Dipl.-Ing. Corinna Odar

In vivo synthesis of non-canonical amino acids for the production of synthetic protein variants in E. coli

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften

eingereicht an der

Technischen Universität Graz

Betreuer

Univ.-Prof. Dipl.-Ing. Dr.techn. Helmut Schwab
Institut für Molekulare Biotechnologie

und

Dipl.-Ing. Dr.techn. Birgit Wiltschi
Austrian Centre of Industrial Biotechnology

Graz, Jänner und 2016

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

Datum

Unterschrift

Abstract

Labelling of proteins with non-canonical amino acids (ncAAs) became a valuable tool in a diverse field of the biosciences. Our main interest in them was their use to engineer proteins with novel traits. Although established protocols are available to produce ncAA variants, the expensiveness of ncAAs themselves often constrains more intense applications of this promising protein engineering technique. Therefore, this thesis focused the *in vivo* synthesis of ncAAs facilitating cheaper production of synthetic protein variants.

Fluoro amino acids in particular are highly valuable compounds, what is mainly due to the extraordinary chemical features delivered by the carbon-fluorine bond. A literature review elaborates on the use of fluoro amino acids as building blocks for fluorinated protein variants in Chapter 1. An insight into different techniques to incorporate fluoro amino acids into proteins is also provided. It highlights prospects and current relevance of fluoro amino acids as a protein engineering tool. Moreover, we detail the biosynthesis of the only fluoro amino acid discovered so far in nature, 4-fluoro-L-threonine (4FT). Although some microbial species realized the enzymatic synthesis of 4FT, it cannot be commercially purchased. The first enzyme of the 4FT pathway, the fluorinase, has attracted special attention because it is the only enzyme class known to date capable of forming the carbon-fluorine bond. The fluorinase catalyzes the synthesis of 5'-deoxy-5'-fluoroadenosine (FDA) from S-(5′-adenosyl)-L-methionine (SAM) and fluoride with L-methionine as the byproduct.

In a long term project, we set out to engineer *E. coli* to biosynthesize 4-fluorothreonine for its subsequent incorporation into proteins to produce fluorinated variants. In chapter 2 we describe the first pathway step in *E. coli* to synthesize FDA by the action of the fluorinase. Fluorinase expression studies served to optimize its expression pattern in *E. coli*. In an assay with lyophilized cells the fluorinase overexpressing strain successfully produced FDA from fluoride. We co-expressed a rat SAM synthase known to increase intracellular SAM levels. Moreover, the SAM synthases recycle the L-methionine to SAM. We showed that the increased SAM levels upon SAM synthase co-expression resulted in elevated FDA conversion. In addition to the experimental section, chapter 2 elaborates on general bottlenecks to realize the long term goal of 4-fluorothreonine synthesis for the expression of fluoro protein variants in *E. coli* and provides potential solutions.

In chapter 3, we established a simple and straight forward procedure for the *in vivo* enzymatic synthesis of non-canonical tryptophans (ncTrp) from indol precursors and their residue-specific incorporation into target proteins in *E. coli*. To enhance the biotransformation of a broad spectrum of indole analogs to the corresponding ncTrp we co-expressed the tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) on top of the tryptophan synthase of the host. The described procedure facilitates the biosynthesis of ncTrp from inexpensive indole precursors and their residue-specific incorporation. The target protein determines if *St*TRPS co-expression is essential or not. Moreover, we

showed that the widely used pET expression system is compatible with only a subset of fluorinated tryptophan analogs.

Chapter 4 explains and illustrates a newly designed Modular Cloning (MoClo) approach for the assembly of *E. coli* expression constructs including complex biological pathways as well as promoter-, RBS- or tag-libraries. It relies on a plasmid platform of basic biological parts that are assembled to multigene constructs in a succession of iterative assembly steps. We provide a systematic nomenclature that unambiguously designates the identity of a part. Moreover, the nomenclature system includes all DNA sequences that are crucial for the standardization of the MoClo assembly approach.

Realization of the presented MoClo strategy will allow the straight-forward as well as time- and cost-saving assembly of complex DNA constructs for a diverse field of applications. The presented genetic tool will tremendously foster engineering of ncAA pathways by allowing high-throughput assembly of DNA constructs encoding multi-enzyme cascades.

Published German Abstract at the Graz University of Technology[a]

Publizierte Kurzfassung auf der Technischen Universität Graz

Der Einbau nicht-kanonischen Aminosäuren (ncAS) in Proteine stellt ein wachsendes Forschungsfeld dar, sowohl in der Synthetischen Biologie als auch im Protein Engineering. Es erlaubt Proteine mit chemischen Eigenschaften auszustatten, die nicht durch das vom genetischen Code vorgegebenen kanonische Repertoire an Aminosäuren zur Verfügung stehen. Obgleich der ncAS-Einbau von großem wissenschaftlichem und industriellem Interesse ist, wird eine extensivere Forschung durch deren Kostspieligkeit sowie kommerzielle Verfügbarkeit behindert. Aus diesem Grund stellt die *in vivo* enzymatische Synthese von ncAS in *E. coli* zum Zweck deren Einbaus in heterolog exprimierte Proteine ein primeres Forschungsziel dieser Dissertation dar.

Wir verwirklichten die Synthese von ncAS ausgehend von einem synthetischen Präkusormolekül, das dem Kulturmedium supplementiert wurde, um Proteinvarianten herzustellen. Darüber hinaus etablierten wir den ersten enzymatischen Schritt einer Multienzymkaskade, die in Zukunft die Synthese einer fluorierten Aminosäure in *E. coli* ermöglichen soll.

Diese Dissertation befasst sich zudem mit einer modularen Assemblierungsstrategy zur Herstellung von Multigenkonstrukten. Die Tatsache, dass jede ncAS ihren eigenen natürlichen oder „engineerten" Syntheseweg besitzt, macht eine robuste und zeitsparende Methode zur Assemblierung der benötigten genetischen Elemente äußerst erstrebenswert. Daher entwarfen wir eine modulare Klonierungsstrategie und entwickelten ein Nomenklatursystem, was die Basis für eine Hochdurchsatz-Konierungsmethode darstellt. Dies soll die Herstellung komplexer Multigenkonstrukte, die mehrere Enzyme und genetische Kontrollelemente umfassen, ermöglichen. Ferner vereinfacht es die Assemblierung unterschiedlicher Expressionskonstrukte zur Herstellung einer bestimmten ncAS. Dadurch wird die Möglichkeit geschaffen ein Screening nach dem effektivsten genetischen Konstrukt zur Synthese der betreffenden ncAS durchzuführen.

a) The published German abstract is different to the English Abstract. At the time of publication of the Germain abstract, no public access was granted to the other content of the thesis. Therefore, the German abstract does not elaborate on any strategic or experimental details.

Table of contents

Scope of the thesis

Incorporation of non-canonical amino acids (ncAAs) into proteins is an emerging research field in Synthetic Biology as well as protein engineering. It allows furnishing a protein with chemical features that are not available within the canonical repertoire of amino acids prescribed by the genetic code. Because fluoro amino acids proved to be expecially attractive to engineer protein traits a detailed literature survey presents the first chapter of this thesis.

Although the incorporation of ncAAs is of high scientific interest, their expensiveness and restricted commercial availability often hampers more extensive research. Therefore, direct *in vivo* enzymatic synthesis of ncAAs in *E. coli* with their subsequent incorporation into heterologously expressed proteins was a primary research objective of this thesis.

The enzymatic route in nature to produce 4-fluorothreonine in microorganisms is the only one for a fluoro amino acid known so far. The 4-fluorothreonine pathway in *Streptomyces cattleya* is well studied and 4-fluorothreonine was already enzymatically synthesized in an *in vitro* approach. We set out to establish this enzymatic route to 4-fluorothreonine in *E. coli*. The second chapter of this thesis deals with establishing the first pathway step of the 4-fluorothreonine pathway to produce 5'-deoxy-5'-fluoroadenosine. Moreover, we elaborate on general requirements and potential bottlenecks to realize 4-fluorothreonie biosynthesis in *E. coli* for the production of fluorinated protein variants in the future.

We also explored the use of synthetic precursor molecules, that is non-canonical indoles, for *in vivo* synthesis of non-canonical tryptophans what is presented in the third chapter. We achieved to develop and characterize an *E. coli* whole cell system that produces non-canonical tryptophans from cheaper indole precursors that were successfully incorporated into target proteins.

Every ncAA has its own synthesis pathway, either natural or engineered, what calls for a robust and time-saving approach to assemble the required genetic elements. Modular Cloning (MoClo) based on Golden Gate cloning presents an effective DNA assembly strategy. However, no published MoClo study could meet our requirements for a modular assembly strategy in *E. coli*. Therefore, we designed a MoClo strategy and developed a nomenclature system. The MoClo design and experimental protocols described in the last chapter of this thesis provide the basis for a high-throughput assembly strategy. This should facilitate the assembly of complex multigene pathway constructs encoding several enzymes and genetic control elements. Moreover, it simplifies the assembly of several different expression constructs for one pathway to screen for the most effective in the synthesis of the non-canonical amino acid in question.

1 Chapter

Fluoro amino acids: A rarity in nature, yet a prospect for protein engineering

Corinna Odar[1,2], Margit Winkler[1], Birgit Wiltschi[1]

[1] Austrian Centre of Industrial Biotechnology, Petersgasse 14, Graz, Austria

[2] Graz University of Technology, Institute of Molecular Biotechnology, Petersgasse 14, Graz Austria

Author contributions

Corinna Odar wrote the manuscript under the guidance of Birgit Wiltschi. Birgit Wiltschi and Margit Winkler provided literature, scientific discussion and proofreading of the manuscript.

1.1 **Abstract**

Fluoro amino acids are highly valuable compounds constantly gaining relevance in diverse fields of the biosciences as well as in the pharmaceutical industry. This can be attributed to the properties of the extremely electronegative fluorine atom. It forms a highly polarized bond of extraordinary strength with carbon. The formation of the fluorine-carbon bond is challenging: Its chemical synthesis demands harsh reaction conditions and to date only one class of enzyme has been found capable of introducing the fluoride ion into an organic compound. Most of these fluorinating enzymes participate in the biosynthesis of 4-fluoro-L-threonine, the only fluoro amino acid of natural origin discovered so far. Despite their scarcity in nature, fluoro amino acids are valuable tools to fluorinate proteins. The fluoro protein variants often show improved stability and folding as well as altered activity and fluorescence characteristics.

This review details the biosynthesis of 4-fluoro-L-threonine with a special focus on the fluorinating enzymes. Moreover, we elaborate on the application of fluoro amino acids as building blocks for fluorinated protein variants. An insight into different techniques to incorporate fluoro amino acids into proteins is also provided. We highlight prospects and current relevance of fluoro amino acids as a tool to engineer proteins with novel traits.

1.2 **Introduction**

The number of man-made organofluorine compounds is ever increasing [1]. Living organisms, however, do not preferentially furnish their biomolecules with fluorine. Only one fluoro amino acid (FAA), 4-fluoro-L-threonine (4F-Thr), is currently known to occur in nature.

Both, the virtual absence of organofluorines from living matter as well as the growing interest in organic compounds containing this halogen originate from its outstanding physicochemical characteristics.

With an electronegativity value of 4.0 on the Pauling scale, fluorine is the most electronegative element in the periodic table [2]. When bound to carbon, fluorine exerts a high polarization rendering the carbon-fluorine bond (C-F) the strongest of all carbon-halogen bonds. The C-F bond even exceeds other covalent carbon bonds in strength, like the carbon-oxygen or carbon-hydrogen bond [3]. Being the smallest of all halogens, fluorine is also an attractive substituent for hydrogen in terms of size (van der Waals radius of H, 1.2 Å [2] *vs* F, 1.47 Å [4]). This applies not specifically to amino acids but to organic compounds in general. Compounds with fluorine mimic their parent compounds with hydrogen well. For example, they fit into the same enzyme binding sites [5], although the C-F bond (1.35 Å [5]) is significantly longer than the C-H bond (1.09 Å [3]).For a comprehensive overview of organofluorine chemistry the review by O'Hagan is recommended [3].

Pharmacologists have recognized the great value of fluorinated compounds for various diagnostic and therapeutic applications [6] and a number of reviews [5, 7-12] and books [13, 14] acknowledge the role of organic fluorine in medicinal chemistry. FAAs are among the organofluorines that find application in this field. For example, free FAAs can serve as mechanism based inhibitors [5, 15] and [18F]-labeled amino acids are frequently used for positron emission tomography (PET) imaging [16-21].

Upon incorporation into proteins, FAAs deliver the extraordinary characteristics of fluorine into the protein matrix while only marginally disturbing its structure [22-27]. FAAs as building blocks represent a convenient way to fluorinate proteins. Often, residue-specific fluorination of proteins improves their stability and folding [22, 24, 26, 28, 29] as well as their activity [23, 30].

Despite its increasing importance for pharma applications and in the biosciences, fluorination chemistry suffers from harsh conditions and largely non-specific reactions [31]. Enzymes capable to selectively fluorinate organic compounds under mild and environmentally friendly conditions would offer a way to tame fluorination chemistry. Indeed, nature evolved fluorinating enzymes, although this enzymatic trait seems rare from a current perspective. Intriguingly, the fluorinating enzyme is involved in the biosynthesis of an FAA.

Here, we review the biosynthesis of the only currently known natural FAA with a special focus on the fluorinating enzyme in its biosynthetic pathway. Moreover, we discuss FAAs as potent building blocks for protein engineering.

1.3 The biosynthesis of 4-fluoro-L-threonine

1.3.1 Naturally occurring organofluorines

Organofluorines are very rare in nature. They constitute the least prominent organohalogen species although fluorine is the most abundant halogen in the earth's crust [32]. In nature, abiotic C-F formation only occurs under harsh conditions like those encountered in geothermal processes [32]. In the biosphere, only a few species including tropical and subtropical plants and some microorganisms carry out organofluorine biosynthesis [33-36].

Products of organofluorine biosynthesis pathways comprise fluorinated fatty acids, nucleosides, fluoroacetone, fluoroacetate (F-acetate) and one FAA [34]. Plants produce most of these compounds and their organofluorine metabolism has been reviewed recently [37]. Also a marine organism, the sponge *Phakellia fusca*, is reported to produce fluorouracil derivatives [36]. However, a very recent review by O'Hagan and Deng points out that fluorouracil and fluoroacetone might not be of biotic origin [38]. F-Acetate is the most abundant fluorinated bioproduct. It is mainly found in plants from Australia and Africa [34, 39].

Microbial fluoride incorporation was unnoted until 1956 when a fluorinated nucleoside in the fermentation broth of *Streptomyces clavus* attracted attention for its antibiotic activity [40, 41].

In an organism of the same genus, *Streptomyces cattleya*, another fluorometabolite antibiotic, 4F-Thr, was identified. Discovered already in the mid-eighties [42], it still constitutes the only naturally FAA known to date. *S. cattleya* does not only secrete 4F-Thr but also the toxic F-acetate [42]. It secretes even more of the latter than of 4F-Thr [43].

Recently, four other microbial species have been found that represent potential fluorometabolite producers: *Streptomyces sp.* MA37, *Nocardia brasiliensis, Actinoplanes sp.* N902-109 [33, 35] and, remarkably, also a marine microorganism, *Streptomyces xinghaiensis* NRRL B-24674 [44]. *Streptomyces sp.* MA37 cultures already proved to synthesize 4F-Thr and F-acetate as well as some so far unidentified fluorometabolites [33]. In cultures of *S. xinghaiensis* F-acetate was found to be the sole fluorinated product [44].

1.3.2 The role of 4-fluoro-L-threonine in the metabolism of *S. cattleya*

Currently, it is unknown how the 4F-Thr producers benefit from its biosynthesis. 4F-Thr was shown to act as an antibiotic agent towards a number of bacteria [42]. As an analog of threonine, 4F-Thr might act as an antagonist as many other amino acid analogs do [45, 46]. However, its mechanism of action has not yet been analyzed in detail.

4F-Thr biosynthesis proceeds at the expense of one threonine for each 4F-Thr molecule formed (Fig. 1A) [47, 48]. This metabolic expense is intriguing as *S. cattleya* already produces the highly toxic F-acetate that has also antimicrobial activity. Therefore, 4F-Thr biosynthesis should account for a selective advantage for *S. cattleya* despite, or rather because, of its relatively high metabolic cost.

Deng *et al.* [47] suggested that 4F-Thr could be incorporated in place of a threonine into a specific protein as a defense mechanism under metabolic stress. However, this hypothesis has not yet been experimentally tackled.

1.3.3 Fluorometabolite biosynthesis in *S. cattleya*

F-Acetate and 4F-Thr are co-products of a specific biosynthesis pathway in *S. cattleya* (see Fig. 1A). Taking great efforts, all reaction intermediates of this pathway have successfully been elucidated since the 1990s [43, 48-58]. Deng *et al.* [47] comprehensively reviewed the deciphering history until 2004. The process employed isotope labeling studies in order to evaluate the involved substrates and ^{19}F-NMR spectroscopy to confirm the identity of the biosynthesized organofluorine compounds [47]. Upon identification of the first product in the pathway, 5'-deoxy-5'-fluoroadenosine (FDA) (Fig. 1A), the enzyme catalyzing the incorporation of fluorine into a metabolite of *S. cattleya* was identified [57]. In an exceptional reaction, the fluorinase (EC 2.5.1.63) activates the fluoride ion for the conversion of S-adenosyl-L-methionine (SAM) into FDA [57]. Three additional enzymes constitute the pathway to the intermediate fluoroacetaldehyde (F-acetaldehyde). At this point, the pathway diverges to yield either F-acetate or 4F-Thr in two independent enzyme catalyzed reactions [59]. The complete biosynthesis pathway is presented in Fig. 1A.

1.3.4 Enzymes for fluorometabolite production and their annotation in the *S. cattleya* genome

The unavailability of a complete *S. cattleya* genome sequence until 2011 (Table S1) [60, 61] impeded gene assignment as well as homology searches to decipher the fluorometabolite pathway enzymes. In 2006, the gene encoding the fluorinase (*flA*) was identified as the first of the pathway enzymes [53]. Examining its genomic surrounding, the FDA phosphorylase gene (*flB*) was identified as the second enzyme in the pathway.

Following the steps catalyzed by the fluorinase and FDA phosphorylase, the next enzyme in the pathway is an isomerase (Mri1, methylthioribose-1-phosphate isomerase 1). It was isolated and shown to convert 5-deoxy-5-fluororibose phosphate (FDRP) to 5-deoxy-5-fluororibulosephosphate (FDRibulP) (Fig. 1A) [48]. This isomerase (Mri1) most likely does not only participate in fluorometabolite biosynthesis but also in the L-methionine salvage pathway [48] as it readily accepts 5-methylthio-ribose-1-phosphate as the substrate *in vitro* [62]. Zhao *et al*. showed that a *Δmri1* mutant lost its ability to synthesize F-acetate and 4F-Thr [61] implying that Mri1 was the only isomerase to convert FDRP in *S. cattleya*. In contrast, Walker *et al.* discovered an additional FDRP isomerase (Mri2). They demonstrated that only the *Δmri1Δmri2* double knockout rendered the strain devoid of F-acetate and 4F-Thr production. If only one of the two isomerase genes was deleted, fluorometabolite production was mutually rescued by the other isomerase [62].

The two studies used a slightly different experimental setup: Both assessed fluorometabolite production after 6 days of cell growth in the presence of 2 mM fluoride. While Walker *et al.* [62]

analyzed the supernatant of *S. cattleya* cultures grown in full media, Zhao *et al.* [61] assessed the biotransformation of added SAM and NaF in cell-free extracts from cells cultivated in defined media.

The differences in the cultivation conditions and in the evaluation of fluorometabolite synthesis might explain the contradicting results. Nevertheless, the way in which the knockout mutants were constructed deserves a more detailed discussion. Walker *et al.* and Zhao *et al.* used the same *S. cattleya* strain (NRRL 8057 (ATCC 35852) and DSM 46488 designate the same strain). Although they used different nomenclatures to identify the *mri1* gene (SCAT_2018 [62] and SCATT_20080 [61], see Table S1) they applied the same strategy to delete the entire open reading frame by the insertion of different selectable resistance markers (the hygromycin [62] or apramycin [61] resistance cassettes). However, Walker *et al.* [62] found that the orientation of the inserted selectable marker was crucial: When inserted on the coding strand of SCAT_2018 it caused a severe growth defect and no organofluorines were produced (supplementary information of [62]). When the marker was inserted on the non-coding strand, the cells grew normally and were capable of organofluorine production. Notably, Zhao *et al.* inserted the apramycin resistance cassette "within the coding region of the gene to be disrupted and in the same direction of transcription" [61], that is on the coding strand. Although the mutant appeared "healthy ... when grown under standard conditions" [61] it was unable to biosynthesize F-acetate or 4F-Thr. This means that the single *Δmri1* mutant of Zhao *et al.* showed a similar phenotype with respect to organofluorine biosynthesis as that of Walker *et al.* where the hygromycin marker had been inserted in the same orientation. These findings call for further experimental studies to reveal the definite role of Mri1 and Mri2 for fluorometabolite synthesis in *S. cattleya*.

In the next pathway step, a Zn^{2+} dependent fuculose aldolase is proposed to convert FDRibulP to F-acetaldehyde (Fig. 1A) [58]. So far, the identification of this enzyme has remained elusive. Genome sequencing of *S. cattleya* [60, 61] and knockout experiments allowed narrowing down putative genes to two nucleotide sequences (Table S1) [61]. Whether one of these sequences encodes the aldolase to convert FDRibulP still needs to be elucidated.

A pyridoxal phosphate dependent transaldolase finally converts F-acetaldehyde into 4F-Thr at the expense of one equivalent of threonine (Fig. 1A) [56]. For gene assignment, the 4F-Thr transaldolase was purified [48]. Finally, knockout studies of the 4F-Thr transaldolase gene verified its indispensable role to produce 4F-Thr in *S. cattleya* [61].

The F-acetaldehyde dehydrogenase was isolated [55], and its genomic identity determined [60]. The enzyme catalyzes the formation of F-acetate, the second end product of the pathway besides 4F-Thr.

The reconstitution of the *S. cattleya* 4F-Thr biosynthesis pathway in an *in vitro* approach confirmed the specific enzyme functions. Except for the still unidentified aldolase, all 4F-Thr/F-acetate pathway genes from *S. cattleya* were heterologously expressed and purified [48, 63]. The isolated enzymes were combined to successfully reconstitute the pathway *in vitro* for 4F-Thr production [48, 63]. An aldolase from *Streptomyces coelicolor* (Table S1) effectively substituted the unidentified *S. cattleya* enzyme in the *in vitro* assay.

Recently, *flA* homologs were identified in *Streptomyces sp.* MA37, *Nocardia brasiliensis, Actinoplanes sp.* N902-109 genomes, and in *Streptomyces xinghaiensis* [33, 35, 44]. These species also harbor homologs of other fluorometabolite pathway genes of *S. cattleya*. Like FlA from *S. cattleya*, the FlA homologs of *Streptomyces sp.* MA37, *N. brasiliensis,* and *Actinoplanes sp.* N902-109 produced FDA from fluoride and SAM *in vitro*. The functions of the other pathway homologs in these newly identified fluorometabolite producers still need to be verified but their involvement in 4F-Thr and F-acetate biosynthesis seems likely. In *S. xinghaiensis* a truncated homolog of the *S. cattleya* 4F-Thr transaldolase was identified; this homolog seems not to have any 4F-Thr synthesis capacity [44]. Cultures of *S. xinghaiensis* only produced F-acetate [44], whereas cultures of *Streptomyces sp.* MA37, harboring a full length 4F-Thr transaldolase homolog produced both, F-acetate as well as 4F-Thr [33].

1.3.5 **The *fl* cluster**

Twelve genes, all of which are most likely involved in fluorometabolite production, are clustered together in the genome of *S. cattleya* (Fig. 1B) [53]. This cluster, termed *fl*, only harbors two enzymes directly involved in fluorometabolite production: The genes encoding the fluorinase (*flA*) and the FDA phosphorylase (*flB*). The role of the remaining 10 genes *flC, flD, flE, flF, flG, flH, flI, flJ, flK,* and *flL* has only been partly deciphered.

FlI is supposed to hydrolyze S-adenosyl-L-homocysteine (SAH). SAH is a competitive inhibitor of the fluorinase. Consequently, FlI might keep cellular levels of SAH low for efficient 4F-Thr/F-acetate formation [53].

The fluoroacetyl-CoA thioesterase FlK is supposed to be involved in F-acetate detoxification [53]. F-acetate is an acetate mimic and can be converted to fluoroacetyl-CoA. Fluoroacetyl-CoA enters the tricarboxylic acid cycle and is converted to a toxic intermediate that leads to the shutdown of the pathway [64]. FlK specifically breaks down fluoroacetyl-CoA to F-acetate and coenzyme A thereby preventing the poisoning effect [53, 65].

The gene products of *flH* and *flC* were proposed to be transmembrane transporters facilitating fluoride uptake or fluorometabolite secretion, respectively [53]. However, the knockout of *flH* had no apparent effect on fluorometabolite production in *S. cattleya* [61]. The role of *flC* was not further examined.

FlE, FlF, FlG and FlL are putative transcription factors that might control the expression of the *fl* genes [53]. The functions of the gene products of *flD* and *flJ* remain speculative [53].

Streptomyces sp. MA37, *Actinoplanes sp.* N902-109, *N. brasiliensis* and *S. xinghaiensis* also show clustering of the putative homologs for fluorometabolite production around the individual *flA* homologs [33, 35, 38, 44]. In these species, the homologs of *flB* are located in close proximity to *flA* as well as the homologs of the 4F-Thr transaldolase, which is not part of the *fl* cluster in *S. cattleya*. In *N. brasiliensis* and *Actinoplanes sp.* N902-109 the homologs of the FDRP isomerase are also located inside the *fl* cluster. Very recently, O'Hagan and Deng analyzed the organization of the genes surrounding the fluorinase genes in the different species [38]. They found that the order and the

direction of transcription of the genes around the *flA* homologs in the Streptomycetes were largely conserved reflecting their close evolutionary relationship. In comparison, the organization of the genes neighboring the *flA* homologs in the evolutionary more distant *N. brasiliensis* and *Actinoplanes sp.* N902-109 was less well conserved [38].

1.4 **Enzymatic fluorination: The fluorinase**

Of all the enzymes involved in the fluorometabolism of *S. cattleya*, the fluorinase has received the most attention so far as it provided the opportunity to reveal for the first time the feat of enzymatic fluoride activation.

What renders fluoride almost inaccessible for enzymatic conversion? What makes it so different compared to other halogens and why is it not susceptible to conventional enzymatic halogenation mechanisms? The properties of fluorine listed in the introduction already partially answer these questions: The outstanding electronegativity of fluorine is a serious challenge. Enzymatic halogenation proceeds predominantly via two oxidative mechanisms that either use H_2O_2 or O_2 as the reducible substrate [66, 67]. The enzymatic activation of chloride, bromide and iodide ions involves the formation of electrophilic halogen intermediates, such as X^+ equivalents or halogen radicals [66]. As F^+ is the strongest oxidizing agent [68] it is not surprising that the known halogenases are unable to muster the redox potential required for its formation [66, 67]. Consequently, carbon-fluorine bond formation has to be distinct from common carbon-halogen bonding that relies on electrophilic halogen intermediates. The only feasible way for enzymatic carbon-fluorine bond formation known to date is the nucleophilic halogenation mechanism [52, 57, 69]. Fluoride is surrounded by a shell of tightly bound water molecules. This water shell renders hydrated fluoride the least nucleophilic halide ion. Thus, the enzyme has to overcome the high solvation barrier to make F^- accessible for bioconversion. Once liberated from its water shell, fluoride is a highly potent nucleophile [66, 67]. In general, the nucleophilic halogenation mechanism occurs rarely as reflected by the small number of reported examples [70-72].

The first reported fluorinating enzymes were two distinct active site mutant glycosidases, an *Agrobacterium sp.* β-glucosidase and a *Cellulomonas fimi* β-mannosidase [69, 73]. The active sites of both enzymes were mutated by replacing the catalytic nucleophile, a glutamate, with alanine, glycine or serine thereby abolishing their glycosidic bond cleavage activity. However, when the enzymes were assayed with different glycosides in the presence of fluoride, substantial glycosidic bond formation activity was restored. Obviously, fluoride was acting in the place of the missing catalytic nucleophile and fluorinated intermediates were formed. Although these intermediates were too reactive for isolation, experimental results argued for a nucleophilic fluorination mechanism that might proceed via hydrogen bonding and desolvation of the fluoride anion [69].

Indeed, co-crystallization studies of the fluorinase from *S. cattleya* with SAM and fluoride supported the notion that the fluoride ion is dehydrated before the C-F bond is formed [52]. The nucleophilic attack of the fluoride on the 5' C of the SAM sugar moiety occurs with an inversion of the 5' C

configuration supporting an S_N2 type of reaction [52, 74]. This was eventually confirmed by stereospecific labeling of the 5'C in the ribose moiety of SAM [75]. Crystallographic data showed the fluorinase structure to be a dimer of trimers (Fig. 2A) [52]. A review by O'Hagan gives a comprehensive overview of the scientific efforts to determine the mechanistic characteristics of the fluorinase until 2006 [59].

How can the fluorinase overcome the kinetic desolvation barrier? This immanent question was tackled by computational studies [76, 77] before the detailed molecular mechanism was unveiled in an experimental study by Zhu *et al.* [78] (Fig. 2B). Polar groups replace the water molecules by hydrogen bonding in the partially desolvated fluoride, which is further stabilized by the positive charge of the sulfur group in SAM. The remaining water molecules are released upon binding of the fluoride to the sugar moiety of the adenosine ring.

A recent study by Lohman *et al.* highlights the power of the SAM-mediated fluoride alkylation catalyzed by the fluorinase [79]. Their experiments show that the rate of alkyl transfer is elevated by a factor of 2×10^{15} compared to the non-enzyme catalyzed reaction.

As the C-F bond formation by fluorinase was recognized as an outstanding catalytic enzyme mechanism, the discovery of a related SAM dependent chlorinase aroused vivid scientific interest [70]. The chlorinase of the marine organism *Salinispora tropica* is involved in the biosynthesis of salinosporamide A (Fig. 3). It uses the same rare nucleophilic substitution strategy as the fluorinase but with 5'-chloro-5'-deoxyadenosine as the corresponding pathway intermediate. Although the chlorinase accepts bromide and iodide as nucleophiles, no reaction occurs with fluoride [70]. Therefore, Eustáquio *et al.* [80] took a metabolic engineering approach to reprogram *S. tropica* for fluorosalinosporamide production. Upon substitution of the chlorinase with the fluorinase, *S. tropica* was capable of producing fluorosalinosporamide (Fig. 3).

Although the chlorinase does not accept fluoride [70], the fluorinase accepts chloride to give 5'-chloro-5'-deoxyadenosine. However, fluoride is clearly preferred over chloride by a factor of 120 [81]. The fluorinase of *S. cattleya* shares 35% amino acid identity with the *S. tropica* chlorinase [82]. The fluorinase is also related to the class of duf-62 enzymes with homologies to selected domains ranging between 25 and 32%. The largest enzyme class within the duf-62s are the SAM hydroxide adenosyltransferases mediating the hydrolytic cleavage of SAM to adenosine and L-methionine. Instead of a halide ion as the nucleophile, the SAM hydroxide adenosyltransferase reaction employs a hydroxide ion to attack the 5'C of SAM [82]. The recently identified fluorinase homologs from *Streptomyces sp.* MA37*, Actinoplanes sp.* N902-109, *N. brasiliensis* and *S. xinghaiensis* show identities between 80% to 87% to the *S. cattleya* enzyme [33, 44] (Fig. 4). All of them share a 21 amino acid loop that seems to be characteristic for this class of fluorination enzymes. Moreover, the crystal structure of the *Streptomyces sp.* MA37 homolog, the only structure available so far for the newly discovered fluorinases, is almost identical to that of the *S. cattleya* fluorinase [33].

Interestingly, in the plant kingdom not a single sequence match was yet identified in homology searches with the fluorinase [78]. Given the presence of organofluoride compounds in plants, this

either suggests enzymatic fluorination as an example of convergent evolution or microbes are the actual producers of the F-acetate found in plants [83].

1.4.1 *In vitro* application of the fluorinase from *S. cattleya*

The *S. cattleya* fluorinase appears to be the perfect candidate for enzymatic fluorination. However, its halogenation ability is strictly restricted to its canonical substrate SAM [84], although the enzyme can use the structurally similar 2'-deoxy SAM [85] and chloride instead of fluoride. This strict substrate specificity impairs the exploitation of the fluorinase as a universal 'fluorinating agent' for biotechnological purposes.

Nevertheless, many enzymes are able to accept fluorinated compounds as substrates or co-substrates. Consequently, several enzymatic methods for the preparation of fluorinated compounds have been developed that exploit the enzymes' blindness towards fluorine substitution. The fluorination pathway in *S. cattleya* was exploited by combining the synthesis of pathway intermediates with a variety of other possible enzymatic conversions [86-89]. The approach proved to be especially attractive for the biocatalytic synthesis of [18F]-labeled compounds for PET. The immobilization of the fluorinase is also a research objective [90, 91]. It facilitates enzyme recycling and makes the application of the fluorinase commercially more attractive.

1.5 FAAs as building blocks for peptides and proteins

Currently, it is unclear whether 4F-Thr, the only naturally occurring FAA known so far, is incorporated into proteins in nature. Neither did we find a report on its successful incorporation into a recombinant protein.

Organochemical synthesis offers a plethora of FAAs that can be applied as building blocks for proteins and peptides to give fluorinated variants. In the next sections, we highlight how the properties of fluorine are used to engineer proteins and peptides by introducing FAAs in place of their non-fluorinated counterparts.

This research field is gaining scientific interest as the number of publications on fluorinated peptides and proteins has increased in recent years (see the examples below). Not least, this can be ascribed to the scientific advances in the protocols for the incorporation into proteins of amino acids that are not prescribed by the genetic code, the so-called non-canonical amino acids [92-96]. In general, the incorporation methods for non-canonical amino acids can be directly transferred to FAAs.

Like other non-canonical amino acids, FAAs can be introduced into proteins and peptides by chemical synthesis [97] or ribosomal translation [95]. Chemical peptide synthesis is restricted to 50-100 amino acid residues [98]. Nevertheless, bigger proteins have been chemically synthesized such as the green fluorescent protein from *Aequorea victoria* (GFP), which is composed of 238 amino acid residues [99]. A way to overcome the size limitations of chemical peptide synthesis are semisynthetic protein synthesis techniques, i.e. protein ligation [100]. Otherwise, FAAs can be channeled into ribosomal

translation *in vitro* and *in vivo*. The reader is referred to the literature [92-96] for a deeper insight into general aspects of the methodology.

In vivo translation of FAAs is currently achieved by two different strategies (Fig. 5A): the residue-specific and the site-specific incorporation. In the residue-specific approach a canonical amino acid is globally replaced by its fluorinated analog. Permissiveness of the translational machinery towards the fluorinated analog is a prerequisite for this approach. In 1959, Munier and Cohen [101] proved that the phenylalanyl-tRNA synthetase (PheRS) of *E. coli* accepts a FAA by demonstrating *in vivo* incorporation of 4-fluoro-L-phenylalanine (4F-Phe) in the place of phenylalanine in *E. coli* proteins. However, full residue-specific incorporation of an FAA is only achieved in a strain that is auxotrophic for the amino acid to be replaced: The auxotrophy facilitates the depletion of the canonical amino acid which is then replaced by an excess of the FAA [102].

Apparently, residue-specific incorporation requires the FAA to be sufficiently similar to its natural analog for recognition by the host aminoacyl-tRNA synthetase (aaRS) [103]. Due to their isostructural characteristics, monofluorinated analogs of canonical amino acids are especially suitable for the residue-specific approach [104].

In contrast, for the site-specific incorporation, sufficient structural difference of the FAA to its natural analog appears to be advantageous (Fig. 5A) [103]. Using an appropriately engineered pair of aaRS and tRNA, an FAA can be inserted into a target protein at a distinct position *in vivo*. An in-frame stop or 4-base (quadruplet) codon, that are both naturally not recognized for amino acid translation, is read by an engineered suppressor tRNA carrying the appropriate anticodon ($tRNA^{aa}_{anticodon}$). An aaRS is engineered such that it charges the suppressor tRNA with the FAA. The FAA is then incorporated at the position of the stop or quadruplet codon thereby expanding the canonical amino acid repertoire [95]. For a successful incorporation, the engineered aaRS/$tRNA^{aa}_{anticodon}$ pair must be orthogonal for the expression host: It should unambiguously recognize the FAA without cross-reacting with any canonical amino acid because promiscuous charging of the tRNA would result in inefficient analog incorporation. The mutant aaRS should also not charge any endogenous *E. coli* tRNA with the FAA because this could produce aberrantly labeled host proteins. Although strains facilitating read-through at multiple amber stop codons have been developed [93, 105, 106], the site-specific incorporation of non-canonical amino acids often focusses at a single in-frame amber position.

In 1998, Furter was the first to incorporate 4F-Phe at an amber stop codon in *E. coli* [107]. He used a non-engineered PheRS/$tRNA^{Phe}_{amber}$ pair from *Saccharomyes cerevisiae*. As the *S. cerevisiae* PheRS/$tRNA^{Phe}_{amber}$ naturally incorporates phenylalanine, the *S. cerevisiae* $tRNA^{Phe}_{amber}$ was charged with 4F-Phe only in the presence of excess 4F-Phe compared to phenylalanine. In order to avoid uniform incorporation of 4F-Phe into all nascent proteins, a 4F-Phe ‘resistant’ host strain was employed. While the endogenous PheRS of wild type *E. coli* accepts 4F-Phe when present in excess, the mutant PheRS of the 4F-Phe ‘resistant’ host strain excludes 4F-Phe from its mutant active site. These experimental conditions enabled site-specific 4F-Phe incorporation with up to 75% efficiency [107].

In 2007, fully site-specific labelling with the FAAs trifluoromethyl-phenylalanine [108] and 4'-[3-(trifluoromethyl)-3H-diazin-3-yl]-L-phenylalanine [109] was achieved. Engineering of aaRS/tRNA$^{aa}_{amber}$ pairs from *Methanococcus jannaschii* enabled orthogonal incorporation of these FAAs [108, 109].

In the years to follow, incorporation of 2-amino-3-(4-(trifluoromethoxy)phenyl)propanoic acid [110], fluorinated benzoylphenylalanine derivatives [111], polyfluorinated 4-methyl phenylalanine derivatives [112] and of *p*-trifluoroacetylphenylalanine [113] was demonstrated. For the evolution of an efficient mutant aaRS for an FAA, sufficient structural difference of the FAA to its natural analog is assumed advantageous [103]. Nevertheless, in 2011, Young *et al.* [113] engineered an aaRS/tRNA$^{aa}_{amber}$ pair with polysubstrate specificity that facilitated the site-specific incorporation of the monofluorinated amino acid 4F-Phe into a target protein. In contrast to the earlier work of Furter described above [107] the engineered aaRS/tRNA$^{aa}_{amber}$ pair did not promiscuously charge phenylalanine. Hence, site-specific incorporation of 4F-Phe could be achieved even in the presence of background phenylalanine during expression. The same approach of engineering an aaRS/tRNA$^{aa}_{amber}$ pair was also successful for 3F-Tyr (3-fluoro-L-tyrosine) [114]. Similar to 4F-Phe, the tyrosyl-tRNA synthetase (TyrRS) of wild type *E. coli* accepts the 3F-Tyr analog. The calculated binding energy of 3F-Tyr (-8.42 kcal/mol) docked in the tyrosine binding pocket of *E. coli* TyrRS is even slightly more favorable for binding than that of the natural substrate tyrosine (-8.36 kcal/mol) [22]. However, 3F-Tyr was specifically inserted at an amber site in the target protein by the engineered 3F-TyrRS/tRNA$^{3F\text{-}Tyr}_{amber}$ pair and incorporation of 3F-Tyr at tyrosine positions by the wild type TyrRS was not reported [114]. Obviously, the orthogonal 3F-TyrRS/tRNA$^{3F\text{-}Tyr}_{amber}$ pair efficiently competes with the *E. coli* TyrRS for the fluorinated tyrosine analog. Moreover, the high specificity of the orthogonal pair for 3F-Tyr allows for sufficient tyrosine in the culture medium to suppress the charging of host tRNAs with 3F-Tyr by the *E. coli* TyrRS.

Another approach to overcome the permissiveness of the *E. coli* translation machinery towards monofluorinated amino acids was demonstrated by Wilkins *et al.* [115]. They structurally modified F-Tyr analogs using a photocleavable protection group to create photocaged F-Tyr. This modification disguised the F-Tyr by rendering it sufficiently different from tyrosine such that it was no longer recognized by the endogenous translation machinery. In parallel, an orthogonal aaRS/tRNA$^{aa}_{amber}$ pair was evolved for the photocaged F-Tyr that facilitated its site-specific introduction into target proteins. After translation, the photo-cage was removed by irradiation with UV light in order to liberate the F-Tyr [115].

Engineering of aaRS facilitates the residue-specific incorporation of FAAs that are not recognized by the endogenous translational machinery. Marginal or absent incorporation of some perfluorinated amino acids (PFAAs) impedes their use for protein engineering [116, 117]. To tackle this issue, Yoo and Tirrel engineered *E. coli* methionyl-tRNA synthetase such that the mutated enzyme facilitated the global incorporation of the Met analog trifluoronorleucine in two model proteins [118].

The incorporation of an FAA at several positions usually affects the target protein in a synergistic manner. In contrast, site-specific incorporation allows studying the effect of one FAA at a distinct position in a target protein. The next chapter focuses on the application of protein engineering with FAAs with a special focus on the residue-specific incorporation approach.

1.6 The effects of FAA incorporation on protein and enzyme characteristics

For an extraordinarily comprehensive collection of data on fluorinated peptides, proteins and amino acid biopolymers, the reader is referred to the recent review by Salwiczek *et al.* [119]. Other reviews focus on fluorinated proteins synthesized *in vivo* [104] and fluorinated biocatalysts [120]. Here, we present important general aspects of FAA incorporation for protein engineering. Reflecting current trends of this research area in the scientific literature we focus mainly on monofluorinated aromatic amino acids (FaAAs) and proline as well as on PFAAs.

1.6.1 Incorporation of monofluorinated amino acid analogs

FaAAs clearly are the preferred building blocks for protein engineering purposes (*vide infra*).

Numerous studies have been devoted to the incorporation of fluorinated proline analogs as they turned out to be useful tools to manipulate the stability and folding of proteins [26-29, 121-126].

So far, other monofluorinated amino acids, whether hydrophobic, polar, or charged, have received less scientific attention for protein engineering. Most of these studies focus on the incorporation of monofluorinated histidine residues [127-131]. Because the pK_a of a histidine residue decreases upon fluorination [132], its substitution by a fluorinated analog can be used to probe the role of its protonation state for protein function [127-131]. Concerning monofluorinated hydrophobic amino acids, studies on the incorporation of 5-fluoro-L-leucine suggest that this FAA is well tolerated as a substitute for its natural analog in native protein structures [133, 134]. Future research to elucidate the effects of monofluorinated amino acids other than FaAAs and fluoro-L-proline (F-Pro) on protein structure and function should provide further insight into their potential as protein engineering tools.

1.6.1.1 Incorporation of monofluorinated aromatic amino acid analogs

Tyrosine and phenylalanine are commonly substituted by *ortho*-, *meta*- or *para*-monofluorinated analogs. Tryptophan residues are targets for fluorination of the aromatic six-ring as well, though to a lesser extent.

Notably, monofluorination of amino acid residues in proteins usually preserves the overall protein structure [22-25, 135, 136]. Even the simultaneous incorporation of three different monofluorinated amino acids ((*2S*,*4S*)-4-fluoroproline ((*4S*) F-Pro), 4F-Phe and 6-fluoro-L-tryptophan (6F-Trp)) into a lipase from *Thermoanaerobacter thermohydrosulfuricus* was possible without significant structural perturbations [25].

Furnishing enzymes with FaAAs can influence their activities. For instance, the incorporation of 3F-Phe increased the activity of *Pvu*II endonuclease by 2-fold [23]. Global incorporation of 3F-Tyr into glutathione S-transferase (GST) yielded a variant with increased acidity of the active site hydroxyl group [137]. This altered the manner in which the side chain participated in catalysis, i.e. the addition of the sulfur of glutathione to electrophilic functional groups. In the fluorinated GST variant, the 3F-Tyr anion appeared to act as a general base while the tyrosyl hydroxyl group in the native enzyme

provided electrophilic stabilization of the thiolate anion of the substrate. Consequently, the fluorinated GST displayed decreased catalytic activity towards 1-chloro-2,4-dinitrobenzene [137]. In contrast, incorporation of 5F-Trp in the same enzyme increased the turnover number of 1-chloro-2,4-dinitrobenzene almost 4-fold compared to the wild type [30]. However, the overall catalytic efficiency was only slightly increased as the 5F-Trp variant displayed a higher K_M than the wild type [30].

Besides that incorporation of FaAAs can modulate enzyme activities it is also used as a tool to improve protein stability and to preserve protein function especially in hostile environments. For instance, the incorporation of 3F-Tyr into GFP made the fluorophore more resistant against alkaline conditions [22]. Several FaAA variants of the lipase from *T. thermohydrosulfuricus* showed improved activities in organic solvents [28]. Global replacement of the tyrosine residues of ω-transaminase by 3F-Tyr enhanced the tolerance of the enzyme towards organic solvents and heat [138]. 4F-Phe incorporation into a phosphotriesterase facilitated enzyme reactions at elevated temperature compared to the wild type [24].

Incorporation of FaAAs was also used to modulate substrate specificities. Upon incorporation of either 3F-Phe or 4F-Phe, a histone acetyltransferase changed its substrate specificity depending on the position of the fluorine in the aromatic ring [139]. 3F-Phe in the lipase from *T. thermohydrosulfuricus* broadened its substrate scope. The fluorinated lipase variant was able to cleave shorter as well as longer fatty acid chains from triglycerides than the non-fluorinated parent enzyme [126].

In addition, the spectroscopic properties of proteins were tackled by FaAA incorporation. For instance, 4F-Trp incorporation was recognized to abolish intrinsic tryptophan fluorescence [140] due to its extremely low quantum yields [141]. Thus, 4F-Trp incorporation can be used to analyze the contribution of tryptophan fluorescence to the fluorescence spectrum of a protein [135, 140, 141]. Other fluorinated tryptophan analogs led to changes in the absorbance and emission maxima [135, 141].

Although canonical phenylalanine residues do not show a significant absorbance, their fluorination can turn proteins into valuable spectroscopic probes [136]. Global substitution of the phenylalanine residues with 2F-, 3F- or 4F-Phe endowed annexin A5 and azurin with two new absorption maxima in the UV range, which are unique to the fluorinated variants. Similar spectroscopic ‘fluorofingers’ were also observed with the 4F-Phe variant of *T. thermohydrosulfuricus* lipase [25].

Several studies aimed at the fluorination of the chromophore in GFP and its mutants. Tyr66 is part of the wild-type GFP chromophore as well as that of the enhanced GFP (EGFP) mutants [142] and the enhanced yellow fluorescent mutant [143, 144]. Another mutant, the enhanced cyan fluorescent protein, harbors tryptophan at position 66 instead of tyrosine (Tyr66Trp) [145]. Fluorination of these aromatic amino acid residues caused several effects: spectral shifts both red or blue, changes in the extinction coefficients and fluorescence intensities as well as changes in the pKa values [146, 147]. The fluorescence of another GFP mutant [148] furnished with 3F-Tyr in its chromophore correlated linearly with pH values from 3 to 8 rendering it a potential pH sensor [22].

Taken together, FaAAs can be applied for a variety of purposes from engineering protein stability or enzyme activity to fluorescence characteristics. Successful protein fluorination studies such as those

highlighted above can encourage future research on the incorporation of FaAAs into other, different protein classes. This implementation of FaAAs as an engineering tool opens up new prospects for protein design.

1.6.1.2 **Incorporation of fluoroproline**

In the set of the 20 canonical amino acids, proline occupies an exceptional place. It is the only amino acid with an alicyclic side chain. Though proteins with catalytic proline residues do exist [149], proline is usually not involved in catalysis and rather plays a pivotal role for the folding and stability of the protein scaffold [150]. An insight into the stereochemistry of proline is essential to understand the effects of its fluorination specifically at the C4 (C^{γ}) carbon atom.

The pyrrolidine ring of proline is not flat but can adopt two conformeric states, either the *exo* or the *endo* pucker (Fig. 5B) [151]. The two conformations differ in the position of the C4 relative to the plane of the pyrrolidine ring. The stereochemistry of a fluorine substituent at C4 promotes one of the two possible proline ring puckers [121, 152]: The (*2S,4R*)-4-fluoroproline ((*4R*)F-Pro) isomer favors *exo* whereas (*4S*)F-Pro favors *endo* puckering (Fig. 5B). This stereochemical bias can be attributed to the gauche effect, which itself is provoked by the inductive effect of the electronegative F substituent [121, 152].

The stereochemistry of a fluorine atom at position C4 of proline not only affects the ring pucker but also the conformation of the peptidyl-proline bond. Consequently, it can influence the folding behavior of a protein. This is because in peptidyl-proline bonds, one of the two proline ring puckering forms is favored and the same peptidyl-proline bond can adopt either the *trans* or the *cis* conformation.

DeRider *et al.* [152] used a series of model compounds containing a single prolyl derivative to analyze the effect of unsubstituted Pro, (*4R*)F-Pro or (*4S*)F-Pro on the *trans*/*cis* conformation of the peptidyl-proline bond. They found that a fluorine substituent in the (*4R*) position on proline stabilized the *trans* conformation relative to unsubstituted proline [152]. This stabilizing effect correlates with the preferred *exo* pucker of the pyrrolidine ring in (*4R*)F-Pro: The *exo* pucker maximizes the overlap between the nucleophilic lone pair (*n*) of the amide oxygen of an X_{aa}-Pro peptide bond (X_{aa} can be any amino acid) and the antibonding π* orbital of the carbonyl group from Pro. Because (*4R*)F-Pro prefers *exo* ring pucker [152] it stabilizes *trans* peptide bonds rather than does unsubstituted proline which exists in *exo* and *endo* puckers. In contrast, the n→π* interaction is less optimal with the *endo* pucker. (*4S*)F-Pro adopts preferably *endo* pucker and, consequently, the *trans* conformation of an X_{aa}-(*4S*)F-Pro bond is less favorable compared to X_{aa}-Pro. This explanation is supported by the observation of DeRider *et al.* who showed that a fluorine atom in the (*4S*) position stabilized the *cis* conformation of their model compound [152]. This finding correlates with the observation that the *cis* conformation of the peptidyl-proline bond in proteins is often associated with *endo* puckering of the proline ring (about 89%) [153]. However, in the *trans* conformation both ring puckering forms are found to the same extent [153]. This suggests that in the context of a complex protein structure several factors contribute to the preferred conformeric state for a given proline residue with respect to its *exo* or the *endo* pucker.

The correlation between proline ring puckering and peptidyl-proline bond conformation raises the question, whether it would be possible to tune the folding of a polypeptide by introducing biased ring

puckering of the 4F-Pro stereoisomers. Experimental studies indeed showed that incorporation of the respective 4F-Pro stereoisomers can promote formation of either the *cis* or *trans* peptide bond. For instance, if a peptidyl-proline bond demands both *trans* isomerization and *exo* proline ring puckering, (*4R*)F-Pro promotes the conformation of the native bond and therefore folding. Likewise, to promote a *cis* peptidyl-proline bond with the proline ring adopting the *endo* pucker, (*4S*)F-Pro is favorable [121, 123, 154, 155].

The ubiquitous structural protein collagen is a paradigm for the stereocontrol of folding and stability by a C4 substituent at the proline. The marked stability of collagen is a consequence of its triple helix structure. The individual polypeptides composing the collagen triple helix contain X_{aa}-Y_{aa}-Gly repeats, where (*2S*)-proline (L-proline) often occupies the X_{aa} position and (*2S*,*4R*)-4-hydroxyproline [(*4R*)OH-Pro] the Y_{aa} position ([156] and references therein). The electronegative hydroxyl group at C4 in (*4R*)OH-Pro imposes an *exo* pucker upon the proline ring. The highly electronegative fluorine substituent increases the inductive effect, thereby promoting proline *exo* puckering even more than the hydroxyl group does. Accordingly, the exchange of (*4R*)OH-Pro by (*4R*)F-Pro conferred hyperstability to a collagen peptide mimic [154]. The bias of (*4R*)F-Pro to adopt *exo* puckering leads to a favorable preorganization of the peptidyl-proline bonds in the denatured state, because the *exo* pucker is the energetically favored conformation for this *trans* peptidyl-proline bond in the native collagen triple helix [155, 157]. The preorganization minimizes the entropic costs of folding, i.e. the difference of entropy of the unfolded *versus* the folded state is comparably low for the (*4R*)F-Pro collagen peptide mimic compared to its canonical counterpart [155, 158]. Studies on an F-Pro elastin mimetic might corroborate the entropy effect as a driving force for enhanced folding kinetics [159]. Elastin fragments, that is tropoelastin, contain proline rich repetitive sequences such as poly-[ValProGlyValGly] [160]. In aqueous solution, elastin fragments undergo a temperature dependent phase transition, which is an essential step for the formation of elastic fibers [159, 161]. This transition (coacervation) is caused by a conformational change of the polypeptides, i.e. the formation of a type II β-turn conformation from a random coil. There is evidence for the abundance of proline *exo* puckering and *trans* conformation of the prolyl-peptidyl bond in the elastin type II β-turn conformation [153, 159]. Consequently, substitution of native proline by the (*4R*)F-Pro analog in the [ValProGlyValGly] repetitive units of the elastin mimetic peptide favored the conformational rearrangement essential for coacervation [159]. The altered conformational thermodynamics became evident in a lower transition temperature compared to the non-fluorinated elastin-mimetic peptide. In contrast, (*4S*)F-Pro incorporation into the same elastin disfavored the type II β-turn conformation. The (*4S*)F-Pro containing elastin showed a higher transition temperature than the non-fluorinated protein, an observation accordant with the (*4S*)F-Pro stereochemical bias for *endo* puckering [153, 159].

F-Pro incorporation also elevated the folded content of the proline rich peptide, ProProProLeuProProLysProLysPhe, in aqueous solution [162]. In the structured state, the peptide adopts a left-handed polyproline II helix with *trans* configured amide bonds. Incorporation of (*4R*)F-Pro at selected positions promoted helix formation [162], corroborating the preorganization principle of the peptidyl-(*4R*)F-Pro bonds in the denatured state.

The examples outlined above clearly demonstrate that the C4-fluorinated stereoisomers of proline can be used to manipulate the folding and stability of peptides. It raises the question whether F-Pro incorporation can be a useful tool to enhance the folding and stability of proteins and enzymes as well.

Indeed, this was demonstrated for simple model proteins harboring only a single proline residue: (*4S*)F-Pro incorporation conferred higher thermostability to a mutant of barstar [121], the inhibitor of the ribonuclease barnase [163]. This mutant contained only a single proline in which the native peptidyl-proline bond was present in *cis* conformation. However, the authors did not comment on the ring puckering of the single proline in the native protein [121], but in general, in the *cis* conformation proline *endo* puckering is highly abundant [153]. Similarly, the incorporation of (*4R*)F-Pro into ubiquitin with a single proline residue in a *trans* peptide bond and *exo* puckering increased the protein stability compared to the wild type [123]. The selective fluorination of the *endo* puckered Pro37 in the human Pin1 WW domain, a compact triple stranded antiparallel β-sheet protein, yielded a (*4S*)F-Pro variant that was more stable than the wild type [122].

Contradicting the trend, (*4R*)F-Pro substitution of the single proline residue adopting a native *exo* pucker impaired the folding and stability of a heat shock protein [164]. For an *E. coli* thioredoxin mutant, Trx1P, replacement of its sole proline residue by either (*4S*)F- or (*4R*)F-Pro resulted in the same stabilization pattern for both stereoisomers: stabilization of its reduced form and destabilization of its oxidized form [165]. These studies exemplify that the biased puckering of the F-Pro is only beneficial if other structural perturbations do not outweigh the positive effect.

This consideration becomes even more relevant with regard to variant proteins with more than one F-Pro residue. Nonetheless, the incorporation of fluorinated prolines is often accompanied by enhanced stability compared to the non-fluorinated protein [26, 27, 29, 125].

For most variant proteins with more than one F-Pro residue, only incorporation of one of the two F-Pro stereoisomers led to functional expression. For the large fragment of *Taq* DNA Polymerase (KlenTaq) only the (*4R*)F-Pro variant resulted in soluble protein. Remarkably, in the (*4R*)F-Pro variant of KlenTaq the preferred *exo* puckering of (*4R*)F-Pro was sacrificed at two positions (out of 32) in order to enable the endogenous *cis* isomerization of the corresponding peptidyl-proline bonds [27].

For F-Pro incorporation into EGFP only the (*4S*)F-Pro variant was soluble and showed enhanced folding characteristics [26]. The *trans* peptidyl-proline bonds are dominant in EGFP (9 from 10) but the resolution of the 3D structure of EGFP does not allow unambiguous assignment of the native proline ring puckers. In the fluorinated EGFP variant, however, nine out of the ten incorporated (*4S*)F-Pro residues adopted the *endo* pucker. The one *exo* pucker, the disfavored puckering form of the (*4S*)F-Pro stereoisomer, might be essential to allow proper protein folding. The biased ring puckering contributed to the folding of the protein variant. It remains to be determined if proline puckering in the naive protein is different to the (*4S*)F-Pro variant [26].

For the Phe63Ala mutant of the mRFP1 red fluorescent protein, incorporation of (*4S*)F-Pro led to insoluble protein whereas the (*4R*)F-Pro variant showed not only faster maturation than the non-fluorinated protein but also increased stability. How these observations are related to the bias for *exo* proline ring puckering introduced by (*4R*)F-Pro was not discussed [29].

F-Pro variants of EGFP as well as of KlenTaq showed enhanced crystallization characteristics. The fluorine substituent led to novel intramolecular interactions, but the overall structural fold of the fluorinated variant proteins was highly similar to their unmodified parents [26, 27].

The incorporation of (*4R*)F-Pro as well as of (*4S*)F-Pro into a lipase of thermophilic origin yielded two variant proteins, both with enhanced resistance towards organic solvents [28, 126]. Unfortunately, no crystal structures are available to examine F-Pro incorporation in this lipase in a structural context.

Although studies on F-Pro variant proteins are still relatively rare, they show that F-Pro incorporation strongly influences protein stability, yet the effects are apparently context dependent. Often, very high resolution 3D structures of fluorinated proteins and their non-fluorinated parent proteins are not available and, therefore, proline ring puckering remains elusive. Currently, we are unable to predict how to improve protein folding and stability using F-Pro. This calls for the systematic development of reliable prediction tools.

1.6.1.3 **Incorporation of perfluorinated amino acid analogs**

For perfluorinated proteins the term 'Teflon proteins' emerged, inspired by the extraordinary characteristics of this fluorocarbon polymer [166]. The nonstick properties of Teflon® (polytetrafluoroethylene) originate from a phenomenon known as the fluorous effect: It describes the tendency of perfluoroalkyl molecules to preferentially interact among each other rather than with hydrophilic or lipophilic molecules [167]. A number of studies aimed at exploiting this 'fluorophilicity' for the design of highly stable peptides and proteins [166, 168]. Because the CF_3 group is twice as hydrophobic as a CH_3 group, the incorporation of PFAAs in proteins increases their hydrophobicity [116]. This raised the question whether the 'fluorous-fluorous' interactions or the increased hydrophobic volume introduced upon PFAA incorporation causes the observed enhancements in protein stabilities. A study by Buer *et al.* [169] examining *de novo* designed proteins of varying fluorination degree argues for the hydrophobic volume as the stabilizing effect.

Protein perfluorination has focused on the incorporation of perfluorinated valine, leucine and isoleucine residues in the hydrophobic core of proteins [168]. While monofluorinated amino acids are structurally well tolerated, PFAAs tend to cause structural perturbations in proteins especially if a higher number of residues are substituted [116]. Mostly, small α-helical proteins are used as model proteins for PFAA incorporation [168]. Presumably, this owes to the intolerance of (larger) protein structures to accommodate PFAAs.

To avoid the negative or even deleterious effect often observed upon residue-specific incorporation of PFAAs, several studies sought to reduce the steric bulkiness in the hydrophobic protein core by reducing the number of the incorporated PFAAs [170-172]. At the bottom line, these studies corroborate the notion that the structures of most native proteins cannot accommodate too many PFAAs because of their highly hydrophobic character and their steric bulkiness. Nevertheless, directed evolution of proteins was successfully applied to optimize the protein scaffolds for the incorporation of trifluoroleucine [173, 174].

In the future, *de novo* design of proteins might render highly-fluorinated protein biosynthesis an attractive strategy [116]. Such protein scaffolds optimized to accommodate PFAAs could allow scientists to exploit the PFAAs' special chemical characteristics for protein engineering.

1.7 **Conclusion**

Fluorine is an outstanding element: The element with the highest electronegativity of all, which is especially attractive when tamed in a C-F bond. The C-F bond is extremely strong but its chemical formation only occurs under harsh conditions. Currently, only one enzyme class is known to cope with the peculiarities of fluoride ions. It efficiently forms C-F bonds between inorganic fluoride and an organic carbon. Most of these fluorinase enzymes seem to be part of a pathway producing 4-fluoro-L-threonine. This fluoro amino acid is one of only three fluorometabolites of microbial origin discovered so far.

Even though scarcely found in nature, organofluorine compounds are of high interest in biotechnological applications, especially in medicinal chemistry. Particularly, fluoro amino acids find use as agents for positron emission tomography due to their metabolic stability arising from fluorination.

Presently, 4-fluoro-L-threonine is the only biosynthesized fluoro amino acid, yet it is unclear whether it is incorporated into natural proteins. On the contrary, a variety of fluoro amino acids have been applied as tools for protein engineering. Established incorporation protocols facilitate their introduction into target proteins in a residue- as well as site-specific manner. It is even feasible to incorporate different fluorinated analogs at the same time. Global fluorination affects the physico-chemical properties of proteins, most notably the stability, the fluorescence characteristics, as well as the catalytic activity of enzymes. Monofluorinated amino acids tend to preserve the overall protein structure owing to the small size of fluorine compared to other functional groups as substituents. Polyfluorination could provoke extraordinary artificial protein properties such as the 'fluorous' effect associated with so-called putative 'Teflon proteins'. However, high-level incorporation of perfluoro amino acids often adversely affects the protein scaffold. To harness the full potential of amino acids with a high fluorination degree in proteins, the engineering of protein scaffolds or their *de novo* design to accommodate these analogs seems essential.

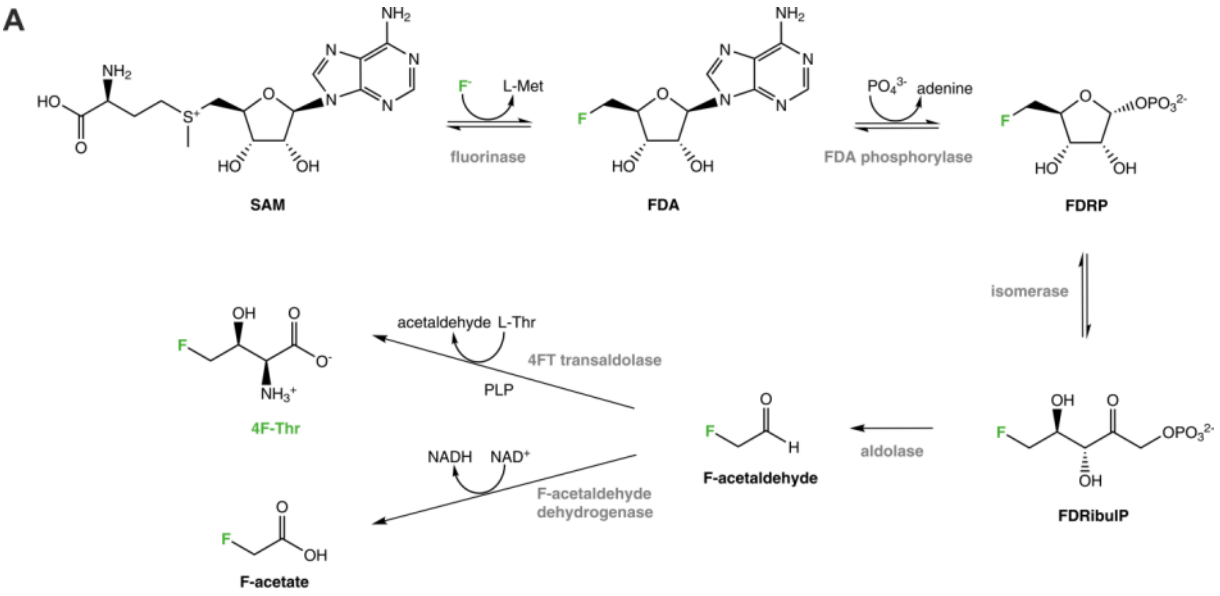
More detailed biochemical studies and high resolution structures of fluorinated proteins in combination with powerful molecular modeling approaches can pave the way to predict the structural and functional effects of translational protein fluorination. As the incorporation techniques improve, fluoro amino acids will eventually arrive as another useful off-the shelf accessory in the protein engineer's toolbox.

1.8 **List of abbreviations**

aaRS: aminoacyl-tRNA synthetase; EGFP: enhanced green fluorescent protein; FAA: fluoro amino acid; FaAA: monofluorinated aromatic amino acid; F-acetaldehyde: fluoroacetaldehyde; F-acetate: fluoroacetate; FDA: 5'-deoxy-5'-fluoroadenosine; FDRibulP: 5-deoxy-5-fluororibulosephosphate; FDRP: 5-deoxy-5-fluororibose phosphate; F-Pro: fluoro-L-proline; F-Thr: fluoro-L-threonine; F-Trp: fluoro-L-tryptophan; F-Tyr: fluoro-L-tyrosine; GFP: green fluorescent protein; GST: glutathione S-transferase; KlenTaq: large fragment of Taq DNA polymerase; Mri: methylthioribose-1-phosphate isomerase; PET: positron emission tomography; PFAA: perfluoro amino acid; PheRS: phenylalanyl-tRNA synthetase; (*4R*)OH-Pro: (*2S*:*4R*)-4-hydroxyproline; (*4R*)F-Pro: (*2S*:*4R*)-4-fluoroproline; SAH: S-adenosyl-L-homocysteine; SAM: S-adenosyl-L-methionine; (*4S*)F-Pro: (*2S*:*4S*)-4-fluoroproline; $tRNA^{aa}_{anticodon}$: tRNA with the appropriate anticodon for a cognate amino acid aa; TyrRS: tyrosyl-tRNA synthetase; X_{aa}: Y_{aa} amino acid at position X or Y

1.9 **Figures and illustrations**

A



B

ORF	Function - Homology/Similarity	(Proposed) function
flE	DNA binding protein	Transcriptional control
flD	Dehalogenase/phosphatase	(Defluorination as a resistance mechanism)
flC	Transmembrane transporter: MFS permease	(Transmembrane transport of F-acetate and 4F-Thr)
flB	FDA phosphorylase (Fig. 1A)	Fluorometabolite biosynthesis
flA	Fluorinase (Fig. 1A)	Fluorometabolite biosynthesis
flF	DNA binding protein	Transcriptional control
flG	DNA binding protein	Transcriptional control
flH	Transmembrane transporter: Na^+/H^+ antiporter	(Facilitation of intracellular F^- accumulation)
flI	Adenosylhomocysteine hydrolase	Prevention of FlA homocysteine inhibition
flJ	Protein of unknown function	-
flK	Fluoroacetyl-CoA thioesterase	Specific hydrolysis of fluoroacetyl-CoA
flL	DNA binding protein	Transcriptional control

Figure 1:

(A) Biosynthesis pathway for 4-fluoro-L-threonine and fluoroacetate in *S. cattleya*.

SAM, S-adenosyl-L-methionine; FDA, 5'-deoxy-5'-fluoroadenosine; FDRP, 5-deoxy-5-fluororibose phosphate; FDRibulP, 5-deoxy-5-fluororibulosephosphate; F-acetaldehyde, fluoroacetaldehyde; F-acetate, fluoroacetate; 4F-Thr, 4-fluoro-L-threonine. Refer to the text for details.

(B) Open reading frames (ORFs) in the *fl* cluster of *S. cattleya* and their proposed functions [53]. A graphical representation of the *fl* cluster is also shown.

The fluorinase (encoded by *flA*) and the FDA phosphorylase (encoded by *flB*) (shown in grey) are the only enzymes of the *fl* cluster that directly participate in the 4F-Thr/F-acetate biosynthesis pathway.

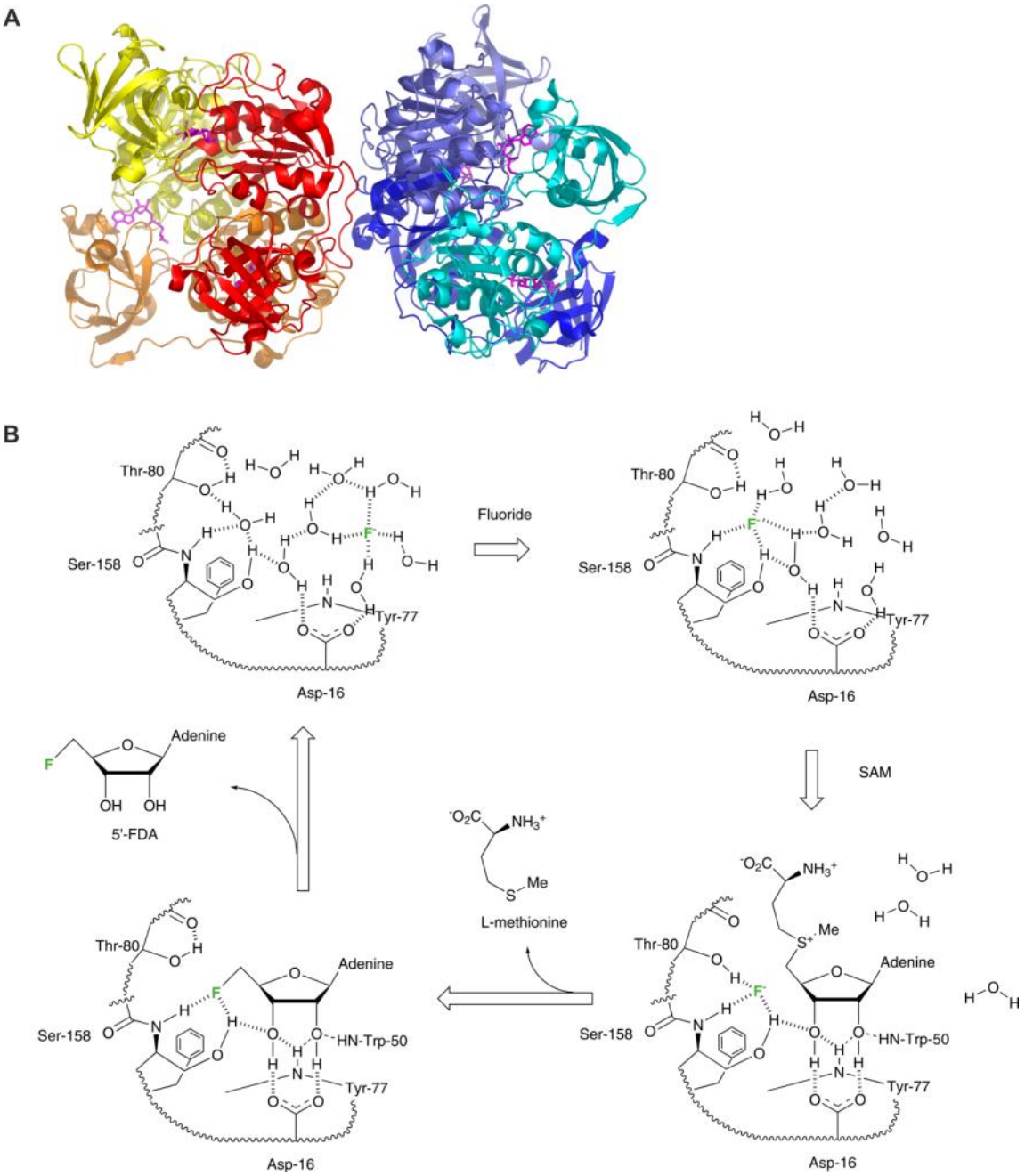


Figure 2:

(A) The crystal structure of the fluorinase from *S. cattleya* in complex with S-adenosyl-L-methionine.

The structure of the fluorinase is composed of six identical subunits that form a dimer of trimers (PDB ID: 1RQP, [52]). The subunits of one trimer are shown in colors of red while the subunits of the other trimer are depicted in blue colors. Three S-adenosyl-L-methionine (SAM) molecules (magenta) are bound by each trimer. The cartoon was generated using PyMOL version 0.96 by DeLano Scientific (www.pymol.sourceforge.net).

(B) Proposed sequential mechanism for the F–C bond formation catalyzed by the fluorinase from *S. cattleya*

Reprinted with permission from Zhu, X., Robinson, D. A., McEwan, A. R., O'Hagan, D., Naismith, J. H., Mechanism of enzymatic fluorination in Streptomyces cattleya. *J. Am. Chem. Soc.* 2007, *129*, 14597-14604. Copyright 2007 American Chemical Society.

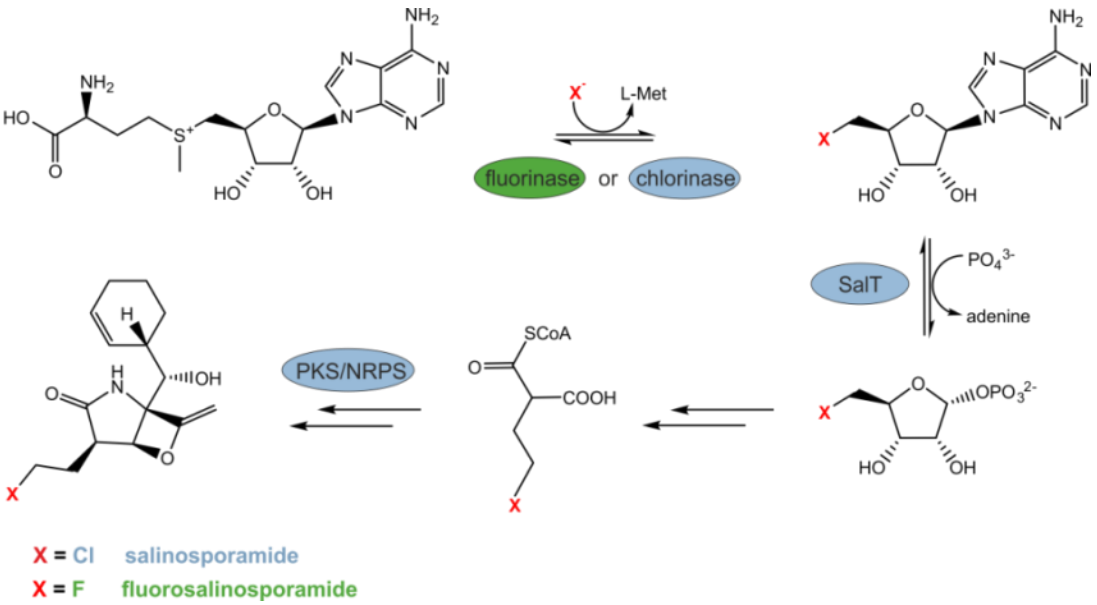


Figure 3: Pathway in *S. tropica* leading to salinosporamide (X=Cl) or fluorosalinosporamide (X=F).

Polyketide synthase (PKS), non-ribosomal peptide synthase (NRPS) (adopted from [80]).

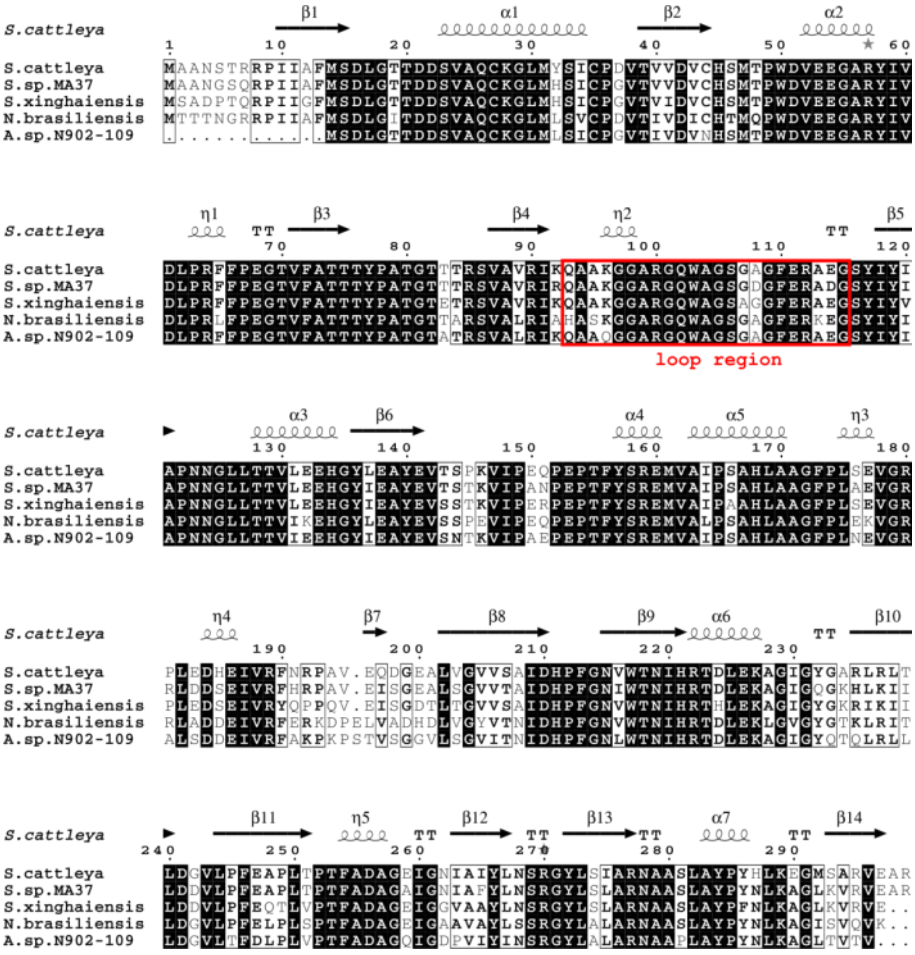


Figure 4: Protein sequence alignment of the fluorinase from *S. cattleya* and its homologs in *Streptomyces sp.* MA37, *S. xinghaiensis*, *N. brasiliensis* and *Actinoplanes sp.* N902-109.

The characteristic loop found in the fluorinase and its homologs is highlighted by a red rectangle. NCBI accession numbers of the fluorinase homologs are [WP_014144878.1] for *S. cattleya*, [CDH39444.1] for *Streptomyces sp.* MA37, [CDP39161.1] for *S. xinghaiensis*, [AHK61118.1] for *N. brasiliensis* and [WP_015619887.1] for *Actinoplanes sp.* N902-109. The secondary structure of the fluorinase from *S. cattleya* (PDB code: 2V7V) is shown at the top. α-helices (α) and 3/10-helices (η) are shown as squiggles, β-strands (β) as arrows, strict β-turns as TT letters. Created in ESPript [175].

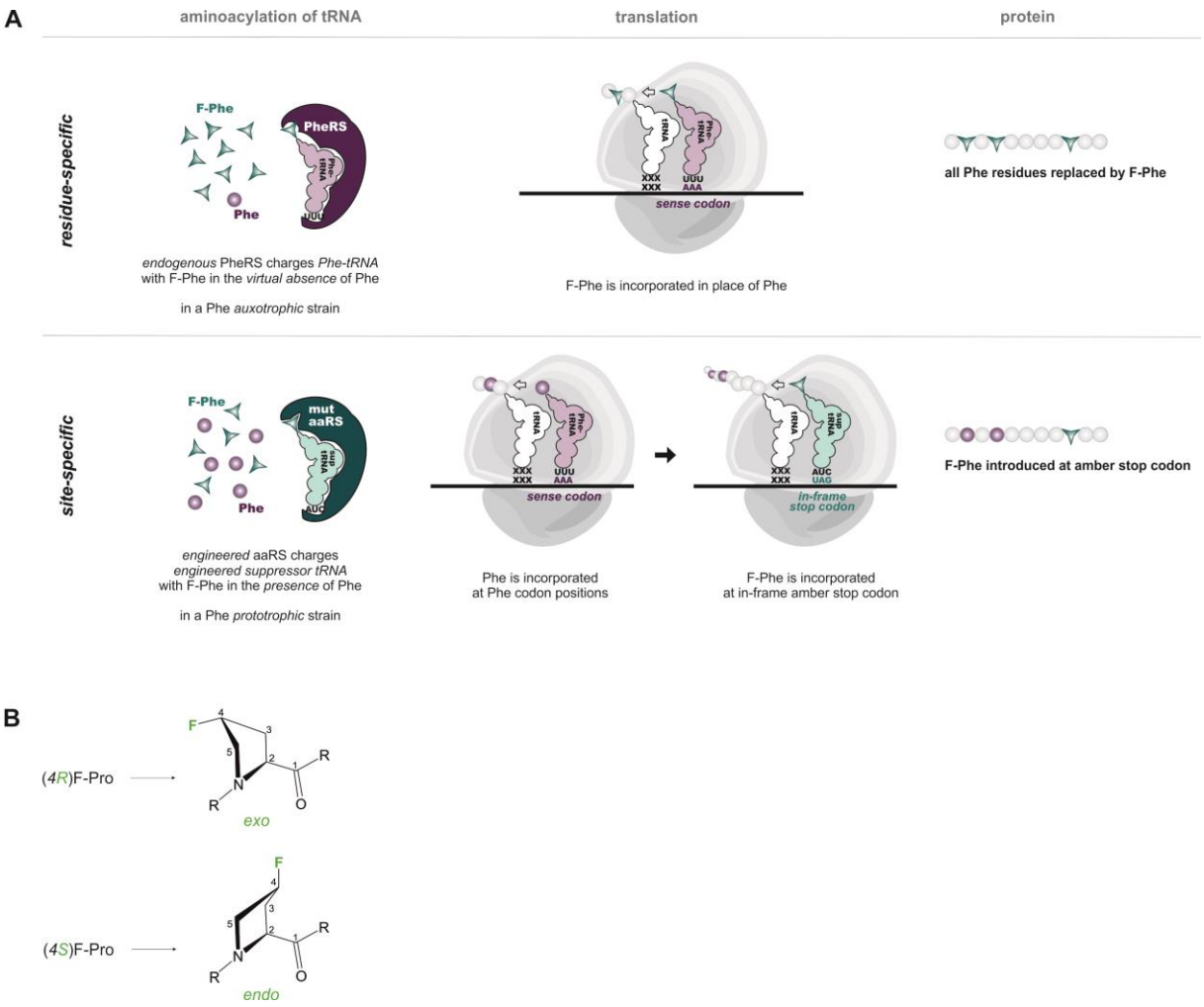


Figure 5:

(A) Schematic representation of the residue- and site-specific incorporation of non-canonical amino acids as exemplified for a fluorophenylalanine analog.

For the residue-specific incorporation of FAAs, it is mandatory that the corresponding canonical amino acid, here Phe, is present only in negligible amounts, as represented by the single Phe molecule. This minimizes competition and ensures efficient charging of the tRNA with the fluorophenylalanine analog (F-Phe) by the phenylalanyl-tRNA synthetase (PheRS). In contrast, the abundance of Phe does not interfere with the site-specific incorporation of F-Phe. An F-Phe specific mutant aminoacyl-tRNA synthetase (mut aaRS) facilitates efficient charging of the suppressor tRNA (sup tRNA) with the fluorinated analog.

(B) The stereochemistry of the fluorine substituent at the C4 carbon of proline (*4R vs 4S*) affects the pucker of the pyrrolidine ring (*exo vs endo*).

1.10 **References**

[1] Zhang, X.-J., Lai, T.-B., Kong, R. Y.-C., Biology of fluoro-organic compounds, in: Horváth, I. T. (Ed.), *Fluorous Chemistry*, Springer Berlin Heidelberg 2012, pp. 365-404.

[2] Pauling, L.*The nature of the chemical bond and the structure of molecules and crystals: an introduction to modern structural chemistry*, Cornell University Press, Ithaca 1960.

[3] O'Hagan, D., Understanding organofluorine chemistry. An introduction to the C-F bond. *Chem. Soc. Rev.* 2008, *37*, 308-319.

[4] Bondi, A., Van der Waals volumes and radii. *J. Phys. Chem.* 1964, *68*, 441-451.

[5] Purser, S., Moore, P. R., Swallow, S., Gouverneur, V., Fluorine in medicinal chemistry. *Chem. Soc. Rev.* 2008, *37*, 320-330.

[6] O'Hagan, D., Fluorine in health care: Organofluorine containing blockbuster drugs. *J. Fluor. Chem.* 2010, *131*, 1071-1081.

[7] Bégué, J.-P., Bonnet-Delpon, D., Recent advances (1995–2005) in fluorinated pharmaceuticals based on natural products. *J. Fluor. Chem.* 2006, *127*, 992-1012.

[8] Bohm, H. J., Banner, D., Bendels, S., Kansy, M., *et al.*, Fluorine in medicinal chemistry. *Chembiochem* 2004, *5*, 637-643.

[9] Hagmann, W. K., The many roles for fluorine in medicinal chemistry. *J. Med. Chem.* 2008, *51*, 4359-4369.

[10] Kirk, K. L., Fluorine in medicinal chemistry: Recent therapeutic applications of fluorinated small molecules. *J. Fluor. Chem.* 2006, *127*, 1013-1029.

[11] Müller, K., Faeh, C., Diederich, F., Fluorine in pharmaceuticals: looking beyond intuition. *Science* 2007, *317*, 1881-1886.

[12] Shah, P., Westwell, A. D., The role of fluorine in medicinal chemistry. *J. Enzyme Inhib. Med. Chem.* 2007, *22*, 527-540.

[13] Ojima, I.*Fluorine in medicinal chemistry and chemical biology*, John Wiley & Son, Chichester 2009.

[14] Tressaud, A., Haufe, G.*Fluorine and health: molecular imaging, biomedical materials and pharmaceuticals*, Elsevier Science 2008.

[15] Berkowitz, D. B., Karukurichi, K. R., de la Salud-Bea, R., Nelson, D. L., *et al.*, Use of fluorinated functionality in enzyme inhibitor development: mechanistic and analytical advantages. *J. Fluor. Chem.* 2008, *129*, 731-742.

[16] Bourdier, T., Shepherd, R., Berghofer, P., Jackson, T., *et al.*, Radiosynthesis and biological evaluation of l- and d-S-(3-[18F]fluoropropyl)homocysteine for tumor imaging using positron emission tomography. *J. Med. Chem.* 2011, *54*, 1860-1870.

[17] Krasikova, R. N., Kuznetsova, O. F., Fedorova, O. S., Belokon, Y. N., *et al.*, 4-[18F]fluoroglutamic acid (BAY 85-8050), a new amino acid radiotracer for PET imaging of tumors: synthesis and in vitro characterization. *J. Med. Chem.* 2010, *54*, 406-410.

[18] McConathy, J., Goodman, M., Non-natural amino acids for tumor imaging using positron emission tomography and single photon emission computed tomography. *Cancer Metastasis Rev.* 2008, *27*, 555-573.

[19] Qu, W., Zha, Z., Ploessl, K., Lieberman, B. P., *et al.*, Synthesis of optically pure 4-fluoro-glutamines as potential metabolic imaging agents for tumors. *J. Am. Chem. Soc.* 2010, *133*, 1122-1133.

[20] Wang, L., Lieberman, B. P., Plössl, K., Qu, W., *et al.*, Synthesis and comparative biological evaluation of L- and D-isomers of 18F-labeled fluoroalkyl phenylalanine derivatives as tumor imaging agents. *Nucl. Med. Biol.* 2011, *38*, 301-312.

[21] Wang, L., Qu, W., Lieberman, B., Ploessl, K., *et al.*, Synthesis and in vitro evaluation of 18F labeled tyrosine derivatives as potential positron emission tomography (PET) imaging agents. *Bioorg. Med. Chem. Lett.* 2010, *20*, 3482-3485.

[22] Ayyadurai, N., Prabhu, N., Deepankumar, K., Kim, A., *et al.*, Biosynthetic substitution of tyrosine in green fluorescent protein with its surrogate fluorotyrosine in Escherichia coli. *Biotechnol. Lett.* 2011, *33*, 2201-2207.

[23] Dominguez, M. A., Thornton, K. C., Melendez, M. G., Dupureur, C. M., Differential effects of isomeric incorporation of fluorophenylalanines into PvuII endonuclease. *Proteins* 2001, *45*, 55-61.

[24] Baker, P. J., Montclare, J. K., Enhanced refoldability and thermoactivity of fluorinated phosphotriesterase. *Chembiochem* 2011, *12*, 1845-1848.

[25] Merkel, L., Schauer, M., Antranikian, G., Budisa, N., Parallel incorporation of different fluorinated amino acids: on the way to "teflon" proteins. *Chembiochem* 2010, *11*, 1505-1507.

[26] Steiner, T., Hess, P., Bae, J. H., Wiltschi, B., *et al.*, Synthetic biology of proteins: tuning GFPs folding and stability with fluoroproline. *PLoS ONE* 2008, *3*, e1680.

[27] Holzberger, B., Obeid, S., Welte, W., Diederichs, K., *et al.*, Structural insights into the potential of 4-fluoroproline to modulate biophysical properties of proteins. *Chem. Sci.* 2012, *3*, 2924-2931.

[28] Acevedo-Rocha, C. G., Hoesl, M. G., Nehring, S., Royter, M., *et al.*, Non-canonical amino acids as a useful synthetic biological tool for lipase-catalysed reactions in hostile environments. *Catal. Sci. Technol.* 2013, *3*, 1198-1201.

[29] Deepankumar, K., Nadarajan, S. P., Ayyadurai, N., Yun, H., Enhancing the biophysical properties of mRFP1 through incorporation of fluoroproline. *Biochem. Biophys. Res. Commun.* 2013, *440*, 509-514.

[30] Parsons, J. F., Xiao, G., Gilliland, G. L., Armstrong, R. N., Enzymes harboring unnatural amino acids: mechanistic and structural analysis of the enhanced catalytic activity of a glutathione transferase containing 5-fluorotryptophan. *Biochemistry* 1998, *37*, 6286-6294.

[31] Murphy, C. D., Schaffrath, C., O'Hagan, D., Fluorinated natural products: the biosynthesis of fluoroacetate and 4-fluorothreonine in Streptomyces cattleya. *Chemosphere* 2003, *52*, 455-461.

[32] Gribble, G. W., Naturally occurring organofluorines, in: Neilson, A. H. (Ed.), *Organofluorines*, Springer, Berlin Heidelberg 2002, pp. 121-136.

[33] Deng, H., Ma, L., Bandaranayaka, N., Qin, Z., *et al.*, Identification of fluorinases from Streptomyces sp MA37, Norcardia brasiliensis, and Actinoplanes sp N902-109 by genome mining. *Chembiochem* 2014, *15*, 364-368.

[34] Harper, D., O'Hagan, D., Murphy, C., Fluorinated natural products: occurrence and biosynthesis, in: Gribble, G. (Ed.), *Natural Production of Organohalogen Compounds*, Springer Berlin Heidelberg 2003, pp. 141-169.

[35] Wang, Y., Deng, Z., Qu, X., Characterization of a SAM-dependent fluorinase from a latent biosynthetic pathway for fluoroacetate and 4-fluorothreonine formation in Nocardia brasiliensis. *F1000Res.* 2014, *3*, 61.

[36] Xu, X.-H., Yao, G.-M., Li, Y.-M., Lu, J.-H., *et al.*, 5-Fluorouracil derivatives from the sponge Phakellia fusca. *J. Nat. Prod.* 2003, *66*, 285-288.

[37] Baunthiyal, M., Pandey, A., Organofluorine metabolism in plants. *Fluoride* 2012, *42*, 78-85.

[38] O'Hagan, D., Deng, H., Enzymatic fluorination and biotechnological developments of the fluorinase. *Chem. Rev.* 2014, DOI 10.1021/cr500209t.

[39] Harper, D. B., O'Hagan, D., The fluorinated natural products. *Nat. Prod. Rep.* 1994, *11*, 123-133.

[40] Hewitt, R., Gumble, A., Taylor, L., Wallace, W., The activity of a new antibiotic, nucleocidin in experimental infections with Trypanosoma equiperdum. *Antibiot. Annu.* 1957, 722-729.

[41] Thomas, S., Singleton, V., Lowery, J., Sharpe, R., *et al.*, Nucleocidin, a new antibiotic with activity against trypanosomes. *Antibiot. Annu.* 1957, 716-721.

[42] Sanada, M., Miyano, T., Iwadare, S., Williamson, J., *et al.*, Biosynthesis of fluorothreonine and fluoroacetic acid by the thienamycin producer, Streptomyces cattleya. *J. Antibiot.* 1986, *39*, 259-265.

[43] Reid, K. A., Bowden, R. D., Dasaradhi, L., Amin, M. R., *et al.*, Biosynthesis of fluorinated secondary metabolites by Streptomyces cattleya. *Microbiology* 1995, *141*, 1385-1393.

[44] Huang, S., Ma, L., Tong, M. H., Yu, Y., *et al.*, Fluoroacetate biosynthesis from the marine-derived bacterium Streptomyces xinghaiensis NRRL B-24674. *Org. Biomol. Chem.* 2014.

[45] Nass, G., Poralla, K., Zahner, H., Biogenetic amino acid antagonists. *Naturwissenschaften* 1971, *58*, 603-610.
[46] Rabinovitz, M., Finkleman, A., Reagan, R. L., Breitman, T. R., Amino acid antagonist death in Escherichia coli. *J. Bacteriol.* 1969, *99*, 336-338.
[47] Deng, H., O'Hagan, D., Schaffrath, C., Fluorometabolite biosynthesis and the fluorinase from Streptomyces cattleya. *Nat. Prod. Rep.* 2004, *21*, 773-784.
[48] Deng, H., Cross, S. M., McGlinchey, R. P., Hamilton, J. T. G., *et al.*, In vitro reconstituted biotransformation of 4-fluorothreonine from fluoride ion: application of the fluorinase. *Chem. Biol.* 2008, *15*, 1268-1276.
[49] Cobb, S. L., Deng, H., Hamilton, J. T. G., McGlinchey, R. P., *et al.*, Identification of 5-fluoro-5-deoxy-D-ribose-1-phosphate as an intermediate in fluorometabolite biosynthesis in Streptomyces cattleya. *ChemInform* 2004, *35*, 592-593.
[50] Cross, S. M., University of St Andrews, Chemistry Theses 2009.
[51] Dong, C., Deng, H., Dorward, M., Schaffrath, C., *et al.*, Crystallization and X-ray diffraction of 5'-fluoro-5'-deoxyadenosine synthase, a fluorination enzyme from Streptomyces cattleya. *Acta Crystallogr. D Biol. Crystallogr.* 2003, *59*, 2292-2293.
[52] Dong, C., Huang, F., Deng, H., Schaffrath, C., *et al.*, Crystal structure and mechanism of a bacterial fluorinating enzyme. *Nature* 2004, *427*, 561-565.
[53] Huang, F., Haydock, S. F., Spiteller, D., Mironenko, T., *et al.*, The gene cluster for fluorometabolite biosynthesis in Streptomyces cattleya: a thioesterase confers resistance to fluoroacetyl-coenzyme A. *Chem. Biol.* 2006, *13*, 475-484.
[54] Moss, S. J., Murphy, C. D., Hamilton, J. T. G., McRoberts, W. C., *et al.*, Fluoroacetaldehyde: a precursor of both fluoroacetate and 4-fluorothreonine in Streptomyces cattleya. *Chem. Commun.* 2000, 2281-2282.
[55] Murphy, C. D., Moss, S. J., O'Hagan, D., Isolation of an aldehyde dehydrogenase involved in the oxidation of fluoroacetaldehyde to fluoroacetate in Streptomyces cattleya. *Appl. Environ. Microbiol.* 2001, *67*, 4919-4921.
[56] Murphy, C. D., O'Hagan, D., Schaffrath, C., Identification of a PLP-dependent threonine transaldolase: a novel enzyme involved in 4-fluorothreonine biosynthesis in Streptomyces cattleya. *Angew. Chem. Int. Ed. Engl.* 2001, *40*, 4479-4481.
[57] O'Hagan, D., Schaffrath, C., Cobb, S. L., Hamilton, J. T. G., *et al.*, Biochemistry: biosynthesis of an organofluorine molecule. *Nature* 2002, *416*, 279.
[58] Onega, M., McGlinchey, R. P., Deng, H., Hamilton, J. T. G., *et al.*, The identification of (3R,4S)-5-fluoro-5-deoxy-d-ribulose-1-phosphate as an intermediate in fluorometabolite biosynthesis in Streptomyces cattleya. *Bioorg. Chem.* 2007, *35*, 375-385.
[59] O'Hagan, D., Recent developments on the fluorinase from Streptomyces cattleya. *J. Fluor. Chem.* 2006, *127*, 1479-1483.
[60] Barbe, V., Bouzon, M., Mangenot, S., Badet, B., *et al.*, Complete genome sequence of Streptomyces cattleya NRRL 8057, a producer of antibiotics and fluorometabolites. *J. Bacteriol.* 2011, *193*, 5055-5056.
[61] Zhao, C., Li, P., Deng, Z., Ou, H.-Y., *et al.*, Insights into fluorometabolite biosynthesis in Streptomyces cattleya DSM46488 through genome sequence and knockout mutants. *Bioorg. Chem.* 2012, *44*, 1-7.
[62] Walker, M. C., Wen, M., Weeks, A. M., Chang, M. C. Y., Temporal and fluoride control of secondary metabolism regulates cellular organofluorine biosynthesis. *ACS Chem. Biol.* 2012, *7*, 1576-1585.
[63] Chan, K. K. J., O'Hagan, D., The rare fluorinated natural products and biotechnological prospects for fluorine enzymology, in: David, A. H. (Ed.), *Meth. Enzymol.*, Academic Press 2012, pp. 219-235.
[64] Goncharov, N. V., Jenkins, R. O., Radilov, A. S., Toxicology of fluoroacetate: a review, with possible directions for therapy research. *J. Appl. Toxicol.* 2006, *26*, 148-161.
[65] Weeks, A. M., Coyle, S. M., Jinek, M., Doudna, J. A., *et al.*, Structural and biochemical studies of a fluoroacetyl-CoA-specific thioesterase reveal a molecular basis for fluorine selectivity. *Biochemistry* 2010, *49*, 9269-9279.
[66] Blasiak, L. C., Drennan, C. L., Structural perspective on enzymatic halogenation. *Acc. Chem. Res.* 2008, *42*, 147-155.
[67] Vaillancourt, F. H., Yeh, E., Vosburg, D. A., Garneau-Tsodikova, S., *et al.*, Nature's inventory of halogenation catalysts: oxidative strategies predominate. *Chem. Rev.* 2006, *106*, 3364-3378.
[68] Lide, D. R.*CRC Handbook of Chemistry and Physics*, CRC Press, Boston 1990-1991.
[69] Zechel, D., Reid, S., Nashiru, O., Mayer, C., *et al.*, Enzymatic synthesis of carbon-fluorine bonds. *J. Am. Chem. Soc.* 2001, *123*, 4350-4351.
[70] Eustaquio, A. S., Pojer, F., Noel, J. P., Moore, B. S., Discovery and characterization of a marine bacterial SAM-dependent chlorinase. *Nat. Chem. Biol.* 2008, *4*, 69-74.
[71] Perrin, C. L., Rodgers, B. L., O'Connor, J. M., Nucleophilic addition to a p-benzyne derived from an enediyne: a new mechanism for halide incorporation into biomolecules. *J. Am. Chem. Soc.* 2007, *129*, 4795-4799.
[72] Wuosmaa, A. M., Hager, L. P., Methyl chloride transferase: a carbocation route for biosynthesis of halometabolites. *Science* 1990, *249*, 160-162.
[73] Nashiru, O., Zechel, D. L., Stoll, D., Mohammadzadeh, T., *et al.*, β-mannosynthase: synthesis of β-mannosides with a mutant β-mannosidase. *Angew. Chem. Int. Ed. Engl.* 2001, *40*, 417-420.
[74] O'Hagan, D., Goss, R. J. M., Meddour, A., Courtieu, J., Assay for the enantiomeric analysis of [2H1]-fluoroacetic acid: insight into the stereochemical course of fluorination during fluorometabolite biosynthesis in streptomyces cattleya. *J. Am. Chem. Soc.* 2003, *125*, 379-387.
[75] Cadicamo, C. D., Courtieu, J., Deng, H., Meddour, A., *et al.*, Enzymatic fluorination in Streptomyces cattleya takes place with an inversion of configuration consistent with an SN2 reaction mechanism. *Chembiochem* 2004, *5*, 685-690.
[76] Vincent, M. A., Hillier, I. H., The solvated fluoride anion can be a good nucleophile. *Chem. Commun.* 2005, 5902-5903.
[77] Senn, H. M., O'Hagan, D., Thiel, W., Insight into enzymatic C−F bond formation from QM and QM/MM calculations. *J. Am. Chem. Soc.* 2005, *127*, 13643-13655.
[78] Zhu, X., Robinson, D. A., McEwan, A. R., O'Hagan, D., *et al.*, Mechanism of enzymatic fluorination in Streptomyces cattleya. *J. Am. Chem. Soc.* 2007, *129*, 14597-14604.
[79] Lohman, D. C., Edwards, D. R., Wolfenden, R., Catalysis by desolvation: the catalytic prowess of SAM-dependent halide-alkylating enzymes. *J. Am. Chem. Soc.* 2013, *135*, 14473-14475.
[80] Eustáquio, A. S., O'Hagan, D., Moore, B. S., Engineering fluorometabolite production: fluorinase expression in Salinispora tropica yields fluorosalinosporamide. *J. Nat. Prod.* 2010, *73*, 378-382.
[81] Deng, H., Cobb, S. L., McEwan, A. R., McGlinchey, R. P., *et al.*, The fluorinase from Streptomyces cattleya is also a chlorinase. *Angew. Chem. Int. Ed. Engl.* 2006, *45*, 759-762.
[82] Deng, H., O'Hagan, D., The fluorinase, the chlorinase and the duf-62 enzymes. *Curr. Opin. Chem. Biol.* 2008, *12*, 582-592.
[83] Walker, M. C., Chang, M. C. Y., Natural and engineered biosynthesis of fluorinated natural products. *Chem. Soc. Rev.* 2014, *43*, 6527-6536.
[84] Li, X.-G., Domarkas, J., O'Hagan, D., Fluorinase mediated chemoenzymatic synthesis of [18F]-fluoroacetate. *Chem. Commun.* 2010, *46*, 7819-7821.
[85] Cobb, S. L., Deng, H., McEwan, A. R., Naismith, J. H., *et al.*, Substrate specificity in enzymatic fluorination. The fluorinase from Streptomyces cattleya accepts 2'-deoxyadenosine substrates. *Org. Biomol. Chem.* 2006, *4*, 1458-1460.
[86] Deng, H., Cobb, S. L., Gee, A. D., Lockhart, A., *et al.*, Fluorinase mediated C-18F bond formation, an enzymatic tool for PET labelling. *Chem. Commun.* 2006, 652-654.
[87] Onega, M., Winkler, M., O'Hagan, D., Fluorinase: a tool for the synthesis of 18F-labeled sugars and nucleosides for PET. *Future Med. Chem.* 2009, *1*, 865-873.
[88] Winkler, M., Domarkas, J., Schweiger, L. F., O'Hagan, D., Fluorinase-coupled base swaps: synthesis of [18F]-5'-deoxy-5'-fluorouridines. *Angew. Chem. Int. Ed. Engl.* 2008, *47*, 10141-10143.

[89] Dall'Angelo, S., Bandaranayaka, N., Windhorst, A. D., Vugts, D. J., *et al.*, Tumour imaging by positron emission tomography using fluorinase generated 5-[18F]fluoro-5-deoxyribose as a novel tracer. *Nucl. Med. Biol.* 2013, *40*, 464-470.
[90] Iwai, N., Tanaka, T., Kitazume, T., Utility of ionic liquid for improvement of fluorination reaction with immobilized fluorinase. *J Mol Catal B Enzym* 2009, *59*, 131-133.
[91] Sergeev, M. E., Morgia, F., Javed, M. R., Doi, M., *et al.*, Polymer-immobilized fluorinase: recyclable catalyst for fluorination reactions. *J Mol Catal B Enzym* 2013, *92*, 51-56.
[92] Liu, C. C., Schultz, P. G., Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* 2010, *79*, 413-444.
[93] Johnson, D. B. F., Xu, J., Shen, Z., Takimoto, J. K., *et al.*, RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat. Chem. Biol.* 2011, *7*, 779-786.
[94] Antonczak, A. K., Morris, J., Tippmann, E. M., Advances in the mechanism and understanding of site-selective noncanonical amino acid incorporation. *Curr. Opin. Struct. Biol.* 2011, *21*, 481-487.
[95] Hoesl, M. G., Budisa, N., Recent advances in genetic code engineering in Escherichia coli. *Curr. Opin. Biotechnol.* 2012, *23*, 751-757.
[96] Chin, J. W., Reprogramming the genetic code. *EMBO J.* 2011, *30*, 2312-2324.
[97] Pandey, A. K., Naduthambi, D., Thomas, K. M., Zondlo, N. J., Proline editing: a general and practical approach to the synthesis of functionally and structurally diverse peptides. Analysis of steric versus stereoelectronic effects of 4-substituted prolines on conformation within peptides. *J. Am. Chem. Soc.* 2013, *135*, 4333-4363.
[98] Kimmerlin, T., Seebach, D., '100 years of peptide synthesis': ligation methods for peptide and protein synthesis with applications to β-peptide assemblies. *J. Pept. Res.* 2005, *65*, 229-260.
[99] Nishiuchi, Y., Inui, T., Nishio, H., Bódi, J., *et al.*, Chemical synthesis of the precursor molecule of the Aequorea green fluorescent protein, subsequent folding, and development of fluorescence. *Proc. Natl. Acad. Sci. USA* 1998, *95*, 13549-13554.
[100] Hackenberger, Christian P. R., Schwarzer, D., Chemoselective ligation and modification strategies for peptides and proteins. *Angew. Chem. Int. Ed. Engl.* 2008, *47*, 10030-10074.
[101] Munier, R., Cohen, G. N., Incorporation d'analogues structuraux d'aminoacides dans les protéines bactériennes au cours de leur synthèse in vivo. *Biochim. Biophys. Acta* 1959, *31*, 378-391.
[102] Wiltschi, B., Expressed protein modifications: Making synthetic proteins, in: Weber, W., Fussenegger, M. (Eds.), *Methods in Molecular Biology - Synthetic Gene Networks*, Humana Press 2012, pp. 211-225.
[103] Zheng, S., Kwon, I., Manipulation of enzyme properties by noncanonical amino acid incorporation. *Biotechnol. J.* 2012, *7*, 47-60.
[104] Merkel, L., Budisa, N., Organic fluorine as a polypeptide building element: in vivo expression of fluorinated peptides, proteins and proteomes. *Org. Biomol. Chem.* 2012, *10*, 7241-7261.
[105] Lajoie, M. J., Rovner, A. J., Goodman, D. B., Aerni, H.-R., *et al.*, Genomically recoded organisms expand biological functions. *Science* 2013, *342*, 357-360.
[106] Mukai, T., Yanagisawa, T., Ohtake, K., Wakamori, M., *et al.*, Genetic-code evolution for protein synthesis with non-natural amino acids. *Biochem. Biophys. Res. Commun.* 2011, *411*, 757-761.
[107] Furter, R., Expansion of the genetic code: site-directed p-fluoro-phenylalanine incorporation in Escherichia coli. *Protein Sci.* 1998, *7*, 419–426.
[108] Jackson, J. C., Hammill, J. T., Mehl, R. A., Site-specific incorporation of a (19)F-amino acid into proteins as an NMR probe for characterizing protein structure and reactivity. *J. Am. Chem. Soc.* 2007, *129*, 1160-1166.
[109] Tippmann, E. M., Liu, W., Summerer, D., Mack, A. V., *et al.*, A genetically encoded diazirine photocrosslinker in Escherichia coli. *Chembiochem* 2007, *8*, 2210-2214.
[110] Cellitti, S. E., Jones, D. H., Lagpacan, L., Hao, X., *et al.*, In vivo incorporation of unnatural amino acids to probe structure, dynamics, and ligand binding in a large protein by nuclear magnetic resonance spectroscopy. *J. Am. Chem. Soc.* 2008, *130*, 9268-9281.
[111] Stokes, A. L., Miyake-Stoner, S. J., Peeler, J. C., Nguyen, D. P., *et al.*, Enhancing the utility of unnatural amino acid synthetases by manipulating broad substrate specificity. *Mol. Biosyst.* 2009, *5*, 1032-1038.
[112] Miyake-Stoner, S. J., Refakis, C. A., Hammill, J. T., Lusic, H., *et al.*, Generating permissive site-specific unnatural aminoacyl-tRNA synthetases. *Biochemistry* 2010, *49*, 1667-1677.
[113] Young, D. D., Young, T. S., Jahnz, M., Ahmad, I., *et al.*, An evolved aminoacyl-tRNA synthetase with atypical polysubstrate specificity. *Biochemistry* 2011, *50*, 1894-1900.
[114] Minnihan, E. C., Young, D. D., Schultz, P. G., Stubbe, J., Incorporation of fluorotyrosines into ribonucleotide reductase using an evolved, polyspecific aminoacyl-tRNA synthetase. *J. Am. Chem. Soc.* 2011, *133*, 15942-15945.
[115] Wilkins, B. J., Marionni, S., Young, D. D., Liu, J., *et al.*, Site-specific incorporation of fluorotyrosines into proteins in Escherichia coli by photochemical disguise. *Biochemistry* 2010, *49*, 1557-1559.
[116] Budisa, N., Pipitone, O., Siwanowicz, I., Rubini, M., *et al.*, Efforts towards the design of 'teflon' proteins: in vivo translation with trifluorinated leucine and methionine analogues. *Chem. Biodivers.* 2004, *1*, 1465-1475.
[117] Tang, Y., Tirrell, D. A., Biosynthesis of a highly stable coiled-coil protein containing hexafluoroleucine in an engineered bacterial host. *J. Am. Chem. Soc.* 2001, *123*, 11089-11090.
[118] Yoo, T. H., Tirrell, D. A., High-throughput screening for methionyl-tRNA synthetases that enable residue-specific incorporation of noncanonical amino acids into recombinant proteins in bacterial cells. *Angew. Chem. Int. Ed. Engl.* 2007, *46*, 5340-5343.
[119] Salwiczek, M., Nyakatura, E. K., Gerling, U. I. M., Ye, S., *et al.*, Fluorinated amino acids: compatibility with native protein structures and effects on protein-protein interactions. *Chem. Soc. Rev.* 2012, *41*, 2135-2171.
[120] Biava, H., Budisa, N., Evolution of fluorinated enzymes: an emerging trend for biocatalyst stabilization. *Eng. Life. Sci.* 2014, *14*, 340–351.
[121] Renner, C., Alefelder, S., Bae, J. H., Budisa, N., *et al.*, Fluoroprolines as tools for protein design and engineering. *Angew. Chem. Int. Ed. Engl.* 2001, *40*, 923-925.
[122] Tang, H.-C., Lin, Y.-J., Horng, J.-C., Modulating the folding stability and ligand binding affinity of Pin1 WW domain by proline ring puckering. *Proteins* 2014, *82*, 67-76.
[123] Crespo, M. D., Rubini, M., Rational design of protein stability: effect of (2S,4R)-4-fluoroproline on the stability and folding pathway of ubiquitin. *PLoS ONE* 2011, *6*, e19425.
[124] Holzberger, B., Marx, A., Replacing 32 proline residues by a noncanonical amino acid results in a highly active DNA polymerase. *J. Am. Chem. Soc.* 2010, *132*, 15708-15713.
[125] Edwardraja, S., Sriram, S., Govindan, R., Budisa, N., *et al.*, Enhancing the thermal stability of a single-chain Fv fragment by in vivo global fluorination of the proline residues. *Mol. Biosyst.* 2011, *7*, 258-265.
[126] Hoesl, M. G., Acevedo-Rocha, C. G., Nehring, S., Royter, M., *et al.*, Lipase congeners designed by genetic code engineering. *ChemCatChem* 2011, *3*, 213-221.
[127] Eichler, J. F., Cramer, J. C., Kirk, K. L., Bann, J. G., Biosynthetic incorporation of fluorohistidine into proteins in E. coli: a new probe of macromolecular structure. *Chembiochem* 2005, *6*, 2170-2173.
[128] Hu, L., Joshi, S. B., Andra, K. K., Thakkar, S. V., *et al.*, Comparison of the structural stability and dynamic properties of recombinant anthrax protective antigen and its 2-fluorohistidine-labeled analogue. *J. Pharm. Sci.* 2012, *101*, 4118-4128.

[129] Jackson, D. Y., Burnier, J., Quan, C., Stanley, M., *et al.*, A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. *Science* 1994, *266*, 243-247.
[130] Wimalasena, D. S., Cramer, J. C., Janowiak, B. E., Juris, S. J., *et al.*, Effect of 2-fluorohistidine labeling of the Anthrax protective antigen on stability, pore formation, and translocation. *Biochemistry* 2007, *46*, 14928-14936.
[131] Wimalasena, D. S., Janowiak, B. E., Lovell, S., Miyagi, M., *et al.*, Evidence that histidine protonation of receptor-bound anthrax protective antigen is a trigger for pore formation. *Biochemistry* 2010, *49*, 6973-6983.
[132] Yeh, H. J. C., Kirk, K. L., Cohen, L. A., Cohen, J. S., 19F and 1H nuclear magnetic resonance studies of ring-fluorinated imidazoles and histidines. *J. Chem. Soc., Perkin Trans. 2* 1975, 928-934.
[133] Alexeev, D., Barlow, P. N., Bury, S. M., Charrier, J. D., *et al.*, Synthesis, structural and biological studies of ubiquitin mutants containing (2S, 4S)-5-fluoroleucine residues strategically placed in the hydrophobic core. *Chembiochem* 2003, *4*, 894-896.
[134] Feeney, J., McCormick, J. E., Bauer, C. J., Birdsall, B., *et al.*, 19F Nuclear Magnetic Resonance Chemical Shifts of Fluorine Containing Aliphatic Amino Acids in Proteins: Studies on Lactobacillus casei Dihydrofolate Reductase Containing (2S,4S)-5-Fluoroleucinell. *J. Am. Chem. Soc.* 1996, *118*, 8700-8706.
[135] Minks, C., Huber, R., Moroder, L., Budisa, N., Atomic mutations at the single tryptophan residue of human recombinant annexin V: effects on structure, stability, and activity. *Biochemistry* 1999, *38*, 10649-10659.
[136] Minks, C., Huber, R., Moroder, L., Budisa, N., Noninvasive tracing of recombinant proteins with "fluorophenylalanine-fingers". *Anal. Biochem.* 2000, *284*, 29-34.
[137] Parsons, J. F., Armstrong, R. N., Proton configuration in the ground state and transition state of a glutathione transferase-catalyzed reaction inferred from the properties of tetradeca(3-fluorotyrosyl)glutathione transferase. *J. Am. Chem. Soc.* 1996, *118*, 2295-2296.
[138] Deepankumar, K., Shon, M., Nadarajan, S. P., Shin, G., *et al.*, Enhancing thermostability and organic solvent tolerance of ω-transaminase through global incorporation of fluorotyrosine. *Adv. Synth. Catal.* 2014, *356*, 993-998.
[139] Mehta, K. R., Yang, C. Y., Montclare, J. K., Modulating substrate specificity of histone acetyltransferase with unnatural amino acids. *Mol. Biosyst.* 2011, *7*, 3050-3055.
[140] Bronskill, P. M., Wong, J. T., Suppression of fluorescence of tryptophan residues in proteins by replacement with 4-fluorotryptophan. *Biochem. J.* 1988, *249*, 305-308.
[141] Wong, C.-Y., Eftink, M. R., Incorporation of tryptophan analogues into Staphylococcal nuclease, its V66W mutant, and Δ137−149 fragment: Spectroscopic Studies. *Biochemistry* 1998, *37*, 8938-8946.
[142] Tsien, R. Y., The green fluorescent protein. *Annu. Rev. Biochem.* 1998, *67*, 509-544.
[143] Llopis, J., McCaffery, J. M., Miyawaki, A., Farquhar, M. G., *et al.*, Measurement of cytosolic, mitochondrial, and golgi pH in single living cells with green fluorescent proteins. *Proc. Natl. Acad. Sci. USA* 1998, *95*, 6803-6808.
[144] Ormö, M., Cubitt, A. B., Kallio, K., Gross, L. A., *et al.*, Crystal structure of the Aequorea victoria green fluorescent protein. *Science* 1996, *273*, 1392-1395.
[145] Heim, R., Tsien, R. Y., Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr. Biol.* 1996, *6*, 178-182.
[146] Budisa, N., Pal, P. P., Alefelder, S., Birle, P., *et al.*, Probing the role of tryptophans in Aequorea victoria green fluorescent proteins with an expanded genetic code. *Biol. Chem.* 2004, *385*, 191-202.
[147] Pal, P. P., Bae, J. H., Azim, M. K., Hess, P., *et al.*, Structural and spectral response of Aequorea victoria green fluorescent proteins to chromophore fluorination. *Biochemistry* 2005, *44*, 3663-3672.
[148] Nagasundarapandian, S., Merkel, L., Budisa, N., Govindan, R., *et al.*, Engineering protein sequence composition for folding robustness renders efficient noncanonical amino acid incorporations. *Chembiochem* 2010, *11*, 2521-2524.
[149] Whitman, C. P., The 4-oxalocrotonate tautomerase family of enzymes: how nature makes new enzymes using a beta-alpha-beta structural motif. *Arch. Biochem. Biophys.* 2002, *402*, 1-13.
[150] Ge, M., Pan, X.-M., The contribution of proline residues to protein stability is associated with isomerization equilibrium in both unfolded and folded states. *Extremophiles* 2009, *13*, 481-489.
[151] Balasubramian, R., Lakshminarayanan, A. V., Sabesan, M. N., G., T., *et al.*, Conformations of proline rings in crystal structures. *Int. J. Pept. Protein Res.* 1971, *3*, 25–33
[152] DeRider, M. L., Wilkens, S. J., Waddell, M. J., Bretscher, L. E., *et al.*, Collagen stability: insights from NMR spectroscopic and hybrid density functional computational investigations of the effect of electronegative substituents on prolyl ring conformations. *J. Am. Chem. Soc.* 2002, *124*, 2497-2505.
[153] Milner-White, E. J., Bell, L. H., Maccallum, P. H., Pyrrolidine ring puckering in cis and trans-proline residues in proteins and polypeptides: Different puckers are favoured in certain situations. *J. Mol. Biol.* 1992, *228*, 725-734.
[154] Holmgren, S. K., Bretscher, L. E., Taylor, K. M., Raines, R. T., A hyperstable collagen mimic. *Chem. Biol.* 1999, *6*, 63-70.
[155] Shoulders, M. D., Kamer, K. J., Raines, R. T., Origin of the stability conferred upon collagen by fluorination. *Bioorg. Med. Chem. Lett.* 2009, *19*, 3859-3862.
[156] Shoulders, M. D., Raines, R. T., Collagen structure and stability. *Annu. Rev. Biochem.* 2009, *78*, 929-958.
[157] Vitagliano, L., Berisio, R., Mazzarella, L., Zagari, A., Structural bases of collagen stabilization induced by proline hydroxylation. *Biopolymers* 2001, *58*, 459-464.
[158] Shoulders, M. D., Satyshur, K. A., Forest, K. T., Raines, R. T., Stereoelectronic and steric effects in side chains preorganize a protein main chain. *Proc. Natl. Acad. Sci. USA* 2009.
[159] Kim, W., McMillan, R. A., Snyder, J. P., Conticello, V. P., A stereoelectronic effect on turn formation due to proline substitution in elastin-mimetic polypeptides. *J. Am. Chem. Soc.* 2005, *127*, 18121-18132.
[160] Muiznieks, L. D., Weiss, A. S., Keeley, F. W., Structural disorder and dynamics of elastin. *Biochem. Cell Biol.* 2010, *88*, 239-250.
[161] Yamaoka, T., Tamura, T., Seto, Y., Tada, T., *et al.*, Mechanism for the phase transition of a genetically engineered elastin model peptide (VPGIG)40 in aqueous solution. *Biomacromolecules* 2003, *4*, 1680-1685.
[162] Borgogno, A., Ruzza, P., The impact of either 4-R-hydroxyproline or 4-R-fluoroproline on the conformation and SH3m-cort binding of HPK1 proline-rich peptide. *Amino Acids* 2013, *44*, 607-614.
[163] Hartley, R. W., Barnase and barstar: Expression of its cloned inhibitor permits expression of a cloned ribonuclease. *J. Mol. Biol.* 1988, *202*, 913-915.
[164] Zheng, T.-Y., Lin, Y.-J., Horng, J.-C., Thermodynamic consequences of incorporating 4-substituted proline derivatives into a small helical protein. *Biochemistry* 2010, *49*, 4255-4263.
[165] Rubini, M., Schärer, M. A., Capitani, G., Glockshuber, R., (4R)- and (4S)-fluoroproline in the conserved cis-prolyl peptide bond of the thioredoxin fold: tertiary structure context dictates ring puckering. *Chembiochem* 2013, *14*, 1053-1057.
[166] Neil, E., Marsh, G., Towards the nonstick egg: designing fluorous proteins. *Chem. Biol.* 2000, *7*, R153-R157.
[167] Cametti, M., Crousse, B., Metrangolo, P., Milani, R., *et al.*, The fluorous effect in biomolecular applications. *Chem. Soc. Rev.* 2012, *41*, 31-42.
[168] Buer, B. C., Marsh, E. N. G., Fluorine: A new element in protein design. *Protein Sci.* 2012, *21*, 453-462.
[169] Buer, B. C., Meagher, J. L., Stuckey, J. A., Marsh, E. N. G., Structural basis for the enhanced stability of highly fluorinated proteins. *Proc. Natl. Acad. Sci. USA* 2012, *109*, 4810-4815.
[170] Panchenko, T., Zhu, W. W., Montclare, J. K., Influence of global fluorination on chloramphenicol acetyltransferase activity and stability. *Biotechnol. Bioeng.* 2006, *94*, 921-930.

[171] Holzberger, B., Rubini, M., Möller, H. M., Marx, A., A highly active DNA polymerase with a fluorous core. *Angew. Chem. Int. Ed. Engl.* 2010, *49*, 1324-1327.
[172] Woll, M. G., Hadley, E. B., Mecozzi, S., Gellman, S. H., Stabilizing and destabilizing effects of phenylalanine --> F5-phenylalanine mutations on the folding of a small protein. *J. Am. Chem. Soc.* 2006, *128*, 15932-15933.
[173] Yoo, T. H., Link, A. J., Tirrell, D. A., Evolution of a fluorinated green fluorescent protein. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 13887-13890.
[174] Montclare, J. K., Tirrell, D. A., Evolving proteins of novel composition. *Angew. Chem. Int. Ed. Engl.* 2006, *45*, 4518-4521.
[175] Robert, X., Gouet, P., Deciphering key features in protein structures with the new ENDscript server. *Nucl. Acids Res.* 2014, *42*, W320-W324.

2 Chapter

5'-deoxy-5'-fluoroadenosine production in *E. coli*: Establishing the first step towards *in vivo* 4-fluorothreonine synthesis

Corinna Odar[1,2], Maria Koshanskaya[1,2], Birgit Wiltschi[1]

[1] Austrian Centre of Industrial Biotechnology, Petersgasse 14, Graz, Austria

[2] Graz University of Technology, Institute of Molecular Biotechnology, Petersgasse 14, Graz Austria

Author contributions

Maria Koshanskaya did the cloning of the codon-optimized versions of FlA and the RLSS expression constructs as well as the main experiments with lyophilized cells with supervision by Corinna Odar.

All other work within this manuscript was done by Corinna Odar with supervision by Birgit Wiltschi.

2.1 **Abstract**

Fluoro amino acids represent a valuable tool for the engineering as well as for biochemical studies of proteins. Some fluoro amino acids, like 4-fluoro-L-threonine (4FT), are commercially unavailable, which limits the repertoire of fluoro amino acids for these studies. The particularity of fluoroorganocompounds such as fluoro amino acids is reflected by their virtual absence in nature. However, although 4FT is not commercially available some microbial species realized its enzymatic synthesis. So far, its biosynthesis represents the only pathway of biological fluoro amino acid formation. The first enzyme of the 4FT pathway, the fluorinase, has attracted special attention because it is the only enzyme class known to date capable of forming the carbon-fluorine bond. The fluorinase catalyzes the synthesis of 5'-deoxy-5'-fluoroadenosine (FDA) from S-(5′-adenosyl)-L-methionine (SAM) and fluoride with L-methionine as the byproduct.

In a long term project, we set out to engineer *E. coli* to biosynthesize 4-fluorothreonine for its subsequent incorporation into proteins to produce fluorinated variants. In this study we aimed at establishing the first pathway step in *E. coli* to synthesize FDA by the action of the fluorinase. Fluorinase expression studies served to optimize its expression pattern in *E. coli*. In an assay with lyophilized cells the fluorinase overexpressing strain successfully produced FDA from fluoride. We co-expressed a rat SAM synthase known to increase intracellular SAM levels. Moreover, the SAM synthases recycle the L-methionine to SAM. We showed that the increased SAM levels upon SAM synthase co-expression resulted in elevated FDA conversion.

In addition to the experimental section, this manuscript elaborates on general bottlenecks to realize the long term goal of 4-fluorothreonine synthesis for the expression of fluoro protein variants in *E. coli* and provides potential solutions.

2.2 **Background**

Incorporation of fluoro amino acids (FAAs) into proteins is of highly scientific and biotechnological interest [1]. However, FAA incorporation to produce fluorinated variants is often hampered by the cost or the availability of the respective fluorinated analogs. Fluorinated metabolites are scarcely found chemical species in the biological realm of life, so it is not surprising that 4-fluorothreonine (4FT) is the only natural FAA found to date.

The first enzyme in the five step reaction cascade to 4FT is the fluorinase (FlA), which forms a carbon-fluorine bond between S-(5′-adenosyl)-L-methionine (SAM) and a fluoride ion to give 5'-deoxy-5'-fluoroadenosine (FDA) [2]. This is the only enzyme class known to date capable of using the fluoride ion to produce a fluoroorganic compound. *Streptomyces cattleya* was the first species to be recognized as a 4FT producer [3]. Its 4FT pathway enzymes (Figure 1) were the subject of several studies with a special focus on the fluorinating enzyme FlA (summarized in [1]).

Nature has evolved an elegant enzymatic strategy to produce 4FT (Figure 1). Chemical approaches, though, seem to be rather inefficient in the economic synthesis of this non-canonical amino acid, because 4FT is not commercially available. This might also explain why 4FT has never been tested for the incorporation into proteins.

Our long term goal is to establish an *E. coli* cell, engineered to produce 4FT for the incorporation into target proteins. Although the 4FT pathway was already reconstituted *in vitro* [4], we had to consider several issues when realizing its synthesis in a whole cell *in vivo* approach.

This study aims at establishing the first pathway step to produce FDA *in vivo* upon the heterologous expression of FlA in *E. coli*. Moreover, it elaborates on general bottlenecks for efficient 4FT *in vivo* synthesis and provides potential solutions (see the Outlook section). The following issues were considered to be relevant for the engineering of an *E. coli* host to produce 4FT for incorporation into proteins:

I) In general, high-level heterologous expression of proteins from *Streptomyces* origins in *E. coli* can be problematic [4]. This can be due to the high GC content in the *Streptomyces* genome [5] resulting in a codon usage often incompatible with the one used in *E. coli* as the heterologous expression host. It was reported that *E. coli* expressed the first three *S. cattleya* 4FT pathway enzymes (Figure 1), FlA, a phosphorylase and an isomerase (15 - 25 mg of purified protein per liter culture [6]). The identity of the fourth enzyme in the pathway, an aldolase (Figure 1), has not been elucidated so far [7]. A surrogate aldolase from *Streptomyces coelicolor* was overexpressed in *E. coli* (protein yield per liter culture not indicated) and shown to convert 5-deoxy-5-fluororibose phosphate to 5-deoxy-5-fluororibulosephosphate (Figure 1) [4]. The 4FT transaldolase in charge of the last reaction step of the pathway (Figure 1) could not be expressed in *E. coli* at all, but only in *Streptomyces lividans* as an alternative expression host [4]. However, expression yields were poor and its successful overexpression in *E. coli* is of high scientific interest to allow for a more detailed functional and structural analysis.

II) Concerning FlA expression, its natural host *S. cattleya* expresses this enzyme at high quantities (Figure 2) [8]. This suggests that a high level of FlA expression represents a prerequisite for the engineering of an efficient FDA and consequently 4FT production host.

III) FlA has a rather high K_M for fluoride, albeit the literature is somewhat controversial about the exact value (K_M (F^-) = 2 mM at 0.8 mM SAM [9], K_M (F^-) = 9 mM at 0.2 mM SAM [8], K_M (F^-) = 10 mM at 0.02 mM SAM [10]). Moreover, Zhu *et al.* reported that the K_M for fluoride increases with higher concentrations of SAM (K_M (F^-) = 10 mM at 0.02 mM SAM, K_M (F^-) = 16 mM at 0.06 mM SAM, K_M (F^-) = 47 mM at 0.3 mM SAM) [10].

Because of the high K_M of FlA for fluoride the intracellular fluoride level must be sufficient to promote an efficient *in vivo* conversion. However, fluoride is cytotoxic to the *E. coli* cell. That is why *E. coli* evolved an active transport system to protect itself from fluoride intoxication [11].

IV) Not only fluoride, but also 4FT is detrimental for *E. coli* cell growth [3]. Its toxicity might be based on its antagonism to threonine [1]. This assumption is supported by the observation that its antimicrobial activity only comes into effect when *E. coli* is cultured in minimal medium without any amino acid supplementations [3].

Fluoroacetaldehyde, the precursor molecule of 4FT in the pathway, is a toxic and unstable compound [12]. Moreover, we cannot rule out that it might be converted to the highly toxic fluoroacetate (Figure 1) by an *E. coli* dehydrogenase. Due to the relatively small size of fluorine compared to other common hydrogen substituents, enzymes often accept the fluorinated analogs for conversion [13]. Fluoroacetate is the second end product of the 4FT pathway in the natural host *S. cattleya* and it is still unclear how *S. cattleya* protects itself from fluoroacetate intoxication[14].

V) Not only fluoroacetaldeyde but also the other pathway intermediates might be accidentally used by endogenous *E. coli* enzymes and retrieved from the 4FT reaction cascade. For example, the *E. coli* nucleosidase MtnN efficiently converted FDA to 5-fluoro-5-deoxy-D-ribose *in vitro* [15].

VI) In the last reaction step, the 4FT transaldolase uses threonine to convert fluoroacetaldehyde to 4FT [16]. In the residue-specific approach for non-canonical amino acid incorporation, the absence of the natural amino acid to be substituted is a prerequisite [17]. Thus, the presence of any threonine renders residue-specific incorporation of 4FT unfeasible and alternative ways for the conversion of fluoroacetaldehyde to 4FT have to be explored.

In this study, we conducted experiments to optimize FlA expression by evaluating the influence of different expression constructs, temperature, expression time, inducer concentrations and an N-terminal His-tag. Moreover, we evaluated the effects of the co-expression of different chaperones on the production of FlA with the intent to apply these chaperone constructs for co-expression with the other pathway enzymes in future studies. As one of the two chaperons, we choose the cognate small heat shock protein LbpB from *E. coli*. A study had demonstrated its positive effect on soluble expression of a protein of *Streptomyces* origin [18]. The second chaperone was the *E. coli*

endogenous periplasmic foldase DsbC which shows chaperone activity [19]. The truncated version of this chaperone ($DsbC_{trunc}$) lacking its periplasmic signal sequence promoted disulfide bond formation in heterologously expressed proteins when co-expressed in the cytoplasm [20]. Thus, it would be of special interest for the overexpression of the 4FT transaldolase with seven predicted disulfide bonds (DIANNA1.1 [21]).As another attempt to improve the FlA expression pattern, we optimized the *flA* gene sequence for the codon usage in *E. coli* with regard to the original codon usage in *S. cattleya*. Therefore, we placed rare *E. coli* codons at the same positions where *S. cattleya* also uses rare codons. The codon usage frequencies of FlA in *S. cattleya* are exemplified for its N-terminal sequence in Figure 3. All gene sequences of the 4FT pathway enzymes were optimized according to this strategy (no experimental data presented herein).

The structural integrity of FlA may be positively influenced by the presence of its substrate S-(5′-adenosyl)-L-methionine (SAM) [9]. *E. coli* cannot assimilate SAM from the culture medium [22], but studies showed that overexpression of the SAM synthase from the liver of *Rattus norvegicus* (RLSS) elevated the intracellular SAM levels [23, 24]. In contrast to *E. coli's* cognate SAM synthase, MetK, for which SAM acts as a feedback inhibitor, RLSS activity appears to be largely unaffected by high SAM concentrations [23, 24]. Together with adenosine triphosphate (ATP), RLSS uses L-methionine, which is the second product of the FlA reaction besides FDA, as a substrate for SAM synthesis. Thus, we argued that the retrieval of methionine from the product side together with an increase in the intracellular SAM pool can push the reaction towards FDA synthesis (Figure 4). This approach was applied for the *in vitro* synthesis of FDA in the presence *E. coli* MetK SAM synthase instead of RLSS [15].

For FDA synthesis we used an *in vivo* approach where FlA overexpressing *E. coli* cultures were supplemented with fluoride. Moreover, an assay with lyophilized cells was applied with substrates for FlA, SAM and fluoride, as well as of substrates for RLSS, L-methionine and ATP.

In summary, the experimental section of this manuscript presents expression studies of the fluorinase from *S. cattleya* and experiments on the enzymatic synthesis of FDA.

2.3 Results and discussion

2.3.1 Expression studies of the fluorinase

In contrast to the use of a cell as a heterologous protein production factory, metabolic engineering aims at high product titers of a desired compound *in vivo* and not necessarily for the high-level expression of (heterologous) enzymes [25]. Thus, we studied the FlA expression pattern from a single copy vector (pKLJ12-His-flA) under an arabinose inducible promoter. We reasoned that the low level expression might result in soluble and metabolically active FlA. Immunodetection of the hexahistidine tagged (His-tagged) FlA, however, showed that most FlA ended up in the insoluble protein fraction (Figure 5). Expression levels increased with the amount of the inducer arabinose but FlA was always pronounced in the insoluble protein fraction compared to the soluble one. Even at the lowest expression regimes only insoluble FlA was detected. These results indicate that the expression pattern

was not influenced by the strength of expression. It appeared that the more FlA was found in the insoluble fraction, the more FlA was also solubly expressed (Figure 5).

To test whether an alternative expression system would enhance soluble FlA levels, we cloned the *flA* gene into different vector constructs. All of them where medium copy plasmids either employing the T7 (pET28a(+)-His-flA, [9]), the TAC (pMS470-His-flA) or the T5 (pQE80L-His-flA) promoter for IPTG inducible expression. SDS PAGE (Figure 6A) and Western blot (Figure 6B) analysis showed that the T5 promoter yielded the most soluble protein. However, most of the overexpressed FlA ended up in the insoluble protein fraction as previously.

In the next experiment we evaluated if the temperature and the time of induction had an influence on the FlA expression pattern. FlA was expressed at 16 °C and 28 °C during an induction period of either 5 h or 16 h (Figure 7). However, the expression pattern was the same for every condition tested. Moreover, we reduced the inducer concentration from 0.5 M to 10 mM IPTG. Less FlA was expressed with 10 mM IPTG, but in contrast to the induction with 0.5 mM IPTG (Figure 7) the SDS gel did not show an overexpression band in the cleared lysate fraction (Figure 8). This observation was consistent with the FlA expression pattern from the single copy vector pKLJ12-His-flA (Figure 5). For this arabinose inducible single copy construct soluble FlA expression was only detected with the highest inducer concentrations of 10 mg/L arabinose. In the experiment with lowered IPTG inducer concentrations we tested whether the His-tag negatively influenced FlA expression. Therefore, the enzyme was expressed without the His-tag from the otherwise identical vector construct (pQE80L-flA). However, the untagged FlA version induced with 10 mM IPTG was present solely in the insoluble protein fraction as well (Figure 8). This result showed that the N-terminal His-tag was not responsible for the mainly insoluble expression of FlA.

In the next step we co-expressed the small heat shock protein LbpB and the foldase DsbC to see whether these chaperones can render FlA expression more soluble. Therefore, we co-transformed the FlA expression strain (pQE80L-His-flA) with either a plasmid harboring only *lbpB* (p15aTAC-lbpB), only *dsbC (*p15aTAC- dsbC) or both of the chaperones (p15aTAC-lbpB-dsbC). In the first expression study the co-expression of LbpB enhanced soluble FlA expression compared to the other expression strains (Figure 9A). However, repetition of these experiments could not verify our first observations and no FlA overexpression bands could be observed again (Figure 9B). The two experiments differed in the FlA induction time that was over night (Figure 9A) or 5 h (Figure 9B). It has to be mentioned that the cell growth was diminished in the cells overexpressing the chaperones, what was most pronounced for the strain overexpressing LbpB (data not shown). Moreover, the main fraction of LbpB was insoluble. The rather strong overexpression of LbpB (Figure 9) might be detrimental to *E. coli* cell growth.

Another approach to improve soluble FlA overexpression was to codon optimize the gene sequence for *E. coli*. To this end, codons were optimized with the program Gene Designer from DNA2.0 Inc. with a codon usage table for highly transcribed genes in *E. coli* (Class1) (Table S1). We manually placed rare *E. coli* codons at the positions where *S. cattleya* uses rare codons for the canonical *flA* sequence (Table 1). Our first attempt for codon optimization failed (FlAco1) as no expression could be detected either in the cleared lysate or in the insoluble protein (pellet) fraction (Figure 10). We reasoned that we

accidentially replaced some codons at the N-terminal *flAco1* sequence, namely Ala2, Arg7 and Pro9, by too frequently used codons in *E. coli* (Table 1). Rare codons are enriched at the N-terminus of many natural proteins. The use of rare codons at the N-terminus was shown to improve heterologous protein expression in *E. coli* [26]. To prove our assumption we replaced the codons for the mentioned three amino acids in the *flAco1* sequence by more rarely used *E. coli* codons (*flAco2*, Table 1). This approach resulted in an effective FlA expression compared to the first optimized version FlAco1, where no expression was detected at all (Figure 10). Notably, for the revised codon-optimized version (*flAco2*) a reasonable higher ratio of FlA was expressed solubly compared to FlA expression with the original *S. cattleya* gene sequence. Insoluble and total FlA protein was less for expressions from *flAco2* (Figure 10).

Crystallographic data suggest that the presence of the FlA substrate SAM might have a positive effect on its structural integrity [9]. To increase intracellular SAM levels we co-expressed RLSS [23, 24] from a compatible low copy plasmid with an arabinose inducible promoter (p15aARA-RLSS). RLSS was induced at a D_{600} of 0.4 and FlA at either a D_{600} of 0.9 (s4) or 4.5 (s5). The later FlA induction at a D_{600} of 4.5 was performed to evaluate if the presumably higher SAM level at this time of induction was decisive for the FlA expression pattern. SDS PAGE analysis of soluble and insoluble protein fractions showed that the overall expression of FlA was slightly enhanced in the RLSS co-expressing strain (Figure 11A). But FlA was mainly found in the insoluble protein fraction for the strain co-expressing RLSS as well as for the strain not expressing it (Figure 11A). After 18 h of expression a very slight FlA band was detected in SDS PAGE analysis for the RLSS co-expression strains (s4 and s5) not present in the strain overexpressing only FlA (s3). However, this can also refer to an unidentified protein band. The staining of the SDS gel was of bad quality and too little cleared lysate was loaded for a conclusive comparison of the FlA expression patterns. Therefore, we applied three times more of the soluble protein samples equaling 1.5 D_{600} units (Figure 11B). Moreover, we used 4-12% Bis-Tris protein gels instead of home-made SDS gels (for details see chapter 5). The band referring to FlA was more pronounced in the RLSS co-expression strains (s4 and s5). In these samples an abundant endogenous protein exactly co-migrates with FlA, which is especially pronounced in the control strain only expressing RLSS and not FlA (s2). Therefore, only a Western blot can give insight into soluble protein expression yields of FlA in the presence or absence of RLSS in future experiments.

Our studies showed that the applied codon optimization approach for the *flA* gene sequence presented the most promising way to improve FlA soluble expression. Moreover, the choice of the vector backbone influenced the FlA expression pattern decisively. Neither expression time, the amount of inducer nor the N-terminal His-tag affected soluble FlA expression in any way. Co-expression of the chaperone LbpB showed a slight positive effect on the soluble FlA ratio, but the result was not reproducible in a biological replicate. Co-expression of the disulfide bond isomerase DsbC did not result in any change of the FlA expression pattern. From the current data it was not possible to clearly deduce if RLSS co-expression has an effect on FlA solubility, but if so, the effect seems to be marginal.

2.3.2 **FDA conversion in *E. coli***

2.3.2.1 ***In vivo* FDA conversion in *E. coli* cultures**

For the *in vivo* formation of FDA we used the strain overexpressing FlA from the single copy arabinose inducible plasmid (pKLJ12-His-flA), because it was the only one available to us at this time of the study.

We designed the experiment based on studies on fluorometabolite synthesis in *S. cattleya*, which was cultured at 30 °C in the presence of 2 mM fluoride [14]. At this time of the study we supposed that a lower expression temperature might improve soluble FlA expression. Therefore, we maintained the temperature after induction of FlA at 25 °C for 3 h before we added fluoride and elevated the temperature to 30 °C.

Fluoride is harmful to *E. coli* cell growth, that is why *E. coli* has evolved an active fluoride transporter to protect itself from fluoride intoxication [11]. A study by Baker *et al.* reported normal growth at fluoride concentrations of up to 1 mM and a slightly diminished cell growth until 10 mM [11]. Nevertheless, for an efficient FDA production, fluoride has to be provided in sufficient amounts because FlA shows a high K_M of 2 mM [9] for this substrate. Other studies reported on even higher K_M values of roughly 9 mM [8] and 10 mM [10] (see also above).

In a first attempt, we cultured the FlA overexpressing strain over night in full medium in the presence of 2 mM fluoride. We prepared cell extracts and lyophilized them together with the culture supernatant since we did not know if FDA was maintained intracellularly or secreted into the culture medium, like by *S. cattleya* [14]. The lyophilisate contained both, the cell extracts together with the culture supernatant (for details see chapter 5.4). HPLC analysis of suspensions of the lyophilisates in buffer did not indicate any FDA synthesis under these culture conditions (Figure 12). Moreover, we expected to see an effect on the growth rate by the addition of 2 mM fluoride as reported in a study by Baker *et al.* [11], but the cells appeared healthy. From our observations, we suspected that the cells might have excluded fluoride. so that the intracellular concentration was presumably too low for the synthesis of detectable FDA amounts. Low intracellular fluoride levels might represent another reason for undetectable FDA synthesis in addition to the low soluble FlA expression yields (Figure 5).

Consequently, in the next experiment we used 10 mM of fluoride and the cells were cultured for three days for FDA production in defined medium with glycerol as the sole carbon source. Lyophilisates were produced the same way as before. However, FDA synthesis was not detected as shown by HPLC analysis (Figure 13).

In addition to the presumably low intracellular fluoride levels, the amount of expressed FlA in *E. coli* could present another limitation (Figure 5). Studies on *S. cattleya* showed that FlA was highly expressed in this fluorometabolite producing organism (Figure 2) [8].

To evaluate if active protein was expressed from pKLJ12-His-flA, we purified it by Ni^{2+} affinity chromatography and conducted an *in vitro* assay with the purified protein. For comparative analysis FlA was also expressed from pET28a(+)-His-flA, a construct that had been reported to express FlA effectively [9]. Like observed before, cleared lysate fractions of the strains carrying the expression

constructs did not show a pronounced overexpression of FlA, but most of the protein was found in its insoluble form (Figure S1A). However, sufficient protein could be purified for *in vitro* conversions (Figure S1B, lanes E). Protein concentrations of purified protein fractions (Figure S1B, lanes E) were determined using the Bradford assay. Because purified protein fractions contained a reasonable amount of unknown protein impurities, the relative content of FlA was calculated by densitometric analysis of SDS gels (Figure S1B, lanes E) with the software ImageJ [27]. For pKLJ12-His-FlA 46% of the protein in the elution fraction was calculated to be FlA and for pET28a(+)-His-flA 73%. This resulted in calculated FlA yields of 0.34 mg for pKLJ12-His-flA and of 0.83 mg for pET28a(+)-His-flA per liter culture. However, we presume that the purified protein fraction of FlA expressed from pET28a(+)-His-flA had a higher relative FlA content than the calculated 73%. Lane E of pET28a(+)-His-flA was clearly overloaded (Figure S1B, lane E) what probably biased the densitometric analysis.

The *in vitro* assays contained 0.28 mg/ml FlA, 1.5 mM SAM and 100 mM fluoride. They were successful in producing FDA confirming the activity of FlA (Figure S2).

As expected more FDA was produced in the *in vitro* reaction with FlA from the pET28a(+) construct (Figure S2, no integrated peak areas could be calculated due to problems with the HPLC device). We assume that it contained more FlA compared to the reaction with FlA expressed from pKLJ12 because of the biased densitometric analysis.

Based on the FlA yields expressed from pKLJ12 calculated in this chapter, the *in vitro* assays contained almost a 1000 times higher concentration of FlA compared to the cell cultures used for the *in vivo* reactions (280 mg/L in the *in vitro* reaction versus 0.34 mg/L culture in the *in vivo* approach). However, we can only speculate about intracellular FlA concentrations.

Moreover, 50 times or 10 times less fluoride was applied in the *in vivo* reactions in full or minimal medium, respectively, than in the *in vitro* reaction (2 mM or 10 mM *in vivo* versus 100 mM *in vitro*).

From these observations we conclude that the FlA expression level as well as fluoride concentrations represent the bottlenecks for a successful FDA synthesis *in vivo*.

2.3.2.2 **FDA conversion with lyophilized *E. coli* cells**

An assay with lyophilized cells was performed, as we suspected too low intracellular fluoride concentrations as one of the potential reasons for the failure of *in vivo* FDA synthesis. Lyophilized cells show preserved enzyme activities [28]. We speculate that the substrates, fluoride and SAM, can pass the cell membrane because it is presumably permeabilized during the lyophilization process. This approach also allowed to evaluate if the co-expression of RLSS increased the SAM levels and thereby affected FDA production. It was previously shown that overexpression of RLSS elevated the intracellular SAM levels [23, 24].

A priori it should be mentioned that decisive conclusions are difficult to draw from the experimental data provided herein as only single measurements from one biological sample are presented. The inherent errors of the experimental procedure and analytical method as well as biological fluctuations in cell cultures were unclear at this point of the study. No internal standard was used to correct for the

experimental error. Moreover, only limited amounts of the FDA standard were available to us, that is why no calibration curve was recorded.

Pre-experimental data to confirm the stability of FDA under the applied reaction conditions (50 mM Tris/Cl buffer, pH 7.8, incubation for 1 h at 37 °C) and the processing of the samples (protein precipitation at 95 °C for 5 min and filtering of the supernatant for HPLC analysis) can be found in the Supplemental (Figure S3).

Based on the FlA expression studies in chapter 2.1 we conducted the following experiments with the ITPG inducible medium copy vector pQE80L-His-flA because this expression construct of the non codon-optimized FlA sequence showed the best soluble FlA expression at this time of the study (Figure 6).

Lyophilisate preparations contained washed cells (for details see chapter 5.5.1) from five different cultures: s1 refers to the empty vector strain (pQE80L and p15aARA); s2 to the strain overexpressing only RLSS induced at a D_{600} of 0.5 (pQE80L empty vector and p15aARA-RLSS); s3 to the strain only overexpressing FlA induced at a D_{600} of 0.9 (pQE80L-His-FlA and p15aARA empty vector); and s4 as well as s5 to the strains overexpressing both FlA and RLSS (pQE80L-His-FlA and p15aARA-RLSS). The difference between s4 and s5 is the expression start of FlA that was induced at a D_{600} of 0.9 for s4 and at a D_{600} of 4.5 for s5. In all strains RLSS was induced at a D_{600} of 0.5.

We analyzed the FDA and SAM synthesis capacity of these lyophilisates under five different reaction conditions further referred to as experiment-A, -B, -C, -D and -E (Table 2). All reactions contained 200 mM fluoride and about 100 mg/mL lyophilized cells with exception of experiment-B.

In experiment-A, the reactions were supplemented with 2 mM SAM in addition to 200 mM fluoride. This experimental condition enabled FDA synthesis with lyophilized cells when FlA substrates are supplemented.

For the evaluation if the amount of lyophilized cells, i.e. the *FlA* enzyme therein, correlates with the synthesized FDA, we applied different concentrations of lyophilized cells in the presence of 2 mM SAM and 200 mM fluoride in experiment-B.

Moreover, we carried out experiment-C, where external SAM was omitted from the reaction mixes. We expected that an increase in SAM levels by the action of RLSS could only be observed in the absence of external SAM.

To further evaluate the SAM synthesis capacity by RLSS, we supplemented 20 mM L-methionine in experiment-D. L-methionine is the second substrate for RLSS besides ATP and a byproduct of FDA synthesis (Figure 4).

In experiment-E both RLSS substrates, 20 mM ATP as well as 20 mm L-methionine, were provided in reactions. Chromatograms of of FDA, SAM and ATP are depicted in Figure S4. L-methionine, fluoride and 50 mM Tris/Cl buffer, pH 7.8 did not show any absorbance at the detection wavelength of 260 nm (data not shown).

Retention times of FDA and SAM in experiment-A were 4.9 min and 11.5 min, respectively (Figure S4A and Figure 14, chromatogram not shown). In experiment-B to E retention times for FDA varied from 4.8 min to 8 min and retention times for SAM from 11.6 min to 14.9 min (Figure 15, chromatogram not shown; Figure 16; Figure 17 and Figure 18). There was a trend in prolongation of retention times in the order the samples were analyzed in the HPLC device.

For experiment-A and -B integrated areas of the relevant peaks were calculated. In experiment B this allowed the comparison of relative levels of FDA or SAM between the reactions with different lyophilisates.

Due to the bad quality of the chromatograms of experiment-C, -D and E we do not show integrated peak areas but manual overlays of the chromatograms. This should facilitate a critical consideration of the data presented herein and preclude any misinterpretation.

FDA or SAM integrated peak areas from reaction mixes with lyophilized cells lacking the substrates, FDA and SAM, served as blanks in experiment-A. Peaks at the retention time of FDA were detected in these blanks (Figure S5) that had not appeared in preliminary experiments (data not shown). Integrated areas of unidentified peaks in these blanks constituted up to 8% of the FDA integrated peak area in the corresponding reaction samples (calculated from data presented in Figure S5 and Figure 14). Interestingly, no peak at the retention time of FDA occurred for the strain only expressing RLSS (s2) (Figure S5). So far, we do not know to which compound these peaks refer or what causes their appearance. Also for the *in vivo* sample taken before fluoride addition and overnight conversion (Figure 12A) a peak at this retention time had occurred.

HPLC analysis of blanks, *i.e.* lyophilized cells, also showed that overexpression of RLSS yielded higher levels of SAM (s2 in Figure S5). Moreover, less SAM was present in the strains overexpressing FlA (s4 and s5) compared to the strain overexpressing RLSS alone (s2) (Figure S5). This might be due to better expression yields of RLSS, i.e. active RLSS enzyme, without co-expression of FlA.

Integrated peak areas at the retention time of FDA in these blanks (s2) (Figure S5) were subtracted from the FDA peak areas in the reactions of experiment-A (Figure 14). In experiment-A we evaluated FDA synthesis with external addition of the FDA substrates. To this end 100 mg/ml lyophilized cells were supplemented with 2 mM SAM and 200 mM fluoride. Reactions with lyophilisates of all strains overexpressing FlA (s3, s4, s5) showed FDA synthesis (Figure 14). The SAM levels where slightly elevated with the RLSS overexpressing strains (s1 versus s2; s3 versus s4 and s5 in Figure 14). SAM levels negatively correlated with FDA levels: reactions with s1 and s2, that do not show any FDA synthesis capacity due to the lack of overexpressed FlA, showed elevated levels of SAM compared to reactions with s3, s4 and s5, which harbor the FlA enzyme (Figure 14). This observation indicates that FlA converted the supplemented SAM into FDA.

In this experiment-A, where 2 mM SAM was added, only marginal differences in FDA production were detected between reactions with lyophilisates of the strain only overexpressing FlA (s3) compared to the strains co-expressing RLSS together with FlA (s4 and s5) (Figure 14). Least FDA was detected in the reaction with the RLSS co-expressing strain with FlA induction at a D_{600} of 4.5 (s5). This indicates

that a longer induction time of FlA *i.e.* a higher expression level, correlates with an elevated FDA synthesis capacity.

In experiment-B we evaluated how the concentration of lyophilized cells, *i.e.* the FlA enzyme therein, influences FDA levels. To this end a lyophilisate stock solution was prepared and diluted to the following final concentrations in the reaction mix: 100 mg/mL, 75 mg/mL, 50 mg/mL, 25 mg/mL, 10 mg/mL and 5 mg/mL. SAM, 2 mM, and fluoride, 200 mM, were added like in experiment-A. The HPLC device was only available to us for a limited time. Therefore no blanks, processed lyophilized cultures, could be measured because the needed detection time to measure 30 samples as blanks would have been more than 7h. Instead, integrated peak areas of strains not overexpressing FlA (s1 and s2), and therefore devoid of any FDA, were subtracted from integrated FDA peak areas of samples overexpressing FlA (s3 and s4, respectively). Only reactions with lyophilized cells devoid of FlA and RLSS (s1) showed unidentified peaks at the retention time of FDA, while no such peak was detected in the reactions with lyophilized cells of the strain only overexpressing RLSS (s2). Integrated FDA peak areas (corrected by subtraction of the described blanks) and SAM peak areas of reactions with lyophilisates of the strain only expressing FlA (s3) and co-expressing RLSS with FlA (s4 and s5) are shown in Figure 15. Experiment-B showed that FDA levels increased with higher concentrations of lyophilisate in the reaction (Figure 15A). The sample with RLSS and FlA induced at a D_{600} of 0.9 (s5) showed lowest FDA levels for all lyophilisate concentrations tested (Figure 15A).

In contrast to FDA levels, SAM levels decreased with higher lyophilisate concentrations in the reaction (Figure 15B). The sample devoid of the enzymes FlA and RLSS (s1) always showed the highest SAM level. The difference of SAM levels between the FlA overexpressing strains (s3 and s4) and the strain devoid of FlA (s1), *i.e.* devoid of FDA synthesis, was more pronounced at lyophilisate concentration of more than 25 mg/mL (Figure 15B). Interestingly, SAM levels in reactions with lyophilisate concentrations between 25 mg/mL and 100 mg/mL do not decrease anymore for the sample devoid of FlA (s1). The reasons behind this effect remained elusive at this point of the study. No integrated SAM peak area could be calculated for s5, because the SAM peak was out of the detection time for these samples. In experiment-A and -B described so far, a high concentration of external SAM (2 mM) was present in reaction mixes. Therefore, an effect of elevated SAM levels in reactions with lyophilisates with RLSS overexpression (Figure S5) would not be visible. How elevated SAM levels by the action of RLSS influenced FDA production became obvious when no external SAM was added to the reaction mix in experiment-C. Reaction mixes contained 92 mg/mL lyophilized cells and 200 mM fluoride. The chromatogram showed that about three times more FDA was produced in the reaction with lyophilisate of the strain with RLSS co-expression and FlA induction at a D_{600} of 0.9 (s4) than in reactions with lyophilisate of the strain only overexpressing FlA (s3) (Figure 16). Reactions with lyophilisate of the strain with RLSS co-expression and FlA induction at a D_{600} of 4.5 (s5) showed lower FDA levels than RLSS co-expression and FlA induction at a D_{600} of 1.9 (s4) (Figure 16). SAM was not detected in the reactions of experiment-C (data not shown), what suggests that it was immediately used by enzymes including FlA to synthesize FDA.

The enzymatic synthesis of FDA from SAM and fluoride is reversible and L-methionine is produced as a byproduct (Figure 4). If L-methionine is removed from the reaction, the equilibrium will shift to the

product side, that is, the production of FDA would probably be stimulated [15] (Figure 4). We reasoned that the condensation of the byproduct L-methionine with ATP by the RLSS to recycle the FlA substrate SAM could be a way to elevate the enzymatically synthesized FDA levels as it was already exemplified in an *in vitro* assay [15]. Like in experiment-C, 92 mg/mL lyophilized cells and 200 mM of fluoride were added to the reaction mix in experiment-D which additionally contained 20 mM of L-methionine but no ATP. The strain overexpressing only RLSS (s2) showed SAM synthesis (Figure 17) what was not observed for the reactions in experiment-C where no L-methionine was added (data not shown). However, addition of L-methionine in experiment-D did not boost FDA synthesis for any reaction (s3, s4 and s5 in Figure 17 compared to Figure 16). If more SAM was present in the reactions due to the addition of L-methionine, it might not have been enough to compensate for the negative effect of the added L-methionine on the FDA reaction equilibrium.

The amount of ATP in the lyophilisate suspensions seemed to be the limiting component for the efficient recycling of SAM, and therefore improved FDA conversions. In addition to 20 mM L-methionine, 20 mM ATP were supplemented in experiment-E to reactions with 92 mg/mL lyophilisated cells. Indeed, the supplementation with ATP resulted in an increase in FDA levels in all reactions (Figure 18 compared to Figure 17). Like in experiment-A and -B where external SAM was present, the strain with FlA induction at a D_{600} of 4.5 (s5) showed least FDA production (Figure 18). FDA levels were the same in reactions with the lyophilisate of the FlA overexpressing strain (s3) and with the lyophilisate of the strain co-overexpressing RLSS with FlA (s4). However, this result is questionable, if the reaction with the lyphilisate of the strain neither expressing FlA nor RLSS (s1) is taken into account. The sample without RLSS and FlA (s1) showed a peak exactly at the retention time of FDA (Figure 18A). The intensity of this peak of unknown identity was much lower in the strain only overexpressing RLSS (s2). Occurrence of this unidentified peak in the sample without RLSS and FlA (s1) was consitent for all experiments but most pronounced in experiment-D, where ATP and L-methionine were added. The presence of a similar peak in the reaction with the FlA expressing strain (s3) cannot be excluded and would lead to an apparently higher FDA content than actually present in the reaction mix.

Taken together we showed that overexpression of RLSS elevated SAM levels (s2 in Figure S5, in Figure 17A and in Figure 18). The positive effect of elevated SAM levels upon RLSS co-expression on FDA synthesis became evident when only fluoride was added to the reaction mix with lyophilized cells (experiment-C; s3 compared to s4 and s5 in Figure 16) as well as in the prescence of L-methionine together with fluoride (experiment-D, s3 compared to s4 and s5 in Figure 17). Addition of ATP and L-methionine to reactions with fluoride provoked elevated SAM levels with lyphilisates of RLSS overexpressing strains (experiment-E; s2, s4, s5 in Figure 18). However, no differences in FDA levels were detected under these reaction conditions in the sample with RLSS and FlA compared to the sample where only FlA was present (s4 compared to s3 in Figure 18). The strain devoid of FlA showed an unidentified peak at the retention time of FlA (s1 in Figure 18). This observation questions if the peak in the sample with FlA really refers to FDA or if it could also correspond to another unidentified substance, what would falsify the result (s3 in Figure 18).

2.4 Conclusions

2.4.1 Expression studies on the fluorinase

For the engineering of an *E. coli* strain effective in FDA production the heterologous expression of the FlA enzyme was a prerequisite. Our studies showed that FlA has a strong tendency to form inclusion bodies in *E. coli*.

Low expression regimes realized with an arabinose inducible single copy expression construct did not improve the FlA expression. The use of a medium copy vector with the T5 promoter resulted in the highest amount of soluble FlA compared to expression from a T7 or TAC promoter.

Intents to improve the expression by lowering the temperature and inducer concentrations failed. The N-terminal His-tag could be excluded as the cause for inclusion body formation since an untagged version of the protein formed predominantly inclusion bodies as well.

The effects of the co-expression of the *E. coli* chaperone LbpB were inconclusive. One experiment showed an improved soluble FlA expression upon LbpB co-expression, but in the repetition of the experiment no effect on FlA expression was observed. The co-expression of $DsbC_{trunc}$ did not result in more soluble FlA either.

RLSS co-expression showed a marginal effect on FlA expression. A slight increase of overall protein titers as well as a slight improvement on the soluble expression were detected, but need to be confirmed in further experiments.

Codon optimization was the most effective way to improve soluble FlA expression. The harmonization of rare codons for the expression in *E. coli*, especially within the translation initiation region of the *flA* gene turned out to be crucial.

2.4.2 FDA conversion in *E. coli*

Attempts to produce FDA *in vivo* in *E. coli* expressing the fluorinase FlA were not successful. We presume that intracellular fluoride concentrations were too low to allow for detectable FDA production. Fluoride is toxic to the *E. coli* cell, which is why it is actively excluded from the intracellular space. Although we confirmed that some active FlA enzyme was expressed, we suspect the FlA expression levels to be too low from the single copy vector used in this experiment.

In the experiments with lyophilized cells we used a medium copy vector with a T5 promoter because it showed the most promising FlA expression pattern at this time of the study (see chapter 3.1). We speculated that fluoride can unconditionally pass the presumably permeabilized membrane of the lyophilized cells. Bioconversions with lyophilized cells supplemented with 2 mM SAM and 200 mM fluoride produced reasonable levels of FDA (experiment-A). Moreover, experiments showed that higher concentrations of lyophilized cells in the reactions correlated with elevated FDA levels (experiment-B: 2 mM SAM, 200 mM fluoride). Differences in FDA levels between reactions with lyophilisates containing FlA alone or RLSS and FlA when supplemented with 2 mM SAM and 200 mM

fluoride were too marginal for interpretation from the single experiments presented in this study (experiment-A and -B).

However, HPLC analysis of lyophilized cells processed equally like reactions without addition of any substrate, *i.e.* blanks, showed that overexpression of RLSS increased SAM levels. Co-expression of RLSS with FlA had a positive effect on FDA synthesis when reaction mixes with lyophilized cells only contained 200 mM fluoride and no SAM (experiment-C). SAM synthases, RLSS as well as the *E. coli* endogenous SAM synthase MetK, use ATP and L-methionine as substrates. L-methionine is a byproduct of FDA synthesis and *via* its recycling to produce SAM it is retrieved from the reaction equilibrium in favor of FDA production. Addition of 20 mM L-methionine to reaction mixes containing 200 mM fluoride did not show an improvement in FDA synthesis, but elevated levels of SAM with the lyophilisate of the strain overexpressing only RLSS (experiment-D). We supposed that the addition of L-methionine pushed the reaction equilibrium of FDA synthesis into the direction of the educts SAM and fluoride. The ATP concentration in the reaction mixes was suspected to be too low for an efficient conversion of the supplemented L-methionine into SAM. Indeed, the addition of 20 mM of both SAM synthase substrates, ATP as well as L-methionine, to the reactions resulted in elevated SAM levels (experiment-E). Also FDA levels increased under these reaction conditions. However, lyophilisates of strains co-expressing RLSS and expressing FlA alone showed an equal FDA synthesis capacity when 20 mM of ATP and L-methionine were supplemented to the reaction.

Taken together, our results indicate that reactions with lyophilisates of strains co-expressing RLSS with FlA showed increased FDA synthesis by providing elevated SAM levels. This effect could only be observed, when no external SAM or SAM synthase substrates, ATP and L-methionine, were supplemented to the reaction mixes.

2.5 **Outlook**

As expression of FlAco2 was most promising, further work can be pursued with this nucleotide sequence. Further codon optimization by using even more rare codons at the start of translation might be beneficial.

Repetition of the LbpB co-expression studies with FlA could verify if it enhances soluble FlA expression or not.Co-expression of the chaperones LbpB and $DsbC_{trunc}$ should be tested for the other pathway enzymes as well. Especially for the 4FT transaldolase predicted to form seven disulfide bonds (DIANNA1.1 [21]) co-expression of $DsbC_{trunc}$ might facilitate protein expression.

Concerning *in vivo* FDA production next experiments should use the pQE80L-His-FlA or pQE80L-His-FlAco2 constructs for enhanced FlA expression. As RLSS co-expression proved to enhance FDA production in the lyophilized cell assay, it should also be tested in the *in vivo* approach. Elevation of intracellular SAM levels by overexpression of RLSS was already reported [23, 24]. This observation was supported in this study by the analysis of lyophilized cells without addition of any substrates that showed increased levels of SAM when RLSS was overexpressed. RLSS co-expression with FlA together with the addition of fluoride in excess and lowering the pH for increased fluoride assimilation might facilitate FDA production *in vivo*.

Results for the FDA production with lyophilized cells need to be confirmed by technical as well as biological replicates. An internal standard, like fluorouracil, as well as a FDA calibration curve would enhance the reliability of the data. Because peaks of unknown identity were detected at the retention time of FDA, the HPLC method should be optimized. This would be especially advisable for the reactions where ATP and L-methionine were added. Instead of the isocratic elution a gradient could be applied to help separation of the different peaks. A gradient elution might also yield reproducible retention times of FDA and SAM. As well, it could shorten the run times.

The established assay with lyophilized cells presents a platform to evaluate the next pathway steps succeeding the fluorinase reaction. However, we could only find literature that applied 19F-NMR to analyze other pathway intermediates than FDA. It has to be considered that 19F-NMR is not very sensitive and that no standards of other pathway intermediates than FDA are available to us. Products of the 4FT pathway are not commercially available. Customized chemical synthesis might be possible, but already failed for 4FT (co-operation with G. Strohmaier, ACIB, Austria).

Concerning the last pathway step to give 4FT from L-threonine and fluoroacetaldehyde by the action of the 4FT transaldolase, a surrogate enzyme that uses another substrate than L-threonine should be explored. As described in the background, the residue-specific approach to incorporate amino acid analogs cannot be applied in the presence of the canonical amino acid to be replaced. The L-threonine aldolase from *Pseudomonas putida* uses L-glycine to synthesize L-threonine. It can be heterologously expressed in *E.* coli and was already shown to produce 4FT from fluoroacetaldehyde and L-glycine in an *in vitro* reaction [29].

2.6 **Materials and Methods**

Chemicals were purchased from Carl Roth, Karlsruhe, Germany unless indicated otherwise.

2.6.1 **Plasmid construction and codon optimization**

Enzymes for cloning and PCR were obtained from Thermo Fisher Scientific (Waltham, MA). DNA fragments were fused by T4 DNA Ligase. PCRs were performed using Phusion® High-Fidelity DNA Polymerase. PCR primers were ordered from IDT Inc. (Coralville, IA). Ligation and Gibson reaction mixes were transformed into *E. coli* TOP10 F' F´{*lac*Iq, Tn*10*(TetR)} *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara leu*) 7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G) (Life Technologies, Carlsbad, CA).

For pKLJ12-His-flA the *His-flA* sequence was PCR amplified from pET28a(+)-His-flA (kindly provided by Margit Winkler, ACIB GmbH, Austria) with the forward primer 5'-CTA TGA GAG GAT CGC ATC ACC ATC ACC ATC ACA GCG GCC GCA TGG CTG CGA ACA GCA CAC-3' and the reverse primer 5'-CTG CAG GTC GAC TCT AGA GGA TCC CCG GGT ACC ATG GTG AAT TCT CAG CGG GCC TCG ACC-3'. The amplicon was Gibson cloned [30] into pKLJ12 [25] cut with NotI & EcoRI.

For Gibson cloning of *His-flA* into pMS470 [31] the vector backbone was digested with EcoRI and HindIII. *flA* was PCR amplified from pKLJ12-His-flA with the respective overhangs provided by the forward primer 5'-ATT GTG AGC GGA TAA CAA TTT CAC ACA GGA AAC AGA ATT CAT TAA AGA GGA GAA ATT AAC-3' and the reverse primer 5'-CAG GCT GAA AAT CTT CTC TCA TCC GCC AAA ACA GCC AAG CTT TCA GCG GGC CTC-3'.

For Gibson cloning of *His-flA* into into pQE80L (Qiagen, Venlo, Netherlands) the vector backbone was cut with BamHI and HindIII. *flA* was PCR amplified from pKLJ12-His-flA with the respective overhangs provided by the forward primer 5'-CTA TGA GAG GAT CGC ATC ACC ATC ACC ATC ACG GAT CCA TGG CTG CGA ACA GCA CAC G-3' and the reverse primer 5'-CTG GAT CTA TCA ACA GGA GTC CAA GCT CAG CTA ATT AAG CTT TCA GCG GGC CTC GAC CC-3'.

For p15aTAC-dsbC the kanamycin resistance gene, the p15a origin of replication and the rrnB terminator were obtained from pBP226 by restriction digest with NotI and EcoRI. The *lacI* gene, the TAC promoter and the truncated version of the *dsbC* gene for cytosolic expression were obtained by PCR amplification from pCm470-DsbC-PLE3-C8P (kindly provided by Christine Winkler, TU Graz, Austria) with the forward primer 5'-CTT GCG GCA GCG TGA AGC TTA TCG ATG CGG CCG CTC ACT GCC CGC TTT CCA GTC-3' and the reverse primer 5'-CTG TTT TAT CAG ACC GCT TCT GCG TTC TGA TTT AAT CGA ATT CTT ATT TAC CGC TGG TCA TTT TTT GG-3' with the overhangs for Gibson assembly.

The *lbpB* sequence was amplified by colony PCR from *E. coli* BL21(DE3) with the forward primer 5'-GTG TGG AAT TGT GAG CGG ATA ACA ATT TCA CAC AGG AAA GGG ATC CTT TAA CTT TAA GAA GGA GAT CAT ATG CGT AAC TTC GAT TTA TCC-3' and the reverse primer 5'-TTT TAT CAG ACC GCT TCT GCG TTC TGA TTT AAT CGA ATT CTT AGC TAT TTA ACG CGG GAC-3'. The

amplicon contained overhangs to Gibson clone it into the p15aTAC backbone, which was obtained by restriction digest of p15aTAC-dsbC with the double cutter EcoRI.

For p15aTAC-lbpB-dsbC lbpB was amplified by colony PCR from *E. coli* BL21(DE3) with the forward primer 5'-GTG TGG AAT TGT GAG CGG ATA ACA ATT TCA CAC AGG AAA GGG ATC CTT TAA CTT TAA GAA GGA GAT CAT ATG CGT AAC TTC GAT TTA TCC-3' and the reverse primer 5'-ATT GCC GCG TCA TCC ATT ATA TCT CCT TCT TAA AGT TAA ACT CGA GTT AGC TAT TTA ACG CGG GAC-3'. The PCR fragment was Gibson cloned into the p15aTAC-dsbC digested with BamHI.

We had the first version of the codon-optimized *flA* (*flAco1*) synthesized as a gBlock (Integrated DNA Technologies, Coralville, IA). For codon optimization we used the program Gene Designer from DNA2.0 Inc. and optimized the *flA* gene sequence according to the codon usage table of *E. coli* Class1 (Table S1). *S. cattleya* uses rare codons for *flA*, what is shown for the N-terminal nucleotide sequence in Figure 3. In the optimized sequence for *E. coli* we manually placed rare codons at the same positions were *S. cattleya* uses rare codons (Table 1). *flAco1* was amplified with the forward primer 5'-CAA TTA TAA TAG ATT CAA TTG TGA GCG GAT AAC AAT TTC ACA CAG AAT TCA TTA AAG AGG AGA AAT TAA CTA TGG CGG CTA ACT CCA CAC-3' and the reverse primer 5'-CTG GAT CTA TCA ACA GGA GTC CAA GCT CAG CTA ATT AAG CTT TCA ACG AGC TTC AAC ACG-3' to provide the nucleotide overhangs for Gibson assembly into pQE80L cut with HindIII and EcoRI.

The second codon-optimized *flA* (*flAco2*) was obtained by a QuikChange® mutagenesis approach [32] with *flAco1* as the PCR template. The forward mutagenic primer was 5'-CTA TGG CCG CTA ACT CCA CAC GGC GTC CCA TTA TTG CAT TTA TGT CTG ACC TGG GTA CC-3' and the reverse primer 5'-CAA TAA TGG GAC GCC GTG TGG AGT TAG CGG CCA TAG TTA ATT TCT CCT CTT TAA TGA ATT CTG TGT GAA ATT GTT ATC CGC TCA C-3'.

RLSS was codon-optimized according to the codon usage table of *E. coli* Class1 (Table S1) with the program Gene Designer from DNA2.0 Inc. and synthesized as a gBlock (Integrated DNA Technologies, Coralville, IA). It was cloned into pJet1.2 with the CloneJET PCR Cloning Kit (Life Technologies, Carlsbad, CA) according to the manufacturer's protocol. The RLSS was amplified by PCR from pJet1.2-RLSS with the forward primer 5'-GCA ACT CTC TAC TGT TTC TCC ATA CCC GTT TTT TTG GTA CCG GAA AAA GGA GAT CTG CAT ATG AAT GGC CCT GTT GAC GG-3' and the reverse primer 5'-GGC AAA TTC TGT TTT ATC AGA CCG CTT CTG CGT TCT GAT TTA ATC TTA ATT AAC TAA AAA ACC AGT TTC TTC GGC ACT TC -3' with overhangs to clone it into p15aARA ([33]) cut with PacI and KpnI by Gibson assembly.

2.6.2 **FlA expression and preparation of cleared cell lysates**

FlA was expressed in the threonine auxotrophic *E. coli* BL21 Gold(DE3) (Δ*thrC*::0 F^- *omp*T *hsd*S_B(r_B^-, m_B^-) *gal dcm* (DE3)) (kindly provided by Niklaus Anderhuber, ACIB GmbH, Austria) for the experiments conducted in chapter 2.1 and 2.2.1. For FDA conversion with lyophilized cells described in chapter 2.2.2 *E. coli* BL21 Gold(DE3) was used. Cells were cultured in full medium (LB medium Lennox) containing 50 mg/L to maintain pET28a(+)-His-flA; 50 mg/L ampicillin (Sigma-Aldrich, St.

Louis, MO) to maintain pKLJ12-His-flA; or 100 mg/L ampicillin to maintain pMS470-His-flA and pQE80L-His-flA. FlA expression was induced at a D_{600} of 0.5-0.8. For expression of FlA from pKLJ12-His-flA we used 10 mg/L or 100 mg/L arabinose (Sigma-Aldrich, St. Louis, MO) for induction. For pET28a(+)-His-flA, pMS470-His-flA, pQE80L-His-flA expression was induced with 0.5 mM IPTG (Biosynth, Staad, Switzerland) unless indicated otherwise. Cells were harvested after 5 h or 16 h of induction.

For chaperone co-expression, the culture medium was additionally supplemented with 50 mg/L kanamycin. FlA and chaperone expression were induced simultaneously by the addition of 0.5 mM IPTG at a D_{600} of 0.5-0.8. Cells were harvested after 5 h or 16 h of induction.

Codon-optimized fluorinase variants were induced with 0.1 mM IPTG for 3.5 h (FlAco2) and 5 h (FlAco1) at 28 °C.

For RLSS co-expression experiments RLSS was induced at a D_{600} of 0.4-0.5 with 0.4% arabinose at 28 °C. At a D_{600} of 0.9 or 4.5 the expression of FlA was induced with 0.1 mM IPTG. Cultures were harvested after 4.5 h or 18 h FlA induction.

Cells were lysed with BugBuster™ Protein Extraction Reagent (Novagen Merck Chemicals Ltd., Nottingham, UK) according to the manufacturer's protocol unless indicated otherwise.

2.6.3 **SDS gels and Western blot**

Coomassie staining of SDS gels was done with a solution containing 2.5 g Brilliant Blue G250 (Sigma-Aldrich, St. Louis, MO), 7.5% (v/v) acetic acid and 50% (v/v) ethanol in water. SDS gels were stained for at least 30 min, rinsed with water and destained with 7.5% (v/v) acetic acid and 20% (v/v) ethanol aqueous solution.

SDS gels were either purchased from Life Technologies (Carlsbad, CA) as NuPAGE® Novex® 4-12% Bis-Tris protein gels or home-made SDS Laemmli gels [34].

For immunodetection of His-tagged proteins all reagents were purchased from Life Technologies (Carlsbad, US). Cells were lysed with CelLyticTM B 2x (Sigma-Aldrich, St. Louis, US) according to the manufacturer's protocol. We used a 6x-His epitope tag antibody from mouse, a goat anti-mouse IgG + IgM (H+L) secondary antibody horseradish peroxidase conjugate and the SuperSignal® West Dura Extended Duration Substrate for chemiluminescent detection.

2.6.4 ***in vivo* synthesis of FDA**

For *in vivo* conversions the threonine auxotrophic *E. coli* BL21 Gold(DE3) strain (Δ*thrC*::0 F^- *ompT* *hsd*S_B(r_B^-, m_B^-) *gal dcm* (DE3)) overexpressing the fluorinase from the pKLJ12-His-flA plasmid was cultured in 350 mL either full (LB medium Lennox, Carl Roth, Karlsruhe, Germany) or minimal medium (Table S2) containing 25 mg/L ampicillin. FlA expression was induced at a D_{600} of 0.5 with 100 mg/L arabinose. After 3 h of FlA expression at 25°C KF was added to final concentrations of either 2 mM for the full medium or of 10 mM for the minimal medium culture and the temperature was elevated to 30 °C. After overnight conversion in full medium or after three days conversion in minimal medium, the

reaction was stopped by freezing 30 mL of the cultures with liquid nitrogen. For cell disruption thawed cell cultures were incubated with 100 µg/ml lysozyme (Carl Roth, Karlsruhe, Germany), 4 µg/ml DNAse (Sigma-Aldrich, St. Louis, MO) and 10 µg/ml RNAse (Sigma-Aldrich, St. Louis, MO) for 30 min on ice before sonication (output control 8, duty cycle 70 - 80%, 6 min) and cell debris was removed by centrifugation (40 000 g, 4 °C, 40 min). Protein was precipitated at 90 °C for 15 min and removed by centrifugation (75 600 g, 4 °C, 40 min). The obtained supernatant was transferred to round bottom flasks, frozen in liquid nitrogen and lyophilized. The lyphilisate was dissolved in 2 mL 100 mM TRIS/Cl, pH 7.45, centrifuged (13 000 g, 5 min) and the supernatant was analysed by HPLC as described in chapter 5.5.3.

2.6.5 Conversions with lyophilized cells for FDA synthesis

2.6.5.1 Preparation of lyophilized cells

For the preparation of lyophilized cells 500 mL cultures were harvested after overnight FlA and RLSS expression as described in chapter 5.2. Cell pellets were washed with 250 mL ice-cold 50 mM Tris/Cl buffer, pH 7.8. Cells were resuspended in 50 mL ice-cold ddH_2O, transferred into pre-chilled round bottom flasks and chilled in liquid nitrogen with slight shaking. Cells were freeze dried over night and stored at -20° C.

2.6.5.2 FDA assay using lyophilized cells

All experiments described in this chapter were carried out in a total reaction volume of 500 µl with 200 mM potassium fluoride.

In experiment-A we applied 100 mg/mL lyophilized cells (350 µl of a 143 mg/mL stock solution in 50 mM Tris/Cl buffer, pH 7.8), 200 mM potassium fluoride (50 µl of a 2 M stock solution in ddH_2O) and 2 mM SAM (100 µl of a 10 mM stock solution in 1 mM HCl). In experiment-B with different amounts of lyophilized cells we prepared a 150 mg/mL lyophilized cell stock solution in 50 mM Tris/Cl buffer, pH 7.8. The stock solution was further diluted with 50 mM Tris/Cl buffer, pH 7.8 to give final concentrations of 100 mg/mL, 75 mg/mL, 50 mg/mL, 25 mg/mL, 10 mg/mL, and 5 mg/mL of lyophilized cells in the final reaction mix like described above (200 mM fluoride and 2 mM SAM).

In experiment-C, -D and -E that omitted SAM, the reaction mixture contained 92 mg/mL of lyophilized cells (320 µl of a 143 mg/mL stock solution in 50 mM Tris/HCl buffer, pH 7.8) and 200 mM potassium fluoride (50 µl of a 2 M stock solution in ddH_2O). Experiment-C did only contain lyophilized cells and 200 mM potassium fluoride. Experiment-D contained 20 mM L-methionine (30 µL of a 335 mM stock solution in 1 mM HCl) in addition to 200 mM potassium fluoride. Experiment-E contained 20 mM ATP (100 µL of a 100 mM stock solution in ddH2O) as well as 20 mM L-methionine (30 µL of a 335 mM stock solution in 1 mM HCl) in addition to 200 mM potassium fluoride.

All reactions were incubated for 1 h at 37 °C with shaking and stopped by protein precipitation at 95 °C for 5 min. The protein precipitate was removed by centrifugation and the supernatant was filtered using a MultiScreen® Filter Plate (Merck Millipore, Billerica, MA.

2.6.5.3 **HPLC-UV measurements and analysis of FDA**

HPLC analysis was conducted as described in Schaffrath *et al.* [8] with modifications. 10 µl of samples were used for analysis by HPLC (unless indicated otherwise) with a ZORBAX Eclipse XDB-C18 column (5 µm, 4.6 x 150 mm cart column) from Agilent Technologies (Santa Clara, CA). Only for analysis of *in vivo* FDA production in full medium (chapter 2.2.2) a Purospher STAR RP-158 LiChroCart 250-4 column (Millipore, Darmstadt, Germany) was used.

The isocratic mobile phase consisted of a 20 mM KH_2PO_4, pH 8.0, buffer and acetonitrile (90:10 v/v). Absorbance was detected at 260 nm at 22°C for 15 min. Retention times of FDA, SAM and ATP were determined by HPLC analysis of pure substances. Moreover, processed reactions with lyophilized cells producing FDA were spiked with FDA and lyophilized cells were spiked with FDA and SAM to verify retention (data not shown).

Data analysis and visualization was done by Microsoft Excel (Microsoft, Redmond, WA). The area of the relevant peaks was automatically integrated by the HPLC software for experiment-A and -B (for experimental details see chapter 5.5.2).

Blanks were subtracted from integrated FDA peak areas in experiment-A and -B. In experiment-A lyophilized cells of the corresponding strain without addition of FlA or RLSS substrate served as blanks. In experiment-B the reaction with the lyophilisate of the strain neither expressing FlA nor RLSS (s1) served as blank for subtraction of the reaction with lyophilisate of the strain only overexpressing FlA (s3). The reaction with the lyophilisate of the strain only overexpressing RLSS (s2) served as a blank for subtraction of the reaction with lyophilisate of the strain co-expressing FlA with RLSS (s4 and s5).

2.7 **List of abbreviations**

ATP: adenosine triphosphate; FDA: 5'-deoxy-5'-fluoroadenosine; FlA: fluorinase; 4FT: 4-fluorothreonine; FAA: fluoro amino acid; His-tag: hexahistidine tag; RLSS: SAM synthase from the liver of *Rattus norvegicus*; SAM: S-(5'-Adenosyl)-L-methionine

2.8 **Figures and illustrations**

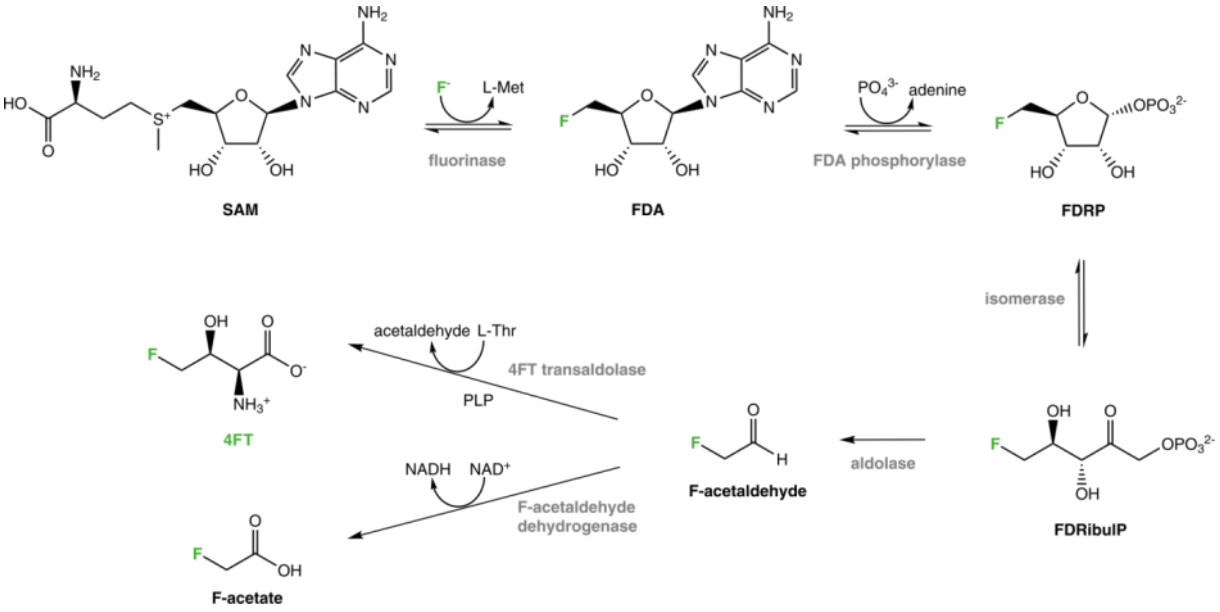


Figure 1: Biosynthesis pathway for 4-fluoro-L-threonine (4FT) and fluoroacetate (F-acetate) in *S. cattleya*

SAM, S-adenosyl-L-methionine; FDA, 5'-deoxy-5'-fluoroadenosine; FDRP, 5-deoxy-5-fluororibose phosphate; FDRibulP, 5-deoxy-5-fluororibulosephosphate; F-acetaldehyde, fluoroacetaldehyde; F-acetate, fluoroacetate.

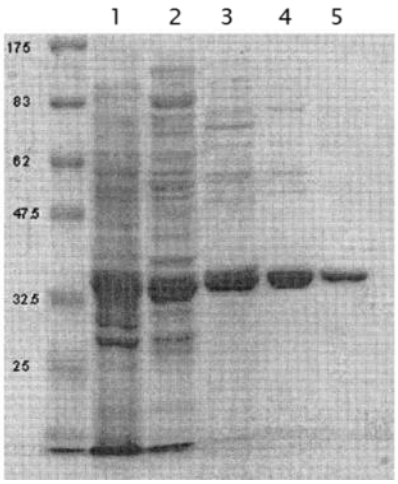


Figure 2: Purification of the fluorinase from *S. cattleya* (reproduced from Schaffrath *et al.* [8])

1, crude cell-free extract; 2, ammonium sulfate precipitate fractionation; 3, hydrophobic interactive chromatography, 4, gel filtration; 5, anion exchange chromatography

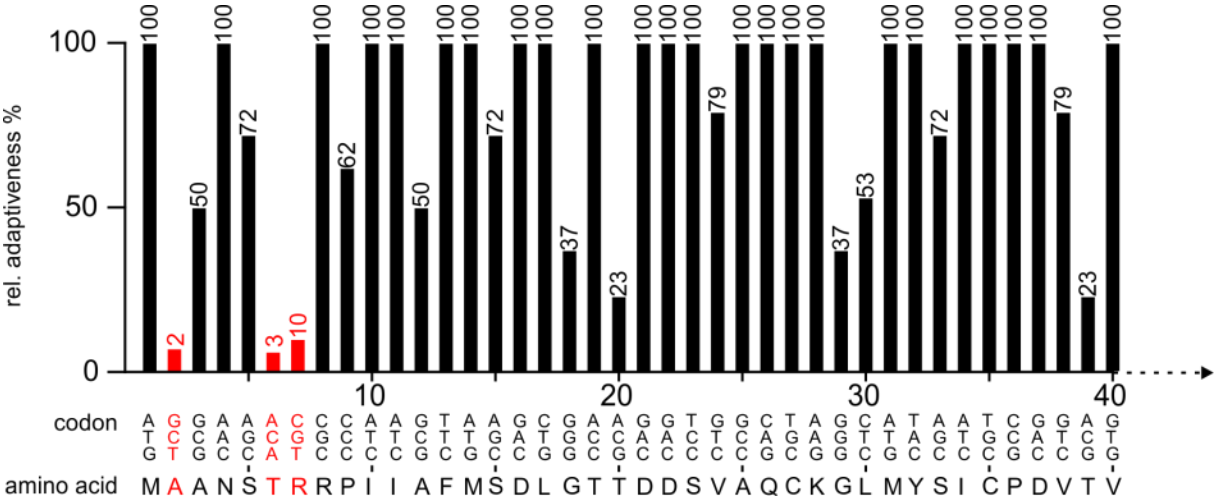


Figure 3: Codon usage for the fluorinase in its natural host *S. cattleya*

For calculation of the relative adaptiveness the most frequently used codon is set to 100 and all other values are calculated using the rule of proportion. The image was obtained from the online program Graphical Codon Usage Analyser (http://gcua.schoedl.de/) [35]. The used *S. cattleya* codon usage table was from the Codon Usage Database Kazusa. http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=29303&aa=1&style=N. Accessed 29 September 2014.

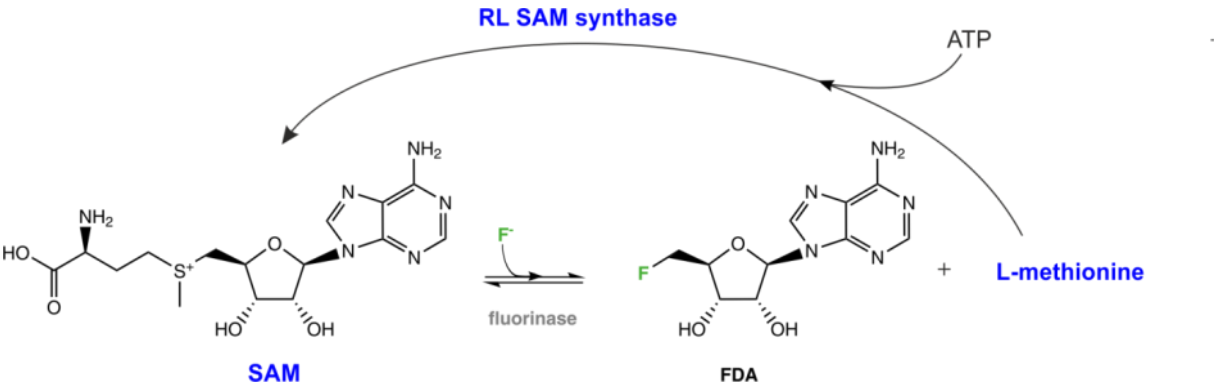


Figure 4: Fluorinase reaction and substrate recycling

Fluorinase reaction to 5'-deoxy-5'-fluoroadenosine (FDA) and recycling of L-methionine to S-adenosyl-L-methionine (SAM) by the action of overexpressed rat liver (RL) SAM synthase

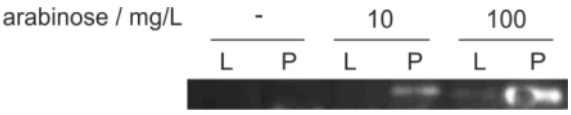


Figure 5: Western blot of the His-tagged fluorinase (FlA) expressed from the single copy vector pKLJ12-flA

Samples were collected after 5 h induction of FlA with arabinose. The amount loaded equals 0.05 D_{600} units. L refers to the cleared lysate fraction, P to the insoluble protein (pellet) fraction, M to protein molecular weight marker.

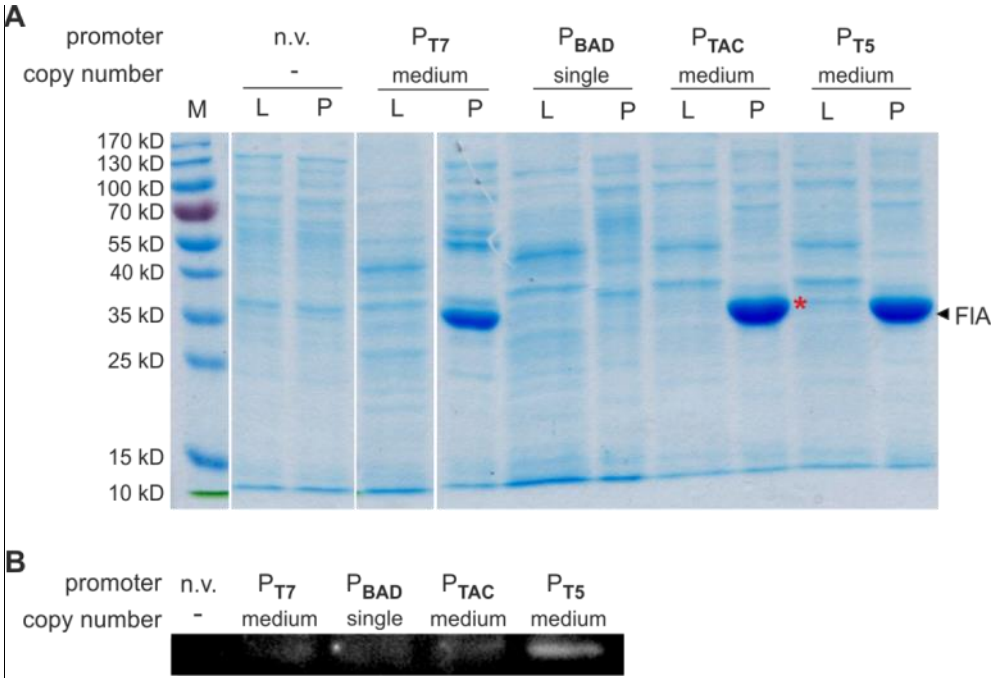


Figure 6: SDS gel (A) and anti-His Western blot (B) of strains overexpressing FlA from different vector constructs

n.v. refers to the untransformed strain not haboring any expression vector; P_{T7} to the T7 promoter on the pET28a(+)-His-flA expression construct; P_{BAD} to the araBAD promoter on pKLJ12-His-flA; P_{T5} to the T5 promoter on pQE80L-His-flA; P_{TAC} to the TAC promoter on pMS470-His-flA; L, cleared lysate fraction; P, insoluble protein (pellet) fraction; and M refers to protein molecular weight marker. The amount loaded equals 0.1 D_{600} units. The calculated molecular weight of His-tagged FlA is 34 kDa. A home-made 12% SDS gel is shown (see chapter 5.3 for details).

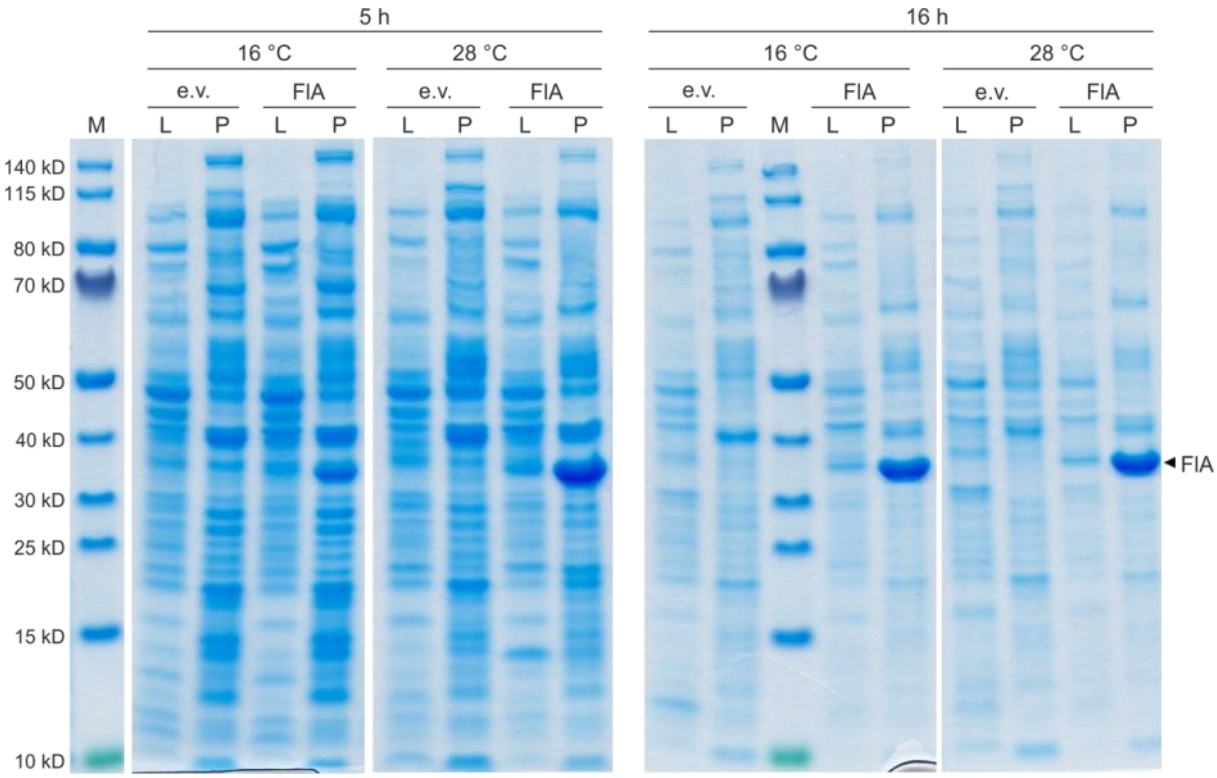


Figure 7: Expression temperature and induction time did not influence the soluble expression of His- FlA

Expression at different temperatures (16 °C and 28 °C) and induction times (5 h ad 16 h) is shown. e.v. refers to empty vector control; M, protein molecular weight marker; L, cleared lysate; and P refers to insoluble protein (pellet) fraction. The amount loaded equaled 0.05 D_{600} units. The calculated molecular weight of His-tagged FlA is 34 kDa. Commercial 4-12% Bis-Tris gels are shown (see chapter 5.3 for details).

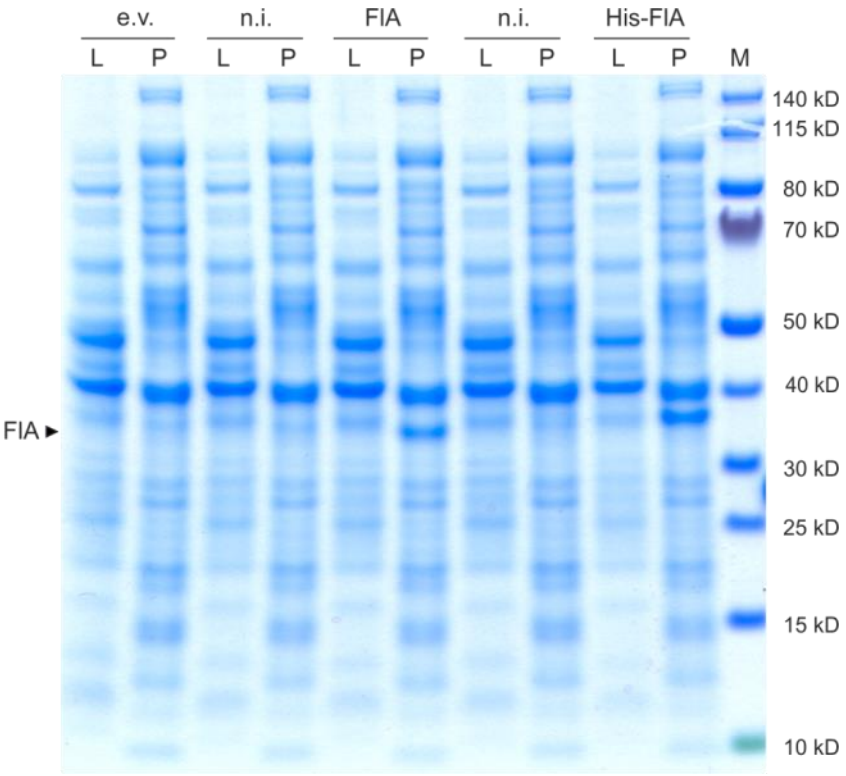


Figure 8: His-tag on FlA and a lowering of the IPTG inducer concentration did not improve soluble FlA expression

Expression was induced with 10 µM IPTG at 16°C for 16 h. e.v. refers to empty vector control; n.i., non-induced sample; M, protein molecular weight marker; L, cleared lysate and P refers to insoluble protein (pellet) fractions.The amount loaded equals 0.1 D_{600} units. The calculated molecular weight of His-tagged FlA and of FlA is is 34 kDa and 32 kDa, respectively. Commercial 4-12% Bis-Tris gels are shown (see chapter 5.3 for details)..

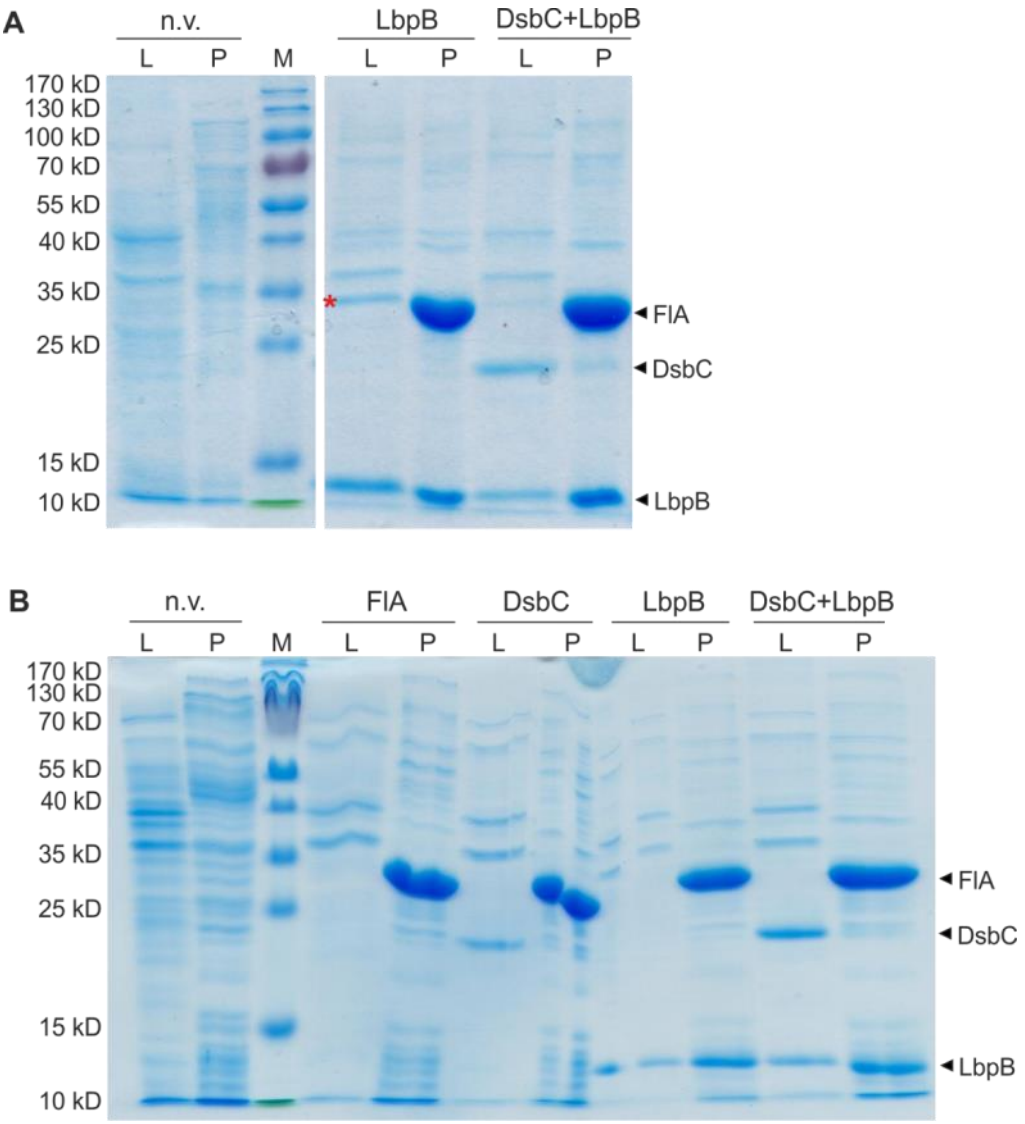


Figure 9: Influence of co-expression of chaperones on soluble expression of His-tagged FlA

Co-expressed chaperones were a truncated version of a disulfide bond isomerase (DsbC) and a small heat shock protein (LbpB). The red star in panel A indicates improved soluble FlA expression with co-expression of LbpB. This effect was not visible for the repetition of the experiment depicted in panel B. n.v. refers to the untransformed strain not haboring any overexpression vector; L, the cleared lysate fraction; P, the insoluble protein (pellet) fraction and M refers to protein molecular weight marker. The amount loaded equals 0.1 D600 units. The calculated molecular weight of His-tagged FlA, DsbC and LbpB is 34 kDa, 24 kDa and 16 kDa, respectively. Home-made 12% SDS gels are shown (see chapter 5.3 for details.

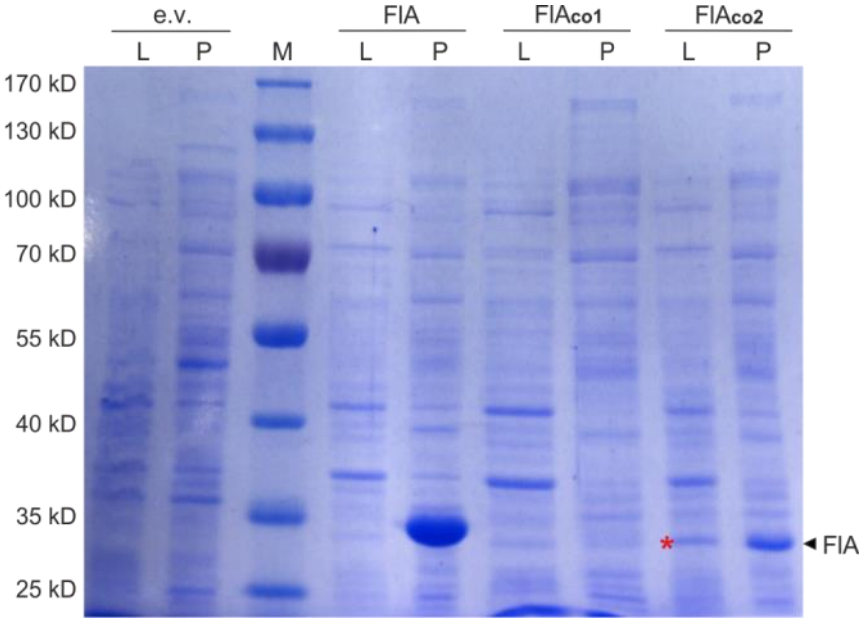


Figure 10: Comparison of FlA expression from canonical (FlA) and codon-optimized FlA sequences (FlAco1 and FlAco2)

The red star indicates improved soluble FlA expression from the codon-optimized FlAco2 sequence (see chapter 5.1 for details on codon optimization). e.v. refers to empty vector control; M, protein molecular weight marker; L, the cleared lysate fraction; P, the insoluble protein (pellet) fraction. The amount loaded equals 0.05 D_{600} units. Home-made 12% SDS gel is shown (see chapter 5.3 for details).

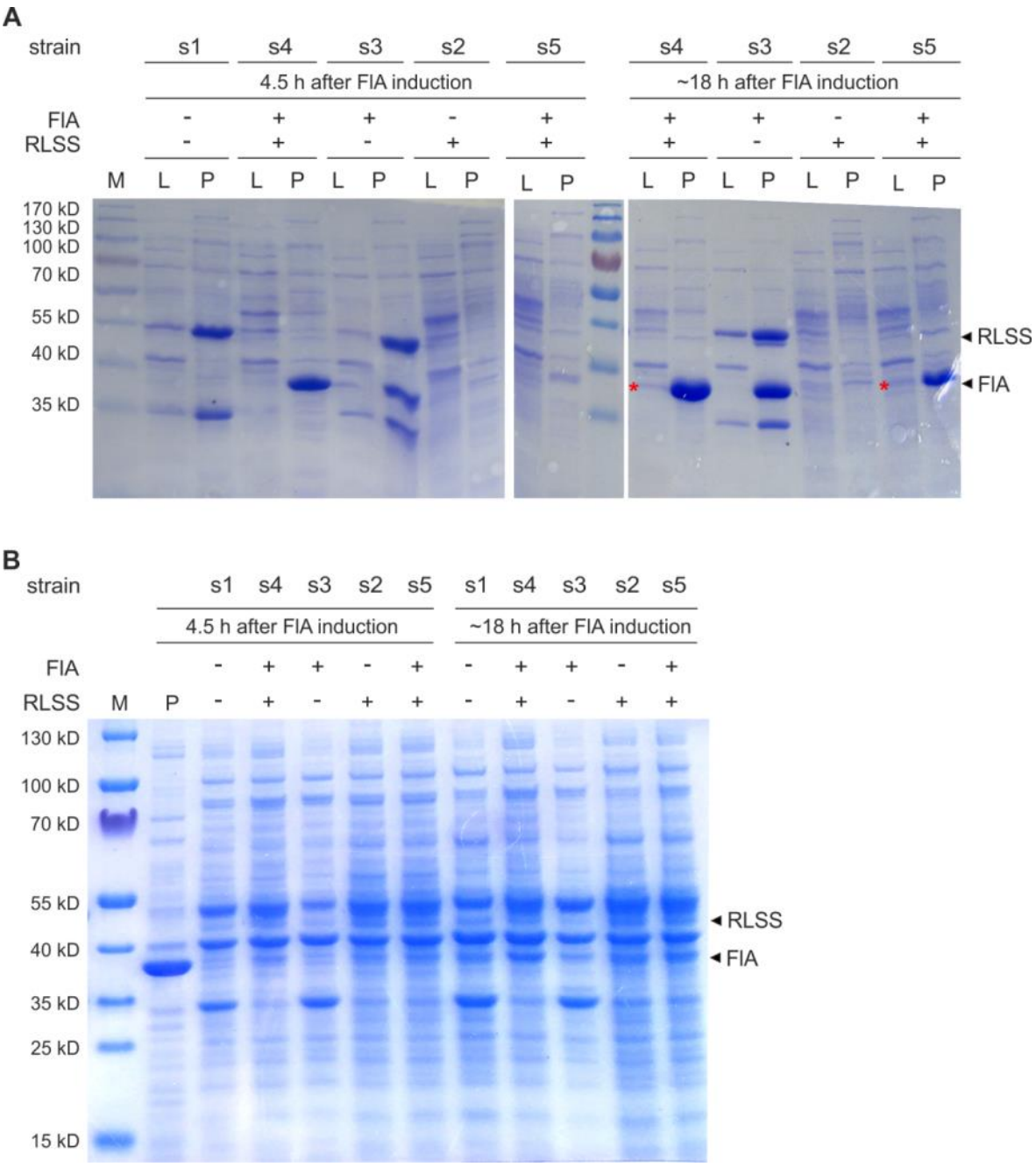


Figure 11: Co-expression of RLSS improved the soluble expression of the FlA

The red star indicates improved soluble FlA expression with RLSS co-expression. Strains overexpressing neither RLSS nor FlA (s1); RLSS and FlA induced at a D_{600} of 0.9 (s4); only FlA (s3); only RLSS (s2); RLSS and FlA induced at a D_{600} of 4.5 (s5) are shown. Panel A shows the cleared lysate (L) and insoluble protein (pellet) fractions and panel B the cleared lysate fractions only. M refers to protein molecular weight marker. The amount loaded equals 0.05 (A) and 0.15 (B) D_{600} units. Home-made 12% SDS gels {(A) and commercial 4-12% Bis-Tris gels (B) are shown (see chapter 5.3 for details).

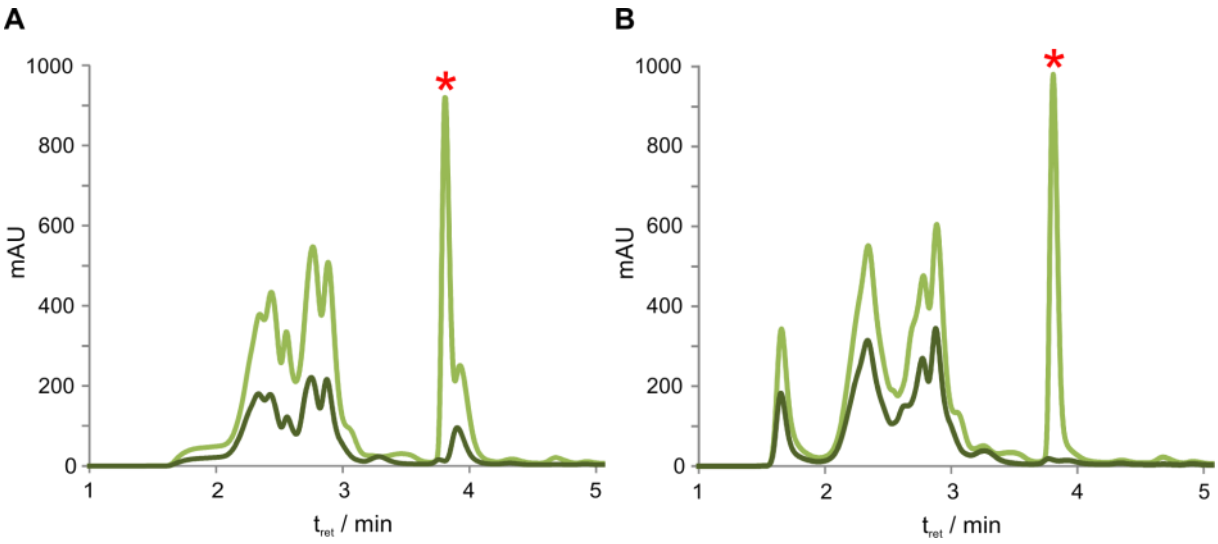


Figure 12: Chromatograms of *in vivo* reactions in full medium do not show any FDA synthesis

.Chromatograms of preparations of the fluorinase overexpressing strain before the addition of 2 mM fluoride (A) and after overnight conversion (B). The cells toghether with the culture medium were prepared for analysis as described in the Materials and Methods section (chapter 5.4). Detection at 260 nm. 5 µL of the suspension (dark green), and 10 µL of the suspension spiked with 50 ng/µL FDA (light green) were injected for HPLC analysis. The FDA peak is indicated with a red star. mAU refers to milli absorbtion units and t_{ret} to the retention time.

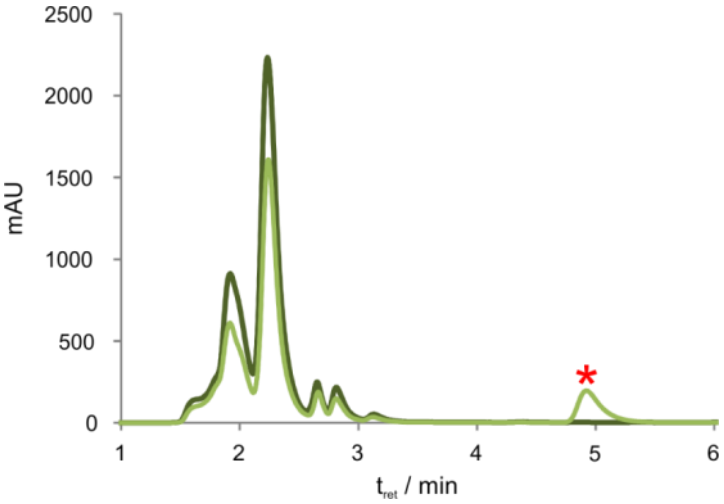


Figure 13: Chromatograms of *in vivo* reactions in minimal medium do not show any FDA synthesis

The cells were cultured for three days in minimal medium with glycerol as the sole carbon source and 10 mM fluoride. The cells together with the culture medium were prepared for analysis as described in the Materials and Methods section (chapter 5.4). Preparations (dark green) were 1:2 diluted and spiked with spiked with 50 ng/µL FDA (light green). The FDA peak is indicated with a red star. mAU refers to milli absorbtion units and t_{ret} to the retention time.

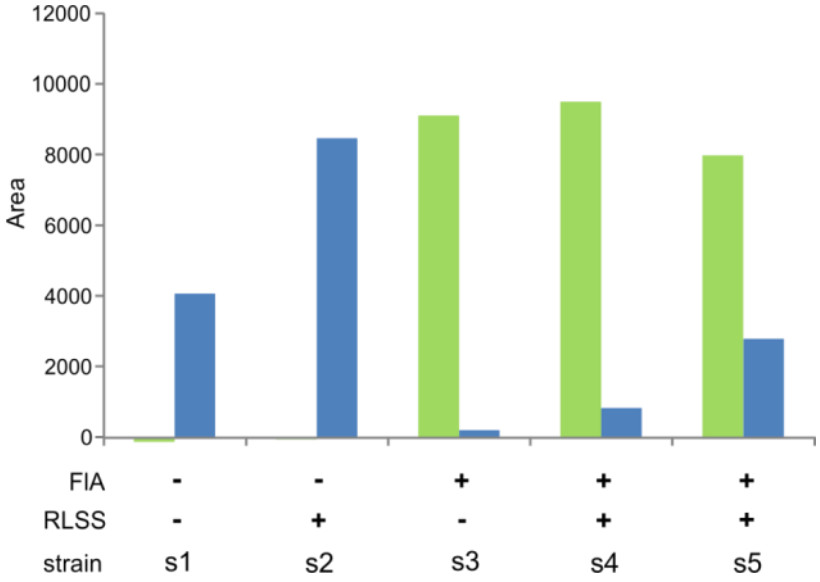


Figure 14: Relative levels of SAM (blue) and FDA (green) in reactions with lyophilized cells expressing different combinations of FIA and RLSS

Integrated areas of the relevant peaks (single experiment) after subtraction of the blanks (for details see chapter 5.5.3) in experiment-A (Table 2) are shown. Detection at 260 nm. FIA refers to the fluorinase and RLSS to the SAM synthase of *R. norvegicus*. Reactions with 200 mM fluoride, 2 mM SAM and 100 mg/mL lyophilized cells of different strains were analysed: the empty vector strain (s1) and the strain only overexpressing RLSS (s2); the strain only overexpressing FIA (s3) and the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 0.9 (s4); the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 4.5 (s5).

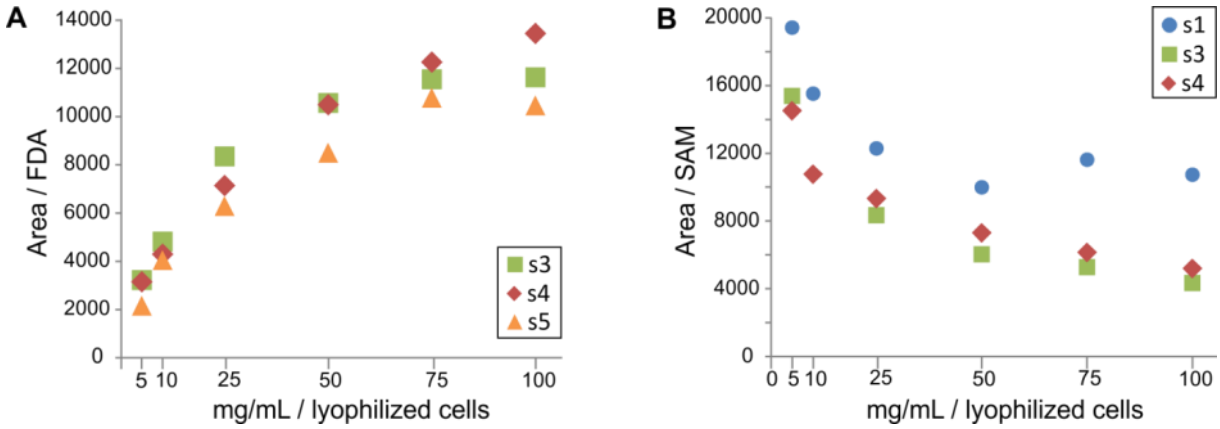


Figure 15: FDA levels increase with the concentration of lyophilized cells (A) and SAM levels decrease with increasing FDA levels (B)

Integrated areas of the relevant peaks (single experiment) of experiment-B (Table 2) after subtraction of blanks (details see chapter 5.5.3) are shown. Detection at 260 nm. Reaction mixes with different concentrations of lyophilized cells (100 mg/mL, 75 mg/mL, 50 mg/mL, 25 mg/L, 10 mg/mL and 5 mg/mL) were supplemented with 200 mM fluoride and 2 mM SAM. In reactions with lyophilized cells the following strains were analysed: the strain neither expressing FIA nor RLSS (s1 in blue, only in panel B); the strain only overexpressing FIA (s3 in green) and the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 0.9 (s4 in red) and the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 4.5 (s5 only in panel A).

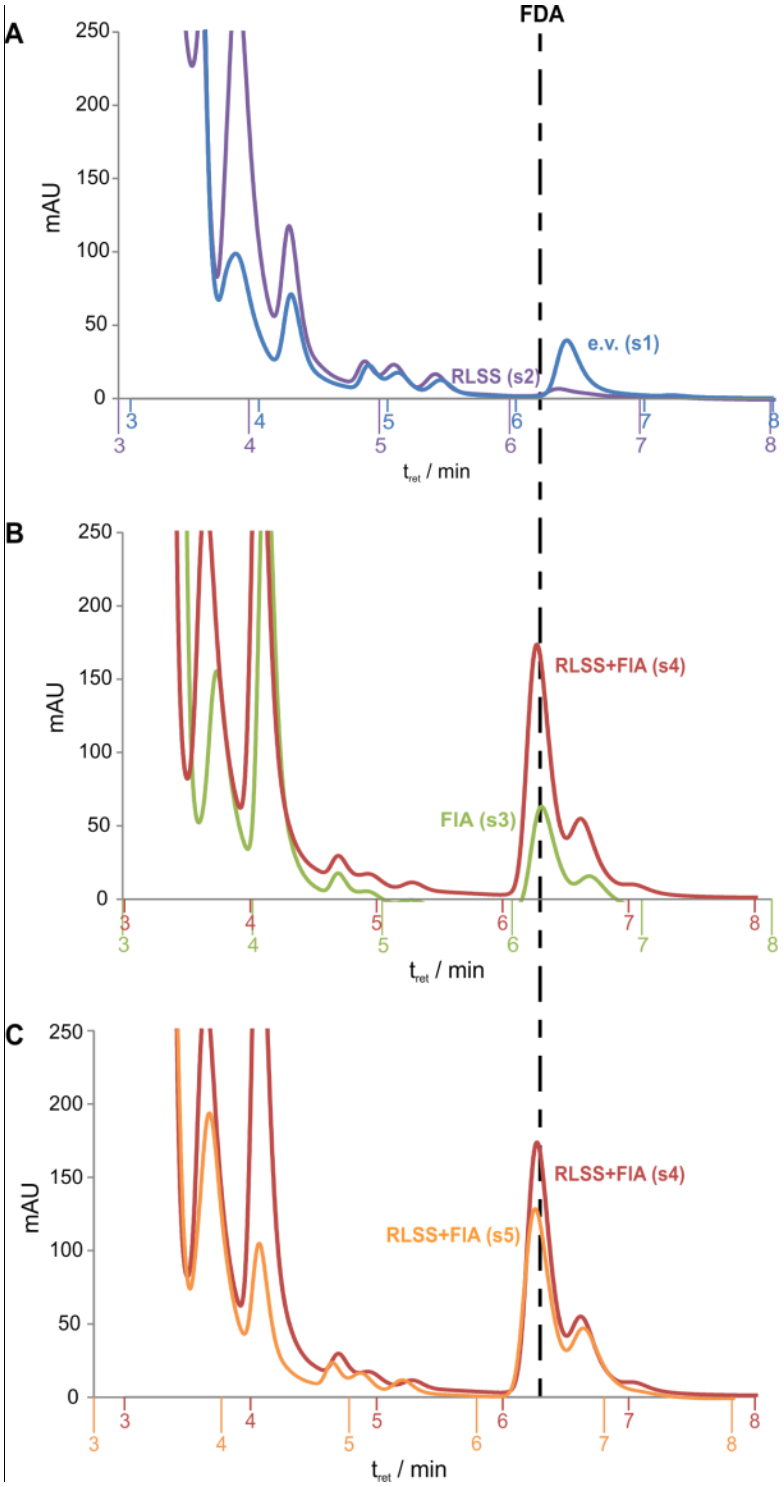


Figure 16: Reactions supplemented with fluoride show elevated FDA synthesis in the presence of RLSS

Overlays chromatograms (detection at 260 nm) of experiment-C (Table 2) are shown. Reactions contained 200 mM fluoride and 100 mg/mL lyophilized cells of the following strains: A, the strain devoid of RLSS and FIA (s1 in blue) and the strain only overexpressing RLSS (s2 in purple); B, the strain only overexpressing FIA (s3 in green) and the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 0.9 (s4 in red); C, the strain overexpressing RLSS and FIA with FIA induced either at a D_{600} of 0.9 (s4 in red) or at a D_{600} of 4.5 (s5 in red). Chromatograms were manually overlayed due to the shifted retention times for FDA and SAM in between different samples (for details see chapter 2.2.2). No SAM was detected in these samples (data not shown). mAU refers to milli absorbtion units and t_{ret} to retention time.

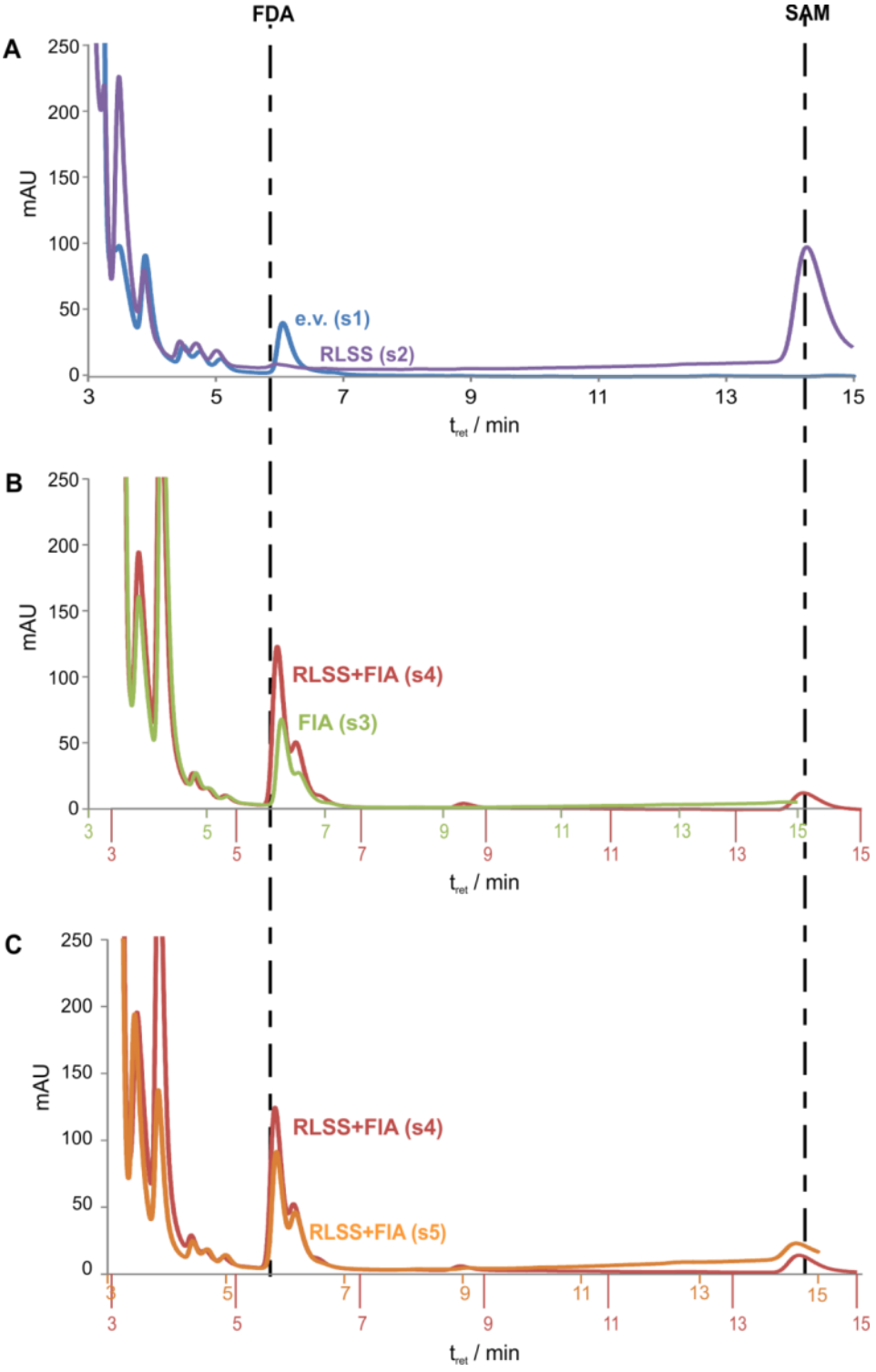


Figure 17: Reactions supplemented with L-methionine and fluoride show elevated FDA and SAM levels in the presence of RLSS

Overlays of chromatograms (detection at 260 nm) of experiment-D (Table 2) are shown. Reactions contained 200 mM fluoride and 100 mg/mL lyophilized cells of the following strains: A, the strain devoid of RLSS and FIA (s1 in blue) and the strain only overexpressing RLSS (s2 in purple); B, the strain only overexpressing FIA (s3 in green) and the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 0.9 (s4 in red); C, the strain overexpressing RLSS and FIA with FIA induced either at a D_{600} of 0.9 (s4 in red) or at a D_{600} of 4.5 (s5 in red). Chromatograms were manually overlayed due to the shifted retention times for FDA and SAM in between different samples (for details see chapter 2.2.2). mAU refers to milli absorbtion units and t_{ret} to retention time.

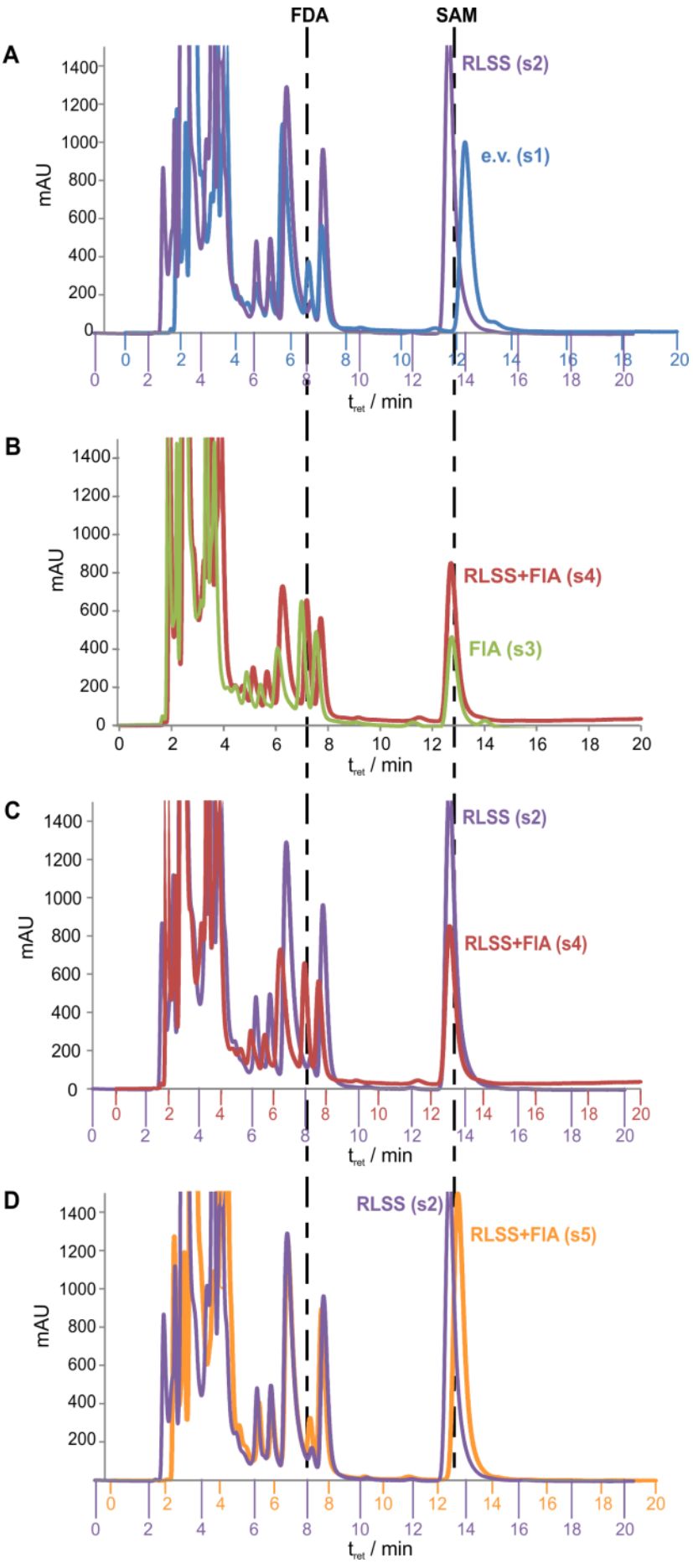

FDA
SAM
A
RLSS (s2)
e.v. (s1)
mAU
t_{ret} / min
B
RLSS+FIA (s4)
FIA (s3)
C
RLSS (s2)
RLSS+FIA (s4)
D
RLSS (s2)
RLSS+FIA (s5)

Figure 18: Reactions supplemented with ATP and L-methionine show elevated levels of FDA and SAM

Overlays of chromatograms (detection at 260 nm) of experiment-E (Table 2) are shown. Reactions contained 200 mM fluoride and 100 mg/mL lyophilized cells of the following strains: A, the strain devoid of RLSS and FlA (s1 in blue) and the strain only overexpressing RLSS (s2 in purple); B, the strain only overexpressing FlA (s3 in green) and the strain overexpressing RLSS and FlA with FlA induced at a D_{600} of 0.9 (s4 in red); C, s2 (in purple) and the strain overexpressing RLSS and FlA with FlA induced either at a D_{600} of 0.9 (s4 in red); D, s2 (in purple) and the strain overexpressing RLSS and FlA with FlA induced at a D_{600} of 4.5 (s5 in red). Chromatograms were manually overlayed due to the shifted retention times for FDA and SAM in between different samples (for details see chapter 2.2.2). mAU refers to milli absorbtion units and t_{ret} to retention time.

2.9 **Tables**

Table 1: Rare codons used for the fluorinase (FlA) in its natural host *S. cattleya* (Sc) and manual codon exchanges for the expression in *E. coli* (Ec)

For *flAco1* underlined codons were exchanged incorrectly and were replace by rare codons in *flAco2*. Codon exchanges for *flAco2* are in bold and italic. Relative adaptiveness (rel. adapt.) values refer to the codon choice for *flA* in its natural host *S. cattleya* and were obtained from the online program Graphical Codon Usage Analyser (http://gcua.schoedl.de/) {Fuhrmann, 2004 #622}. For the calculation of the relative adaptiveness the most frequently used codon is set to 100 and all other values are calculated using the rule of proportion. Freq. refers to the codon usage frequency according to the codon usage table of *E. coli* Class1 (Table S1)

position	amino acid	*S. cattleya* rel. adapt. / %	automatically assigned codon	freq. / % in Sc	*flAco1*	freq. / % in Ec	*flAco2*	freq. / % in Ec
2	A	2	GCA	24	GCG	32	***GCC***	16
6	T	3	ACT	29	ACA	4		
7	R	10	CGT	64	CGC	33	***CGG***	1
9	P	9	CCC	2	CCG	72	***CCC***	2
62	L	2	CTG	77	TTG	5		
69	G	8	GGC	43	GGT	51		
77	Y	4	TAC	65	TAT	35		
98	G	9	GGC	43	GGA	2		
180	R	10	CGC	33	CGT	64		
258	G	9	GGT	51	GGC	43		

Table 2: Components in reactions with lyophilized cells

For details see chapter 5.5.2.

experiment	lyophilized cells / mg/mL	fluoride / mM	SAM / mM	L-methionine / mM	ATP / mM
A	100	200	2	-	-
B	5 to 100	200	2	-	-
C	92	200	-	-	-
D	92	200	-	20	-
E	92	200	-	20	-

2.10 Supporting information

2.10.1 Supplementary Materials and Methods

2.10.1.1 Expression of FlA for purification and *in vitro* synthesis of FDA

Chemicals were purchased from Carl Roth, Karlsruhe, Germany unless indicated otherwise.

FlA was expressed in the threonine auxotrophic *E. coli* BL21 Gold(DE3) (Invitrogen, Carlsbad, US) (Δ*thrC*::0 F^- *omp*T *hsd*S_B(r_B^-, m_B^-) *gal dcm* (DE3). Cells were cultured in 1 L of full medium (LB medium Lennox in the presence of 50 mg/L kanamycin for expression from pET28a(+)-His-flA and in the presence of 50 mg/L ampicillin (Sigma-Aldrich, St. Louis, MO) for expression from pKLJ12-His-flA. For FlA induction at a D_{600} of 0.5 we used 1 mM IPTG (Biosynth, Staad, Switzerland) for pET28a(+)-His-flA and 250 mg/L arabinose (Sigma-Aldrich, St. Louis, MO) for pKLJ12-His-flA. Cells were harvested after 18 h of induction at 28 °C. Cell pellets were resuspended in 50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole (Merck Chemicals Ltd., Nottingham, UK), pH 8. 17 mg/mL lysozyme, 0.003 mg/mL DNAse (Sigma-Aldrich, St. Louis, MO) were added to the resuspension buffer and cells were lysed by sonication (output control 8, duty cycle 70 - 80%, 6 min).

FlA was purified by its His-tag with Ni^{2+}-affinity chromatography with a nickel-nitrilotriacetic acid (Ni-NTA) resin (Qiagen, Venlo, Netherland) according to the manufacturer's protocol.

Protein concentrations were determined by the Bradford method with the BioRad® protein assay (BioRad, Hercules, CA) as described in the manufacturer's protocol.
Elution buffer was exchanged to reaction buffer (100 mM TRIS/Cl, pH 7.5) using PD-10 desalting columns (GE Healthcare, Chalfont Saint Giles, Great Britain) according to the manufacturer's protocol.

in vitro reactions were done like described in Chan *et al.* [6] with modifications. The reaction mix contained 1.5 mM SAM (30 µL of a 10 mM stock solution in 1 mM HCl), 100 mM potassium fluoride (10 µl of a 2 M stock solution in ddH_2O) and 0.28 mg/mL FlA. After incubation for at 37 °C for 1 h protein was precipitated at 95 °C for 3 min and removed by centrifugation. For HPLC analysis see chapter 5.5.3.

Nucleotide sequence 1: pKLJ12-His-FlA

*ara*C	146-1024
*ara*BAD promoter	1298-1326
His-tag	1389-1406
flA	1416-2315
rrnB T2	2454-2620
*amp*R	2890-3750
mini-F fragment	3919-13351

GGCCACGATGCGTCCGGCGTAGAGGATCTGCTCATGTTTGACAGCTTATCATCGATGCATAATGTGCCTGTCAAATGGACGAAGCAGGGATTCTGCAAACCCTATGCTACTCCGTCAAGCCGTCA
ATTGTCTGATTCGTTACCAATTATGACAACTTGACGGCTACATCATTCACTTTTTCTTCACAACCGGCACGGAACTCGCTCGGGCTGGCCCCGGTGCATTTTTTAAATACCCGCGAGAAATAGAG
TTGATCGTCAAAACCAACATTGCGACCGACGGTGGCGATAGGCATCCGGGTGGTGCTCAAAAGCAGCTTCGCCTGGCTGATACGTTGGTCCTCGCGCCAGCTTAAGACGCTAATCCCTAACTGCT
GGCGGAAAAGATGTGACAGACGCGACGGCGACAAGCAAACATGCTGTGCGACGCTGGCGATATCAAAATTGCTGTCTGCCAGGTGATCGCTGATGTACTGACAAGCCTCGCGTACCCGATTATCC
ATCGGTGGATGGAGCGACTCGTTAATCGCTTCCATGCGCCGCAGTAACAATTGCTCAAGCAGATTTATCGCCAGCAGCTCCGAATAGCGCCCTTCCCCTTGCCCGGCGTTAATGATTTGCCCAAA
CAGGTCGCTGAAATGCGGCTGGTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGCAAATATTGACGGCCAGTTAAGCCATTCATGCCAGTAGGCGCGCGGACGAAAGTAAACCCACTGGTGAT
ACCATTCGCGAGCCTCCGGATGACGACCGTAGTGATGAATCTCTCCTGGCGGGAACAGCAAAATATCACCCGGTCGGCAAACAAATTCTCGTCCCTGATTTTTCACCACCCCCTGACCGCGAATG
GTGAGATTGAGAATATAACCTTTCATTCCCAGCGGTCGGTCGATAAAAAAATCGAGATAACCGTTGGCCTCAATCGGCGTTAAACCCGCCACCAGATGGGCATTAAACGAGTATCCCGGCAGCAG
GGGATCATTTTGCGCTTCAGCCATACTTTTCATACTCCCGCCATTCAGAGAAGAAACCAATTGTCCATATTGCATCAGACATTGCCGTCACTGCGTCTTTTACTGGCTCTTCTCGCTAACCAAAC
CGGTAACCCCGCTTATTAAAAGCATTCTGTAACAAAGCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACA
CTTTGCTATGCCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGGCTAGCATTAAAGAGGAGAAATTAAC
TATGAGAGGATCGCATCACCATCACCATCACAGCGGCCGCATGGCTGCGAACAGCACACGTCGCCCCATCATCGCGTTCATGAGCGACCTGGGGACCACGGACGACTCCGTCGCCCAGTGCAAGG
GGCTCATGTACAGCATCTGCCCGGACGTCACGGTGGTGGACGTCTGCCACTCGATGACGCCCTGGGACGTCGAGGAGGGCGCCCCGCTACATCGTGGACCTTCCGCGCTTCTTCCCCGAGGGAACG
GTCTTCGCCACCACCACCTATCCGGCGACCGGCACCACCACCCGCTCGGTGGCGGTGCGCATCAAGCAGGCCGCCAAGGGCGGTGCCCGCGGCCAGTGGGCGGGCTCGGGGGCCGGCTTCGAGCG
CGCCGAGGGCTCGTACATCTACATCGCGCCCAACAACGGGCTGCTGACCACCGTGCTGGAGGAGCACGGCTACCTGGAGGCGTACGAGGTCACCTCGCCGAAGGTCATCCCCGAGCAGCCCGAAC
CGACCTTCTACAGCCGGGAGATGGTGGCCATCCCCTCCGCGCACCTGGCCGCCGGCTTCCCGCTGTCCGAGGTCGGCCGTCCGCTGGAGGACCACGAGATCGTCCGCTTCAACCGCCCGGCCGTC
GAGCAGGACGGGGAGGCGCTGGTGGGCGTGGTCTCCGCCATCGACCACCCGTTCGGCAACGTGTGGACCAACATCCACCGCACCGACCTGGAGAAGGCGGGCATCGGCTACGGCGCCCGGCTGCG
GCTGACGCTGGACGGCGTGCTGCCGTTCGAGGCGCCGCTGACCCCGACGTTCGCCGACGCCGGTGAGATCGGCAACATCGCCATCTACCTCAACAGCCGCGGTTACCTGTCCATCGCGCGCAACG
CGGCCAGCCTCGCCTACCCGTACCACCTCAAGGAGGGCATGTCCGCCCGGGTCGAGGCCCGCTGAGAATTCACCATGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGCT
GTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAAC
TCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTT
GTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAG
CAGAAGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATT
GAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAG
TTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATT
ATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTAT
GCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGT
TGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACA
ATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAG
CACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAA
CTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGC
CCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCTCGACAGCGACACACTTGCATCGGATGCAGCCCGGTTAACGTGCCGGCACGGCCTGGGTAACCAGGTATTTTGTCCACATA
ACCGTGCGCAAAATGTTGTGGATAAGCAGGACACAGCAGCAATCCACAGCAGGCATACAACCGCACACCGAGGTTACTCCGTTCTACAGGTTACGACGACATGTCAATACTTGCCCTTGACAGGC
ATTGATGGAATCGTAGTCTCACGCTGATAGTCTGATCGACAATACAAGTGGGACCGTGGTCCCAGACCGATAATCAGACCGACAACACGAGTGGGATCGTGGTCCCAGACTAATAATCAGACCGA
CGATACGAGTGGGACCGTGGTCCCAGACTAATAATCAGACCGACGATACGAGTGGGACCGTGGTTCCAGACTAATAATCAGACCGACGATACGAGTGGGACCGTGGTCCCAGACTAATAATCAGA
CCGACGATACGAGTGGGACCATGGTCCCAGACTAATAATCAGACCGACGATACGAGTGGGACCGTGGTCCCAGTCTGATTATCAGACCGACGATACGAGTGGGACCGTGGTCCCAGACTAATAAT
CAGACCGACGATACGAGTGGGACCGTGGTCCCAGACTAATAATCAGACCGACGATACGAGTGGGACCGTGGTCCCAGTCTGATTATCAGACCGACGATACAAGTGGAACAGTGGGCCCAGAGAGA
ATATTCAGGCCAGTTATGCTTTCTGGCCTGTAACAAAGGACATTAAGTAAAGACAGATAAACGTAGACTAAAACGTGGTCGCATCAGGGTGCTGGCTTTTCAAGTTCCTTAAGAATGGCCTCAAT
TTTCTCTATACACTCAGTTGGAACACGAGACCTGTCCAGGTTAAGCACCATTTTATCGCCCTTATACAATACTGTCGCTCCAGGAGCAAACTGATGTCGTGAGCTTAAACTAGTTCTTGATGCAG
ATGACGTTTTAAGCACAGAAGTTAAAAGAGTGATAACTTCTTCAGCTTCAAATATCACCCCAGCTTTTTTCTGCTCATGAAGGTTAGATGCCTGCTGCTTAAGTAATTCCTCTTTATCTGTAAAG
GCTTTTTGAAGTGCATCACCTGACCGGGCAGATAGTTCACCGGGGTGAGAAAAAAGAGCAACAACTGATTTAGGCAATTTGGCGGTGTTGATACAGCGGGTAATAATCTTACGTGAAATATTTTC
CGCATCAGCCAGCGCAGAAATATTTCCAGCAAATTCATTCTGCAATCGGCTTGCATAACGCTGACCACGTTCATAAGCACTTGTTGGGCGATAATCGTTACCCAATCTGGATAATGCAGCCATCT
GCTCATCATCCAGCTCGCCAACCAGAACACGATAATCACTTTCGGTAAGTGCAGCAGCTTTACGACGGCGACTCCCATCGGCAATTTCTATGACACCAGATACTCTTCGACCGAACGCCGGTGTC
TGTTGACCAGTCAGTAGAAAAGAAGGGATGAGATCATCCAGTGCGTCCTCAGTAAGCAGCTCCTGGTCACGTTCATTACCTGACCATACCCGAGAGGTCTTCTCAACACTATCACCCCGGAGCAC
TTCAAGAGTAAACTTCACATCCCGACCACATACAGGCAAAGTAATGGCATTACCGCGAGCCATTACTCCTACGCGCGCAATTAACGAATCCACCATCGGGGCAGCTGGTGTCGATAACGAAGTAT
CTTCAACCGGTTGAGTATTGAGCGTATGTTTTGGAATAACAGGCGCACGCTTCATTATCTAATCTCCCAGCGTGGTTTAATCAGACGATCGAAAATTTCATTGCAGACAGGTTCCCAAATAGAAA
GAGCATTTCTCCAGGCACCAGTTGAAGAGCGTTGATCAATGGCCTGTTCAAAAACAGTTCTCATCCGGATCTGACCTTTACCAACTTCATCCGTTTCACGTACAACATTTTTTAGAACCATGCTT
CCCCAGGCATCCCGAATTTGCTCCTCCATCCACGGGGACTGAGAGCCATTACTATTGCTGTATTTGGTAAGCAAAATACGTACATCAGGCTCGAACCCTTTAAGATCAACGTTCTTGAGCAGATC
ACGAAGCATATCGAAAAACTGCAGTGCGGAGGTGTAGTCAAACAACTCAGCAGGCGTGGGAACAATCAGCACATCAGCAGCACATACGACATTAATCGTGCCGATACCCAGGTTAGGCGCGCTGT
CAATAACTATGACATCATAGTCATGAGCAACAGTTTCAATGGCCAGTCGGAGCATCAGGTGTGGATCGGTGGGCAGTTTACCTTCATCAAATTTGCCCATTAACTCAGTTTCAATACGGTGCAGA
GCCAGACAGGAAGGAATAATGTCAAGCCCCGGCCAGCAAGTGGGCTTTATTGCATAAGTGACATCGTCCTTTTCCCCAAGATAGAAAGGCAGGAGAGTGTCTTCTGCATGAATATGAAGATCTGG
TACCCATCCGTGATACATTGAGGCTGTTCCCTGGGGGTCGTTACCTTCCACGAGCAAAACACGTAGCCCCTTCAGAGCCAGATCCTGAGCAAGATGAACAGAAACTGAGGTTTTGTAAACGCCAC
CTTTATGGGCAGCAACCCCGATCACCGGTGGAAATACGTCTTCAGCACGTCGCAATCGCGTACCAAACACATCACGCATATGATTAATTTGTTCAATTGTATAACCAACACGTTGCTCAACCCGT
CCTCGAATTTCCATATCCGGGTGCGGTAGTCGCCCTGCTTTCTCGGCATCTCTGATAGCCTGAGAAGAAACCCCAACTAAATCCGCTGCTTCACCTATTCTCCAGCGCCGGGTTATTTTCCTCGC
TTCCGGGCTGTCATCATTAAACTGTGCAATGGCGATAGCCTTCGTCATTTCATGACCAGCGTTTATGCACTGGTTAAGTGTTTCCATGAGTTTCATTCTGAACATCCTTTAATCATTGCTTTGCG
TTTTTTTATTAAATCTTGCAATTTACTGCAAAGCAACAACAAAATCGCAAAGTCATCAAAAAACCGCAAAGTTGTTTAAAATAAGAGCAACACTACAAAAGGAGATAAGAAGAGCACATACCTCA
GTCACTTATTATCACTAGCGCTCGCCGCAGCCGTGTAACCGAGCATAGCGAGCGAACTGGCGAGGAAGCAAAGAAGAACTGTTCTGTCAGATAGCTCTTACGCTCAGCGCAAGAAGAAATATCCA
CCGTGGGAAAAACTCCAGGTAGAGGTACACACGCGGATAGCCAATTCAGAGTAATAAACTGTGATAATCAACCCTCATCAATGATGACGAACTAACCCCCGATATCAGGTCACATGACGAAGGGA
AAGAGAAGGAAATCAACTGTGACAAACTGCCCTCAAATTTGGCTTCCTTAAAAATTACAGTTCAAAAAGTATGAGAAAATCCATGCAGGCTGAAGGAAACAGCAAAACTGTGACAAATTACCCTC
AGTAGGTCAGAACAAATGTGACGAACCACCCTCAAATCTGTGACAGATAACCCTCAGACTATCCTGTCGTCATGGAAGTGATATCGCGGAAGGAAAATACGATATGAGTCGTCTGGCGGCCTTTC
TTTTTCTCAATGTATGAGAGGCGCATTGGAGTTCTGCTGTTGATCTCATTAACACAGACCTGCAGGAAGCGGCGGCGGAAGTCAGGCATACGCTGGTAACTTTGAGGCAGCTGGTAACGCTCTAT
GATCCAGTCGATTTTCAGAGAGACGATGCCTGAGCCATCCGGCTTACGATACTGACACAGGGATTCGTATAAACGCATGGCATACGGATTGGTGATTTCTTTTGTTTCACTAAGCCGAAACTGCG
TAAACCGGTTCTGTAACCCGATAAAGAAGGGAATGAGATATGGGTTGATATGTACACTGTAAAGCCCTCTGGATGGACTGTGCGCACGTTTGATAAACCAAGGAAAAGATTCATAGCCTTTTTCA
TCGCCGGCATCCTCTTCAGGGCGATAAAAAACCACTTCCTTCCCCGCGAAACTCTTCAATGCCTGCCGTATATCCTTACTGGCTTCCGCAGAGGTCAATCCGAATATTTCAGCATATTTAGCAAC
ATGGATCTCGCAGATACCGTCATGTTCCTGTAGGGTGCCATCAGATTTTCTGATCTGGTCAACGAACAGATACAGCATACGTTTTTGATCCCGGGAGAGACTATATGCCGCCTCAGTGAGGTCGT
TTGACTGGACGATTCGCGGGCTATTTTTACGTTTCTTGTGATTGATAACCGCTGTTTCCGCCATGACAGATCCATGTGAAGTGTGACAAGTTTTTAGATTGTCACACTAAATAAAAAAGAGTCAA
TAAGCAGGGATAACTTTGTGAAAAAACAGCTTCTTCTGAGGGCAATTTGTCACAGGGTTAAGGGCAATTTGTCACAGACAGGACTGTCATTTGAGGGTGATTTGTCACACTGAAAGGGCAATTTG

TCACAACACCTTCTCTAGAACCAGCATGGATAAAGGCCTACAAGGCGCTCTAAAAAAGAAGATCTAAAAACTATAAAAAAAAATAATTATAAAAATATCCCCGTGGATAAGTGGATAACCCCAAGG
GAAGTTTTTTCAGGCATCGTGTGTAAGCAGAATATATAAGTGCTGTTCCCTGGTGCTTCCTCGCTCACTCGAGGGCTTCGCCGTCGCTCGACTGCGGCGAGCCTACTGGCTGTAAAAGGACAGAC
CACATCATGGTTCTGTGTTCATTAGGTTGTTCTGTCCATTGCTGACATAATCCGCTCCACTTCAACGTAACACCGCACGAAGATTTCTATTGTTCCTGAAGGCATATTCAAATCGTTTTCGTTAC
CGCTTGCAGGCATCATGACAGAACACTACTTCCTATAAACGCTACACAGGCTCCTGAGATTAATAATGCGGATCTCTACGATAATGGGAGATTTTCCCGACTGTTTCGTTCGCTTCTCAGTGGAT
AACAGCCAGCTTCTCTGTTTAACAGACAAAAACAGCATATCCACTCAGTTCCACATTTCCATATAAAGGCCAAGGCATTTATTCTCAGGATAATTGTTTCAGCATCGCAACCGCATCAGACTCCG
GCATCGCAAACTGCACCCGGTGCCGGGCAGCCACATCCAGCGCAAAAACCTTCGTGTAGACTTCCGTTGAACTGATGGACTTATGTCCCATCAGGCTTTGCAGAACTTTCAGCGGTATACCGGCA
TACAGCATGTGCATCGCATAGGAATGGCGGAACGTATGTGGTGTGACCGGAACAGAGAACGTCACACCGTCAGCAGCAGCGGCGGCAACCGCCTCCCCAATCCAGGTCCTGACCGTTCTGTCCGT
CACTTCCCAGATCCGCGCTTTCTCTGTCCTTCCTGTGCGACGGTTACGCCGCTCCATGGGTATTTTCAGTTGTGCCACCATCGTCTGCAGCTGGCTGACGTACCAGGAGTCAGAGAGCGGAACCA
GCCGGTGAGTCTGCTGACCGGCGGGCATTCTCCCCGCCGTCCTGGCAGCTTTTTCGGTCCGTTGTTTCAGGGTCGCAAGCTGCACAAACGGATACGGAGGCGCAAGCGAAAAATCCCCCCGCGTC
AGCGCCAGTGCTTCATTAATGCGTGCTCCGGTGTTCCACAGTGTGGCCAGCAGCATCTTGCGGTGCAGATCCGGGACGTAATGGAGCAGGGCACTCACTTCCGGAGCCAGCAGATATTTTGGCAG
TTCATCATGGACCATCGACATCTGGCGAAGTGCCAGAGCTGCCGGATAATCAATGGCAACAGGCAGCGATGCAGGCTGCCCGGCAGAATACACTGCCGGTACCATGACTGCAGACTGGCTGTGTA
TAACGGAGCCTGACATTTATATTCCCCAGAACATCAGGTTAATGGCGTTTTTGATGTCATTTTCGCGGTGGCTGAGATCAGCCACTTCTTCCCCGATAACGGAGACCGGCACACTGGCCATATCG
GTGGTCATCATGCGCCAGCTTTCATCCCCGATATGCACCACCGGGTAAAGTTCACGGGAGACTTTATCTGACAGCAGACGTGCACTGGCCAGGGGGATCACCATCCGTCGCCCGGGCGTGTCAAT
AATATCACTCTGTACATCCACAAACAGACGATAACGGCTCTCTCTTTTATAGGTGTAAACCTTAAACTGCATTTCACCAGTCCCTGTTCTCGTCAGCAAAAGAGCCGTTCATTTCAATAAACCGG
GCGACCTCAGCCATCCCTTCCTGATTTTCCGCTTTCCAGCGTTCGGCACGCAGACGACGGGCTTCATTCTGCATGGTTGTGCTTACCAGACCGGAGATATTGACATCATATGCCTTGAGCAACTG
ATAGCTGTCGCTGTCAACTGTCACTGTAATACGCTGCTTCATAGCACACCTCTTTTTGACATACTTCGGGTATACATATCAGTATATATTCTTATACCGCAAAAATCAGCGCGCAAATACGCATA
CTGTTATCTGGCTTTTAGTAAGCCTTATGTATTTTACCTTTCGTTATGTTAACCAATAAAAATTAAAATCTGACTTATAAAAACAAAGCGTAATTACCGCATTCCCGTTTCGTATGGATCCGACT
TTATTTGAATAAACGACTTTCCGGAAGTGTACAACGCTGAGAAAATCATCATATCGCTGTTTTCCGCCCCTGCGCGGCGCATCACAGTCCAGAGATACAACACAACATATAGTGCCATGGGCAAC
TGTATGAGGCACTACATGCTGTGCTGACGGGATGTAACGCTGTGCGGAGGGAGAGGAAGATCCGTAACTTGGGGTTTTCGGATCTCCGGAGAAATAAGAAGAGGCAGGCATTGTTGTAGCATTAA
AATCCTCAATTTTAGTAATTTTAGAAGCATCACAATTTTCAAAAACTAAAAAACTGATTAATAGCTAATAATAAAATGCTAGGTTAGCATTAAGAACAAAACAAACATTAAAGGATCGATGATGC
TAAGCCAGCTTAACCTGCGTTTTCACAAAAAACTTATCGAGGCGCTGAAAACGCGTGCCGGTCGGGAAAATACTTCGGTCAACGCCCTAGCCGAACGCTTCCTCGATGACGGCCTGAAAACGGTC
GCGCCCGGTGACGGGTATTTTCAGCTCATTGCCGATCCGGAGGCCACGGTCCGGCAGCTGTACCGGCATATCATTCTGGGCCAGACCTTCGGCACGTCAGCGCTCTCCCGCGACGAACTGCGCTT
TGTCCTGGTTCACGTCAGGGAGGCCTTCCTGCGCGGGCATAATAGGCTGGCCACACTCCCGGCCCTGGACACGCTGCTGGACATCACTGGCAACCTGCTGGCCTGGCAGGTGGAGCACGATCGCC
CTGTGGACGGTCACTACCTGAAGGGGATTTTCCGCCTGGCCGGGAAAAACTGGACGGAAGAGTTTGAAGCCTTCCGGGCAGCGCTGCGCCCCGTGGTGGATCAGATGTATGCCGAACACCTCCTG
CGTCCGCTCGAGAGCGACTGTTTTGGCCTGGCGGAGGTGCCGGACGCGGTGCTGGCAGAAATCTTCACCCTGCCGCGGCTGAAAGCTGTTTTCCCGTTGATGCTGCGCGGCCTGGACTGGAACAC
AGAACAGGCAAGAACGCTGGCACAGGAGCTGCGCCCGGTTATTTCTGCCGTCACGGAAACCATTGAGGCCGGCACGCTGCGCCTGGAGATCCGCGTTGACGGCCAGCACCCCGGAGAACGCCCGG
GGGCCTGGTACACCACACCCCGTCTGCACCTGCTGATCACCGGCCAGGACTTCGTGGTGCCGTACGGCTGGGAAGCGTTATCCGAACTGCTGGGCCTGTTCACGCTGTATGCCCGGCATCCGGAA
GCCCTGACGCACGGTCACCAGGGGGAGCGAGTGATGTTCTCCCCGCCCGGAAACGTCACGCCGGAGGCTTCTTCGGGATAGACGGCCTGCGGATTTTTATGCCGGCGGAGGCTTTGAGACGCTGG
TCCGGGACCTCGCCACGCGTGTCAAGAAGGCCCGCTGGCGGAGGCCTGACCGGGCTGCGCTGCCTGTACGGAGATCTGTAATATGAAGATACTGAGGCAGCTGTGGAACCAGAAAGGGCTTGATG
CGGCAGTTGAAGATGTACCTGAGGATCGATACGGTTTCGGTAATATCGCAGAAAATATATCCCGTTCCATCCTTACTCTTCCTCTGGAGGCTTCAAATGTTGTCGGAATAGAAGGTGCATGGGGA
TCGGGTAAAACAAGCCTCCTTAATCTGATCCTGAGAAACCTTGCCCTGAAAAAAGATGCTCACACACATGTGCTGCATATTTCTCCCTGGCTGAGTGGCGGCAGTCCGGTTGAAGCGCTCTTTCT
TCCGGTTGCCACGGTTATCCAGCAGGAAATGGAAATACGCTATCCCCCGAAGGGCTTCAAAAAGCTCTGGCGTAAATATCTGTTGTCACCGGAAGCTCAAAAAGTGATTGAGTATGCTCAGGATA
CTTCATCGCGCGTCCTTCCACTGGTTCAATATATCGGACAGTTTTCCAGCATTATTAACTGGATAGCAGGCGGAATAAAGGTATTTTCAGACAGCCGTCTGGCTGTTGATCAGAAAACCACGACA
AAGCTTCGGGCTGAAATTGCAGGACAACTGGTGAGTCTGGATCTGAAGTTCATTGTTGTCATGGATGATCTGGACCGACTGGAGCCATCCCAGGTGGCGGAAGTGTTCAGGCTTGTGCGTGCAGT
AGCCGATCTGCCCCGCTTTACCCATATTCTCTGTTATGACAGGCAGATTATCACTCATGCCGTTGAACATGCGCTGAATATCGAAGATGGCAGCCGTTATCTCCAGAAAATCATTCAGCTTAGTT
TTAAATTACCCCGACCTGAAGCCTTTGATTTACGTAATGAATTTCGCCAGCGGGCTGAGGCTCTATATCAGCAAATTAATAATCAACCGCCAGACTCTGGAATGGTAAGGGATCTCATCGCGGTG
ACTGATACCTATGGTGCCGCACTTTCGACGCCACGGGAAATCCATCAGGCCATTAATTCCCTGATTTTTCTTTATCCGGGGATGCGGGATTTTGTTTATTTCCCTGATTTGTGCCTGCTTCAGCT
TATACGGGTGACAAACCCGGCTCTGTATGACTGGACAGAGCATTACCTGACAGAACGGTCCGTGATTGAAACCGGTCAGGGTATGCTTTCTGACGGAGAGAAAGCAGACTTCCGGGAGGGGCTTA
TCAGATGTATGAAGACGTTCAGGGCATCAAATGCAGACTCGTTTCTGACACTTGCATACTGGATCCCAGGTATCAGTGGACATAATGATGAATATCTGAATCTTTTTGAGCCCGTCAGTGAGGAT
TTTCGTCATATCCAGACGACCGGTAAACGACTCAGCAGCCTGACGCACTGGCGATATTATTTTGCCTTCTCCTCACCACAGAATGTTCTGCCACCTGAATT

Nucleotide sequence 2: pMS470-His-flA

TAC promoter	1-206
His-tag	253-270
flA	280-1179
rrnB	1268-1593
*amp*R	1705-2565
ColE1	2660-3342
lacI	3635-4726

CAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAACGGTTCTGGCA
AATATTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGAATTCATTAAAGAGGAGAAATTAACTATGAGAGGAT
CGCATCACCATCACCATCACAGCGGCCGCATGGCTGCGAACAGCACACGTCGCCCCATCATCGCGTTCATGAGCGACCTGGGGACCACGGACGACTCCGTCGCCCAGTGCAAGGGGCTCATGTAC
AGCATCTGCCCGGACGTCACGGTGGTGGACGTCTGCCACTCGATGACGCCCTGGGACGTCGAGGAGGGCGCCCGCTACATCGTGGACCTTCCGCGCTTCTTCCCCGAGGGAACGGTCTTCGCCAC
CACCACCTATCCGGCGACCGGCACCACCACCCGCTCGGTGGCGGTGCGCATCAAGCAGGCCGCCAAGGGCGGTGCCCGCGGCCAGTGGGCGGGCTCGGGGGCCGGCTTCGAGCGCGCCGAGGGCT
CGTACATCTACATCGCGCCCAACAACGGGCTGCTGACCACCGTGCTGGAGGAGCACGGCTACCTGGAGGCGTACGAGGTCACCTCGCCGAAGGTCATCCCCGAGCAGCCCGAACCGACCTTCTAC
AGCCGGGAGATGGTGGCCATCCCCTCCGCGCACCTGGCCGCCGGCTTCCCGCTGTCCGAGGTCGGCCGTCCGCTGGAGGACCACGAGATCGTCCGCTTCAACCGCCCGGCCGTCGAGCAGGACGG
GGAGGCGCTGGTGGGCGTGGTCTCCGCCATCGACCACCCGTTCGGCAACGTGTGGACCAACATCCACCGCACCGACCTGGAGAAGGCGGGCATCGGCTACGGCGCCCGGCTGCGGCTGACGCTGG
ACGGCGTGCTGCCGTTCGAGGCGCCGCTGACCCCGACGTTCGCCGACGCCGGTGAGATCGGCAACATCGCCATCTACCTCAACAGCCGCGGTTACCTGTCCATCGCGCGCAACGCGGCCAGCCTC
GCCTACCCGTACCACCTCAAGGAGGGCATGTCCGCCCGGGTCGAGGCCCGCTGAAAGCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAAATCAGAACGCAGAAGCGG
TCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGA
ACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGC
GAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTTGTTTATTTTT
CTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGT
TTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGA
CTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCG
GAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTNCA
GCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCC
GGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGG
CAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTT
AAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTT
TCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAA
ATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTT
ACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGA
GCATTGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTC
CTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTT
GCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGC
GGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCACGAACGCCAGCAAGACGTAGCCCAGCGCGTCGGCCAGCTTGCAATTCGCGCTAACTTACATTAATT
GCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCA
GTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGCTGGTGGTTAACGGCGGGATATAA
CATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGT
GGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCA
GACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTG
ATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAG
AAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCA
GGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCA
GAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCG
CTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCAACGTAAATGCCGCTTCGCCTTCGCGCGCGAATTGCAAGCTGATCCGGGCTTATCGACTGCACGGTGCACCAATGCTTCT
GGCGT

Nucleotide sequence 3: pQE80L-His-flA

T5 promoter	7-87
His-tag	127-144
flA	151-1050
lambda t0	1072-1166
rrnB T1	1928-2025
lacI	2113-3195
ColE1	3773-4455
*amp*R	4550-5410

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCATTAAAGAGGAGAAATTAACTATGAGAGGATC
GCATCACCATCACCATCACGGATCCATGGCTGCGAACAGCACACGTCGCCCCATCATCGCGTTCATGAGCGACCTGGGGACCACGGACGACTCCGTCGCCCAGTGCAAGGGGCTCATGTACAGCA
TCTGCCCGGACGTCACGGTGGTGGACGTCTGCCACTCGATGACGCCCTGGGACGTCGAGGAGGGCGCCCGCTACATCGTGGACCTTCCGCGCTTCTTCCCCGAGGGAACGGTCTTCGCCACCACC
ACCTATCCGGCGACCGGCACCACCACCCGCTCGGTGGCGGTGCGCATCAAGCAGGCCGCCAAGGGCGGTGCCCGCGGCCAGTGGGCGGGCTCGGGGGCCGGCTTCGAGCGCGCCGAGGGCTCGTA
CATCTACATCGCGCCCAACAACGGGCTGCTGACCACCGTGCTGGAGGAGCACGGCTACCTGGAGGCGTACGAGGTCACCTCGCCGAAGGTCATCCCCGAGCAGCCCGAACCGACCTTCTACAGCC
GGGAGATGGTGGCCATCCCCTCCGCGCACCTGGCCGCCGGCTTCCCGCTGTCCGAGGTCGGCCGTCCGCTGGAGGACCACGAGATCGTCCGCTTCAACCGCCCGGCCGTCGAGCAGGACGGGGAG
GCGCTGGTGGGCGTGGTCTCCGCCATCGACCACCCGTTCGGCAACGTGTGGACCAACATCCACCGCACCGACCTGGAGAAGGCGGGCATCGGCTACGGCGCCCGGCTGCGGCTGACGCTGGACGG
CGTGCTGCCGTTCGAGGCGCCGCTGACCCCGACGTTCGCCGACGCCGGTGAGATCGGCAACATCGCCATCTACCTCAACAGCCGCGGCTACCTGTCCATCGCGCGCAACGCGGCCAGCCTCGCCT
ACCCGTACCACCTCAAGGAGGGCATGTCCGCCCGGGTCGAGGCCCGCTGAAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCA
GAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAAGCTAGCTTGGCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATC
CCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCACAAGTTTT
ATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAG
CAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAAGGGTTTAT
TGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATATTATACGCAAGGCG
ACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTT
TAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACG
CTCTCCTGAGTAGGACAAATCCGCCCTCTAGATTACGTGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTT
TCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATT
GCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGT
CGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGC
ATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGA
ACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACAT
CAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTA
CAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAAC
GCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCA
CCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAG
GTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTC
CCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGT
GTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCT
CACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGC
AAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGA
TACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAG
GTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACT
TATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGC
GCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATC
TCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAAT
GAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCC
GTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGA
GCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCG
TGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTC
AGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGA
ATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGA
TCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAG
GGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACA
AATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

Nucleotide sequence 4: p15aTAC-lbpB

lacI	15-1106
TAC promoter	1260-1466
lbpB	1502-1930
rrnB	1978-2303
p15a origin	2331-3242
*kan*R	3651-4445

ATCGATGCGGCCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTT
CACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGCTGGTGGTTAACGGCGGGA
TATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATC
GCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCC
AGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATAC
TGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGC
GCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGC
GTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTT
TCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCC
GGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCAACGTAAATGCCGCTTCGCCTTCGCGCGCGAATTGCAAGCTGATCCGGGCTTATCGACTGCACGGTGCACCAATG
CTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAAC
GGTTCTGGCAAATATTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAAGGGATCCTTTAACTTTAAGAAGGAGATCA
TATGCGTAACTTCGATTTATCCCCACTGATGCGTCAATGGATCGGTTTTGACAAACTGGCCAACGCACTGCAAAACGCCGGTGAAAGCCAGAGCTTCCCGCCGTACAACATTGAGAAAAGCGACG
ATAACCACTACCGCATTACCCTTGCGCTGGCAGGTTTCCGTCAGGAAGATTTAGAGATTCAACTGGAAGGTACGCGCCTGAGCGTAAAAGGCACGCCGGAGCAGCCAAAAGAAGAGAAAAAATGG
CTGCATCAAGGGCTTATGAATCAGCCATTTAGCCTGAGCTTTACGCTGGCTGAAAATATGGAAGTCTCTGGCGCAACCTTCGTAAACGGTTTACTGCATATTGATTTAATTCGTAATGAGCCTGA
ACCCATCGCAGCGCAGCGTATCGCTATCAGCGAACGTCCCGCGTTAAATAGCTAAGAATTCGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTG
GTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAG
ACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCA
TAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGGATAAGCTGTCAAACATGAGAATTAACAACTTATATCGTATGGGGCTGACTTCAGGTGCTACATTTGAAG
AGATAAATTGCACTGAAATCTAGAAATATTTTATCTGATTAATAAGATGATCTTCTTGAGATCGTTTTGGTCTGCGCGTAATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTT
TCGAAGGTTCTCTGAGCTACCAACTCTTTGAACCGAGGTAACTGGCTTGGAGGAGCGCAGTCACCAAAACTTGTCCTTTCAGTTTAGCCTTAACCGGCGCATGACTTCAAGACTAACTCCTCTAA
ATCAATTACCAGTGGCTGCTGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGC
TTGGAGCGAACTGCCTACCCGGAACTGAGTGTCAGGCGTGGAATGAGACAAACGCGGCCATAACAGCGGAATGACACCGGTAAACCGAAAGGCAGGAACAGGAGAGCGCACGAGGGAGCCGCCAG
GGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCACTGATTTGAGCGTCAGATTTCGTGATGCTTGTCAGGGGGGCGGAGCCTATGGAAAAACGGCTTTGCCGCGGCCCTCTCAC
TTCCCTGTTAAGTATCTTCCTGGCATCTTCCAGGAAATCTCCGCCCCGTTCGTAAGCCATTTCCGCTCGCCGCAGTCGAACGACCGAGCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCC
TGTATCACATATTCTGCTGACGCACCGGTGCAGCCTTTTTTCTCCTGCCACATGAAGCACTTCACTGACACCCTCATCAGTGCCAACATAGTAAGCCAGTATACACTCCGCTAGCGCTGATGTCC
GGCGGTGCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGGGTCGACCAAAGCGGCCATCGTGCCTCCCCACTCCTGCAGTTCGGGGGCATGGATGCGCGGATAGCCGCT
GCTGGTTTCCTGGATGCCGACGGATTTGCACTGCCGGTAGAACTCCGCGAGGTCGTCCAGCCTCAGGCAGCAGCTGAACCAACTCGCGAGGGGATCGAGCCCGGGGTGGGCGAAGAACTCCAGCA
TGAGATCCCCGCGCTGGAGGATCATCCAGCCGGCGTCCCGGAAAACGATTCCGAAGCCCAACCTTTCATAGAAGGCGGCGGTGGAATCGAAATCTCGTGATGGCAGGTTGGGCGTCGCTTGGTCG
GTCATTTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTGCGAATCGGGAGCGGCGATACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCA
AGCTCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGCCACACCCAGCCGGCCACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCAGGC
ATCGCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCCTGGCGAACAGTTCGGCTGGCGCGAGCCCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTT
CCATCCGAGTACGTGCTCGCTCGATGCGATGTTTCGCTTGGTGGTCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGCATCAGCCATGATGGATACTTTCTCGGCAGGAGCA
AGGTGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAGCCAGTCCCTTCCCGCTTCAGTGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGATAGCCG
CGCTGCCTCGTCCTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGCCCCTGCGCTGACAGCCGGAACACGGCGGCATCAGAGCAGCCGATTGTCTGTTGTGCCC
AGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAAGCATGCGAAACGACCGTCATCCTGTCTCTTGATCAGATCTTGATCCCCTGCGCCATCAG
ATCCTTGGCGGCAAGAAAGCCATCCAGTTTACTTTGCAGGGCTTCCCAACCTTACCAGAGGGCGCCCCAGCTGGCAATTCCGGTT

Nucleotide sequence 5: p15aTAC-dsbC

lacI	15..1106
TAC promoter	1260..1466
truncated *dsbC*	1524..2177
rrnB	2225..2550
p15a origin	2578..3489
*kan*R	3898..4692

ATCGATGCGGCCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTT
CACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGCTGGTGGTTAACGGCGGGA
TATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATC
GCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCC
AGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATAC
TGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGC
GCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGC
GTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTT
TCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCC
GGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCAACGTAAATGCCGCTTCGCCTTCGCGCGCGAATTGCAAGCTGATCCGGGCTTATCGACTGCACGGTGCACCAATG
CTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAAC
GGTTCTGGCAAATATTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGAATTCGAGCTCGGTACCCGGGGATCC
TTTAACTTTAAGAAGGAGATATAATGGATGACGCGGCAATTCAACAAACGTTAGCCAAAATGGGCATCAAAAGCAGCGATATTCAGCCCGCGCCTGTAGCTGGCATGAAGACAGTTCTGACTAAC
AGCGGCGTGTTGTACATCACCGATGATGGTAAACATATCATTCAGGGGCCAATGTATGACGTTAGTGGCACGGCTCCGGTCAATGTCACCAATAAGATGCTGTTAAAGCAGTTGAATGCGCTTGA
AAAAGAGATGATCGTTTATAAAGCGCCGCAGGAAAAACACGTCATCACCGTGTTTACTGATATTACCTGTGGTTACTGCCACAAACTGCATGAGCAAATGGCAGACTACAACGCGCTGGGGATCA
CCGTGCGTTATCTTGCTTTCCCGCGCCAGGGGCTGGACAGCGATGCAGAGAAAGAAATGAAAGCTATCTGGTGTGCGAAAGATAAAAACAAAGCGTTTGATGATGTGATGGCAGGTAAAAGCGTC
GCACCAGCCAGTTGCGACGTGGATATTGCCGACCATTACGCACTTGGCGTCCAGCTTGGCGTTAGCGGTACTCCGGCAGTTGTGCTGAGCAATGGCACACTTGTTCCGGGTTACCAGCCGCCGAA
AGAGATGAAAGAATTCCTCGACGAACACCAAAAAATGACCAGCGGTAAATAAGAATTCGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTC
CCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACT
GGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAA
ACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGGATAAGCTGTCAAACATGAGAATTAACAACTTATATCGTATGGGGCTGACTTCAGGTGCTACATTTGAAGAGA
TAAATTGCACTGAAATCTAGAAATATTTTATCTGATTAATAAGATGATCTTCTTGAGATCGTTTTGGTCTGCGCGTAATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTTTCG
AAGGTTCTCTGAGCTACCAACTCTTTGAACCGAGGTAACTGGCTTGGAGGAGCGCAGTCACCAAAACTTGTCCTTTCAGTTTAGCCTTAACCGGCGCATGACTTCAAGACTAACTCCTCTAAATC
AATTACCAGTGGCTGCTGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGCTTG
GAGCGAACTGCCTACCCGGAACTGAGTGTCAGGCGTGGAATGAGACAAACGCGGCCATAACAGCGGAATGACACCGGTAAACCGAAAGGCAGGAACAGGAGAGCGCACGAGGGAGCCGCCAGGGG
AAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCACTGATTTGAGCGTCAGATTTCGTGATGCTTGTCAGGGGGGCGGAGCCTATGGAAAAACGGCTTTGCCGCGGCCCTCTCACTTC
CCTGTTAAGTATCTTCCTGGCATCTTCCAGGAAATCTCCGCCCCGTTCGTAAGCCATTTCCGCTCGCCGCAGTCGAACGACCGAGCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCCTGT
ATCACATATTCTGCTGACGCACCGGTGCAGCCTTTTTTCTCCTGCCACATGAAGCACTTCACTGACACCCTCATCAGTGCCAACATAGTAAGCCAGTATACACTCCGCTAGCGCTGATGTCCGGC
GGTGCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGGGTCGACCAAAGCGGCCATCGTGCCTCCCCACTCCTGCAGTTCGGGGGCATGGATGCGCGGATAGCCGCTGCT
GGTTTCCTGGATGCCGACGGATTTGCACTGCCGGTAGAACTCCGCGAGGTCGTCCAGCCTCAGGCAGCAGCTGAACCAACTCGCGAGGGGATCGAGCCCGGGGTGGGCGAAGAACTCCAGCATGA
GATCCCCGCGCTGGAGGATCATCCAGCCGGCGTCCCGGAAAACGATTCCGAAGCCCAACCTTTCATAGAAGGCGGCGGTGGAATCGAAATCTCGTGATGGCAGGTTGGGCGTCGCTTGGTCGGTC
ATTTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTGCGAATCGGGAGCGGCGATACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCAAGC
TCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGCCACACCCAGCCGGCCACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCAGGCATC
GCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCCTGGCGAACAGTTCGGCTGGCGCGAGCCCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTTCCA
TCCGAGTACGTGCTCGCTCGATGCGATGTTTCGCTTGGTGGTCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGCATCAGCCATGATGGATACTTTCTCGGCAGGAGCAAGG
TGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAGCCAGTCCCTTCCCGCTTCAGTGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGATAGCCGCGC
TGCCTCGTCCTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGCCCCTGCGCTGACAGCCGGAACACGGCGGCATCAGAGCAGCCGATTGTCTGTTGTGCCCAGT
CATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAAGCATGCGAAACGACCGTCATCCTGTCTCTTGATCAGATCTTGATCCCCTGCGCCATCAGATC
CTTGGCGGCAAGAAAGCCATCCAGTTTACTTTGCAGGGCTTCCCAACCTTACCAGAGGGCGCCCCAGCTGGCAATTCCGGTTCGCTTGCTGTCCATAAAACCGCCCAGTCTAGCTATCGCCATGT
AAGCCCACTGCAAGCTACCTGCTTTCTCTTTGCGCTTGCGTTTTCCCTTGTCCAGATAGCCCAGTAGCTGACATTCATCCGGGGTCAGCACCGTTTCTGCGGACTGGCTTTCTACGTGTTCCGCT
TCCTTTAGCAGCCCTTGCGCCCTGAGTGCTTGCGGCAGCGTGAAGCTT

Nucleotide sequence 6: p15aTAC-lbpB-dsbC

lacI	15..1106
TAC promoter	1260..1466
lbpB	1502..1930
truncated *dsbC*	1960..2613
rrnB	2661..2986
p15a origin	3014..3925
*kan*R	4334..5128

ATCGATGCGGCCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTT
CACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGCTGGTGGTTAACGGCGGGA
TATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATC
GCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCC
AGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATAC
TGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGC
GCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGC
GTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTT
TCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCC
GGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCAACGTAAATGCCGCTTCGCCTTCGCGCGCGAATTGCAAGCTGATCCGGGCTTATCGACTGCACGGTGCACCAATG
CTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAAC
GGTTCTGGCAAATATTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAAGGGATCCTTTAACTTTAAGAAGGAGATCA
TATGCGTAACTTCGATTTATCCCCACTGATGCGTCAATGGATCGGTTTTGACAAACTGGCCAACGCACTGCAAAACGCCGGTGAAAGCCAGAGCTTCCCGCCGTACAACATTGAGAAAAGCGACG
ATAACCACTACCGCATTACCCTTGCGCTGGCAGGTTTCCGTCAGGAAGATTTAGAGATTCAACTGGAAGGTACGCGCCTGAGCGTAAAAGGCACGCCGGAGCAGCCAAAAGAAGAGAAAAAATGG
CTGCATCAAGGGCTTATGAATCAGCCATTTAGCCTGAGCTTTACGCTGGCTGAAAATATGGAAGTCTCTGGCGCAACCTTCGTAAACGGTTTACTGCATATTGATTTAATTCGTAATGAGCCTGA
ACCCATCGCAGCGCAGCGTATCGCTATCAGCGAACGTCCCGCGTTAAATAGCTAACTCGAGTTTAACTTTAAGAAGGAGATATAATGGATGACGCGGCAATTCAACAAACGTTAGCCAAAATGGG
CATCAAAAGCAGCGATATTCAGCCCGCGCCTGTAGCTGGCATGAAGACAGTTCTGACTAACAGCGGCGTGTTGTACATCACCGATGATGGTAAACATATCATTCAGGGGCCAATGTATGACGTTA
GTGGCACGGCTCCGGTCAATGTCACCAATAAGATGCTGTTAAAGCAGTTGAATGCGCTTGAAAAAGAGATGATCGTTTATAAAGCGCCGCAGGAAAAACACGTCATCACCGTGTTTACTGATATT
ACCTGTGGTTACTGCCACAAACTGCATGAGCAAATGGCAGACTACAACGCGCTGGGGATCACCGTGCGTTATCTTGCTTTCCCGCGCCAGGGGCTGGACAGCGATGCAGAGAAAGAAATGAAAGC
TATCTGGTGTGCGAAAGATAAAAACAAAGCGTTTGATGATGTGATGGCAGGTAAAAGCGTCGCACCAGCCAGTTGCGACGTGGATATTGCCGACCATTACGCACTTGGCGTCCAGCTTGGCGTTA
GCGGTACTCCGGCAGTTGTGCTGAGCAATGGCACACTTGTTCCGGGTTACCAGCCGCCGAAAGAGATGAAAGAATTCCTCGACGAACACCAAAAAATGACCAGCGGTAAATAAGAATTCGATTAA
ATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTC
CCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGGG
AGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTTTGGATAAGCTGTCA
AACATGAGAATTAACAACTTATATCGTATGGGGCTGACTTCAGGTGCTACATTTGAAGAGATAAATTGCACTGAAATCTAGAAATATTTTATCTGATTAATAAGATGATCTTCTTGAGATCGTTT
TGGTCTGCGCGTAATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTTTCGAAGGTTCTCTGAGCTACCAACTCTTTGAACCGAGGTAACTGGCTTGGAGGAGCGCAGTCACCAA
AACTTGTCCTTTCAGTTTAGCCTTAACCGGCGCATGACTTCAAGACTAACTCCTCTAAATCAATTACCAGTGGCTGCTGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTGGACTCAAGACGATA
GTTACCGGATAAGGCGCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGCTTGGAGCGAACTGCCTACCCGGAACTGAGTGTCAGGCGTGGAATGAGACAAACGCGGCCATAACAGC
GGAATGACACCGGTAAACCGAAAGGCAGGAACAGGAGAGCGCACGAGGGAGCCGCCAGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCACTGATTTGAGCGTCAGATTTCG
TGATGCTTGTCAGGGGGGCGGAGCCTATGGAAAAACGGCTTTGCCGCGGCCCTCTCACTTCCCTGTTAAGTATCTTCCTGGCATCTTCCAGGAAATCTCCGCCCCGTTCGTAAGCCATTTCCGCT
CGCCGCAGTCGAACGACCGAGCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCCTGTATCACATATTCTGCTGACGCACCGGTGCAGCCTTTTTTCTCCTGCCACATGAAGCACTTCACTG
ACACCCTCATCAGTGCCAACATAGTAAGCCAGTATACACTCCGCTAGCGCTGATGTCCGGCGGTGCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGGGTCGACCAAAG
CGGCCATCGTGCCTCCCCACTCCTGCAGTTCGGGGGCATGGATGCGCGGATAGCCGCTGCTGGTTTCCTGGATGCCGACGGATTTGCACTGCCGGTAGAACTCCGCGAGGTCGTCCAGCCTCAGG
CAGCAGCTGAACCAACTCGCGAGGGGATCGAGCCCGGGGTGGGCGAAGAACTCCAGCATGAGATCCCCGCGCTGGAGGATCATCCAGCCGGCGTCCCGGAAAACGATTCCGAAGCCCAACCTTTC
ATAGAAGGCGGCGGTGGAATCGAAATCTCGTGATGGCAGGTTGGGCGTCGCTTGGTCGGTCATTTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTG
CGAATCGGGAGCGGCGATACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCAAGCTCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGCCACACCCAGCCGGC
CACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCAGGCATCGCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCCTGGCGAACAGTTCG
GCTGGCGCGAGCCCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTTCCATCCGAGTACGTGCTCGCTCGATGCGATGTTTCGCTTGGTGGTCGAATGGGCAGGTAGCCGGATC
AAGCGTATGCAGCCGCCGCATTGCATCAGCCATGATGGATACTTTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAGCCAGTCCCTTCCCGCTTCAG
TGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGATAGCCGCGCTGCCTCGTCCTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGC
CCCTGCGCTGACAGCCGGAACACGGCGGCATCAGAGCAGCCGATTGTCTGTTGTGCCCAGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAAG
CATGCGAAACGACCGTCATCCTGTCTCTTGATCAGATCTTGATCCCCTGCGCCATCAGATCCTTGGCGGCAAGAAAGCCATCCAGTTTACTTTGCAGGGCTTCCCAACCTTACCAGAGGGCGCCC
CAGCTGGCAATTCCGGTTCGCTTGCTGTCCATAAAACCGCCCAGTCTAGCTATCGCCATGTAAGCCCACTGCAAGCTACCTGCTTTCTCTTTGCGCTTGCGTTTTCCCTTGTCCAGATAGCCCAG
TAGCTGACATTCATCCGGGGTCAGCACCGTTTCTGCGGACTGGCTTTCTACGTGTTCCGCTTCCTTTAGCAGCCCTTGCGCCCTGAGTGCTTGCGGCAGCGTGAAGCTT

Nucleotide sequence 7: First codon-optimized version of the fluorinase (FlAco1)

ATGGCGGCTAACTCCACACGCCGTCCGATTATTGCATTTATGTCTGACCTGGGTACCACTGATGACTCTGTAGCTCAATGCAAAGGTCTGATGTACAGCATCTGTCCGGACGTAACCGTCGTTGA
TGTTTGCCACTCTATGACCCCGTGGGACGTGGAAGAAGGTGCCCGCTACATCGTTGATTTGCCGCGTTTCTTTCCGGAAGGTACCGTTTTTGCGACCACCACTTATCCAGCAACTGGTACTACCA
CTCGTTCTGTGGCGGTTCGTATCAAACAGGCCGCGAAGGGTGGAGCACGCGGCCAGTGGGCTGGCTCCGGTGCTGGTTTCGAACGCGCTGAAGGTAGCTACATTTACATCGCACCAAACAACGGC
CTGCTGACCACCGTTCTGGAGGAACATGGTTATCTGGAAGCATATGAAGTAACTTCCCCGAAAGTGATCCCGGAACAGCCGGAGCCGACCTTCTACTCCCGTGAAATGGTAGCAATTCCGAGCGC
GCACCTGGCGGCGGGTTTCCCGCTGTCTGAGGTTGGTCGTCCGCTGGAAGATCACGAGATTGTACGTTTTAACCGTCCGGCGGTCGAGCAGGATGGTGAAGCTCTGGTAGGTGTTGTATCTGCAA
TCGATCACCCGTTCGGTAACGTTTGGACCAACATCCATCGTACCGACCTGGAGAAAGCGGGTATCGGTTACGGTGCGCGCCTGCGTCTGACCCTGGATGGTGTGCTGCCATTCGAGGCTCCGCTG
ACTCCGACCTTTGCAGATGCTGGCGAGATCGGTAACATCGCTATCTATCTGAACTCTCGCGGTTACCTGTCTATTGCTCGTAACGCCGCAAGCCTGGCTTACCCGTACCATCTGAAGGAAGGTAT
GTCTGCACGTGTTGAAGCTCGTTGA

2.10.2 **Supplementary Figures**

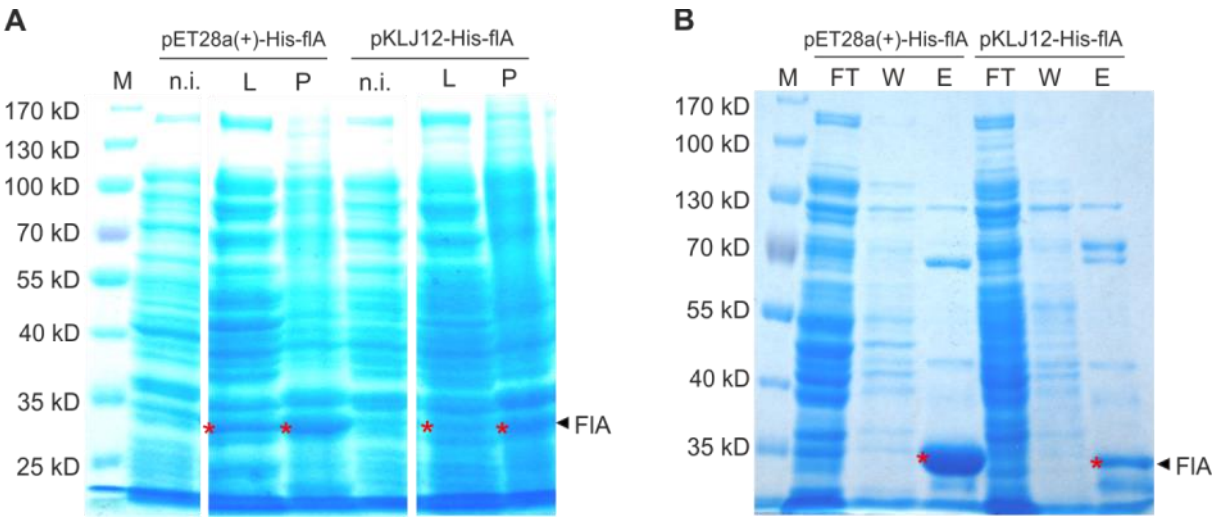


Figure S1: FlA expression from two vector constructs yields different amounts of purified protein

12% home-made SDS gels (see chapter 5.3 for details) of the FlA overexpressing strains (A) and the steps of FlA Ni-affinity purification (B) are shown. The red star refers to the presumed FlA protein band. The calculated molecular weight of His-tagged FlA is 34 kDa. M refers to protein molecular weight marker, n.i. to non-induced sample at the time of induction, L to the cleared lysate fraction, P to the insoluble protein (pellet) fraction, FT to flow through, W to wash fraction and E to elution fraction. 10% SDS Laemmli gel [34] is shown.

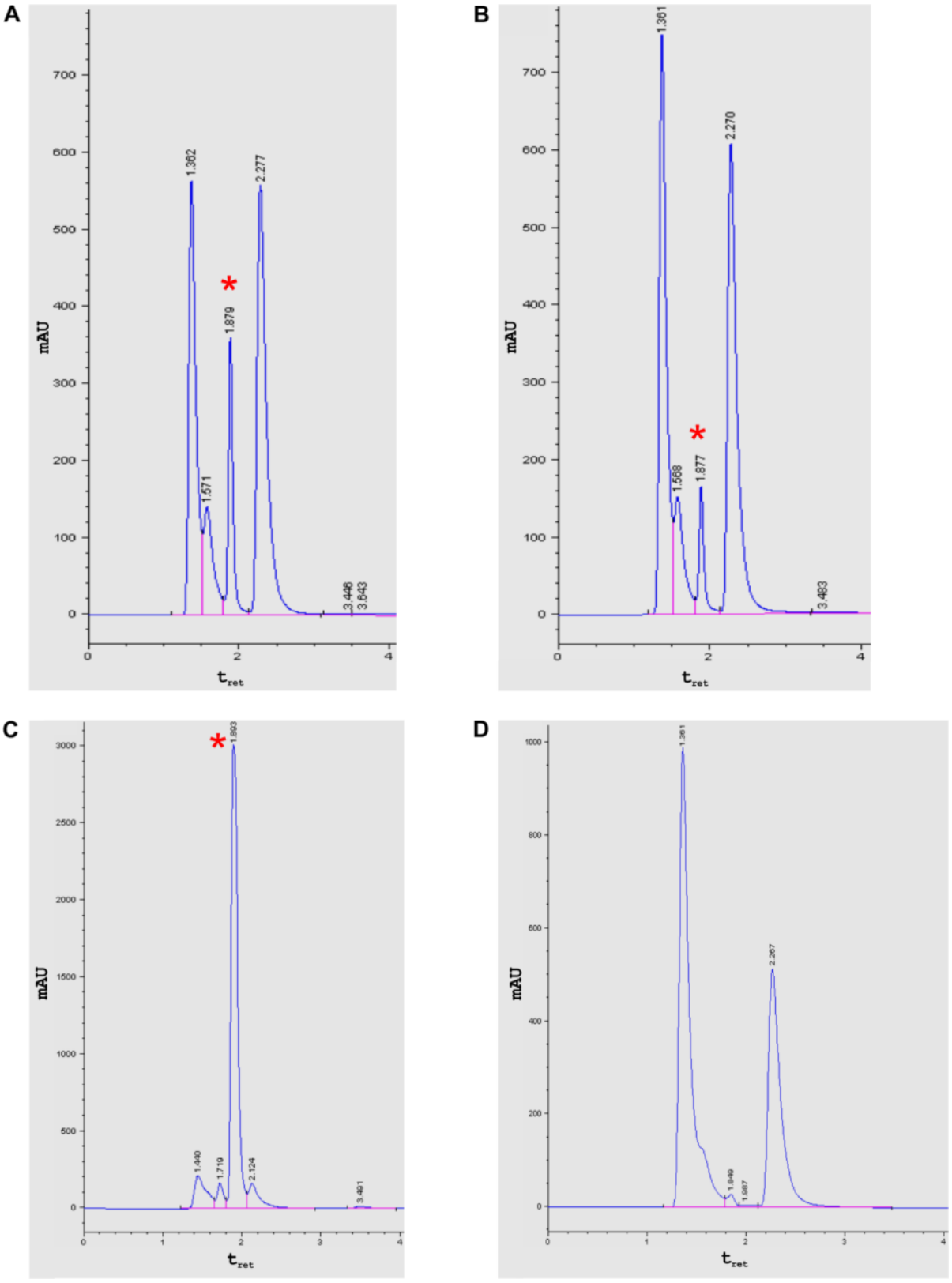


Figure S2: Chromatograms of *in vitro* reactions show FDA synthesis

Detection at 260 nm. The FDA peak at a retention time (t_{ret}) of about 1.9 is indicated with a red star. Panel A shows the reaction of the fluorinase purified from the strain haboring pET28a(+)-His-flA and panel B from the strain haboring pKLJ12-His-flA. Panel C shows the chromatogram of FDA and panel D the reaction without addition of FlA. The red star indicates the FDA peak. FDA shows another t_{ret} in

these measurements compared to others presented in this study due to technical problems of the HPLC device.

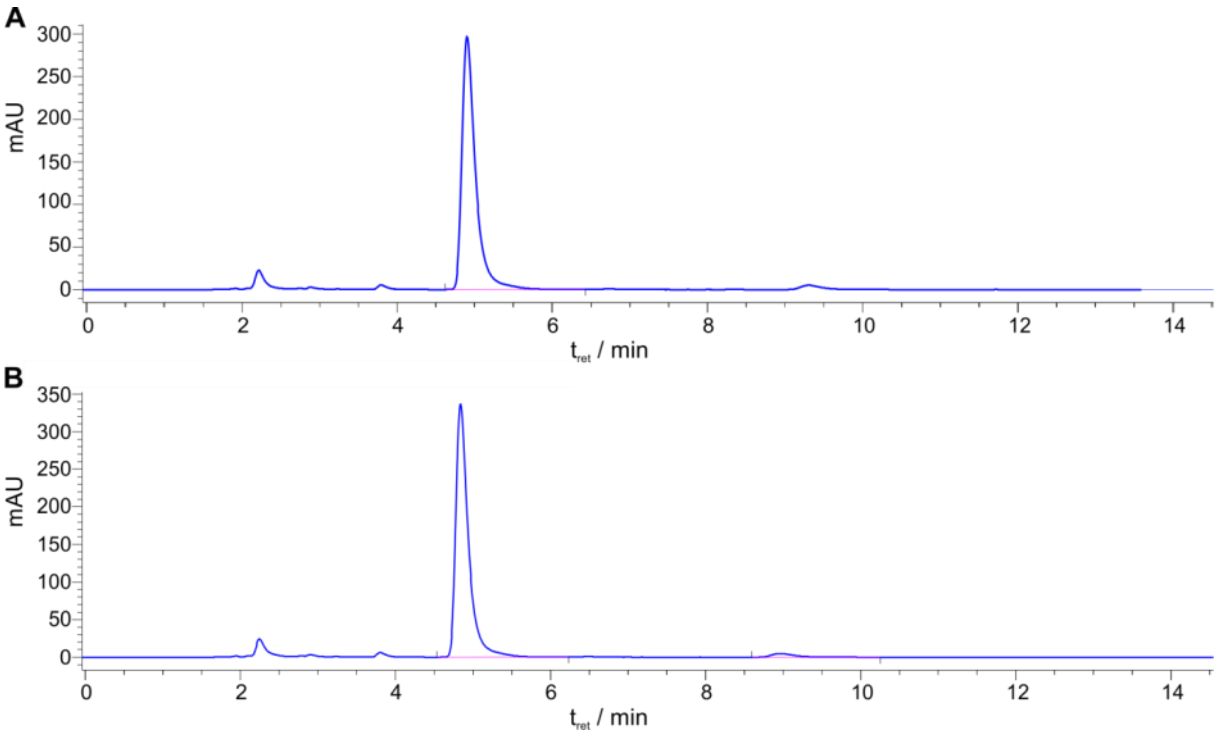


Figure S3: Chromatograms of FDA showing its stability under the applied reaction conditions

Detection at 260 nm. 5 µl of 50 µg/mL FDA was injected for HPLC analysis. Panel A shows the chromatogram before and panel B after processing FDA the same way as reactions with lyophilized cells (incubation for 1 h at 37 °C followed by protein precipitation at 95 °C for 5 min; for details see Materials and Methods, chapter 5.5.2).

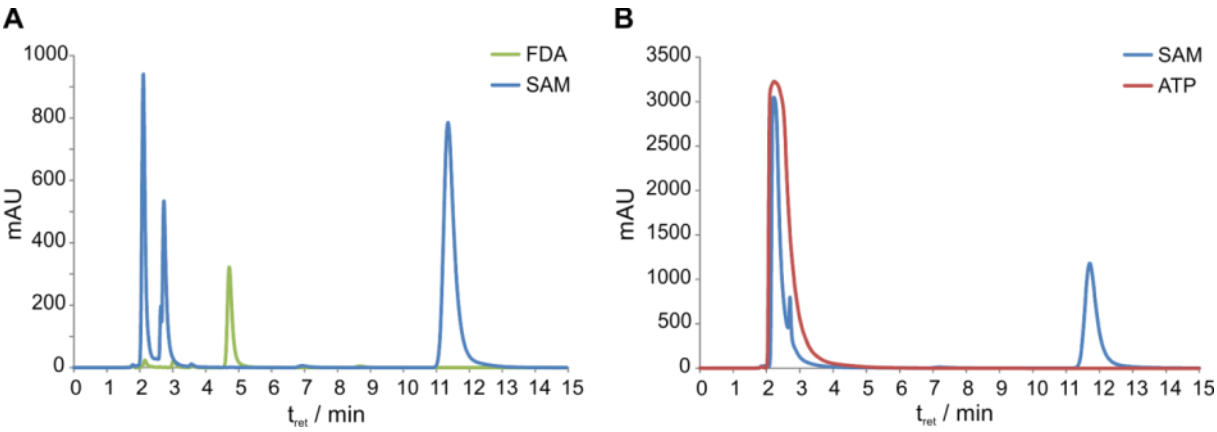


Figure S4: Chromatogram of FDA, SAM und ATP

HPLC analysis of 50 µg/ml FDA (A), 10 mM SAM (A and B) and 100 mM ATP (B). Detection at 260 nm.

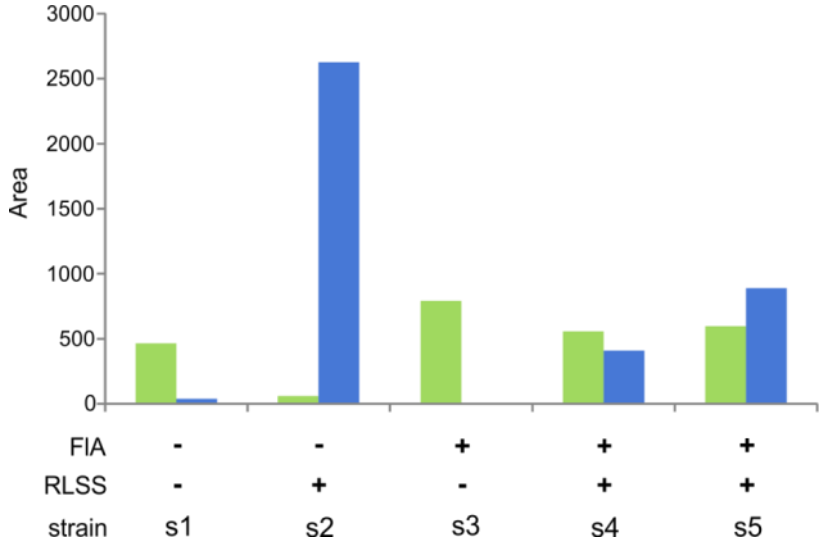


Figure S5: Relative levels of SAM (blue) and of an unidentified substance (green) at the retention time of FDA in lyophilized cells expressing different combinations of FIA and RLSS

Integrated areas of the relevant peaks (single experiment) are shown. Detection at 260 nm. FIA refers to the fluorinase and RLSS to the SAM synthase of *R. norvegicus*. Suspensions of 100 mg/mL lyophilized cells were processed like reaction mixes without addtion of substrates (for details see chapter 5.5.2). Lyophilized cells of different strains were analysed: the empty vector strain (s1) and the strain only overexpressing RLSS (s2); the strain only overexpressing FIA (s3) and the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 0.9 (s4); the strain overexpressing RLSS and FIA with FIA induced at a D_{600} of 0.9 (s4) or at a D_{600} of 4.5 (s5).

2.10.3 **Supplementary Tables**

Table S1: Codon usage table of *E. coli* Class1 used to optimize the fluorinase gene sequence with the program Gene Designer

Amino Acid	Codon	Frequency	Amino Acid	Codon	Frequency
A	GCG	0.32	N	AAC	0.83
A	GCT	0.28	N	AAT	0.17
A	GCA	0.24	P	CCG	0.72
A	GCC	0.16	P	CCA	0.15
C	TGC	0.61	P	CCT	0.11
C	TGT	0.39	P	CCC	0.02
D	GAC	0.54	Q	CAG	0.81
D	GAT	0.46	Q	CAA	0.19
E	GAA	0.75	R	CGT	0.64
E	GAG	0.25	R	CGC	0.33
F	TTC	0.71	R	AGA	0.01
F	TTT	0.29	R	CGG	0.01
G	GGT	0.51	R	CGA	0.01
G	GGC	0.43	R	AGG	0.0
G	GGG	0.04	S	TCT	0.33
G	GGA	0.02	S	TCC	0.27
H	CAC	0.7	S	AGC	0.24
H	CAT	0.3	S	TCG	0.07
I	ATC	0.66	S	TCA	0.05
I	ATT	0.33	S	AGT	0.04
I	ATA	0.01	T	ACC	0.54
K	AAA	0.79	T	ACT	0.29
K	AAG	0.21	T	ACG	0.13
L	CTG	0.77	T	ACA	0.04
L	CTC	0.08	V	GTT	0.4
L	CTT	0.06	V	GTG	0.27
L	TTG	0.05	V	GTA	0.2
L	TTA	0.03	V	GTC	0.13
L	CTA	0.01	W	TGG	1.0
M	ATG	1.0	Y	TAC	0.65
			Y	TAT	0.35

Table S2: Minimal medium composition

Final concentrations are indicated. Chemicals were purchased from Carl Roth, Karlsruhe, Germany unless indicated otherwise.

Description	Component	c [unit as indicated]
Trace elements	$FeSO_4 \cdot 7\ H_2O$	9 µM
	$MnSO_4 \cdot H_2O$	3.5 µM
	$AlCl_3 \cdot 6\ H_2O$	2.5 µM
	CoCl2·6 H2O	2 µM
	$ZnSO_4 \cdot 7\ H_2O$	0.4 µM
	$Na_2MoO_4 \cdot 2\ H_2O$	0.5 µM
	$CuCl_2 \cdot 2$ H2O	0.4 µM
	H_3BO_3	0.5 µM
Salts	Na_2HPO_4	47.8 mM
	KH_2PO_4	22.0 mM
	NaCl	8.6 mM
	NH_4Cl	18.6 mM
Other ingredients	Glucose	20.0 mM
	$MgSO_4 \cdot 7\ H_2O$	10.0 mM
	$CaCl_2 \cdot 2$ H2O	7 nM $CaCl_2$
	19 amino acid supplementation (without Trp)	0.5 mg/L
	Trp	18 µM
	D(+)-Biotine	4.0 µM
	Thiamine hydrochloride	3.3 µM
	Ampicillin (Sigma-Aldrich, St. Louis, MO)	25 mg/L

2.11 References

1. Odar C, Winkler M, Wiltschi B: **Fluoro amino acids: a rarity in nature, yet a prospect for protein engineering.** *Biotechnol J* 2015, **10:**427-446.
2. O'Hagan D, Schaffrath C, Cobb SL, Hamilton JTG, Murphy CD: **Biochemistry: biosynthesis of an organofluorine molecule.** *Nature* 2002, **416:**279.
3. Sanada M, Miyano T, Iwadare S, Williamson J, Arison B, Smith J, Douglas A, Liesch J, Inamine E: **Biosynthesis of fluorothreonine and fluoroacetic acid by the thienamycin producer, Streptomyces cattleya.** *J Antibiot* 1986, **39:**259-265.
4. Deng H, Cross SM, McGlinchey RP, Hamilton JTG, O'Hagan D: **In vitro reconstituted biotransformation of 4-fluorothreonine from fluoride ion: application of the fluorinase.** *Chem Biol* 2008, **15:**1268-1276.
5. Wright F, Bibb MJ: **Codon usage in the G+C-rich Streptomyces genome.** *Gene* 1992, **113:**55-65.
6. Chan KKJ, O'Hagan D: **The rare fluorinated natural products and biotechnological prospects for fluorine enzymology.** In *Meth Enzymol. Volume* Volume 516. Edited by David AH: Academic Press; 2012: 219-235
7. Zhao C, Li P, Deng Z, Ou H-Y, McGlinchey RP, O'Hagan D: **Insights into fluorometabolite biosynthesis in Streptomyces cattleya DSM46488 through genome sequence and knockout mutants.** *Bioorg Chem* 2012, **44:**1-7.
8. Schaffrath C, Deng H, O'Hagan D: **Isolation and characterisation of 5′ -fluorodeoxyadenosine synthase, a fluorination enzyme from Streptomyces cattleya.** *FEBS Lett* 2003, **547:**111-114.
9. Dong C, Huang F, Deng H, Schaffrath C, Spencer JB, O'Hagan D, Naismith JH: **Crystal structure and mechanism of a bacterial fluorinating enzyme.** *Nature* 2004, **427:**561-565.
10. Zhu X, Robinson DA, McEwan AR, O'Hagan D, Naismith JH: **Mechanism of enzymatic fluorination in Streptomyces cattleya.** *J Am Chem Soc* 2007, **129:**14597-14604.
11. Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB, Breaker RR: **Widespread genetic switches and toxicity resistance proteins for fluoride.** *Science* 2012, **335:**233-235.
12. Saunders BC, Stacey GJ: **Toxic fluorine compounds containing the C-F link; methyl fluoroacetate and related compounds.** *J Chem Soc* 1948, **70:**1773-1779.
13. Purser S, Moore PR, Swallow S, Gouverneur V: **Fluorine in medicinal chemistry.** *Chem Soc Rev* 2008, **37:**320-330.
14. Walker MC, Wen M, Weeks AM, Chang MCY: **Temporal and fluoride control of secondary metabolism regulates cellular organofluorine biosynthesis.** *ACS Chem Biol* 2012, **7:**1576-1585.
15. Iwai N, Kitahara Y, Kitazume T: **One-pot three-step continuous enzymatic synthesis of 5-fluoro-5-deoxy-d-ribose.** *J Mol Catal B Enzym* 2011, **73:**1-4.
16. Murphy CD, O'Hagan D, Schaffrath C: **Identification of a PLP-dependent threonine transaldolase: a novel enzyme involved in 4-fluorothreonine biosynthesis in Streptomyces cattleya.** *Angew Chem Int Ed Engl* 2001, **40:**4479-4481.
17. Wiltschi B: **Expressed protein modifications: Making synthetic proteins.** In *Methods in Molecular Biology - Synthetic Gene Networks. Volume* 813. Edited by Weber W, Fussenegger M: Humana Press; 2012: 211-225: *Methods in Molecular Biology*].
18. Su X, Zhang S, Wang L, Dong Z: **Overexpression of IbpB enhances production of soluble active Streptomyces olivaceovirdis XynB in Escherichia coli.** *Biochem Biophys Res Commun* 2009, **390:**673-677.
19. Chen J, Song J-l, Zhang S, Wang Y, Cui D-f, Wang C-c: **Chaperone activity of DsbC.** *J Biol Chem* 1999, **274:**19601-19605.
20. Nozach H, Fruchart-Gaillard C, Fenaille F, Beau F, Ramos OHP, Douzi B, Saez NJ, Moutiez M, Servent D, Gondry M, et al: **High throughput screening identifies disulfide isomerase DsbC as a very efficient partner for recombinant expression of small disulfide-rich proteins in E. coli.** *Microb Cell Fact* 2013, **12:**37-37.
21. Ferrè F, Clote P: **DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification.** *Nucl Acids Res* 2006, **34:**W182-W185.
22. Paoni NF, Koshland DE, Jr.: **Permeabilization of cells for studies on the biochemistry of bacterial chemotaxis.** *Proc Natl Acad Sci U S A* 1979, **76:**3693-3697.
23. Posnick LM, Samson LD: **Influence of S-adenosylmethionine pool size on spontaneous mutation, Dam methylation, and cell growth of escherichia coli.** *J Bacteriol* 1999, **181:**6756-6762.
24. Alvarez L, Mingorance J, Pajares MA, Mato JM: **Expression of rat liver S-adenosylmethionine synthetase in Escherichia coli results in two active oligomeric forms.** *Biochem J* 1994, **301 (Pt 2):**557-561.
25. Jones KL, Keasling JD: **Construction and characterization of F plasmid-based expression vectors.** *Biotechnol Bioeng* 1998, **59:**659-665.
26. Goodman DB, Church GM, Kosuri S: **Causes and effects of N-terminal codon bias in bacterial genes.** *Science* 2013, **342:**475-479.
27. Schneider CA, Rasband WS, Eliceiri KW: **NIH Image to ImageJ: 25 years of image analysis.** *Nat Meth* 2012, **9:**671-675.
28. Gurney T, Jr.: **Preparation of lyophilized cells to preserve enzyme activities and high molecular weight nucleic acids.** In *Nucleic Acids. Volume* 2. Edited by Walker J: Humana Press; 1984: 35-42: *Methods in Molecular Biology*].
29. Steinreiber J, Fesko K, Mayer C, Reisinger C, Schürmann M, Griengl H: **Synthesis of γ-halogenated and long-chain β-hydroxy-α-amino acids and 2-amino-1,3-diols using threonine aldolases.** *Tetrahedron* 2007, **63:**8088-8093.
30. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO: **Enzymatic assembly of DNA molecules up to several hundred kilobases.** *Nat Meth* 2009, **6:**343-345.
31. Balzer D, Ziegelin G, Pansegrau W, Kruft V, Lanka E: **KorB protein of promiscuous plasmid RP4 recognizes inverted sequence repetitions in regions essential for conjugative plasmid transfer.** *Nucl Acids Res* 1992, **20:**1851-1858.
32. Zheng L, Baumann U, Reymond J-L: **An efficient one-step site-directed and site-saturation mutagenesis protocol.** *Nucl Acids Res* 2004, **32:**e115-e115.
33. Odar C, Fladischer P, Niklaus A, Schöffmann H, Murgu O, Wiltschi B: **The incorporation of in vivo synthesized tryptophan analogs into target proteins in E. coli is dependent on the expression system.** in preparation.
34. Laemmli UK: **Cleavage of structural proteins during the assembly of the head of bacteriophage T4.** *Nature* 1970, **227:**680-685.
35. Fuhrmann M, Hausherr A, Ferbitz L, Schödl T, Heitzer M, Hegemann P: **Monitoring dynamic expression of nuclear genes in Chlamydomonas reinhardtii by using a synthetic luciferase reporter gene.** *Plant Molecular Biology* 2004, **55:**869-881.

3 Chapter

Residue-specific incorporation of biosynthesized tryptophan analogs: The protein expression system matters

Corinna Odar[1,2], Carina Sommer[1,2], Patrik Fladischer[1,2], Niklaus Anderhuber[1], Heidemarie Schöffmann[1], Octavian Murgu[1], Birgit Wiltschi[1]

[1] Austrian Centre of Industrial Biotechnology, Petersgasse 14, Graz, Austria

[2] Graz University of Technology, Institute of Molecular Biotechnology, Petersgasse 14, Graz Austria

Authors' contributions

Corinna Odar designed the study under the guidance of Birgit Wiltschi. Corinna Odar and Carina Sommer performed the experiments. Nikolaus Anderhuber and Octavian Murgu designed and constructed the Δ*trpC*::0 expression strain. Patrik Fladischer and Heidemarie Schöffmann designed and performed the measurements to assess the limiting tryptophan concentration.

3.1 **Abstract**

Background: Residue specific labeling of proteins with non-canonical amino acids has become a promising tool in the protein engineer's toolbox. The replacement of tryptophan by its non-canonical counterparts attracts special attention due to tryptophan's extraordinary status in the canonical set of amino acids: It is encoded by a single triplet codon (UGG); it is a relatively rare amino acid in proteins; it largely contributes to a protein's spectroscopic properties; and it is involved in molecular interactions like hydrogen bonding, π-stacking and cation-π interactions. However, most-tryptophan analogs are not commercially available or are forbiddingly costly for large-scale applications.

Results: We established a simple and straight forward procedure for the *in vivo* enzymatic synthesis of non-canonical tryptophans (ncTrp) from indol precursors and their residue-specific incorporation into target proteins in *E. coli*. To enhance the biotransformation of a broad spectrum of indole analogs to the corresponding ncTrp we co-expressed the tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) on top of the tryptophan synthase of the host. A tryptophan auxotrophic strain equipped with *St*TRPS facilitated the full labeling of the enhanced cyan fluorescent protein with ncTrp synthesized from a panel of indole analogs. Using exclusively the host TRPS, fluorinated indole analogs were converted to the corresponding ncTrp and efficiently introduced into the hydroxynitril lyase from *Granulicella tundricola (Gt*HNL). However, *Gt*HNL was not expressed with indoles carrying polar substituents. *Gt*HNL was encoded on a pET vector whose strong T7*lac* promoter is transcribed by the T7 RNA polymerase. Since the expression of the T7 RNA polymerase is co-regulated with the target protein, we evaluated whether the enzyme could cope with the accidental incorporation of the biosynthesized ncTrp. All tested indole analogs except the fluorinated ones either impaired the expression of the variant proteins from the pET vector or abolished it entirely. Yet, the expression from the alternative pQE system remained unaffected.

Conclusion: The described procedure facilitates the biosynthesis of ncTrp from inexpensive indole precursors and their residue-specific incorporation. The target protein determines if *St*TRPS co-expression is essential or not. The widely used pET expression system is compatible with only a subset of fluorinated tryptophan analogs.

3.2 **Background**

Non-canonical tryptophan (ncTrp) derivatives were successfully incorporated into proteins to engineer spectroscopic traits [1-3], to modulate enzyme activity [4], or to elucidate protein structure and function [5, 6]. The commonly used ncTrp are structural analogs of the canonical amino acid tryptophan (Trp). Replacement of Trp by these ncTrp often affects the physicochemical properties of proteins while the overall protein structure is preserved [2]. Replacement of Trp by any of the 19 other canonical amino acids prescribed by the genetic code can be deleterious as Trp is unique in size and chemistry [7]. Phenylalanine (Phe) and tyrosine (Tyr) are sterically similar aromatic amino acids [3]. However, the side chain of Trp is bulkier than that of Phe and Tyr. Trp contains a bicyclic indole moiety as compared to the monocyclic phenyl ring of Phe and Tyr. In addition, Tyr carries a hydrophilic hydroxyl group, which is absent from Trp. Histidine is polar and less aromatic than Trp and the hydrophobic aliphatic amino acids such as glycine, alanine, valine, leucine or isoluecine are substantially smaller. The exchange of Trp by another canonical amino acid can therefore lead to structural and/or functional perturbations.

Usually, ncTrp carry (small) substituents in or attached to the indole side chain, such as sulfur, selenium, nitrogen or halogen atoms and amino- hydroxy- or methyl groups [8]. They modulate the physicochemical properties of Trp without grossly changing its structure. This is why ncTrp can be smart substitutes for Trp.

Non-canonical amino acids (ncAAs) can be introduced into proteins by substituting a structurally and/or chemically similar canonical amino acid (residue-specific incorporation) or in response to a stop or quadruplet codon (site-specific incorporation) [9]. The latter technique makes use of an orthogonal pair consisting of an engineered aminoacyl-tRNA synthetase and a suppressor tRNA [10]. Orthogonal pairs for the site-specific incorporation of Trp analogs at an amber stop codon were described recently [11]. For the residue specific-incorporation of ncTrp instead of Trp into a protein expressed in *E. coli*, a Trp auxotrophic strain is required [12]. The Trp auxotrophy facilitates control of the intracellular Trp and ncTrp levels by supplementing the medium with the respective amino acid. In contrast to the site-specific incorporation, an orthogonal pair is not necessary as the host tryptophanyl-tRNA synthetase (TrpRS) accepts derivatives of Trp as substrates if they are chemically and structurally similar. However, it only does so in the virtual absence of its natural substrate Trp, which can be achieved by depriving the auxotroph of Trp in the medium. If the auxotroph is fed with an ncTrp instead, the TrpRS can charge the Trp-specific transfer-RNA ($tRNA_{CCA}^{Trp}$) with the ncTrp by virtue of its substrate tolerance. Consequently, the ncTrp will be incorporated at all Trp codons (UGG) instead of Trp (codon reassignment). Different protocols were devised for this supplementation based incorporation (SPI) of ncAAs [13-15] and the translational permissivity of Trp analogs containing fluoro-, chalcogen-, aza-, hydroxy-, or amino substituents as well as of 4-methyltryptophan was previously shown [1, 16]

However, a major obstacle for a routine and large scale application of ncTrp is the unavailability or high cost of the Trp analogs, as for instance compared to indole analogs. 4-aminotryptophan ($4NH_2$-Trp), 4-azatryptohan (4aza-Trp), 4-hydroxytryptohan (4OH-Trp) and 7-fluorotryptohan (7F-Trp) (the chemical structures are shown in Figure 1A) are not available or available by custom synthesis. 5-

fluoro-L-tryptohane (5F-Trp) is 33-fold and 5-fluoro-D,L-tryptohan four times more expensive than 5-fluoroindole (5F-indole) (for details see Additional file 1).

The chemical synthesis of Trp analogs involves multi-step specialized procedures. Recently, the biosynthesis of 5-hydroxytryptophan from glucose was described, albeit the ncTrp was produced at relatively low titers [17]. For this reason, the direct enzymatic condensation of simple indole precursors with L-serine by the tryptophan synthase (EC 4.2.1.20; TRPS) is a convenient alternative [18]. TRPS consists of an α- and a β-subunit that form a dimer of dimers [19, 20]. It catalyzes the last step of the Trp biosynthesis (Additional file 2: Figure S1). The active enzyme contains a hydrophobic intramolecular tunnel for the retroaldol cleavage of indole-3-glycerol to indole and D-glyceraldehyde-3-phosphate, and the subsequent condensation of indole with serine to Trp (Figure 1B) [21].

TRPS was previously demonstrated to synthesize ncTrp from indole analogs and L-serine: *E. coli* lysates overexpressing TRPS from *Salmonella enterica* (*Se*TRPS) efficiently produced Trp analogs [22, 23]. Engineered *E. coli* biofilms only showed ncTrp synthesis when overexpressing *Se*TRPS [24]. As well, the cognate TRPS of *E. coli* is capable of converting a range of indole analogs [25-28]. Several studies on protein variants containing ncTrp used the *in vitro* conversion by TRPS to produce the desired Trp analogs [3, 27, 29-31]. This required the heterologous overexpression of TRPS, often of *Salmonella typhimurium* (*St*TRPS), in *E. coli* and the purification of the enzyme. The TRPS homologs mentioned above share high sequence similarities on the amino acid level: TRPS of the two *Salmonella* species only differs in one amino acid residue in the α-subunit (α- and β-subunit of *St*TRPS from pEBA-10 [32] *versus* α*Se*TRPS [NCBI:WP_023236840] and β*Se*TRPS [NCBI:WP_023236840]). The α-subunit of *St*TRPS is 85% identical to the α-subunit [NCBI:WP_038339275.1] of the *E. coli* TRPS while the β-subunit [NCBI:WP_021550874.1] shows even 97% identity (C. Odar and B. Wiltschi, our own finding).

Budisa and co-workers reported the direct addition of azaindoles to the growth medium which were then intracellularly converted into the corresponding azatryptophan analogs by the TRPS of *E. coli*, and were translated into human annexin A5 [1]. This approach represents a time- and cost-saving alternative as it bypasses the enzyme purification step necessary for the *in vitro* reaction and it avoids the procurement of expensive Trp analogs. However, the group did not analyze further whether the approach could be extended to other ncTrp nor did they elaborate a generally applicable protocol.

In this study, we conducted a systematic analysis of the combined *in vivo* synthesis of different ncTrp and their incorporation into proteins in *E. coli*. We employed the TRPS from *Salmonella typhimurium* or the host TRPS for the formation of aza-, fluoro-, hydroxy-, and aminotryptophans. In our study, we included two different target proteins with varying numbers of Trp residues. Enhanced cyan fluorescent protein (ECFP) contained two Trp residues and the hydroxynitril lyase from *Granulicella tundricola* (*Gt*HNL) contained seven. Finally, we assessed the incorporation efficiency of selected ncTrp into the target proteins with the widely used pET system for protein expression [33, 34] in comparison to pQE, which is commonly used for the residue-specific incorporation of ncAAs [14, 35, 36].

We found that the co-expression of *St*TRPS was crucial to fully exchange the two Trp residues of ECFP with ncTrp. In contrast, the activity of the host TRPS was sufficient to achieve the full substitution of three times more Trp residues in *Gt*HNL. Moreover, we demonstrated that the pET expression system was only compatible with fluorinated indole analogs. Here, we suggest a procedure for the labeling of target proteins with any ncTrp that can be biosynthesized *in vivo* from a cheap indole precursor and cellular L-serine.

3.3 Results and discussion

3.3.1 *In vivo* enzymatic synthesis of ncTrp and the residue-specific incorporation

Previously, the endogenous expression levels of the *E. coli* TRPS had sufficed to support efficient biotransformation of azaindoles to the corresponding ncTrp for their subsequent translation into a target protein [1]. Azaindoles are very similar to indole, in fact, the exchange of one carbon for a nitrogen represents an atomic modification of the indole ring (Figure 1A). We hypothesized that other less similar indole analogs would require elevated levels of TRPS for their efficient *in vivo* bioconversion to the corresponding ncTrp. Since *St*TRPS had been used to synthesize various ncTrp *in vitro* [2, 3, 18, 23, 27, 29, 37, 38] we decided to integrate its co-expression into our established procedure for the residue-specific incorporation of ncTrp into a target protein [15].

This incorporation procedure is separated into three phases: First, Trp auxotrophic cells are grown in the presence of Trp ("growth phase"). Only so much Trp is supplied in the synthetic medium that the Trp auxotroph consumes it entirely and stops growing in the mid-log phase ("starvation phase"). During the growth phase, the expression of the target protein is off. Subsequently, the Trp-starved cells are supplemented with the ncTrp and the expression of the target protein is induced ("production phase"). This three-phase procedure ensures that enough cells accumulate before the production of the target protein with the ncTrp is initiated. The cell accumulation is necessary because the auxotroph does not grow during the protein production phase where the medium contains the ncTrp instead of the canonical Trp that it requires for growth. Nevertheless, ribosomal translation is still active. Provided the efficient uptake of the ncTrp from the medium, which is the case for many Trp analogs [16, 27], the intracellular ncTrp level raises above the threshold that allows its charging onto $tRNA_{CCA}^{Trp}$ by the host TrpRS and the incorporation into the target protein. The incorporation is more efficient the higher the ncTrp level is relative to any residual canonical Trp.

To tap the full potential of the *in vivo* synthesis of Trp analogs from the corresponding indole precursors and cellular L-serine we supplemented the Trp auxotrophic cells with the indole analog instead of a Trp analog. Indole is readily taken up by *E. coli* cells [39]. We organized the *in vivo* synthesis such that ncTrp would be present in adequate amounts when the target protein expression was switched on. As soon as the intracellular ncTrp pool exceeded the threshold for its activation by the TrpRS it would be incorporated into any protein newly translated at the same time. This was the desired event for the target protein, however, also *St*TRPS would be labeled with the ncTrp if it was

expressed at that instant. The accidental incorporation of an ncTrp into *St*TRPS might have impaired its function and, consequently, the cellular supply with more ncTrp. We anticipated that an early onset of *St*TRPS expression when Trp was still present would equip the cells with functional enzyme for an efficient *in vivo* synthesis of the ncTrp. In other words, the expression of the *St*TRPS had to start before that of the target protein (Figure 2A). To uncouple the expression of the *St*TRPS and the target protein, we chose an arabinose- and an IPTG inducible promoter, respectively, to directedly drive the expression of the individual genes. In addition, we added the indole at the onset of the starvation phase, shortly before we induced target protein expression (Figure 2A). This strategy allowed the formation of a starter pool of ncTrp for incorporation following shortly after. Yet, it reduced the potentially harmful premature incorporation of the ncTrp into essential cellular proteins to a negligible extent.

A Trp auxotrophic *E. coli trpC::0* deletion mutant that carried a functional genomic copy of the TRPS served as the expression host in all experiments. We titrated the Trp supplementation of the culture medium such that cell growth stopped due to Trp depletion at an attenuance at 600 nm (D_{600}) of 1.8 (data not shown, see the Methods section for details). We sought to keep the metabolic burden of the heterologous expression of *St*TRPS low yet to provide enough enzyme for the efficient transformation of the indole precursors. Therefore, *St*TRPS was expressed under the arabinose inducible *araBAD* promoter from a medium copy plasmid. To evaluate whether the expression of *St*TRPS affected the depletion of Trp we monitored the growth pattern when *St*TRPS was induced at cell densities equaling D_{600} 0.5, 1 and 1.5 with 200 mg/L, 20 mg/L and 2 mg/L arabinose as the inducer (Additional file 3: Figure S2). Growth stalled at D_{600} 1.8 or slightly below as expected when we induced the expression of *St*TRPS at D_{600} 1 or 1.5 with 2 mg/L or 20 mg/L arabinose (Table 1 and Additional file 3: Figure S2). The growth arrest was caused by the depletion of Trp and not by any other medium component as the cells resumed to grow when they were supplemented with Trp (Additional file 3: Figure S2). However, when *St*TRPS was induced earlier (D_{600} 0.5) or with a higher inducer concentration (20 and 200 mg/L arabinose) the growth arrest became less pronounced or disappeared entirely (Table 1 and Additional file 3: Figure S2). Surprisingly, these results indicate that a strong and long overexpression of *St*TRPS might enable the cells to bypass their Trp auxotrophy. So far, we have not been able to explain this effect. In the Trp auxotrophic strain, the entire *trpC* gene was deleted except for the first six nucleotides of the coding sequence and 21 nucleotides at the 3'-terminus (*trpC::0* mutant; see the Methods section for strain generation). Therefore, a reversion of the mutation appears highly unlikely. TrpC catalyzes the formation of the precursor molecule for indole synthesis, indole-3-glycerol-phosphate, which is used in the subsequent step by TRPS to synthesize Trp (Additional file 2: Figure S1). The deletion of *trpC* should completely abolish the strain's capability for Trp biosynthesis. Indeed, it did under all conditions except when high amounts of arabinose were applied to induce *St*TRPS at a comparably low cell density (Additional file 3: Figure S2A, 200 mg/L arabinose, blue diamonds). Since we observed the same growth behavior in three independent experiments, and to avoid this phenomenon, we induced *St*TRPS expression at D_{600} 1 in all following experiments.

We next assessed the influence of the *St*TRPS expression level on the titer of a synthetic protein containing a Trp analog. We used 4-aminoindole ($4NH_2$-indole) as the model substrate for *St*TRPS

and the enhanced cyan fluorescent protein (ECFP) as the model protein. ECFP contains two Trp residues (W57 and W66; Additional file 4: Amino acid sequence 1) that were successfully exchanged for $4NH_2$-Trp before [2]. The residue-specific incorporation of $4NH_2$-Trp into the fluorophore of ECFP causes a significant red shift of the fluorescence. This spectroscopic change affords a straight-forward assessment of the incorporation efficiency of the analog.

We transformed the tryptophan auxotrophic strain that carried the *St*TRPS expression plasmid with an IPTG-inducible high copy expression vector for ECFP. The strain was induced for *St*TRPS expression at D_{600} 1 with 200 mg/L, 20 mg/L or 2 mg/L arabinose. Induction with 2 mg/L arabinose showed the expected growth arrest at a D_{600} 2, while D_{600} values were slightly above (D_{600} ~2.3) when 200 and 20 mg/L arabinose were applied for *St*TrpS induction (Figure 2B). We supplemented the cells with 1 mM $4NH_2$-indole and 30 min later we induced the expression of ECFP with IPTG. The β-subunit of *St*TRPS and ECFP were N-terminally hexahistidine-tagged for their immunodetection (Figure 2C). The expression of *St*TRPS under the arabinose inducible promoter was leaky (Figure 2C, no supplementations, samples a, c and d). In comparison, no ECFP expression could be detected before the induction with IPTG (Figure 2C, samples b). While the low leaky expression of *St*TRPS can be tolerated, tight expression of the target protein is important to ensure efficient labeling with the ncTrp. The immunoblot showed that slightly more [$4NH_2$-Trp]ECFP was produced after 3 h of expression when less arabinose was used for the induction of *St*TRPS (Figure 2C, samples c, 2 mg/L and 20 mg/L *versus* 200 mg/L arabinose). However, the expression level of *St*TRPS did not visibly affect the [$4NH_2$-Trp]ECFP variant protein levels after overnight induction (Figure 2C, samples d). Relative to the titers after 3 h of induction, the accumulation of *St*TRPS, specifically with 200 mg/L arabinose for induction, was more pronounced than that of [$4NH_2$-Trp]ECFP. Based on these observations we applied 2 mg/L arabinose for *St*TRPS induction at D_{600} 1 (Figure 2A and B).

We assessed the effect of $4NH_2$-indole addition either at Trp starvation (protocol A) or earlier during the process, when we induced the expression of *St*TRPS (protocol B, Additional file 5: Figure S3). We reasoned that a prolonged conversion of $4NH_2$-indole to $4NH_2$-Trp by *St*TRPS for 3.5 h as in protocol B *versus* 0.5 h as in protocol A could lead to higher levels of the ncTrp and consequently to an enhanced expression of [$4NH_2$-Trp]ECFP. However, the accumulation of $4NH_2$-Trp during the growth phase could lead to its accidental incorporation into *St*TRPS and cellular proteins despite the presence of Trp, which might affect the enzyme function and/or growth. Indeed, we observed that the addition of $4NH_2$-indole before the depletion of Trp led to a growth pattern without a visible growth arrest (Additional file 5: Figure S3). The absence of a growth arrest with protocol B suggests that $4NH_2$-Trp was produced and that it supplemented the Trp auxotrophy of the strain, substituting for Trp after its depletion or maybe even before. Likewise, growth resumed after the later addition of $4NH_2$-indole (Additional file 5: Figure S3, protocol A). The cell growth on Trp analogs is discussed in more detail in chapter 2.2.

Replacement of W66 in ECFP by $4NH_2$-Trp red-shifts the emission wavelength, which is visible as gold fluorescence [2]. This change of the spectroscopic traits served as a proof for the incorporation of $4NH_2$-Trp into ECFP at the W66 position (Figure 3A). Applying protocol A we produced [$4NH_2$-Trp]ECFP with a fluorescence spectrum typical for the gold fluorescent protein [2]. Protocol B resulted

in only partial labelling of ECFP with $4NH_2$-Trp as was apparent from the emission spectrum showing fluorescence characteristics of both, the [$4NH_2$-Trp]ECFP as well as ECFP. The fluorescence measurements where validated by mass spectrometric analysis of the incorporation efficiencies (Figure 3B and Additional file 6: Table S1). The incorporation of $4NH_2$-Trp at both Trp positions with protocol A was confirmed whereas with protocol B only 28.7% of the variant preparation showed full incorporation (Figure 3B). We speculate that the induction of *St*TRPS and the concomitant addition of $4NH_2$-indole in protocol B caused an accumulation of $4NH_2$-Trp while Trp was still available. From previous observations [35] it is reasonable to presume that both amino acids sustained the growth of the Trp auxotrophic expression strain. The absence of the growth arrest in protocol B apparently confirms this assumption (Additional file 5: Figure S3). Consequently, the mixture of tryptophans resulted in the mixed variant species that we observed with protocol B. In protocol A, $4NH_2$-indole was added *after* the exhaustion of Trp. Thus, we expect that the intracellular level of $4NH_2$-Trp had been substantially higher than that of Trp, which facilitated the observed full labelling of ECFP with the Trp analog.

Based on these observations, we devised the following ultimate procedure for the *in vivo* synthesis of an ncTrp and its residue-specific incorporation into a target protein (Figure 2A): *St*TRPS was induced with 2 mg/L arabinose at D_{600} 1. The minimal medium was supplemented with 18 µM Trp so that depletion occurred at a D_{600} between 1.8 and 2. At the same time, the indole analog was added. 30 minutes later the target protein was induced and expressed over night.

3.3.2 The effect of the co-expression of the tryptophan synthase on the incorporation efficiency of different tryptophan analogs

We validated the procedure described in section 2.1 with a number of different indole analogs. In addition to $4NH_2$-indole, we supplemented the Trp auxotrophic *E. coli* cells with 4aza-, 7aza-, 4OH-, 5OH-, 5F-, 6F- and 7F-indole (Figure 1A).

Many non-canonical amino acids are cytotoxic compounds [40] and usually they do not support the growth of auxotrophic cells. In the present study, we observed that the cells were able to grow on the Trp analogs after the depletion of Trp, albeit to different extents (Figure 2B, other data not shown). We observed cell growth with D_{600} values between 2.1 and 3.7 after overnight cultivation when we supplemented the cells with the hydroxylated indole analogs and with those containing aza- and amino-substituents. On the fluorinated indoles, the cells reached densities of up to D_{600} 5.2. Apparently, *E. coli* tolerated the Trp analogs very well. Budisa and co-workers previously reported that a thiophene-derivative of Trp could support the growth of a Trp auxotrophic *E. coli* strain when mixed with Trp [35]. The same group and others demonstrated that auxotrophic *E. coli* can even be evolved to grow on ncTrp as the sole Trp source [41-43].

In the setup experiments described in section 2.1, we had used ECFP as the model protein. ECFP is composed of 251 amino acid residues, two of which are Trp (Trp content 0.8%) (Additional file 4: Amino acid sequence 1). We suspected that the incorporation efficiency of ncTrps was dependent on the relative Trp content of a protein, and therefore we included *Gt*HNL as a second, different model protein in our analysis. *Gt*HNL contains seven tryptophans in a total of 186 amino acids (Trp content

3.8%) (Additional file 7: Amino acid sequence 2). *Gt*HNL is well characterized and the wild type can be expressed at high levels in *E. coli* [44, 45]. In addition to the *St*TRPS expressing strain ((+)*St*TRPS strain) described in chapter 2.1, the same *E. coli* strain without the heterologously expressed *St*TRPS activity ((-)*St*TRPS strain) served as the second synthetic protein production host. While the latter strain relied on its endogenous TRPS for ncTrp synthesis, the co-expression of *St*TRPS in the first strain added extra Trp synthase activity and broadened the indole scope.

Using the procedure described in section 2.1, we expressed both model proteins in the presence of the indoles listed Figure 1A. We determined the relative expression levels of ncTrp variants in the cleared lysate fraction, *i.e.* the soluble protein fraction (Additional file 8: Table S2) by the densitometric analysis of the relevant bands on the SDS gels shown in Figure 4A and B and Figure 5. For most variants, the (-)*St*TRPS strain showed higher relative expression titers. This was especially pronounced with cells that expressed ECFP in the presence of indole, 7aza-, 4OH- and 5OH-indole (Figure 4A and B), as well as for the expression of *Gt*HNL with the fluorinated indoles (Figure 5). Apparently, the co-expression of *St*TRPS in (+)*St*TRPS consumed metabolic energy which was otherwise used for synthetic protein production.

ECFP variants were produced with all indoles with 40 to 260 mg expressed protein per liter culture (determined by total protein concentrations of lysates, data not shown and densitometric analysis of SDS gels displayed in Figure 4A and B). The alpha-subunit of *St*TRPS (calculated molecular weight 29 kDa) and ECFP (calculated molecular weight 28 kDa) are very similar in size. Nevertheless, the SDS-PAGE resolved the two proteins sufficiently well (Figure 4A and B) for the densitometric analysis. To unambiguously confirm the identity of the [ncTrp]ECFP bands we performed an immunodetection of their N-terminal hexahistidine-tag, which confirmed the findings of the SDS-PAGE analysis and refined them. We observed that the variants containing fluorinated Trp analogs were expressed in soluble form while the incorporation of the polar analogs $4NH_2$-Trp, 4- and 7aza-Trp as well as 4- and 5OH-Trp led to a substantial fraction of insoluble ECFP (Figure 4C). Although the expression of ECFP with 7aza-, 5OH- and 7F-indole in the (+)*St*TRPS strain was very low and only detectable by silver staining (Figure 4), we were able to purify enough variant protein each (Additional file 9: Figure S4) for the analysis of the efficiency of the ncTrp incorporation by mass spectrometry. The majority of the variant proteins isolated from the (+)*St*TRPS expression host showed full labeling with the Trp analog (Table 2 and Additional file 10: Table S3). [4aza-Trp]ECFP, [4OH-Trp]ECFP and [7F-Trp]ECFP were only partially labeled. In contrast, the [ncTrp]ECFP variants expressed in the (-)*St*TRPS were very inhomogeneously labeled and consisted of a mix of unlabeled, partially and fully labeled protein species. Apparently, the indole to ncTrp conversion in the (-)*St*TRPS strain had been insufficient for the full labeling of ECFP.

Previously, it was reported that hydroxytryptophans could not be introduced into fluorescent proteins from *Aequorea victoria* although they had been successfully incorporated into other proteins before [2, 16]. The titers of [4OH-Trp]ECFP and [5OH-Trp]ECFP were low, still we were able to produce these hydroxytryptophan labeled ECFP variants using the indole precursors with our newly established procedure. Apparently, the *in vivo* synthesis of the hydroxytryptophans by the *St*TRPS was beneficial for their subsequent translation. The incorporation efficiency could possibly be further improved by the

co-overexpression of the TrpRS as was already demonstrated for Trp analogs [46]. The co-overexpression of the corresponding aminoacyl-tRNA synthetases also proved successful for the incorporation Met [47-49] and Leu analogs [50].

In contrast to ECFP, which could be produced with all indoles used in this study, only the parent *Gt*HNL with Trp and the fluorinated [ncTrp]*Gt*HNL variants were expressed (0.7 to 1 g expressed protein per liter culture in the (+)*St*TRPS strain, determined by total protein concentrations of lysates and densitometric analysis of SDS gels displayed in Figure 5). We were unable to detect the expression of any other [ncTrp]*Gt*HNL by SDS-PAGE analysis (Figure 5). One explanation for the observed expression pattern is that the *Gt*HNL structure tolerated only the fluorotryophans while the other Trp analogs impeded the proper translation and/or folding of the protein. Alternatively, selected ncTrp might have interfered with the transcription of the *Gt*HNL gene as will be discussed in detail in chapter 2.3. A substantial portion of the expressed *Gt*HNL variants, even of the parent protein, ended up in the insoluble fraction and *Gt*HNL expressed with 6F-indole was predominantly insoluble. For the fluorinated [ncTrp]*Gt*HNL variants the improved expression titers in the (-)*St*TRPS strain over the (+)*St*TRPS strain were even more pronounced than for [ncTrp]ECFP (Additional file 8: Table S2). Mass analysis of the purified [ncTrp]*Gt*HNL variants (Additional file 11: Figure S5) confirmed the full labeling of [5F-Trp]*Gt*HNL, [6F-Trp]*Gt*HNL and [7F-Trp]*Gt*HNL in the (-)*St*TRPS strain (Additional file 12: Table S4). The *Gt*HNL variant proteins that had been produced in the (+)*St*TRPS strain were not analyzed since full labeling ocurred already without the co-expression of *St*TRPS. This is again in contrast to the findings with the ECFP variants, which showed complete labeling only when *St*TRPS was co-expressed. The results with *Gt*HNL are surprising since the number of Trp residues to be exchanged to ncTrp is more than three times higher in *Gt*HNL than in ECFP (seven *versus* two). Obviously, the ncTrp incorporation efficiency not only depends on the co-expression of *St*TRPS, but is dependent on the target protein as well.

3.3.3 Incorporation of biosynthesized tryptophan analogs using two different expression systems

We expressed *Gt*HNL using the pET expression system [44, 45]. This system is frequently used for heterologous protein production in *E. coli*. However, its application for SPI, specifically for the procedure described in this study can be inappropriate. On selected pET vectors, gene expression is driven by a strong, IPTG-inducible T7*lac* promoter [51, 52]. The T7 promoter is selectively transcribed by the RNA polymerase of bacteriophage T7 (T7 RNAP) and T7 promoters do not occur naturally in *E. coli* [53]. The Trp auxotrophic descendant of *E. coli* BL21(DE3) Gold that was used in this study carries a single chromosomally integrated copy of the T7 RNAP gene under the control of the IPTG-inducible *lacUV5* promoter [33]. Because IPTG is the inducer of the T7 RNAP as well as of the target gene, the presence of non-canonical amino acids during IPTG induction leads to their incorporation not only into the target protein but also into the T7 RNAP [15]. This is equally true for the biosynthesized Trp analogs described in this study. As soon as Trp in the medium is exhausted the indole analog is added and 30 minutes later IPTG is added to induce the expression of the target protein (see section 2.1). The Trp analogs that have accumulated at this point cannot only be

incorporated into the target but also into T7 RNAP. The enzyme shows a comparably high Trp content of 2.2% (19 Trp in 883 total amino acid residues). We hypothesized that the incorporation of some Trp analogs used in this study might impair the activity of the T7 RNAP or impede its functional expression. This could explain why only a fraction of the Trp analogs, specifically the fluorinated ones, led to the expression of labeled *Gt*HNL as described in the previous section. To shed light on this issue, we systematically evaluated our incorporation procedure for *in vivo* synthesized ncTrp in combination with the pET expression system.

For the incorporation of different ncTrp, we had expressed ECFP under control of the strong bacteriophage T5 promoter (see section 2.2). In contrast to the T7 promoter, transcription from the T5 promoter is initiated by the DNA-dependent *E. coli* RNA polymerase (*Ec*RNAP) [54]. *Ec*RNAP is an essential enzyme. Though the accidental incorporation of ncTrp into the enzyme might occur and render it non-functional, *Ec*RNAP produced during the growth phase in the presence of Trp is functional and can compensate for a potentially non-functional variant. Therefore, T5-based expression systems such as the pQE vectors are preferred for SPI. To test whether the misincorporation of certan ncTrp into T7 RNAP caused the failure of *Gt*HNL to express with indoles other than the fluorinated analogs, we subcloned *Gt*HNL from a pET vector (T7*lac* promoter) into a pQE vector (T5 *lacO* promoter). However, to our surprise and for unknown reasons the expression of the parent *Gt*HNL under control of the T5 promoter was extremely low (Additional file 13: Figure S6). Expression of any labelled *Gt*HNL from these vector construct is highly unlikely, because yields of most variant protein expressions are estimated to be by far lower compared to the yield of the parent protein (Figure 4).To rule out an accidental mutation of the T5 *lacO* promoter we confirmed its correct sequence. The low expression yields from the T5 promoter appears to be a peculiar trait of some members of the cupin protein family, such as *Gt*HNL (B. Wiltschi, personal communication). Consequently, we were unable to directly compare the expression of *Gt*HNL under the control of the two different promoters and with different indole analogs.

Next, we subcloned ECFP from the pQE vector into a pET vector. To elucidate whether the ncTrp synthesized from $4NH_2$-, 4aza-, 7aza-, 4OH- and 5OH-indole (Figure 1A) impaired the function of T7 RNAP we expressed ECFP from the T7*lac* as well as the T5 *lacO* promoter in the presence of these indole analogs. Immunodetection of the hexahistidine-tagged [ncTrp]ECFP variants in whole cell extracts using a hexahistidine-specific antibody showed that the ncTrp species clearly influenced the expression patterns from the two different promoters (Figure 6). Expression of the parent ECFP was elevated with the T7*lac* promoter as compared to the T5 *lacO* promoter (Figure 6, indole), whereas the majority of the [ncTrp]ECFP variants showed an opposing trend. [$4NH_2$-Trp]ECFP and [4aza-Trp]ECFP showed only low expression with the T7*lac* promoter while [4OH-Trp]ECFP was undetectable. In contrast, all these variants were well expressed under control of the T5 *lacO* promoter (Figure 6, $4NH_2$-ind, 4aza-ind, 4OH-ind). The expression level of [7aza-Trp]ECFP was comparably low with both promoters and [5OH-Trp]ECFP apparently was not expressed from either promoter (Figure 6, 7aza-ind, 5OH-ind). The latter finding appears to contradict the earlier results shown in Figure 4 where we were able to immunodetect the [5OH-Trp]ECFP variant. However, the immunoblots in Figure 4 and Figure 6 cannot be compared directly since different amounts of total protein were

analyzed each: In Figure 4 3 µg of total protein were loaded of each sample while Figure 6 shows the expression of the [ncTrp]ECFP variants in whole cell extracts. The latter were normalized as described in the Methods section to ensure equal relative protein amounts in each lane while the absolute protein amount was not determined. We therefore cannot rule out that [5OH-Trp]ECFP was expressed under control of the T5 *lacO* as well as the T7*lac* promoter, albeit definitely at a much lower level than the other variants shown in Figure 6.

Taken together, our results discourage the use of the pET expression system for the incorporation of ncTrp other than monofluorinated analogs. For the comparison of the promoters we cloned N-terminally hexahistidine-tagged ECFP into two different expression vectors, pET21a(+) (T7*lac* promoter) and pQE80L (T5 *lacO* promoter), otherwise the expression strain and the expression conditions were identical. Apparently, the incorporation of ncTrp, specifically of polar analogs such as aza-, amino-, and hydroxy-substituted tryptophans impaired the function of T7 RNAP. This notion is supported by the improved expression of ECFP under control of the T7*lac* promoter as compared to T5 *lacO* in the presence of indole (Figure 6, indole). Since expression of *Gt*HNL under control of the T5 *lacO* promoter was too low for variant production, we cannot rule out that ncTrp other than the fluorinated analogs were incorporated into *Gt*HNL yet caused rapid degradation of the variant proteins, for instance due to misfolding. However, our results with the expression of ECFP from the T7*lac* and T5 *lacO* promoters strongly suggest that the failure to express these *Gt*HNL variants was caused by the formation of a non-functional T7 RNAP with certain ncTrp. Along the same lines, Ayyadurai *et al.* reported the superiority of the pQE expression system over the pET system for the incorporation of Met analogs [55]. Other groups showed the compatibility of the T7-based expression system with 5F-Trp [4, 5, 56], which we confirmed by the successful expression of the [5F-Trp]*Gt*HNL from the T7*lac* promoter (section 2.2). Monofluorinated Trp analogs apparently do not impair the function of T7 RNAP since we were able to incorporporate also 6F-Trp and 7F-Trp into *Gt*HNL (Figure 5).

An alternative way to facilitate the production of ncTrp variants with the T7 promoter/T7 RNAP system might be the use of alternative expression hosts where the T7 RNAP expression is controled by a different inducer than the target protein. Strains with salt- (BL21-SI^{TM}) or arabinose- (BL21-AI^{TM}) inducible promotors are available. These strains would allow to uncouple the expression of the T7 RNAP and the target protein. This might represent a strategy to separate the expression of functional T7 RNAP from the synthesis of the tryptophan analogs and their subsequent incorporation into the target. However, the approach implies that the T7 RNAP is already present before the target protein expression is induced from T7*lac* by IPTG. This bears the danger of leaky target protein expression leading to inhomogeneously labeled variant proteins. Hence, the preferable expression systems to use with the procedure described in this study are independent of the co-expression of T7 RNAP or any other RNAP with the target gene.

3.4 **Conclusions**

We have devised a simple and straight forward procedure for the *in vivo* enzymatic synthesis of various tryptophan analogs from indoles for their subsequent residue-specific incorporation into a target protein. The approach includes the co-expression of the tryptophan synthase from *Salmonella typhimurium* on top of the TRPS activity of the *E. coli* host to enhance the transformation of a broad spectrum of indole analogs to tryptophan. In our study, we evaluated aza-, amino-, hydroxy- and fluoroindoles, which were all transformed into the corresponding tryptophan analogs. The need for the additional *St*TRPS activity depends on the target protein and should be assessed individually. We observed that the co-expression of *St*TRPS slightly reduced the titer of the variant proteins. Although the procedure can be applied for the expression of a target protein using the pET system, one has to keep in mind that the accidental incorporation of an ncTrp, which occurs under these conditions, can impair the function of the T7 RNAP and therefore affect protein expression. In our hands, only monofluorinated Trp analogs were compatible with the pET expression system.

We systematically evaluated the procedure with two target proteins of different Trp content (0.8%, low-Trp *versus* 3.8%, high-Trp). The titers of the expressed variant proteins varied with the incorporated ncTrp. Contrary to our expectations, *St*TRPS co-expression improved the incorporation efficiency with the low-Trp target protein but the host TRPS activity was sufficient for the homogeneous incorporation of fluorinated Trp analogs into the high-Trp target. We anticipate that the *in vivo* synthesis of ncTrp as describe in this study could be combined with the site-specific incorporation of ncTrp [11].

3.5 **Methods**

3.5.1 **Plasmid and strain construction**

All enzymes for cloning and PCR were obtained from Thermo Fisher Scientific (Waltham, MA) unless indicated otherwise. DNA fragments were ligated using T4 DNA Ligase. PCRs were performed using Phusion® High-Fidelity DNA Polymerase. PCR primers were ordered from IDT Inc. (Coralville, IA). *E. coli* TOP10 F' (Life Technologies, Carlsbad, CA) was used for cloning and plasmid amplification.

All chemicals for Gibson reaction mixes were purchased from Carl Roth (Karlsruhe, Germany) unless indicated otherwise. Gibson reaction mixes contained 4 U/mL T5 exonuclease (BioZym, Hessisch Oldendorf, Germany), 25 U/mL Phusion® High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA), 4000 U/mL Taq DNA ligase (New England Biolabs, Ipswich, MA), 5% PEG-8000 (Sigma-Aldrich, St. Louis, MO), 10 mM $MgCl_2$, 10 mM DTT, and 1 mM NAD, 0.2 mM dNTPs each in 100 mM TRIS/Cl.

The p15aARA vector backbone was obtained by restriction cloning with EcoRI and NotI of pLEU* (N. Anderhuber, manuscript in preparation) providing the rrnB terminator, the p15a origin and a kanamycin resistance marker gene and of pQEara-gbd (Marzluf T., 2011, "Gamma Hydroxybutyrate Dehydrogenase (GHBdH) congeners", Bachelor Thesis) providing the P_{araBAD} operon from pKD46 [57].

To construct the expression plasmid for *St*TRPS, p15aARA-*St*TRPS (Additional file 14: Nucleotide Sequence 1) the His-β*St*TRPS-α*St*TRPS sequence was PCR amplified from pSTWS2H (Hoffman M., 2010, "Constitutive expression of the tryptophan synthase of *Salmonella typhimurium* in *E. coli*", Project Report) providing the *St*TRPS sequence from pEBA-10 [32] by PCR with the forward primer 5'-CGCAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGTACCGGAAAAAGGAGATCTGCATATG AG-3' and the reverse primer 5'-TGTTTTATCAGACCGCTTCTGCGTTCTGATTTAATCTTAATTAATT ATGCGCGGCTGGCG-3' to provide the overhangs for Gibson cloning [58] into the p15aARA vector backbone.

Construct pQE80L-His-ECFP carries the ECFP coding sequence with an N-terminal hexahistidine tag under the IPTG-inducible T5 *lacO* promoter (Additional file 15: Nucleotide sequence 2). ECFP was amplified from pET30a-ECFP (kindly provided by Steiner K., ACIB GmbH, Austria) with the forward primer 5'-CGCATCACCATCACCATCACGGATCCATGGTGAGCAAGGGCGAG-3' and the revers primer 5'-CAAGCTCAGCTAATTAAGCTTTTATCACTTGTACAGCTCGTCCATGC-3'. By Gibson isothermal assembly the amplicon was inserted into XhoI and XbaI cut pQE80L (Qiagen, Venlo, Netherlands) which contained the His-tag.

To construct pET21a(+)-His-ECFP (Additional file 16: Nucleotide sequence 3) the His-ECFP sequence was PCR amplified from pQE80L-His-ECFP with the forward primer 5'-GATAACAATTCCCCTCTAGA AATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGAGAGGATCGCATCACCATCAC-3' and the reverse primer 5'-GTGGTGGTGGTGGTGGTGCTCGAGTGCGGCCGCAAGCTTTCACTTGTACAGC TCGTCCATG-3'. The fragment was inserted into pET21a(+) (Novagen Merck Chemicals Ltd., Nottingham, UK) cut with XhoI and XbaI using Gibson isothermal assembly.

Plasmid pET30a-SBP-*Gt*HNL (*Gt*HNL codon-optimized for *E. coli* [44], kindly provided by Steiner K., ACIB GmbH, Austria) could not be used directly because it contained the kanamycin resistence marker, which is the same, as on p15aARA-*St*TRPS that was used for co-expression studies. The pET21a(+)-SBP-*Gt*HNL construct (Additional file 17: Nucleotide sequence 4) was obtained by digestion of pET30a-SBP-*Gt*HNL with XhoI and XbaI. The released SBP-*Gt*HNL fragment was then ligated into pET21a(+) (Novagen Merck Chemicals Ltd., Nottingham, UK) cut with the same enzymes. In this construct, the coding sequence of *Gt*HNL is preceded by 162 nt that encode a TEV protease cleavage site and the streptavidin-binding peptide tag (SBP-tag).

For the construction of pQE80L-SBP-*Gt*HNL (Additional file 18: Nucleotide sequence 5) the SBP-*Gt*HNL sequence was PCR amplified from pET30a-SBP-*Gt*HNL with the forward primer 5'-GTGAGC GGATAACAATTTCACACAGAATTCATTAAAGAGGAGAAATTAACTATGGACGAGAAGACCACCGG-3' and the reverse primer 5'-CAACAGGAGTCCAAGCTCAGCTAATTAAGCTTGGATCCTCATTAATTA GCGACGATACTGTTCATCGGTAAC-3'. The PCR fragment was introduced into the pQE80L (Qiagen, Venlo, Netherlands) vector backbone cut with HindIII and EcoRI.

All plasmids were sequence verified (Microsynth, Balgach, Switzerland).

To construct a Trp auxotrophic expression strain the *trpC* gene (encodes the indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase; EC 4.1.1.48) of BL21(DE3) Gold (Agilent Technologies, Santa Clara, CA) was deleted by λ Red recombineering [57, 59]. The λ Red genes were provided on pSIM6 [59]. The DNA fragment to integrate into the *E. coli* genome was obtained by PCR-amplification of the kanamycin resistance cassette encoded on the pKD13 template [57] with the forward primer 5'-CGACAGAGTTACCGCACTGGCGGCACGAGGGTAAATGATGATTCCGGGGATC CGTCGACC-3' and the reverse primer 5'-GTTGTCATTGTTCCTTTCCTTAATATGCGCGCAGCGTCT GTGTAGGCTGGAGCTGCTTCG-3' (sequences priming on pKD13 are underlined). The primers contained 5'-flanking sequences for recombination that encompassed bases -34 to +6 (where A of the start ATG is +1) upstream and bases +1342 to +1362 plus 19 bp downstream of the *trpC* gene. That is, all but the first six 5'-terminal and the 21 3'-terminal bases of the *trpC* sequence were deleted from the *E. coli* genome. Genomic integration of this PCR amplicon conferred kanamycin resistance that was flanked by flippase recognition target (FRT) sites. The FRT sites facilitated removal of the antibiotic resistance gene by FLP-FRT recombination. The recombinase FLP was provided on pCP20 [57]. The removal of the marker resulted in BL21(DE3) Gold Δ*trpC*::0. The deletion of *trpC* was sequence verified.

3.5.2 **Cell culture, [ncTrp] variant expression and cell disruption**

All chemicals were purchased from Sigma-Aldrich, St. Louis, MO unless indicated otherwise.

For ncTrp variant expression the Trp auxotrophic *E. coli* BL21(DE3) Gold Δ*trpC*::0 strain transformed with the appropriate expression plasmid(s) was cultured in 100 mL synthetic minimal medium (for details on media components see Additional file 19: Table S5). The culture medium contained excess amounts (0.5 mg/L) of all amino acids except of Trp.

We performed a titration experiment in order to assess the limiting concentration of Trp that facilitates a growth arrest of the Trp auxotrophic *E. coli* strain in minimal medium at D_{600} 2. 100 ml cultures were supplemented with 0.1 µM, 7 µM or 18 µM of Trp. Growth medium with an excess of 0.5 mM Trp served as the positive growth control. Cells were grown in baffled shake flasks at 37 °C with vigorous shaking. D_{600} was recorded over time in three technical replicates. Clearly, growth of the Trp auxotrophic *E. coli* strain was dependent on the amount of Trp present in the medium. The auxotrophic cells stopped growing when they ran out of Trp as indicated by a stagnating D_{600}. The limiting concentration of 18 µM Trp sustained cell growth to a D_{600} of 1.8 - 2. We did not observe growth arrest if the cultures were supplemented with excess of Trp. Growth arrest occurred upon the depletion of Trp and not of any other nutrient in the medium. When excess Trp was added after depletion, the cells readily resumed growth and grew to stationary phase.

For ncTrp variant expressions cells were cultured at 37 °C until ncTrp variant induction and then the temperature was lowered to 28 °C for overnight expression.

*St*TRPS expression was induced at D_{600} values of 0.5, 1 or 1.5 with 200 mg/L, 20 mg/L or 2 mg/L arabinose for induction. The final expression protocols A and B used 2 mg/ml arabinose at a D_{600} of 1 for *St*TRPS induction.

With protocol A the indole analog was added at the time of Trp depletion at a D_{600} of 1.8 - 2. With protocol B the indole analog was added at the time of *St*TRPS induction at a D_{600} of 1.

The relative mean deviation for D_{600} values was calculated with the formula provided in the Additional file 20.

ECFP and *Gt*HNL expression was induced using 0.1 mM IPTG (Biosynth, Staad, Switzerland) after the depletion of Trp and 30 min after the addition of 1 mM of the indole analog (all from Molekula GmbH, München, Germany). For *Gt*HNL expression 0.1 mM $MnCl_2$ was added at the time of induction. Harvested cells expressing ECFP variants were resuspended in 20 mM Tris/Cl, 150 mM NaCl, pH 7.8 and *Gt*HNL variants in 50 mM Bis-Tris, 30 mM NaCl, pH 6.8. 17 mg/mL lysozyme (Carl Roth, Karlsruhe, Germany), 0.003 mg/mL DNAse were added to the resuspension buffer and cells were lysed by sonication (output control 8, duty cycle 70 - 80%, 6 min) unless indicated otherwise (Figure 2C). Soluble and insoluble protein fractions were determined from these samples.

For whole cell extracts BugBuster™ Protein Extraction Reagent (Novagen Merck Chemicals Ltd., Nottingham, UK) was used according to the manufacturer's protocol with the following modifications: For these samples soluble and insoluble protein fractions were not separated. Instead SDS loading buffer was added, the samples were heated for 5 min at 90 °C and thoroughly mixed before applying them on the SDS gel.

3.5.3 **Purification of ncTrp protein variants, SDS gels and Western blots**

[$4NH_2$-Trp]ECFP variants were purified by an extraction protocol described in Samarkina *et al.* [60]. All other [ncTrp]ECFP variants were purified by their His-tag with Ni^{2+}-affinity chromatography. We performed Ni^{2+}-affinity chromatography in batch mode using nickel-nitrilotriacetic acid (Ni-NTA) resin

(Qiagen, Venlo, Netherland) according to the manufacturer's protocol. For the ECFP variants produced in the (-)*St*TRPS strain 500 µL of Ni-NTA resin (250 µL bed volume) were used for 30 mL of cleared lysate obtained from 200 to 500 D_{600} units of harvested cells. For the ECFP variants produced in the (+)*St*TRPS strain 500 µL of Ni-NTA resin (250 µL bed volume) were used for 15 mL of cleared lysates obtained from 100 to 250 D_{600} units of harvested cells. In the (+)*St*TRPS strain the His-tag on the β-subunit of *St*TRPS led to its co-purification with the His-tagged [ncTrp]ECFP variants. Moreover, the β-subunit forms a heterotetramer with the α-subunit of *St*TRPS. We diminished heterotetramer formation, i.e. additional impurities by the α-subunit of *St*TRPS in the purification process, by adding 1 M urea to cleared cell lysates of the (+)*St*Trp strain and 3.75 M urea to the wash buffer.

For all Ni^{2+}-affinity chromatography purifications we implemented a final centrifugation step to remove any remaining Ni-NTA beads before exchanging the elution buffer to the storage buffer (50 mM TRIS/HCl, pH 7.8) by dialysis (Snake Skin Dialysis Tubing, 3500 MWCO, #68035, Thermo Scientific, Waltham, MA). The [ncTrp]*Gt*HNL variants were purified by their SBP-tag with a StrepTactin® resin (IBA, Göttingen, Germany) according to the manufacturer's protocol. EDTA was omitted from buffer solutions, as *Gt*HNL contains Mn^{2+} in its active site. Elution buffer was exchanged to storage buffer (20 mM TRIS/Cl, 200 mM NaCl, pH 7.5) using PD-10 desalting columns (GE Healthcare, Chalfont Saint Giles, Great Britain) according to the manufacturer's protocol.

Protein concentrations of cleared cell lysates, insoluble protein (pellet) fractions and purified protein were determined by the Bradford method with the BioRad® protein assay (BioRad, Hercules, CA) as described in the manufacturer's protocol.

Cleared cell lysates of all variants as well as Ni^{2+} affinity chromatography purified [ncTrp]ECFP variants were applied on NuPAGE® Novex® 4-12% Bis-Tris protein gels (Life Technologies™, Carlsbad, CA). All other purified [ncTrp]ECFP as well as [ncTrp]*Gt*HNL variants were applied on 12% and 14% Laemmli SDS gels [61], respectively.

SDS gels were stained with Coomassie brillant blue or silver stain as described in the Additional file 21.

For the determination of protein titers (Additional file 8: Table S2) SDS gels (Figure 4 and Figure 5) were analyzed densitometrically by the software ImageJ [62]. A factor to correct for different protein amounts loaded on the gel was calculated from three endogenous *E. coli* protein bands (indicated for each gel, exemplified in the Additional file 22: Figure S7, boxes 1-3). We determined relative protein amounts for each [ncTrp]ECFP variant in the (-)*St*TRPS strain compared to the (+)*St*TRPS (Additional file 8: Table S2). In the same way we calculated a correction factor from SDS gels from three endogenous *E. coli* protein bands to apply equal protein amounts from whole cell extracts in Western blots (Figure 6).

For immunodetection of hexahistidine-tagged proteins all reagents were purchased from Life Technologies (Carlsbad, CA). We used a 6x-His epitope tag antibody from mouse, a goat anti-mouse IgG + IgM (H+L) secondary antibody horseradish peroxidase conjugate and the SuperSignal® West Dura Extended Duration Substrate for chemiluminescent detection.

3.5.4 **Fluorescence spectrometry**

Fluorescence emission of purified ECFP (0,6 mg/mL) and [4NH_2-Trp]ECFP (0.4 mg/mL) samples was detected as described in Bae *et al.* [2] with excitation wavelengths of 434 nm and 466 nm, respectively. Fluorescence spectra of ECFP and [4NH_2-Trp]ECFP were recorded on a Synergy MX spectrometer (BioTek, Winooski, VT) at 20 °C in 50 mM Tris/Cl, pH 7.5 buffer.

3.5.5 **Mass spectrometry**

The protein solutions were desalted using Amicon Ultra 0.5 mL centrifugal filter units (Millipore, Billerica, MA). A final protein concentration of 30 pmol/µL was obtained with water containing 5% acetonitrile and 0.1% trifluoracetic acid. The separation of possible protein variants was carried out on a capillary HPLC system (1200 Agilent, Santa Clara, CA) using a PepSwift RP monolithic column (50 x 0.5 mm, Thermo, Germering, Germany) at a flow rate of 20 µL/min and a column temperature of 60 °C. The gradient of solution A (water + 0.05% TFA) and B (ACN + 0.05% TFA) was performed as follows: 10% B for 5 min, 10%-100% B for 50 min, 100%-10% B for 1 min, 10% B for 15 min. Injection volume was 5 µL. The Thermo LTQ-FT mass spectrometer (Thermo Fisher Scientific, Waltham, MA) was operated with an ESI source in positive mode with following settings: mass range: 300-2000 m/z, resolution 400000, 500 ms injection time, 1 microscan, source voltage 5 kV, capillary voltage 35 V, sheath gas flow 15. The protein mass spectra were deconvoluted by the Thermo Fisher Scientific software Protein Deconvolution 2.0, using the Xtract algorithm. The following main parameters were applied: charge carrier, H^+; m/z range, minimal 800 to maximal 2000; minimal detected charge state, 4; s/n threshold, 5; relative abundance threshold, 20%.

Occasionally, parent protein contaminates preparations of variant proteins that were produced by SPI [2]. In this study we did not detect contamination with parent protein, however, trace amounts of unlabeled species might be present but fall below the detection limit of the mass spectrometry method (2-5%).

Abundances of non-canonical protein species, *i.e.* incorporation efficiency, is the relative sum intensity of each species in the sample.

3.6 List of abbreviations

4aza-indole: 4-azaindole; 4aza-indole: 4-azaindole; 7aza-Trp: 7-aza-L-tryptophan; D_{600}: optical density; ECFP: enhanced cyan fluorescent protein; 5F-indole: 5-fluoroindole; 6F-indole: 6-fluoroindole; 7F-indole: 7-fluoroindole; 5F-Trp: 5-fluoro-L-tryptophan; 6F-Trp: 6-fluoro-L-tryptophan; 7F-Trp: 7-fluoro-L-tryptophan; *Ec*RNAP, *E. coli* RNA polymerase; *Gt*HNL: hydroxynitril lyase from *Granulicella tundricola*; ncAA: non-canonical amino acid; ncTrp: non-canonical L-tryptophan; $4NH_2$-indole: 4-aminoindole; $4NH_2$-Trp: 4-amino-L-tryptophan; 4OH-indole: 4-hydroxyindole; 4OH-indole: 4-hydroxyindole; 5OH-Trp: 5-hydroxy-L-tryptophan; 5OH-Trp: 5-hydroxy-L-tryptophan; SBP-tag: streptavidin-binding peptide tag; SPI: supplementation based incorporation; *St*TRPS: tryptophan synthase from *Salmonella typhimurium*; (+)*St*TRPS strain: tryptophan auxotrophic *E. coli* host strain overexpressing the tryptophan synthase from *Salmonella typhimurium*; (-)*St*TRPS strain: tryptophan auxotrophic *E. coli* host strain only expressing its cognate tryptophan synthase; T7 RNAP: T7 RNA polymerase; $tRNA_{CCA}^{Trp}$: tryptophan transfer RNA; Trp: L-tryptophan; TrpRS: tryptophanyl-tRNA synthetase; TRPS: tryptophan synthase

3.7 **Figures and illustrations**

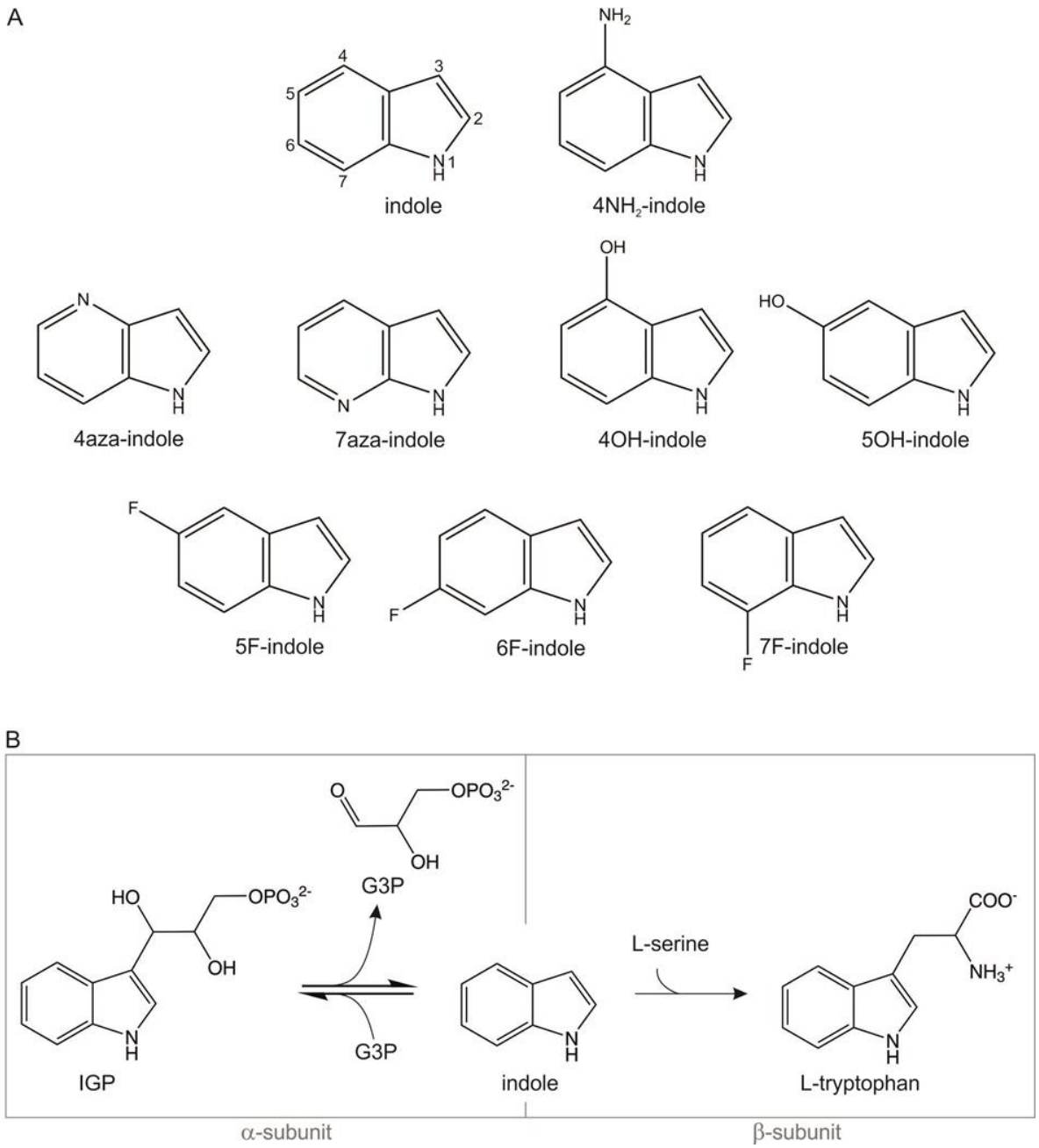


Figure 1:

A) Indole and indole analogs used in this study

$4NH_2$-indole, 4-aminoindole; 4aza-indole, 4-azaindole; 7aza-indole, 7-azaindole; 4OH-indole, 4-hydroxyindole; 5OH-indole, 5-hydroxyindole; 5F-indole, 5-fluoroindole; 6F-indole, 6-fluoroindole; 7F-indole, 7-fluoroindole.

B) Conversion of indole to L-tryptophan by the action of TRPS [21]

IGP, indole-3-glycerol-phosphate; G3P, glyceraldehyde-3-phosphate

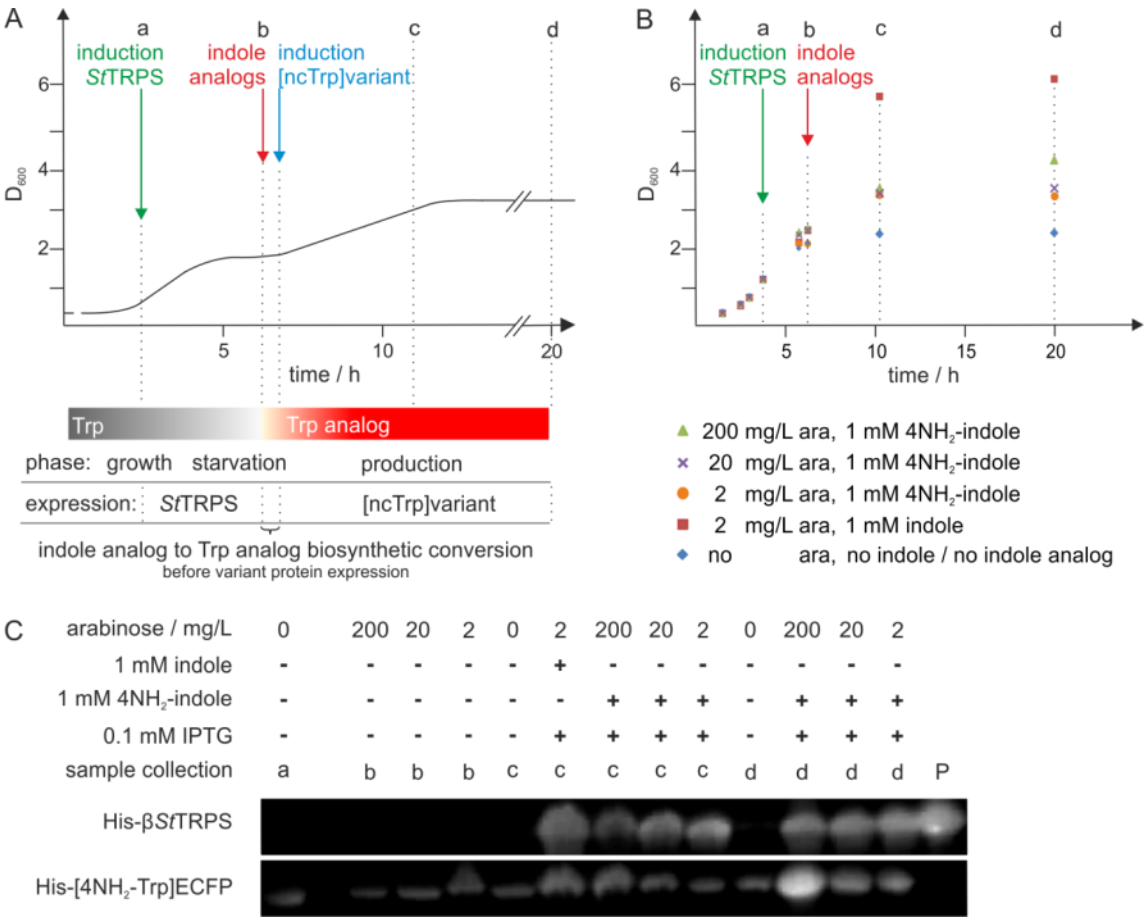


Figure 2:

A) Schematic illustration of the process for ncTrp synthesis and incorporation into a target protein

The growth pattern and variant protein production in relation to the tryptophan consumption and enzymatic non-canonical tryptophan (ncTrp) analog synthesis is shown. The induction of the tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) with arabinose, the addition of indole analogs, and the induction of the target protein with IPTG are indicated by arrows. Samples for immunoblotting were collected at the instances marked with letters (a, before *St*TRPS induction; b, before addition of indole analog; c, after 3 h of variant protein expression; d, after overnight expression).

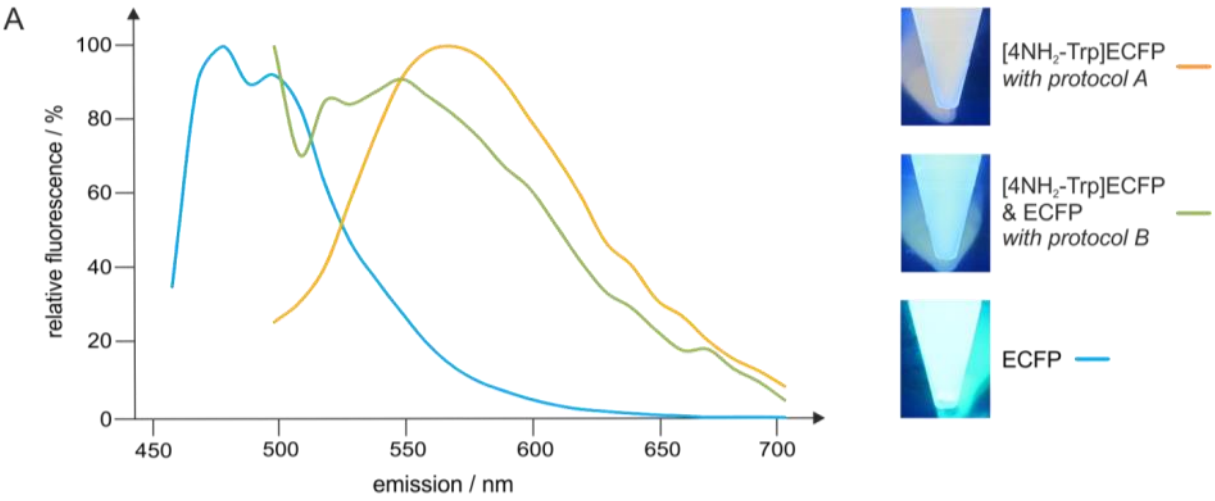
B) Growth curves of indole- and $4NH_2$-indole fed *E. coli* cultures

Different arabinose concentrations, 200 mg/L, 20 mg/L and 2 mg/L were applied for the induction of the tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) at a D_{600} of 1. Samples were collected before induction of *St*TRPS with arabinose (a), before addition of indole / $4NH_2$-indole and induction with IPTG (b), after 3 h of expression (c), and after expression over night (d) according to the scheme in Figure 2A. D_{600} values at the times found in Additional file 3: Figure S2 were determined in triplicates and the relative mean deviation was calculated (Additional file 23: Table S6, other D_{600} values were single point measurements).

C) Western blot of the His-tagged β-subunit of *St*TRPS and the His-tagged ECFP variant

Cleared cell lysates are shown. Samples were collected before induction of *St*TRPS with arabinose (a), before addition of indole/$4NH_2$-indole and induction with IPTG (b), after 3 h of expression (c), and after expression overnight (d) according to the scheme in Figure 2A. P refers to purified His-tagged ECFP as positive control. 0.075 D_{600} units were loaded. Samples were prepared with CelLytic™ B 2x according to the manufacturer's protocol. For the Western blot of the corresponding samples of the insoluble protein (pellet) fraction refer to Additional file 24: Figure S8.

A



B

variant species *i.e.* number of $4NH_2$-Trp	abundance of species / % *protocol A*	 *protocol B*
0	n.d.	12.3
1	n.d.	58.9
2	100	28.7

Figure 3:

A) Fluorescence emission spectra and fluorescing samples of isolated ECFP and [$4NH_2$-Trp]ECFP

The [$4NH_2$-Trp]ECFP variant was produced according to the expression protocols A or B. The excitation wavelength was 434 nm for ECFP and 466 nm for [$4NH_2$-Trp]ECFP.

B) Protocol A facilitated full substitution of Trp by [$4NH_2$-Trp] in ECFP

Data was obtained by mass spectrometric analysis and variants were produced using the expression protocols A or B; n.d., not detected. For found masses refer to Additional file 6: Table S1. Abundance of variant species was calculated as described in the Methods section. Protocol A, addition of $4NH_2$-indole after Trp starvation; protocol B, addition of $4NH_2$-indole at the induction of *St*TRPS (Additional file 5: Figure S3).

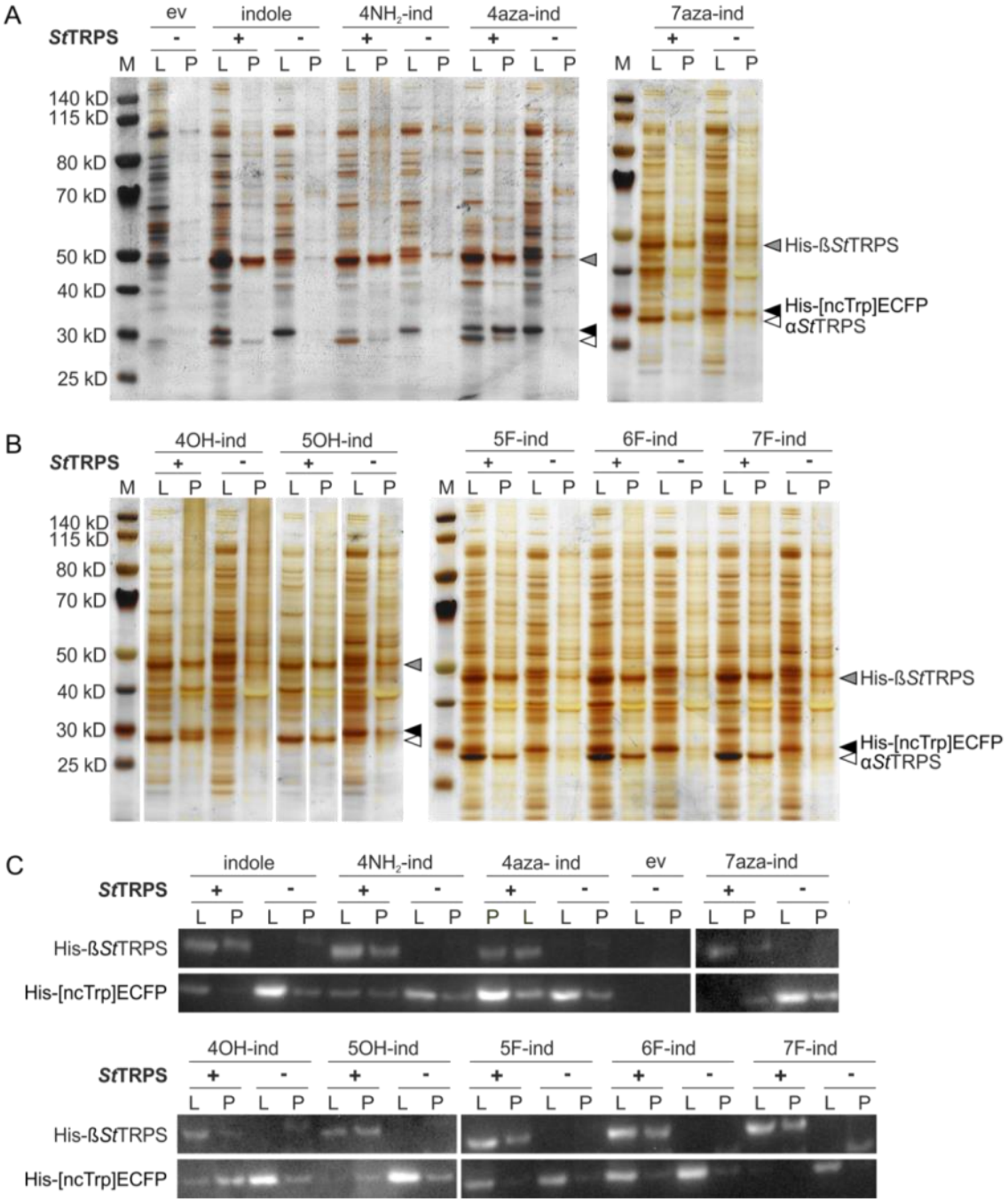


Figure 4:

A), B) Expression and solubility of different ncTrp variants of ECFP

The soluble (L, cleared lysate) and insoluble (P, pellet) protein fractions are shown. *St*TRPS was co-expressed (+) or not (-). His-β*St*TRPS (calculated molecular weight 45 kDa) and α*St*TRPS (calculated molecular weight 29 kDa) designate the beta (grey arrow) and alpha (white arrow) subunits of *St*TRPS; ECFP (black arrow), calculated molecular weight 28 kDa. M, molecular size marker; x-ind, indole analog; ev, empty vector control. 3 µg of of total protein were loaded per lane. Silver stained 4-12% Bis-Tris SDS gels are shown.

C) Immunoblot of the [ncTrp]ECFP variants

Immunodetection of the hexahistidine-tagged β-subunit of the *S. typhimurium* tryptophan synthase (β*St*TRPS) and the [ncTrp]ECFP variants with an anti-His antibody. The soluble (L, cleared lysate) and insoluble (P, pellet) protein fractions are shown; x-ind, indole analog. 4 µg of of total protein were loaded per lane.

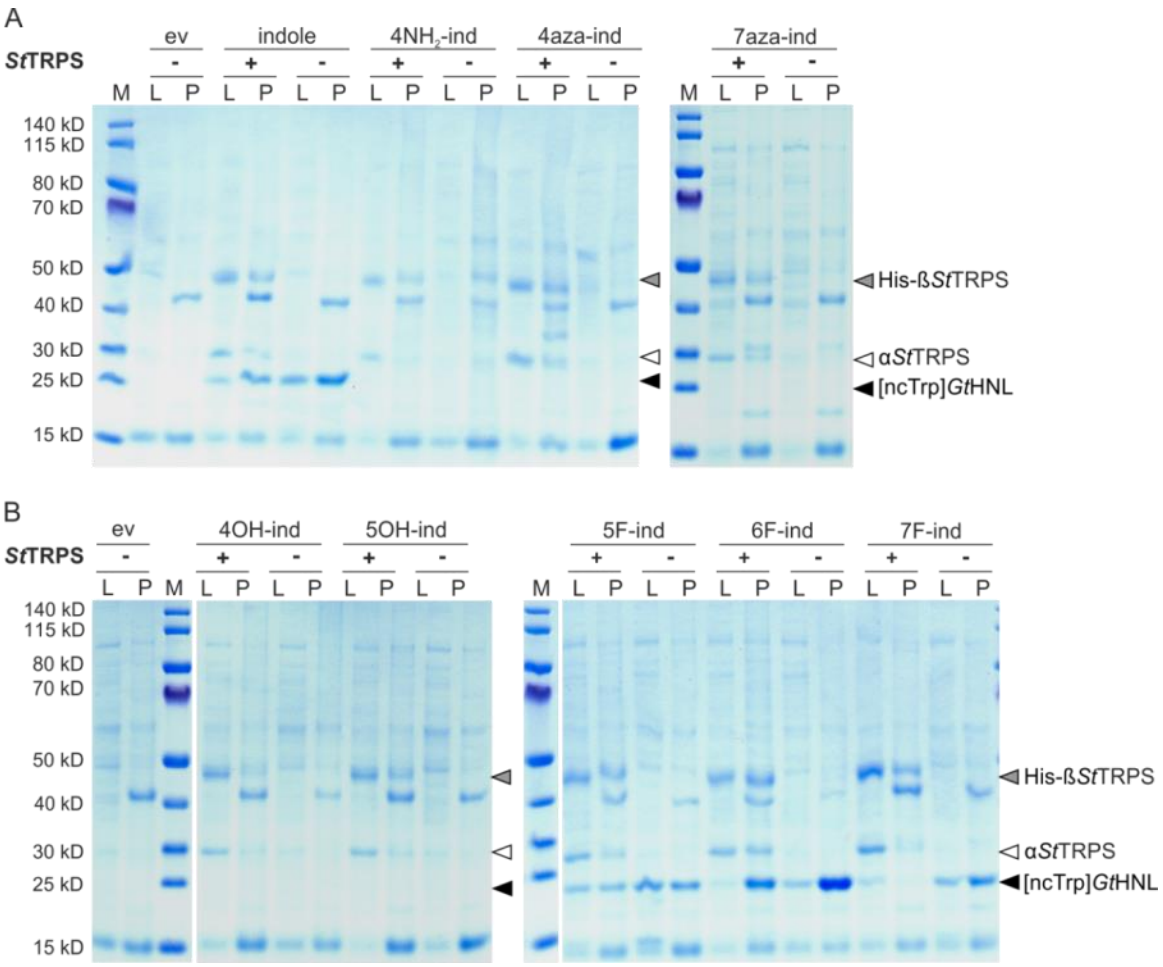


Figure 5: Expression and solubility of different ncTrp variants of *Gt*HNL

The soluble (L, cleared lysate) and insoluble (P, pellet) protein fractions are shown in A) and B). *St*TRPS was co-expressed (+) or not (-). His-β*St*TRPS (grey arrow, calculated molecular weight 45 kDa) and α*St*TRPS (white arrow, calculated molecular weight 29 kDa) designate the two subunits of *St*TRPS; *Gt*HNL (black arrow), calculated molecular weight 20 kDa. M, molecular size marker; x-ind, indole analog; ev, empty vector control. 4 µg of of total protein were loaded per lane. Coomassie stained 4-12% Bis-Tris SDS gels are shown.

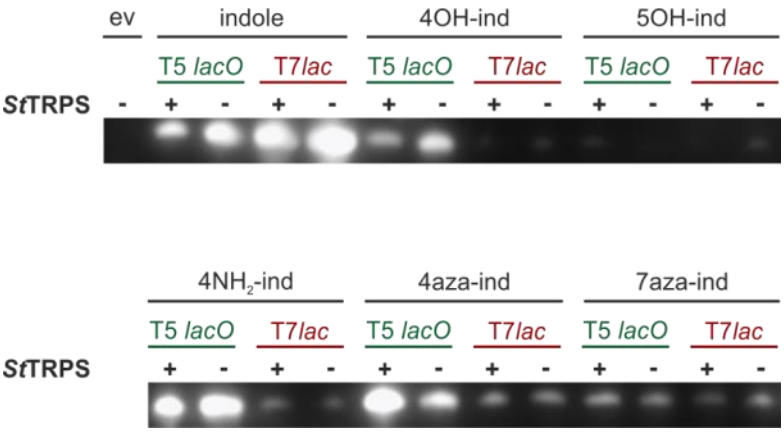


Figure 6: Comparative expression of different [ncTrp]ECFP variants under the control of the T7*lac* and T5 *lacO* promoters

The hexahistidine-tagged [ncTrp]ECFP variants were immunodected in whole cell extracts using an anti-His antibody. Expression occurred in the presence (+) or absence (-) of *St*TRPS. T7*lac*, expression using the pET21(+)-His-ECFP construct; T5 *lacO*, expression using the pQE80L-His-ECFP construct; ev, the empty vector control; x-ind, indole analog as listed in Figure 1A. Equal amounts of total protein were loaded in each lane (see Methods section for details). For the corresponding SDS gel with the normalized amounts of total protein refer to the Additional file 25: Figure S9.

3.8 **Tables**

Table 1: Cell density at tryptophan depletion as a function of *St*TRPS induction time and arabinose concentration

Tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) was induced at D_{600} 0.5, 1 or 1.5 with arabinose (ara) concentrations of 200 mg/L, 20 mg/L and 2 mg/L. D_{600} values > 2.0 are highlighted in dark grey, values 1.9 ≤ D_{600} ≤ 2.0 are highlighted in grey and values ≤ 1.8 are in light grey. D_{600} was determined in triplicates and the mean values are shown. For the calculated relative mean deviation refer to the Additional file 26: Table S7. For the representation of the growth pattern refer to Additional file 3: Figure S2.

ara / mg/L	200		20		2	
time / h	6.75	8	6.75	8	6.75	8
D_{600} at *St*TRPS induction	D_{600} before restoring cell growth upon Trp addition					
0.5	2.8	3.4	1.9	2.1	2.0	2.1
1	1.9	2.0	1.8	1.8	1.8	1.8
1.5	1.8	1.9	1.7	1.8	1.7	1.8

Table 2: Abundance of the different [ncTrp]ECFP variant species

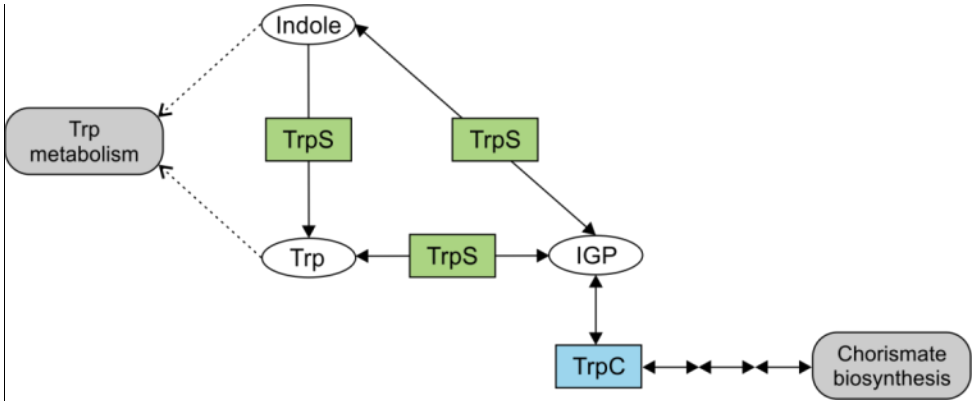
(+)*St*TRPS co-expresses the *Salmonella typhimurium* tryptophan synthase while (-)*St*TRPS does not. The found masses are listed in Additional file 10: Table S3. Incorporation efficiencies were calculated as described in the Methods section.

variant protein	number of incorporated ncTrp	abundance of species / %	
		(+)*St*TRPS strain	(-)*St*TRPS strain
[4NH_2-Trp]ECFP	0	n.d.	n.d.
	1	n.d.	79
	2	100	21
[4aza-Trp]ECFP	0	n.d.	39
	1	36	23
	2	64	38
[7aza-Trp]ECFP	0	n.d.	n.d.
	1	n.d.	100
	2	100	n.d.
[4OH-Trp]ECFP	0	n.d.	13
	1	38	71
	2	62	16
[5OH-Trp]ECFP	0	n.d.	82
	1	n.d.	18
	2	100	n.d.
[5F-Trp]ECFP	0	n.d.	n.d.
	1	n.d.	68
	2	100	32
[6F-Trp]ECFP	0	n.d.	n.d.
	1	n.d.	34
	2	100	66
[7F-Trp]ECFP	0	n.d.	n.d.
	1	100	44
	2	n.d.	55

3.9 Additional files / Supporting information

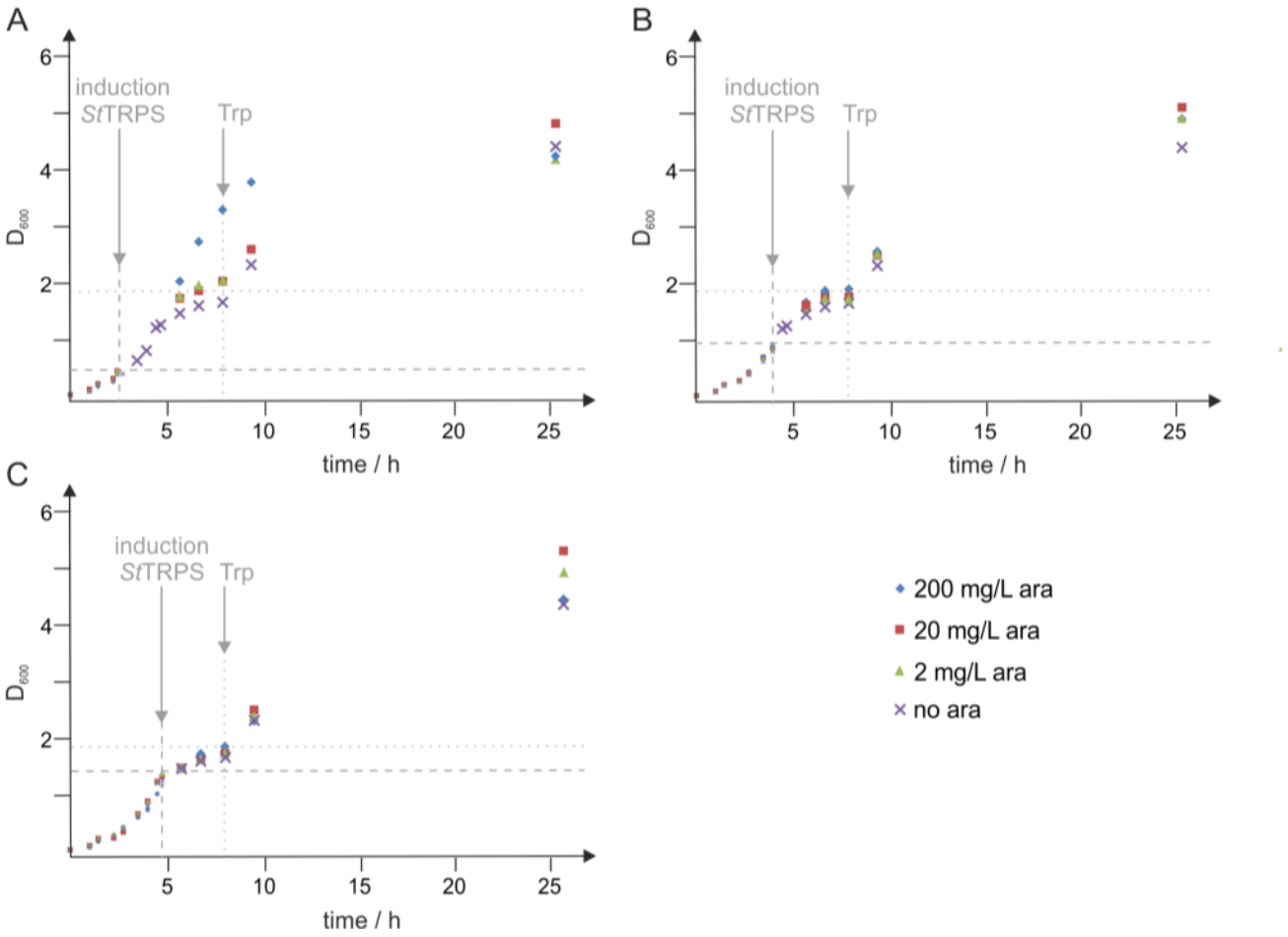
Additional file 1: Availability of Trp analogs

We searched the online databases ChemSpider (www.chemspider.com), ChemBook (www.chemicalbook.com) and PubChem (https://pubchem.ncbi.nlm.nih.gov/) for the commercial availability of $4NH_2$-Trp, 4aza-Trp, 4OH-Trp and 7F-Trp. Prices of 5F-Trp and 5F-indole were compared at Sigma-Aldrich (St. Louis, MO).



Additional file 2: Figure S1: Schematic presentation of the involvement of the tryptophan synthase (TRPS) and of the indole-3-glycerol-phosphate (IGP) synthase (TrpC) in the tryptophan metabolism of *E. coli*

Pathway information was obtained from the kyoto encyclopedia of genes and genomes (KEGG) database [63-65].



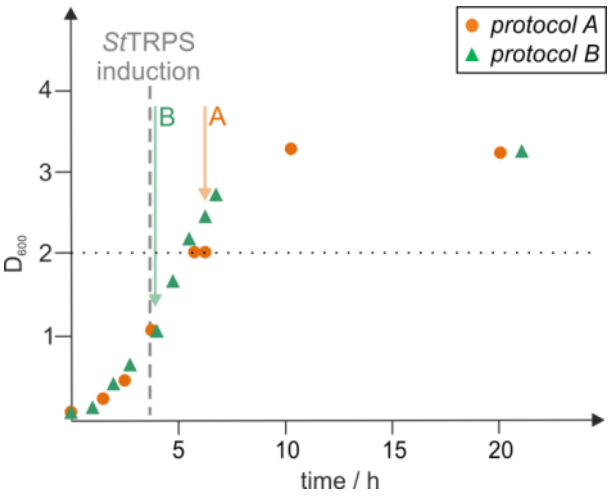
Additional file 3: Figure S2: **Influence of the co-expression of the tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) on growth behavior of the Trp auxotrophic strain**

*St*TRPS induction (vertical dashed line with grey arrow) with different arabinose concentrations (200 mg/L, 20 mg/L, 2 mg/L) at a D_{600} of 0.5 (A) 1 (B) and 1.5 (C). A grey dashed horizontal line indicates the cell density at the induction of *St*TRPS. 25 µM Trp (vertical dotted line with grey arrow) was added to confirm that the growth arrest was caused by the depletion of Trp. The grey dotted horizontal line indicates D_{600} 1.8 at growth arrest in the same strain without the expression of *St*TRPS. The D_{600} values at the times found in Table 1 were determined in triplicates and the relative mean deviation was calculated (Additional file 26: Table S7). Other D_{600} values were single measurements.

Additional file 4: Amino acid sequence 1: Enhanced cyan fluorescent protein (ECFP)

Protein sequence in one letter amino acid code of the N-terminally hexahistidine-tagged enhanced cyan fluorescent protein (251 amino acids, 2 Trp). The hexahistidine-tag is shown in grey. Tryptophan residues are highlighted in grey with the corresponding amino acid residue numbers in subscript.

MRGSHHHHHHGSM_1VSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPW_{57}PTL
VTTLTW_{66}GVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDG
NILGHKLEYNYISHNVYITADKQKNGIKANFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSK
DPNEKRDHMVLLEFVTAAGITLGMDELYK



Additional file 5: Figure S3: Growth pattern of the Trp auxotrophic expression strain upon the addition of 4NH_2-indole at different cell densities

The time of induction of the tryptophan synthase from *S. typhimurium* (*St*TRPS) with 2 mg/L arabinose at D_{600} 1 is indicated with a dashed vertical line in grey. Trp depletion is indicated with a dotted horizontal line. Addition of 4NH_2-indole is indicated by arrows. In protocol A (orange) 4NH_2-indole was added after Trp depletion, whereas in protocol B (green) 4NH_2-indole was added when we induced *St*TRPS. Unexpectedly, the expression strain did not show a growth arrest due to Trp depletion when protocol B was applied. The D_{600} values at the times listed in Additional file 27: Table S8 were determined in triplicates and the relative mean deviation was calculated. Other D_{600} values were single measurements.

Additional file 6: Table S1: Calculated and found masses for the [4NH_2-Trp]ECFP variant described in chapter 2.1

Masses were calculated with the MassXpert software [66] and 4NH_2-Trp was manually integrated. Monoisotopic intact protein masses were determined by ESI-MS (see Methods section). The number of incorporated 4NH_2-Trp is suggested; n.d., not detected.

variant	chromophore formation	$mass_{calculated}$ / Da	number of incorporated 4NH_2-Trp	$mass_{found}$ protocol A / Da	$mass_{found}$ protocol B / Da
[4NH_2-Trp]ECFP	no	28,286.12	0	n.d.	n.d.
		28,301.13	1	n.d.	n.d.
		28,316.14	2	n.d.	n.d.
	yes	28,266.09	0	n.d.	28,265.20
		28,281.10	1	n.d.	28,280.29 [a]
		28,296.11	2	28,296.26	28,296.28

[a] The mass of 28,280.29 Da found for the expression with protocol B could refer to some unidentified protein, because a mass peak at 28,281.19 Da was also detected for the unlabeled ECFP sample (data not shown). No peak in this mass range was detected for the [4NH_2-Trp]ECFP expression using protocol A.

Additional file 7: Amino acid sequence 2: Hydroxynitril lyase from *Granulicella tundricola* (*Gt*HNL)

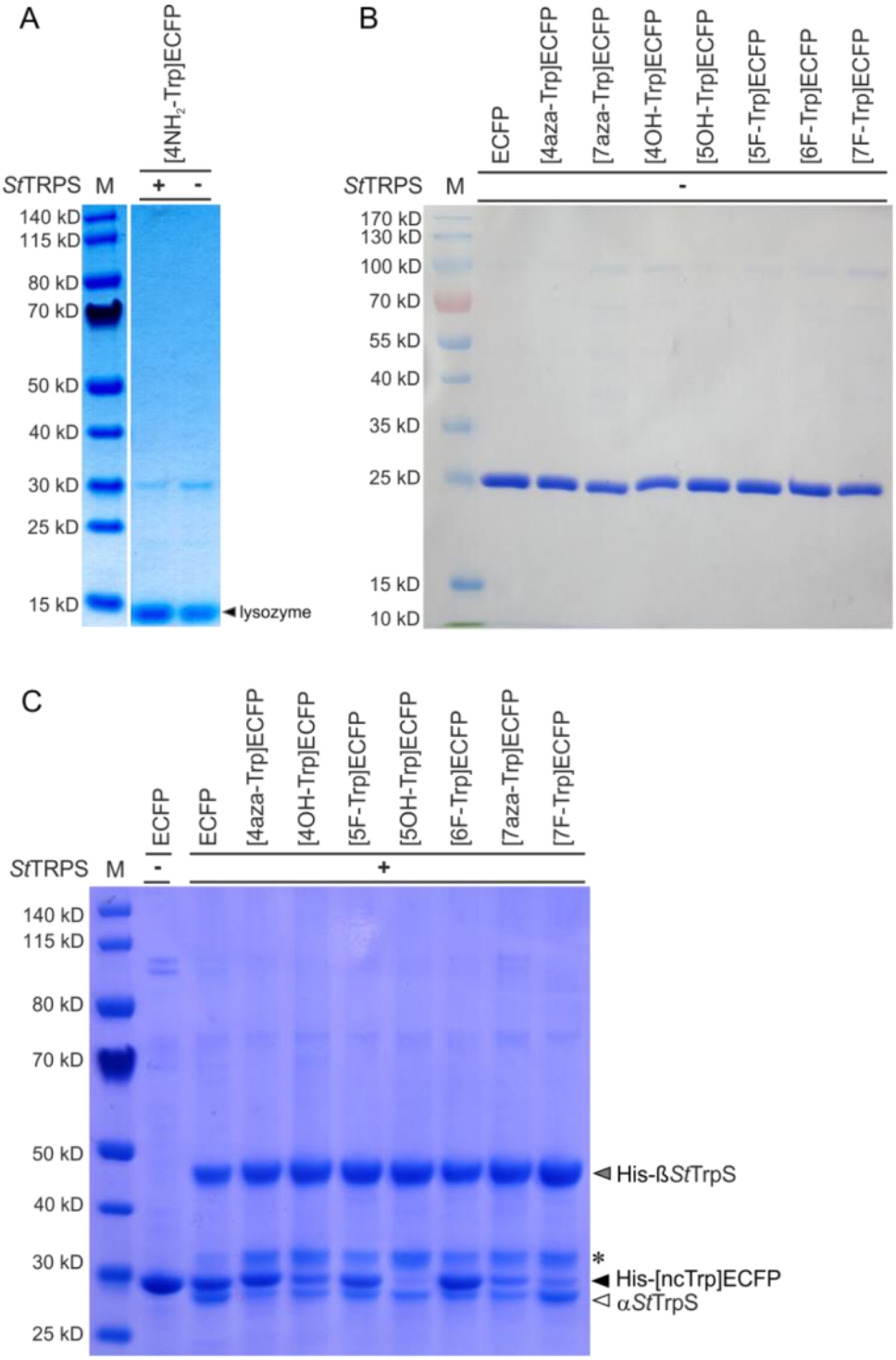
Protein sequence in one letter amino acid code of N-terminally streptavidin-binding-protein (SBP) -tagged hydroxynitril lyase from *Granulicella tundricola* (186 amino acids; 7 Trp). The SBP-tag and the TEV protease cleavage site are shown in grey. Tryptophan residues are highlighted in grey.

MDEKTTGWRGGHVVEGLAGELEQLRARLEHHPQGQREPSGGCKLGLGTENLYFQSMEIKRVGSQASGKGPADWFT
GTVRIDPLFQAPDPALVAGASVTFEPGARTAWHTHPLGQTLIVTAGCGWAQREGGAVEEIHPGDVVWFSPGEKHW
HGAAPTTAMTHLAIQERLDGKAVDWMEHVTDEQYRR

Additional file 8: Table S2: Relative expression levels of the [ncTrp]ECFP and [ncTrp]*Gt*HNL variants in the absence and presence of *St*TRPS.

The expression levels of the variant proteins in the soluble protein fraction in the absence ((-)*St*TRPS) and in the presence ((+)*St*TRPS) of the *Salmonella typhimurium* tryptophan synthase were compared. The corresponding band intensities were determined densitometrically from the SDS gels shown in Figure 4A and B and Figure 5 using the program ImageJ [62]. Relative protein amounts were calculated as detailed in the Methods section; n.d., not detected because protein was not expressed; n.c., not calculated because [6F-Trp]*Gt*HNL was not expressed in the (+)*St*TRPS strain.

	relative band intensity (-)*St*TRPS / (+)*St*TRPS	
$ncTrp_{incorporated}$	**ECFP variants**	***Gt*HNL variants**
Trp (canonical)	0.9	4.5
[4NH_2-Trp]	1.7	n.d.
[4aza-Trp]	1.5	n.d.
[7aza-Trp]	2.3	n.d.
[4OH-Trp]	1.5	n.d.
[5OH-Trp]	4.3	n.d.
[5F-Trp]	1.0	4,6
[6F-Trp]	0.9	n.c.
[7F-Trp]	1.4	12.1



Additional file 9: Figure S4: SDS-PAGE analysis of the purified [ncTrp]ECFP variants

A) Purification by extraction [60]. B) and C) Purification by Ni^{2+}-affinity chromatography of His-[ncTrp]ECFP (in B and C, black arrow) and His-β*St*TRPS (in C, grey arrow). α*St*TRPS (in C, white arrow) was co-purified because of binding to the His-β*St*TRPS subunit. For details on purification see the Methods section. Expression of [ncTrp]ECFP variants (calculated molecular weight 28 kDa) in the presence (+) or absence (-) of *St*TRPS. His-β*St*TRPS (calculated molecular weight 45 kDa) and α*St*TRPS (calculated molecular weight 29 kDa) designate the subunits of *St*TRPS. M, molecular weight marker; x-Trp; incorporated ncTrp; the nomenclature is the same as for the corresponding indoles shown in Figure 1A. The grey asterisk refers to an unknown protein. 0.8 µg (A), 3 µg (B) and 4 µg (C) of protein were loaded on 4-12% Bis-Tris protein gels (A, C) and a 12% SDS Laemmli gel (B). The gels were stained with Coomassie brillant blue.

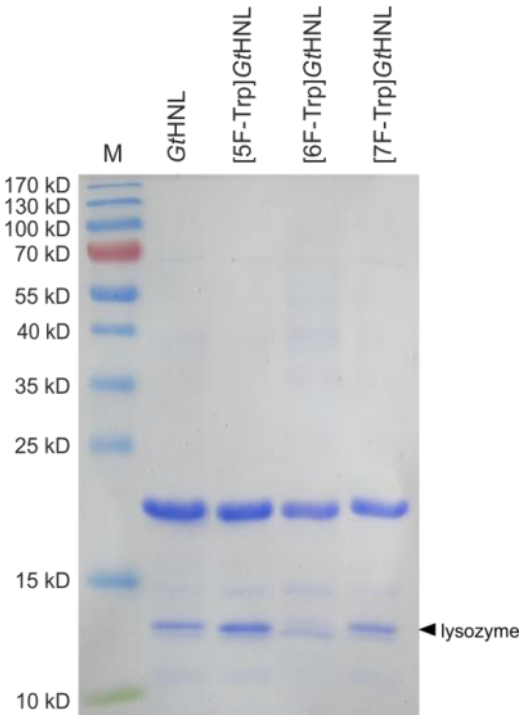
Additional file 10: Table S3: Calculated and found masses of the [ncTrp]ECFP variants described in chapter 2.2

Masses were calculated with the MassXpert software [66]. The ncTrp used in this study were manually integrated. Intact protein masses were determined by ESI-MS (see Methods section). The number of incorporated ncTrp is suggested; n.d., not detected.

[ncTrp] ECFP variant	chromophore formation	$mass_{calculated}$ / Da	number of incorporated ncTrp	$mass_{found}$ (+)*St*TRPS / Da	$mass_{found}$ (-)*St*TRPS / Da
[4NH_2-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,301.13	1	n.d.	28,300.39
		28,316.14	2	28,315.38	n.d.
	yes	28,266.09	none	n.d.	n.d.
		28,281.10	1	n.d.	n.d.
		28,296.11	2	28,296.50	28,296.40
[4aza-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,287.11	1	n.d.	n.d.
		28,289.11	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	28,265.65
		28,267.08	1	28,267.46	28,266.34 [a]
		28,269.08	2	28,268.45	28,269.33
[7aza-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,287.11	1	n.d.	28,287.49
		28,289.11	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	n.d.
		28,267.08	1	n.d.	28,266.47 [a]
		28,269.08	2	28,268.50	n.d.
[4OH-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,302.11	1	n.d.	n.d.
		28,318.11	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	28,266.45
		28,282.08	1	28,281.36	28,282.33
		28,298.077	2	28,298.28	28,297.32
[5OH-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,302.11	1	n.d.	n.d.
		28,318.11	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	28,266.47
		28,282.08	1	n.d.	28,282.44
		28,298.08	2	28,297.48	n.d.

[5F-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,304.11	1	n.d.	28304.36
		28,322.100	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	n.d.
		28,284.08	1	n.d.	28,284.40
		28,302.07	2	28302.46	28,301.65
[6F-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,304.11	1	n.d.	n.d.
		28,322.10	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	n.d.
		28,284.08	1	n.d.	28,284.45
		28,302.07	2	28302.40	28,301.48
[7F-Trp]ECFP	no	28,286.12	none	n.d.	n.d.
		28,304.11	1	n.d.	28,304.57
		28,322.10	2	n.d.	n.d.
	yes	28,266.09	none	n.d.	n.d.
		28,284.08	1	28283.39	28,283.44
		28,302.07	2	n.d.	28,302.32

[a] Mass peaks 28,266.34 Da in the [4aza-Trp]ECFP sample and 28,266.47 Da in the [7aza-Trp]ECFP sample might refer to unlabeled parent ECFP as well. Because a different mass of 28,265.65 Da referring to unlabeled parent ECFP was found in the [4aza-Trp]ECFP sample, we assumed that the mass of 28,266.34 Da in the [7aza-Trp]ECFP sample has to be assigned to the single labelled variant.



Additional file 11: Figure S5: SDS-PAGE analysis of the purified [ncTrp]*Gt*HNL variants

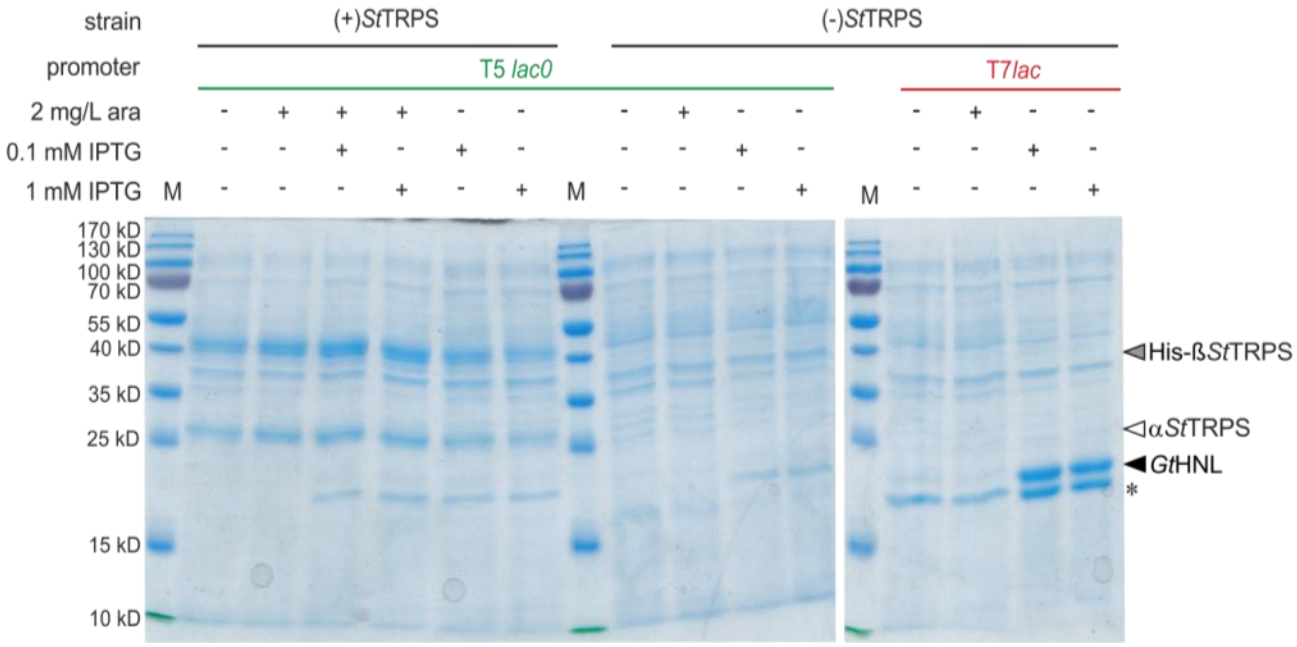
The calculated molecular weight of the *Gt*HNL variants is 20 kDa. 4 µg of protein were loaded per lane. M, molecular weight marker; 14% Coomassie stained SDS gel.

Additional file 12: Table S4: Calculated and found masses of the [ncTrp]*Gt*HNL variants described in chapter 2.2

Masses were calculated with the MassXpert software [66]. The ncTrp used in this study were manually integrated. Intact protein masses were determined by ESI-MS (see Methods section). The number of incorporated ncTrp is suggested. SS, disulfid bond; n.d., not detected.

			[ncTrp]*Gt*HNL variant		
disulfide bond formation	**$mass_{calculated}$**	**number of incorporated ncTrp**	**[5F-Trp]**	**[6F-Trp]**	**[7F-Trp]**
monomer with 1 SS	20,278.97	0	n.d.	n.d.	n.d.
	20,296.96	1	n.d.	n.d.	n.d.
	20,314.95	2	n.d.	n.d.	n.d.
	20,332.94	3	n.d.	n.d.	n.d.
	20,350.93	4	n.d.	n.d.	n.d.
	20,368.92	5	n.d.	n.d.	n.d.
	20,386.91	6	n.d.	n.d.	n.d.
	20,404.91	7	20,404.72	20,404.75	20,404.75
dimer with 1 SS	40,559.96	0	n.d.	n.d.	n.d.
	40,595.94	1	n.d.	n.d.	n.d.
	40,631.92	2	n.d.	n.d.	n.d.
	40,667.90	3	n.d.	n.d.	n.d.
	40,703.88	4	n.d.	n.d.	n.d.
	40,739.86	5	n.d.	n.d.	n.d.
	40,775.84	6	n.d.	n.d.	n.d.
	40,811.84	7	40,811.53	n.d.	40,810.56 [a]
dimer with 2 SS	40,557.95	0	n.d.	n.d.	n.d.
	40,593.93	1	n.d.	n.d.	n.d.
	40,629.91	2	n.d.	n.d.	n.d.
	40,665.89	3	n.d.	n.d.	n.d.
	40,701.87	4	n.d.	n.d.	n.d.
	40,737.85	5	n.d.	n.d.	n.d.
	40,773.83	6	n.d.	n.d.	n.d.
	40809.83	7	n.d.	n.d.	40810.56 [a]

[a] The found mass of 40,810.56 Da in the [7F-Trp]*Gt*HNL might refer to the fully labelled dimer species forming either one or two disulfide bonds.



Additional file 13: Figure S6: Expression of *Gt*HNL under control of the T5 and T7 promoters

Whole cell extracts of strains expressing *Gt*HNL in the presence of Trp (calculated molecular weight 20 kDa) were analyzed by SDS-PAGE. T5, expression under control of the T7*lacO* promoter from the pQE80L-SBP-*Gt*HNL expression construct; T7, expression under control of the T7*lac* promoter of the pET21a(+)-SBP-*Gt*HNL expression construct; (+)*St*TRPS, *E. coli* strain co-expressing the tryptophan synthase from *S. typhimurium*; (-)*St*TRPS *E. coli* strain without co-expression of *St*TRPS; M, molecular size marker. Arabinose (ara) was used to induce *St*TRPS expression and IPTG induced the expression of *Gt*HNL. The grey asterisk refers to an unknown protein. Cells equaling 0.05 D_{600} units were loaded per lane. Coumassie stained 14% SDS gels are shown.

Additional file 14: Nucleotide sequence 1: p15aARA-*St*TRPS

rrnB (T1&T2)	50-375
p15A ori	403-1314
kanR	1723-2655
rrnB T2	2894-2921
araC	2964-4021
araBAD promoter	4118-4145
His-tag	4203-4220
β*St*TRPS	4233-5426
α*St*TRPS	5426-6232

TTAATTAAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGG
TAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGG
ACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATCCTGACGGATGGCCTTTT
TGGATAAGCTGTCAAACATGAGAATTAACAACTTATATCGTATGGGGCTGACTTCAGGTGCTACATTTGAAGAGATAAATTGCACTGAAATCTAGAAATATTTTATCTGATTAATAAGATGATCT
TCTTGAGATCGTTTTGGTCTGCGCGTAATCTCTTGCTCTGAAAACGAAAAAACCGCCTTGCAGGGCGGTTTTTCGAAGGTTCTCTGAGCTACCAACTCTTTGAACCGAGGTAACTGGCTTGGAGG
AGCGCAGTCACCAAAACTTGTCCTTTCAGTTTAGCCTTAACCGGCGCATGACTTCAAGACTAACTCCTCTAAATCAATTACCAGTGGCTGCTGCCAGTGGTGCTTTTGCATGTCTTTCCGGGTTG
GACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGACTGAACGGGGGGTTCGTGCATACAGTCCAGCTTGGAGCGAACTGCCTACCCGGAACTGAGTGTCAGGCGTGGAATGAGACAAAC
GCGGCCATAACAGCGGAATGACACCGGTAAACCGAAAGGCAGGAACAGGAGAGCGCACGAGGGAGCCGCCAGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCACTGATTTG
AGCGTCAGATTTCGTGATGCTTGTCAGGGGGGCGGAGCCTATGGAAAAACGGCTTTGCCGCGGCCCTCTCACTTCCCTGTTAAGTATCTTCCTGGCATCTTCCAGGAAATCTCCGCCCCGTTCGT
AAGCCATTTCCGCTCGCCGCAGTCGAACGACCGAGCGTAGCGAGTCAGTGAGCGAGGAAGCGGAATATATCCTGTATCACATATTCTGCTGACGCACCGGTGCAGCCTTTTTTCTCCTGCCACAT
GAAGCACTTCACTGACACCCTCATCAGTGCCAACATAGTAAGCCAGTATACACTCCGCTAGCGCTGATGTCCGGCGGTGCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGAC
AGGGTCGACCAAAGCGGCCATCGTGCCTCCCCACTCCTGCAGTTCGGGGGCATGGATGCGCGGATAGCCGCTGCTGGTTTCCTGGATGCCGACGGATTTGCACTGCCGGTAGAACTCCGCGAGGT
CGTCCAGCCTCAGGCAGCAGCTGAACCAACTCGCGAGGGGATCGAGCCCGGGGTGGGCGAAGAACTCCAGCATGAGATCCCCGCGCTGGAGGATCATCCAGCCGGCGTCCCGGAAAACGATTCCG
AAGCCCAACCTTTCATAGAAGGCGGCGGTGGAATCGAAATCTCGTGATGGCAGGTTGGGCGTCGCTTGGTCGGTCATTTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCAAGAAGGCGATAG
AAGGCGATGCGCTGCGAATCGGGAGCGGCGATACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCAAGCTCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGC
CACACCCAGCCGGCCACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCAGGCATCGCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCC
TGGCGAACAGTTCGGCTGGCGCGAGCCCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTTCCATCCGAGTACGTGCTCGCTCGATGCGATGTTTCGCTTGGTGGTCGAATGGG
CAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGCATCAGCCATGATGGATACTTTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAGCCAGTC
CCTTCCCGCTTCAGTGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGATAGCCGCGCTGCCTCGTCCTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAA
AAAGAACCGGGCGCCCCTGCGCTGACAGCCGGAACACGGCGGCATCAGAGCAGCCGATTGTCTGTTGTGCCCAGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAAT
CCATCTTGTTCAAGCATGCGAAACGACCGTCATCCTGTCTCTTGATCAGATCTTGATCCCCTGCGCCATCAGATCCTTGGCGGCAAGAAAGCCATCCAGTTTACTTTGCAGGGCTTCCCAACCTT
ACCAGAGGGCGCCCCAGCTGGCAATTCCGGTTCGCTTGCTGTCCATAAAACCGCCCAGTCTAGCTATCGCCATGTAAGCCCACTGCAAGCTACCTGCTTTCTCTTTGCGCTTGCGTTTTCCCTTG
TCCAGATAGCCCAGTAGCTGACATTCATCCGGGGTCAGCACCGTTTCTGCGGACTGGCTTTCTACGTGTTCCGCTTCCTTTAGCAGCCCTTGCGCCCTGAGTGCTTGCGGCAGCGTGAAGCTTAT
CGATGCGGCCGCCTCGAGAAAAGGCCATCCGTCAGGATGGCCTTCTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGCATCGATTTATTATGACAACTTGACGGCTACATCATTCACTTTTTCT
TCACAACCGGCACGGAACTCGCTCGGGCTGGCCCCGGTGCATTTTTTAAATACCCGCGAGAAATAGAGTTGATCGTCAAAACCAACATTGCGACCGACGGTGGCGATAGGCATCCGGGTGGTGCT
CAAAAGCAGCTTCGCCTGGCTGATACGTTGGTCCTCGCGCCAGCTTAAGACGCTAATCCCTAACTGCTGGCGGAAAAGATGTGACAGACGCGACGGCGACAAGCAAACATGCTGTGCGACGCTGG
CGATATCAAAATTGCTGTCTGCCAGGTGATCGCTGATGTACTGACAAGCCTCGCGTACCCGATTATCCATCGGTGGATGGAGCGACTCGTTAATCGCTTCCATGCGCCGCAGTAACAATTGCTCA
AGCAGATTTATCGCCAGCAGCTCCGAATAGCGCCCTTCCCCTTGCCCGGCGTTAATGATTTGCCCAAACAGGTCGCTGAAATGCGGCTGGTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGC
AAATATTGACGGCCAGTTAAGCCATTCATGCCAGTAGGCGCGCGGACGAAAGTAAACCCACTGGTGATACCATTCGCGAGCCTCCGGATGACGACCGTAGTGATGAATCTCTCCTGGCGGGAACA
GCAAAATATCACCCGGTCGGCAAACAAATTCTCGTCCCTGATTTTTCACCACCCCCTGACCGCGAATGGTGAGATTGAGAATATAACCTTTCATTCCCAGCGGTCGGTCGATAAAAAAATCGAGA
TAACCGTTGGCCTCAATCGGCGTTAAACCCGCCACCAGATGGGCATTAAACGAGTATCCCGGCAGCAGGGGATCATTTTGCGCTTCAGCCATACTTTTCATACTCCCGCCATTCAGAGAAGAAAC
CAATTGTCCATATTGCATCAGACATTGCCGTCACTGCGTCTTTTACTGGCTCTTCTCGCTAACCAAACCGGTAACCCCGCTTATTAAAAGCATTCTGTAACAAAGCGGGACCAAAGCCATGACAA
AAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGACGCTTT
TTATCGCAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGGTACCGGAAAAAGGAGATCTGCATATGAGAGGATCGCATCACCACCACCACCATAGCAGCGGCCATATGACAACACTTCTCAAC
CCCTACTTTGGTGAATTCGGCGGCATGTATGTGCCGCAGATCCTGATGCCTGCGCTGAACCAGCTTGAAGAGGCCTTCGTCAGCGCGCAAAAAGATCCTGAATTTCAGGCGCAATTCGCCGATCT
GCTAAAAAACTACGCGGGACGCCCCACCGCGCTGACGAAATGCCAGAACATTACCGCCGGTACGCGTACCACGTTGTATTTAAAGCGCGAAGATCTACTGCACGGCGGCGCGCACAAAACCAATC
AGGTACTGGGTCAGGCGCTGCTGGCCAAACGGATGGGTAAAAGCGAGATTATCGCTGAAACCGGCGCCGGTCAGCACGGCGTCGCCTCTGCGCTCGCCAGCGCCCTGCTGGGTCTGAAATGCCGT
ATCTATATGGGCGCCAAAGACGTTGAGCGCCAGTCGCCGAACGTCTTCCGTATGCGTCTGATGGGCGCTGAGGTCATCCCGGTTCATAGCGGCTCCGCTACGCTAAAAGATGCCTGTAACGAGGC
GCTGCGCGACTGGTCCGGTAGTTACGAAACCGCGCACTATATGCTCGGCACGGCGGCAGGACCGCATCCCTATCCCACCATCGTTCGCGAGTTCCAGCGCATGATTGGCGAAGAGACGAAAGCGC
AAATCCTCGACAAAGAGGGCCGTCTGCCAGATGCCGTTATCGCTTGCGTCGGTGGCGGCTCAAACGCTATCGGGATGTTTGCCGATTTTATTAATGATACCAGCGTCGGGCTAATAGGCGTTGAA
CCTGGTGGTCATGGTATTGAAACCGGCGAGCATGGCGCGCCGCTTAAACATGGTCGCGTTGGCATCTATTTCGGGATGAAAGCGCCGATGATGCAAACAGCAGACGGGCAAATTGAAGAGTCCTA
TTCCATTTCCGCCGGGCTCGATTTCCCGTCGTTGGGCCGCAGCATGCGTACCTGAACAGCATCGGACGCGCGGATTATGTCTCCATTACCGATGATGAGGCGCTGGAAGCCTTCAAAACGTTGT
GCCGCCATGAGGGAATTATCCCGGCGCTGGAGTCCTCCCACGCGTTGGCGCACGCTCTGAAAATGATGCGCGAGCAGCCGGAAAAAGAGCAACTGCTGGTGGTCAATCTCTCTGGCCGCGGAGAT
AAAGACATCTTTACCGTACACGATATCCTGAAAGCGCGAGGGGAAATCTGATGGAACGCTACGAAAATTTATTTGCCCAACTCAACGATCGCCGGGAAGGCGCTTTTGTCCCCTTCGTGACCCTG
GGCGACCCTGGCATTGAACAGTCACTGAAAATTATTGACACACTGATTGACGCCGGCGCCGACGCTCTAGAACTGGGGGTTCCCTTCTCCGATCCGCTGGCCGATGGCCCTACCATCCAGAATGC
GAACTTACGCGCCTTCGCCGCTGGCGTCACGCCGGCTCAGTGTTTTGAAATGCTGGCGCTGATTCGTGAAAAACACCCGACCATTCCGATTGGCCTGCTAATGTACGCGAATCTGGTGTTCAATA
ACGGCATAGATGCGTTCTATGCCCGTTGTGAACAGGTTGGCGTAGATTCCGTGCTGGTCGCAGATGTCCCGGTTGAAGAATCGGCCCCCTTCCGCCAGGCAGCGTTACGGCATAATATCGCGCCG
ATCTTCATCTGCCCGCCAAATGCGGATGACGATCTTCTGCGCCAGGTCGCATCTTACGGCCGCGGTTACACCTACCTGCTTTCGCGTTCGGGTGTCACCGGCGCGGAAAACCGTGGCGCATTGCC
GTTGCATCATCTCATTGAGAAGCTTAAAGAGTACCATGCCGCGCCTGCGTTACAGGGCTTCGGTATCTCCTCGCCGGAACAGGTGTCTGCGGCCGTGCGTGCCGGGGCGGCTGGCGCTATCTCCG
GCTCAGCCATTGTCAAGATTATCGAGAAAAACCTCGCGTCTCCCAAACAGATGTTGGCGGAGCTCAGGTCCTTTGTCTCAGCCATGAAAGCCGCCAGCCGCGCATAA

Additional file 15: Nucleotide sequence 2: pQE80L-His-ECFP

T5 promoter	7-87
His-tag	127-144
ECFP	151-870
lambda t_0	895-989
rrnB T1	1751-1848
lacI	1936-3018
ColE1	3596-4278
ampR	4373-5233

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCATTAAAGAGGAGAAATTAACTATGAGAGGATC
GCATCACCATCACCATCACGGATCCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGG
GCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTGGGGCGTGCAGTGCTTCAGCCGCTAC
CCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGG
CGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACATCAGCCACAACGTCTATATCACCGCCGACAAGCAGA
AGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCAC
TACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGATAAAA
GCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAAGCTAGCT
TGGCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGT
ACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATT
TCGTATGGCAATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGC
AGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCACCAGTTTT
GATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGG
CTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTG
AGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCCTCTAGATTACGTGCAGTCGATGATAA
GCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGC
GGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTT
TGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAAT
GGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTT
CCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGA
TGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATC
CTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCA
GTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGA
ATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTA
TAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCC
ACGGGTGCGCATGATCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCC
GTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACC
ATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCAC
TCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCT
CCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCC
TGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAA
CCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGT
AGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCG
GCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAA
AACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTA
CCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTG
CAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGT
TGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACG
ATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATA
ATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAAT
ACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTG
ATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAAT
ATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAA
ACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

Additional file 16: Nucleotide sequence 3: pET21a(+)-His-ECFP

f1 origin	12-467
ampR	599-1456
ColE1	1554-2263
lacI	3651-4730
T7 promoter/lac operator	5117-5160
His-tag	5217-5234
ECFP	5241-5960
T7 terminator	6067-6113

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTC
TCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGG
CCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGG
GATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAA
CCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCC
CTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAG
CGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCA
TACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAAC
TTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCG
TGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCAC
TTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATC
TACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTT
AAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGAT
CTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTC
AGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAG
TGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAAC
TGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAAC
GCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCT
GGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCG
AGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAA
GCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGA
CCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCG
TCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGG
GGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAA
AAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCC
AGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCA
ACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGG
CTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGC
ATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTA
ACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTT
TTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAA
CGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAA
CCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTA
TGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAA
AATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGA
CGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGC
GACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTC
CCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGAC
TCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCC
GTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCC
GATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGAC
TCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGAGAGGATCGCATCACCATCACCATCACGGATCCATGGTGAGCA
AGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTG
AAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTGGGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTC
CGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCA
TCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACATCAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCAC
AACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCC
CAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTGAAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACT
GAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGA
ACTATATCCGGAT

Additional file 17: Nucleotide sequence 4: pET21a(+)-SBP-*Gt*HNL

f1 origin	12-467
ampR	599-1456
ColE1	1554-2263
lacI	3651-4730
T7 promoter / lac operator	5117-5160
SBP	5208-5339
TEV site	5349-5366
*Gt*HNL	5367-5762
T7 terminator	5873-5918

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTC
TCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGG
CCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGG
GATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAA
CCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCC
CTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAG
CGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCA
TACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAAC
TTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCG
TGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCAC
TTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATC
TACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTT
AAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGAT
CTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTC
AGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAG
TGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAAC
TGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAAC
GCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCT
GGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCG
AGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAA
GCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGA
CCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCG
TCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGG
GGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAA
AAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCC
AGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCA
ACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGG
CTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGC
ATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTA
ACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTT
TTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAA
CGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAA
CCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTA
TGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAA
AATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGA
CGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGC
GACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTC
CCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGAC
TCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCC
GTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCC
GATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGAC
TCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGGACGAGAAGACCACCGGCTGGCGGGGCGGCCACGTGGTGGAGG
GCCTGGCCGGCGAGCTGGAGCAGCTGCGGGCCAGGCTGGAGCACCACCCTCAGGGCCAGCGGGAGCCCTCCGGCGGCTGCAAGCTGGGCCTGGGTACCGAGAACCTGTACTTCCAATCCATGGAA
ATTAAACGTGTTGGTAGCCAGGCAAGCGGTAAAGGTCCGGCAGATTGGTTTACCGGCACCGTTCGTATTGATCCGCTGTTTCAGGCACCGGATCCGGCACTGGTTGCCGGTGCAAGCGTTACCTT
TGAACCGGGTGCACGTACCGCATGGCATACCCATCCGCTGGGTCAGACCCTGATTGTTACCGCAGGTTGTGGTTGGGCACAGCGTGAAGGTGGTGCAGTTGAAGAAATTCATCCGGGTGATGTTG
TTTGGTTTAGTCCGGGTGAAAAACATTGGCATGGTGCAGCACCGACCACCGCAATGACCCATCTGGCAATTCAGGAACGTCTGGACGGTAAAGCAGTTGATTGGATGGAACATGTTACCGATGAA
CAGTATCGTCGCTAAAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC
ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

Additional file 18: Nucleotide sequence 5: pQE80L-SBP-*Gt*HNL

T5 promoter	7-87
SBP	115-249
TEV site	259-276
*Gt*HNL	277-675
lambda t_0	710-804
rrnB T1	1566-1663
lacI	1751-2866
ColE1	3411-4093
ampR	4188-5048

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCATTAAAGAGGAGAAATTAACTATGGACGAGAA
GACCACCGGCTGGCGGGGCGGCCACGTGGTGGAGGGCCTGGCCGGCGAGCTGGAGCAGCTGCGGGCCAGGCTGGAGCACCACCCTCAGGGCCAGCGGGAGCCCTCCGGCGGCTGCAAGCTGGGCC
TGGGTACCGAGAACCTGTACTTCCAATCCATGGAAATTAAACGTGTTGGTAGCCAGGCAAGCGGTAAAGGTCCGGCAGATTGGTTTACCGGCACCGTTCGTATTGATCCGCTGTTTCAGGCACCG
GATCCGGCACTGGTTGCCGGTGCAAGCGTTACCTTTGAACCGGGTGCACGTACCGCATGGCATACCCATCCGCTGGGTCAGACCCTGATTGTTACCGCAGGTTGTGGTTGGGCACAGCGTGAAGG
TGGTGCAGTTGAAGAAATTCATCCGGGTGATGTTGTTTGGTTTAGTCCGGGTGAAAAACATTGGCATGGTGCAGCACCGACCACCGCAATGACCCATCTGGCAATTCAGGAACGTCTGGACGGTA
AAGCAGTTGATTGGATGGAACATGTTACCGATGAACAGTATCGTCGCTAATTAATGAGGATCCAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCAT
CTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAAGCTAGCTTGGCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCA
CCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAAT
AAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACAC
CGTTTTCCATGAGCAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCC
CTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATAT
TATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGG
GGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGT
TTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCCCTCTAGATTACGTGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGC
TCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGG
GCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTG
TCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGAT
GCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGAC
GCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTC
TGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTG
CACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGAC
TGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGG
CTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGC
CATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCGGAATTTCGGGCAGCGTTGGGTCCTGGCCACGGGTGCGCATGATCTAGAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTG
ACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTA
GCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTT
CCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGA
GCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGAC
AGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATA
GCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCG
GTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGT
ATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCA
GAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTT
TTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGT
TGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAG
CCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATT
GCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCC
TCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCA
AGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGA
AAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAA
TGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTT
AGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTT
CAC

Additional file 19: Table S5: Minimal medium composition

Final concentrations are indicated. All chemicals were from Carl Roth, Karlsruhe, Germany unless indicated otherwise.

Description	Component	c [unit as indicated]
Trace elements	$FeSO_4 \cdot 7\ H_2O$	9 µM
	$MnSO_4 \cdot H_2O$	3.5 µM
	$AlCl_3 \cdot 6\ H_2O$	2.5 µM
	CoCl2·6 H2O	2 µM
	$ZnSO_4 \cdot 7\ H_2O$	0.4 µM
	$Na_2MoO_4 \cdot 2\ H_2O$	0.5 µM
	$CuCl_2 \cdot 2$ H2O	0.4 µM
	H_3BO_3	0.5 µM
Salts	Na_2HPO_4	47.8 mM
	KH_2PO_4	22.0 mM
	NaCl	8.6 mM
	NH_4Cl	18.6 mM
Other ingredients	Glucose	20.0 mM
	$MgSO_4 \cdot 7\ H_2O$	1.0 mM
	$CaCl_2 \cdot 2$ H2O	7 µM $CaCl_2$
	19 amino acid supplementation (without Trp)	0.5 mg/L
	Trp	18 µM
	D(+)-Biotine	4.0 µM
	Thiamine hydrochloride	3.3 µM
	Ampicillin (Sigma-Aldrich, St. Louis, MO)	100 mg/L
	Kanamycin	50 mg/L

Additional file 20: Formula 1: Formula used in this study to calculate the relative mean deviation (RMD) of D_{600} values

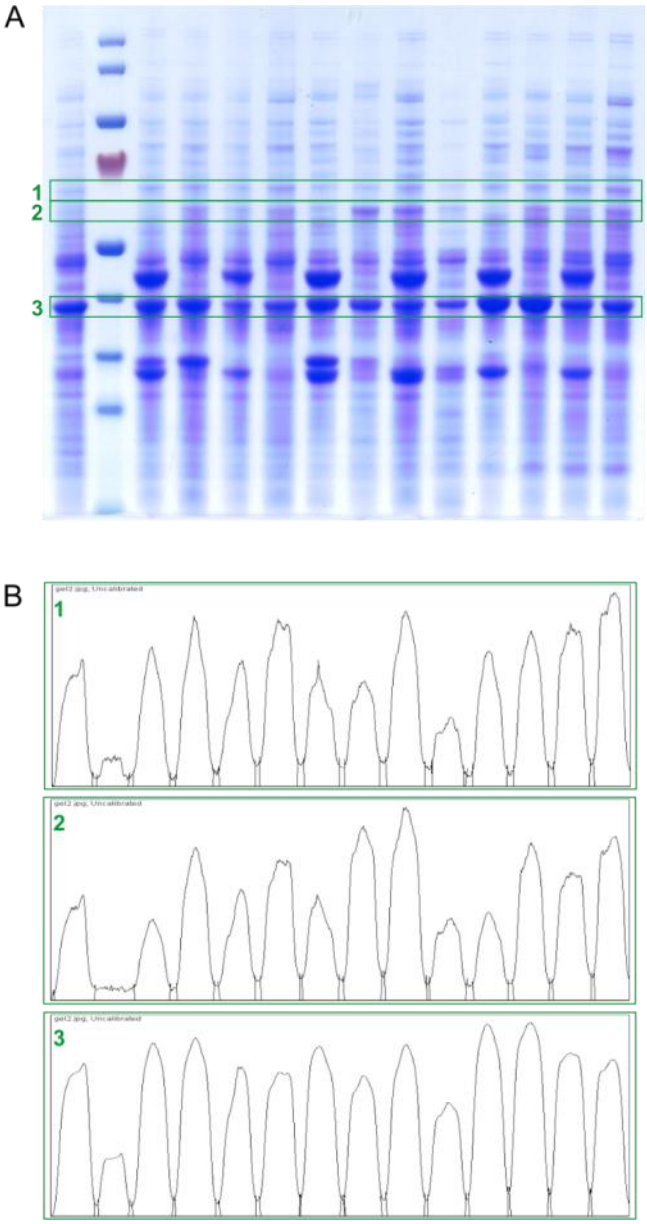
n refers to the number of replicates, x to the D_{600} of a replicate and $\bar{x}$ to the mean of x

$$RMD = \frac{\Sigma|x - \bar{x}|}{n} * \frac{100}{\bar{x}}$$

Additional file 21: Staining of SDS gels

Coomassie staining solution contained 2.5 g Brilliant Blue G250, 7.5% (v/v) acetic acid and 50% (v/v) ethanol in water. Gels were stained for at least 30 min, rinsed with water and destained with 7.5% (v/v) acetic acid and 20% (v/v) ethanol aqueous solution.

Silver staining of SDS gels was originally described in Merril et al. [67] and modified accordingly. Briefly, silver staining included incubation with 50% (v/v) water, 40% (v/v) ethanol, and 10% (v/v) acetic acid in doubly distilled water (ddH_2O) for 20 min and then with a 0.2% (w/v) sodium thiosulfate, 0.5 M sodium acetate, 30% (v/v) ethanol aqueous solution for 20 min. Gels were rinsed with ddH_2O and incubated with ddH_2O 4 times for 5 min before incubation with a 0.2% (w/v) silver nitrate aqueous solution for 15 min. Gels were washed once with water. For color development they were incubated with a 3% (w/v) sodium carbonate, 0.01% (v/v) formaldehyde aqueous solution until the staining reaction was stopped by exchanging the staining solution for 1.5% (w/v) ethylenediaminetetraacetic acid (EDTA) aqueous solution for 10 min. Finally, the gels were washed with water.

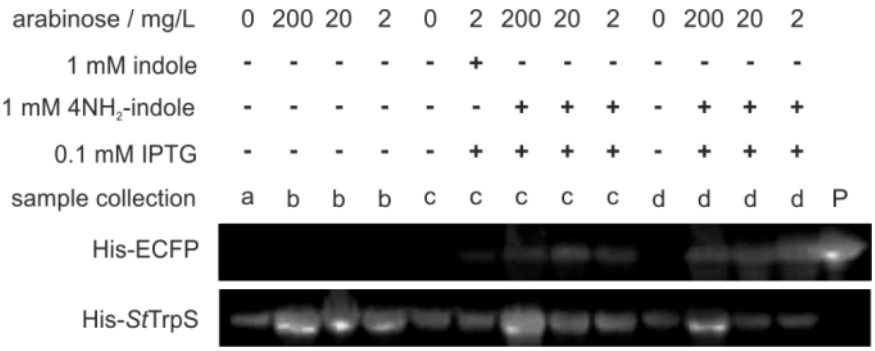


Additional file 22: Figure S7: Example for the densitometric analysis of SDS gels

Whole cell extracts are shown in a Coomassie stained 4-12% Bis-Tris SDS gel. Cells equaling 0.05 D_{600} units were loaded per lane. Three endogenous *E. coli* protein bands (1, 2 and 3 in A) were analyzed densitometrically (B) with the program ImageJ [62]. Integrated peak areas were calculated from B. From the relative integrated peak area of each protein band a normalization factor was calculated to correct for different amounts of whole cell extracts. This normalization factor was used to apply the same amount of whole cell extract in a SDS gel (Additional file 25: Figure S9) and in a Western blot (Figure 6).

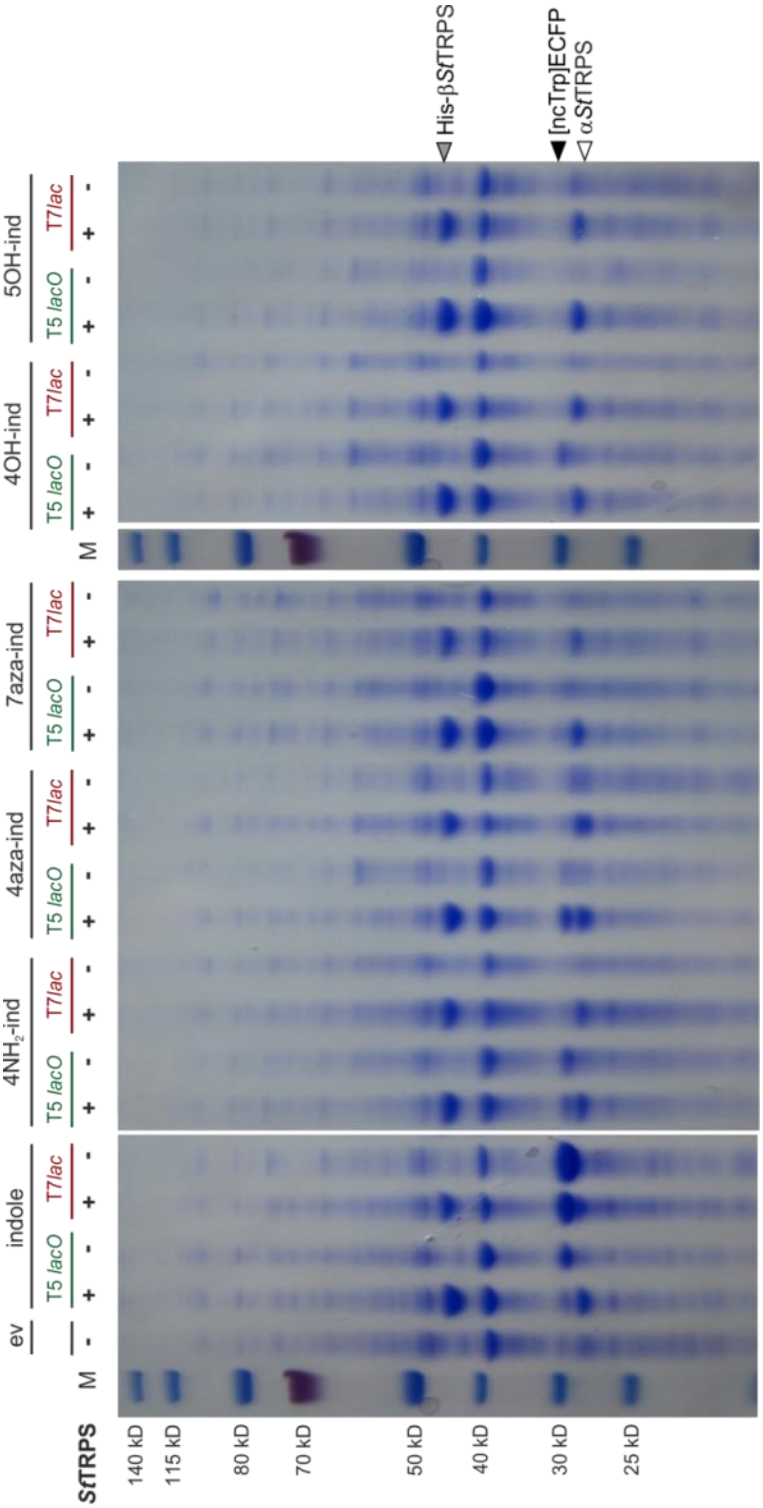
Additional file 23: Table S6: Relative mean deviation for average D_{600} values in Figure 2B calculated with the Additional file 20: Formula 1 for n=3

time / h sample	3.75	5.75	6.25	10.25	20
200 mg/L ara + $4NH_2$-indole	3.5	2.5	2.0	1.4	2.3
20 mg/L ara + $4NH_2$-indole	0.9	2.2	1.4	1.0	1.2
2 mg/L ara + $4NH_2$-indole	1.1	3.0	2.7	1.5	2.1
2 mg/L ara + indole	0.9	1.7	3.3	0.7	1.9
indole	2.6	1.0	1.8	0.8	1.3



Additional file 24: Figure S8: Western blot of the His-tagged β-subunit of *St*TRPS and the His-tagged ECFP variant in the insoluble protein (pellet) fraction

Samples were collected before induction of *St*TRPS with arabinose (a), before addition of indole / $4NH_2$-indole and induction with IPTG (b), after 3 h of expression (c), and after expression overnight (d) according to the scheme in Figure 2A. P refers to purified His-tagged ECFP as positive control. For the Western blot of the corresponding samples in the cleared lysate fraction refer to Figure 2C.



Additional file 25: Figure S9: Comparative expression of different [ncTrp]ECFP variants under control of the T7*lac* and T5 *lacO* promoters

Whole cell extracts of strains expressing various [ncTrp]ECFP variants from the T7*lac* (pET21a(+)-His-ECFP expression construct) or T5 *lacO* promoter (pQE80L-His-ECFP expression construct) in the presence (+) or absence (-) of *St*TRPS are shown. The 4-12% Bis-Tris gel was stained with Coomassie brillant blue. Calculated molecular weights: [ncTrp]ECFP, 28 kDa; α*St*TRPS, 29 kDa; His-β*St*TRPS, 45 kDa; α*St*TRPS and His-β*St*TRPS designate the α- and β-subunits of the *S. typhimurium* tryptophan synthase *St*TRPS. M, molecular size marker; ev, empty vector control; x-ind, indole analog

as listed in Figure 1A. The total protein concentrations of the whole cell extracts were normalized as described in the Methods section and equal amounts of protein were loaded in each lane.

Additional file 26: Table S7: Relative mean deviation for average D_{600} values in Table 1 and Additional file 3: Figure S2 calculated with the Additional file 20: Formula 1 for n=3

D_{600}	time / h ara / mg/L	1.45	2.25	2.45	3.5	4	5.75	6.75	8	9.5	25.5
	200	0.99	1.34	1.18	-	-	2.44	2.15	0.33	0.35	1.55
0.5	20	4.39	3.68	3.62	-	-	0.98	0.92	1.28	0.00	2.28
	2	1.0	1.0	1.8	-	-	0.7	1.7	2.6	0.9	0.0
	200	-	-	-	1.36	0.12	0.51	0.91	2.25	0.68	3.44
1	20	-	-	-	3.68	1.34	0.92	0.60	2.41	0.52	2.62
	2	-	-	-	0.79	0.54	0.27	3.35	1.35	1.63	2.32
	200	-	-	-	1.36	0.12	0.51	0.91	2.25	0.68	3.44
1.5	20	-	-	-	3.68	1.34	0.92	0.60	2.41	0.52	2.62
	2	-	-	-	0.79	0.54	0.27	3.35	1.35	1.63	2.32

Additional file 27: Table S8: Relative mean deviation for average D_{600} values in (Additional file 5: Figure S3) calculated with the Additional file 20: Formula 1 for n=3

time / h sample	3,75	4	4,75	5,5	5,75	6,25	6,75	10,25	20	21
protocol A	1,09				2,99	2,65		1,51	2,08	
protocol B		3,31	3,19	2,04		1,27	3,62			2,83

3.10 **References**

1. Lepthien S, Hoesl MG, Merkel L, Budisa N: **Azatryptophans endow proteins with intrinsic blue fluorescence.** *Proc Natl Acad Sci USA* 2008, **105:**16095-16100.
2. Bae JH, Rubini M, Jung G, Wiegand G, Seifert MHJ, Azim MK, Kim J-S, Zumbusch A, Holak TA, Moroder L, et al: **Expansion of the genetic code enables design of a novel "gold" class of green fluorescent proteins.** *J Mol Biol* 2003, **328:**1071-1081.
3. Budisa N, Pal PP, Alefelder S, Birle P, Krywcun T, Rubini M, Wenger W, Bae JH, Steiner T: **Probing the role of tryptophans in Aequorea victoria green fluorescent proteins with an expanded genetic code.** *Biol Chem* 2004, **385:**191-202.
4. Parsons JF, Xiao G, Gilliland GL, Armstrong RN: **Enzymes harboring unnatural amino acids: mechanistic and structural analysis of the enhanced catalytic activity of a glutathione transferase containing 5-fluorotryptophan.** *Biochemistry* 1998, **37:**6286-6294.
5. Acchione M, Lee Y-C, DeSantis ME, Lipschultz CA, Wlodawer A, Li M, Shanmuganathan A, Walter RL, Smith-Gill S, Barchi JJ: **Specific fluorine labeling of the HyHEL10 antibody affects antigen binding and dynamics.** *Biochemistry* 2012, **51:**6017-6027.
6. Boles JO, Henderson J, Hatch D, Silks LAP: **Synthesis and incorporation of [6,7]-selenatryptophan into dihydrofolate reductase.** *Biochem Biophys Res Commun* 2002, **298:**257-261.
7. Betts MJ, Russell RB: **Amino acid properties and consequences of substitutions.** In *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data.* Edited by Barnes MR. Chichester, West Sussex, England: John Wiley & Sons, Ltd; 2007: 289-316.[England (Series Editor)
8. Budisa N: *Engineering the genetic code - expanding the amino acid repertoire for the design of novel proteins.* Weinheim: Wiley-VCH Verlag GmbH & Co KGaA; 2006.
9. Hoesl MG, Budisa N: **Recent advances in genetic code engineering in Escherichia coli.** *Curr Opin Biotechnol* 2012, **23:**751-757.
10. Liu CC, Schultz PG: **Adding new chemistries to the genetic code.** *Annu Rev Biochem* 2010, **79:**413-444.
11. Chatterjee A, Sun SB, Furman JL, Xiao H, Schultz PG: **A versatile platform for single and multiple unnatural amino acid mutagenesis in Escherichia coli.** *Biochemistry* 2013.
12. Minks C, Huber R, Moroder L, Budisa N: **Atomic mutations at the single tryptophan residue of human recombinant annexin V: effects on structure, stability, and activity.** *Biochemistry* 1999, **38:**10649-10659.
13. Budisa N, Steipe B, Demange P, Eckerskorn C, Kellermann J, Huber R: **High-level biosynthetic substitution of methionine in proteins by its analogs 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in Escherichia coli.** *Eur J Biochem* 1995, **230:**788-796.
14. van Hest JC, Kiick KL, Tirrell DA: **Efficient incorporation of unsaturated methionine analogues into proteins in vivo.** *J Am Chem Soc* 2000, **122:**1282-1288.
15. Wiltschi B: **Expressed protein modifications: Making synthetic proteins.** In *Methods in Molecular Biology - Synthetic Gene Networks. Volume* 813. Edited by Weber W, Fussenegger M: Humana Press; 2012: 211-225: *Methods in Molecular Biology*].
16. Budisa N, Pal PP: **Designing novel spectral classes of proteins with a tryptophan-expanded genetic code.** *Biol Chem* 2004, **385:**893-904.
17. Sun X, Lin Y, Yuan Q, Yan Y: **Precursor-directed biosynthesis of 5-hydroxytryptophan using metabolically engineered E. coli.** *ACS Synthetic Biology* 2015, **4:**554-558.
18. Goss RJM, Newill PLA: **A convenient enzymatic synthesis of l-halotryptophans.** *Chem Commun* 2006**:**4924-4925.
19. Crawford IP, Yanofsky C: **On the separation of the tryptophan synthetase of Escherichia coli into two protein components.** *Proc Natl Acad Sci USA* 1958, **44:**1161-1170.
20. Adachi O, Kohn LD, Miles EW: **Crystalline α2β2 complexes of tryptophan synthetase of Escherichia coli : A comparison between the native complex and the reconstituted complex.** *J Biol Chem* 1974, **249:**7756-7763.
21. Dunn MF: **Allosteric regulation of substrate channeling and catalysis in the tryptophan synthase bienzyme complex.** *Arch Biochem Biophys* 2012, **519:**154-166.
22. Smith DRM, Willemse T, Gkotsi DS, Schepens W, Maes BUW, Ballet S, Goss RJM: **The first one-pot synthesis of L-7-iodotryptophan from 7-iodoindole and serine, and an improved synthesis of other L-7-halotryptophans.** *Org Lett* 2014, **16:**2622-2625.
23. Winn M, Roy AD, Grüschow S, Parameswaran RS, Goss RJM: **A convenient one-step synthesis of l-aminotryptophans and improved synthesis of 5-fluorotryptophan.** *Bioorg Med Chem Lett* 2008, **18:**4508-4510.
24. Perni S, Hackett L, Goss RJM, Simmons MJ, Overton TW: **Optimisation of engineered Escherichia coli biofilms for enzymatic biosynthesis of l-halotryptophans.** *AMB Express* 2013, **3:**66-66.
25. Wilcox M: **The enzymatic synthesis of L-tryptophan analogues.** *Anal Biochem* 1974, **59:**436-440.
26. Phillips RS: **Synthetic applications of tryptophan synthase.** *Tetrahedron Asymmetry* 2004, **15:**2787-2792.
27. Giese C, Lepthien S, Metzner L, Brandsch M, Budisa N, Lilie H: **Intracellular uptake and inhibitory activity of aromatic fluorinated amino acids in human breast cancer cells.** *ChemMedChem* 2008, **3:**1449-1456.
28. Crowley PB, Kyne C, Monteith WB: **Simple and inexpensive incorporation of 19F-Tryptophan for protein NMR spectroscopy.** *Chem Commun* 2012, **48:**10681-10683.
29. Bae JH, Alefelder S, Kaiser JT, Friedrich R, Moroder L, Huber R, Budisa N: **Incorporation of β-selenolo[3,2-b]pyrrolyl-alanine into proteins for phase determination in protein X-ray crystallography.** *J Mol Biol* 2001, **309:**925-936.
30. Budisa N, Rubini M, Bae JH, Weyher E, Wenger W, Golbik R, Huber R, Moroder L: **Global replacement of tryptophan with aminotryptophans generates non-invasive protein-based optical pH sensors.** *Angew Chem Int Ed Engl* 2002, **41:**4066-4069.
31. Hoesl MG, Larregola M, Cui H, Budisa N: **Azatryptophans as tools to study polarity requirements for folding of green fluorescent protein.** *J Pept Sci* 2010, **16:**589-595.
32. Yang L-h, Ahmed SA, Wilson Miles E: **PCR mutagenesis and overexpression of tryptophan synthase from Salmonella typhimurium: on the roles of beta2 subunit Lys-382.** *Protein Expr Purif* 1996, **8:**126-136.
33. Studier FW, Moffatt BA: **Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes.** *J Mol Biol* 1986, **189:**113-130.
34. Rosenberg AH, Lade BN, Dao-shan C, Lin S-W, Dunn JJ, Studier FW: **Vectors for selective expression of cloned DNAs by T7 RNA polymerase.** *Gene* 1987, **56:**125-135.
35. Budisa N, Alefelder S, Bae JH, Golbik R, Minks C, Huber R, Moroder L: **Proteins with β-(thienopyrrolyl)alanines as alternative chromophores and pharmaceutically active amino acids.** *Protein Science : A Publication of the Protein Society* 2001, **10:**1281-1292.

36. Hoesl MG, Acevedo-Rocha CG, Nehring S, Royter M, Wolschner C, Wiltschi B, Budisa N, Antranikian G: **Lipase congeners designed by genetic code engineering.** *ChemCatChem* 2011, **3:**213-221.
37. Phillips RS, Cohen LA, Annby U, Wensbo D, Gronowitz S: **Enzymatic synthesis of thia-L-tryptophans.** *Bioorg Med Chem Lett* 1995, **5:**1133-1134.
38. Sloan MJ, Phillips RS: **Enzymatic synthesis of aza-L-tryptophans: The preparation of 5- and 6-aza-L-tryptophan.** *Bioorg Med Chem Lett* 1992, **2:**1053-1056.
39. Piñero-Fernandez S, Chimerel C, Keyser UF, Summers DK: **Indole transport across Escherichia coli membranes.** *J Bacteriol* 2011, **193:**1793-1798.
40. Fowden L, Lewis D, Tristram H: **Toxic amino acids: their action as antimetabolites.** In *Adv Enzymol Relat Areas Mol Biol.* John Wiley & Sons, Inc.; 1967: 89-163
41. Pratt EA, Ho C: **Incorporation of fluorotryptophans into proteins of Escherichia coli.** *Biochemistry* 1975, **14:**3035-3040.
42. Bacher JM, Ellington AD: **Selection and characterization of Escherichia coli variants capable of growth on an otherwise toxic tryptophan analogue.** *J Bacteriol* 2001, **183:**5414-5425.
43. Hoesl MG, Oehm S, Durkin P, Darmon E, Peil L, Aerni H-R, Rappsilber J, Rinehart J, Leach D, Söll D, Budisa N: **Chemical Evolution of a Bacterial Proteome.** *Angew Chem Int Ed Engl* 2015, **54:**10030-10034.
44. Hajnal I, Łyskowski A, Hanefeld U, Gruber K, Schwab H, Steiner K: **Biochemical and structural characterization of a novel bacterial manganese-dependent hydroxynitrile lyase.** *FEBS J* 2013, **280:**5815-5828.
45. Wiedner R, Kothbauer B, Pavkov-Keller T, Gruber-Khadjawi M, Gruber K, Schwab H, Steiner K: **Improving the properties of bacterial R-selective hydroxynitrile lyases for industrial applications.** *ChemCatChem* 2015, **7:**325-332.
46. Petrović DM, Leenhouts K, van Roosmalen ML, Broos J: **An expression system for the efficient incorporation of an expanded set of tryptophan analogues.** *Amino Acids* 2013, **44:**1329-1336.
47. Link AJ, Vink MKS, Tirrell DA: **Presentation and detection of azide functionality in bacterial cell surface proteins.** *J Am Chem Soc* 2004, **126:**10598-10602.
48. Kiick KL, van Hest JCM, Tirrell DA: **Expanding the scope of protein biosynthesis by altering the methionyl-tRNA synthetase activity of a bacterial expression host.** *Angew Chem Int Ed Engl* 2000, **39:**2148-2152.
49. Kiick KL, Tirrell DA: **Protein engineering by in vivo incorporation of non-natural amino acids: Control of incorporation of methionine analogues by methionyl-tRNA synthetase.** *Tetrahedron* 2000, **56:**9487-9493.
50. Tang Y, Tirrell DA: **Biosynthesis of a highly stable coiled-coil protein containing hexafluoroleucine in an engineered bacterial host.** *J Am Chem Soc* 2001, **123:**11089-11090.
51. Studier WF, Rosenberg AH, Dunn JJ, Dubendorff JW: **Use of T7 RNA polymerase to direct expression of cloned genes.** In *Meth Enzymol. Volume* Volume 185. Edited by David VG: Academic Press; 1990: 60-89
52. Dubendorf JW, Studier FW: **Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor.** *J Mol Biol* 1991, **219:**45-59.
53. Chamberlin M, McGrath J, Waskell L: **New RNA polymerase from Escherichia coli infected with bacteriophage T7.** *Nature* 1970, **228:**227-231.
54. Gentz R, Bujard H: **Promoters recognized by Escherichia coli RNA polymerase selected by function: highly efficient promoters from bacteriophage T5.** *J Bacteriol* 1985, **164:**70-77.
55. Ayyadurai N, Neelamegam R, Nagasundarapandian S, Edwardraja S, Park H, Lee S, Yoo T, Yoon H, Lee S-G: **Importance of expression system in the production of unnatural recombinant proteins in Escherichia coli.** *Biotechnol Bioprocess Eng* 2009, **14:**257-265.
56. Al-Abdul-Wahid MS, DeMill CM, Serwin MB, Prosser RS, Stewart BA: **Effect of juxtamembrane tryptophans on the immersion depth of Synaptobrevin, an integral vesicle membrane protein.** *Biochim Biophys Acta* 2012, **1818:**2994-2999.
57. Datsenko KA, Wanner BL: **One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.** *Proc Natl Acad Sci USA* 2000, **97:**6640-6645.
58. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO: **Enzymatic assembly of DNA molecules up to several hundred kilobases.** *Nat Meth* 2009, **6:**343-345.
59. Datta S, Costantino N, Court DL: **A set of recombineering plasmids for gram-negative bacteria.** *Gene* 2006, **379:**109-115.
60. Samarkina ON, Popova AG, Gvozdik EY, Chkalina AV, Zvyagin IV, Rylova YV, Rudenko NV, Lusta KA, Kelmanson IV, Gorokhovatsky AY, Vinokurov LM: **Universal and rapid method for purification of GFP-like proteins by the ethanol extraction.** *Protein Expr Purif* 2009, **65:**108-113.
61. Laemmli UK: **Cleavage of structural proteins during the assembly of the head of bacteriophage T4.** *Nature* 1970, **227:**680-685.
62. Schneider CA, Rasband WS, Eliceiri KW: **NIH Image to ImageJ: 25 years of image analysis.** *Nat Meth* 2012, **9:**671-675.
63. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Res* 2014, **42:**D199-205.
64. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
65. **Kyoto Encyclopedia of Genes and Genomes (KEGG) database.** http://www.genome.jp/kegg-bin/show_pathway?org_name=ebl&mapno=00400&mapscale=&show_description=hide. Accessed 23 November 2015.
66. Rusconi F: **massXpert 2: a cross-platform software environment for polymer chemistry modelling and simulation/analysis of mass spectrometric data.** *Bioinformatics* 2009, **25:**2741-2742.
67. Merril CR, Goldman D, Van Keuren ML: **[17] Silver staining methods for polyacrylamide gel electrophoresis.** In *Meth Enzymol. Volume* Volume 96. Edited by Sidney Fleischer BF: Academic Press; 1983: 230-239

4 Chapter

Modular assembly of multi-part DNA constructs by type IIS restriction cloning: A MoClo manual

Corinna Odar[1,2] , Maria Koshanskaya[1,2], Birgit Wiltschi[1]

[1] Austrian Centre of Industrial Biotechnology, Petersgasse 14, Graz, Austria

[2] Graz University of Technology, Institute of Molecular Biotechnology, Petersgasse 14, Graz Austria

Authors' contributions

Corinna Odar conceived the MoClo design, developed the nomenclature and wrote the manuscript. Maria Koshanskaya did all the cloning work with supervision by Corinna Odar and helped with preparation of Figure 9 and Figure 11. Birgit Wiltschi supervised the design and the work.

4.1 **Abstract**

The design of complex DNA constructs gains increasing relevance in the field of molecular biosciences and especially in Synthetic Biology. Up-to-date assembly strategies exploit the modular nature of the information encoded on the DNA. These Modular Cloning (MoClo) approaches are hierarchical assembly strategies that use Golden Gate cloning for the directional, scarless assembly of standardized DNA modules. A decisive advantage of the MoClo strategy is its independence of the target sequences to be assembled what allows standardization and automation of the assembly process. Along with the enhanced efficiency for complex DNA assemblies evolved the need for well-designed cloning approaches to guarantee a standardized widely applicable cloning strategy.

Here, we explain and illustrate a newly designed MoClo approach for the assembly of *E. coli* expression constructs including complex biological pathways as well as promoter-, RBS- or tag-libraries. It relies on a plasmid platform of basic biological parts that are assembled to multigene constructs in a succession of iterative assembly steps. According to the biological function of a part, it is assigned with a specified standardized identity. This identity directly refers to the parts position in a directed arrangement to give multipartite constructs. We provide a systematic nomenclature that unambiguously designates the identity of a part. Moreover, the nomenclature system includes all DNA sequences that are crucial for the standardization of the MoClo assembly approach.

Realization of the presented MoClo strategy will allow for the straight-forward as well as time- and cost-saving assembly of complex DNA constructs for a diverse field of applications.

4.2 **Background**

In a strategic vision from ERASynBio “synthetic biology is defined as the engineering of biology” [1]. With DNA as the basic instruction set of any organism, the engineering of this instruction set presents an inherent part of synthetic biology and an essential tool in molecular biology. Tremendous progress in the field of genetic engineering brought forth advanced strategies for the faster and more reliable assembly of DNA fragments [2]. The modular nature of the information encoded on DNA facilitates the design of standardized and combinatorial DNA assembly strategies for the routine engineering of new biological pathways or the construction of libraries.

Golden Gate cloning has been recognized as a highly useful tool providing the basis for the modular cloning (MoClo) approach [3]. Golden Gate cloning exploits the nature of type IIS endonucleases that cleave DNA distal to their recognition sequences (Figure 1) [4]. With rare exceptions, such as MlyI, type IIS endonucleases produce cohesive ends. The sequence of the single-stranded DNA overhangs is not dependent on the enzyme but is determined by the nucleotides that happen to be present between the cleavage sites on the two opposite DNA strands (Figure 1; N indicates any nucleotide). Consequently, the sequence of these nucleotides can be freely chosen and it is revealed when the enzyme cuts. Conveniently, overhangs with different nucleotide sequences can be generated by the same type IIS endonuclease. Complementary sequence overhangs function as fusion sites by which DNA molecules can be ligated and carefully designed fusion sites allow scarless links between adjacent fragments. With the free choice of specific fusion sites the directional assembly of up to nine DNA fragments is possible in a single Golden Gate cloning reaction [5]. Moreover, recognition sequences are cleaved off during the assembly process and consequently new constructs are no subject of re-digestion. These are decisive advantages over classical cloning techniques with type II endonucleases, which cut inside a palindromic recognition sequence. In this case, the fusion sites are pre-defined by the enzyme. The ligation of complementary fusion sites of type II endonucleases usually regenerates the recognition sequence, which can be re-cut by the same enzyme.

Golden Gate cloning works in a one-pot reaction, where several undigested DNA fragments or plasmids are incubated in the presence of a type IIS restriction enzyme and ligase. The application of sequence-verified plasmids in Golden Gate cloning simplifies cloning work by abolishing the need for PCRs, gel extraction of DNA fragments and sequence verification in each assembly step.

The MoClo approach presented in this study is a bottom-up, hierarchical assembly strategy that uses several successive Golden Gate reactions for the assembly of DNA constructs consisting of several modules. MoClo abstracts a DNA sequence into modular elements based on the encoded genetic information [6]. These modular DNA fragments assemble in a standardized directional manner to give higher functional units. Modularity and standardization allow for the interchangeable use of DNA modules in the MoClo system. The universal character of module types guarantees their reuse for the combinatorial assembly of several DNA constructs. The hierarchical MoClo assembly approach applies several successions of assembly steps, *i.e.* Golden Gate reactions, each giving rise to a specific module type. On each level of assembly unique fusion sites are generated on the respective

modules upon restriction so that they can be combined in an order pre-designed by the MoClo user. The DNA fragments to be assembled have to be devoid of any recognition sites of endonucleases used in this multi-step assembly approach. User-specific multigene constructs can be assembled from pre-made plasmids that are deposited in the combinatorial library collection.

Careful design of MoClo library plasmids and of the hierarchical assembly steps is a prerequisite for the establishment of a functional MoClo system. Here we describe the MoClo system that was set up in the Junior Group Synthetic Biology at the Austrian Centre of Industrial Biotechnology – acib. It is based on studies by Weber *et al.* [3] and Lee *et al.* [6]. However, its conceptional details as well as nomenclature of DNA fragments deviate from the mentioned studies. Moreover, the presented cloning strategy is designed for the assembly of *E. coli* expression constructs and not for eukaryotic systems. However, due to its flexible but standardized modularity, the assembly system described in this study can be extended to DNA constructs for any expression host.

4.3 The conceptional design

We designed the MoClo assembly strategy with the aim to meet the following requirements: Each level of assembly and module used therein has to be standardized to maintain interchangeability within each level and among the three levels. The system has to provide enough flexibility for “out of the box” experiments, without the need to deviate from the standardized design. Modularity must be designed in a logical way to be self-explanatory for a user familiar with the assembly strategy. This should guarantee that the module library can be used for a wide range of experiments. The application of the assembled DNA constructs ranges from simple expression experiments and studies on the influence of tags on protein expression over libraries for promoters, ribosomal binding sites (RBSs), or target protein mutants to complex pathway assemblies. Moreover, a less experienced user should be able to apply the assembly system knowing its basic rules and features without a deeper insight into its complex details.

4.3.1 The three levels of assembly

The assembly procedure was abstracted into three levels of Golden Gate reactions each giving rise to a standardized circular vector construct harboring a defined module (Figure 2). The modules are integrated into destination vectors, which are specific for each assembly level. A plasmid construct from one assembly level becomes the donor plasmid for the higher order level.

On level°0, part plasmids are assembled from DNA fragments obtained by a PCR reaction or directly by DNA synthesis (Figure 2A). These basic DNA parts are either regulatory DNA sequences, like promoters or terminators, or the coding sequence of a gene of interest (GOI) (for further definition of parts see chapter 2.2.). They constitute the basic DNA modules deposited in the MoClo plasmid collection. The part plasmids are used on level°1 to assemble a multipartite DNA construct, termed a

device, in a new destination vector (Figure 2B). A device can already be a fully functional transcriptional unit, when it consists at least of three parts comprising a promoter, a GOI and a terminator. Connectors are the second module type at level°1 in addition to parts. Connectors flank each device and contain the fusion sites for their assembly in the destination vector on level°1 (Figure 2B). Moreover, additional fusion sites inside the connector sequence facilitate the assembly of several devices at level°2 to obtain a multigene/pathway plasmid (Figure 2C) (for further description on the connector design see chapter 2.2.7).

The part generated at level°0 can be conceived as a genetic module, the device at level°1 as a transcriptional module and the pathway/multigene at level°2 as a metabolic module.

Before level°0 assembly the parts have to be furnished with flanking overhangs that contain the restriction and fusion sites for level°0 as well as for level°1 (Figure 2A). As indicated by arrows in Figure 2 and explained in the background section above, type IIS restriction enzymes work in a directional manner (Figure 1). The type IIS recognition sequence is placed so that restriction creates a fusion site directly flanking the module sequence, while cutting off the recognition sequence itself. BsaI is used in the first round of assembly at level°0 to generate a part plasmid or connector plasmid. Digestion with BsaI liberates the part or connector with universal fusion sites protruding like sticky ends while the BsaI recognition sequence is cleaved off. Likewise, cleavage of the level°0 destination vector exposes compatible fusion sites and excises the *lacZ* reporter together with the BsaI recognition sites. The parts and/or connectors anneal to the destination vector via their compatible fusion sites and the resulting DNA constructs are stably linked by ligation (Figure 2A). Since the recognition sites are removed during the assembly, the resulting DNA constructs are immune to re-digestion by BsaI. This is the reason why the restriction enzyme and the ligase can be combined in one reaction.

On level°1, BpiI is used to cleave the parts and connectors leaving unique and specific fusion sites on these modules. This facilitates the assembly of the modules in a certain order in the new destination vector on level°1 (Figure 2B). Again, BpiI cuts all part and connector plasmids as well as the destination vector in a one-pot reaction which also contains the ligase.

On level°2, the generated devices are digested with BsaI with its recognition sequence found inside the connectors (Figure 2C). Each connector flanking the devices contains unique fusion sites for the directional assembly of the devices to give a multigene/pathway plasmid on level°2.

The levels can be further expanded to higher orders. For example, a level°3 could be added by furnishing several level°2 destination vectors with additional flanking BpiI restriction sites generating specific fusion sites. Several pathway modules could be assembled on a level°3 by BpiI digest. However, the addition of further levels is limited by the size and stability of the resulting constructs.

4.3.2 **Nomenclature and special considerations**

A consistent systematic nomenclature is an inherent component for the design of and work with complex assembly strategies. It is obligatory that all users comply with one unambiguous nomenclature concept to maintain the standardized modularity of the system as well as the interchangeability of the modules.

The nomenclature should be self-explanatory and should inherently reveal the position as well as the function of a module in the standardized assembly framework. Users of the system must be familiar with its nomenclature rules what implies that they have understood the design of the assembly system.

The nomenclature designates **modules**, **overhangs** and **fusion sites** (Figure 3). All module names are listed in Table 1 and will be explained in detail in the following chapter.

4.3.2.1 **Modules**

Modules comprise basic biological parts, multipartite devices, or entire multigene systems.

Lower case characters define the **identity** of a certain module:

On all assembly levels, the abbreviation “ve” refers to the vector backbone of the destination plasmids and “in” to the inserts. On level°0 “in” is a part or a connector, on level°1 “in” is a device flanked by connectors and on level°2 “in” is a multigene construct.

Part plasmids generated on level°0 either have an “a”, “b” or “c” identity according to their relative position in a transcriptional unit. The regulatory sequences driving gene expression, like promoter/operator sequence, are termed a-parts. a-Parts are always followed downstream by b-parts which always include several subparts like RBS, the coding sequence and an optional N- or C-terminal tag (for further description on subparts see chapter 2.2.5). The transcriptional terminator is a c-part that follows downstream of a b-part.

Connectors are abbreviated by “con”; ”dev” refers to a device generated on level°1 (for further description on connectors and devices see chapter 2.2.7).

4.3.2.2 **Overhangs**

Each part/subpart, connector or vector backbone is flanked by an **upstream 5'-overhang** and by a **downstream 3'-overhang**. The overhangs are composed of the type IIS recognition sites corresponding to the level°0 and level°1 assembly reactions as well as the fusion sites (Figure 3A).

The nomenclature of parts/subparts can be directly deduced from the overhangs they are furnished with on level°0 and *vice versa* (Figure 2A and Figure 3). Overhangs always include the identity of the module (lower case letter ve, a, b, c or con) they belong to (Figure 3B). Moreover, an S (start) or even number in the overhang name refers to its position at the 5' terminus on the part/subpart. An E (end) or odd number refers to the 3' terminus on the part/subpart (for further description see chapter 2.2.4). The 5' terminus is indicated first, then the 3' terminus. For instance, aSE, bSE and cSE parts designate a promoter, a GOI (with RBS and optionally tags) and a terminator in a monocistronic

transcription unit (Figure 3B and Figure 4). The assembly of polycistronic expression constructs is mediated by connectors in level°2, see chapter 2.2.8 for details.

4.3.2.3 **Fusion sites**

Also the nomenclature of the four nucleotide long fusion sites generated upon type IIS restriction has to be defined (Table 2). Names of fusion sites always start with “f-“.

The fusion sites at level°0 assemblies are always f-veE-inS (in the 5'-overhang) and f-inE-veS (in the 3' overhang) for ligation of any part/subpart or connector into pMC0.

Fusion site names joining parts/subparts at level°1 always consist of: f-(3’-overhang name of the first/upstream part) - (5’-overhang name of the second/downstream part). In this way the name unambiguously shows which parts/subparts are assembled with the respective fusion site. The fusion site between the two parts aSE and bSE is f-aE-bS (Figure 4). It fuses the 3’-terminus of the a-part with the 5’-terminus of the b-part. Parts can consist of several subparts like bS1, b23, b45 and b6E (Figure 4B). Because assembled subparts have the same identity, the identity designator can be omitted for the 5'-overhang in the fusion site name. Fusion site names between b-subparts would be as follows: f-b1-2 between bS1 and b23, f-b3-5 between b34 and b45, etc.

On level°1, connectors are assembled so that they flank the multipartite device (Figure 2B). Consequently, the fusion site for ligation into the vector backbone pMC1 at level°1 is **always** inside the connectors. These fusion sites are named f-veE-con for the connector at the 5’-terminus of the device and f-con-veS for the connector at the 3’-terminus of the device. **Each** connector ligates to the vector backbone in **level°1** independent of which device it is flanking. Therefore, the connector type (explained in chapter 2.2.7) is not indicated in the level°1 fusion site names, f-veE-con and f-con-veS.

Moreover connectors contain the fusion sites for level°2 to assemble the devices they are flanking in a pre-designed order (Figure 2C). Fusion site names on level°2 contain “con” and the number of the connector they fuse together: f-con1-2 is the fusion site to assemble con1 at the 3’-terminus with con2 at the 5’-terminus; f-con3-4 is the fusion site to assemble con3 with con4; etc. (for further description on connector design see chapter 2.2.7).

4.3.2.4 **Parts: order and designation**

Upper case characters, S or E, and/or numbers, 1 to ∞, in part names refer to their position and order when they are combined at level°1 (Figure 3). Consequently, a part/subpart always contains two of these upper case characters/numbers defining its 5’- and 3’-overhang in addition to the lower case character defining its module identity. The upper case characters, either S or E, refer to the start, or end of a part. S refers to the 5’-overhang of the first subpart and E to the 3’-overhang of the last subpart (Figure 3A). Thus, only the fusion site in an E-overhang is compatible with the fusion site in an S-overhang to assemble two parts of different identities.

If a part is further split into subparts, numbers refer to the order in which the subparts will be assembled. Even numbers refer to the 5’-overhang of a part while odd numbers refer to its 3’-

overhang. For example, if there are four subparts to be assembled into a b-part they are named: bS1, b23, b45 and b6E (Figure 3C and Figure 4B). This principle can be applied for every subpart (Figure 3D). Nomenclature and standardization of subparts is described in chapter 2.2.5.

A module only works in the hierarchical assembly scheme if it has a start S and an end E as well as a sequential numbering of its submodules.

The designation of the module is added after the module identity and position/order, like for example: aSE-T5 promoter, b23-His tag, b23-FLAG tag, b4E-esterase, b4E-lipase, etc.

Example: Nomenclature of parts with overhangs to assemble a transcriptional unit

The following parts are required for the assembly of a simple transcriptional unit consisting of promoter/operator, RBS/GOI and terminator for the expression of one gene, for instance the IPTG-inducible expression of an esterase:

aSE-T5/*lacO* (a, IPTG-inducible T5/*lacO* promoter; S, 5'-fusion site, anneals to E of the destination vector; E, 3'-fusion site, anneals to S of b; this part is flanked by the overhangs aS and aE);

bSE-RBS-esterase (b, esterase with RBS; S, 5'-fusion site, anneals to E of the promoter a; E, 3'-fusion site, anneals to S of the termintor c; this part is flanked by the overhangs bS and bE);

cSE-*rrnB*T1 (c, terminator *rrnB*T1; S, 5'-fusion site, anneals to E of esterase b; E, 3'-fusion site, anneals to S of the destination vector; this part is flanked by the overhangs cS and cE). Nucleotide sequences of 3'- and 5'-overhangs for these parts can be found in Table 3.

4.3.2.5 b-Subpart design on level°1

Parts can be further split into subparts to provide full flexibility in the level°1 assembly. Theoretically, the MoClo user can invent any subpart design to assemble a full part consisting of several subparts (Figure 3D). However, it is recommended to develop standardized designs to ensure the general applicability. This should guarantee standardization of the assembly reactions at the different levels and interchangeability of the parts in the plasmid collection.

Subparts of the b-parts are predefined as the RBS, the GOI and an optional tag (Figure 3C and Figure 4). For instance, it would be possible to screen an RBS library for the most efficient expression of a GOI. Otherwise, one could assess the influence of diverse N- or C-terminal tags on the expression of a GOI. **Subparts of the b-part have to be chosen according to the MoClo design described in this chapter** (Figure 3C, Figure 4 and Figure 5).

The RBS is the first subpart of the b-part. It is standardized to be either bS1 (Figure 4A and Figure 5A) or bS3 (Figure 4B and Figure 5A) depending on whether an N-terminal tag is introduced or not. Special attention must be payed to the modular design of the RBS, because the distance between the Shine Dalgarno sequence in the RBS and the start ATG of the GOI has to be maintained. Therefore, special fusion sites were designed: The fusion site f-b1-2, CAAT, of the bS1-RBS contains the nucleotides A and T of the start ATG (Figure 6A). The fusion site f-b3-4, AATG, of the bS3-RBS includes all nucleotides for the start ATG (Figure 6B). This design guarantees that the distance

between the Shine Dalgarno sequence and start ATG as well as the nucleotide identities are preserved.

If an N-terminal tag is introduced, the RBS to use is a bS1-part (Figure 6A, Table 4). To its 3'-end ligates the tag being a b23-part followed by the GOI being a b4E part. The final assembly order without connectors at level°1 would look like: aSE-promoter $_{[f\text{-}aE\text{-}bS]}$ bS1-RBS $_{[f\text{-}b1\text{-}2]}$ b23-tag $_{[f\text{-}b3\text{-}4]}$ b4E-GOI $_{[f\text{-}bE\text{-}cS]}$ cSE-terminator. If no N-terminal tag is introduced, the RBS has to ligate directly to the b4E-GOI. Therefore, part b23-tag is omitted and the RBS to use for this assembly is a bS3-part (Figure 4A, Figure 5A, Table 5). It directly has the f-b3-4 fusion site on its 3'-terminus to ligate to the b4E part: aSE-promoter $_{[f\text{-}aE\text{-}bS]}$ bS3-RBS $_{[f\text{-}b3\text{-}4]}$ b4E-GOI $_{[f\text{-}bE\text{-}cS]}$ cSE-terminator. If a C-terminal tag is desired the GOI has to be designed as a b45 part that assembles with the downstream tag being a b6E part (Figure 5B, Table 6): aSE-promoter $_{[f\text{-}aE\text{-}bS]}$ bS3-RBS $_{[f\text{-}b3\text{-}4]}$ b45-GOI $_{[f\text{-}b5\text{-}6]}$ b6E-tag $_{[f\text{-}bE\text{-}cS]}$ cSE-terminator. For an untagged version a spacer with a b6E-part identity can be introduced (Figure 5B, Table 6). The fusion site between the GOI and tag, f-b5-6 GGAT, contains a glycin codon, GGA. The last nucleotide of the f-b5-6 is a T that should be complemented with CG preceding the tag sequence. The GGT TCG sequence is the glycin-serine linker between the GOI and the tag. For the spacer the last nucleotide T, proceeding the GGA glycine codon of the f-b5-6, should be complemented with GA at the beginning of the spacer sequence to form a TGA stop codon. So the untagged version of a GOI with b45 part identity contains an additional glycine codon at its C-terminus. If the GOI is chosen to be a b45 part, a non-tagged, N-terminally tagged or C-terminally tagged versions can be constructed and tested (Figure 5B).

As described above, the fusion sites with b-identity, apart from f-aE-bS and f-bE-cS, contain coding sequences to facilitate modularity inside the transcribed sequence. Thus, the individual design of the b-subparts must take into account that certain nucleotides are already provided in the fusion sites. Sequence requirements for subparts are listed in Table 7.

4.3.2.6 **Removal of internal type IIS recognition sequences**

To assure the assembly of the parts on level 1 in the desired order, all DNA sequences have to be devoid of any internal recognition sites of those type IIS endonucleases that are used for their further assembly. At level°0, internal recognition sites can be removed by introducing silent mutations into the target DNA sequence with mutagenic primers (Figure 7). To this end, the original sequence is split up, so that the endonuclease recognition site to be removed is close to one terminus of one of the DNA fragments. In a PCR an outer primer introduces a silent mutation thereby "destroying" the type IIS recognition sequence in this DNA fragment. Also the other DNA fragments that are devoid of recognition sequences are PCR amplified. All primers in these PCRs furnish the DNA fragments with flanking nucleotides containing BsaI recognition sites (Figure 7). Upon BsaI restriction at level°0 fusion sites are generated that facilitate the scarless assembly of the (mutated) DNA fragments and the insertion of the resulting part into the level°0 destination vector pMC0. Nomenclature of PCR fragments to be assembled into one part consists of the name of the part, e.g. *lacI*, esterase, maltose

binding protein (MBP-) tag etc, followed by the large character S or E and consecutive numbering likewise to the nomenclature scheme for subparts (for details refer to chapter 2.2.5). If an esterase is assembled from three PCR fragments into one b-part at level°0, PCR fragments are termed esterase-S1, esterase-23 and esterase-4E. No certain module identity is ascribed to these fragments because their internal fusion sites, f-1-2 and f-3-4, are dependent on the sequence to be assembled and cannot be standardized.

4.3.2.7 **Connector design and devices**

As already described, the level°1 assembly of basic biological parts in a directed order generates a device. Two connectors participate in the level°1 assembly reaction and provide the fusion sites for ligation of this device into the level°1 destination vector (Figure 2B, Figure 8A and Figure 9). Moreover, connectors contain the BsaI recognition sequence to assemble several devices on level°2 (Figure 2C). Because BsaI, which is normally used at level°0, would cut inside the connector sequence, BsmBI is used instead. BsmBI sites flank the connector instead of BsaI to allow for ligation into BsaI-linearized and purified pMC0 (for details refer to chapter 3.2)

The connectors flanking the devices follow the same nomenclature system used for overhangs flanking parts (described in chapter 2.2.2). 5'-overhang of a part is either S or an even number. Likewise, the connector at the 5' terminus of a device (5'-connector, Figure 9A) is either a conS or a connector with an even number like con2, con4, etc (Figure 8B). The 3'-connector (Figure 9B) downstream of the device is either conE or a connector with an uneven number like con1, con3, etc (Figure 8B). Connectors to ligate several devices follow a consecutive numbering: con1 ligates to con2, con3 ligates with con4, etc.

A second character in the connector name describes to which part the connector ligates at the **level°1** assembly. If the device is a full transcriptional unit the 5'-connector ligates with its 3'-terminus to an a-part (f-con-aS on level°1) and is therefore termed either conSa (Figure 5A), con2a or con4a, etc. The 3'-connector ligates with its 5' terminus to a c-part (f-cE-con in level°1) resulting in either conEc, con1c or con3c, etc. (for the polycistronic assembly see chapter 2.2.8)

In the level°1 assembly reaction, the 5'-connector always ligates to the destination vector *via* its 5'-fusion site f-veE-con (Figure 8A: 5'-connector fusion site °1 in green; Table 8). The 3'-connector always ligates with the level°1 destination vector *via* its 3'-fusion site f-con-veS (Figure 8A: 3'-connector fusion site °1 in green; Table 9).

On the following level°2, the fusion site in the connector is responsible for the directional assembly of the devices to give a multigene construct. Only conS and conE ligate with the level°2 destination vector. Thus, conS and conE always flank the multigene module at the 5' and 3' terminus, respectively. In between, (theoretically) an indefinite number of devices can be assembled using connectors of consecutive numbering (Figure 8B). In the level°2 assembly, it does not matter which part the connector is fused to. A con1 always ligates with con2 with the fusion site f-con1-2

independent if it is a con1b or con1c and a con2a or con2b, respectively. Because the level°2 fusion site is independent of part identities, the part identity is not indicated in the level°2 fusion site name.

Devices should be termed according to the described nomenclature system for their unambiguous identification: For instance, in a multigene module that consists of three devices, the first device is termed devS1, the next downstream device dev23 and the following downstream device dev4E (Figure 8B). A pathway consisting of three devices representing full transcriptional units would be denoted as follows: conSa-devS1-con1c $_{[f\text{-}con1\text{-}2]}$ con2a-dev23-con3c $_{[f\text{-}con3\text{-}4]}$ con4a-dev4E-conEc. With this standardized nomenclature, a device plasmid library can be generated for the interchangeable and combinatorial use to construct multigene-plasmids in a single assembly step.

4.3.2.8 Connector design for polycystronic or bi-directional assemblies

In *E. coli*, the expression of several proteins from one promoter is possible *via* the polycistronic arrangement of genes in one operon. For the realization of such an arrangement in the hierarchical MoClo assembly approach, connectors that directly ligate to the b-part, which presents the GOI, are provided (Table 8 and Table 9). A polycistronic assembly on level°2 with three devices would be: conSa-devS1-con1b $_{[f\text{-}con1\text{-}2]}$ con2b-dev23-con3b $_{[f\text{-}con3\text{-}4]}$ con4b-dev4E-conEc. At level°1, con1b was ligated to the 3'-terminus of the b-part in devS1; con2b was ligated to the 5'-terminus of the second b-part in the dev23. On level°2, the fusion site f-con1-2 ligates the two devS1 and dev23 into a polycistronic arrangement.

As described in chapter 2.2.7, the part the connector was fused to on level°1 does not influence the fusion site at level°2 and is therefore not indicated in the level°2 fusion site name.

Moreover, the bi-directional arrangement of two transcriptional units can be realized by a special connector design. In a bi-directional arrangement two devices have an opposite transcriptional orientation to each other (Figure 10A). To achieve this, one of the two devices is fused to a connector on level°1 that flips its orientation (Figure 10B, Table 10 and Table 11). RV in connector names indicate that they assemble a device in a reverse orientation compared to standard devices at level°1 (Figure 10A). Thus, the fusion sites in RV-connectors are the reverse complement sequence of the fusion sites normally used on level°1 (Figure 10B). Level°2 fusion sites are the same for RV-connectors and standard connectors. Therefore, numbering of devices and connectors always starts with the first device after the conS connector independent of a uni- or bi-directional assembly. A bi-directional assembly of two devices would have the following structure: conScRV-devS1-con1aRV $_{[f\text{-}con1\text{-}2]}$ con2a-dev2E-conEc.

4.3.3 Destination plasmids and selection of positive clones

Selection of positive clones in the multi-step assembly process is based on a color reporter and an antibiotic marker.

Screening for successful integration of the DNA module into the destination vector is realized by color selection. In the presented design we describe a blue/white selection [7] with the destination vectors containing a *lacZ* transcriptional unit (Table 12). *lacZ* encodes the β-galactosidase, which can convert colorless 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) to blue 5-bromo-4-chloro-3-hydroxyindole (Figure 11). X-Gal has to be supplemented in plate media to screen the transformants. Clones that carry a functional *lacZ* gene, e.g. on the destination vector, are blue. Once *lacZ* is replaced by a DNA module during the assembly process, the clones appear white (Figure 11). The use of *lacZ* as the color selection marker implies that the cloning strain is devoid of a functional genomic copy of the *lacZ* gene.

The concept of color selection can also be realized with other genes than *lacZ*, e.g. with fluorescent proteins [6]. These have the advantage that no substrate has to be added to the transformation media for color development and that the cloning strain does not require any specific mutations. However, the clones must be irradiated, usually with UV light, to excite fluorescence.

For each level, the destination plasmid carries another antibiotic resistance marker (Figure 2). Chloramphenicol is used on level°0 for the selection of part plasmids (Figure 2A). These plasmids with chloramphenicol resistance can be used at level°1 where the device plasmid is selected on ampicillin (Figure 2B). At the last level°2, the destination vector carries a kanamycin resistance marker to select against the device plasmids used in this assembly reaction (Figure 2C).

Destination plasmids are abbreviated by pMC followed by the number referring to the level of assembly. Because several destination vectors exist for each level, a second number is added to distinguish between them. Names and features of the respective destination plasmids designed so far are summarized in Table 12. Level°0 part plasmids are only used for levelˆ1 assembly reactions, while level°1 device plasmids and level°2 multigene-plasmids are also used for protein expression. Therefore, a higher copy origin of replication (pMB1) was chosen for pMC0 than for pMC1 and pMC2 (ColE1) (Table 12) [8]. All destination vectors can be assembled *de novo* from DNA sequences in a single Golden Gate cloning reaction. A pMC vector consists of the following modules (Figure 12): veS1-ori, ve23-terminator of antibiotic resistance marker, ve4E-antibiotic resistance marker gene and promoter, inSE-lacZ or any other gene for color selection. Overhangs to generate those modules can be found in Table 13 for pMC0, Table 14 for pMC1 and Table 15 for pMC2. Each of these modules can be replaced in a new Golden Gate assembly reaction. For example, if a pMC should be generated with a p15a ori, a p15a sequence with flanking veS and ve1 overhangs is constructed. The new veS1-p15a module is combined with the other three existing modules ve23, ve4E and inSE in an assembly reaction.

4.4 Experimental and *status quo* of the JG-SynBio plasmid platform

Experimental conditions were evaluated in the Master Thesis of M. Koshanskaya [9]. In this manuscript, we only give an experimental guideline. In addition, we report on the *status quo* of the MoClo plasmid platform available in the Junior Group Synthetic Biology at acib. We achieved to assemble the basic biological parts presented in Table 16. We observed that assembly efficiency strongly depends on the parts' nucleotide sequences. For more detail refer to the Master Thesis of M. Koshanskaya [9].

4.4.1 Primer design

Flanking regions on basic biological parts must contain the restriction and fusion sites for the assembly process. These are usually created by PCR of a part template with primers containing the additional nucleotides (Figure 2A). An alternative is custom DNA systhesis of the part sequences including the required flanking regions. Flanking nucleotide sequences to create a specific part identity are provided in Table 3, Table 4, Table 5 and Table 6. For the flanking sequences on connectors see Table 8, Table 9, Table 10 and Table 11.

PCRs should be performed with a proofreading polymerase like Phusion® High-Fidelity DNA Polymerase (Life Technologies, Carlsbad, US). We recommend to use primers with homologous sequences to the part template corresponding to a melting temperatur of about 55-60 °C. This should result in a melting temperature of the full length primer on the genereated amplicon of about 70-75°C (melting temperatures were calculated with the DNA software SNAP gene, GSL Biotech, Chicago, US). A two step PCR can be applied with five cycles of 50-55°C annealing temperature and 35 cycles of 65-70°C annealing temperature.

4.4.2 Connector design

Because the connectors' sole role is the directional assembly of devices to create multigene constructs, its DNA sequence only has to meet the requirement of being neutral and orthogonal. These spacer DNA sequences can be created with the online tool R2O DNA (http://www.r2odna.com/) using standard settings (Table S1) with BpiI and BsmBI recognition sites as additional forbidden sequences (Table S2). Connectors are designed to have a total length of 160 bp (nucleotide length recommended by Dueber lab, University of California, Berkeley, US, personal communication). As the program does not give any sequence output for nucleotide length of 160 bp, 80 bp long sequences were created. In one run the program can only generate up to 8 sequences. Therefore, 15 sequences with a length of 80 bp were created in two runs (Table S3). Because the sequences originated from two separate runs, they were aligned in ClustalW to evaluate their sequence homologies (Table S4). Two of the generated sequences were used each for one connector as summarized in Table 17. Four

connectors were ordered as gBlocks (Integrated DNA Technologies, Coralville, IA) and cloned into pJET1.2 (Thermo Fisher Scientific, Waltham, MA).

Because of their internal BsaI recognition site, level°0 assemblies of connector plasmids apply purified BsaI-linearized pMC0 in the presence of BsmBI to generate the °0-fusion sites on the connector (Figure 9). Assembly of connectors into the level°0 destination plasmids failed until now for unknown reasons. For further elaboration on experimental details refer to the Master Thesis of M. Koshanskaya [9].

4.4.3 **Design of destination plasmids**

For explanation on the principle of plasmid selection and blue/white screening refer to chapter 2.3. Here we elaborate on the construction of the destination plasmids listed in Table 12.

Our first designed vector contained only a portion of *lacZ*, which was used for α-complementation [7]. The cloning strain, e.g. TOP10F', contained a *lacZ*Δ*M15* mutation that lacks the N-terminal sequence of *lacZ* responsible for oligomerization of the enzyme. Therefore, a functional β-galactosidase is only formed when complemented with the lacking N-terminal sequence, *lacZα*, encoded on the transformed plasmid. For pMC0.1 we used the P_{em7} promoter for the constitutive expression of *lacZα* (Table 12). However, this vector construct showed only poor color development even in the presence of isopropyl-β-D-1-thiogalactopyranoside (IPTG) to induce the genomic *lacZ* fragment, *lacZ*Δ*M15,* which is under the control of the endogenous P_{lac} promoter (data not shown). We reasoned this was due to low expression levels of *lacZα* with the P_{em7} as well as of the genomic part of *lacZ, lacZ*Δ*M15*. Therefore, we decided to use the whole *lacZ* gene, instead of *lacZα*, under control of a strong promoter. We did assemblies of destination plasmids with P_{tacII} either with a *lac* operator, *lacO*, or without. As the destination plasmids do not carry the *lacI* repressor gene, only expression of the endogenous *lacI* copy attenuates *lacZ* expression. Because the endogenous *lacI* expression levels are too low for complete repression of *lacZ* encoded on high copy destination plasmids, *lacZ* expression is only diminished in presence of the repressor binding region, *lacO*.

Only the assemblies with promoters including *lacO* were successful, while assemblies without *lacO* lacked the promoter sequence at all. This observation suggests that high expression levels of *lacZ* are detrimental to *E. coli*.

P_{tacII} with *lacO* was used with the destination vectors of all levels (Table 12).

4.4.4 **RBS design**

As described in chapter 2.2.5 the last two nucleotides of the RBSs are included in their 3'-terminal fusion sites in the level°1 assembly. Therefore, all RBS sequences used in this study were changed to contain a cytosine and adenine at their 3' end. A list of useable RBSs changed according to the assembly strategy can be found in Table 18. Because the RBS sequence alone is too short to form an

autonomous part, a 30 bp long spacer sequence was included upstream of the RBS before the promoter-operator region (Nucleotide sequence 11 to 14). Spacer sequences were created with the online tool R2O DNA (http://www.r2odna.com/) as described in chapter 3.2.

4.4.5 **Assembly reaction conditions**

Restriction enzymes can be purchased as FastDigest versions from Thermo Scientific (Waltham, MA) or as High Fidelity versions from New England Biolabs (Ipswich, MA). T4 or T7 DNA ligase can be purchased from New England Biolabs (Ipswich, MA). PCRs were purified with the Wizard SV Gel and PCR Clean-Up Kit (Promega, Madison, WI) according to the manufacturer's protocol. Chemicals were purchased from Carl Roth (Karlsruhe, Germany) unless indicated otherwise.

10 mM DTT should be added to the assembly reactions with BpiI and BsmBI to enhance restriction efficiency.

BsaI was used in level°0 and level°2 assembly reactions and BpiI in the level°1 assembly reaction. Concerning assemblies of pMC destination vectors, BpiI was used for pMC0 and pMC2, while BsmBI was used for pMC1.The assembly reactions of the destination vectors included 100 ng of each PCR-amplified part, 10 U restriction enzyme, 25 mM ATP, 10 U T4 or T7 DNA ligase in a total volume of 20 µl of the appropriate reaction buffer.

For the assembly of plasmids containing basic biological parts we recommend to use the pMC0.3 on level°0, because it showed best color development of all the pMC0 vectors presented in Table 12. Our reaction contained about 100 ng of pMC0.3, 500 ng of PCR-amplified biological part, 10 U BsaI, 25 mM ATP and 10 U T7 DNA ligase in a total reaction volume of 20 µl of the reaction buffer of the used restriction enzyme.

For the assembly of connectors in the level°0 destination vector we commend to use BsmBI. The level°0 plasmid backbone, pMC0, was obtained by linearization of pMC0.1 with BsaI and agarose gel purification of the backbone fragment. Connectors can be directly applied as gBlocks (Integrated DNA Technologies, Coralville, IA) or after subcloning into pJET1.2 (Thermo Scientific, Waltham, MA). Since gBlocks contain a mixture of sequences the latter procedure is preferred. For unknown reasons connector assembly reactions were unsuccessful so far. For further elaboration on the experimental details refer to the Master Thesis of M. Koshanskaya [9].

No level°1 and level°2 assembly reactions have been tested yet.

We recommend to use a cyclic restriction-ligation protocol. It applies 24 cycles of restriction at 37°C for 2 min and ligation at 22°C with T4 or 25°C with T7 DNA ligase for 5 min. As an alternative, assembly reactions can be incubated at 37°C for 2 h for restriction and ligation each. For higher transformation rates we always heat inactivated the samples at 65°C for 10 min.

After transformation *E. coli* cells were plated on LB-agar containing 40 mg/mL X-Gal for color development and the antibiotic of the corresponding assembly level.

4.5 **Conclusions**

A MoClo strategy was developed for the directional assembly of DNA modules to form multigene constructs. Our design focused on expression constructs for the use in *E. coli*. However, the concept is applicable to any expression host of choice.

The presented MoClo approach comprises three standardized assembly levels each giving rise to a specified module: parts, devices and multigene construct. Level°0 generates plasmids containing basic biological parts. With standardized fusion sites specific for a certain part, they are assembled into devices in a directional manner on level°1. A device is a multipartite module that can be a transcriptional unit consisting of the three parts: promoter, gene of interest and terminator. Devices generated at level°1 are flanked by connectors that direct their order of assembly on the level°2. In this way several devices are used to create a multigene construct.

Destination plasmids provide the backbone for the modules of each assembly step. They also consist of standardized modules that are assembled in a Golden Gate cloning reaction. This facilitates the straight forward assembly of user-customized destination vectors.

Our intent was to create a plasmid platform of basic biological parts that can be universally used for the assembly of multipartite DNA constructs. Nomenclature of part plasmids in this library was standardized to guarantee that the specified identity of the parts remains preserved. Moreover, the standardized nomenclature system does not only include parts but also (i) the 5’ and 3’ flanking regions (overhangs), that have to be provided on parts for their further assembly, (ii) the fusion sites generated upon restriction and (iii) the connectors. The nomenclature of these DNA fragments directly refers to their designation in the MoClo strategy.

The MoClo concept presented in this study provides a universally applicable cloning tool on the basis of a plasmid platform of basic biological parts.

4.6 List of abbreviations

For module abbreviations see Table 1. MoClo: Modular Cloning

4.7 Figures and illustrations

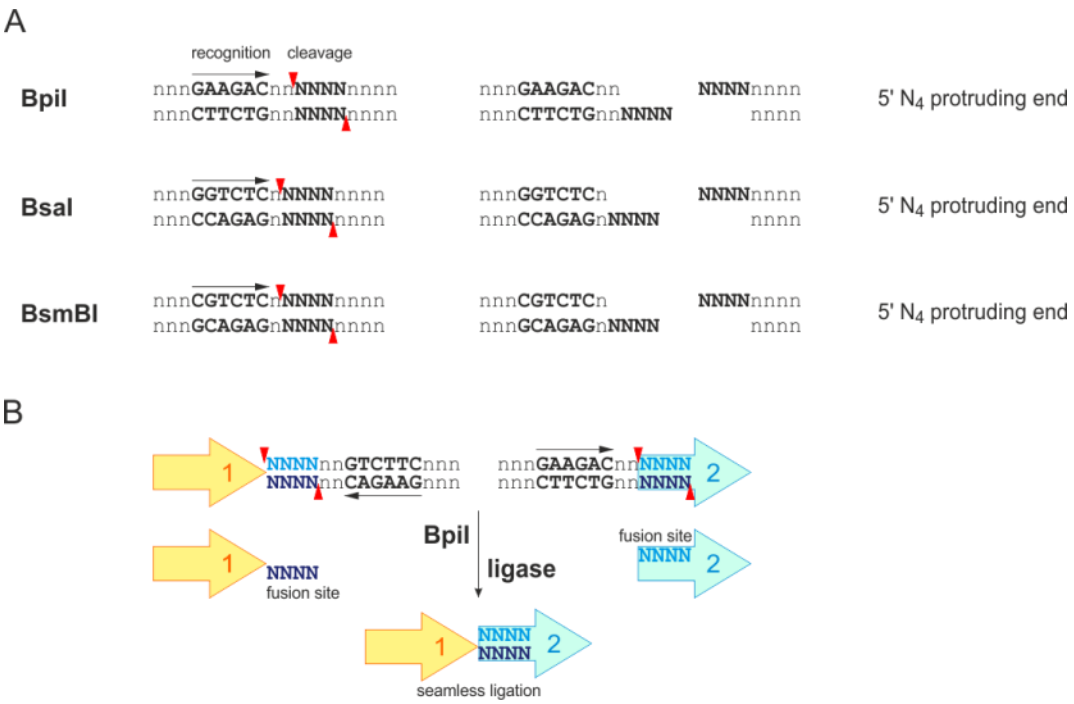


Figure 1: Recognition sequences and cleavage sites of type IIS endonucleases used in this study

Black arrows indicate the direction of enzyme cleavage and red arrows the cleavage site. (A) Type IIS restriction enzymes used in this study cutting distal to their recognition sequence to produce four nucleotide long fusion sites (NNNN). (B) Golden Gate reaction with BpiI and ligase for the seamless ligation of two DNA fragments, 1 and 2 in a yellow and blue arrow, respectively.

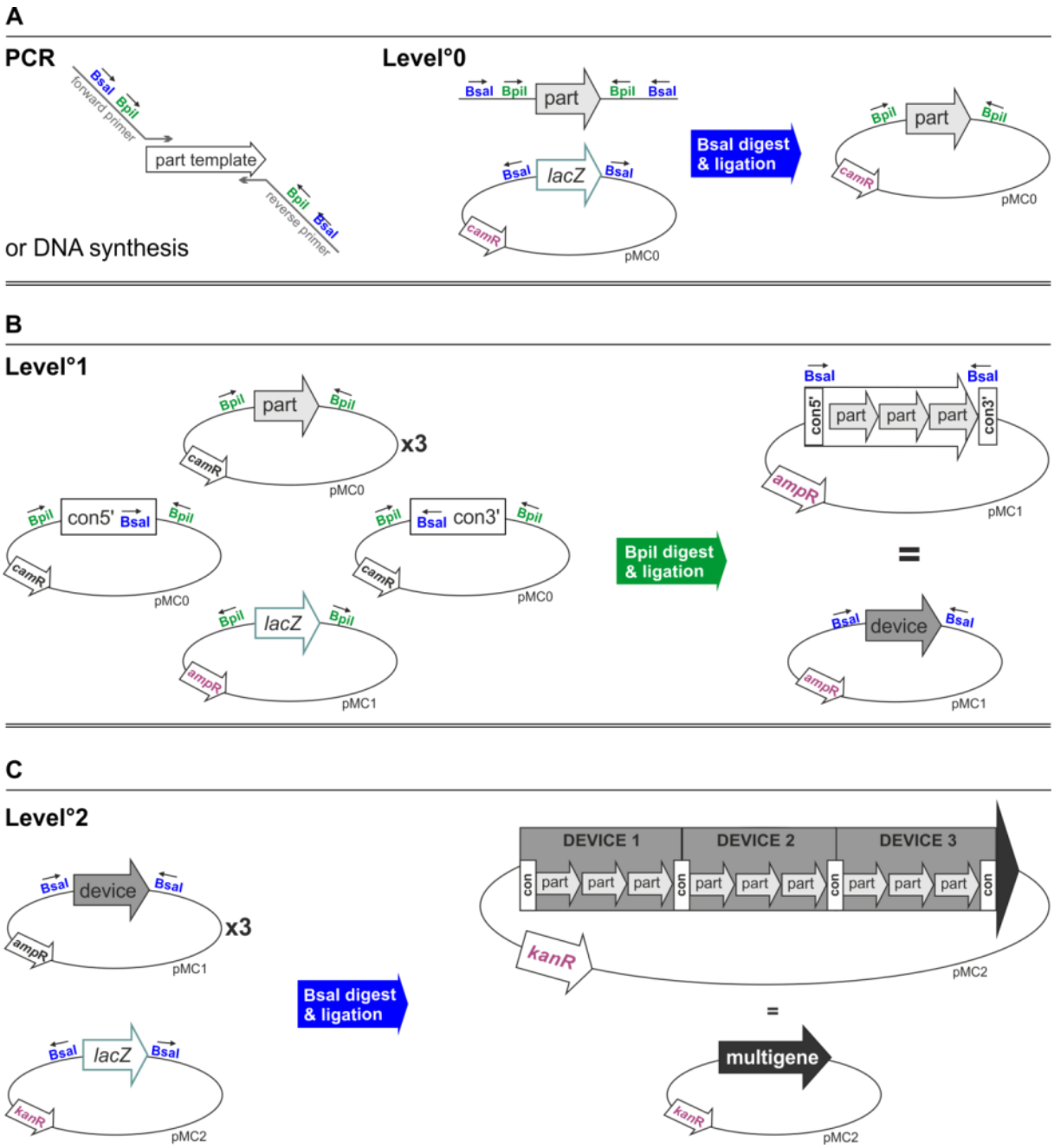


Figure 2: The three levels of the MoClo assembly strategy presented in this study

(A) PCR with primers that contain the overhangs to furnish a PCR template with recognition and cleavage sites of BpiI and BsaI. This amplicon, that can also be obtained directly by custom DNA synthesis, is assembled with the level°0 destination vector (pMC0) to a part plasmid upon BsaI restriction and ligation. Part plasmids are selected by chloramphenicol restistance (camR). (B) On level 1, connectors (con) and part plasmids are combined with a level°1 destination vector (pMC1). Upon BpiI restriction and ligation a device plasmid is assembled that confers ampicillin resistance (ampR). (C) On level°2, device plasmids are combined with a level°2 destination vector (pMC2), digested with BsaI and ligated. This assembly yields a multigene plasmid conferring kanamycin resistence (kanR).

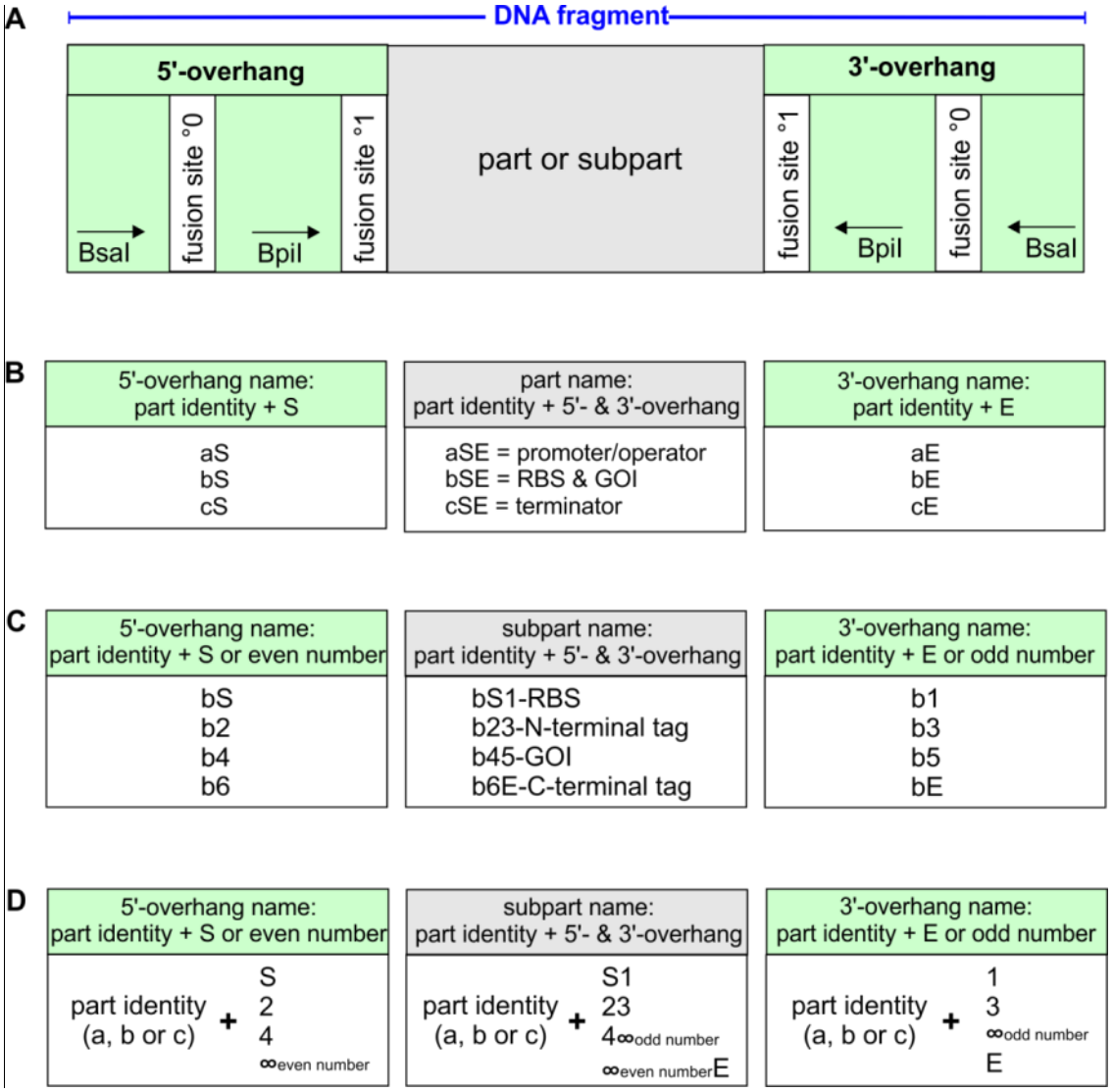


Figure 3: Representation of overhangs, parts and subparts for which a MoClo nomenclature was developed

(A) Schematic representation of the part design. Parts are flanked by overhangs containing type IIS endonuclease recognition sites,BsaI and BpiI, that generate fusion sites for the level°0 (fusion site°0) and level°1 (fusion site°1), respectively, upon restriction. Black arrows indicate the direction of restriction distal to the endonuclease recognition sequence. (B) Nomenclature of parts (for details refer to chapter 2.2.4.) (C) Nomenclature of b subparts (for details refer to chapter 2.2.5) (D) General nomenclature pattern for subparts.

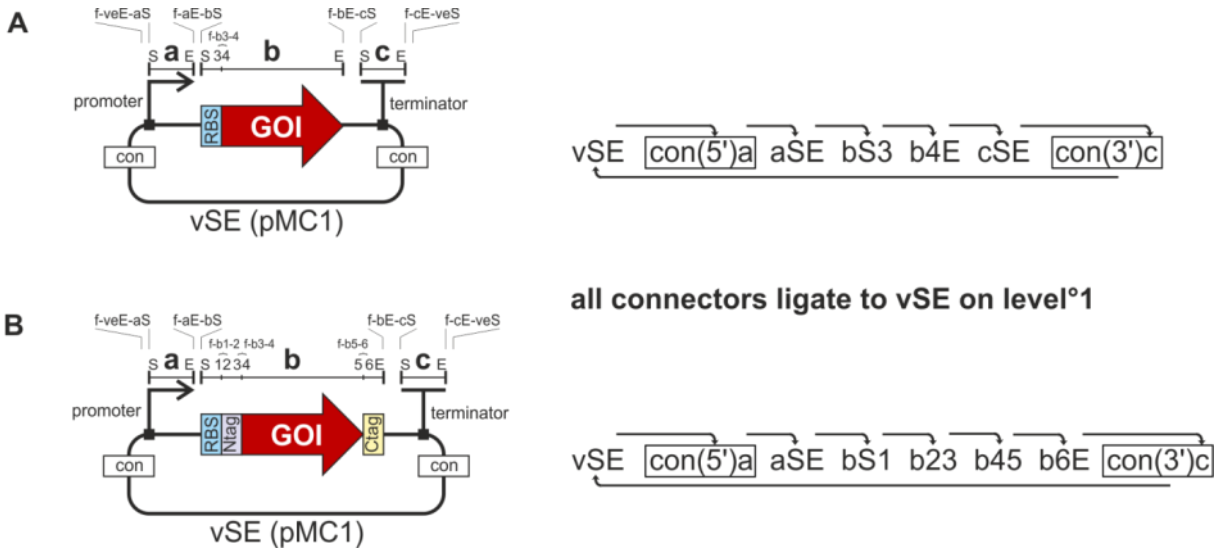


Figure 4: Schematic representation of the nomenclature for parts/subparts and fusion sites at level°1

RBS: ribosomal binding site, Ntag: N-terminal tag, GOI: gene of interest, Ctag: C-terminal tag. For description of nomenclature refer to chapter 2.2.

Figure 5: Modular architecture of b-subparts

(A) Design of the gene of interest (GOI) as a b4E part provides the oportunity for N-terminally tagging. (B) Design of the GOI as a b45 part provides the the oportunity for N-terminally as well as C-terminal tagging.

Figure 6: Subpart design of the ribosomal binding site (RBS)

(A) For N-terminal tagging the fusion site f-b1-2, CAAT, between bS1-RBS and the b23-tag contains the nucleotides A and T of the tag's start ATG. (B) If no N-terminal tag is introduced, the fusion site f-b3-4 provides the start ATG for the gene of interest (GOI), which can either be a b4E or a b45 part. Because the RBS sequence alone is too short to form an autonomous part a biologically neutral orthogonal DNA spacer sequence (spacer I) is provided (for details refer to chapter 3.4).

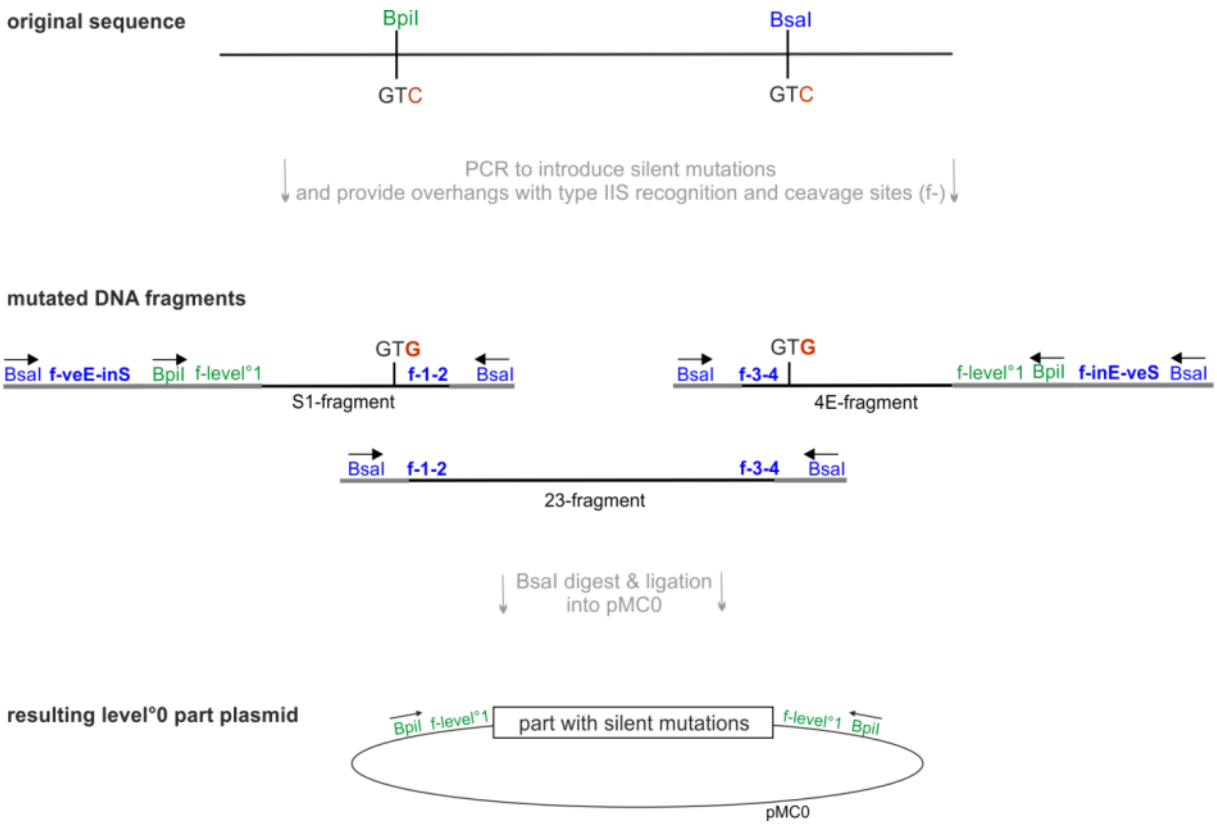


Figure 7: Removal of internal restriction sites at level°0

A valin codon, GTC, inside the recognition sites of BpiI (in green) and BsaI (in blue) is shown. Primers provide 5' and 3' extensions (overhangs, in grey) to furnish fragments with enzyme recognition sequences and fusion sites for the level°0 (in blue) and level°1 (in green) assembly reactions. Mutagenic primers introduce the silent mutation C>G (in brown), thereby removing the internal recognition sequences. PCR fragments, S1-, 23- and 4E-fragment, are used to generate a part/subpart plasmid on the level°0 using BsaI and ligase. Direction of type IIS restriction is indicated with a black arrow.

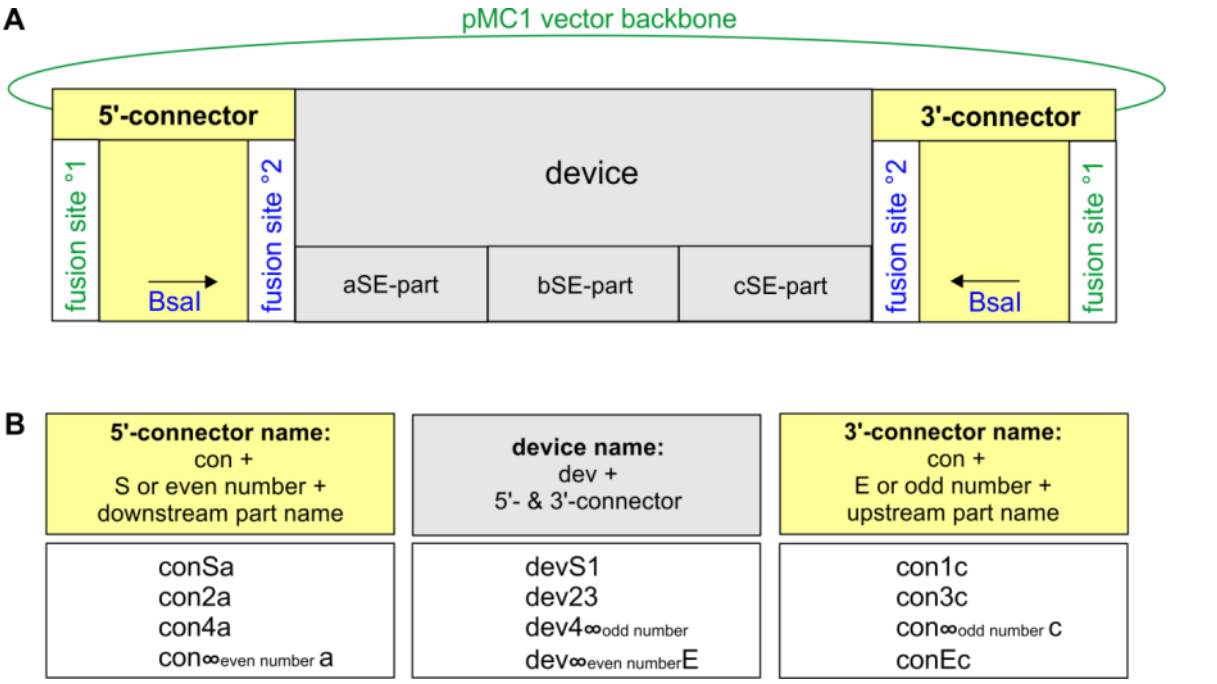


Figure 8: Schematic representation of connector and devices for which a MoClo nomenclature was developed

(A) Devices are flanked by connectors that provided the fusion sites on level°1 (in green) for the assembly of several parts into the destination vector pMC1 (in green). Connectors contain BsaI recognition sequences (in blue, direction indicated with a black arrow) that generate fusion sites (in blue) for the assembly of several devices on level°2. (B) Schematic representation of nomenclature for connectors and devices for the assembly of several transcritional units (for details refer to chapter 2.2.7).

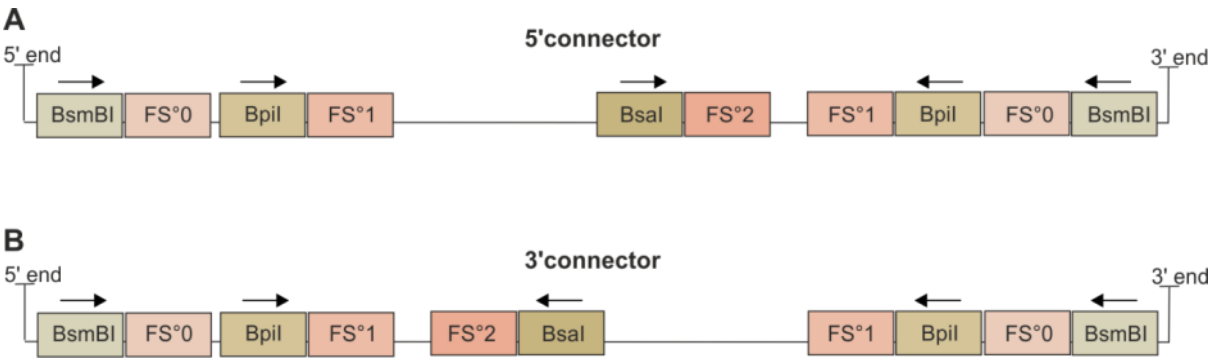


Figure 9: Design of connectors

Endonuclease type IIS, BsmBI, BpiI and BsaI, recognition and cut sites in 5'-connectors (A) and 3'-connectors (B) to provide the fusion sites (FS) on the three levels of assembly, °0, °1 and °2.

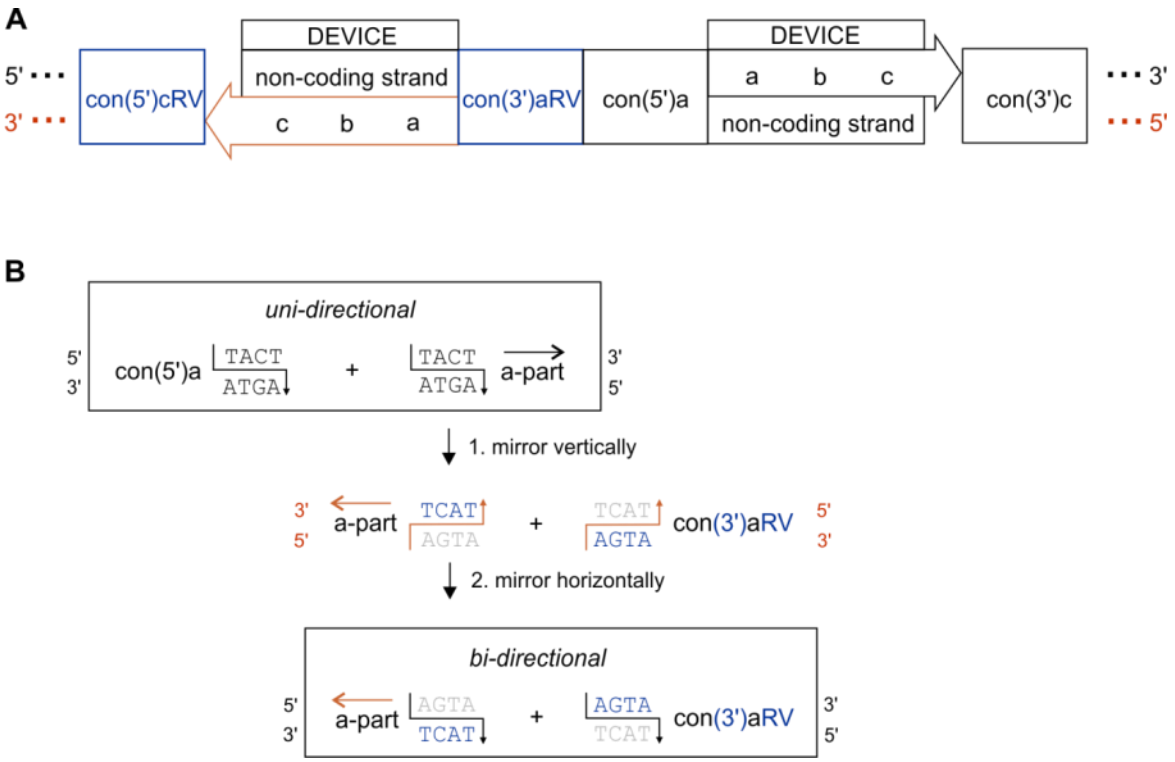


Figure 10: Level°1 connector design for the bi-directional assembly approach

(A) Schematic presentation of the bi-directional assembly between two devices. RV-connectors, con(3')aRV and con(5')cRV, flip the orientation of a device (a, b, c, always on the coding strand) in the level°1 assembly. The con(3')aRV connector ligates with the con(5')a connector at level°2. (B) Descriptive presentation, how con(3')aRV connectors flip the device orientation on level°1. con(3')aRV provides a fusion site that is the reverse complement sequence of standard con(5')a connectors used in the uni-directional assembly. The same principle can be applied for all connector types.

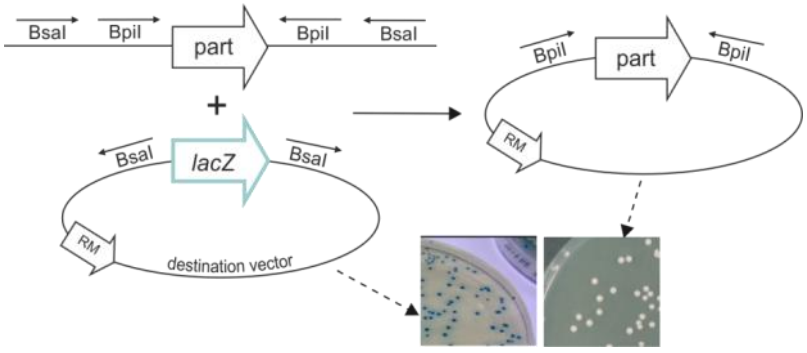


Figure 11: Color selection in the level°0 assembly

A transcriptional unit for the expression of the β-galactosidase gene *lacZ*, confers a blue colour to negative transformants. Upon replacment of *lacZ* by a part in the level°1 assembly positive clones do not show color formation. RM refers to resistence marker.

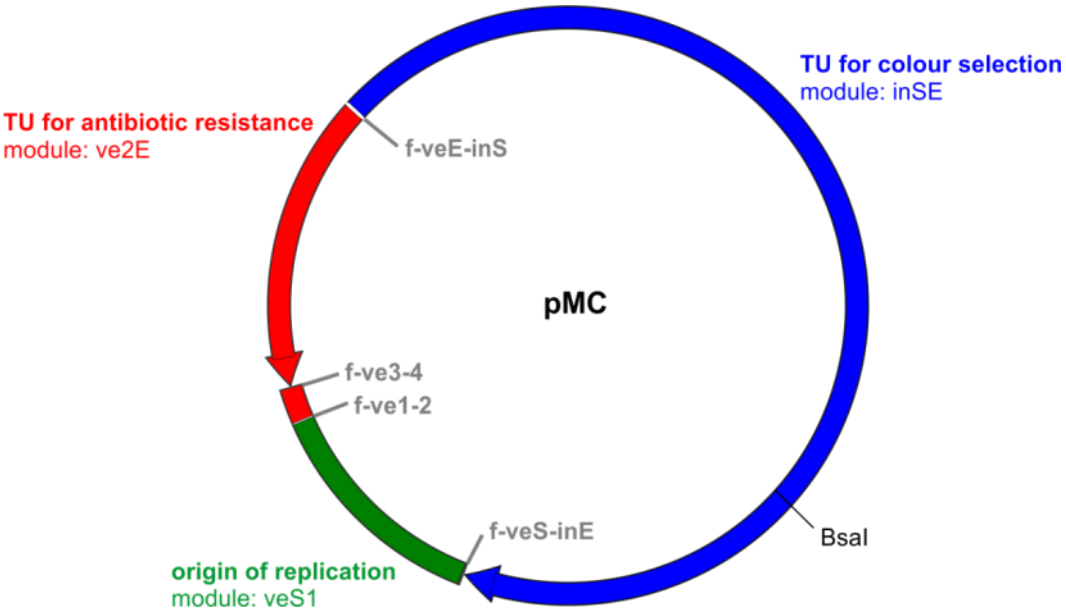


Figure 12: Modular architecture of pMC destination plasmids

All modules and the applied fusion sites in the assembly process of pMC plasmids are shown. TU refers to transcriptional unit.

4.8 **Tables**

Table 1: Abbreviations for modules and DNA sequences

RBS refers to ribosomal binding site and GOI to gene of interest.

abbreviation	description	synthesis	used on level
in	transcriptional unit for color selection	PCR, DNA synthesis	pMC assemblies
ve	destination vector after type IIS restriction	pMC assemblies	all
a	promoter-operator part	level°0	°1
b	RBS-GOI part	level°0	°1
c	terminator part	level°0	°1
dev	device	level°1	°2
con(5')	connector at 5' terminus of device	level°0	°1 and °2
con(3')	connector at 3' terminus of device	level°0	°1 and °2

Table 2: Fusion sites used in this study

Fusion sites (f-) are generated upon restriction with type IIS restriction endonucleases

sequence	pMC vectors	level°0	level°1	level°2
GAAC	f-veE-inS	f-veE-inS	f-veE-con	f-veE-conS
TCGC	f-inE-veS	f-inE-veS	f-con-veS	f-conE-veS
TACT	f-ve1-2		f-con-aS	f-con1-2
GCTT	f-ve3-4		f-aE-bS f-con-bS	f-con3-4
CAAT			f-b1-2	
AATG			f-b3-4	
GGAT			f-b5-6	
GGAG			f-bE-cS f-bE-con	
GCCT			f-con-cE	
AGGC			f-cE-con	
AGTA			f-aS-con	

Table 3: Overhangs on DNA sequences to form part plasmids at level°0 that are assembled to a transcriptional unit at level°1

BsaI recognition sites are coloured blue and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at

level°1 are bold and underlined. RBS refers to ribosomal binding site and GOI to gene of interest. Four nucleotide long fusion sites used at level°1 and fusion site nomenclature are shown in brown.

overhang with fusion sites at 5'			overhang with fusion sites at 3'	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
catgcaGGTCTCtGAACtaGAAGACatTACT TACT (f-con-aS)	aS	promoter & operator	aE	GCTTatGTCTTCagTCGCtGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatGCTT GCTT (f-aE-bS)	bS	RBS & GOI	bE	GGAGatGTCTTCagTCGCgGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatGGAG GGAG (f-bE-cS)	cS	terminator	cE	AGGCatGTCTTCagTCGCgGAGACCtgttca AGGC (f-cE-con)

Table 4: Overhangs of b-subparts for an N-terminal tag

The fusion site f-b1-2, CAAT between bS1-RBS (ribosomal binding site) and the b23-tag, contains the nucleotides A and T (yellow background), of the tag's start ATG. The only nucleotide G (in red) of the start ATG has to be provided with the tag sequence. The fusion site f-b3-4, AATG between b23-tag and b4E-GOI (gene of interest) provides an ATG (yellow background). BsaI recognition sites are coloured blue and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined. Four nucleotide long fusion sites used at level°1 and fusion site nomenclature are shown in brown. w/o refers to without.

overhang with fusion sites at 5'			overhang with fusion sites at 3'	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
catgcaGGTCTCtGAACtaGAAGACatGCTT GCTT (f-aE-bS)	bS	RBS	**b1**	CAATatGTCTTCagTCGCtGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatCAAT CAAT (f-b1-2)	b2	G+ tag$_{w/o\ ATG}$	b3	AATGatGTCTTCagTCGCtGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatAATG AATG (f-b3-4)	b4	GOI$_{w/o\ ATG}$	bE	GGAGatGTCTTCagTCGCgGAGACCtgttca

Table 5: Overhangs of b-subparts without any tag

The fusion site f-b3-4, AATG between bS3-RBS (ribosomal binding site) and b4E-GOI (gene of interest) provides the start ATG (orange background). BsaI recognition sites are coloured blue and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined. Four nucleotide long fusion site used at level°1 and fusion sites nomenclature are shown in brown. w/o refers to without.

overhang with fusion sites at 5'			overhang with fusion sites at 3'	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
catgcaGGTCTCtGAACtaGAAGACatGCTT GCTT (f-aE-bS)	bS	RBS	**b3**	AATGatGTCTTCagTCGCtGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatAATG AATG (f-b3-4)	b4	GOI$_{w/o\ ATG}$	bE	GGAGatGTCTTCagTCGCgGAGACCtgttca

Table 6: Overhangs of b-subparts with a C-terminal tag

The fusion site f-b5-6, GGAT between b45-GOI (gene of interest) and b6E-tag provides a glycin codon and (pink background) and the nucleotide T (in red) to form a serin codon with the nucleotide CG (in

red) in the tag sequence. If a spacer is introduced instead of a tag the nucleotide T (in red) is complemented with a GA to form a stop codon after the GGA glycin codon. BsaI recognition sites are coloured blue and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined. Four nucleotide long fusion site used at level°1 and fusion site nomenclature are shown in brown. w/o refers to without.

overhang with fusion sites at 5'			**overhang with fusion sites at 3'**	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
catgcaGGTCTCtGAACtaGAAGACatGCTT **GCTT (f-aE-bS)**	bS	RBS	**b3**	AATGatGTCTTCagTCGCtGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatAATG **AATG (f-b3-4)**	b4	GOI$_{w/o\ ATG}$	b5	GGATatGTCTTCagTCGCtGAGACCtgttca
catgcaGGTCTCtGAACtaGAAGACatGGAT **GGAT (f-b5-6)**	b6	CG+ tag or GA+ spacer	bE	GGAGatGTCTTCagTCGCgGAGACCtgttca

Table 7: Requirements on sequence design for b-subparts

subpart identity	subpart description	requirement
b23	N-terminal tag	only G of start ATG
b4E	gene of interest	no start ATG
b45	gene of interest	no start ATG & no stop codon
b6E	C-terminal tag	additional CG at 5' terminus to form Gly-Ser linker with fusion site (GGA TCG)
b6E	spacer	additional GA at 5' terminus to give a stop codon with the T in the fusion site (GGA TGA)

Table 8: Overhangs on 5'-connectors

5'-connectors include conS, con2, con4, con6, etc. Only 3'-overhang changes for 5'-connectors because the 5'-fusion site always ligates with the destination vector pMC1 in the level°1 assembly. BsmBI recognition sites are coloured orange and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined. 5'-connectors designed so far are conSa, con2a and con4b (Table 17).

overhang with fusion sites at 5'		**overhang with fusion sites at 3'**
5' → 3' sequence	**module**	**5' → 3' sequence**
catgcaCGTCTCtGAACtaGAAGACatGAAC	con(5')a	TACTatGTCTTCagTCGCtGAGACGtgttca
catgcaCGTCTCtGAACtaGAAGACatGAAC	con(5')b	GCTTatGTCTTCagTCGCtGAGACGtgttca

Table 9: Overhangs on 3'-connectors

BsmBI recognition sites are coloured orange and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined. 3'-connectors designed so far are conEc and con3b (Table 17).

overhang with fusion sites at 5'		overhang with fusion sites at 3'
5' → 3' sequence	**module**	**5' → 3' sequence**
catgcaCGTCTCtGAACtaGAAGACat**AGGC**	con(3')c	**TCGC**atGTCTTCagTCGCtGAGACGtgttca
catgcaCGTCTCtGAACtaGAAGACat**GGAG**	con(3')b	**TCGC**atGTCTTCagTCGCtGAGACGtgttca

Table 10: Overhangs on 5'-connectors for the bidirectional approach (RV-connectors)

RV-5'-connectors include conS, con2, con4, con6, etc. Only 3'-overhang changes for RV-5'-connectors because the 3'-fusion site always ligates with the destination vector pMC1 in the level°1 assembly. BsmBI recognition sites are coloured orange and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined.

overhang with fusion sites at 5'		overhang with fusion sites at 3'
5' → 3' sequence	**module**	**5' → 3' sequence**
catgcaCGTCTCtGAACtaGAAGACat**GAAC**	con(5')cRV	**GCCT**atGTCTTCagTCGCtGAGACGtgttca

Table 11: Overhangs on 3'-connectors for a bidirectional assembly (RV-connectors)

RV-3'-connectors include conE, con1, con3, con5, etc. Only 5'-overhang changes for 3'-connectors because the 3'-fusion site always ligates with the destination vector pMC1 in the level°1 assembly. BsmBI recognition sites are coloured orange and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Fusion sites generated at level°0 are underlined and fusion sites generated at level°1 are bold and underlined.

overhang with fusion sites at 5'		overhang with fusion sites at 3'
5' → 3' sequence	**module**	**5' → 3' sequence**
catgcaCGTCTCtGAACtaGAAGACat**AGTA**	con(3')aRV	**TCGC**atGTCTTCagTCGCtGAGACGtgttca

Table 12: Features on pMC destination plasmids

For sequence information refer to the Master Thesis of M. Koshanskaya [9].

pMC	promoter and gene for color selection	ori	resistance	restriction enzyme used for pMC assembly	use with higher order pMC	in libary
0.1	P_{em7} - lacZ**α**	pMB1	chloramphenicol	BpiI	1.x	yes
0.2	P_{em7} - lacZ	pMB1	chloramphenicol	BpiI	1.x	yes
0.3	$P_{tacII/lacO}$ - lacZ	pMB1	chloramphenicol	BpiI	1.x	yes
1.1	$P_{tacII/lacO}$ - lacZ	ColE1	ampicillin	BsmBI	2.1	yes
1.2	$P_{tacII/lacO}$ - lacZ	ColE1	kanamycin	BsmBI	2.2	no
2.1	$P_{tacII/lacO}$ - lacZ	ColE1	kanamycin	BpiI	-	yes
2.2	$P_{tacII/lacO}$ - lacZ	ColE1	ampicillin	BpiI	-	no

Table 13: Overhangs for the assembly of level°0 destination plasmids pMC0

BsaI recognition sites are coloured blue and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Generated fusion sites are bold and underlined. RM refers to resistence marker gene and P_{RM} to the promoter of the resistence marker gene.

overhang with fusion sites at 5'			overhang with fusion sites at 3'	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
ctGAAGACta**TCGC**	veS	ori	ve1	**TACT**taGTCTTCag
ctGAAGACta**TACT**	ve2	[RM terminator	ve3	**GCTT**taGTCTTCag
ctGAAGACta**GCTT**	ve4	< RM gene < $P_{RM}]_{rv}$	veE	**GAAC**taGTCTTCag
ctGAAGACta**GAAC**tGAGACC	inS	[TU for color selection]	inE	GGTCTCg**TCGC**atGTCTTCga

Table 14: Overhangs for the assembly of level°1 destination plasmids pMC1

BsmBI recognition sites are coloured orange and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Generated fusion sites are bold and underlined. RM refers to resistence marker gene and PRM to the promoter of the resistence marker gene.

overhang with fusion sites at 5'			overhang with fusion sites at 3'	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
catgcaCGTCTCa**TCGC**	veS	ori	ve1	**TACT**aGAGACGtgttca
catgcaCGTCTCa**TACT**	ve2	[RM terminator	ve3	**GCTT**aGAGACGtgttca
catgcaCGTCTCa**GCTT**	ve4	< RM gene < $P_{RM}]_{rv}$	veE	**GAAC**aGAGACGtgttca
catgcaCGTCTCa**GAAC**ctGTCTTC	inS	[TU for color selection]	inE	GAAGACta**TCGC**aGAGACGtgttca

Table 15: Overhangs for the assembly of level°2 destination plasmids pMC2

BsaI recognition sites are coloured blue and BpiI recognition sites green. Dark colours refer to restriction in 5' to 3' direction while light colors refer to restriction in 3' to 5' direction distal from the recognition sites. Generated fusion sites are bold and underlined. RM refers to resistence marker gene and P_{RM} to the promoter of the resistence marker gene.

overhang with fusion sites at 5'			**overhang with fusion sites at 3'**	
5' → 3' sequence	**name**	**module**	**name**	**5' → 3' sequence**
ctGAAGACta**TCGC**	veS	ori	ve1	**TACT**taGTCTTCag
ctGAAGACta**TACT**	ve2	[RM terminator	ve3	**GCTT**taGTCTTCag
catgcaGAAGACta**GCTT**	ve4	< RM gene < $P_{RM}]_{rv}$	veE	**GAAC**taGTCTTCag
ctGAAGACta**GAAC**tGAGACC	inS	[TU for color selection]	inE	GGTCTCg**TCGC**atGTCTTCga

Table 16: Basic biological parts in the MoClo plasmid platform described in this study

Parts do not include genes of interest. Parts marked with a star could no be assembled. For DNA sequences of listed parts refer to Nucleotide sequence 9 to Nucleotide sequence 19 in the Supplemental. RBS: ribosomal binding site, GST: glutathion-S-transferase, MBP: maltose binding protein.

part identity	**part sequence**
aSE	araC-P_{BAD}
aSE	lacI-P_{T5}-*lacO**
bS1	RBS III
bS1	RBS IV
bS3	RBS III
bS3	RBS IV*
b23	FLAG
b23	GST
b23	MBP
cSE	rrnB T1
cSE	T7

Table 17: Connectors designed in this study

For DNA sequences of the listed connectors refer to Nucleotide sequence 1 to Nucleotide sequence 8 in the Supplemental. Sequence numbers refer to Table S3. 5' FS°1: fusion site generated at level°1 at the 5' terminus, 3' FS°1: fusion site generated at level°1 at the 3' terminus, FS°2: fusion site generated at level°2, polycis.: for the use in a polycistronic arrangement of genes.

5'-connector	**sequence number**	***status quo***	**5' FS°1**	**3' FS°1**	**FS°2**
conSa	1 + 3	gBlock cloned into pJET1.2	GAAC	TACT	GAAC
conScRV	1 + 3	designed	GAAC	GCCT	GAAC
con2a	8 + 9	gBlock cloned into pJET1.2	GAAC	TACT	TACT
con2b (polycis.)	8 + 9	proposed design	GAAC	GCTT	TACT
con4a	12 + 14	proposed design	GAAC	TACT	GCTT
con4b (polycis.)	12 + 14	designed	GAAC	GCTT	GCTT
3'-connector	**sequence number**	***status quo***	**5' FS°1**	**3' FS°1**	**FS°2**
con1c	6 + 15	gBlock cloned into pJET1.2	AGGC	TCGC	TACT
con1aRV	6 + 15	designed	AGTA	TCGC	TACT
con3c	10 + 11	proposed design	AGGC	TCGC	GCTT
con3b (polycis.)	10 + 11	designed	GGAG	TCGC	GCTT
conEc	4 + 5	gBlock cloned into pJET1.2	AGGC	TCGC	TCGC

Table 18: RBS sequences for the MoClo assembly approach described in this study

Last two codons of the original RBS (source) were changed to CA

given name	**sequence**	**source**
RBS I	CACAGGAAACAGGAATTCA	unknown, contains EcoRI
RBS II	TTAAGGAGGTAAAAACA	BioBrick
RBS III	ATTAAAGAGGAGAAATTAACA	pMS570, pQE80L
RBS IV*	AAGATAAGAAGGAGATATACA	pET28a(+)
RBS V	AACTTTAAGAAGGAGATCCA	unknown, contains NdeI
RBS VI	AACTTTAAGAAGGAGATACA	RBS V without NdeI
RBS VII	CACAGGAAACAGGAATCCA	RBS I without EcoRI

4.9 Supporting information

4.9.1 Supplementary sequences

Nucleotide sequence 1: conSa cloned into pJET1.2

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATGAACATTTCGTTTTATTCCCGTTACCAGAACTACAACACCTGCTTTTCAATCCTGAGATAGACCTTT
ACGACTTTCGGAAGAGATACGCCTGATTGAATAACTGTGTAGAAACAACCGAACGCAGAAGGATACGATAAAAGTGGGTCTCAGAACTTGGTAG
CGATACTATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 2: conScRV designed

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATGAACATTTCGTTTTATTCCCGTTACCAGAACTACAACACCTGCTTTTCAATCCTGAGATAGACCTTT
ACGACTTTCGGAAGAGATACGCCTGATTGAATAACTGTGTAGAAACAACCGAACGCAGAAGGATACGATAAAAGTGGGTCTCAGAACTTGGTAG
CGAGCCTATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 3: con1c cloned into pJET1.2

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATAGGCATTTTCCTTTTACTGGAGACCTTACCGACAGGGAGTGATTCTTCCACAGTAGCGTGTAGGTTC
TATCCGATTTTATTTACCTATCGCCCACAGAAGTATTTCAGATTGGTGTAGCCTATTGTTGATTTTCCGTTGTAACTCTTATTCGTGGCGTAAA
GTATCGCATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 4: con1aRV designed

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATAGTAATTTTCCTTTTACTGGAGACCTTACCGACAGGGAGTGATTCTTCCACAGTAGCGTGTAGGTTC
TATCCGATTTTATTTACCTATCGCCCACAGAAGTATTTCAGATTGGTGTAGCCTATTGTTGATTTTCCGTTGTAACTCTTATTCGTGGCGTAAA
GTATCGCATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 5: con2a cloned into pJET1.2

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATGAACGACGGTAGAAGTATTTATCCACGGCTGTATCTGATTTCACTGAGACACCTATTGAGTTGTTTT
ACCCTTCCTTTCGTTATTACGAATAAAACAGTGAAGTTGATTGGGAACCTCTCCTGAACGGATTACCTTACTATCTGGTCTCTTACTGTCACCG
AAATACTATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 6: con3b designed

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATGGAGGATAATCAGTGCTTTGAGACCGGTTTTCAGGTAATAAGGTTCGTGTAAGCAACAATCTCAGTA
TCGGACAACTATCCAGTATAATCTGTAAGAGTAATAGTGCCGATAAGTTCCTCGTAAAAGTCCGTTGATAGGTGATTTCAGGTCAGTCTCTGTC
TACTCGCATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 7: con4b designed

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATGAACACTCAATCCCGTATCGTTGCTCAGAATCTACTATCAGTTGGAGAAACAGCGGAAAGGTGTAAA
AGCGTAACAATAAAGTTGTTATTTCAGGGTTCACTACAGACGCAACAACGAAGCAGATTGAGTCCAACTTTCTTACGGTCTCGGCTTATTTTAT
CAGGCTTATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 8: conEc cloned into pJET1.2

Sequence before level°0 assembly

CATGCACGTCTCTGAACTAGAAGACATAGGCAGAGGAATCTTCGCGGAGACCTTTATTTACGCAACACTCAGGTCAACTTACTTCTACGAATCA
CGGAAACTATTGTGGTTAGGGTTACTGTTATTTCTCTACCTGACGACAATCCTACTTCGGAGTGTGGCTATTCAACTACAACGATAATCTTCTT
ACATCGCATGTCTTCAGTCGCTGAGACGTGTTCA

Nucleotide sequence 9: bSE-araC-P_{BAD}

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC

GAACTAGAAGACATTACTTTATGACAACTTGACGGCTACATCATTCACTTTTTCTTCACAACCGGCACGGAACTCGCTCGGGCTGGCCCCGGTG
CATTTTTTAAATACCCGCGAGAAATAGAGTTGATCGTCAAAACCAACATTGCGACCGACGGTGGCGATAGGCATCCGGGTGGTGCTCAAAAGCA
GCTTCGCCTGGCTGATACGTTGGTCCTCGCGCCAGCTTAAGACGCTAATCCCTAACTGCTGGCGGAAAAGATGTGACAGACGCGACGGCGACAA
GCAAACATGCTGTGCGACGCTGGCGATATCAAAATTGCTGTCTGCCAGGTGATCGCTGATGTACTGACAAGCCTCGCGTACCCGATTATCCATC
GGTGGATGGAGCGACTCGTTAATCGCTTCCATGCGCCGCAGTAACAATTGCTCAAGCAGATTTATCGCCAGCAGCTCCGAATAGCGCCCTTCCC
CTTGCCCGGCGTTAATGATTTGCCCAAACAGGTCGCTGAAATGCGGCTGGTGCGCTTCATCCGGGCGAAAGAACCCCGTATTGGCAAATATTGA
CGGCCAGTTAAGCCATTCATGCCAGTAGGCGCGCGGACGAAAGTAAACCCACTGGTGATACCATTCGCGAGCCTCCGGATGACGACCGTAGTGA
TGAATCTCTCCTGGCGGGAACAGCAAAATATCACCCGGTCGGCAAACAAATTCTCGTCCCTGATTTTTCACCACCCCCTGACCGCGAATGGTGA
GATTGAGAATATAACCTTTCATTCCCAGCGGTCGGTCGATAAAAAAATCGAGATAACCGTTGGCCTCAATCGGCGTTAAACCCGCCACCAGATG
GGCATTAAACGAGTATCCCGGCAGCAGGGGATCATTTTGCGCTTCAGCCATACTTTTCATACTCCCGCCATTCAGAGAAGAAACCAATTGTCCA
TATTGCATCAGACATTGCCGTCACTGCGTCTTTTACTGGCTCTTCTCGCTAACCAAACCGGTAACCCCGCTTATTAAAAGCATTCTGTAACAAA
GCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTG
CTATGCCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATACCCGTTTTTTTGC
TTATGTCTTCAGTCGC

Nucleotide sequence 10: aSE-lacI-P_{T5}*lacO*

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC.

GAACTAGAAGACATTACTATTCTCACCAATAAAAAACGCCCGGCGGCAACCGAGCGTTCTGAACAAATCCAGATGGAGTTCTGAGGTCATTACT
GGATCTATCAACAGGAGTCGCGGCCGCTTACGTGCAGTCGATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATT
AATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTG
CGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAG
CGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTATCTTCGGTATCGTCGTATC
CCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGC
AGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATT
TGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGAC
CCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCCTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAA
TAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCG
AGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATT
TAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGC
CACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGG
GAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGC
GCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGTCCAACTTACGAGTGATAAATCATAAAAAATTTATTTGCTTTGTGAGCGG
ATAACAATTATAATAGATTCAATTGTGAGCGGATAACAATTTCACACAGCTTATGTCTTCAGTCGC

Nucleotide sequence 11: bS1-RBS III

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC. Spacer sequence in bold.

GAACATGAAGACTCGCTTAATACTGTTCGGACACTCAGAAGGCTTTATATTAAAGAGGAGAAATTAACAATATGTCTTCAGTCGC

Nucleotide sequence 12: bS1-RBS IV

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC. Spacer sequence in bold.

GAACATGAAGACTCGCTT**TTCTTACGAATCAACTTATTGGAGCGGAAC**AAGATAAGAAGGAGATATACAATATGTCTTCAGTCGC

Nucleotide sequence 13: bS3-RBS III

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC. Spacer sequence in bold.

GAACATGAAGACTCGCTT**AATACTGTTCGGACACTCAGAAGGCTTTAT**ATTAAAGAGGAGAAATTAACAATGATGTCTTCAGTCGC

Nucleotide sequence 14: bS3-RBS IV

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC. Spacer sequence in bold.

GAACATGAAGACTCGCTT**TTCTTACGAATCAACTTATTGGAGCGGAAC**AAGATAAGAAGGAGATATACAATGATGTCTTCAGTCGC

Nucleotide sequence 15: b23-FLAG tag

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC.

GAACATGAAGACTCCAATGGGTGACTACAAGGACGACGATGACAAAGGTTCAATGATGTCTTCAGTCGC

Nucleotide sequence 16: b23-GST tag

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC.

GAACATGAAGACTCCAATGTCCCCTATACTAGGTTATTGGAAAATTAAGGGCCTTGTGCAACCCACTCGACTTCTTTTGGAATATCTTGAAGAA
AAATATGAAGAGCATTTGTATGAGCGCGATGAAGGTGATAAATGGCGAAACAAAAAGTTTGAATTGGGTTTGGAGTTTCCCAATCTTCCTTATT
ATATTGATGGTGATGTTAAATTAACACAGTCTATGGCCATCATACGTTATATAGCTGACAAGCACAACATGTTGGGTGGTTGTCCAAAAGAGCG
TGCAGAGATTTCAATGCTTGAAGGAGCGGTTTTGGATATTAGATACGGTGTTTCGAGAATTGCATATAGTAAAGACTTTGAAACTCTCAAAGTT
GATTTTCTTAGCAAGCTACCTGAAATGCTGAAAATGTTCGAAGATCGTTTATGTCATAAAACATATTTAAATGGTGATCATGTAACCCATCCTG
ACTTCATGTTGTATGACGCTCTTGATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTTAAAAAACGTAT
TGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATAGCATGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGAC
CATCCTCCAAAAGGTTCAATGATGTCTTCAGTCGC

Nucleotide sequence 17: b23-MBP tag

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC.

GAACATGAAGACTCCAATGAAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGACTCGCTGAAGTCGGTAAG
AAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCC
CTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAA
GCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAA
GATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACC
TGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGG
CGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCA
GAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCGACACCAGCAAAGTGAATTATGGTGTAA
CGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGC
AAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTTAC
GAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTT
TCTGGTATGCCGTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTAATGGTTCAGG
AAGTGGCTCAATGATGTCTTCAGTCGC

Nucleotide sequence 18: cSE-rrnB T1

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC.

GAACTAGAAGACATGGAGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTC
TCCTGAGTAGGACAAATCCGCCAGGCATGTCTTCAGTCGC

Nucleotide sequence 19: cSE-T7

Sequence presents biological part cloned into pMC0 including level°1 fusion sites GAAC and TCGC.

GAACTAGAAGACATGGAGAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGAGGCATGTCTTCAGTCGC

4.9.2 Supplementary tables

Table S1: Standard settings of the online tool R2O DNA

Parameter	Setting
GC content	40%
Forbidden sequences	see Table S3
Selected genomes	saccharomyces_cerevisiae_genome escherichia_coli_k12_dh10b escherichia_coli_k12_mg1655 escherichia_coli_k12_w3110 bacillus_subtilis_168 igem_all_parts_082013
Blast e-value	1
Temperature for DNA free folding energy calculations (C):	50
Minimum intra-molecular folding free energy (kcal/mol)	-4
Minimum inter-molecular folding free energy (kcal/mol)	-9
Maximum Smith-Waterman alignment score (EDNAFULL matrix)	90
Maximum allowed exact sub-sequence match length	10

Table S2: Standard and user specified forbidden sequences for the online tool R2O DNA

Sequence	Sequence description
TTGACA	*E. coli* sig70 -30 site
TATAAT	*E. coli* sig70 -10 site
TTGACA	*E. coli* sig70 -30 site
TTGNNNNNNNNNNNNNNNNNNNNTATNNT	*E. coli* sig70 promoter weak consensus
TGGCACGNNNNTTGC	*E. coli* sig54 promoter consensus
TCNCCCTTGAANNNNNNNNNNNNNNNNNCCCCATTTA	*E. coli* sig32 promoter consensus
GAACTNNNNNNNNNNNNNNNNGTCNNA	*E. coli* sig24 promoter consensus
AAAGA	RBS
AGGAGG	Shine-Dalgarno sequence or 2xArg bad codon
ATG	translation initiation sites
TTATNCACA	DnaA binding sites
TGTGANNNNNNTCACANT	CAP binding sites
CTAG, CTAA, CAAA, CAAG	IS5 insertion site
NGCTNAGCN	IS10 insertion site
GGGNNNNNCCC	IS231 insertion site
G{3,}[ATGC]{1,7}G{3,}[ATGC]{1,7}G{3,}[ATGC]{1,7}G{3,}	G-quadruplex
GGGG	G-quadruplex
GGTCTC	BsaI
AAAAA, TTTTT, GGGGG, CCCCC, ATATAT, ACACAC, AGAGAG, TATATA, TCTCTC, TGTGTG, CACACA, CTCTCT, CGCGCG, GAGAGA, GTGTGT, GCGCGC	run
GAAGAC	BpiI (specified by user)
CGTCTC	BsmBI (specified by user)

Table S3: Sequences generated in the online tool R2O DNA

number	sequence
1	ATTTCGTTTTATTCCCGTTACCAGAACTACAACACCTGCTTTTCAATCCTGAGATAGACCTTTACGACTTTCGGAAGAGA
3	TACGCCTGATTGAATAACTGTGTAGAAACAACCGAACGCAGAAGGATACGATAAAAGTGAGTAAGAACTATTGGTAGCGA
4	AGAGGAATCTATCTGCTGGACTTTATTTACGCAACACTCAGGTCAACTTACTTCTACGAATCACGGAAACTATTGTGGTT
5	AGGGTTACTGTTATTTCTCTACCTGACGACAATCCTACTTCGGAGTGTGGCTATTCAACTACAACGATAATCTTCTTACA
6	ATTTTCCTTTATCTGGTTGACTTACCGACAGGGAGTGATTCTTCCACAGTAGCGTGTAGGTTCTATCCGATTTTATTTAC
8	GACGGTAGAAGTATTTATCCACGGCTGTATCTGATTTCACTGAGACACCTATTGAGTTGTTTTACCCTTCCTTTCGTTAT
9	TACGAATAAAACAGTGAAGTTGATTGGGAACCTCTCCTGAACGGATTACCTTACTATCTCGTGTATTGTTGTCACCGAAA
10	GATAATCAGTTCCGTTCTTCTGGTTTTCAGGTAATAAGGTTCGTGTAAGCAACAATCTCAGTATCGGACAACTATCCAGT
11	ATAATCTGTAAGAGTAATAGTGCCGATAAGTTCCTCGTAAAAGTCCGTTGATAGGTGATTTCAGGTCAGTCTCTGTCTAC
12	ACTCAATCCCGTATCGTTGCTCAGAATCTACTATCAGTTGGAGAAACAGCGGAAAGGTGTAAAAGCGTAACAATAAAGTT
13	AGTAGTATTATCCAACGCTTACCAGTTATTGACGATTTCTGTAGGACTTCCGTAGCACCCTTATCTGAGAGTTGTTGAAT
14	GTTATTTCAGGGTTCACTACAGACGCAACAACGAAGCAGATTGAGTCCAACTTTCTTACCTTTACGGAGTATTTTATCAG
15	CTATCGCCCACAGAAGTATTTCAGATTGGTGTAGCCTATTGTTGATTTTCCGTTGTAACTCTTATTCGTGGCGTAAAGTA
16	GTAATAGATAGGATAAGTCAGAAGCAGTGTCGTGTAGAGTGAAAAGGCTCCAGATTCGTAGTATTGTCCCGATTATTCTT
17	ATTGATAGGCTCTTATTGTAGTAAAGTTTTCTGACCAGGTGTGAATCCTCTCCGTGACTGAACGAAGTGTTGTAAGTAAC

Table S4: ClustalW pairwise alignment scores between sequences used for connectors

Pairwise scores are the number of identities between the two sequences, divided by the length of the alignment.

sequence number	sequence number	Score / %
1	3	50.0
1	4	45.0
1	5	42.5
1	6	50.0
1	8	45.0
1	9	42.5
1	10	38.75
1	11	45.0
1	12	42.5
1	13	51.25
1	14	48.75
1	15	53.75
1	16	25.0
1	17	51.25
3	4	38.75
3	5	38.75
3	6	42.5
3	8	35.0
3	9	52.5
3	10	43.75
3	11	42.5
3	12	46.25
3	13	42.5
3	14	40.0
3	15	43.75
3	16	46.25
3	17	55.0
4	5	48.75
4	6	47.5
4	8	48.75
4	9	45.0
4	10	48.75
4	11	43.75
4	12	46.25
4	13	43.75
4	14	48.75
4	15	51.25
4	16	40.0
4	17	51.25

sequence number	sequence number	Score/ %
5	6	50.0
5	7	45.0
5	8	52.5
5	9	41.25
5	10	43.75
5	11	50.0
5	12	41.25
5	13	47.5
5	14	50.0
5	15	41.25
5	16	45.0
5	17	51.25
6	8	42.5
6	9	42.5
6	10	51.25
6	11	52.5
6	12	41.25
6	13	51.25
6	14	52.5
6	15	43.75
6	16	45.0
6	17	37.5
8	9	51.25
8	10	38.75
8	11	45.0
8	12	40.0
8	13	56.25
8	14	52.5
8	15	52.5
8	16	42.5
8	17	47.5

sequence number	sequence number	Score / %
9	10	45.0
9	11	50.0
9	12	37.5
9	13	47.5
9	14	50.0
9	15	50.0
9	16	48.75
9	17	48.75
10	11	48.75
10	12	53.75
10	13	47.5
10	14	45.0
10	15	41.25
10	16	51.25
10	17	46.25
11	12	38.75
11	13	46.25
11	14	26.25
11	15	41.25
11	16	53.75
11	17	48.75
12	13	35.0
12	14	38.75
12	15	40.0
12	16	43.75
12	17	51.25
13	14	38.75
13	15	48.75
13	16	51.25
13	17	50.0
14	15	50.0
14	16	51.25
14	17	51.25
15	16	43.75
15	17	50.0
16	17	47.5

4.10 **References**

1. Participants of the ERASynBio 2nd Strategic Conference and Strategic Vision focus groups: **Next steps for European synthetic biology: a strategic vision from ERASynBio.** pp. 1-32: European Research Area Network for the development and coordination of synthetic biology in Europe (ERASynBio); 2014:1-32.
2. Valla S, Lale R: *DNA cloning and assembly methods.* Springer Protocols: Springer; 2014.
3. Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S: **A modular cloning system for standardized assembly of multigene constructs.** *PLoS ONE* 2011, **6:**e16765.
4. Szybalski W, Kim SC, Hasan N, Podhajska AJ: **Class-IIS restriction enzymes — a review.** *Gene* 1991, **100:**13-26.
5. Engler C, Marillonnet S: **Golden Gate cloning.** In *DNA Cloning and Assembly Methods. Volume* 1116. Edited by Valla S, Lale R: Humana Press; 2014: 119-131: *Methods in Molecular Biology*].
6. Lee ME, DeLoache WC, Cervantes B, Dueber JE: **A highly characterized yeast toolkit for modular, multipart assembly.** *ACS Synthetic Biology* 2015.
7. Ream W, Geller B, Trempy J, Field K: *Molecular microbiology laboratory: a writing-intensive course.* Elsevier Science; 2012.
8. Casali N, Preston A: *E. coli plasmid vectors: methods and applications.* Humana Press; 2003.
9. Koshanskaya M: **Combinatorial cloning of multipart expression constructs using modular cloning.** *Master Thesis.* Graz University of Technology, Institute of Molecular Biotechnology; 2015.

5 Summarising conclusion

This thesis elaborates on the *in vivo* synthesis of non-canonical amino acids (ncAAs) in *E. coli* to produce synthetic protein variants. It focuses on fluoro amino acids as this class of ncAAs proved to be especially attractive for protein engineering.

A literature review on fluoro amino acids presenting the first chapter of this thesis recognizes their high value to engineer protein traits, which is mainly due to the outstanding features of the carbon-fluorine bond. Moreover, it details the biosynthesis of 4-fluoro-L-threonine produced by several *Streptomyces* species. It constitutes the only representative of fluorinated amino acid analogs discovered so far in nature. Interestingly, 4-fluoro-L-threonine is not commercially available showing the intricacy of its biosynthesis.

In the first reaction of the five-step 4-fluoro-L-threonine pathway, fluoride and S-adenosyl-L-methionine (SAM) are converted to 5'-deoxy-5'-fluoroadenosine (FDA) by the fluorinase enzyme. In the second chapter of this thesis we set out to establish FDA conversion in *E. coli* with the future goal to produce 4-fluoro-L-threonine. Our first aim was to improve soluble fluorinase expression in *E. coli* as our experiments showed a strong tendency for inclusion body formation. To this end we tested several vector contructs: an arabinose inducible single copy plasmid and three medium copy pasmids, which were all IPTG inducible either containing the T5, T7 or TAC promoter to drive gene expression. Suprisingly, low expression regimes realized with the arabinose inducible single copy vector did not prevent protein aggregation. Although none of the tested plasmids distinctly improved the fluorinase expression pattern, the vector construct with the T5 promoter turned out to be the most suitable to express this enzyme. We observed that varying expression conditions like temperature, induction time or inducer concentration had no effect on soluble fluorinase expression. We also ruled out that the N-terminal His-tag on the enzyme did account for protein aggregation.

In another approach we co-expressed the chaperones LbpB and a truncated version of DsbC. Only LbpB co-expression showed a slightly positive effect on soluble FlA levels in one experiment. However, this finding needs to be confirmed as we obtained contradicting results.

It has been reported that the presence of its substrate SAM is important for the structural integrity of the fluorinase [1]. Therefore, we co-expressed the rat liver SAM synthase (RLSS) to elevate intracellular SAM levels [2, 3]. Although this led to some increase of soluble fluorinase, most of the protein still accumulated in the insoluble fraction.

A codon optimization strategy turned out to be the most effective approach to reduce inclusion body formation: We mimicked the codon usage of the fluorinase in its original *Streptomyces* host, also decribed as codon harmonization, and additionally placed rare *E. coli* codons at the region of translation initiation. The codon-opimized expression construct showed a decisively lower tendency for fluorinase aggregation in *E. coli*.

The second chapter also elaborates on FDA production in *E. coli* by the action of overexpressed fluorinase. We used the original fluorinase gene sequence in these experiments, because the codon-optimized version has not been availabe yet at this time of the study. Although we verified the prescence of active enzyme in cells, we could not detect any FDA synthesis activity in an *in vivo* approach, where cultures were supplemented with fluoride. We reasoned that FDA synthesis failed due to an exclusion of fluoride from the intracellular space, because it is toxic to *E. coli* [4]. Therefore, lyophilized cells of cultures overexpressing FlA alone as well as together with RLSS to increase intracellular SAM levels were used for FDA bioconversions. We assumed that the lyophilisation process permeabilized the cell membrane what should allow the unconditional passage of external fluoride as well as SAM. For FDA conversions we supplemented different substrates to reactions with lyophilized cells. Reactions contained either only fluoride or fluoride together with SAM. Moreover, we added only L-methionine or L-methionine together with ATP, which are both substrates for SAM synthases. In all experiments we could observe FDA synthesis proving FlA activity in lyophilized cells. RLSS overexpression elevated SAM levels, what resulted in an FDA increase. This was only the case when fluoride and no external SAM or SAM synthase substrates were added to reaction mixes.

In chapter 3 we describe the use of synthetic precursor molecules, that is indole analogs, to produce non-canonical tryptophans (ncTrp) *in vivo* for the expression of protein variants in *E. coli*. The approach includes the co-expression of the tryptophan synthase from *Salmonella typhimurium* (*St*TRPS) on top of the TRPS activity of the *E. coli* host to enhance the transformation of a broad spectrum of indole analogs to tryptophan.

For the residue-specific incorporation of ncTrp we established a two-step protocol. In the early growth phase we induced *St*TRPS expression in the prescence of limiting concentrations of tryptophan. When a growth arrest occurred due to tryptophan depletion, we added the indole analog and shortly thereafter induced target gene expression. If tryptophan would be still present during variant expression, it is preferably incorporated into the target protein leading to unlabeled protein species. We observed that expression of *St*TRPS influences Trp depletion and that only its low level expression was compatible with the protocol for residue-specific ncTrp incorporation.

We systematically evaluated the procedure with two target proteins of different Trp content (0.8%, low-Trp *versus* 3.8%, high-Trp) with and without *St*TRPS co-expression. The titers of the expressed variant proteins varied with the incorporated ncTrp. Contrary to our expectations, *St*TRPS co-expression improved the incorporation efficiency with the low-Trp target protein but the host TRPS activity was sufficient for the homogeneous incorporation of fluorinated Trp analogs into the high-Trp target.

We also evaluated the procedure for the expression of a target protein using the pET system. One has to keep in mind that the accidental incorporation of ncTrp, which occurs under these conditions, can impair the function of the T7 RNA polymerase and therefore affect protein expression. In our hands, only monofluorinated Trp analogs were compatible with the pET expression system.

Chapter 4 of this thesis addresses the establishment of a cloning strategy that allows the straight-forward assembly of multigene constructs to express enzyme cascades for ncAA synthesis. Based on the Modular Cloning (MoClo) method [5] we designed a standardized assembly strategy. It comprises three hierarchical assembly levels: The first level yields standardized biological parts like genetic control elements or protein tags as well as the genes of interest. In the next level these parts are assembled in a predefined order to give devices. Devices can already be fully functional transcriptional units of one gene of interest. Multigene or pathway constructs are assembled in the last assembly level from devices. The cloning concept worked in our hands as we succeeded in the assembly of several part plasmids and destination vectors. Moreover, we developed a nomenclature system that should guarantee the standardized use of modules. The described MoClo design provides the basis for a high-throughput cloning method to assemble complex pathway constructs.

Taken together, the thesis deals with genetic and metablic engineering approachs to facilitate the *in vivo* synthesis of ncAAs for variant protein production in *E. coli*. NcAA pathways are often complex and consist of several enzymatic steps. The assembly of the corresponding genes in the most suitable order and stoichiometry can constitute the first bottleneck. Therefore, advanced genetic tools are a prerequisite to realize *in vivo* ncAA synthesis. NcAAs can be synthesized from scratch in the metabolically engineered host or synthetic precursor molecules are added to the expression culture and converted to ncAAs. Synthesis of ncAAs inside the expression host should circumvent the need for expensive amino acid analogs to produce protein variants. Cost reduction would tremendously foster research on and industrial interest in ncAAs as valuable tools for protein engineering. Not least, metabolic engineering approaches broaden the spectrum of ncAAs currently available to the scientist.

5.1 References

1. Dong C, Huang F, Deng H, Schaffrath C, Spencer JB, O'Hagan D, Naismith JH: **Crystal structure and mechanism of a bacterial fluorinating enzyme.** *Nature* 2004, **427:**561-565.
2. Alvarez L, Mingorance J, Pajares MA, Mato JM: **Expression of rat liver S-adenosylmethionine synthetase in Escherichia coli results in two active oligomeric forms.** *Biochem J* 1994, **301 (Pt 2):**557-561.
3. Posnick LM, Samson LD: **Influence of S-adenosylmethionine pool size on spontaneous mutation, Dam methylation, and cell growth of escherichia coli.** *J Bacteriol* 1999, **181:**6756-6762.
4. Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB, Breaker RR: **Widespread genetic switches and toxicity resistance proteins for fluoride.** *Science* 2012, **335:**233-235.
5. Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S: **A modular cloning system for standardized assembly of multigene constructs.** *PLoS ONE* 2011, **6:**e16765.