Structure-function relations of RNA molecules involved in gene expression and host defence

Sjoerd C.A. Creutzburg

Propositions

1. No part of mRNA has no influence on translation.
   (this thesis)
2. When applying directed evolution, it is a challenge to avoid undirected evolution.
   (this thesis)
3. Language and life share most of their evolutionary principles.
4. A Bezier-curve can only approximate a circle.
5. Illusions exploit the pattern recognition that is trained by experience.
6. Falsely held beliefs obstruct the quest for truth.
7. Singing and whistling reduce one's anxiety and stress.

Propositions belonging to the thesis entitled:

Structure-function relations of RNA molecules involved in gene expression and host defence

Sjoerd Creutzburg
Wageningen, 20th November 2020

Structure-function relations of RNA molecules involved in gene expression and host defence

Sjoerd C.A. Creutzburg

Thesis committee

Promotor

Prof. Dr John van der Oost

Personal chair at the Laboratory of Microbiology

Wageningen University & Research

Co-promotor

Dr Servé W.M. Kengen

Assistant Professor at the Laboratory of Microbiology

Wageningen University & Research

Other members

Prof. Dr Michiel Kleerebezem, Wageningen University & Research

Prof. Dr Ruud A. Weusthuis, Wageningen University & Research

Prof. Dr Niels Geijsen, Hubrecht Instituut, Utrecht

Dr Stan J.J. Brouns, Delft University of Technology

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Structure-function relations of RNA molecules involved in gene expression and host defence

Sjoerd C.A. Creutzburg

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 20 November 2020
at 11 a.m. in the Aula.

Sjoerd C.A. Creutzburg
Structure-function relations of RNA molecules involved in gene expression and host defence, 210 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2020)
With references, with summary in English

ISBN: 978-94-6395-572-0
DOI: https://doi.org/10.18174/532325

Contents

Chapter 1

General introduction and thesis outline

The PhD project described in this thesis focused on different roles RNA molecules can play: (I) control of gene expression, (II) efficiency of translation and (III) efficacy of host defence systems.

A direct way of RNA to control protein expression is via riboswitches. Riboswitches are specific parts of the mRNA and are most abundant in prokaryotes, where they may modulate either transcription or (functional) translation. Apart from adapting naturally occurring riboswitches, also RNA molecules with catalytic activity, called ribozymes, can be engineered and re-purposed to function as a control element. The engineering as well as the subsequent screening of riboswitches is an arduous task, and several techniques have been used for this, including fluorescence activated cell sorting (FACS). Besides completely inhibiting protein formation, RNA can also regulate the gene expression in a more subtle way. Tuning the rate of translation is potentially possible by adjusting a number of factors including the strength of the ribosome binding site (RBS), the codon use, the secondary structure of the mRNA and the mRNA stability. This thesis also describes an investigation on the role of mRNA design, especially with respect to polycistronic messengers. It focuses on how differences in codon use affects the translation of the whole mRNA and on the role of the RNA downstream of the stop codon. Last but not least, small RNAs are frequently being used as guides of proteins. These ribonucleoprotein (RNP) complexes have distinct functions, ranging from control of gene expression to anti-viral defence. Initially, Argonaute proteins of eukaryotes were demonstrated to function as specific RNA-guided RNA targeting systems (1). More recently, it was discovered that CRISPR-Cas nucleases of prokaryotes also use RNA guides to target complementary nucleotide sequences. In its simplest form, as represented in the Class 2 systems such as Cas9 and Cas12a, it concerns a big multi-domain protein that binds a CRISPR RNA (crRNA) guide. Subsequently, this guide allows the RNP complex to find its target DNA, and the target DNA is cleaved. (2)

RNA IN GENE EXPRESSION - CODING & NON-CODING

Following the central dogma in molecular biology (Figure 1.1) (3), expression and functioning of proteins can be controlled on three basic levels: transcription, translation and post-translation (e.g. modification, complex formation and localisation). For many years the only known molecules that were composed of ribonucleotide building blocks were the RNA molecules directly involved in gene expression: the short living, DNA-derived transcripts (mRNA), and the so-called stable RNAs that are directly involved in translating the code of nucleotide triplets to amino acids (tRNA) and catalysing the polymerisation of these amino acids to (poly)peptides (rRNA). Despite the fact that the overall mechanism of gene expression has been revealed more than half a century ago, there are still details that remain to be discovered. Especially for heterologous gene expression, relevant details are provided on the impact of both the coding region (codon bias) and the untranslated

regions (5′ UTR, 3′ UTR) have been provided in the here-described research. These new insights are anticipated to be useful for optimising future designs of gene and operon expression.

RNA IN CONTROL - RIBOSWITCHES

In general terms, a riboswitch is an RNA element that controls the functionality of the mRNA it is part of. The riboswitch is composed of a ligand binding domain (aptamer) that, upon binding of the ligand, is the prime cause of a conformational change, and an expression platform that translates the conformational change into the riboswitch action. Natural riboswitches respond to various molecules including metal ions (Mg, Ni, Co) (4, 5), amino acids (glycine, lysine) or their derivatives (S-adenosylmethionine (SAM), S-adenosylhomocysteine (SAH)), and co-enzymes co-enzyme B12, (flavin mononucleotide (FMN), thiamine pyrophosphate (TPP)) (6). RNA thermometers could also be considered riboswitches (7, 8). Engineering of these natural riboswitches is challenging, as they are co-evolved with the gene it controls and may not work with another coding sequence. Therefore, the binding domain (aptamer) is generally used to engineer a novel riboswitch instead. Because not all aptamers respond to ligand binding in the same way, some riboswitches employ a communication module, which transduces the signal of the aptamer to the platform.

Aptamers

Aptamers are short single-stranded oligonucleotides that can bind various molecules with high affinity and specificity. The aptamer is the key sensory part of the riboswitch. It can form a docking site for a ligand by secondary and tertiary interactions. In a functional riboswitch, the conformation of the ready-to-bind aptamer has to compete with other mutually exclusive RNA conformations. When the ligand binds, it cements the structure of the aptamer, favouring a certain conformational state of the region. Because of the flexibility of RNA and the potential for alternative base interactions, an aptamer can have

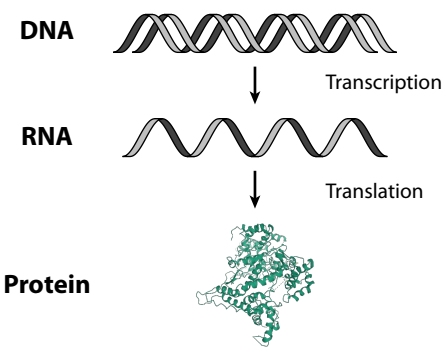


Figure 1.1. Central dogma of molecular biology.

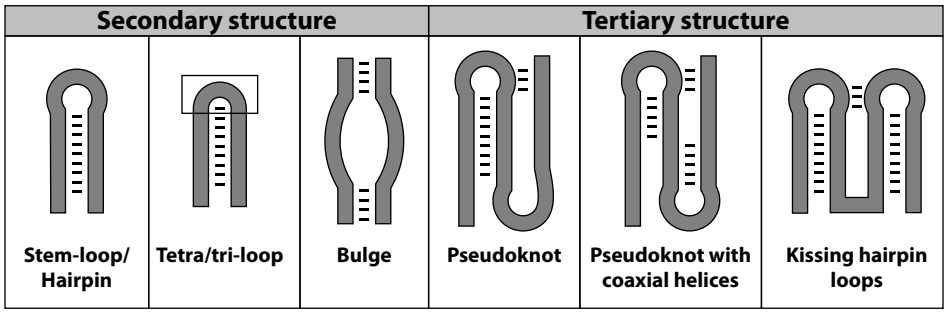


Figure 1.2. List RNA Structures. This list is by no means exhaustive, but covers the most important structures referenced in this thesis.

quite exotic structural motifs like tetraloops, kink loops, pseudoknots and kissing hairpin loops (Figure 1.2) (9). Binding of ligand molecules occurs via interactions like hydrogen bonding or aromatic stacking. The interaction between the aptamer and its ligand can be quite strong (theophylline aptamer has a Kd of 300-400 nM for theophylline (10, 11)) and quite selective against some related compounds (caffeine has a Kd in the mM range).

Transcriptional riboswitches

Inhibition of transcription is a common way of preventing protein expression from a certain gene. This type of control is well known in molecular biology and usually consists of a DNA binding protein that requires a very specific DNA motive to bind to. Upon binding of a ligand, this protein is either released from the DNA (e.g. LacI, TetR), bound to the DNA (e.g. CRP, RhaRS) or it changes from a repressor to an activator (e.g. AraC). Alternatively, transcription can be controlled by riboswitches. Such a transcription-based riboswitch (Figure 1.3A) involves the formation of a potential terminator in the 5' UTR of the mRNA. This terminator usually consists of a strong, high GC stem-loop structure directly followed by a polyU stretch (rho independent terminator). Depending on the type of riboswitch, this terminator is either formed in the presence of a ligand (repressor riboswitch) or disrupted in the presence of a ligand (activator riboswitch). The transcription is still initiated, but then prematurely terminated and no protein is synthesised from that mRNA. Fast folding into the ligand-bound conformational state is imperative. When the terminator formation is completed before the aptamer-ligand folding, transcription will cease. Depending on the aptamer and platform in question, the aptamer folding may be sufficiently fast, or transcription retardation motifs are introduced to allow more time for the aptamer folding to occur. Theoretically, a transcriptional riboswitch could be placed in the intergenic regions of an operon as well, inhibiting the transcription from that point forward.

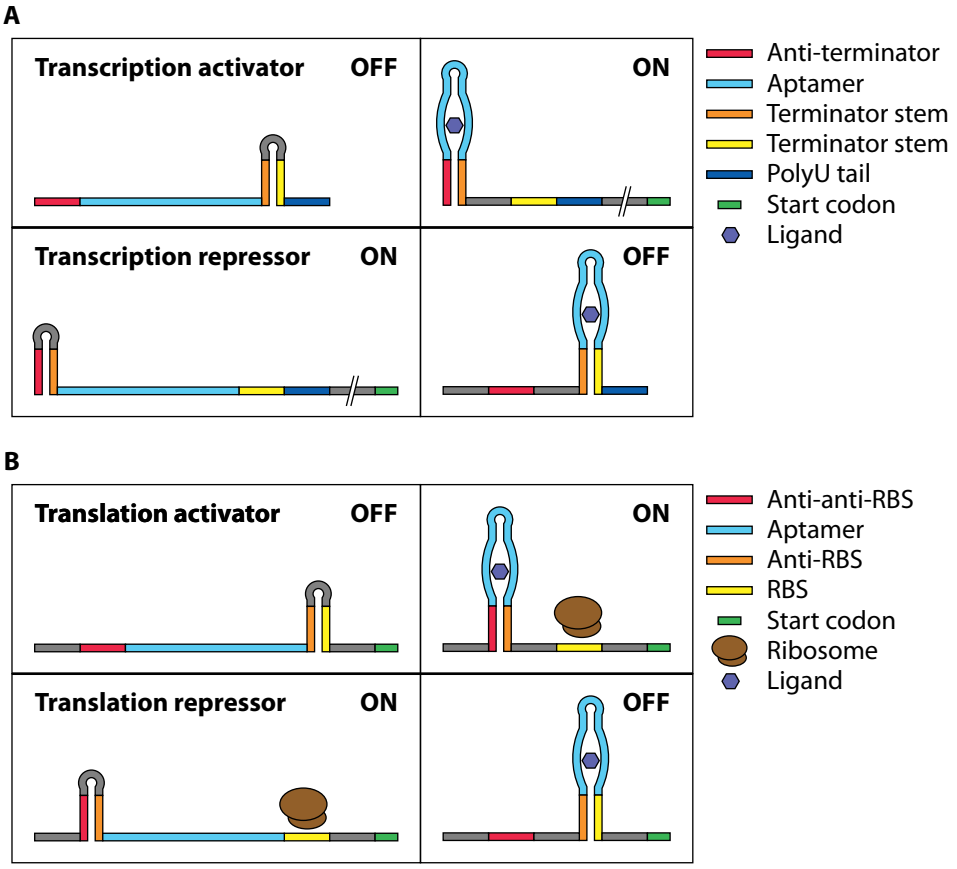


Figure 1.3. General mechanisms of a riboswitch. (A) The transcriptional riboswitch causes the formation of a terminator in its OFF state which is either prevented or induced by a ligand. (B) The translational riboswitch blocks the ribosome from binding to the RBS, thereby preventing translation initiation. The ligand either releases or causes this block.

Translational riboswitches

Translational riboswitches (Figure 1.3B) work in a different way. Transcription occurs completely, but translation cannot be initiated or elongated. Most commonly, the translational riboswitch involves masking of the ribosome binding site (RBS), like the co-enzyme-B12 riboswitch of *Salmonella typhimurium* (12). A part of the 5′ UTR base-pairs with the RBS and prevents the ribosome from binding, thus inhibiting translation initiation. The 5′ region of the coding sequence can be employed for this function as well. When the RBS, start codon and coding region form a stem-loop, the translation initiation is also impeded. Some of the natural riboswitches can be engineered to control other genes. However, this is not true for all translational riboswitches, as some include the coding sequence.

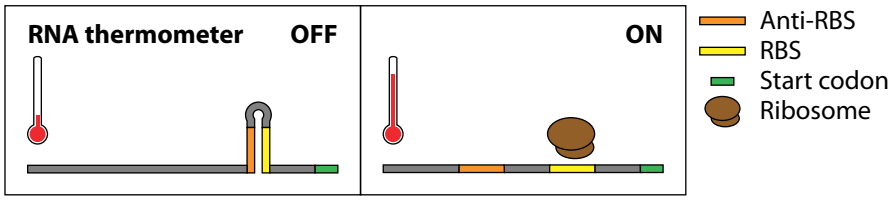


Figure 1.4. RNA thermometer. In its simplest form, the RNA thermometer is a stem-loop involving the RBS, which melts at higher temperature. The melted form is then subject to translation initiation.

RNA thermometers

A group of ligand-independent riboswitches is the RNA thermometers (Figure 1.4). The RNA thermometer is a temperature-sensitive RNA molecule that regulates gene expression. It is mostly observed to be in control of heat-shock and cold-shock protein production. Compared to transcription control by a protein, a faster response is achieved when the encoding RNA is already produced, which is essential in shock situations. The heat-induced riboswitch consists of a stem-loop that masks the RBS and sometimes the start codon as well. These structures are finely tuned to mask the RBS at low temperature, but open up at a few degrees higher temperature. This fine-tuning is achieved by base-pairing, alternative base-pairing, bulges, stem size and loop size. Besides the control of heat-shock proteins, pathogens use the thermometer also to switch on virulence genes when they have arrived in a living host. Cold-shock riboswitches work in a similar fashion, but the structure masking the RBS is now favoured at a higher temperature and the competition structure with exposed RBS is favoured at a lower temperature.

Ribozymes

While most catalytic functions are performed by proteins, ribonucleic enzymes (ribozymes) also play a key role in some cellular processes. For example, RNaseP is a ribozyme involved in the maturation of tRNAs. As an engineering tool, ribozymes are especially suited as many of them function in a variety of hosts. Next to using a ribozyme in its native form, it can be engineered to respond to a signal by coupling the ribozyme to an aptamer. An example of such an "aptazyme" is based on the hammerhead ribozyme, which is involved in the maturation of CRISPR arrays. Cas9 activity requires mature crRNA, which is generated by the aptazyme cleaving off a part of the pre-crRNA. This makes the activity dependent on theophylline (13). A more classical approach of the hammerhead ribozyme as a riboswitch involves the masking of the RBS by the 5′ UTR (14). The 5′ UTR was cleaved off by the theophylline-dependent ribozyme to release the RBS for ribosome binding. An advantage that ribozymes have over more conventional riboswitches, is that the activity can be observed *in vitro* as well. This allows for the riboswitch to be created *in vitro* instead of *in vivo*, which facilitates the screening of larger libraries of riboswitch variants.

A group of ribozymes that plays a major role in this thesis is the Group I self-splicing introns. Compared to simple ribozymes as the hammerhead (~40 nucleotides), the introns are quite large (~300 nucleotides core ribozyme). They are found across the domains of life, including some bacteria, eukaryotic micro-organisms and chloroplasts (15). Also, both prokaryotic and eukaryotic viruses sometimes harbour group I introns. While in viruses, for example, the introns are found in coding sequences, this is not the case for eukaryotic micro-organisms. There, they have only been found in ribosomal DNA (rDNA) (16). Some of the group I introns also harbour a homing endonuclease gene (HEG), which could label them selfish elements. The HEG facilitates transposon behaviour excising the intron from one location and integrating it into another. However, since the presence of the intron in DNA will not abolish the function of the RNA transcribed from it, the phenotypical cost is limited. Other introns are dependent on host factors for their splicing, indicating co-evolution with their host. The most extreme version is an obligatory intron in plant chloroplasts. This intron cannot splice on its own and may serve as a reservoir for non-coding RNA. The HEG typically splits the ribozyme and can be replaced with other sequences as well, without interfering with the splicing. The presence of the HEG also provides a prime target site for engineering the group I intron. The site can be truncated into a simple stem-loop to shorten the intron, or it can be outfitted with an aptamer to engineer it into a riboswitch.

Engineering of riboswitches

Several strategies have been developed to engineer novel riboswitches. The general procedure starts by generating a ligand binding domain (aptamer). Next, a design is made that couples the aptamer to the platform, which usually involves randomisation of the communication module and part of the platform. Finally, the aptamer/platform combination variants are screened for combinations that actually act as a riboswitch *in vivo*.

Finding new aptamers

Generation of a new aptamer starts with a random library of DNA (20-60 nucleotides) flanked by known sequences. This comes in the form of oligonucleotides with a length of around 100 nt. If the aptamer is to be used in a riboswitch, it is transcribed into RNA via *in vitro* transcription first. There are several methods to capture variants that bind to the ligand of interest. The classical systematic evolution of ligands by exponential enrichment (SELEX) (17) procedure has the ligand bound to a column matrix (Figure 1.5A). The RNA library (~10^{15} or nmol order of magnitude) is run through the column and the unbound variants are washed away with increasingly high salt buffers. If the aptamer needs to discriminate between closely related compounds, a wash step with the unwanted ligand is included. Eventually, bound RNA is eluted from the column with the ligand of interest

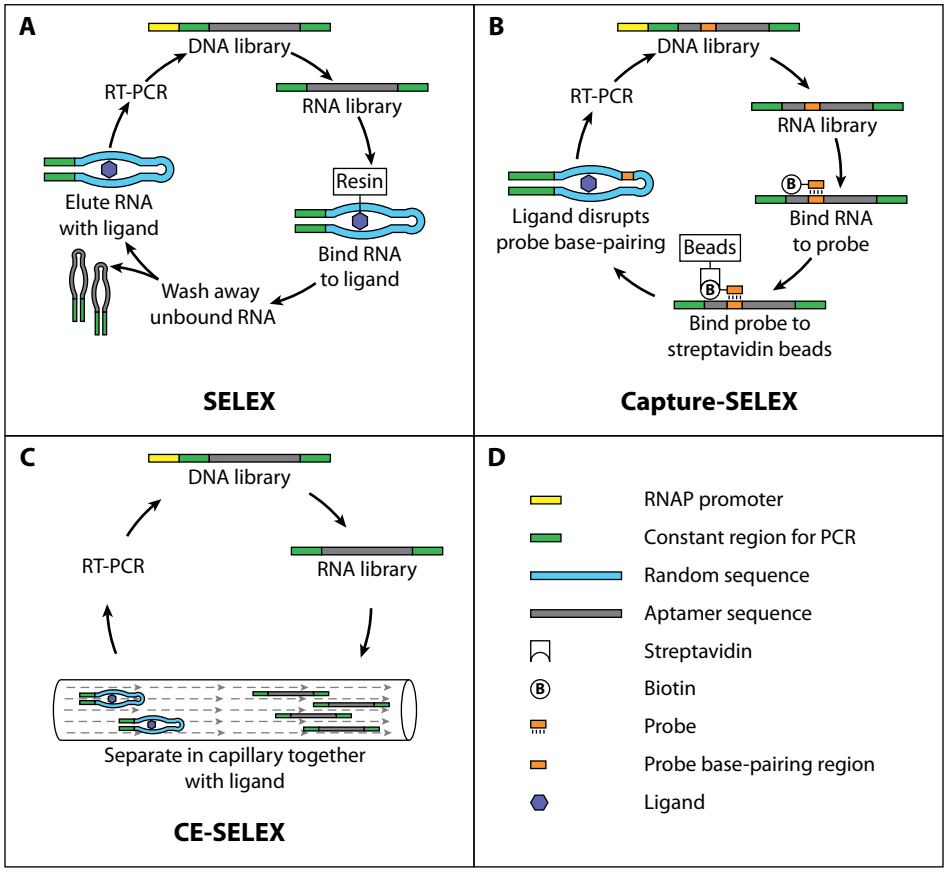


Figure 1.5 Aptamer selection procedures. (A) The classical SELEX procedure consists of several rounds of enrichment. The FluMag-SELEX uses magnetic beads instead of resin. DNA aptamer selection also uses fluorescent labels for DNA quantification. (B) The Capture-SELEX is based on the conformational change of the RNA molecule, so it cannot base-pair with a probe any longer. (C) The Capillary Electrophoresis SELEX (CE-SELEX) separates the heavier ligand-bound RNA molecules from the lighter unbound molecules. (D) Legend of A-C.

and subjected to RT-PCR to re-obtain the now enriched DNA library. With the RT-PCR the SELEX round ends and the next round starts again with *in vitro* transcription. The vast number of combinations possible with for example 60 nt of randomised sequence (10^{36} theoretical variants) greatly exceeds the number of variants in the library. Therefore, instead of high-fidelity PCR after reverse transcription, error prone PCR can be used as well to increase the variation in the RNA that already binds to some extent. After several rounds of SELEX, the aptamer sequences are determined and a consensus is made from that. A variant of the classical SELEX is the FluMag-SELEX (18), which uses magnetic beads instead of a column to immobilise the target and fluorescent labels for DNA quantification (if used for DNA aptamers).

Another variant is the Capture-SELEX (19) (Figure 1.5B), which differs from the previous methods in the way that it does not require the target ligand to be coupled to a matrix. Especially very small target molecules are hard to covalently couple to a matrix and besides that, the coupled target is not exactly the same as the uncoupled target. The Capture-SELEX library consists of a probe target sequence, which is flanked by randomised sequences. The randomised sequences are in turn flanked by PCR annealing sequences. The principle behind the Capture-SELEX is that the binding of the ligand to the oligonucleotide causes a conformational change that cannot allow the biotinylated probe to bind to the probe target sequence. It basically means that ligand-bound nucleic acid is no longer binding the biotinylated probe and therefore stays in solution when exposed to streptavidin beads. The nucleic acid is amplified by (RT)-PCR and subjected to several other rounds of selection.

A method that stands apart from the others is based on capillary electrophoresis (CE-SELEX) (20) (Figure 1.5C). When the ligand binds the oligonucleotide, the mass increases slightly. This mass increase can be used to separate the oligonucleotides bound by their ligand from the unbound ones by capillary electrophoresis. Contrary to the other methods, the authors claim, it can be performed in 3-4 rounds instead of in 6-8 rounds.

Combining aptamer and platform

After the minimal consensus sequence of the aptamer has been determined, it has to be combined with a riboswitch platform and turned into a functional riboswitch. In general terms, the aptamer in its ligand bound form should have the 5′ and 3′ ends in different interactions from the unbound form. If the ends are base-pairing in the same way with approximately the same strength, it cannot be turned into a riboswitch. Judging how the RNA will interact with itself is very difficult, especially when aptamers and ligands are involved. Therefore, randomised libraries of potential riboswitches are made and screened. The targets for randomisation differ per type of riboswitch. For transcriptional riboswitches (Figure 1.3A), the anti-terminator and loop are prime candidates for randomisation. Depending on the screening and selection power, also the terminator stem can be (partly) randomised. For translational riboswitches (Figure 1.3B), the anti-anti-RBS, anti-RBS and loop sequences are randomised. Ribozymes do not require competition between key elements like an RBS or a terminator stem. The aptamer is mostly located on the ribozyme instead of one of the stem-loops. Sometimes, the stem-loop has a specified sequence or length, but if that does not extend to more a few base-pairs, the stem-loop can be replaced by an aptamer. The aptamer takes over the function of the stem-loop if it is ligand bound, but does not do so without ligand. The riboswitch function in these kinds of constructs lies in the way the aptamer is attached to the ribozyme and therefore it is the sequence and size of the interface that is varied. It may be the case that the stem can adopt two conformations. The inactive form is the most stable, but does not result in the

correct base-pairing at the base of the stem and therefore renders the ribozyme inactive. The aptamer can correct the base-pairing at the base of the stem, when it is ligand-bound. However, if the inactive form is too stable, the aptamer cannot correct it. Vice versa, if the stem of the active form is too stable, the contribution of the aptamer is not needed and also no riboswitch is obtained.

Screening and selection of functional riboswitches

The last step in obtaining a functional riboswitch is the screening and selection of the randomised variants for variants that show riboswitch activity. One of the issues with randomisation is the exponential increase in variants for each nucleotide that is randomised (10^3 per 5 nucleotides). This makes screening of individual colonies an arduous task. Several solutions have been proposed to mitigate labour intensity. The most obvious solution is using fluorescence activated cell sorting (FACS) (21). The riboswitch variants control a fluorescent protein gene like *gfp* and the cells harbouring these constructs are grown with or without target ligand. Depending on what the ligand should do, the cells are sorted in a FACS machine and pooled in fluorescent or non-fluorescent populations. In case of an activating riboswitch, addition of ligand should cause fluorescence, so the fluorescent population is retained. Without addition of ligand, the cells should not fluoresce, and the non-fluorescent population is retained. This process is repeated several times until the best variants have been selected. Of course, this procedure requires access to an expensive FACS machine for an extended period of time, so also more low-tech solutions have been proposed.

One of the more low tech solutions is the motility assay (22). In this assay, the riboswitch is made to control the flagellum of *E. coli*. When the bacterial suspension is spotted on an agar plate with target ligand, the bacteria activate their flagellum and move away from the centre. This population is collected and then (after regrowth) spotted on plates without ligand. Cells harbouring functional activating riboswitches should not move in this case and the centre population if collected. After several rounds of repetition, a functional riboswitch is obtained. The most elegant way would be a selection and counter-selection method based on growth. This is the least expensive and labour-intensive way, but it does have some major concerns. The most important thing to keep in mind is that the method does not select for cells harbouring functional riboswitches. Instead, it selects for the cells best surviving the selection and counter-selection, which may or may not be the cells with a functional riboswitch. Escape mutants even may have a growth advantage over the riboswitch carrying cells. This thesis highlights a method that mitigates the danger posed by escape mutants by using the same gene for selection and counter-selection and amplified signal from the riboswitch via T7 RNA polymerase.

RNA AS MESSENGER

RNA is most well-known for its role as messenger (mRNA), bridging the gap between DNA and protein (Figure 1.1). However, there is more to the mRNA than just the coding sequence. The sequence of the RBS influences the translation initiation rate, Whether the ribosome can subsequently bind to the RBS depends on the structure of the mRNA. Sequestering the RBS is a common mechanism of riboswitches, but can be caused by the coding sequence as well. The codons the used in the coding sequence heavily influence its translation rate (23, 24). In an attempt to elucidate how small changes in codon use affect the overall translation, we observed that the expression of both RFP (internal standard) and GFP (codon variants) changed with the codon use of only GFP. This observation led to the investigation of the mRNA as a whole and, finally, it led to the conclusion that all parts of the mRNA can influence its translation.

RNA IN DEFENCE - CRISPR

Since the 1990s, many examples have been discovered of small RNA molecules that are involved in control/regulation (25). A range of different proteins/enzymes use the RNA molecules as guide molecules, using an anti-sense mechanism to specifically associate with target mRNA molecules, resulting either in inhibition of translation, or in destabilizing the transcripts and enhancing their decay. For example, in eukaryotes Argonaute RNA guides allow for RNA interference. Several RNAi applications have been described to reveal gene function, but also to cure certain human genetic diseases (26). In addition, CRISPR-Cas has been discovered 15 years ago in prokaryotes. Different strategies are used to use their crRNA guides to target either RNA of DNA has been. Several CRISPR nucleases such as Cas9 and Cas12a have been re-purposed for a range of applications in biotechnology and in health care. Although several algorithms have been developed to screen for off-target effects (27–29), the functionality of crRNA guides is still rather unpredictable (30). In the course of this PhD project, the crRNA-dependent efficiency of Cas12a is investigated. The Cas12a recognises the repeat-derived pseudoknot structure on the pre-crRNA. However, if this structure is disturbed, Cas12a activity is lost. Several guidelines have been established to improve the folding of the pre-crRNA and enhance Cas12a activity.

Thesis outline

As is introduced in **Chapter 1**, the structure-function relations of RNA are very important for understanding and engineering of key biological processes. It is described how both coding sequences (codon bias) and non-coding sequences contribute substantially to the overall efficiency of gene expression, both of stand-alone genes and genes clustering in operons. While gene expression control by RNA structure in the form of riboswitches is not ubiquitous in nature, it does play an important role in several metabolic pathways. These include the production of cofactors, amino acids and DNA precursors. Naturally occurring riboswitches generally can switch between two states (on/off) due to binding/release of a specific ligand molecule; the riboswitch' conformational change in its turn causes a conformational shift of their target RNA, resulting either in masking or de-masking essential elements involved in gene expression, such as a ribosome binding site, start codon or splice site. Transcriptionally regulating riboswitches either cause formation of terminator stems or prevent their formation upon ligand binding. Naturally occurring ribozymes, like group I and group II introns or the hammerhead ribozyme, do not function as a riboswitch *per se*, but can be engineered to control genes of interest. Crucial for gene editing by using CRISPR-Cas nucleases is the understanding of the design principles of CRISPR RNA guides, which is very important for predictability of guide functionality.

Chapter 2 | Engineering of a group I aptazyme to fit biosensor requirements

This chapter describes the expansion of the applicability of the theophylline dependent group I intron from the phage T4 *td* gene. By optimising the transcription of an intron interrupted *thyA* gene, the induction by theophylline was required for growth of a *thyA* knock-out strain in the absence of thymidine. In addition, the requirement for fine-tuning of the transcription was mostly eliminated by introducing a second intron into the *thyA* gene.

Chapter 3 | Insertion of a theophylline-dependent riboswitch into RNAP

In this chapter the effect of the intron flanking regions was determined. The 5' flank is part of the internal guide sequence and plays a key role in the splicing efficiency. An intron library with several possible interactions of the P1 and P2 stems was made between D6 and S7 of *E. coli lacZ*. β-galactosidase assays revealed that minor changes in the 5' flank may be both detrimental and beneficial for splicing activity. At the 3' flank, diverting from wild-type yielded reduced splice rates. A test case was made of the commonly used T7 RNA polymerase. By controlling the T7 RNAP transcription with rhamnose and its translation with theophylline, the GFPuv transcribed by T7 RNAP was almost completely absent without induction by both rhamnose and theophylline. Moreover, flow cytometry analysis revealed a population-wise induction by theophylline.

Chapter 4 | *In vivo* selection of riboswitches with an altered specificity

Several strategies have been proposed for fitting new aptamers on riboswitch platforms. In this chapter, a selection and counter-selection strategy is described, which makes use of the population-wise induction by theophylline and the use of *thyA* both as a selection and counter-selection marker. The theophylline dependent intron in T7 RNAP was modified into a library of introns that incorporated a random nucleotide in eleven positions simultaneously, yielding a potential variation of 4 million different sequences. The T7 RNAP controlled both *thyA* and *gfpuv* for selection/counter-selection and monitoring. The goal of this experiment was to enrich intron variants that could respond to 3-methylxanthine and not to theophylline. By alternating selection and counter-selection, several intron variants were obtained that met these requirements, eventually dominating the bacterial culture.

Chapter 5 | Grafting of a citrulline aptamer onto the phage T4 td group I intron

In this chapter an attempt has been made to graft a citrulline aptamer onto the group I intron. A strategy similar to chapter 4 was applied and several variants of aptamers were identified. In a Δ*argFGI* background, the citrulline could only be produced via plasmid borne *argF*. The citrulline dependent introns were tested in two ways: the intron interrupting T7 RNAP, which in turn controls the expression of GFP, and the direct interruption of *lacZ*. Both systems show an increased signal when a functional *argF* gene was introduced.

Chapter 6 | Translational feed-forward and feed-back control

This chapter focusses on the mRNA transcript as a whole, rather than on a riboswitch. The codon usage in mRNA transcripts differs significantly between organisms. A 'wild-type' *gfp* gene (encoding Green Fluorescent Protein, GFP) and a codon harmonised *gfp* were placed in a polycistronic mRNA together with a gene encoding a Red Fluorescent Protein (RFP). The order of the genes appeared to have a big impact on their respective translation rate only when the transcription is high. In that case, the first gene in the operon is translated most efficiently, which is not observed at low transcription. The codon harmonisation of *gfp* impacts the translation rate significantly and the impact itself is also dependent on the transcription rate. A low transcription rate allows for larger differences between the codon harmonised and the original *gfp*. A peculiar finding is that the expression of RFP, regardless of its position in the operon, is coupled to the expression of GFP. This observation suggests a continuous train of ribosomes for the whole mRNA. Moreover, addition of a strong synthetic terminator was found to cause a major increase in functional gene expression, at least partly due to enhanced mRNA stability.

Chapter 7 | Medium-throughput *in vitro* detection of DNA cleavage by CRISPR-Cas12a

A method is described for *in vitro* cleavage assays for CRISPR-Cas12a. The method is based on the separation of a biotin and fluorophore on either side of a DNA target molecule. DNA cleavage by Cas12a, or any endonuclease for that matter, causes the fluorophore to stay in solution when streptavidin beads are added. The fluorescence of the solution is then a measure of how much of the DNA has been cleaved. It is most suited for kinetics studies or tens to hundreds of samples, hence medium-throughput.

Chapter 8 | Good guide, bad guide: Spacer sequence-dependent cleavage efficiency of Cas12a

In this chapter, the effect of RNA folding on CRISPR-Cas12a cleavage efficiency is studied. The CRISPR-Cas12a effector proteins use a crRNA guide to find a complementary target DNA molecule. The pre-crRNA is transcribed from a CRISPR array consisting of repeats and spacers. The repeat-derived sequences of the pre-crRNAs are bound by the Cas12a protein and processed to mature guides. It was found that the folding of the pre-crRNA may have a major impact on Cas12a recognising the pre-crRNA. In some cases, imposing an appropriate structure on the pre-crRNA caused significant improvement of the guide functionality, as reflected by cleavage activity.

Chapter 9 | Summary and general discussion

In the last chapter, the results of the different studies of this PhD project on RNA structure-function relations are recapitulated. Moreover, an outlook is presented, and potential future research lines and potential applications are discussed.

Chapter 2

Engineering of a group I aptazyme to fit biosensor requirements

Sjoerd C.A. Creutzburg, Servé W. Kengen, John van der Oost

Abstract

The *td* gene of bacteriophage T4 is interrupted by a group I intron. During infection of its *Escherichia coli* host, the phage DNA is injected and the *td* gene is transcribed. Maturation of the transcript occurs through self-splicing of the group I intron. Previously, a theophylline-binding RNA aptamer has been functionally integrated in the phage intron, resulting in a riboswitch in which the splicing activity of the intron has been made theophylline-dependent. The riboswitch has been transplanted to the *thyA* gene, the *E. coli* homolog of td. Consequently, the mRNA encoding *E. coli* thymidylate synthase (ThyA) was interrupted by the theophylline-dependent intron. Since *thyA* is an essential gene in the pyrimidine synthesis, the disruption of the *thyA* gene is toxic in the absence of thymidine in the medium. In this study, the intron was inserted into the *E. coli thyA* gene at two sites. One at the position as it is found in the phage T4 *td* gene and/or the other at a new position upstream. Since the ribozyme-flanking sequences in both the 5' exon and the 3' exon directly influence the splicing efficiency, the ribozyme insertion site was carefully selected to alter as little in the RNA sequence of the ribozyme as possible, while the amino acid-encoding sequences of the flanking exons were to remain intact. The insertion of a second group I intron into the coding sequence of *thyA* resulted in an increase of stringency compared to a single insertion of either intron. No background growth was observed with two introns, and full growth was obtained by induction with theophylline.

Introduction

A biosensor is generally composed of a sensor, a signal transducing element and a reporter. In *in vivo* biosensors, the reporter may be a gene that encodes a protein that acts either as a selection marker (antibiotic resistance, auxotrophy complementation) or as a screening marker (chromogen, fluorescence, luminescence). The control element (sensor and signal transducing unit) can act at different stages in the protein production process. Protein-based control elements like the Lac repressor (LacI) typically intervene at DNA level with the transcription of the gene, while inteins and post-translational modification systems act by activation at protein level. In addition, control at RNA level is possible by riboswitches through control of translation.

Natural riboswitches are ligand-dependent, gene regulating RNAs that are typically located in the 5'-untranslated region (5'-UTR) of an mRNA transcript. In the absence of the ligand, in its unbound state, the ribosome binding site (RBS) of the transcript is sequestered by a complementary anti-RBS sequence, thereby preventing translation. Upon binding of the ligand to the aptamer domain of the riboswitch, the anti-RBS is released from the RBS, allowing translation (6). Alternatively, the unbound state could involve a free RBS that is sequestered upon ligand binding (6). Another possible mechanism is the formation of a terminator structure either in the presence or absence of a ligand (31). It may be difficult to alter the ligand specificity of these riboswitches, when the anti-RBS or anti-terminator is part of the aptamer domain. In such cases, altering the aptamer domain will also change the anti-RBS or anti-terminator rendering the riboswitch inactive. A solution to that problem would be to randomise the 5'-UTR and to screen for riboswitch activity (32). Another possibility would be to engineer a ribozyme, either a synthetic or a natural one, into a riboswitch by introducing an aptamer domain. This would create an allosterically controlled ribozyme, also referred to as an aptazyme.

A well-characterized example of a functional aptazyme design is based on the hammerhead ribozyme, which is an endonuclease. It relies on the 5'-UTR sequestering the RBS to block translation initiation. The ribozyme has been coupled to an aptamer to cleave off the part of the 5'-UTR sequestering the RBS only when a ligand is bound. (33, 34). A similar synthetic construct has been applied in the 3'-UTR of eukaryotic transcripts, where the ribozyme cleaves off the poly-A tail upon binding by a ligand. For a review on synthetic aptazymes, see (35).

A distinct type of ribozyme concerns group I introns. Upon transcription of a group I intron-containing gene, maturation of the mRNA transcript occurs through self-splicing of the group I intron due its typical catalytic RNA structure. Previously, a theophylline aptamer has been functionally integrated in the group I intron that resides in the *td* gene of the T4 bacteriophage, resulting in a riboswitch in which the splicing activity of the intron has

been made theophylline-dependent (36).

The intron-based aptazyme system has several properties that makes it suitable as an *in vivo* biosensor. For instance, when the intron would be unfolded, no splicing will occur as the ribozyme activity will be lost, implying that it still prevents functional translation of the marker gene. On the contrary, in case of riboswitches that block the RBS, unfolding of the riboswitch part of the mRNA would release the obstruction that would result in undesired leakage. A potential general problem of aptamers is the fact that leakage can occur when, in addition to the desired on/off secondary structures, intermediate aptamer conformations are formed when no ligand is present, as this would result in undesired functional translation of the marker.

In the present study, we describe transplantation of the aforementioned synthetic aptamer/intron-based riboswitch (36) to the *td* homolog of *E. coli*, the *thyA* gene at the previously used location as well as at a new upstream location. Two variants with single introns, and one with an intron pair were analysed. As *thyA* is an essential gene that encodes an enzyme in the pyrimidine synthesis, the disruption of the *thyA* gene is generally toxic. Complementation of the *thyA* disruption is possible either by adding thymidine to the medium or by supplying the gene on a plasmid. In addition, survival may correlate with the efficiency of the theophylline-induced splicing of the intron. The latter strategy has been used in this study to optimize the experimental conditions and the design of integration with respect to riboswitch functionality.

Results and discussion

Auxotrophy complementation with *thyA* under control of a theophylline-dependent riboswitch

The best type of biosensor shows no signal without the presence of ligand, full signal in the presence of a high concentration of ligand, and medium signal with intermediate concentrations of ligand. The translation of theophylline-controlled thymidylate synthase (ThyA) in the absence of theophylline is never exactly zero, but it can be insufficient to support growth in an auxotroph. Depending on the transcription rate, the amount of ThyA may never exceed the required minimum, always exceed it, or exceed it only in the presence of theophylline. As only a low amount of thymidine is present in most widely used *E. coli* media, all experiments can be performed on rich media like LB.

ThyA catalyses a crucial step in the pyrimidine synthesis. It catalyses the reductive methylation reaction that converts dUMP into dTMP, using 5,10 dimethylene-THF as methyl-donating cofactor. dTMP is a precursor of dTTP which is required for DNA synthesis. dTMP can also be derived from thymidine when this is present in the medium. In case of *E. coli* DH10B, thymine cannot be used as dTMP precursor, as this strain lacks the enzyme

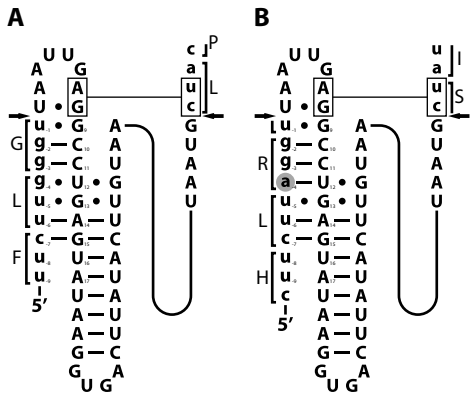


Figure 2.1. Secondary structure of natural and engineered T4 self-splicing introns. Exon sequences are in lower case, intron sequences in upper case and the arrow indicates the splice sites. (A) The natural phage T4 *td* intron as it is inserted into the *E. coli thyA* gene at position F171-P175 (36). (B) The engineered T4 *td* intron to fit a second location in the *thyA* gene at position H51-I55. A single point mutation, which does not abolish the intron's splicing, is made in the 5' exon ("g" to "a" at position -4).

that catalyses the conversion of thymine to thymidine. Hence, ThyA-deficient bacteria can only grow on (minimal or rich) medium to which thymidine has been added and are therefore ideal auxotrophs for developing a *thyA*-based biosensor.

A synthetic theophylline-dependent riboswitch, composed of the self-splicing intron of phage T4 and an integrated theophylline aptamer, has been described before (36). In the latter study, the intron has been engineered in the *thyA* gene of *E. coli*; although the exon boundaries (i.e. the flanking regions of the intron) participate in the functional ribozyme's secondary structure (Figure 2.1A), the intron could be integrated in the bacterial gene by introducing silent mutations only (36).

Reporter plasmids were constructed that carry a p15A origin of replication derived from pACYC184, a kanamycin resistance gene from pET24d and both the 5'-UTR and CDS of the *E. coli thyA* gene (with the aforementioned riboswitch intron). A terminator (37) and promoters of different strength were placed upstream of the 5'-UTR (82). *E. coli* DH10B-Δ*thyA* carrying the reporter constructs were grown with different amounts of theophylline and monitored for 20h. The growth rate was determined in biological triplicates of each construct and for each theophylline concentration (Figure 2.2).

Since the inoculate was grown on medium containing thymidine, the cells always grow on the residual thymidine. However, no growth above OD600 of 0.040 was observed for bacteria expressing *thyA* under control of the Para, Pbla, Pcat and Plac promoters. These promoters are the weakest of the set and therefore do not support exponential growth. Interestingly, they do not stop growing at the same density. Depending on the promoter strength and amount of theophylline, the bacteria extend their growth by supplying a little ThyA in different quantities. However, when the thymidine is depleted, growth cannot be supported by ThyA alone.

Theophylline-dependent exponential growth was observed for PlacUV5, PtacI and Ptet. These promoters more closely resemble the consensus sequence of the -35 and -10 regions (Table 1.1). Better growth is to be expected when the auxotrophy complementation is

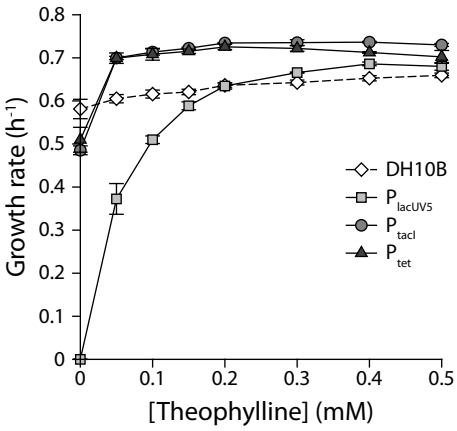


Figure 2.2. Growth rate of single intron constructs with constitutive promoters of different strengths at different theophylline concentrations. The different strains were cultivated in microtiter plates. The constructs with promoters PtacI (circle) and Ptet (triangle) show near maximum growth while not being induced and maximum growth with slight theophylline induction. The PlacUV5 (square) construct shows optimal theophylline dependence, having no growth without induction and a dynamic growth rate range (0 h^{-1} – 0.69 h^{-1}) at theophylline concentrations between 0 mM – 0.4 mM. The promoters Para, Pbla, Pcat and Plac do not support exponential growth nor does the negative control (frame shifted *thyA*) (not shown). *E. coli* DH10B containing the frame-shifted *thyA* serves as positive control (diamond).

under control of a stronger promoter. The promoters PtacI and Ptet are strong enough to support exponential growth without induction by theophylline. Contrary, the PlacUV5 promoter does not (Figure 2.2) support growth without theophylline. The positive control, *E. coli* DH10B with the plasmid borne frame-shifted *thyA* gene, reveals a minor growth stimulation by theophylline (Figure 2.2).

Auxotrophy complementation indirectly depends on the concentration of mature mRNA. The concentration of mature mRNA in its turn depends on both the concentration of immature mRNA and the maturation rate. The concentration of immature mRNA is mostly dependent on the transcription rate (promoter strength) and degradation rate, while the maturation rate is dependent on the intron splicing. The intron splicing, in turn, is enhanced by theophylline binding. Apparently, the maturation rate does not equal zero when no inducer is present. This leakage is observed when analysing the constructs having a strong promoter (Ptet, PtacI). When there is no uninduced splicing, the promoter strength should have no effect without the theophylline inducer. The weak promoter constructs do not generate enough mature mRNA even when the maturation rate is high. In those cases, the amount of ThyA is not enough to reach the minimal concentration of dTTP required for the cells to grow. A dTTP concentration below the minimal requirement will result in what is called thymineless death. It appears there is a narrow range between never enough ThyA and always enough ThyA. The balance is matched rather well with the PlacUV5 promoter: on microtiter plates no growth is observed when no inducer is present and maximum growth is observed at full induction (Figure 2.2).

Although uninduced growth of *E. coli* DH10B-Δ*thyA* carrying the PlacUV5 construct was not observed in microtiter plates, it was observed in 5-mL cultures in 50-mL Greiner tubes. Evaporation is a serious issue in the microtiter plate, especially after several hours

of growth. By that time, all exponential growth had stopped already and carry-over thymidine was consumed, preventing further growth. The bacteria cultivated in Greiner tubes, however, did not suffer from evaporation, so a very small subpopulation having a slightly increased *thyA* expression may become dominant overnight. It indicates that the background expression of ThyA is only just below the minimal requirement sometimes exceeding it. This very precise tuning weakens the robustness of the system as a biosensor.

Introduction of a second intron into the *thyA* coding sequence

The strong promoters PtacI and Ptet caused exponential growth in the absence of theophylline. This suggests that apparently some mRNA matured in the absence of theophylline, leading to a sufficient high level of functional ThyA. To reduce this undesired mRNA maturation, a second intron was introduced into the coding region of *thyA*. No other part of the *thyA* gene matches the native flanking regions of the phage T4 *td* intron, so another strategy was applied. To find a matching insertion site presents a challenge as the intron ribozyme is composed not only of the intron region itself, but of the 3' end of exon 1 and the 5' end of exon 2 as well. The intron flanks are therefore part of both the ribozyme and the coding region (Figure 2.1).

The flanking sequences do not need to match the native flanking sequence perfectly for a functional phage T4 *td* intron (38). Next to some tolerance in the intron flanking regions, the coding sequence can be composed of different codons. An algorithm was written to analyse the *thyA* coding sequence for possible locations for the intron to be inserted. The position had to meet several requirements: (I) mutations in the coding sequence were to be translationally silent for both the intron flanking regions and the restriction sites to clone the second intron into the *thyA* gene, (II) the flanking regions of the second intron had to match the flanking regions of the native intron as closely as possible, (III) no mutations in the flanking regions were allowed other than described by Pichler et al. (38), and (IV) the possibility of introduction a (translationally silent) restriction site next to either flank was preferred.

The best option was an insertion identified as HLRSI (amino acids 51-55) with only one mutation in the intron flanking region changing a G•U wobble base pair (Figure 2.1A) into a A•U base pair (Figure 2.1B). Two unique restriction sites could be introduced close to the insertion site: Psp1406I upstream and PstI downstream. A construct with a tandem intron at (H51 - I55) and (F171 - P175), and a construct with only the (H51 - I55) intron were made. These were tested and compared to the original single intron (F171 - P175) construct. In all cases, the *thyA* gene was under control of the PtacI promoter. The constructs were tested in *E. coli* DH10B-Δ*thyA* according to the same protocol as used for the F171-P175 single intron construct (Figure 2.3).

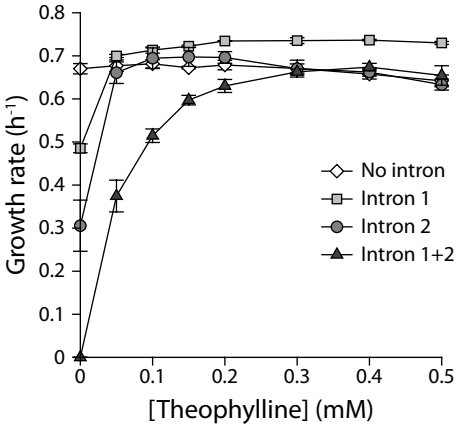


Figure 2.3. Growth rate of single and double intron constructs under control of the PtacI promoter at different theophylline concentrations. Intron 1 (square) is the intron at (F171 - P175), intron 2 (circle) is the intron at (H51 - I55). Introns 1 and 2 in tandem (triangle)s shows a theophylline dependence with a dynamic growth rate range of 0 h-1 – 0.67 h-1 between 0 mM – 0.4 mM theophylline.

The goal of reducing the background growth as a result of splicing in the absence of theophylline was accomplished by introducing the second intron. The tandem intron completely prevented the background growth even with the PtacI promoter (Figure 2.3). Contrary to the single intron construct, the tandem intron construct also did not grow in Greiner tubes in the absence of both thymidine and theophylline. The second intron (at H51 - I55) alone shows a slightly lower ThyA expression compared to the first intron (at F171 - P175), suggesting that the background intron splicing is less. This result may be caused by the difference in position, the difference in sequence or both. The secondary and tertiary structure of the RNA surrounding the intron as well as the actual sequence of the intron flanking regions may affect the splicing rate. Both are different for intron 1 and 2. Furthermore, by introducing the silent mutations for the restriction sites and the intron flanking regions, the amino acid sequence may not be altered, but the codon usage is. The difference in codon usage may of course affect the ThyA expression.

Conclusions

The holy grail of biotechnological improvement of production strains, would be a laboratory evolution approach in which variants with elevated production yields could be selected by a growth advantage. For that reason, a riboswitch composed of the phage T4 *td* intron combined with a ligand-binding aptamer can be a useful tool for selection of *E. coli* variants that survive/grow faster because they produce a desired target molecule (the ligand) in their cytoplasm. The ability of this system to completely select against bacteria that have no ligand makes it relatively straightforward, provided that the riboswitch is stringently controlled. A property the *td* intron-based riboswitch shares with many other riboswitches is its transferability to other organisms, as riboswitches are not affected by post-translational modification. Group I introns have the extra advantage that no species-specific elements like an RBS sequence or a poly-A tail are involved.

Uninduced and fully induced splicing must be carefully balanced so the bacteria without the ligand do not grow at all, while the bacteria with the ligand do. We here demonstrated that the PlacUV5 promoter does balance the leakage and the induced expression so that the dynamic range is between 0 mM and 0.4 mM theophylline, resulting in a growth rate between 0 h^{-1} and 0.69 h^{-1} in microtiter plates. When the balance is hard to find, there is a possibility of introducing a second intron at an upstream position. Although the tandem introns will require elevated transcription rates, they are also more effective in reducing background splicing, while maintaining the dynamic range in both inducer concentration and growth rate.

Acknowledgments

The authors are grateful to Teunke van Rossum for stimulating discussions and aid in generation of the *E. coli* DH10B-Δ*thyA* strain.

Materials and methods

Chemicals and plasmids

Thymidine and theophylline and were purchased from Sigma-Aldrich. A plasmid containing the *E. coli thyA* gene interrupted by a modified phage T4 *td* intron between G173 and L174 was commissioned from GeneArt (pMA-ThyA-SI001) as well as an intron version containing a theophylline responsive aptamer (pMA-ThyA-Theo). Plasmid pET24d was purchased from Novagen. Plasmid pRham C-His was purchased from Lucigen. Enzymes were purchased from Thermo Scientific and used according to the manufacturer's instructions, unless stated otherwise.

Bacterial strains and media

E. coli DH10B T1R was purchased from Invitrogen (C6400-03) and used for plasmid propagation and standard molecular techniques, as well as parent strain for the *thyA* deficient *E. coli* DH10B-Δ*thyA* strain. Bacteria were generally grown at 37°C on LB medium (10 g/L peptone, 5 g/L yeast extract, 10 g/L NaCl) containing the appropriate antibiotics: kanamycin (50 mg/L), ampicillin (100 mg/L), chloramphenicol (35 mg/L) and tetracycline (15 mg/L). In addition, the auxotrophic *E. coli* DH10B-Δ*thyA* was complemented with thymidine (100 mg/L) when necessary. Transformation was performed with an ECM 63 electroporator (BTX) at 2500 V, 200 Ω and 25 µF, 2 mm cuvettes, 20-40 µL of electro-competent cells and recovery in LB.

Construction of reporter plasmids

The reporter plasmids pSC018a-g - Theo were constructed using pACYC184 as a base. The steps include exchange of the chloramphenicol acetyltransferase (cat) for the aminoglycoside 3'-phosphotransferase (kan) from pET24d (Novagen), exchanging the TetA(C) for the *thyA* gene encoded on the pMA-ThyA-SI001 plasmid and exchanging the 6b hairpin for the theophylline responsive aptamer from pMA-ThyA-Theo).

Promoter variants were made by polymerase chain reaction (PCR) and ligating the PCR product into pSC018f-Theo between the KpnI and BcuI sites. pSC022f-Theo was constructed by cloning a second theophylline responsive intron into pSC018f-Theo between R53 and S54 using the Psp1406I and PstI sites. The second intron was generated by PCR using pMA-ThyA-Theo as a template. pSC024f and pSC026f-Theo were constructed by using pSC018f-Theo and pSC022f-Theo respectively as template for PCR. Ligation of the PCR products into pSC018f between the Acc65I and MluI site removed the intron between G173 and L174, leaving no intron in pSC024f and one intron between R53 and S54 in pSC026f-Theo. A frame-shift construct in the *thyA* gene of pSC024f was constructed by digestion with MluI, Klenow fragment 5' fill-in and re-ligation of the plasmid.

DNA purification was performed with the DNA Clean & Concentrator-5 kit of Zymo Research (D4004) or the Zymoclean™ Gel DNA Recovery Kit (D4002). Plasmid was isolated with the Plasmid Miniprep kit of Thermo Scientific (#K0503). Ligation was performed at 22°C for 1h, followed by 10 min heat inactivation. All plasmids were verified by PCR and/or restriction analysis and sequencing by GATC Biotech (Konstanz, Germany).

Construction of the thymidine synthase deficient strain

The *thyA* deficient strain DH10B-Δ*thyA* was made according to Datsenko and Wanner, 2000 (39) with the exception of the PCR template and the competent cells protocol and the PCR template for the insertion cassette.

Electro-competent cells were made by growing DH10B T1R (Invitrogen) containing pKD46 at 30°C on 16 g/L peptone, 10 g/L yeast extract and appropriate antibiotic to an OD600 of 0.4 and cooled down to 4°C, washed with ultrapure water once and 10% glycerol twice. Finally, the bacteria were concentrated 250x in 10% glycerol.

DH10B containing pKD46 was transformed with a PCR product generated from pMA-RQ-Lox71-kan-Lox66, kindly provided by Teunke van Rossum, containing a kanamycin resistance gene flanked by Lox71 and Lox66. The Lox sites can be recombined by cre recombinase removing kanamycin resistance, but do not form a functional Lox site. Transformed bacteria were recovered in LB medium containing thymidine (100 mg/L) for 2.5 h at 37°C and plated on LB agar plates containing kanamycin (50 mg/L) and thymidine (20 mg/L). Colonies were verified for *thyA* deficiency by plating on LB agar

plates containing kanamycin (50 mg/L). Plasmid curation was assessed by growing on LB agar plates containing ampicillin (100 mg/L) and thymidine (20 mg/L).

Electro-competent cells were made from DH10B-Δ*thyA*-kan growing on medium containing kanamycin (50 mg/L) and thymidine (100 mg/L) at 37°C and transformed with pJW168 containing the cre recombinase. Auxotrophy, recombination of the Lox sites and plasmid curation were assessed by plating on LB agar medium, LB agar containing kanamycin (50 mg/L) and thymidine (20 mg/L) and plating on LB agar medium containing ampicillin (50 mg/L) and thymidine (20 mg/L). Electro-competent cells were made of the knock-out strain and transformed with the auxotrophy reporter constructs.

E. coli DH10B-Δ*thyA* growth assays

E. coli DH10B-Δ*thyA* containing a reporter construct of the pSC series were grown overnight at 37°C on LB medium containing kanamycin (50 mg/L) and thymidine (100 mg/L). A 10^{-4} dilution was made and grown with different concentrations of theophylline in a 96-well microtiter plate (Greiner) in a final volume of 200 µL. Culture plates were incubated under continuous shaking for 20h at 37°C and the OD600 was measured every 10 minutes in a Synergy MX plate reader. As carry-over thymidine allows the knock-out strain to grow without ThyA, the OD600 had to exceed 0.040 AU to be counted as growth. Growth rate (µ) was calculated from at least 1h of exponential growth exceeding an OD600 of 0.040 according to

$$\ln(\mathrm{OD}_{600\_t}) = \ln(OD_{600\_0}) \cdot \mu \cdot t$$

Table 1.1. Promoter sequences.

Name	Sequence
Consensus	TTGACANNNNNNNNNNNNNNNN-TATAAT
P_ara	CTGACGCTTTTTATCGCAACTC--TCTACT
P_bla	TTCAAATATGTATCCGCTCATGA-GACAAT
P_cat	ATGAAATAAGATCACTACCGGGCGTATTTT
P_lac	TTTACACTTTATGCTTCCGGCTCGTATGTT
P_lacUV5	TTTACACTTTATGCTTCCGGCTCGTATAAT
P_tacI	TTGACAATTAATCATCGGCTCG--TATAAT
P_tet	TTGACAGCTTATCATCGATAAGC-TTTAAT

Chapter 3

Insertion of a theophylline-dependent riboswitch into RNAP

Sjoerd C.A. Creutzburg, Evans Asamoah Gyimah,

Servé W.M. Kengen, John van der Oost

Abstract

Allosteric ribozymes are a type of riboswitch responsive to several compounds. These tuneable catalytic RNAs are also referred to as aptazymes. One such aptazyme is the theophylline-dependent group I intron. This intron can be inserted in a gene of interest, and only after theophylline-induced splicing, functional expression of the gene will occur. Instead of using this riboswitch to control the expression of an individual gene, we used it here to control the expression of the T7 RNA polymerase (RNAP), that, in turn, controls the expression of a gene of interest. Advantages of this approach are that low levels of RNAP will result in large amounts of protein, and that multiple genes can be controlled with only one riboswitch. However, the introduction of a group I intron into a gene is not straightforward, because the bases flanking the intron are part both of the coding sequence and of the ribozyme. Changing part of the coding sequence may result in a non-functional protein, while changing the ribozyme might impair the splicing activity. We used the *E. coli lacZ* gene as model for determining the effect of changing the flanking regions on the functional gene expression. Depending on the actual combination, changing the flanking regions gave an 80% decrease to 40% increase in protein level. This suggests that although many combinations allow splicing, the flanking regions determine the splicing rate markedly. Based on the flanking region analysis with *lacZ*, an algorithm was made to identify possible intron introduction sites. This tool was used to introduce the theophylline-dependent group I intron at three selected positions. The gene encoding the T7-RNAP was under control of the *E. coli* rhaBAD promoter. After integration into the *E. coli* DH10B genome, the functional expression of T7-RNAP was tested by expression of GFPuv under control of the T7 promoter. Without induction by both rhamnose and theophylline, the GFPuv fluorescence is hardly detectable. Induction with either rhamnose or theophylline results in minimal functional expression of RNAP. Only when both inducers are present a good expression was obtained. Analysis by flow cytometry revealed a fluorescent population that increases with increasing concentration of theophylline, while the non-fluorescent population decreases.

Introduction

In vivo biosensors have a wide range of potential applications. The biosensor can be used to monitor intracellular concentrations of metabolites and enzyme activity (40–42). Apart from monitoring, a biosensor can be used as a switch, for instance to turn on the production of enzymes for novel pathways in metabolic engineering or synthetic biology (43). Different applications require different genes to be controlled. If a set of stand-alone genes (not clustered in an operon) is to be induced with a xenobiotic, the genes all need a biosensor that controls them. Riboswitches make for excellent artificial biosensors (41, 44, 45). To be widely applicable, biosensors require binding of different ligands. These ligand binding domains, or aptamers, can be found using high throughput methods like systematic evolution of ligands by exponential enrichment (SELEX) (11), or by allosteric selection (46).

One potential biosensor is derived from the phage T4 *td* group I self-splicing intron (Figure 3.1A). By grafting an aptamer onto the P6 stem, the intron can be modified into an allosteric ribozyme, or group I aptazyme. In previous work, the phage T4 *td* group I self-splicing intron was successfully adapted to respond to the binding of theophylline and splicing itself out of the *thyA* gene of *E. coli* (36). Adapting the intron insertion site of *E. coli thyA* is rather straightforward. The *td* and *thyA* genes are homologues and have identical amino acid sequences surrounding the intron insertion site. However, turning the intron into a general, widely applicable synthetic riboswitch requires the transferability to other organisms and other genes.

Introducing the group I aptazyme biosensor into other genes is not straightforward, as part of the ribozyme is also part of the gene coding region. Comparing the phage T4 *td* intron with its relatives in the group IA reveals regions that may tolerate nucleotide substitutions. The intron flanking region is at least ten nucleotides long with several nucleotides that are part of conserved structures (47–50). Downstream of the 5′ splice site is a highly conserved wobble base pair and domain P1 extends to three nucleotides upstream of the 5′ splice site, i.e. these three nucleotides are part of the 5′ exon (Figure 3.1C). Changing these nucleotides and retaining the P1 structure requires changing the internal guide sequence (IGS), which risks rendering the intron defective. A similar case is the domain P10 on the 3′ end of the intron. Two nucleotides of the intron interact with the 3′ exon. Changing the interactions of P10 might also cause impediment of splicing (Figure 3.1C). For the group IA introns, the length of P10 is generally at least two base pairs, but the effect of extending it is unknown. Since the *td* intron has only two pairs, shortening it even further is undesired.

Apart from the highly conserved structures of P1 and P10, the P1 shows possible competition with P2 extending to seven nucleotides upstream of the 5′ splice site (Figure

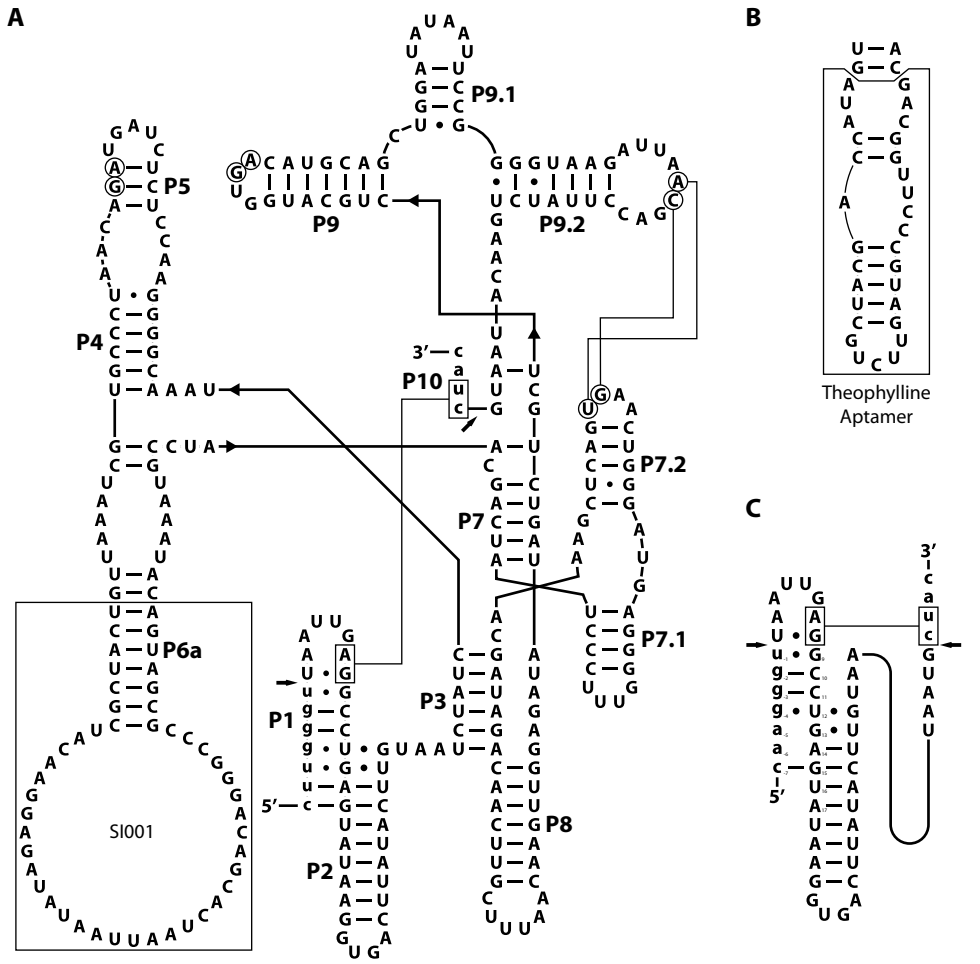


Figure 3.1. Predicted secondary and tertiary structures of the SI001 intron and the theophylline dependent intron. (A) T4 *td* intron inserted into the *lacZ* gene. Structure follows the format of Cech et al. (51). Uppercase letters indicate the intron, lowercase letters the exons. Arrows indicate the splice site. Boxed is the part that can be replaced by the theophylline aptamer to generate a theophylline-dependent aptazyme. (B) Theophylline aptamer grafted onto the intron inserted into the T7 rnap gene. (C) Flanking regions of the T7 rnap introns.

3.1A). Retaining the exact sequence for all of these exon-carried nucleotides does not leave much freedom for the insertion site. One solution is identifying sites in the protein where inserting a couple of amino acids has no significant effect, for instance in a loop region on the exterior of the protein structure. The amino acids that will be inserted depend on the frame. With the splice site in frame 1, this sequence is 'XLGL'. In frame 2 it is 'XWVX' and in frame 3 it reads 'LGSX'. In these instances, the 'X' represents a restricted choice of amino acids. Thus, inserting an intron including its flanking region requires knowledge about the protein it will be inserted into. A second solution is to search the protein-of-interest for a

sequence that can be encoded by the DNA sequence of the flanking regions. The protein sequence should then exactly match one of the possible protein sequences encoded by the flanking regions.

A more interesting option is finding potential insertion sites that do not, or only partially, match with the wild type intron flanking regions. The intron flanking regions partially interact with the internal guide sequence (IGS) of the intron and therefore not all mutations in the flanking regions will be allowed. Some of the boundaries have been elucidated previously (38). However, in order to predict the effect of the flanking regions accurately, the flanks need to be studied in more detail. The 5' flank and the IGS are part of P1, while the IGS can also base pair downstream to be part of P2 (Figure 3.1A). Changing the interactions in P1 will influence the pairing of the IGS. Weakening P1 will favour the IGS pairing in P2 and vice versa. Since the IGS is either part of the 3' end of P1 or the 5' end of P2, the position of the nucleotide that is changed will determine the impact on the splicing. In this study, we set out to investigate the tolerance of the ribozyme regarding variations in the 5' and 3' intron flanking regions. By using the *E. coli lacZ* gene as phenotypic reporter, we analysed how certain combinations of nucleotides in the flanking regions affect splicing. After establishing the rules for appropriate riboswitch insertion, as a proof of principle we inserted the theophylline responsive group I aptazyme into the RNAP gene. To allow for convenient functional analysis, we used a system in which the RNAP controls the expression of the fluorescent GFPuv.

Results and discussion

Analysis of the intron insertion site

The flanking regions of the group I introns are part of the coding sequence as well as of the ribozyme. When inserting the intron into another gene it is almost impossible to retain both the intron flanking regions and the CDS. Applying minor changes to the CDS with synonymous codons may create a site that resembles the wild type intron flanking regions. However, it is not clear to which extent the flanking region determines the splicing efficiency. To investigate the effect of the flanking regions of the T4 *td* intron on its splicing efficiency and on the expression of the target gene, we made a series of constructs containing the *lacZ* gene from *E. coli* with the intron in between amino acids D6 and S7. The *lacZ* gene was used because its functional expression can be easily monitored by a colorimetric assay and its high tolerance for modification of the 5' end.

The intron was flanked by 5 amino acids from phage T4 *td* upstream and downstream as well as a BspTI and PstI site on the 5' and 3' end respectively. Mutations were made by PCR in the 5' flank and 3' flank. To allow for further engineering of the riboswitch, the domain P6, which in the wild type phage T4 *td* intron contains the I-TevI homing endonuclease, is

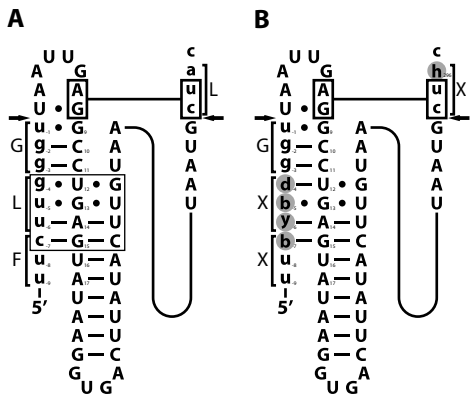


Figure 3.2. Detail of the phage T4 *td* intron. The arrows indicate the splicing site. The box indicates the region were the IGS may interact with both the P1 (left hairpin) and P2 (right hairpin). Nucleotides in capital letters are part of the intron and non-capital nucleotides are part of either the 5′ or 3′ exon. (A) The natural T4 *td* intron is inserted into *lacZ* between D6 and S7. (B) Nucleotides marked by a dark circle have been modified into "b" (G/U/C), "y" (C/U), "d" (G/A/U) or "h" (A/U/C). Positions -7 and 296 were changed in single nucleotide mutants only where the other positions conformed to the natural intron. Positions -4 to -6 were changed in all possible combinations of match, wobble and mismatch.

replaced by a hairpin containing two BtgZI sites for golden gate cloning and several other restriction endonuclease sites (Figure 3.1A).

The wild type interactions are shown in Figure 3.2A. All position numbers are relative to the 5′ splice site. Positions 9 to 11 do not show competition for pairing between P1 and P2. The same is true for positions 16 and 17. Positions 12 to 15 may either pair to form P1 or P2. It is expected that mutating these will yield the largest difference in splicing efficiency. The P1 suggested by Thompson et al. (36) does not show involvement of position -7 in base pairing. Since there is the possibility of base pairing, it is interesting to assay the effect of disallowing base pairing. If there is indeed no interaction between position -7 and the IGS, no difference in LacZ activity should be observed. Preferably all combinations are tested, but this will lead to exponential expansion of the number of mutants to test. To narrow down the possible combinations, position -7 was mutated to form either a mismatch or a wobble pair, while all other nucleotides remained wild type. The positions -6 to -4 were changed into all possible combinations of pair, wobble pair and mismatch. The suggested interaction between the intron and the 5′ end of exon 2 consists of only two base pairs. This length is not conserved within the group I introns and may affect the splicing as well. The Tetrahymena intron has as much as six base pairs. Position 296, which is the third nucleotide of the 3′ exon and not base pairing with the P1 loop, was mutated to form either a pair or a wobble pair with the P1 loop. Like position -7, the other nucleotides remained wild type. All mutants were assayed for β-galactosidase activity after overnight growth.

Position -7 and position 296 clearly show a decrease in LacZ activity as the pairing deviates more from the wild type situation (Figure 3.3A/B). Position -7 preferably pairs with position 15. A weaker interaction in the form of a wobble base pair does impede the intron splicing, but not as severely as having no interaction at all. The exact opposite is found for position 296 where a mismatch allows for the highest intron splicing activity. The weak wobble

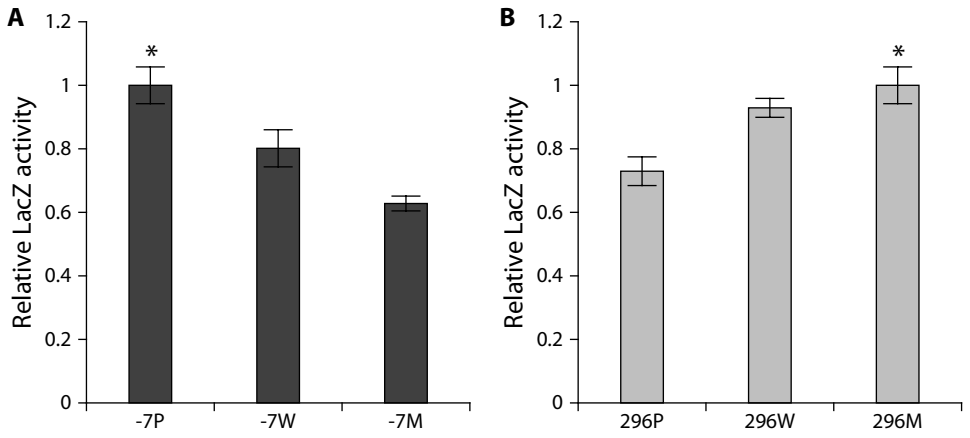


Figure 3.3. LacZ activity of position -7 mutants and the position 296 mutants. The asterisk indicates the wild-type intron. (A) Changes at the -7 position negatively affect the intron splicing. Both the wobble pair (-7W) and the mismatch (-7M) show a decreased activity compared to the wild type (-7P). (B) Stabilising the interactions between the 5′ end of the intron and the 5′ end of exon 2 does not aid the splicing as both the pair (296P) and the wobble pair (296W) exhibit a lower LacZ activity than the wild type (296M).

base pair impedes splicing to some extent, while the stronger pair decreases the splicing to a significantly larger extent. It shows that the positions -7 and 296 strongly influence the splicing efficiency by their interaction with the intron. For position -7 it appears that decreasing the interaction with the intron may decrease the efficiency by some 40%, while increasing the interaction of position 296 with the intron may decrease the efficiency by 25%.

Positions -4 to -6 were tested in all possible combinations of pair (P), mismatch (M) and wobble pair (W) if applicable (Figure 3.4). Unfortunately, there is no sequence for pair-mismatch-pair (PMP) that does not yield either the UAA or UGA stop codon. Selecting a different frame will yield a UAG stop codon for PWP and MWP or a stop codon in the 3′ flanking region. The wild type combination PWW mirrors the interactions of P2 with the IGS (Figure 3.2). Strengthening the interactions of P1 will lessen the interactions of P2. A mismatch of position -4 with the IGS will favour P2 instead of P1. Figure 3.4 shows that a -4 mismatch is preferred in almost all variants, except for those in which both -6 and -5 are mismatched too. A wobble base-pair at position -5 negates to a large extent the effect that -6 and -4 have on the splicing. In contrast, a pair or a mismatch at position -5 means that depending on -6 and -4 the splicing efficiency may be very high or very low. Position -6 in general appears in favour of being paired, however, strengthening the P1 to full extent (PPP) is detrimental for the splicing, almost inhibiting splicing completely. Completely mismatching positions -6 to -4 impedes the splicing significantly, but not to a very large extent. The cumulative effect of changes in the intron flanking regions remains

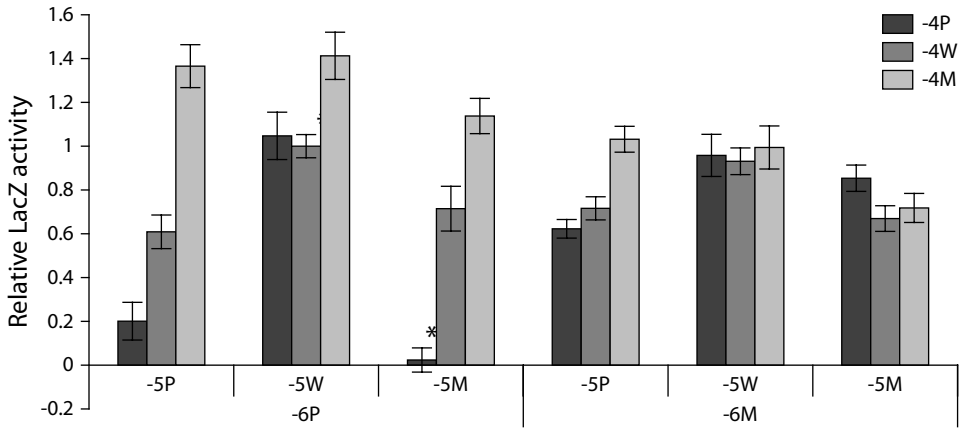


Figure 3.4. LacZ activity of all possible combinations for pair (P), wobble pair (W) and mismatch (M) at positions -6 to -4. PMP reads either UAA or UGA both being stop codons. The wild type intron (*) is PWW (UUG) and set to 1. All other LacZ activities are a fraction of the wild type activity.

currently unknown, but for many genes an insertion position with at least decent splicing can be found already. Retaining the wild type interactions of positions -1 to -3, -7 and 295 to 296 and possibly changing positions -4 to -6 will yield an active intron with a good splicing efficiency.

Intron functionality in RNAP

When the group I aptazyme is used as a biosensor, the signal needs to be quantifiable. An obvious target for inserting the theophylline-dependent intron would be GFP or another fluorescent protein. Compared to LacZ, GFP is even easier to measure and compatible with automated processes like fluorescence activated cell sorting (FACS). However, inserting the intron into a gene may impede the expression significantly (Supplementary figure 3.1). For enzymes this poses no problem, since the substrate conversion will take place in time. Low enzyme concentrations will cause the reaction to be slower, but they can still be measured by extending the reaction time. Fluorescent proteins differ in that respect. The signal is as strong as the number of fluorescent proteins present. To be able to visualise the induction using GFP, the effect of induction must be amplified. Otherwise the signal would be too low. Such amplification can be achieved by creating a cascade, which consists of an intron-containing RNAP and a GFP under control of the T7 promoter (Figure 3.5). The RNAP gene is under control of the rhamnose promoter. RNAP is a well characterised enzyme with a known promoter sequence, which is recognised by RNAP only. This ensures minimal background expression of GFP. Another advantage of using a cascade with RNAP is the possibility to control multiple genes with one intron riboswitch. Integrating RNAP on the genome eliminates the gene dose effect caused by variance in plasmid copy number.

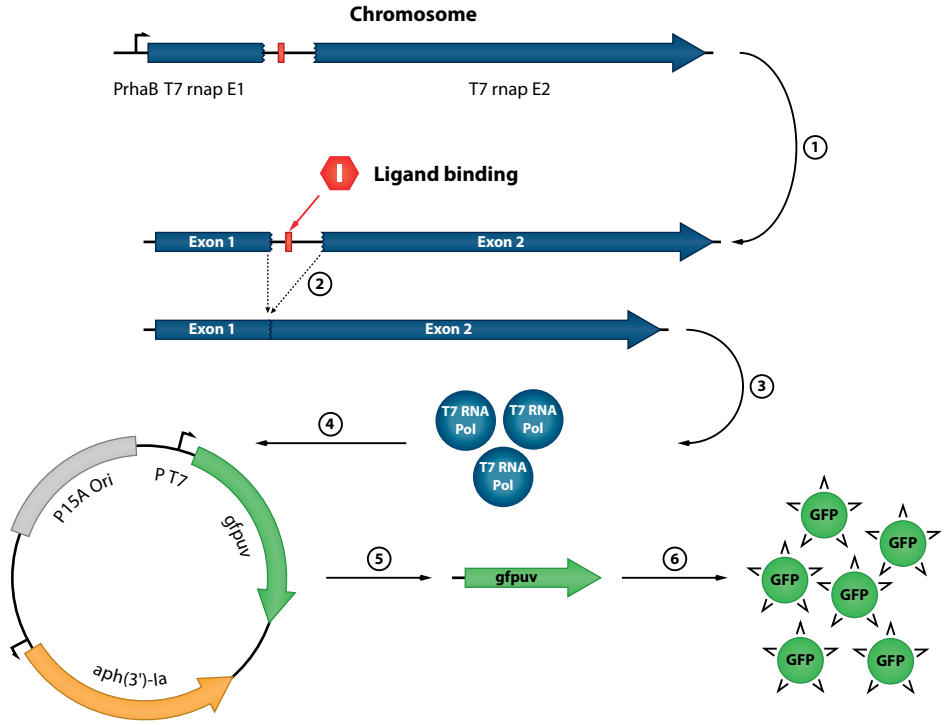


Figure 3.5. RNAP cascade to GFPuv. (1) RNAP is transcribed via induction by L-rhamnose. (2) Upon ligand binding the aptamer changes conformation and the intron can splice out of the mRNA, yielding a functional RNAP mRNA. (3) The mRNA is translated into RNAP, which in turn (4) binds to the T7 promoter and (5) transcribes the *gfpuv* gene. After the (6) translation, the fluorescence can be measured at 395 nm excitation and 508 nm emission.

A gene the size of RNAP has a high probability of containing a possible intron insertion site. Several insertion sites were identified using the python script based on the data obtained from the *lacZ* experiments. The RNAP gene was interrupted at three locations with a theophylline-dependent intron. The flanking regions were near identical for all three locations and the locations were spread along the gene to determine the effect of intron positioning.

The three different strains of *E. coli* DH10B with integrated RNAP were tested for their response to both L-rhamnose and theophylline. The results are depicted in Figure 3.6. The fluorescence shows a strong response to both L-rhamnose and theophylline. The fluorescence observed without induction does not significantly differ from the fluorescence observed for the GFPuv reporter plasmid alone. This implies that the dual control on the RNAP – transcription control by L-rhamnose and translation control by theophylline – succeeds to a large extent in keeping the RNAP inactive. The transcription of RNAP itself dictates the dynamic range in fluorescence caused by theophylline. This

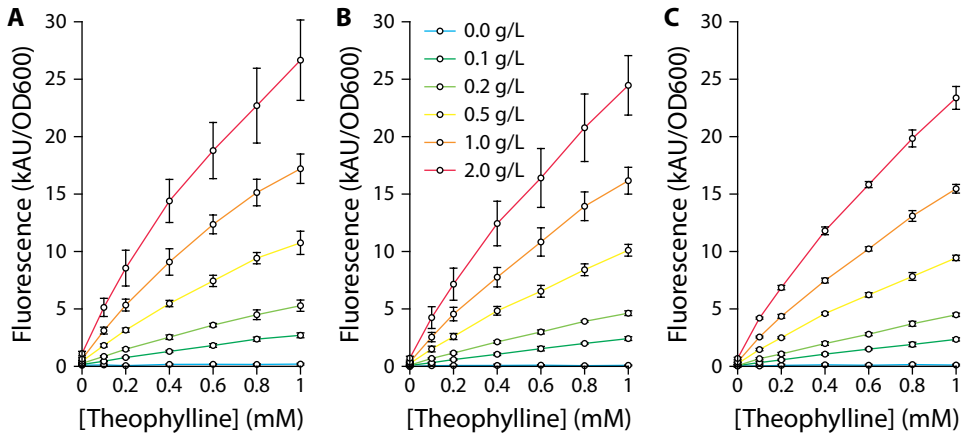


Figure 3.6. Response to theophylline and L-rhamnose of the integrated RNAP controlled GFPuv. The intron is inserted between G201 and L202 (A), G449 and L450 (B), and between G671 and L672 (C) based on the wild type amino acid sequence. Rhamnose concentration is varied from 0 g/L to 1 g/L. Error bars indicate the standard deviation of three independent biological replicates.

dynamic range can be adjusted according to the requirements of the application. At any L-rhamnose concentration, the fold-change caused by theophylline is around 15 times. The cascade being a multi-component system with two types of control makes it very hard to model the dependence on both L-rhamnose and theophylline.

The location of the aptazyme in the RNAP does appear to have an influence on the expression. Several factors may play a role. For instance, while the immediate surroundings of the intron are the same, the possible interactions further upstream or downstream are different, changing the splicing efficiency along with it. Also, the group I introns are known to require active translation to splice effectively and translation speed is not equal throughout the mRNA. Slow or fast translation speed locally can influence the intron splicing. Finally, the RNAP is native to an *E. coli* phage, but insertion of restriction endonuclease sites required a few silent mutations changing the codon bias. Although a difference in fluorescence of this magnitude is not readily expected from a few silent mutations, the possibility cannot be ignored.

Flow cytometry analysis of the bacterial cultures showed a low and high fluorescence population (Figure 3.7). Within the same population, the difference in fluorescence ranges about 10^2 fold. The control without any induction shows an induced population size of <0.1%. With transcription induction by L-rhamnose, increasing the theophylline concentration resulted in an increased size of the high fluorescence population, but it did not elevate its actual fluorescence intensity by much. This suggests that the populations represent bacteria that have no functional RNAP mRNA versus at least one functional RNAP mRNA. The chance of having at least one functional RNAP mRNA is then determined by the theophylline concentration. Induction works on individual mRNA molecules, so

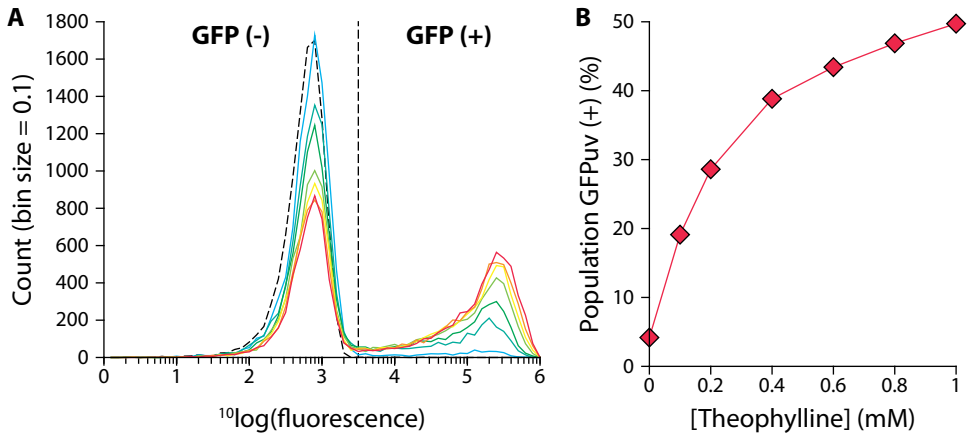


Figure 3.7. Flow cytometry analysis of different inductions with theophylline. (A) The fluorescence of DH10B with integrated RNAP. The theophylline dependent intron is inserted between G201 and L202 (Figure 3.6A). Cells are grouped by their fluorescence where Δ^{10}log(Fluorescence) = 0.1. The control has no induction with L-rhamnose or TP, while the others have 1 g/L L-rhamnose induction and different concentrations of TP. For each of the inducer concentrations, two populations can be distinguished. With increasing inducer concentration, the high fluorescent population increases in size and a little in fluorescence as well, while the low fluorescent population decreases in size. (B) The population size from (A) depends on theophylline approximately following first order binding kinetics with a basal level of 4% of the cells being induced at 1 g/L L-rhamnose.

increased induction must mean that bacteria are likely to harbour multiple copies of functional mRNA, resulting in even more copies of RNAP. Since the copy number of the reporter plasmid is rather low (10-12 copies per cell), it may well be saturated with a low concentration of RNAP, so further increase in RNAP might not have a large effect on the GFP production. Measuring the population size of the various cultures shows a first order binding kinetics response to theophylline. Since the cultures have been induced with L-rhamnose, there is a tiny population (about 4%) of GFPuv positive bacteria without theophylline, that increases with about a factor 12 upon full induction by theophylline.

For a final application as a biosensor, the transcription of RNAP may be fixed with a constitutive promoter, while the functional translation will be ligand dependent. The intron lowers the maximum translation quite severely, so not all reporter genes will show enough signal when put under control of a ligand-dependent intron directly. Enzymes like LacZ can handle the lower translation efficiency, but all genes that need a decent expression to function, like GFPuv, can now be put under control of one enzyme. Enriching for the ligand-induced bacteria can still be done with population-wise induction, since the induction makes it more likely to end up in the fraction expressing RNAP. If overexpression of the selection gene is toxic, selection can also be performed with direct interruption of an antibiotic resistance gene or auxotrophy complementing gene.

The RNAP cascade may see another application altogether. The intron controlled RNAP shows very low RNAP background, so it can be employed as a generic tool for well controlled gene expression. The double control serves as tight regulation already, but if necessary, the reporter gene c.q. gene of interest can have a control in addition e.g. lacI. The population-wise induction behaviour has consequences for the application. Gradually increasing the induction does not result in increased levels of protein in individual cells, so low expression per cell, does not work. The reverse is also true. Leakage in RNAP is not caused in all bacteria a little bit, but in a few bacteria a lot. Toxic proteins will cause harm in a minor population, but are not transcribed by RNAP in the vast majority.

Conclusions

For insertion of a tuneable splicing intron (aptazyme) into a gene-of-interest, the flanking regions of the insertion site of that intron are very important for splicing activity. The splicing rate for several flanking regions of the T4 *td* intron-based aptazyme was estimated, and found to depend heavily on the actual sequence and ranges from 20% to 140% of wild type splicing. Fixing some of the nucleotides in the intron finding algorithm and leaving positions -6 to -4 to be mutated, yields a functional though not always optimal silent intron insertion site. This method will facilitate the introduction of multiple introns into a gene, thereby increasing the stringency of the riboswitch. Most reasonably sized proteins will have the possibility of having an intron introduced through silent mutations, i.e. without changing the amino acid sequence. Changing the codon use is mostly inevitable.

The theophylline-dependent phage T4 *td* intron was successfully inserted into the gene encoding the T7 RNAP. This aptazyme enables high-level expression of GFPuv upon theophylline induction. Under transcription control by the rhamnose promoter and translation control by the group I intron, the system barely exhibits background expression. The induction with theophylline works by raising the number of bacteria that exhibit RNAP expression. While this behaviour results in more fluorescence for the culture as a whole, individual bacteria were not shown to have increased fluorescence upon further induction.

Materials and methods

Bacterial strains and media

E. coli DH10B T1R was purchased from Invitrogen (C6400-03) and used for plasmid propagation and standard molecular techniques. It was also used for expression of *lacZ* and as base strain for the RNAP knock-in strains. Transformation was performed with a ECM 63 electroporator (BTX) at 2500 V, 200 Ω and 25 µF, 2 mm cuvettes, 20-40 µL of electro-competent cells and recovery in LB.

Bacteria were generally grown at 37°C on LB medium (Miller) containing the appropriate antibiotics: kanamycin (50 mg/L), ampicillin (100 mg/L) and chloramphenicol (35 mg/L).

Construction of LacZ reporter plasmids

The *lacZ* reporter plasmid series were constructed from pEA001 (Supplementary sequence 3.1; Supplementary table 3.1; Supplementary table 3.2). The plasmid contains the *E. coli lacZ* gene under control of the lacUV5 promoter. Ten amino acids flanking the phage T4 *td* intron (five from each side) were introduced between D6 and S7 of LacZ, omitting the intron itself. For cloning purposes, the ten amino acids were in turn flanked by a BspTI and PstI restriction site. Generating the complete series was performed by PCR from pMA-ThyA-SI001, digestion with BspTI and PstI and ligation into pEA001. Negative control pEA001 [-1P] has base-pairing one nucleotide upstream of the 5′ splice site, rendering the intron dysfunctional.

β-galactosidase activity assay

Permeabilisation solution contains 100 mM Na2HPO4, 20 mM KCl, 2 mM MgSO4, 0.8 g/L CTAB, 0.4 g/L sodium deoxycholate and 5.4 mL/L β-mercaptoethanol. Substrate solution contains 60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 1 g/L o-nitrophenyl-β-D-galactopyranoside (ONPG) and 2.7 mL/L β-mercaptoethanol. Stop solution consists of 1 M Na_2CO_3.

LacZ activity was assayed in *E. coli* DH10B T1R in triplicate. After overnight growth at 37°C , 20 µL of culture was mixed with 80 µL of permeabilisation solution and incubated at 30°C for 30 min. 600 µL of pre-warmed substrate solution was added and incubated at 30°C until sufficient colour had developed. 700 µL of stop solution was added to quench the reaction. The reaction was filtered through a 0.2 µM filter and measured in a spectrophotometer at 420 nm in a 1 cm cuvette.

LacZ activity was calculated according to

$$LacZ = \frac{A_{420}}{t} \cdot \frac{Vtotal}{Vculture \cdot OD_{600}}$$

The LacZ activities of all clones were divided by the LacZ activity exhibited by the wild type intron.

Integration of GH_6-RNAP variants

Construction of pSC020

pKD46 (39) was digested with PdmI and NcoI and ligated into pJW168 (Lucigen) between SmaI and NcoI. The resulting plasmid contains the lambda red genes and cre recombinase.

RNAP/GFPuv cascade

The pRham-T7His-Lox (Supplementary sequence 3.2) uses plasmid pRham-CHis (Lucigen) as a base. It contains a 5' MGH_6 tagged T7 rnap located downstream of the pRham-CHis NdeI site. T7 rnap is followed by a cassette containing chloramphenicol acetyltransferase (cat) from pACYC184, which is under control of PtacI and flanked by BglII – Lox71 and Lox66 – NotI. The pMA-*thyA*-Theo (Supplementary sequence 3.3) was used as PCR template for insertion of the theophylline responsive intron with respective restriction sites. Three fragments were generated by PCR: T7 rnap exon1 [RE1], [RE1]-TP intron-[RE2] and [RE2]-T7 rnap exon2. The rnap fragments were generated using pAR1219 as template. Exon1 fragments used BG4467 as forward primer, while the exon2 fragments used BG4681 as reverse primer. For the Theo4/5/6 constructs, the exon1 reverse and exon2 forward primers were respectively BG4677 & BG4678 / BG4732 & BG4733 / BG4679 & BG4680. The Theo4/5/6 intron fragments were derived from pMA-*thyA*-Theo using primers BG4907 & BG4683 / BG4908 & BG4735 / BG4909 & BG4685 (Supplementary table 2). Fragments were ligated into pRham-T7His-Lox replacing the T7His. The intron positions are based on the wild type RNAP gene and are between G201 and L202 flanked by PscI and HindIII (pRham-T7His-Theo4-Lox), between G449 and L450 flanked by Bsu15I and XagI (pRham-T7His-Theo5-Lox) and between G671 and L672 flanked by Eco88I and PstI (pRham-T7His-Theo6-Lox). All have CAAGGGT as 5' intron flank instead of wild type CTTGGGT. The 3' intron flanks are CTAC, CTAC and CTAA respectively.

E. coli DH10B harbouring the pSC020 plasmid were made electro-competent while being induced with 10 mM L-arabinose. A PCR reaction was performed on EcoO109I linearized pRham-T7His-TheoX-Lox plasmids with BG4792 and BG4794. The PCR reaction was purified and transformed into *E. coli* DH10B (pSC020) and recovered for 2.5 h at 30°C. The bacteria were plated on LB agar plates containing 100 mg/L ampicillin and 35 mg/L chloramphenicol and incubated at 30°C overnight. Four LB cultures containing 1 mM IPTG were inoculated with DH10B-T7His-TheoX-Lox and incubated at 30°C for 7h to allow cre recombination. Then it was incubated at 37°C overnight to cure the plasmid pSC020. Cultures were streaked on LB plates and grown at 37°C overnight. Single colonies were tested by PCR for insertion of RNAP and removal of the antibiotics cassette as well as striped on LB plates containing 100 mg/L ampicillin to ensure plasmid loss.

The reporter plasmid pSC028-GFPuv-Term (Supplementary sequence 4) features the pACYC184 p15A origin of replication, kanamycin resistance derived from pET24d and GFPuv under control PT7 derived from pGFPuv. Downstream of GFPuv is the T7 terminator also derived from pET24d.

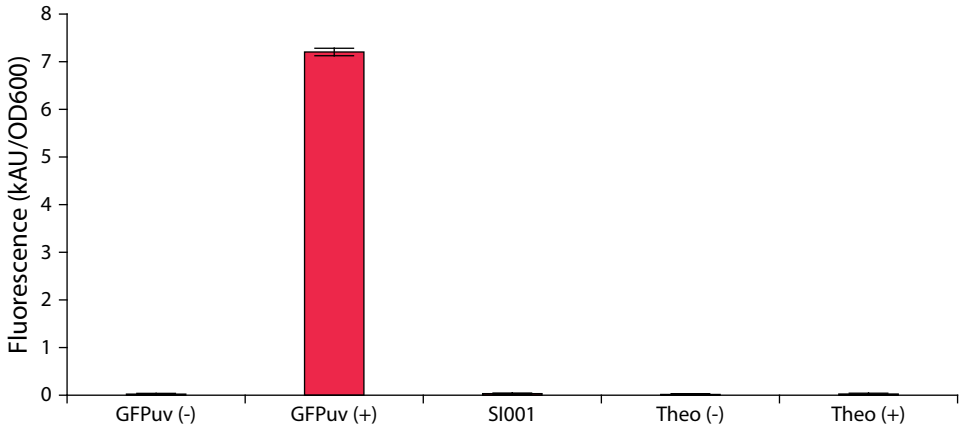
GFPuv fluorescence

E. coli DH10B-T7His-TheoX harbouring the pSC028-GFPuv-term plasmid was grown overnight at 37ºC in LB medium containing kanamycin (50 mg/L). A 96-well 2-mL culture plate (Greiner) was filled with a concentrate of theophylline and L-rhamnose. LB medium containing kanamycin and overnight grown bacteria were added so that the final concentration of kanamycin was 50 mg/L, the bacteria had a final dilution of 10^{-3} and the theophylline and L-rhamnose were diluted to 1x in 500 µL total volume. Culture plates were incubated at 37ºC overnight under continuous shaking. The bacteria were centrifuged for 10 min at 4800 g in a Sorval Legend centrifuge. The supernatant was cleared and the cell pellet was resuspended in 500 µL 50 mM Tris-HCl pH 7.5. After resuspension, the plates were incubated at 37ºC for 1 h to allow maturation of the GFPuv. 100 µL of suspension was pipetted into a 96-well black plate with clear bottom (Perkin Elmer) and measured with a Synergy MX plate reader. The cell density was measured by scattering at 600 nm and the fluorescence was measured at an excitation wavelength of 395 nm with a width of 20 nm and an emission wavelength of 508 nm with 20 nm width with a gain of 50. The background fluorescence and background scattering were subtracted and the fluorescence was divided by the scattering at 600 nm. The background fluorescence of bacteria without either GFPuv or GH6-RNAP was negligible, but the fluorescence caused by other components than GFPuv in the bacteria was still subtracted.

Flow cytometry

E. coli DH10B-T7His-Theo4 harbouring the pSC028-GFPuv-term plasmid was grown overnight at 37ºC in LB medium containing kanamycin (50 mg/L). 5 mL of LB supplemented with kanamycin, 1 g/L of L-rhamnose and theophylline (0-1 mM) was inoculated with 1 µL of the pre-culture and grown overnight at 37°C. Cultures were diluted 100x in 1x PBS with 10 mM EDTA and measured with a Life Technologies Attune NxT flow cytometer, using the blue laser. Based on the control with neither L-rhamnose nor theophylline, a non-fluorescent population was established.

SUPPLEMENTARY DATA



Supplementary Figure 3.1. Direct interruption of GFPuv with a T4 *td* intron variant. The T7 promoter of pSC028-GFPuv-Term is replaced by the PtacI promoter. GFPuv (-) has no GFP gene, GFPuv (+) has constitutive expression of GFPuv, SI001 has GFPuv interrupted by the SI001 intron, which does not requires no induction to splice, and Theo (-) and Theo (+) have GFPuv interrupted by the theophylline dependent intron without (-) or with (+) 1mM theophylline as inducer. None of the GFPuv genes interrupted by an intron has fluorescent signal above the autofluorescence of GFPuv (-)

Supplementary sequence 3.1. pEA001

TCTAGATTTC AGTGCAATTT ATCTCTTCAA ATGTAGCACC TGAAGTCAGC CCCATACGAT ATAAGTTGTA
ATTCGGTACC CCGCTTCGGC GGGGTTTTTT CAAGTTTACA CTTTATGCTT CCGGCTCGTA TAATGTGTGG
GGAGACCACA ACGGTTTCCC TCTAGAAATA ATTTTGTTTA ACTATAAGAA GGAGATATAC ATATGACCAT
GATTACGGAT CTTAAGGATG TTTTCTTGGG TCTACCGTTT AATATTCTGC AGTCACTGGC CGTCGTTTTA
CAACGTCGTG ACTGGGAAAA CCCTGGCGTT ACCCAACTTA ATCGCCTTGC AGCACATCCC CCTTTCGCCA
GCTGGCGTAA TAGCGAAGAG GCCCGCACCG ATCGCCCTTC CCAACAGTTG CGCAGCCTGA ATGGCGAATG
GCGCTTTGCC TGGTTTCCGG CACCAGAAGC GGTGCCGGAA AGCTGGCTGG AGTGCGATCT TCCTGAGGCC
GATACTGTCG TCGTCCCCTC AAACTGGCAG ATGCACGGTT ACGATGCGCC CATCTACACC AACGTGACCT
ATCCCATTAC GGTCAATCCG CCGTTTGTTC CCACGGAGAA TCCGACGGGT TGTTACTCGC TCACATTTAA
TGTTGATGAA AGCTGGCTAC AGGAAGGCCA GACGCGAATT ATTTTTGATG GCGTTAACTC GGCGTTTCAT
CTGTGGTGCA ACGGGCGCTG GGTCGGTTAC GGCCAGGACA GTCGTTTGCC GTCTGAATTT GACCTGAGCG
CATTTTTACG CGCCGGAGAA AACCGCCTCG CGGTGATGGT GCTGCGCTGG AGTGACGGCA GTTATCTGGA
AGATCAGGAT ATGTGGCGGA TGAGCGGCAT TTTCCGTGAC GTCTCGTTGC TGCATAAACC GACTACACAA
ATCAGCGATT TCCATGTTGC CACTCGCTTT AATGATGATT TCAGCCGCGC TGTACTGGAG GCTGAAGTTC
AGATGTGCGG CGAGTTGCGT GACTACCTAC GGGTAACAGT TTCTTTATGG CAGGGTGAAA CGCAGGTCGC
CAGCGGCACC GCGCCTTTCG GCGGTGAAAT TATCGATGAG CGTGGTGGTT ATGCCGATCG CGTCACACTA
CGTCTGAACG TCGAAAACCC GAAACTGTGG AGCGCCGAAA TCCCGAATCT CTATCGTGCG GTGGTTGAAC
TGCACACCGC CGACGGCACG CTGATTGAAG CAGAAGCCTG CGATGTCGGT TTCCGCGAGG TGCGGATTGA
AAATGGTCTG CTGCTGCTGA ACGGCAAGCC GTTGCTGATT CGAGGCGTTA ACCGTCACGA GCATCATCCT
CTGCATGGTC AGGTCATGGA TGAGCAGACG ATGGTGCAGG ATATCCTGCT GATGAAGCAG AACAACTTTA
ACGCCGTGCG CTGTTCGCAT TATCCGAACC ATCCGCTGTG GTACACGCTG TGCGACCGCT ACGGCCTGTA

TGTGGTGGAT GAAGCCAATA TTGAAACCCA CGGCATGGTG CCAATGAATC GTCTGACCGA TGATCCGCGC
TGGCTACCGG CGATGAGCGA ACGCGTAACG CGAATGGTGC AGCGCGATCG TAATCACCCG AGTGTGATCA
TCTGGTCGCT GGGGAATGAA TCAGGCCACG GCGCTAATCA CGACGCGCTG TATCGCTGGA TCAAATCTGT
CGATCCTTCC CGCCCGGTGC AGTATGAAGG CGGCGGAGCC GACACCACGG CCACCGATAT TATTTGCCCG
ATGTACGCGC GCGTGGATGA AGACCAGCCC TTCCCGGCTG TGCCGAAATG GTCCATCAAA AAATGGCTTT
CGCTACCTGG AGAGACGCGC CCGCTGATCC TTTGCGAATA CGCCCACGCG ATGGGTAACA GTCTTGGCGG
TTTCGCTAAA TACTGGCAGG CGTTTCGTCA GTATCCCCGT TTACAGGGCG GCTTCGTCTG GGACTGGGTG
GATCAGTCGC TGATTAAATA TGATGAAAAC GGCAACCCGT GGTCGGCTTA CGGCGGTGAT TTTGGCGATA
CGCCGAACGA TCGCCAGTTC TGTATGAACG GTCTGGTCTT TGCCGACCGC ACGCCGCATC CAGCGCTGAC
GGAAGCAAAA CACCAGCAGC AGTTTTTCCA GTTCCGTTTA TCCGGGCAAA CCATCGAAGT GACCAGCGAA
TACCTGTTCC GTCATAGCGA TAACGAGCTC CTGCACTGGA TGGTGGCGCT GGATGGTAAG CCGCTGGCAA
GCGGTGAAGT GCCTCTGGAT GTCGCTCCAC AAGGTAAACA GTTGATTGAA CTGCCTGAAC TACCGCAGCC
GGAGAGCGCC GGGCAACTCT GGCTCACAGT ACGCGTAGTG CAACCGAACG CGACCGCATG GTCAGAAGCC
GGGCACATCA GCGCCTGGCA GCAGTGGCGT CTGGCGGAAA ACCTCAGTGT GACGCTCCCC GCCGCGTCCC
ACGCCATCCC GCATCTGACC ACCAGCGAAA TGGATTTTTG CATCGAGCTG GGTAATAAGC GTTGGCAATT
TAACCGCCAG TCAGGCTTTC TTTCACAGAT GTGGATTGGC GATAAAAAAC AACTGCTGAC GCCGCTGCGC
GATCAGTTCA CCCGTGCACC GCTGGATAAC GACATTGGCG TAAGTGAAGC GACCCGCATT GACCCTAACG
CCTGGGTCGA ACGCTGGAAG GCGGCGGGCC ATTACCAGGC CGAAGCAGCG TTGTTGCAGT GCACGGCAGA
TACACTTGCT GATGCGGTGC TGATTACGAC CGCTCACGCG TGGCAGCATC AGGGGAAAAC CTTATTTATC
AGCCGGAAAA CCTACCGGAT TGATGGTAGT GGTCAAATGG CGATTACCGT TGATGTTGAA GTGGCGAGCG
ATACACCGCA TCCGGCGCGG ATTGGCCTGA ACTGCCAGCT GGCGCAGGTA GCAGAGCGGG TAAACTGGCT
CGGATTAGGG CCGCAAGAAA ACTATCCCGA CCGCCTTACT GCCGCCTGTT TTGACCGCTG GGATCTGCCA
TTGTCAGACA TGTATACCCC GTACGTCTTC CCGAGCGAAA ACGGTCTGCG CTGCGGGACG CGCGAATTGA
ATTATGGCCC ACACCAGTGG CGCGGCGACT TCCAGTTCAA CATCAGCCGC TACAGTCAAC AGCAACTGAT
GGAAACCAGC CATCGCCATC TGCTGCACGC GGAAGAAGGC ACATGGCTGA ATATCGACGG TTTCCACATG
GGGATTGGTG GCGACGACTC CTGGAGCCCG TCAGTATCGG CGGAATTCCA GCTGAGCGCC GGTCGCTACC
ATTACCAGTT GGTCTGGTGT CAAAAATAAA CTAGTGCAAG TGGCACTTTT CGGGGAAATG TGCGCGGAAC
CCCTATTTGT TTATTTTTCT AAATACATTC AAATATGTAT CCGCTCATGA ATTAATTCTT AGAAAAACTC
ATCGAGCATC AAATGAAACT GCAATTTATT CATATCAGGA TTATCAATAC CATATTTTTG AAAAAGCCGT
TTCTGTAATG AAGGAGAAAA CTCACCGAGG CAGTTCCATA GGATGGCAAG ATCCTGGTAT CGGTCTGCGA
TTCCGACTCG TCCAACATCA ATACAACCTA TTAATTTCCC CTCGTCAAAA ATAAGGTTAT CAAGTGAGAA
ATCACCATGA GTGACGACTG AATCCGGTGA GAATGGCAAA AGTTTATGCA TTTCTTTCCA GACTTGTTCA
ACAGGCCAGC CATTACGCTC GTCATCAAAA TCACTCGCAT CAACCAAACC GTTATTCATT CGTGATTGCG
CCTGAGCGAG ACGAAATACG CGGTCGCTGT TAAAAGGACA ATTACAAACA GGAATCGAAT GCAACCGGCG
CAGGAACACT GCCAGCGCAT CAACAATATT TTCACCTGAA TCAGGATATT CTTCTAATAC CTGGAATGCT
GTTTTCCCGG GGATCGCAGT GGTGAGTAAC CATGCATCAT CAGGAGTACG GATAAAATGC TTGATGGTCG
GAAGAGGCAT AAATTCCGTC AGCCAGTTTA GTCTGACCAT CTCATCTGTA ACATCATTGG CAACGCTACC
TTTGCCATGT TTCAGAAACA ACTCTGGCGC ATCGGGCTTC CCATACAATC GATAGATTGT CGCACCTGAT
TGCCCGACAT TATCGCGAGC CCATTTATAC CCATATAAAT CAGCATCCAT GTTGGAATTT AATCGCGGCC
TAGAGCAAGA CGTTTCCCGT TGAATATGGC TCATACTCTT CCTTTTTCAA TATTATTGAA GCATTTATCA
GGGTTATTGT CTCATGAGCG GATACATATT TGAATGTATT TAGAAAAATA AACAAATAGG CTGTCCCTCC
TGTTCAGCTA CTGACGGGGT GGTGCGTAAC GGCAAAAGCA CCGCCGGACA TCAGCGCTAG CGGAGTGTAT
ACTGGCTTAC TATGTTGGCA CTGATGAGGG TGTCAGTGAA GTGCTTCATG TGGCAGGAGA AAAAAGGCTG
CACCGGTGCG TCAGCAGAAT ATGTGATACA GGATATATTC CGCTTCCTCG CTCACTGACT CGCTACGCTC
GGTCGTTCGA CTGCGGCGAG CGGAAATGGC TTACGAACGG GGCGGAGATT TCCTGGAAGA TGCCAGGAAG
ATACTTAACA GGGAAGTGAG AGGGCCGCGG CAAAGCCGTT TTTCCATAGG CTCCGCCCCC CTGACAAGCA

TCACGAAATC TGACGCTCAA ATCAGTGGTG GCGAAACCCG ACAGGACTAT AAAGATACCA GGCGTTTCCC
CCTGGCGGCT CCCTCGTGCG CTCTCCTGTT CCTGCCTTTC GGTTTACCGG TGTCATTCCG CTGTTATGGC
CGCGTTTGTC TCATTCCACG CCTGACACTC AGTTCCGGGT AGGCAGTTCG CTCCAAGCTG GACTGTATGC
ACGAACCCCC CGTTCAGTCC GACCGCTGCG CCTTATCCGG TAACTATCGT CTTGAGTCCA ACCCGGAAAG
ACATGCAAAA GCACCACTGG CAGCAGCCAC TGGTAATTGA TTTAGAGGAG TTAGTCTTGA AGTCATGCGC
CGGTTAAGGC TAAACTGAAA GGACAAGTTT TGGTGACTGC GCTCCTCCAA GCCAGTTACC TCGGTTCAAA
GAGTTGGTAG CTCAGAGAAC CTTCGAAAAA CCGCCCTGCA AGGCGGTTTT TTCGTTTTCA GAGCAAGAGA
TTACGCGCAG ACCAAAACGA TCTCAAGAAG ATCATCTTAT TAATCAGATA AAATATT

Supplementary sequence 3.2. pRham-T7His-Lox

CACCACAATT CAGCAAATTG TGAACATCAT CACGTTCATC TTTCCCTGGT TGCCAATGGC CCATTTTCCT
GTCAGTAACG AGAAGGTCGC GAATTCAGGC GCTTTTTAGA CTGGTCGTAG GGAGACCACA ACGGTTTCCC
TCTAGAAATA ATTTTGTTTA ACTATAAGAA GGAGATATAC ATATGGGTCA TCACCATCAC CATCACAACA
CGATTAACAT CGCTAAGAAC GACTTCTCTG ACATCGAACT GGCTGCTATC CCGTTCAACA CTCTGGCTGA
CCATTACGGT GAGCGTTTAG CTCGCGAACA GTTGGCCCTT GAGCATGAGT CTTACGAGAT GGGTGAAGCA
CGCTTCCGCA AGATGTTTGA GCGTCAACTT AAAGCTGGTG AGGTTGCGGA TAACGCTGCC GCCAAGCCTC
TCATCACTAC CCTACTCCCT AAGATGATTG CACGCATCAA CGACTGGTTT GAGGAAGTGA AAGCTAAGCG
CGGCAAGCGC CCGACAGCCT TCCAGTTCCT GCAAGAAATC AAGCCGGAAG CCGTAGCGTA CATCACCATT
AAGACCACTC TGGCTTGCCT AACCAGTGCT GACAATACAA CCGTTCAGGC TGTAGCAAGC GCAATCGGTC
GGGCCATTGA GGACGAGGCT CGCTTCGGTC GTATCCGTGA CCTTGAAGCT AAGCACTTCA AGAAAAACGT
TGAGGAACAA CTCAACAAGC GCGTAGGGCA CGTCTACAAG AAAGCATTTA TGCAAGTTGT CGAGGCTGAC
ATGCTCTCTA AGGGTCTACT CGGTGGCGAG GCGTGGTCTT CGTGGCATAA GGAAGACTCT ATTCATGTAG
GAGTACGCTG CATCGAGATG CTCATTGAGT CAACCGGAAT GGTTAGCTTA CACCGCCAAA ATGCTGGCGT
AGTAGGTCAA GACTCTGAGA CTATCGAACT CGCACCTGAA TACGCTGAGG CTATCGCAAC CCGTGCAGGT
GCGCTGGCTG GCATCTCTCC GATGTTCCAA CCTTGCGTAG TTCCTCCTAA GCCGTGGACT GGCATTACTG
GTGGTGGCTA TTGGGCTAAC GGTCGTCGTC CTCTGGCGCT GGTGCGTACT CACAGTAAGA AAGCACTGAT
GCGCTACGAA GACGTTTACA TGCCTGAGGT GTACAAAGCG ATTAACATTG CGCAAAACAC CGCATGGAAA
ATCAACAAGA AAGTCCTAGC GGTCGCCAAC GTAATCACCA AGTGGAAGCA TTGTCCGGTC GAGGACATCC
CTGCGATTGA GCGTGAAGAA CTCCCGATGA AACCGGAAGA CATCGACATG AATCCTGAGG CTCTCACCGC
GTGGAAACGT GCTGCCGCTG CTGTGTACCG CAAGGACAAG GCTCGCAAGT CTCGCCGTAT CAGCCTTGAG
TTCATGCTTG AGCAAGCCAA TAAGTTTGCT AACCATAAGG CCATCTGGTT CCCTTACAAC ATGGACTGGC
GCGGTCGTGT TTACGCTGTG TCAATGTTCA ACCCGCAAGG TAACGATATG ACCAAAGGAC TGCTTACGCT
GGCGAAAGGT AAACCAATCG GTAAGGAAGG TTACTACTGG CTGAAAATCC ACGGTGCAAA CTGTGCGGGT
GTCGATAAGG TTCCGTTCCC TGAGCGCATC AAGTTCATTG AGGAAAACCA CGAGAACATC ATGGCTTGCG
CTAAGTCTCC ACTGGAGAAC ACTTGGTGGG CTGAGCAAGA TTCTCCGTTC TGCTTCCTTG CGTTCTGCTT
TGAGTACGCT GGGGTACAGC ACCACGGCCT GAGCTATAAC TGCTCCCTTC CGCTGGCGTT TGACGGGTCT
TGCTCTGGCA TCCAGCACTT CTCCGCGATG CTCCGAGATG AGGTAGGTGG TCGCGCGGTT AACTTGCTTC
CTAGTGAAAC CGTTCAGGAC ATCTACGGGA TTGTTGCTAA GAAAGTCAAC GAGATTCTAC AAGCAGACGC
AATCAATGGG ACCGATAACG AAGTAGTTAC CGTGACCGAT GAGAACACTG GTGAAATCTC TGAGAAAGTC
AAGCTGGGCA CTAAGGCACT GGCTGGTCAA TGGCTGGCTT ACGGTGTTAC TCGCAGTGTG ACTAAGCGTT
CAGTCATGAC GCTGGCTTAC GGGTCCAAAG AGTTCGGCTT CCGTCAACAA GTGCTGGAAG ATACCATTCA
GCCAGCTATT GATTCCGGCA AGGGTCTGAT GTTCACTCAG CCGAATCAGG CTGCTGGATA CATGGCTAAG
CTGATTTGGG AATCTGTGAG CGTGACGGTG GTAGCTGCGG TTGAAGCAAT GAACTGGCTT AAGTCTGCTG
CTAAGCTGCT GGCTGCTGAG GTCAAAGATA AGAAGACTGG AGAGATTCTT CGCAAGCGTT GCGCTGTGCA
TTGGGTAACT CCTGATGGTT TCCCTGTGTG GCAGGAATAC AAGAAGCCTA TTCAGACGCG CTTGAACCTG
ATGTTCCTCG GTCAGTTCCG CTTACAGCCT ACCATTAACA CCAACAAAGA TAGCGAGATT GATGCACACA

AACAGGAGTC TGGTATCGCT CCTAACTTTG TACACAGCCA AGACGGTAGC CACCTTCGTA AGACTGTAGT
GTGGGCACAC GAGAAGTACG GAATCGAATC TTTTGCACTG ATTCACGACT CCTTCGGTAC GATTCCGGCT
GACGCTGCGA ACCTGTTCAA AGCAGTGCGC GAAACTATGG TTGACACTTA TGAGTCTTGT GATGTACTGG
CTGATTTCTA CGACCAGTTC GCTGACCAGT TGCACGAGTC TCAATTGGAC AAAATGCCAG CACTTCCGGC
TAAAGGTAAC TTGAACCTCC GTGACATCTT AGAGTCGGAC TTCGCGTTCG CGTAAAGATC TTACCGTTCG
TATAATGTAT GCTATACGAA GTTATGAGCT GTTGACAATT AATCATCGGC TCGTATAATG TGTGGGCAAT
GAGCTTGCAC TGCAGAACTT TCTCGAGGAT ATACCATGGA GAAAAAAATC ACTGGATATA CCACCGTTGA
TATATCCCAA TGGCATCGTA AAGAACATTT TGAGGCATTT CAGTCAGTTG CTCAATGTAC CTATAACCAG
ACCGTTCAGC TGGATATTAC GGCCTTTTTA AAGACCGTAA AGAAAAATAA GCACAAGTTT TATCCGGCCT
TTATTCACAT TCTTGCCCGC CTGATGAATG CTCATCCGGA ATTCCGTATG GCAATGAAAG ACGGTGAGCT
GGTGATATGG GATAGTGTTC ACCCTTGTTA CACCGTTTTC CATGAGCAAA CTGAAACGTT TTCATCGCTC
TGGAGTGAAT ACCACGACGA TTTCCGGCAG TTTCTACACA TATATTCGCA AGATGTGGCG TGTTACGGTG
AAAACCTGGC CTATTTCCCT AAAGGGTTTA TTGAGAATAT GTTTTTCGTC TCAGCCAATC CCTGGGTGAG
TTTCACCAGT TTTGATTTAA ACGTGGCCAA TATGGACAAC TTCTTCGCCC CCGTTTTCAC TATGGGCAAA
TATTATACGC AAGGCGACAA GGTGCTGATG CCGCTGGCGA TTCAGGTTCA TCATGCCGTT TGTGATGGCT
TCCATGTCGG CAGAATGCTT AATGAATTAC AACAGTACTG CGATGAGTGG CAGGGCGGGG CGTAAATAAC
TTCGTATAAT GTATGCTATA CGAACGGTAG CGGCCGCCAC CGCTGAGCAA TAACTAGCAT AACCCCTTGG
GGCCTCTAAA CGGGTCTTGA GGGGTTTTTT GCTGAAAGGA GGAACTATAT CCGGGTAACG AATTCAAGCT
TGATATCATT CAGGACGAGC CTCAGACTCC AGCGTAACTG GACTGCAATC AACTCACTGG CTCACCTTCA
CGGGTGGGCC TTTCTTCGGT AGAAAATCAA AGGATCTTCT TGAGATCCTT TTTTTCTGCG CGTAATCTGC
TGCTTGCAAA CAAAAAAACC ACCGCTACCA GCGGTGGTTT GTTTGCCGGA TCAAGAGCTA CCAACTCTTT
TTCCGAGGTA ACTGGCTTCA GCAGAGCGCA GATACCAAAT ACTGTTCTTC TAGTGTAGCC GTAGTTAGGC
CACCACTTCA AGAACTCTGT AGCACCGCCT ACATACCTCG CTCTGCTAAT CCTGTTACCA GTGGCTGCTG
CCAGTGGCGA TAAGTCGTGT CTTACCGGGT TGGACTCAAG ACGATAGTTA CCGGATAAGG CGCAGCGGTC
GGGCTGAACG GGGGGTTCGT GCACACAGCC CAGCTTGGAG CGAACGACCT ACACCGAACT GAGATACCTA
CAGCGTGAGC TATGAGAAAG CGCCACGCTT CCCGAAGGGA GAAAGGCGGA CAGGTATCCG GTAAGCGGCA
GGGTCGGAAC AGGAGAGCGC ACGAGGGAGC TTCCAGGGGG AAACGCCTGG TATCTTTATA GTCCTGTCGG
GTTTCGCCAC CTCTGACTTG AGCATCGATT TTTGTGATGC TCGTCAGGGG GGCGGAGCCT ATGGAAAAAC
GCCAGCAACG CAGAAAGGCC CACCCGAAGG TGAGCCAGGT GATTACATTT GGGCCCTCAT CAGAGGTTTT
CACCGTCATC ACCGAAACGC GCGAGGCAGC TGCGGTAAAG CTCATCAGCG TGGTCGTGAA GCGATTCACA
GATGTCTGCC TGTTCATCCG CGTCCAGCTC GTTGAGTTTC TCCAGAAGCG TTAATGTCTG GCTTCTGATA
AAGCGGGCCA TGTTAAGGGC GGTTTTTTCC TGTTTGGTCA TTTAGAAAAA CTCATCGAGC ATCAAGTGAA
ACTGCAATTT ATTCATATCA GGATTATCAA TACCATATTT TTGAAAAAGC CGTTTCTGTA ATGAAGGAGA
AAACTCACCG AGGCAGTTCC ATAGGATGGC AAGATCCTGG TATCGGTCTG CGATTCCGAC TCGTCCAACA
TCAATACAAC CTATTAATTT CCCCTCGTCA AAAATAAGGT TATCAAGTGA GAAATCACCA TGAGTGACGA
CTGAATCCGG TGAGAATGGC AAAAGCTTAT GCATTTCTTT CCAGACTTGT TCAACAGGCC AGCCATTACG
CTCGTCATCA AAATCACTCG CACCAACCAA ACCGTTATTC ATTCGTGATT GCGCCTGAGC GAGACGAAAT
ACGCGATCGC CGTTAAAAGG ACAATTACAA ACAGGAATCG AATGCAACCG GCGCAGGAAC ACTGCCAGCG
CATCAACAAT ATTTTCACCT GAATCAGGAT ATTCTTCTAA TACCTGGAAT GCTGTTTTCC CTGGGATCGC
AGTGGTGAGT AACCATGCAT CATCAGGAGT ACGGATAAAA TGCTTGATGG TCGGAAGAGG CATAAATTCC
GTCAGCCAGT TTAGCCTGAC CATCTCATCT GTAACATCAT TGGCAACGCT ACCTTTGCCA TGTTTCAGAA
ACAACTCTGG CGCATCGGGC TTCCCATACA ATCGATAGAT TGTCGCACCT GATTGCCCGA CATTATCGCG
AGCCCATTTA TACCCATATA AATCAGCATC CATGTTGGAA TTTAATCGCG GCCTCGAGCA AGACGTTTCC
CGTTGAATAT GGCTCATAGC TCCTGAAAAT CTCGATAACT CAAAAAATAC GCCCGGTAGT GATCTTATTT
CATTATGGTG AAAGTTGGAA CCTCTTACGT GCCGATCAAG TCAAAAGCCT CCGGTCGGAG GCTTTTGACT
TTCTGCTATG GAGGTCAGGT ATGATTTAAA TGGTCAGTAT TGAGCGATAT CTAGAGAATT CGTC

Supplementary sequence 3.3. pMA-ThyA-Theo (relevant part)

GGCCGTCAAG GCCACGTGTC TTGTCCAGAG CTCGGGTTAA TTGAGGCCTG AGTATAAGGT GACTTATACT
TGTAATCTAT CTAAACGGGG AACCTCTCTA GTAGACAATC CCGTGCTAAA TTGATACCAG CATCGTCTTG
ATGCCCTTGG CAGCATAAAT GCCTAACGAC TATCCCTTTG GGGAGTAGGG TCAAGTGACT CGAAACGATA
GACAACTTGC TTTAACAAGT TGGAGATATA GTCTGCTCTG CATGGTGACA TGCAGCTGGA TATAATTCCG
GGGTAAGATT AACGACCTTA TCTGAACATA ATGCTACCGT TTAATATTGC CAGCTACGCG TGGTACCTGG
AGCACAAGAC TGGCCTCATG GGCC

Supplementary sequence 3.4. pSC028-GFPuv-Term

TCTAGATTTC AGTGCAATTT ATCTCTTCAA ATGTAGCACC TGAAGTCAGC CCCATACGAT ATAAGTTGTA
ATTCGGTACC CCGCTTCGGC GGGGTTTTTT CAAGTAATAC GACTCACTAT AGGGAGACCA CAACGGTTTC
CCTCTAGAAA TAATTTTGTT TAACTATAAG AAGGAGATAT ACATATGAGT AAAGGAGAAG AACTTTTCAC
TGGAGTTGTC CCAATTCTTG TTGAATTAGA TGGTGATGTT AATGGGCACA AATTTTCTGT CAGTGGAGAG
GGTGAAGGTG ATGCAACATA CGGAAAACTT ACCCTTAAAT TTATTTGCAC TACTGGAAAA CTACCTGTTC
CATGGCCAAC ACTTGTCACT ACTTTCTCTT ATGGTGTTCA ATGCTTTTCC CGTTATCCGG ATCACATGAA
ACGGCATGAC TTTTTCAAGA GTGCCATGCC CGAAGGTTAT GTACAGGAAC GCACTATATC TTTCAAAGAT
GACGGGAACT ACAAGACGCG TGCTGAAGTC AAGTTTGAAG GTGATACCCT TGTTAATCGT ATCGAGTTAA
AAGGTATTGA TTTTAAAGAA GATGGAAACA TTCTCGGACA CAAACTGGAG TACAACTATA ACTCACACAA
TGTATACATC ACGGCAGACA AACAAAAGAA TGGAATCAAA GCTAACTTCA AAATTCGCCA CAACATTGAA
GATGGATCCG TTCAACTAGC AGACCATTAT CAACAAAATA CTCCAATTGG CGATGGCCCT GTCCTTTTAC
CAGACAACCA TTACCTGTCG ACACAATCTG CCCTTTCGAA AGATCCCAAC GAAAAGCGTG ACCACATGGT
CCTTCTTGAG TTTGTAACTG CTGCTGGGAT TACACATGGC ATGGATGAGC TCTACAAATA AACTAGTGAT
CCGGCTGCTA ACAAAGCCCG AAAGGAAGCT GAGTTGGCTG CTGCCACCGC TGAGCAATAA CTAGCATAAC
CCCTTGGGGC CTCTAAACGG GTCTTGAGGG GTTTTTTGCT GAAAGGAGGA ACTATATCTA GTGCAAGTGG
CACTTTTCGG GGAAATGTGC GCGGAACCCC TATTTGTTTA TTTTTCTAAA TACATTCAAA TATGTATCCG
CTCATGAATT AATTCTTAGA AAAACTCATC GAGCATCAAA TGAAACTGCA ATTTATTCAT ATCAGGATTA
TCAATACCAT ATTTTTGAAA AAGCCGTTTC TGTAATGAAG GAGAAAACTC ACCGAGGCAG TTCCATAGGA
TGGCAAGATC CTGGTATCGG TCTGCGATTC CGACTCGTCC AACATCAATA CAACCTATTA ATTTCCCCTC
GTCAAAAATA AGGTTATCAA GTGAGAAATC ACCATGAGTG ACGACTGAAT CCGGTGAGAA TGGCAAAAGT
TTATGCATTT CTTTCCAGAC TTGTTCAACA GGCCAGCCAT TACGCTCGTC ATCAAAATCA CTCGCATCAA
CCAAACCGTT ATTCATTCGT GATTGCGCCT GAGCGAGACG AAATACGCGG TCGCTGTTAA AAGGACAATT
ACAAACAGGA ATCGAATGCA ACCGGCGCAG GAACACTGCC AGCGCATCAA CAATATTTTC ACCTGAATCA
GGATATTCTT CTAATACCTG GAATGCTGTT TTCCCGGGGA TCGCAGTGGT GAGTAACCAT GCATCATCAG
GAGTACGGAT AAAATGCTTG ATGGTCGGAA GAGGCATAAA TTCCGTCAGC CAGTTTAGTC TGACCATCTC
ATCTGTAACA TCATTGGCAA CGCTACCTTT GCCATGTTTC AGAAACAACT CTGGCGCATC GGGCTTCCCA
TACAATCGAT AGATTGTCGC ACCTGATTGC CCGACATTAT CGCGAGCCCA TTTATACCCA TATAAATCAG
CATCCATGTT GGAATTTAAT CGCGGCCTAG AGCAAGACGT TTCCCGTTGA ATATGGCTCA TACTCTTCCT
TTTTCAATAT TATTGAAGCA TTTATCAGGG TTATTGTCTC ATGAGCGGAT ACATATTTGA ATGTATTTAG
AAAAATAAAC AAATAGGCTG TCCCTCCTGT TCAGCTACTG ACGGGGTGGT GCGTAACGGC AAAAGCACCG
CCGGACATCA GCGCTAGCGG AGTGTATACT GGCTTACTAT GTTGGCACTG ATGAGGGTGT CAGTGAAGTG
CTTCATGTGG CAGGAGAAAA AAGGCTGCAC CGGTGCGTCA GCAGAATATG TGATACAGGA TATATTCCGC
TTCCTCGCTC ACTGACTCGC TACGCTCGGT CGTTCGACTG CGGCGAGCGG AAATGGCTTA CGAACGGGGC
GGAGATTTCC TGGAAGATGC CAGGAAGATA CTTAACAGGG AAGTGAGAGG GCCGCGGCAA AGCCGTTTTT
CCATAGGCTC CGCCCCCCTG ACAAGCATCA CGAAATCTGA CGCTCAAATC AGTGGTGGCG AAACCCGACA
GGACTATAAA GATACCAGGC GTTTCCCCCT GGCGGCTCCC TCGTGCGCTC TCCTGTTCCT GCCTTTCGGT
TTACCGGTGT CATTCCGCTG TTATGGCCGC GTTTGTCTCA TTCCACGCCT GACACTCAGT TCCGGGTAGG

CAGTTCGCTC CAAGCTGGAC TGTATGCACG AACCCCCCGT TCAGTCCGAC CGCTGCGCCT TATCCGGTAA CTATCGTCTT GAGTCCAACC CGGAAAGACA TGCAAAAGCA CCACTGGCAG CAGCCACTGG TAATTGATTT AGAGGAGTTA GTCTTGAAGT CATGCGCCGG TTAAGGCTAA ACTGAAAGGA CAAGTTTTGG TGACTGCGCT CCTCCAAGCC AGTTACCTCG GTTCAAAGAG TTGGTAGCTC AGAGAACCTT CGAAAAACCG CCCTGCAAGG CGGTTTTTTC GTTTTCAGAG CAAGAGATTA CGCGCAGACC AAAACGATCT CAAGAAGATC ATCTTATTAA TCAGATAAAA TATT

Supplementary table 3.1. T4 *td* intron variants in *lacZ*. Positions are modified to form a pair (P), wobble pair (W) or mismatch (M). pEA001 [WT] is the same as pEA001 [PWW], pEA001 [-7P] and pEA001 [296M].

Name	-7	-6	-5	-4	Forward	Reverse	Comment
pEA001 [WT]	P	W	W	M	BG5038	BG5040	
pEA001 ΔI							Intron partially removed
pEA001 [-7W]	P	W	W	M	BG5304	BG5040	
pEA001 [-7M]	P	W	W	M	BG5039	BG5040	
pEA001 [PPP]	P	P	P	M	BG5206	BG5040	
pEA001 [PPW]	P	P	W	M	BG5207	BG5040	
pEA001 [PPM]	P	P	M	M	BG5208	BG5040	
pEA001 [PWP]	P	W	P	M	BG5209	BG5040	
pEA001 [PWW]	P	W	W	M	BG5210	BG5040	
pEA001 [PWM]	P	W	M	M	BG5211	BG5040	
pEA001 [PMP]	P	M	P	M	BG5212	BG5040	
pEA001 [PMW]	P	M	W	M	BG5213	BG5040	
pEA001 [PMM]	P	M	M	M	BG5214	BG5040	
pEA001 [MPP]	M	P	P	M	BG5215	BG5040	
pEA001 [MPW]	M	P	W	M	BG5216	BG5040	
pEA001 [MPM]	M	P	M	M	BG5217	BG5040	
pEA001 [MPW]	M	W	P	M	BG5218	BG5040	
pEA001 [MWW]	M	W	W	M	BG5219	BG5040	
pEA001 [MWM]	M	W	M	M	BG5220	BG5040	
pEA001 [MMP]	M	M	P	M	BG5221	BG5040	
pEA001 [MMW]	M	M	W	M	BG5222	BG5040	
pEA001 [MMM]	M	M	M	M	BG5223	BG5040	
pEA001 [296P]	P	W	W	P	BG5210	BG5305	
pEA001 [296W]	P	W	W	W	BG5210	BG5306	
pEA001 [-1P]	P	W	W	M	BG5224	BG5040	Pair at position -1

Supplementary table 3.2. Primers

Name	Sequence
BG4467	TACTCATATGGGTCATCACCATCACCATCACAACACGATTAACATCGCTAAGAACGACTTC
BG4677	AGTAAAGCTTCGCCACCGAGTAGACCCTTAGACAACATGTCAGCCTCGACAACTTGC
BG4678	TATCAAGCTTGGTCTTCGTGGCATAAGGAAGACTC
BG4679	CGCGCCCGAGTCAATAGCTGGCTGAATGGTATCTTCCAG
BG4680	TACTAAGCTTTACTACCTCGGGCAAGGGTCTGATGTTCACTCAGCCGAATCAGGCTGCAGGATAC ATGGCTAAGCTGATTTGGGAATCTG
BG4681	AGTACCATGGTCTAGAAGATCTTTACGCGAACGCGAAGTCCGAC
BG4683	TACCAAGCTTCGCCACCGAGTAGCATTATGTTCAGATAAGGTCGTTAATC
BG4685	TATCCTGCAGCCTGATTCGGCTGAGTGAACATTAGCATTATGTTCAGATAAGGTCGTTAATC
BG4732	GAACATCGATACAGCGTAAACACGACCGCG
BG4733	CTGTATCGATGTTCAACCCGCAAGGTAACGATATGACCAAGGGTCTACTTACCCTGGCGAAGGGT AAACCAATCGGTAAGGAAGGTTACTAC
BG4735	TTACCCTTCGCCAGGGTAAGTAGCATTATGTTCAGATAAGGTCGTTAATC
BG4907	gctgACATGTTAtccAAGGGTTAATTGAGGCCTgaGTATAAGGTGACTTATACTTGTAATCTATC TAAAC
BG4908	CTGTATCGATGTTCAACCCGCAAGGTAACGATATGACCAAGGGTTAATTGAGGCCTgaGTATAAG GTGACTTATACTTGTAATCTATCTAAAC
BG4909	TAGACTCGGGCAAGGGTTAATTGAGGCCTgaGTATAAGGTGACTTATACTTGTAATCTATCTAAA C
BG5038	GATCTTAAGGATGTTCTcttgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5039	GATCTTAAGGATGTTTTgttgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5040	TGActgcagAATATTAAACGGTAGCATTATGTTCAGATAAGGTCG
BG5206	GATCTTAAGGATGTTTTctcaGGTTAATTGAGGCCTGAGTATAAGGTG
BG5207	GATCTTAAGGATGTTTTctcgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5208	GATCTTAAGGATGTTTTctctGGTTAATTGAGGCCTGAGTATAAGGTG
BG5209	GATCTTAAGGATGTTTTcttaGGTTAATTGAGGCCTGAGTATAAGGTG
BG5210	GATCTTAAGGATGTTTTcttgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5211	GATCTTAAGGATGTTTTctttGGTTAATTGAGGCCTGAGTATAAGGTG
BG5212	GATCTTAAGGATGTTTTctgaGGTTAATTGAGGCCTGAGTATAAGGTG
BG5213	GATCTTAAGGATGTTTTctggGGTTAATTGAGGCCTGAGTATAAGGTG
BG5214	GATCTTAAGGATGTTTTctgtGGTTAATTGAGGCCTGAGTATAAGGTG
BG5215	GATCTTAAGGATGTTTTcccaGGTTAATTGAGGCCTGAGTATAAGGTG
BG5216	GATCTTAAGGATGTTTTcccgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5217	GATCTTAAGGATGTTTTccctGGTTAATTGAGGCCTGAGTATAAGGTG
BG5218	GATCTTAAGGATGTTTTcctaGGTTAATTGAGGCCTGAGTATAAGGTG
BG5219	GATCTTAAGGATGTTTTcctgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5220	GATCTTAAGGATGTTTTccttGGTTAATTGAGGCCTGAGTATAAGGTG

BG5221	GATCTTAAGGATGTTTTccgaGGTTAATTGAGGCCTGAGTATAAGGTG
BG5222	GATCTTAAGGATGTTTTccggGGTTAATTGAGGCCTGAGTATAAGGTG
BG5223	GATCTTAAGGATGTTTTccgtGGTTAATTGAGGCCTGAGTATAAGGTG
BG5224	GATCTTAAGGATGTTTTccgtGGCTAATTGAGGCCTGAGTATAAGGTG
BG5304	GATCTTAAGGATGTTCTtttgGGTTAATTGAGGCCTGAGTATAAGGTG
BG5305	TGActgcagAATATTAAACGGgAGCATTATGTTCAGATAAGGTCG
BG5306	TGActgcagAATATTAAACGGaAGCATTATGTTCAGATAAGGTCG

Chapter 4

In vivo selection of riboswitches with an altered specificity

Sjoerd C.A. Creutzburg, Teunke van Rossum,

Servé W.M. Kengen, John van der Oost

Abstract

Aptazymes are ribozymes with a ligand responsive aptamer that can be deployed as synthetic riboswitches. The performance of the riboswitch is determined by the properties of the aptamer (e.g. the kd value), the location of the aptamer on the platform as well as the sequences of the platform, communication module and aptamer. Unfortunately, however, many grafted hybrids do not yield a functional riboswitch and not all aptamer variants bind ligands equally well. Screening and selection *in vivo* is the best way to obtain a riboswitch with the desired properties for *in vivo* application. We developed an approach to select for functional riboswitches from a large aptamer library (4×10^6 theoretical variants) based on the theophylline aptamer using selection and counter-selection based on thymidylate synthase (ThyA). Using this method, we were able to select for functional riboswitches, responding to 3-methylxanthine (3-MX) and not to theophylline (TP). Apart from the general applicability of the described approach, the selected aptamer variants provided some fundamental insights into the roles of specific nucleotide residues in riboswitch activity.

Introduction

Natural riboswitches are regulatory RNAs that are often located in the 5' untranslated regions (5' UTRs) of prokaryotic mRNAs. Riboswitches are typically composed of two domains, a ligand-binding aptamer domain and an expression platform domain, that together allow for allosteric regulation of gene expression. They typically form complex 3-dimensional structures that enable them to recognise small ligands such as metal ions (52), vitamins (53) and amino acids (54). Binding of these small molecules triggers a conformational change in the RNA of the aptamer domain, which is transduced to the platform, and either positively or negatively influences expression of the associated gene as an ON/OFF switch (55). Grafting new aptamers on the platform of a natural riboswitch potentially generates a synthetic riboswitch with novel ligand specificity.

Alternatively, synthetic riboswitches can be obtained by engineering natural ribozymes. For example, the self-splicing (group I) introns are ribozymes that are integrated in bacterial or viral genes (56). After their transcription, these ribozymes can splice themselves out of the hybrid mRNA and restore the original, intron-free open reading frame of the host gene. Grafting an aptamer onto the group I intron may yield an aptazyme that can be used as a riboswitch. An example of this approach was the grafting of the theophylline aptamer (11) onto the phage T4 *td* intron (36).

For each ligand a new aptamer must be developed. Synthetic aptamers are usually isolated from a library of random RNA molecules (15 – 60 nucleotides) by 'systematic evolution of ligands by exponential enrichment' (SELEX) (11), or by allosteric selection (57, 58). These *in vitro* procedures deliver multiple variants of RNA molecules that bind the target ligand with different affinity and specificity. Subsequent functional coupling of these aptamers to the riboswitch expression platform domain appears to be a major challenge. Grafting methods may differ from platform to platform, but they always involve a communication module between the aptamer and the platform to allow for signal transduction from the aptamer to the platform (32, 59, 60). The sequence of a communication module that allows for effective signal transduction depends on its surroundings, i.e. the platform and aptamer sequence. Therefore, the communication module is generally randomised and subjected to a screening or selection method. During screening or selection, binding of the ligand may cause the riboswitch to turn ON or OFF, or it may even cause no response at all, leaving it in either the ON or the OFF state. Since binding of a ligand should cause an allosteric change in the riboswitch conformation and, as a result, change in expression of the associated gene, selection of functional riboswitches from variant libraries requires the efficient elimination of non-responsive and incorrectly responding constructs. Changing the ligand specificity can be performed in a similar fashion, randomising part of the aptamer and screening or selecting for or against binding of related compounds.

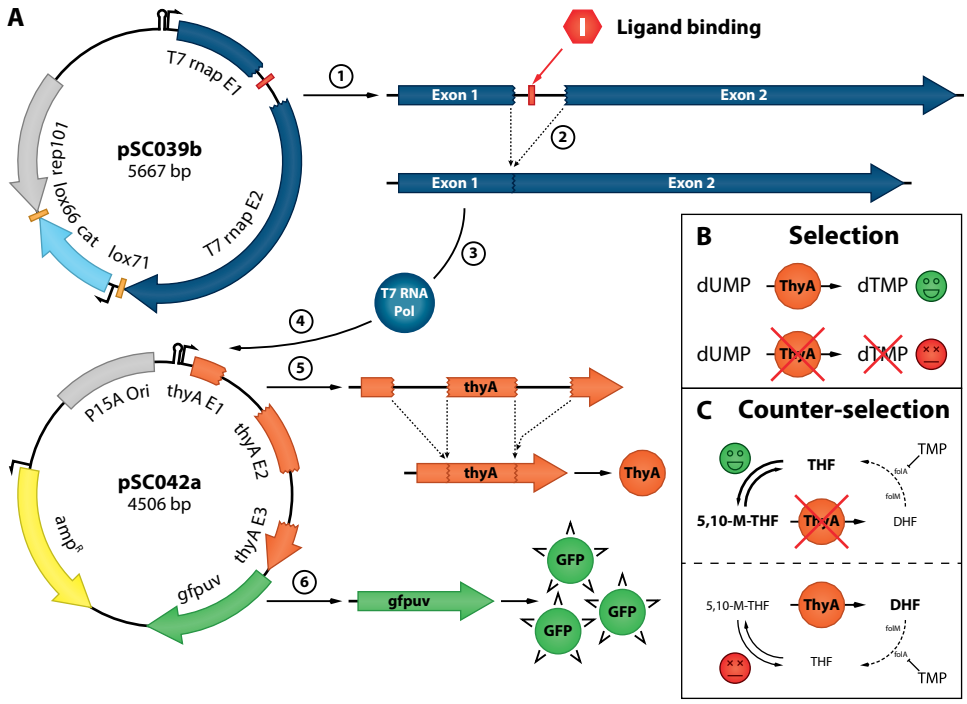


Figure 4.1. Overview of the RNAP sensor-reporter cascade. (A) RNAP cascade. (1) Transcription of the RNAP mRNA. (2) Ligand-dependent or independent splicing of the intron from the RNAP mRNA thereby restoring the open reading frame. (3) Translation of RNAP. (4) Binding of RNAP to the T7 promoter. (5) Transcription of *thyA* is followed by spontaneous maturation and translation to ThyA. In the presence of both TMP and dT, ThyA causes cell death. In the absence of both TMP and dT, ThyA is required for growth. (6) Transcription of *gfpuv* is followed by translation to GFPuv. (B) Mechanism of selection. ThyA is the only enzyme in *E. coli* converting dUMP into dTMP, so it is required for growth in the absence of thymidine. (C) Mechanism of the counter-selection. In the presence of thymidine, ThyA is not required for growth and will deplete the THF derivatives when trimethoprim (TMP) inhibits FolA.

Randomising nucleotides of either the communication module for grafting of the aptamer or part of the aptamer for changing the ligand specificity, will lead to an exponential increase in the number of variants. Moreover, only a small fraction of the cells expressing the generated library will harbour a functional riboswitch with the desired ligand specificity. To establish an approach that allows for selecting functional riboswitches from a variant library, the riboswitch activity is coupled to the expression of a reporter, like the gene encoding the green fluorescent protein (GFP). The fluorescence indicates whether the riboswitch is ON or OFF under the assay conditions. No exposure to ligand should yield no signal, while presence of the ligand should induce functional expression of the reporter gene. After the transformation of bacterial cells with the riboswitch library (including the GFP reporter), most of the cells will not react to the presence of the intended ligand and yield a signal somewhere between ON and OFF. Because of the high

number of variants, isolation of a functional riboswitch is only feasible when screening large numbers of bacteria with a high-throughput technique like fluorescence-activated cell sorting (FACS) (21). Although this can be done, it still takes a lot of processing time. To make isolation easier, it is preferable to use a selection/counter-selection system instead of a screening system.

In this study, we developed an *in vivo* selection/counter-selection method that allows cells with a functional riboswitch to outcompete the rest of the library variants. The system exploits the thymidylate synthase (ThyA) enzyme that catalyses the synthesis of the essential DNA precursor deoxythymidine monophosphate (dTMP). It allows for iteration of selection (ON in the presence of inducer) and counter-selection (OFF in the absence of inducer). Bacteria harbouring a functional riboswitch will gain dominance over the rest of the bacteria upon iteration of selection and counter-selection. The *thyA* gene is under control of the T7 promoter. The T7 RNA polymerase (RNAP) is interrupted by the T4 *td* intron with a library of aptamers grafted onto it. The T7 polymerase also controls the expression of a GFPuv-encoding gene (*gfpuv*), so the response to the inducer can be easily monitored. Using the theophylline aptamer grafted onto the phage T4 *td* group I intron as a base, we were able to select for riboswitches that respond to 3-methylxanthine (3-MX) and not to theophylline (TP).

Results

Design of the selection/counter-selection system

The phage T4 *td* intron has previously been converted into an aptazyme by introduction of the theophylline (TP) aptamer (36). Here we aim to make a first step in developing this aptazyme into a universal platform for ligand detection. However, substituting the TP aptamer for different aptamers or altering the specificity of the TP aptamer in place requires a high-throughput screening or selection method to fish out functional aptazymes. Preferably, this is done by using iterative selection/counter-selection combined with screening.

Combining a selection/counter-selection approach with a screening approach requires a system in which at least two genes are controlled by the same aptamer variant. Introducing the same aptamer twice or more into a set of genes in a random fashion is impossible. To solve this, we made a design in which (I) the functional expression of T7 RNA polymerase (RNAP) is controlled by the group I intron-based riboswitch, and (II) the polymerase controls the expression of both ThyA for selection/counter-selection, and GFPuv for screening the response to its target ligand (Figure 4.1).

Selection occurs through synthesis of dTMP by the plasmid-encoded ThyA, in a strain auxotrophic for dTMP. Without a functional ThyA, cells are unable to synthesise DNA and

therefore will not grow. Counter-selection is performed by adding trimethoprim (TMP), which inhibits FolA, causing ThyA to be toxic (Supplementary figure 4.1A), and thymidine (dT), which is converted into dTMP in a ThyA-independent manner (Supplementary Figure 4.1B), to complement the auxotrophy. Selection on ThyA activity is performed in medium containing the desired new riboswitch ligand, 3-methylxanthine (3-MX). ThyA expression is controlled by functional expression of the T7 RNA polymerase (RNAP) which requires splicing of the group I intron (Figure 4.1). Splicing may or may not be induced by 3-MX. When splicing is not induced and does not occur by itself, no ThyA is expressed and under these conditions the cells will die. Counter-selection, on the other hand, is performed on medium containing TP, dT and TMP. Combining TMP with dT will only allow growth of cells that do not express ThyA, i.e. cells that do not splice out the intron from the RNAP transcript. Cells having active ThyA in the absence of 3-MX (e.g. cells that are induced by TP or cells that have splicing without induction) are killed by TMP. Alternating the selection and counter-selection will favour growth of cells that, in the ideal case, carry an intron that splices out in reaction to binding of 3-MX and that does not respond to TP.

Using the RNAP as intermediate has some implications for the system. Apart from controlling the expression of multiple genes, the high processivity of the RNAP causes it to act as a signal amplifier, which may result in major differences in transcription level of the gene(s) it controls, even at very low RNAP expression levels. This large effect would not have been possible were the aptazyme introduced directly into either the *gfpuv* or *thyA* gene. Because the intron needs time to splice out before the mRNA is degraded, it causes a serious drop in gene expression of the gene it interrupts. This effect of the intron on gene expression was also exploited to impede the expression of ThyA. Two inducer-independent introns interrupt the *thyA* gene so transcription by RNAP is required and it does not survive without the high transcription by the RNAP. Low expression levels may not be problematic for many enzymes, as a longer reaction time may still suffice to measure their activity. GFPuv on the other hand needs a decent expression to distinguish it from the auto-fluorescence background. Hence, direct interruption of GFPuv not a suitable approach (Supplementary figure 3.1). Instead, placing GFPuv under control of the RNAP solves this problem.

Another feature of the RNAP cascade is a population-wise induction. This means that some cells do express GFPuv and ThyA and others do not. Since both genes are controlled by RNAP, the presence of functional RNAP dictates the expression of GFPuv and ThyA. Once the RNAP mRNA, containing an intron, has been made, it has limited time for maturation (i.e. intron splicing) and translation before it is degraded. When none of the mRNAs is matured and translated before degradation, there is no RNAP protein. This is population number one. A single RNAP mRNA will generate multiple RNAP protein molecules, which in turn will generate huge amounts of transcript of *gfpuv* and *thyA*. Multiple functional RNAP mRNAs will not contribute as much to the RNAP activity as the first and cannot

easily be distinguished, so a RNAP concentration larger than zero is the other population. Presence of RNAP protein causes high transcription of *gfpuv* and *thyA*. This means that those cells not expressing RNAP will not be green fluorescent and will not survive the selection. They are, however, now resistant to TMP and will survive the counter-selection. When the RNAP concentration is not zero, the vast amounts of transcript for *gfpuv* will turn the cell green fluorescent, while the high transcription of *thyA* causes the cell to produce amounts of ThyA high enough to survive selection and to die during counter-selection despite the impediment by the two introns in the *thyA* gene. The splice rate of the intron in the RNAP mRNA determines how many mRNAs are matured before degradation. The chance of that amount being larger than zero increases with splice rate. The number of cells expressing the RNAP and, as a consequence, GFPuv and ThyA, increases with the splice rate. If the splice rate is ligand dependent as it would be in a functional riboswitch, the fraction of cells expressing the ThyA is also ligand dependent. In case of introns not responding to the inducer compound the fraction of cells with growth advantage during the selection is the same fraction with disadvantage during the counter-selection. This is not true for riboswitch containing cells. Those not only have a large population of cells not expressing ThyA in the absence of ligand during counter-selection, but they also have a large population of cells expressing ThyA in the presence of ligand during selection.

The importance of the population-wise induction is best illustrated by comparing it to direct interruption of the *thyA* gene by a possibly ligand-dependent intron. The population-wise induction causes high expression of ThyA or almost none which means there is no "small" amount of ThyA in a cell. Direct interruption of the *thyA* gene by an intron may lead to variants that have low expression of ThyA. A small amount of ThyA being produced may lead to generating enough dTMP to survive selection, but not enough depletion of 5,10-dimethylene-THF to cause toxicity of TMP during counter-selection. How small the amount of ThyA actually is, may differ between the selection conditions and the counter-selection conditions, so cells might grow at almost full rate during both selection and counter-selection. Distinguishing between the cells harbouring a riboswitch and the cells balancing their ThyA expression is hard, since the riboswitches will not establish dominance in the culture.

As a proof of principle for the screening and selection/counter-selection system, we made a design based on the previously generated 3-MX and TP aptamers, allowing both aptamers to be present in the library. Whereas the TP aptamer can bind both 3-MX and TP (11), the 3-MX aptamer can only bind 3-MX (61). Selection is therefore performed in the presence of 3-MX on LBG medium. Counter-selection is performed on LBG medium supplemented with TP, dT and TMP. While in theory there is no bias towards generating the TP aptamer, counter-selection against TP should prevent its enrichment.

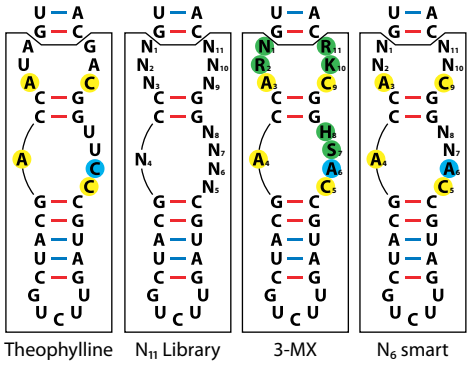


Figure 4.2. Structure of the aptamer domain. The TP aptamer was randomised for all 11 nucleotides not involved in Watson-Crick base pairing to yield the N11 library. For all riboswitches reacting to 3-MX (3-MX), the yellow nucleotides are conserved and identical to the TP aptamer. The blue adenine is conserved for 3-MX aptamers, but differs from the TP aptamer. The non-conserved nucleotides are marked green. Although not conserved, there is preference for certain combinations. This was further assessed with the N6 smart library. IUPAC codes are N=G/A/U/C, R = A/G, D = G/A/U, K=G/U, H=A/U/C, S=G/C. CG pairs are indicated in red and AU pairs in blue. Wobble base pairs are indicated with a dot.

Library generation

The TP aptamer was randomised in such a way that the basic structure was retained (Figure 4.2). All 11 non-base pairing nucleotides were made completely degenerate. The theoretical number of variants is 4^{11}, about 4.2×10^6. To ensure that the vast majority of the variants was included in the selection, a large enough library had to be constructed containing at least 4×10^7 colony forming units (cfu). The library containing 11 random nucleotides was constructed by "'round-the-horn" site directed mutagenesis (62) of pSC039-Theo4. The initial library contained 2.36×10^8 cfu giving a theoretical 56-fold coverage.

N_{11} enrichment on 3-MX

After library generation, the cells were subjugated to the selection and counter-selection for the enrichment of functional riboswitches. The enrichments were monitored for their fluorescence response to 3-MX, TP and ultrapure water (control). Measurements were performed in triplicate for each condition. For the initial 5 transfers no enrichment of 3-MX responsive variants was visible (Figure 4.3). Only after counter-selection round 6 (CS6), it became apparent that the selection/counter-selection worked and that 3-MX-responsive bacteria were enriched. As anticipated, these bacteria were not responsive to TP (Figure 4.3).

Individual colonies were picked once the fluorescence showed significant difference between the selection and counter-selection agent at counter-selection round 7. The clones reacting to 3-MX and not to TP were sequenced. A total of 31 colonies was picked, of which 25 were responsive to 3-MX. Out of these 25 responsive clones, 17 proved unique after sequencing.

Sequence analysis revealed that five of the randomised nucleotides are conserved in all of the 3-MX-responsive clones (Figure 4.2). The conserved nucleotides match with those of

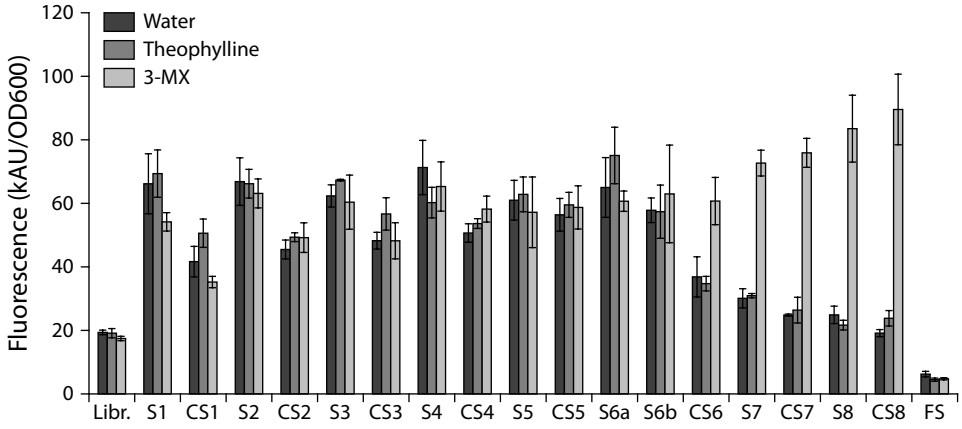


Figure 4.3. Selection of riboswitches with improved specificity from the N_{11} library. The enrichment performed with 3-MX selection and TP counter-selection was grown on 3-MX and TP. A control with ultrapure water was included to assess whether the counter-selection agent acts as an antagonist. The x-axis represents alternating rounds of selection (S) and counter-selection (CS). A RNAP-frameshift (FS) control indicates the fluorescence background of the bacteria containing the pSC042a reporter plasmid. S6a and S6b represent the two 10^{-2} dilutions where S6a was incubated at 37°C overnight instead of 7h.

the previously described 3-MX aptamer (36, 61, 63). Nucleotide A6 was found to be a key nucleotide discriminating between TP and 3-MX. At position 6, 3-MX binding requires A or C, while TP binding requires C. Counter-selection on TP eliminated the variants with a C at position 6. Altogether, the results indicated the following consensus sequence: NNA-A-CANN-CNN. Six positions allowed several nucleotide options and a large difference in response to 3-MX was found for individual clones. Aptamers conforming to the consensus sequence typically have a response ratio 3-MX/TP between 2.7 and 9.9 for 1 mM of 3-MX versus 1 mM TP. This small dataset does already show some correlations (Figure 4.4). Position 7 has mostly S (G/C) for the highly responsive variants, while W (A/T) is more prevalent in the poorly responsive variants. Position 8 is preferably not a G.

Variants with a high ratio between 3-MX and TP occur more frequently than variants with a low ratio – average of minimum and maximum equals 6.3, while the median is 7.3. However, this method does result in enrichment of variants with an induction ratio as low as 2.7 as well, indicating that given the presence of a functional riboswitch in the library, it may become enriched even though it has no optimal riboswitch properties. Eventually, the variants with a high ratio will become dominant, depending on the number of cycles of the enrichment procedure.

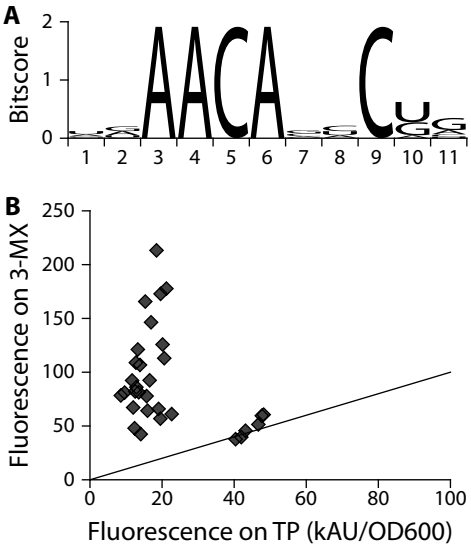


Figure 4.4. Statistics of the N_{11} library. Individual clones from counter-selection round 7 were tested for their response to 3-MX and TP and sequence analysed. (A) Weblogo of all clones responsive to 3-MX and not to TP. 5 positions are clearly conserved and some others have a preference. (B) The fluorescence of 3-MX and TP indicates whether certain clones are associated with high or low response to 3-MX. Most of the clones from the selection are responsive to 3-MX and not to TP, albeit in a vast range of 3-MX/TP ratios.

N_6 smart library

Analysis of the N11 library revealed that six nucleotide positions in the 3-MX responsive aptamers appear to be less conserved resulting in a wide range of 3-MX/TP fluorescence ratios. Based on this observation, a smaller library was generated and screened. This library had positions 1, 2, 7, 8, 10 and 11 randomised, while position 6 was kept A to counter binding of TP (Figure 4.2). The use of a lower 3-MX concentration (0.1 mM) is anticipated to separate the poorly responsive riboswitches from the highly responsive ones. The library has 4096 theoretical variants and the cfu count at the start was 4×10^6, so it has a coverage of roughly 10^3. Since the smaller library is expected to have a larger fraction of 3-MX responsive clones, a total of 55 clones, derived from selection round 2 (S2) and 4 (S4), was tested for their response to 3-MX and sequenced. The 3-MX and TP responses are shown in Figure 4.5. While the S2 has already a good number of 3-MX responsive clones, the S4 is clearly further enriched, yet also still has unresponsive clones.

Based on the crystal structure of the TP aptamer (63) and the analysis of the N6 smart library, several interactions can be distinguished that are influenced by the type of nucleotide. These interactions are position 1 – 2 (stacking), 10 – 11 (stacking), 1 – 11 (hydrogen bonds), 2 – 7 – 10 hydrogen bonds and stacking with the ligand (TP). Position 8 is involved in stacking with the base pairing G nucleotide two nucleotides downstream and hydrogen bonding with TP. Assuming 3-MX is positioned in the aptamer similar to TP, the interactions that hold TP in place might play a role in 3-MX coordination as well. It implies that the positions 1-2-7-10-11 might all be dependent on each other. Position 8 has more independence as it is not known to bind any other variable factor in the N6 smart library. To analyse the interactions within the aptamer, boxplots of the 3-MX/TP

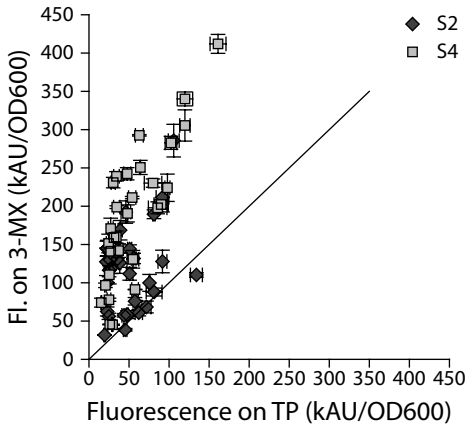


Figure 4.5. Statistics of the N6 library. Individual clones from selection rounds 2 and 4 were tested for their response to 3-MX and TP in triplicate. While selection round 2 (S2) still has a lot of clones with hardly any response, S4 is already more enriched for the highly responsive clones.

fluorescence ratio were made for several combinations of nucleotide positions (Figure 4.6). Pairs of nucleotides associated with a strong 3-MX response will have a high value median, while a low value median corresponds to a weak response.

Positions 1 and 2 are preferably N and R (G/A), but only UA has high responsiveness to 3-MX consistently (Figure 4.6A). Medium response would be GA, GG, AG, UG, CA. Combinations AA and CG mostly have low response to 3-MX. Positions 1 and 10 are preferably N and K (G/U) (Figure 4.6B). Combinations with position 10 being U are all rather good regardless of position 1, but position 10 being G requires position 1 to be U. Positions 1 and 11 are preferably N and D (G/A/U) (Figure 4.6C). Closer inspection reveals that the combinations are more like RR, KU, UG and CA instead of all possibilities. Positions 2 and 11 are preferably R and D (Figure 4.6D). GG and AU are highly preferred, GA and AR are decent, but GU shows hardly any response to 3-MX. Positions 10 and 11 are preferably K and D (Figure 4.6F). Again, not all combinations respond to 3-MX similarly. GG and UD can have a good response, but GA and GU do not. Positions 2 and 10 are preferably R and K (Figure 4.6I). The combination of A and G does not work, while the other three of RK do. The three most prevalent combinations of positions 1, 2, 10 and 11 associated with high response to 3-MX are UA-UU, UG-GG and CA-UA.

Positions 1 and 11 are important for multiple functions. These nucleotides are involved in both stabilising the interactions between the aptamer and 3-MX and influencing the splicing activity of the intron. Elongating the stem by introduction of another base pair directly adjacent to the existing stem (positions 1 and 11) could force the riboswitch in the ON state. A CG pair exhibits moderate non-induced splicing, but shows low induction too. A GC pair results in high non-induced splicing and low induction. AU shows high non-induced splicing with hardly any induction. For these three pairs, the riboswitches appear to be good enough to survive a couple of rounds of selection and counter-selection. A riboswitch with the UA pair does not at all occur in this data set, suggesting that it either

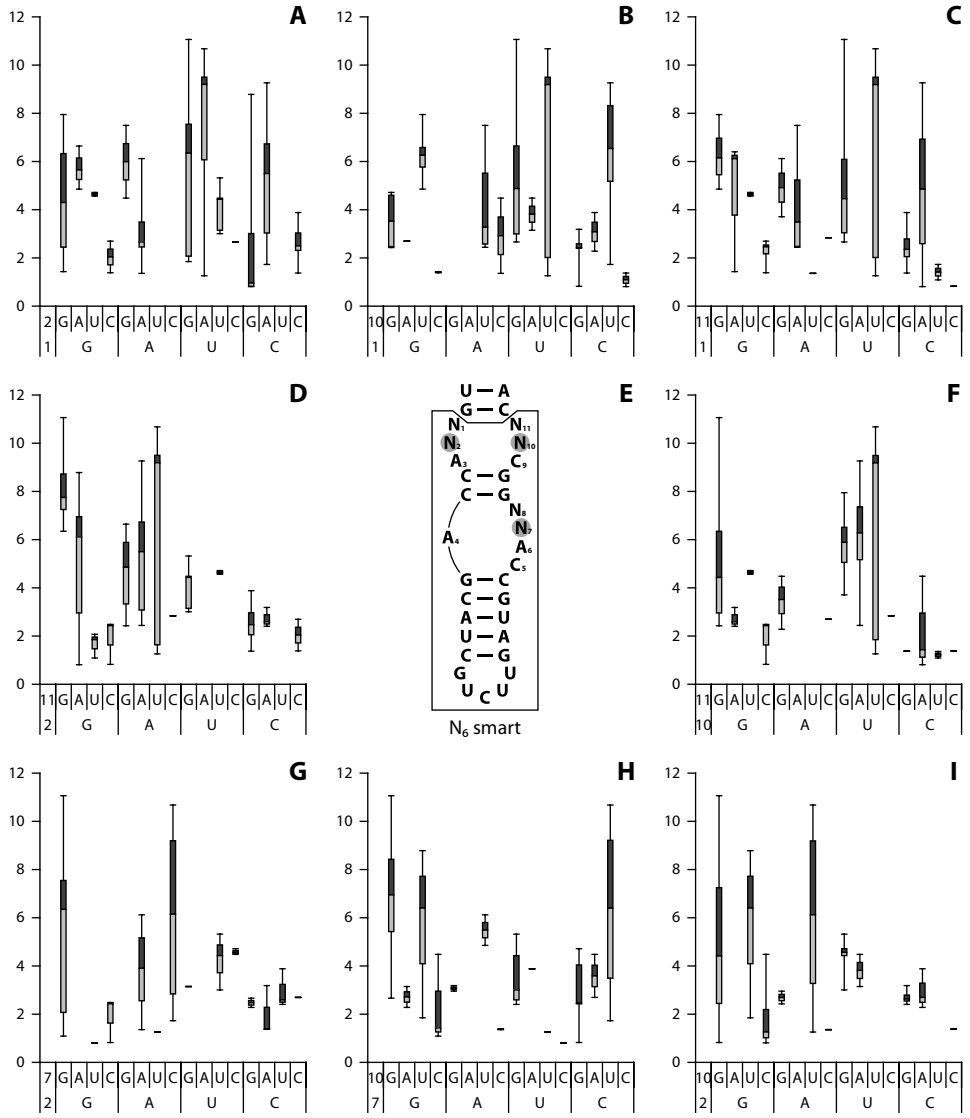


Figure 4.6. Analysis of pairs of nucleotides in the aptamer. Y-axis shows the quotient of fluorescence of bacteria exposed to 3-MX and to TP analysed by position pair of all individual clones combined (N11 CS7, N6 S2, N6 S4). a-d and f-h give response to 3-MX compared to TP for several position combinations. The positions involved are indicated at the bottom left. The floor of the binding pocket is indicated with shaded circles in panel e.

exhibits very high non-induced splicing or prevents binding of 3-MX altogether. Wobble base pairs UG and GU do not cause high non-induced splicing per se and they do allow for high 3-MX/TP response ratios. However, the performance of the riboswitch with positions 1 and 11 forming a wobble base pair depends highly on the other nucleotides. Generally speaking, mismatching 1 and 11 or formation of a UG wobble pair seems to be most favoured. Especially UU is associated with high 3-MX/TP ratios.

The positions 2, 7 and 10 form the floor of the ligand binding pocket. Since all three interact with each other through hydrogen bonding, favoured and disfavoured combinations are likely. Positions 2 and 7 show preference for GG, AC and possibly AA (Figure 4.6G). Positions 2 and 10 show preference for GG, GU and AU (Figure 4.6I). Positions 7 and 10 show preference for GG, GU and CU (Figure 4.6H). If 2 is G, then 7 is G and 10 is G or U, so GGK. If 2 is A, 7 is C or A and 10 is U, so AMU. So, the stacking is performed by rather strict and distinct combinations of nucleotides: GGK or AMU on positions 2, 7 and 10. The aptamer variants have predominantly either ACU (29% of total) or GGK (18% of total). The nucleotide bases at positions 1 and 2 as well as 10 and 11 are stacking too. This results in division into groups. In high fold-change variants the 2 – 7 – 10 triplet ACU is associated with 1 – 11 UU yielding UA-CH-UU and 1 – 11 CA yielding CA-CH-UA. The other variants are UG-G-H-KG.

Discussion

Riboswitches can be used to regulate the expression of specific genes and as such may find applications as biosensor in synthetic biology. However, available riboswitches generally will not have the desired specificity, hence novel switches need to be developed that are responsive to a ligand of interest. Here, we developed an *in vivo* selection system for finding novel aptazymes with an altered ligand specificity. The system is based on the T7 RNA polymerase that controls the expression of a dual reporter system consisting of an auxotrophic selection marker (ThyA) and a fluorescent screening marker (GFPuv). The expression of the T7 polymerase is controlled by an interrupting group I aptazyme that is responsive to theophylline (TP). Here, we used the system to successfully change the specificity of the aptazyme from TP to 3-methylxanthine (3-MX) by alternating rounds of selection and counter-selection.

Eleven nucleotides of the aptamer sequence that are not involved in base pairing were initially randomised. During the first selection rounds of the N_{11} library it appeared as if nothing was happening (Figure 4.3). First after counter-selection round 6, the culture was suddenly taken over by 3-MX-responsive bacteria. This apparent delay can be explained by assuming that a small number of highly fluorescent cells does not increase the overall fluorescence of the entire culture significantly. Only when the fraction of responsive bacteria reaches a certain level, the difference in fluorescence will become apparent.

For instance, while there is a large difference between a fraction of 10^{-5} and 10^{-2}, neither of them will significantly impact the fluorescence. Here, the 3-MX responsive bacteria gain advantage over one competing group each round. During selection (on 3-MX) they gain advantage over the "mostly OFF" group and during counter-selection (on TP) over the "mostly ON" group. After selection round 6, the 3-MX-responsive group may have outcompeted the "mostly OFF" group, but not the "mostly ON" group. During counter-selection round 6, the "mostly ON" group is being outcompeted while the "mostly OFF" group cannot recuperate quickly enough. At this tipping point the functional riboswitches have become the dominant fraction and the fluorescence caused by the induction (functional RNAP as a result of 3-MX-induced intron splicing) is having an impact on the total fluorescence.

Most of the variants obtained from the enrichment show a response to the inducer compound 3-MX and not to TP. Five nucleotides appeared to be completely conserved in all variants responding to 3-MX. However, concerning the other six nucleotides there is a large variety between the different clones, although some patterns can still be distinguished. A third Watson-Crick base pair after the stem connecting the aptamer to the intron is highly disfavoured for 3-MX responsiveness. A wobble base pair or mismatch pair after this stem can yield both high and low induction, which means that the response to 3-MX is not solely depending on the connection point between the aptamer and the platform. We can see the influence of the rest of the aptamer as well.

Although a high-resolution structure is required to draw solid conclusions, it is tempting to speculate on the molecular basis of the obtained aptamer variants. According to the known structure of the TP aptamer, the cytosine at position 6 is hydrogen bonding with the O6 and H7 of TP (Supplementary Figure 4.2) (63). The observed substitution of the cytosine with adenine in case of the 3-MX aptamer (Figure 4.2), may suggest that the positioning of the TP and 3-MX ligands is not the same in the two aptamers, as the purine adenine is larger than the pyrimidine cytosine. The difference between the two ligands is the methylation of the N1 position in TP that is absent in 3-MX. This N1 position on 3-MX or TP does not interact with the cytosine or the adenine at position 6 of the aptamer. However, changing the C into A may have consequences for the rest of the binding pocket. There are some highly conserved nucleotides in the aptamer equal for TP and 3-MX, but for example the 2-7-10 triplet of the TP aptamer (U-U-A) is hardly found at all when enriching for 3-MX responsivity (preferably A-C-U or G-G-K). It means that the positions selected for the N6 smart library might have a different nucleotide preference for TP compared to 3-MX, possibly caused by a slightly different positioning of 3-MX in the aptamer. This hypothesis can only be confirmed by crystallography.

While the screening/selection/counter-selection method was employed for altering the specificity of an aptamer, it could also be used for engineering of novel aptamer-platform

combinations. This usually requires a communication module between the aptamer and the platform. Also, aptamers are not always confined to a specific sequence as was observed in this study. Obtaining the best possible combination of communication module and aptamer requires the randomisation of the communication module as well as (part of) the aptamer. Since randomising nucleotides causes exponential increase of possibilities for each nucleotide randomised, an efficient selection and counter-selection system is required. A proof of principle for such a selection and counter-selection was previously obtained for the protein switch LacI (64). It takes advantage of the tetA gene conferring tetracycline resistance (selection) and sensitivity to Ni^{2+} or several chelator compounds (counter-selection) (65). However, there is no prior knowledge on how a novel switch will behave and overexpression of the tetA gene appears to be toxic to *E. coli*, which is often the case for membrane proteins (66, 67). Moderate expression on the other hand, does neither render the bacteria very resistant to tetracycline nor very sensitive to Ni^{2+} or a chelator compound (65). The system we describe here does not have these drawbacks. The ThyA protein does not become toxic in an RNAP-based expression system, and due to the population-wise induction, counter-selection is more effective.

In conclusion, we here describe an *in vivo* selection and counter-selection system that can be used to enrich for aptazymes with altered ligand specificity. The functionality of the system has been demonstrated by generating a library of millions of theophylline aptamer variants, from which a couple dozens of aptazymes were selected that lost their ability to bind theophylline, while retaining the binding of 3-methylxanthine. Importantly, the here described screening and selection system can easily be adapted to enrich for an *in vivo* biosensor that is controlled by another type of riboswitch, or even by a protein-based sensor/regulator like LacI.

Materials and methods

Plasmids

The pSC039b-Theo4 plasmid (Supplementary sequence 4.1) encodes the actual riboswitch, involving a T7 RNA polymerase (RNAP) derived from pAR1219 (Sigma-Aldrich) under control of a Pbla promoter. This RNAP is interrupted by a TP-dependent group I aptazyme. Upon binding of TP to the aptamer of the riboswitch the group I self-splicing intron is excised and a mature mRNA is formed. The plasmid has a very low copy number due to the SC101 origin of replication from pFU98 (68), and contains a chloramphenicol resistance gene from pACYC184 (69). A derivative plasmid, pSC039b-FS was made by digesting pSC039b-Theo4 with HindIII, Klenow fill-in and religation. This generates a frameshift in RNAP.

Plasmid pSC042a (Supplementary sequence 4.2) is the reporter plasmid derived from pACYC184 (69) and harbours the *E. coli thyA* gene and the *gfpuv* gene from pGFPuv (Clontech Laboratories). Both are under control of the same T7 promoter. It was observed that weak background expression of ThyA already nullifies the auxotrophy. Therefore, we introduced a weak ribosome binding site and interrupted *thyA* by two T4 *td* introns, which are not dependent on any inducer but slow down functional expression, to make its expression completely dependent on T7 polymerase activity. Its ampicillin resistance marker is derived from pUC19 (Invitrogen).

Strains and media

E. coli DH10B T1R (Invitrogen C6400-03) was used for plasmid propagation and general cloning techniques. *E. coli* DH10B-Δ*thyA*, made according to van Rossum et al. 2017 (70) with primers BG4510 and BG4511 (Supplementary table 4.1) and pMA-RQ_lox71_kan_lox66 as template, was used for the selection and counter-selection. Bacteria were cultured on LB (10 g/L bacto peptone (Oxoid), 5 g/L yeast extract (BD), 10 g/L NaCl (Acros)) or LBG (LB with 5 g/L glycerol). Antibiotics were used if appropriate: chloramphenicol (cam) (35 mg/L), ampicillin (amp) (100 mg/L). Trimethoprim (TMP), theophylline (TP) and 3-methylxanthine (3-MX) were purchased from Sigma-Aldrich.

Library generation

The library with eleven randomised nucleotides (N11; Figure 4.2) in the aptamer domain of the self-splicing intron, was generated by "round-the-horn" PCR with forward primer BG6768 (Supplementary table 4.1), reverse primer BG6243 and pSC039b-Theo4 as template. This technique amplifies the whole plasmid and introduces random nucleotides on the end of the PCR product, which is the aptamer domain (62). The PCR product was purified with a Zymo clean and concentrator kit (D4004) and circularised by T4 ligase. The circularised PCR product was purified again and about 5 µg was used to transform *E. coli* DH10B-Δ*thyA* harbouring the pSC042a reporter plasmid. Plating of several dilutions on LB agar plates containing cam and amp provided an estimation for the library size. The library that has six out of eleven positions randomised (N6 smart; Figure 4.2) was created in a similar fashion using BG7741 as forward primer instead of BG6768. All oligonucleotides were purchased from Sigma-Aldrich.

Selection and counter-selection

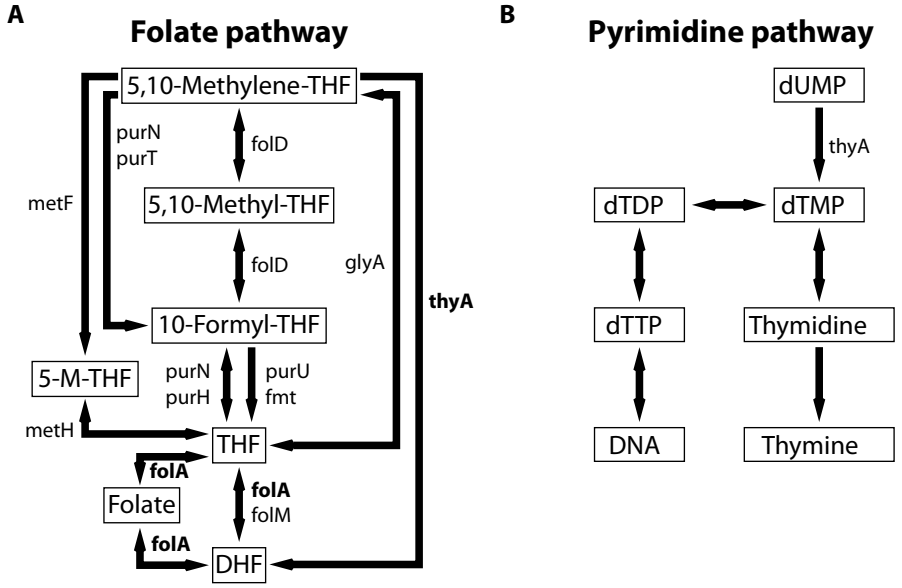
The library was transferred in a dilution of 1×10^{-4} for each transfer between selection and counter-selection. To prevent the use of large volumes of medium while retaining the full library, 100 µL of culture was transferred to 10 mL first and grown for >7h at 37°C. Then, this culture was transferred another time in the same dilution and grown at 37°C overnight. The selection medium consisted of LBG supplemented with cam (35 mg/L),

amp (100 mg/L) and the intended aptamer ligand 3-MX (1 mM). The counter-selection medium consisted of LBG supplemented with cam (35 mg/L), amp (100 mg/L), thymidine (dT) (40 mg/L) and the possible, but unintended, ligand TP (1 mM). Both selection and counter-selection cultures were grown at 37°C overnight, but the counter-selection culture was only supplemented with 5 mg/L TMP after an incubation of 2.5 h at 37°C. Since the N6 smart library should contain many 3-MX responsive variants, the selection was performed more stringently with 0.1 mM 3-MX instead of 1 mM. Transfers from selection to counter-selection and vice versa were done by diluting 1×10^{-3}.

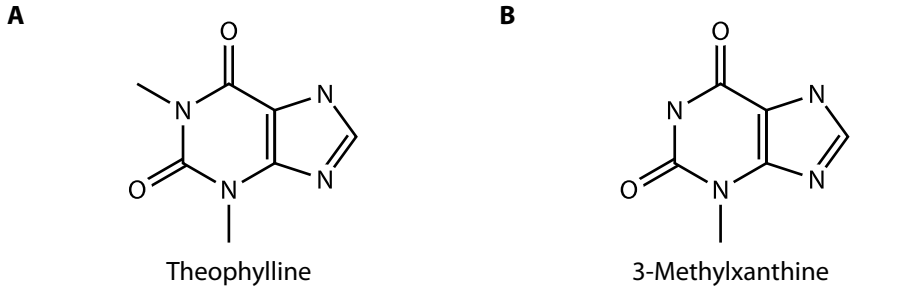
Inducer response assay

To monitor the progress of the enrichment and to assay the performance of individual clones, the GFPuv fluorescence was measured. The fluorescence assay was carried out in 500 µL LB supplemented with cam (35 mg/L) and amp (100 mg/L) in a Masterblock 96 deep well microplate (Greiner). dT (100 mg/L) was added to avoid selection and TMP was omitted to avoid counter-selection. Cultures were re-inoculated in triplicate by diluting 1×10^{-3} from an overnight culture and the cells were grown at 37°C overnight in the presence of either ultrapure water, TP or 3-MX at 1 mM. The bacteria were harvested by centrifugation (3000 × g; 10 min) and resuspended in 500 µl of 50 mM Tris-HCl (pH 7.5). Since the oxygen supply is not sufficient for maturation of the GFPuv, the resuspended cells were left to mature for 1h at 37°C. Then, 100 µL of cell suspension was measured using a BioTek Synergy MX microplate reader. Fluorescence was then calculated as raw fluorescence per OD600 for 100 µL. Auto-fluorescence was estimated by introduction of a frameshift in RNAP (pSC039-FS).

SUPPLEMENTARY DATA



Supplementary figure 4.1. Folate and pyrimidine pathways. (A) FolA is inhibited by TMP, which causes ThyA to exhaust THF and its derivative metabolites. These derivatives are needed as substrates for among others the essential Fmt. De novo synthesis can be done via FolM, which is not TMP sensitive, but it cannot supply enough THF to overcome the exhaustion by ThyA. (B) As dTMP is an essential metabolite, counter-selection with TMP can only be done with an alternate source. The thymidine is added to negate the auxotrophy. During selection the thymidine is not added and the bacteria rely on ThyA activity for their dTMP supply.



Supplementary figure 4.2. Structures of theophylline and 3-Methylxanthine. Theophylline (A) has an additional methylation on the N1 position compared to 3-methylxanthine (B).

Supplementary sequence 4.1. pSC039-Theo4

GGTACCCCGC TTCGGCGGGG TTTTTTCAAG TTCAAATATG TATCCGCTCA TGAGACAATG TGTGGGGAGA
CCACAACGGT TTCCCTCTAG AAATAATTTT GTTTAACTAT AAGAAGGAGA TATACATATG GGTCATCACC
ATCACCATCA CAACACGATT AACATCGCTA AGAACGACTT CTCTGACATC GAACTGGCTG CTATCCCGTT
CAACACTCTG GCTGACCATT ACGGTGAGCG TTTAGCTCGC GAACAGTTGG CCCTTGAGCA TGAGTCTTAC
GAGATGGGTG AAGCACGCTT CCGCAAGATG TTTGAGCGTC AACTTAAAGC TGGTGAGGTT GCGGATAACG
CTGCCGCCAA GCCTCTCATC ACTACCCTAC TCCCTAAGAT GATTGCACGC ATCAACGACT GGTTTGAGGA
AGTGAAAGCT AAGCGCGGCA AGCGCCCGAC AGCCTTCCAG TTCCTGCAAG AAATCAAGCC GGAAGCCGTA
GCGTACATCA CCATTAAGAC CACTCTGGCT TGCCTAACCA GTGCTGACAA TACAACCGTT CAGGCTGTAG
CAAGCGCAAT CGGTCGGGCC ATTGAGGACG AGGCTCGCTT CGGTCGTATC CGTGACCTTG AAGCTAAGCA
CTTCAAGAAA AACGTTGAGG AACAACTCAA CAAGCGCGTA GGGCACGTCT ACAAGAAAGC ATTTATGCAA
GTTGTCGAGG CTGACATGTT ATCCAAGGGT TAATTGAGGC CTGAGTATAA GGTGACTTAT ACTTGTAATC
TATCTAAACG GGGAACCTCT CTAGTAGACA ATCCCGTGCT AAATTGATAC CAGCATCGTC TTGATGCCCT
TGGCAGCATA AATGCCTAAC GACTATCCCT TTGGGGAGTA GGGTCAAGTG ACTCGAAACG ATAGACAACT
TGCTTTAACA AGTTGGAGAT ATAGTCTGCT CTGCATGGTG ACATGCAGCT GGATATAATT CCGGGGTAAG
ATTAACGACC TTATCTGAAC ATAATGCTAC TCGGTGGCGA AGCTTGGTCT TCGTGGCATA AGGAAGACTC
TATTCATGTA GGAGTACGCT GCATCGAGAT GCTCATTGAG TCAACCGGAA TGGTTAGCTT ACACCGCCAA
AATGCTGGCG TAGTAGGTCA AGACTCTGAG ACTATCGAAC TCGCACCTGA ATACGCTGAG GCTATCGCAA
CCCGTGCAGG TGCGCTGGCT GGCATCTCTC CGATGTTCCA ACCTTGCGTA GTTCCTCCTA AGCCGTGGAC
TGGCATTACT GGTGGTGGCT ATTGGGCTAA CGGTCGTCGT CCTCTGGCGC TGGTGCGTAC TCACAGTAAG
AAAGCACTGA TGCGCTACGA AGACGTTTAC ATGCCTGAGG TGTACAAAGC GATTAACATT GCGCAAAACA
CCGCATGGAA AATCAACAAG AAAGTCCTAG CGGTCGCCAA CGTAATCACC AAGTGGAAGC ATTGTCCGGT
CGAGGACATC CCTGCGATTG AGCGTGAAGA ACTCCCGATG AAACCGGAAG ACATCGACAT GAATCCTGAG
GCTCTCACCG CGTGGAAACG TGCTGCCGCT GCTGTGTACC GCAAGGACAA GGCTCGCAAG TCTCGCCGTA
TCAGCCTTGA GTTCATGCTT GAGCAAGCCA ATAAGTTTGC TAACCATAAG GCCATCTGGT TCCCTTACAA
CATGGACTGG CGCGGTCGTG TTTACGCTGT GTCAATGTTC AACCCGCAAG GTAACGATAT GACCAAAGGA
CTGCTTACGC TGGCGAAAGG TAAACCAATC GGTAAGGAAG GTTACTACTG GCTGAAAATC CACGGTGCAA
ACTGTGCGGG TGTCGATAAG GTTCCGTTCC CTGAGCGCAT CAAGTTCATT GAGGAAAACC ACGAGAACAT
CATGGCTTGC GCTAAGTCTC CACTGGAGAA CACTTGGTGG GCTGAGCAAG ATTCTCCGTT CTGCTTCCTT
GCGTTCTGCT TTGAGTACGC TGGGGTACAG CACCACGGCC TGAGCTATAA CTGCTCCCTT CCGCTGGCGT
TTGACGGGTC TTGCTCTGGC ATCCAGCACT TCTCCGCGAT GCTCCGAGAT GAGGTAGGTG GTCGCGCGGT
TAACTTGCTT CCTAGTGAAA CCGTTCAGGA CATCTACGGG ATTGTTGCTA AGAAAGTCAA CGAGATTCTA
CAAGCAGACG CAATCAATGG GACCGATAAC GAAGTAGTTA CCGTGACCGA TGAGAACACT GGTGAAATCT
CTGAGAAAGT CAAGCTGGGC ACTAAGGCAC TGGCTGGTCA ATGGCTGGCT TACGGTGTTA CTCGCAGTGT
GACTAAGCGT TCAGTCATGA CGCTGGCTTA CGGGTCCAAA GAGTTCGGCT TCCGTCAACA AGTGCTGGAA
GATACCATTC AGCCAGCTAT TGATTCCGGC AAGGGTCTGA TGTTCACTCA GCCGAATCAG GCTGCTGGAT
ACATGGCTAA GCTGATTTGG GAATCTGTGA GCGTGACGGT GGTAGCTGCG GTTGAAGCAA TGAACTGGCT
TAAGTCTGCT GCTAAGCTGC TGGCTGCTGA GGTCAAAGAT AAGAAGACTG GAGAGATTCT TCGCAAGCGT
TGCGCTGTGC ATTGGGTAAC TCCTGATGGT TTCCCTGTGT GGCAGGAATA CAAGAAGCCT ATTCAGACGC
GCTTGAACCT GATGTTCCTC GGTCAGTTCC GCTTACAGCC TACCATTAAC ACCAACAAAG ATAGCGAGAT
TGATGCACAC AAACAGGAGT CTGGTATCGC TCCTAACTTT GTACACAGCC AAGACGGTAG CCACCTTCGT
AAGACTGTAG TGTGGGCACA CGAGAAGTAC GGAATCGAAT CTTTTGCACT GATTCACGAC TCCTTCGGTA
CGATTCCGGC TGACGCTGCG AACCTGTTCA AAGCAGTGCG CGAAACTATG GTTGACACTT ATGAGTCTTG
TGATGTACTG GCTGATTTCT ACGACCAGTT CGCTGACCAG TTGCACGAGT CTCAATTGGA CAAAATGCCA
GCACTTCCGG CTAAAGGTAA CTTGAACCTC CGTGACATCT TAGAGTCGGA CTTCGCGTTC GCGTAAAGAT
CTTACCGTTC GTATAATGTA TGCTATACGA AGTTATGAGC TGTTGACAAT TAATCATCGG CTCGTATAAT

GTGTGGGCAA TGAGCTTGCA CTGCAGAACT TTCTCGAGGA TATACCATGG AGAAAAAAAT CACTGGATAT
ACCACCGTTG ATATATCCCA ATGGCATCGT AAAGAACATT TTGAGGCATT TCAGTCAGTT GCTCAATGTA
CCTATAACCA GACCGTTCAG CTGGATATTA CGGCCTTTTT AAAGACCGTA AAGAAAAATA AGCACAAGTT
TTATCCGGCC TTTATTCACA TTCTTGCCCG CCTGATGAAT GCTCATCCGG AATTCCGTAT GGCAATGAAA
GACGGTGAGC TGGTGATATG GGATAGTGTT CACCCTTGTT ACACCGTTTT CCATGAGCAA ACTGAAACGT
TTTCATCGCT CTGGAGTGAA TACCACGACG ATTTCCGGCA GTTTCTACAC ATATATTCGC AAGATGTGGC
GTGTTACGGT GAAAACCTGG CCTATTTCCC TAAAGGGTTT ATTGAGAATA TGTTTTTCGT CTCAGCCAAT
CCCTGGGTGA GTTTCACCAG TTTTGATTTA AACGTGGCCA ATATGGACAA CTTCTTCGCC CCCGTTTTCA
CTATGGGCAA ATATTATACG CAAGGCGACA AGGTGCTGAT GCCGCTGGCG ATTCAGGTTC ATCATGCCGT
TTGTGATGGC TTCCATGTCG GCAGAATGCT TAATGAATTA CAACAGTACT GCGATGAGTG GCAGGGCGGG
GCGTAAATAA CTTCGTATAA TGTATGCTAT ACGAACGGTA GCGGCCGCTC AGATCCTTCC GTATTTAGCC
AGTATGTTCT CTAGTGTGGT TCGTTGTTTT TGCGTGAGCC ATGAGAACGA ACCATTGAGA TCATACTTAC
TTTGCATGTC ACTCAAAAAT TTTGCCTCAA AACTGGTGAG CTGAATTTTT GCAGTTAAAG CATCGTGTAG
TGTTTTTCTT AGTCCGTTAT GTAGGTAGGA ATCTGATGTA ATGGTTGTTG GTATTTTGTC ACCATTCATT
TTTATCTGGT TGTTCTCAAG TTCGGTTACG AGATCCATTT GTCTATCTAG TTCAACTTGG AAAATCAACG
TATCAGTCGG GCGGCCTCGC TTATCAACCA CCAATTTCAT ATTGCTGTAA GTGTTTAAAT CTTTACTTAT
TGGTTTCAAA ACCCATTGGT TAAGCCTTTT AAACTCATGG TAGTTATTTT CAAGCATTAA CATGAACTTA
AATTCATCAA GGCTAATCTC TATATTTGCC TTGTGAGTTT TCTTTTGTGT TAGTTCTTTT AATAACCACT
CATAAATCCT CATAGAGTAT TTGTTTTCAA AAGACTTAAC ATGTTCCAGA TTATATTTTA TGAATTTTTT
TAACTGGAAA AGATAAGGCA ATATCTCTTC ACTAAAAACT AATTCTAATT TTTCGCTTGA GAACTTGGCA
TAGTTTGTCC ACTGGAAAAT CTCAAAGCCT TTAACCAAAG GATTCCTGAT TTCCACAGTT CTCGTCATCA
GCTCTCTGGT TGCTTTAGCT AATACACCAT AAGCATTTTC CCTACTGATG TTCATCATCT GAGCGTATTG
GTTATAAGTG AACGATACCG TCCGTTCTTT CCTTGTAGGG TTTTCAATCG TGGGGTTGAG TAGTGCCACA
CAGCATAAAA TTAGCTTGGT TTCATGCTCC GTTAAGTCAT AGCGACTAAT CGCTAGTTCA TTTGCTTTGA
AAACAACTAA TTCAGACATA CATCTCAATT GGTCTAGGTG ATTTTAATCA CTATACCAAT TGAGATGGGC
TAGTCAATGA TAATTACTAG TCCTTTTCCC GGGAGATCTG GGTATCTGTA AATTCTGCTA GACCTTTGCT
GGAAAACTTG TAAATTCTGC TAGACCCTCT GTAAATTCCG CTAGACCTTT GTGTGTTTTT TTTGTTTATA
TTCAAGTGGT TATAATTTAT AGAATAAAGA AAGAATAAAA AAAGATAAAA AGAATAGATC CCAGCCCTGT
GTATAACTCA CTACTTTAGT CAGTTCCGCA GTATTACAAA AGGATGTCGC AAACGCTGTT TGCTCCTCTA
CAAAACAGAC CTTAAAACCC TAAAGGCTTA AGTAGCACCC TCGCAAGCTC GGGCAAATCG CTGAATATTC
CTTTTGTCTC CGACCATCAG GCACCTGAGT CGCTGTCTTT TTCGTGACAT TCAGTTCGCT GCGCTCACGG
CTCTGGCAGT GAATGGGGGT AAATGGCACT ACAGGCGCCT TTTATGGATT CATGCAAGGA AACTACCCAT
AATACAAGAA AAGCCCGTCA CGGGCTTCTC AGGGCGTTTT ATGGCGGGTC TGCTATGTGG TGCTATCTGA
CTTTTTGCTG TTCAGCAGTT CCTGCCCTCT GATTTTCCAG TCTGACCACT TCGGATTATC CCGTGACAGG
TCATTCAGAC TGGCTAATGC ACCCAGTAAG GCAGCGGTAT CATCAACAGG CTTACCCGTC TTACGTGCCG
ATCAAGTCAA AAGCCTCCGG TCGGAGGCTT TTGACTTTCT GCTATGGAGG TCAGGTATGA TTTTACT

Supplementary sequence 4.2. pSC042a

TCTAGATTTC AGTGCAATTT ATCTCTTCAA ATGTAGCACC TGAAGTCAGC CCCATACGAT ATAAGTTGTA
ATTCGGTACC CCGCTTCGGC GGGGTTTTTT CAAGTAATAC GACTCACTAT AGGGAGACCA CAACGGTTTC
CCATATCGTC GCAGCCCACA GCAACACGTT TCCTGATAAC ATATGAAACA GTATTTAGAA CTGATGCAAA
AAGTGCTCGA CGAAGGCACA CAGAAAAACG ACCGTACCGG AACCGGAACG CTTTCCATTT TTGGTCATCA
GATGCGTTTT AACCTGCAAG ATGGATTCCC GCTGGTGACA ACTAAACGTT GCCACCTTAG GTTAATTGAG
GCCTGAGTAT AAGGTGACTT ATACTTGTAA TCTATCTAAA CGGGGAACCT CTCTAGTAGA CAATCCCGTG
CTAAATTGTC ATCGCTACAA GGAGATATAA TTAATCACGA CAGGGCCCGC GATGACATAA ATGCCTAACG
ACTATCCCTT TGGGGAGTAG GGTCAAGTGA CTCGAAACGA TAGACAACTT GCTTTAACAA GTTGGAGATA

TAGTCTGCTC TGCATGGTGA CATGCAGCTG GATATAATTC CGGGGTAAGA TTAACGACCT TATCTGAACA
TAATGCTATA ATACATGAAC TGCTGTGGTT TCTGCAGGGC GACACTAACA TTGCTTATCT ACACGAAAAC
AATGTCACCA TCTGGGACGA ATGGGCCGAT GAAAACGGCG ACCTCGGGCC AGTGTATGGT AAACAGTGGC
GCGCCTGGCC AACGCCAGAT GGTCGTCATA TTGACCAGAT CACTACGGTA CTGAACCAGC TGAAAAACGA
CCCGGATTCG CGCCGCATTA TTGTTTCAGC GTGGAACGTA GGCGAACTGG ATAAAATGGC GCTGGCACCG
TGCCATGCAT TCTTCCAGTT CTATGTGGCA GACGGCAAAC TCTCTTGCCA GCTTTATCAG CGCTCCTGTG
ATGTTTTCTT GGGTTAATTG AGGCCTGAGT ATAAGGTGAC TTATACTTGT AATCTATCTA AACGGGGAAC
CTCTCTAGTA GACAATCCCG TGCTAAATTG TCATCGCTAC AAGGAGATAT AATTAATCAC GACAGGGCCC
GCGATGACAT AAATGCCTAA CGACTATCCC TTTGGGGAGT AGGGTCAAGT GACTCGAAAC GATAGACAAC
TTGCTTTAAC AAGTTGGAGA TATAGTCTGC TCTGCATGGT GACATGCAGC TGGATATAAT TCCGGGGTAA
GATTAACGAC CTTATCTGAA CATAATGCTA CCGTTTAATA TTGCCAGCTA CGCGTTATTG GTGCACATGA
TGGCGCAGCA GTGCGATCTG GAAGTGGGTG ATTTTGTCTG GACCGGTGGC GACACGCATC TGTACAGCAA
CCACATGGAT CAAACTCATC TGCAATTAAG CCGCGAACCG CGTCCGCTGC CGAAGTTGAT TATCAAACGT
AAACCCGAAT CCATCTTCGA CTACCGTTTC GAAGACTTTG AGATTGAAGG CTACGATCCG CATCCGGGCA
TTAAAGCGCC GGTGGCTATC TAAACTAGAA ATAATTTTGT TTAACTATAA GAAGGAGATA TACATATGAG
TAAAGGAGAA GAACTTTTCA CTGGAGTTGT CCCAATTCTT GTTGAATTAG ATGGTGATGT TAATGGGCAC
AAATTTTCTG TCAGTGGAGA GGGTGAAGGT GATGCAACAT ACGGAAAACT TACCCTTAAA TTTATTTGCA
CTACTGGAAA ACTACCTGTT CCATGGCCAA CACTTGTCAC TACTTTCTCT TATGGTGTTC AATGCTTTTC
CCGTTATCCG GATCACATGA AACGGCATGA CTTTTTCAAG AGTGCCATGC CCGAAGGTTA TGTACAGGAA
CGCACTATAT CTTTCAAAGA TGACGGGAAC TACAAGACGC GTGCTGAAGT CAAGTTTGAA GGTGATACCC
TTGTTAATCG TATCGAGTTA AAAGGTATTG ATTTTAAAGA AGATGGAAAC ATTCTCGGAC ACAAACTGGA
GTACAACTAT AACTCACACA ATGTATACAT CACGGCAGAC AAACAAAAGA ATGGAATCAA AGCTAACTTC
AAAATTCGCC ACAACATTGA AGATGGATCC GTTCAACTAG CAGACCATTA TCAACAAAAT ACTCCAATTG
GCGATGGCCC TGTCCTTTTA CCAGACAACC ATTACCTGTC GACACAATCT GCCCTTTCGA AAGATCCCAA
CGAAAAGCGT GACCACATGG TCCTTCTTGA GTTTGTAACT GCTGCTGGGA TTACACATGG CATGGATGAG
CTCTACAAAT AAACTAGTGA TCCGGCTGCT AACAAAGCCC GAAAGGAAGC TGAGTTGGCT GCTGCCACCG
CTGAGCAATA ACTAGCATAA CCCCTTGGGG CCTCTAAACG GGTCTTGAGG GGTTTTTTGC TGAAAGGAGG
AACTATATCT AGTGCAAGTG GCACTTTTCG GGGAAATGTG CGCGGAACCC CTATTTGTTT ATTTTTCTAA
ATACATTCAA ATATGTATCC GCTCATGAGA TTATCAAAAA GGATCTTCAC CTAGATCCTT TTAAATTAAA
AATGAAGTTT TAAATCAATC TAAAGTATAT ATGAGTAAAC TTGGTCTGAC AGTTACCAAT GCTTAATCAG
TGAGGCACCT ATCTCAGCGA TCTGTCTATT TCGTTCATCC ATAGTTGCCT GACTCCCCGT CGTGTAGATA
ACTACGATAC GGGAGGGCTT ACCATCTGGC CCCAGTGCTG CAATGATACC GCGAGACCCA CGCTCACCGG
CTCCAGATTT ATCAGCAATA AACCAGCCAG CCGGAAGGGC CGAGCGCAGA AGTGGTCCTG CAACTTTATC
CGCCTCCATC CAGTCTATTA ATTGTTGCCG GGAAGCTAGA GTAAGTAGTT CGCCAGTTAA TAGTTTGCGC
AACGTTGTTG CCATTGCTAC AGGCATCGTG GTGTCACGCT CGTCGTTTGG TATGGCTTCA TTCAGCTCCG
GTTCCCAACG ATCAAGGCGA GTTACATGAT CCCCCATGTT GTGCAAAAAA GCGGTTAGCT CCTTCGGTCC
TCCGATCGTT GTCAGAAGTA AGTTGGCCGC AGTGTTATCA CTCATGGTTA TGGCAGCACT GCATAATTCT
CTTACTGTCA TGCCATCCGT AAGATGCTTT TCTGTGACTG GTGAGTACTC AACCAAGTCA TTCTGAGAAT
AGTGTATGCG GCGACCGAGT TGCTCTTGCC CGGCGTCAAT ACGGGATAAT ACCGCGCCAC ATAGCAGAAC
TTTAAAAGTG CTCATCATTG GAAAACGTTC TTCGGGGCGA AAACTCTCAA GGATCTTACC GCTGTTGAGA
TCCAGTTCGA TGTAACCCAC TCGTGCACCC AACTGATCTT CAGCATCTTT TACTTTCACC AGCGTTTCTG
GGTGAGCAAA AACAGGAAGG CAAAATGCCG CAAAAAAGGG AATAAGGGCG ACACGGAAAT GTTGAATACT
CATACTCTTC CTTTTTCAAT ATTATTGAAG CATTTATCAG GGTTATTGTC TCATGAGCGG ATACATATTT
GAATGTATTT AGAAAAATAA ACAAATAGGC TGTCCCTCCT GTTCAGCTAC TGACGGGGTG GTGCGTAACG
GCAAAAGCAC CGCCGGACAT CAGCGCTAGC GGAGTGTATA CTGGCTTACT ATGTTGGCAC TGATGAGGGT
GTCAGTGAAG TGCTTCATGT GGCAGGAGAA AAAAGGCTGC ACCGGTGCGT CAGCAGAATA TGTGATACAG

GATATATTCC GCTTCCTCGC TCACTGACTC GCTACGCTCG GTCGTTCGAC TGCGGCGAGC GGAAATGGCT
TACGAACGGG GCGGAGATTT CCTGGAAGAT GCCAGGAAGA TACTTAACAG GGAAGTGAGA GGGCCGCGGC
AAAGCCGTTT TTCCATAGGC TCCGCCCCCC TGACAAGCAT CACGAAATCT GACGCTCAAA TCAGTGGTGG
CGAAACCCGA CAGGACTATA AAGATACCAG GCGTTTCCCC CTGGCGGCTC CCTCGTGCGC TCTCCTGTTC
CTGCCTTTCG GTTTACCGGT GTCATTCCGC TGTTATGGCC GCGTTTGTCT CATTCCACGC CTGACACTCA
GTTCCGGGTA GGCAGTTCGC TCCAAGCTGG ACTGTATGCA CGAACCCCCC GTTCAGTCCG ACCGCTGCGC
CTTATCCGGT AACTATCGTC TTGAGTCCAA CCCGGAAAGA CATGCAAAAG CACCACTGGC AGCAGCCACT
GGTAATTGAT TTAGAGGAGT TAGTCTTGAA GTCATGCGCC GGTTAAGGCT AAACTGAAAG GACAAGTTTT
GGTGACTGCG CTCCTCCAAG CCAGTTACCT CGGTTCAAAG AGTTGGTAGC TCAGAGAACC TTCGAAAAAC
CGCCCTGCAA GGCGGTTTTT TCGTTTTCAG AGCAAGAGAT TACGCGCAGA CCAAAACGAT CTCAAGAAGA
TCATCTTATT AATCAGATAA AATATT

Supplementary table 4.1. Primer list

Name	Sequence
BG6768	TGNNNCCNGCATCGTCTTGATGCNNNNGGNNNCATAAATGCCTAACGACTATCCCTTTG
BG6243	[Phos]ATTTAGCACGGGATTGTCTACTAG
BG7741	TGNNACCAGCATCGTCTTGATGCCANNGGCNNCATAAATGCCTAACGACTATCCCTTTG
BG4510	TCTGGGCATATCGTCGCAGCCCACAGCAACACGTTTCCTGAGGAACCATGGGTGTCTTTTTTACC TGTTTGACC
BG4511	GGCGTCGGCTCTGGCAGGATGTTTCGTAATTAGATAGCCACCGGCGCTTTTTCCTACCTCTGGTG AAGGAGTTG

Chapter 5

Grafting of a citrulline aptamer onto the phage T4 *td* group I intron

Sjoerd C.A. Creutzburg, Servé W.M. Kengen, John van der Oost

Abstract

Biosensors can be used among other applications to monitor intracellular metabolites. This applies to both metabolites that the cell makes naturally and metabolites resulting from engineered enzyme(s). We made a first attempt to engineer the phage T4 *td* intron into a citrulline responsive aptazyme. Since citrulline is a natural *E. coli* metabolite, the pathway generating citrulline was knocked out. A library of intron/aptamer combinations was made with a set of randomised communication modules. This library was then screened for functional variants using a previously developed cascade system based on RNAP, containing the intron-library and *thyA*, enabling selection and counter-selection. After the sequencing revealed that a sole communication module had taken over the enrichment, the aptamer was further refined. As proof of concept that the newly formed aptazyme can discriminate between cells harbouring a functional and dysfunctional citrulline synthesis pathway, the aptazyme was cloned into *lacZ*. Although further refinement of the aptazyme is required, the difference in LacZ activity confirms that the aptazyme indeed responded to the activity of the citrulline producing enzyme ArgF.

Introduction

In vivo biosensors are used to indicate the presence of a certain compound by induction of a reporter gene, resulting in a detectable signal. For many natural metabolites, like sugars and amino acids, natural sensing molecules (protein or riboswitch) already exist. However, when aiming for monitoring of a xenobiotic compound, the natural biosensors often fail to recognise and detect it. In this case, novel synthetic biosensors need to be engineered. The biosensor can be protein or riboswitch based. Proteins can bind with high affinity, but cannot be easily engineered towards binding completely unrelated compounds. This is different for riboswitches, where a new RNA molecule capable of binding the compound can be selected *in vitro* from a large library of variants. However, since the binding of the compound does not necessarily translate to a conformational change in the connected platform, a communication module should be inserted between them. This module comprises a set of random nucleotides, which generates a library of variants that can be screened or selected for the desired *in vivo* activity.

As a proof of principle that a synthetic riboswitch can discriminate between the presence of a functional and dysfunctional enzyme, we engineered a riboswitch that responds to L-citrulline. Citrulline is an intermediate of the arginine biosynthesis and is naturally produced in *E. coli* from ornithine via *argF* or *argI*,. The knockouts can easily be identified by growing with or without citrulline in the absence of arginine. An aptamer for L-citrulline has previously been generated *in vitro* by Famulok et al. (71). The aptamer needs to be attached to the riboswitch platform, which is the group I self-splicing intron in the phage T4 *td* gene. This group I intron has earlier been engineered to respond to theophylline and 3-methylxanthine (36). In the present study, the citrulline aptamer is therefore grafted onto the same site of the intron, the P6. The aptamer and platform are linked by a randomised communication module and selected by alternating rounds of *in vivo* selection and counter-selection using ThyA (thymidylate synthase). The newly formed aptazyme from the enrichment procedure was then inserted into the *E. coli lacZ* gene. Based on LacZ activity, one should be able to discriminate between the presence of a functional or dysfunctional *argF* gene, where the functional *argF* should yield higher LacZ activity.

Results

Library design and generation

The citrulline aptamer consists of two strands of RNA forming a stem on either side (Figure 5.1). To fit as an aptamer on a riboswitch, one of those sides will be attached to the platform, while the other side is capped with a hairpin. Since it is unknown in which orientation the aptamer needs to be grafted on the platform, two sets of libraries were

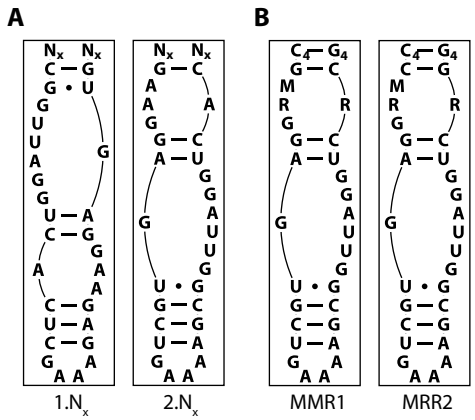


Figure 5.1. Design of the citrulline communication modules (Citrulline 1.Nx and 2.Nx) and aptamer libraries. (A) The grafting of the citrulline aptamer onto the intron was based on the position the aptamer has in the theophylline dependent aptazyme (36). Several variants of the citrulline aptamer were designed, each with a different size of communication module. Nx indicates the number of random nucleotides from 1 up to 4 (N1 - N4) between the intron platform and the citrulline aptamer. (B) The smaller MRR1 and MRR2 libraries were designed based on the enrichment of the Nx libraries and have the orientation of Citrulline 2.N4

designed. The libraries are either orientation 1 or 2 and contain a communication module consisting of one to four nucleotides between the platform and the aptamer. Library Citrulline 1.N_1 would be orientation 1 and one random nucleotide pair spacing between platform and aptamer. Both strands of the aptamer have one nucleotide spacing so the total number of variants is 16. N_2, N_3 and N_4 would have 256, 4×10^3 and 7×10^4 theoretical variants, respectively. The colony forming unit (cfu) count of the largest libraries were estimated by plating. The library size of both the 1.N_4 and 2.N_4 libraries was estimated to be 2×10^7 cfu. Libraries are combined into two enrichment starting cultures with all N_1 and N_2 variants combined, and all N_3 and N_4 variants as well.

Enrichment of the citrulline grafts

Since we want to generate a riboswitch, induced cells should activate the functional expression of the gene of interest, while non-induced cells should not. We set up a selection and counter-selection method with a T7 RNA polymerase (RNAP) cascade (see chapter 4). The RNAP is interrupted by a phage T4 *td* intron, of which the splicing rate should be dependent on an inducer compound. The T7 polymerase in turn controls the expression of two marker genes. The *thyA* gene is used for selection and counter-selection, while *gfpuv* is used for monitoring of enrichment progress and quantification of the response to the inducer compound.

The ThyA enzyme converts deoxyuridine monophosphate (dUMP) into deoxythymidine monophosphate (dTMP), which is an essential metabolite for DNA synthesis. dTMP can be synthesised from dUMP, but also from thymidine (dT). Thymidine is reversibly converted to dTMP and irreversibly to thymine. ThyA also requires 5, 10-methylene-THF for the reaction, which is part of the folate cycle. 5, 10-methylene-THF is produced via a cascade starting at DHF which is converted to THF by FolA. Trimethoprim (TMP) is a FolA inhibitor and thereby blocks the majority of synthesis of all THF related compounds. If FolA is inhibited, the ThyA will deplete the THF metabolite pool and the depletion will disturb

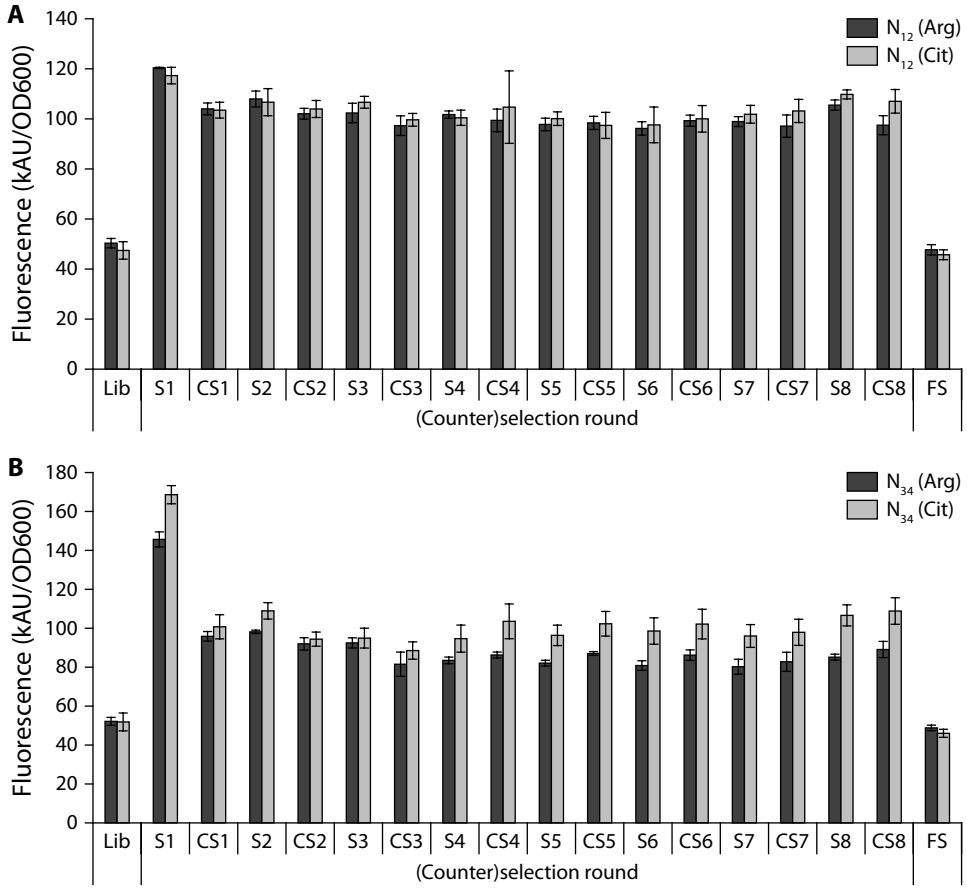


Figure 5.2. Enrichment of the citrulline communication module libraries. The aptamer is grafted onto the platform in two orientations and have 1 to 4 nucleotides stem length. The two orientations are mixed 1:1 and the stem lengths 1 and 2 (A), as well as 3 and 4 (B) are mixed 1:16. Starting at S(election round) 1 the cultures were enriched to C(ounter) S(election round) 8. The enrichment of the N1.2 libraries is not very successful, while the $N_{3,4}$ libraries start separating arginine (arg) from citrulline (cit) at CS3 or S4. The latter enrichment does not advance distinguishing between arginine and citrulline, having a similar pattern from CS4 onward.

the function of genes relying on those metabolites, like fmt. During the counter-selection, FolA is inhibited, but not FolM. FolM carries out a similar function as FolA, but is not as effective. FolM is strong enough to keep the folate cycle running when bacterial growth dilutes the metabolites, but it cannot supply enough THF to counter the depletion by ThyA. Therefore, bacteria not expressing ThyA can grow in the presence of TMP and dT and bacteria that do express ThyA cannot.

The aptamer we tried to graft onto the phage T4 *td* intron has been reported to bind to citrulline. Citrulline is an essential metabolite in the arginine synthesis. Provided there is no

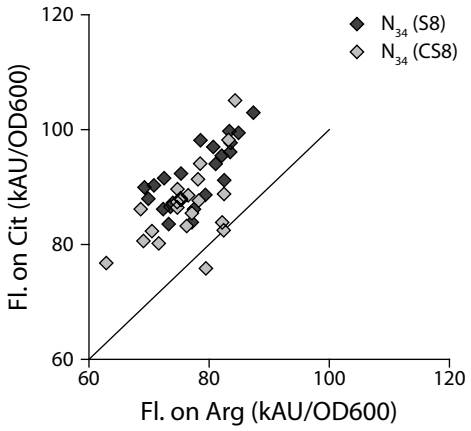


Figure 5.3. Response to arginine and citrulline of N3.4 S8 and N3.4 CS8 individual clones. S(election round) 8 (dark) and C(ounter) S(election round) 8 (light) cluster together for the most part. The clones that had a fluorescence citrulline:arginine ratio of at least 1.25 and a two population Z-score of over 2 were selected and sequenced.

other source of arginine, knocking out both *argF* and *argI* will be lethal. Supplying citrulline in the medium will allow for growth and if the bacteria grow, the citrulline must have entered the cell. Without intracellular citrulline, there will be no activation of a riboswitch, and in this way the bacteria require the uptake of extracellular citrulline. During counter-selection, the citrulline will cause the bacteria that do respond to the compound to have a fitness disadvantage. Therefore, the citrulline must have been metabolised at the end of the selection phase. Counter-selection is performed with arginine, which is required not to have an inductive effect on the riboswitch. Preventing carry-over of compounds is not as imperative for the selection phase. If there is a preference for the uptake of arginine over citrulline, synthesis of ThyA is delayed. While this will become a problem eventually, as Δ*thyA* mutants cannot grow in the absence of both dT and ThyA, the cells can last a while on the dT that is passed from the counter-selection phase. Arginine is made limiting for growth and TMP does not cause toxicity at a concentration of 5 µg/L.

Starting with a round of selection on citrulline the enrichments were alternatingly subjected to selection and counter-selection until round 8. The progress was monitored by total fluorescence (Figure 5.2). Since the RNAP controls both *thyA* and *gfpuv*, GFPuv fluorescence is an indication of the intron splice rate. By culturing on either citrulline or arginine one can estimate the progress of the enrichment. If bacteria harbouring a functional riboswitch have become a significant fraction of the culture, the fluorescence on citrulline will exceed the fluorescence on arginine. After the plating of the selection round 8 and the counter-selection round 8, 22 individual clones were picked from the enriched N3.4 library and tested for their response to citrulline (Figure 5.3). The clones from the N3 + N4 enrichment showed very similar fluorescence and a consistent, though small, response to citrulline. Clones with a fold-change over 1.25 and a Z-score from a standard two-sample z-test over 2 were sequenced. Sequencing revealed that all clones were in fact the same. The communication module of the enriched clone consisted of 5′ – CCCC – aptamer orientation 2 – GGGG – 3′.

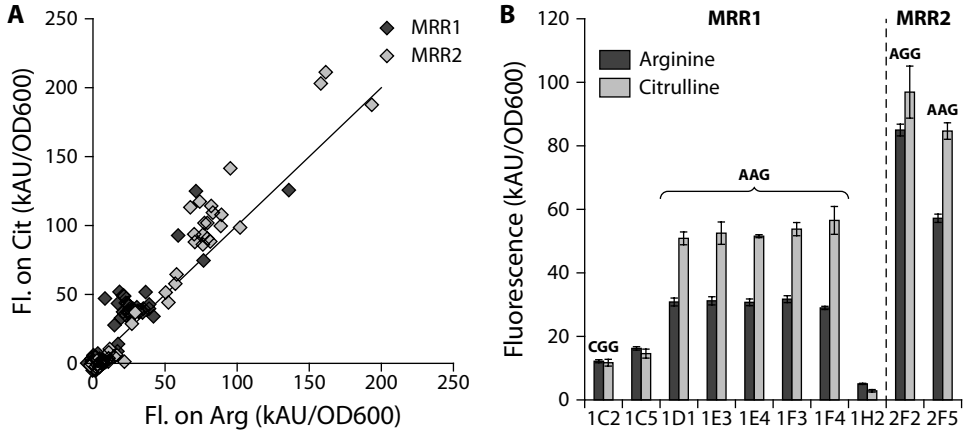


Figure 5.4. Response of citrulline MRR1 and MRR2 individual clones to citrulline and arginine. (A) The MRR1 library (C4G-MR//R-CG4) shows response for nearly all clones, while the MRR2 library (C4C-MR//R-GG4) does for a couple only. The 10 clones with the highest fluorescence cit/arg ratio were reassessed in triplicate. (B) Assessment of the top scoring MRR clones. Fluorescence of the RNAP-FS control was subtracted from the fluorescence of the MRR clones, error bars indicate the SD of biological quadruplicates. AAG1 is the most prevalent and has the largest response to citrulline. AAG2 also shows a response to citrulline, but has a much larger background. Sequencing of the 1C5 and 1H2 clones had failed, but they do not show any response to citrulline.

Since the induction effect was small, but significant, we tried to further improve the aptamer. The aptamer itself does not have a completely fixed sequence (71). To enhance the response to citrulline, we made aptamer variants with orientation 2 and the CCCC/GGGG communication module. The aptamer would then look like cccc-SMR – ggagugcugaaaagcgguuagguc-RS-gggg. S is either G or C, M is A or C, and R is A or G. The aptamer library was split in GMR//RC (MRR1) and CMR//RG (MRR2). The libraries each contain 8 different variants, so selection is not necessary. The libraries were plated and 46 colonies were screened for their response to citrulline compared to arginine (Figure 5.4A). 10 colonies that performed the best were reassessed in triplicate (Figure 5.4B). The citrulline AAG1 aptamer (GAA//GC) showed the strongest response to citrulline.

Another riboswitch that showed a significant response had the AAG2 aptamer (CAA//GG). While both responded to citrulline, their uninduced splicing differs by close to a factor 2.

In the end we want the riboswitch to indicate the presence of a certain enzyme activity. In this case the enzyme activity would be the synthesis of citrulline from ornithine by ArgF (or ArgI). The riboswitch should then respond to intracellularly synthesised citrulline. To avoid citrulline being metabolised, the *argG* gene is knocked out. ArgG converts citrulline into argininosuccinate, which is processed further into arginine by ArgH. The complete arginine pathway is regulated by ArgR. Among others, it inhibits the production of ornithine, the precursor of citrulline. Knocking out this gene will ensure the presence of the citrulline

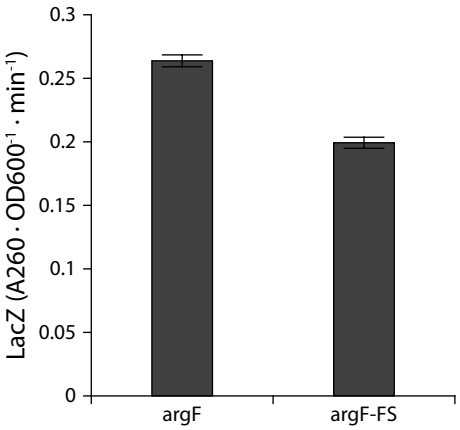


Figure 5.5. Response of the citrulline dependent *lacZ* to functional and dysfunctional ArgF. The *argF* gene produces citrulline from ornithine. Citrulline accumulates because of the knockout of *argG*. Citrulline induces splicing of the intron in the *lacZ* gene and causes more LacZ activity.

precursor. Changing from induction by citrulline in the medium to citrulline synthesis by ArgF would, however, mean a three-plasmid system for GFPuv measurements. To avoid the three-plasmid system and to show the riboswitch functions in another setting as well, the intron with the citrulline AAG1 aptamer/communication module was cloned into the *lacZ* gene of pSC046e. The pSC046e-citrulline AAG1 was co-transformed with either pRham-*argF* or pRham-*argF*-FS (with a frameshift in the *argF* gene). LacZ activity in an *E. coli* DH10B-Δ*argFGIR* background indicates the presence of ArgF.

Figure 5.5 shows the activity of LacZ with a functional or dysfunctional *argF* gene. The functional *argF* gene clearly causes higher LacZ activity than the dysfunctional *argF*, albeit by a small margin. The difference in LacZ activity indeed indicates that the riboswitch is active, but also that its induction is not yet sufficient to turn it into a screening or selection system for enzyme activity.

Discussion

Riboswitches are powerful tools for detecting small molecules, including metabolites. Contrary to protein sensors, the riboswitch can be adapted to bind a compound that has no structural similarity to its native ligand. The phage T4 *td* intron has no ligand of its own and has been adapted to bind theophylline and 3-methylxanthine. Obtaining new synthetic riboswitches, however, is still a challenging process. Binding of the ligand to an aptamer should result in a structural change of the ribozyme, which is required for the system to function as a riboswitch. We showed that a citrulline aptamer can be grafted onto the phage T4 *td* intron, yielding a functional riboswitch. The strategy using ThyA as selection and counter-selection tool is effective in selecting for improved riboswitches that do not have a strong response to the target ligand, citrulline in this case.

For a stronger response to citrulline, the aptamer had to be adapted. Either the adapted aptamer has a higher affinity for citrulline, or the aptamer/communication module

combination displays less ribozyme activity in the absence of citrulline. The fluorescence observed when grown on arginine is much higher for the AAG2 compared to AAG1 (Figure 5.4B). The lower background when switching the GC pair is caused by the riboswitch staying in the OFF state more often. Affinity for citrulline does not play a role here, so it has most likely to do with the balance between the OFF state and ON state. CGG1 is expected to be stuck in the active state, since it forms the strongest stem with six consecutive GC pairs when unbound to citrulline. However, the CGG1 is not responsive and has a very low background fluorescence, indicating it must be stuck in the OFF state instead. The citrulline aptamers drawn in (Figure 5.1B) are in fact in the OFF state, while shifting the base pairing along the stem causes the ON state. This is also why the AGG2 has such high background. The OFF state is stabilised by five consecutive GC pairs in the same orientation, while the shifted state has either four or five GC pairs depending on the way it is shifted. If the OFF state is just as stable as the ON state, the added value of binding citrulline is minimal. Shifting the stem in AAG1 will cost two GC base pairs and needs to be stabilised by citrulline.

To exclude that the increase in production of GFPuv could be derived from citrulline directly instead of via the riboswitch, the *argG* gene was knocked out. This knockout strain cannot process citrulline, so it must always be supplemented with arginine. The different environment of the intron, interrupting *lacZ*, and the inability to process citrulline, combined with the functional or dysfunctional ArgF showed that the intron responds to citrulline when it is made intracellularly. Although selection of ArgF analogues can be performed much more easily with auxotrophy complementation, a system like this can be used to trace intracellular enzyme activity or provide selective advantage for enzymes that do not perform a key role in the growth.

Materials and methods

Strains and media

Cloning and plasmid propagation was performed with *E. coli* DH10B T1R (Invitrogen: C6400-03). This strain was used as basis for knock-out strains *E. coli* DH10B-Δ*argFI*, DH10B-Δ*argFGI* and DH10B-Δ*argFI*-Δ*thyA*. Culturing was generally done in LB (10 g/L peptone, 5 g/L yeast extract, 10 g/L NaCl) at 37°C. If high density was required, M9TG medium was used (10 g/L tryptone, 1x M9 salts, 5 g/L glycerol). Antibiotics and other supplements were added when required: chloramphenicol (cam) (35 mg/L), ampicillin (amp) (100 mg/L), kanamycin (kan) (50 mg/L), trimethoprim (TMP) (5 mg/L), thymidine (dT) (100 mg/L), citrulline (cit) (50 mg/L), arginine (arg) (50 mg/L).

SSCM Medium composition

Fe(III) citrate	200	µM	$CaCl_2$	0.100	µM	Ile	0.035	g/L
$Na_2MoO_4.2H_2O$	10	µM	EDTA	0.010	µM	Leu	0.250	g/L
$CoCl_2.6H_2O$	10	µM	Ala	0.060	g/L	Lys	0.033	g/L
$MnCl_2.4H_2O$	75	µM	Arg	-	g/L	Met	0.017	g/L
$CuCl_2.2H2O$	10	µM	Asn	0.012	g/L	Phe	0.034	g/L
H_3BO_3	50	µM	Asp	0.050	g/L	Pro	0.117	g/L
$ZnAc_2.2H_2O$	150	µM	Cys	0.035	g/L	Ser	0.056	g/L
Glycerol ($C_3H_8O_3$)	100	µM	Glu	0.204	g/L	Thr	0.044	g/L
$(NH4)_2HPO_4$	30.3	µM	Gln	0.005	g/L	Trp	0.008	g/L
KH_2PO_4	97.7	µM	Gly	0.073	g/L	Tyr	0.013	g/L
Citric Acid	10	µM	His	0.020	g/L	Val	0.061	g/L
$MgSO_4$	5	µM						

The medium was supplemented with dT (40 mg/L) and TMP (5 mg/L), amp (50 mg/L), cam (17.5 mg/L), citrulline (50 mg/L) and arginine (50 mg/L) when necessary.

To exclude the production of citrulline during counter-selection, the *argF* and *argI* genes needed to be eliminated. A gene knock-out of *argF* and *argI* was made using the pSC020 plasmid in *E. coli* DH10B. This double knock-out strain was used to generate a strain with an additional *argG* knock-out (DH10B-Δ*argFGI*) and a strain with an additional *thyA* knock-out (DH10B-Δ*argFI*-Δ*thyA*). The *argFGI* knock-out was used in turn as a basis for a knock-out in argR to generate DH10B-Δ*argFGIR*.

The knockouts were generated according to the method of Datsenko and Wanner, 2000 (39) adapted to the pSC020 plasmid. The pSC020 contains both the λ-red genes and the cre recombinase, so the plasmid is not cured after homologous recombination. Instead, the bacteria are plated on medium selective for both the pSC020 plasmid and the integrated marker. The plasmid is cured only after the cre recombination has completed. Most of the time, the colonies obtained from this method are mixed. In some bacteria the cre recombination has occurred, while in others it has not. However, the colonies could only have been formed if the initial cell contained the integrated selection marker and, assumingly, the gene knockout.

The knockout of the genes leaves a scar of one broken lox site flanked by primer annealing sites on either side of about 25 bp. This creates several homologous regions. To minimise the chance of the integration taking place in the wrong location, two distinct knockout cassettes were used. PCR products were generated for integration replacing *argF* (primers BG6118 and BG6119, kanR cassette), *argI* (primers BG6772 and BG6773, kanR cassette), *argG* (primers BG6388 and BG6389, camR cassette), argR (primers BG8251 and BG8252, camR cassette) and *thyA* (primers BG4510 and BG4511, kanR cassette). Knockouts were

verified by PCR, sequencing and auxotrophy selection (Δ*thyA* and Δ*argFI*).

Generation of the citrulline aptamer library

E. coli DH10B-Δ*argFI*-Δ*thyA* cells were made competent by growing on 16 g/L peptone, 10 g/L yeast extract to an OD600 of 0.4. Cells were rapidly cooled on ice-water and harvested by centrifugation at 3000 g for 10 min. Three washes were applied: one culture volume of ice-cold ultrapure water and two times half a culture volume of ice-cold 10% glycerol in ultrapure water. Finally the cells were resuspended in a total volume of 1/250 culture volume of 10% glycerol and stored in aliquots at -80°C.

A "'round-the-horn" polymerase chain reaction (PCR) was performed with Q5 DNA polymerase (NEB) using pSC039b-Theo4 as a template. Reverse primer BG6243 was phosphorylated to facilitate the ligation further on. BG6422 to BG6430, the forward primers, generate a series of citrulline aptamer grafts. The PCR fragments were purified with the Zymo clean and concentrator kit, ligated with T4 ligase (NEB) and purified again. *E. coli* DH10B-Δ*argFI*-Δ*thyA* competent cells were transformed with an amount of ligation reaction dependent on the theoretical variant number of the library. The amounts of DNA were 50ng (N_1), 50ng (N_2), 100ng (N_3) and 500ng (N_4). The volume of competent cells also varied. Except for the N4 libraries, which had 200 µL of cell suspension transformed, the reactions used 40 µL. Transformation reactions were diluted 250x in M9TG supplemented with cam and amp. The more refined MRR libraries were made with primers BG8099 (MRR1) or BG8100 (MRR2) in combination with reverse primer BG6243. The transformation was done with 40 µL of *E. coli* DH10B-Δ*argFI*-Δ*thyA* cell suspension, 20 ng of ligation reaction and a final culture volume of 10 mL LB supplemented with cam, amp and dT.

Selection and counter-selection of citrulline aptamer grafts

After the library generation, several of the libraries were combined to reduce the number of cultures. The libraries Citrulline 1.N_1, 1.N_2, 2.N_1 and 2.N_2 were combined as were the libraries Citrulline 1.N_3, 1.N_4, 2.N_3 and 2.N_4. The libraries were mixed based on their theoretical variance. Since, for example, the 1.N_2 library has 16 times as many combinations of nucleotides compared to the 1.N_1 library, the libraries were mixed 1:16:1:16 (1.N_1:1.N_2:2.N_1:2.N_2 or 1.N_3:1.N_4:2.N_3:2.N_4).

To enrich for the functional citrulline riboswitches, the selection and counter-selection were alternated. This approach prevents the escape mutants from becoming dominant. Knocking out the *thyA* gene is highly beneficial during counter-selection, but is detrimental during selection. Constitutive expression of *thyA* causes the exact opposite effect. Strains having Δ*argFI* need either citrulline or arginine in the medium to survive. Selection was performed on SSCM supplemented with cam, amp and cit, and counter-selection on SSCM supplemented with cam, amp and arg. Cultures were diluted 1/1000

from the previous culture and incubated at 37°C overnight. 2.5h post inoculation, the counter-selection medium was supplemented additionally with TMP.

GFPuv fluorescence assays

Monitoring of progress was performed by diluting the overnight cell culture 1/5000 in fresh SSCM supplemented with cam, amp, dT, and either arginine (150 mg/L) or citrulline (150 mg/L) and incubated at 37°C overnight. When the progress was sufficient, the enrichment cultures were plated on LB plates supplemented with cam, amp and dT. Individual clones were picked and cultured on SSCM supplemented with cam, amp, dT and arg. The cultures were incubated at 37°C overnight and used for inoculation of 200 µL SSCM supplemented with cam, amp, dT and either arginine (150 mg/L) or citrulline (150 mg/L) on a 96 wells plate (Greiner) covered with a gas permeable membrane. The plates were incubated at 37°C at 900 rpm overnight.

After the exposure to either citrulline or arginine, the cultures were resuspended in an equivalent volume of 50 mM Tris-HCl pH 7.5 and 100 µL was transferred to a µClear 96 wells plate (Greiner). Fluorescence was measured on a BioTek Synergy MX microplate reader with a gain of 75. Fluorescence was calculated as raw fluorescence per OD600 for 100 µL.

LacZ assays

As a proof of principle for indicating citrulline synthesis by the bacteria, the citrulline AAG1 aptamer was cloned into the intron of the *lacZ* gene of pSC046e. The *lacZ* gene is constitutively expressed under control of the lacUV5 promoter. The *E. coli argF* gene was cloned into pRham (Lucigen). To ensure the accumulation of citrulline, the assays were carried out in the *E. coli* DH10B-Δ*argFGIR* strain. This strain cannot produce citrulline, but does synthesise its precursor ornithine, and cannot convert citrulline to arginine. Resupplying the *argF* gene on a plasmid will cause the accumulation of citrulline. Since the strain cannot process citrulline, the supplementation of arginine is a necessity. Therefore, the culturing during the assays was carried out on LB medium supplemented with cam and kan. Bacteria carrying the pSC046e-citrulline AAG1 plasmid and either the pRham-*argF* or pRham-*argF*-FS were grown at 37°C overnight in 6-fold. The cultures were diluted 5x and 20 µL of culture was mixed with 80 µL of permeabilisation solution (100 mM Na_2HPO_4, 20 mM KCl, 2 mM $MgSO_4$, 0.8 g/L CTAB, 0.4 g/L sodium deoxycholate and 5.4 mL/L β-mercaptoethanol) and incubated at 30°C for 30 min. 600 µL of pre-warmed substrate solution (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 1 g/L o-nitrophenyl-β-D-galactopyranoside (ONPG) and 2.7 mL/L β-mercaptoethanol) was added and incubated at 30°C until sufficient colour had developed. 700 µL of stop solution (1 M Na_2CO_3) was added to quench the reaction. The reaction was filtered through a 0.2 µM filter and measured in a spectrophotometer at 420 nm in a 1 cm cuvette. Activity was calculated as:

$$LacZ = \frac{A_{420}}{t} \cdot \frac{Vtotal}{Vculture \cdot OD_{600}}$$

Chapter 6

Translational feed-forward and feed-back control

Sjoerd C.A. Creutzburg, Thijs Nieuwkoop, Thijmen Zegers, John van der Oost

Abstract

Genes that are co-expressed as polycistronic mRNAs are not necessarily translated with the same efficiency. Differences in protein synthesis in such cases generally are caused by distinct rates of translation initiation and/or elongation, which in turn are governed by their ribosome binding site, their codon usage, and/or their mRNA secondary structure. Translational coupling of downstream genes and their upstream counterparts is a well-established feed-forward phenomenon, in which the translation of an upstream cistron influences that of a downstream one. In contrast, we here describe different types of feed-back control of gene expression. First, we demonstrate that a downstream gene may influence the expression of an upstream gene. In addition, we show a major impact of the sequence of the 3′ UTR, including the spacer between the coding sequence and the terminator. Moreover, we show that the ratio between the translation of the genes in an operon is also dependent on the transcription rate. It is concluded that, even after half a century of intense research, the sequences of the translated and untranslated regions of genes and operons still have unpredictable impact on the relative rates of the transcription and translation processes, and hence are crucial determinants for the efficiency of gene expression. The here-presented results may contribute to elucidating the molecular basis of these phenomena, which is crucial for fundamental understanding as well as for applications that rely on operon design.

Introduction

Prokaryotes often generate polycistronic mRNA for the concerted transcription of functionally related genes, for instance for enzymes that compose metabolic pathways and for subunits of protein complexes (72, 73). While transcription of the genes clustered as an operon is generally equal, differences in translation rate may cause differential expression of these genes. Differential translation may arise from differences in ribosome binding efficiency, codon usage and impediments like strong secondary structures. Between successive genes in the operon, translational coupling has already been observed several decades ago (74–76). More recently, a systematic and quantitative characterisation of *E. coli* operons has shown that the expression of an upstream gene can influence the expression a downstream gene, depending on the length of the intergenic region (77). Increased translation of an upstream gene by incorporating a range of ribosomal binding sites (RBS), each having different ribosome binding strength, has a direct or indirect effect on the translation rate of the downstream genes. Furthermore, it was found that translational coupling is also affected by so-called polar mutations in the upstream gene's coding sequence (74). For instance, a point mutation in the upstream gene trpE affects the expression of the downstream trpD. Two main models have been proposed to explain this phenomenon.

The first model is based on ribosomal "flow-through". Ribosomes terminating translation at the upstream gene's stop codon are in the direct vicinity of the downstream gene's initiation sites, thus a direct feed-forward control may occur (78). This model was further confirmed by increasing the distance between the stop codon and the downstream start codon (77). A decrease of translational coupling was found by increasing the length of the intergenic region. The second model is based on the helicase activity of 70S ribosome. Increased translation of an upstream gene can dissolve secondary structures throughout the operon, potentially removing inhibitory structures present on the downstream gene's initiation region. Strong secondary structures upstream of the atpA gene's translation initiation region were indeed found to inhibit the translation rate of atpA (79). A model was proposed in which the secondary structure within the upstream atpH cistron is dissolved by the processive ribosome activity, also resulting in unfolding the downstream mRNA to improve accessibility for ribosomal binding to allow translation initiation of atpA. This translational feed-forward coupling phenomenon has important implications for designing operon synthesis as well as for operon reduction. The alterations of gene order and sequence not only affect translation rates of the altered gene but can also affect expression ratios throughout the operon, resulting in differential stoichiometries (24).

In this study, we explored the relation between local sequence mutations, secondary RNA structures and, consequently, gene expression levels throughout an operon and found a new form of coupling. We conclude that gene expression is not only influenced by

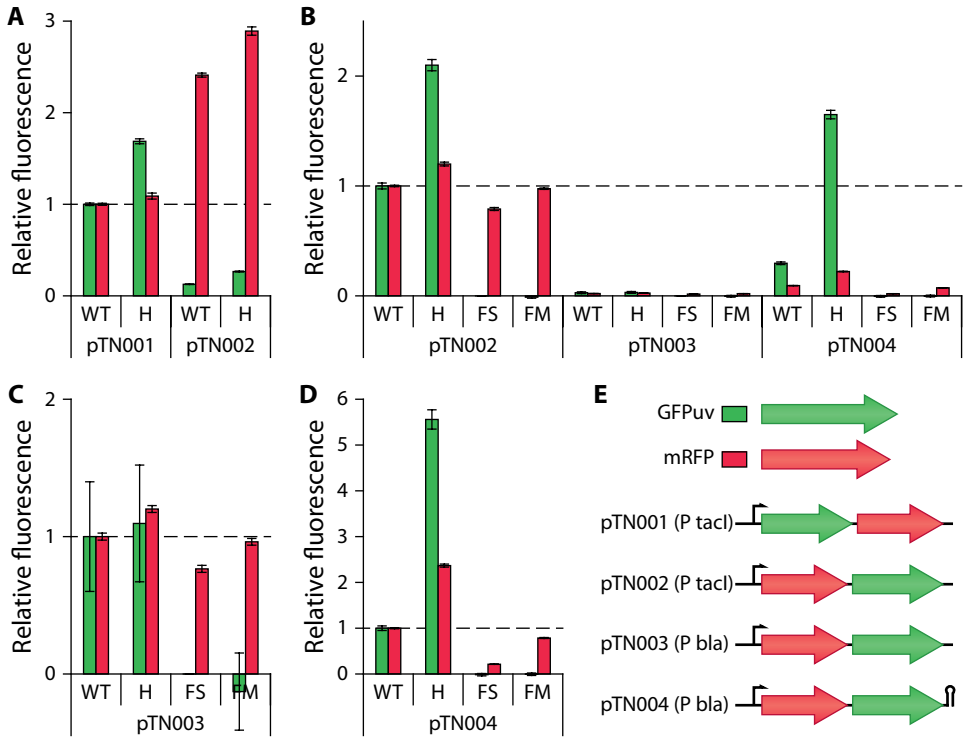


Figure 6.1. Coupled expression of RFP (red) and GFP (green) in the same operon. (A) A small increase in RFP can be seen with a better expression of GFP. pTN001 has the GFP in front of the RFP, while pTN002 has it reversed. The expression of both GFP and RFP is highly dependent on their respective position in the operon, but in all cases the wild-type GFP (WT) shows less fluorescence than the harmonised GPFuv (H). Error bars represent the standard deviation. (B) Compared to pTN002, the pTN003 and pTN004 have a weak promoter. The pTN004 has, in addition, a terminator behind the GFP. The weak promoter diminishes the expression of pTN003 severely and can be partially compensated for by addition of the terminator in pTN004. The terminator enhances stability of the mRNA, but not all genes profit from that to the same extend. The RFP/GFP ratio is vastly reduced in pTN004 compared to pTN002, while the H/WT ratio is increased. (C) pTN003 dataset of (B) normalised for pTN003-WT. (D) pTN004 dataset of (B) normalised for pTN004-WT. (E) Construct overview of the pTN001-4 plasmids. PtacI is about 17 times as strong as Pbla. Cloning a terminator behind pTN002 was unsuccessful.

upstream sequences as previously described (feed-forward control), but also vice versa by downstream sequences through feed-back control. Whilst keeping RBS sequences constant, differences in expression of the upstream gene can be observed when altering downstream gene sequences. The translation rate of an upstream monomeric red fluorescent protein (mRFP, hereafter called RFP) gene is altered when the downstream green fluorescent protein (GFPuv, hereafter called GFP) gene sequence is altered. The effect has been assayed of co-expressing a upstream gene encoding a single RFP variant with a downstream gene encoding one of four types of GFP (wild-type GFP, harmonised

GFP, frameshifted GFP and a functional mutant of GFP). To have a homogenous transcript size, a strong rho-independent terminator was added, which increased protein expression drastically. Moreover, substantial differences in expression were detected when the 30 bp linker between the stop codon and the terminator hairpin was varied (post stop, ante terminator region or PSAT region). The most obvious explanation for the observed fluctuations in gene expression is a change in stability of the corresponding mRNA, although the increase in protein expression does not match the increase in mRNA levels. Furthermore, the contributing effects of this region are not associated to a specific open reading frame sequence and therefore offer a new generic means of controlling gene expression.

RESULTS

Translational coupling occurs regardless of ORF order in the mRNA

The difference in codon usage between the wild-type (WT) and harmonised (H) GFP-encoding gene (80) causes a difference in expression level, where the harmonised gene is expressed significantly more than the wild-type. To allow accurate quantification of this effect, we cloned a monomeric red fluorescent protein (RFP) downstream the GFP variants as an internal control (Figure 6.1A; pTN001). The GFP fluorescence indeed shows that the harmonised GFP is better than the wild-type, but, unfortunately, the RFP fluorescence also fluctuates (1.09-fold; $p = 0.0487$), most likely reflecting translational coupling (feed-forward control). This implies that the *rfp*

gene in these constructs cannot be used as internal control. Reversing the order of the genes, *rfp* upstream and *gfp* downstream (Figure 6.1A; pTN002), RFP foremost lowers the expression of GFP in favour of RFP. However, instead of diminishing the translational coupling, expression of the *gfp* gene influences that of the *rfp* gene even more when it is located downstream the RFP. To further investigate this reciprocal translational coupling (feed-back control), a GFP frameshift mutant (FS) with a 4-nucleotide insertion half way, and a GFP functional mutant (FM) with a Tyr66Ser mutation that prevents formation of the fluorophore, were made (Figure 6.1B; pTN002). In addition to this set, in order to make all of the transcripts the same size, a strong synthetic terminator was inserted downstream of the GFP with a 45-nucleotide spacing sequence between the stop codon and first nucleotide of the terminator stem. This spacing sequence was generated by a random number generator and selected for approximately 50% GC content and lack of secondary structure as predicted by mFold (81). While the WT-GFP showed significantly higher GFP fluorescence, we failed to obtain correct clones with the terminator behind the H-GFP; we only obtained clones with mutations in the h-*gfp* coding region or in the terminator stem, and of a h-*gfp* gene disrupted by the insertion of a transposon. Given that the H-GFP has more expression than the WT-GFP and the terminator increases expression, the bacteria

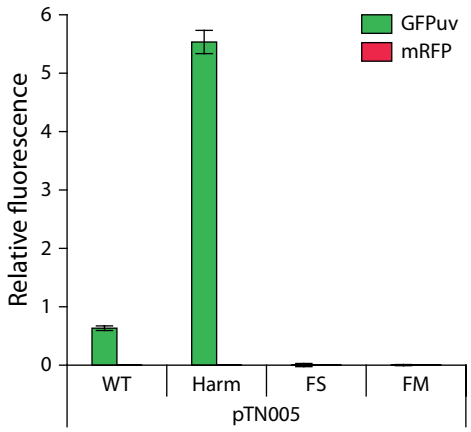


Figure 6.2. Fluorescence of GFP alone. The GFP variants are under control of the Pbla promoter and transcription is terminated by a terminator. Hence, fluorescence values are normalised for pTN004. Green fluorescence of the WT is less than in pTN004, while the others are roughly the same. Red fluorescence is totally absent, excluding the GFP contributing to red fluorescence in the other constructs.

most likely could not cope with the burden or internal GFP concentration in case of the H-GFP construct. Therefore, this set of constructs was abandoned.

A new set of constructs was made with, instead of the PtacI promoter, the weaker Pbla promoter (82) controlling the operon without terminator (pTN003) and with terminator (pTN004). In the absence of the terminator, the weak promoter diminishes the expression of both GFP and RFP to very low levels (Figure 6.1B; pTN003), while including the terminator (Figure 6.1B; pTN004) restores the H-GFP expression almost to the level of the PtacI promoter (Figure 6.1B; pTN002). When the pTN003 and pTN004 constructs are normalised to their respective WT-GFP constructs (Figure 6.1CD), it becomes clear that the interdependency of GFP, RFP and their surroundings is severe. Figure 6.1C shows the weak promoter without terminator. The large error bars in the GFP is because the expression of GFP is so low that the total fluorescence is only 10% above that of the frameshift variant (FS-GFP) and just 4x the fluorescence background of the medium. The RFP has no detectable auto-fluorescence from either the medium or the cells, so it can still be measured accurately at low expression levels. Since the expression of GFP (and RFP) is higher with the terminator, fluorescence of the pTN004 construct (Figure 6.1D) can be measured accurately.

Regardless of the constructs' promoter or presence of a terminator, the basic pattern for the translational coupling is the same. The RFP expression is highest when co-expressed with H-GFP, followed by WT-GFP, FM-GFP and lastly FS-GFP. This strongly suggests the occurrence of translational coupling independent of the gene order, either feed-forward (Figure 6.1A) or feed-back (Figure 6.1CD). The WT-GFP and FM-GFP have almost the same sequence, as is reflected in the expression level of RFP. The FS-GFP also has almost the same sequence as the WT-GFP, but the key difference is that the frameshift mutant yields a truncated GFP (half the size of the WT) due to a premature stop codon. The terminator appears to amplify the effect of translation efficiency of GFP and the coupled RFP. The

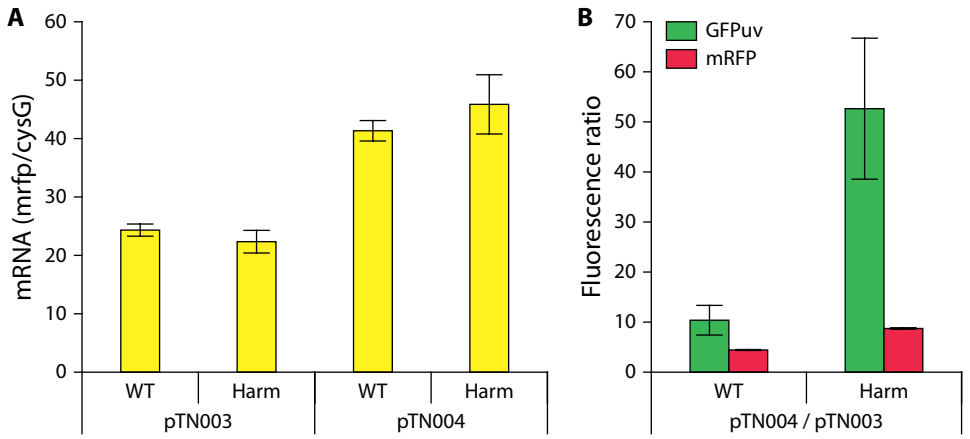


Figure 6.3. RT-qPCR data for pTN003 and pTN004 compared to the fluorescence. (A) mRNA abundance was estimated by qPCR on the RFP gene. The internal standard cysG (83) was used for normalisation. (B) Fluorescence ratio of pTN004 and pTN003.

pattern of GFP in pTN004 is more pronounced compared to pTN002 (Figure 6.1B); pTN002 H-GFP/WT-GFP is just over 2, while pTN004 H-GFP/WT-GFP is over 5 (Figure 6.1BD). GFP fluorescence of pTN003 (Figure 6.1C) is only 10% of the background fluorescence, so the H-GFP and WT-GFP are not significantly different. Comparing the RFP expression, while the actual RFP expression is vastly different, compared to their respective WT-GFP, the pTN002 constructs and pTN003 constructs are very similar (pTN002 - Figure 6.1B; pTN003 – Figure 6.1C; RFP). The constructs of pTN004 (Figure 6.1D) show amplification of that pattern. The amplification is possibly also true for the FM-GFP, but since its RFP expression similar to that of, perhaps slightly lower than, the WT-GFP, it is difficult to ascertain.

To exclude the possibility that the fluorescence at 607 nm (normally attributed to RFP) is influenced by GFP directly, we made the constructs with a Pbla promoter, GFP only and a terminator (pTN005). The GFP fluorescence is comparable to the GFP fluorescence in pTN004, while the fluorescence in the RFP spectrum cannot be detected at all (Figure 6.2), showing that GFP itself cannot be responsible for changes in the fluorescence at 607 nm.

The effect of the 3'UTR on expression

We then looked into the major effect on expression by the addition of a synthetic terminator. RT-qPCR analysis of the pTN003 and pTN004 constructs (Figure 6.3A) shows a clear stabilising effect of the terminator. The terminator increases the mRNA abundance during the mid-log phase by approximately a factor 2, regardless of codon use. Contrary, the increase in protein expression (measured as fluorescence) is significantly higher than that (Figure 6.3B), with clear influence on the codon usage. The H-GFP fluorescence increases by a factor 50 when adding a terminator, while the WT-GFP increases "only" a factor 10. The RFP increases by a factor 4 and 9 respectively. Since the codon use has no

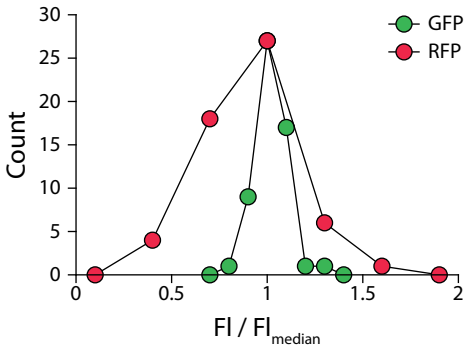


Figure 6.4. PSAT region library. Green indicates the PSAT sequences found for GFP, and red are the PSAT sequences for RFP.

significant effect on the mRNA concentration (Figure 6.3A), it is not likely that ribosome shielding through more efficient translation contributes to enhanced mRNA stability.

Next, we looked into a possible effect on expression by the 45-nucleotide post stop-codon, ante terminator (PSAT) region in the 3'UTR. To this end, a GFP and an RFP library were generated containing a completely randomised 30 bp PSAT region, replacing the 45 bp original. The length of PSAT is based on a recent study that reported that a sequence smaller than 30 nucleotides can have a negative effect on the terminator's termination efficiency, while increasing the size above 30 nucleotides did not show any effect (84). After randomisation and transformation of the plasmid libraries to *E. coli*, transformant cells with a range of fluorescence were obtained for both RFP (5.4-fold difference) and GFP (2.7-fold difference). The associated PSAT region sequences were obtained via Sanger sequencing (Figure 6.4, Supplementary table 6.2). A selected sequence set was interchanged between the two reporters to determine whether the effect of the PSAT sequence on the fluorescent level is protein specific (pTN006 series; Supplementary table 6.1). Five sequences with a representing fluorescent range were selected for both the RFP and the GFP library (Figure 6.5A, PSAT 1-10). The sequences were cloned behind the original CDS, to serve as a control, and behind the alternative CDS. In addition, both sets of five were cloned behind *lacZ*, for an independent verification (Figure 6.5A). In an attempt to elucidate which part of the sequence is responsible for the high translation efficiency of PSAT region 9, a series of truncations was made from both the 5' and 3' end. Truncation even to 15 bp from either side does not appear to change the translation efficiency by much. Only the PSAT region 12, which ends in CCC, lowers the expression of all reporter proteins, indicating that interactions with the terminator might play a role. Surprisingly, there is a very good correlation (Figure 6.5B-D) between the observed relative GFP, RFP and LacZ levels, suggesting that the effect of the PSAT sequence is not CDS dependent, but rather a generic phenomenon.

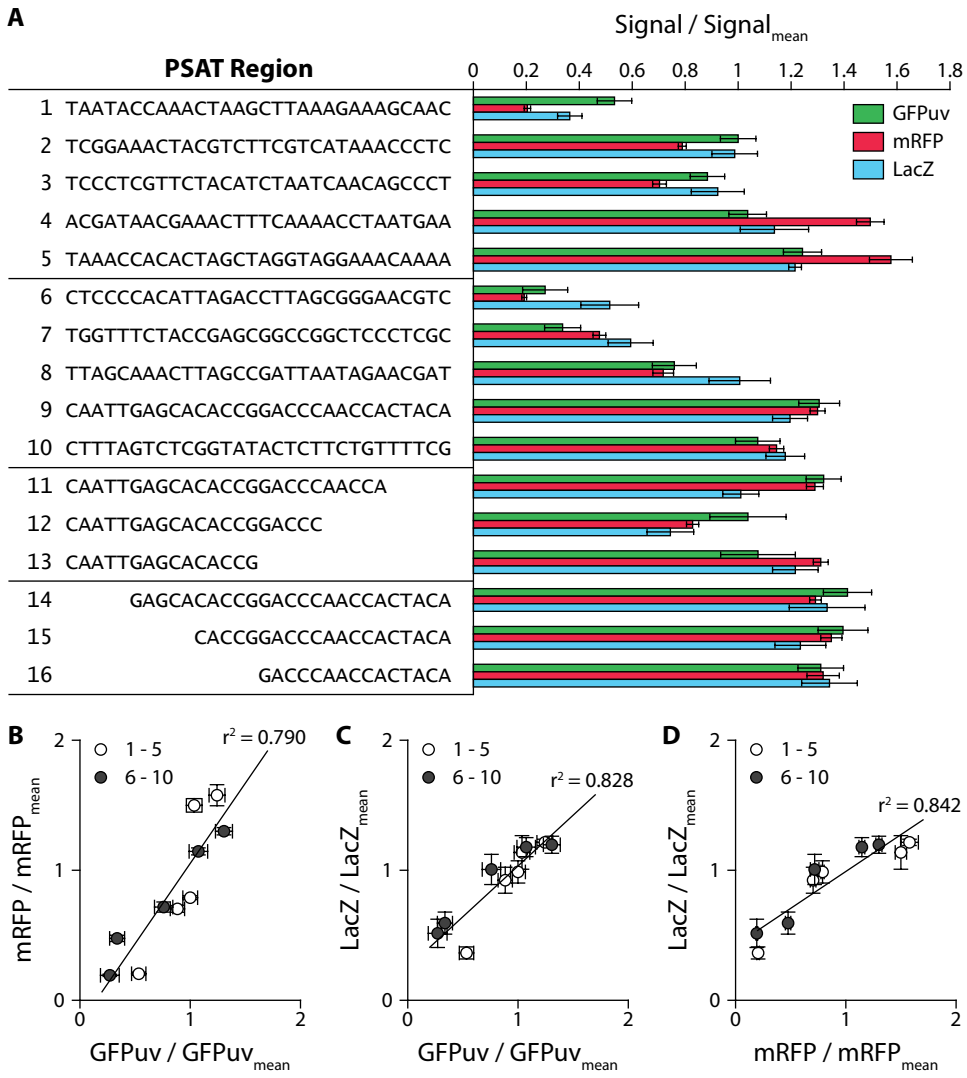


Figure 6.5. Selected PSAT sequences and truncations of PSAT-9 with GFP, RFP and LacZ. (A) PSAT sequence and relative signal. For GFP and RFP fluorescence was measured, for LacZ the hydrolysis of ONPG and the extinction at 420 nm. (B-D) Correlations between GFP/RFP, GFP/LacZ and RFP/LacZ. All have decent correlation, indicating that the effect the 3'UTR has on translation is mostly ORF independent.

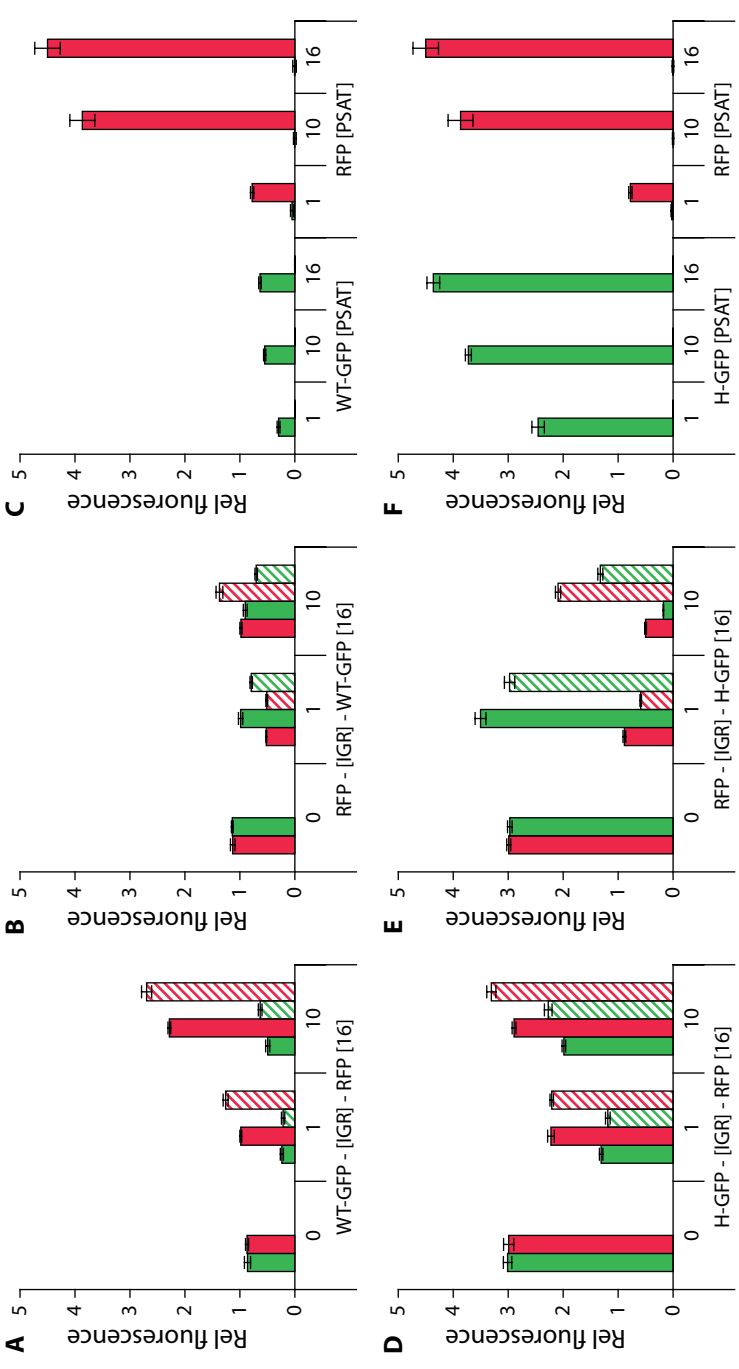


Figure 6.6. GFP and RFP fluorescence in an operon with different intergenic regions (IGRs). The numbers on the x axis (0, 1, 10, 16) indicate the PSAT region added to the IGR directly behind the stop codon (sequences as in Figure 6.5), where 0 is a control without any extra nucleotides. Solid: No terminator stem behind the PSAT region. Dashed: With terminator stem behind the PSAT region. Green: GFP fluorescence. Red: RFP fluorescence. All fluorescence has been normalised for the raw average of WT-GFP IGR-0. Error bars indicate the SD. (A) The WT-GFP followed by an IGR and RFP. (B) RFP followed by an IGR and WT-GFP. (C) Controls with either WT-GFP or RFP. (D) The H-GFP followed by an IGR and RFP. (E) RFP followed by an IGR and H-GFP. (F) Controls with either H-GFP or RFP.

Operon intergenic regions

The effect of the 3' UTR on translation may be interesting for tuning expression of the genes in operons. In addition to tuning through sequences at the 3' end of the operon, we set out to analyse the effect of the same PSAT sequences between the two coding sequences in the operon. Hence, a set of constructs was made with an intergenic region (IGR) either derived from PSAT region 1 (low translation; IGR-1) or PSAT region 10 (high translation; IGR-10) (Figure 6.5A; pTN007 series; Supplementary table 1). Since it is unclear whether the terminator is involved in the

aforementioned modulating effects, a version with and without terminator stem was designed. IGR-0 is the control without any addition to the intergenic region, and the 3'UTR is the same as in construct IGR-16. RFP, WT-GFP and H-GFP were used in different orders (Figure 6.6). The fluorescence was normalised for the average of both of the IGR-0 constructs (either a GFP variant or RFP in the first position). While GFP and RFP values can be compared, the values do not represent an equivalent in protein molecules.

Remarkably, when no PSAT is inserted into the IGR (IGR-0), the first position is no longer favoured (Figure 6.6AB; Figure 6.6DE). This is in sharp contrast to the operons as depicted in Figure 6.1A, where expression of the gene at the first position is highly favoured. An explanation for this discrepancy might be differences of transcription rates, where high transcription (tac promoter, pTN001/pTN002) might cause limited ribosome availability. In that case, the first gene can already be translated while the second gene is not even transcribed, causing relatively high expression the first gene in these operon constructs (Figure 6.1A), and much less so in case of the less efficient bla promoter (pTN004 and derivatives, Figure 6.1B-D, Figure 6.6). In case of the WT-GFP, the order does influence the total expression of GFP and RFP. For example, this may be caused by secondary structures around the ribosome binding site associated with *gfp*. Both IGR-1 and IGR-10 in WT-GFP-[IGR]-RFP (Figure 6.6A) show a discrepancy between the poorly translated WT-GFP and the more efficiently translated RFP, indicating loss of translational coupling. Including neither IGR-1 nor IGR-10 suggest a strong influence on the translational coupling, but the substantial impact it has on the overall translation indicates that the coupling does persist. On the other hand, the constructs in the reverse order (RFP – WT-GFP; Figure 6.6B) do not exhibit this behaviour. Introduction of IGR-1 has the predicted effect on RFP translation, but WT-GFP translation appears to be largely unaffected. Extending the IGR with the terminator stem lowers the translational coupling somewhat, but not nearly as much as is seen in the WT-GFP-RFP constructs. Interestingly, the WT-GFP translation in several of the operons exceeds the translation of only WT-GFP (compare to Figure 6.6C).

Using the H-GFP instead of WT-GFP (Figure 6.6DE) results in highly increased GFP fluorescence. The pattern of H-GFP – RFP is similar to the WT-GFP – RFP, but far less pronounced. In contrast, the RFP – H-GFP constructs have a more pronounced pattern. WT-

GFP translation acts as rate-limiting factor, so a poor IGR cannot attenuate the translation much further (Figure 6.6B). This is not the case for the H-GFP (Figure 6.6E), so the poor IGR1 becomes the limiting factor.

Discussion

We here describe that expression of genes in an operon can be coupled regardless of the order of those genes, either through feed-forward or feed-back control by both coding and non-coding sequences. The addition of a strong terminator at the 3'-end had a major influence on the expression of both GFP and RFP. Moreover, varying a spacer sequence (PSAT region) between the stop codon and terminator resulted in up to a 7.7-fold difference in expression, which appears to be largely ORF independent. Translational coupling has previously been observed in operons where the translation rate of the first gene influences the expression of at least two downstream genes (feed-forward) (77). It was explicitly stated that the accumulation of the downstream encoded protein had no influence on the upstream translation. In the literature, two models for translational coupling have been proposed. The first model is based on the disruption of secondary structures (79, 85). Translation initiation can be severely hampered by a secondary structure masking the RBS. The helicase activity of the ribosome can disrupt these structures, but only when it is already bound (86). If the coding region of gene A forms an RNA structure with the RBS of gene B, thereby preventing ribosome binding, the frequency of ribosomes passing ORF-1 influences the availability of the RBS for the translation initiation of ORF-2. In feed-forward translation coupling (first ORF-1, then ORF-2), the secondary structure consists of two parts that are in relatively close proximity. An extreme example of this principle is the bi-cistronic design, in which a strong first RBS attracts a ribosome, after which the ribosome translates a short peptide that overlaps with the second RBS. The second RBS is now available for ribosomes to bind (87). The second model for feed-forward translational coupling argues that the ribosome is in close proximity of the second RBS when it is released from the mRNA after the first gene has been translated. This model is supported by the introduction of a premature stop codon in the first gene diminishing the translational coupling, where the distance between the stop codon and the SD of the second gene is negatively correlated with the translational coupling (75, 77). A combination of both models appears to be in good agreement with feed-forward translation coupling observations.

Based on the here presented experimental data, it is tempting to propose a model as well for the feed-back translational coupling phenomena. The aforementioned 'proximity model' does not work for feed-back translational coupling. The downstream gene stop codon could be in the proximity of the upstream gene RBS if their respective 3' UTR and 5' UTR were to interact with each other, basically forming a loop between the beginning and end of the transcript. This is virtually impossible for the constructs in this study,

because the part of the 5′ UTR available for this interaction is short (28 nt) with a very low GC content (18%). We observed that the introduction of a premature stop codon – via a frameshift – also decreased the expression of the upstream gene. Barring the long-range interactions between the beginning and the end of the transcript, this observation seems to contradict this model.

A more likely hypothesis is that the H-GFP construct is less prone to degradation by the degradosome due to increased ribosome shielding (88–91). What determines the translation yield, is a combination of several factors during initiation (the RBS availability) and elongation (many factors, including the transcript's codon usage and the availability of matching charged tRNAs, and the transcript's secondary structure (23)). The combination of these factors may cause a huge difference in coverage of the mRNA by the ribosomes, up to a 100 fold (92, 93). In contrast, the effect of mRNA stability appears to be limited and the codon usage has no significant effect on it (Figure 6.3).

Addition of a strong terminator improves the mRNA stability, as it forms a stem loop which improves mRNA stability at the 3′ end protecting it from 3′-5′ exonuclease attacks (94), and has been shown to increase protein expression (2-fold (in *E. coli*) (95) and over 30-fold (HeLa cell line) (96)). We have found a rough approximation of 4- to 50-fold increase in protein expression with the addition of a terminator, but that improvement is much larger than the improvement in mRNA concentration (lower than 2-fold). This indicates that the terminator does not only improve the stability of the mRNA, but also improves the translation rate of the mRNA. This is corroborated by the fact that the H-GFP profits much more from the terminator than the WT-GFP does. If the lack of a terminator causes a bottleneck after ORF-2, ribosomes stack up into ORF-2, unable to progress, and then the effect of codon usage in ORF-2 is diminished. This is likely also the case in pTN001 and pTN002, where the number of ribosomes may be limiting the translation, due to the high transcription of the operon. The terminator releases the bottleneck and the ORFs that were most hampered by this bottleneck profit the most. It is hard to see how the terminator located after ORF-2 directly influences the translation of ORF-1, since the effect on mRNA stability is so limited. A possibility is that the translation of ORF-1 is directly influenced by the translation of ORF-2. The best explanation would be a continuous ribosome train. Normally, after peptide chain termination, the ribosome stays bound to the mRNA awaiting recycling (97). The ribosome is about 20 nm, while an RNA base spans about 0.34 nm. This means that the ribosome may cover as many as 60 nucleotides. The intergenic region is rather short (42 nt), so likely, the main source for ribosomes for ORF-2 is the recycled ribosomes of ORF-1 (proximity effect). On the other hand, a bottleneck in ORF-2 is directly transduced to ORF-1. The distances are so short that a ribosome in the process of being recycled can block the way of ribosomes in ORF-1. Ribosome profiling (93) may reveal whether there are bottlenecks after the stop codon of ORF-2 that are simply not transcribed because of the terminator.

Randomising the PSAT region resulted in variable fluorescence levels. A sequence that gives rise to high GFP expression also results in relatively high RFP expression and vice versa. Secondary structures between the ORF and the PSAT region are unlikely to be the reason for the variability in fluorescence. It is more likely that the PSAT region interacts with the terminator in some way, either strengthening or weakening the stem structure, or that it affects how the ribosome behaves during the last stages of translation. We analysed 160 sequences for nucleotide preference for each position after the stop codon, and searched for patterns comparing the top tier and the bottom tier. However, no correlation could be found. Hence, the exact mechanism is still unknown at this point, but it appears that this region influences protein expression rather independently (the same sequence has a similar effect on three unrelated preceding ORFs, and therefore is highly interesting in protein expression studies that require high protein yield. Since we do not know by what mechanism the sequences of the 3' UTR (and by extension the IGR) influence the translation rate, it is difficult to explain their rather irregular behaviour in polycistronic mRNA. From what we gathered, a rapidly translated ORF can be severely hindered by a poorly translated ORF both upstream and downstream. Poor codon usage of the upstream ORF and a poor choice of IGR sequence cause translational impediment of the downstream ORF, which can be alleviated by increasing the distance (Figure 6.6AD). Poor codon usage of the downstream ORF impedes the upstream ORF translation severely regardless of the IGR and the impediment is not solved by increasing the distance moderately (30-45 bp) (Figure 6.6B). However, the choice of IGR still influences the upstream ORF. The addition of the terminator stem to the IGR has a very moderate effect as well.

Altogether the results here show the considerations that must be taken into account when designing and studying polycistronic mRNAs. Besides the previously established forward translational coupling, we now show that reverse, feed-back translational coupling exists. A knockout of a single gene within an operon will affect the expression of both upstream and downstream genes, which might result in a phenotype that cannot be exclusively attributed to the absence of said gene. Instead of knocking out the gene, we advise opting for a functional mutant. If the gene's active site is unknown frameshifts could be introduced, however early stop codons should then be avoided.

Materials and methods

Strains and media

Throughout this study we used *E. coli* DH10B T1R (Invitrogen C6400-03). Bacterial cultures used for cloning were grown in LB medium (10 g/L Bacto peptone, 5 g/L yeast extract, 10 g/L NaCl in demineralised water), with 50 mg/L kanamycin when appropriate. An additional 15 g/L agar is added for standard medium plates. The fluorescence assays were performed on M9TG (1x M9 salts (Sigma), 10 g/L tryptone, 5 g/L glycerol), which has

allowed for high cell density and has low auto-fluorescence. All cultures were grown at 37°C.

Plasmids

All plasmids used have the same backbone containing a kanamycin resistance gene and the P15A replication origin. pTN001 and pTN002 feature a strong PtacI promoter, while pTN003 and pTN004 have a weak Pbla promoter. The pTN001 places the GFP in front of RFP, while all others do the reverse. pTN004 is the only construct with a strong terminator almost directly behind the operon. Sequences can be found in the supplementary data. Harmonisation was performed according to .

Fluorescence assay

To ensure an equal growth start, bacteria harbouring different constructs were grown in a pre-culture of 200 µL M9TG, supplemented with 50 mg/L kanamycin in a 96 wells 2 mL Masterblock (Greiner). The Masterblock was covered with a gas-permeable membrane and incubated overnight at 37°C. The pre-cultures were diluted 10000x in fresh 200 µL M9TG and grown in the same way as the pre-cultures. The cultures (and blank medium) were then cooled down to room temperature and diluted 5x in 1x PBS pH 7.4. Finally, 100 µL of the dilution was measured with a BioTek Synergy MX microplate reader at excitation 395/20, emission 508/20, gain 75 (GFP), and excitation 584/9, emission 607/9, gain 100 (RFP). Fluorescence was calculated as raw fluorescence per OD600 for 100 µL 5x dilution. Auto-fluorescence, estimated by introduction of a frameshift in the GFP of pTN002 (pTN002-FS), was subtracted and samples were normalised by dividing by one of the wild-type (WT) samples.

RT-qPCR analysis

10 mL LB with 50 mg/L kanamycin was inoculated 1:1000 from an LB kanamycin preculture. Cells were grown to an OD600 of 0.6 and cooled down on ice-water. Cells were pelleted and resuspended in 250 µL of 50 mM Tris-HCl pH8, 10 mM EDTA and 10 mM DTT. Cells were then lysed with 250 µL of [0.2 M NaOH and 1% SDS]. Protein, genomic DNA and SDS were precipitated by adding 250 µL [1.8 M potassium acetate and 1.2 M acetic acid]. Debris was pelleted in a microcentrifuge tube and 650 µL was transferred to a new Eppendorf tube. RNA was precipitated by adding 650 µL isopropanol and centrifuging for 5 minutes at maximum speed. RNA pellets were washed with 500 µL of [10 mM Tris-HCl pH8 and 70% ethanol], and dried in a laminar flow cabinet. Pellets were dissolved in 100 µL DNAseI buffer (NEB) with 0.25 µL DNAse I (NEB) and incubated at 37 °C for 30 minutes. First, 300 µL of DNAseI buffer was added and then 200 µL of Roti aqua phenol (Roth). The phases were separated by centrifugation and 300 µL of the aqueous phase was transferred to a new Eppendorf tube. 300 µL of isopropanol was added to the aqueous phase and the

mixture was loaded on a silica column (Thermo K0702). The RNA was washed twice with 400 µL [10 mM Tris-HCl pH8, 70% ethanol and 100 mM NaCl]. Finally, the RNA was eluted into 50 µL of [1 mM Tris-HCl pH8, 0.1 mM EDTA]. The RNA was diluted to 1 g/L in water and cDNA was generated with the Maxima H minus (Thermo) reverse transcriptase. RT-qPCR was performed with the SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad) using cDNA derived from 10 ng of total RNA in a 10 µL reaction.

PSAT region library generation

ssDNA containing 30 degenerate nucleotides flanked on both sides with 4 nucleotide overhang and BsaI recognition sites is converted to double stranded DNA using PCR and a primer that binds in the fixed region. 200 pmol ssDNA and 400 pmol primer is used in a 50 µl OneTaq® (NEB) reaction. 99 Cycles of a 5 second primer binding phase and 5 second elongation phase was performed. The dsDNA is purified and concentrated using a silica column (Zymo D4004). The backbone is prepared by first inserting a substantial piece of nonsense DNA flanked by outward facing BsaI sites between the stop codon and terminator which allows for more precise gel separation later on. The plasmid is sequence verified and pre-digested using BsaI-HF®v2 (NEB) to reduce transformation background. The digested backbone is purified from agarose gel (Zymo D4002). The dsDNA is inserted into the backbone using a NEB® Golden Gate Assembly Kit (BsaI-HF®v2) with a 3:1 ratio. 300 colonies were picked and the fluorescence quantified using a Attune NxT Flow Cytometer (Thermo). 96 cultures covering the full fluorescent range were reinoculated. The cultures were measured again and the associated DNA send for sanger sequencing.

SUPPLEMENTARY DATA

Supplementary sequence 6.1. Backbone

GCAAGTGGCA CTTTTCGGGG AAATGTGCGC GGAACCCCTA TTTGTTTATT TTTCTAAATA CATTCAAATA
TGTATCCGCT CATGAATTAA TTCTTAGAAA AACTCATCGA GCATCAAATG AAACTGCAAT TTATTCATAT
CAGGATTATC AATACCATAT TTTTGAAAAA GCCGTTTCTG TAATGAAGGA GAAAACTCAC CGAGGCAGTT
CCATAGGATG GCAAGATCCT GGTATCGGTC TGCGATTCCG ACTCGTCCAA CATCAATACA ACCTATTAAT
TTCCCCTCGT CAAAAATAAG GTTATCAAGT GAGAAATCAC CATGAGTGAC GACTGAATCC GGTGAGAATG
GCAAAAGTTT ATGCATTTCT TTCCAGACTT GTTCAACAGG CCAGCCATTA CGCTCGTCAT CAAAATCACT
CGCATCAACC AAACCGTTAT TCATTCGTGA TTGCGCCTGA GCGAGACGAA ATACGCGGTC GCTGTTAAAA
GGACAATTAC AAACAGGAAT CGAATGCAAC CGGCGCAGGA ACACTGCCAG CGCATCAACA ATATTTTCAC
CTGAATCAGG ATATTCTTCT AATACCTGGA ATGCTGTTTT CCCGGGGATC GCAGTGGTGA GTAACCATGC
ATCATCAGGA GTACGGATAA AATGCTTGAT GGTCGGAAGA GGCATAAATT CCGTCAGCCA GTTTAGTCTG
ACCATCTCAT CTGTAACATC ATTGGCAACG CTACCTTTGC CATGTTTCAG AAACAACTCT GGCGCATCGG
GCTTCCCATA CAATCGATAG ATTGTCGCAC CTGATTGCCC GACATTATCG CGAGCCCATT TATACCCATA
TAAATCAGCA TCCATGTTGG AATTTAATCG CGGCCTAGAG CAAGACGTTT CCCGTTGAAT ATGGCTCATA
CTCTTCCTTT TTCAATATTA TTGAAGCATT TATCAGGGTT ATTGTCTCAT GAGCGGATAC ATATTTGAAT
GTATTTAGAA AAATAAACAA ATAGGCTGTC CCTCCTGTTC AGCTACTGAC GGGGTGGTGC GTAACGGCAA
AAGCACCGCC GGACATCAGC GCTAGCGGAG TGTATACTGG CTTACTATGT TGGCACTGAT GAGGGTGTCA
GTGAAGTGCT TCATGTGGCA GGAGAAAAAA GGCTGCACCG GTGCGTCAGC AGAATATGTG ATACAGGATA
TATTCCGCTT CCTCGCTCAC TGACTCGCTA CGCTCGGTCG TTCGACTGCG GCGAGCGGAA ATGGCTTACG
AACGGGGCGG AGATTTCCTG GAAGATGCCA GGAAGATACT TAACAGGGAA GTGAGAGGGC CGCGGCAAAG
CCGTTTTTCC ATAGGCTCCG CCCCCCTGAC AAGCATCACG AAATCTGACG CTCAAATCAG TGGTGGCGAA
ACCCGACAGG ACTATAAAGA TACCAGGCGT TTCCCCCTGG CGGCTCCCTC GTGCGCTCTC CTGTTCCTGC
CTTTCGGTTT ACCGGTGTCA TTCCGCTGTT ATGGCCGCGT TTGTCTCATT CCACGCCTGA CACTCAGTTC
CGGGTAGGCA GTTCGCTCCA AGCTGGACTG TATGCACGAA CCCCCCGTTC AGTCCGACCG CTGCGCCTTA
TCCGGTAACT ATCGTCTTGA GTCCAACCCG GAAAGACATG CAAAAGCACC ACTGGCAGCA GCCACTGGTA
ATTGATTTAG AGGAGTTAGT CTTGAAGTCA TGCGCCGGTT AAGGCTAAAC TGAAAGGACA AGTTTTGGTG
ACTGCGCTCC TCCAAGCCAG TTACCTCGGT TCAAAGAGTT GGTAGCTCAG AGAACCTTCG AAAAACCGCC
CTGCAAGGCG GTTTTTTCGT TTTCAGAGCA AGAGATTACG CGCAGACCAA AACGATCTCA AGAAGATCAT
CTTATTAATC AGATAAAATA TTTCTAGATT TCAGTGCAAT TTATCTCTTC AAATGTAGCA CCTGAAGTCA
GCCCCATACG ATATAAGTTG TAATTCGGTA CCCCGCTTCG GCGGGGTTTT TTCAAG

Supplementary sequence 6.2. WT-GFP

ATGAGTAAAG GAGAAGAACT TTTCACTGGA GTTGTCCCAA TTCTTGTTGA ATTAGATGGT GATGTTAATG
GGCACAAATT TTCTGTCAGT GGAGAGGGTG AAGGTGATGC AACATACGGA AAACTTACCC TTAAATTTAT
TTGCACTACT GGAAAACTAC CTGTTCCATG GCCAACACTT GTCACTACTT TCTCTTATGG TGTTCAATGC
TTTTCCCGTT ATCCGGATCA CATGAAACGG CATGACTTTT TCAAGAGTGC CATGCCCGAA GGTTATGTAC
AGGAACGCAC TATATCTTTC AAAGATGACG GGAACTACAA GACGCGTGCT GAAGTCAAGT TTGAAGGTGA
TACCCTTGTT AATCGTATCG AGTTAAAAGG TATTGATTTT AAAGAAGATG GAAACATTCT CGGACACAAA
CTGGAGTACA ACTATAACTC ACACAATGTA TACATCACGG CAGACAAACA AAAGAATGGA ATCAAAGCTA

ACTTCAAAAT TCGCCACAAC ATTGAAGATG GATCCGTTCA ACTAGCAGAC CATTATCAAC AAAATACTCC
AATTGGCGAT GGCCCTGTCC TTTTACCAGA CAACCATTAC CTGTCGACAC AATCTGCCCT TTCGAAAGAT
CCCAACGAAA AGCGTGACCA CATGGTCCTT CTTGAGTTTG TAACTGCTGC TGGGATTACA CATGGCATGG
ATGAGCTCTA CAAATAA

Supplementary sequence 6.3. H-GFP

ATGTCGAAAG GTGAAGAACT GTTTACTGGT GTGGTTCCGA TTCTGGTGGA ATTGGATGGG GATGTGAATG
GGCATAAATT CTCCGTTTCG GGTGAGGGGG AAGGGGATGC TACCTATGGT AAACTGACTC TGAAATTCAT
TTGTACTACT GGTAAACTAC CGGTGCCGTG GCCGACCCTG GTTACTACTT TTTCCTACGG GGTGCAGTGT
TTCAGCCGTT ATCCGGATCA CATGAAAAGG CACGACTTCT TTAAGTCGGC TATGCCCGAA GGGTACGTAC
AAGAACGTAC TATATCGTTT AAAGATGACG GGAATTATAA GACCCGAGCA GAAGTTAAGT TCGAAGGGGA
TACTCTGGTG AATCGTATTG AGTTGAAAGG GATTGATTTC AAAGAAGATG GTAATATTCT GGGTCATAAA
TTAGAATATA ATTACAATAG CCATAATGTA TATATTACCG CTGACAAACA GAAGAATGGT ATTAAAGCTA
ATTTTAAAAT TCGTCATAAT ATTGAAGATG GTTCGGTGCA GCTAGCTGAC CACTACCAGC AGAATACTCC
GATTGGGGAT GGGCCGGTTC TGTTGCCGGA CAATCACTAT CTATCGACCC AGTCCGCTCT GTCGAAAGAT
CCCAATGAAA AGCGTGACCA TATGGTTCTG CTGGAGTTCG TAACCGCAGC AGGGATTACC CACGGGATGG
ATGAACTATA TAAATAA

Supplementary sequence 6.4. RFP

ATGGCTTCCT CCGAAGACGT TATCAAAGAG TTCATGCGTT TCAAAGTTCG TATGGAAGGT TCCGTTAACG
GTCACGAGTT CGAAATCGAA GGTGAAGGTG AAGGTCGTCC GTACGAAGGT ACACAGACCG CTAAACTGAA
AGTTACCAAA GGTGGCCCGC TGCCGTTCGC TTGGGACATC CTGTCCCCGC AGTTCCAGTA CGGTTCCAAA
GCTTACGTTA AACACCCGGC TGACATCCCG GACTACCTGA AACTGTCCTT CCCGGAAGGT TTCAAATGGG
AACGTGTTAT GAACTTCGAA GACGGTGGTG TTGTTACCGT TACCCAGGAC TCCTCCCTGC AAGACGGTGA
GTTCATCTAC AAAGTTAAAC TGCGTGGTAC CAACTTCCCG TCCGACGGTC CGGTTATGCA GAAAAAAACC
ATGGGTTGGG AAGCTTCCAC CGAACGTATG TACCCGGAAG ACGGTGCTCT GAAAGGTGAA ATCAAAATGC
GTCTGAAACT GAAAGACGGT GGTCACTACG ACGCTGAAGT TAAAACCACC TACATGGCTA AAAAACCGGT
TCAGCTGCCG GGTGCTTACA AAACCGACAT CAAACTGGAC ATCACCTCCC ACAACGAAGA CTACACCATC
GTTGAACAGT ACGAACGTGC TGAAGGTCGT CACTCCACCG GTGCTTAA

Supplementary sequence 6.5. LacZ

ATGACCATGA TTACGGATTC ACTGGCCGTC GTTTTACAAC GTCGTGACTG GGAAAACCCT GGCGTTACCC
AACTTAATCG CCTTGCAGCA CATCCCCCTT TCGCCAGCTG GCGTAATAGC GAAGAGGCCC GCACCGATCG
CCCTTCCCAA CAGTTGCGCA GCCTGAATGG CGAATGGCGC TTTGCCTGGT TTCCGGCACC AGAAGCGGTG
CCGGAAAGCT GGCTGGAGTG CGATCTTCCT GAGGCCGATA CTGTCGTCGT CCCCTCAAAC TGGCAGATGC
ACGGTTACGA TGCGCCCATC TACACCAACG TGACCTATCC CATTACGGTC AATCCGCCGT TTGTTCCCAC
GGAGAATCCG ACGGGTTGTT ACTCGCTCAC ATTTAATGTT GATGAAAGCT GGCTACAGGA AGGCCAGACG
CGAATTATTT TTGATGGCGT TAACTCGGCG TTTCATCTGT GGTGCAACGG GCGCTGGGTC GGTTACGGCC
AGGACAGTCG TTTGCCGTCT GAATTTGACC TGAGCGCATT TTTACGCGCC GGAGAAAACC GCCTCGCGGT
GATGGTGCTG CGCTGGAGTG ACGGCAGTTA TCTGGAAGAT CAGGATATGT GGCGGATGAG CGGCATTTTC
CGTGACGTCT CGTTGCTGCA TAAACCGACT ACACAAATCA GCGATTTCCA TGTTGCCACT CGCTTTAATG
ATGATTTCAG CCGCGCTGTA CTGGAGGCTG AAGTTCAGAT GTGCGGCGAG TTGCGTGACT ACCTACGGGT
AACAGTTTCT TTATGGCAGG GTGAAACGCA GGTCGCCAGC GGCACCGCGC CTTTCGGCGG TGAAATTATC

GATGAGCGTG GTGGTTATGC CGATCGCGTC ACACTACGTC TGAACGTCGA AAACCCGAAA CTGTGGAGCG
CCGAAATCCC GAATCTCTAT CGTGCGGTGG TTGAACTGCA CACCGCCGAC GGCACGCTGA TTGAAGCAGA
AGCCTGCGAT GTCGGTTTCC GCGAGGTGCG GATTGAAAAT GGTCTGCTGC TGCTGAACGG CAAGCCGTTG
CTGATTCGAG GCGTTAACCG TCACGAGCAT CATCCTCTGC ATGGTCAGGT CATGGATGAG CAGACGATGG
TGCAGGATAT CCTGCTGATG AAGCAGAACA ACTTTAACGC CGTGCGCTGT TCGCATTATC CGAACCATCC
GCTGTGGTAC ACGCTGTGCG ACCGCTACGG CCTGTATGTG GTGGATGAAG CCAATATTGA AACCCACGGC
ATGGTGCCAA TGAATCGTCT GACCGATGAT CCGCGCTGGC TACCGGCGAT GAGCGAACGC GTAACGCGAA
TGGTGCAGCG CGATCGTAAT CACCCGAGTG TGATCATCTG GTCGCTGGGG AATGAATCAG GCCACGGCGC
TAATCACGAC GCGCTGTATC GCTGGATCAA ATCTGTCGAT CCTTCCCGCC CGGTGCAGTA TGAAGGCGGC
GGAGCCGACA CCACGGCCAC CGATATTATT TGCCCGATGT ACGCGCGCGT GGATGAAGAC CAGCCCTTCC
CGGCTGTGCC GAAATGGTCC ATCAAAAAAT GGCTTTCGCT ACCTGGAGAG ACGCGCCCGC TGATCCTTTG
CGAATACGCC CACGCGATGG GTAACAGTCT TGGCGGTTTC GCTAAATACT GGCAGGCGTT TCGTCAGTAT
CCCCGTTTAC AGGGCGGCTT CGTCTGGGAC TGGGTGGATC AGTCGCTGAT TAAATATGAT GAAAACGGCA
ACCCGTGGTC GGCTTACGGC GGTGATTTTG GCGATACGCC GAACGATCGC CAGTTCTGTA TGAACGGTCT
GGTCTTTGCC GACCGCACGC CGCATCCAGC GCTGACGGAA GCAAAACACC AGCAGCAGTT TTTCCAGTTC
CGTTTATCCG GGCAAACCAT CGAAGTGACC AGCGAATACC TGTTCCGTCA TAGCGATAAC GAGCTCCTGC
ACTGGATGGT GGCGCTGGAT GGTAAGCCGC TGGCAAGCGG TGAAGTGCCT CTGGATGTCG CTCCACAAGG
TAAACAGTTG ATTGAACTGC CTGAACTACC GCAGCCGGAG AGCGCCGGGC AACTCTGGCT CACAGTACGC
GTAGTGCAAC CGAACGCGAC CGCATGGTCA GAAGCCGGGC ACATCAGCGC CTGGCAGCAG TGGCGTCTGG
CGGAAAACCT CAGTGTGACG CTCCCCGCCG CGTCCCACGC CATCCCGCAT CTGACCACCA GCGAAATGGA
TTTTTGCATC GAGCTGGGTA ATAAGCGTTG GCAATTTAAC CGCCAGTCAG GCTTTCTTTC ACAGATGTGG
ATTGGCGATA AAAAACAACT GCTGACGCCG CTGCGCGATC AGTTCACCCG TGCACCGCTG GATAACGACA
TTGGCGTAAG TGAAGCGACC CGCATTGACC CTAACGCCTG GGTCGAACGC TGGAAGGCGG CGGGCCATTA
CCAGGCCGAA GCAGCGTTGT TGCAGTGCAC GGCAGATACA CTTGCTGATG CGGTGCTGAT TACGACCGCT
CACGCGTGGC AGCATCAGGG GAAAACCTTA TTTATCAGCC GGAAAACCTA CCGGATTGAT GGTAGTGGTC
AAATGGCGAT TACCGTTGAT GTTGAAGTGG CGAGCGATAC ACCGCATCCG GCGCGGATTG GCCTGAACTG
CCAGCTGGCG CAGGTAGCAG AGCGGGTAAA CTGGCTCGGA TTAGGGCCGC AAGAAAACTA TCCCGACCGC
CTTACTGCCG CCTGTTTTGA CCGCTGGGAT CTGCCATTGT CAGACATGTA TACCCCGTAC GTCTTCCCGA
GCGAAAACGG TCTGCGCTGC GGGACGCGCG AATTGAATTA TGGCCCACAC CAGTGGCGCG GCGACTTCCA
GTTCAACATC AGCCGCTACA GTCAACAGCA ACTGATGGAA ACCAGCCATC GCCATCTGCT GCACGCGGAA
GAAGGCACAT GGCTGAATAT CGACGGTTTC CACATGGGGA TTGGTGGCGA CGACTCCTGG AGCCCGTCAG
TATCGGCGGA ATTCCAGCTG AGCGCCGGTC GCTACCATTA CCAGTTGGTC TGGTGTCAAA AATAA

Supplementary table 6.1. Construct design

5′-UTRs	Sequence
5′-UTR (tac)	TTGACAATTA ATCATCGGCT CGTATAATGT GTGGGGAGAC CACAACGGTT TCCCTCTAGA AATAATTTTG TTTAACTATA AGAAGGAGAT ATACAT
5′-UTR (bla)	TTCAAATATG TATCCGCTCA TGAGACAATG TGTGGGGAGA CCACAACGGT TTCCCTCTAG AAATAATTTT GTTTAACTAT AAGAAGGAGA TATACAT

IGRs	Sequence
IGR [0]	ACTAGAAATA ATTTTGTTTA ACTATAAGAA GGAGATATAC AT
IGR [PSAT]	[PSAT] ACTAGAAATA ATTTTGTTTA ACTATAAGAA GGAGATATAC AT
IGR [PSAT] stem	[PSAT] CCCCGCTTCG GCGGGGACTA GAAATAATTT TGTTTAACTA TAAGAAGGAG ATATACAT
3'UTRs	**Sequence**
3' UTR No Term	ACTAGT
3' UTR [0] Term	ACTAGTATAA TGATGTGTTA TCATTGATGC GAGGCGCCTA TACCTCCCCG CTTCGGCGGG GTTTTTTT
3' UTR [PSAT] Term	[PSAT] CCCCGCTTCG GCGGGGTTTT TTT
3' UTR [16] Term	GACCCAACCA CTACACCCCG CTTCGGCGGG GTTTTTTT
PSAT	
PSAT [1]	TAATACCAAA CTAAGCTTAA AGAAAGCAAC
PSAT [2]	TCGGAAACTA CGTCTTCGTC ATAAACCCTC
PSAT [3]	TCCCTCGTTC TACATCTAAT CAACAGCCCT
PSAT [4]	ACGATAACGA AACTTTCAAA ACCTAATGAA
PSAT [5]	TAAACCACAC TAGCTAGGTA GGAAACAAAA
PSAT [6]	CTCCCCACAT TAGACCTTAG CGGGAACGTC
PSAT [7]	TGGTTTCTAC CGAGCGGCCG GCTCCCTCGC
PSAT [8]	TTAGCAAACT TAGCCGATTA ATAGAACGAT
PSAT [9]	CAATTGAGCA CACCGGACCC AACCACTACA
PSAT [10]	CTTTAGTCTC GGTATACTCT TCTGTTTTCG
PSAT [11]	CAATTGAGCA CACCGGACCC AACCA
PSAT [12]	CAATTGAGCA CACCGGACCC
PSAT [13]	CAATTGAGCA CACCG
PSAT [14]	GAGCA CACCGGACCC AACCACTACA
PSAT [15]	CACCGGACCC AACCACTACA
PSAT [16]	GACCC AACCACTACA
Layout	
pTN001 Series	Backbone - 5'-UTR (tac) – GFP – IGR [0] – RFP – 3' UTR No Term
pTN002 Series	Backbone – 5'-UTR (tac) – RFP – IGR [0] – GFP – 3' UTR No Term
pTN003 Series	Backbone – 5'-UTR (bla) – RFP – IGR [0] – GFP – 3' UTR No Term
pTN004 Series	Backbone – 5'-UTR (bla) – RFP – IGR [0] – GFP – 3' UTR [0] Term
pTN005	Backbone – 5'-UTR (bla) – GFP – 3' UTR [0] Term
pTN006 Series	Backbone – 5'-UTR (bla) – GFP/RFP/LacZ – 3' UTR [PSAT] Term
pTN007 Series	Backbone – 5'-UTR (bla) – [see Figure 6.6] – 3' UTR [PSAT] Term

Supplementary table 6.2. PSAT library sequences with observed fluorescence.

PSAT sequence (GFP lib)	GFP	PSAT sequence (RFP lib)	RFP
CTTTAGTCTCGGTATACTCTTCTGTTTTCG	38257	TAAACCACACTAGCTAGGTAGGAAACAAAA	98994
AGTTGTCCGTGCGTCTCTTAAACTGGTAAA	34595	GCGAAGTCCAACACTCCACCAAGAATCTAC	90809
CAAGTAAGCTTGAGGCCTAACTACAACAAA	34401	TCAAACACAATTCAATCTACAGCAAACAAG	89434
ACCTTACTTCGCTTAAAACTCTGTATCTAA	33026	AAGAGAGCAGTAGAGGCGTCAGCAGGACAC	85051
CGAACGGGTTAGACGTATATAAACTGAAAA	32869	TAACCAAGCAAAGAAACCACATCCCACTAA	84717
GACCTCGCCCACCCAAGTTGCACCTATGTA	32795	TAACGACAATCACGGTGTGAGAAATCTATC	82093
CAATTGAGCACACCGGACCCAACCACTACA	32143	TCTGGCATTCCTCCCGGCGTGCTACCTCAT	81512
TCCATGTCCCGCACCTTTCCCTATTCTACT	31964	ACGATAACGAAACTTTCAAAACCTAATGAA	77448
CCTGATCAAATTATGAAATAAACCTCTGAA	31957	TAAAACCGCGAAAGCATCACAACAAACCAA	76806
CCTGTAAAGCGAACGAGCAAAACTCATACA	31923	GGAGGTAAAGATAGTCAAACACAACAAGAA	75679
CGCCACCAGACATGCCCGTTCTTACTAACC	31792	ACAAAACTCAGAGGAAAAGAAGAAAACAAA	74712
CCCTCCACTTAATGGCAAGCAGTCCTTCCA	31775	GGAAACCGCAGACGATAAATAGGAGCAAAA	74281
ACCGGGCCCCACCCCGTTTGCAATACCTCA	31770	AGGACGTTCGAAGGTAACAATATGAGAAAT	73571
CTCATTTGCTACTCACTTTATAGCACTGTA	31694	AGTGGCAGCTCAGCATCCTTTGTACCCTAA	72939
TCTTGAATGTTAAGCATGCAGTTAATACAG	31690	AATGGCACAACATCCAAAATCTAAAACCAC	72748
AGCCTGACCTTGATTATGAGAGTGAACAAA	31668	CCGACACGAATGGCCGACCGAAGTAATACA	72500
CTTCACGTTCACGCCTTTACAATTGATTTA	31649	TCCTAAGATCACCTTTCCATCCTAACCGAC	71527
CCCCTATCATCGCTTCTGGTACATCACCTA	31542	TTACCAAAATCCAACTCAACAAAGAAATAT	70909
AGGGCCTTCCCTTCCCTACCTCTGTCCAAC	31467	AACACATTATCTCACTTTTAATCACGTTAA	70063
ATCTCATTCGTGATTGTATATGATTGAGTC	31345	AAAACCGCACAAATATCCAATAGGCGCAAA	70046
AGGTTCGTTGGTGTTTATAACACTTTGTTT	31229	ATAAGAAATAAAACAAAGTAAGTAAGATCA	70022
CTCAGAGCCCCTGTAGTCAGCACTTTGCCC	31106	TCGCACGGCTCAATGTGCAAATTACACCCA	69908
ACCACACGATTAAATCCCAGAAACATCATA	31043	CAAATGTTCGCGCGCAGTGCGCGTTGGTAG	69856
ACTAAGCTTTACATAAGGCTGATTGTGCAC	30955	GCACATATGAGCAGTACGAGAGCAATATAA	69460
ACACGTTTCCGGTGTTGGCCCATCACGATA	30935	CAAAACAAAGGACACGCCAAAATAATTTAC	69409
ACATGAGCGGAATCGCTAACTAAGTTAAAC	30902	TAACTCTCAAGACGCGATACCAAAACATAA	68868
AATTGTCGATGTATGCTAAAACTTCAAATT	30835	ATAACCTGACATACCCCTAAGATAACCGTG	68078
CTCAATCCATAGACCAATCCAACCAATTCT	30833	GACACCTCCCAAACCACTGCACCTTGAACC	68042
ATCGATATTCCGCTAGATATATGTTCATTT	30812	AAGCAATACACAGATAATAAACACACAAAT	68006
CTGGGCGTCCAACTAAGGCCCCACGGACCT	30601	AAGAAATAAACCATAACCAAAATCAGTGTG	67633
ATCGACACCCTACCGGACAACTTTGTCTGC	30495	ACGGCTCTTTGAGACGTCATGTTATACTCC	64287
TCCTTTCTGCAGTAAGAAGTAAACGAGAAT	30487	TGGGCCTGGAGCGCCACCGTACCGGAGGAG	63889
TCCATGTCCCGCACCTTTCCCTATTCTACT	30410	CCGTTATGATATCCCTCTTAAACATTCTAC	63853
CTCAGAGCCCCTGTAGTCAGCACTTTGCCC	30372	AACATGGCTACGGATCCAATGCCACATGTT	62205
TAGCTAGGTTGCTTGACAATCTGTCCCTTC	30331	TCCCTCGTTCTACATCTAATCAACAGCCCT	57534

AATTTTAGGCTATTACGAAGACTTGTTATT	30163	ACATGCTGAGTTTTCGAATCGGATCGAAAA	54974
CTGAAATCACTAATGTTTCGGTAAAACGCT	29730	TACCCTTCTTCAGCTGCTTTCCAACCCTCC	52651
TCCTGCCGCGCGCACGGTTGGGCAACGGCA	29702	ACAGAATGCGTGGGCGCGGAAGGAGCAAGC	50340
TCCAGGCAAAGGCACCCTTCGAAACGCACT	29652	CCATGAATCGTCTGCGTCGCTGGGGCTCTC	50039
ACTAGTTCTGGTATCATTAGGTCTAGTTGC	29584	CCTCTCCTATATTCCTCCACCGCATGCTAT	49081
AATCTTCCTGCACTCACTAGCCGTCTTATT	29553	TAATGAGGTCACGGTGTGCTGGAAGGGTGT	49015
CTCACAGACCTTCCTCACCCATCTGACTCC	29152	ACGAACCTTTCCACTCCCCAATTCTTAAGT	48558
CTGAAATCACTAATGTTTCGGTAAAACGCT	29076	ATGAAAACTACAAAATTCATCAACTAAACT	48462
TCGCCTTCAACAGGGCCTATCCAGTACCCC	28840	ATCCTGCCCTTCAGCCCATGCTCCTCCTGT	47859
ATCAGACACCTTATGACTACCAGTAAAGTC	28709	CCTAGCTGGGAGCGCGCGGTGTGCGTTGCC	47590
CTTGTTGAGTGGTGCCTCGGGGAGCGAGGG	28626	AGTGAAATGCTAGCTACGCCGTTCTCCTTC	46727
TGCCTGCGCGGCGCCCCGGGCGACAGGCCC	28293	TCCGATCCTTTGGACTCCGCGGCGGCTCGT	45352
CCACGGTGAAGCTAATCACCTCCCGGTGCC	28148	ATTCTCTCACCTCGCACCCAGGCGGGCGCC	44284
CCAAAGCCATGGCTGTCTGCAACCAAAGAC	27866	TAGTACAAATTTGCTAACATAAATACAATC	42824
CCGCTCTTCAGAGCGCCTAATCTCCGAGCC	27341	ACGAACCTTTCCACTCCCCAATTCTTAAGT	41525
ATCAATGGTTAGCGTAATCGACACTATACC	27157	TCTTCCTACTCTCTCCTACTGTCTCCTTTC	41331
TTCTTTGCGTTCTAACATGTTGCGTATACT	26959	TCGGAAACTACGTCTTCGTCATAAACCCTC	39852
CATCACAGACCCAAGAGCCGCAAAATCGTC	26574	TAACTAATACATACCACTGAGATCTCCTCC	36239
TCATACCTAGCTCCTAGTACATTCCCCGCG	26496	TAGCACACAAAACGAAATAAATCAACCAAG	35361
TTAGCAAACTTAGCCGATTAATAGAACGAT	26413	AATATACCAAGCAAATACAGGTGACGCAAT	30280
ACCCGAGAACCTCCCCGCCCGCTCCCCACT	22798	TAATACCAAACTAAGCTTAAAGAAAGCAAC	18374
TCGAGGAGGGTGTGGCGAAATCTCAGTGCT	22104		
CGATTAAACGCCATAGCAGCGTGGGGCGGC	22018		
CGGGACCCACAGGGCCCAGTACCCTGTGGG	21922		
CGGGCGTAAGGGCGCGTGCGGCGGTGTGTG	21553		
ACAACAGAGTCCGACCGAGAGGGGCCGAGT	21261		
CCAGTTGTGCGTCACCTTGTCATTGTTTGT	21059		
ACTACAGCTATGCTCCGAATCTACAGGAAA	20677		
TCTTTCGGCTCGTGGCGCGCGCTCCCCGCG	20622		
TGGTTTCTACCGAGCGGCCGGCTCCCTCGC	19503		
TCGTGCTCGGCATATGCGGGCGGGGCAATA	18815		
GCGTGTTCCTATTTCCATTCATGTAGGTAT	17980		
CCGGCGGGTGTGCGCGCGTGTTCCGCCGCT	16654		
TCGCAGAGCTGTATTAAGCCCATTGCAATC	16098		
CACCATCCACACGTCGAAGCATATGTTAAT	15087		
CGCGCGTGCCAGTGTGGGTGGGCGGGCGGC	13991		
CTCCCCACATTAGACCTTAGCGGGAACGTC	13722		

Chapter 7

Medium-throughput *in vitro* detection of DNA cleavage by CRISPR-Cas12a

Sjoerd C.A. Creutzburg, Thomas Swartjes, John van der Oost

Abstract

Quantifying DNA cleavage by CRISPR-Cas nucleases is usually done by separating the cleaved products from the non-cleaved target by agarose gel electrophoresis. We devised a method that eliminates the quantification from band intensity on agarose gel, and uses a target with a fluorescent dye on the one end and a biotin on the other. Cleavage of the target will separate the dye from the biotin, and cause the dye to stay in solution when streptavidin beads are introduced. All non-cleaved target will be eliminated from solution and no longer contribute to detectable fluorescence. Cleavage will therefore increase the fluorescent signal. A control, which has no streptavidin treatment, is taken along to correct for any errors that might have been introduced by pipetting, inactivation of the fluorescent dye or release of the biotin during several steps of the procedure. With this method we were able to quantify the fraction of active Cas12a in a purification sample and assess the cleavage rate.

Overview

Quantifying DNA cleavage by CRISPR-Cas nucleases is key in studying their kinetics (98). We found the need for a method that is more high-throughput than analysis by a commonly used method like agarose of polyacrylamide gel electrophoresis and subsequent quantification (2), and does not require expensive equipment like the Agilent Fragment Analyzer. The basic principle behind this technique is the separation of a fluorophore and a biotin upon cleavage of the nuclease. The target DNA features a fluorophore on the one end and a biotin on the other. After an appropriate incubation period with the nuclease, the biotin is bound to immobilised streptavidin on magnetic beads and removed from solution. Therefore, any DNA molecule that retains its biotin label will no longer contribute to the fluorescence of the solution. By comparing that fluorescence to the fluorescence when no streptavidin was added, one can estimate the fraction of DNA target that is cleaved by the nuclease. An overview of the principle is depicted in Figure 7.1A and an overview of the workflow is described in Figure 1B. While this method was developed for the assessment of CRISPR-Cas nucleases, it could be applied to other DNAses as well (argonaute (1, 99), restriction enzymes), as long as it separates the fluorophore from the biotin. Note that the CRISPR-Cas nuclease cleavage activity observed in these assays may not resemble the *in vivo* situation perfectly, since more factors are involved *in vivo* (e.g. chromatin structure (100, 101)).

Equipment

A microtiter plate reader that can measure fluorescence with the appropriate wavelengths for the fluorescent dye is required. We used a Biotek Synergy MX microplate reader. We made our own magnet plate to pull down the streptavidin beads. The rims were sawn off a flat-bottom 96 wells plate (Greiner). Neodymium disc magnets with a diameter of 6 mm and a height of 1 mm were used (https://www.supermagnete.nl (S-06-01-N)). The magnets on the bottom of the plate are held in place by another magnet inside the well (Figure 7.2). Magnets in a horizontal or vertical line on the plate alternate their polarity, while all diagonal lines have the same magnetic polarity (like a chess board).

Target design and generation

Targets can be made either by "'round the horn" overhang extension PCR and subsequent re-circularisation, or by ligating annealed oligos into a Golden Gate entry site. Alternatively, an oligo with a target flanked by two primer annealing sites (one for the biotin primer and the other for the Alexa Fluor primer) can be used as a template. A 770 bp amplicon that harbours the protospacer target was made by polymerase chain reaction (PCR), with a biotin on the 5'-end of one primer and an Alexa Fluor 594 dye on the 5'-end of the other primer. In our experience, the Alexa Fluor 594 can withstand the PCR denaturation steps

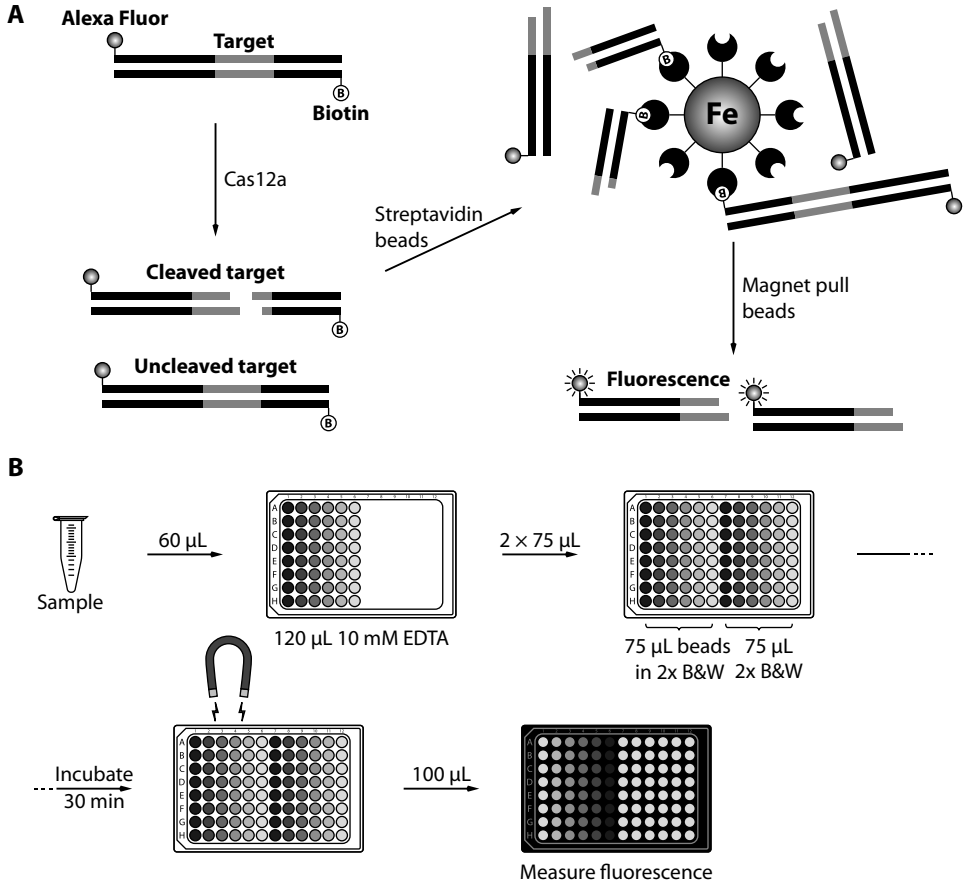


Figure 7.1. Medium-throughput in vitro detection of DNA cleavage. (A) The principle of the here-described method is that the DNA nuclease of interest separates the biotin from the fluorophore, which results in different fluorescence intensities when the biotin-labelled target DNA molecules are removed from solution. (B) Practical overview of the technique. The colour intensity indicates the amount of cleaved target DNA, which can be caused by a higher concentration of Cas protein, or a longer reaction time. Half of the plate is dedicated to the controls. Control samples without beads do not have fluorescence removed, so show maximum fluorescence always.

to a sufficient degree for detecting the DNA, but the integrity is somewhat compromised (by about 4%). Target PCRs were purified by the Zymo clean and concentrator kit (Zymo Research) and quantified with a Qubit dsDNA BR Assay Kit (Thermo Fischer Scientific). The Qubit quantifies DNA mass, so the concentrations were recalculated to molar.

Guide design and *in vitro* transcription

A template oligonucleotide was designed for *in vitro* transcription (IVT). It is the reverse complement of the T7 promoter that is followed by a guanine for IVT efficiency, the mature repeat and the spacer derived part of the guide. The T7 RNA polymerase only requires the

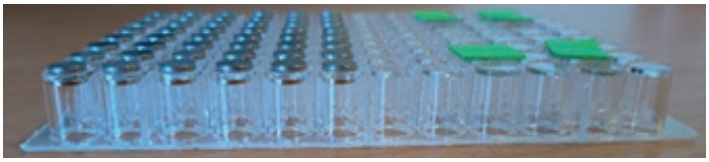


Figure 7.2. Magnet plate to pull down the streptavidin beads. Only half of the plate will contain beads so there is no need for a full plate of magnets.

promoter to be dsDNA, so the non-template oligo only comprises the T7 promoter and no part of the actual transcript. IVT was performed with 1 mM rNTPs, 0.5 µM T7 promoter sense oligo, 0.25 µM template oligo and 10 U/µL home-made T7 polymerase in T7 buffer (NEB) at 37°C for 1h. The gRNAs were purified from denaturing polyacrylamide gel by the crush and soak method in 50 mM Tris-HCl, 1 mM EDTA and 10 mM DTT and subsequently concentrated and purified by ultrafiltration using an Amicon 3k filter unit (Merck-Millipore) and washing with ultrapure water. gRNAs were quantified with the Qubit RNA BR Assay kit (Thermo Fischer Scientific).

Titrating the streptavidin beads

The target was titrated with Dynabeads MyOne Streptavidin C1 (Thermo Fischer Scientific) to determine their specific capacity. Target molecules may not carry the biotin label, while they do carry the Alexa Fluor label. These target molecules cannot be bound to the streptavidin beads and cause a minimum residual fluorescence after streptavidin binding the subsequent removal of bound target molecules. The concentration of beads at which the minimal residual fluorescence was observed was used in the cleavage assays. The size of the target will largely determine the binding capacity of the beads, where larger DNA molecules reduce the binding capacity. This means that for other target sizes, the titration should be performed again. The following was used: 50 µL of a 1 ng/µL solution of 770 bp target was bound to different amounts of Dynabeads® MyOne™ Streptavidin C1 that was diluted 40x in 1x bind and wash solution.

Nuclease titration

Protein quantification assays like the Bradford assay can estimate the total amount of protein, but cannot distinguish between active and inactive protein. As a test case, we used the Cas12a nuclease (formerly known as Cpf1 (102)) from *Francisella tularensis subsp. novicida* U112. The binding of Cas12a is not easily reversed, so it will stay bound to the target, making it a single-turnover enzyme (98). This makes the activity quantification different from a multi-turnover enzyme, as the concentration of Cas12a needs to be in the same range as the target concentration for full cleavage to occur.

10 nM of target was cleaved by different amounts of Cas12a in the presence of high amounts of gRNA (>100 nM) in NEBuffer4 to determine the concentration of Cas12a after

purification. Full cleavage can be observed when the amount of Cas12a is equal or higher than the amount of target, but since the cleavage rate slows down when the target is depleted, the best estimates can be derived from reactions where 5-8 nM of target is cleaved.

Cleavage assay

The target DNA solution was diluted to 8 nM in ultrapure water (equals 2x concentrated). The 2x RNP solution was composed of 6 nM Cas12a protein and 18 nM gRNA in 2x NEBuffer4 (NEB). The 2x target and 2x RNP solutions were pre-warmed at 37°C in a climate chamber. The solutions were mixed 1:1 while leaving enough sample for a non-cleaved control and incubated for up to 30 minutes at 37°C. The cleavage reactions were quenched by pipetting 60 µL of cleaved sample in 120 µL of 10 mM EDTA. The non-cleaved control was made by pipetting 30 µL of 2x target solution in the 120 µL of 10 mM EDTA first and then adding 30 µL of 2x RNP solution. The quenched reactions were cooled down to room temperature for further processing.

Streptavidin treatment

75 µL of quenched reaction was pipetted into 75 µL 2x B&W buffer (10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl) and another 75 µL was pipetted into 75 µL 2x B&W buffer containing 2 µL of streptavidin beads. The samples were then incubated at room temperature whilst shaking mildly for 45 minutes. The beads were pulled down by a magnet and 100 µL of sample was transferred to a black 96 wells plate for fluorescence measurement at 585 nm with a bandwidth of 20 nm as excitation wavelength and 626 nm with a bandwidth of 20 nm as emission wavelength.

Calculating the cleavage

The cleavage can be estimated by the liberation of fluorescent dye. Provided that all targets are biotinylated, no fluorescence should be observed without cleavage. However, in practice, there is always a fraction of primers that have no biotin on the 5'-end. Therefore, it is important to take that fraction into account, when accurately determining the cleavage.

Each cleavage assay is comprised of a dsDNA that is targeted by Cas12a ("cleaved") and a non-targeted control ("uncleaved"). Both of these are split in two; one is treated with streptavidin ("strep") and the other is not ("total"). The fluorescence (FI) of all samples is measured.

Estimating the fraction of non-biotinylated targets is done by

$$f_{nonBtn} = \frac{Fl_{uncleaved\_strep}}{Fl_{uncleaved\_total}}$$

We assumed that the biotinylation of the target does not influence the cleavage rate. However, the fraction that is not biotinylated does not lose fluorescence upon streptavidin treatment. The fluorescence in the cleaved sample that is still observed after streptavidin treatment can either be derived from target that is biotinylated and cleaved, or target that is not biotinylated.

$$f_{Fl} = \frac{Fl_{cleaved\_strep}}{Fl_{cleaved\_total}} = f_{cleaved} \cdot (1 - f_{nonBtn}) + f_{nonBtn}$$

Rewriting this formula gives:

$$f_{cleaved} = \frac{f_{Fl} - f_{nonBtn}}{1 - f_{nonBtn}}$$

Or in its expanded form:

$$f_{cleaved} = \frac{\frac{Fl_{cleaved\_strep}}{Fl_{cleaved\_total}} - \frac{Fl_{uncleaved\_strep}}{Fl_{uncleaved\_total}}}{1 - \frac{Fl_{uncleaved\_strep}}{Fl_{uncleaved\_total}}}$$

Tips and tricks

- While pipetting accuracy is always important, from the step where the quenched sample is split into a streptavidin treated part and an untreated part and onwards, accurate pipetting is paramount. Using reverse pipetting is a good way to keep the volumes exactly the same.
- Air displacement pipettes are sensitive to differences in temperature. In a climate chamber this will not be an issue when all equipment is acclimatised.
- Alexa Fluor dyes may be slightly light sensitive, so keep everything in the dark.

Appendix 1. Equipment and supply list

Equipment

- Microplate reader that can measure fluorescence
- 96 wells magnet plate
- Multichannel pipettes

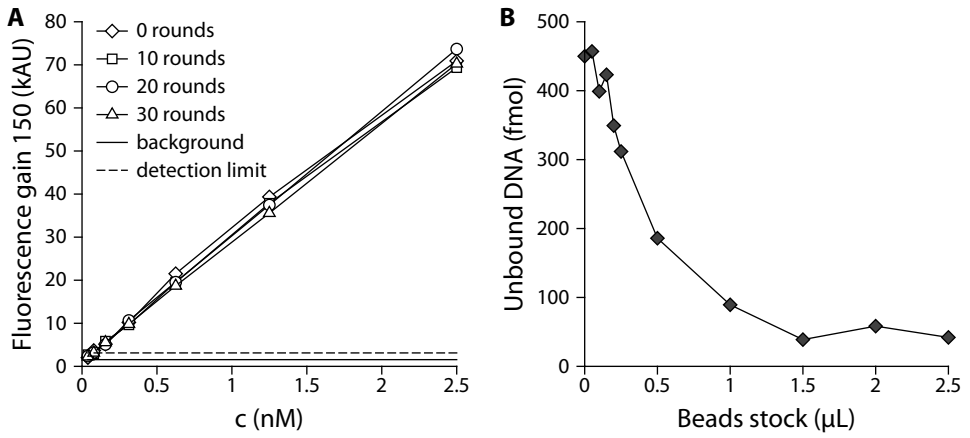


Figure 7.3. Alexa dye stability and streptavidin bead capacity. (A) Thermostability and detection limit of the Alexa Fluor 594. The solid line indicates the basal fluorescence of the buffer without any fluorescent dye, while the dotted line is two times that value. (B) Streptavidin beads titration. Without beads, the fluorescence is at its maximum. Maximum binding is reached at 1.5 µL streptavidin beads stock suspension.

Supplies

- 96 wells plates
- Black 96 wells plates for fluorescence
- Dynabeads® MyOne™ Streptavidin C1
- DNA oligo with Alexa Fluor 594 attached to the 5′ end
- DNA oligo with biotin attached to the 5′ end

Appendix 2. Assay setup

Alexa Fluor 594 heat stability and detection limit.

100 nM of the oligonucleotide with the Alexa Fluor 594 attached to it was subjected to several rounds of PCR (98°C 20s, 60°C 30s, 72°C 30s) after an initial 98°C 60s step. The reaction was carried out in 1x Q5 master mix (NEB). The reactions were then serial diluted by a factor 2 starting at 10 nM and 100 µL was measured in a Biotek Synergy MX microplate reader (Figure 7.3A). The Alexa dye is at about 96% of its initial fluorescence after 30 rounds of PCR. The detection limit of this fluorescent dye is about 0.1 nM.

Streptavidin bead titration

457 fmol of a 770 bp linear DNA fragment with on one side a biotin and on the other side an Alexa Fluor 594 was incubated with different amounts of streptavidin beads. After

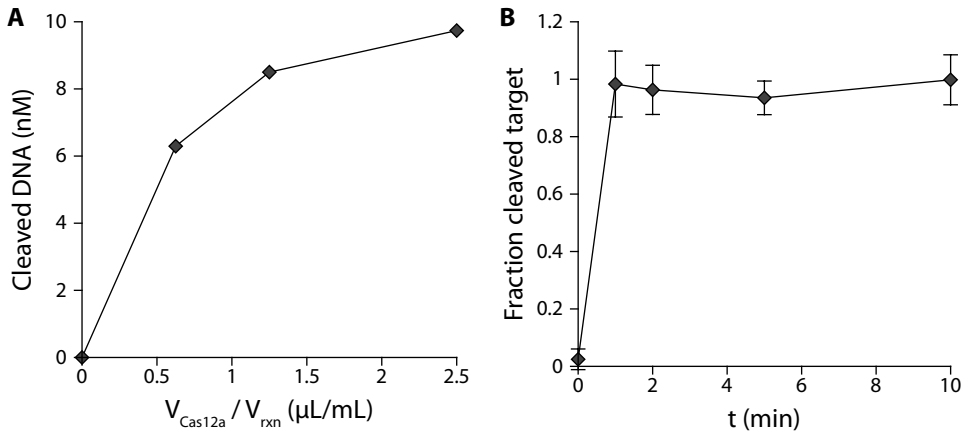


Figure 7.4. FnCas12a titration and kinetic assay. (A) Titration of FnCas12a. On the y-axis is the amount of cleaved DNA, while the inverse dilution factor is on the X-axis. (B) Kinetic assay. A 770 bp target was cleaved by FnCas12a in the presence of corresponding gRNA. The error bars indicate the SD for 6 replicates.

magnetic pulldown, supernatant fluorescence was measured and recalculated to unbound DNA (Figure 7.3B). The minimum fluorescence is reached at about 1.5 uL of streptavidin beads stock suspension. From this experiment, the amount of non-biotinylated DNA is estimated at about 8%.

Cas12a titration

FnCas12a was purified following an established protocol (103). The final concentration was about 200 µM as estimated by Roti nanoquant (Roth) analysis. 10 nM of DNA target was digested with different amounts of FnCas12a at 37°C for 30 minutes (Figure 7.4A). Since this procedure is carried out to estimate the amount of active FnCas12a in the batch, a significant fraction of target must be cleaved, while the non-cleaved DNA is in excess. The reaction of FnCas12a with its target lowers both of their concentrations and will slow down the reaction. This is prevented by an excess of target. Just over 6 nM of DNA is cleaved at a factor 1600 dilution of the stock (0.625 µL/mL), which indicates that the amount of active protein in the batch is actually quite low: about 10 µM.

Kinetic assay

3 nM (final concentration) of FnCas12a was mixed with 9 nM crRNA (final concentration), equilibrated at 37°C and mixed with 4 nM target (final concentration). Samples were taken at 1, 2, 5, 10 and 30 minutes and analysed (Figure 7.4B). The cleaved fraction was then normalised for the maximum cleaved fraction at 30 minutes. In this case, cleavage is quite fast and the reaction is already over after only 1 minute.

Chapter 8

Good guide, bad guide: Spacer sequence-dependent cleavage efficiency of Cas12a

Sjoerd C.A. Creutzburg, Wen Y. Wu, Prarthana Mohanraju, Thomas Swartjes, Ferhat Alkan, Jan Gorodkin, Raymond H.J. Staals, John van der Oost

Abstract

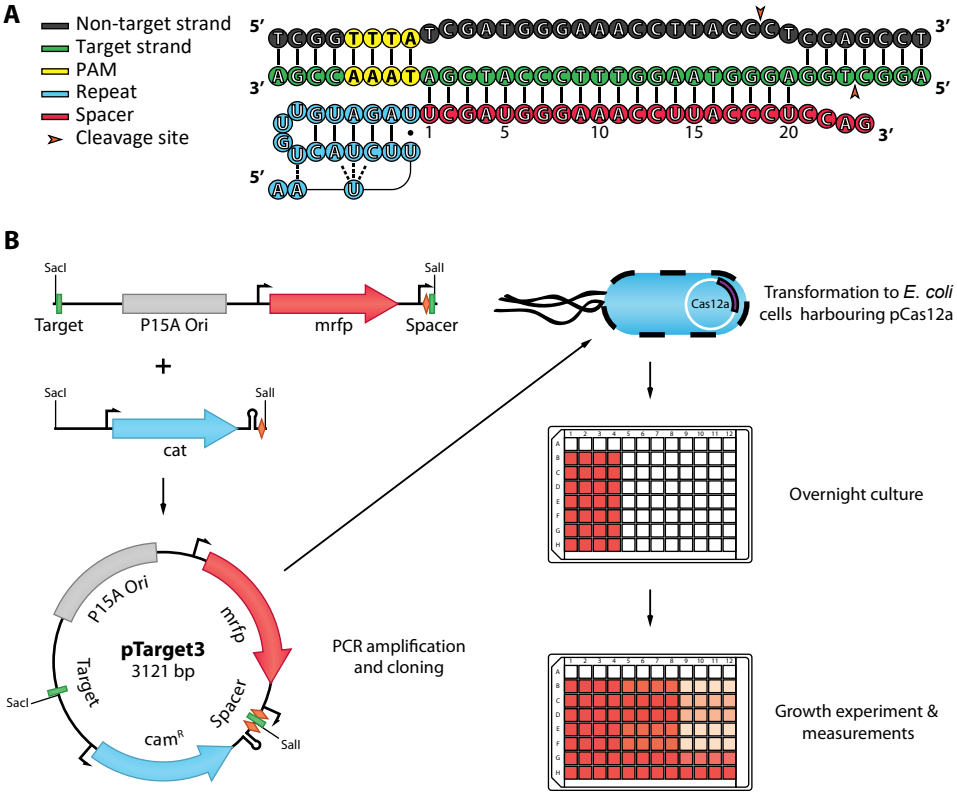
Genome editing has recently made a revolutionary development with the introduction of the CRISPR-Cas technology. The programmable CRISPR-associated Cas9 and Cas12a nucleases generate specific dsDNA breaks in the genome, after which host DNA-repair mechanisms can be manipulated to implement the desired editing. Despite this spectacular progress, the efficiency of Cas9/Cas12a-based engineering can still be improved. Here, we address the variation in guide-dependent efficiency of Cas12a, and set out to reveal the molecular basis of this phenomenon. We established a sensitive and robust *in vivo* targeting assay based on loss of a target plasmid encoding the red fluorescent protein (mRFP). Our results suggest that folding of both the precursor guide (pre-crRNA) and the mature guide (crRNA) have a major influence on Cas12a activity. Especially, base pairing of the direct repeat, other than with itself, was found to be detrimental to the activity of Cas12a. Furthermore, we describe different approaches to minimise base-pairing interactions between the direct repeat and the variable part of the guide. We show that design of the 3' end of the guide, which is not involved in target strand base pairing, may result in substantial improvement of the guide's targeting potential and hence of its genome editing efficiency.

Introduction

The CRISPR-associated nucleases Cas9 and Cas12a (formerly known as Cpf1 (2)) are distinct types of crRNA-guided DNA endonucleases that have been developed into powerful genome editing tools (104–108). Cas9 and Cas12a have rapidly become popular tools for a broad spectrum of genetic engineering applications (109–113), based on the successful heterologous expression of these Cas nucleases and on the relatively easy adjustment of their specificity through exchanging their crRNA guides. The formation of functional crRNAs relies on the conversion of precursor RNA (pre-crRNA) to mature crRNA. In the case of Cas9, the repeat parts of the pre-crRNA are recognised by partly complementary trans-acting crRNAs (tracrRNA). In the presence of Cas9, base pairing between the pre-crRNA repeats and the tracrRNA anti-repeats results in local dsRNA fragments that are specifically cleaved by RNaseIII. After cleavage, the crRNA-tracrRNA pair remains stably bound by Cas9 (114) (Supplementary figure 8.1A). To allow for easy crRNA adjustment for Cas9, a synthetic loop has been introduced to connect the crRNA repeat part with the tracrRNA anti-repeat fragment, resulting in a single-guide RNA (sgRNA) (115). In case of Cas12a, however, tracrRNA and RNaseIII are not involved in crRNA maturation. Cas12a directly associates with the pre-crRNA, most likely through recognising the typical pseudoknot-type hairpin structure of the repeat fragments, after which maturation of the crRNA is catalysed by a dedicated catalytic ribonuclease domain of Cas12a (2, 116–118) (Figure 8.1A).

A general issue for the application of both Cas9 and Cas12a nucleases appears to be the unpredictable success of crRNA design and target selection, often resulting in designing 3-4 crRNAs for target a single gene. On the one hand, this problem may be caused by differences in local chromatin structure that may severely affect the accessibility of chromosomal targets (100, 101). On the other hand, it may be caused by the nucleotide composition of the variable parts of the crRNAs. Based on genome-wide guide library screens, different algorithms and scoring systems have been developed to predict crRNA performance of Cas9 (101, 119–126). The secondary structure of the crRNA has been proposed to be a major player in crRNA performance (127), potentially resulting in poor cleavage activity (128). Also, in case of Cas12a, the editing efficiency varies substantially depending on the design of the crRNA (129). In an attempt to predict the guide functionality, a recent analysis of crRNA-dependent targeting activity of Cas12a from *Acidaminococcus spec.* (AsCas12a) and Lachnospiraceae bacterium (LbCas12a) has been used to compose an algorithm for crRNA design (129).

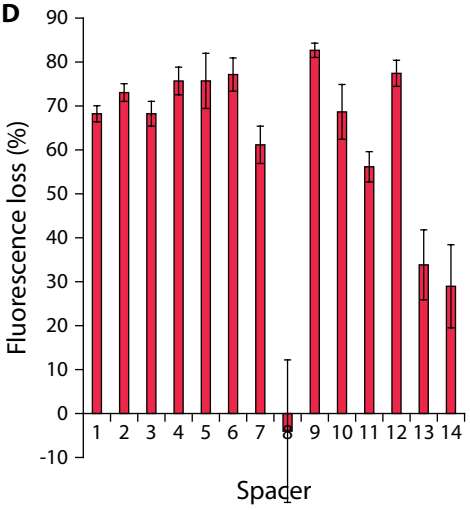
Although it is known that the spacer sequence of the crRNA may affect target cleavage efficiency, the molecular basis of this phenomenon remains unclear. An important feature of the Cas9 and Cas12 crRNAs is the formation of well-conserved secondary structures of their invariable sequences, that most likely allows for specific protein-RNA recognition



C

	Seed																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Sp1	U	C	G	A	U	G	G	G	A	A	A	C	C	U	U	A	C	C	C	U	C	C	A	G
Sp2	G	A	C	U	U	A	C	C	A	A	U	G	A	G	C	A	C	G	U	G	G	A	G	U
Sp3	A	A	U	A	U	U	C	C	C	C	A	G	G	U	U	C	A	G	U	G	G	A	G	U
Sp4	A	C	A	C	A	C	U	G	C	A	A	U	U	C	A	G	G	U	U	G	G	A	G	U
Sp5	A	C	C	A	G	U	C	C	A	U	C	A	A	U	U	G	U	A	G	U	G	G	A	G	U
Sp6	A	C	G	C	C	U	A	C	G	A	G	A	U	C	A	U	U	G	U	G	G	A	G	U
Sp7	U	G	C	U	C	C	C	G	U	A	A	C	A	C	U	A	G	C	U	G	G	A	G	U
Sp8	C	C	A	G	U	C	U	A	A	A	U	C	G	C	U	U	A	C	U	G	G	A	G	U
Sp9	A	C	U	A	U	A	U	G	C	A	C	C	G	G	A	G	C	U	U	G	G	A	G	U
Sp10	C	A	C	A	A	U	U	A	C	C	G	G	U	C	C	C	A	A	U	G	G	A	G	U
Sp11	C	C	A	G	U	C	U	A	A	C	C	C	G	C	U	U	A	C	G	U	G	A	G	U
Sp12	C	C	A	G	U	C	U	A	A	C	A	C	G	C	U	U	A	C	G	U	G	A	G	U
Sp13	C	C	A	G	C	C	A	A	C	C	C	C	G	C	U	U	A	C	G	U	G	A	G	U
Sp14	C	C	A	G	C	C	A	A	C	C	A	C	G	C	U	U	A	C	G	U	G	A	G	U

D



E

	Seed																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Sp8	C	C	A	G	U	C	U	A	A	**A**	**U**	C	G	C	U	U	A	C	**U**	**G**	G	A	G	U
Sp12	C	C	A	G	**U**	C	**U**	A	**A**	**C**	**A**	C	G	C	U	U	A	C	**G**	**U**	G	A	G	U
Sp14	C	C	A	G	**C**	C	**A**	A	**C**	C	A	C	G	C	U	U	A	C	G	U	G	A	G	U

Figure 8.1. Analysis of Cas12a cRNA performance. Cas12a guide bound to target, workflow schematic and measurement of Cas12a activity using crRNAs with 14 different spacers. (A) Cas12a binds the repeat of the pre-crRNA, which forms a typical pseudoknot structure (blue). It recognises a TTTV PAM (Yellow) and forms an R-loop with the spacer part of the crRNA (red) and the target strand of the DNA (green). Cleavage occurs mostly after positions 18 on the non-target strand and 23 on the target strand. (B) Target plasmid construction starts with a PCR, to incorporate a certain spacer (green box labelled "Spacer") downstream from the leader and the first repeat (orange diamond) and a matching target sequence (green box labelled "Target") on opposite sides of the origin of replication (ori) and *mrfp* gene. A second SacI-SalI fragment consists of the chloramphenicol resistance marker (cat) and the second repeat (orange diamond). The target plasmid is obtained by digestion of both fragments (SacI and SalI) and ligation. The plasmid is then transformed to an *E. coli* strain containing a plasmid pCas12a that allows for expression of FnCas12a with an *ssrA* tag upon induction by L-rhamnose. After transformation, cells are grown overnight in liquid medium, followed by inoculation in fresh medium containing L-rhamnose to induce Cas12a. Fluorescence of mRFP, which is expressed constitutively, is then measured to assess cleavage efficiency. (C) The different spacer sequences (Sp1-Sp14). The seed sequence is indicated by the green shade. (D) Cleavage efficiency of Cas12a shown in terms of fluorescence loss for 14 different spacers in pTarget3. Average values from three biological replicates are shown, with error bars representing SD. (E) Alignment highlighting differences between the sequence of Sp8 and Sp12 (red), and between Sp12 and Sp14 (blue).

and eventually for stable association of the crRNA and its partner nuclease. In case of Cas12a, perturbations in the hairpin/pseudoknot at the 5' part of the crRNA (Figure 8.1A) most likely interfere with the complex formation of Cas12a and its crRNA guide. Hence, predicting cleavage efficiency solely based on spacer-target complementarity is insufficient, as also potential disruption of the pseudoknot should be taken into account. Unfortunately, the reliability of currently available tools for predicting the secondary structure of individual small RNA molecules is relatively low.

In this study, we aimed to reveal the molecular basis of the aforementioned variability of crRNA guide performance of Cas12a. We initially established a sensitive and robust mRFP-based fluorescence-loss assay in *Escherichia coli* to monitor the *in vivo* targeting efficiency of Cas12a from *Francisella tularensis subsp. novicida* (FnCas12a). This system was used to analyse how different spacer sequences and (pre-)crRNA variations affect target cleavage efficiency. We found that the effect on target cleavage by a single nucleotide change in a spacer often depends on its surrounding nucleotides. This observation suggests that these effects are not caused by direct nucleotide-protein interactions as previously proposed (129), but rather by the formation of distinct secondary structures of the closely related crRNA variants. Interestingly, we found that efficient targeting requires only 19 nucleotides of base pairing between the crRNA and the target strand (Supplementary figure 8.2), even though position 20 can base pair as well (118). We also found that the last nucleotides of the spacer (position 20 and onwards) can be rationally modified to shift the folding equilibrium from an inappropriate fold, which decreases its efficiency, towards the optimal pseudoknot structure, resulting in the conversion of poorly-performing crRNAs to crRNAs with improved target cleavage efficiency. Our findings contribute to a better understanding of spacer-sequence dependent cleavage efficiencies, and provide design strategies to improve crRNA performance in general, and in Cas12a in particular.

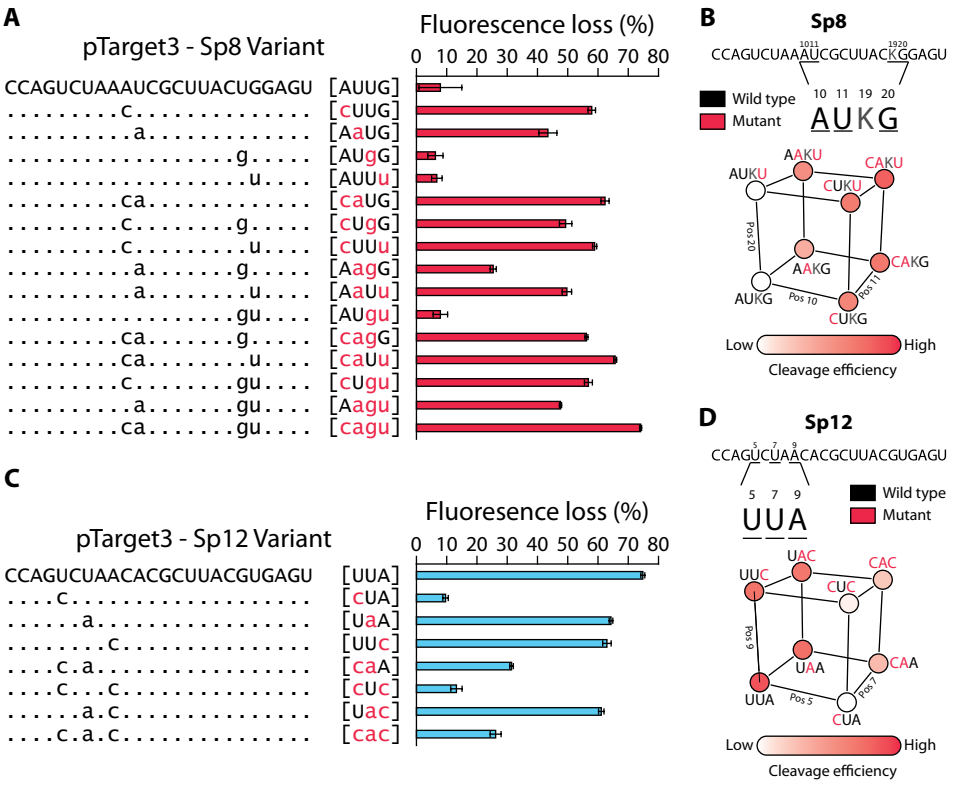

A
pTarget3 - Sp8 Variant
Fluorescence loss (%)
0 10 20 30 40 50 60 70 80
CCAGUCUAAAUCGCUUACUGGAGU [AUUG]
.........c.............. [cUUG]
.........a.............. [AaUG]
.................g..... [AUgG]
..................u.... [AUUu]
.........ca............. [caUG]
.........c........g..... [cUgG]
.........c.........u.... [cUUu]
.........a.......g..... [AagG]
.........a........u.... [AaUu]
..................gu.... [AUgu]
.........ca.......g..... [cagG]
.........ca........u.... [caUu]
.........c........gu.... [cUgu]
.........a.......gu.... [Aagu]
.........ca.......gu.... [cagu]
B
Sp8
CCAGUCUAAAUCGCUUACKGGAGU
1011
1920
10 11 19 20
AUKG
Wild type
Mutant
AAKU
AUKU
CAKU
CUKU
AAKG
CAKG
AUKG
CUKG
Pos 20
Pos 11
Pos 10
Low
High
Cleavage efficiency
C
pTarget3 - Sp12 Variant
Fluoresence loss (%)
0 10 20 30 40 50 60 70 80
CCAGUCUAACACGCUUACGUGAGU [UUA]
....c.................. [cUA]
......a................ [UaA]
........c.............. [UUc]
....c.a................ [caA]
....c...c.............. [cUc]
......a.c.............. [Uac]
....c.a.c.............. [cac]
D
Sp12
CCAGUCUAACACGCUUACGUGAGU
5 7 9
UUA
Wild type
Mutant
UAC
CAC
UUC
CUC
UAA
CAA
UUA
CUA
Pos 6
Pos 5
Pos 7
Low
High
Cleavage efficiency
E
Active guide
Leader
Repeat
Repeat
Spacer
SalI
Sp8
Sp12
Sp14

Figure 8.2. Comparison of Sp8, Sp12 and Sp14. (A) Cleavage efficiency shown in terms of fluorescence loss for different Sp8 variants in pTarget3. Average values from three biological replicates are shown, with error bars representing SD. The wild type and mutated Sp8 nucleotide sequences are shown in upper case black and lowercase red letters, respectively. In brackets [NNNN] are the nucleotides in position 10, 11, 19 and 20. (B) A 3-D representation of the data in panel A. Cleavage efficiency is represented by the intensity of the colour red for different spacers shown in a 4 letter code, which are the nucleotides in positions 10, 11, 19 and 20. K (G or U) at position 19 is constant and the average was taken to generate the red intensity for each spacer variant. The wild type and mutated Sp8 nucleotide sequences are shown in black and red, respectively. Each corner of the cube represents a certain spacer sequence, and moving along either of the 3 axes changes the sequences at one nucleotide position only. (C) Cleavage efficiency shown in terms of fluorescence loss for different Sp12 variant in pTarget3. In brackets [NNN] are the nucleotides at positions 5, 7 and 9. (D) A 3-D representation of the data in panel C similar to panel B. (E) Predicted crRNA structures for Sp8, Sp12 and Sp14. Folding of an active crRNA is given on the left and the folding based on prediction are given on the right. Each spacer is in equilibrium with its active state and inactive state as indicated by the arrow. Thicker arrows represent that the equilibrium is shifted more towards a certain state.

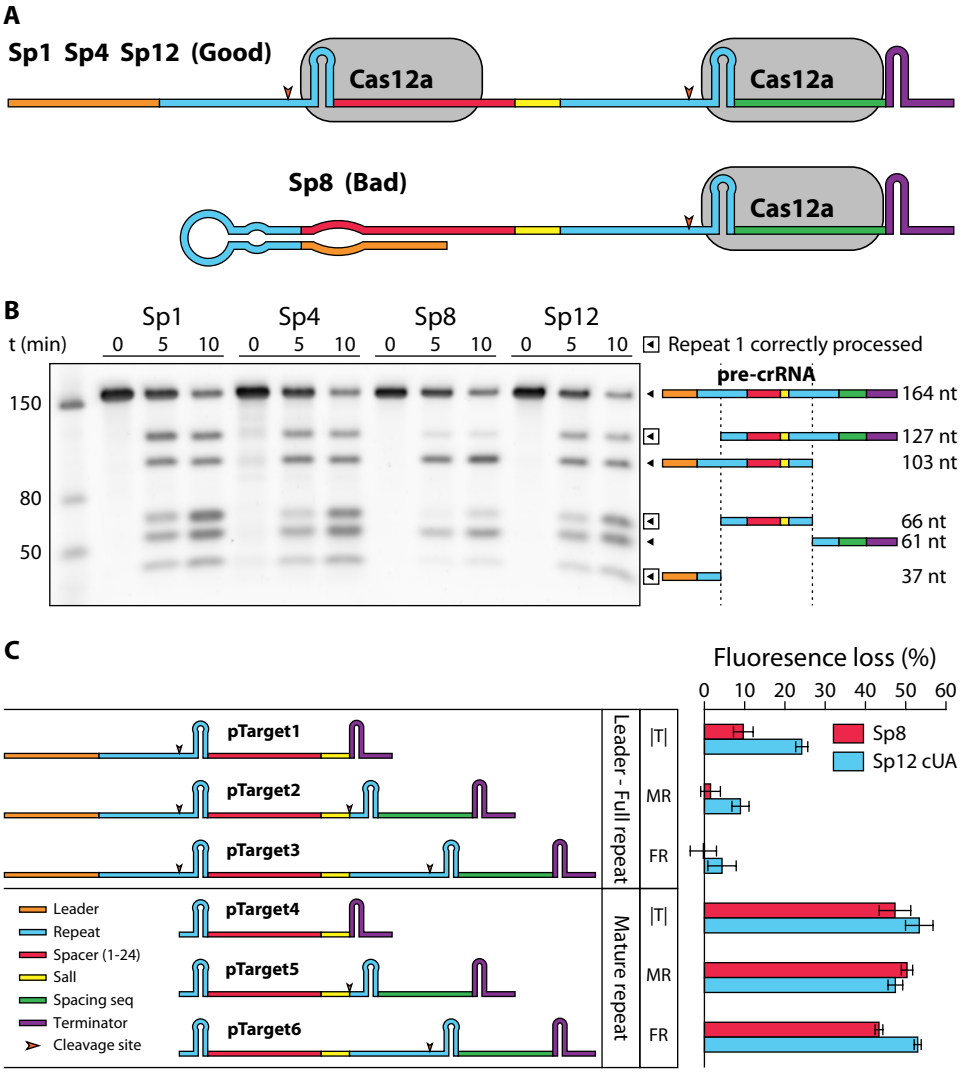
Results

Sequence of the spacer affects cleavage efficiency

A set of 10 crRNAs with different spacers of similar composition (Sp1-10; Figure 8.1C) was initially used for their *in vivo* functionality to allow Cas12a to target complementary plasmid-borne sequences in *E. coli*. The spacer transcripts had a 5' leader sequence, followed by the spacer that was flanked by a full repeat on both ends (L-FR-Sp-FR; pTarget3; Table 1), which most closely resembles the original, native array (2). The observed targeting functionality (fluorescence loss) of nine spacers was in the range of 60-85%, with the exception of the poorly performing Sp8 (0%) (Figure 8.1D). To reveal the molecular basis of this phenomenon, four related spacers were designed (Sp11-14). This resulted in an interesting set of three closely related spacers with major differences in performance (Figure 8.1D/1E). The most dramatic difference was observed between Sp8 and Sp12 that, although they differ only by four nucleotides, perform either very badly (Sp8, no fluorescence loss) or very well (Sp12, among the fastest). It should be noted that, with this set of crRNAs, similar trends in targeting efficiencies were observed for FnCas12a and AsCas12a (Supplementary figure 8.5).

Fluctuations in cleavage efficiency with single nucleotide changes

Since Sp8 and Sp12 have the same seed (nucleotides 1-5), the composition of the seed sequence appears not to play a role in the observed differences in cleavage efficiency. Sp8 and Sp12 differ in only four nucleotides (at positions 10, 11, 19, 20; Figure 8.1E). To reveal which of these nucleotides are responsible for the major difference in targeting efficiency, we made a library to systematically test all 16 combinations (four nucleotide positions and two possible nucleotides per position) between Sp8 and Sp12. In addition, to shed light



on the influence of the 3' end of the pre-crRNA, libraries were generated in two different CRISPR designs. In one library the L-FR-Sp-FR (pTarget3; Table 1) crRNA design (Figure 8.2A) was used, whereas another library was made using the L-FR-Sp-|T| (pTarget1; Table 1) crRNA design (Supplementary figure 8.6A), which minimises the influence of the 3' end. Overall, similar trends were observed for Sp8 variants in the two libraries. For the Sp8 variants, position 19 appeared of least influence on crRNA performance, and in Figure 8.2B the base at this position is referred to as "K" (G or U) for clarity purposes. Screening of the library indicated that positions 10 and 11 are almost solely responsible for the low activity

Figure 8.3. Cleavage efficiency for spacers with different upstream and downstream sequences. (A) According to our hypothesis, the Sp1, Sp4 and Sp12 will have efficient processing for both repeats. The Sp8 causes misfolding of the pseudoknot and thereby impedes the processing of the first repeat. (B) In vitro pre-crRNA processing of Sp1, Sp4, Sp8 and Sp12 at different time intervals. The marker lane on the left shows three bands from the low range ssRNA ladder (NEB). The bands of 127 nt, 66 nt and 37 nt indicate processing of the pre-crRNA that would lead to a targeting RNP. The processing of the second repeat (103 nt, 61 nt) gives rise to a non-targeting RNP. (C) Two relatively inefficient spacers were tested in different pre-crRNA architectures (pTarget1 - pTarget6): Sp8 (red bars) and Sp12-variant [cUA] (blue bars). Average values from three biological replicates are shown, with error bars representing SD. A 20 nt leader-end is shown in orange, a full or mature repeat in blue, a 24 nt spacer sequence in red, a SalI restriction site in yellow, a spacing sequence in green and the terminator is shown in purple. Spacers were flanked on the 5' end with a leader-end sequence and full repeat, or they were flanked by a mature repeat only. Within each category were spacers containing various downstream sequences. Sequences such as, only a terminator (|T|), a mature repeat-spacing sequence-terminator (MR) and a full repeat-spacing sequence-terminator (FR).

observed for Sp8 (Figure 8.2B). Surprisingly, only spacers containing the combination of A10 and U11 showed very low efficiency, independent of variation at position 20. Even changing either one of the positions, A10C or U11A, yielded moderate to high cleavage efficiency. The presence of the A10-U11 pair in the efficient Sp2 (Figure 8.1C/1D), strongly suggests that its negative impact in Sp8 is not position dependent (interaction between the crRNA and the Cas12a protein), but rather context dependent (intramolecular interactions in the crRNA).

Another major difference in cleavage efficiency was observed for the related spacers Sp12 and Sp14 (Figure 8.1D). These sequences differ at only three positions (5, 7 and 9; Figure 8.1E), so we made a variant library to pinpoint the determining nucleotides for both the L-FR-Sp-FR (pTarget3) design (Figure 8.2C) and the L-FR-Sp-|T| (pTarget1; Supplementary figure 8.6B). Both in Sp12 and Sp14, position 5 is most important, while the impact of position 7 depends on its surrounding nucleotides, and position 9 appears to be the least influential on activity (Figure 8.2B). The highly efficient Sp12 was severely affected by changing position 5 (U5C) (Figure 8.2D). The analysis of both libraries revealed similar trends for most variants, indicating that cleavage efficiencies can be substantially influenced by a single nucleotide change within a given position in a spacer. As noted earlier, however, the presence of C5 in the efficient Sp6 (Figure 8.1C/1D), suggests that the negative impact of C5 in Sp14 is not position dependent. Rather, this effect may depend on the surrounding nucleotides. These examples of context dependence indicate that differences in crRNA secondary structure correlate with fluctuating Cas12a cleavage efficiencies.

To check the potential involvement of crRNA secondary structure, attempts were made to predict the folding of the pre-crRNAs of Sp8, Sp12 and Sp14 (Figure 8.2E) with RNA secondary structure prediction tools RNAfold (130) and mFold (81). While the folding

energies obtained by both tools are not exactly the same, the generated structures are similar. Instead of the pseudoknot that is required for an active Cas12a-crRNA complex, the analysis suggested a strong alternative structure of the pre-crRNA (Sp8) that is formed by base pairing between the leader and the spacer. The alternative structure of the Sp8 pre-crRNA is stabilised by a long stem, composed of mainly A•U pairs. Interrupting this stem, immediately causes the equilibrium to shift to the active fold (A10C or U11A) (Figure 8.2C/2D). In Sp12, the alternative structure is neither stabilised by a very long nor a very strong stem, but the U5C mutation changes that drastically. While the Sp12 [cUA] variant had almost no activity left, the activity could be partially restored by U7A (Figure 8.2C/2D). Without the U5C mutation, the U7A should sufficiently destabilise a stem that requires no further destabilisation, and indeed we saw limited effect of U7A in that case (Figure 8.2C/2D/2E). The A9C mutation has a counter-intuitive effect on Sp12. While the A9C will shorten the stem and destabilise it in the Sp12 [cUA] variant, it has a slight negative effect on the other three variants ([UUA], [UaA] and [caA]) (Figure 8.2C/2D). Whereas our current understanding is inadequate to explain this phenomenon, it might be related either to alternative base pairing or to other inactive folds.

Sequences directly flanking the spacer affect pre-crRNA processing

When the mature crRNA is bound to the Cas12a protein, the repeat has a pseudoknot structure (118). Disrupting this structure is anticipated to affect the binding of the pre-crRNA to the Cas12a protein. To demonstrate this, an *in vitro* processing assay was performed (Figure 8.3A/3B), revealing that the 164 nt pre-crRNA of pTarget3 is not cleaved very well at the first repeat for Sp8. Other spacers (Sp1, Sp4, Sp12) showed good processing of both repeats. This agrees well with the model that the lack of pseudoknot formation does affect binding and processing of the pre-crRNA, resulting in reduced levels of mature crRNA and, hence, impaired cleavage efficiency by Cas12a.

We then set out to systematically test the influence of different upstream and downstream sequences on the spacer cleavage efficiency of two bad spacers, the original Sp8 (Figure 8.1C) and a poorly performing Sp12-variant ([cUA]; Figure 8.2C). For both spacers, six different constructs were made and tested (pTarget1 to pTarget6; Figure 8.3C). The presence of a leader sequence upstream of the repeat-spacer resulted in a major reduction of the Cas12a cleavage activity, with similar trends for both spacers. In contrast, different sequences downstream of the spacer led to relatively minor fluctuations in cleavage efficiency for both spacers (Figure 8.3C). This indicated that the sequence context of the precursor and/or mature crRNA may seriously impact its performance. In this case, omitting the upstream leader sequence that probably disturbs the formation of the desired pseudoknot structure (Figure 8.2E) resulted in substantial restoration of crRNA performance.

crRNA folding can cause unstable pseudoknot formation

Assuming that at the pre-crRNA stage, correct folding of the pseudoknot is key for appropriate docking in Cas12a and hence for eventual cleavage efficiency, functionality of a guide correlates with its potential to form a pseudoknot. In the case of Sp8 where the leader is impeding pseudoknot formation, the nucleotides that caused the impediment, could be substituted or even omitted. When the crRNA nucleotides causing pseudoknot disruption are actually involved in base pairing with the target strand (Figure 8.1A), they cannot be changed or omitted. An alternative strategy to enhance pseudoknot formation nonetheless, would be to force those nucleotides to base pair with nucleotides at the crRNA 3' end instead. Since adding secondary structure may cause issues of its own, it was assessed whether crRNA performance could be influenced by masking certain parts of the crRNA through designed intra-molecular base pairing with its own 3'-sequence. To increase the chance of back-folding, the distance between the 3' tail and the masked part should be as small as possible. Therefore, instead of adding 5 nucleotides to the spacer, we replaced the last 5 nucleotides. Although it is known that that Cas12a does not need the full 23 nt spacer (2), it was important to ensure that shortening the base pairing of spacer and target to 19 base pairs did not influence the activity. Hence, we conducted a pilot experiment with pTarget3 - Sp4, where the target (not the crRNA) was complementary to the crRNA up to position 19. Under these conditions, 19 base pairs appeared to be sufficient for maximal cleavage efficiency (Supplementary figure 8.2). This was confirmed by a more elaborate experiment with different crRNA lengths (Supplementary figure 8.8).

Next, we generated a new spacer that had as few (predicted) base pairs in the first 19 nucleotides as possible, approximately 50% GC-content and no single base stretches longer than 3 (Back-fold lib A). We then created a library of crRNAs in pTarget1 (Figure 8.4A) with variable spacer tail sequences to mask specific positions, ranging from the direct repeat (DR; position -3) to position 11 of the spacer (selected examples are depicted in Figure 8.4C/4D/4E).

Back-fold lib A [DR] (Figure 8.4C) results in base pairs between positions 1-6 with positions 20-25. The SalI site starting at position 26 then base pairs with the last three nucleotides of the direct repeat (DR). The rest of the library started masking at position [n] and had the last base pair 4 nucleotides downstream. The nucleotide at position 25 was designed to force a mismatch with the nucleotide at [n-1], in an attempt not to extend the base pairing beyond 20-24; for example, the design of crRNA Back-fold lib [2] was such that positions 2-6 base paired with positions 20-24, and that the nucleotide at position 1 [2-1] mismatched with that at position 25. Likewise, in Back-fold lib A [4] the masking started at position 4, ended with position 8, and the nucleotide at position 3 [4-1] had a mismatch with the nucleotide at position 25. To minimise the chance of the spacer 3' end base pairing with the 3' end of the transcript, the crRNA design was L-FR-Sp-|T| (pTarget1). In the generated

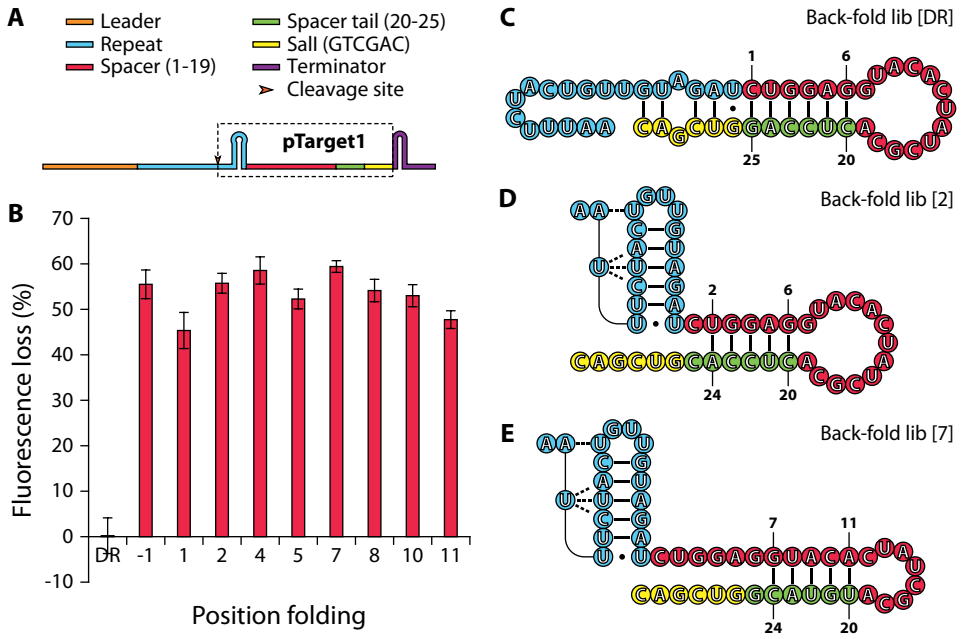


Figure 8.4. Cleavage efficiency for folding library. (A) A library of designs has been made in pTarget1. Panels C-E will only show the boxed part. (B) Cleavage efficiency shown in terms of fluorescence loss for different crRNA variants from Back-folding library A (Back-fold lib A). Average values from three biological replicates are shown, with error bars representing SD. (C-E) The intended folds of crRNAs of Back-fold lib [DR], Back-fold lib [2] and Back-fold lib [7]. The number in brackets indicates the starting position of the base pairing.

library, spacers were designed from random sequences and selected for showing minimal predicted secondary structure, roughly 50% GC and unique library members.

For Back-fold lib A [DR] (Figure 8.4C), we observed that base pairing with the direct repeat (almost) abolished Cas12a activity (Figure 8.4B). Mismatching of position 25 and position 1, as is the case in Back-fold lib A [2], negated this effect completely (Figure 8.4B/4D). Unexpectedly, there appears to be no trend in cleavage efficiency between the different designs of Back-fold lib A [-1] to [11] (Figure 8.4B). Cleavage efficiencies fluctuated between 40-60% for all constructs within one library. In particular, Back-fold lib A [-1] showed no diminished activity even though the -1 position is considered to be part of the pseudoknot with a U•U pair (118). Possibly, the likelihood of base pairing between the 5′ and 3′ ends of the crRNA is reduced because of the distance between the positions. We also tested two other folding libraries made from the selection of spacer bases (Supplementary figure 8.7), and again did not observe a trend between masking of a specific position and cleavage efficiency, apart from the [DR] constructs.

Hence, base pairing with the direct repeat does completely abolish Cas12a activity, supporting the model that inadequate pseudoknot formation is detrimental for crRNA

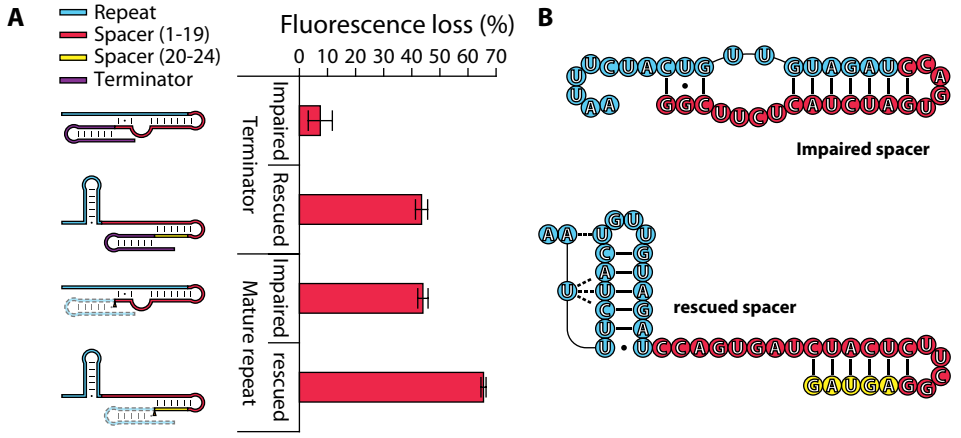


Figure 8.5. Rescuing "impaired" spacers. (A) Cleavage efficiency is shown in terms of fluorescence loss for various crRNA variants. The mature repeat is shown in blue, the 1-19 nt of spacer sequence in red, the 20-24 nt positions are shown in yellow. The crRNAs are categorised as "impaired" and "rescued" either with (pTarget7) or without (pTarget8) a 3' terminator sequence. Average values from three biological replicates are shown, with error bars representing SD. (B) Detail of the structures in (A).

functionality. On the other hand, these results show that crRNA structures that mask spacer positions [-1] to [11] do not abolish Cas12a activity; as discussed below, this may be useful for rescue of bad crRNAs.

Rescuing "impaired" spacers

Certain spacer sequences may cause crRNAs to adopt a fold that disrupts the pseudoknot structure, resulting in hampered Cas12a binding and poor targeting activity. In an effort to rescue such spacers, we adjusted the 3' end of the crRNA sequences such that any base pairing with the direct repeat is mitigated. This should favour pseudoknot formation. To test this, two additional crRNAs were designed. One crRNA has a spacer length of 19 nucleotides and base pairing between spacer positions 6-12 and the complementary nucleotides in the direct repeat, which cause an alternate fold that is unlikely to associate properly with the Cas12a protein, resulting in the "impaired" spacer (Figure 8.5B; impaired). The other crRNA has the same spacer of 19 nucleotides with an additional five nucleotides at the 3' end, that are designed such that they may allow for intramolecular base pairing with the spacer sequence from 9-14, avoiding the pseudoknot disruption, and thus converting the "impaired" spacer into a "rescued" one (Figure 8.5B; rescued). Since every nucleotide in the transcript may influence the folding, the spacers were flanked with a mature repeat at the 5' end. Downstream of the spacer, a flanking sequence was included either with a terminator (pTarget7-IS and pTarget7-IS-rescued), or with a second repeat and a terminator (pTarget8-IS and pTarget8-IS-rescued). While the former will retain its terminator, the latter will be recognised and further processed by Cas12a, so it does not

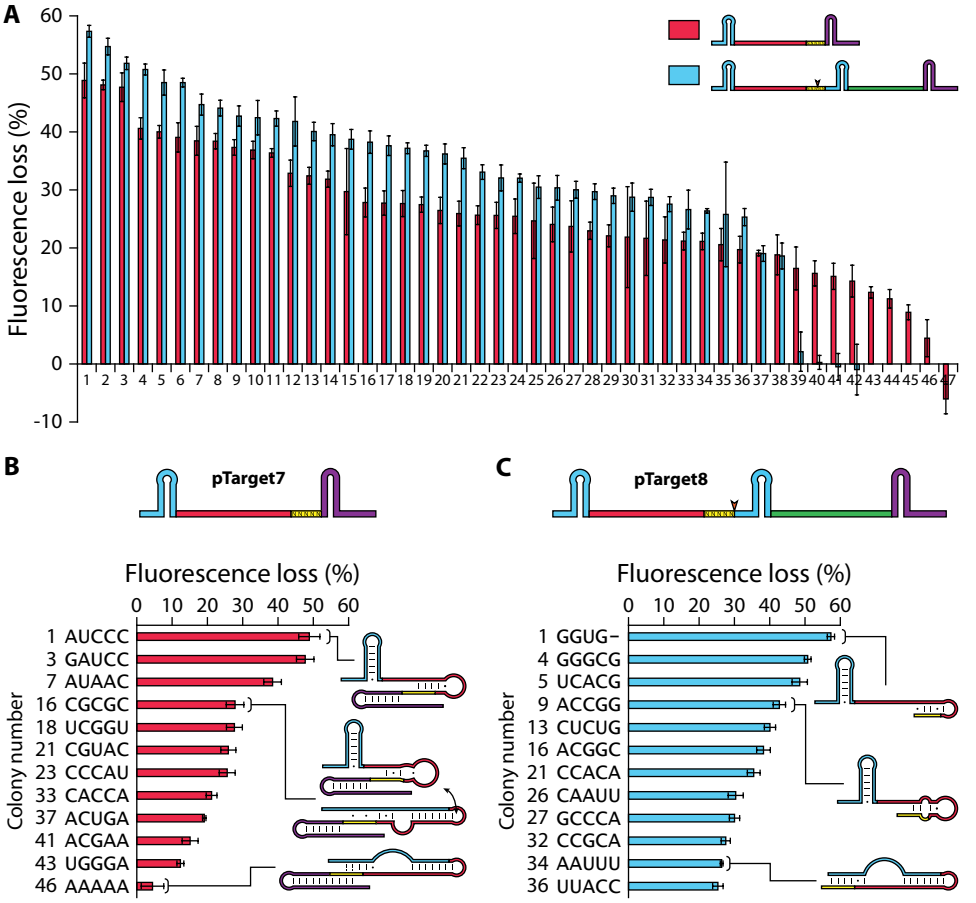


Figure 8.6. Cas12a activity of two "impaired" spacer (IS) libraries varying position 20-24. (A) Cleavage efficiency shown in terms of fluorescence loss of 47 randomly selected colonies for pTarget7-IS-N5 and pTarget8-IS-N5. Non-fluorescent clones were omitted. The target and first 19 nucleotides of the spacer are the same as Figure 8.5. (B) Cleavage efficiency shown in terms of fluorescence loss from sequenced variants of pTarget7-IS-N5 with the given possible crRNA structure of variants 1, 16 and 46. The crRNAs contain a mature direct repeat (blue), a spacer (red), a variable sequence at position 20-24 (yellow) and a terminator (purple). (C) Cleavage efficiency shown in terms of fluorescence loss from sequenced variants of pTarget8-IS-N5 with the given possible crRNA structure of variants 1, 9 and 34. The crRNAs contain a mature direct repeat (blue), a spacer (red), a variable sequence at position 20-24 (yellow), a spacing sequence (green) and a terminator (purple). The cleavage position of the crRNA processing is indicated with an orange arrow.

contain a mature repeat-terminator sequence during DNA targeting.

The "impaired" spacer, either with or without terminator, has a lower cleavage efficiency than its "rescued" counterpart (Figure 8.5A). Unlike aforementioned designs (Figure 8.3B), the spacers with terminator overall had a much lower efficiency than the ones with a mature repeat. The difference is that the terminator was positioned slightly closer to the

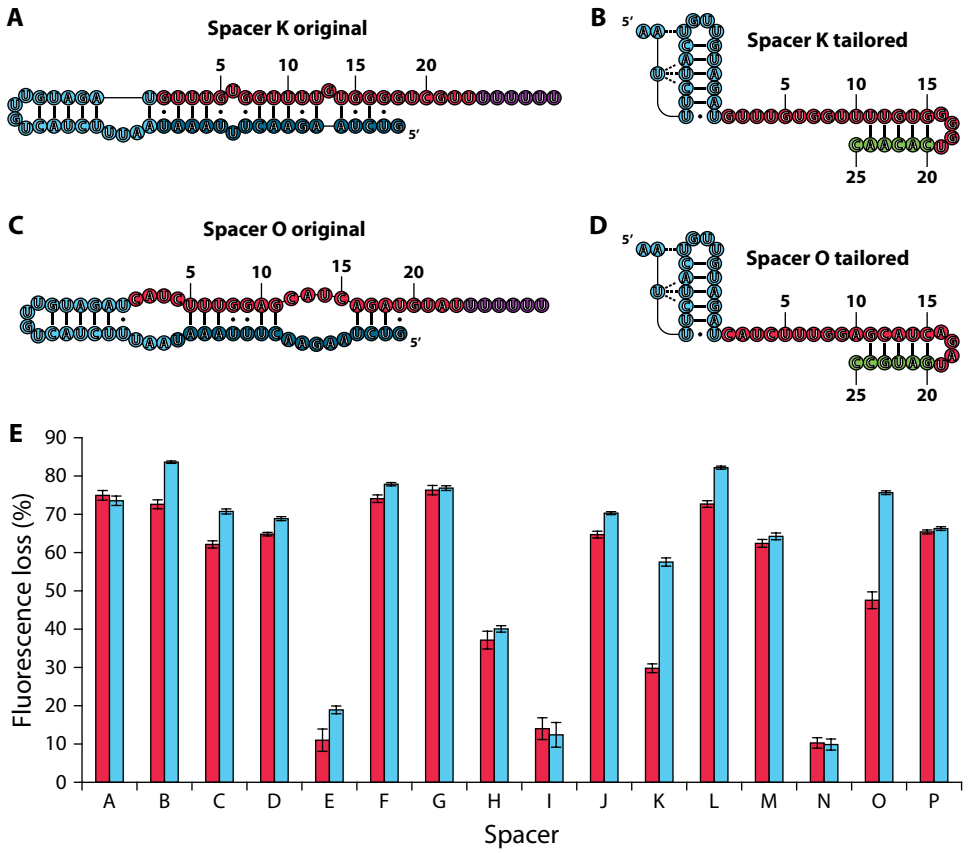


Figure 8.7. Tailoring pre-crRNA design for reported "bad" spacers. (A) Predicted structure of spacer K pre-crRNA as used by Kim et al. 2018. (B) Tailored pre-crRNA design of spacer K. (C) Predicted structure of spacer O pre-crRNA as used by Kim et al. 2018. (D) Tailored pre-crRNA design of spacer O. (E) Performance of the reported "bad" spacers in the fluorescence loss assay. Original versions are in red and tailored versions are blue.

R-loop in pTarget7 (Figure 8.5A) than it was in the pTarget1 or pTarget4 constructs (Figure 8.3B), due to the SalI site in the latter two designs. The terminator was even closer for the "impaired" spacer than it is for the "rescued" spacer, since the former was 5 nucleotides shorter.

To further prove that the observed increase in cleavage efficiency (Figure 8.5A) was not just caused by an increase of spacer length but rather by improved folding, we constructed two libraries with randomised tails using the "impaired" spacer as a base. Each library contained the "impaired" spacer of 19 fixed nucleotides and 5 variable nucleotides (NNNNN) at positions 20-24, followed by either the terminator (pTarget7-IS-N5) or a mature repeat (pTarget8-IS-N5). Within each library, 47 colonies were randomly selected and their cleavage efficiency was assessed (Figure 8.6A). Out of the 47, we selected 12 spacers that

covered the whole range of cleavage efficiencies. Those were sequenced and their RNA structures were predicted by RNAfold (130). Again, some general trends are observed that support the correlation between pseudoknot formation and crRNA performance (fluorescence loss). In case of low efficiency spacers, base pairing with the direct repeat appears to be enhanced, hampering formation of the pseudoknot. In case of relatively efficient spacers, the proper pseudoknot structure is not challenged due to intra-spacer base pairing involving the variable positions (20-24) and (part of the) positions 6-12 (Figure 8.6B/6C). The sequences that show intermediate efficiency seemed to correspond to spacers of which the pseudoknot/alternate states are not favoured either way.

To estimate the contribution of pre-crRNA misfolding to the complete Cas12a cleavage efficiency, we selected 16 spacers from a previous study (131) that performed poorly as judged from relatively low indel formation, despite the fact that the accessibility of corresponding human target genes was good. Apart from the original guides, tailored pre-crRNA designs were made (as described above) with nucleotides 20-24 base pairing with nucleotides 11-15 on every spacer and with a mismatch between nucleotides 10 and 25. Under control of the U6 promoter, transcription starts at the beginning of the full repeat and it ends at a polyU tail. A direct mimic in bacteria was impossible as the *E. coli* RNA polymerase does not stop at polyU when it is not preceded by a strong stem-loop. The closest approximation is adding the polyU tail to the spacer and mimic the termination by processing of a second mature repeat (FR-Sp-MR; pTarget9). This way, the crRNAs were most comparable to the mammalian situation. The tailored versions were made in pTarget8 (MR-Sp-MR; Figure 8.5). Most of the spacers performed well in the assay (Figure 8.7E). Spacers I, N, and to a certain extend also Spacer H, underperformed, and were not rescued by our tailoring approach. On the other hand, spacers K and O (Figure 8.7A/7C) did show substantial improvement after tailoring (Figure 8.7B/7D). While poor crRNA performance is clearly not explained solely by secondary structure in the pre-crRNA, implementing our design of folding the spacer tail back onto the spacer itself does not impede Cas12a activity (0-5%-point). Sometimes it results in a moderate improvement (crRNAs B,C,E,L: 5-20%-point), and sometimes in a major one (crRNAs K,O: >20%-point). While the targeting by spacer E was not impressive even after tailoring, the relative increase in efficiency was substantial (72%).

Limits of imposing structure onto the crRNA

Varying the flanking sequences has significantly improved the efficiency (plasmid targeting / fluorescence loss) of the poor performing Sp8 from 0-8% (pTarget1-3) to 40-50% (pTarget4-6) (Figure 8.3B). Still, it did not have the efficiency of well performing spacers Sp4 or Sp9 (75-85% fluorescence loss; Figure 8.1D). In the aforementioned designs, strong indications for undesired secondary structures were found in precursors of poorly performing crRNA variants. However, incompatible secondary structures that disrupt

association with Cas12a and/or RNA-DNA binding, might also occur within mature crRNAs. As an example, we again focused on Sp8, the mature crRNA of which can potentially form two alternative structures that would disrupt pseudoknot formation (Figure 8.8A). By changing the tail of the spacer, we could mask the spacer nucleotides involved in the unfavourable alternate structures. Whereas minimal base pairing of the spacer nucleotides occurred in the aforementioned Back-fold lib (Figure 8.4), the nucleotides in the tail of these new designs have to compete with nucleotides in the direct repeat for base pairing with the spacer. In contrast to the aforementioned "impaired" spacer (Figure 8.6), the tail's target nucleotides are positioned closer to the direct repeat. Short distance interactions are more likely to occur than long distance interactions, so five nucleotides in the spacer tail are not expected to compete effectively with the nucleotides in the direct repeat. Instead, these tail nucleotides are more likely to interact with matching nucleotides in the adjacent spacer part. Therefore, additional designs were made (MR-Sp-MR construct in pTarget8) in an attempt to improve the masking, and hence the performance of Sp8 (Figure 8.8B/8C). Compared to aforementioned designs with long a loop (8 nt) and a short stem (5 nt) (Figure 8.4E), we here used pTarget8 to systematically test designs for intramolecular spacer/spacer-tail interactions, by varying both the loop size (4, 5, or 6 nt) and the stem size (4 x 3 bp, 3 x 4 bp, or 2 x 5 bp) (Figure 8.8C). Because the spacer folding should be reversible to allow for base pairing with a complementary DNA target, all stems are interrupted by single nucleotide bulges and end with a mismatch. In case of the 4 and 5 nt loops, alternative designs (Figure 8.8C; alternative (alt)) were made to analyse to effect of base pairing to the seed region.

Compared to the control (Sp8 without tail, in pTarget5) that results in 50% fluorescence loss (Figure 8.8B ("A")), 10 out of 14 tested designs indeed showed improvement in cleavage efficiency of up to 14 percent point (Figure 8.8B). Two designs with relatively weak folds (e.g. [stem 3/loop 6]) did not result in improved efficiency, probably due to pseudoknot disruption. Likewise, the two designs that result in the strongest secondary structures [stem-5/loop-5] and [stem-4/loop-5] did not result in enhanced activity levels (43% and 53%), most likely because the stable fold of the crRNA's spacer hampers efficient DNA targeting. Although no solid correlation was observed between Cas12a activity and the overall folding energy of the spacer/spacer-tail part of the crRNA, the distribution of base pairing strength (G•C-pairs versus A•U-pairs) did seem to be important. Hairpin formation is more likely to occur when the loop is small, and when there is stronger base pairing near the loop (130). This is in agreement with the difference in performance of designs [stem-3/loop-5; alternative] and [stem-3/loop-6] (Figure 8.8C). Although the overall folds are very similar, the former has improved activity (61% versus 52%) probably due to a smaller loop (5 nt versus 6 nt) and a stronger loop-adjacent stem (3 G•C-pairs versus 2 G•C pairs). Hence, the loop-size and stem composition determine the equilibrium of different crRNA folds, reflecting the overall cleavage activity.

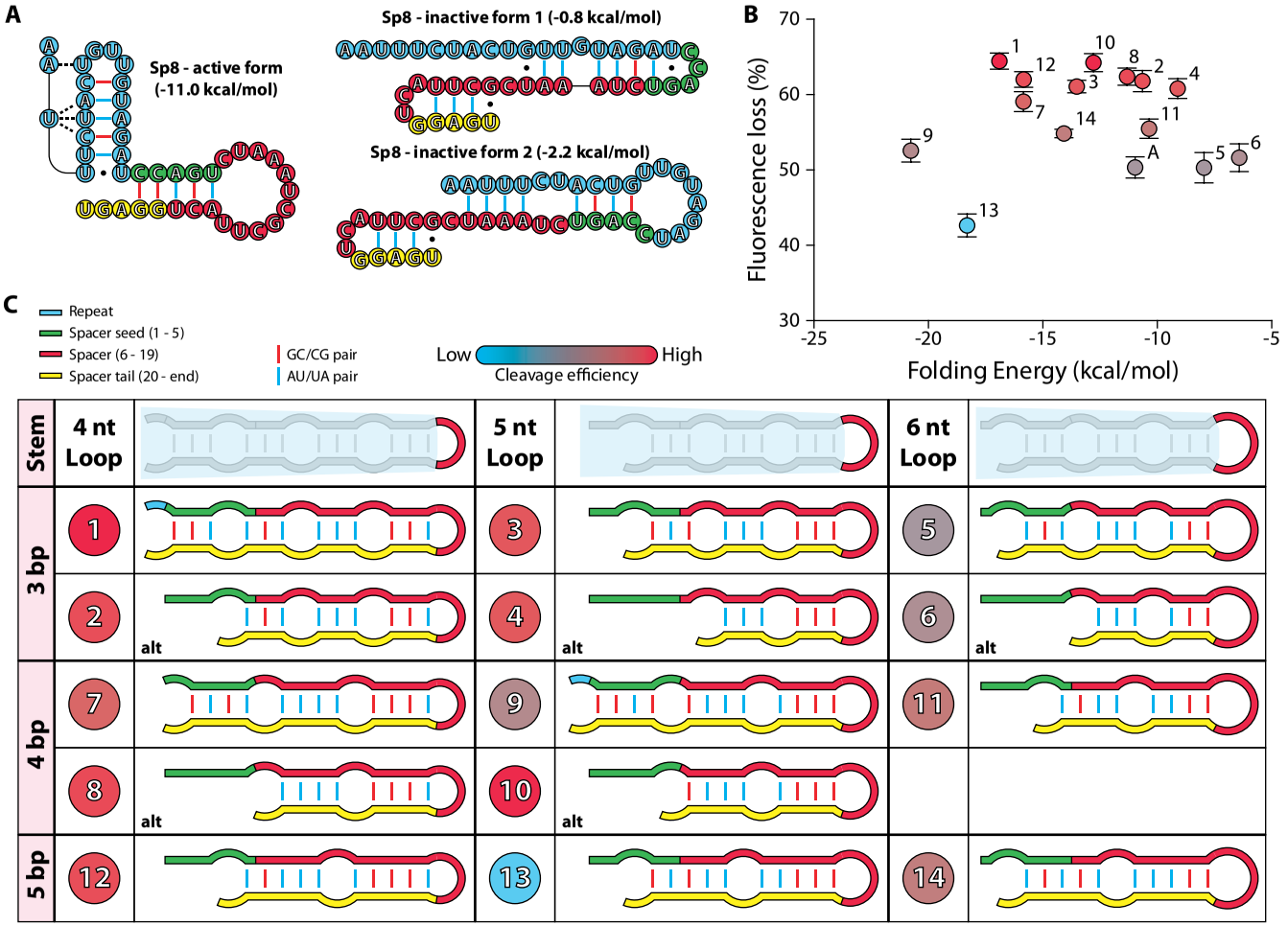

A
Sp8 - active form
(-11.0 kcal/mol)
Sp8 - inactive form 1 (-0.8 kcal/mol)
Sp8 - inactive form 2 (-2.2 kcal/mol)
B
Fluorescence loss (%)
Folding Energy (kcal/mol)
C
Repeat
Spacer seed (1 - 5)
Spacer (6 - 19)
Spacer tail (20 - end)
GC/CG pair
AU/UA pair
Low
High
Cleavage efficiency
Stem
4 nt Loop
5 nt Loop
6 nt Loop
3 bp
4 bp
5 bp
alt

Figure 8.8. Imposing structure onto the Sp8 crRNA. (A) The Sp8 has two alternate, inactive forms involving the first 11 nucleotides of the spacer and part of the 5′ repeat. The mature repeat (blue) is followed by 5 nucleotides of seed region (green) and 14 residual nucleotides of Sp8 (red). The tail (yellow) were replaced by other nucleotides depending on the construct. (B) Scatterplot of folding energy of the crRNA against the fluorescence loss. Numbers indicate the construct depicted in Figure 8.8C. The colour gradient corresponds to the fluorescence loss and correspond to the gradient colours of Figure 8.8C.. Value "A" corresponds with the Sp8 active form of Figure 8.8A. (C) Fluorescence loss in pTarget8 and structure of the designs for the rescue of Sp8. The fluorescence loss, depicted by a circle with gradient colour, and the numbers are linked to Figure 8.8B. The tail (yellow) added to the spacer folds back onto the spacer with different loop and stem sizes, always intermitted by a single nucleotide bulge. The tails that cover the seed extensively are also shortened by one stem to yield the alternative (alt) form. The grey circle indicates the activity of Cas12a in observed fluorescence loss, while the blue bar graph depicts the theoretical folding free energy of the ensemble omitting the repeat. To judge the stem strength at a glance, at the G•C pairs are marked in red and the A•U pairs in blue.

Comparison of designs that only differ in their tail length (Figure 8.8C; alt), might reflect an effect of reduced masking of the seed region. In case of stem-3 variants no differences are observed, but in case of the stem-4 variants there seems to be a difference, possibly due to the longer tails covering the seeds more extensively (4 out of 5 seed bases). Relatively weak base pairing between the tail and the seed in [stem-4/loop-4] is rather well tolerated (59% in seed-masked versus 62% in seed-free), but stronger base pairing (3 G•C pairs instead of 2) in [stem-4/loop-5] appears to be penalised (53% in seed-masked versus 64% in seed-free) (Figure 8.8C).

Discussion

High efficiency is an important criterium for genome editing applications by CRISPR-associated nucleases. The efficiency of genome editing depends on (I) delivery of Cas effector proteins and their crRNAs (132), (II) crRNA performance (this study), (III) accessibility (chromatin structure) of the target site (100), (IV) accuracy of the Cas nuclease (98), and (V) the host's mechanism to repair the generated double stranded break (DSB) (133). In this study, we set out to reveal the molecular basis of variable crRNA functionality in targeting a plasmid in *E. coli* by Cas12a.

As has been reported previously (129, 131), we observed major fluctuation in crRNA functionality when analysing a small set of Cas12a crRNAs (Figure 8.1D) and derived variants (Figure 8.2A/2C). Our working hypothesis is that the pseudoknot structure of the pre-crRNA is required for appropriate binding by Cas12a, and hence for processing to mature crRNA and formation of a cleavage-compatible Cas12a/crRNA complex. The hairpin within this pseudoknot structure is formed through intramolecular base pairing of the palindromic part of the repeat. Structural variation can occur when flanking sequences (the upstream leader or the downstream spacer) compete for base pairing with the nucleotides in the repeat, resulting in alternative secondary structures. The relative

strength of these competing structures will determine the equilibrium of the mixture of functional and non-functional crRNAs. In agreement with this assumption, it was found that the leader sequence combined with the sequence of the bad crRNA (Sp8) potentially adopts an alternative fold in which base pairing occurs between the Sp8 spacer and part of the direct repeat (Figure 8.2E). In the broader sense, any sequence adjacent to the repeat may cause the pre-crRNA to fold into a structure that is not recognised by Cas12a. In case of a CRISPR array, this includes interactions with other upstream and downstream sequences (Figure 8.2E).

Based on these insights, we then assessed whether it would be possible to avoid disruption of the pseudoknot structure by a competing spacer sequence, through designing base pairing of the spacer with a spacer-tail (i.e. positions 20-24, which are not involved in DNA targeting). We designed a library of constructs with a variable spacer tail that potentially could fold back onto the spacer at different positions (Figure 8.4). As expected, cleavage activity is abolished if the repeat is targeted with a strong competing complementary sequence, disrupting the pseudoknot structure. Interestingly, base pairing of the spacer-tail with the spacer-seed appeared not to affect Cas12a activity, indicating that after association with Cas12a, the designed stem-loop of the spacer part is 'dissolved' again. This spacer-tail design strategy can be very useful to rescue bad spacers, especially in editing efforts in which there is little flexibility with respect to the target site (e.g. base editing). To test this, a poorly-performing "impaired" spacer was designed such that base pairing may occur between 7 nucleotides of the spacer (positions 6-12) and the repeat, potentially resulting in formation of an undesired alternative structure (Figure 8.5). In two different crRNA constructs, the addition of a short spacer-tail (positions 20-24) that is complementary to spacer positions 9-14 indeed resulted in increased Cas12a cleavage efficiency, most likely through specific folding the tail back onto the spacer, thereby destabilising the alternate structure and enhancing pseudoknot formation.

Further improvement of the Sp8 was achieved in a similar way as the “impaired” spacer. By folding the tail of the spacer back onto the spacer itself, the efficiency was increased from below 50% (Figure 8.3B) to 64% (Figure 8.8). However, the targeting efficiency did not reach the top levels of e.g. Sp9 or Sp12 in pTarget3, of which detected fluorescence loss is 75-80% (Figure 8.1D). How far the improvement can be extended should be addressed in further studies.

By systematically analysing the functionality of (pre-)crRNA variants, we have shown that in the context of both precursor and mature crRNA, the repeat can base pair with upstream and downstream sequences. Overall, our experimental findings and computational analyses (Supplementary data) support that the repeat's hairpin structure is a key requirement for high cleavage efficiencies. We introduced three crRNA features that are significantly correlated with experimentally determined plasmid loss efficiencies:

pseudoknot stem-forming (high base pairing potential), pseudoknot loop-accessibility (low base pairing potential), and reversible spacer folding (base pairing potential of the first 19 nucleotides). Unfortunately, the currently available RNA folding algorithms can neither make an accurate prediction of all potential base pair possibilities, nor can they predict the (equilibrium constants between) different possible folding states. Nevertheless, we have tried to predict whether spacers get trapped in an undesired inactive state (no canonical pseudoknot), or rather in a desired active state (with canonical pseudoknot) (see Supplementary Figure 8.9). By doing so, we hoped to make a first step towards developing an algorithm for designing functional spacers for Cas12a. Although analysis of the used procedure mainly resulted in an indication of potentially incorrect folding of the crRNAs, it did not provide the desired robust estimation of their functionality (targeting efficiency).

In conclusion, although it may be difficult to pinpoint exactly why one spacer is more efficient than another, we can use our findings to propose some general guidelines for design of Cas12a-associated crRNAs. (I) Keep the pseudoknot structure intact. Folding of the pre-crRNA has a major impact on the DNA cleavage activity of Cas12a. Most importantly, the pseudoknot that corresponds to the palindromic part of the repeat should be formed correctly. Any stretch of RNA that is complementary to (part of) this repeat sequence may interfere with the pseudoknot formation. This includes any region upstream and downstream of the crRNA, as well as the spacer itself. Upstream and downstream regions are to be omitted as much as possible: the shorter the crRNA design the better (117). (II) Avoid base pairing between pseudoknot and spacer by imposing a structure onto the pre-crRNA. In case RNA structure prediction programs suggest base pairing between the repeat and the spacer, potential problems of such undesired base pairing can be reduced by introducing a 5-16 nt tail at the 3' end of the 19 nt spacer. Too strong base pairing by the tail may result in an irreversible fold that will impede proper docking of Cas12a and/or formation of the R-loop configuration; on the other hand, too weak base pairing does not allow for the intended re-structuring of the pre-crRNA. As a rule of thumb, the overall strength of the spacer structure (not including pseudoknot) should range from -5 to -11 kcal/mol. Locally, high-GC stems towards the loop are less problematic than high-GC stems at the seed. Base pairing with the seed is not advised, unless the strength of the structure requires it. Tuning can be accomplished by introducing mismatches. Based on our observations, well-performing crRNAs may be obtained by designing stems of 3-4 base pairs, including 1-2 GC pairs, interspaced with a single mismatch. The hairpin loop should preferably be 4-5 nucleotides long, while the seed is kept free from base pairing. The shortest construct that can meet these guidelines is 19 nucleotides of target sequence of which the last 4-5 constitute the loop. Positions 20-29 should then form two interrupted stems of 4 base pairs, ending with a mismatch. Although spacers of which nucleotides 1-19 have the potential to base pair with the repeat or the seed may never reach the efficiency of other spacers, it should be possible to design a spacer-tail to gain sufficient

functionality. (III) Introduce a well-structured RNA at the 3' end. Ending with a terminator allows for the most accurate prediction of secondary structures, as the nucleotides of the terminator are unlikely to base pair anywhere else. Terminators should be separated from the direct repeat by at least 24 nucleotides of spacer to avoid steric hindrance. It should be kept in mind that the 3'tail of the terminator may still elongate the terminator stem, which should be avoided. Ending with the highly structured mature repeat (last 19 nt of the direct repeat), a separation sequence that forms a hairpin, as well as the aforementioned highly structured terminator, maximises the predictability of the pre-crRNA without the need for extra nucleotides attached to the spacer. Since the separation sequence used in this study does not form a strong hairpin, unlike the terminator, it could end up back-folding to the spacer. A terminator directly following the second mature repeat might solve such an issue, although we have not assessed the impact on the processing of the second repeat. (IV) Avoid intra-array complementarity. In case of a multiplex approach to target different sequences simultaneously, multiple spacers can be combined in a single array. However, such a design adds to the unpredictability of its fold and, therefore, of the efficiency of individual spacers. Should the application of the Cas12a demand an array, one needs to take care that the consecutive spacers do not interact with each other. Imposing a fold onto the pre-crRNA may be utilised to keep the spacers from interacting, but this does not work for two consecutive spacers with high homology.

Material and methods

Strains and media

E. coli DH10B T1R (Invitrogen) was used as host for cloning, plasmid propagation and fluorescence assays. Bacteria were generally cultured on LB medium (10 g/L peptone (Oxoid), 5 g/L yeast extract (BD), 10 g/L NaCl (Acros)) at 37°C. When required, media were supplemented with kanamycin (kan; 50 mg/L) and/or chloramphenicol (cam; 35 mg/L). Fluorescence assays were performed on M9TG medium (1x M9 salts (Sigma), 10 g/L tryptone (Oxoid), 5 g/L glycerol (Acros)). Induction of the FnCas12a was done with L-rhamnose (2 g/L).

Plasmids

The commercial pRham N-His SUMO Kan from Lucigen was made compatible for Ligation Independent Cloning (LIC) by polymerase chain reaction (PCR) (BG7802 and BG7803). The fncas12a gene was PCR amplified (BG7709 and BG7710) and cloned into pRham_ LIC using a standard LIC protocol. The pRham-FnCas12a-DAS was made using pRham-FnCas12a as a base. The pRham-Cas12a was digested with BamHI-HF (NEB) and SpeI-HF (NEB). A fragment was created with a variant of the ssrA tag behind the Cas12a coding sequence (AANDENYADAS; see below) by PCR with Q5 polymerase (NEB), using BG8998

and BG10140 as primers and pRham-FnCas12a as template. The PCR fragment was digested with BamHI-HF and SpeI-HF and ligated into the pRham-FnCas12a digest using T4 ligase (NEB) to generate pCas12a.

Target plasmids pTarget1 to pTarget8 are generated from two fragments. The fragments are generated by PCR with Q5 polymerase (NEB). The first fragment contains the cat gene (chloramphenicol resistance) from pACYC184 flanked by a SalI site – terminator (Target1 – F1)/mature repeat (Target2 – F1)/full repeat (Target3 – F1) on one side and SacI on the other. The second fragment contains the P15A ori from pACYC184 and an *mrfp* gene. A SacI site – PAM – target is attached by PCR on the one end while the other end the crRNA is attached followed by a SalI site. Depending on whether a full repeat is required (Target1 – F2) or a mature repeat (Target4 – F2), a different template is used. Ligating the Target1 – F2 to Target1/2/3 – F1 yields pTarget1/2/3 respectively and ligating Target4 – F2 to Target1/2/3 – F1 yields pTarget4/5/6. Fragments were digested with SalI-HF (NEB) and SacI-HF (NEB), and ligated with T4 ligase (NEB). pTarget7 is generated by ligating Target7 – F1 to Target7 – F2, while pTarget8 is a ligation of Target8 – F1 and Target8 – F2. pTarget9 is a ligation of Target8 – F1 and Target9 – F2. These plasmids are assembled by Golden Gate cloning with SapI (NEB) and T4 ligase (NEB). An overview of the cloning details, including primer sequences, can be found in the supplementary data (Supplementary sequence 1, Supplementary table 1, Supplementary table 2).

Fluorescence loss assay

The targeting activity of programmable nucleases in bacteria is generally measured either by transformation efficiency assays (based on the recovery of viable transformants), or by plasmid loss assays (based on loss of plasmid over time). In both cases, the fraction of bacteria harbouring a plasmid is assessed by plating in parallel on both selective and non-selective medium. However, apart from being labour intensive, we consider these methods not accurate enough to distinguish small differences in cleavage efficiency. Major drawbacks of transformation efficiency assays include inconsistency of bacterial competence and differences in plasmid purity and concentration, resulting in low accuracy. Plasmid loss assays do not suffer from these artefacts, but still require plating and colony counting. The duration of the expression of the Cas12a nuclease is quite essential, as resolution is lost either at high plasmid clearance rates or during extended Cas12a exposure. While the extended Cas12a exposure increases sensitivity, the information on plasmid cleavage efficiency is lost after full plasmid clearance.

For these reasons, we developed a robust screening approach that allows for accurate detection of variations in the copy number of a target plasmid. Apart from a chloramphenicol resistance marker (cat), the target plasmid (pTarget) contains a constitutively expressed reporter gene (*mrfp*), a short CRISPR array with a single spacer sequence, and a matching target sequence downstream of a 5'-TTTV PAM motif (hereafter

referred to as "target") (Figure 8.1B; Table 1; Supplementary table 2). To ensure the differences in mRFP fluorescence are a direct result of Cas12a cleavage activity, and not, for example, caused by blocking of read-through transcription, the *mrfp* gene is isolated by two terminators. In addition, targeting occurs outside of the *mrfp* transcription region. The target plasmids were individually transformed to *E. coli* cells harbouring a second plasmid (pCas12a) that encodes FnCas12a (hereafter referred to as Cas12a). The rate of Cas12a-mediated clearance of the target plasmid is detected as loss of mRFP fluorescence, directly reflecting the spacer-based targeting efficiency.

Compared to plasmid loss or transformation assays, the here-established fluorescence-loss assay allows for distinguishing between efficient (good) and less-efficient (bad) crRNAs with high accuracy and ease. The fluorescence builds up as long as the plasmid is retained and the bacteria are still producing protein. The higher the activity of the Cas12a, the higher the loss of fluorescence. The effect of exposure to Cas12a is terminated by bacteria reaching the stationary phase where protein production eventually stops, so timing is less essential compared to a plasmid loss assay. Even when all the plasmid is lost, differences in targeting activity can still be extrapolated from fluorescence values. Important to note, however, is that the dilution of the preculture to the final culture dictates the exposure time. It is therefore essential that the dilution is carried out very precisely. If more sensitivity is required, the cells can be further diluted to allow for a longer period of plasmid loss.

Since targeting of the plasmid may cause escape mutants, we needed to limit the exposure of the target plasmid to Cas12a. Chemically competent *E. coli* DH10B harbouring the pCas12a plasmid were transformed with target plasmid and recovered in LB. After recovery, the bacteria were diluted 1:100 to 200 µL M9TG medium supplemented with kan/cam from the transformation mix, and grown in a 2 mL 96-well masterblock (Greiner) covered with a gas-permeable membrane at 37°C overnight. Presence of both kan and cam will ensure the bacteria retain both plasmids. After overnight growth, the bacteria were diluted 10^{-4} (two steps of 10^{-2}) into 200 µL fresh M9TG medium supplemented with kan/cam, kan, or kan and with 2 g/L L-rhamnose and grown at 37°C overnight in a master block covered with a gas-permeable membrane. The cultures were cooled down to room temperature and diluted 5x in 1x PBS pH 7.4. As controls, non-targeted plasmid with mRFP (pTS001) and non-targeted plasmid without mRFP (pACYC184) were used alongside non-inoculated M9TG medium. The latter served both as a negative control for growth and a blank for fluorescence and light scattering. 100 µL of the diluted cultures was measured on a Synergy MX microplate reader. Fluorescence measurements were performed with an excitation at 584 nm with a bandwidth of 9 nm, emission at 607 nm with a bandwidth of 9 nm and a gain of 120. Fluorescence loss of (x) was calculated as follows:

$$Fluorescence\ loss$$
$$= 100 \times (1 - \frac{\text{avg}((Fl[x,rham] - Fl[blank])/(OD_{600}[x,rham] - OD_{600}[blank]))}{\text{avg}((Fl[x,cam] - Fl[blank])/(OD_{600}[x,cam] - OD_{600}[blank]))})$$

$$\sigma_{[x,Fl.loss]} = \left|\frac{\mu_{[x,rham]}}{\mu_{[x,cam]}}\right| \times \sqrt{(\frac{\sigma_{[x,rham]}}{\mu_{[x,rham]}})^2 + (\frac{\sigma_{[x,cam]}}{\mu_{[x,cam]}})^2}$$

Fine-tuning of the assay

To allow for accurate comparative analyses of crRNA performance, fine-tuning of the assay has been performed at three levels. (I) Synchronising cells - The time available for the plasmid clearance is crucial for the final fluorescence. For the best possible comparison of targeting performance of different crRNAs, the bacteria harbouring both plasmids (pCas12a and pTarget) were synchronised by growing them to the stationary phase in a pre-culture in the presence of antibiotics to select for maintenance of both plasmids. (II) Minimise targeting when undesired - Simultaneous selection of pTarget maintenance (through presence of chloramphenicol) and targeting of pTarget by background Cas12a activity, would allow for growth of escape mutants and hence selection of false negatives in the fluorescence loss assay. Indeed, under these conditions we observed sabotage of Cas12a activity in different ways: by deletion of the entire spacer through recombination of the two flanking repeats, by recombination of the ribosome binding site of the cas12a gene, and by introduction of a transposon into the cas12a coding region (not shown). To control the timing of Cas12a targeting, the cas12a gene expression is controlled by the rhamnose-inducible PrhaBAD promoter. (III) Limit the lifespan of Cas12a - To further reduce leaky expression of Cas12a, an ssrA degradation tag is fused to the C-terminus of the protein. While the native ssrA tag ("AANDENYALAA") almost completely abolishes the Cas12a activity, a less efficient tag variant ("AANDENYADAS") (134, 135) was found to limit both the Cas12a residence time and its leaky expression (Supplementary figure 8.3). It also reduces the protein levels of Cas12a in the cell while being induced. A large excess of Cas12a would cause crRNAs to bind because of the high effector protein concentration rather than a high affinity; in that case moderate affinity might cause the fluorescence to drop to background levels, resulting in loss of resolution for the high efficiency spacers. When the Cas12a concentration is limited, we can distinguish moderate from good spacers.

Without Cas12a induction and without antibiotics pressure on the target plasmid, a very low basic level of fluorescence loss is observed for most spacers (Supplementary figure 8.4). A low, non-induced fluorescence loss indicates that the plasmid is not targeted severely during the synchronisation, and therefore, escape mutants have no significant growth advantage. Some of the highly efficient spacers do show substantial fluorescence loss

without induction of Cas12a and without addition of cam. The presence of cam enforces target plasmid maintenance, and although we see a reduction in fluorescence compared to less efficient spacers, the reduction is only marginally under these conditions, and no escape mutants were observed.

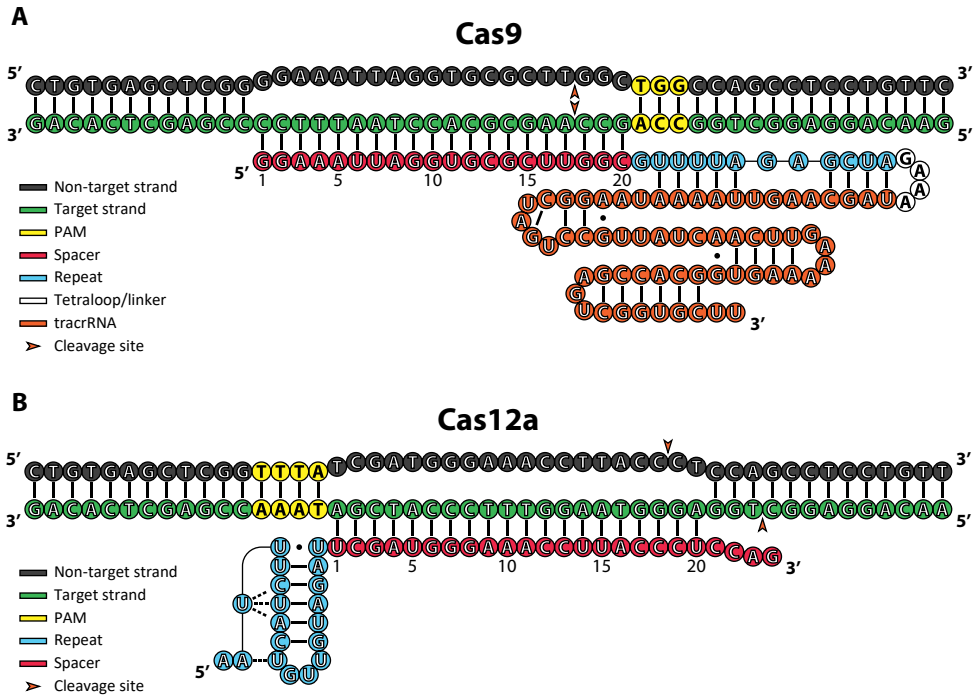
In vitro cleavage assay

Pre-crRNA was made by *in vitro* transcription with the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). FnCas12a was diluted to 0.4 nM in 2x NEBuffer 4 with 2 mM Mg2+, instead of 20 mM Mg2+. Reactions were started by adding 0.2 nM pre-crRNA in a 1:1 ratio. Reactions were incubated at 37°C and sampled at t = 0, t = 5 and t = 10 min. Sampled reactions were quenched by adding 1 µL of quenching solution (9% SDS, 50 mM EDTA) to 10 µL of sample. Samples were heated to 95°C for 5 min and cooled to 12°C. Potassium dodecyl sulphate was pelleted by centrifugation and the 10 µL of supernatant was mixed with 2x RNA Loading Dye (NEB) and analysed on a 1.5 mm 5% acrylamide gel containing 7M urea. The gel was run on a Bio-Rad Mini-PROTEAN Tetra Cell system at 15 mA until the bromophenol blue was at the bottom. RNA cleavage products were stained by SYBR gold and visualised on a Bio-Rad Gel Doc XR+.

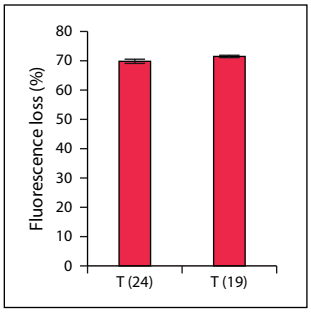
Acknowledgements

The authors would like to thank Jorik Bot for experimental support at an early stage of this project.

Supplementary data

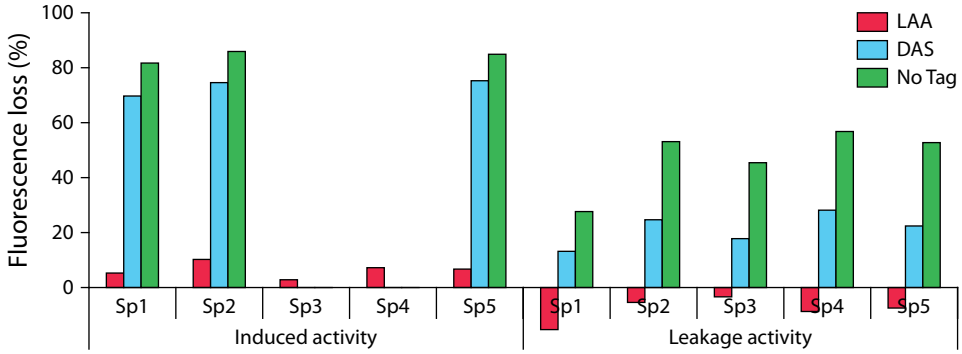


Supplementary figure 8.1. Spacer target binding of Cas9 and Cas12a. (A) Cas9 has an R-loop of 20 nucleotides, and a three nucleotide 3′ PAM. Attached to the spacer is a 3′ repeat with tracrRNA, which is required for Cas9 activity. This design couples the tracrRNA (orange) to the repeat (blue), which is not the case in the natural situation. Cleavage is PAM proximal and it leaves a blunt end. (B) Cas12a also has a 20 base pair R-loop. The PAM is four nucleotides long and located on the 5′ side of the target. The cleavage is PAM distal and leaves a 5′ overhang of, in most occasions, 5 nucleotides.

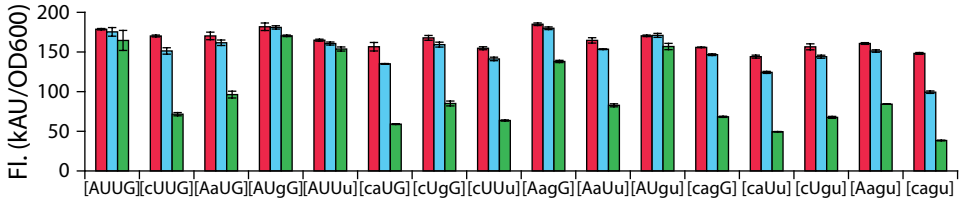


	Seed																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
gRNA Sp4	A	C	A	C	A	C	U	G	C	A	A	U	U	C	A	G	G	U	U	G	G	A	G	U
Target (24)	A	C	A	C	A	C	T	G	C	A	A	T	T	C	A	G	G	T	T	G	G	A	G	T
Target (19)	A	C	A	C	A	C	T	G	C	A	A	T	T	C	A	G	G	T	T	A	T	C	G	T

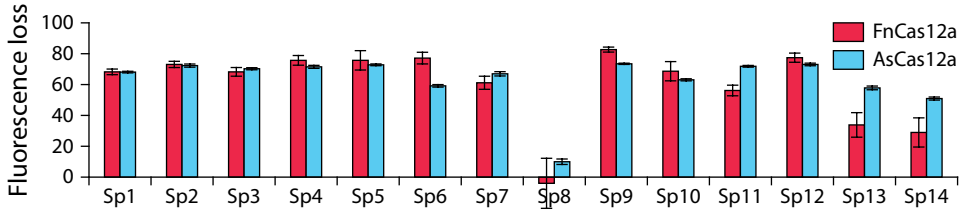
Supplementary figure 8.2. Pilot experiment with 19 nucleotides spacer and target complementarity. Fluorescence loss is given for two different targets with the error bars indicating S.D. The spacer is the same in both instances, but the target allows for either 19 or 24 base pairs. The R-loop of the 24 nucleotides target is restricted to 20 nucleotides by the Cas12a itself. There appears to be little difference between the 19 and 20 base pair R-loop, which justifies the choice for a 19-nucleotide spacer and target complementarity.



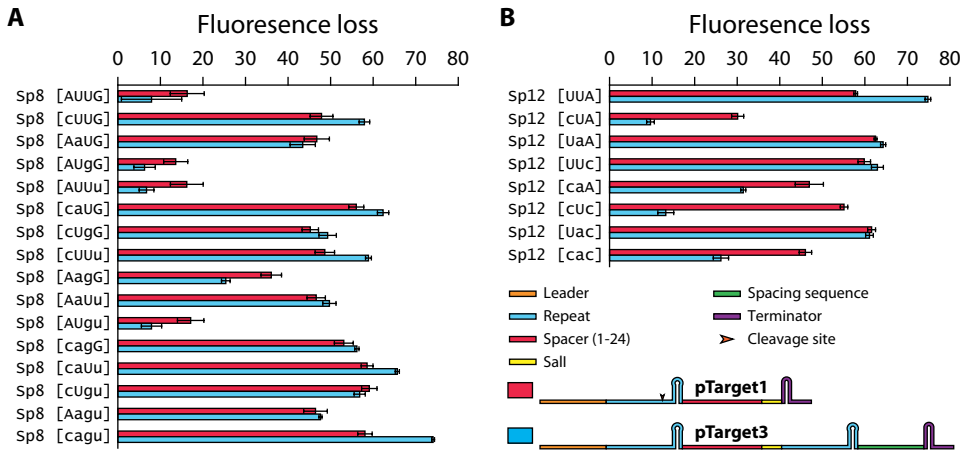
Supplementary figure 8.3. Degradation tag comparison. Degradation tags were compared and show that the "DAS" tag retains a lot of activity, but not as much as the untagged Cas12a. The benefit of using the tag nonetheless is the avoidance of escape mutants, which were sometimes observed with the untagged Cas12a. This is seen by the leakage activity, which is much higher for the untagged compared to the DAS tagged Cas12a. The ":LAA" tag abolishes activity almost completely and is considered too strong.



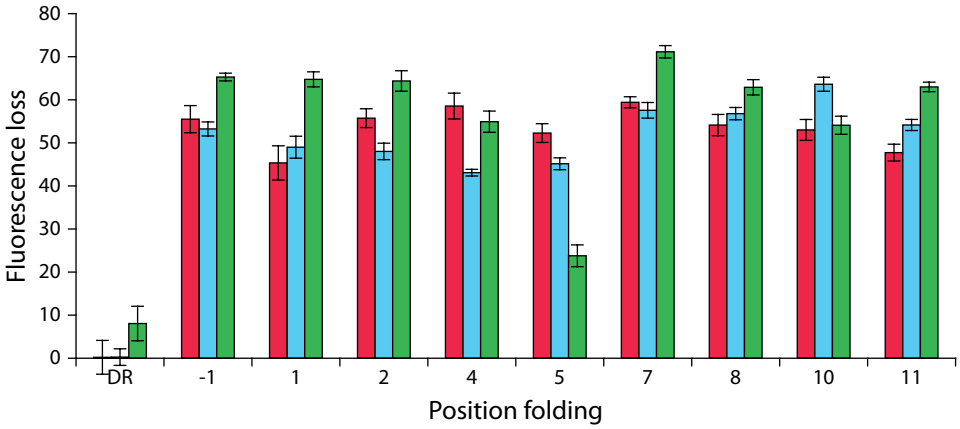
Supplementary figure 8.4. Raw fluorescence of the Sp8 variants in L-FR-Sp-FR (pTarget3) (See Figure 2A). Fluorescence is calculated as raw fluorescence per OD600 and averaged from three replicates, with the error bars representing the S.D.. Kanamycin selects for the pCas12a and chloramphenicol for the pTarget3. Bacteria are grown on kanamycin and chloramphenicol (red), kanamycin alone (blue) or kanamycin and 2 g/L L-rhamnose (green).



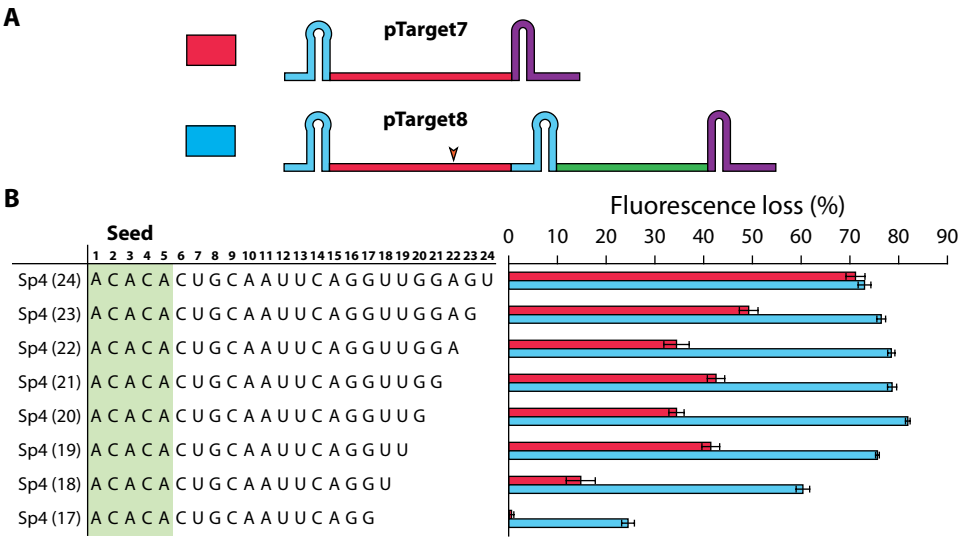
Supplementary figure 8.5. Activity of FnCas12a compared to AsCas12a. The constructs of pTarget3 were also tested for AsCas12a (blue) and compared to FnCas12a (red). In general, good spacers for FnCas12a are good spacers for AsCas12a, but the latter appears more forgiving for some of them (Sp11, Sp13, Sp14), while some others have slightly diminished activity (Sp6). Possibly, the AsCas12a can better cope with lower concentrations of correctly folded guide than FnCas12a does. Since the inactive structure for e.g. Sp14 is not as strong as it is for Sp8, more guide will be correctly folded. Higher affinity for correctly folded guide then might still increase the activity.



Supplementary figure 8.6. Sp8 and Sp12 variants with terminator or second full repeat. (A) Sp8 variants leading to Sp12 (Sp8 [cagu]). The variants with terminator directly behind the Sall site (red) show a similar pattern compared to the variants with a second full repeat (blue). However, the best spacers (cagu e.g.) with terminator do not perform as well as the variant without, while the worst spacers (AUUG e.g) have better activity with the terminator directly behind it. (B) Variants of Sp12 leading to Sp14 (Sp12 [cac]). Some of the spacers show a large discrepancy between the terminator and the second full repeat. The difference between terminator and full repeat is quite a number of nucleotides. These change the context of the spacer and may influence the folding and which is its most stable structure significantly. Context dependence was also shown with the removal of the leader-end and replacing the full repeat with a mature repeat on the 5', so the 3' sequence affecting the efficiency can be expected.



Supplementary figure 8.7. Back-folding libraries. The fluorescence loss was determined for 3 libraries in L-FR-Sp-|T| (pTarget1) with the same intended folding and three replicates with the error bars representing S.D..



Supplementary figure 8.8. Cleavage efficiency for different spacer lengths. (A) Schematic of pTarget7 and pTarget8. (B) Cleavage efficiency is shown in terms of fluorescence loss for various spacer lengths (17-24) with the error bars representing the SD.. Blue bars are cleavage efficiency shown for spacers containing a mature repeat and a terminator at the 3' end and in red bars are for spacers that only contain a terminator at the 3' end.

Cleavage efficiency for different spacer lengths

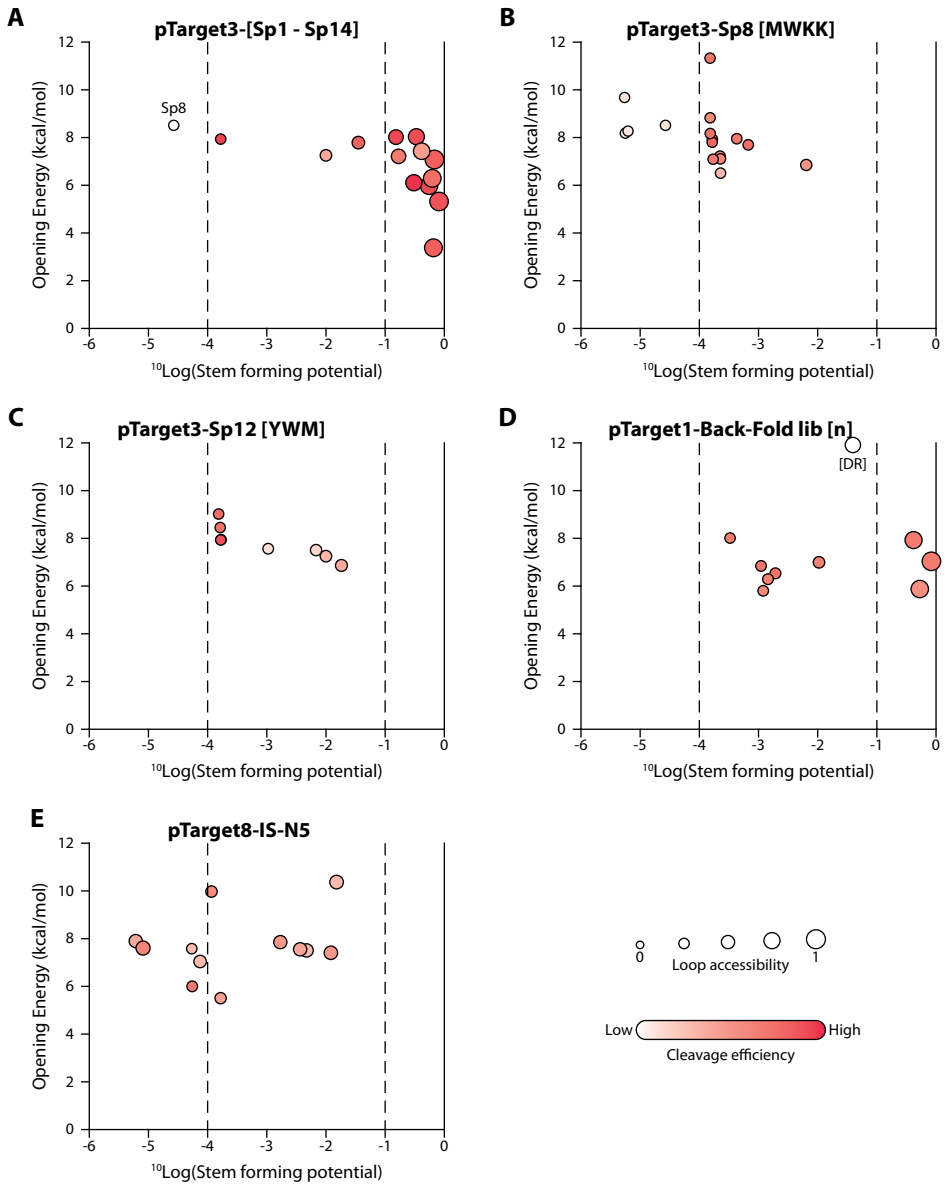
As intramolecular base pairing between spacer and repeat should be avoided, the minimal spacer length in the mature guide sequence was determined. Two guide expression constructs were compared, pTarget7 and pTarget8 (Figure 8.7A). In case of pTarget8, with a mature repeat and terminator at the 3' end, Cas12a requires a guide sequence of 19-24 nucleotides for optimal *in vivo* cleavage efficiency (70-80%). A drop to 60% and 25% is observed when spacer length is reduced to 18-nt and 17-nt, respectively (Supplementary figure 8.8B). In case of pTarget7, with only a terminator at the 3' end, optimal activity is observed with a 24 nt spacer (70%), and gradually decreasing efficiencies (50-1%) upon reduction of spacer length (23-17 nt) (Figure 7B). The pTarget7 constructs have the terminator directly at the 3' end of the spacer, which appears to be an additional limitation upon spacer length reduction, most likely due to steric hindrance. Additionally, the terminator stem is actually longer in the constructs shorter than 24 nt. This is because the terminator ends in a polyU tail, which potentially binds to the GGAG at the end of Sp4. Since position 24 is "U", the base pairing is interrupted for that length. The G-U wobble pair is not particularly strong, but it can add some strength and length to the already strong terminator stem. Altogether, these findings agree with a previously reported minimal spacer length (16-17 nt) for *in vitro* cleavage analysis of FnCas12a (2), and with the requirement of a "minimal" R-loop for binding (and cleavage) of DNA targets by Cas12a

(98). The latter study revealed that a mismatch at position 20 has no significant effect on the binding rate, while mismatches at positions 19 and 18 do result in decreased binding rates.

Computational analysis of pre-crRNA folding and its relationship with spacer efficiency

To analyse the folding characteristics of pre-crRNA, we introduced three measures: stem-forming potential, loop-accessibility and spacer opening energy. We computed all these measures using the RNAfold and RNAplfold programs from the Vienna RNA package v2.4.1 (130). In all our computations, we provided the full sequence of the pre-crRNA as input and used the programs in the default settings, except for the parameters window length W and base-pair distance L in RNAplfold. These were set to the length of the pre-crRNA. Using RNAfold, we computed all possible base-pair probabilities in pre-crRNA, but subsequently we only focused on the five base-pairs that forms the stem of the hairpin loop (U[-15]:A[-2], C[-14]:G[-3], U[-13]:A[-4], A[-12]:U[-5], C[-11]:G[-6] base-pairs). The stem-forming potential is equal to the product of these 5 base-pair probabilities. The other two measures, hairpin loop accessibility and spacer opening energy, were computed using the RNAplfold program. The hairpin loop-accessibility measure corresponds to the mean probability of UGUU region of the hairpin loop to be unpaired and the spacer opening energy corresponds to the approximate opening energy computed for the first 19 nts of the spacer. Both measures are parsed from RNAplfold output with no further change.

As described above, a major challenge for Cas12a crRNA design is to avoid that the spacer sequence affects the pseudoknot structure, at the level of either precursor crRNA or mature crRNA. Here, we set out to investigate whether disruptions on the pseudoknot structure can be computationally predicted through RNA secondary structure prediction tools. Since the pseudoknot hairpin consists of a stem and a loop, we defined two measures describing the robustness of the hairpin structure in the context of the folded pre-crRNA: the chance of the stem being formed (stem-forming potential) and the chance of the loop being unpaired (loop-accessibility). In addition, a third measure is introduced to assess how difficult it is to abolish the folding of the spacer part of the crRNA (spacer opening energy). High spacer opening energy could lead to inefficient targeting due to the potential strong base pairing within the spacer region that would restrain the target DNA recognition. This is also in parallel with energy models proposed for Cas9 DNA recognition. Details on how these three measures are calculated are described in the Materials and Methods section. Several sets of constructs that are presented throughout the study have been analysed with the in silico approach, as depicted in Supplementary figure 9. The plots reveal that spacers with a stem-forming potential $>10^{-1}$ have high cleavage efficiency, whereas spacers with a stem forming potential $<10^{-4}$ tend to be very inefficient. The same is true for the hairpin loop accessibility, as spacers having unpaired



Supplementary figure 8.9. Computational analysis of pre-crRNA. (A) pTarget3 Sp1-Sp14. (B) pTarget3-Sp8 variants (C) pTarget3 Sp12 variants (D) pTarget1 Back-Fold library (E) pTarget8-IS-N5

hairpin loops are mostly efficient. A notable exception is the pTarget1-Back-fold lib [DR] (Supplementary figure 8.9D). Since the 3' spacer tail of this construct only targets the base of the hairpin stem, the loop accessibility is high and the stem forming potential is also rather high. Evidently, the opening energy is very high, so the pre-crRNA is unlikely to shift towards the active state, and therefore, Back-fold lib [DR] construct potentially gets

trapped in the inactive state. This also holds true for the Sp8 (Supplementary figure 8.9A) and 4 left-most spacers in Supplementary figure 8.9B. They have smaller opening energies than Back-fold lib [DR] but stronger signal on pseudoknot structure disruption (low loop accessibility and very low stem forming potential). On the other hand, in Supplementary figure 8.9E, five spacers with low stem stability ($<10^{-4}$) have medium-low efficiencies which might be due to alternative competing structures. Overall, although the individual RNA folding-related criteria cannot explain spacer performance alone, they all have significant correlation with the fluorescence loss measurements (p-value<0.01, Pearson's ρ=0.31 (stem forming potential), 0.29 (loop-accessibility), -0.38 (spacer opening energy)), indicating that this analysis may be useful for identifying potential dysfunctional gRNAs.

Supplementary sequence 1. pTarget3-DNMT1

TCTAGATTTC AGTGCAATTT ATCTCTTCAA ATGTAGCACC TGAAGTCAGC CCCATACGAT ATAAGTTGTA
ATTCGGTACC CCGCTTCGGC GGGGTTTTTT CAAGTTCAAA TATGTATCCG CTCATGAGAC AATGTGTGGG
GAGACCACAA CGGTTTCCCT CTAGAAATAA TTTTGTTTAA CTATAAGAAG GAGATATACA TATGGCTTCC
TCCGAAGACG TTATCAAAGA GTTCATGCGT TTCAAAGTTC GTATGGAAGG TTCCGTTAAC GGTCACGAGT
TCGAAATCGA AGGTGAAGGT GAAGGTCGTC CGTACGAAGG TACACAGACC GCTAAACTGA AAGTTACCAA
AGGTGGCCCG CTGCCGTTCG CTTGGGACAT CCTGTCCCCG CAGTTCCAGT ACGGTTCCAA AGCTTACGTT
AAACACCCGG CTGACATCCC GGACTACCTG AAACTGTCCT TCCCGGAAGG TTTCAAATGG GAACGTGTTA
TGAACTTCGA AGACGGTGGT GTTGTTACCG TTACCCAGGA CTCCTCCCTG CAAGACGGTG AGTTCATCTA
CAAAGTTAAA CTGCGTGGTA CCAACTTCCC GTCCGACGGT CCGGTTATGC AGAAAAAAAC CATGGGTTGG
GAAGCTTCCA CCGAACGTAT GTACCCGGAA GACGGTGCTC TGAAAGGTGA AATCAAAATG CGTCTGAAAC
TGAAAGACGG TGGTCACTAC GACGCTGAAG TTAAAACCAC CTACATGGCT AAAAAACCGG TTCAGCTGCC
GGGTGCTTAC AAAACCGACA TCAAACTGGA CATCACCTCC CACAACGAAG ACTACACCAT CGTTGAACAG
TACGAACGTG CTGAAGGTCG TCACTCCACC GGTGCTTAAG CGGCCGCATA ATGATGTGTT ATCATTGATG
CGAGGTCGCC TATACCTCCC CGCTTCGGCG GGGTTTTTTC CCGGGTTTAC ACTTTATGCT TCCGGCTCGT
ATAATGTGTG GCTGATTTAG GCAAAAACGG GTCTAAGAAC TTTAAATAAT TTCTACTGTT GTAGATAGGA
GTGTTCAGTC TCCGTGAACG GTCGACGTCT AAGAACTTTA AATAATTTCT ACTGTTGTAG ATAGATACCG
GACAACGTGT CTCCCCGCTT CGGCGGGGTT TTTTCCTAGG ACTAGTCTTA TTCAGGCGTA GCACCAGGCG
TTTAAGGGCA CCAATAACTG CCTTAAAAAA ATTACGCCCC GCCCTGCCAC TCATCGCAGT ACTGTTGTAA
TTCATTAAGC ATTCTGCCGA CATGGAAGCC ATCACAAACG GCATGATGAA CCTGAATCGC CAGCGGCATC
AGCACCTTGT CGCCTTGCGT ATAATATTTG CCCATGGTGA AAACGGGGGC GAAGAAGTTG TCCATATTGG
CCACGTTTAA ATCAAAACTG GTGAAACTCA CCCAGGGATT GGCTGAGACG AAAAACATAT TCTCAATAAA
CCCTTTAGGG AAATAGGCCA GGTTTTCACC GTAACACGCC ACATCTTGCG AATATATGTG TAGAAACTGC
CGGAAATCGT CGTGGTATTC ACTCCAGAGC GATGAAAACG TTTCAGTTTG CTCATGGAAA ACGGTGTAAC
AAGGGTGAAC ACTATCCCAT ATCACCAGCT CACCGTCTTT CATTGCCATA CGGAATTCCG GATGAGCATT
CATCAGGCGG GCAAGAATGT GAATAAAGGC CGGATAAAAC TTGTGCTTAT TTTTCTTTAC GGTCTTTAAA
AAGGCCGTAA TATCCAGCTG AACGGTCTGG TTATAGGTAC ATTGAGCAAC TGACTGAAAT GCCTCAAAAT
GTTCTTTACG ATGCCATTGG GATATATCAA CGGTGGTATA TCCAGTGATT TTTTTCTCCA TTTTAGCTTC
CTTAGCTCCT GAAAATCTCG ATAACTCAAA AAATACGCCC GGTAGTGATC TTATTTCATT ATGGTGAAAG
TTGGAACCTC TTACGTGCCG ATCAACGTCT CATTTTCGCC AAAAGTTGGC CCAGGGCTTC CCGGTATCAA
CAGGGACACC AGGATTTATT TATTCTGCGA AGTGATCTTC CGTCACAGGT ATTTATTCGG CGCAAAGTGC
GTCGGGTGAT GCTGCCAACT TACTGATTTA GTGTATGATG GTGTTTTTGA GGTGCTCCAG TGGCTTCTGT
TTCTATCAGC TGTGAGCTCG GTTTGAGGAG TGTTCAGTCT CCGTGAACGC CTCCTGTTCA GCTACTGACG
GGGTGGTGCG TAACGGCAAA AGCACCGCCG GACATCAGCG CTAGCGGAGT GTATACTGGC TTACTATGTT

GGCACTGATG AGGGTGTCAG TGAAGTGCTT CATGTGGCAG GAGAAAAAAG GCTGCACCGG TGCGTCAGCA
GAATATGTGA TACAGGATAT ATTCCGCTTC CTCGCTCACT GACTCGCTAC GCTCGGTCGT TCGACTGCGG
CGAGCGGAAA TGGCTTACGA ACGGGGCGGA GATTTCCTGG AAGATGCCAG GAAGATACTT AACAGGGAAG
TGAGAGGGCC GCGGCAAAGC CGTTTTTCCA TAGGCTCCGC CCCCCTGACA AGCATCACGA AATCTGACGC
TCAAATCAGT GGTGGCGAAA CCCGACAGGA CTATAAAGAT ACCAGGCGTT TCCCCCTGGC GGCTCCCTCG
TGCGCTCTCC TGTTCCTGCC TTTCGGTTTA CCGGTGTCAT TCCGCTGTTA TGGCCGCGTT TGTCTCATTC
CACGCCTGAC ACTCAGTTCC GGGTAGGCAG TTCGCTCCAA GCTGGACTGT ATGCACGAAC CCCCCGTTCA
GTCCGACCGC TGCGCCTTAT CCGGTAACTA TCGTCTTGAG TCCAACCCGG AAAGACATGC AAAAGCACCA
CTGGCAGCAG CCACTGGTAA TTGATTTAGA GGAGTTAGTC TTGAAGTCAT GCGCCGGTTA AGGCTAAACT
GAAAGGACAA GTTTTGGTGA CTGCGCTCCT CCAAGCCAGT TACCTCGGTT CAAAGAGTTG GTAGCTCAGA
GAACCTTCGA AAAACCGCCC TGCAAGGCGG TTTTTTCGTT TTCAGAGCAA GAGATTACGC GCAGACCAAA
ACGATCTCAA GAAGATCATC TTATTAATCA GATAAAATAT T

Supplementary table 8.1. Primer list

Name	Sequence
BG7709	TTTAAGAAGGAGATATAAGTCATGTCAATTTATCAAGAATTTGTTAATAAATATAG
BG7710	TTATGGAGTTGGAGTCTTATTATTAGTTATTCCTATTCTGCACG
BG7802	ACTCCAACTCCATAAGGATCCTAGAGCGGCCGCCAC
BG7803	ACTTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGG
BG8998	CATGGCGAATGTATCAAAGCAGC
BG9443	TACGTCGACGTCTAAGAACTTTAAATAATTTCTACTGTTG
BG9446	TACGTCGACCTGGAGGGTAAGGTTTCCCATCGAATCTACAACAGTAGAAATTATTTAAAG
BG9448	TACGTCGACACTCCACGTGCTCATTGGTAAGTCATCTACAACAGTAGAAATTATTTAAAG
BG9450	TACGTCGACACTCCACTGAACCTGGGGAATATTATCTACAACAGTAGAAATTATTTAAAG
BG9452	TACGTCGACACTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTATTTAAAG
BG9454	TACGTCGACACTCCACTACAATGATGGACTGGTATCTACAACAGTAGAAATTATTTAAAG
BG9456	TACGTCGACACTCCACAATGATCTCGTAGGCGTATCTACAACAGTAGAAATTATTTAAAG
BG9458	TACGTCGACACTCCAGCTAGTGTTACGGGAGCAATCTACAACAGTAGAAATTATTTAAAG
BG9460	TACGTCGACACTCCAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG9462	TACGTCGACACTCCAAGCTCCGGTGCATATAGTATCTACAACAGTAGAAATTATTTAAAG
BG9464	TACGTCGACACTCCATTGGGACCGGTAATTGTGATCTACAACAGTAGAAATTATTTAAAG
BG9558	TACGTCGACACTCACGTAAGCGGGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG9560	TACGTCGACACTCACGTAAGCGTGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG9562	TACGTCGACACTCACGTAAGCGGGGTTGGCTGGATCTACAACAGTAGAAATTATTTAAAG
BG9564	TACGTCGACACTCACGTAAGCGTGGTTGGCTGGATCTACAACAGTAGAAATTATTTAAAG
BG9821	GGCGAGCTCACAGCTGATAGAAACAGAAGCCAC
BG9822	GGCGAGCTCGGTTTATCGATGGGAAACCTTACCCTCCAGCCTCCTGTTCAGCTACTGACG
BG9823	GGCGAGCTCGGTTTAGACTTACCAATGAGCACGTGGAGTCCTCCTGTTCAGCTACTGACG
BG9824	GGCGAGCTCGGTTTAAATATTCCCCAGGTTCAGTGGAGTCCTCCTGTTCAGCTACTGACG
BG9825	GGCGAGCTCGGTTTAACACACTGCAATTCAGGTTGGAGTCCTCCTGTTCAGCTACTGACG

BG9826	GGCGAGCTCGGTTTAACCAGTCCATCATTGTAGTGGAGTCCTCCTGTTCAGCTACTGACG
BG9827	GGCGAGCTCGGTTTAACGCCTACGAGATCATTGTGGAGTCCTCCTGTTCAGCTACTGACG
BG9828	GGCGAGCTCGGTTTATGCTCCCGTAACACTAGCTGGAGTCCTCCTGTTCAGCTACTGACG
BG9829	GGCGAGCTCGGTTTACCAGTCTAAATCGCTTACTGGAGTCCTCCTGTTCAGCTACTGACG
BG9830	GGCGAGCTCGGTTTAACTATATGCACCGGAGCTTGGAGTCCTCCTGTTCAGCTACTGACG
BG9831	GGCGAGCTCGGTTTACACAATTACCGGTCCCAATGGAGTCCTCCTGTTCAGCTACTGACG
BG9832	GGCGAGCTCGGTTTACCAGTCTAACCCGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG9833	GGCGAGCTCGGTTTACCAGTCTAACACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG9834	GGCGAGCTCGGTTTACCAGCCAACCCCGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG9835	GGCGAGCTCGGTTTACCAGCCAACCACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG10140	ACTGGATCCTTAAGAAGCGTCAGCGTAGTTTTCGTCGTTAGCAGCGTTATTCCTATTCTGCACGA ACTC
BG10196	GGCGAGCTCGGTTTGAGGAGTGTTCAGTCTCCGTGAACGCCTCCTGTTCAGCTACTGACG
BG10468	TACGTCGACCCCCGCTTCGGCGGGGTTTTTTCCTAGGACTAGTCTTATTCAG
BG10471	GGCGAGCTCGGTTTCCTGGAGGTACACTATCGCATGAGTCCTCCTGTTCAGCTACTGACG
BG10472	GCGTCGACCTGGAGTGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10473	GCGTCGACTTCTGGTGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10474	GCGTCGACGTGGAGTGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10475	GCGTCGACCAGGTATGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10476	GCGTCGACCTACACTGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10477	GCGTCGACGACTATTGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10478	GGCGAGCTCGGTTTCCCTGAGGAATGCATCGCTATGAGTCCTCCTGTTCAGCTACTGACG
BG10479	GCGTCGACCCTGAGTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10480	GCGTCGACTTCCTGTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10481	GCGTCGACGCTGAGTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10482	GCGTCGACCAGGAATAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10483	GCGTCGACCAATGCTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10484	GCGTCGACAGCATCTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10485	GGCGAGCTCGGTTTCCTTGGAAGAGTCACCATGCTGAGTCCTCCTGTTCAGCTACTGACG
BG10486	GCGTCGACCTTGGAGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10487	GCGTCGACTTCTTGGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10488	GCGTCGACGTTGGAGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10489	GCGTCGACCGAAGAGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10490	GCGTCGACTGAGTCGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10491	GCGTCGACCTCACCGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10509	GCGTCGACACTGGATGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10510	GCGTCGACCGAGGTTGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG

BG10511	GCGTCGACCGTACATGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10512	GCGTCGACTCACTATGCGATAGTGTACCTCCAGATCTACAACAGTAGAAATTATTTAAAG
BG10513	GCGTCGACACCTGATAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10514	GCGTCGACAGAGGATAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10515	GCGTCGACCGAATGTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10516	GCGTCGACTTGCATTAGCGATGCATTCCTCAGGATCTACAACAGTAGAAATTATTTAAAG
BG10517	GCGTCGACACTTGGGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10518	GCGTCGACAGGAAGGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10519	GCGTCGACTAGAGTGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG10520	GCGTCGACTGTCACGCATGGTGACTCTTCCAAGATCTACAACAGTAGAAATTATTTAAAG
BG11244	GGCGAGCTCGGTTTACCAGTCTAACTCGCTTACTGGAGTCCTCCTGTTCAGCTACTGACG
BG11245	TACGTCGACACTCCAGTAAGCGAGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11246	GGCGAGCTCGGTTTACCAGTCTAAAACGCTTACTGGAGTCCTCCTGTTCAGCTACTGACG
BG11247	TACGTCGACACTCCAGTAAGCGTTTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11248	GGCGAGCTCGGTTTACCAGTCTAAATCGCTTACGGGAGTCCTCCTGTTCAGCTACTGACG
BG11249	TACGTCGACACTCCCGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11250	GGCGAGCTCGGTTTACCAGTCTAAATCGCTTACTTGAGTCCTCCTGTTCAGCTACTGACG
BG11251	TACGTCGACACTCAAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11252	GGCGAGCTCGGTTTACCAGTCTAACACGCTTACTGGAGTCCTCCTGTTCAGCTACTGACG
BG11253	TACGTCGACACTCCAGTAAGCGTGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11254	GGCGAGCTCGGTTTACCAGTCTAACTCGCTTACGGGAGTCCTCCTGTTCAGCTACTGACG
BG11255	TACGTCGACACTCCCGTAAGCGAGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11256	GGCGAGCTCGGTTTACCAGTCTAACTCGCTTACTTGAGTCCTCCTGTTCAGCTACTGACG
BG11257	TACGTCGACACTCAAGTAAGCGAGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11258	GGCGAGCTCGGTTTACCAGTCTAAAACGCTTACGGGAGTCCTCCTGTTCAGCTACTGACG
BG11259	TACGTCGACACTCCCGTAAGCGTTTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11260	GGCGAGCTCGGTTTACCAGTCTAAAACGCTTACTTGAGTCCTCCTGTTCAGCTACTGACG
BG11261	TACGTCGACACTCAAGTAAGCGTTTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11262	GGCGAGCTCGGTTTACCAGTCTAAATCGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11263	TACGTCGACACTCACGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11264	GGCGAGCTCGGTTTACCAGTCTAACACGCTTACGGGAGTCCTCCTGTTCAGCTACTGACG
BG11265	TACGTCGACACTCCCGTAAGCGTGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11266	GGCGAGCTCGGTTTACCAGTCTAACACGCTTACTTGAGTCCTCCTGTTCAGCTACTGACG
BG11267	TACGTCGACACTCAAGTAAGCGTGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11268	GGCGAGCTCGGTTTACCAGTCTAACTCGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11269	TACGTCGACACTCACGTAAGCGAGTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11270	GGCGAGCTCGGTTTACCAGTCTAAAACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG

BG11271	TACGTCGACACTCACGTAAGCGTTTTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11435	GGCGAGCTCGGTTTACCAGCCTAACACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11436	TACGTCGACACTCACGTAAGCGTGTTAGGCTGGATCTACAACAGTAGAAATTATTTAAAG
BG11437	GGCGAGCTCGGTTTACCAGTCAAACACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11438	TACGTCGACACTCACGTAAGCGTGTTTGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11439	GGCGAGCTCGGTTTACCAGTCTACCACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11440	TACGTCGACACTCACGTAAGCGTGGTAGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11441	GGCGAGCTCGGTTTACCAGCCAAACACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11442	TACGTCGACACTCACGTAAGCGTGTTTGGCTGGATCTACAACAGTAGAAATTATTTAAAG
BG11443	GGCGAGCTCGGTTTACCAGCCTACCACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11444	TACGTCGACACTCACGTAAGCGTGGTAGGCTGGATCTACAACAGTAGAAATTATTTAAAG
BG11445	GGCGAGCTCGGTTTACCAGTCAACCACGCTTACGTGAGTCCTCCTGTTCAGCTACTGACG
BG11446	TACGTCGACACTCACGTAAGCGTGGTTGACTGGATCTACAACAGTAGAAATTATTTAAAG
BG11629	TACTGTCGACCGTTCACGGAGACTGAACACTCCTACCTACAACAGTAGGAATTAATTATTTAAAG TTCTTAGACCCGTTTTTGC
BG11630	TACTGTCGACCGTTCACGGAGACTGAACACTCCTATCCACAACAGTGGAAATTAATTATTTAAAG TTCTTAGACCCGTTTTTGC
BG11631	TACTGTCGACCGTTCACGGAGACTGAACACTCCTATCTGCAACAGCAGAAATTAATTATTTAAAG TTCTTAGACCCGTTTTTGC
BG11632	TACTGTCGACCGTTCACGGAGACTGAACACTCCTATCTACCGAAGTAGAAATTAATTATTTAAAG TTCTTAGACCCGTTTTTGC
BG11633	TACTGTCGACCGTTCACGGAGACTGAACACTCCTACCCGCCGAAGCGGGAATTAATTATTTAAAG TTCTTAGACCCGTTTTTGC
BG11634	TACTGTCGACCGTTCACGGAGACTGAACACTCCTACCCCGCCGAAGCGGGGAATTAATTATTTAA AGTTCTTAGACCCGTTTTTGC
BG11635	TACTGTCGACCGTTCACGGAGACTGAACACTCCTATCTACAACAGTAGAAATTCCACACATTATA CGAGCCGGAAG
BG12039	TACGTCGACAATTTCTACTGTTGTAGATAGATACCGGAC
BG13467	GGCTCTTCACCCCGCTTCGGCGGGGTTTTTTCCTAGGACTAGTCTTATTCAG
BG13468	GGCTCTTCAAATTTCTACTGTTGTAGATAGATACCGGAC
BG13469	GGCTCTTCAACCGAGCTCACAGCTGATAG
BG13470	GGCTCTTCAGGTTTACCAGTGATCTACTCTTCGGCCTCCTGTTCAGCTACTGACG
BG13471	GGCTCTTCAGGGCCGAAGAGTAGATCACTGGATCTACAACAGTAGAAATTCCAC
BG13472	GGCTCTTCAGGGCTACTCCGAAGAGTAGATCACTGGATCTACAACAGTAGAAATTCCAC
BG13473	GGCTCTTCAATTCCGAAGAGTAGATCACTGGATCTACAACAGTAGAAATTCCAC
BG13474	GGCTCTTCAATTCTACTCCGAAGAGTAGATCACTGGATCTACAACAGTAGAAATTCCAC
BG13475	GGCTCTTCAGGTTTAACACACTGCAATTCAGGTTGGAGTCCTCCTGTTCAGCTACTGACG
BG13476	GGCTCTTCAGGGACTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13477	GGCTCTTCAGGGCTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13478	GGCTCTTCAGGGTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC

BG13479	GGCTCTTCAGGGCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13480	GGCTCTTCAGGGCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13481	GGCTCTTCAGGGAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13482	GGCTCTTCAGGGACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13483	GGCTCTTCAGGGCCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13484	GGCTCTTCAATTACTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13485	GGCTCTTCAATTCTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13486	GGCTCTTCAATTTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13487	GGCTCTTCAATTCCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13488	GGCTCTTCAATTCAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13489	GGCTCTTCAATTAACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13490	GGCTCTTCAATTACCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13491	GGCTCTTCAATTCCTGAATTGCAGTGTGTATCTACAACAGTAGAAATTCCAC
BG13678	GGCTCTTCAGGGNNNNNCCGAAGAGTAGATCACTGGATCTACAACAGTAGAAATTCCAC
BG13679	GGCTCTTCAATTNNNNNCCGAAGAGTAGATCACTGGATCTACAACAGTAGAAATTCCAC
BG13936	GGCTCTTCAGGTTTACCAGTCTAAATCGCTTACTCCTCCTGTTCAGCTACTGACG
BG13937	GGCTCTTCAATTACCACTCTTAATGGCTAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAAT TCCAC
BG13938	GGCTCTTCAATTTGTCAAAAACGCAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCA C
BG13939	GGCTCTTCAATTGAGTGTAATTCGAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCA C
BG13940	GGCTCTTCAATTGCAGTGTAAAACGCTAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATT CCAC
BG13941	GGCTCTTCAATTACCAGACTAATTCGCAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATT CCAC
BG13942	GGCTCTTCAATTCTCTATATCGAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCAC
BG13943	GGCTCTTCAATTCTCTAATTCGCTAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCA C
BG13944	GGCTCTTCAATTTGTCTATATCGCAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCA C
BG13945	GGCTCTTCAATTGAGTCTTAATCGAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCA C
BG13946	GGCTCTTCAATTATCTTAATGGCTAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCA C
BG13947	GGCTCTTCAATTAAAAACGCAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCAC
BG13948	GGCTCTTCAATTATAATTCGAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCAC
BG13949	GGCTCTTCAATTATAAAACGCTAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCAC
BG13950	GGCTCTTCAATTACTAATTCGCAGTAAGCGATTTAGACTGGATCTACAACAGTAGAAATTCCAC
BG19146	CATCTACAACAGTAGAAATTATTTAAAGTTCTTAGACCACACATTATACGAGCCGGAAGC
BG19150	GGCTCTTCAGGTTTGAAATAAAACTTTAAAACATGCCACCTCCTGTTCAGCTACTGACG

BG19151	GGCTCTTCAGGTTTAAAACATGCCATAAAAAGCAACTACCTCCTGTTCAGCTACTGACG
BG19152	GGCTCTTCAGGTTTAACATTTTGGAATGTCCTCGTTGGCCTCCTGTTCAGCTACTGACG
BG19153	GGCTCTTCAGGTTTCACTCATTTGACTGTTTTTGGAGCCCTCCTGTTCAGCTACTGACG
BG19154	GGCTCTTCAGGTTTTCTTATGCTCCATGTATGCGCCCGCCTCCTGTTCAGCTACTGACG
BG19155	GGCTCTTCAGGTTTAAGATGCTGACTGTTTTTGAAATACCTCCTGTTCAGCTACTGACG
BG19156	GGCTCTTCAGGTTTGGAGCCATGACTTTTAACATTTTGCCTCCTGTTCAGCTACTGACG
BG19157	GGCTCTTCAGGTTTCGGGGGCTCCGTAACTTTCTATCCCCTCCTGTTCAGCTACTGACG
BG19158	GGCTCTTCAGGTTTAAGGGCCGCTCTAGGCCAGGATGGCCTCCTGTTCAGCTACTGACG
BG19159	GGCTCTTCAGGTTTGGAGCATCAGATGTATTCCCTAAGCCTCCTGTTCAGCTACTGACG
BG19160	GGCTCTTCAGGTTTGGTTTGTGGTTTTGTGGGGTCGTTCCTCCTGTTCAGCTACTGACG
BG19161	GGCTCTTCAGGTTTGAAATGTGCTCCCAGCTGCAGGCCCCTCCTGTTCAGCTACTGACG
BG19162	GGCTCTTCAGGTTTCCCATAGTTACAATCAAAGACACTCCTCCTGTTCAGCTACTGACG
BG19163	GGCTCTTCAGGTTTCGCGCGCCACCAGGCCGTAGTGTCCCTCCTGTTCAGCTACTGACG
BG19164	GGCTCTTCAGGTTTCCATCTTTGGAGCATCAGATGTATCCTCCTGTTCAGCTACTGACG
BG19165	GGCTCTTCAGGTTTAGCTGTCACTGAAAGAGATTTATTCCTCCTGTTCAGCTACTGACG
BG19166	GGCTCTTCAATTaaaaaatggcatgttttaaagttttatttATCTACAACAGTAGAAATTATTTA AAG
BG19167	GGCTCTTCAATTaaaaaatagttgctttttatggcatgtttATCTACAACAGTAGAAATTATTTA AAG
BG19168	GGCTCTTCAATTaaaaaaccaacgaggacattccaaaatgtATCTACAACAGTAGAAATTATTTA AAG
BG19169	GGCTCTTCAATTaaaaaagctccaaaaacagtcaaatgagtATCTACAACAGTAGAAATTATTTA AAG
BG19170	GGCTCTTCAATTaaaaaacgggcgcatacatggagcataagATCTACAACAGTAGAAATTATTTA AAG
BG19171	GGCTCTTCAATTaaaaaatatttcaaaaacagtcagcatctATCTACAACAGTAGAAATTATTTA AAG
BG19172	GGCTCTTCAATTaaaaaacaaaatgttaaaagtcatggctcATCTACAACAGTAGAAATTATTTA AAG
BG19173	GGCTCTTCAATTaaaaaaggatagaaagttacggagcccccATCTACAACAGTAGAAATTATTTA AAG
BG19174	GGCTCTTCAATTaaaaaaccatcctggcctagagcggccctATCTACAACAGTAGAAATTATTTA AAG
BG19175	GGCTCTTCAATTaaaaaacttagggaatacatctgatgctcATCTACAACAGTAGAAATTATTTA AAG
BG19176	GGCTCTTCAATTaaaaaaaacgaccccacaaaaccacaaacATCTACAACAGTAGAAATTATTTA AAG
BG19177	GGCTCTTCAATTaaaaaaggcctgcagctgggagcacatttATCTACAACAGTAGAAATTATTTA AAG
BG19178	GGCTCTTCAATTaaaaaaagtgtctttgattgtaactatggATCTACAACAGTAGAAATTATTTA AAG
BG19179	GGCTCTTCAATTaaaaaagacactacggcctggtggcgcgcATCTACAACAGTAGAAATTATTTA AAG

BG19180	GGCTCTTCAATTaaaaaaatacatctgatgctccaaagatgATCTACAACAGTAGAAATTATTTA AAG
BG19181	GGCTCTTCAATTaaaaaaaataaatctctttcagtgacagcATCTACAACAGTAGAAATTATTTA AAG
BG19182	GGCTCTTCAATTgttaaaatgttttaaagttttatttATCTACAACAGTAGAAATTCCAC
BG19183	GGCTCTTCAATTgtaaaatgctttttatggcatgtttATCTACAACAGTAGAAATTCCAC
BG19184	GGCTCTTCAATTgatgtccgaggacattccaaaatgtATCTACAACAGTAGAAATTCCAC
BG19185	GGCTCTTCAATTgctgttcaaaaacagtcaaatgagtATCTACAACAGTAGAAATTCCAC
BG19186	GGCTCTTCAATTgatgtacgcatacatggagcataagATCTACAACAGTAGAAATTCCAC
BG19187	GGCTCTTCAATTgtgttttcaaaaacagtcagcatctATCTACAACAGTAGAAATTCCAC
BG19188	GGCTCTTCAATTgttttaatgttaaaagtcatggctcATCTACAACAGTAGAAATTCCAC
BG19189	GGCTCTTCAATTctaactagaaagttacggagcccccATCTACAACAGTAGAAATTCCAC
BG19190	GGCTCTTCAATTgtaggccctggcctagagcggcccctATCTACAACAGTAGAAATTCCAC
BG19191	GGCTCTTCAATTgtgtatgggaatacatctgatgctcATCTACAACAGTAGAAATTCCAC
BG19192	GGCTCTTCAATTgttgtgaccccacaaaaccacaaacATCTACAACAGTAGAAATTCCAC
BG19193	GGCTCTTCAATTgccagctgcagctgggagcacatttATCTACAACAGTAGAAATTCCAC
BG19194	GGCTCTTCAATTgaatcatctttgattgtaactatggATCTACAACAGTAGAAATTCCAC
BG19195	GGCTCTTCAATTgaggccctacggcctggtggcgcgcATCTACAACAGTAGAAATTCCAC
BG19196	GGCTCTTCAATTggcatcatctgatgctccaaagatgATCTACAACAGTAGAAATTCCAC
BG19197	GGCTCTTCAATTcaaagaaatctctttcagtgacagcATCTACAACAGTAGAAATTCCAC

Supplementary table 2. Construct list. Combinations of fragments yield different pTarget plasmids (see materials and methods)

Name	Type	Template	Fwd	Rev
Target1 - F1	SalI-|T|-cat-SacI	pTarget3 - DNMT1 (2)	BG10468	BG9821
Target2 - F1	SalI-MR-cat-SacI	pTarget3 - DNMT1 (2)	BG12039	BG9821
Target3 - F1	SalI-FR-cat-SacI	pTarget3 - DNMT1 (2)	BG9443	BG9821
Target7 - F1	SapI-|T|-cat-SapI	pTarget3 - DNMT1 (2)	BG13467	BG13469
Target8 - F1	SapI-MR-cat-SapI	pTarget3 - DNMT1 (2)	BG13468	BG13469
Target1 - Sp1 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9822	BG9446
Target1 - Sp2 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9823	BG9448
Target1 - Sp3 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9824	BG9450
Target1 - Sp4 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9825	BG9452
Target1 - Sp5 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9826	BG9454
Target1 - Sp6 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9827	BG9456
Target1 - Sp7 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9828	BG9458
Target1 - Sp8 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9829	BG9460
Target1 - Sp9 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9830	BG9462
Target1 - Sp10 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9831	BG9464

Target1 - Sp11 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9832	BG9558
Target1 - Sp12 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9833	BG9560
Target1 - Sp13 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9834	BG9562
Target1 - Sp14 - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG9835	BG9564
Target1 - Backfold lib A (DR) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10472
Target1 - Backfold lib A (-1) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10473
Target1 - Backfold lib A (1) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10509
Target1 - Backfold lib A (2) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10474
Target1 - Backfold lib A (4) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10510
Target1 - Backfold lib A (5) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10475
Target1 - Backfold lib A (7) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10511
Target1 - Backfold lib A (8) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10476
Target1 - Backfold lib A (10) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10512
Target1 - Backfold lib A (11) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10471	BG10477
Target1 - Backfold lib B (DR) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10479
Target1 - Backfold lib B (-1) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10480
Target1 - Backfold lib B (1) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10513
Target1 - Backfold lib B (2) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10481
Target1 - Backfold lib B (4) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10514
Target1 - Backfold lib B (5) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10482
Target1 - Backfold lib B (7) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10515
Target1 - Backfold lib B (8) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10483
Target1 - Backfold lib B (10) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10516
Target1 - Backfold lib B (11) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10478	BG10484
Target1 - Backfold lib C (DR) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10486
Target1 - Backfold lib C (-1) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10487
Target1 - Backfold lib C (1) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10517
Target1 - Backfold lib C (2) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10488
Target1 - Backfold lib C (4) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10518
Target1 - Backfold lib C (5) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10489
Target1 - Backfold lib C (7) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10519
Target1 - Backfold lib C (8) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10490
Target1 - Backfold lib C (10) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10520
Target1 - Backfold lib C (11) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10485	BG10491
Target1 - Sp8 [cUUG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11244	BG11245
Target1 - Sp8 [AaUG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11246	BG11247
Target1 - Sp8 [AUgG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11248	BG11249
Target1 - Sp8 [AUUu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11250	BG11251
Target1 - Sp8 [caUG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11252	BG11253

Target1 - Sp8 [cUgG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11254	BG11255
Target1 - Sp8 [cUUu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11256	BG11257
Target1 - Sp8 [AagG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11258	BG11259
Target1 - Sp8 [AaUu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11260	BG11261
Target1 - Sp8 [AUgu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11262	BG11263
Target1 - Sp8 [cagG] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11264	BG11265
Target1 - Sp8 [caUu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11266	BG11267
Target1 - Sp8 [cUgu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11268	BG11269
Target1 - Sp8 [Aagu] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11270	BG11271
Target1 - Sp12 [cuA] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11435	BG11436
Target1 - Sp12 [uaA] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11437	BG11438
Target1 - Sp12 [uuc] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11439	BG11440
Target1 - Sp12 [caA] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11441	BG11442
Target1 - Sp12 [cuc] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11443	BG11444
Target1 - Sp12 [uac] - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG11445	BG11446
Target1 - DNMT1 (2) (R_+2) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11629
Target1 - DNMT1 (2) (R_+4) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11630
Target1 - DNMT1 (2) (R_+5) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11631
Target1 - DNMT1 (2) (R_TL) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11632
Target1 - DNMT1 (2) (R_Max) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11633
Target1 - DNMT1 (2) (R_Term) - F2	SacI-FR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11634
Target4 - DNMT1 (2) - F2	SacI-MR-Sp-SalI	pTarget3 - DNMT1 (2)	BG10196	BG11635
Target4 - Sp12 [cUA] - F2	SacI-MR-Sp-SalI	pTarget4 - DNMT1 (2)	BG11435	BG12041
Target4 - Sp8 - F2	SacI-MR-Sp-SalI	pTarget4 - DNMT1 (2)	BG9829	BG12042
Target7 - "Impaired" - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13470	BG13471
Target7 - "Rescued" - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13470	BG13472
Target8 - "Impaired" - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13470	BG13473
Target8 - "Rescued" - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13470	BG13474
Target7 - Sp4 (24 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13476
Target7 - Sp4 (23 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13477
Target7 - Sp4 (22 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13478
Target7 - Sp4 (21 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13479
Target7 - Sp4 (20 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13480
Target7 - Sp4 (19 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13481
Target7 - Sp4 (18 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13482
Target7 - Sp4 (17 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13483
Target8 - Sp4 (24 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13484
Target8 - Sp4 (23 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13485
Target8 - Sp4 (22 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13486

Target8 - Sp4 (21 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13487
Target8 - Sp4 (20 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13488
Target8 - Sp4 (19 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13489
Target8 - Sp4 (18 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13490
Target8 - Sp4 (17 nt) - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13491
Target7 - IS - N5 - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13678
Target8 - IS - N5 - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13475	BG13679
Target8 Sp8 [S3.L4] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13937
Target8 Sp8 [S3.L5] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13938
Target8 Sp8 [S3.L6] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13939
Target8 Sp8 [S4.L4] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13940
Target8 Sp8 [S4.L5] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13941
Target8 Sp8 [S4.L6] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13942
Target8 Sp8 [S5.L4] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13943
Target8 Sp8 [S5.L5] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13944
Target8 Sp8 [S5.L6] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13945
Target8 Sp8 [S3.L4.alt] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13946
Target8 Sp8 [S3.L5.alt] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13947
Target8 Sp8 [S3.L6.alt] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13948
Target8 Sp8 [S4.L4.alt] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13949
Target8 Sp8 [S4.L5.alt] - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG13936	BG13950
Target9 F2 Template	SapI-FR-Sp-SapI	pTarget3 - DNMT1 (2)	BG19150	BG19146
Target9 Kim original A - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19150	BG19166
Target9 Kim original B - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19151	BG19167
Target9 Kim original C - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19152	BG19168
Target9 Kim original D - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19153	BG19169
Target9 Kim original E - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19154	BG19170
Target9 Kim original F - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19155	BG19171
Target9 Kim original G - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19156	BG19172
Target9 Kim original H - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19157	BG19173
Target9 Kim original I - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19158	BG19174
Target9 Kim original J - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19159	BG19175
Target9 Kim original K - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19160	BG19176
Target9 Kim original L - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19161	BG19177
Target9 Kim original M - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19162	BG19178
Target9 Kim original N - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19163	BG19179
Target9 Kim original O - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19164	BG19180
Target9 Kim original P - F2	SapI-FR-Sp-SapI	Target9 F2 Template	BG19165	BG19181
Target8 Kim tailored A - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19150	BG19182

Target8 Kim tailored B - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19151	BG19183
Target8 Kim tailored C - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19152	BG19184
Target8 Kim tailored D - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19153	BG19185
Target8 Kim tailored E - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19154	BG19186
Target8 Kim tailored F - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19155	BG19187
Target8 Kim tailored G - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19156	BG19188
Target8 Kim tailored H - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19157	BG19189
Target8 Kim tailored I - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19158	BG19190
Target8 Kim tailored J - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19159	BG19191
Target8 Kim tailored K - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19160	BG19192
Target8 Kim tailored L - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19161	BG19193
Target8 Kim tailored M - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19162	BG19194
Target8 Kim tailored N - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19163	BG19195
Target8 Kim tailored O - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19164	BG19196
Target8 Kim tailored P - F2	SapI-MR-Sp-SapI	pTarget4 - DNMT1 (2)	BG19165	BG19197

Chapter 9

Summary and general discussion

After the seminal discoveries that led to the Central Dogma of Molecular Biology (DNA > RNA > Protein) (3), RNA was initially considered only as mediator between DNA and protein (mRNA, tRNA), and as structural part of the ribosome (rRNA). Proteins were then thought to to be responsible for catalysing all biological reactions, including the transcription and translation processes. In the mid-1980's, however, some RNA molecules have been demonstrated to possess catalytic activity (47, 136), just like some proteins (the enzymes). These so-called ribozymes were demonstrated to be responsible for self-splicing of introns during maturation of mRNA transcripts as well as for the maturation of tRNAs by the RNA-component of bacterial RNaseP (137). Even more spectacular, after revealing the details of the ribosome structure, also the key machinery responsible for protein synthesis appeared to rely on ribozyme activity for linking the amino acid building blocks (138). Moreover, in the early 1990s, *in vitro* selection methods demonstrated that RNA, again just like proteins, could potentially bind a variety of target molecules (11, 17, 139, 140). Last but not least, small RNA molecules have been demonstrated to have the potential to act as guide molecules for a range of proteins (e.g. Hfq, Argonaute, CRISPR-Cas), that target complementary RNA or DNA sequences, and as such allow for control of biological processes ranging from regulating gene expression to protecting hosts from invaders (108, 113, 141–143). These ground-breaking discoveries have led to a major upgrade of the impact of RNA in biology, and have provided interesting ideas on an RNA-world in the early evolution of life on earth (144–146). In this thesis, several aspects of fundamental RNA structure-function relations are addressed, resulting in insights with potential for innovative applications.

Aptamers

The ligand-binding RNAs, the so-called RNA aptamers, are capable of binding targets ranging from small molecules to proteins and with an affinity comparable to that of proteins binding their respective substrates. In 2002, the first riboswitches were described (147–149). At the same time, the group I intron from phage T4 located inside the *td* gene was engineered to have the intron splicing dependent on the presence of theophylline (36). Since the intron exhibits catalytic activity and is made of RNA, it is a ribonucleic enzyme, in short ribozyme; and outfitted with an aptamer, it becomes an aptazyme. The first chapters describe how this synthetic intron-based riboswitch can be used to function as a biosensor with different output signals. In chapter 2, the output signal is growth of *E. coli* in response to the presence of theophylline. The strain of *E. coli* used here has a genomic knockout of the vital *thyA* gene, which is responsible for the generation of the DNA precursor deoxythymidine monophosphate (dTMP), and is complemented by a plasmid borne *thyA* gene. The theophylline responsive intron (Figure 9.1B, 9.1C) interrupts the plasmid borne *thyA* gene. Splicing of the intron (mRNA maturation) restores the open reading frame (Figure 9.1A), and enables synthesis of dTMP; thus, dTMP production

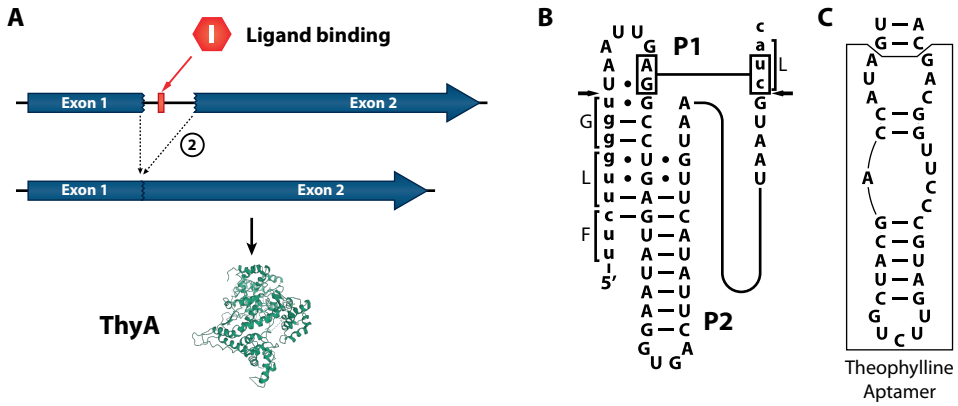


Figure 9.1. Overview of the riboswitch in *thyA*. (A) The ligand (theophylline) binds to the intron, which splices out of the mRNA restoring the open reading frame and allows for translation into a functional protein. (B) The 5′ and 3′ regions of the intron ribozyme. Arrows indicate the splice site. The intron interrupts the protein sequence FLGLP. (C) The aptamer is depicted in the boxed region binds to theophylline and some related compounds. When the shaded C is exchanged for A, the aptamer cannot bind theophylline anymore, but it still binds 3-methylxanthine.

depends on the presence of theophylline. However, the growth does not only depend on the presence of theophylline. dTMP production depends on ThyA, and ThyA depends on the mature mRNA and the translation rate. The concentration of mature mRNA in turn depends on the rate of transcription of non-mature mRNA, the rate of maturation (mol · mol-1 · s-1), and the rate of degradation of both mature and non-mature mRNA.

Tuning the biosensor

While the splicing of the intron heavily depends on the presence of its ligand theophylline, some splicing also occurs in its absence. The most ideal biosensor would show no signal without target ligand. Presence of the inducer should then increase the signal above the detection limit. In this case, the concentration of active ThyA should not exceed the minimal required in the absence of theophylline, while it should exceed it in its presence. In this system, there are two parameters that can be easily adjusted: transcription and translation. The transcription of *thyA* was selected as parameter to be varied and, therefore, several constructs were made with promoters of different strengths. The expected outcome was that for some of the promoters the transcription would be insufficient to support growth even in the presence of theophylline. Others would be so strong that uninduced splicing may already support growth and yet some promoters would show the desired outcome of supporting growth only in the presence of theophylline. The different expected outcomes were indeed observed, but only one of the constructs showed the theophylline being required for growth. This clearly indicated that the system has to be meticulously balanced to act as a proper biosensor. Modulating the translation rate

would likely enhance the robustness, but a more attractive option was to make the level of functional *thyA* mRNA more dependent on theophylline. The dTMP level, as a result of uninduced splicing, would be further below the minimum requirement for growth, while induction with theophylline would still be able to cause the dTMP level to rise above this minimum. Therefore, a second intron was introduced upstream of the first. The introduction of the second intron had major ramifications for the maturation, both in the presence and absence of theophylline. While the uninduced maturation rate was lowered more than the induced maturation rate, the transcription had to be increased to allow for enough ThyA with induction.

Introduction of the intron in other genes

Introduction of the second intron did make the system more robust, but caused another problem. The ribozyme is actually larger than the intron alone and includes a couple of nucleotides of the exons flanking it. The *thyA* coding sequence had to be altered slightly and the second intron did not behave in the exact same way the first one did. If the riboswitch were to be introduced in other genes, dissimilar from *thyA*, it should still work. There is an average of three codons per amino acid to encode proteins. So, there is a great likelihood of the mRNA being able to harbour an intron but still retain its encoded protein sequence. In chapter 3, the consequences for changing the intron flanking regions are determined. An intron that does not respond to theophylline was inserted into the *E. coli lacZ* gene. This gene was already widely used as reporter and was known to be able to harbour large insertions between the D6 and S7 amino acid. The intron with 15 nucleotide flanks was introduced and the flanks were altered from position -7 to -4 and on position +3 counting from the splice site (Figure 9.1B). The -7 to -4 positions base-pair in the P1 stem. Their respective binding partners in the intron also have the possibility to extend the P2 stem instead of the P1 stem (see Figure 1.1). This means that varying the -7 to -4 positions might not disrupt the whole structure of the ribozyme to render it inactive. The same is true for the +3 position which can extend the interactions with the loop of P1.

The results indicated that the wild-type intron was not the most active version, but that it performed well. Alterations made to the +3 position all had a negative impact on the splicing rate, as did the alterations on the -7 position. Changing the -6 to -4 positions yielded variable splice rates from almost inactive to 40% better than wild-type. In addition to the insertion of the intron in *lacZ*, an attempt was made to insert it into GFP. Inserting the intron into an enzyme like ThyA or LacZ worked well, but direct introduction in GFP proved problematic. The fluorescent signal does not exceed the detection limit. To still be able to prove that the insertion in other proteins works, another enzyme was selected. The T7 RNA polymerase (RNAP) was a good candidate because of its size and the possibility to use GFP as a reporter, albeit indirectly. Three positions were identified in RNAP that would allow for decent splicing of the intron. These positions also had the same intron flanks and

were distributed evenly over the ORF. The theophylline responsive intron was inserted in all three positions individually, integrated into the *E. coli* chromosome and tested for their theophylline response. The results showed that after splicing a functional polymerase was formed, and that theophylline caused a solid response with no difference between the positions of the intron. Flowcytometry analysis revealed that the GFP fluorescence is mostly binary. The cells divide into two populations – one fluorescent and the other not – and the concentration of theophylline mostly changes the population sizes. Increased theophylline concentrations only affected the fluorescence of the fluorescent population marginally.

Other types of riboswitches

The group I aptazyme described in chapters 2 and 3 contains an intron that is rather large and has a complex structure. Attempts were made to work also with simpler riboswitches including a translational riboswitch obtained by Topp and Gallivan (150) and a hammerhead aptazyme (14). These types of riboswitches rely on competition between several elements in the mRNA. These elements include the 5' UTR, RBS, and coding sequence. For different organisms and different genes, these elements are not the same. If the riboswitch is selected to block a certain RBS of GFP in *E. coli*, it may not be transferable to organisms with a different RBS consensus sequence. And if the GFP coding sequence is important for the structure of the riboswitch, it may not be exchanged for another gene.

The group I aptazyme does not suffer from these drawbacks, but since it needs to interrupt a coding sequence to work, the options for insertion are somewhat limited. While transcriptional riboswitches require knowledge of the promoter and start of a coding sequence, translational riboswitches require knowledge of the coding sequence only. As the intron may be inserted into the coding sequence itself, and hence there is no issue with potential alternative start sites. In addition to the aforementioned integration of the intron (taking into account matching flanking sequences), the intron can be added as a tag at the 5' end of the gene of interest. Again, because of the required flanks at DNA/RNA level, this will result in eventually adding 3 to 5 amino acids.

Changing the aptamer

One of the advantages of using a riboswitch as a biosensor, is that the generation of a binding domain is relatively straightforward. A protein has 20 possible amino acids, while nucleic acids only have 4 possibilities (excluding modifications like inosine). Contrary to protein synthesis, DNA synthesis is well-established and cheap and RNA is easy to make from DNA. Also, RNA aptamers can be quite short (tens of nucleotides) and still bind its target ligand with good affinity. However, the aptamer is not a riboswitch and significant screening and selection has to be performed to find aptamer/platform combinations that

are riboswitches. Transferring *in vitro* results to *in vivo* applications not always works out the way it was intended. The *in vivo* selection method described in Chapters 4 has no such drawback. It uses selection (selection gene ON in the presence of ligand supports growth) and counterselection (counterselection gene OFF in the absence of ligand supports growth) to enrich for ligand responsive intron variants. The selection method takes advantage of two properties: the T7 RNAP induction causes two distinct populations differing in size depending on the presence of ligand, and *thyA* can be used as a selection and counterselection gene. Selection is caused by the cell's requirement for dTMP, which is made by ThyA, and counterselection is based on the depletion of the folate cycle in the presence of trimethoprim. Using the same gene ensures the integrity of the coding sequence. Escaping the counterselection by mutating the *thyA* gene will not allow growth in the subsequent selection round and vice versa. Amplification by T7 RNAP means that the ThyA expression does not need to be balanced. When it is under control of T7 RNAP, it is produced in vast quantities or not at all. An additional benefit is the possibility to indirectly screen the progress with GFP that is also under control of RNAP. The sizes of the populations do matter somewhat. The larger the difference in populations between presence and absence of ligand, the easier it is to enrich for them. Bacteria harbouring an intron variant that always produces RNAP will grow well during selection, but will also suffer significantly during counterselection. An intron variant that never produces RNAP will have the exact opposite effect. Introns that act as riboswitches will allow the bacteria to grow during both the selection and counterselection stages.

The selection method was tested with the theophylline aptamer. The theophylline aptamer can also bind the related compound 3-methylxanthine. The selection was designed to enrich for variants that could still bind 3-methylxanthine (or better than the original theophylline aptamer) and could not bind theophylline. 11 nucleotides in the theophylline aptamer (Figure 9.1C) are present in a bulge and have a possible interaction with the ligand. All of them were randomised yielding 4×10^6 theoretical variants. The library was subjected to selection with 3-methylxanthine and counterselection with theophylline for 8 rounds each and progress was monitored by GFP expression with and without addition of 3-methylxanthine or theophylline. After 6 rounds of selection and counterselection, the riboswitch containing bacteria began to impact the total fluorescence significantly and after 8 rounds, this population was dominant. Sequencing analysis and 3-methylxanthine response assays demonstrated that several positions could only have one possible nucleotide. 4 of these were identical to the theophylline aptamer and 1 was different, exchanging C for A. The other nucleotides showed little preference as individual positions, but distinct combinations of nucleotides were identified associated with the best riboswitches.

The conversion of theophylline to 3-methylxanthine

Various organisms can degrade caffeine-like molecules. They are typically isolated from coffee plantations. One of these organisms is *Pseudomonas putida* CBB5 (151, 152). The N1 methyl group of theophylline is removed by the N-demethylase NdmA in collaboration with the redox enzyme NdmD. The 3-methylxanthine is formed from theophylline by demethylation of the N1, so adding theophylline should result in 3-methylxanthine in the presence of NdmAD. Attempts in *E. coli* and *P. putida* KT2440 did not yield positive results. *P. putida* was responsive to external 3-methylxanthine, but the conversion of theophylline to 3-methylxanthine either did not happen at all or was insufficient to generate a response above the detection limit.

Grafting a novel aptamer

Both the theophylline and 3-methylxanthine aptamers were known to work with the intron. The method was next tested with a novel aptamer (Chapter 5). Since a biosensor has to potentially measure internal metabolites, this aptamer should preferably bind something that is made by the bacteria. A known aptamer was selected that can bind citrulline, an intermediate in the arginine synthesis. Two genes are responsible for the conversion of ornithine to citrulline (*argF*, *argI*) and either of them can perform this role. Citrulline is then further converted into L-arginosuccinate by *argG*. Auxotrophy testing confirmed that *E. coli* with both *argF* and *argI* knocked out requires external arginine or citrulline. The aptamer was grafted onto the intron with a random communication module of different lengths. The selection and counterselection was performed in a similar fashion as earlier. The basic medium did not contain arginine or citrulline and was supplemented with these during the selection (citrulline) and counterselection (arginine) rounds. After 5 rounds of enrichment, one of the libraries showed response to citrulline, but it did not progress after that. The enrichment was analysed by sequencing and assaying the citrulline response for individual clones and it was determined that the culture was dominated by a single clone. This clone was used as the basis for the next set of experiments. The citrulline aptamer, like the theophylline aptamer, is not just one sequence. It has several possible nucleotides at certain positions close to the communication module. These positions were semi-randomised and individual clones were analysed for their citrulline response. The best performing clone did respond to citrulline, but the response was weak (under 2-fold). To verify the functionality of the intron, the intron was transferred to *lacZ*. It also responded to citrulline in *lacZ*, albeit weakly.

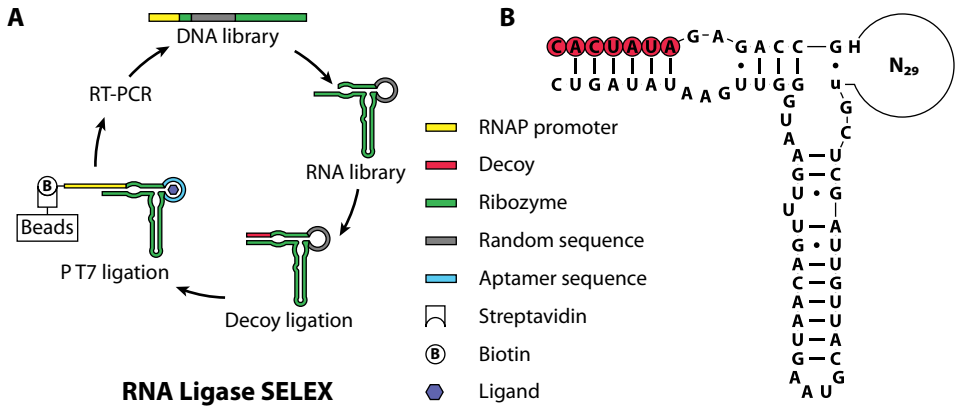


Figure 9.2. Overview of the RNA ligase SELEX. (A) One selection round of RNA ligase SELEX starting with a DNA library. (B) Sequence and structure of the transcribed library ligated to a decoy RNA (red), which makes up the last 7 nucleotides of the T7 RNAP promoter sequence.

IN VITRO ENRICHMENT OF RIBOSWITCHES

In an effort to generate new aptamer sequences *in vitro* that would fit a ribozyme like the group I intron, a strategy was devised based on an RNA ligase ribozyme (153–155) (Figure 9.2A, 9.2B). This ribozyme can ligate itself to another RNA that is complementary to its 3'-end. In the DNA library, one of the hairpins was exchanged for a random sequence of 30 nt, where the first was not allowed to be G. After transcription from the DNA template, the ribozymes that have ligase activity without ligand bind and ligate a decoy RNA. Next, a biotinylated RNA resembling the T7 RNAP promoter is added in excess along with the ligand. When the ligand induces ligase activity, the T7 RNAP promoter is ligated to the ribozyme, which allows the ligand responsive ligases to be bound to streptavidin beads. The RT-PCR subsequently regenerates the DNA library to start a new round. Unfortunately, the method did not enrich the correct variants and was abandoned in favour of the *in vivo* selection.

Another approach that was not actualised uses the group I intron itself. This approach would follow the Capture-SELEX routine (Figure 1.5B). The RNA would consist of the intron and two flanking regions and instead of binding the RNA in the middle with a probe, the flanking regions are bound. The introns that cannot splice out on their own will be bound to the magnetic bead and separated from the bead again with the addition of ligand. The RT-PCR then generates the DNA for the next round. One of the challenges to overcome would be the library generation, since the random nucleotides need to form the aptamer are located far from the start of the transcript, so it is difficult to add with PCR.

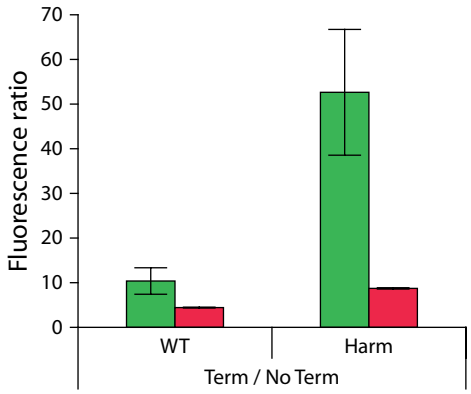


Figure 9.3. Comparison of two RFP-GFP operons with and without terminator. The wild-type (WT) and harmonised (Harm) both refer to the GFP. The values represent the ratio between the constructs with and without terminator.

Effects of RNA sequence on translation

RNA in the form of riboswitches affects the translation depending on external factors. The same mechanism that makes the riboswitch block translation (i.e. sequestering key elements like the RBS), can also be present permanently. The sequences surrounding the RBS, either the upstream 5'-UTR or the downstream part of the coding sequence, may form a secondary structure that includes the RBS, resulting in a drop of its availability for ribosome recruitment. This sequestering of the RBS can be avoided either by adjusting the codon use at the start of the coding sequence, or by using a bi-cistronic design in which a well-expressed small ORF is positioned upstream the gene of interest (87). Avoiding sequestering the RBS is not the only effect that codon use has. Hyper-thermophilic bacteria and archaea may benefit from high GC codons, for example, especially in the tRNA and rRNA genes, because of enhanced structural stability. The codon use of protein-coding genes is thought to be correlated to the availability of tRNAs, where the most abundant proteins are encoded by genes that make use of codons that correspond to the most abundant tRNAs. The more a codon is used, the more the tRNAs should be able to cope with the demand. This principle is reflected by the codon adaptivity index (CAI) (156). Not all organisms use codons at the same frequency, so transferring genes from one organism to the other, may result low production levels of these heterologous proteins. Several approaches have been suggested to do adapt the coding sequence to the new organism. These include 'codon optimization' by designing genes with the highest possible CAI, where all codons are encoded by the most abundant codon for the amino acid, or the codon distribution generally follows the codon usage of the new host. A more recent approach matches the codon frequency in the native organism with the codon frequency in the new host by individual positions and is called 'codon harmonisation' (80). Multidomain proteins may benefit from an occasional slow translation to allow for proper folding, increasing the functional protein content. To assay the effect of exchanging one codon for the other, a sensitive assay is required that can distinguish small differences

in translation. For this purpose, GFP codon variants were used with RFP as an internal standard. The two variants of GFP were either wild-type or harmonised and to ensure equal transcription, both GFP and RFP were located on the same mRNA. The use of RFP as internal control as part of the same operon proved impossible as it changed too much depending on the variant of GFP in front of it. A solution would be to have the GFP and RFP separately transcribed from the same plasmid. However, the equal transcription of the standard and the tested protein is not guaranteed. Also, to avoid overburdening the ribosomes, the transcription should be low.

Chapter 6 describes some of the ramifications that the operon design has on the translation of the genes of interest. Feed-forward translational coupling was already known (74–77), so the RFP and GFP were switched. The feed-back translational coupling we observed was even more pronounced than the feed-forward coupling. To generate mRNAs of uniform size, a strong rho-independent terminator was added to the construct separated 45 random nucleotides from the stop codon of the second ORF of the 2-gene operon (GFP).

The wild-type GFP showed more expression of both RFP and GFP, while we were not able to create the construct with the harmonized GFP, likely due to toxicity of the overexpression. The toxicity issue was mitigated by the introduction of a much weaker promoter (bla promoter) to lessen the transcription. This new set of constructs hardly showed any fluorescence, but the introduction of the terminator improved the fluorescence dramatically. Translational coupling was again observed for the two GFP variants, but also creating a frameshift in GFP resulted in a different expression level for RFP. A functional mutant, i.e. a point mutation in the fluorophore of GFP, had almost no effect on the RFP. The frameshift in GFP having such a profound effect on the expression of RFP has serious consequences. Bacterial genomes are notoriously dense and many genes are located in operons. Disruption of a gene may affect the neighbouring genes, leading to false conclusions about the disrupted gene. Therefore, an active site mutant or short in frame deletion to impact the protein structure is advised when doing genomic studies.

The addition of a terminator strongly impacted the translation rate, increasing GFP over 50-fold in one instance (Figure 9.3). The translation efficiency depends on the translation rate of an individual mRNA and the abundance of that mRNA. The terminator did affect the abundance of mRNA as well, but the RT-qPCR demonstrated that the increase was minor compared to the increase in translation. The codon use in GFP did not affect the mRNA abundance at all, so change in RFP cannot be solely caused by a change in mRNA abundance. It leads to another model. The intergenic region (IGR) is short (42 nt; 14 nm). A ribosome that has finished translating the first ORF stays attached to the RNA awaiting recycling (97). The ribosome is about 20 nm large, which is enough to block the whole IGR. The ribosomes waiting for recycling then complete the train of ribosomes from the second ORF to the first and the translation of the first ORF now depends on the translation

of the second ORF.

When the terminator was first introduced, a region of random nucleotides was added to space the end of the coding region and the terminator stem apart. An in-depth analysis of this region with GFP and RFP revealed that the sequence of the post stop-codon, ante terminator (PSAT) region can change the translation rate significantly (over 5-fold). However, the impact of the PSAT region follows a bell curve, where most PSAT sequences yield average GFP and RFP expression. Analysis of the PSAT region behind three different reporters (GFP, RFP, LacZ), demonstrated that the effect this region has on the translation is generally independent from the reporter it follows. This means that secondary structure formation with the ORF is unlikely, although the PSAT may still influence the terminator formation. The terminator proved vital in the expression, so if the terminator is compromised, the translation may be lowered as a result.

To exclude the interactions with the terminator as the prime mechanism, the PSAT regions were cloned behind the first ORF in an operon, increasing the IGR by 30 nt. Variants with and without terminator stem were made, to exclude the necessity of the stem for the PSAT mechanism. The insertion of the PSAT into the IGR does decouple the translations of the first and second ORF to some extent. Furthermore, the PSAT has the same effect on the translation when no terminator or terminator stem is present, ruling out the terminator interactions as prime mechanism. A possible explanation is that the PSAT influences the location and residence time of the ribosome after translation, but that needs to be ascertained.

The influence of RNA folding on CRISPR-Cas efficiency

The Clustered Regularly Interspaced Short Palindromic Repeats with associated proteins (CRISPR-Cas) technology is a very potent tool for genome editing and metabolic modulation in a variety of organisms. The CRISPR-Cas systems can target DNA or RNA by matching an RNA guide (crRNA) bound to the effector complex (Class 1) or effector protein (Class 2) (157). Many of the CRISPR-Cas systems exhibit DNase or RNase activity and can be exploited for counterselection after homologous recombination, transcription suppression and RNA degradation. The pre-crRNA is derived from an array of repeats and spacers. The effector recognises the repeat-derived part of the pre-crRNA and binds to it. Depending on the type, the effector can either auto-process the pre-crRNA into single crRNAs, or it requires other nucleases to do so. The spacer part of the mature crRNA guide is then used by the effector to find a matching protospacer (target) that is flanked by a Protospacer Adjacent Motif (PAM). This motif is important to distinguish the array on the chromosome (self) from the actual invading target (non-self). The CRISPR-Cas system used in this thesis is FnCas12a, which is a Class 2 type V system from *Francisella tularensis subsp. novicida* (2).

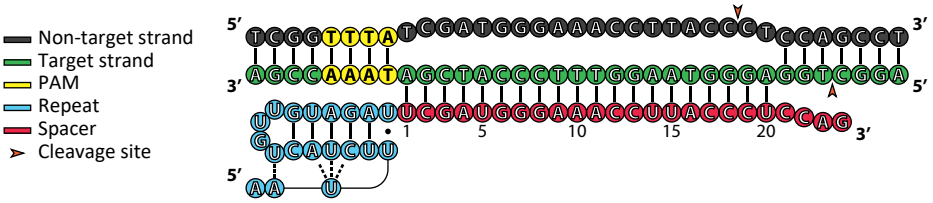


Figure 9.4. Cas12a bound with its crRNA to a target DNA. Cas12a binds to the pseudoknot (blue) of the crRNA. After locating a PAM (yellow) the R-loop is formed by base-pairing of the spacer-derived sequence (red) to the target strand (green). After base-pairing to 19-20 nucleotides, the non-target and target strands are cleaved.

Chapter 7 describes a method for *in vitro* measurement of DNase activity, which is not strictly limited to CRISPR-Cas proteins, but uses FnCas12a as test case. The method is based on the separation of a biotin from a fluorophore. A PCR fragment harbouring the target sequence is made with one biotinylated primer and one primer with a fluorophore. This fragment is incubated with the nuclease and the reaction is quenched with EDTA after a certain period of time. The biotin is then bound to magnetic streptavidin beads. When the fluorophore is still connected to the biotin, it will be removed from solution, but if the nuclease separated the two, it will remain there. The fluorescence is now a measurement for the nuclease activity. The test case showed that FnCas12a can fully cleave the target in less than a minute of reaction time.

In Chapter 8 the effect of RNA structure on the FnCas12a activity is investigated. When Cas12 binds to the pre-crRNA, it can process it into crRNA without aid of other proteins. The protein binds to a pseudoknot structure that is formed by the repeat located upstream of the spacer (Figure 9.4). This pseudoknot is essential for the recognition of (pre-)crRNA. The pseudoknot from the FnCas12a can be disrupted by RNA upstream and downstream from the repeat. If this happens, the repeat is not recognised by FnCas12a. Without crRNA, the FnCas12a cannot find targets nor exhibit its DNase activity. *In vitro* processing assays indeed demonstrated that the protein cannot bind the pre-crRNA in the case where there is also no *in vivo* activity. We proposed several solutions to this issue. Shortening the repeat takes away some possible alternative structures that compete with the pseudoknot. Alternatively, the spacers can also be forced into a certain structure, that at least does not affect the pseudoknot structure. The part of the spacer that base-pairs with the target is 20 nt long, but the spacer itself can be longer than that without impeding the Cas12a binding. A strategy base-pairing the 3' end of the spacer back onto the 5' end of the spacer indeed diminished the pseudoknot being disrupted. Care must be taken, however, since the extended spacer does need to bind to the target as well. When the intended structure of the spacer to rescue the pseudoknot is too strong, the spacer cannot bind the target efficiently. As everything in life, it is all about balance.

Conclusion

Where the storage of genetic information has largely been taken over by DNA, the RNA is crucial in the synthesis of proteins, both as a carrier for genetic information and the coupling of amino acids. As a mRNA, it can control the protein synthesis in response to internal and external factors via riboswitches. The codon use and overall layout of the mRNA can synthesise proteins according to the stoichiometry of pathways and protein complexes via differential translation. And last but not least, the RNA plays a crucial role in the defence against viruses and selfish mobile elements. In conclusion, RNA is the glue that binds the pieces of life together.

References

1. Höck,J. and Meister,G. (2008) The Argonaute protein family. Genome Biol., 9.
2. Zetsche,B., Gootenberg,J.S., Abudayyeh,O.O., Slaymaker,I.M., Makarova,K.S., Essletzbichler,P., Volz,S.E., Joung,J., Van Der Oost,J., Regev,A., et al. (2015) Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell, 163, 759–771.
3. Crick,F. (1970) Central dogma of molecular biology. Nature, 10.1038/227561a0.
4. Regulski,E., Moy,R.H. and Weinberg,Z. (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. Chemtracts.
5. Spinelli,S. V., Pontel,L.B., García Véscovi,E. and Soncini,F.C. (2008) Regulation of magnesium homeostasis in Salmonella: Mg2+ targets the mgtA transcript for degradation by RNase E. FEMS Microbiol. Lett., 10.1111/j.1574-6968.2008.01065.x.
6. Barrick,J.E. and Breaker,R.R. (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol., 10.1186/gb-2007-8-11-r239.
7. Morita,M., Kanemori,M., Yanagi,H. and Yura,T. (1999) Heat-induced synthesis of σ32 in *Escherichia coli*: Structural and functional dissection of rpoH mRNA secondary structure. J. Bacteriol., 10.1128/jb.181.2.401-410.1999.
8. Morita,M.T., Tanaka,Y., Kodama,T.S., Kyogoku,Y., Yanagi,H. and Yura,T. (1999) Translational induction of heat shock transcription factor σ32: Evidence for a built-in RNA thermosensor. Genes Dev., 10.1101/gad.13.6.655.
9. Chojnowski,G., Waleń,T. and Bujnicki,J.M. (2014) RNA Bricks - A database of RNA 3D motifs and their interactions. Nucleic Acids Res., 10.1093/nar/gkt1084.
10. Zimmermann,G.R., Wick,C.L., Shields,T.P., Jenison,R.D. and Pardi,A. (2000) Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. RNA, 10.1017/S1355838200000169.
11. Jenison,R.D., Gill,S.C., Pardi,A. and Polisky,B. (1994) High-resolution molecular discrimination by RNA. Science (80-.)., 263, 1425–1429.
12. Nahvi,A., Barrick,J.E. and Breaker,R.R. (2004) Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. Nucleic Acids Res., 10.1093/nar/gkh167.
13. Tang,W., Hu,J.H. and Liu,D.R. (2017) Aptazyme-embedded guide RNAs enable ligand-responsive genome editing and transcriptional activation. Nat. Commun., 10.1038/ncomms15939.

14. Ogawa,A. and Maeda,M. (2007) Aptazyme-based riboswitches as label-free and detector-free sensors for cofactors. Bioorganic Med. Chem. Lett., 10.1016/j.bmcl.2007.03.033.

15. Hedberg,A. and Johansen,S.D. (2013) Nuclear group i introns in self-splicing and beyond. Mob. DNA, 10.1186/1759-8753-4-17.

16. Bioinformatics,B., Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D 'souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L. V, et al. (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics.

17. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science (80-.)., 10.1126/science.2200121.

18. Stoltenburg,R., Reinemann,C. and Strehlitz,B. (2005) FluMag-SELEX as an advantageous method for DNA aptamer selection. Anal. Bioanal. Chem., 10.1007/s00216-005-3388-9.

19. Lauridsen,L.H., Doessing,H.B., Long,K.S. and Nielsen,A.T. (2018) A capture-SELEX strategy for multiplexed selection of RNA aptamers against small molecules. In Methods in Molecular Biology.

20. Mendonsa,S.D. and Bowser,M.T. (2004) In Vitro Evolution of Functional DNA Using Capillary Electrophoresis. J. Am. Chem. Soc., 10.1021/ja037832s.

21. Ghazi,Z., Fowler,C.C. and Li,Y. (2014) Artificial riboswitch selection: A FACS-based approach. Methods Mol. Biol., 1111, 57–75.

22. Topp,S. and Gallivan,J.P. (2007) Guiding bacteria with small molecules and RNA. J. Am. Chem. Soc., 10.1021/ja0692480.

23. Quax,T.E.F., Claassens,N.J., Söll,D. and van der Oost,J. (2015) Codon Bias as a Means to Fine-Tune Gene Expression. Mol. Cell, 10.1016/j.molcel.2015.05.035.

24. Quax,T.E.F., Wolf,Y.I., Koehorst,J.J., Wurtzel,O., vanderOost,R., Ran,W., Blombach,F., Makarova,K.S., Brouns,S.J.J., Forster,A.C., et al. (2013) Differential translation tunes uneven production of operon-encoded proteins. Cell Rep., 10.1016/j.celrep.2013.07.049.

25. Zhang,C. (2009) Novel functions for small RNA molecules. Curr. Opin. Mol. Ther.

26. Leung,R.K.M. and Whittaker,P.A. (2005) RNA interference: From gene silencing to gene-specific therapeutics. Pharmacol. Ther., 10.1016/j.pharmthera.2005.03.004.

27. Xie,K., Zhang,J. and Yang,Y. (2014) Genome-wide prediction of highly specific guide

RNA spacers for CRISPR-Cas9-mediated genome editing in model plants and major crops. Mol. Plant, 10.1093/mp/ssu009.

28. Briner,A.E., Henriksen,E.D. and Barrangou,R. (2016) Prediction and validation of native and engineered cas9 guide sequences. Cold Spring Harb. Protoc., 10.1101/pdb.prot086785.

29. Concordet,J.P. and Haeussler,M. (2018) CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res., 10.1093/nar/gky354.

30. Chari,R., Yeo,N.C., Chavez,A. and Church,G.M. (2017) SgRNA Scorer 2.0: A Species-Independent Model to Predict CRISPR/Cas9 Activity. ACS Synth. Biol., 10.1021/acssynbio.6b00343.

31. Mandal,M. and Breaker,R.R. (2004) Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol., 5, 451–463.

32. Gallivan,J.P. (2007) Toward reprogramming bacteria with small molecules and RNA. Curr. Opin. Chem. Biol., 11, 612–619.

33. Zhang,S., Stancek,M. and Isaksson,L.A. (1997) The efficiency of a cis-cleaving ribozyme in an mRNA coding region is influenced by the translating ribosome in vivo. Nucleic Acids Res., 25, 4301–4306.

34. Goler,J.A., Carothers,J.M. and Keasling,J.D. (2014) Dual-selection for evolution of in vivo functional aptazymes as riboswitch parts. Methods Mol. Biol., 1111, 221–235.

35. Berens,C., Groher,F. and Suess,B. (2015) RNA aptamers as genetic control devices: The potential of riboswitches as synthetic elements for regulating gene expression. Biotechnol. J., 10, 246–257.

36. Thompson,K.M., Syrett,H.A., Knudsen,S.M. and Ellington,A.D. (2002) Group I aptazymes as genetic regulatory switches. BMC Biotechnol., 2, 21.

37. Huang,H. (2007) Design and Characterization of Artificial Transcriptional Terminators. Thesis, 10.1177/0272989x10386800\n10.1177/0272989X10364845; Mortimer, D., Segal, L., Sturm, J., Can we derive an 'exchange rate' between descriptive and preference-based outcome measures for stroke? Results from the transfer to utility (TTU) technique (2009) Health Qual Life Outcomes, 7, p. 33.

38. Pichler,A. and Schroeder,R. (2002) Folding problems of the 5′ splice site containing the P1 stem of the group I thymidylate synthase intron. Substrate binding inhibition in vitro and mis-splicing in vivo. J. Biol. Chem., 277, 17987–17993.

39. Datsenko,K.A. and Wanner,B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A., 97, 6640–6645.

40. Barrozo,A., Borstnar,R., Marloie,G. and Kamerlin,S.C.L. (2012) Computational protein engineering: Bridging the gap between rational design and laboratory evolution. Int. J. Mol. Sci., 13, 12428–12460.

41. Jang,S. and Jung,G.Y. (2018) Systematic optimization of L-tryptophan riboswitches for efficient monitoring of the metabolite in *Escherichia coli*. Biotechnol. Bioeng., 115, 266–271.

42. Koizumi,M., Kerr,J.N., Soukup,G.A. and Breaker,R.R. (1999) Allosteric ribozymes sensitive to the second messengers cAMP and cGMP. Nucleic Acids Symp. Ser., 10.1093/nass/42.1.275.

43. De Las Heras,A., Carreño,C.A., Martínez-García,E. and De Lorenzo,V. (2010) Engineering input/output nodes in prokaryotic regulatory circuits. FEMS Microbiol. Rev., 34, 842–865.

44. Breaker,R.R. (2002) Engineered allosteric ribozymes as biosensor components. Curr. Opin. Biotechnol., 13, 31–39.

45. Müller,S., Strohbach,D. and Wolf,J. (2006) Sensors made of RNA: Tailored ribozymes for detection of small organic molecules, metals, nucleic acids and proteins. IEE Proc. Nanobiotechnology, 153, 31–40.

46. Piganeau,N. (2012) Selecting allosteric ribozymes. Methods Mol. Biol., 848, 317–328.

47. Been,M.D. and Cech,T.R. (1986) One binding site determines sequence specificity of Tetrahymena pre-rRNA self-splicing, trans-splicing, and RNA enzyme activity. Cell, 47, 207–216.

48. Doudna,J.A., Cormack,B.P. and Szostak,J.W. (1989) RNA structure, not sequence, determines the 5′ splice-site specificity of a group I intron. Proc. Natl. Acad. Sci. U. S. A., 86, 7402–7406.

49. Brion,P., Schroeder,R., Michel,F. and Westhof,E. (1999) Influence of specific mutations on the thermal stability of the *td* group I intron in vitro and on its splicing efficiency in vivo: A comparative study. Rna, 5, 947–958.

50. Guo,F. and Cech,T.R. (2002) In vivo selection of better self-splicing introns in *Escherichia coli*: The role of the P1 extension helix of the Tetrahymena intron. Rna, 8, 647–658.

51. Cechu,T.R., Damberger,S.H. and Guteli,R.R. (1994) Representation of the secondary and tertiary structure of group I introns. Nat. Struct. Biol., 10.1038/nsb0594-273.

52. Dambach,M., Sandoval,M., Updegrove,T.B., Anantharaman,V., Aravind,L., Waters,L.S. and Storz,G. (2015) The Ubiquitous yybP-ykoY Riboswitch Is a Manganese-Responsive Regulatory Element. Mol. Cell, 57, 1099–1109.

53. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by a metabolite binding mRNA. Chem. Biol., 9, 1043–1049.

54. Sudarsan,N., Wickiser,J.K., Nakamura,S., Ebert,M.S. and Breaker,R.R. (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. Genes Dev., 17, 2688–2697.

55. Groher,F. and Suess,B. (2014) Synthetic riboswitches - A tool comes of age. Biochim. Biophys. Acta - Gene Regul. Mech., 1839, 964–973.

56. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J., et al. (2015) Rfam 12.0: Updates to the RNA families database. Nucleic Acids Res., 43, D130–D137.

57. Koizumi,M., Soukup,G.A., Kerr,J.N.Q. and Breaker,R.R. (1999) Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. Nat. Struct. Biol., 6, 1062–1071.

58. Piganeau,N., Jenne,A., Thuillier,V. and Famulok,M. (2000) An allosteric ribozyme regulated by doxycyline. Angew. Chemie - Int. Ed., 39, 4369–4373.

59. Soukup,G.A. and Breaker,R.R. (1999) Nucleic acid molecular switches. Trends Biotechnol., 17, 469–476.

60. Suess,B., Fink,B., Berens,C., Stentz,R. and Hillen,W. (2004) A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. Nucleic Acids Res., 32, 1610–1614.

61. Soukup,G.A., Emilsson,G.A.M. and Breaker,R.R. (2000) Altering molecular recognition of RNA aptamers by allosteric selection. J. Mol. Biol., 298, 623–632.

62. http://www.openwetware.org/wiki/'Round-the-horn_site-directed_mutagenesis (2017) 'Round-the-horn site directed mutagenisis.

63. Zimmermann,G.R., Jenison,R.D., Wick,C.L., Simorre,J.P. and Pardi,A. (1997) Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. Nat. Struct. Biol., 4, 644–649.

64. Nomura,Y. and Yokobayashi,Y. (2007) Reengineering a natural riboswitch by dual genetic selection. J. Am. Chem. Soc., 129, 13814–13815.

65. Stavropoulos,T.A. and Strathdee,C.A. (2000) Expression of the tetA(C) tetracycline efflux pump in *Escherichia coli* confers osmotic sensitivity. FEMS Microbiol. Lett., 190, 147–150.

66. Eckert,B. and Beck,C.F. (1989) Overproduction of transposon Tn10-encoded tetracycline resistance protein results in cell death and loss of membrane potential. J.

Bacteriol., 171, 3557–3559.

67. Wagner,S., Bader,M.L., Drew,D. and de Gier,J.W. (2006) Rationalizing membrane protein overexpression. Trends Biotechnol., 24, 364–371.

68. Uliczka,F., Pisano,F., Kochut,A., Opitz,W., Herbst,K., Stolz,T. and Dersch,P. (2011) Monitoring of gene expression in bacteria during infections using an adaptable set of bioluminescent, fluorescent and colorigenic fusion vectors. PLoS One, 6.

69. Chang,A.C.Y. and Cohen,S.N. (1978) Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J. Bacteriol., 134, 1141–1156.

70. van Rossum,T., Muras,A., Baur,M.J.J., Creutzburg,S.C.A., van der Oost,J. and Kengen,S.W.M. (2017) A growth- and bioluminescence-based bioreporter for the in vivo detection of novel biocatalysts. Microb. Biotechnol., 10, 625–641.

71. Famulok,M. (1994) Molecular Recognition of Amino Acids by RNA-Aptamers: An L-Citrulline Binding RNA Motif and Its Evolution into an L-Arginine Binder. J. Am. Chem. Soc., 116, 1698–1706.

72. Galperin,M.Y. and Koonin,E. V. (2000) Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol., 10.1038/76443.

73. Huynen,M., Snel,B., Lathe,W. and Bork,P. (2000) Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Res., 10.1101/gr.10.8.1204.

74. Oppenheim,D.S. and Yanofsky,C. (1980) Translational coupling during expression of the tryptophan operon of *Escherichia coli*. Genetics.

75. Schümperli,D., McKenney,K., Sobieski,D.A. and Rosenberg,M. (1982) Translational coupling at an intercistronic boundary of the *Escherichia coli* galactose operon. Cell, 30, 865–871.

76. Aksoy,S., Squires,C.L. and Squires,C. (1984) Translational coupling of the trpB and trpA genes in the *Escherichia coli* Tryptophan operon. J. Bacteriol., 157, 363–367.

77. Levin-Karp,A., Barenholz,U., Bareia,T., Dayagi,M., Zelcbuch,L., Antonovsky,N., Noor,E. and Milo,R. (2013) Quantifying translational coupling in *E. coli* synthetic operons using RBS modulation and fluorescent reporters. ACS Synth. Biol., 2, 327–336.

78. Govantes,F., Andújar,E. and Santero,E. (1998) Mechanism of translational coupling in the nifLA operon of Klebsiella pneumoniae. EMBO J., 17, 2368–2377.

79. Rex,G., Surin,B., Besse,G., Schneppe,B. and McCarthy,J.E.G. (1994) The mechanism of translational coupling in *Escherichia coli*. Higher order structure in the atpHA mRNA

acts as a conformational switch regulating the access of de novo initiating ribosomes. J. Biol. Chem., 269, 18118–18127.

80. Claassens,N.J., Siliakus,M.F., Spaans,S.K., Creutzburg,S.C.A., Nijsse,B., Schaap,P.J., Quax,T.E.F. and Van Der Oost,J. (2017) Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS One, 12.

81. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31, 3406–3415.

82. Deuschle,U., Kammerer,W., Gentz,R. and Bujard,H. (1986) Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. EMBO J., 5, 2987–2994.

83. Zhou,K., Zhou,L., Lim,Q., Zou,R., Stephanopoulos,G. and Too,H.P. (2011) Novel reference genes for quantifying transcriptional responses of *Escherichia coli* to protein overexpression by quantitative PCR. BMC Mol. Biol., 10.1186/1471-2199-12-18.

84. Li,R., Zhang,Q., Li,J. and Shi,H. (2016) Effects of cooperation between translating ribosome and RNA polymerase on termination efficiency of the Rho-independent terminator. Nucleic Acids Res., 44, 2554–2563.

85. Mossey,P. and Das,A. (2013) Expression of Agrobacterium tumefaciens octopine Ti-plasmid virB8 gene is regulated by translational coupling. Plasmid, 69, 72–80.

86. Takyar,S., Hickerson,R.P. and Noller,H.F. (2005) mRNA helicase activity of the ribosome. Cell, 120, 49–58.

87. Mutalik,V.K., Guimaraes,J.C., Cambray,G., Lam,C., Christoffersen,M.J., Mai,Q.A., Tran,A.B., Paull,M., Keasling,J.D., Arkin,A.P., et al. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. Nat. Methods, 10.1038/nmeth.2404.

88. Vytvytska,O., Moll,I., Kaberdin,V.R., Von Gabain,A. and Bläsi,U. (2000) Hfq (HF1) stimulates ompA mRNA decay by interfering with ribosome binding. Genes Dev., 14, 1109–1118.

89. Braun,F., Le Derout,J. and Régnier,P. (1998) Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of *Escherichia coli*. EMBO J., 17, 4790–4797.

90. Edri,S. and Tuller,T. (2014) Quantifying the effect of ribosomal density on mRNA stability. PLoS One, 9.

91. Deneke,C., Lipowsky,R. and Valleriani,A. (2013) Effect of ribosome shielding on mRNA stability. Phys. Biol., 10.

92. Oh,E., Becker,A.H., Sandikci,A., Huber,D., Chaba,R., Gloge,F., Nichols,R.J., Typas,A.,

Gross,C.A., Kramer,G., et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell, 147, 1295–1308.

93. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science (80-.)., 324, 218–223.

94. Newbury,S.F., Smith,N.H., Robinson,E.C., Hiles,I.D. and Higgins,C.F. (1987) Stabilization of translationally active mRNA by prokaryotic REP sequences. Cell, 48, 297–310.

95. Vasquez,J.R., Evnin,L.B., Higaki,J.N. and Craik,C.S. (1989) An expression system for trypsin. J. Cell. Biochem., 39, 265–276.

96. West,S. and Proudfoot,N.J. (2009) Transcriptional Termination Enhances Protein Expression in Human Cells. Mol. Cell, 33, 354–364.

97. Kiel,M.C., Kaji,H. and Kaji,A. (2007) Ribosome recycling: An essential process of protein synthesis. Biochem. Mol. Biol. Educ., 10.1002/bmb.6.

98. Strohkendl,I., Saifuddin,F.A., Rybarski,J.R., Finkelstein,I.J. and Russell,R. (2018) Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a. Mol. Cell, 71, 816-824.e3.

99. Ender,C. and Meister,G. (2010) Argonaute proteins at a glance. J. Cell Sci., 123, 1819–1823.

100. Uusi-Mäkelä,M.I.E., Barker,H.R., Bäuerlein,C.A., Häkkinen,T., Nykter,M. and Rämet,M. (2018) Chromatin accessibility is associated with CRISPR-Cas9 efficiency in the zebrafish (Danio rerio). PLoS One, 13.

101. Chari,R., Mali,P., Moosburner,M. and Church,G.M. (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat. Methods, 12, 823–826.

102. Shmakov,S., Smargon,A., Scott,D., Cox,D., Pyzocha,N., Yan,W., Abudayyeh,O.O., Gootenberg,J.S., Makarova,K.S., Wolf,Y.I., et al. (2017) Diversity and evolution of class 2 CRISPR-Cas systems. Nat. Rev. Microbiol., 15, 169–182.

103. Mohanraju,P., Oost,J., Jinek,M. and Swarts,D. (2018) Heterologous Expression and Purification of the CRISPR-Cas12a/Cpf1 Protein. Bio-Protocol, 8.

104. Doudna,J.A. and Charpentier,E. (2014) The new frontier of genome engineering with CRISPR-Cas9. Science (80-.)., 346.

105. Kim,H. and Kim,J.S. (2014) A guide to genome engineering with programmable nucleases. Nat. Rev. Genet., 15, 321–334.

106. Sander,J.D. and Joung,J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol., 32, 347–350.

107. Cox,D.B.T., Platt,R.J. and Zhang,F. (2015) Therapeutic genome editing: Prospects and challenges. Nat. Med., 21, 121–131.

108. Mohanraju,P., Makarova,K.S., Zetsche,B., Zhang,F., Koonin,E. V. and Van Der Oost,J. (2016) Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science (80-.)., 353, 556–568.

109. Hart,T., Chandrashekhar,M., Aregger,M., Steinhart,Z., Brown,K.R., MacLeod,G., Mis,M., Zimmermann,M., Fradet-Turcotte,A., Sun,S., et al. (2015) High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell, 163, 1515–1526.

110. Shalem,O., Sanjana,N.E. and Zhang,F. (2015) High-throughput functional genomics using CRISPR-Cas9. Nat. Rev. Genet., 16, 299–311.

111. Wang,W., Ye,C., Liu,J., Zhang,D., Kimata,J.T. and Zhou,P. (2014) CCR5 gene disruption via lentiviral vectors expressing Cas9 and single guided RNA renders cells resistant to HIV-1 infection. PLoS One, 9.

112. Zhou,H., Liu,B., Weeks,D.P., Spalding,M.H. and Yang,B. (2014) Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. Nucleic Acids Res., 42, 10903–10914.

113. Wu,W.Y., Lebbink,J.H.G., Kanaar,R., Geijsen,N. and Van Der Oost,J. (2018) Genome editing by natural and engineered CRISPR-associated nucleases. Nat. Chem. Biol., 14, 642–651.

114. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., Pirzada,Z.A., Eckert,M.R., Vogel,J. and Charpentier,E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 471, 602–607.

115. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-.)., 337, 816–821.

116. Fonfara,I., Richter,H., BratoviÄ‚M., Le Rhun,A. and Charpentier,E. (2016) The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature, 532, 517–521.

117. Zetsche,B., Heidenreich,M., Mohanraju,P., Fedorova,I., Kneppers,J., Degennaro,E.M., Winblad,N., Choudhury,S.R., Abudayyeh,O.O., Gootenberg,J.S., et al. (2017) Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nat. Biotechnol., 35, 31–34.

118. Swarts,D.C., van der Oost,J. and Jinek,M. (2017) Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Mol. Cell, 66, 221-233.e4.

119.Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science (80-.)., 343, 80–84.

120.Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat. Biotechnol., 32, 1262–1267.

121.Ren,X., Yang,Z., Xu,J., Sun,J., Mao,D., Hu,Y., Yang,S.J., Qiao,H.H., Wang,X., Hu,Q., et al. (2014) Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. Cell Rep., 9, 1151–1162.

122.Malina,A., Katigbak,A., Cencic,R., Maïga,R.I., Robert,F., Miura,H. and Pelletier,J. (2014) Adapting CRISPR/Cas9 for functional genomics screens. Methods Enzymol., 546, 193–213.

123.Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat. Methods, 12, 982–988.

124.Xu,H., Xiao,T., Chen,C.H., Li,W., Meyer,C.A., Wu,Q., Wu,D., Cong,L., Zhang,F., Liu,J.S., et al. (2015) Sequence determinants of improved CRISPR sgRNA design. Genome Res., 25, 1147–1157.

125.Wong,N., Liu,W. and Wang,X. (2015) WU-CRISPR: Characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol., 16.

126.Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R., et al. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol., 34, 184–191.

127.Chu,V.T., Graf,R., Wirtz,T., Weber,T., Favret,J., Li,X., Petsch,K., Tran,N.T., Sieweke,M.H., Berek,C., et al. (2016) Efficient CRISPR-mediated mutagenesis in primary immune cells using CrispRGold and a C57BL/6 Cas9 transgenic mouse line. Proc. Natl. Acad. Sci. U. S. A., 113, 12514–12519.

128.Thyme,S.B., Akhmetova,L., Montague,T.G., Valen,E. and Schier,A.F. (2016) Internal guide RNA interactions interfere with Cas9-mediated cleavage. Nat. Commun., 7.

129.Kim,H.K., Song,M., Lee,J., Menon,A.V., Jung,S., Kang,Y.M., Choi,J.W., Woo,E., Koh,H.C., Nam,J.W., et al. (2017) In vivo high-throughput profiling of CRISPR-Cpf1 activity. Nat. Methods, 14, 153–159.

130.Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. Algorithms Mol. Biol., 6.

131.Kim,H.K., Min,S., Song,M., Jung,S., Choi,J.W., Kim,Y., Lee,S., Yoon,S. and Kim,H. (2018) Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat.

Biotechnol., 36, 239–241.

132.Hu,P., Zhao,X., Zhang,Q., Li,W. and Zu,Y. (2018) Comparison of various nuclear localization signal-fused Cas9 proteins and Cas9 mRNA for genome editing in Zebrafish. G3 Genes, Genomes, Genet., 8, 823–831.

133.Aird,E.J., Lovendahl,K.N., St. Martin,A., Harris,R.S. and Gordon,W.R. (2018) Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. Commun. Biol., 1.

134.Flynn,J.M., Levchenko,I., Seidel,M., Wickner,S.H., Sauer,R.T. and Baker,T.A. (2001) Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. Proc. Natl. Acad. Sci. U. S. A., 98, 10584–10589.

135.McGinness,K.E., Baker,T.A. and Sauer,R.T. (2006) Engineering Controllable Protein Degradation. Mol. Cell, 22, 701–707.

136.Cech,T.R. (1987) The chemistry of self-splicing RNA and RNA enzymes. Science (80-.)., 10.1126/science.2438771.

137.Baer,M.F., Reilly,R.M., McCorkle,G.M., Hai,T.Y., Altman,S. and RajBhandary,U.L. (1988) The recognition by RNaseP of precursor tRNAs. J. Biol. Chem.

138.Cech,T.R. (2000) The ribosome is a ribozyme. Science (80-.)., 10.1126/science.289.5481.878.

139.Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. Nature, 10.1038/346818a0.

140.Robertson,D.L. and Joyce,G.F. (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. Nature, 10.1038/344467a0.

141.Fire,A. and Mello,C. (1998) RNAi : the review. Control.

142.Hegge,J.W., Swarts,D.C. and Van Der Oost,J. (2018) Prokaryotic argonaute proteins: Novel genome-editing tools? Nat. Rev. Microbiol., 10.1038/nrmicro.2017.73.

143.Vogel,J. and Luisi,B.F. (2011) Hfq and its constellation of RNA. Nat. Rev. Microbiol., 10.1038/nrmicro2615.

144.Cech,T.R. (2015) RNA World research - Still evolving. RNA, 10.1261/rna.049965.115.

145.Cech,T.R. (2009) Crawling Out of the RNA World. Cell, 10.1016/j.cell.2009.02.002.

146.Cech,T.R. (2012) The RNA worlds in context. Cold Spring Harb. Perspect. Biol., 10.1101/cshperspect.a006742.

147.Mironov,A.S., Gusarov,I., Rafikov,R., Lopez,L.E., Shatalin,K., Kreneva,R.A., Perumov,D.A. and Nudler,E. (2002) Sensing small molecules by nascent RNA: A mechanism to

control transcription in bacteria. Cell, 10.1016/S0092-8674(02)01134-0.

148. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by a metabolite binding mRNA. Chem. Biol., 10.1016/S1074-5521(02)00224-7.

149. Winkler,W., Nahvi,A. and Breaker,R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature, 10.1038/nature01145.

150. Topp,S. and Gallivan,J.P. (2008) Riboswitches in unexpected places - A synthetic riboswitch in a protein coding region. Rna, 14, 2498–2503.

151. Summers,R.M., Louie,T.M., Yu,C.L., Gakhar,L., Louie,K.C. and Subramanian,M. (2012) Novel, highly specific N-demethylases enable bacteria to live on caffeine and related purine alkaloids. J. Bacteriol., 10.1128/JB.06637-11.

152. Summers,R.M., Seffernick,J.L., Quandt,E.M., Yu,C.L., Barrick,J.E. and Subramanian,M. V. (2013) Caffeine junkie: An unprecedented glutathione S-transferase-dependent oxygenase required for caffeine degradation by pseudomonas putida CBB5. J. Bacteriol., 10.1128/JB.00585-13.

153. Rogers,J. and Joyce,G.F. (2001) The effect of cytidine on the structure and function of an RNA ligase ribozyme. RNA, 10.1017/S135583820100228X.

154. Lam,B.J. and Joyce,G.F. (2011) An isothermal system that couples ligand-dependent catalysis to ligand-independent exponential amplification. J. Am. Chem. Soc., 10.1021/ja111136d.

155. Paul,N. and Joyce,G.F. (2002) A self-replicating ligase ribozyme. Proc. Natl. Acad. Sci. U. S. A., 10.1073/pnas.202471099.

156. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res., 10.1093/nar/15.3.1281.

157. Makarova,K.S., Wolf,Y.I. and Koonin,E. V. (2018) Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? Cris. J., 10.1089/crispr.2018.0033.

Acknowledgements

And so ends this chapter in my life. It has been a rollercoaster full of excitement, brain racking, disappointment and eureka moments. A journey I luckily did not travel alone, but together with many inspiring people.

Allereerst natuurlijk **John**. Dank voor je onophoudelijke steun en vertrouwen. Ik denk dat wij elkaar prima vonden in alle wilde plannen waar we mee op de proppen kwamen. En die werden ook nog uitgevoerd onder het motto: "ik weet niet of het gaat werken, maar er is maar één manier om daar achter te komen".

Servé, dank voor je interessante discussies en scherpe inzichten. Ik kon altijd rekenen op een nieuwe invalshoek om de uitdagingen te lijf te gaan (en dat was ook wel nodig met dit project).

Dan komen we bij mijn paranimfen **Thijs** en **Thomas**, die niet alleen mij op het laatste moment hebben bijgestaan, maar ook zelf bloed, zweet en tranen in dit boekje hebben zitten. Jullie waren twee geweldige studenten en hebben die PhD positie bij Microbiologie echt verdiend.

Many thanks also to my office mates, especially **Teunke**. I really enjoyed our collaboration on the Hotzyme project, tackeling the project from two sides. We had very fun discussions on a vast variety of topics. **Melvin**, you could really give me a run for my money with those difficult questions of yours. **John R.**, I probably would not have ended up here if not for you. Thank you for being my supervisor during my MSc thesis and it was fun being your office mate. **Nico**, thank you for the time we had and the collaboration on the codon project. **Yifan**, it was an honour being your paranymph and thank you for all the great times. Especially when that included some rare expensive whisky. **Jorrit**, thanks for the great stories on how not to fabricate an NgAgo paper. **Marnix**, thanks for the chats we had in the office. **Despoina**, you really have the cutest way of saying my name and that game of chess against you was one of the most nerve-wrecking moments of my life. **Prarthana**, **Wen**, I had a great time working with you on the Cas12a and Mmu projects. And we always had plenty of opportunity to crack a joke at each other. **Costas**, **Adini**, it was great working with you on SIBR-Cas. I hope we can pull this project through. And **Belén**,thanks for your involvement whenever we needed to move towards eukaryotes.

Thanks to all of members of the Bacterial Genetics group throughout the years: Mark, Vincent, Bram, Amos, Tom vd W, Elleke, Tessa, Slavtscho, Tijn, Hanne, Edze, Matthijs, Daan, Tim, Patrick, Raymond, Jochem, Rebecca, Franklin, Bas, Joyshree, Mihris, Jeroen, Joep, Max, Lorenzo, Mamou, Enrico, Janneke, Ismael, Rob, Carina, Guus, James, Lione, Eric, Jurre, Yannis, Aleks, Eugenios, Eric, Maartje and everyone I am forgetting.

Thanks to all my students whom I have supervised and were not mentioned earlier: Haftu, Evans, Lucia, Vanessa, Jorik, Sophie, Judith, Timon, Efthymios, Daan and Ricardo.

Willem, Thijs, I thank you for the great environment that the Micriobiology group is. And also thanks to all the staff. Anja, Hannie, Wim, Carolien, Sjon, Philippe, Merlijn, Ton, Tom S, Monika and Steven.

Dan wil ik mijn vrienden van Canzone bedanken. **Elysa**, **Janna**, **Jos**, **Joost** en **Wouter**, en de invalsopraan **Sanne**. Zoals Janna en ik allebei gesteld hebben (sorry Janna, ik heb je stelling gejat), zingen is goed voor de mentale gezondheid en ik heb genoten van de repetities met jullie en de vele concerten die we hebben gedaan. Ik zie ernaar uit om weer eens bijeen te komen na deze vreselijke pandemie. Natuurlijk mogen ook mijn vrienden van Vivavoce en de WSKOV niet vergeten worden. In het bijzonder mijn Viva bestuursgenoten Pieter, David, Antke, Marijke, Sonja, Esmer en Gemma.

Als laatste wil ik graag mijn lieve familie bedanken voor alle interesse en support door de vele jaren van dit PhD project heen. **Diederik**, **Suzanna**, bedankt voor de geweldige vakanties en natuurlijk het heerlijke eten. **Paps**, **mams**, dank jullie wel dat jullie altijd voor me klaar staan met raad en daad. Ik had dit nooit zonder jullie kunnen doen.

About the cover

The pattern on the cover is a so-called Penrose tiling. It uses kites and darts derived from a pentagon. Following the rules on which sides may and may not touch, you get a pattern that is largely 5-fold symmetrical. While true 5-fold symmetry cannot be achieved for an infinite tiling, this strategy results in a good approximation. The amount of kites and darts follows the Fibonacci sequence, ultimately resulting in the kite to dart ratio becoming $0.5 + 0.5 \times 5^{0.5} = \varphi$ (i.e. the golden ratio).

About the author

Sjoerd Constantijn Arnoud Creutzburg was born on 10 July 1987 in Nieuwegein, The Netherlands. After graduating from the Christelijk Gymnasium Utrecht in 2005, he started his BSc in Biotechnology at Wageningen University. His batchelor's thesis was performed at the department of Physical Chemistry and Colloid Science (now Physical Chemistry and Soft Matter) under the supervision of Dr Saskia Lindhoud. It encompassed research on the formation of particles consisting of block co-polymers with different charges. After obtaining his BSc degree, he continued his studies in Biotechnology at Wageningen University, with a dual major in Cellular and Molecular Biotechnology, and Medical Biotechnology. His first major thesis was performed at the Laboratory of Microbiology under the supervision of Dr John Raedts, optimising the production of functional murine D-Glucuronyl C5-epimerase. The second major thesis was done at the Laboratory of Virology under the supervision of Dr Afshin Mehraban, studying the production and formation of virus-like particles. These virus-like particles consisted of the capsid protein of cowpea chlorotic mottle virus and had potential as carriers for epitopes in vaccine development. To complete his MSc, he did an internship at Synthon in Nijmegen, studying translation efficiency of monoclonal antibodies in animal cells. After obtaining his MSc degree, he started a PhD at the Laboratory of Microbiology at Wageningen University under the supervision of Prof. Dr John van der Oost and Dr Servé W.M. Kengen. Here, he worked on several projects including the development of a riboswitch selection system, optimisation of translation in prokaryotes, and targeting efficiency of CRISPR-Cas effector proteins.

List of publications

Creutzburg, S.C.A.*, Nieuwkoop, T.*, Zegers, T. and van der Oost, J. (2020) Translational feed-forward and feed-back control. *Manuscript in preparation*

Creutzburg, S.C.A., van Rossum, T, Kengen, S.W.M. and van der Oost, J. (2020) In vivo selection of riboswitches with an altered specificity. *Submitted in Nucleic Acids Research*

Creutzburg, S.C.A., Wu, W.Y., Mohanraju, P., Swartjes, T., Alkan, F., Gorodkin, J., Staals, R.H.J. and van der Oost, J. (2020) Good guide, bad guide: spacer sequence-dependent cleavage efficiency of Cas12a. Nucleic Acids Res., 10.1093/nar/gkz1240.

Creutzburg, S.C.A., Swartjes, T. and van der Oost, J. (2020) Medium-throughput in vitro detection of DNA cleavage by CRISPR-Cas12a. Methods, 10.1016/j.ymeth.2019.11.005.

Claassens, N.J., Siliakus, M.F., Spaans, S.K., **Creutzburg, S.C.A.**, Nijsse, B., Schaap, P.J., Quax, T.E.F. and Van Der Oost, J. (2017) Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS One, 12.

van Rossum, T., Muras, A., Baur, M.J.J., **Creutzburg, S.C.A.**, van der Oost, J. and Kengen, S.W.M. (2017) A growth- and bioluminescence-based bioreporter for the in vivo detection of novel biocatalysts. Microb. Biotechnol., 10, 625–641.

Hassani-Mehraban, A., **Creutzburg, S.C.A.**, Heereveld, L. and Kormelink, R. (2015) Feasibility of Cowpea chlorotic mottle virus-like particles as scaffold for epitope presentations. BMC Biotechnol., 15, 80.

*equal contribution

Patent applications

C. Patinios, **S.C.A. Creutzburg**, J. van der Oost and R.H.J Staals. Universal CRISPR Tool (Filed in October, 2020)

T. Nieuwkoop, **S.C.A. Creutzburg** and J. van der Oost. 3' UTR (Filed in October, 2020)

J. van der Oost, P. Mohanraju, W.Y. Wu, **S.C.A. Creutzburg**. CRISPR type V-U1 (Filed in July 2019)

S.C.A. Creutzburg and J. van der Oost (2016) An intronic self-splicing riboswitch for use in regulating synthesis of a reporter protein in the screening and selection of enzyme variants. (Abandoned)

S.C.A. Creutzburg and J. van der Oost (2016) Intronic, self-splicing riboswitch for inducible gene expression. (Abandoned)

Co-author affiliations

Laboratory of Microbiology, Department of Agrotechnology and Food Sciences.

Wageningen University, 6703 HB Wageningen, The Netherlands

Evans Asamoah Gyimah, Servé W.M. Kengen, Prarthana Mohanraju, Thijs Nieuwkoop, John van der Oost, Raymond H.J. Staals, Thomas Swartjes, Wen Y. Wu, Thijmen Zegers

Center for non-coding RNA in Technology and Health,

Department of Veterinary and Animal Sciences, University of Copenhagen, Frederiksberg C, Denmark

Ferhat Alkan, Jan Gorodkin

Division of Oncogenomics

Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands

Ferhat Alkan

Completed training activities

Discipline specific activities

Meetings and conferences

- Hotzyme meeting, Wageningen, The Netherlands (2012) *
- Annual Molecular Genetics meeting, Lunteren, The Netherlands (2013) **
- Hotzyme meeting, Athens, Greece (2013) *
- VLAG Symposium *
- Annual Molecular Genetics meeting, Lunteren, The Netherlands (2014)
- Regulatory RNAs in microbes, Stockholm, Sweden (2014)
- Hotzyme meeting, Exeter, England (2014) *
- Annual Molecular Genetics meeting, Lunteren, The Netherlands (2015) *
- Hotzyme meeting, Copenhagen, Denmark (2015) *
- Zing Conference: Regulating with RNA in Bacteria and Archaea, Cancun, Mexico (2015) **
- Annual Molecular Genetics meeting, Wageningen, The Netherlands (2016)
- Scientific Spring Meeting KNVM & NVMM, Papendal, The Netherlands (2016) *
- Host microbe genetics meeting, Wageningen, The Netherlands (2017)
- Laboratory of Microbiology centennial, Wageningen, The Netherlands (2017)

*oral presentation **poster presentation

General courses

- Competence assessment (2013)
- Time course management (2013)
- Scientific writing (2015)
- Scientific publishing (2017)
- Working on your PhD research in times of crisis (2020)
- Career Perspectives (2020)

Optionals

- Preparation research proposal
- Bacterial Genetics group meetings, Wageningen, The Netherlands
- Microbiology PhD/Post-Doc meetings, Wageningen, The Netherlands
- PhD study trip to the USA (2015)

The research described in this thesis was financially supported by the European Union via the project "Hotzyme" (GA: 265933), and the Netherlands Organization for Scientific Research (NWO) via a TOP grant (714.015.001) and a TTW grant (15804).

Printed by proefschriftmaken.nl || https://www.proefschriftmaken.nl