UNIVERSITY OF HAMBURG

DOCTORAL THESIS

Rewiring *E. coli* as an optimal vitamin production platform

Author:
Dóra VITAY

Supervisors:
Dr. Luisa GRONENBERG,
Dr. Hans GENEE,
Prof. Dr. Zoya IGNATOVA

Dissertation submitted to the Universität Hamburg in fulfillment of the requirements for the degree Dr. rer. nat.

2024

Biosyntia ApS
AG Ignatova

Defense Committee

Reviewer 1: Prof. Dr. Zoya Ignatova
Reviewer 2: Carlos G. Acevedo-Rocha, PhD
Oral Defense Committee:
1. Prof. Dr. Zoya Ignatova
2. Dr. Carlos G. Acevedo-Rocha
3. Prof. Dr. Paul Hudson
4. Prof. Dr. Ralph Holl
5. Prof. Dr. Peter Heisig
Date of Defense: 2025.01.20

Dissertation timeline

The disseration's experimental work was carried out mainly at the facilities of Biosyntia Aps, in Copenhagen, Denmark between 2018 July and 2021 July. Some experiments were carried out at the University of Hamburg, in Hamburg, Germany in 2019 September. Additional experiments were carried out at a CRO (Bioneer ApS) in Copenhagen in 2019 and 2021. The data analysis and thesis writing was carried out between 2020 July and 2023 December in Copenhagen and Stockholm.

Contents

List of Abbreviations

5-HTP	5-Hydroxytryptophan
ACP	Acyl Carrier Protein
Amp	Ampicillin
AMUSER	Automated DNA Modifications with USER cloning
ATP	Adenosine triphosphate
B. subtilis	Bacillis subtilis
CFU	Colony forming unit
CoA	Coenzyme A
Cas	CRISPR associated protein
CRISPR	clustered regularly interspaced short palindromic repeats
CRO	Contract research organization
DAPA	7,8-diaminononanoate
DBTL	Design Build Test Learn
ddPCR	dropdlet digital polymerase chain reaction
DNA	Deoxyribonucleic acid
DTB	desthiobiotin
E. coli	Escherichia coli
FACS	Fluorescence activated cell sorting
GFP	Green fluorescent protein
Glu	Glutamate
GMO	Genetically Modified Organism
HABA	4′-hydroxyazobenzene-2-carboxylic acid
IDT	Integrated DNA Technologes
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kan	Kanamycin
KAPA	8(S)-Amino-7-Oxononanoic Acid
kDa	kilo Dalton
L-trp	L-tryptophan
LB	Lysogeny Broth
LC-MS	Liquid chromatography - mass spectrometry
mMOPS	minimal 3-(N-morpholino)propanesulfonic acid media
MQ	ultrapure water
NADPH	Nicotinamide adenine dinucleotide phosphate
NGS	Next Generation Sequencing
OD	Optical density
P. putida	Pseudomonas putida

PCN	Plasmid copy number
PCR	Polymerase chain raction
pDNA	plasmid DNA
qPCR	quantitative PCR
R&D	Research and Development
RBS	Ribosome Binding Site
RNA	Ribonucleic acid
ROS	Reactive Oxygen Species
*S. cerevisia*e	Saccharomyces cerevisiae
SAH	S-Adenosyl homocysteine
SAM	S-Adenosyl methionine
SEVA	Standard European Vector Architecture
Spec	Spectinomycin
TE	Tris EDTA Buffer
TLC	Thin Layer Chromatography
USER	uracil-specific excision reagent
WGS	Whole genome sequencing

UNIVERSITY OF HAMBURG

Abstract

Biochemistry and Molecular Biology

AG Ignatova

Dr. rer. nat.

Rewiring E.coli as an optimal vitamin production platform

by Dóra VITAY

As DNA synthesis and metabolic engineering are becoming more accessible and modular, the usage of plasmid-based expression vectors in bacterial systems are more wide-spread than ever. This carries the question of the effect of the expression system on the performance of the created strains. In this thesis we investigated the current usage of multi-plasmid systems across various publications, as well as the methods used to report the potential effect of these systems. We investigated the performance of 2 and 3-plasmid systems consisting of the most widely used plasmid backbones with the expression of fluorescent proteins modeling the expression of cargo from the vectors, and investigated the change in their copy numbers. Lastly we conducted experiments on the effect of said plasmids in an industrial strain development scenario where we focused on the by-product formation in the biotin production pathway, which is expressed from a multi-plasmid system.

UNIVERSITY OF HAMBURG

Abstract

Biochemistry and Molecular Biology

AG Ignatova

Dr. rer. nat.

Neuausrichtung von *E. coli* als optimale Plattform zur Vitaminproduktion

by Dóra VITAY

Da DNA-Synthese und Stoffwechseltechnik immer zugänglicher und modularer werden, ist die Verwendung von plasmidbasierten Expressionsvektoren in bakteriellen Systemen weiter verbreitet als je zuvor. Dies wirft die Frage nach der Wirkung des Expressionssystems auf die Leistung der erstellten Stämme auf. In dieser Arbeit haben wir die aktuelle Verwendung von Multiplasmidsystemen in verschiedenen Veröffentlichungen sowie die Methoden untersucht, mit denen die potenzielle Wirkung dieser Systeme beschrieben wird. Wir haben die Leistung von 2- und 3-Plasmidsystemen untersucht, die aus den am häufigsten verwendeten Plasmid-Rückgraten bestehen, wobei die Expression von fluoreszierenden Proteinen die Expression der Fracht aus den Vektoren modelliert, und die Änderung ihrer Kopienzahlen untersucht. Zuletzt haben wir die Wirkung der genannten Plasmide in einem Szenario der industriellen Stammentwicklung untersucht, wobei wir uns auf die Bildung von Nebenprodukten im Biotinproduktionsweg konzentrierten, der von einem Multiplasmidsystem exprimiert wird.

1 Introduction

1.1 Cell factories and metabolic engineering

There is an increasing demand for sustainably produced ingredients across various industries, as many current production processes still rely on non-renewable resources. Fermentation, a process driven by microbial metabolism, has long been used for producing a wide range of chemicals, with its industrial applications gaining prominence over the past century. While fermentation has been integral to food production since ancient times—evident in early civilizations' use of fermented beverages and foods—the application of engineered microbes for industrial-scale chemical production has only advanced significantly in recent decades.

The concept of using cell factories to produce small molecules and chemicals has evolved alongside breakthroughs in genetics, metabolic engineering, and synthetic biology. Metabolic engineering aims to improve the production of economically valuable molecules through the genetic manipulation of microbial metabolism. While the discipline is a little over 30 years old, advancements in metabolic engineering have led to industrial-level molecule production benefiting multiple industries, including chemical, agriculture, food, pharmaceutical, and energy sectors. (Volk et al., 2023)

Despite significant advancements, many fermentation-based processes remain economically and industrially challenging due to factors such as yield limitations, substrate costs, and scalability issues. Shared environmental impact hotspots have been identified across industrial fermentation-based products, including biomass production, energy consumption, and end-of-life fate. Practical considerations for improving the sustainability performance of bio-based products made via industrial fermentation are essential to address these challenges. (Jullesson et al., 2015)

The fields of metabolic engineering and synthetic biology continue to play a crucial role in optimizing microbial strains, but the full realization of economically viable fermentation processes for many chemicals is still an ongoing pursuit. The next decade holds great excitement and challenges for metabolic engineers and synthetic biologists, with expectations for a growing repertoire of bio-derived products and emerging technologies to modulate and control cellular processes. (Yadav et al., 2012)

1.2 Vitamins

Vitamins, by definition, are essential compounds that are required in small quantities for the organism to grow and function. They are generally supplied in the diet since they are not synthesized or the level of synthesis is not sufficient to supply the necessary amounts to the organism. These compounds can be found in various plants and microbes, including those in the diet is the basic method of ensuring the necessary intake. Vitamins are used in the food and feed industry to counteract deficiencies in diet and enhance the nutritional value of products. Not only the food industry but the pharmaceutical and cosmetics industry has also been interested in adding vitamins to their products. Extraction from natively producing plants and microbes has been a limiting factor, therefore, the various fields turned to the chemical industry to be able to produce vitamins in industrially feasible quantities.

1.3 Industrial vitamin production

Bulk production of various vitamins has been historically carried out via chemical production, which can be an efficient method in the case of some vitamins, but mostly carries the burden of non-renewable chemicals and hazardous waste as by-products. The global market for fermented vitamins has been on the rise since the early 2000s and is expected to grow approximately 4% annually (Acevedo-Rocha et al., 2019).

B1 vitamin (thiamine) is a great example of limited cell factories for industrial production. The pathway participants in its biosynthesis in various microbes have been extensively studied (Jurgenson, Begley, and Ealick, 2009; Berstenhorst, Hohmann, and Stahmann, 2009) and understood. Even with

a deep understanding of the pathway participants, biosynthesis is tightly regulated by the translation and transcription levels. This tight regulation makes it necessary to engineer the expression of genes that potentially have organism-wide effects, which can negatively affect cell viability. Riboswitch-based biosensors (Genee et al., 2016) and development of transporters (Bali, Genee, and Sommer, 2018) based on said biosensors made it possible to increase the production of thiamine in *E. coli*. Some of the identified respective biosynthesis genes are dependent on the presence of iron-sulfur clusters and others are inhibited by various by-products, such as 5-deoxyadenosine and methionine. This shows that the challenge of industrially viable cell factories lies in engineering a balanced expression system. Due to the low price of the chemical production of thiamine, it is a challenge to create an industrially viable cell-based production platform. The target yield is relatively high compared to the currently available production strains and the extensive engineering of the organism carries a large development cost.

While vitamin B1 needs metabolic engineering breakthroughs to be viable for large-scale industrial production, there are other B vitamins that are successfully produced by fermentation in an industrial level. A great example of sustainable industrial vitamin production with cell factories is vitamin B2, riboflavin. Its chemical production uses toxic organic solvents (Eggersdorfer et al., 2012) but after the application of various mutagenesis methods on native producer microbes, they showed to be good candidates for industrial production of vitamin B2. The change of process to fermentation from chemical production reduced the production costs by 43% as well as created a more sustainable process (Revuelta et al., 2016). The improved strains made it possible for the fermentation process to compete with the chemical process in cost and feasibility. This led to the fact that in 2012, 75% of the worldwide produced vitamin B2 was produced via fermentation. Even though the riboflavin production enzymes are considered slow, the introduction of mutations in the flavokinase/flavin adenine dinucleotide synthase gene and additional stabilization of mRNA made it possible that the production strain is capable of stable expression of the genes of interest. The production strains are efficient, but there are still possible bottlenecks in the production. One of the main tools used in the production strain is riboswitches, to stabilize the previously mentioned expression levels. There are recently identified regulatory proteins, which could potentially interfere with the riboswitches and decrease their production. Another challenge in the production of riboflavin

is its reactivity. If there is no stable transportation system that would direct the produced compound through the cytoplasm onto the media, the intracellular concentration will result in ROS-related stress.

As shown in the examples above, there are multiple challenges in the industrial production of vitamins. In most cases, the chemical process achieves a higher yield than production via fermentation.. Even though the current yield is higher, the chemical production in the case of some vitamins is still costly and limited in volume. Chemical production is currently used for vitamins B7 and B1 is a high-cost low-yield process, therefore industrial and academic efforts are both utilized to create efficient cell factories.

Other than the cost and yield of production, different factors make fermentative production of chemicals a more attractive production method. One of these is the possibly lowered environmental impact. To measure the environmental impact of the process, one needs to look into the Life Cycle Assessment of the whole process from production to waste handling. Such analysis of processes was done for the production of riboflavin (vitamin B2) and showed that the fermentative process scores better than the chemical production in most variables. Switching to fermentation from chemical production reduces the impact on global warming, acidification, and ozone creation. It also uses 40% less energy than chemical production. On the other hand, eutrophication potential increased by 40% (Hohmann et al., 2016). Not all current bio-based production of vitamins are sustainable. Vitamin B12 is generally produced via fermentation, buts its downstream chemical processing includes the use of potassium cyanide. The use of such a hazardous substance demands stringent safety protocols to protect workers and prevent accidental exposure. Additionally, its presence complicates waste disposal and environmental remediation, as even small amounts can have significant detrimental effects on ecosystems. Regulatory compliance also becomes more challenging, as the use of potassium cyanide is heavily regulated to mitigate risks to human health and the environment.

As these production strains are considered genetically modified organisms (GMOs), there are strict regulatory restrictions on their use. Vitamins are considered additives for the food and feed industry. Therefore, the regulations of their product differ from the use of GMOs for food and feed in the European Union (Paracchini et al., 2017). The reason for the differences in regulation is the possibility of completely removing the GMO from the

final product, and that it is possible to ensure that the chemical structure is not changed. The challenge is rather to find a sustainable and efficient downstream processing method. During downstream processing, the product needs to be purified from non-desirable impurities, based on its target market regulation.

1.4 Design-Build-Test-Learn engineering cycle

The Design-Build-Test-Learn (DBTL) engineering cycle is widely used in most fields of engineering. Its use in the field of metabolic engineering gives a systematic and modular method for improving cell factories (Carbonell et al., 2018). With the development of the synthetic biology toolbox, it is possible to create systematic and product-independent strategies for improving strains for the production of our compounds of interest (Nora et al., 2019).

The increased understanding of various elements of the molecular machinery, as well as an increase in the throughput of most analytical methods caused a shift in the bottlenecks within the DBTL cycle in the last decades. With the development of genetic engineering tools, cloning of DNA constructs has become a high-throughput step of the experimental workflow. More widespread use of various omics methods has made it possible for scientists to have a broader view of the investigated organism. This abundance of information has also brought challenges. To be able to find the needle in the growing haystack of data, one needs to take the design of experiments and the effort to draw meaningful conclusions into consideration when planning a study.

In Chapter 2, the importance of the design step is in focus as we emphasize the various possibilities during the design phase, and which aspects can be affected in the later cycle steps. In Chapter 3, we demonstrate the effect of the variables described in Chapter 2 via examining the variability in the most used tools of synthetic biology, plasmids. Chapter 4 demonstrates two rounds of a strain development cycle aimed to understand metabolic changes in *E. coli* strains engineered to produce dethiobiotin (DTB), a biotin precursor molecule.

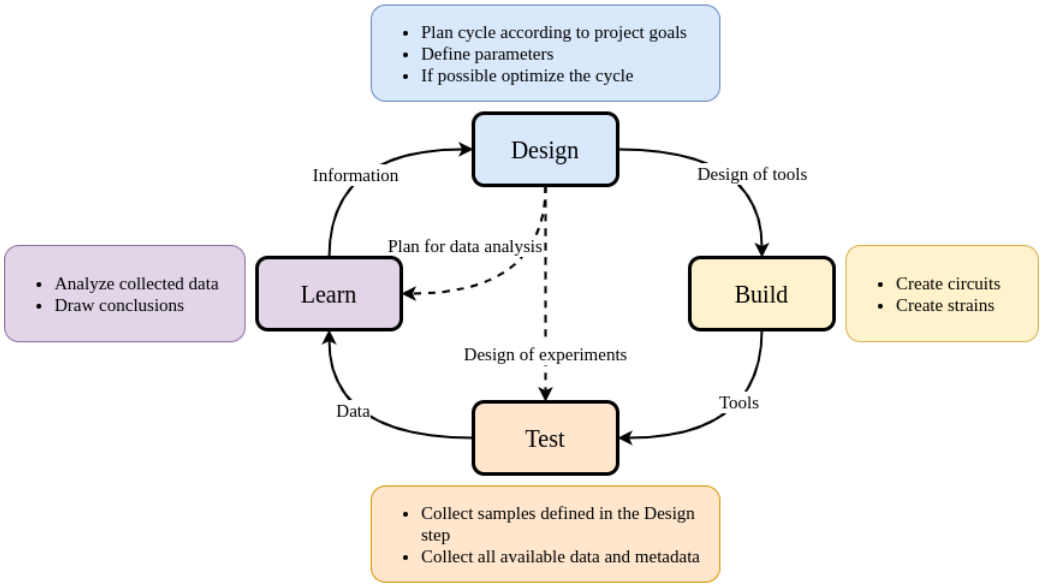


FIGURE 1.1: Strain development DBTL cycle

1.4.1 Design

When starting a new project, there are multiple decisions that the development team/researcher working on the project must make to have an efficient DBTL cycle (Figure 1.1). Using the cycle in metabolic engineering projects, the overall goal is generally to achieve or optimize the production of a target compound. First and foremost, one needs to have a target molecule, which is aimed to be produced via the cell factory. Once the compound of interest is identified, the possible pathways for bioproduction need to be assessed. This step can be done with an extensive literature search combined with metabolic modeling approaches. Choosing the production organism is also an important step in the initial design. The most common choices for organisms are native producers, which could be engineered to produce the compound in higher than native amounts to reach commercially viable yields. In most cases, choosing the native producer can be a great starting point, but the availability of genetic engineering tools can be limited. For production strains needing extensive engineering and/or does not have a native producer candidate, model organisms are the choice of scientists. Model organisms, such as *E. coli*, *S. cerevisiae* or *B. subtilis* have a developed toolbox and large amount of information about their regulatory systems. The advantage of this knowledge is extremely beneficial for the researcher, as most tools are well-documented and readily available. Other than toolboxes available, most of the model organisms have well-developed metabolic models (Passi

et al., 2021).

Metabolic models can be used to narrow down or discover targets for mutations without the need for time-consuming and expensive wet-lab experiments. During the design phase, not just the target genes, but the expression systems are also needed to be chosen. Depending on the project and on the chosen organism, one can have limited opportunities. For example, one can use plasmids or chromosomal integration to introduce genetic circuits and modifications. As we will describe in detail in Chapter 2 and 3 both methods have benefits and drawbacks.

Other than the aforementioned factors, multiple other aspects need to be taken into consideration during the Design step of the DBTL cycle. These factors usually depend on the goal of the DBTL cycle and the possibilities of the project. Decisions made in the design step are defined throughout the engineering cycle, therefore it deserves the time and effort. A well-planned study is straightforward to carry out in all steps and results in clear learnings that can be used to continue the development. One can be assessing the available methods that can be used during the various steps, or limitations in the number of samples analyzed per cycle. The availability of personnel is also a key factor for most studies, especially in industrial R&D settings.

Designing the experiments based on information from previous cycles and gathering information from outside the cycle are only parts of the design step. To be able to draw meaningful conclusions from the obtained data, it is essential to plan the analysis of the expected data. As we decided on the methods during the Design step, the collected data type should be already known. One can prepare a data analysis pipeline before the experiments start, and with a mock dataset can test the performance of the analysis. Making it possible to adjust the design of the experiments or the pipeline accordingly, before committing to the experiments.

In this thesis, Chapter 2 and Chapter 3 are targeted to strain developers planning to use multi-plasmid systems in their development workflow to aid their decision-making during the Design step. In 2 we describe the current state of multi-plasmid systems in metabolic engineering of *E. coli* cell factories. We show the development of expression vectors and examples of various multi-plasmid system challenges. Focusing on how one can ensure minimal noise from the expression system and what to monitor to assess the behavior of the multi-plasmid system.

In Chapter 3 we introduce an experimental example of multi-plasmid system instability and noise by investigating 3 of the most commonly used plasmids in *E. coli* metabolic engineering. We investigate the expression of fluorescent proteins from pBR322, p15A, and pSC101 backbones and monitor their copy number throughout multiple generations, to show the possible stability issues that could occur between DBTL cycles. We also emphasize the various methods used in the chapter to assess the behavior of the multi-plasmid system.

1.4.2 Build

In the next step of the engineering cycle, Build, the main task of the researcher is to prepare all described genetic elements and strains from the previous step. Since numerous designs can be generated with computational and modeling efforts in the design step, it was necessary to increase the throughput of the Build step. The possible throughout of this step increased in the last decades due to advances in metabolic engineering tools. A great example of such developments is the use of various methods to introduce modifications to the genetic elements, such as Gibson cloning and the CRISPR-Cas9 system. As new methods emerge, some workhorse methods are still used with little modification since the 1950s. One of these methods is the expression of genes from plasmid vectors.

Other than the possibility of creating precise modifications, the availability of automation of methods in the build step has also increased the throughput significantly. With the help of the integration of multiple modules on liquid handling modules, such as heating modules to carry out PCR or magnetic modules for automated purification of the genetic construct, cloning and transformation can be fully automated. Providing possibilities for automation can increase the walkaway time from the lab, while the experiment is still progressing. Giving more opportunities for the researcher to focus on other tasks in the DBTL cycle.

1.4.3 Test

The test step of the engineering cycle consists of exploring the behavior of the elements created in the previous step, with methods described in the Design

step. This step can be time-consuming, as it involves the characterization of strains, possibly in various experimental setups. All methods that were described in the Design phase need to be carried out in this step, from cultivation, through sample preparation for the various analytics methods, until the collection of the data originating from the measurements. The emerging field of lab automation has made it possible to increase the throughput of the step as well. There are various liquid handling technologies available to replace the laborious steps of sampling or sample preparation. In some experiments, such as directed evolution experiments or dilution-heavy assays, utilization of automation solutions can not only increase the throughput of the step but can also ensure reproducibility between experiments by eliminating the human error factor.

Aside from optimizing the throughput of the step, the inclusion of various controls in the measurements is also essential. With various controls used, a larger amount of samples can be handled, therefore the workload could increase within the step, but at the same time increase the reliability of the generated data. By comparing the unknown samples to the controls, the effects of multiple factors in the study can be investigated. In Chapter 2 we discuss the importance of controls when using multi-plasmid systems in more detail.

All data generated from the samples will be used in the next, Learn, step. To create an efficient workflow between these steps, all the measured data must be saved in an organized way, with all the metadata generated. By storing this data in a secure, organized, and traceable way, it is possible to use it outside this DBTL cycle. By creating a collection of data, future Design steps can be more efficient as more data is available during the Design phase of a different cycle.

1.4.4 Learn

In the closing step of the DBTL cycle, the goal is to conclude if the hypothesis described in the Design step could be investigated during the cycle and with what result. During this step, one analyzes the data collected in the previous steps. If high-throughput methods were utilized during the cycle, there is a possibility that a large amount of raw data was collected. To find

meaningful differences in the dataset, that are caused by the introduced varieties of genetic circuits or process conditions, one can prepare a data pipeline, as mentioned in the Design step. Depending on the available tools for the analyst, the data analysis can be done with vendor software available with most analytical machines, spreadsheet editors, and by writing custom scripts in various programming languages. In reality, these methods are generally combined during the data analysis and can be delegated to different members of the research team, alternatively outsourced. The utilization of reproducible data analysis has been the focus of research and development as the amount of data is increasing per publication. It is routine that in addition to describing it in detail in the Methods sections in case of a publication, the data pipeline is provided with the paper and the raw data can be requested from the authors. This also gave rise to sharing data pipelines between research groups and made it possible for projects like Galaxy (Afgan et al., 2018) or De Novo DNA (LLC, 2020) to become standards and be usable by scientists with little to no knowledge of coding. As for software tools for analyzing biological data, there are large well-maintained libraries for most programming languages. Such as Biopython (Cock et al., 2009) for Python and Bioconductor (Gentleman et al., 2004) for R.

2 Multi-plasmid systems in *E.coli*

2.1 Abstract

From their identification in the 1950s to their widespread use in genetic engineering since the 1970s, plasmids became vital tools in synthetic biology and metabolic engineering projects. Their widespread use can be attributed to the fast development of straightforward cloning frameworks which made it possible to build and combine plasmids for any use. High-level expression of proteins is also easily achievable by using plasmids, as their copy number provides more copies of the gene in the cell than chromosomal integration. By combining multiple plasmids in a single strain, multi-plasmid systems became a tool to facilitate modular engineering of long or complex pathways. Modular plasmid systems and the possibility of high-throughput strain engineering provides a platform for increased screening efficiency. However, one of the underlying challenges is the predictability of plasmid systems. Understanding plasmids and predicting the level of noise and uncertainty in their use alone and combination introduced to a cell factory is valuable information. Studies can be even more robust by considering this knowledge during the design of studies and experimental workflows. In this review chapter, we aim to show examples of multi-plasmid systems and tools for assessing their behavior.

2.2 Introduction

As synthetic biology pushes the boundaries of cell physiology, many complex genetic circuits are created for different production purposes. There is an increasing need for efficient and predictable expression of these complex systems. Plasmids can be a great tool to effectively modularize complex pathways. The ability to separate pathway elements for characterization and get

a more detailed picture of cell physiology has been shown in multiple studies (Jeschek, Gerngross, and Panke, 2017; Li, Wang, and Zhang, 2019; Papin, Reed, and Palsson, 2004; Xu et al., 2013). The modules can be combined with multiple strain engineering strategies to achieve the target phenotype. One of these strategies is co-transforming the strains with modules, which reside on separate, but compatible plasmids, into the same strain and assessing their added effect. When combining two or more separately optimized modules, the production yields of the combined strain might be lower than expected from their individual performance. In order to optimize the modules for the production of the final product, multiple variables need to be considered. In case of a linear pathway, where the modules are producing the enzymes for the subsequent reaction, such variables can be the burden expressing multiple proteins, or even the maximum possible flux through the reaction. This aspect can be addressed via engineering of the expression levels via promoter engineering, or even protein engineering, which could solve possible substrate specificity problems. These variables need to be tuned during additional experiments in order to create an industrially viable production strain.

To be able to design studies where the combined effect can be confidently predicted, we need deep understanding of the target pathway as well as knowledge of how the coexisting of plasmids affect the performance of the pathway in the host organism. An often overlooked factor in the optimization strategy is the stability of plasmids. When compatible plasmids are present in the same cell, their copy number can differ from the literature value or their copy number when expressed alone. This unwanted effect can be caused by the additional pressure on the replication machinery or the additional burden of expressing proteins to accommodate the replication of the new plasmid. To minimize this effect, we need to understand the effect of multi-plasmid expression systems on the cell. In the following sections, we aim to describe the behavior of multi-plasmid systems with reported examples.

2.3 Plasmids as tools for strain development

In order to be able to confidently design experiments, we need the experimental tools to perform predictably. Otherwise, additional rounds of experiments are needed to confirm that the performance increase was not due to

noise from the utilized tools. When engineering strains for optimal production of a target compound, the metabolism of the strain is pushed towards an extreme, therefore, the burden of the tools utilized also needs to be taken into consideration when optimizing such a system. Understanding the exact burden the tools are creating for the engineered organism is essential to be able to assess the viability of the production strain. Many factors can play a role in how a synthetic genetic circuit performs in a host organism. Different interactions with the host can influence the design and application of the genetic circuit (Cardinale and Arkin, 2012).

Plasmids, were first defined in 1952 by Joshua Lederberg (Lederberg, 1952). With the discovery of the double-helix structure of DNA in 1953 (J. D. Watson and F. H. C. Crick, 1953), the interest for such extrachromosomal units bloomed. Naturally occurring plasmids are mostly non-essential for the organism, but provide some sort of fitness benefit. Therefore, the organism does not actively try to get rid of the additional burden, as it provides a positive effect, while still taking resources from the organism. One of the most common positive effect for the organism that can be obtained by plasmids is the availability to resist various toxic compounds that are present in the media, such as antibiotics. Organisms have been shown to share plasmids with horizontal gene transfer, ensuring the survival of not only the individual organism but the population. Naturally occurring plasmids vary largely in size and usually also carry virulence factors, machinery that ensures their ability to be transferred to other organisms. In the 1960s, a wide variety of plasmids were isolated. In 1973, Cohen and Chang transformed in vitro generated plasmids into bacteria (Cohen et al., 1973), which opened the door for a wide variety of applications.

Plasmids can be used to express all necessary elements to introduce a heterologous pathway for producing the target of interest. However, there is an additional element that need to be introduced to the strain in order for the plasmid to be maintained across generations. The selection pressure, which ensures that only cells which contain the plasmid are remaining in the culture. There are multiple methods to ensure such behavior, but the most common is the use of antibiotic selection markers. By the addition of an antibiotic agent to the growth media, only cells which produce the protein to neutralize the antibiotic agent, which is encoded on the plasmid, can survive. However, during industrial-scale fermentation, the use of antibiotics is

generally avoided. The reason for their unpopularity is the need for an additional potentially expensive element in the production media, increasing the cost of the broth. Furthermore, when antibiotics are used, there may be strict regulations on treating the waste after fermentation, as well as the need to remove all traces of antibiotics from the product.

On the other hand, in research and development of production strains, the use of plasmids is standard practice in case of bacterial expression systems such as *E. coli*. In other cases such as *S. cerevisiae*, due to its ability for heterologous recombination, modifications to the genome are not as challenging and time-consuming to carry out as it is in *E. coli*. Therefore, in case of *S. cerevisiae* the standard practice is more towards creating direct genome modifications. Plasmids are still used to deliver various tools such as sensors or plasmids encoding CRISPR/Cas systems for editing or regulatory purposes. Introduction of genes via plasmids is still significantly faster and higher throughput than genomic integration, speeding up the DBTL cycle and increasing the design space that can be constructed.

Learning and understanding more about plasmid segregation and the reason for the incompatibility between plasmids would make it possible for engineers to design their expression system with more confidence and less noise (Paulsson, 2004; Tal and Paulsson, 2012; Wong Ng et al., 2010), increasing data quality and ability to learn and predict from data. The main challenge in engineering a strain for industrial fermentation is its efficiency in transforming raw material to product. Since most target compounds are either not natively produced or only present in small amounts, the engineering can be extensive. The amount of expressed protein in the cell largely depend on the copy number of the gene present in the cell. As well as on the levels of transcription, which are mainly regulated by the promoter, RBS sequences and their regulators. One of the great advantages of plasmids is that they quickly allow for changes in copy number, relative to achieving it with genome integration. Commonly denoted as high, medium, and low copy numbers, the plasmid copy number per cell (PCN) usually averages > 300+, 300 > PCN > 10, and 10 >.

Popular plasmids that we use are a product of vigorous engineering in the 1970s (Sutcliffe, 1979). A great example is one of the most popular plasmids used for protein expression in *E.coli*, pBR322 (Sutcliffe, 1979). As shown

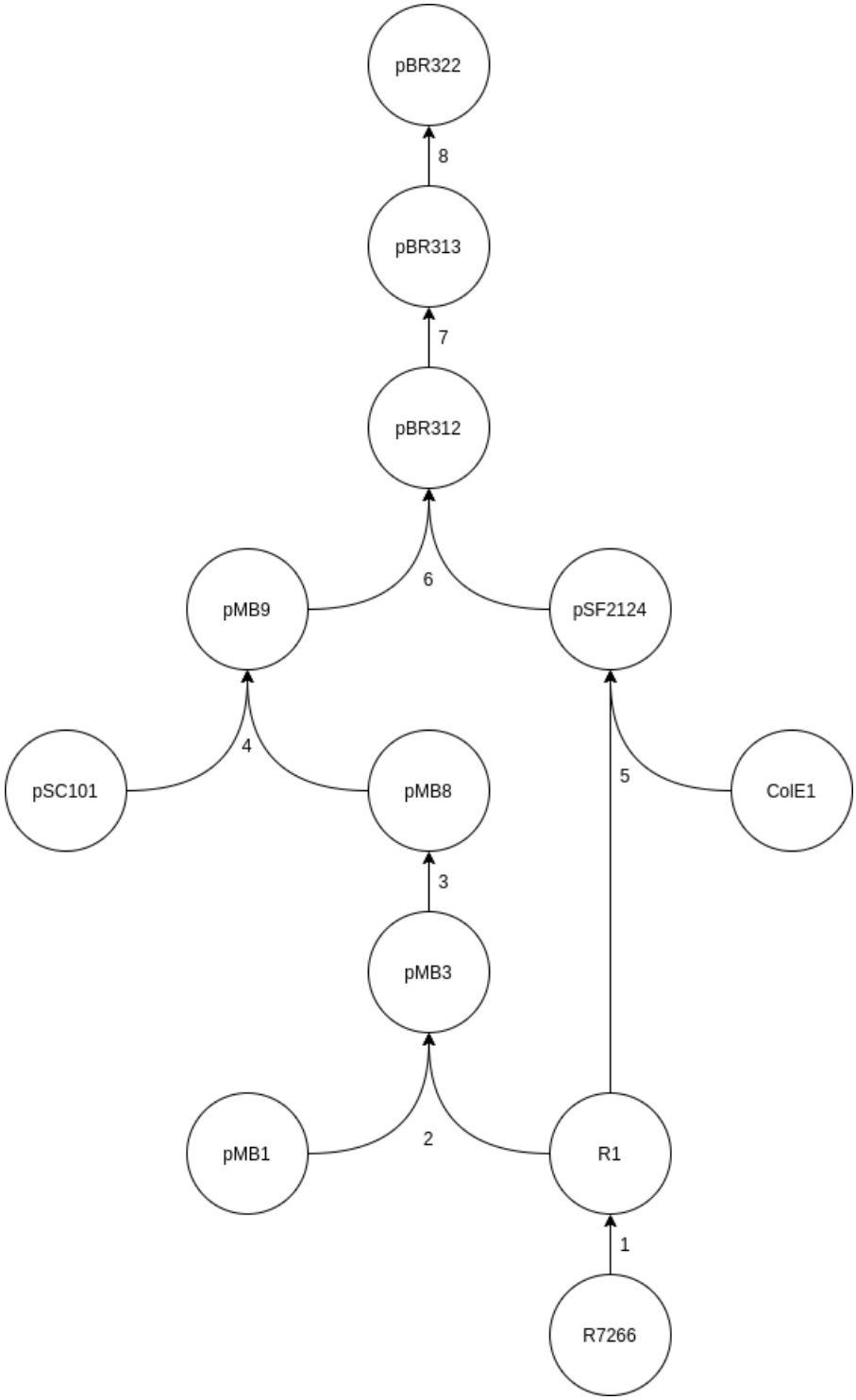


FIGURE 2.1: Development of pBR322 cloning vector. Figure adapted from Sutcliffe, 1979. 1: Meynell and Datta, 1967, 2: Betlach M et al., 1976, 3: Rodriguez et al., 1976, 4:, 5:, 6, 7: Bolivar et al., 1977b, 8:

in Figure 2.1, pBR322 was created in 1977 (Bolivar et al., 1977a), but its ancestors were isolated more than a decade before. In 1965, the R7268 plasmid was isolated, then renamed to R1 (Datta and Kontomichalou, 1965). A variant of R1, R1drd19, was isolated in 1967, which had de-repressed mating transfer (Meynell and Datta, 1967), hence the plasmid lost its ability to be transferred between cells other than during replication. Later, the same year they combined the Th3 transposon, present on the R1 variant with pMB1 plasmid to create pMB3 (Betlach M et al., 1976). A miniaturized version of pMB3, pMB8, was created. EcoRI* fragments from pSC101 were combined with pMB8 into pMB9 (Rodriguez et al., 1976). Based on pMB9, Bolivar et al. created a range of plasmids, combined with the Tn3 transposon, originating from R1 variant R1drd19, such as pBR313 (Bolivar et al., 1977b). Fragmentation and religation of pBR313 resulted in the widely used pBR322 (Bolivar, 1978). From the 11 mentioned predecessor plasmids, pBR322 can be reconstructed using only 3 of them. pMB1 and pSC101 containing the origin of replication and the *Tn3* transposon containing the Ampicillin resistance genes (Sutcliffe, 1979).

Plasmids have developed to be the backbone of the synthetic biology toolbox (Nora et al., 2019) as they facilitate testing and application of other synthetic biology tools. They are easy to handle during experiments, as there are ready-made kits for every step of the workflow. The streamlined methods make it possible to focus on the cloning efforts to ensure a wide range of plasmids available. With the development of the cloning toolbox, it is possible to create a wide variety of different plasmids, even libraries within a short period of time. Automation of the plasmid cloning and transformation can be easily achieved with the help of liquid handling robots. The utilization of laboratory automation on various scales can make the plasmid engineering a high-throughput step in the engineering cycle (Chao et al., 2017). In order to create the target strains, it is essential for the tool (in this case of plasmid) to be easily integrate to the production system (the cell).

2.3.1 Standardization and nomenclature

A major issue when designing plasmid systems is the lack of standardized nomenclature for different iterations and variations of the same origin. Due to the lack of unified nomenclature and the wide variety of derived origins

with different plasmid copy numbers, just the nominal annotation of the origin is not enough to identify a plasmid's expected copy number. This problem has already been identified in the early days of synthetic biology. As a solution, Novick proposed a standard nomenclature in 1976 (Novick et al., 1976). In many publications, since the widespread availability of DNA sequencing, standard practice was attaching the sequence of each used plasmid as supplementary material. This practice has given an opportunity to collect published plasmids into open databases. Recently, the practice of submitting the plasmids to AddGene repository, where the samples undergo independent sequencing to verify the submitted sequence and are later available for researchers. Standardization would make it possible to share and validate results between laboratories reliably. Creating databases of the standardized parts would make automated assembly based on design of experiment approaches the standard way of plasmid assembly. Regardless of the development made with plasmids since the 1970s, most publications still adhere to the arguably outdated nomenclature described by Novick. There are also recent efforts to standardize plasmids such as BioBricks (Shetty, Endy, and Knight, 2008) or SEVA (Silva-Rocha et al., 2013; Martínez-García et al., 2015).

As of today, there are a wide variety of practices utilized to publish plasmid sequences in order to ensure reproducible science. One of the most community driven approaches is the documentation and deposition of the plasmid into the AddGene database (Addgene, 2024). Where scientists can also request the deposited plasmid and obtain them as an agar-swab in order to be able to propagate and purify the requested plasmid in-house. In order to be able to estimate the performance of the plasmid in the target system, one needs to be able to estimate the plasmid copy number and the strength of the promoter and ribosome binding site (RBS). Even if the promoter and RBS strength is known, and the expression levels can be estimated with various models, plasmid copy number depends on various factors. Therefore, in databases, where plasmids are available for purchase, there is mostly a broad definition of copy numbers, namely low, medium and high. This categorization is rather subjective, but generally low corresponds to 1-5 copies, medium to 10-50 copies and high copy numbers generally are in the 100+ copies per cell range.

The challenge in standardization and description of plasmids in detail

lies in that a large variety of plasmids exist, and new varieties can be created easily. Therefore, one can estimate the behavior of individual plasmids when they are in the cell without the presence of another plasmid. As the use of modular design in metabolic engineering approaches became more widespread, mainly in order to increase throughput of the DBTL cycle, there are more examples of combinatory use of plasmids. Although there is no openly available database where the plasmid compatibility is mentioned in detail. It would be beneficial for scientists who are planning to use modular engineering to be able to assess which plasmid combinations to use in order to reach a stable and predictable system.

A great example for the result of standardization efforts is the pLannotate software published in 2021, which utilizes multiple feature databases in order to annotate plasmids purely from its DNA sequence (McGuffie and Barrick, 2021). The authors combined the usage of GenoLIB genetic parts' database, FPBase fluorescent protein database, SwissProt database for coding sequences and Rfam for noncoding RNA sequence annotations. Based on their combination of searches and alignments to various databases, after establishing a complex scoring system, the authors created a tool that can reliably annotate the plasmid sequence. Although the annotations are present, it still will not give detailed information about the plasmid copy number to be expected when using the annotated plasmid. pLannotate was not the first plasmid annotation tool that have emerged. There are predecessors for the solution, that required the creation of a feature database from already known sequences (Roosaare et al., 2018, Wishart et al., 2023).

For designing plasmids *in silico* based on the publicly available repositories, Timmons et al. created a repository-based plasmid design tool, which utilizes these deposited plasmid parts and creates protocols for assembly of the designed plasmid (Timmons and Densmore, 2020) It is a great tool to use but solely relies on the correct submission of metadata and sequences to the utilized databases. Also, this solution only applies the backbones based on annotation and does not give any information about the predicted performance of the designed plasmid. In order to fully automate the plasmid design and be able to utilize these aforementioned tools for *in silico* plasmid design, the curation of these databases is essential.

2.3.2 Understanding of replication machinery

In order to understand the possible variability introduced by plasmids, one needs to understand the replication of the utilized plasmid. It has been concluded that the main factor is the origin of replication (Nordström and Dasgupta, 2006). However, the main burden on the cell originates from the expression of the cargo regardless of copy number (Corchero and Villaverde, 1998). As plasmids have been heavily modified during development, genetic elements that are responsible for the regulation of plasmid copy numbers has been identified. The copy number is also dependent on the host (Wegryzin 1999. Santos-Lopez 2017) and growth conditions (Lin-Chao and Bremer, 1986).

Novic suggested the use of incompatibility groups to differentiate between types of segregational machinery (Novick, 1987). As he pointed out, incompatibility means that the ratio of cohabited cells to cells containing only one type of plasmid will decrease significantly faster per generation. It is possible to keep plasmids that carry the same origin but different cargo in the same strain. However, extensive selection pressure is needed (Alnoch et al., 2018; Wierzbicki et al., 2016; Yan et al., 2017).

In order to create vectors with desired copy numbers, regulatory elements in the origin of replication need to be fine-tuned. A great example is creating copy number variants of the pSC101 backbone. Point mutation of in pSC101 in amino acid position 93 (Glu) has been identified in 1991 to increase the copy number of the plasmid 4-5 fold. In a study in 2008, the authors further investigated the modifications in *repA* position 93 in order to modulate the copy number of pSC101. When Glutamate was changed to Arginine, the plasmid copy number increased to around 240, while still retaining the compatibility while expressed in the same cell as ColE1-type plasmids. Interestingly, in the study they only investigated if the two plasmids are retained in the cell, but their copy number was not measured in co-expressed strains. The authors suspect that these mutant pSC101 plasmids have a defective dimerization mechanism, therefore their replication is not well regulated (Peterson and Phillips, 2008).

In another recent publication, authors reported an induction system where the copy number of pUC19 could be modulated between 30-300 copies per cell (Rouches et al., 2021). The authors utilized the previously collected

knowledge about the PCN control in pUC19 plasmids, and targeted the machinery in order to modulate the copy number. They targeted the expression of the priming RNA with switching it to an IPTG-inducible promoter. In the absence of the inducible promoter, the plasmid is present in 270 copies per cell. With the inducible promoter introduced, they modulated the plasmid copy numbers to 1-50 copies per cell. They also showed that with the help of site directed mutagenesis on the -35 -10 region of the priming RNA promoter, that the behavior of the inducible plasmid copy numbers are reproducible. The authors successfully showed the possibilities of copy number modulation while maintaining the structure of the origin of replication and only modifying the expression levels of the participants. It is suspected that the modifications did not affect the compatibility of the plasmid, but the authors do not mention this aspect of the newly created vectors in the article.

2.3.3 Practical use of plasmids

There are multiple ways of introducing plasmids to cells. Two of the most used methods are chemical transformation and electroporation. In both methods, the cells are prepared to receive the plasmids, and we rely on diffusion to deliver the plasmid from the surrounding liquid media. These competent cells can be prepared in-house if the background strain already carries chromosomal modifications. There are also commercially available competent cells, which can be useful for various purposes, such as protein production, creation of large plasmid stocks, where the host is not the main concern. Depending on the preparation of the competent cells, the transformation efficiency can differ by magnitudes. Therefore, it is important to know if transformation efficiency is important for the study and plan the transformation mixture accordingly.

In the case of chemical transformation or in other name 'heat-shock' transformation, the cell membrane is made permeable during the preparation (Mandel and Higa, 1970; Cohen et al., 1973) by incubating them in calcium chloride. In order to increase the membrane permeability, other compounds can be added to the incubation mixture other than calcium chloride. Some of these chemicals such as calcium, manganese, potassium, cobalt, rubidium, dimethyl sulfoxide and dithiothreitol are described in (Hanahan and

Harbor, 1983). During transformation plasmid DNA and the prepared competent cells are incubated together, then the mixture is exposed to high temperatures for a short period, hence the name 'heat-shock'. In this period, the membrane becomes more permeable and the plasmids can enter the cells via diffusion. After the heat-shock the cells are placed on ice to make the cell membrane more rigid and limit the uptake of DNA, then rescued in rich media for a short time without selective pressure in order to support the division of the cells surviving the transformation effect. The rescued culture is grown in solid or liquid media with selective pressure in order to select for the transformants which took up the plasmid from the transformation mixture.

In order to prepare the cells for transformation with electroporation, they are washed multiple times with deionized water in order to remove all possible salts from the cell surface and media, which could be interfering with the electroporation. During electroporation, the cells are treated with an electrical field to make plasmid diffusion into the cell possible (Fiedler and Wirth, 1988; Dower, Miller, and Ragsdale, 1988). Short, high-voltage electrical field pulse is applied to the competent cell - plasmid mixture. The electrical field creates transient pores on the cell membrane, where the plasmid molecules can enter via diffusion. The same rescuing method is used for electroporated cells as for chemically transformed cells.

In order to ensure that the plasmid has been taken up and propagating in the target strain, the use of selection pressure is generally applied. There is usually a marker gene on the engineered plasmid in order to be able to use the selection pressure to select those colonies and samples that have taken up the plasmid. One of the most common selection pressure used in metabolic engineering are antibiotic resistance genes. By adding the antibiotics to the growth media, only cells that harbor the selection marker can grow and therefore ensure that the cells present in the growth media are most likely successfully taken up the plasmid. Antibiotics are practical and easy to use additions to a system, but there are applications where their usage is not feasible or allowed by regulations. There are multiple other selection systems that can be used in order to ensure that the plasmid has been taken up by the cell during the transformation event, and it retained. One of them is using strains that are auxotrophic and the marker encoded on the plasmid would rescue the growth of the cell via providing the missing gene or genes

that are necessary for the auxotrophy. As this method requires to match the marker and the strain where the plasmid is included, the method is generally utilized in production strains that are further in development.

Plasmid engineering solved several gaps in experimental procedures, where cloning is troublesome. Multi-host plasmids have enabled the scientists to assemble the circuit in a model organism, and when the circuit is ready, transform it into the target species. Maximum possible expression of a heterologous protein is a key use-case in many experiments. Therefore, the development of plasmid-based expression systems for this purpose has been an important field of engineering. For other use-cases, plasmids are engineered to modulate the expression of one or multiple target sequences. Besides plasmid copy numbers, which are mainly defined by the origin of replication carried by the vector, there are multiple ways of regulating the expression of the cargo. The expression of the cargo protein can be fine-tuned with various strength promoters/RBS regions and regulatory systems. By inducible or constitutive promoters of different strengths or additional regulatory circuits, which can provide steady expression levels regardless of copy number (Segall-Shapiro, Sontag, and Voigt, 2018). There are plasmids that are only part of the strain temporarily, while they are expressing their cargo in order to facilitate the engineering of the strain. Such as biosensors, which only reside in the production strain while the evaluation of the strain is ongoing. Other tools that use plasmids in order to express their machinery for modification of the strain are plasmids carrying the CRISPR/Cas9 system or lambda red genes. Lambda red technology utilizes the phage lambda red recombination system—comprising the Exo, Beta, and Gam proteins—to enable efficient homologous recombination in bacteria. This method facilitates precise gene editing, allowing for targeted insertions, deletions, or modifications by promoting recombination between linear DNA fragments and the host genome. Its utility in streamlining genetic engineering in E. coli and other organisms was notably demonstrated by 2000. In case of temporary plasmids, an important feature of the tool is the easy removal from the strain, also known as curing. The curing step is necessary to efficiently continue with the strain development with different plasmid tools without the possible interference from the sensor. Several methods have been developed to cure plasmids from bacterial cells. One common approach is chemical curing, which involves agents such as ethidium bromide or acridine orange that intercalate with DNA and inhibit plasmid replication (Summers, 1991).

Thermal curing is another method where bacterial cultures are grown at elevated temperatures that are non-permissive for plasmid replication, particularly if the plasmid contains a temperature-sensitive replication origin (Novick et al., 1976). Detergent-based treatments, such as exposure to SDS, can also disrupt plasmid maintenance. In addition, incompatibility-based curing uses a second plasmid carrying a similar replication origin to displace the resident plasmid. More recently, CRISPR-Cas systems have been harnessed to target and selectively degrade plasmid DNA, offering a precise and programmable method for plasmid curing (Bikard et al., 2013).

2.4 Possibilities and challenges of multi-plasmid systems

In metabolic engineering studies, one might need to test vastly different expression levels between the pathway genes. In order to be able to combine the pathway genes into modules that are easy to modify and introduce to the strain, the usage of multiple plasmids with compatible backbones are beneficial. By including the genes on different plasmid backbones, it is possible to cover a wide range of copy numbers and hence expression levels. It is also easier to construct multiple operons that can reside on separate plasmids and fine-tune their expression than to create a large plasmid with all genes and individually tune their expression. Also, as previously mentioned, the size of the plasmid has an effect on its copy number. Therefore, when the cargo of a plasmid is extended with new pathway genes, it is always necessary to characterize the plasmid. When using various plasmids with different copy numbers, the expected behavior is that the protein abundance in the cell correlates with the copy number of the plasmid expressing it. This assumption allows for balancing multiple genes of a heterologous pathway in a wide range of expressions. One of the main advantages of modularization is that the development can be parallel.

Plasmid backbones and some genetic circuits are readily available for order from various vendors, such as GenScript. AddGene, BioRad, etc. There are also services where the pre-designed plasmids can be ordered and therefore there is no need for in-house cloning. When one tries to combine these

off-the-shelf plasmids in the same strain, the expectation is that their presence will not interfere with each other, and they fulfill their function as they would be present lone in the cell. In case there is an effect of the combination, that is generally appointed to be caused by the combination of the cargos on the introduced plasmids and not to the possible difference in copy numbers. As these parts are generally developed independently of each other and require optimization when combined.

The introduction of multiple plasmids into the strain can cause unexpected behaviors. For example, not all plasmids can be used together. As previously described, the segregation system is the main factor in deciding if two plasmids will be able to coexist in the cell throughout generations. Based on their segregation systems, plasmid origins can be classified into incompatibility groups. As a rule of thumb, members of the same group should not be used in the same strain, as there is a high possibility of losing one of the plasmids after a few generations. This is generally followed in most studies where the goal is to create efficient cell factories. Even though the plasmids will be retained in the same strain if they are from different incompatibility groups, their plasmid copy number and ratio can still be subject to change depending on the combination of backbones used. The incompatibility groups recommend the use of plasmid pairs that have drastically different segregational systems. Regardless of which incompatibility group they belong to, all plasmids pose a burden on the cell. Therefore, the additional burden of including a further plasmid can influence the resources available for replication.

One method to easily modulate protein expression levels is to use inducible promoters. When the same inducible promoter system is present on multiple plasmids in the same cell, one must choose the inducer concentration carefully. As previously observed (Guido et al., 2006; Brewster et al., 2014) with the increased copy number of the promoter, higher concentration of inducer in the media was necessary to reach production levels than in case when of a single-plasmid system (Cardinale and Arkin, 2012; Nora et al., 2019). Interestingly this phenomenon is not always addressed or observed. In another reported case, there was no effect on the inducer concentration when an additional inducible plasmid was added to the strain (Han and Eiteman, 2017).

In metabolic engineering, promoters such as T7 and T5 are crucial for

driving high levels of gene expression to optimize metabolic pathways. The T7 promoter, recognized by T7 RNA polymerase, offers robust transcriptional activity and has been widely utilized to overexpress genes in microbial systems (William Studier et al., 1990). Similarly, the T5 promoter is known for its strong performance and, when combined with regulatory elements, provides precise control over gene expression. A key regulatory mechanism in these systems involves the lac operator-lacI repressor module; the lacI protein binds to operator sites, effectively blocking RNA polymerase binding and preventing transcription until an inducer is added (Baneyx, 1999). This tight regulation allows metabolic engineers to fine-tune gene expression, thereby optimizing metabolic fluxes and enhancing yields while minimizing metabolic burden and potential toxicity (2014).

These observations, where the addition of promoter copy does not affect the induction level, could be explained with plasmid instability or pathway imbalance. As new copies of the inducible promoters are added to the strain, it is expected that the amount of inducer needed per promoter copy does not change and hence the inducer concentration necessary to reach the same expression levels should be higher. When promoter titration is not observed, it is possible that the amount of promoters did not increase with the expected amount due to decreased plasmid copy numbers. Similar phenomena appear in transcription factor titrating, where the stronger the binding site introduced on a plasmid, the higher copy number could be measured (Brewster et al., 2014). A unique problem arises in cases where protein expression from two plasmids needs to be balanced in equimolar amounts. As compatible plasmids rarely have identical copy numbers, protein expression needs to be tuned by promoter strength. As shown in a recent study (Battesti and Bouveret, 2012), a balanced system could not be reliably achieved with current compatible plasmid pairs.

Biosensors are an example of a tool that is commonly plasmid-based and often combined with other plasmids libraries. Hence, the addition of the biosensor plasmid can influence the production of the target compound, and it is important to set up controls to be able to confidently measure the output from the sensor.

2.5 Examples of multi-plasmid systems in strain development

Here we describe real-world examples of use of multi-plasmid systems in strain development and highlight both benefits of this approach and problems as well as possible pitfalls. In Table 2.1, we collected a list of examples.

2.5.1 Use of plasmids to separately express different modules of a branched metabolic pathway

As previously described, metabolic pathways can be divided into modules based on various principals. A modular multi-plasmid system was employed to develop an E. coli 5-HTP production strain (Wang et al., 2018), and the effect of plasmid origins was taken into account. Initially, the two branches of the functional pathway were expressed using a pBR322 and p15A plasmid pairing. Later, it was discovered that expression from the pBR322 plasmid caused the accumulation of a pathway intermediate. As the backbone was changed from pBR322 to pSC101, the accumulation was less significant, and 5-HTP production increased by 25%. In a follow-up study, (Xu et al., 2020) made even more effort in alleviating the metabolic burden of multi-plasmid systems by integrating the genes carried on the pSC101 backbone into the genome. The motivation to move away from a multi-plasmid system in the final production strain was that the p15A-pSC101 combination became unstable. At 24h of fermentation, only 23% of the cells contained the p15A plasmid. The authors confirmed that the p15A plasmid loss only appears when the two compatible plasmids are expressed together, and the stability is rescued when they are residing in separate strains. They decided to integrate the low copy number plasmid and keep the medium-copy p15A backbone. This new strain architecture gave lower overall 5-HTP production rates but lowered the aforementioned L-trp accumulation and a more beneficial product/byproduct ratio, which makes the downstream processing easier to carry out.

2.5.2 Plasmid-based biosensors

Plasmid-based biosensors where the signal from the sensor circuit is correlating with the intracellular level of the compound of interest are considered a high-throughput screening method. During development, the quantification of compounds with analytical methods can be time-consuming and expensive. Instead of sampling all strains and measuring the concentration of compounds with various chromatography methods, it is convenient for the researcher to use a fluorescent readout and estimate the production capability of the created strain. This method of screening will not give detailed view on the production, but is a great indicator during the development and can save time and cost in screening. A great example of such system is Yu et al,2019 where they investigate the production of isobutanol, which is a candidate for transport biofuel (Yu et al., 2019). The authors created a biosensor based on a transcription factor, which is sensitive and selective enough for their purposes. They modulated the expression of the biosensor by placing it on different backbones. The authors observed that the high copy number plasmid (ColE1) was not suitable as a biosensor due to leaky expression of the reporter, in this case GFP. Low copy number backbone (pSC101) was also not suitable for their purposes as the expression of the reporter was too low in order to create the necessary dynamic range for the sensor system. A medium copy number backbone (ColA) was finally chosen as a suitable carrier for the biosensor. It provided the authors with the expected fluorescent response and therefore was suitable as a screening sensor. During the development of the production strain, the authors combined the medium copy number plasmid with a high copy number production plasmid (ColE1). They successfully increased the production of the strain and scaled-up the strain to small bioreactor. This study is a great example of successful biosensor development, proven successful with the production example. As there was no identified mismatch between the expected results with or without the production plasmid in any strains, the aspect of plasmid stability and compatibility was not mentioned in the article.

As it has been reviewed in Lim H 2018, plasmid copy number modulation is a commonly used method for tuning the dynamic range of the engineered sensory circuit (Kang et al., 2018). High throughput screening use of biosensors are generally applied to assess the relative performance of the

strains in order to choose the best possible producers from the current engineering cycle, there has been little focus given to the segregational instability of sensory plasmid systems. Even though the sensor plasmids will only reside in the strain for one development cycle, it is important to take the burden posed on the cell.

2.5.3 Scaling-up to fermentation scale

Because the initial results are generally small-scale batch cultures, a new challenge arises when the strains are scaled up to perform in a larger scale fermentation. In fermentation, the conditions significantly differ from the lab-scale tests. When strains present unexpected behavior during the scale-up, adjusting the process conditions or fine-tuning the genetic circuits are the mainly used tools. In most cases, the aim is to remove one or all plasmids from the final production strain. Therefore, plasmid stability is rarely assessed during the development of strains compared to the efforts taken to assess protein expression.

As previously discussed, another challenge is to scale up the production from microtiter plates to multi-liter fermentors during strain development. Depending on the techniques, the strains will go through more generations during fermentation than during a small-scale screening test. More generations allow for plasmid instability to appear, such as in the 5-HTP production example. During the optimization of the fermentation process, plasmid stability should be monitored. A recent study for resveratrol scale-up (Afonso et al., 2015) reported plasmid copy numbers in different tested conditions. The authors used a dual plasmid system for resveratrol production, where they expressed the two enzymes responsible for resveratrol production on two separate plasmids. They utilized pBR322, as a high copy number plasmid with inducible expression system to produce stilbene synthase and p15A with a constitutive promoter to express 4-coumaroyl CoA ligase in E. coli. As a great example of taking plasmid segregation into account during the characterization of production strains, the authors sampled multiple conditions and time points for PCN determination with qPCR. The measured copy numbers were higher at the later time point (30h) compared to the earlier sampling (22h), which correlated with the higher resveratrol production at 30h.

But after inspecting the separate copy numbers of the plasmids, it was observed that there is no correlation between the copy number and the resveratrol production. This observation was followed by the fact that the measured copy numbers for pBR322 were in most cases lower than the literature reported values. When expressed in the same cell, pBR322 copy numbers were even lower than the p15A plasmid copy numbers in most cases. The conclusion the authors draw is that the bottleneck of the resveratrol production could be the expression levels of stilbene synthase from the expectedly higher copy number plasmid. As reasons for the diverging results the authors name the possibility of the metabolic burden posing on the cell due to multiple plasmids being expressed and the race for the shared resources, even when the backbones are compatible. Other explanation could be the segregational instability of plasmids. The authors rightfully emphasize the importance of plasmid segregation monitoring during the strain development and suggest it as a routine method. This backbone combination was also problematic during the scale-up step in the 5-HTP example (Xu et al., 2020).

2.5.4 Importance of assessment

In many cases, instability using multi-plasmid systems is acknowledged but not further addressed. The goal of most projects is to optimize the production of the final product of interest and not optimize the plasmid expression (Alnoch et al., 2018; Dubois et al., 2019; Furuno et al., 2005; Hsia et al., 2016; Kim et al., 2014; Liu et al., 2018; McNerney and Styczynski, 2017; Moon et al., 2010; Munro et al., 2019; Thuronyi, Privalsky, and Chang, 2017; Wierzbicki et al., 2016; Yan et al., 2017; Yim et al., 2011). In cases where multi-plasmid systems introduce uncertainty or unexpected low yields, most strain developers avoid multi-plasmid systems in the next generation of strains. They generally switch to a single plasmid system by creating a plasmid that carries all necessary genes, or they integrate the cargo into the genome (Afonso et al., 2015; Alonso-Gutierrez et al., 2013; Han and Eiteman, 2017; Xu et al., 2020; Zhang et al., 2017). When switching to one plasmid system, there is a possibility that the created plasmid will become large and its replication and expression pose a more considerable burden on the cell than the burden of two plasmid systems.

When plasmid performance relating copy number is not assessed, there is a chance that false conclusions will be drawn regarding the performance and possibilities of the modified pathway genes. Collecting data about the plasmid copy number and/or plasmid stability would ensure that there is a clear distinction between the changes caused by the differences in the plasmid copy numbers and between the changes that are the result of the introduced genetic modifications. By routinely collecting data about plasmid metric, it would also be possible to gather more information about multi-plasmid system behaviors regarding various combinations of plasmids. This additional data could be used to better understand the behavior of multi-plasmid systems and could provide valuable information for planning multi-plasmid systems.

In an example study (Kim et al., 2019), precursor production was optimized separately but did not perform as expected once combined in the production strain. The authors assumed that the background strain could not bear the burden of phenol production. Nevertheless, they did not mention the possibility of plasmid instability. After adding a second "production" plasmid on top of the optimized precursor production plasmid, the actual copy number of the plasmids can be drastically different from their literature-reported copy numbers. As this aspect was not assessed by the authors, it is possible that some conclusions drawn from the experimental results did not fully explain the observed phenomena and the authors did not consider certain engineering strategies for further improvement of the strain. We can see the same phenomena in the resveratrol production example (Afonso et al., 2015). Therefore, it is worth assessing the changes in plasmid stability and copy number in cases like this. Another case where multi-plasmid systems can be hard to assess is cases like (Munro et al., 2019), where the information about the plasmid backbones, hence their compatibility, is missing from the report. This makes it hard to reproduce the data and build on the already gathered knowledge about the plasmid. As large data collection and AI-based learning becomes more prevalent it is important to not forget to collect data on this very important variable (plasmid copy number).

2.6 Methods to assess the behavior of multi - plasmid systems

In metabolic engineering approaches, choosing the measurements for the Test step of the engineering cycle during the Design step is essential. One needs to assess what measurements are the best representatives of the system variables. The main variables that are generally measured in studies are the product of interest, which is a direct measurement for the productivity and the growth of the strain, which can be a proxy for the wellness of the production strain. It is possible to use different medium- to high-throughput analytical assays or analytical chemistry methods to measure the produced compound. In order to assess the viability of the strain in the later stages of scale-up, it is common to assess the same variables of the strain in multiple conditions Afonso et al., 2015. When these conditions are assessed, such as media additions or process condition changes, the observed change is generally attributed to the effect of the change on the pathway. In some cases the observed change could be the result of a shift in plasmid copy numbers. If this aspect is not taken into consideration, the researcher will draw the conclusions based on the available methods and might miss the opportunity to exhaust all options for optimization of the system.

When utilizing a modular approach to assemble a large pathway in a strain, one should assess the productivity and stability of the strain regarding plasmid retention and copy numbers per module as well as in a combined state. If one can define performance per module, the combination of different modules could be predicted. When assessing plasmid copy number or any gene copy number, it is important to take the cell cycle and doubling time into consideration. Copy numbers can drastically differ before replication, where genome integrated genes can be present in 2 copies, and plasmids are also enriched. It is recommended to sample the investigated strains at the same growth phase for comparison. Due to differences in growth rate, these times can differ per strain.

2.6.1 Choosing controls

Choosing the correct control strains for comparison is also vital for generating reliable and reproducible results. It is recommended to always include a

strain with the same origin and selection marker without cargo as a control to decouple the effect of plasmid addition from the effect of a cargo gene.

2.6.2 Measuring plasmid stability

There are multiple ways one can measure plasmid stability and copy numbers. Indirect measurements are generally easy to carry out, but as their name suggests, they are rather a measurement of the product of the plasmid. This can involve quantification of the product of response from the plasmid cargo. One example is measuring the fluorescent signal when the cargo is a fluorescent protein. The most common test for plasmid loss is plating some cultures on selective plates and noting the CFU. This will give an estimate of the fraction of the population that contains the plasmid. Indirect measurements will give a proxy of the expression from the plasmid but will not give information about the stability of the plasmid copy number. In order to assess plasmid copy numbers and stability directly, one must target the plasmid itself.

Measuring the copy number via qPCR-based methods on purified cell pellet is the most common method to measure plasmid stability and copy number directly. qPCR primers target a unique sequence in the plasmid, which is not present in the genome, then another set of primers target a genome-specific region that only exists in one copy in the genome. By measuring the abundance of these PCR products in the samples, one can directly correlate how many copies of the plasmid is present per chromosome. When utilizing this direct qPCR measurement, one must be wary of how the DNA is extracted from the cell culture. If purified with commercial kits, the obtained PCN may be higher than the expected/literature values due to enrichment of the plasmids relative to chromosomal DNA during extraction. To avoid bias, one can prepare the samples with crude boiling. The method highly depends on the PCR efficiency and in most cases it is assumed 100% after careful design of the probes. In most cases in order to reach quantifiable amount of plasmid DNA in the samples, multiple series of dilutions are needed for the same sample. This increases the amount of reactions to be run, which is already doubles for every plasmid that is contained in the strain. As to be able to measure a 2-plasmid system with qPCR, one needs to create 3 different reactions from the same template. One for the housekeeping gene quantification in order to be able to normalize the PCN based on the chromosomal

copy number, and one reaction for each plasmid contained. The measurements could be parallelized with multiplexed qPCR reactions, which need further optimization and development for each assay.

In a recent article (Jahn et al., 2014) a new method for plasmid copy number determination was described for *P. putida*. The method is based on the above described qPCR method, but provides a more robust method, without the need for tedious dilution steps. It uses digital droplet PCR method, which uses a single PCR reaction divided into smaller amounts and carries out the PCR reaction separately. In a follow-up article (Jahn et al., 2016) the authors investigated widely used and available plasmid backbones from the SEVA collection and expressed a fluorescent protein as cargo. The standard plasmids were expressed in E.coli and were undergone FACS in order to classify the population into subpopulations based on their fluorescence levels. The authors also measured the PCN of each subpopulation with ddPCR and observed that heterogeneity was present in most cases regarding the fluorescent signal. For most investigated plasmid backbones, the measured values were towards the low end of previously reported plasmid copy numbers. Interestingly, heterogeneity was present in most backbones. Even plasmids that showed normal distribution in their fluorescent values (ColE1), their plasmid copy number was ranging from 9 to 123.

Plasmid copy numbers can also be measured via NGS methods. Although the extraction method can again introduce significant bias, the method is still used to evaluate relative levels between samples. With crude boiling as an extraction method and taking the growth rate into account, plasmid copy numbers determined via ddPCR and NGS sequencing showed the same fold changes.

2.7 Recommended factors to consider when designing multi-plasmid systems

In this section we aim to collect the factors that one should consider while designing a multi-plasmid systems with the purpose of monitoring the behavior of the plasmid system. We aim to give guidance for strain engineers and collect examples with references that can be useful to consider during the Design step of the DBTL cycle.

In strain development, the goal is to balance the expression of genes to produce the compound of interest in an optimal yield. The burden of the introduced modifications is also aimed to be minimal, to create a viable cell factory for industrial production. The reason to use multi-plasmid systems in the development of a production strain can be varying. Pathways can be modularized chronologically (George et al., 2014) or based on various principles (Jeschek, Gerngross, and Panke, 2017; Papin, Reed, and Palsson, 2004). Modules are isolated elements of the pathway. Therefore, their behavior can be individually assessed in the strain. Understanding the module behavior gives a greater chance to improve their performance in the system.

During parallel development of the modules, one can quickly create various variations and libraries of plasmids to test. Introducing a similar amount of modifications in the chromosome is challenging and often more time-consuming. In order to ensure the presence of multiple plasmids throughout generations, plasmids with compatible segregation systems should be used. The stability and copy number of compatible plasmids are usually assumed stable throughout the development.

2.7.1 Origin of replication

To scale plasmid copy numbers to the desired amount is the primary way of creating new vectors. Usually, these modifications are targeted to regulatory regions on the origin of replication. For example, ColE1 origin contains a Rom gene, which is responsible for regulating the copy number. When the gene is deleted, the copy number of the modified plasmid increases. Since these plasmids are still in the same incompatibility group and contain the same essential origin, they might be mentioned as "ColE1-plasmid" in publications, without mentioning the lack or presence of Rom. In some cases, it is beneficial to be able to modulate the plasmid copy number during growth.

When choosing the origins of replication, one tends to use what is at hand in the lab. Using readily available plasmid modules in the lab creates a standard for that backbone in that lab, which is excellent for reproducibility within the research group. However, reusing the same module throughout the years and multiple plasmid preparations might introduce unwanted changes in the sequence. Such as leftover padding and overhang modules around the necessary origin of replication elements. Due to these hereditary

modifications, when one annotates the origin of replication in a publication, the length and sequence of the flanking regions can be different.

Generally, the rule of thumb is that one should not use two different plasmids with the same origin in the same strain. Their stability decreases drastically, and the production of the target compound will also be unstable throughout multiple generations. However, there are examples of combining identical or incompatible origins in the same strain and maintaining stable expression (Alnoch et al., 2018; Wierzbicki et al., 2016; Yan et al., 2017). This can be achieved with severe selection pressure, which poses its metabolic burden. Such harsh selective pressure can be a high concentration of multiple antibiotics, where the antibiotic marker is placed on the plasmids separately (Schmidt et al., 2012). With this strategy, the surviving cells will have to contain both plasmids to produce the antibiotic markers. This method only works in situations where the marker functions intracellularly and the cells cannot rescue its surrounding cells via diffusion of the product of the marker.

2.7.2 Expression system

Expressing the gene of interest from plasmids in E.coli can be carried out with various expression systems. The least engineered version of an expression system consists of the native promoter of the expressed gene on a plasmid backbone (with selection system and origin of replication). As engineering of genetic circuits developed, various promoters with known relative strengths were developed. These constitutive, but varying strength promoters allowed to fine-tune the expression of cargo from the plasmid and allowed for a lower level of modulation than changing the plasmid origin of replication to a different (higher or lower) copy number version. In the 1980s researchers developed multiple inducible promoter systems. One of the first was the lac promoter system, where the addition of lactose or IPTG regulated the expression level of the protein that were encoded after the promoter structure on the plasmid. Similar inducible systems were developed for various inducer molecules, which allowed to include multiple inducers in the same cell factory. Hence, allow the use of less cloning in order to create a wide range of expression levels for the same elements of the vector. Other than using external molecules to regulate the expression level of the cargo, the use of transcription factors have also been investigated and used for various

purposes. In case of using transcription factor regulated promoters, one can allow for self-regulation from the organism. Alternatively, one can also encode the transcription factor on the plasmid with an inducible promoter and by modulating the level of the transcription factor not just affect the expression of the plasmid-based proteins, but also the expression of native genes on the chromosome. Recently the use of various CRISPR-based techniques have also been developed in order to regulate the expression of heterologous and native genes. In the case of expressing a CRISPR-based system in the strain, one generally encodes the machinery on a plasmid, which is then expressed on top of the additional modification of the system. Therefore, it is essential to be able to estimate the expression of various plasmids in the same cell. In a recent example, where Tong et al. described a versatile toolkit for E. coli engineering using CRISPR-prime editing, they used a 3-4 plasmid system in order explore the effects of mutations on a single nucleotide level both on plasmid and chromosome-based targets. The authors mentioned that the multi-plasmid system must consist of compatible backbones, therefore their choice of 3 compatible plasmid backbones were p15A, ColE1 and CloDF13. These plasmids are in separate incompatibility groups, but there is no further investigation in the study about their stability together. This example also highlights that the plasmid instability and compatibility is not investigated routinely, but rather only in cases where its effect is already disturbing the development of a technique (Tong et al., 2021).

2.7.3 Selection systems

In order for the plasmid to be maintained in the cell regardless of its carried cargo, a selection pressure is usually applied during the design of the experiments. In metabolic engineering applications the main method of providing selection pressure is encoding an antibiotic resistance gene on the plasmid and later adding that antibiotics to the growth media. Hence, ensuring that only cell that carry the resistance gene and express the necessary protein marker survive. In some cases, this selection method can be leaking, as the produced metabolite that ensures the survival can be secreted by the cells and a neighboring cell can take it up and survive without its ability to produce the molecule. In industrial fermentation and products created for medical use, the use of antibiotic markers is not always acceptable. While in the pharmaceutical industry, the reasoning is mainly the possibility of antibiotic

carryover to the final product, which could cause allergies in some patients. For fermentation industry, the possibility of antibiotic-resistant bacteria escaping the fermentation vessel and entering the environment poses a risk of horizontal gene transfer. Therefore, there are strict regulations on how the biomass is treated after the fermentation in order to minimize its possible negative effects on the environment.

Even thought the industrial use of antibiotic markers are frowned upon, their use in the development of the production strain is still significant. It is easy to create plasmids with encoding an antibiotic resistance marker and later add the antibiotic to the growth media. The effect on the native metabolism of the strain is minimal, and any strain can be used with the same plasmid. Therefore, their use in high-throughput strain development is still a cheap and secure way of providing selection pressure.

There are other methods for creating selection pressure towards strains bearing plasmids compared to non-plasmid-bearing strains. Some of these selection technologies require the modification of the background strain in order to be functional. The most common way of creating selection pressure without antibiotics is to knock out an essential gene from the chromosome of the engineered organism, then include that essential gene on the plasmid in order to ensure the survivability of the cell. It has been shown an effective method for ensuring plasmid presence in the cell. Although effective in most cases, when one needs to include multiple different plasmids in the cell, that also increases the amount of genetic modifications needed to be done on the chassis. One essential gene knocked out per plasmid inserted. A different method, that can be also utilized is the toxin-antitoxin system which has been used since the 1990s. The goal of such system is that the cell encodes a gene that is toxic, and the plasmid codes a gene which eliminated the introduced toxicity. This ensures that cells which do not bear plasmids are killed.

2.8 Conclusions

Multi-plasmid systems have emerged as a powerful tool in metabolic engineering and synthetic biology, enabling modular pathway engineering and increased screening efficiency. However, the behavior and interactions of multiple plasmids within a single cell can be complex and unpredictable.

This review has highlighted several key considerations and challenges when working with multi-plasmid systems.

Plasmid copy number and stability can vary significantly from expected values when multiple plasmids are present, impacting expression levels and strain performance. Routine monitoring of plasmid metrics is recommended but often overlooked in many studies. Careful selection of compatible plasmid backbones is critical, as even supposedly compatible plasmids can show unexpected interactions or instability over time. Expression systems, including promoters and regulatory elements, need to be optimized in the context of multiple plasmids to account for resource competition. Selection systems must be carefully designed to maintain multiple plasmids without imposing excessive metabolic burden. Furthermore, scale-up to fermentation conditions can reveal plasmid instability issues not apparent in small-scale experiments.

To realize the full potential of multi-plasmid systems, several areas warrant further research and development. Improved methods for rapid, high-throughput assessment of plasmid copy number and stability in multi-plasmid strains are needed, as current techniques like qPCR are labor-intensive for multiple plasmids. Expanded databases and predictive models of plasmid compatibility and behavior would greatly aid in the rational design of multi-plasmid systems. Development of novel plasmid backbones and regulatory elements optimized for stable co-existence and tunable expression could address many current limitations. Additionally, standardized reporting of plasmid characteristics and behavior in multi-plasmid strains would help build community knowledge and improve reproducibility across studies.

With continued advances in these areas, multi-plasmid systems will become an increasingly powerful and predictable tool for complex pathway engineering. However, careful experimental design and thorough characterization of plasmid behavior remain essential. By routinely assessing plasmid metrics alongside other strain characteristics, metabolic engineers can gain deeper insights into their systems and avoid drawing incorrect conclusions about pathway performance. This approach will ultimately accelerate the development of optimized microbial cell factories for diverse applications, from biofuel production to pharmaceutical synthesis. As the field continues to evolve, a more nuanced understanding of multi-plasmid dynamics will

be crucial in pushing the boundaries of what is possible in synthetic biology and metabolic engineering.

Author	Year	Product	Plasmid 1			Plasmid 2			Plasmid 3		
Moon	2010	D-glutaric acid	p15A	Plac		ColE1	Ptet				
Yim	2011	1,4-butanediol	p15A	pLac	Chl	ColE1	PLac	Kan			
			pSC101	pLAC	Amp						
Battesti	2012	two-hybrid system	p15A	pLac	Kan	ColE1	pLAc	Amp			
Juminaga	2012	L-tyrosine	p15A	pLac	Amp	pSC101	pTet	Kan			
						pBBR1	pLac	Chl			
Alonso-Gutierrez	2013	limonene	p15A	pLac	Chl	ColE1	pTrc	Amp			
		perillyl alcohol	p15A	pLac	Chl / Amp	pBBR1	pRAD	Kan			
Kim	2014	phenol	p15A	various	Kan	ColE1	pTac	Amp			
Afonso	2015	resveratrol	p15A	constitutive	Chl	pBR322	pLac	Amp			
Hsia	2016	mCherry	p15A	pBad	Kan	ColE1	constitutive	Chl			
Wierzbicki	2016	fatty acid ethyl ester	pBR322	T7	Amp	pBR322	T7	Kan			
		fatty acid isobutyl ester				ColE1	T7	Kan			
Zhang	2017	mCherry	p15A	pBad	Amp	WH32-5	pchnB	Kan			
Thurany	2017	fluorine	CloDF13	T7		ColE1	pTrc				
McNerney	2017	lycopene	p15A	varied	Kan/Chl	pBR322	varied	Tet			
Yan	2017	arsenic resistange genes	pBR322	pLac	Kan	pBR322	pLac	Amp			
Han	2017	L-xylulose	ColE1	pLAc	Amp	p15A	pLac	Kan			
Walther	2017	2,4-dihydroxybutiric acid	p15A	lac	Kan	ColE1	tac	Amp			
Wang	2017	pyrogallol	p15A	lac	Kan	pSC101	lac	Cm			
						ColE1	lac	Amp			
Chen	2017	cis-4-hydroxyl-l-proline	pBR322	T7	Km	p15A	T7	Cm	pCDF	trc	Sm
Liu	2017	biofuel	pSC101	T7	Chl	p15A	lac	Spec	ColE1	lac	Amp
Thompson	2017	muconic acid	p15A	lac	Amp	ColA	lac	Kan			
			pBR322	lac	Amp						
Wang	2018	5-HTP	p15A	T7	Chl	pBR322	T7	Tet			
						pSC101	T7	Amp			
Alnoch	2018	lipase, foldase	pBR322	T7	Kan	pBR322	T7	Amp			
Liu	2018	sesquiterpene	p15A	varied	Chl	ColE1	varied	Amp			
Liu	2018	3-hydroxypropionate	pSC101	araBAD	Str	R6k	araBAD	Kan			
Li	2018	baicalein	CloDF13	T7	Spd	pBR322	T7	Amp	p15A	T7	Cm
Chen	2018	tryptophan	p15A	tet	Cm	pSC101	tac	Kan			
Kim	2018	palmitic acid	CloDF13	T7	Sm	p15A	T7	Km			
Nemr	2018	(R)-1,3-butanediol	pBR322		Amp	p15A		Kan			
Seok	2018	3-hydroxypropionic acid	p15A	tac	Cm	ColE1	T5	Amp			
Wang	2018	5-hydroxytryptophan	p15A	T7	Cm	pBR322	T7	Amp			
						pBR322	tac	Tet			
						pSC101	tac	Amp			
Goyal	2018	method	p15A	UV5	Cm	pBR322	trc	Amp			
Yishai	2018	glycine	pSC101	pgi	Str	p15A	pgi	tet			
Luo	2018	2-pyrone-4,6-dicarboxylic acid	p15A	tac	Km		lac	Cm			
Dubois	2019	DKP alkaloids	ColE1	T5	Amp	RSF101	T7	Kan			
Munro	2019	proteprhodopsin	p15A	constitutive	kan	?	Plac, Para	Amp			
Wang	2019	anthranilate	p15A	lac	Kan	pSC101	lac	Cl			
						ColE1	lac	Amp			
Yang	2019	method	ColE1		Cm	p15A		Km	ColA	Pbad	SM
Noda	2019	isobutanol	pBR322	trc	Amp	p15A	trc	Km			
						pSC101	AlacO1	Cm			
Yu	2019	isobutanol	p15A	lac	Kan	ColE1	lac	Amp	ColA	lac	Amp
Li	2019	tropone	pSC101	lac	Cm	ColE1	lac	Amp	p15A	lac	Kan
Luo	2019	2-amino-1,3-propanediol	RSF	T7	Kn	p15A	T7	Cm			
Tian	2019	isopentenol	p15A	UV5	Cm	ColE1	trc	Amp		tet	
Luo	2020	benzoic acid	p15A		Km	pBR322		Amp	CloDF13		Sm
Liu	2020	succinate	p15A	T7	Cm	pBR322	T7	Amp			
						pSC101	araBAD	Amp			
						ColE1	T7	Kan			
						MB1	lac	Amp			
Xie	2020	method	pSC101	constitutive		pMB1			p15A		
Park	2020	GABA	p15A	T7	Cm	ColE1	tac	Amp			
Zhang	2020	aglycolysated immuniglobulin g	pBR322	UV5	Amp	p15A	pBAD	Cm			
Long	2020	L-proline	pSC101		Sm	pBR322					
Ko	2020	acrylic acid	p15A	lac	Km	pBR322	constitutive	Ap			
Zhan	2020	glycolate	p15A	constitutive	Cm	pSC101		Amp			

TABLE 2.1: Examples of multi-plasmid systems in publications

3 Investigating plasmid behaviour in multi-plasmid systems in E.coli

3.1 Introduction

As described in the previous chapter, modularizing a pathway of interest to optimize separately and then, in a later stage, test combinations of the modules is a common practice in metabolic engineering. While the advantages of such a system have been described in detail, one needs to be aware of the variability it can introduce to the measured data. In this study, we investigate the variability in a strain engineering workflow. Additionally, we aim to explore the behavior of multi-plasmid systems for high-throughput screening purposes.

3.1.1 Variability in screening data with multi-plasmid strains

In a metabolic engineering project with a modular approach, significant time and resources must be allocated to fine-tune the modules. The combined modules usually show a slightly undesired phenotype in the production strain; hence an extensive tuning step is needed. Such undesired behavior can be lower than expected production, unstable production of the target or in extreme cases, even growth defect.. Therefore, the plasmid used to carry the different modules must be well known and exert predicted behavior. As discussed in the previous chapter, multiple plasmid combinations are commonly used in modular metabolic engineering approaches. Amongst the most widely used expression origins are pBR322, p15A and pSC101, which plasmid backbones were also chosen as a basis for this study. These plasmids bear the advantage that they are compatible, as their replication machinery does not interfere with each other when utilized in the same cell. Therefore, they can be co-transformed in any combination, while the cell retains them in

the presence of different markers. Generally, pBR322 is used as a high copy number plasmid (20-200), p15A is considered a stable medium copy number plasmid (5-30) and pSC101 is a plasmid with a low copy number (1-5).

Biotin is a crucial vitamin in most forms of life and has a large market in the food and cosmetics industry. Biosyntia aims to create an industrially viable biotin cell factory and is developing strains to fulfil various criteria.

In order to parallelize the development of the biotin production pathway in *E. coli*, the necessary genes were divided into two modules. This modular strategy offers two main advantages. First, it enables the independent optimization of the pathway's distinct segments, the upstream reactions leading to desthiobiotin (DTB) production and the downstream conversion of DTB to biotin. Second, it allows us to decouple challenges associated with enzyme kinetics and flux control in the pathway, thereby simplifying troubleshooting and targeted improvements. Module 1 contains the genes encoding the enzymes responsible for the early steps of biotin synthesis: bioG, bioC, bioF, bioA, and bioD. These genes work together to enhance the production of DTB, the direct precursor of biotin. Notably, although *E. coli* naturally employs bioH in its biotin pathway, our design incorporates bioG, an enzyme not natively present in *E. coli*. This substitution was made because bioG may offer distinct catalytic properties or improved flux characteristics compared to bioH, which could be critical for achieving the desired DTB levels. Module 2 is dedicated to the final step of the pathway, which involves the conversion of DTB to biotin catalyzed by bioB (biotin synthase). BioB is recognized as a challenging enzyme to engineer due to its relatively slow kinetics, and it often represents a bottleneck that limits the overall flux through the biotin pathway. The separation into two modules not only streamlines the optimization process by allowing each module to reach a target yield before integration, but also facilitates the study of their interactions. For instance, an accumulation of DTB in Module 1 could affect the efficiency of bioB in Module 2, making it essential to balance the flux between the two segments. Understanding and managing this interplay is key to developing an effective production strain.

In order to test the effect of various auxiliary genes, which might help the production while not being directly connected to the pathway of interest, were also included in most engineering efforts. Therefore, when biotin production from glucose is investigated in a strain where these modules are

combined, the production strain usually contains 2 to 3 different plasmid backbones.

When these modules are combined, one can expect differences in performance from their expressions. These differences can be the result of various effects. One of the most investigated amongst these possible effects is an imbalance in the expression of the modules. For example, if the DTB pathway module is highly expressed, the strain is expected to produce large amounts of DTB. In the case of low *bioB* expression, in the combined production strain, the produced biotin is an indicator of the capabilities of the *bioB* module at its current expression level. The accumulation of DTB in that strain would show that the upstream pathway is capable of producing more precursor, which accumulates due to the imbalance between the modules. Another, mostly overlooked possibility is the differences in the plasmid copy number when compatible plasmids carrying the modules are combined within the same strain. We aimed to investigate the effect of plasmid backbone choices on the variability of production results. We investigated data collected from two different assays for measuring the biotin production performance of strains. Biotin assay measures the concentration of biotin in the spent media, assessing the secreted target product (Bali et al., 2020). The HABA assay measures all biotin vitamers in the supernatant of the culture and is used mainly to determine the performance of the pathway upstream of the *bioB* module (Buerger, 2019). This assay measures a wide variety of compounds, and we expected to have more considerable variation between biological replicates.

We investigated the variation between technical and biological replicates of the same strain, which contains at least two compatible plasmid backbones. We were interested if certain combinations of plasmids show larger variability than other combinations. The data points shown in Figure 3.1 are standard deviations of technical triplicates of various strains.

As summarized in Figure 3.1, assay data shows that in the presence of both pBR322 and pSC101 backbones, high variability is introduced to the measurements of biotin and biotin vitamers. The variability was significantly reduced in strains where pSC101 and p15A were used. It raises the question of why do the different combinations of plasmids provide more stable production than others. Do they introduce the same amount of variability, and therefore the effect can be assigned as noise, or does the combination affect

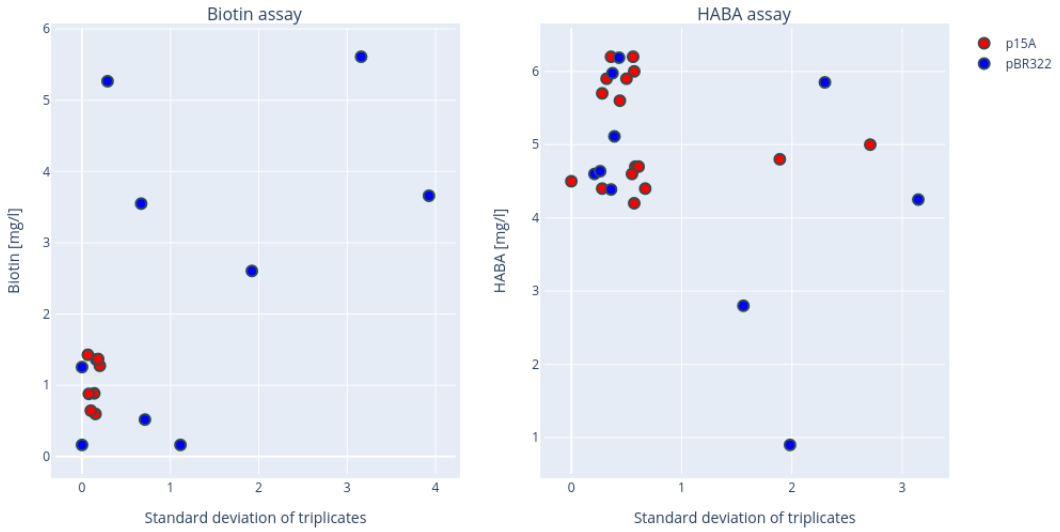


FIGURE 3.1: Standard deviation of technical triplicates against the measured biotin or HABA vitamer concentration. Data was collected from experiments carried out at Biosyntia in the timeframe 2018 January - May. The data points visualized in the graphs are from various experiments, where bioB is expressed from a plasmid utilizing pSC101 backbone while the rest of the biotin production genes are expressed from either a pBR322 or p15A backbone There is also variation in the induction level of the genes, which could also introduce variation. It can be observed that the pBR322 backbone shows the most extreme standard deviations in both assays.

production stability. Important to note that in Figure 3.1 we do not differentiate between the expression system carried by the plasmids, but we compare the variability between biological replicates of the same strain.

3.1.2 Experimental design

In order to explore the effects of variation, we aimed to create a minimalist system to model the development cycle in strain engineering. As shown in Figure 3.2, we transformed the strains with the created plasmid(s). The transformation mixtures were plated on selective LB media in serial dilution to monitor transformation efficiency. We randomly picked individual colonies and used triplicates of each colony for growth and fluorescence signal measurement over time in a fluorescent plate reader. The growth and fluorescence data provide information about the production capabilities of the strain, while growth indicates the overall fitness of the strain. In order to acquire a closer view of the plasmid presence in the production strains, cell pellet samples were taken to undergo plasmid copy number determination.

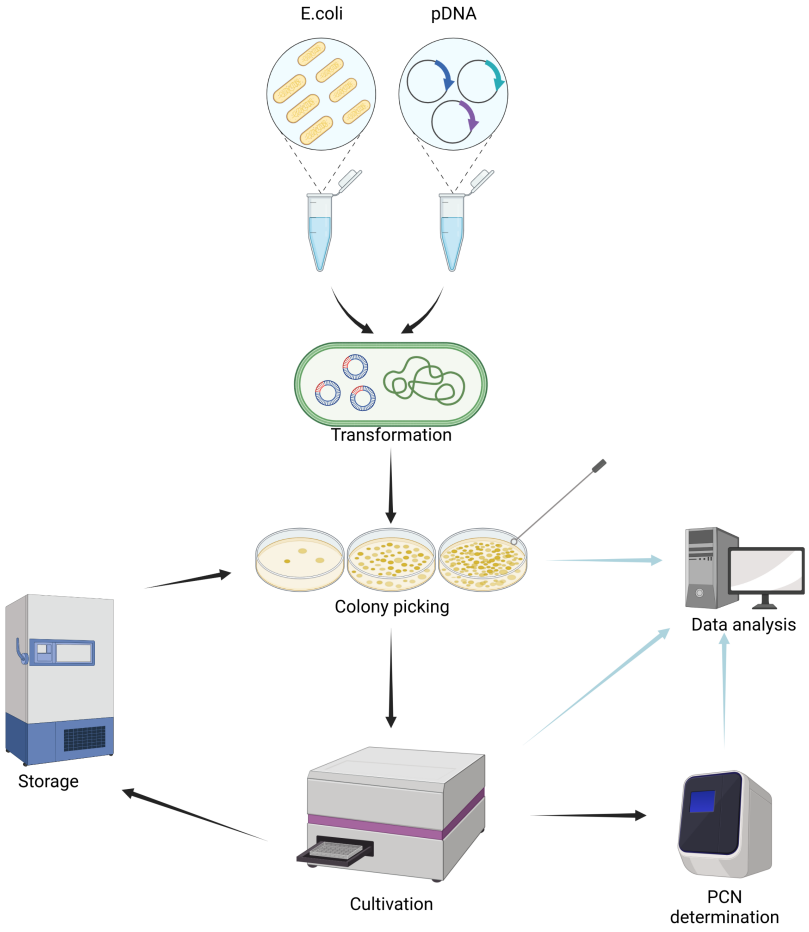


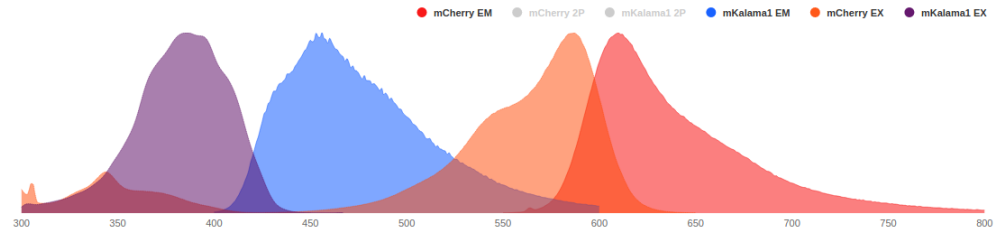
FIGURE 3.2: Overview of the experimental setup. Chemically competent *E. coli* was transformed with the target plasmid(s). The transformation mixture was plated in serial dilutions to assess transformation efficiency. The picked colonies were cultivated and their growth rate, fluorescent signal was measured with a plate reader. Plasmid copy number was determined from the samples taken at the end of the cultivation. The cultures were stored at -80 °C between plating to assess the noise between generations and to model the practical use of strains in an industrial research and development environment.

3.2 Methods

3.2.1 Choice of model cargo

In order to mimic the production strains and have a quantifiable signal about the production of cargo from the various plasmids, we designed the plasmids present in this study with fluorescent proteins. When designing the

standards plasmids, it was essential to be able to separately quantify the fluorescence signal when both fluorescent proteins are expressed in the same cell. To achieve this experimental setup, there was the need for two fluorescent proteins without or just minimally overlapping emission and excitation spectra. Fluorescent proteins mCherry and mKalama were chosen for this purpose. These proteins are close in size (26.72 kDa and 26.79 kDa, respectively) and brightness, meaning that the possible burden of producing these proteins could be similar, and we can expect the fluorescent signal to be in similar range (Figure 3.3).



Attribute Comparison

Name	λ_{ex}	λ_{em}	Stokes	EC	QY	Brightness	pKa	Aggregation	Maturation	Lifetime	kDa
mCherry	587	610	23	72,000	0.22	15.84	4.5	m	15.0	1.4	26.72
mKalama1	385	456	71	36,000	0.45	16.2	5.5	m			26.79

FIGURE 3.3: Fluorescence profile of mCherry and mKalama1 [Source: FPBase.org]

3.2.2 Plasmids and strains

All plasmids in the study were created by USER cloning. Primers were designed with the help of AMUSER. Building blocks for the plasmids were cloned with the reported primers from in-house plasmids unless stated otherwise. mKalama gene sequence was ordered from IDT in the form of a gene block. The gene block was amplified with specific primers and used as a fragment in USER cloning. The chosen backbones were pBR322, p15A and pSC101, as these are compatible plasmids that can be maintained in the same cell throughout generations according to literature and previous experiences. As antibiotic markers, we used genes that encode resistance for Kanamycin, Carbenicillin and Spectinomycin, respectively, in the various backbones. A type of backbone will always carry the same antibiotic marker in this study. While there are reported effects of the marker on the plasmid copy numbers, all conclusions in this chapter are drawn about the combination of origin of replication and marker combinations. The study was aimed to model

an industrial, automatable Research and Development environment; hence the methods have been chosen with the possibility of scaling up to high-throughput processes.

Backbone	Cargo			Marker
	empty	mCherry	mKalama	
pBR322	pBS1730	pBS1731	pBS1732	Kan
p15A	pBS1733	pBS1734	pBS1735	Amp
pSC101	pBS1736	pBS1737	pBS1738	Spec

TABLE 3.1: Plasmids used in the study, indicating their backbone, cargo and antibiotic resistance.

The created USER cloning mixtures were transformed into chemically competent *E. coli* BW25113 strain. Colonies from the transformed strains were picked and underwent colony PCR to verify the presence of the plasmids. From the positive hits of the colony PCR, the plasmid was extracted with Miniprep and sent for Sanger sequencing with primers to verify the assembled sequence. The obtained purified plasmids were stored at -20 °C in TE buffer until their use in transformation of target strains. Due to confidentiality restrictions, the DNA sequences cannot be included in this thesis. These sequences are considered confidential information of the company and cannot be publicly disclosed.

3.2.3 Strains

All strains created in the study originate from *E. coli* BW25113 of the Keio Collection. All transformations were carried out as chemical transformations. 1 ml competent cells were prepared from 50 ml culture of OD600 0.6 with the RbCl method (1983) and stored at -80 °C until their use. The transformation was carried out in all cases with negative controls (samples with competent cells without pDNA) to determine the average viable cells in the transformation mixture. For strains used in the initial study, 100 µl prepared, and competent cells were mixed with 100 ng pDNA. During the later stage of the study, we aimed for a more controlled transformation mixture, as we are interested in its effect on the production performance of the transformed strain. In the chapter, we refer to ratios regarding plasmid to cell numbers. These ratios take the number of cells and the number of plasmid copies in the transformation mixture (Table 3.2). After each transformation, the transformation

Strain	Cell : SumP	P1 : P2	P1	P2
1	1:1	-	pBS1730	-
2	1:100	-	pBS1730	-
3	1:1000	-	pBS1730	-
4	1:1	-	pBS1734	-
5	1:100	-	pBS1734	-
6	1:1000	-	pBS1734	-
7	1:1	1:1	pBS1731	pBS1733
8	1:100	1:1	pBS1731	pBS1733
9	1:1000	1:1	pBS1731	pBS1733
10	1:1	1:1	pBS1734	pBS1730
11	1:100	1:1	pBS1734	pBS1730
12	1:1000	1:1	pBS1734	pBS1730
13	1:1000	1:10	pBS1731	pBS1733
14	1:1000	1:1000	pBS1731	pBS1733
15	1:1000	1:10	pBS1734	pBS1730
16	1:1000	1:1000	pBS1734	pBS1730

TABLE 3.2: Strains used in the final part of the study, indicating the ratio of cell to plasmid and plasmid to plasmid in the transformation mixture.

efficiency was determined by plating serial dilutions (0 x, 10^1x, 10^3x, 10^9x) of the transformation mixture on selective LB plates. The same dilution plating was carried out with the negative control too. Transformation efficiency was estimated by counting the colonies on each plate after 24 h of incubation at 37 °C. Strains were also verified via colony PCR with primers defined in the supplementary material.

3.2.4 Cultivation and sampling

Individual colonies were taken from the transformation plates and used to inoculate 200 µl minimal media with the appropriate antibiotics. This pre-culture was incubated at 37 °C for 24 hours in an orbiting shaker with the speed of 250 rpm in a 96-well plate sealed with a sterile BreathEasy seal. The cultivated pre-cultures were diluted 1:100 with fresh selective mMOPS media and separated technical triplicates into a black 96-well plate with clear bottom and sealed with sterile see through BreathEasy foil. The plate was then placed in a Thermo Fisher Varioskan Lux plate reader. In the plate reader, the temperature was set to 37 °C, and continuous shaking was set between measurements, which were 10 minutes apart for 24 hours. Optical density

at 600 nm was measured alongside the mKalama (excitation: 385 nm, emission: 456 nm) and mCherry (excitation: 587 nm, emission: 610 nm) signals. At the end of the plate reader-run, 10 µl of each sample was taken for qPCR measurements. 75 µl of culture and 75 µl of 20% glycerol was mixed in a new 96-well plate and the resulted stock was sealed with aluminum foil and stored at -80 °C. These glycerol stocks were streaked onto LB plates with the appropriate antibiotics for the next generation of strains. For the next generation pre-cultures, colonies from the freshly streaked plates were taken. For next generation sequencing, samples were prepared in LB media with the appropriate antibiotics. After 24h of incubation at 37 °C, genomic extraction was carried out, based on the external NGS service provider's suggestion.

3.2.5 Plasmid copy number determination

In order to obtain data in a high-throughput manner, we needed an automatable process for determining the plasmid copy number in each strain. Sequencing-based approaches rely on the assumption that plasmids maintain the same ratio to the genomic DNA during sample preparation. By accepting this assumption, we can gather valuable information about the possibly occurred mutations in the strain, which could explain differences in phenotype. Other popular methods of choice are qPCR based-methods. Contrary to sequencing-based methods, qPCR based methods require less sample preparation. However, there is a need for careful primer design to ensure the same efficiency across primer pairs.

Sequencing-based approach

The method's principle is to quantify the number of reads in a specific region typical to the plasmid, generally backbone or resistance marker gene, and compare it to the number of reads from a housekeeping gene, which is present in a known amount of copies in the chromosome. We attempted to identify plasmid copy numbers using two methods based on the NGS reads. A k-mer based method with the application of PlasmidSeeker software (Roosaare et al., 2018). In this method, a list of read of a k-mer list is created (k length strings of reads) from the reads.

And the approximate genome coverage is calculated by the median abundance of these k-mers. Plasmid copy numbers are estimated by dividing the median abundance of plasmid originated k-mers with the median abundance of k-mers from the chromosome. A different approach was to use the software, QIAGEN CLC Genomics Workbench 21.0, where we used the coverage of a housekeeping dene (*dxs*) and the coverage of the antibiotic marker gene on the plasmids. The logic behind evaluating the results with the two different software was the same, compare the number of reads on target sequences.

In this study, the first choice of plasmid copy number determination was with the help of whole-genome sequencing, as we were aiming to gather additional information about possible mutations. DNA Shotgun sequencing of samples prepared with Prepared samples were sent to LGC Biotechnology for NGS Illumina sequencing. Estimating plasmid copy numbers via PCR- or sequencing-based methods have the same principle. From a sample, we aim to see the ratio of the plasmid to chromosomal DNA. We can estimate the average plasmid copy number per cell in the investigated sample from that ratio. The target sequences are similar in the sequencing and the PCR-based approach.

qPCR-based approach

As described in multiple publications, qPCR is a robust method of evaluating plasmid copy numbers. In most cases, where the investigated organism is *E. coli* the resistance marker gene is used to quantify the plasmids, while the housekeeping gene is generally *dxs*. *dxs* codes the 1-deoxyxylulose-5-phosphate synthase enzyme, which facilitates the first and rate-limiting step of the isoprenoid biosynthesis and also a key element of the thiamine biosynthesis pathway. The gene is present in only one copy on the chromosome; therefore, it is an excellent proxy for the number of cells in the sample.

Cell pellet samples underwent crude boiling sample extraction, which consisted of heating and shaking 100µl of the sample at 90C for x minutes. After the boiling step, the samples were centrifuged, and the supernatant was used as template in the qPCR reaction. The reaction used SybrGreen dye, and the primers targeting the housekeeping gene and the resistance markers are described in Supplementary material

The first qPCR measurements were carried out at the University of Hamburg, during 2019 autumn, as the first of many attempts to assess the copy number of a large number of samples. The measurements were aimed to gain knowledge about the method and verify that all strains can be quantified. qPCR measurements were carried out with a SensiFAST SYBR Hi-ROX Kit on a Biometra TOPtical qPCR machine according to the manufacturer's recommendations. The rest of the measurements were impossible as travel restrictions were applied due to the COVID-19 pandemic.

We aimed to rent a qPCR machine from a local CRO (Bioneer ApS) and re-run some samples from the previous sample set to verify that any new measurement can be compared with the previous results. Unfortunately, we could only measure two 96-well plates due to the limited budget, which surprisingly resulted in significant difference in the measured PCN compared to the previously obtained results. The qPCR master mix and machine was different from the previous measurements, but the cycling protocol, primers and sample preparation were the same.

In an attempt to quantify plasmid copy numbers from the later stage of the study, we aimed to carry out three 368-well plate measurements due to budget constraints. The first plate was aimed to optimize the assay conditions while keeping the primers and sample preparation method the same. It gave inconclusive results; therefore, a second plate had to be used for optimization. The second plate included a smaller amount of biological replicates and several dilutions of the template DNA and a modified cycling protocol. The new setup showed results in the literature range reported copy numbers, and the results were consistent between technical replicates. The cycling conditions were the following: The third plate was used to measure a subset of strains. Plasmid copy numbers were calculated with the $\Delta\Delta C_t$ method.

3.3 Results

In the first round of experiments, we were interested in the efficiency and precision of the method of quantifying plasmid copy numbers via whole genome sequencing read quantification, as described in the methods section. Plasmid copy numbers showed an expected trend for the strains carrying only one type of plasmid (Figure 3.4a). Even though the trends were similar

to the literature values, the absolute values were lower than expected. For pBR322 backbone plasmids, the expected plasmid copy number is above 50, around 15 for p15A and 1 for pSC101. Results observed were an average of 5.83 copies per cell with values ranging between 3.1 - 9.6, depending on the carried cargo. pSC101 backbone plasmids seem to have a low presence in the cultures, as the measured copy numbers for each strain carrying was below 1. Which could indicate that there are cells in the culture which do not possess the plasmid.

In case of the different plasmid combinations, most strains were in the approximate range of the solo strains with a few exceptions. When one combines pBR322 and p15A (Figure 3.4b), there were strains which do not show any reads for the p15A backbone plasmid, but can still grow in the presence of both antibiotics. An other interesting observation is the increased plasmid copy number of p15A when carrying mCherry instead of mKalama, or even its empty counterpart. Strains where pBS1734 was present, showed approximately 3-times the amount of p15A plasmid presence compared to strains containing pBSS1733 (empty) or pBS1734 (mKalama), regardless of which plasmid was it paired with (Figure 3.4).

These observations led us to doubt the results as they were inconsistent with literature and strains which have been grown on a certain antibiotic did not show any reads for the region of the antibiotic resistance marker. These doubts led the study into the direction of characterizing the plasmid copy number of the strains with a different, qPCR based method.

The following fluorescence and growth rate results were obtained from the plate reader experiments performed, and samples of each well were collected and prepared to be measured with qPCR. In Figure 3.5 we can observe that the growth rate of the various strains does change significantly in some cases, but never dropped below 0.8 1/h. Regarding the measured fluorescence results, one can observe a large variation between wells. Interestingly for the solo strains (Figure 3.5a), the largest mKalama production was shown by the strain carrying the protein's encoding sequence on a p15A plasmid. pBS1735 is high performing mKalama plasmid, which shows more fluorescence results than its pBR322 counterpart. The reason for this behavior can be that the large copy number plasmid has created many of the amino acid chains for the protein, but they either formed as inclusion bodies, although there was no indication it on the colonies when frown is solid or liquid media,

or does not fold properly and hence the fluorescent signal does not correlate with the protein amount in the cell. We have taken this into consideration when planning the following experiments and decided not to change the expression systems, as we are also interested in if all proteins are possible to express with pBR322, which do not seem to be the case. We did not change the expression systems, as we would have liked to simulate a lab where not all tools are available.

The above-mentioned strains' plasmid copy numbers were measured with qPCR (Figure 3.6) to be able to quantify the amount of plasmids present in the cell. As these samples were measured in the plate reader, we overlapped the fluorescence results with the freshly obtained plasmid copy number results. As shown in Figure 3.6a, we can observe that contrary to the previously obtained plasmid copy number results, the highest copy number was obtained from the strains carrying the plasmid encoding mKalama, in each plasmid backbone. In this case the PCN of pBR322 was still lower than the expected literature value, while p15A matched the range of previously reported values and pSC101 slightly above the literature-reported values. Interestingly, we can observe that for the strain carrying only pBS1732 does have the most number of plasmid copies in the cells and also the lowest amount of active mKalama fluorescence, which indicates that the high copy number is not a favored expression method for mKalama.

In case of the double strains (Figures 3.6b, 3.6c, 3.6d) we can observe that the plasmid copy numbers are in the same range as the solo strains. While the fluorescence results show some interesting patterns. In case of the pBR322 and p15A pairings (Figure 3.6b), we can observe that the expression of mCherry has been the highest when both plasmids expressed the protein (pBS1731 and pBR1734), and that strain also showed the most plasmid presence. While the strain carrying mCherry on p15A backbone and mKalama on pBR322 (pBS1732, pBS1734) shows similar mCherry production with minimal mKalama production. Similar effect can be observed when the pairing's pBR322 is carrying mKalama, while there is no cargo on the p15A. In case of the p15A and pSC101 pairings (Figure 3.6c), we can observe that the strains carrying mCherry or mKalama on p15A show highest expression of the fluorescent protein amongst them, but expression of any other protein from the pSC101 interferes with its production. Interestingly, the strains carrying mCherry or mKalama and paired with an empty pSC101

vector showed higher fluorescence readings than the strains carrying those plasmids alone. For strains with the pBR322 and pSC101 pairings (Figure 3.6d) the plasmid copy numbers remain simialr to the expected values. While the fluorescence results show an interesting pattern for mKalama being expressed from pBR322 with the presence of a pSC101 plasmid. In those strains, the mKalama production is seemingly lower based on its fluorescence signal, while in cases where the pBR322 plasmid was carrying mCherry, the production have not been affected to the same extent. This result can also validate the previous assumption, that expressing mKalama with the pBR322 backbone is not ideal.

As the previous qPCR measurements were done in Hamburg, and due to the developed travel restrictions due to the pandemic, repeating experiments in Hamburg was out of scope. Therefore, we aimed to find a danish CRO, where the rental of a qPCR machine was organized. A subset of samples from the previous dataset was prepared as a verification plate for the method ran on the qPCR at the CRO, Bioneer ApS.

As we can see on the results in Figure 3.7, the results were inconsistent with the previous plasmid copy number measurements. As the measurements were more than 10-fold higher than the ones measured in Hamburg. We attempted to investigate the difference with including samples from the same experiment with multiple timepoints. Assuming that the plasmid copy number would be the highest at the latest timepoint, especially for the backbones without a regulated origin, such as pBR322. We can observe the expected trend in case the plasmids were individually part of the strain. Such as in Figure 3.7a, where pBS1730 is an empty pBR322 plasmid. Although when the same backbone was expressing mKalama the opposite effect was observed, together with significantly lower plasmid copy numbers than the empty vector. As the chosen cargo for this experiment was mKalama, we can again assume that the expression of this fluorescent protein from this backbone is not ideal, not just on the expression level, but for plasmid replication either. For double strains, there was even higher variability between the plasmid copy numbers on p15A (Figure 3.7b).

In order to investigate the variability in these samples that were aimed to validate our previous results, we have started to pay close attention to the transformation event during th preparation of the strains. In the first round of experiments, we have generated 16 different strains as described in

Table 3.2. We measured their transformation efficiency and investigated their fluorescence signals at 24 hours in multiple generations.

When assessing their transformation efficiency, (Figure 3.8) we can see that there is a large variability between strains. In case of the first 3 strains, the plasmid was an empty pBR322 vector, and the only variability in the experiment was the cell to plasmid ratio, which was established by the amount of plasmid copies per cell. In strain 1, this ratio is 1:1 meaning that for each cell, there was one plasmid present in the transformation mixture, same logic with 1:100 and 1:1000 ratio for strains 2 and 3. We can observe that the abundance of plasmids, even in case of this large copy number plasmid has not increased, but rather decreased the transformation efficiency. For a p15A plasmid, carrying mCherry (Strain 4, 5, 6) the transformation efficiency was similar to the pBR322, but the drop of efficiency was only observed at the highest plasmid load. When multiple plasmids were introduced in the transformation mixture, the efficiencies in general dropped compared to the single plasmid containing strains, which is expected. Although there was one outlier, Strain 11, which contained 100 plasmids each per cell when the pBR322 backbone was empty, and the p15A backbone was carrying the mCherry protein. Strain 7 did not provide transformants. The second worst transformation efficiency was provided by strain 13, where the plasmids were overloaded 1:1000 ratio.

When assessing the fluorescence of these strains in different generations, we can observe an interesting trend. In the experiments strain 7 and 10 are missing, as 7 did not provide a successful transformation, while strain 10 did not grow in liquid media. These strains both contained the least amount of plasmids (1:1 cell to plasmid ratio) and equal amount of plasmids per cell (1:1 plasmid to plasmid ratio). This shows that there is a need to have more than 1 plasmid present per cell in the transformation mixture when transforming multiple plasmids at the same time to obtain a viable strain.

In this experiment we have only expressed mCherry from either pBR322 or p15A backbone, therefore, the fluorescence values defined here are originating from mCherry, and mKalama was not assessed in this part of the study. If we assess the fluorescence results seen in Figure 3.9, we can see that there is significant variance between the generations in most cases. Interestingly, we even measured mCherry expression in strains where no expression was expected, such as Strain 1, 2 and 3. Although no fluorescence was

observed in the wild type control included. In case of strains where we expected the expression and fluorescence signal, we cannot observe any repeating trend. In case of strains 11 and 12, we can see an increase of fluorescence with each generation. These strains contained an empty pBR322 plasmid and mCherry was carried on the p15A backbone, although there was no significant difference between the fluorescence provided by them. The highest and most stable mCherry production was provided by strains 13 and 14, where mCherry was expressed from the high copy number pBR322 and they also contained an empty p15A backbone. Both of these strains have contained 1000 copies of plasmids per cell, and 1:10 and 1:1000 ratio of pPR322 to p15A respectively. When in a similar setup, Strains 15 and 16, with switched cargo, the strains performed similarly stable and showed an expectedly lower fluorescence signal, as the mCherry was carries on the lower copy number p15A plasmid.

When assessing the growth rate of these strains, we can see that in most cases they are either stable throughout the generations, or mostly increase (Figure 3.10). The generally lowest growth rate was shown by the strains only containing pBR322 plasmids, which supports the hypothesis of the highest copy number plasmid posing the most burden on the growth.

We also assessed the possible correlation between the plasmid copy number and the growth rate, as well as the fluorescence signal. In case of Strain 3 (Figure 3.11a) where the empty pBR322 plasmid was present in the strain, the growth rate correlated with the plasmid copy number, the lower the copy number was in the strain, the higher the growth rate was, which is the expected behaviour, especially for high copy number plasmids, even without cargo. In case of Strain 4 (Figure 3.11b), where the medium copy number p15A backbone plasmid contained the mCherry protein, the plasmid copy number did not show such clear correlation to the growth rate. This can be explained by the fact that the p15A backbone does not pose as much burden on the strain as a pBR322 backbone, even if their estimated copy numbers were similar. There was no clear correlation or trend observable between various generations in case of these two strains

For strains where multiple plasmids were expressed, we could see various trends, but nothing that would point to the same directions. In case of Strain 9 (Figure 3.12a), we can observe that generation 2 and 3 showed generally higher plasmid copy number than generation 1. With stable growth rate

in generation 2. Although this stability was not observed in generations 1 and 3. Interestingly in Strain 13, where the plasmid pBR322 was carrying the mCherry protein, while p15A was empty, and in the transformation mixture the cell to plasmid ratio was 1:100 and plasmid to plasmid was 1 to 10, the first generation showed the highest copy numbers for both backbones, and the highest growth rates as well. While the third generation showed the lowest copy numbers, and the lowest growth rate. As well as the copy numbers of the two backbones were very similar. In case of strain 15, where the plasmid and cell ratios were the same as in Strain 13, but the cargo and empty plasmid was switched, the same trend could not been observed. As all the generations and copy numbers grouped together.

We were also interested in investigating how the fluorescence signal and plasmid copy number correlated, which is an indication of the efficiency of the strain in producing the target protein, in this case mCherry. When assessing strain 3, we did not expect fluorescence signal, but still observed similar levels as in case of Strain 4, where the mCherry was carried on the p15A backbone (Figure 3.13).

For strains where multiple plasmids were expressed, we could not observe any trends, and the fluorescence signal was also similar to Strain 3, where no mCherry was expressed, therefore we did not draw major conclusions from these results (Figure 3.14).

3.4 Discussion

3.4.1 Effect of transformation event

As described in the previous chapter, plasmids can be introduced via various methods. When plasmids are introduced to the strain, the efficiency of the transformation event is generally measured in colony forming units per nanograms plasmid introduced (CFU/ng). In many cases, to increase the transformation efficiency, when co-transforming plasmids we use the same mass of plasmids regardless of their copy number. When co-transforming plasmids with ether chemical transformation or electroporation, we rely on diffusion to ensure that plasmids enter the cells. As diffusion is not regulated by genetic elements or copy number regulation that resides on the plasmid

backbones, we hypothesize that the initial copy number after the transformation event is different than the literature reported individual copy number of the plasmids. Therefore, we hypothesized that the variation seen in the screening results could be the result of the different copy number distribution in the transformation mixture. As plasmids are frequently co-transformed, we were interested in if the cell-to-plasmid ratio or, in the case of multi-plasmid strains, plasmid-to-plasmid ratio influences the expression of our protein of interest and could influence the variability between colonies after transformation. We aimed to investigate the following factors:

- Plasmid backbones
- Presence of cargo
- Cell-to-Plasmid ratio of transformation mixture
- Plasmid-to-Plasmid ratio (for multi-plasmid strains)

3.4.2 NGS dataset

As in many cases to establish the exact genotype of a strain, one uses whole genome sequencing as part of the strain engineering. Estimating the plasmid copy numbers of the engineered strains from the same dataset would be ideal, as one would only need to include an extra step in the data analysis pipeline without additional steps in the lab workflow. We aimed to investigate this possibility, as it has been used to identify plasmids in various applications. Most of these applications are used to identify plasmids from environmental samples and rather focuses on the qualitative than quantitative aspect of plasmid presence in the strains. We observed that samples showed plasmid copy numbers are significantly differ from the literature values (Figure 3.3). And although the plasmid copy numbers were following the trends of the fluorescent signals recorded from the plasmid, the method was not sufficiently quantitative for our purposes (Figure 3.15). The genomic extraction method could be a reason why the plasmids seem to be enriched in all samples, regardless of their backbone or literature reported copy numbers. As it has been shown in most genomic extraction methods enrich plasmids, especially plasmids which are small, relative to the genome, as chromosomal and plasmid dDNA purifies with these kits with a different efficiency. The extraction method's effect on the introduced bias also depends on which

commercial kit was used and what type of extraction does that kit use. Generally low copy number plasmids are enriched in the samples, resulting in increased amounts of for example pSC101 compared to pBR322.

It is also important to note that the samples cannot be assumed homogenous regarding the plasmid copy numbers, due to segregational variability as explained in Chapter 2. As we can observe, the copy number values widely vary among the samples and strains (Figure 3.3). The copy numbers among samples are provided in the supplementary material and the median, minimum and maximum values are provided in 3.3. Plasmid copy numbers determined with this method are at least 10-fold higher for pSC101 than the reported literature values. As pSC101 has active partitioning mechanism, which ensures that there are only 1-4 plasmid per daughter cell, observing copy numbers ranging 18 - 58 suggest at least 10-fold more than previously reported. In case of p15A and pBR322, which have passive plasmid segregation systems, the values are still generally above literature reported plasmid copy numbers, but their variation is not strictly 10-fold increased to literature reported values. This supports the findings in that due to the whole genome purification, small and low copy number plasmid reads are enriched in the sequencing data.As this method is sensitive to bias introduced via the genome purification method and also requires the samples to be grown in complex media, in this thesis we are not continuing to apply this quantification method. Since the screening of production strains is routinely carried out in minimal media, it is essential that the sample strains carrying the standard plasmids alone or in combination are grown as close to the screening conditions as possible.

3.4.3 Hamburg dataset

Further investigating methods to quantify average plasmid copy numbers in screening conditions, we aimed to use qPCR in order to quantify the amount of DNA copies per chromosome. With this method, we do not gain information about the possible mutations that could occur on the plasmids. This lack of identification is not necessary for our purposes since we aiming to investigate cultures that contain a heterologous population regarding plasmid copy number distribution and the possible mutations could also be in so few of the plasmids per culture, that it would not show up in the NGS dataset either.

Plasmid	Min	Avg	Max
pBS1730	153.08	424.52	2251.5
pBS1731	87.11	533.39	1299.2
pBS1732	189.26	666.00	1046.6
pBS1733	40.38	71.00	410.3
pBS1734	0	310.6	1363.3
pBS1735	0	66.9	223.0
pBS1736	20.74	37.1	90.2
pBS1737	16.96	31.46	60.6
pBS1738	18.41	29.87	58.7

TABLE 3.3: Plasmid copy numbers calculated from NGS data gene coverage of dxs vs marker gene, samples were grown in complex media to stationary phase

With the qPCR-based method, we are targeting the same genes as in the NGS method, with creating specific primers that amplify a 100-120 bp region from the gene of interest. As a housekeeping gene, *dxs* was chosen, as it only has one single copy on the chromosome, hence it is a great standard to relate the plasmid coverage to. For the plasmid quantification, we targeted the marker regions of the plasmids to be able to compare strains with the same backbone but different cargos. The housekeeping gene should be carefully chosen, and should not be part of the engineered genes.

Results generated with this method gave similar, but still varying copy numbers compared to literature values. We aimed to run more samples on the same machine with the same method, but due to the COVID-19 related lockdowns in Germany and Denmark, the travels and measurements became impossible to carry out. Hence, we rented a different model qPCR machine from a Copenhagen-based CRO, and ran a verification plate to ensure that the runs on the two machines are comparable. Unfortunately, as it can be observed, the copy numbers varied significantly. The dataset was not investigated further with qPCR methods, assumptions and hypothesis was deducted from the single-replicate indication of the "Hamburg dataset".

3.4.4 Transformation study dataset

We aimed to return to Hamburg in order to carry out the qPCR measurements for the transformation study, but due to the prolonged lockdowns and travel restrictions due to the European COVID-19 situation, the travels were

impossible to carry out. Hence, we returned to the CRO in Copenhagen to optimize the method and study copy numbers throughout multiple generations in small scale screening conditions. The method optimization was done on 2 plates to ensure reproducibility, then one 384-plate was used to quantify selected samples due to financial reasons.

PCN quantification method of choice can influence the results as well as the sample preparation method, hence it is important to take these into consideration when choosing a method. The samples cannot be considered homogenous regarding copy numbers in the cells. In order to estimate the average copy numbers in the cells, there is a need for technical replicates from the same sample. The variation between technical replicates cannot be defined in the NGS or in the Hamburg dataset, but in the Bioneer dataset we can observe variation among technical replicates, but it is significantly lower than between the biological replicates. Therefore, we can show that the differences in copy numbers shown in the transformation study are not only due to noise. Although noise still present significantly in the dataset, it can be explained with the non-regulated segregation of p15A and pBR322 backbones, as variation is significantly less in the pSC101 plasmid copy numbers, which backbone is equipped with actively regulated segregational system.

3.5 Conclusions

This chapter explored the efficiency and precision of quantifying plasmid copy numbers using whole genome sequencing read quantification and qPCR methods, while also investigating the effects of various plasmid combinations on gene expression and cell growth. The results revealed a complex interplay between plasmid copy numbers, protein expression, and cellular burden.

Initially, the whole genome sequencing method yielded lower-than-expected plasmid copy numbers, particularly for pBR322 backbone plasmids. This discrepancy with literature values raised concerns about the method's accuracy and led to the adoption of qPCR-based quantification. The qPCR measurements provided a different perspective, showing higher copy numbers for plasmids encoding mKalama, which contradicted the sequencing results. This discrepancy highlights the challenges in accurately

quantifying plasmid copy numbers and emphasizes the need for multiple validation methods in such studies.

A significant finding was the suboptimal expression of mKalama from the pBR322 backbone. Despite high copy numbers, the fluorescence signal was lower than expected, suggesting issues with protein folding or the formation of inclusion bodies. This observation underscores the importance of considering not only the copy number but also the compatibility between the expression system and the target protein.

The study of plasmid combinations revealed complex interactions affecting both copy numbers and protein expression levels. In some cases, the presence of one plasmid influenced the copy number or expression of another, demonstrating the intricate balance within the cellular environment. These findings have important implications for the design of multi-plasmid systems in synthetic biology applications.

Transformation efficiency experiments showed significant variability between strains and were influenced by the plasmid-to-cell ratios. It is important to note that the number of colony forming units (CFUs) obtained from a transformation is not solely determined by plasmid properties but also by factors such as the optical density (OD) of the cell culture and the amount of DNA used. As documented in the literature (2020), increasing the OD and DNA quantity generally results in more CFUs up to an optimal point; beyond that point, further increases can actually lead to fewer CFUs. This non-linear dependency should be carefully considered when optimizing transformation protocols, particularly in experiments involving multiple plasmids.

Growth rate analysis revealed that, in general, rates were stable or increased across generations. However, strains carrying pBR322 plasmids showed the lowest growth rates, confirming the higher metabolic burden imposed by high-copy-number plasmids. This observation is crucial for balancing protein production with cellular health in biotechnological applications.

The study also attempted to correlate plasmid copy number with growth rate and fluorescence signal. However, no consistent trends were observed across different strains and generations, highlighting the complexity of these relationships. This lack of clear correlation emphasizes the need for case-by-case optimization in plasmid-based expression systems.

In conclusion, this chapter provides valuable insights into the challenges of plasmid copy number quantification and the complex dynamics of plasmid behavior in bacterial cells. The results underscore the importance of carefully considering plasmid backbone, cargo, and combinations when designing expression systems. The inconsistencies observed between different quantification methods and the unexpected behaviors of some plasmid combinations, as well as the nuanced dependency of transformation efficiency on OD and DNA quantity, highlight the need for further research in this area. Future studies should focus on developing more reliable methods for plasmid copy number quantification and on unraveling the mechanisms behind the observed plasmid interactions, as well as optimizing transformation conditions. Such research will be crucial for advancing our understanding of plasmid biology and improving the design of synthetic biological systems.

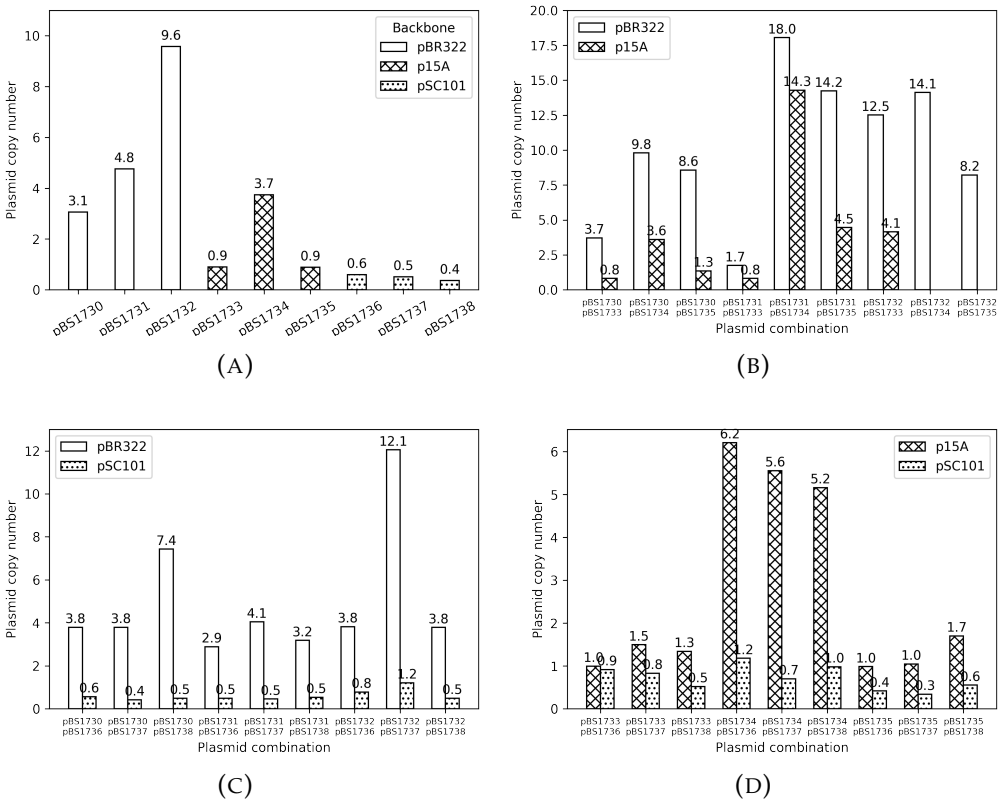


FIGURE 3.4: Plasmid copy numbers from the WGS dataset using the k-mer based PlasmidSeeker method (n=1). Description of plasmids can be found in Table 3.1. Plasmids present in the strains as labels on the x axis in each subplot. (A) : Strains carrying only one type of plasmid without cargo, with mCherry and with mKalama expressed respectively in each backbone. (B) : Strains carrying pBR322 and p15A plasmids with all possible cargo combinations. (C) : Strains carrying pBR322 and pSC101 plasmids with all possible cargo combinations. (D) : Strains carrying p15A and pSC101 plasmids with all possible cargo combinations.

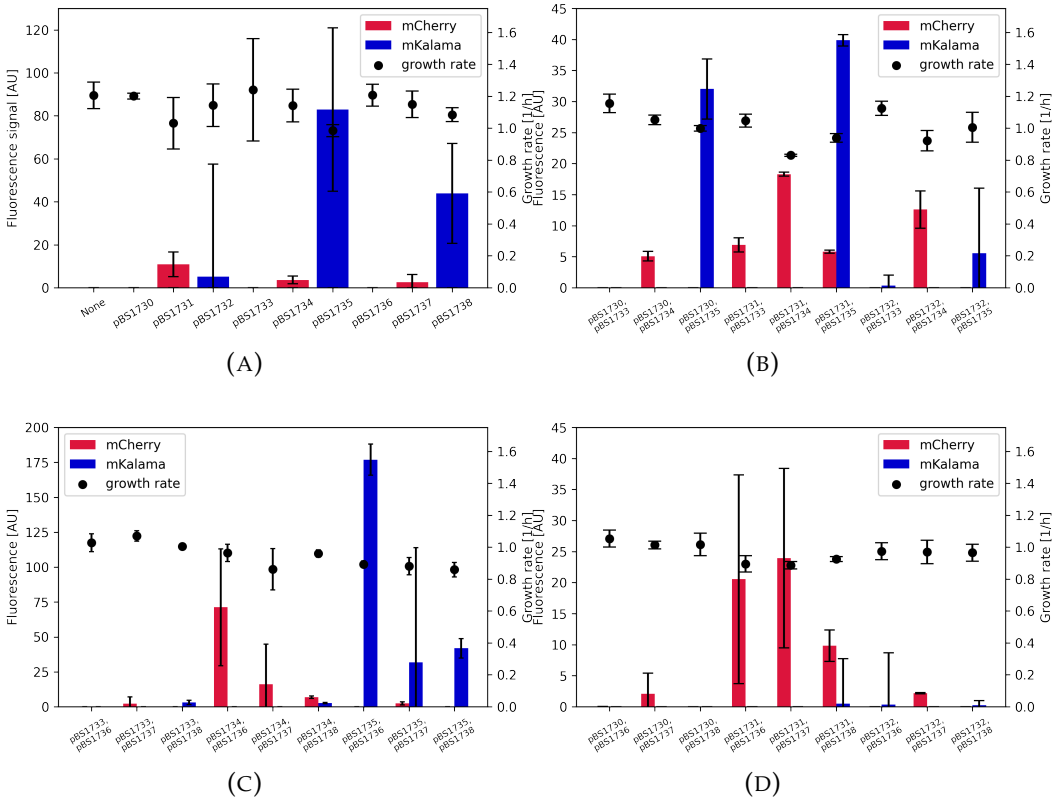


FIGURE 3.5: Fluorescence and growth rate results from the "Hamburg" dataset (n=3). Error bars representing the standard deviation between the replicates. Fluorescence represented by bar graphs and growth rate by dots. Plasmids present in the strains are indicated in the x axis in each subplot. (A) : Strains carrying only one type of plasmid without cargo, with mCherry and with mKalama expressed respectively in each backbone. (B) : Strains carrying pBR322 and p15A plasmids with all possible cargo combinations. (C) : Strains carrying p15A and pSC101 plasmids with all possible cargo combinations. (D) : Strains carrying pBR322 and pSC101 plasmids with all possible cargo combinations.

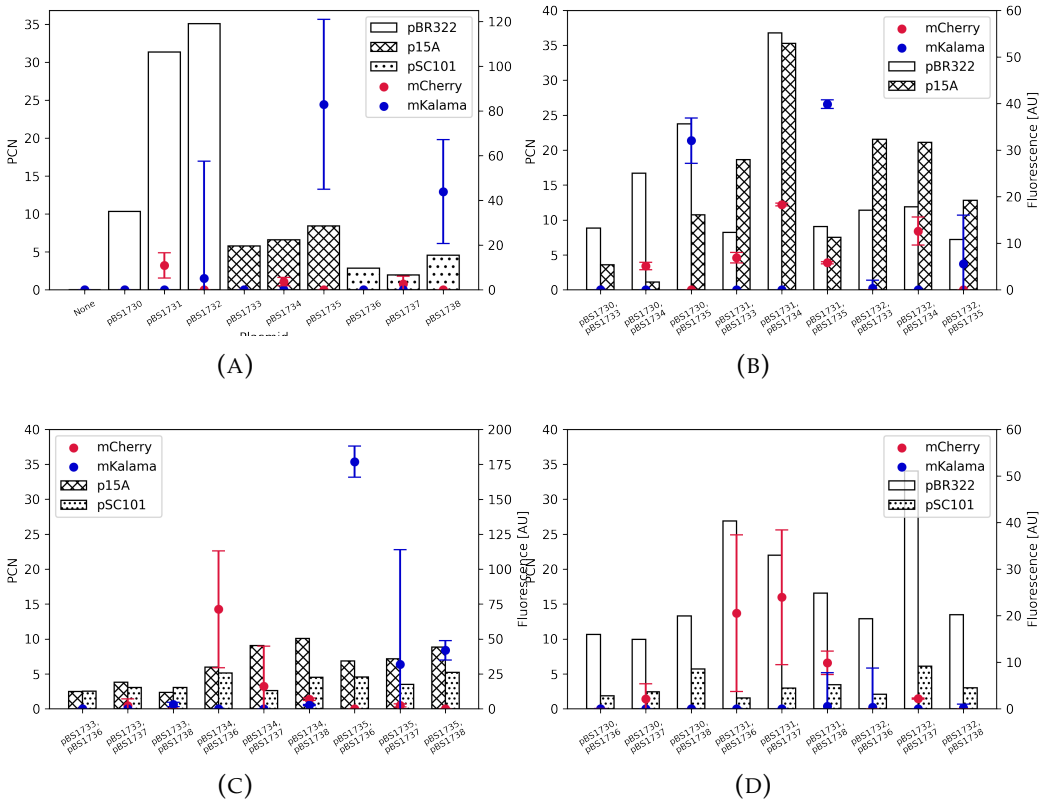
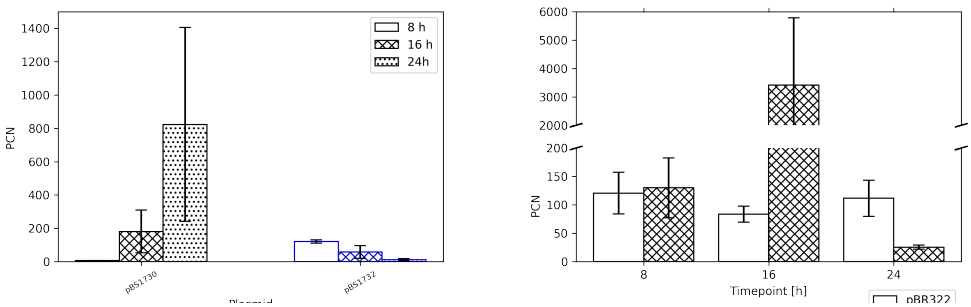


FIGURE 3.6: Fluorescence and plasmid copy number results from the "Hamburg" dataset. (n = 3) Error bars representing the standard deviation between the replicates. Fluorescence represented by dots and plasmid copy number by bars. Plasmids present in the strains are indicated in the x axis in each subplot. (A) : Strains carrying only one type of plasmid without cargo, with mCherry and with mKalama expressed respectively in each backbone. (B) : Strains carrying pBR322 and p15A plasmids with all possible cargo combinations. (C) : Strains carrying p15A and pSC101 plasmids with all possible cargo combinations. (D) : Strains carrying pBR322 and pSC101 plasmids with all possible cargo combinations.



(A) (B)

FIGURE 3.7: verification trial of the "Hamburg" dataset. (n=3) Error bars representing the standard deviation between the replicates. (A) : Plasmid copy numbers of strains carrying empty pBR322 (pBS1730) or mKalama (pBS1732) plasmids at different timepoints. (B): Plasmid copy number of strains carrying empty pBR322 (pBS1730) and mKalama (pBS1732) plasmids at different timepoints. With solid bar representing pBR322 copy number and chequered bar representing p15A copy number.

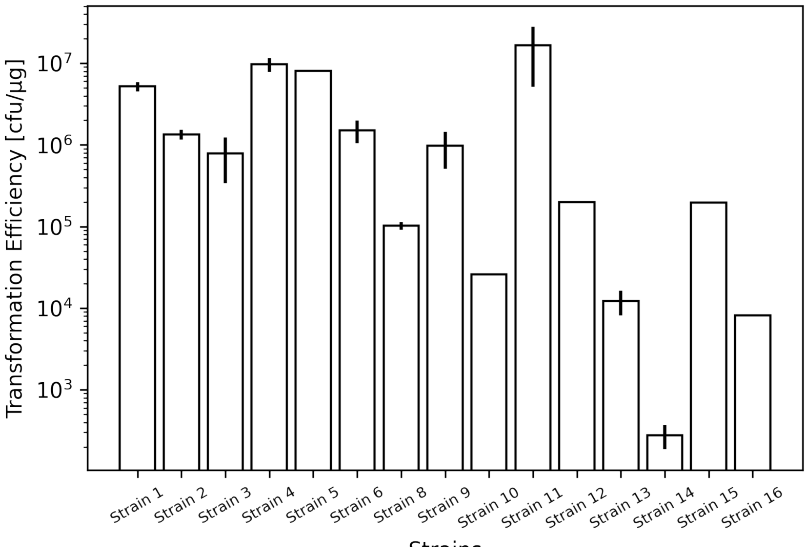


FIGURE 3.8: Transformation efficiency [cfu/µg DNA] of each successfully transformed strain in the study (n=3). Error bars representing the standard deviation between the replicates.

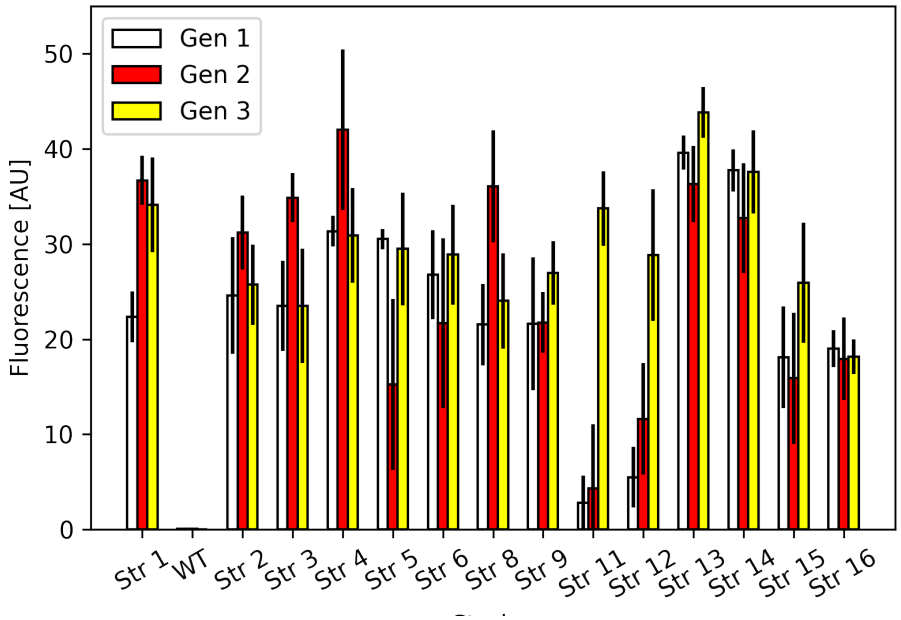


FIGURE 3.9: Fluorescence signal of strains from the transformation study in each generation (n=3). Error bars representing the standard deviation between the replicates.

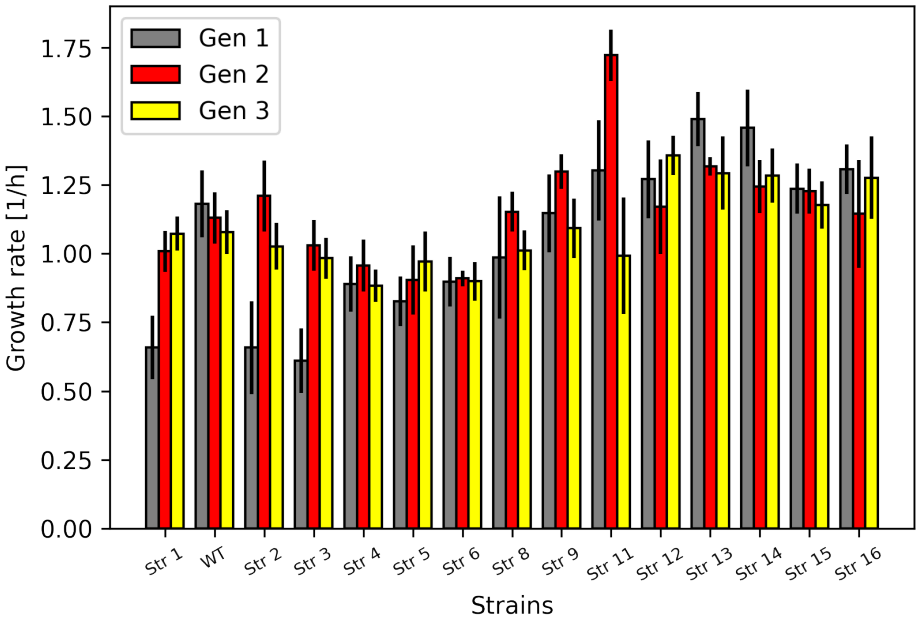
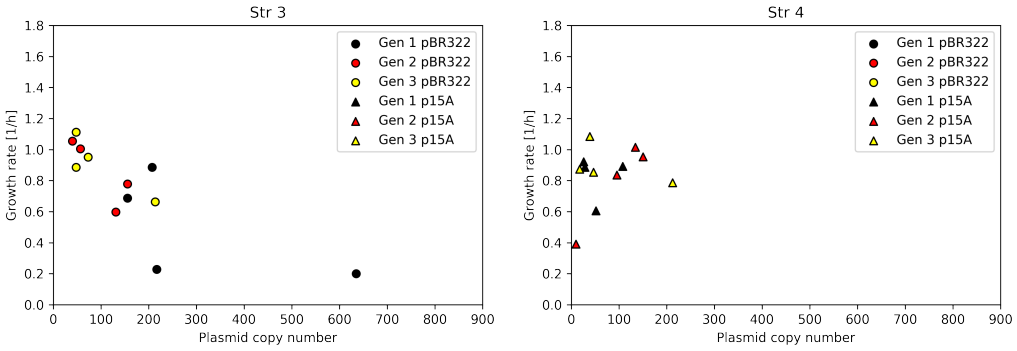
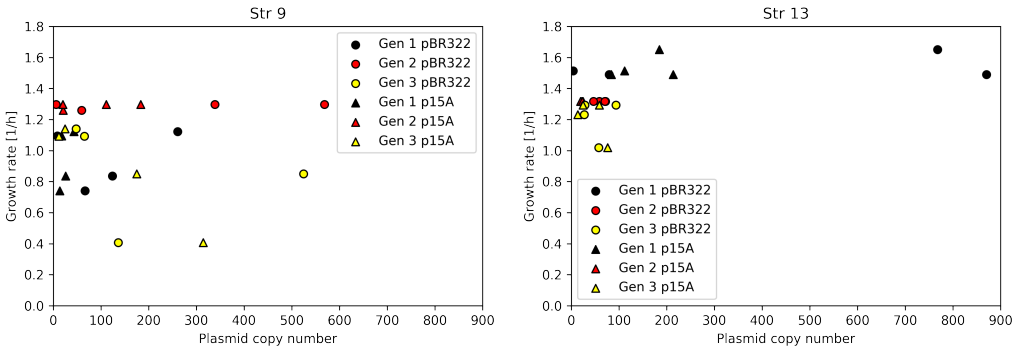


FIGURE 3.10: Growth rate of stains from the transformation study in each generation (n=3). Error bars representing the standard deviation between the replicates.

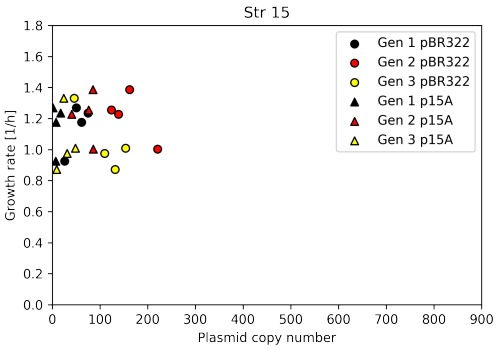


(A) (B)

FIGURE 3.11: PCN vs growth rate for transformation study strains in detail part 1

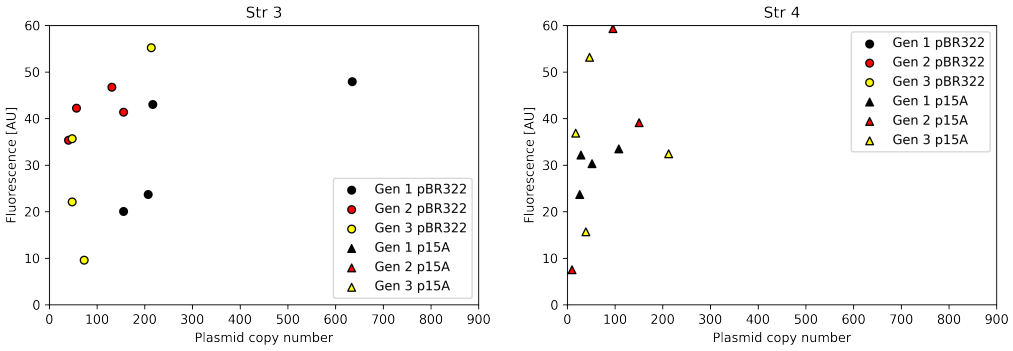


(A) (B)



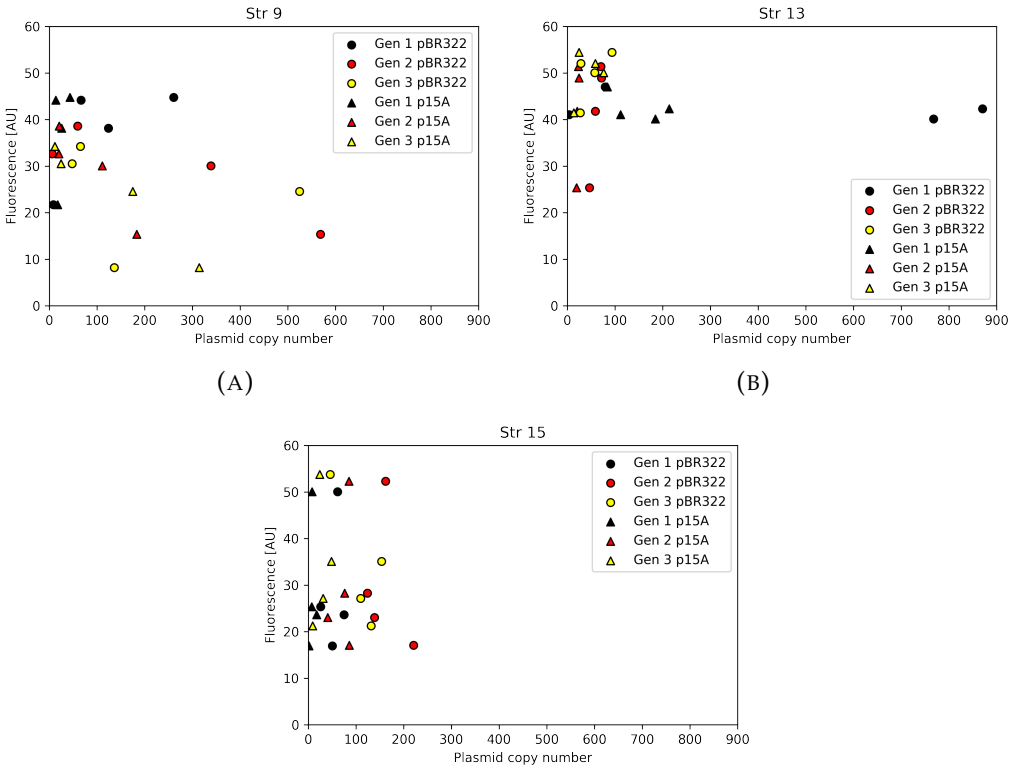
(C)

FIGURE 3.12: PCN vs growth rate for transformation study strains in detail part 2



(A) (B)

FIGURE 3.13: PCN vs fluorescence for transformation study strains in detail part 1



(A) (B)

(C)

FIGURE 3.14: PCN vs fluorescence for transformation study strains in detail part 2

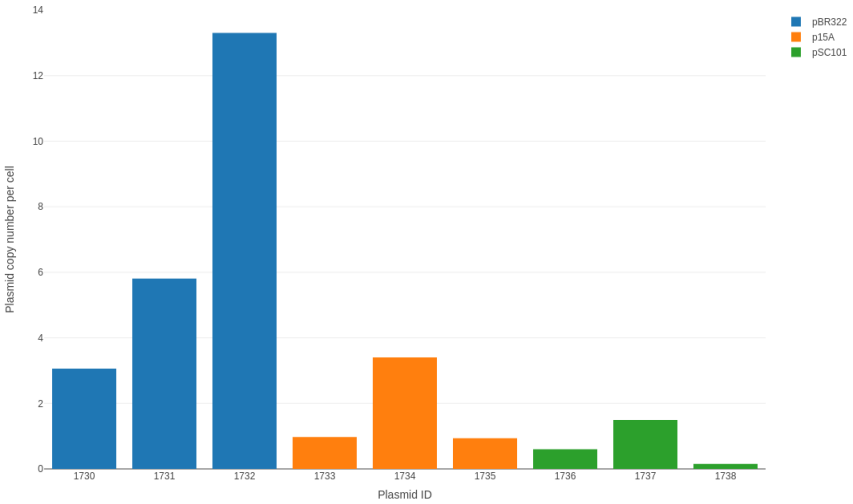


FIGURE 3.15: Plasmid copy numbers from the WGS dataset using the built-in quantification method of QIAGEN CLC Genomics Workbench 21.0, (n=1). Description of plasmids can be found in Table 3.1. Plasmids present in the strains as labels on the x axis

4 Investigating by-product formation in the biotin pathway with metabolomics

4.1 Biotin pathway

Biotin is a water-soluble organic heterobicyclic compound, which is present in small amounts in most living cells. Biotin was first discovered in 1901, and its role in various metabolic pathways has been extensively studied in all forms of life (Lin, Hanson, and Cronan, 2010). It has a wide variety of uses in the cells, but only fungi, bacteria and plants can synthesize it. Mammals need to take in biotin from the diet or via the microbiota (Lin and Cronan, 2011). In enzymes such as pyruvate carboxylase, biotin functions as a prosthetic group because it is covalently attached to the enzyme, ensuring that the carboxyl transfer process is both efficient and tightly controlled. In contrast, within acetyl-CoA carboxylase, biotin acts as a coenzyme by transiently accepting and transferring an activated carboxyl group between distinct active sites, thus facilitating the catalytic cycle without being permanently bound. Additionally, biotin's essential presence in these carboxylases highlights its role as a cofactor, where it is indispensable for maintaining proper enzyme conformation and activity. This multifaceted functionality of biotin underscores its vital contribution to metabolic pathways by carrying carboxyl groups, whether it is permanently anchored as a prosthetic group, transiently operating as a coenzyme, or generally required as a cofactor for enzymatic activity.

Also known as vitamin B7 or vitamin H, it is widely used in the food and feed industry and the cosmetics industry. Its widespread use creates a 1000 ton/year production need to satisfy the needs of various industrie (Acevedo-Rocha et al., 2019). Hence efficient and profitable production of biotin is of interest for the chemical industry. Currently, all commercial biotin

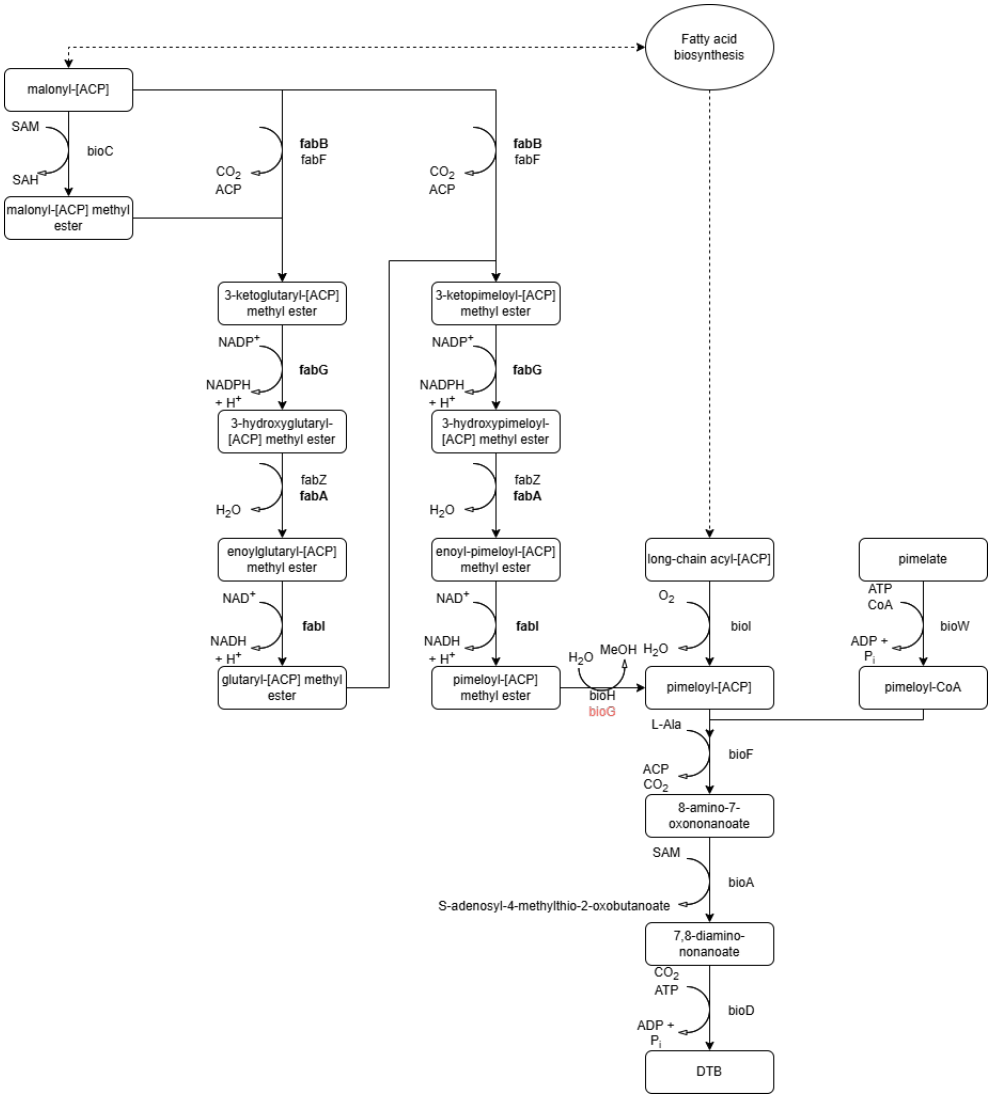


FIGURE 4.1: Schematics of the *E.coli* DTB biosynthesis pathway. Bold genes are under control of *fadR* transcription factor. bioG, which is non-native in *E.coli* is shown in red.

is produced chemically by a costly multistep process that uses significant amounts of organic solvents and generate toxic waste streams (Chen et al., 2007; Xiong et al., 2016).

Production of biotin by fermentation would provide a sustainable alternative and could ease the environmental pressure created from the chemical production. Therefore, the engineering of a biotin cell factory is of interest. In this thesis, we are investigating *E.coli* strains engineered towards efficient biotin production. Our goal is to create a strain that produces sufficient biotin precursors as part of a larger project, with the final goal of creating an efficient cell factory for biotin production from glucose. The precursor we are aiming to produce is desthiobiotin, the substrate of bioB, the last step in the biotin pathway. Apart from some dedicated enzymes (BioC,H,F,A,D,B), the biotin biosynthetic pathway in *E.coli* also borrows the fatty acid biosynthetic enzymes to act on biotin specific substrates (Figure 4.1).

The first reaction in the biotin biosynthesys pathway is a methylation, where bioC transfers a methyl group to the malonyl-ACP from a *S*-adenosyl-L-methionine (SAM). This reaction was challenging to characterize as the *E. coli* enzyme is difficult to investigate in vitro due to its instability. In the lack of bioC, the strain cannot produce biotin, therefore it is considered an essential gene. Overexpression of the methyltransferase can result in toxicity due to the accumulation of malonyl-ACP methyl ester.

The following reactions are part of the elongation cycle of the fatty acid biosynthesis. The second step of the biotin biosynthesis can be carried out with two enzymes, fabB and fabF. Both enzymes are 3-oxoacyl-[acyl carrier protein] synthases which are parts of the fatty acid elongation cycle. fabF requires ACP as the acyl donor, while fabB doe not. *FabB* is under the regulation of the transcription factor fadR. The ratio of fabA and fabB is a key factor in the regulation of saturated and unsaturated fatty acid balance in the cell. fabB is under the regulation of transcription factor fadR.

The next step in the fatty acid elongation cycle is carried out by fabG. fabG is a reductase, which applies a 2 hydrogens from the cofactor NADPH + H to saturate the elongated ACP-bound fatty acid methyl ester. In previous studies it was shown to be specific for NADPH and favored ACP-bound substrates over CoA derivates. But fabG was not selective to the chain-length of the substrate. fabG is an essential gene in *E. coli* and is under the regulation of transcription factor fadR.

Both fabA and fabZ are dehydratases that catalyze the next step in the fatty acid elongation cycle. fabA is mainly active on intermediate chain-length substrate but the two has overlapping substrate specificities. Out of the two fabZ has the broader substrate specificity, and while it is mainly involved in the unsaturated branch of the fatty acid synthesis (Heath and Rock, 1996b). fabA does not only harbors the dehydratase functionality but it also shows isomerase activity. It has an essential role in the synthesis of unsaturated fatty acids. fabA is regulated under the transcriptional activator fadR and transcriptional repressor fabR.

fabI is an enoyl-acyl carrier protein reductase, which is the last step in the fatty acid elongation cycle. It catalyzes an essential reaction in the fatty acid biosynthesis, as it reduces the 2,3-double bond. The enzyme can be product inhibited, therefore its balancing is an essential factor for optimization of flux through the elongation cycle (Heath and Rock, 1996a; Heath and Rock, 1995).

bioH catalyzes first reaction, which branches out from the fatty acid elongation cycle part of the biotin biosynthesis. This reaction is catalyzed via an esterase, which hydrolyses the methyl ester from the ACP-bound pimleoyl methyl ester. With this reaction the product is removed from the elongation cycle and is channeled to the biotin biosynthesis pathway. The promiscuity of bioH has been suspected based on various studies (Song, 2009; Lin, Hanson, and Cronan, 2010). The physiological substrate is pimeloyl-ACP methyl ester, but it might have catalytic activity in hydrolyzing the methyl ester from shorted chained fatty acid metabolites. If we hypothesize that this enzyme can act on most ACP-bound methyl ester intermediates in the fatty acid elongation cycle, we can expect to have accumulation of ACP-bound intermiediates without methyl ester groups.

BioF encodes the enzyme 8-Amino-7-oxononanoate synthase, which catalyzes the decarboxylative condensation pimeloyl-ACP and L-alanine and removes the ACP from the molecule. In the target biotin pathway, this reaction from KAPA to DAPA. This enzyme is known to be promiscuous and take pimeloyl-ACP and pimeloyl-CoA as its substrate. When pimeloyl-CoA is the substrate, the catalyzed reaction removes the coenzymeA group. Therefore, we can suspect that the possible promiscuity of this enzyme could act on ACP-bound fatty acid intermediates creates by the promiscuous bioH. In case the enzyme can also act on ACP-bound methyl ester intermediates, it could give rise to various methyl esters branching out from the ACP-bound

methyl esters of the fatty acid elongation cycle. *BioF* is not an essential gene in either complex or minimal media for *E. coli*.

BioA encodes 7,8-diaminopelargonic acid synthase, which is in the standard biotin pathway is the reaction from DAPA to DTB. In its catalyzed reaction, it transfers an amino group from SAM to the *8S*-8-amino-7-oxononanoate. This reaction is the only known example of SAM acting as an amino-donor instead of a carboxyl-donor. Assuming that the enzyme is promiscuous, we can expect it to be acting on the products of bioF, which are not ACP-bound, and either methylated or not. Depending on the first enzyme that acted on the ACP-bound methylated fatty acid biosynthesis intermediate.

BioD encodes the dethiobiotin synthase, which closes the first ring in the bicicylic structure of biotin. This is a defining, irreversible step in the biotin pathway. The reaction also needs ATP as substrate and is inhibited by its product, ADP. If we assume that the enzyme is promiscuous, we can expect it to acting on the product of bioA, and to create slight variations of DTB based on the promiscuity of bioH and bioF branching out from the fatty acid elongation cycle metabolites.

4.2 Hypothesis based on previous data

Previous observations suggest that the imbalanced pathway can lead to promiscuous activity of the biotin synthesis enzymes on earlier intermediates of the fatty acid elongation pathway (Fig 4.2). Accumulation of off-pathway metabolites generated by promiscuous activity decreases the efficiency of a cell factory. Carbon from the feed will be trapped in these dead-end metabolites and decrease the yield of the production. The by-products might cause toxicity or increase the effect of product inhibition. Investigating the levels of various metabolites throughout the biosynthetic pathway could identify possible bottlenecks and optimize the cell factory. It could also provide information about novel enzymatic activities of the biotin biosynthesis enzymes that have previously not been reported.

An unknown byproduct appeared regularly on the TLC assay for biotin vitamers, the intensity of the spot showed correlation to the accumulation of DTB and leveling out of the biotin production. The unknown byproduct

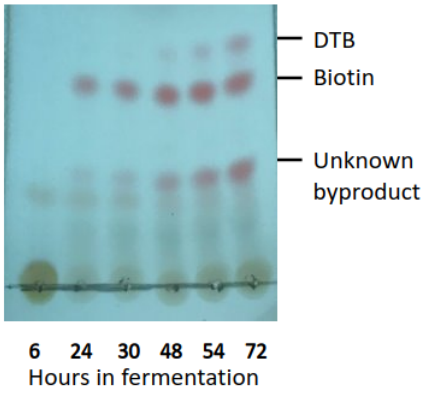


FIGURE 4.2: Thin layer chromatography of fermentation supernatant samples

was purified from the fermentation broth and underwent LC-MS and NMR analysis. The compound was identified as 3-hydoxydesthiobiotin. Its accumulation over time in the fermentation broth suggests that it is a dead-end metabolite. Therefore, we hypothesize that its presence could be inhibiting reactions upstream. In order for 3-hydroxydesthiobiotin to be present and accumulate, we suspect that bioD is promiscuous and closes the ureido ring on compounds close in structure to DAPA. We also assume that therefore other participants of the biotin pathway are also promiscuous, in order for the precursors of 3-hydroxydesthiobiotin to exist. This would suggest that the fabA / fabZ reaction is a bottleneck in the system. As 3-hydroxydesthiobiotin accumulates in quantities that can be visualized on TLC from fermentation samples, there is a possibility that other intermediates also accumulate at bottlenecks, which we cannot visualize with the current in-house analytical methods. The investigation and identification of the unknown compound was led by Markus Jondelius Hederos (Biosyntia ApS) and Linda Ahonen (Biosyntia ApS).

4.3 First DBTL cycle

In order to investigate our hypothesis and improve on the DTB production of the engineered strain, we aimed to set up a new DBTL cycle. The DBTL cycle contains varying factors that could affect the production of the hypothesized by-products. During this cycle, we would identify and quantify possible bottlenecks in the Test step and understand the possible reasons for their

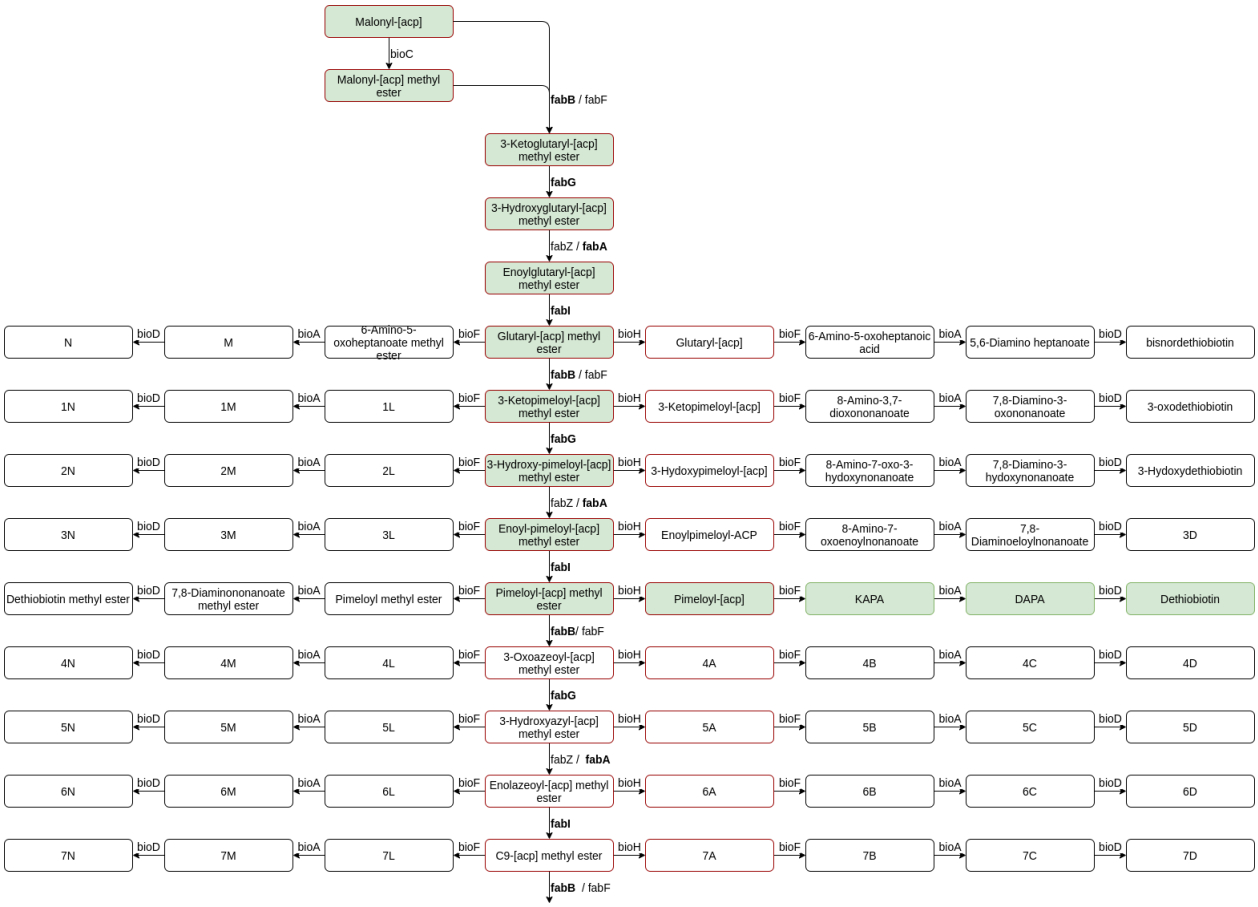


FIGURE 4.3: DTB pathway in *E.coli* with possible by-products due to enzyme promiscuity of *bioH* and *bioF*. Compounds with green background are part of the preferred DTB production pathway. Compounds with red border are not possible to detect with the high-throughput untargeted metabolomics method. Bold genes are under control of fadR transcription factor

accumulation in the Learn step.

During this experimental DBTL cycle, we chose a high-throughput analytics method, which allows us to carry out quick iterations of the cycle. A drawback of using such a "quick and dirty" analytics method is that we are unable to identify quantitatively or qualitatively the ACP-bound intermediates and possible by-products. The ACP-bound intermediates could also accumulate and there are possible methods such as (Greg Bokinsky's proteomics method) to investigate this aspect of the strains. In a different DBTL cycle, we encourage to include those methods, but they were outside of the scope of this study. One of the most critical steps during the first cycle is to create an initial library of possible compounds that could accumulate. As we

hypothesize that the biotin biosynthesis enzymes can act on different intermediates of the fatty acid elongation cycle, we created a library of hypothetical compounds that branch out from these intermediates (Figure: 4.3).

As our hypothesis involves multiple genes in the fatty acid metabolism, which are under the regulation of transcription factor fadR, we included the expression of the transcription factor. FadR is a global dual regulator, which can act as a switch between fatty acid beta-oxidation and fatty acid biosynthesis. Overexpression of the transcription factor results in 5 - 7.5-fold increase in fatty acid production The additional expression of FadR could increase the general availability of fatty acids and the flux into the fatty acid biosynthesis. As *fabD* is also under the regulation of FadR transcription factor, malonyl-CoA could also be more available in the cell as a substrate for bioC. This additional flux could benefit the biotin biosynthesis via increasing the availability of substrate into the pathway.

For initialization of the DBTL cycle, our starting strain was an already engineered *E.coli* strain. The biotin biosynthesis genes, which are responsible for the production of biotin and its precursors, were deleted from the engineered strain in order to be able to measure the effect of biotin production efficiency and minimize the background signal. Table 4.1 summarizes the strains planned during the initial Design step of the DBTL cycle. Strains A, B, and C fulfill the purpose of controls, as strain A is expected to show the effect of the production of the enzymes without the burden of their activity, while strain B is expected to show the burden originating from carrying the plasmid backbone with the same antibiotic marker. Strain C is our base production strain that we are aiming to improve on. The remaining strains are expected to show differences in their DTB production and metabolic profile. The DTB plasmids in the strain are varying in the strength of the constitutive expression of *BioF*, decreasing from v0 to v2. The data is expected to give valuable information about the possible accumulating pathway by-products.

4.3.1 Build step

Plasmids

Plasmids pBS2145, pBS2095, pBS2242, pBS2246, pBS2162 were created by Nils Myling-Petersen (Biosyntia ApS) and Bo Salomonsen (Biosyntia ApS).

Strain	Background strain	Plasmids	Description
A	MG1655 Δbio	pBS2145	Defective DTB plasmid
B	MG1655 Δbio	pBS1733	Empty p15A plasmid
C	MG1655 Δbio	pBS2095	DTB plasmid v0
D	MG1655 Δbio	pBS2242	DTB plasmid v1
E	MG1655 Δbio	pBS2246	DTB plasmid v2
F	MG1655 Δbio	pBS2095, pBS2162	DTB plasmid, fadR plasmid

TABLE 4.1: Strains used in the first DBTL cycle.

Plasmid pBS1733 was created by Dóra Vitay. All genes were amplified from the genome with PCR, using PhusionU polymerase, and plasmids were created from the amplified fragments with USER cloning. Backbones of plasmids pBS2145, pBS1733, pBS2095, pBS2242, pBS2246 were obtained from in-house p15A backbones with kanamycin resistance marker. The backbone of pBS2161 was amplified from in-house PSC101 with a spectinomycin resistance marker. Due to confidentiality restrictions, the DNA sequences cannot be included in this thesis. These sequences are considered confidential information of the company and cannot be publicly disclosed.

Strains

Background strain BS4755 was created by Bo Salomonsen (Biosyntia ApS). Strains present in the study were created via chemical transformation of the background strain with the plasmids mentioned above.

4.3.2 Test step

Cultivation and sampling

All experiments were carried out in minimal media (mMOPS), which is described in the supplementary materials. Strains created were stored as 1% glycerol stocks at -80°C. During the test step, strains were plated on mMOPS agar plates supplemented with the correct combination of antibiotics to retain the plasmids and with 1nM biotin to overcome the auxotrophy of the strain. Cultivation for sampling for DTB production and metabolomics measurements were carried out in 24-well deep-well plates with the final volume of 4 ml. For all cultivations, mMOPS complemented with 1 nM biotin was used with the addition of antibiotics to ensure that plasmids are

retained in the strain. Cultivations were run for 24 hours at 37°C with 250 rpm agitation, and plates were sealed with sterile, breathable seal. Sampling was carried out at 4h and 24h timepoints. 10µl of the taken samples were diluted with mMOPS 10- and 100-fold to assess their OD600 with a plate reader. The samples were then normalized to 1 OD/ml with MQ dH2O. The samples were spun down with tabletop microcentrifuge (2 minutes at 10000 x g). 100µl of the supernatant were saved for the in-house DTB assay described below. 200µl of the supernatant were sent to ETH Zurich for untargeted metabolomics analysis. The cell pellets were prepared for untargeted metabolomics analysis with hot water extraction.

Assessment of DTB production

Supernatants of the samples were assessed for the presence of DTB with a bioassay method, where the bioassay strain is auxotrophic for DTB, via a deletion in the *bioD* gene. Therefore it can only grow from taking up DTB from the environment. The strain and the assay was created by Daniel Bonifácio at Biosyntia and the details are confidential.

Metabolomics

The untargeted metabolomics measurements of the samples extracted with hot water were carried out by Karin Meier, Ph.D. in the research group of Prof. Dr. Uwe Sauer at ETH Zurich (Sauer, 2024).

4.3.3 Results of Test step

Assessment with in-house DTB assay

With the in-house bioassay, we measure the produced DTB that is secreted into the media during the cultivation. The method does not take intracellular DTB accumulation into consideration. This in-house method is used routinely, as previous experiments showed that the metabolites of interests (DTB and biotin) were not present in the cell pellet samples in a significant amount. The cells tend to secrete most of the produced excess of these compounds. We used this method as a proxy for efficient DTB production, as the

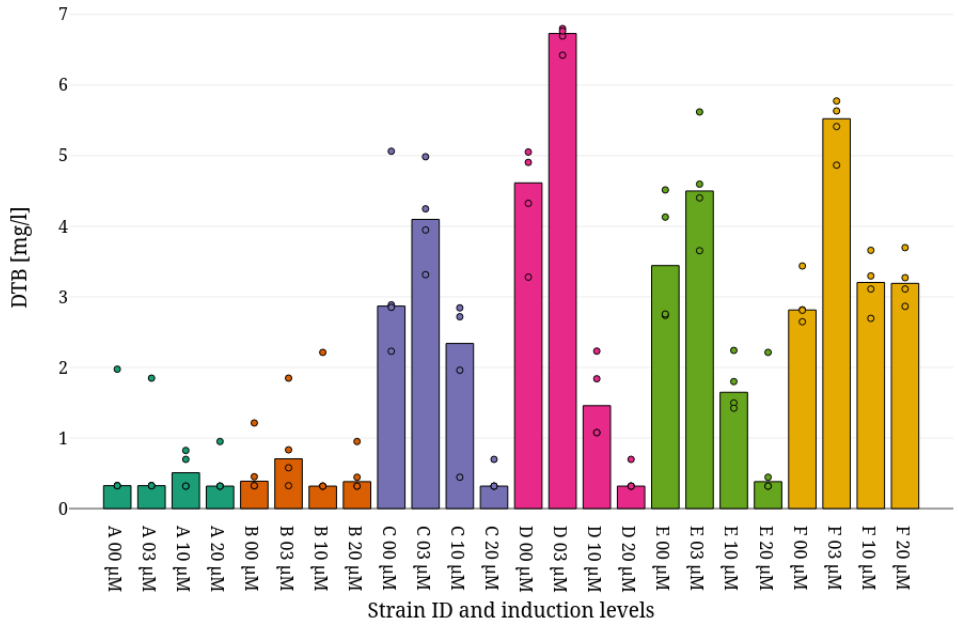


FIGURE 4.4: DTB production of investigated strains at various IPTG levels in minimal media at 24h. (n=4)

downstream processing of the final product from fermentation broth is preferred in industrial production. In case there is a need for cell lysis in order to gain higher yields during fermentation, the downstream processing cost of the product will significantly increase.

We can observe little to no production in strain A and B, which provides us with a good baseline to assess the performance of the production strains. All production strains harbor inducible *BioC* and constitutively expressed *BioH, BioF, BioA, BioD*. The variability between the DTB plasmids in this DBTL cycle is the strength of the RBS in front of *bioF*. Strains were grown without inducer and with low (3 µM), medium (10µM) and high (20µM) induction levels. All strains intended for production (C, D, E, F) are generating significantly more DTB than the control strains (A, B) (Figure 4.4). Strain C produces 4 mg/l DTB at most at low induction levels. Production decreases with increasing induction levels and at high induction level it drops to the wild-type levels. The decreased production at high induction levels could be the result of materials trapped in dead-end by-products. The same toxicity can be observed for strains that harbor the other variations of the DTB plasmid. Although the production profile is similar in strain C, D and E, the varying expression levels of bioF still has an effect on the level of production. The decreased expression of bioF in strain D showed the highest production

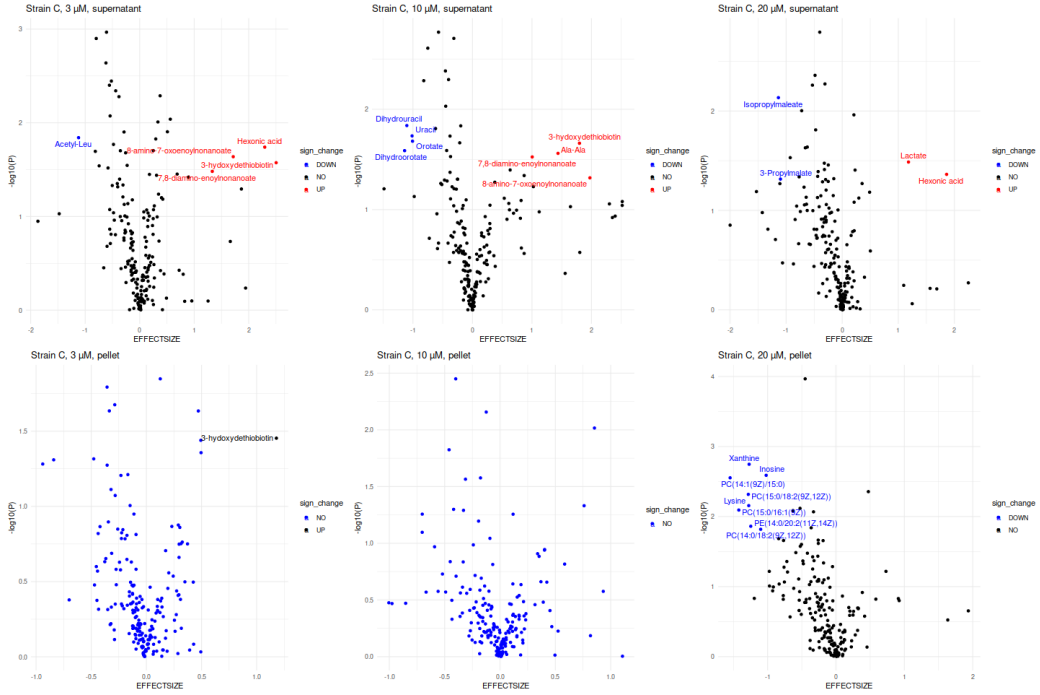


FIGURE 4.5: Volcano plot of strain C vs strain A (n=4)

amongst the investigated strains. By decreasing the amount of available bioF, the chance of the non-occupied enzymes carrying out side-reactions is also lower. This also supports the hypothesis that the bottleneck could be in the fatty acid biosynthesis enzymes, and that ACP-bound intermediates are accumulated in the production strains due to slow flux through the fatty acid elongation cycle. There is a difference in the production profile of strain F, where the production is still the best at the low induction level, but the strain sill produces significant amount of DTB at medium and high induction levels. This observation can suggest that the expression of fadR does increase the availability of precursors for the DTB pathway compared to strain C. Its expression does not increase the production at medium and high induction levels to the low induction levels. Hence we can suspect that increasing the flux into and through the fatty acid biosynthesis is a good step towards a better DTB producer strain, but requires fine-tuning along with the expression of the biotin biosynthesis genes.

Metabolic profiles

We searched for the byproducts shown in Figure: 4.3 by exact mass. Indeed, masses corresponding to 15 out of the 54 compounds were detected in at least

some of the strains overexpressing a DTB pathway. This indicates that the hypothesized off-pathway metabolites are produced and that at least some promiscuous activity of the bioH,F,A and/or D is present. As the method used for this is not quantitative we cannot estimate the amount of these compounds produced.

Although 15 different peaks with the same molecular weight as the hypothetical by-products were observed in the metabolic profile, only a few of them presented significant changes in their levels across samples. The untargeted metabolomics method, as mentioned before, does not include a chromatography step, therefore compounds with the same molecular weight cannot be separated and identified. A clear example of this limitation is compound KAPA, a substrate of bioA in the biotin pathway. As its molecular weight is equal to the molecular weight of polyethylene glycol, we cannot differentiate between KAPA and polyethylene glycol levels. We can only observe the overall change in the peak intensity and can deduct that changes in that peak can be the effect of changing KAPA levels in the strain. In order to be able to measure these critical metabolites in the pathway quantitatively, we need to apply a different method, such as analytical methods containing a chromatography step.

When comparing the metabolic profile of strain C (DTB pathway) to strain A (defective DTB pathway), we can see the accumulation of 8-amino-7-oxoenoylnonaoate 7,8-diamino-enoylnonaoate and 3-hydroxydesthiobiotin at 3µM and 10µM induction levels in the supernatant samples (Figure 4.5). At 20µM, where the production diminishes, we can observe decreased levels of various phospholipids in the pellet sample. This observation suggests that the current DTB production strain could have a bottleneck at fabI, as the by-products 8-amino-7-oxoenoylnonaoate and 7,8-diamino-enoylnonaoate are hypothetical by-products branching out if bioH would act on enoyl-pimeloy-[ACP]-methyl ester instead of its native substrate pimeloyl-[ACP]-methyl ester. The accumulation of 3-hydroxydesthiobiotin could also result from the promiscuous activity of bioH and the rest of the downstream enzymes present in the DTB pathway (bioFAD). These results confirm our hypothesis that compared to the wild-type strain, the production strain engineered to produce DTB in excess produces by-products due to the promiscuous activity of the bioH enzyme. The decreased abundance of various phospholipids in the highly induced samples confirms the hypothesis that when the

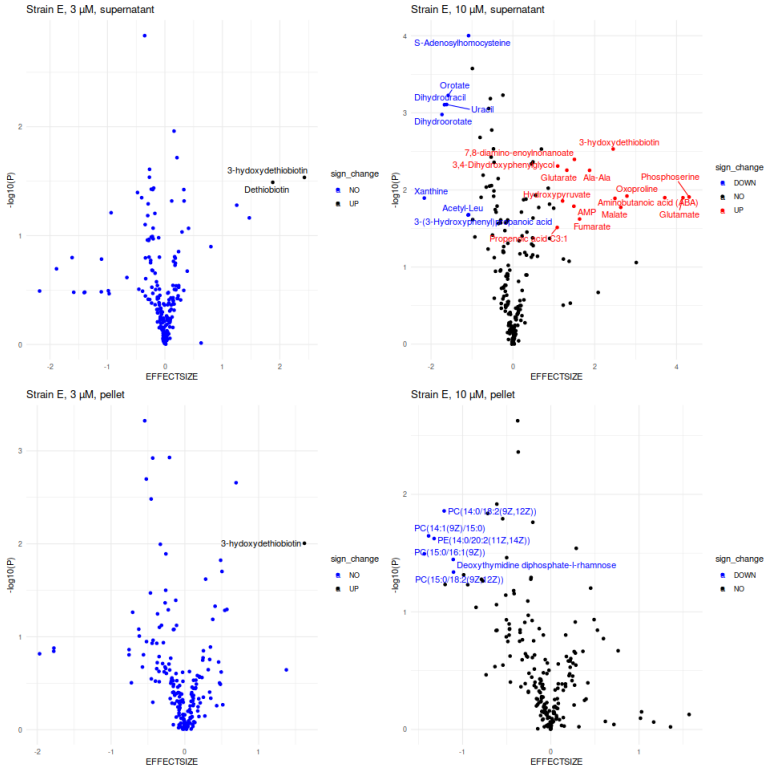


FIGURE 4.6: Volcano plot of strain E vs strain A (n=4)

engineered pathway is induced to toxicity, the primary source of toxicity is the lack of lipids. The presence and significant increase in abundance of 3-oxodethiobiotin supports the hypothesis that the unidentified compound on the TLC and MS spectra could be the same compound.

Since strain C and E had similar production profiles, their metabolic profiles are comparable, with few significant changes. When we look at the metabolic profile of strain E compared to control strain A, we can see the accumulation of 3-hydroxydethiobiotin in both the pellet and supernatant

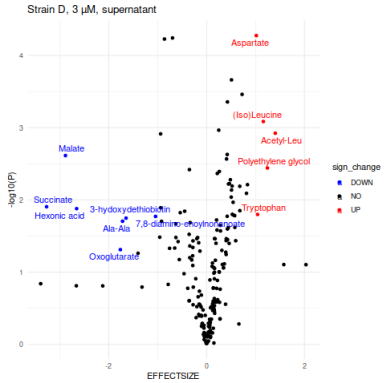


FIGURE 4.7: Volcano plot of strain D vs strain C (n=4)

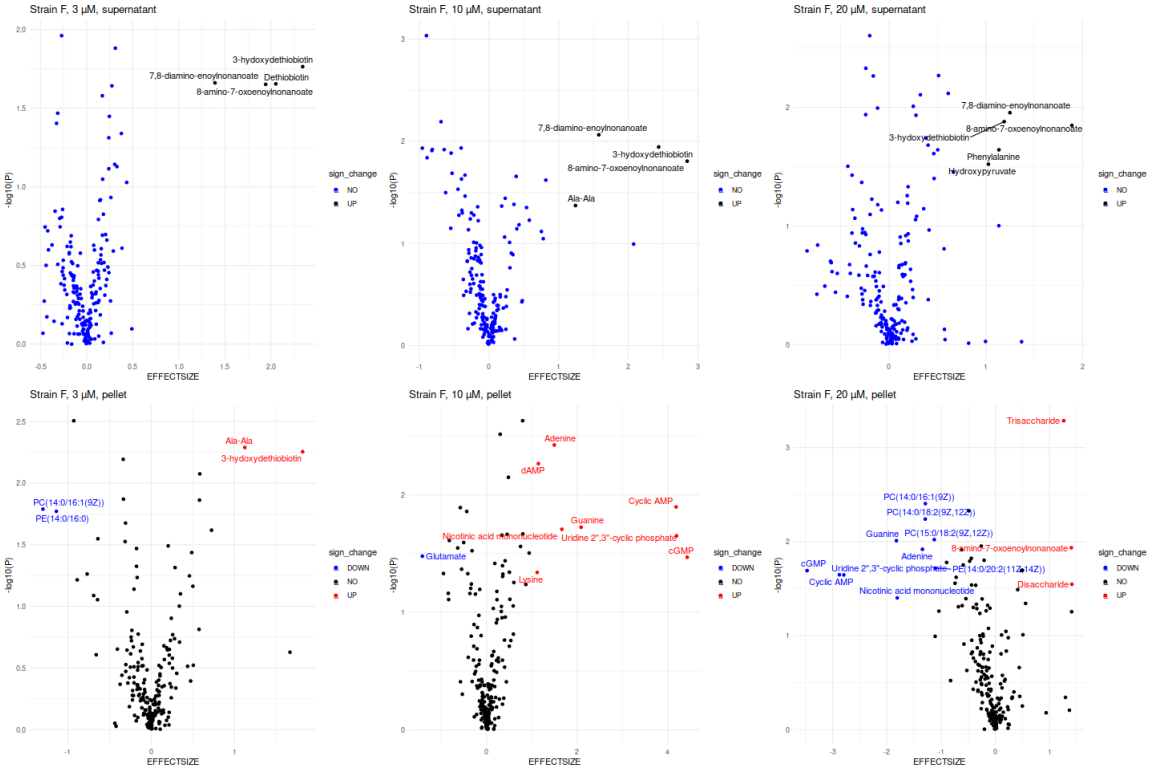


FIGURE 4.8: Volcano plot of strain F vs strain C (n=4)

samples at 3µM induction, but no other by-product is more abundant than in the control strain (Figure 4.6). This observation does fall in line with the presence of the same compound in strain C and confirms that these strains might have a bottleneck in the fabA and fabI catalyzed reactions. We can see the lack of lipids in the pellet sample at 20µM induction level, as well as the increased abundance of by-products 3-hydroxydesthiobiotin and 7,8-diamino-enoylnonanoate. It is indicating the presence of the same bottleneck as in strain C. Hence we can conclude that the decreased *BioF* RBS on plasmid pBS2242 compared to pBS2095 did not improve the flux through the DTB pathway. S-Adenosyl homocysteine also shows decreased abundance in strain E than in strain A.

Since strain D was the best performing strain in the in-house DTB production assay, its metabolic profile was compared to the metabolic profile of strain C (Figure 4.7). Interestingly, there was a significant difference at 3µM, where strain D shows decreased abundance in the by-products 3-oxodethiobiotin and 7,8-diamino-enoylnonanoate. These two metabolites are hypothesized by-products originating from the promiscuous activity of

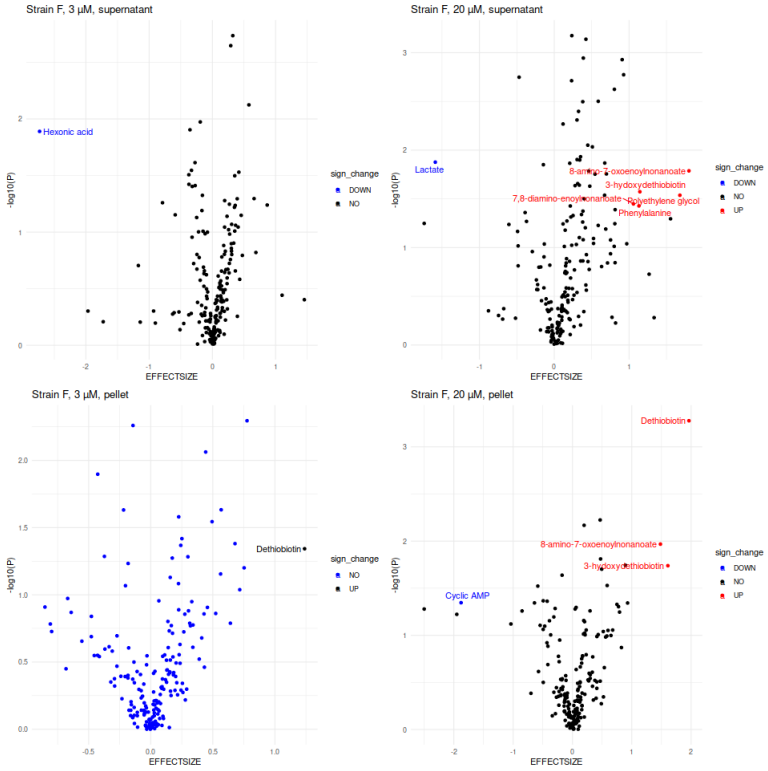


FIGURE 4.9: Volcano plot of strain F vs strain C (n=4)

bioH and Bio genes downstream of bioH. The decreased abundance compared to the DTB productions train to be improved confirms that decreasing the *BioF* expression does have a beneficial effect on the DTB production. As strain D and E only differ in the RBS strength of *BioF*, this result also highlights the need for fine-tuning the expression of the genes present in the DTB production pathway. Polyethylene glycol is in increased abundance in strain D, which indicates that KAPA could be accumulating in the samples, indicating that there could be a bottleneck at bioA. This is in line with more flux being on-pathway in this strain.

Strain F showed a widened induction window (Figure 4.4). When compared to control strain A (Figure 4.9 Top) strain F shows the accumulation of the same by-products as strain C at 3µM and 10µM. Therefore we cannot conclude that the expression of fadR transcription factor elevates the previously observed possibility of bottlenecks in the fatty acid and DTB biosynthesis pathways. At 20µM, strain F is significantly more abundant in hypothetical by-products than strain C at the same level of induction. We can also observe the increased abundance of our target product, DTB, which makes strain F an up-and-coming candidate to follow up. This result could indicate that the

DTB bioproduction pathway is still active at high induction levels, unlike in the production strain where fadR is not expressed in excess. The strain also shows the decreased abundance of phospholipids, just as strain C at 20µM, indicating that the availability of the lipids could still be a limiting factor for growth and production.

4.3.4 Learn step

We can conclude that some of the suspected by-products are present in strains showing DTB production. The main by-products accumulating were 3-hydroxydesthiobiotin, 7,8-diamino-enoylnonanoate and 8-amino-7-oxoenoylnonaoate. Based on the accumulation of these by-products, we can suspect a bottleneck in fabZ / fabA and fabI reactions. In case there is a bottleneck in fabA / fabZ reaction, there could be an accumulation of the product of fabG, 3-hydroxy-pimeloy-ACP methyl ester. We were not able to investigate the possible accumulation of ACP-bound by-products with the high-throughput metabolomics method.

In case of accumulation of 3-hydroxy-pimeloyl-ACP methyl ester, the promiscuous bioH could act on this pathway intermediate and create substrate for bioF and the rest of the DTB pathway enzymes. Accumulation of 7,8-diamino-enoylnonanoate and 8-amino-7-oxoenoylnonanoate could be the result of a bottleneck in the fabI reaction. If the substrate of fabI (enoyl-pimeloyl-ACP methyl ester) is in excess, the promiscuous bioH might act on it and create the by-products via the bio enzymes.

Expression of fadR is beneficial for DTB production, as the strain expressing fadR from a plasmid along a DTB production system present showed an increased production window in the induction gradient. As the by-product presence is similar in high-producing induction levels in strain C and F, it is likely that the effect of fadR on the overall fatty acid availability is more prominent than its effect on the genes present in the pathway of interest.

In conclusion, to improve DTB production in E.coli, it will be beneficial to include overexpression of fadR in the production strain. Decreasing the expression of bioF also has a positive effect on the production levels. In the future, a more precise balancing of the suspected bottleneck reactions (fabZ,

fabA, fabI) could decrease the accumulation of the by-products by creating an optimal flux through the DTB pathway.

4.4 Second DBTL cycle

4.4.1 Design step of the next iteration of the DTBL cycle

To better accommodate the *FadR* overexpression to our strains as an auxiliary gene for increasing production capacity, we increased its chromosomal expression by creating a new background strain. We hypothesized that the increased chromosomal expression would have a similar effect as the plasmid-based expression and would allow for testing other pathway and auxiliary plasmid combinations. In order to investigate the suspected bottlenecks, we included a new set of auxiliary plasmids, which carry *FabI, FabA, and FabZ* separately.

4.4.2 Build step

To increase the expression of fadR from the chromosome, we knocked in a strong promoter in front of the native *FadR* in the previously used background strain (BS4755) to create a new background strain (BS5973). With this method, we aim to increase the expression slightly in a way that the production burden of the protein does not become hindering for the cell.

4.4.3 Results of the second Test step

Assessment with in-house DTB assay

We can observe (Figure 4.10) that in strains without fadR overexpressed from the genome, the same trend can be observed as in the previous cycle. The DTB pathway consisting of the lower *BioF* RBS shows better overall DTB production. Expression of the fab auxiliary genes increased the DTB production in all cases, regardless of the DTB plasmid. The combination of low RBS *BioF* and auxiliary *FadR* plasmid showed that the strains perform similarly, with fabZ providing the highest production. While when the auxiliary plasmids

Strain	Background strain	Plasmid	Description
A	BS4755	pBS2145	'WT' background, defective DTB
C	BS4755	pBS2095	'WT' background, standard DTB
D	BS4755	pBS2242	'WT' background, low RBS bioF DTB
F	BS4755	pBS2095, pBS2162	'WT' background, standard DTB, fadR
G	BS4755	pBS2145, pBS2162	'WT' background, defective DTB, fadR
H	BS4755	pBS2095, fabI	'WT' background, standard DTB, fabI
I	BS4755	pBS2095, fabA	'WT' background, standard DTB, fabA
J	BS4755	pBS2095, fabZ	'WT' background, standard DTB, fabZ
K	BS4755	pBS2242, pBS2162	'WT' background, low RBS bioF DTB
L	BS4755	pBS2242, fabI	'WT' background, low RBS bioF DTB, fabI
M	BS4755	pBS2242, fabA	'WT' background, low RBS bioF DTB, fabA
N	BS4755	pBS2242, fabZ	'WT' background, low RBS bioF DTB, fabZ
O	BS5973	pBS2145, pBS2162	fadR background defective DTB, fadR
P	BS5973	pBS2095	fadR background standard DTB
Q	BS5973	pBS2095, pBS2162	fadR background standard DTB, fadR
R	BS5973	pBS2095, fabI	fadR background standard DTB, fabI
S	BS5973	pBS2095, fabA	fadR background standard DTB, fabA
T	BS5973	pBS2095, fabZ	fadR background standard DTB, fabZ
U	BS5973	pBS2242	fadR background low RBS bioF DTB
V	BS5973	pBS2242, pBS2162	fadR background low RBS bioF DTB, fadR
W	BS5973	pBS2242, fabI	fadR background low RBS bioF DTB, fabI
X	BS5973	pBS2242, fabA	fadR background low RBS bioF DTB, fabA
Y	BS5973	pBS2242, fabZ	fadR background low RBS bioF DTB, fabZ

TABLE 4.2: Strains used in the study

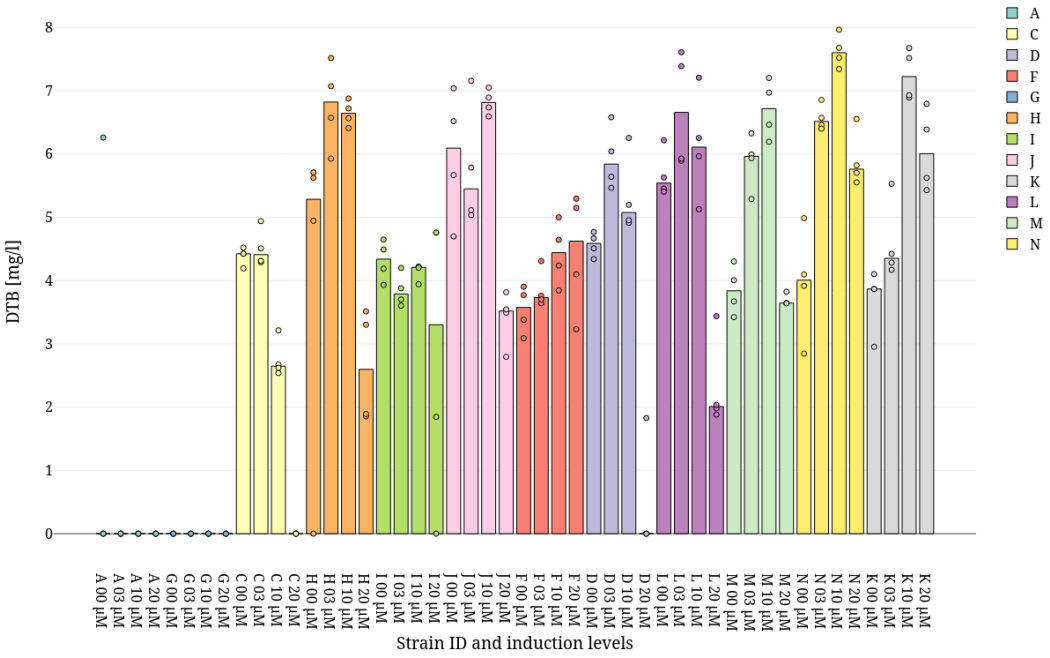


FIGURE 4.10: DTB production of strains with BS4755 background in the follow-up study

are combined with the pBS2095 DTB plasmid, the highest DTB producers are strains with plasmids containing fabI and fabZ. This observation suggests that the increased expression of the bottleneck enzymes is beneficial for DTB production. Hence, the hypothesis that the engineered strains have insufficient fatty acid biosynthesis was correct and in the future it is recommended to tune the expression of fab enzymes alongside the enzymes present in the DTB biosynthetic pathway.

Expression of fabA shows no change in the DTB production profile when combined with pBS2095, but shows an increase in production when combined with the pBS2242 DTB production plasmid. Interestingly, when fadR was expressed from a plasmid as in the previous cycle, the strain's production profile was showing increased production compared to both strain D (low bioF DTB pathway) and strain F (pBS2095 and fadR on a plasmid).

When investigating the strains where fadR is overexpressed from the chromosome (Figure 4.11), we observed some interesting production profiles. Without fadR expressed in excess from the chromosome, all investigated strains with a functional DTB production system showed DTB production. This changed with the fadR integrated strains as two of these strains showed no DTB production. Strain X and Strain V both harbored the low

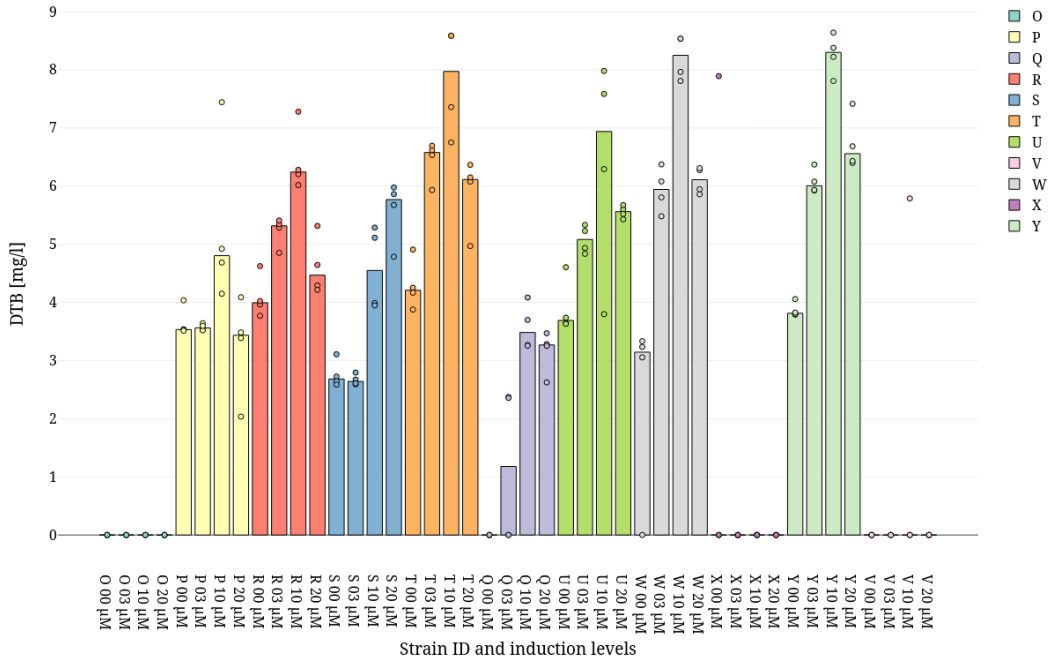


FIGURE 4.11: DTB production of strains with BS5973 background in the follow-up study (n=4)

bioF DTB plasmid and an auxiliary plasmid, fabA for strain X and fadR for strain V.

In the fadR expressing background strain, we can observe the same trend as in our original background strain. The addition of low bioF DTB production plasmid resulted in higher production of DTB. When the new background is combined with the DTB production plasmid and the fadR auxiliary plasmid (Strains Q and V), we observe a drop in production compared to their non-fadR background pairs (Strain F and K).

Regarding the fab auxiliary plasmids, in the case of pBS2095, we can observe the most improvement in DTB production when fabZ is coded on the auxiliary plasmid (Strains T and Y). The effect of fadR on a plasmid is larger than increasing its expression levels from the chromosome. (Strains C, F, P).

Metabolic profile

Regarding metabolic profiles, similar trends can be observed between strains carrying a fab gene as an auxiliary helper. This observation suggests that all of these three suspected enzymes have a beneficial effect when expressed in

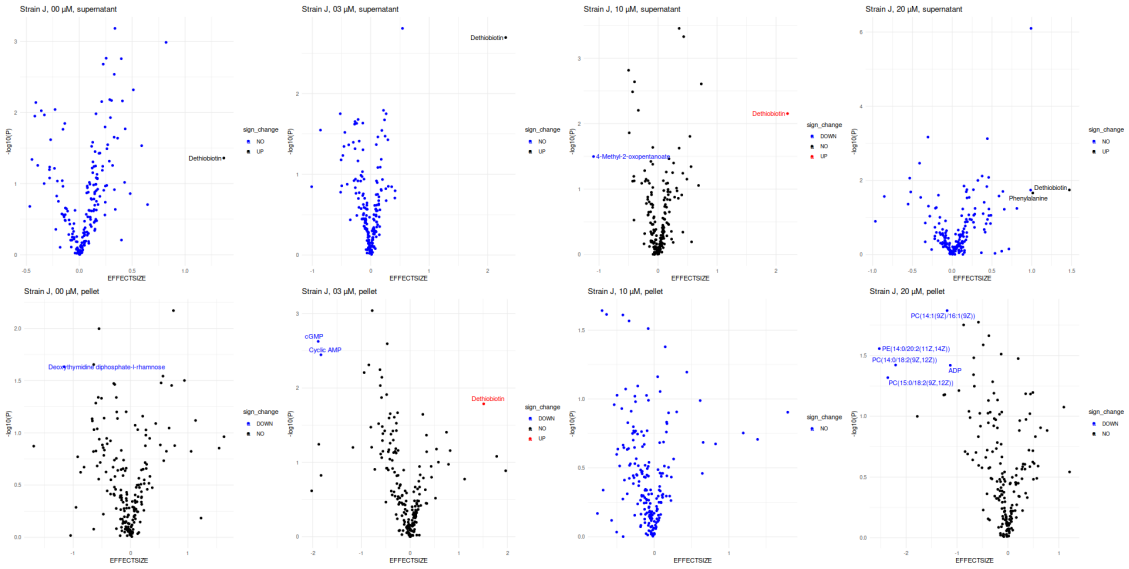


FIGURE 4.12: Volcano plot of strain J vs A (n=4)

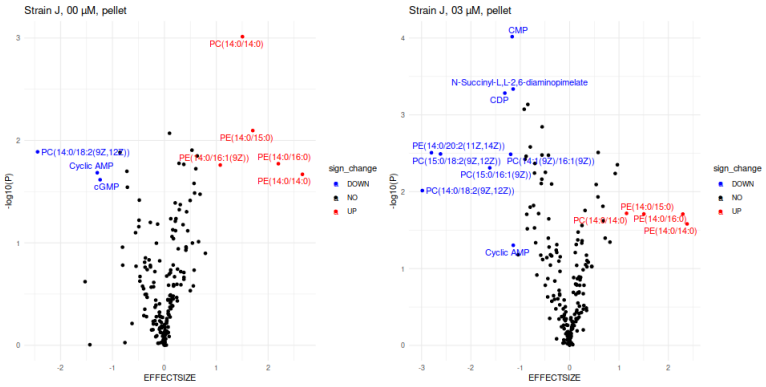


FIGURE 4.13: Volcano plot of strain J vs C (n=4)

excess. As we can deduct for the next cycle, these genes need to be combined and their expression tuned together with the DTB biosynthesis pathway.

Strain J (Figure 4.12 and 4.13), which has the original background strain, carries the pBS2095 "standard" DTB plasmid and has *FabZ* as auxiliary gene shows large positive fold changes when compared to the non-prducing control strain. The lack of lipids compared to the control strains shows up at 20µM, which is similar to strain C, which is the basis of the improvement. This observation suggests that the expression of fabZ is beneficial for DTB production but does not solve the lipid availability alone. When this promising production strain is compared to the strain to be improved (C), we can see differences in lipid abundance already at 0 and 3 µM induction levels. This clearly shows that the overexpression of fabZ does change the lipid composition of the cell. These early, more abundant lipids can be the reason for the

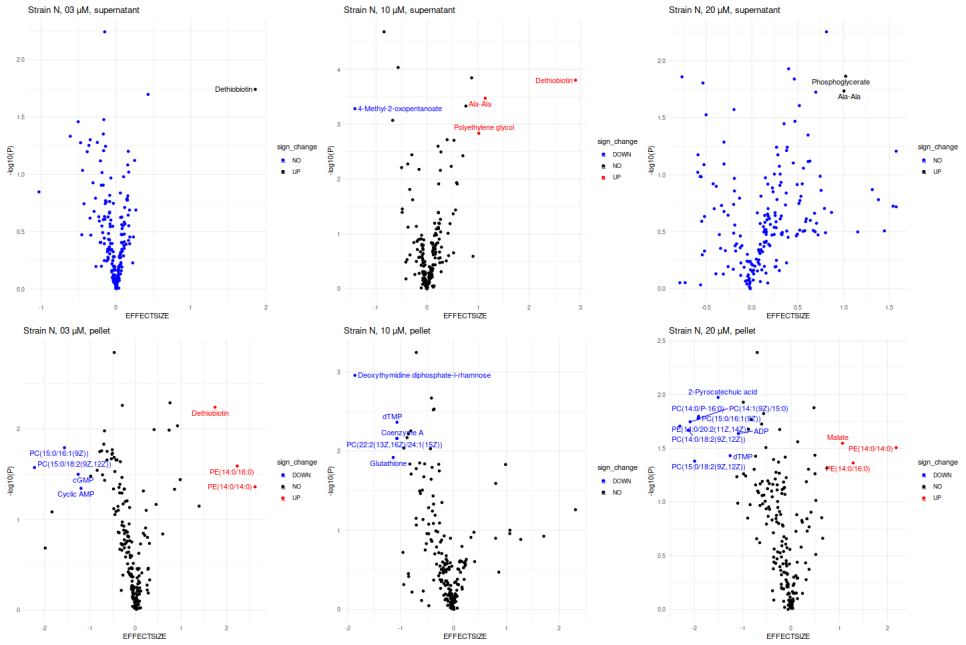


FIGURE 4.14: Volcano plot of strain N vs A (n=4)

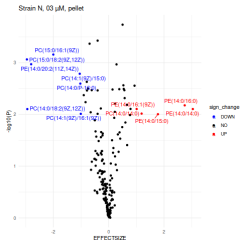


FIGURE 4.15: Volcano plot of strain N vs C (n=4)

overall higher production of DTB, which suggests that this engineering route has more potential.

Strain N (Figure 4.14 and 4.15), which is similar to strain J, but consists of the more efficient low bioF RBS DTB plasmid from the previous round. When compared to the control strain A, it still presents less abundant lipids already at 3µM. This strain was the highest producer amongst the strains without increased fadR expression from the chromosome. Similar differences in lipid abundance show up at 3µM when compared to strain C. This suggests that the combination of low bioF RBS and overexpression of fabZ does accumulate beneficial effects on the DTB production without largely affecting its metabolome.

Strain T (Figure 4.16 and 4.17), which is one of the 3 highest producing strains in this round of the cycle, harbors the identical plasmids as strain J, but the background strain does carry the knocked-in promoter for fadR.

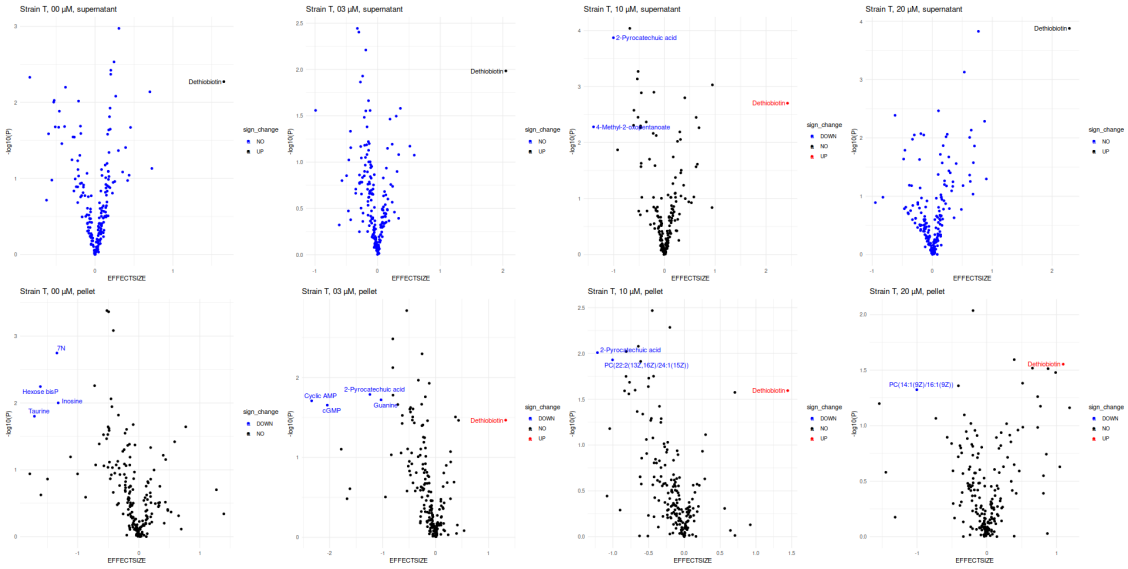


FIGURE 4.16: Volcano plot of strain T vs A (n=4)

The strain does not show significant differences in lipids when compared to strain A. While its production levels are comparable with strain J, based on the metabolomics measurements, the strain seems to deal with the stress better than J, as there is a more minor difference in lipid abundance when compared to the non-producing control strain A. When compared to the to-be-improved strain C, it produces significantly more DTB in most IPTG concentrations, and its metabolomic profile is closer to strain C than for strain J.

Strain Y (Figure 4.18 and 4.19), which was also one of the top producers, carries the new backbone and harbors the low bioF plasmid alongside the fabZ auxiliary plasmid. Its metabolic profile shows that compared to the non-producing control strain A, the lipid imbalance first shows up on the 20µM induced pellet samples. This observation shows that up until that induction level, the fatty acid metabolism did not show significant differences from the native levels. From this observation, we can say that the addition of fadR and fabZ was beneficial not only for the DTB production yield but also for the fitness of the cell factory. When we compare the strain to the production strain to be improved, we can see differences between the lipid metabolism already at early induction levels.

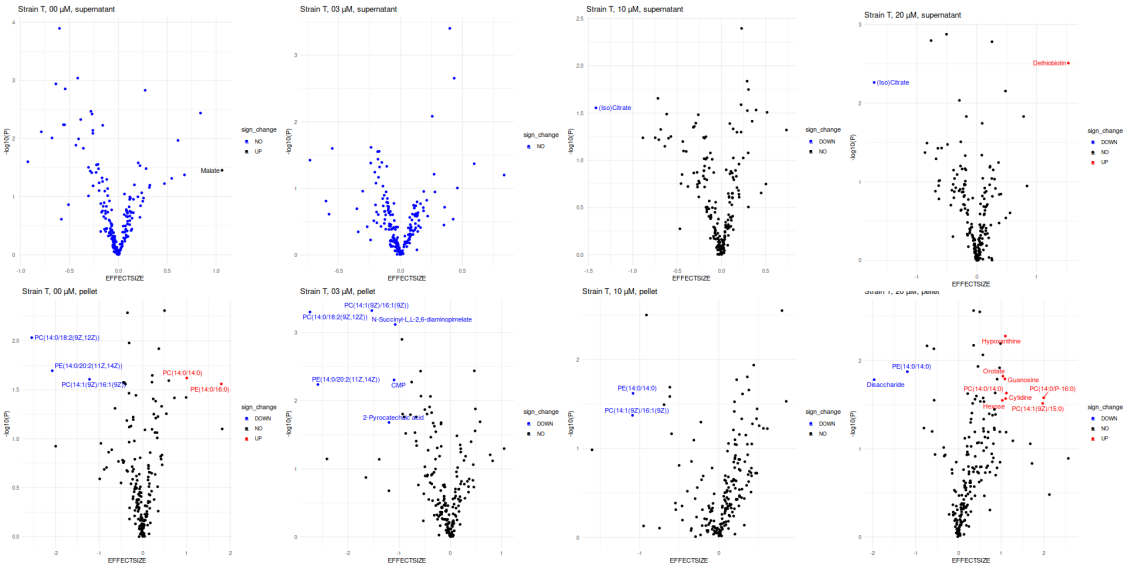


FIGURE 4.17: Volcano plot of strain T vs C (n=4)

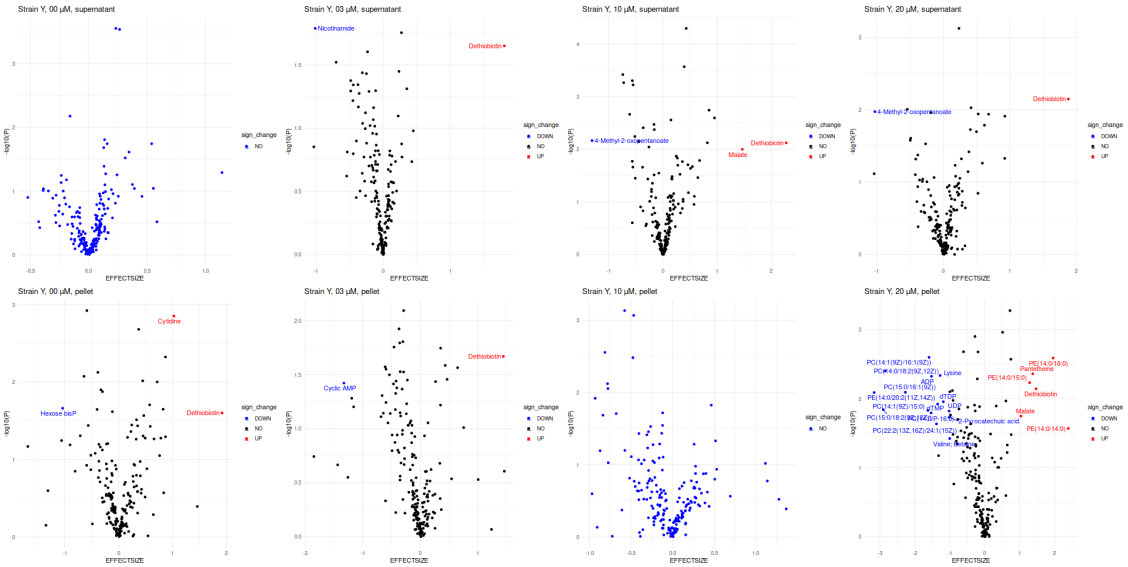


FIGURE 4.18: Volcano plot of strain Y vs A (top), and vs C (bottom) (n=4)

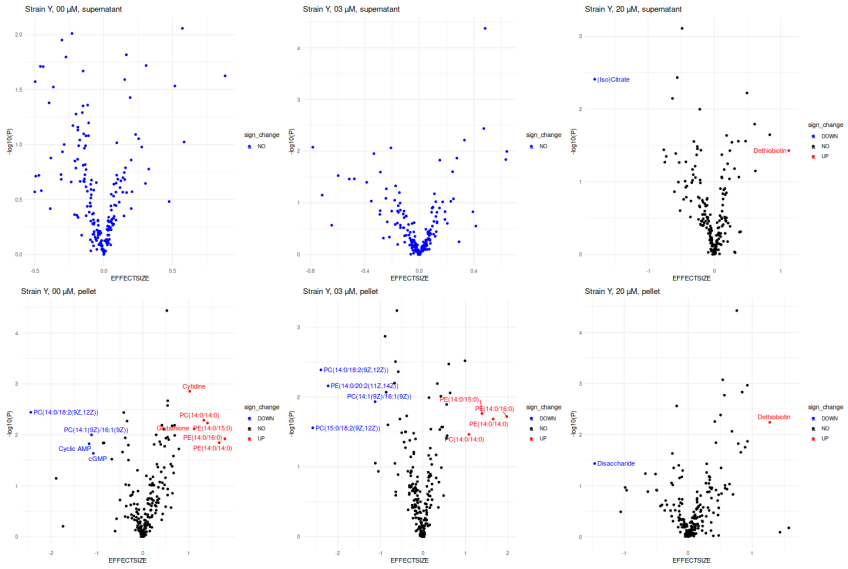


FIGURE 4.19: Volcano plot of strain Y vs A (top), and vs C (bottom) (n=4)

4.4.4 Learn step

We can conclude from this round of experiments that the increased expression of fadR from the chromosome has similar effects on the DTB production and metabolic profile of the strains as the plasmid-based expression. This suggests that including the fadR gene in the overexpression pool of a DTB producing cell factory is of interest, and one should continue the optimization of the strain with this modification. Fine-tuning of the fadR transcription factor is still necessary alongside of the DTB production pathway and other auxiliary genes.

Regarding the expression of various fab genes that were suspected bottlenecks in the previous round of the DBTL cycle, their individual expression was beneficial for DTB production yield as well as for the fitness of the cell factory.Expression of the borrowed fatty acid genes did show to be beneficial in order to maintain a healthy lipid balance in the cell. The presence of by-product metabolites was less prominent in the samples taken in the second round of the DBTL cycle. Metabolites were detected with the method, but as seen on the volcano plots of the second round of the samples, there were no significant fold changes between samples. Even in the case of strains present in both rounds, the by-product metabolites only show up on the first round of experiments is worrisome regarding the reproducibility of the data. There were differences between the two rounds, which could explain the difference

in metabolic profiles. The culture volume was significantly larger during the first round (4ml) compared to the second round (400µl). While during the first round, the raw data was annotated with the fermentation samples. Annotation can then be biased towards the fermentation samples and give a relatively higher value for these peaks. At the same time, the same by-products were present in the samples in the second round, but their intensity values were lower than in the first round. In order to confidently identify and quantify these hypothesized by-products, this high throughput method is not sensitive or selective enough. The authors suggest carrying out targeted metabolomics measurements on the compounds, using a wide range of standards when available. A limiting factor in carrying out these specific measurements can be that these compounds are hypothetical, and most do not have a commercially available standard. In order to be able to verify their existence in the samples, one needs to either synthesize and add them to a specific analytical method. Alternatively, one can also try to purify the suspected compounds from the culture broth and carry out various analytical methods such as NMR to clarify their structure and identify them.

4.5 Discussion

As presented in this chapter, the use of high-throughput analytics methods with a developed DBTL cycle can generate large amount of information about the engineered cell factory. In order to carry on with more iterations of the cycle, one does not necessairly need to analyze all aspects of the gathered data. It is sufficient to store the data and most importantly the metadata, in order to be able to later return to the unanalyzed dataset at a later cycle.

We can conclude that the excess expression of the fadR transcription factor had a beneficial effect on the production of DTB, which makes this modification a good auxiliary gene candidate for an efficient biotin cell factory. In all cases, expression of the transcription factor resulted in the broader production window on the induction gradient. The increased production window proves that the initial strains were defective in their lipid metabolism, and therefore, their lipid availability needs to be balanced in the future. It is suggested that a larger range of inducer concentrations to be tested in the future, up to 100µM IPTG or more, depending on the expression level of the transcription factor

Decreased RBS of *BioF* interestingly also showed beneficial effects on the production of DTB, especially while combining it with auxiliary genes such as *FabI/A/Z*. Decrease RBF means that the expression of the protein is lowered. Therefore it is less abundant in the cell. This observation shows that the current expression of bioF is in excess; therefore it could be prone to non-specific by-product formation as it has been shown in the first round of untargeted metabolomics experiments. We can suggest for a follow-up experimental round to lower the expression of the targeted bio genes in order to re-balance the fatty acid biosynthesis genes. Furthermore, in the following step, increase their expression to match the lipids that might be available in excess.

Identified possible bottlenecks in the first round of experiments were expressed in excess in the second round to investigate the effect of their increased abundance in the cell factory. In order to be able to differentiate the effect of various bottlenecks, the created strains only harbored one of the suspected bottleneck genes. As it is shown, all of them had a beneficial effect on the production of DTB, and in some cases on the fitness of the cell factory. Therefore, for the future engineering of a biotin cell factory, it is advised to take these possible bottlenecks into consideration. Similar to the increased chromosomal expression of fadR, one can create a strain with an increased chromosomal expression of the investigated bottlenecks.

We confirmed that there are by-products present in the engineered cells. However, their identification and detailed quantification were outside of the scope of this project. As suggested in the last Learn step, their identification and quantification could be a separate project and could highlight the promiscuity of the biotin biosynthesis pathway enzymes. If promiscuity is proven and found problematic, one can turn to protein engineering to create more specific enzymes and circumvent their non-specific binding of products.

4.6 Conclusions

This chapter demonstrated the application of a Design-Build-Test-Learn (DBTL) cycle to investigate and optimize desthiobiotin (DTB) production in engineered E. coli strains. Through two iterations of the cycle, we gained

valuable insights into the metabolic bottlenecks and byproduct formation in the biotin pathway.

In the first DBTL cycle, we identified several key findings. The accumulation of byproducts such as 3-hydroxydesthiobiotin, 7,8-diaminoenoylnonanoate, and 8-amino-7-oxoenoylnonanoate suggested promiscuous activity of biotin pathway enzymes. This observation led us to investigate potential bottlenecks in the fabZ/fabA and fabI reactions of the fatty acid biosynthesis pathway. Additionally, we discovered beneficial effects of fadR overexpression on DTB production, as well as improved DTB production with decreased bioF expression. These initial results provided a foundation for further optimization in the subsequent cycle.

The second DBTL cycle built upon these findings and explored more targeted interventions. We investigated chromosomal overexpression of fadR, as well as individual overexpression of fabI, fabA, and fabZ. Furthermore, we examined combinations of optimized DTB pathways with these auxiliary genes. This round of experiments yielded several important outcomes. We confirmed that fadR overexpression, whether chromosomal or plasmid-based, consistently improves DTB production. Individual overexpression of fabI, fabA, and fabZ all showed beneficial effects on DTB production and overall cellular fitness. Notably, the combination of low bioF expression with auxiliary genes further improved DTB yields. We also observed reduced byproduct formation and improved lipid metabolism in these optimized strains.

These results highlight the complex interplay between the biotin pathway, fatty acid biosynthesis, and overall cellular metabolism. The study demonstrates the power of iterative DBTL cycles in metabolic engineering, allowing for rapid identification and addressing of bottlenecks and off-target effects. By systematically testing hypotheses and incorporating new findings into subsequent designs, we were able to achieve significant improvements in DTB production and strain performance.

Looking ahead, several avenues for future work have been identified. Fine-tuning the expression levels of fadR, biotin pathway genes, and auxiliary fatty acid biosynthesis genes will be crucial for further optimization. Exploring higher induction levels in the optimized strains may reveal the full potential of DTB production. A more detailed characterization and quantification of byproducts using targeted metabolomics approaches could provide

deeper insights into pathway flux and enzyme promiscuity. This might lead to opportunities for protein engineering of biotin pathway enzymes to reduce off-target activities. Finally, integrating the optimized pathway components into the genome could result in stable, plasmid-free production strains suitable for industrial applications.

In conclusion, this study has significantly advanced our understanding of DTB production in E. coli and laid the groundwork for developing highly efficient biotin-producing cell factories. The approach demonstrated here, combining high-throughput screening with untargeted metabolomics, can be broadly applied to other metabolic engineering projects. This methodology accelerates the development of microbial cell factories for various valuable compounds by providing a systematic framework for strain optimization. As we continue to refine these techniques and apply them to increasingly complex metabolic systems, we move closer to realizing the full potential of synthetic biology in industrial biotechnology.

5 Conclusions

5.1 Overall suggestions for metabolic engineers on how to use multi-plasmid systems

5.1.1 Importance of plasmid compatibility

As highlighted and discussed in the review of various findings from literature in Chapter 2, the utilization of plasmids is a key method in metabolic engineering. Their widespread and easy way of manipulation to fit the researchers' needs in any given experiment is ecompassing their biggest strength and weakness. In order to design a multi-plasmid system, the researcher need to take into consideration the replication machinery and therefore the compatibility of the plasmid backbones that are planned to be present in the expression system. One needs to make sure that the segregarion machineries are compatible to maintain the backbones throughout multiple generations in the strains, providing a robust and predictable expression system. Luckily there are many possibilities to choose from in terms of plasmid backbones. If one is not taking plasmid backbone compatibility into account when planning their expression system, it is still a possibility to recognize the issue downstream with the above mentioned methods. But it is important that during the evaluation of non-expected data, when the expression system can be deemed unstable, the possibility of plasmid backbone incompatibility has been taken into account.

5.1.2 Impact of copy number variability

As plasmid backbones with various replication machineries can present variation in plasmid copy numbers in strains during fermentation, it is important

that the variability is accounted for during data analysis. The level of expression from the vector can be greatly dependent on many factors, discussed in Chapter 3. Expression from a non-regulated plasmid, also considered high copy number plasmid, can introduce large variability in the expression, therefore utilization of such plasmids in multi-plasmid systems can introduce even larger differences in the final protein levels during fermentation. In this thesis, we only utilized small scale and high throughput expression, in order to be able to screen a multitude of combinations. Therefore the findings would need to be validated with larger scale fermentations too, even though most industrially viable strains only use a single or no plasmids in their production strain, especially due to these limitations. But in order to ease the later scale-up of promising strains from high-throughput strain development, data regarding larger scale experiments would be beneficial to build a model that could predict the behaviour of the designed multi-plasmid system.

In order to be able to monitor the plasmid copy numbers to enhance the reproducibility of screening experiments, there are measures in the experimental design that can be taken as precautions. These methods will increase the labor and data output that has been generated in the research and development process, but with automated tools, can be used as a quality control measure during development. Such methods include monitoring the plasmid copy numbers via either sequencing or qPCR methods. As both methods can be deemed expensive to carry out, these measures can be utilized in case of uncertain data, or established as part of a pre-study while developing the expression system to be used. Then a model can be built from the data generated, and utilized to predict the behaviour of the screening experiments. When large deviation from the model has been observed, it can be beneficial to repeat the measurements on the problematic strains. But screening every strain for plasmid copy number would not be necessary.

5.1.3 Practical recommendations

When working with multi-plasmid systems, researchers should prioritize plasmid compatibility and carefully consider copy numbers, as these factors significantly impact gene expression and host metabolic burden. It is crucial to provide comprehensive reports on plasmid backbones, compatibility, and

copy numbers in publications to ensure reproducibility and facilitate further research. Regular assessment of plasmid stability is essential, particularly for long-term experiments or industrial applications. Implementing thorough data collection practices, including information on plasmid dynamics and interactions, will support future AI-based learning and predictive modeling efforts.

To enhance the accuracy and reliability of results, researchers should employ targeted metabolomics with a wide range of standards when investigating metabolic pathways and potential by-products. This approach offers improved sensitivity and selectivity compared to high-throughput methods, especially for hypothetical compounds or those lacking commercial standards. Additionally, maintaining consistency in experimental conditions, such as culture volumes and annotation methods in metabolomics studies, is crucial for ensuring reproducibility. By adhering to these recommendations, scientists can improve the design, implementation, and analysis of multi-plasmid systems in metabolic engineering projects, ultimately leading to more robust and reproducible outcomes.

5.2 Advances and innovation in biotin synthesis engineering

The development and optimization of biotin production pathways in E. coli has been significantly advanced through the use of modular assembly techniques. We divided the biotin biosynthesis pathway into two main modules: the first comprising genes *BioG*, *BioC*, *BioF*, *BioA*, and *BioD* for producing desthiobiotin (DTB), and the second focusing on the final step catalyzed by bioB. This modular approach allowed for parallel optimization of different pathway segments, enabling more efficient testing and fine-tuning of various configurations.

Significant enhancements in biotin yield were achieved through targeted metabolic engineering interventions. One key strategy involved increasing the chromosomal expression of fadR, a transcriptional regulator that indirectly boosts biotin production by enhancing the availability of malonyl-CoA, a critical precursor in the biotin pathway. While previous studies have focused on bioB as a rate-limiting step, future work could profitably explore

engineering the substrate specificities of BioG and BioH (Acevedo-Rocha et al., 2019), which are key enzymes in the upstream segment of the pathway. By addressing these bottlenecks and integrating such refinements with a modular assembly approach, it may be possible to develop more efficient biotin cell factories with enhanced production capabilities.

5.2.1 Key challenges in scale-up

Several key hurdles were identified during the scale-up process of biotin production. One significant challenge was enzyme inefficiency, particularly with the bioB enzyme, which is considered slow and a major limiting factor in the biotin biosynthesis pathway. This enzyme inefficiency creates a bottleneck that restricts the overall flux through the pathway, limiting biotin production rates. Another hurdle encountered was related to substrate availability, which can become a limiting factor as production scales up.

To overcome these challenges, several strategies were proposed. One approach is to focus on further genetic modifications to improve enzyme efficiency and stability. This could involve protein engineering to create more specific enzymes, particularly if enzyme promiscuity is found to be problematic. The increased chromosomal expression of key genes, similar to the successful approach with *FadR*, could be applied to other identified bottlenecks in the pathway.

Another strategy involves exploring the use of alternative substrates. While specific alternative substrates were not explored as part of this thesis, the approach could help address issues of substrate availability or cost-effectiveness in large-scale production.

The identification and quantification of by-products present in the engineered cells was suggested as a separate project that could provide valuable insights. This could highlight the promiscuity of biotin biosynthesis pathway enzymes and inform further optimization efforts.

Lastly, it was advised to consider these potential bottlenecks in future engineering of biotin cell factories. By anticipating and addressing these challenges early in the development process, researchers can design more efficient and scalable biotin production systems. This may involve creating

strains with increased chromosomal expression of key enzymes or implementing other genetic modifications to enhance pathway efficiency and robustness during scale-up.

5.2.2 Industrial viability

The findings from the biotin pathway research have significant potential for improving commercial biotin production. The modular approach used in developing the biotin production pathway in E. coli can be directly applied to industrial settings. This method allows for more efficient optimization of different pathway segments, which can lead to higher yields and more cost-effective production processes.

The specific metabolic engineering interventions, such as increased chromosomal expression of fadR and optimization of the bioB enzyme, can be implemented in industrial strains to enhance biotin yields. These improvements address key bottlenecks in the production pathway and can significantly increase the efficiency of commercial biotin production. The findings also highlight the importance of considering enzyme inefficiencies and substrate availability during scale-up. Industrial partners can use this knowledge to anticipate and address potential challenges in large-scale fermentation processes, leading to more robust and reliable production systems.

To encourage collaborations between academic research and industrial partners, it's crucial to emphasize the mutual benefits of such partnerships. Academic researchers can provide cutting-edge knowledge and innovative approaches, while industrial partners can offer practical insights into large-scale production challenges and commercialization processes. Collaborations could focus on further optimizing the engineered strains for industrial-scale fermentation, addressing challenges specific to commercial production environments. Joint research projects could also explore the use of alternative substrates or develop new genetic modifications to enhance pathway efficiency and robustness. By bridging the gap between laboratory research and industrial application, these collaborations can accelerate the development and commercialization of improved biotin production methods. This could lead to more sustainable and cost-effective production of biotin, benefiting both the industry and consumers.

Encouraging open communication and knowledge sharing between academia and industry, while respecting intellectual property rights, can foster an environment of innovation and rapid advancement in biotin production technology. This collaborative approach can help translate scientific advancements into real-world applications, ultimately contributing to more efficient and sustainable production of this essential vitamin for various industries.

5.3 Future directions

5.3.1 Enhancements in designing multi-plasmid systems

Future research should focus on developing more robust and versatile plasmid systems that can maintain stability and functionality across different host organisms. This involves a deeper investigation into plasmid compatibility and copy number variability. This could involve engineering new synthetic origins of replication or modifying existing ones to enhance their performance in multi-plasmid systems. Another important aspect to investigate is the impact of selection systems on plasmid stability and host organism viability. While antibiotic resistance markers are commonly used, they may not be suitable for all applications.

Studies should also focus on understanding and mitigating the metabolic burden imposed by multi-plasmid systems on host organisms. This could involve developing plasmids with more efficient gene expression systems or exploring ways to minimize unnecessary gene expression. Furthermore, research into plasmid segregation systems could lead to the development of plasmids that are more stably maintained across generations in various host organisms. This is particularly important for industrial applications where long-term stability is crucial.

5.3.2 Biotin production

We should aim to fully understand and control the regulatory mechanisms governing the biotin synthesis pathway. This comprehensive approach

would allow for more precise manipulation of the pathway to maximize biotin production rates.

One key area of focus should be the complete elucidation of the regulatory elements controlling biotin biosynthesis. A deeper understanding of these components and their responses to external signals could enable more effective strategies for upregulating beneficial genes while suppressing those that limit production. For example, future experiments could employ targeted mutagenesis of the BirA transcription factor, building on previous work showing that specific BirA mutations lead to increased biotin production (Chakravartty and Cronan, 2011), to directly assess its regulatory impact. Additionally, studies could be designed to examine the effects of other global regulators, such as fadR, on precursor availability and overall cellular metabolism, thereby identifying further targets for optimizing biotin production.

Another crucial area for investigation is the feedback mechanisms within the biotin pathway. Understanding how intermediates and end-products of the pathway regulate enzyme activity and gene expression could provide insights into breaking feedback inhibition loops and maintaining high production rates. Future studies should also focus on the bioB enzyme, which has been identified as a major bottleneck in the pathway. Research into the regulatory mechanisms controlling bioB expression and activity, as well as protein engineering efforts to improve its efficiency, could significantly enhance overall biotin production rates. The identification and quantification of by-products in the engineered cells, as suggested in the thesis, could provide valuable insights into the promiscuity of biotin biosynthesis pathway enzymes. This information could be used to design more specific enzymes or to develop strategies for redirecting metabolic flux away from undesired by-products and towards biotin production.

Lastly, research should explore the integration of advanced techniques such as metabolic flux analysis and genome-scale modeling to gain a more comprehensive understanding of how the biotin pathway interacts with overall cellular metabolism. This systems-level approach could reveal new strategies for maximizing biotin production rates by optimizing not just the pathway itself, but also its integration with the host organism's metabolism. By combining these approaches, researchers can develop more efficient and sustainable biotin production systems that meet the growing demand for

this essential vitamin.

5.4 Final remarks

Reflecting on the integral role of synthetic biology and metabolic engineering in advancing our capability to produce critical compounds efficiently and sustainably, it's clear that these fields have revolutionized our approach to creating sustainable solutions for various industries. The research presented in this thesis exemplifies how these disciplines can be leveraged to address the increasing need for sustainably produced ingredients, moving away from non-renewable solutions.

Synthetic biology and metabolic engineering have enabled the development of customized microbes that serve as efficient cell factories for producing small molecules and chemicals. This approach builds upon the ancient technology of fermentation, but with the added power of modern genetic and molecular biology knowledge. The ability to engineer microbes for specific purposes has opened up new possibilities for sustainable production of a wide range of compounds, including biotin.

The research presented in this thesis, focusing on the optimization of biotin production pathways, demonstrates the potential of these fields to contribute innovative solutions to global challenges. The ability to produce biotin and other essential compounds more efficiently and sustainably can lead to more accessible and affordable supplements. This is particularly important in addressing global health challenges and ensuring equitable access to vital nutrients.

For industry, this research showcases how metabolic engineering can be used to create more efficient and sustainable production processes. The modular approach to pathway engineering, combined with planned data analysis and optimization techniques, provides a blueprint for developing production systems for a wide range of valuable compounds. This has the potential to transform industrial production, making it more environmentally friendly and economically viable. Moreover, the emphasis on collaboration between academic research and industrial partners highlighted in this work is crucial

for translating scientific advancements into real-world applications. This collaborative approach can accelerate the development and implementation of innovative solutions across various sectors.

In conclusion, this research contributes to the broader goal of creating sustainable, bio-based solutions to global challenges. By advancing our understanding of metabolic pathways and developing tools for their optimization, it paves the way for more efficient production of critical compounds. This not only addresses immediate needs in various industries but also contributes to the long-term goal of transitioning to more sustainable and renewable production methods, aligning with global efforts to address climate change and resource scarcity.

Bibliography

Acevedo-Rocha, Carlos G. et al. (2019). "Microbial cell factories for the sustainable manufacturing of B vitamins". In: *Current Opinion in Biotechnology* 56, pp. 18–29. ISSN: 18790429. DOI: 10.1016/J.COPBIO.2018.07.006. URL: https://doi.org/10.1016/j.copbio.2018.07.006.

Addgene (2024). *Addgene*. URL: https://www.addgene.org/ (visited on 02/11/2024).

Afgan, Enis et al. (May 2018). "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update". In: *Nucleic Acids Research* 46.W1, W537–W544. ISSN: 0305-1048. DOI: 10.1093/nar/gky379. eprint: https://academic.oup.com/nar/article-pdf/46/W1/W537/25110642/gky379.pdf. URL: https://doi.org/10.1093/nar/gky379.

Afonso, Margarida S. et al. (2015). "Resveratrol production in bioreactor: Assessment of cell physiological states and plasmid segregational stability". In: *Biotechnology Reports* 5.1, pp. 7–13. ISSN: 2215017X. DOI: 10.1016/j.btre.2014.10.008. URL: http://dx.doi.org/10.1016/j.btre.2014.10.008.

Alnoch, Robson Carlos et al. (2018). "Co-expression, purification and characterization of the lipase and foldase of Burkholderia contaminans LTEB11". In: *International Journal of Biological Macromolecules* 116, pp. 1222–1231. ISSN: 18790003. DOI: 10.1016/j.ijbiomac.2018.05.086. URL: https://doi.org/10.1016/j.ijbiomac.2018.05.086.

Alonso-Gutierrez, Jorge et al. (2013). "Metabolic engineering of Escherichia coli for limonene and perillyl alcohol production". In: *Metabolic Engineering* 19, pp. 33–41. ISSN: 10967176. DOI: 10.1016/j.ymben.2013.05.004. URL: http://dx.doi.org/10.1016/j.ymben.2013.05.004https://www.sciencedirect.com/science/article/pii/S109671761300058X?via%3Dihub.

Bali, Anne P., Hans J. Genee, and Morten O.A. Sommer (2018). "Directed Evolution of Membrane Transport Using Synthetic Selections". In: *ACS Synthetic Biology* 7.3, pp. 789–793. ISSN: 21615063. DOI: 10.1021/ACSSYNBIO.7B00407/SUPPL_FILE/SB7B00407_SI_001.PDF. URL: https://pubs.acs.org/doi/abs/10.1021/acssynbio.7b00407.

Bali, Anne P. et al. (July 2020). "Improved biotin, thiamine, and lipoic acid biosynthesis by engineering the global regulator IscR". In: *Metabolic Engineering* 60, 97–109. ISSN: 1096-7176. DOI: 10.1016/j.ymben.2020.03.005. URL: http://dx.doi.org/10.1016/j.ymben.2020.03.005.

Baneyx, F (Oct. 1999). "Recombinant protein expression in Escherichia coli". In: *Current Opinion in Biotechnology* 10.5, 411–421. ISSN: 0958-1669. DOI: 10.1016/s0958-1669(99)00003-8. URL: http://dx.doi.org/10.1016/s0958-1669(99)00003-8.

Battesti, Aurélia and Emmanuelle Bouveret (2012). "The bacterial two-hybrid system based on adenylate cyclase reconstitution in Escherichia coli". In: *Methods* 58.4, pp. 325–334. ISSN: 10462023. DOI: 10.1016/j.ymeth.2012.07.018.

Berstenhorst, S. M. zu, H. P. Hohmann, and K. P. Stahmann (2009). "Vitamins and Vitamin-like Compounds: Microbial Production". In: *Encyclopedia of Microbiology*, pp. 549–561. DOI: 10.1016/B978-012373944-5.00161-9.

Betlach M et al. (1976). "A restriction endonuclease analysis of the bacterial plasmid controlling the ecoRI restriction and modification of DNA". In: *Fed Proc* 9.35.

Bikard, David et al. (June 2013). "Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system". In: *Nucleic Acids Research* 41.15, 7429–7437. ISSN: 0305-1048. DOI: 10.1093/nar/gkt520. URL: http://dx.doi.org/10.1093/nar/gkt520.

Bolivar, F. (1978). "Construction and characterization of new cloning vehicles III. Derivatives of plasmid pBR322 carrying unique Eco RI sites for selection of Eco RI generated recombinant DNA molecules". In: *Gene* 4.2. ISSN: 03781119. DOI: 10.1016/0378-1119(78)90025-2.

Bolivar, Francisco et al. (1977a). "Construction and characterization of new cloning vehicle. II. A multipurpose cloning system". In: *Gene* 2.2. ISSN: 03781119. DOI: 10.1016/0378-1119(77)90000-2.

Bolivar, Francisco et al. (1977b). "Construction and characterization of new cloning vehicles I. Ampicillin-resistant derivatives of the plasmid pMB9". In: *Gene* 2.2. ISSN: 03781119. DOI: 10.1016/0378-1119(77)90074-9.

Brewster, Robert C. et al. (2014). "The transcription factor titration effect dictates level of gene expression". In: *Cell* 156.6, pp. 1312–1323. ISSN: 10974172. DOI: 10.1016/j.cell.2014.02.022.

Buerger, Josephine (2019). "Screening and selection: Applications for vitamin biosynthesis in Escherichia coli". PhD thesis. URL: https://api.semanticscholar.org/CorpusID:91389954.

Carbonell, Pablo et al. (2018). "An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals". In: *Communications Biology 2018 1:1* 1.1, pp. 1–10. ISSN: 2399-3642. DOI: 10.1038/s42003-018-0076-9. URL: https://www.nature.com/articles/s42003-018-0076-9.

Cardinale, Stefano and Adam Paul Arkin (2012). "Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems". In: *Biotechnology Journal* 7.7, pp. 856–866. ISSN: 18606768. DOI: 10.1002/biot.201200085.

Chai, Dafei et al. (May 2020). "The optimization system for preparation of TG1 competent cells and electrotransformation". In: *MicrobiologyOpen* 9.7. ISSN: 2045-8827. DOI: 10.1002/mbo3.1043. URL: http://dx.doi.org/10.1002/mbo3.1043.

Chakravartty, V. and J. E. Cronan (Dec. 2011). "Altered Regulation of Escherichia coli Biotin Biosynthesis in BirA Superrepressor Mutant Strains". In: *Journal of Bacteriology* 194.5, 1113–1126. ISSN: 0021-9193. DOI: 10.1128/jb.06549-11. URL: http://dx.doi.org/10.1128/JB.06549-11.

Chao, Ran et al. (2017). "Engineering biological systems using automated biofoundries". In: *Metabolic Engineering* 42. ISSN: 10967176. DOI: 10.1016/j.ymben.2017.06.003.

Chen, Fen-Er et al. (Nov. 2007). "An improved synthesis of a key intermediate for (+)-biotin from d-mannose". In: *Carbohydrate Research* 342.16, 2461–2464. ISSN: 0008-6215. DOI: 10.1016/j.carres.2007.06.029. URL: http://dx.doi.org/10.1016/j.carres.2007.06.029.

Cock, Peter J. A. et al. (Mar. 2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11, pp. 1422–1423. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp163. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/11/1422/944180/btp163.pdf. URL: https://doi.org/10.1093/bioinformatics/btp163.

Cohen, S. N. et al. (1973). "Construction of Biologically Functional Bacterial Plasmids In Vitro". In: *Proceedings of the National Academy of Sciences* 70.11. ISSN: 0027-8424. DOI: 10.1073/pnas.70.11.3240.

Corchero, José L. and Antonio Villaverde (1998). "Plasmid maintenance in Escherichia coli recombinant cultures is dramatically, steadily, and specifically influenced by features of the encoded proteins". In: *Biotechnology and Bioengineering* 58.6, pp. 625–632. ISSN: 00063592. DOI: 10.1002/(SICI)1097-0290(19980620)58:6<625::AID-BIT8>3.0.CO;2-K.

Datsenko, Kirill A. and Barry L. Wanner (May 2000). "One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products". In: *Proceedings of the National Academy of Sciences* 97.12, 6640–6645. ISSN: 1091-6490. DOI: 10.1073/pnas.120163297. URL: http://dx.doi.org/10.1073/pnas.120163297.

Datta, Naomi and Polyxeni Kontomichalou (1965). "Penicillinase Synthesis Controlled By Infectious R Factors In Enterobacteriaceae". In: *Nature* 208.5007. ISSN: 0028-0836. DOI: 10.1038/208239a0.

Dower, William J, Jeff F Miller, and Charles W Ragsdale (1988). "High efficiency transformation of E.coli by high voltage electroporation". In: *Nucleic Acids Research* 16.13, pp. 6127–6145. ISSN: 03051048. DOI: 10.1093/nar/16.13.6127. URL: https://academic.oup.com/nar/article-abstract/16/13/6127/1046461.

Dubois, Pavlina et al. (2019). "Author Correction: Reprogramming Escherichia coli for the production of prenylated indole diketopiperazine alkaloids (Scientific Reports, (2019), 9, 1, (9208), 10.1038/s41598-019-45519-y)". In: *Scientific Reports* 9.1, pp. 1–13. ISSN: 20452322. DOI: 10.1038/s41598-019-51404-5.

Eggersdorfer, Manfred et al. (2012). "One Hundred Years of Vitamins—A Success Story of the Natural Sciences". In: *Angewandte Chemie International Edition* 51.52, pp. 12960–12990. ISSN: 1521-3773. DOI: 10.1002/ANIE.201205886. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201205886https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201205886https://onlinelibrary.wiley.com/doi/10.1002/anie.201205886.

Fiedler, Stefan and Reinhard Wirth (1988). "Transformation of bacteria with plasmid DNA by electroporation". In: *Analytical Biochemistry* 170.1, pp. 38–44. ISSN: 10960309. DOI: 10.1016/0003-2697(88)90086-3.

Furuno, Sigekazu et al. (2005). "Negative control of plasmid pSC101 replication by increased concentrations of both initiator protein and iterons." In: *The Journal of General and Applied Microbiology* 46.1, pp. 29–37. ISSN: 0022-1260. DOI: 10.2323/jgam.46.29.

Genee, Hans J. et al. (2016). "Functional mining of transporters using synthetic selections". In: *Nature Chemical Biology 2016 12:12* 12.12, pp. 1015–1022. ISSN: 1552-4469. DOI: 10.1038/nchembio.2189. URL: https://www.nature.com/articles/nchembio.2189.

Gentleman, Robert C et al. (2004). "Open Access Bioconductor: open software development for computational biology and bioinformatics". In: *Genome Biology* 5.2. URL: http://genomebiology.com/2004/5/10/http://genomebiology.com/2004/5/10/R80.

George, Kevin W. et al. (2014). "Correlation analysis of targeted proteins and metabolites to assess and engineer microbial isopentenol production". In: *Biotechnology and Bioengineering* 111.8, pp. 1648–1658. ISSN: 00063592. DOI: 10.1002/bit.25226. URL: http://doi.wiley.com/10.1002/bit.25226.

Guido, Nicholas J. et al. (2006). "A bottom-up approach to gene regulation". In: *Nature* 439.7078. ISSN: 0028-0836. DOI: 10.1038/nature04473.

Han, Qi and Mark A. Eiteman (2017). "Coupling xylitol dehydrogenase with NADH oxidase improves L-xylulose production in Escherichia coli culture". In: *Enzyme and Microbial Technology* 106.July, pp. 106–113. ISSN: 18790909. DOI: 10.1016/j.enzmictec.2017.07.010. URL: http://dx.doi.org/10.1016/j.enzmictec.2017.07.010.

Hanahan, Douglas and Cold Spring Harbor (1983). "Studies on Transformation of Escherichia coli with Plasmids". In: *J. Mol. Biol* 0, pp. 557–580. ISSN: 0361-5995. DOI: 10.2136/sssaj1998.03615995006200050032x. arXiv: arXiv:1011.1669v3.

Heath, Richard J. and Charles O. Rock (Nov. 1995). "Enoyl-Acyl Carrier Protein Reductase (fabI) Plays a Determinant Role in Completing Cycles of Fatty Acid Elongation in Escherichia coli". In: *Journal of Biological Chemistry* 270.44. ISSN: 0021-9258. DOI: 10.1074/jbc.270.44.26538. URL: http://dx.doi.org/10.1074/jbc.270.44.26538.

— (Jan. 1996a). "Regulation of Fatty Acid Elongation and Initiation by Acyl-Acyl Carrier Protein in Escherichia coli". In: *Journal of Biological Chemistry* 271.4. ISSN: 0021-9258. DOI: 10.1074/jbc.271.4.1833. URL: http://dx.doi.org/10.1074/jbc.271.4.1833.

Heath, Richard J. and Charles O. Rock (Nov. 1996b). "Roles of the FabA and FabZ beta-Hydroxyacyl-Acyl Carrier Protein Dehydratases in Escherichia coli Fatty Acid Biosynthesis". In: *Journal of Biological Chemistry* 271.44. ISSN: 0021-9258. DOI: 10.1074/jbc.271.44.27795. URL: http://dx.doi.org/10.1074/jbc.271.44.27795.

Hohmann, Hans-Peter et al. (2016). "Host Organisms: Bacillus subtilis". In: *Industrial Biotechnology*, pp. 221–297. DOI: 10.1002/9783527807796.CH7. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/9783527807796.ch7https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527807796.ch7https://onlinelibrary.wiley.com/doi/10.1002/9783527807796.ch7.

Hsia, Justin et al. (2016). "Modular synthetic inverters from zinc finger proteins and small RNAs". In: *PLoS ONE* 11.2, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0149483.

J. D. Watson and F. H. C. Crick (1953). "Molecular structure of nucleic acids". In: *Nature* 171.4356, pp. 737–738. DOI: https://doi.org/10.1038/171737a0.

Jahn, Michael et al. (2014). "Accurate Determination of Plasmid Copy Number of Flow-Sorted Cells using Droplet Digital PCR". In: *Analytical Chemistry* 86.12. PMID: 24842041, pp. 5969–5976. DOI: 10.1021/ac501118v. eprint: https://doi.org/10.1021/ac501118v. URL: https://doi.org/10.1021/ac501118v.

Jahn, Michael et al. (Dec. 2016). "Copy number variability of expression plasmids determined by cell sorting and Droplet Digital PCR". In: *Microbial Cell Factories* 15.1. ISSN: 1475-2859. DOI: 10.1186/s12934-016-0610-8. URL: http://dx.doi.org/10.1186/s12934-016-0610-8.

Jeschek, Markus, Daniel Gerngross, and Sven Panke (2017). *Combinatorial pathway optimization for streamlined metabolic engineering*. DOI: 10.1016/j.copbio.2017.06.014.

Jullesson, David et al. (Nov. 2015). "Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals". In: *Biotechnology Advances* 33.7, 1395–1402. ISSN: 0734-9750. DOI: 10.1016/j.biotechadv.2015.02.011. URL: http://dx.doi.org/10.1016/j.biotechadv.2015.02.011.

Jurgenson, Christopher T., Tadhg P. Begley, and Steven E. Ealick (2009). "The Structural and Biochemical Foundations of Thiamin Biosynthesis". In: *http://dx.doi.org/10.1146/annurev.biochem.78.072407.102340* 78, pp. 569–

603. ISSN: 00664154. DOI: 10.1146/ANNUREV.BIOCHEM.78.072407.102340. URL: https://www.annualreviews.org/doi/abs/10.1146/annurev.biochem.78.072407.102340.

Kang, Chae Won et al. (July 2018). "Synthetic auxotrophs for stable and tunable maintenance of plasmid copy number". In: *Metabolic Engineering* 48, 121–128. ISSN: 1096-7176. DOI: 10.1016/j.ymben.2018.05.020. URL: http://dx.doi.org/10.1016/j.ymben.2018.05.020.

Kim, Byoungjin et al. (2014). "Metabolic engineering of Escherichia coli for the production of phenol from glucose". In: *Biotechnology Journal* 9.5, pp. 621–629. ISSN: 18607314. DOI: 10.1002/biot.201300263.

Kim, Da Sol et al. (2019). "A new prokaryotic expression vector for the expression of antimicrobial peptide abaecin using SUMO fusion tag". In: *BMC Biotechnology* 19.1, pp. 1–12. ISSN: 14726750. DOI: 10.1186/s12896-019-0506-x.

Lederberg, Joshua (1952). "Cell Genetics and Hereditary Symbiosis". In: *Physiological Reviews* 32.1952 October, pp. 403–430. DOI: 10.1152/physrev.1952.32.4.403. URL: www.physiology.org/journal/physrev.

Li, Zhenghong, Xiaonan Wang, and Haoran Zhang (2019). "Balancing the non-linear rosmarinic acid biosynthetic pathway by modular co-culture engineering". In: *Metabolic Engineering* 54, pp. 1–11. ISSN: 10967184. DOI: 10.1016/j.ymben.2019.03.002.

Lin, Steven and John E. Cronan (2011). "Closing in on complete pathways of biotin biosynthesis". In: *Molecular BioSystems* 7.6, p. 1811. ISSN: 1742-2051. DOI: 10.1039/c1mb05022b. URL: http://dx.doi.org/10.1039/C1MB05022B.

Lin, Steven, Ryan E Hanson, and John E Cronan (Aug. 2010). "Biotin synthesis begins by hijacking the fatty acid synthetic pathway". In: *Nature Chemical Biology* 6.9. ISSN: 1552-4469. DOI: 10.1038/nchembio.420. URL: http://dx.doi.org/10.1038/nchembio.420.

Lin-Chao, S. and H. Bremer (1986). "Effect of the bacterial growth rate on replication control of plasmid pBR322 in Escherichia coli". In: *Molecular and General Genetics MGG* 203.1. ISSN: 0026-8925. DOI: 10.1007/BF00330395.

Liu, Chun Li et al. (2018). "Renewable production of high density jet fuel precursor sesquiterpenes from Escherichia coli". In: *Biotechnology for Biofuels* 11.1, pp. 1–15. ISSN: 17546834. DOI: 10.1186/s13068-018-1272-z. URL: https://doi.org/10.1186/s13068-018-1272-z.

LLC, De Novo DNA (2020). *De Novo DNA*. URL: https://www.denovodna.com/software/ (visited on 12/30/2021).

Mandel, M. and A. Higa (1970). "Calcium-dependent bacteriophage DNA infection". In: *Journal of Molecular Biology* 53.1. ISSN: 00222836. DOI: 10.1016/0022-2836(70)90051-3.

Martínez-García, Esteban et al. (2015). "SEVA 2.0: an update of the Standard European Vector Architecture for de-/re-construction of bacterial functionalities". In: *Nucleic Acids Research* 43.D1. ISSN: 1362-4962. DOI: 10.1093/nar/gku1114.

McGuffie, Matthew J and Jeffrey E Barrick (May 2021). "pLannotate: engineered plasmid annotation". In: *Nucleic Acids Research* 49.W1, W516–W522. ISSN: 1362-4962. DOI: 10.1093/nar/gkab374. URL: http://dx.doi.org/10.1093/nar/gkab374.

McNerney, Monica P. and Mark P. Styczynski (2017). "Precise control of lycopene production to enable a fast-responding, minimal-equipment biosensor". In: *Metabolic Engineering* 43, pp. 46–53. ISSN: 10967184. DOI: 10.1016/j.ymben.2017.07.004. URL: http://dx.doi.org/10.1016/j.ymben.2017.07.004.

Meynell, Elinor and Naomi Datta (1967). "Mutant Drug Resistant Factors of High Transmissibility". In: *Nature* 214.5091. ISSN: 0028-0836. DOI: 10.1038/214885a0.

Moon, Tae Seok et al. (2010). "Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered E. coli". In: *Metabolic Engineering* 12.3, pp. 298–305. ISSN: 10967176. DOI: 10.1016/j.ymben.2010.01.003. URL: http://dx.doi.org/10.1016/j.ymben.2010.01.003.

Munro, Rachel A. et al. (2019). "Biosynthetic production of fully carbon-13 labeled retinal in E. coli for structural and functional studies of rhodopsins". In: *Journal of Biomolecular NMR* 73.1-2, pp. 49–58. ISSN: 15735001. DOI: 10.1007/s10858-019-00225-9.

Nora, Luísa Czamanski et al. (2019). "Recent advances in plasmid-based tools for establishing novel microbial chassis". In: *Biotechnology Advances*, p. 107433. ISSN: 07349750. DOI: 10.1016/j.biotechadv.2019.107433. URL: https://doi.org/10.1016/j.biotechadv.2019.107433https://linkinghub.elsevier.com/retrieve/pii/S0734975019301338.

Nordström, Kurt and Santanu Dasgupta (2006). "Copy-number control of the Escherichia coli chromosome: A plasmidologist's view". In: *EMBO Reports* 7.5, pp. 484–489. ISSN: 1469221X. DOI: 10.1038/sj.embor.7400681.

Novick, R P et al. (1976). "Uniform nomenclature for bacterial plasmids: a proposal". In: *Bacteriological Reviews* 40.1. ISSN: 0005-3678. DOI: 10.1128/br.40.1.168-189.1976.

Novick, Richard P (1987). *Plasmid Incompatibility*. Tech. rep. 4, pp. 381–395.

Papin, Jason A., Jennifer L. Reed, and Bernhard O. Palsson (2004). *Hierarchical thinking in network biology: The unbiased modularization of biochemical networks*. DOI: 10.1016/j.tibs.2004.10.001.

Paracchini, Valentina et al. (2017). "Molecular characterization of an unauthorized genetically modified Bacillus subtilis production strain identified in a vitamin B2 feed additive". In: *Food Chemistry* 230, pp. 681–689. ISSN: 0308-8146. DOI: 10.1016/J.FOODCHEM.2017.03.042.

Passi, Anurag et al. (Dec. 2021). "Genome-Scale Metabolic Modeling Enables In-Depth Understanding of Big Data". In: *Metabolites* 12.1, p. 14. ISSN: 2218-1989. DOI: 10.3390/metabo12010014. URL: http://dx.doi.org/10.3390/metabo12010014.

Paulsson, Johan (2004). "Summing up the noise in gene networks". In: *Nature* 427.6973, pp. 415–418. ISSN: 00280836. DOI: 10.1038/nature02257. arXiv: /www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006164&tool=pmcentrez&rendertype=abstract. [Figures, S., 2010. Supplementary information. Nature, 1(c), pp.1-7. Available at: http:].

Peterson, James and Gregory J. Phillips (2008). "New pSC101-derivative cloning vectors with elevated copy numbers". In: *Plasmid* 59.3, pp. 193–201. ISSN: 0147619X. DOI: 10.1016/j.plasmid.2008.01.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0147619X0800005X.

Revuelta, Jose L. et al. (2016). "Microbial biotechnology for the synthesis of (pro)vitamins, biopigments and antioxidants: challenges and opportunities". In: *Microbial Biotechnology* 9.5, pp. 564–567. ISSN: 1751-7915. DOI: 10.1111/1751-7915.12379. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1751-7915.12379https://onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.12379https://sfamjournals.onlinelibrary.wiley.com/doi/10.1111/1751-7915.12379.

Rodriguez, Raymond L. et al. (1976). "CONSTRUCTION AND CHARACTERIZATION OF CLONING VEHICLES". In: *Molecular Mechanisms in*

the Control of Gene Expression. Elsevier. DOI: 10.1016/B978-0-12-518550-9.50052-4.

Roosaare, Märt et al. (2018). "PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads". en. In: *PeerJ* 6.e4588, e4588.

Rosano, GermÃ¡n L. and Eduardo A. Ceccarelli (Apr. 2014). "Recombinant protein expression in Escherichia coli: advances and challenges". In: *Frontiers in Microbiology* 5. ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00172. URL: http://dx.doi.org/10.3389/fmicb.2014.00172.

Rouches, Miles V et al. (2021). "A Plasmid System with Tunable Copy Number". In: *bioRxiv*, p. 2021.07.13.451660. DOI: 10.1101/2021.07.13.451660. URL: http://biorxiv.org/content/early/2021/07/13/2021.07.13.451660.abstract.

Sauer (2024). URL: https://imsb.ethz.ch/research/sauer.html.

Schmidt, Calvin M. et al. (2012). "Stable maintenance of multiple plasmids in E. coli using a single selective marker." In: *ACS synthetic biology* 1.10, pp. 445–450. ISSN: 21615063. DOI: 10.1021/sb3000589.

Segall-Shapiro, Thomas H., Eduardo D. Sontag, and Christopher A. Voigt (2018). "Engineered promoters enable constant gene expression at any copy number in bacteria". In: *Nature Biotechnology* 36.4, pp. 352–358. ISSN: 15461696. DOI: 10.1038/nbt.4111.

Shetty, Reshma P, Drew Endy, and Thomas F Knight (2008). "Engineering BioBrick vectors from BioBrick parts". In: *Journal of Biological Engineering* 2.1. ISSN: 1754-1611. DOI: 10.1186/1754-1611-2-5.

Silva-Rocha, Rafael et al. (2013). "The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes". In: *Nucleic Acids Research* 41.D1. ISSN: 0305-1048. DOI: 10.1093/nar/gks1119.

Song, Jae Kwang (Feb. 2009). "Gene Cloning, Expression, and Characterization of a New Carboxylesterase from Serratia sp. SES-01: Comparison with Escherichia coli BioHe Enzyme". In: *Journal of Microbiology and Biotechnology* 19.2. ISSN: 1738-8872. DOI: 10.4014/jmb.0804.287. URL: http://dx.doi.org/10.4014/jmb.0804.287.

Summers, D (Jan. 1991). "The kinetics of plasmid loss". In: *Trends in Biotechnology* 9.1, 273–278. ISSN: 0167-7799. DOI: 10.1016/0167-7799(91)90089-z. URL: http://dx.doi.org/10.1016/0167-7799(91)90089-z.

Sutcliffe, J. G. (1979). "Complete Nucleotide Sequence of the Escherichia coli Plasmid pBR322". In: *Cold Spring Harbor Symposia on Quantitative Biology* 43.0. ISSN: 0091-7451. DOI: 10.1101/SQB.1979.043.01.013.

Tal, Shay and Johan Paulsson (2012). "Evaluating quantitative methods for measuring plasmid copy numbers in single cells". In: *Plasmid* 67.2, pp. 167–173. ISSN: 0147619X. DOI: 10.1016/j.plasmid.2012.01.004.

Thuronyi, Benjamin W., Thomas M. Privalsky, and Michelle C.Y. Chang (2017). "Engineered Fluorine Metabolism and Fluoropolymer Production in Living Cells". In: *Angewandte Chemie - International Edition* 56.44, pp. 13637–13640. ISSN: 15213773. DOI: 10.1002/anie.201706696.

Timmons, Joshua J. and Doug Densmore (Jan. 2020). "Repository-based plasmid design". In: *PLOS ONE* 15.1. Ed. by Ruslan Kalendar, e0223935. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0223935. URL: http://dx.doi.org/10.1371/journal.pone.0223935.

Tong, Yaojun et al. (Sept. 2021). "A versatile genetic engineering toolkit for E. coli based on CRISPR-prime editing". In: *Nature Communications* 12.1. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25541-3. URL: http://dx.doi.org/10.1038/s41467-021-25541-3.

Volk, Michael J. et al. (2023). "Metabolic Engineering: Methodologies and Applications". In: *Chemical Reviews* 123.9. PMID: 36584306, pp. 5521–5570. DOI: 10.1021/acs.chemrev.2c00403. eprint: https://doi.org/10.1021/acs.chemrev.2c00403. URL: https://doi.org/10.1021/acs.chemrev.2c00403.

Wang, Haijiao et al. (2018). "Metabolic pathway engineering for high-level production of 5-hydroxytryptophan in Escherichia coli". In: *Metabolic Engineering* 48, pp. 279–287. ISSN: 10967184. DOI: 10.1016/j.ymben.2018.06.007. URL: https://doi.org/10.1016/j.ymben.2018.06.007.

Wierzbicki, Michael et al. (2016). "Engineering an Escherichia coli platform to synthesize designer biodiesels". In: *Journal of Biotechnology* 224, pp. 27–34. ISSN: 18734863. DOI: 10.1016/j.jbiotec.2016.03.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0168165616300992http://dx.doi.org/10.1016/j.jbiotec.2016.03.001.

William Studier, F. et al. (1990). "[6] Use of T7 RNA polymerase to direct expression of cloned genes". In: *Gene Expression Technology*. Elsevier, 60–89. DOI: 10.1016/0076-6879(90)85008-c. URL: http://dx.doi.org/10.1016/0076-6879(90)85008-c.

Wishart, David S et al. (2023). "PlasMapper 3.0-a web server for generating, editing, annotating and visualizing publication quality plasmid maps". en. In: *Nucleic Acids Res.* 51.W1, W459–W467.

Wong Ng, Jérôme et al. (2010). "Plasmid copy number noise in monoclonal populations of bacteria". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 81.1. ISSN: 15393755. DOI: 10.1103/PhysRevE.81.011909.

Xiong, Fei et al. (2016). "Preparation of (+)-Biotin: Process Development and Scale-up". In: *HETEROCYCLES* 92.3, p. 544. ISSN: 0385-5414. DOI: 10.3987/com-15-13379. URL: http://dx.doi.org/10.3987/COM-15-13379.

Xu, Da et al. (2020). "Enhanced production of 5-hydroxytryptophan through the regulation of L-tryptophan biosynthetic pathway". In: *Applied Microbiology and Biotechnology* 104.6, pp. 2481–2488. ISSN: 14320614. DOI: 10.1007/s00253-020-10371-y.

Xu, Peng et al. (2013). "Modular optimization of multi-gene pathways for fatty acids production in E. coli". In: *Nature Communications* 4.1, p. 1409. ISSN: 2041-1723. DOI: 10.1038/ncomms2425. URL: www.nature.com/naturecommunicationshttp://www.nature.com/articles/ncomms2425.

Yadav, Vikramaditya G. et al. (May 2012). "The future of metabolic engineering and synthetic biology: Towards a systematic practice". In: *Metabolic Engineering* 14.3, 233–241. ISSN: 1096-7176. DOI: 10.1016/j.ymben.2012.02.001. URL: http://dx.doi.org/10.1016/j.ymben.2012.02.001.

Yan, Yu et al. (2017). "Co-expression of cyanobacterial genes for arsenic methylation and demethylation in escherichia coli offers insights into arsenic resistance". In: *Frontiers in Microbiology* 8.JAN, pp. 1–9. ISSN: 1664302X. DOI: 10.3389/fmicb.2017.00060.

Yim, Harry et al. (2011). "Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol". In: *Nature Chemical Biology* 7.7, pp. 445–452. ISSN: 15524469. DOI: 10.1038/nchembio.580. URL: http://dx.doi.org/10.1038/nchembio.580.

Yu, Huan et al. (Feb. 2019). "Establishment of BmoR-based biosensor to screen isobutanol overproducer". In: *Microbial Cell Factories* 18.1. ISSN: 1475-2859. DOI: 10.1186/s12934-019-1084-2. URL: http://dx.doi.org/10.1186/s12934-019-1084-2.

Zhang, Jingwei et al. (2017). "Development of a Transcription Factor-Based Lactam Biosensor". In: *ACS Synthetic Biology* 6.3, pp. 439–445. ISSN: 21615063. DOI: 10.1021/acssynbio.6b00136.

A Appendix

A.1 Figures

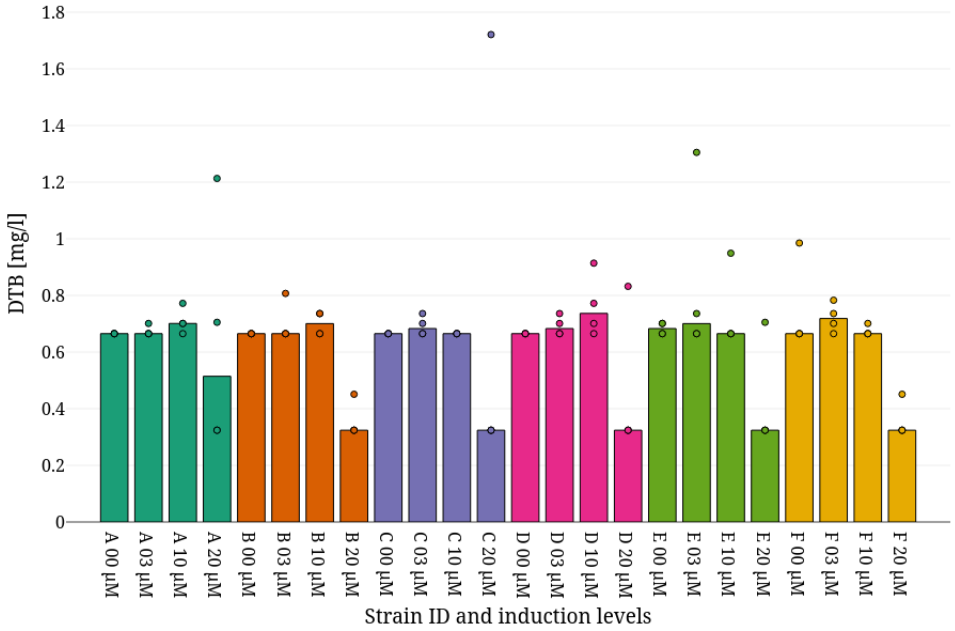


FIGURE A.1: DTB production of investigated strains at various IPTG levels in minimal media at 4h

A.2 Tables

Compound name	Formula
KAPA	C9H17NO3
DTB	C10H18N2O3
6-amino-5-oxoheptanoic acid	C7H13NO3
8-amino-3,7-dioxononanoate	C9H15NO4
3-oxodethiobiotin	C10H16N2O4

8-amino-7-oxo-3-hidoxynonanoate	C9H17NO4
3-hydoxydethiobiotin	C10H18N2O4
8-amino-7-oxoenoylnonanoate	C9H15NO3
bisnordethiobitin	C8H14N2O3
4B	C11H19NO4
4C	C11H22N2O3
4D	C12H20N2O4
5B	C11H21NO4
5C	C11H24N2O3
5D	C12H22N2O4
6B	C11H19NO3
6C	C11H22N2O2
N	C9H16N2O3
6-amino-5-oxoheptanoate-ME	C8H15NO3
DTB-ME	C11H20N2O3
5,6-diamino-heptanoate	C7H16N2O2
DAPA	C9H20N2O2
7,8-diamino-3-oxononanoate	C9H18N2O3
7,8-diamono-3-hidorxynonanoate	C9H20N2O3
7,8-diamino-enoylnonanoate	C9H18N2O2
3D	C10H16N2O3
6D	C12H20N2O3
pimeloyl-ME	C10H19NO3
7,8-diaminononanoate-ME	C10H22N2O2
7B	C11H213NO
7C	C11H24N2O2
7D	C12H22N2O3
M	C8H18N2O2
1L	C10H17NO4
1M	C10H20N2O3
1N	C11H18N2O4
2L	C10H19NO4
2M	C10H22N2O3
2N	C11H20N2O4
3L	C10H17NO3
3M	C10H20N2O2

3N	C11H18N2O3
4L	C12H21NO4
4M	C12H24N2O3
4N	C13H22N2O4
5L	C12H23NO4
5M	C12H26N2O3
5N	C13H24N2O4
6L	C12H21NO3
6M	C12H24N2O2
6N	C13H22N2O3
7L	C12H23NO3
7M	C12H26N2O2
7N	C13H24N2O3

A.3 Materials

A.3.1 minimal MOPS media

Ingredient	Mass or molarity	Volume (dissolved in water)
MOPS buffer	83.72 g	
Tricine	7.17 g	
FeSO4	0.028 g	10 mL
NH4Cl	1.9 M	50 mL
K2SO4	0.276 M	10 mL
CaCl2.2H2O	1 M	5 uL
MgCl2	2 M	2.625 uL
NaCl	5 M	100 mL
(NH4)6Mo7O24.4H2O	0.00018 g	
H3BO3	0.00124 g	
CoCl2	0.00036 g	
CuSO4	0.00012 g	
MnCl2	0.0008 g	
ZnSO4	0.00014 g	

TABLE A.2: 10 x MOPS stock solution

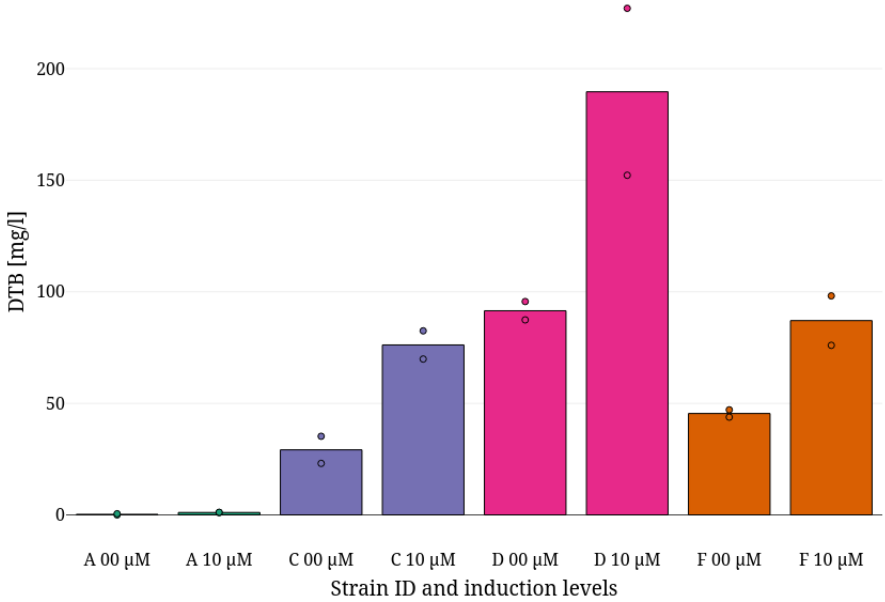


FIGURE A.2: DTB production of investigated strains at various IPTG levels in complex media at 24h

Vitamin	**Mass**	**Volume**	
calcium pantohetane	0.238 g	25 mL	dissolve in de-ionized water
p-aminobenzoic acid	0.067 g	25 mL	dissolve in 0.02 M KOH
p-hydroxybenzoic acid	0.069 g	25 mL	dissolve in 0.02 M KOH
2,3-dihydroxybenzoic acid	0.077 g	25 mL	dissolve in 0.02 M KOH

TABLE A.3: 500 x vitamin solution

Ingredient	**Volume (mL) for 1000 mL**
10 X MOPS (see table A.2)	100
0.132 M K2HPO4	10
500 x Vitamin solution (see table A.3)	2
20% D-Glucose	10
Sterile de-ionized water	up to 1000 mL

TABLE A.4: minimal MOPS medium

A.4 Protocols

A.4.1 Chemical transformation of *E. coli* cells

Materials

- Chemically competent cells
- DNA in solution
- SOC media
- LB agar plates with appropriate antibiotic

Procedure

1. Add 1-5 μL of DNA to 50 μL of competent cells in a 1.5 mL microcentrifuge tube (values can be changed according to needs).
2. Keep on ice for 20 minutes
3. Heat-shock the samples in 42°C water bath for 45 seconds.
4. Keep samples on ice for 2 minutes
5. Recover cells by adding 1 ml SOC media and incubate at 37°C for 2 hours.
6. Centrifuge samples at 1000 rpm for 5 minutes and remove supernatant.
7. Resuspend cells in 100 μL of SOC media.
8. Plate 50 μL of cells on LB agar plates with appropriate antibiotic.

Acknowledgements

I would like to express my appreciation to my suprtvisors, Dr Luisa Gronenberg, who supported my ideas and guided me through all obstacles of a research project with various hurdles, such as a global pandemic event. Thank you for your understanding, and encouragement in topics within and outside of scientific topics. I would also like to thank Dr. Hans Genee for providing an environment where I could find my passion for laboratory automation and data science. This thesis and the experimental work would not have been possible without all the incredible people I have had the chance to work with at Biosyntia during these years. And of course Prof. Dr. Zoya Ignatova, who made the whole journey possible, and providing me with the possibility of working on this project with more than generous extensions.

As part of an ITN, I am very grateful for my fellow doctoral students at SynCrop. We had a great start of the consortium with meetings all over Europe, and a prospective plan for great long-term secondment stays. Unfortunately due to the pandemic, most of those plans could not materialize, but we all found our own solutions in these troubling times to keep up the spirits and the collaboration.

And this whole journey would never have started without the continuous support and encouragement of my family and friends. Thank you Zrean for always supporting and believing in me, even if I was just mixing random liquids in the lab at odd hours. I can not express the gratitude I have for my friends, Livi and Saci for keeping up my spirits from thousands of kilometers away, your friendship is invaluable.

Lastly, I would also like to express my gratitude to my colleagues at EnginZyme. They supported me through a transition from lab to data science, while still making it possible to finish this thesis.

Declaration of Authorship

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated in the dissertation.

I, the undersigned, declare that this bound copy of the dissertation and the dissertation submitted in electronic form (via the Docata upload) and the printed bound copy of the dissertation submitted to the faculty (responsible Academic Office or the Doctoral Office Physics) for archiving are identical.

Signed: Dóra Vitay

Date: 2025.03.26

List of Hazardous Chemicals

Chemical	Hazard statement	Explanation	Pictogram
Ampicillin	H317	May cause an allergic skin reaction	
	H334	May cause allergy or asthma symptoms or breathing difficulties if inhaled	
	P261	Avoid breathing dust	
	P272	Contaminated work clothing should not be allowed out of the workplace	
	P280	Wear protective gloves	
	P284	Wear respiratory protection	
	P302 + P352	If on skin: Wash with plenty of water	
	P333 + P313	If skin irritation or rash occurs: Get medical advice/attention	

Chemical	Hazard statement	Explanation	Pictogram
Kanamycin	H360	May damage fertility or the unborn child	
	P201	Obtain special instructions before use	
	P202	Do not handle until all safety precautions have been read and understood	
	P280	Wear protective gloves/clothing/ eye protection/face protection	
	P308+ P313	If exposed or concerned: Get medical advice/attention	
	P405	Store locked up	
	P501	Dispose of contents/container to an approved waste disposal plant	

Chemical	Hazard statement	Explanation	Pictogram
Spectinomycin	H315	Causes skin irritation	
	H319	Causes serious eye irritation	
	H335	May cause respiratory irritation	
	P261	Avoid breathing dust	
	P264	Wash skin thoroughly after handling	
	P271	Use only outdoors or well ventilated-area	
	P280	Wear protective gloves/eye protection/ face protection	
	P302 + P352	If on skin: Wash with plenty of water	
	P305 + P351 + P338	If in eyes: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing	

Chemical	Hazard statement	Explanation	Pictogram
Carbenicillin	H317	May cause allergic skin reaction	
	H334	May cause allergy or asthma symptoms or breathing difficulties	
	P261	Avoid breathing dust/fume/gas/mist/ vapors/spray	
	P272	Contaminated work clothing should not be allowed out of the workplace	
	P280	Wear protective gloves	
	P284	Wear respiratory protection	
	P302 + P352	If on skin: Wash with plenty of water	
	P304 + P340 + P312	If inhaled: Remove person to fresh air and keep comfortable for breathing. Call a poison center/doctor if you feel unwell	