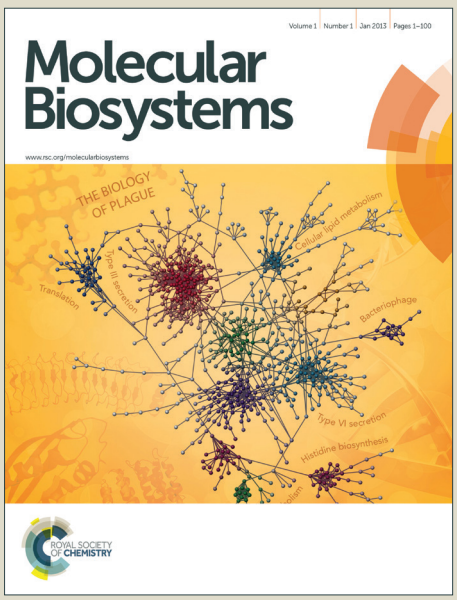
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange

Ralf Beer*[a], Konrad Herbst*[a],Nikolaos Ignatiadis*[a], Ilia Kats*[a], Lorenz Adlung*[b], Hannah Meyer*[c], Dominik Niopek*[a,d], Tania Christiansen[a], Fanny Georgi[a], Nils Kurzawa[a], Johanna Meichsner[a], Sophie Rabe[a], Anja Riedel[a], Joshua Sachs[a], Julia Schessner[a], Florian Schmidt[a], Philipp Walch[a], Katharina Niopek[e], Tim Heinemann[f], Roland Eils[a,g,h] and Barbara Di Ventura[#a,d]; iGEM team Heidelberg 2013

a) Center for Quantitative Biology (Bioquant), Division of Integrative Bioinformatics and Systems Biology, University of Heidelberg, Germany
b) German Cancer Research Center (DKFZ), Division of Systems Biology of Signal Transduction, Heidelberg, Germany
c) European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom
d) German Cancer Research Center (DKFZ), Division of Theoretical Bioinformatics, Synthetic Biology Group, Heidelberg, Germany
e) German Cancer Research Center (DKFZ), Division of Molecular Metabolic Control, Heidelberg, Germany
f) German Cancer Research Center (DKFZ), Division of Theoretical Systems Biology, Heidelberg, Germany
g) German Cancer Research Center (DKFZ), Division of Theoretical Bioinformatics
h) Institute for Pharmacy and Molecular Biotechnology (IPMB), Division of Bionformatics and Systems Biology, Heidelberg, Germany

* equally contributing authors

E-mail: barbara.diventura@bioquant.uni-heidelberg.de

Non-ribosomal peptide synthetases (NRPSs) are enzymes that catalyze ribosome-independent production of small peptides, most of which are bioactive. NRPSs act as peptide assembly lines where individual, often

interconnected modules each incorporate a specific amino acid to the nascent chain. The modules themselves consist of several domains that function in the activation, modification and condensation of the substrate. NRPSs are evidently modular, yet experimental proof of the ability to engineer desired permutations of domains and modules is still sought. Here, we use a synthetic-biology approach to create a small library of engineered NRPSs, in which the domain responsible for carrying the activated amino acid (T domain) is exchanged with natural or synthetic T domains. As a model system, we employ the single-module NRPS IndC from *Photorhabdus luminescens* that produces the blue pigment indigoidine. As chassis we use *Escherichia coli*. We demonstrate that heterologous T domain exchange is possible, even for T domains derived from different organisms. Interestingly, substitution of the native T domain with a synthetic one enhanced indigoidine production. Moreover, we show that selection of appropriate inter-domain linker regions is critical for functionality. Taken together, our results extend the engineering avenues for NRPSs, as they point out the possibility of combining domain sequences coming from different pathways, organisms or from conservation criteria. Moreover, our data suggest that NRPSs can be rationally engineered to control the level of production of the corresponding peptides. This could have important implications for industrial and medical applications.

Introduction

The ribosome is undoubtedly the molecular machinery universally used by cells across all domains of life to produce complex polypeptides. Yet, not all peptides stem from the ribosome. Bacteria and fungi – and perhaps

even other organisms such as sponges, tunicates, gastropods, and plants[1] – have an additional pathway for synthesizing peptides that relies on enzymes entirely independent of the ribosome: the non-ribosomal peptide synthetases (NRPSs)[2]. Peptides assembled by NRPSs, termed non-ribosomal peptides (NRPs), comprise a wide range of secondary metabolites, including commonly used antibiotics, as well as metabolic and detoxifying enzymes (reviewed in [3,4]). They are usually short but of remarkable structural diversity[5]. NRPS can produce peptides consisting of a range of 500 different monomers, including non-proteinogenic, N-methylated and D-enantiomeric amino acids[1].

NRPSs seem to lend themselves to rational engineering, as they are intrinsically modular in nature[6,7]. Indeed, NRP synthesis mostly follows the co-linearity rule[6], whereby the number and order of the substrate-specific modules NRPSs are composed of, corresponds precisely to the number and order of the amino acids in the synthesized NRP. NRPSs' modules are, in turn, composed of domains, some of which are required for the functionality of the module, while others are optional and introduce various modifications into the amino acids[1]. Crucial for the functionality of each module are the condensation (C), the adenylation (A) and the thiolation (T) domains. The A domain is responsible for the substrate specificity of the module and acts by activating its specific amino acid substrate. The T domain serves as an anchor point for the activated amino acid as well as for the growing peptide chain (in case of multi-modular NRPSs). The generation of the latter is facilitated by the C domain that couples activated amino acids from the preceding T domain to the succeeding one[7]. Prior to the attachment of amino acids, the T domain itself has to become activated by the transfer of 4'-phosphopantetheine (4'-PP) from Coenzyme A to a conserved serine via

phosphopantetheinyl transferases (PPTases)[8]. Neighboring domains and modules are connected by catalytically inactive linker regions, which are important for proper functionality of the NRPS. Most NRPS terminate with a thioesterase domain (TE) that cleaves the final NRP from the enzyme.

Several hybrid NRPSs obtained by module fusions have been successfully created[9,10], pointing to the possibility of modifying linker regions connecting two domains. However, these module fusions were limited to exchanges with domains from the same or a related organism or lacked *in vivo* validation. Direct substitution of a domain within a module has also been attempted. While exchanging an initial A domain and a terminal A-T fragment was successful[10,11], exchanging a central T domain was not, and random mutagenesis was needed to recover the NRPS function[12]. No insight into the mechanism by which these mutations could restore function was given. Therefore, despite the great advances done in this direction, further analysis is needed to understand the rules for engineering functional NRPSs through permutations of domains and modules.

To assess the interchangeability of T domains derived from different pathways, bacterial species, or with entirely synthetic ones, we selected the single-module NRPS IndC, encoded by *Photorhabdus luminescens*, responsible for the production of the blue pigment indigoidine[13] (Fig. 1a). We created a small library of IndC variants (Fig. 1b) that was quantitatively tested for *in vivo* production of indigoidine using the well-characterized bacterium *Escherichia coli* as chassis (Fig. 1c). We show that the IndC T domain can be successfully exchanged with natural and synthetic T domains, with the latter giving rise to new synthetases that outperform the native enzyme in *in vivo* indigoidine

production. Interestingly, we found that linker regions that connect the T domain to its neighboring domains must be carefully chosen to yield functional NRPSs.

Figure 1: Engineering the indigoidine synthetase IndC by T domain exchange. (a) Schematic representation of intracellular indigoidine production. For simplicity the activating reaction by the PPTase is omitted. A-Ox-A is the adenylation domain with an internal oxidation (Ox) domain, T is the thiolation domain and TE is the thioesterase domain. See also Fig. S1 for the chemical reactions leading to indigoidine production. **(b)** Schematic representation of IndC T domain exchange. **(c)** Indicated *E. coli* strains were co-transformed with constructs encoding IndC and the PPTase Sfp and grown in LB medium at room temperature. After 40 h, cultures were transferred into cuvettes and photographed using a Canon EOS 600D digital camera. (-): *E. coli* TOP10 without indigoidine synthetase.

Results

Selection of T domains for the small library of engineered indigoidine synthetases

To build our small library of engineered indigoidine synthetases, we first selected T domains belonging to unrelated NRPS modules present in different host organisms (Table 1). We also included the T domain of the closely related indigoidine synthetase BpsA, encoded by *Streptomyces*

lavendulae lavendulae (ATCC11924)[14], which presumably has a high chance of yielding a functional hybrid enzyme (Table 1). Moreover, we chose two T domains derived from an NRPS of unknown function (plu2642 and plu2670) (Table 1).

Table 1. Natural T domains used in this study.

Origin	NRPS module	NRPS function
S. lavendulae	BpsA	Blue pigment synthetase[14]
E. coli	EntF	Enterobactin synthetase[15]
B. parabrevis	TycA1	Tyrocidine synthetase (first module) [16]
B. parabrevis	TycC6	Tyrocidine synthetase (last module) [16]
D. acidovorans	DelH4	Delftibactin synthetase (second last module) [17]
D. acidovorans	DelH5	Delftibactin synthetase (last module) [17]
P. luminescens	Plu2642	unknown function[18]
P. luminescens	Plu2670	unknown function[18]

The bottom-up approach of synthetic biology consists in constructing biological entities (*e.g.* DNA regulatory elements, proteins, gene networks) "from scratch", putting together well-characterized functional units. Following this approach, we further thought of designing synthetic T domains to learn more about potential sequence features important for NRPS functionality. Thus, we created four sequence libraries with high homology to a) the entire IndC NRPS, the IndC T domain either b) restricted to sequences found in its host organism *P. luminescens* or c) without organism restriction and d) T domains from NRPSs unrelated to IndC (Fig. 2a). The amino acids at each position in the synthetic T domain sequence were determined using three different methods ("Consensus", "Guided Random" and "Randomized Generation" methods; see ESI section 2 for details). In total, seven synthetic T domains were created (Fig. 2a). Among these, some are highly homologous to the native T domain (those obtained with the "Consensus"

method, SynT1-T4), some are of lower homology (those obtained with the "Guided Random" method, SynT5-T6), and one is very distantly related to it (that obtained with the "Randomized Generation" method, SynT7). We used Pfam to predict the IndC T domain boundaries[19]. A multiple sequence alignment of all T domains used in this study is shown in Fig. 2b.

T domains within NRPSs must be activated by phosphopantheteinyl transferases (PPTases) (Fig. S2a). To select an appropriate PPTase for the native IndC, we cloned four PPTases of different origins (Table S1) and qualitatively assessed them for their ability to activate IndC, thus making bacterial colonies turn blue on agar plates. We found that all considered PPTases yielded blue colonies (Fig. S2b), indicating that the PPTases could, at least to some extent, interact with IndC. For all subsequent studies, we chose the well-characterized PPTase Sfp from *Bacillus subtilis*[20], whose broad activation spectrum renders it particularly suitable for biotechnological applications[8,21]. We also tested different *E. coli* strains for indigoidine production, namely the wild type MG1655 strain, the engineered BL21(DE3) strain carrying a single chromosomal copy of the sfp gene (BAP1 strain[22]) and the TOP10 strain. We co-expressed IndC and Sfp in each strain and found that TOP10 was the most suitable for indigoidine production under the tested conditions (Fig. 1c).

Figure 2: Workflow for the design of synthetic IndC T domains. (a) Schematic view of the workflow used to create synthetic T domains. The indicated sequences were used as queries in the BLAST algorithm to identify similar sequences. The 50 sequences with the highest identity score with respect to the query were considered for the following procedure. In query 4, for each of the five selected T domains the 50 best hits were determined and

combined. These libraries of closely related sequences were then subjected to multiple sequence alignment (MSA) by ClustalO. Seven synthetic T domains were created from either the consensus sequences of the MSA or amino acid profiles determined thereof. Generation of the synthetic sequences was based on three different algorithms: The first one is favoring the conservation of the native IndC T domain at a certain position unless the frequency of a new amino acid within the MSA was >50% ("Consensus" method), the second is building novel T domain sequences based on the amino acid frequencies in the sequence profile ("Guided Random" method) and the third is randomly choosing amino acids from the MSA at a certain position independent of their frequency ("Randomized Generation" method). **(b)** MSA of the native IndC T domain and all natural and synthetic T domains used in this study (amino acid color code according to ClustalX[23]). Numbers at the beginning and end of the sequence describe positions of the T domains in the respective NRPSs.

Optimized indigoidine production by an engineered synthetase containing a synthetic T domain

Indigoidine production by IndC variants was first qualitatively assessed by the presence of blue colonies on agar plates (Fig. 3). Overall, we found that one third of the engineered synthetases were functional.

Figure 3: Phenotypic comparison of *E. coli* Top10 cells expressing IndC variants engineered by T domain exchange. Bacteria co-transformed with constructs encoding for the indicated IndC variants and the PPTase Sfp were grown on agar plates for 48 h at 30°C. Colonies were then photographed using a Canon EOS 600D digital camera. Scale bar = 0.5 cm. wt, wild type IndC synthetase.

Indigoidine has been shown to inhibit bacterial growth through an unknown mechanism[24] and, while small colonies effectively produce more pigment per cell, bigger colonies produce more pigment in total and appear bluer (Fig. 3, compare SynT1 and Plu2642). To quantitatively compare the level of indigoidine produced by different IndC variants, we used a recently described method, in which the absorbance of bacterial cultures at 590 and 800 nm is measured[25] (Fig. 4, a and b). The variant containing the natural Plu2642 T domain and four variants containing synthetic T domains that yielded blue colonies on agar plates were chosen for the quantitative indigoidine production assay. In addition, we included three variants that did not show blue colonies on agar plates to assess whether different growth conditions (on solid agar versus liquid medium) might affect indigoidine production (Fig. 4a). Measurements were referenced to indigoidine production from native IndC. *E. coli* TOP10 cells transformed with an empty plasmid were used as negative control. Quantitative measurements confirmed the results obtained with the agar plates, with constructs corresponding to white colonies showing little to no indigoidine production (BpsA, TycC6 and SynT7 in Fig. 3 and Fig. 4, a and c). Interestingly, we found that the synthetic T domains T1, T3 and T4 yielded higher indigoidine levels than native IndC (Fig. 4c). Due to its antimicrobial effects, a negative correlation between indigoidine and

bacterial growth is expected. Surprisingly, we found that the SynT1 variant not only grew better but also produced more indigoidine (Fig. 4d).

Figure 4: Tuned indigoidine production by engineered IndC variants. (a) Representative picture (acquired with Canon EOS 600D) of *E. coli* TOP10 liquid cultures co-expressing the indicated constructs and the PPTase Sfp in a 96-well plate format. Cultures were inoculated from single colonies and grown in falcon tubes for 31 h at 30°C. 200 µL of each culture were then transferred into a 96-well plate for measurement of indigoidine production in a plate reader. The three replicates correspond to three different colonies picked from the same plate for each construct. **(b)** Graph showing the contribution of indigoidine to the absorbance of the bacterial culture at 590 nm. Indigoidine amounts can be calculated with a formula that considers optical densities (OD) at 590 and 800 nm (see ESI section 3 for details**).** Inset shows the linear relation between OD590 and OD800 for a bacterial culture. **(c)** Bar plot showing indigoidine amounts (blue) and OD800 (grey) of *E. coli* cultures co-expressing the indicated IndC variants and Sfp after 31 h of incubation at 30°C. **(d)** Bar plot showing the indigoidine/OD800 ratio for the data shown in (c). *** indicates a P value <0.0001. P value was calculated using the Welch's t test. (c-d) Error bars: s.d. from n = 3 independent experiments. In each experiment three independent colonies were measured for each construct.

Quantitative dynamic modeling of indigoidine production

To understand which reactions in the indigoidine production pathway might be affected by the T domain exchange, we developed a mathematical model based on ordinary differential equations (Fig. 5a and ESI section 4). In the model, bacterial growth is represented by a logistic equation[26] (with β being the bacterial growth rate and B_{max} the maximum growth capacity), while indigoidine production is decomposed into three

steps: synthesis of cyclic glutamine, dimerization to indigoidine and degradation (with rates k_{syn}, k_{dim} and k_{deg}, respectively).

As the quantitative production assay (Fig. 4) only represented an end-point measurement, dynamic data were acquired to calibrate the model and estimate kinetic parameters (see ESI section 4). The time-resolved data qualitatively confirmed the main trends seen in the quantitative indigoidine production assay after 31 hours (Fig. 4c): the variant SynT5 produced less indigoidine than native IndC but grew better; in terms of both, growth and indigoidine production, SynT3 performed better as compared to native IndC; and SynT1 exhibited the overall best indigoidine production (Fig. 5b). However, it is worth to note that the time-course data did not exactly match the end-point measurements, as, for instance, there was some variability in the curves corresponding to cultures of the same construct started from three independent colonies picked from the plate. Another interesting difference was the absence of a negative effect of indigoidine on bacterial growth as all variants grew well independent of indigoidine production (Fig. 5b). The differences between the quantitative results in the two assays could be due to the culturing conditions (*e.g.* total volume, shaking trajectories).

Nevertheless, the model could capture the experimentally measured dynamics (Fig. 5b), thus information about kinetic parameters could be inferred. We estimated the five parameters β, B_{max}, k_{syn}, k_{dim} and k_{deg} (Fig. 5a) for four indigoidine-producing variants (wild type, SynT1, SynT3 and SynT4) and found that bacterial growth parameters (β, B_{max}) as well as indigoidine degradation rate (k_{deg}) were very similar in all cases (Fig. 5c). On the contrary, synthesis of cyclized glutamine (k_{syn}) and its dimerization into indigoidine (k_{dim}) varied considerably, suggesting that the corresponding reactions significantly impact indigoidine production

and might be responsible for the observed difference among synthetase variants.

To understand the influence of each reaction on indigoidine production, we simulated the latter over a range of the two parameters k_{syn} and k_{dim} covering the values found for the different variants (Fig. 5d). Notably, SynT1 and SynT3 are rather close to each other in parameter space with fast dimerization and rather low synthesis rate. Native IndC is characterized, instead, by rapid cyclization of glutamine but delayed dimerization (Fig. S3), resulting in only moderate indigoidine yield. The model, therefore, suggests that to improve the synthetase functionality one has to push it into the parameter space with both high synthesis and dimerization rates. By exchanging the native T domain with the synthetic T1 and T3 domains we might be allowing for a better dimerization of cyclized glutamine, thus increasing indigoidine production, albeit not to the maximum attainable.

Figure 5: Quantitative dynamic modeling of indigoidine production. (a) Reaction scheme of indigoidine production. The first reaction represents bacterial (Bac) growth (rates β, B_{max}), the second reaction represents conversion of glutamine (Gln) to cyclized glutamine (cGln) (rate k_{syn}), the third reaction represents dimerization of cGln to indigoidine (Ind) (rate k_{dim}) and the fourth reaction represents degradation of indigoidine (rate k_{deg}). **(b)** Model calibration with time-resolved data for bacterial density and indigoidine concentration for the indicated variants. Solid lines depict model trajectories, with shaded regions showing the 95% confidence interval. Dashed lines depict experimental data. For all panels, one exemplary data set is shown with χ^2 as a measure for the goodness of the fit. The number of data points for calibration was N=714. **(c)** Boxplot of the distribution of best estimates for parameter values obtained for the four IndC variants. **(d)** Surface plot color-coding indigodine yield simulated for k_{syn} and k_{dim}. White circles represent 95% confidence regions for the two parameters of the indicated variants.

T domain boundary selection affects the functionality of engineered synthetases

We were surprised to see that inter-species exchange of the native IndC T domain with the one derived from the closely related indigoidine synthetase BpsA yielded a dysfunctional enzyme (Fig. 3, compare "wild type" and "BpsA", and Fig. 4, a, c and d). This could be due to poor selection of domain boundaries. One of the most commonly applied NRPS domain prediction tools is Pfam[19], which was used to determine the boundaries of the IndC T domain to be exchanged (Fig. 6a). However, results from multiple online NRPS domain and module boundary prediction tools[27-30] often do not match. The predictions from any single tool may therefore be suboptimal. To create a library of constructs representing different boundary selections, we aligned the sequences of IndC and BpsA (Fig. 6b). Using Pfam, the position of the T domains and linker regions within these NRPSs was determined. We then generated combinations in which the linker sequence (connecting A to T and T to TE) was derived from either IndC or BpsA (Fig. 6b) to see whether we could recover a functional indigoidine synthetase. We also tested whether moving the boundaries inwards would restore the functionality of the synthetase ("C" and "1" in Fig. 6a and b). Fig. 6c shows that defining different boundaries for the T domain affects the functionality of the resulting enzymes. Interestingly, we found that retention of the A-T linker region from the incoming BpsA T domain (domain boundaries A-1 and A-2) resulted in a functional indigoidine synthetase comparable to the wild type IndC. The innermost T domain boundaries (boundaries C-1) resulted instead in a very weak indigoidine synthetase.

Figure 6: Impact of T domain boundaries selection on engineered synthetase functionality. (a) Sequence alignment of IndC and BpsA. Indicated are Pfam-predicted

domain borders (above) and alternative boundaries experimentally tested. Sequence numbering is indicated for the first and last amino acid. **(b)** Zoom-in of the sequence alignment shown in (a). Amino acid color code according to ClustalX[23]. **(c)** Qualitative assessment of functionality of engineered synthetases with different T domain boundaries. Three single colonies of TOP10 cells co-expressing the PPTase Sfp and the BpSA variant with the indicated boundaries were used to inoculate liquid cultures. Cultures were grown for 31 h at 30°C in falcon tubes and then transferred into a 96-well plate for easy visual comparison of blue coloring. Picture was taken with a Canon EOS 600D digital camera.

Discussion

In this work, we successfully engineered the indigoidine synthetase IndC from *P. luminescens* by exchanging its T domain with various natural and synthetic ones (Fig. 3, Fig. 4 and Fig. 6). A similar attempt had previously failed in the context of the indigoidine synthetase BpsA and only random mutagenesis followed by screenings could restore pigment production[12]. However, in that study, the reasons why the mutations could turn a dysfunctional synthetase in a functional one were not investigated – given that the focus of their study was on large-scale recovery of PPTases. It would be interesting to see whether the mutations affected the linker regions, as in this study we show that appropriate selection of domain boundaries is a critical parameter to adjust. Indeed, we find that T domains from unrelated organisms and pathways as well as synthetic T domains can take the place of the endogenous IndC T domain, as far as there is enough sequence homology to it and as far as the boundaries are appropriately chosen (Fig. 2b, Fig. 6 and Fig. S4). In particular, the IndC T domain core motif DDFFESGGNSL, important for 4'-PP co-factor binding and substrate acetylation is highly conserved among functional engineered variants or even completely identical in case of the improved SynT1. In fact, this latter has only four altered

amino acid residues located outside of the core motif, namely G955A, M986L, F998L and N1008T. Their role in enhancing indigoidine production is to be investigated. Especially, it will be interesting to see whether these mutations enhance the interaction of the engineered synthetase with the amino acid substrate facilitating further dimerization into indigoidine, as predicted by the model. In general, detailed mutational analysis of functional and non-functional IndC variants and structural data will be invaluable in elucidating the effect of different residues on the overall NRPS structure and function.

Experimenting with different boundaries allowed us to extrapolate what seems to be a general "design principle" for creating functional T domains, that stems from the intimate interconnection between the preceding linker (to the A domain) and the T domain itself. Indeed, when exchanging a T domain together with its own preceding linker, we had a greater rate of success (Fig. 6). Bringing in also the following T-TE linker reduced this success (Fig. 6). Based on these findings we recommend that only the preceding linker should be carried along when exchanging natural T domains. Importantly, this rule for finding optimal linker sequences is implemented in the computer-aided design software (NRPSDesigner) we created for allowing users to easily assemble NRPSs of desired composition (manuscript in preparation).

It is striking that the SynT1 variant can lead to the highest amount of blue pigment, all the while allowing the bacterial host cells to grow better than those expressing native IndC (Fig. 4, c and d). Indeed – at least under certain growth conditions – we observed an anti-correlation between growth and indigoidine production for most other engineered

IndC variants (Fig. 4d), which is in agreement with recent observations that indigoidine has antimicrobial properties[24]. So far, the mechanism of this antimicrobial effect is poorly understood. Interestingly, our modeling results suggest that SynT1 does not exhibit a higher L-glutamine cyclization rate as compared to native IndC. The key optimized parameter is rather related to the dimerization of cyclized glutamine to form the blue pigment (Fig. 5 and Fig. S3). We therefore hypothesize that cyclized glutamine (indigoidine precursor) might be also a toxic species alongside with indigoidine. This would explain why the SynT1 variant produces more indigoidine and grows better than the wild type, which has to deal with more precursor than SynT1 (Fig. S3). Interestingly, this hypothesis seems to be corroborated by findings that show how cyclized glutamine (also known as pyroglutamic acid) inhibits the growth of the hyperthermophilic archaeon *Sulfolobus solfataricus*[31] and energy production and lipid synthesis in rat cerebral cortex[32]. Further experiments are needed to clarify the role of cyclized glutamine on *E. coli* growth.

In conclusion, we were able to demonstrate that it is possible to replace the T domain of the indigoidine synthetase IndC with other natural as well as synthetic T domains, while preserving or even enhancing its functionality. Our success rate was particularly high when applying the "Consensus" method for construction of synthetic T domains, leading to three out of four novel synthetase variants with maintained or improved functionality. We would argue that this approach is not limited to the T domain of IndC or to single-module NRPSs, thereby expanding the NRPS engineering avenues. However, a rigorous experimental verification is needed to prove this hypothesis.

This study sets the basis for future rational engineering of NRPSs

that could improve the production of important bioactive peptides such as antibiotics[16,33,34]. Indeed, the availability of antibiotics towards which bacteria have not yet developed resistance is becoming a major concern[35,36]. Creating novel synthetases by domain swapping could, for instance, allow generating libraries of modified natural antibiotics with better properties than the original molecule[37,38]. On a more speculative line, a future in which entirely synthetic, tailor-made peptides can be produced is envisaged, provided we are able to put together modules in a desired way[39,40]. Creation of novel peptides from scratch by module combination has been shown to be possible, but until now only short peptides could be created[41]. This suggests that, despite the intrinsic modularity of the non-ribosomal peptide synthesis pipeline, complex synthetic NRPSs cannot easily be engineered by putting together modules *à la carte*, for example due to structural constraints. Yet, a successful example of multiple module exchanges within the daptomycin synthetase scaffold[38] points to the possibility of combining different modules in a way that their interaction remains functional even if these modules have not been put together by natural selection. We speculate that domain swapping – especially using synthetic domains as exemplified by this study – might be used to reduce the evolutionary distance between modules to be fused, thus increasing the chances of getting functional engineered enzymes.

Acknowledgments

We thank Blaine A. Pfeifer for supplying us with the *E. coli* BAP1 strain, Mohamed A. Marahiel for the *B. parabrevis* strain, and Martin Fussenegger for plasmid pMM64 containing the *bpsA* gene. We also thank Janet Lei for critical reading of this manuscript and for useful

discussions. This work was supported by the Klaus Tschira Foundation, Boehringer Ingelheim, the Helmholtz Initiative in Synthetic Biology, the BioQuant, the DKFZ and the University of Heidelberg.

Notes and references

1 Caboche, S., Leclere, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *Journal of bacteriology* **192**, 5143-5150, doi:10.1128/JB.00315-10 (2010).

2 Marahiel, M. A. Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. *Journal of peptide science : an official publication of the European Peptide Society* **15**, 799-807, doi:10.1002/psc.1183 (2009).

3 Vining, L. C. Functions of secondary metabolites. *Annual review of microbiology* **44**, 395-427, doi:10.1146/annurev.mi.44.100190.002143 (1990).

4 Marahiel, M. A., Nakano, M. M. & Zuber, P. Regulation of peptide antibiotic production in Bacillus. *Molecular microbiology* **7**, 631-636 (1993).

5 Sieber, S. A. & Marahiel, M. A. Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chemical reviews* **105**, 715-738, doi:10.1021/cr0301191 (2005).

6 Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chemical reviews* **97**, 2651-2674 (1997).

7 Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chemical reviews* **106**, 3468-3496, doi:10.1021/cr0503097 (2006).

8 Lambalot, R. H. *et al.* A new enzyme superfamily - the phosphopantetheinyl transferases. *Chemistry & biology* **3**, 923-936 (1996).

9 Doekel, S. *et al.* Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. *Microbiology* **154**, 2872-2880, doi:10.1099/mic.0.2008/020685-0 (2008).

10 Doekel, S. & Marahiel, M. A. Dipeptide formation on engineered hybrid peptide synthetases. *Chemistry & biology* **7**, 373-384 (2000).

11 Stachelhaus, T., Schneider, A. & Marahiel, M. A. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science* **269**, 69-72 (1995).

12 Owen, J. G., Robins, K. J., Parachin, N. S. & Ackerley, D. F. A functional screen for recovery of 4'-phosphopantetheinyl transferase and associated natural product biosynthesis genes from metagenome libraries. *Environmental microbiology* **14**, 1198-1209, doi:10.1111/j.1462-2920.2012.02699.x (2012).

13 Brachmann, A. O. *et al.* Triggering the production of the cryptic blue pigment indigoidine from Photorhabdus luminescens. *Journal of biotechnology* **157**, 96-99, doi:10.1016/j.jbiotec.2011.10.002 (2012).
14 Takahashi, H. *et al.* Cloning and characterization of a Streptomyces single module type non-ribosomal peptide synthetase catalyzing a blue pigment synthesis. *The Journal of biological chemistry* **282**, 9073-9081, doi:10.1074/jbc.M611319200 (2007).
15 Ehmann, D. E., Shaw-Reid, C. A., Losey, H. C. & Walsh, C. T. The EntF and EntE adenylation domains of Escherichia coli enterobactin synthetase: sequestration and selectivity in acyl-AMP transfers to thiolation domain cosubstrates. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 2509-2514, doi:10.1073/pnas.040572897 (2000).
16 Mootz, H. D. & Marahiel, M. A. The tyrocidine biosynthesis operon of Bacillus brevis: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *Journal of bacteriology* **179**, 6843-6850 (1997).
17 Johnston, C. W. *et al.* Gold biomineralization by a metallophore from a gold-associated microbe. *Nature chemical biology* **9**, 241-243, doi:10.1038/nchembio.1179 (2013).
18 Duchaud, E. *et al.* The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. *Nature biotechnology* **21**, 1307-1313, doi:10.1038/nbt886 (2003).
19 Punta, M. *et al.* The Pfam protein families database. *Nucleic acids research* **40**, D290-301, doi:10.1093/nar/gkr1065 (2012).
20 Nakano, M. M., Corbell, N., Besson, J. & Zuber, P. Isolation and characterization of sfp: a gene that functions in the production of the lipopeptide biosurfactant, surfactin, in Bacillus subtilis. *Molecular & general genetics : MGG* **232**, 313-321 (1992).
21 Reuter, K., Mofid, M. R., Marahiel, M. A. & Ficner, R. Crystal structure of the surfactin synthetase-activating enzyme sfp: a prototype of the 4'-phosphopantetheinyl transferase superfamily. *The EMBO journal* **18**, 6823-6831, doi:10.1093/emboj/18.23.6823 (1999).
22 Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E. & Khosla, C. Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science* **291**, 1790-1792, doi:10.1126/science.1058092 (2001).
23 Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. Multiple sequence alignment with Clustal X. *Trends in biochemical sciences* **23**, 403-405 (1998).
24 Cude, W. N. *et al.* Production of the antimicrobial secondary metabolite indigoidine contributes to competitive surface colonization by the marine roseobacter Phaeobacter sp. strain Y4I. *Applied and environmental microbiology* **78**, 4771-4780, doi:10.1128/AEM.00297-12 (2012).
25 Myers, J. A., Curtis, B. S. & Curtis, W. R. Improving accuracy of cell and chromophore concentration measurements using optical density. *Bmc Biophys* **6**, doi:10.1186/2046-1682-6-4 (2013).
26 Pruitt, K. M. & Kamau, D. N. Mathematical-Models of Bacterial-Growth, Inhibition and Death under Combined Stress Conditions. *J Ind Microbiol* **12**, 221-231, doi:Doi 10.1007/Bf01584194 (1993).

27 Ansari, M. Z., Yadav, G., Gokhale, R. S. & Mohanty, D. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic acids research* **32**, W405-413, doi:10.1093/nar/gkh359 (2004).

28 Bachmann, B. O. & Ravel, J. Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods in enzymology* **458**, 181-217, doi:10.1016/S0076-6879(09)04808-3 (2009).

29 Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* **39**, W339-346, doi:10.1093/nar/gkr466 (2011).

30 Blin, K. *et al.* antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research* **41**, W204-212, doi:10.1093/nar/gkt449 (2013).

31 Park, C. B., Lee, S. B. & Ryu, D. D. L-pyroglutamate spontaneously formed from L-glutamate inhibits growth of the hyperthermophilic archaeon Sulfolobus solfataricus. *Applied and environmental microbiology* **67**, 3650-3654, doi:10.1128/AEM.67.8.3650-3654.2001 (2001).

32 Silva, A. R. *et al.* L-pyroglutamic acid inhibits energy production and lipid synthesis in cerebral cortex of young rats in vitro. *Neurochemical research* **26**, 1277-1283 (2001).

33 Keller, U., Lang, M., Crnovcic, I., Pfennig, F. & Schauwecker, F. The actinomycin biosynthetic gene cluster of Streptomyces chrysomallus: a genetic hall of mirrors for synthesis of a molecule with mirror symmetry. *Journal of bacteriology* **192**, 2583-2595, doi:10.1128/JB.01526-09 (2010).

34 van Liempt, H., von Dohren, H. & Kleinkauf, H. delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase from Aspergillus nidulans. The first enzyme in penicillin biosynthesis is a multifunctional peptide synthetase. *The Journal of biological chemistry* **264**, 3680-3684 (1989).

35 Walsh, C. Where will new antibiotics come from? *Nature reviews. Microbiology* **1**, 65-70, doi:10.1038/nrmicro727 (2003).

36 Urgently needed: new antibiotics. *Lancet* **374**, 1868, doi:10.1016/S0140-6736(09)62076-6 (2009).

37 Crawford, J. M., Portmann, C., Kontnik, R., Walsh, C. T. & Clardy, J. NRPS substrate promiscuity diversifies the xenematides. *Organic letters* **13**, 5144-5147, doi:10.1021/ol2020237 (2011).

38 Nguyen, K. T. *et al.* Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17462-17467, doi:10.1073/pnas.0608589103 (2006).

39 Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282**, 63-68 (1998).

40 Planson, A. G., Carbonell, P., Grigoras, I. & Faulon, J. L. Engineering antibiotic production and overcoming bacterial resistance. *Biotechnology journal* **6**, 812-825, doi:10.1002/biot.201100085 (2011).

41 Mootz, H. D., Schwarzer, D. & Marahiel, M. A. Construction of hybrid peptide synthetases by module and domain fusions. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5848-5853, doi:10.1073/pnas.100075897 (2000).

42 Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PloS one* **4**, e6441, doi:10.1371/journal.pone.0006441 (2009).
43 Quan, J. & Tian, J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nature protocols* **6**, 242-251, doi:10.1038/nprot.2010.181 (2011).
44 Raue, A. *et al.* Lessons learned from quantitative dynamical modeling in systems biology. *PloS one* **8**, e74335, doi:10.1371/journal.pone.0074335 (2013).

Author contribution

All authors conceived and designed the study and discussed results. R.B., K.H., I.K., L.A. and D.N. performed experiments under the supervision of H.M., L.A., D.N. and B.D.V. N.I., I.K., L.A., T.H. and H.M. performed bioinformatics analyses. L.A. and I.K. conceived and performed mathematical modeling. B.D.V. wrote the manuscript with input from L.A., H.M., T.H., R.B. and D.N.

Experimental

Bacterial strains and growth conditions

Escherichia coli TOP10 cells (Invitrogen) were used for indigoidine expression experiments. *E. coli* TOP10, co-transformed with PPTase and IndC expression constructs were grown on LB agar plates or in LB media containing 50 µg ml^{-1} kanamycin and 30 µg ml^{-1} chloramphenicol at 30°C.

Plasmid construction

Details on the cloning procedure, bacterial strains used for PCR amplification of genes, primer sequences (Supplementary Table 2) and a list of all constructs used and created in this study (Supplementary Table 3) can be found in ESI section 1. In brief, all plasmids constructed in this study were created using an adapted circular polymerase extension cloning CPEC[42,43] protocol. Sequences encoding the PPTases Sfp, Svp, EntD and DelC were amplified from the corresponding bacterial genomes and cloned into an expression vector carrying a *lacI* promoter, a double terminator and a kanamycin resistance cassette, thereby producing constructs pRB1-4, respectively. Furthermore, a native IndC expression construct named pRB5 comprising a *lacI* promoter, the *indC* coding sequence, a double terminator and a chloramphenicol resistance cassette was created. For T domain exchange, the native IndC T domain in pRB5 was replaced by the positive selection marker ccdB, yielding the IndC scaffold construct pRB6. T domains were either amplified from the corresponding bacterial genomes, or T domains sequences were obtained as gBlocks (Integrated DNA Technologies, Coralville, Iowa). Subsequently, these T domains were introduced into pRB6 by CPEC, thereby replacing the ccdB selection marker, producing constructs pRB7 to pRB21.

Creation of synthetic T domains

Four libraries of homologous protein sequences were generated using NCBI protein BLAST (http://blast.ncbi.nlm.nih.gov) with standard parameters. Query sequences were, respectively: IndC (1), the IndC T domain with different restrictions in the query space (2+3), and five T domains from IndC-unrelated pathways (4). In query (4) for each T domain, the 50 best hits were determined and combined. Based on

multiple sequence alignment by ClustalO (http://www.ebi.ac.uk/Tools/msa/clustalo/), a consensus sequence was calculated for each homology library. Synthetic T domains were derived thereof by applying three different methods ("Consensus", "Guided Random", "Randomized Generation"), which lead to final sequences of varying homology to the wild type IndC. For details on query sequence generation, selection of hits, and methods for obtaining the synthetic T domain sequences, see ESI section 2.

Indigoidine production assay

E. coli Top10 cells were co-transformed with Sfp expression vector pRB1 and one of the IndC expression constructs pRB5 or pRB7-21. Cells co-transformed with pRB1 and an empty pSB1C3 vector served as negative control. Two days post transformation, 2 mL LB liquid cultures were inoculated in 14 ml snap-cap falcon tubes (BD) from single colonies (3 cultures for each IndC variant or control) and incubated at 30°C at 220 rpm shaking for 31 h. Afterwards, 200 µL of each culture were transferred into the well of a transparent 96-well plate (Greiner) and absorbances at 590 and 800 nm were measured in an Infinite M1000 PRO plate reader (Tecan). The absolute absorption of indigoidine per measurement was obtained as described by Myers *et al.*[25] Details can be found in ESI section 3.

Time-resolved data acquisition and mathematical modeling

Liquid overnight cultures of indigoidine-producing *E. coli* cells were diluted 1:1,000 in LB medium and grown in transparent 96-well plates

(Greiner) at 30°C inside an Infinite M1000 PRO plate reader (Tecan) in six technical replicates for up to 60 h, measuring absorbances at 590 nm and 800 nm every 5 minutes upon shaking. Data were pre-processed using a combination of Perl and R scripts to determine the indigoidine concentration (see ESI section 3). Only data sets with recognizable bacterial growth were used for model calibration. The mathematical model was based on ordinary differential equations. A detailed explanation of the model can be found in ESI section 4.

Parameter estimations and identifiability analyses were performed in MATLAB (MathWorks) using the D2D software package (https://bitbucket.org/d2d-development/d2d-software/)[44].

Statistical Analysis

To compare mean values obtained from the indigoidine production assay for the different variants, we performed one-sided, unpaired Welch's t test. Analysis was conducted with R.

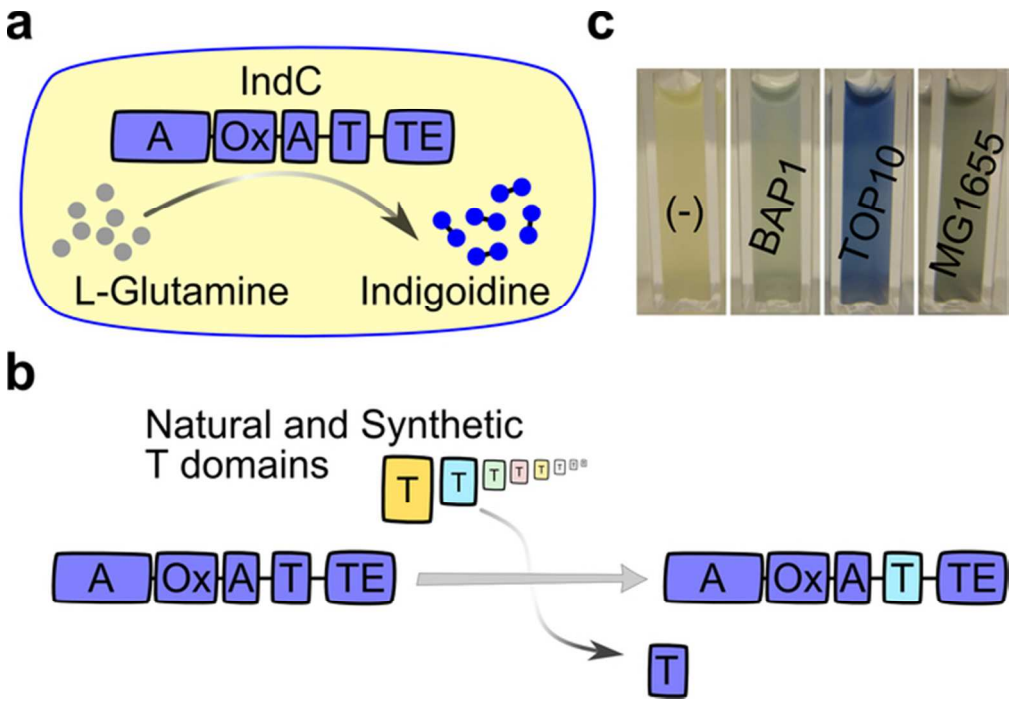


Figure 1: Engineering the indigoidine synthetase IndC by T domain exchange. (a) Schematic representation of intracellular indigoidine production. For simplicity the activating reaction by the PPTase is omitted. A-Ox-A is the adenylation domain with an internal oxidation (Ox) domain, T is the thiolation domain and TE is the thioesterase domain. See also Fig. S1 for the chemical reactions leading to indigoidine production. (b) Schematic representation of IndC T domain exchange. (c) Indicated E. coli strains were co-transformed with constructs encoding IndC and the PPTase Sfp and grown in LB medium at room temperature. After 40 h, cultures were transferred into cuvettes and photographed using a Canon EOS 600D digital camera. (-): E. coli TOP10 without indigoidine synthetase.
58x40mm (300 x 300 DPI)

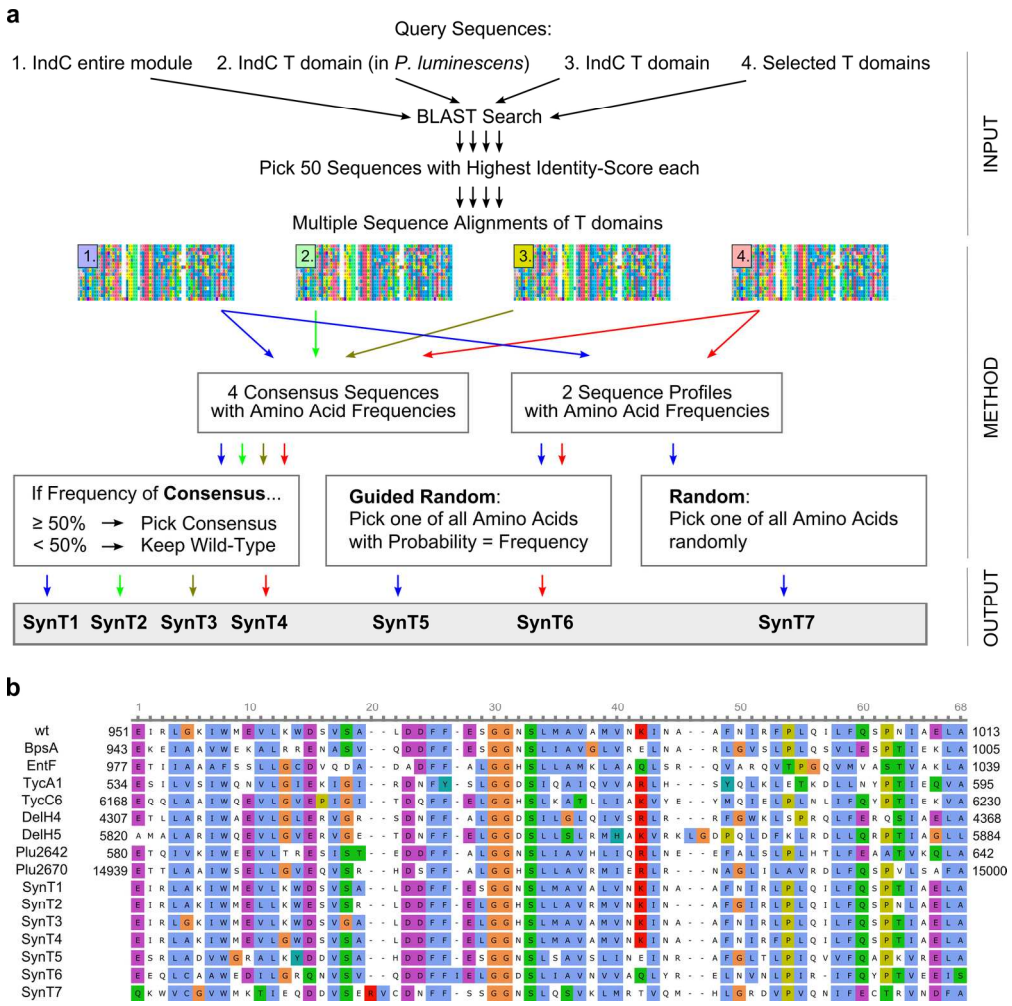


Figure 2: Workflow for the design of synthetic IndC T domains. (a) Schematic view of the workflow used to create synthetic T domains. The indicated sequences were used as queries in the BLAST algorithm to identify similar sequences. The 50 sequences with the highest identity score with respect to the query were considered for the following procedure. In query 4, for each of the five selected T domains the 50 best hits were determined and combined. These libraries of closely related sequences were then subjected to multiple sequence alignment (MSA) by ClustalO. Seven synthetic T domains were created from either the consensus sequences of the MSA or amino acid profiles determined thereof. Generation of the synthetic sequences was based on three different algorithms: The first one is favoring the conservation of the native IndC T domain at a certain position unless the frequency of a new amino acid within the MSA was >50% ("Consensus" method), the second is building novel T domain sequences based on the amino acid frequencies in the sequence profile ("Guided Random" method) and the third is randomly choosing amino acids from the MSA at a certain position independent of their frequency ("Randomized Generation" method). (b) MSA of the native IndC T domain and all natural and synthetic T domains used in this study (amino acid color code according to ClustalX23). Numbers at the beginning and end of the sequence describe positions of the T domains in the respective NRPSs.
176x174mm (300 x 300 DPI)

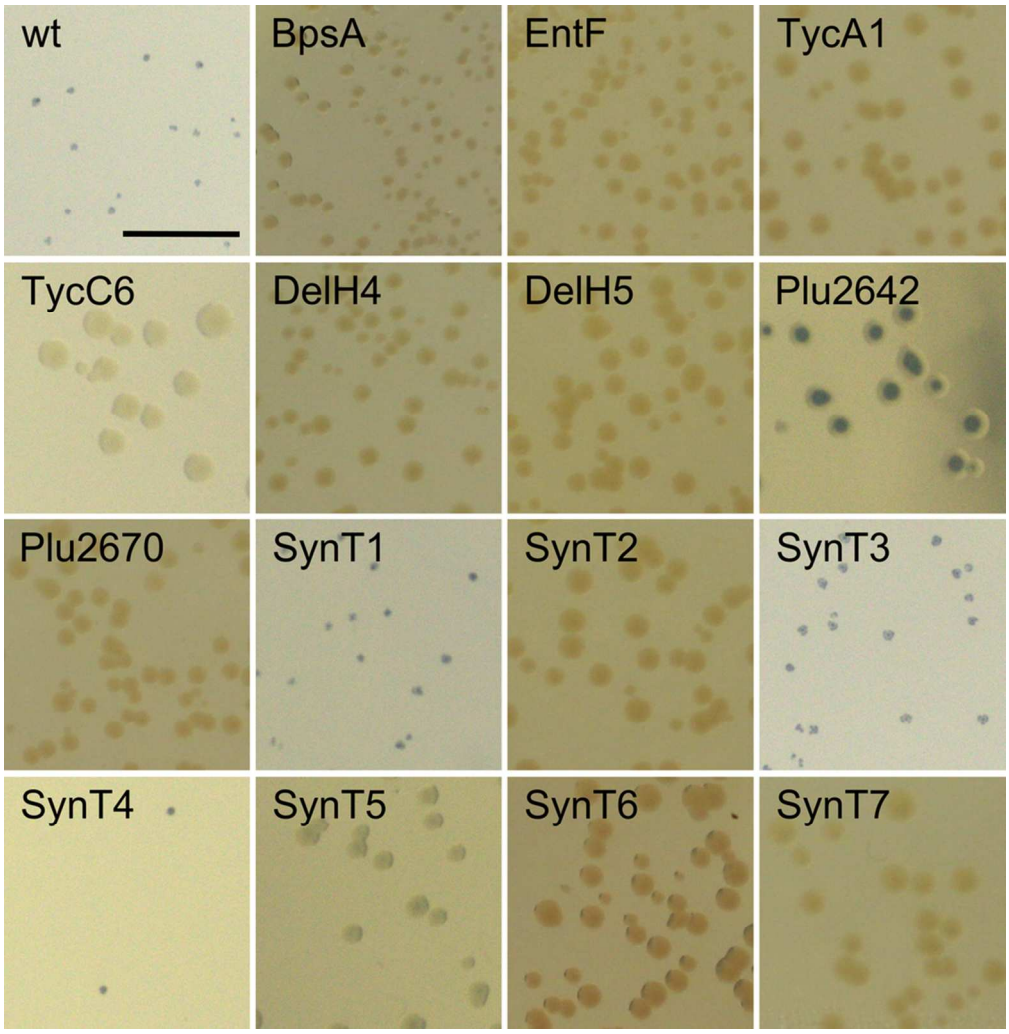


Figure 3: Phenotypic comparison of E. coli Top10 cells expressing IndC variants engineered by T domain exchange. Bacteria co-transformed with constructs encoding for the indicated IndC variants and the PPTase Sfp were grown on agar plates for 48 h at 30°C. Colonies were then photographed using a Canon EOS 600D digital camera. Scale bar = 0.5 cm. wt, wild type IndC synthetase.
89x91mm (300 x 300 DPI)

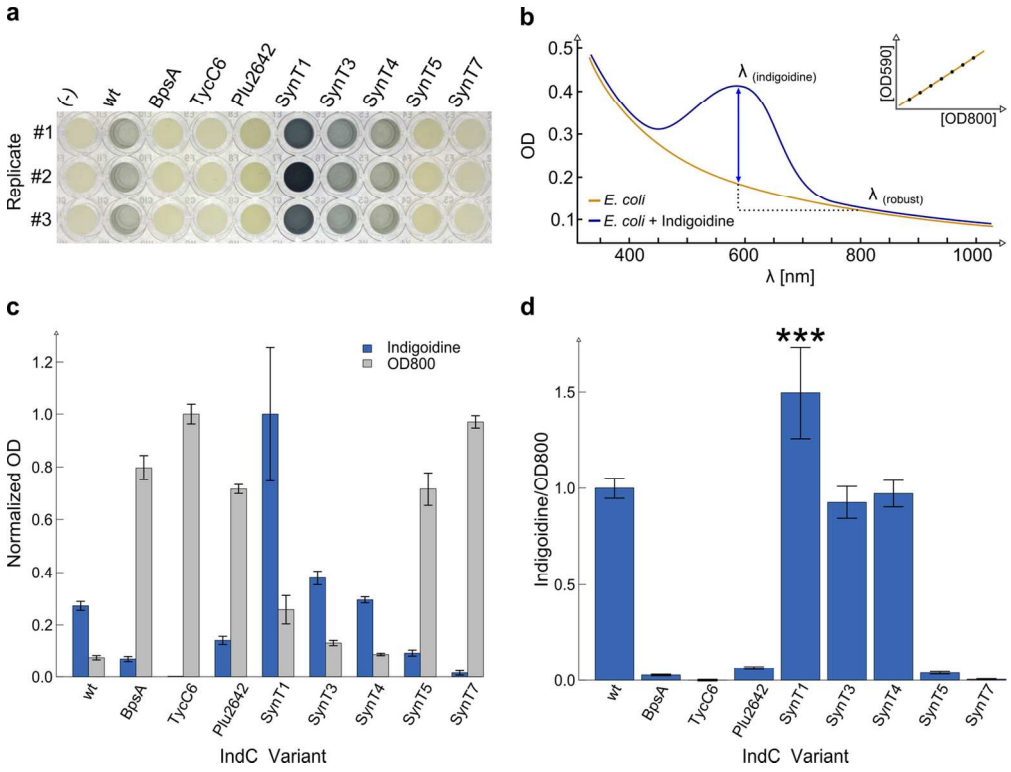


Figure 4: Tuned indigoidine production by engineered IndC variants. (a) Representative picture (acquired with Canon EOS 600D) of E. coli TOP10 liquid cultures co-expressing the indicated constructs and the PPTase Sfp in a 96-well plate format. Cultures were inoculated from single colonies and grown in falcon tubes for 31 h at 30°C. 200 µL of each culture were then transferred into a 96-well plate for measurement of indigoidine production in a plate reader. The three replicates correspond to three different colonies picked from the same plate for each construct. (b) Graph showing the contribution of indigoidine to the absorbance of the bacterial culture at 590 nm. Indigoidine amounts can be calculated with a formula that considers optical densities (OD) at 590 and 800 nm (see ESI section 3 for details). Inset shows the linear relation between OD590 and OD800 for a bacterial culture. (c) Bar plot showing indigoidine amounts (blue) and OD800 (grey) of E. coli cultures co-expressing the indicated IndC variants and Sfp after 31 h of incubation at 30°C. (d) Bar plot showing the indigoidine/OD800 ratio for the data shown in (c). *** indicates a P value < 0.0001. P value was calculated using the Welch's t test. (c-d) Error bars: s.d. from n = 3 independent experiments. In each experiment three independent colonies were measured for each construct.
136x102mm (300 x 300 DPI)

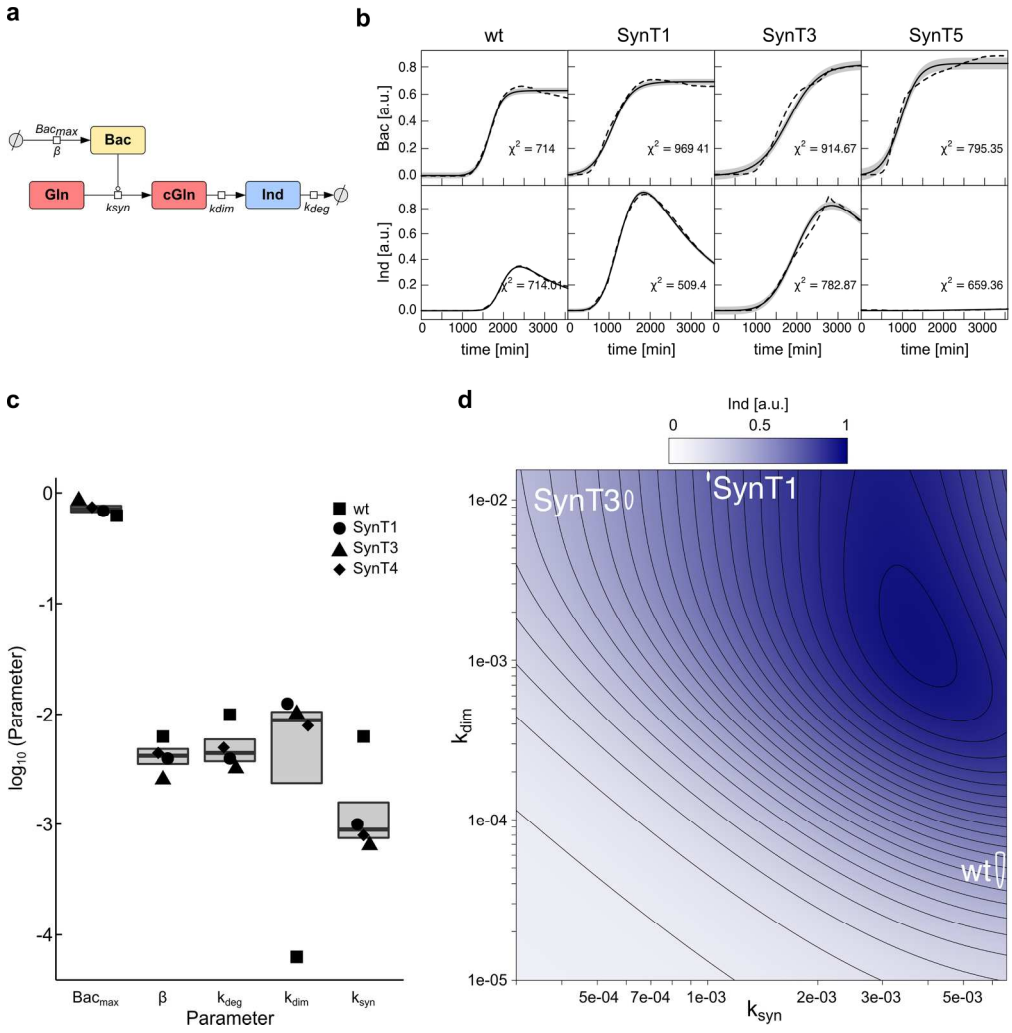


Figure 5: Quantitative dynamic modeling of indigoidine production. (a) Reaction scheme of indigoidine production. The first reaction represents bacterial (Bac) growth (rates β, Bmax), the second reaction represents conversion of glutamine (Gln) to cyclized glutamine (cGln) (rate ksyn), the third reaction represents dimerization of cGln to indigoidine (Ind) (rate kdim) and the fourth reaction represents degradation of indigoidine (rate kdeg). (b) Model calibration with time-resolved data for bacterial density and indigoidine concentration for the indicated variants. Solid lines depict model trajectories, with shaded regions showing the 95% confidence interval. Dashed lines depict experimental data. For all panels, one exemplary data set is shown with χ2 as a measure for the goodness of the fit. The number of data points for calibration was N=714. (c) Boxplot of the distribution of best estimates for parameter values obtained for the four IndC variants. (d) Surface plot color-coding indigodine yield simulated for ksyn and kdim. White circles represent 95% confidence regions for the two parameters of the indicated variants.
182x185mm (300 x 300 DPI)

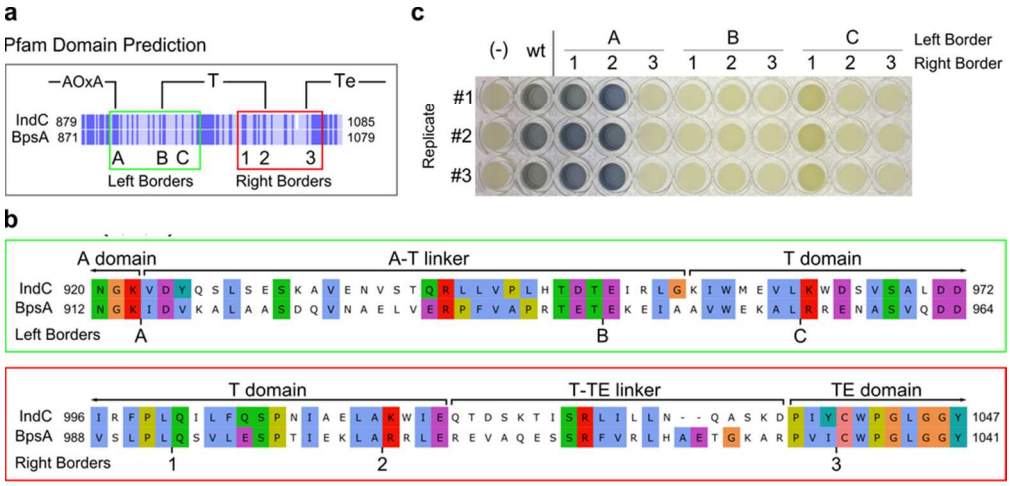


Figure 6: Impact of T domain boundaries selection on engineered synthetase functionality. (a) Sequence alignment of IndC and BpsA. Indicated are Pfam-predicted domain borders (above) and alternative boundaries experimentally tested. Sequence numbering is indicated for the first and last amino acid. (b) Zoom-in of the sequence alignment shown in (a). Amino acid color code according to ClustalX23. (c) Qualitative assessment of functionality of engineered synthetases with different T domain boundaries. Three single colonies of TOP10 cells co-expressing the PPTase Sfp and the BpSA variant with the indicated boundaries were used to inoculate liquid cultures. Cultures were grown for 31 h at 30°C in falcon tubes and then transferred into a 96-well plate for easy visual comparison of blue coloring. Picture was taken with a Canon EOS 600D digital camera.
86x41mm (300 x 300 DPI)