylline concentration.

| Parameter      | Value     | Reasoning                                                                       |
|----------------|-----------|---------------------------------------------------------------------------------|
| $\%C_T^{(0+)}$ | 0.050861  | Response of sequence with $0.1117 \text{ min}^{-1}$ rate constant (0.9428 mins) |
| $\%C_T^{(0-)}$ | 0.8006    | Response of sequence with $0.1117 \text{ min}^{-1}$ rate constant (44.58 mins)  |
| $\%C_{i(+)}$   | 0.807956  | fraction cleaved Theo Con, 3.16 mM theo (0.9428 mins)                           |
| $\%C_{i(-)}$   | 0.8006    | Estimated* fraction cleaved Theo Con, 0 mM theo (44.58 mins)                    |
| $I$            | $10^{14}$ | Maximum throughput of method                                                    |

Table 9-3: Parameters used to generate the simulation illustrated in Figure 4-4C when inactivation delayed during negative selection – Optimal delay during negative selection calculated at 44.58 minutes while positive selection delay based on TRT ribozyme inactivation time; calculated at 0.9428 mins (refer to Section 9.3.1). \*Cleavage rate constant for Theophylline Control under 0 mM theophylline estimated at  $0.1117 \text{ min}^{-1}$  (Section 9.3.1).

## 9.2 Appendix for Chapter 2

### 9.2.1 Candidate bait-oligo toehold sequences

Of the 10 identified sequences in Table 9-4, the 2<sup>nd</sup> sequence was arbitrarily chosen.

| Seq no. | Sequence |
|---------|----------|
| 1       | GAAACCC  |
| 2       | GAACACC  |
| 3       | GAACCAC  |
| 4       | GACAACC  |
| 5       | GACACAC  |
| 6       | GACCAAC  |
| 7       | GCAAACC  |
| 8       | GCAACAC  |
| 9       | GCACAAC  |
| 10      | GCCAAAC  |

Table 9-4: Candidate bait-oligo toehold sequences

### 9.2.2 Candidate RT primer sequences

Table 9-5 illustrates the 14 RT primer sequences and their properties. The corresponding Trp library contains the reverse complement of each sequence at the 3' end, enabling reverse transcription.

| Sequence no. | Sequence      | Length | T <sub>m</sub> (°C) |
|--------------|---------------|--------|---------------------|
| 1            | GGGTGTTTCT    | 10     | 42.07705            |
| 2            | GGTGGTTTCT    | 10     | 42.07705            |
| 3            | GGGTGTTTTCT   | 11     | 45.14303            |
| 4            | GGTTGTTTCT    | 11     | 45.14303            |
| 5            | GGGTTTTTCT    | 11     | 42.13404            |
| 6            | GGTGGTTTTCT   | 11     | 45.14303            |
| 7            | GGTGTGTTTCT   | 11     | 44.89325            |
| 8            | GGTTGGTTTCT   | 11     | 45.14303            |
| 9            | GGGTTTTTTTCT  | 12     | 44.99495            |
| 10           | GGTGTTTTTCT   | 12     | 44.77099            |
| 11           | GGTTGTTTTTCT  | 12     | 44.77099            |
| 12           | GGTTTGTTTTCT  | 12     | 44.77099            |
| 13           | GGTTTTGTTTCT  | 12     | 44.77099            |
| 14           | GGTTTTTTTTTCT | 13     | 44.68574            |

Table 9-5: Candidate RT primer sequences.

### 9.2.3 Timing of selection method

| Step                                                   | Time (minutes) |
|--------------------------------------------------------|----------------|
| Prepare IVT/RT*                                        | 30             |
| Incubate IVT/RT*                                       | 50             |
| Non-specific cDNA purification                         | 135            |
| Ligation of selected cDNA using <i>Tth</i> DNA Ligase  | 90             |
| Pull-down & purification of ligation reaction products | 50             |
| Semi-qPCR                                              | 60             |
| PCR purification                                       | 50             |
|                                                        |                |
| Total (hours)                                          | 7.75           |

*Table 9-6: Time required to implement the selection method described in this thesis – The description of each step is given in the Materials & methods chapter. Times are accurate for 8 samples conducted in parallel. \*IVT/RT equivalent to the TRT reaction (refer to Chapter 3.; Section 3.3.2.5).*

## 9.3 Appendix for Chapter 4

### 9.3.1 Maximum number of sequences sampled during each round

To calculate the number cDNA molecules which can be generated during selection, it was necessary to know three parameters. These were the volume of the RT reaction ( $V_{RT}$ ), the concentration of RNA during reverse transcription ( $C_{RT}$ ) and the efficiency of the RTase ( $E_{RT}$ ). With these parameters, the number of cDNA molecules can be calculated using the following expression:

$$C_{RT}V_{RT}E_{RT}N_A$$

Where  $N_A$  is Avogadro's constant.

To determine the concentration of RNA and efficiency of the reverse transcriptase, the manufacturers recommended conditions were consulted. According to the manufacturer, up to 50 ng/μL RNA can be added to a NEB Protoscript® II RT reaction. Furthermore, 80 - 90 % yields can be achieved even with an RNA concentration of 60 ng/μL and using random primers<sup>173</sup>. Given the recommended concentration of RNA by the manufacturer was less than this and that specific RT primers were used,

it was assumed that the RT reaction was 100 % efficient with 50 ng/μL RNA ( $E_{RT} = 1$ ). Using the molecular weight of the Theophylline Control F RNA sequence, this yields  $C_R = 1.34 \mu\text{M}$ . Note that the TRT reaction conditions used could match or exceed the yield of cDNA achieved under the manufacturer's recommended conditions; refer to Chapter 3, Figure 3-8.

To determine the volume of the RT reaction, the subsequent purification of the RT reaction was considered. According to the protocol given in Section 7.2.5 of the Materials & methods chapter, 1.8x the volume of SPRI beads must be added to the RT reaction to purify the resulting cDNA. Based on the tolerance of approximately 200 μL per well in a microtiter plate, 70 μL RT reaction could be purified during selection. Using  $V_{RT} = 70 \mu\text{L}$ , approximately  $10^{14}$  cDNA molecules could be generated during reverse transcription and hence selection.

### 9.3.2 Theo Con rate constants and TRT inactivation time

An expression to calculate a cleavage rate constant ( $k_C$ ) given a % cleaved value (%C) after a given IVT time ( $t$ ) was generated by rearranging Equation (9-3), yielding:

$$k_C = W(0, \frac{100e^{\left(\frac{100}{\%C-100}\right)}(10^4 - 100\%C)}{t(\%C - 100)^2}) \quad (9-1)$$

where  $W$  is the lambert W function. Equation (9-1) was then used to calculate a cleavage rate constant of  $0.1117 \text{ min}^{-1}$  for the Theophylline Control in the presence of 0 mM theophylline using the following parameters\*<sup>31</sup>:  $\%C = 39.77665873$  and  $t = 10$ .

An expression to calculate the time ribozyme inactivation occurs during the TRT reaction was similarly calculated by rearranging Equation (9-3), yielding:

---

\*<sup>31</sup>Parameters acquired from Figure 3-5.

$$t = \frac{1}{k_c} W \left( 0, \frac{100 e^{\left( \frac{100}{\%C - 100} \right)}}{\%C - 100} \right) - \frac{100}{k_c (\%C - 100)} \quad (9-2)$$

Equation (9-2) was then used to calculate an inactivation time of 0.9428 minutes using the following parameters<sup>\*32</sup>:  $k_c = 0.1117$  and  $\%C = 5.085208536$ .

Equation (9-1) was then used to calculate the cleavage rate constant of  $5.4922 \text{ min}^{-1}$  for the Theophylline Control in the presence of 3.16 mM theophylline using the following parameters<sup>\*32</sup>:  $\%C = 80.79652 \%$  and  $t = 0.9428$ .

### 9.3.3 % cleaved and fitness from cleavage rate constant and inactivation/IVT time

An expression for the fraction cleaved of a sequence ( $\%C$ ) given a cleavage rate constant ( $k_c$ ) and an IVT or inactivation time ( $t$ ) was derived based on the work of Long and Uhlenbeck<sup>141</sup>.

$$\%C = 1 - \frac{1 - e^{(-k_c t)}}{k_c t} \quad (9-3)$$

To calculate the fitness of a sequence from cleavage rate constants and inactivation times, Equation (9-3) was substituted into Equation (3-14), yielding:

$$fitness_i = \left( 1 - \frac{1 - e^{(-k_{+c} t_{+})}}{k_{+c} t_{+}} \right) \left( \frac{1 - e^{(-k_{-c} t_{-})}}{k_{-c} t_{-}} \right) \quad (9-4)$$

where  $k_{+c}$  and  $k_{-c}$  are the cleavage rate constants of sequence  $i$  under the selection conditions, respectively, while  $t_{+}$  and  $t_{-}$  are the inactivation times under positive and negative selection conditions.

### 9.3.4 Most-fit ligand-unresponsive sequences

For each inactivation time, the cleavage rate constant which yielded the most-fit ligand-unresponsive sequence was determined using a direct search method in MATLAB. The fitness of each ligand-

---

<sup>\*32</sup> $\%C$  Parameters acquired from Figure 3-7.

unresponsive sequence was calculated using Equation (9-4) for  $10^5$  cleavage rate constants between 0 and 6 min<sup>-1</sup>.

## 9.4 Appendix for Chapter 5

### 9.4.1 Algorithm to remove obvious cheaters

The MATLAB script listed at the bottom of this section was used to remove sequences from in Figure 5-8 that did not have at least 32 nucleotides in either stem-loops I or II. The parameters listed in Table 9-7 are required to implement this script:

| Parameter         | Description                                                                                                        |
|-------------------|--------------------------------------------------------------------------------------------------------------------|
| MSA45con.Sequence | Consensus sequence generated from the multiple sequence alignment in Figure 5-8 using the “multialign” function.   |
| IS(1).Sequence    | Theophylline Library DNA sequence (coding sequence)                                                                |
| MSA45             | Vector of structures generated from the multiple sequence alignment in Figure 5-8 using the “multialign” function. |

*Table 9-7: Parameters required to remove obvious cheaters using algorithm.*

```
%Index library residues to consensus sequence
[~, alignlcon] = nwalignment(MSA45con.Sequence, IS(1).Sequence(17 +
1:length(IS(1).Sequence)));
lcongapi = regexp(alignlcon(3, :), '-');
lconi = 1:length(IS(1).Sequence(17 + 1:length(IS(1).Sequence)));
for i = 1:length(lcongapi)
    lconi(find(lconi == lcongapi(i) - 1):length(lconi)) = lconi(find(lconi ==
lcongapi(i) - 1):length(lconi)) + 1;
end
% Delete sequences which don't have required amount of nucleotides in given
% domains.
IF45IDdel = zeros(numel(MSA45), 1);
for i = 1:numel(MSA45)
    MSA45i = MSA45(i).Sequence;
    if lconi(79) - lconi(19) - numel(regexp(MSA45i(lconi(19):lconi(79)), '-')) >=
32
        IF45IDdel(i) = 1;
    elseif lconi(96) - lconi(83) - numel(regexp(MSA45i(lconi(83):lconi(96)), '-'))
>= 32
        IF45IDdel(i) = 1;
    end
end
IF45IDb = IF45ID(logical(IF45IDdel)); % Indexes of non-obvious cheaters.
```

#### 9.4.2 Conditions used to measure fitness of selected sequences

Negative selection conditions: Individual sequences were incubated under TRT reaction conditions in the presence of 0 mM theophylline.

Positive selection conditions: Individual sequences were incubated under TRT reaction conditions in the presence of 4.9 mM theophylline. A higher concentration of theophylline was used compared to the 3.16 mM theophylline said to have been used during the Theophylline Library selection experiment (Figure 3-15). The explanation for this relates to analysis of the theophylline stock solution used during this experiment. Specifically, theophylline precipitate was identified in this stock solution and when the theophylline concentration was measured as previously described<sup>221</sup>, it was found to be present at 4.9 mM rather than 3.16 mM.

#### 9.4.3 Flow cytometry histograms

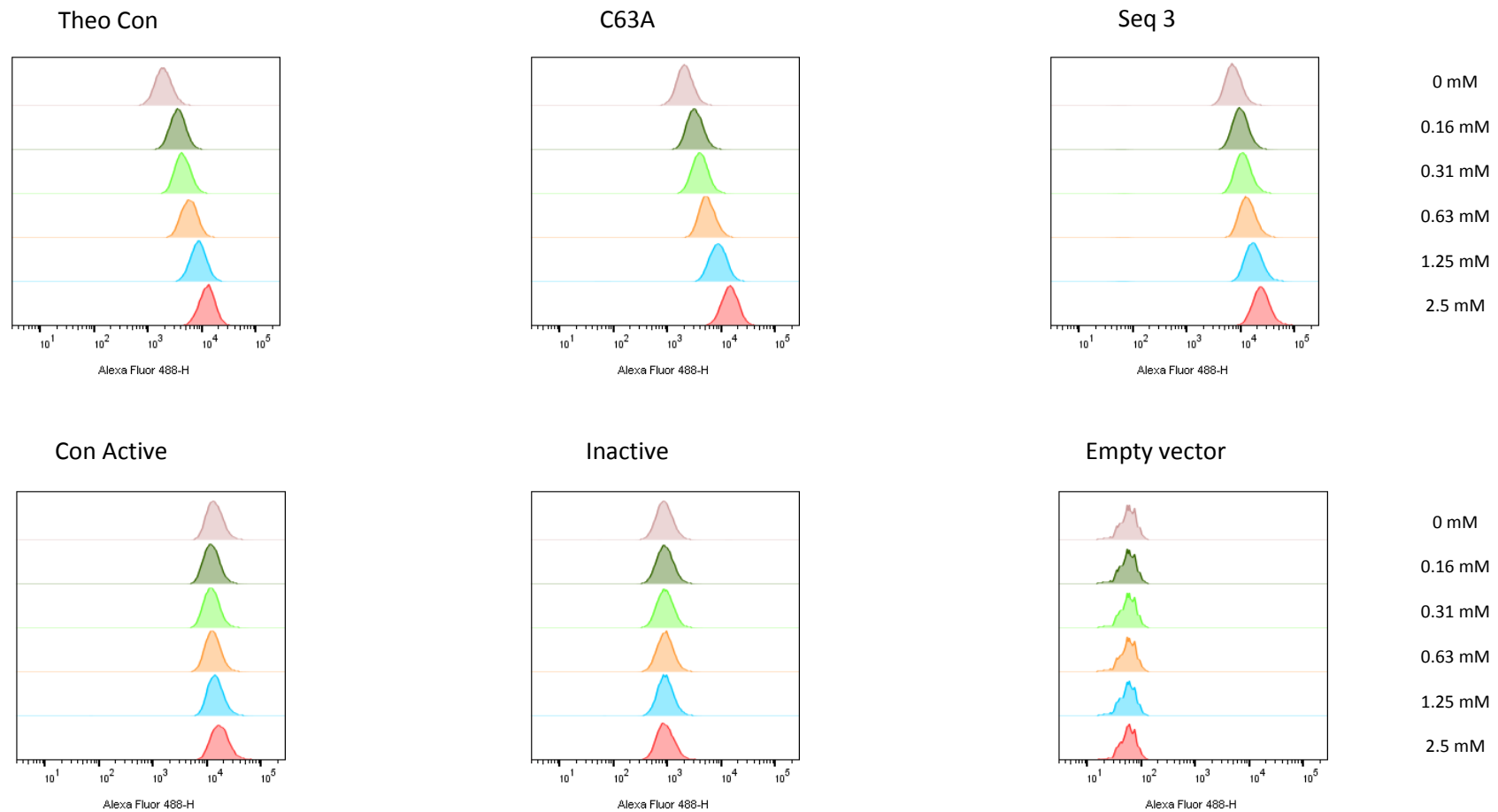


Figure 9-1: Flow cytometry histograms relating to the data in Figure 5-14 - Cells harbouring an empty vector or constructs containing Theophylline Control (Theo Con), C63A variant, Sequence 3, constitutively active or inactive sequences were analysed under variable theophylline (0 – 2.5 mM). The following histograms were plotted on a log scale.