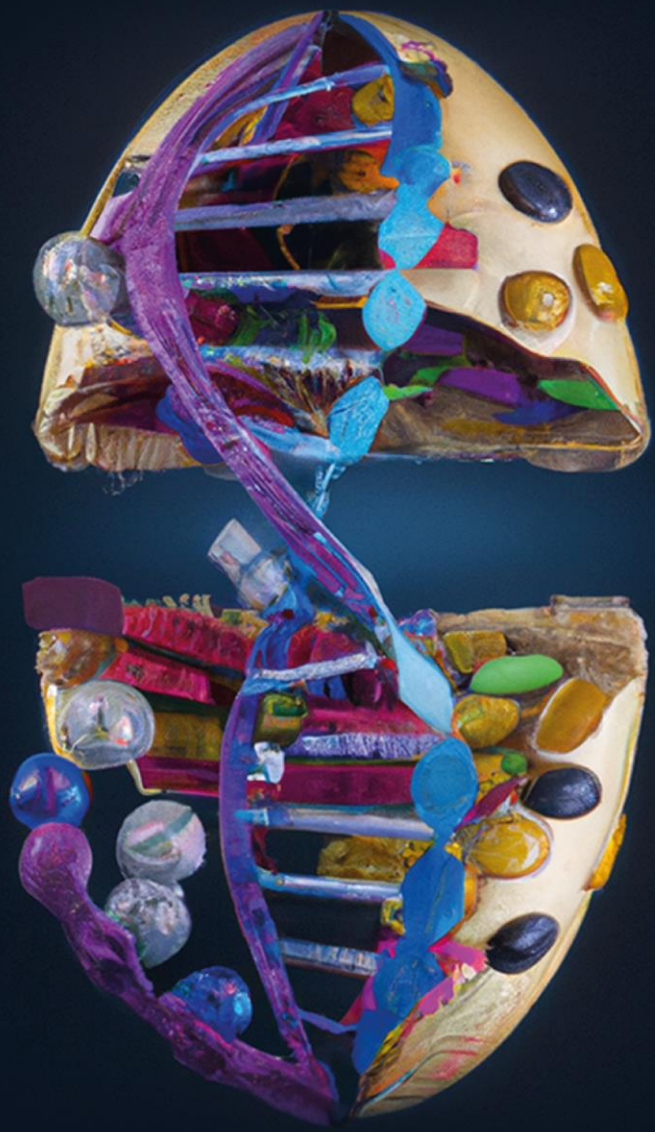
Towards a synthetic cell: designing, testing and optimizing functional genomic modules in *E. coli*



Joep Houkes

Propositions

1. A synthetic cell cannot be made without the help of evolution.
   (this thesis)

2. Essentiality of a gene is not an innate property but rather a contextual one.
   (this thesis)

3. Governments that do not allow genetically modified crops contribute to starvation.

4. Pharmaceutical companies charge excessive prices for (life-saving) medicines.

5. At institutions with high tuition fees, one pays for the brand, and not for the education quality.

6. It is undesirable that propositions about the King and Queen are forbidden.

Propositions belonging to the thesis, entitled

Towards a synthetic cell: designing, testing and optimizing functional genomic modules in *E. coli*

Joep Houkes

Wageningen, 11 January 2023

Towards a synthetic cell: designing, testing and optimizing functional genomic modules in *E. coli*

Joep Houkes

Thesis committee

Promotor

Prof. Dr John van der Oost

Personal chair at the Laboratory of Microbiology

Wageningen University & Research

Co-promotor

Dr Nico J.P.H Claassens

Assistant Professor, Laboratory of Microbiology

Wageningen University & Research

Other members

Prof. Dr Vitor A.P. Martins dos Santos, Wageningen University & Research

Prof. Dr Pascale A.S. Daran-Lapujade, Delft University of Technology

Prof. Dr Bert Poolman, Groningen University

Dr Lennart Schada von Borzyskowski, Leiden University

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

Towards a synthetic cell: designing, testing and optimizing functional genomic modules in *E. coli*

Joep Houkes

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Wednesday 11 January 2023

at 4 p.m. in the Omnia Auditorium.

Joep Houkes
Towards a synthetic cell: designing, testing and optimizing functional genomic modules in *E. coli*
174 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN: 978-94-6447-510-4
DOI: 10.18174/582244

Table of contents

Chapter 1
Towards a synthetic cell

Building a synthetic cell

Probably the biggest challenge in synthetic biology is to construct a functional synthetic cell (1–4). There are two very different approaches by which such a synthetic cell can be built: top-down and bottom-up (5–8). The top-down approach is characterized by taking a natural organism and deleting all non-essential genes. Thi s leaves one with a minimal functional cell. Several examples of these minimal synthetic cells have been described in recent years, although one can argue whether deleting non-essential genes actually creates a "synthetic" cell.

The top-down approach to create a minimal cell starts with finding out which genes are essential and can be knocked out (9). A well-studied approach to find essential genes is transposon mutagenesis, where transposons are randomly inserted into the genome (10, 11). Often these transposons contain a selection marker, to obtain only strains with integrated transposons. Transposons can only integrate into non-essential parts of the genome as they still allow the cell to survive. Of course, essentiality of a gene is dependent on the environment, in certain conditions it might be possible to delete a gene, which becomes essential in other conditions.

Gene essentiality

It used to be very time intensive to find integration sites of transposons, which had to be done by extensive genetic foot printing techniques by PCR-based sequence analyses or by using a microarray (10, 12–14). Nowadays, deep sequencing of a population of cells that have undergone transposon mutagenesis can quickly identify essential and non-essential genes (15–19). For *Escherichia coli* the Keio collection was made, a library of thousands of transposon mutagenesis mutants, each of which contains a knockout mutation of a non-essential gene in *E. coli* (20). As said before, essentiality of a gene is very environmentally dependent (21, 22). Studies have been done to try and find a minimum number of core genes needed for life (23, 24). The KEIO collection, and many other studies for gene essentiality test essentiality in rich media (20).

E. coli K-12, one of the most studied strains of this model organism, has about 4400 genes (25, 26). Of these genes, almost 4000 single knockouts are present in the KEIO collection (20). Theoretically, one could use these numbers to calculate that *E. coli* K-12 has approximately 400 essential genes. Even though only ~10% of genes cannot be knocked out by single knockout mutations, the *E. coli* cells with the most reduced genomes have a 38,9% reduction of genome size (27–29). Similarly, in the gram-positive model bacterium *Bacillus*

subtilis, a maximum genome reduction of 36% has been reached so far (30). Although the aim of these studies was not necessarily to make the smallest possible genome, but also to have a relatively normal growing cell, it shows that much more of the genome is essential than can be theoretically calculated from determining gene essentiality alone.

Top-down synthetic cells

Perhaps the most prominent example of a top-down synthetic cell was created by Hutchison III and colleagues (31). Their first approach was to synthesize a bacterium with the smallest genome known: *Mycoplasma genitalium* which naturally only has 525 genes in a 580 kb genome (32). This organism became the first organism with a chemically synthesized genome (33). However, the growth rate of *M. genitalium* was too low to efficiently minimize this organism (31). Instead, *Mycoplasma mycoides* was used for genome minimization. First, a copy of the genome was re-built by oligonucleotide synthesis of small parts and then further assembled by PCR (1.4 kb parts), which were assembled into longer parts (7 kb) by Gibson assembly (34). These parts were then assembled in two steps to a full genome (1.08 Mb) using *in vivo* assembly in yeast. Finally, the genome was transplanted into a host cell of a different organism (*Mycoplasma capricolum)* leading to the first iteration of this synthetic cell (JCVI-syn1.0) (34). Next, a minimal genome for this organism was computationally designed, by using the transposon integration approach to identify the essential genes. This genome was constructed in the same way, transplanted into *M. capricolum* and tested by selection for the synthetic genome. This design-built-test cycle was repeated until a cell with 479 genes and a genome of 531 kb was created. Remarkably, this minimal synthetic cell, named JCVI-syn3.0, still had 149 essential genes with unknown function, 79 of which could not even be placed into a broad functional category (18). However, even if the function of these genes was still unknown, deleting one of them, led to a non-functioning synthetic cell, indicating that essential mechanistic details of the bacterial cell remain yet to be discovered (31).

To gain a better understanding of this minimal synthetic cell, the vast amount of experimental experience gained from building JCVI-syn3.0 was used to construct JCVI-syn3A. The main difference between these synthetic cells is a 19-gene chromosomal insertion in section 6 of the 8-part assembly. These 19 genes contain some of the cell partitioning genes *ftsZ* and *sepF*, leading to more consistent morphological features, comparable to its ancestor, *M. mycoides*. The resulting synthetic cell, consisting of 493 genes in a 543 kb genome might not be the most minimal cell ever created, but 98% of the enzymatic reactions performed within the organism are captured in a metabolic network, supported by annotation or experimental

evidence (18). Additionally, this synthetic cell was subjected to transposon mutagenesis (11, 14). Interestingly, this screening classified 118 genes as non-essential (18). However, this population of cells that underwent transposon mutagenesis was passaged four times, this reduced the amount of non-essential genes to 76. The other 42 genes were classified as quasi-essential: not essential for survival, but essential for robust growth; strains harboring transposons in these 42 genes were outcompeted by the other strains in the population (18).

To further test the essentiality of these non-essential genes, triple knockout experiments of various sets of genes all led to cells with greatly impaired growth (18). In this context, it is virtually not possible to design a genome with only quasi-essential and essential genes. Deleting non-essential genes leads to multiple other non-essential genes to become (quasi-) essential (22). This is further proven by a similar transposon mutagenesis study in JCVI-syn1.0 where some genes were classified as quasi-essential, which were classified as non-essential in JCVI-syn3A (18, 31).

In all instances of using a top-down method to build a synthetic cell, there has always been a unknown factor. One can calculate, reason, or model a set of essential genes that in theory could be enough for life, experimentally though, this set is never enough, and other quasi-essential or non-essential genes, often of unknown function, are in fact essential for making a synthetic cell (14, 31, 34). To build a synthetic cell in which all genes are of known function, perhaps it should not start from a living organism, but from scratch: bottom-up.

Creating a bottom-up synthetic cell

There are a number of international consortia that are independently aiming to construct a synthetic cell using the bottom-up approach (35–37). These groups are aiming to combine well-studied cellular modules of different origins together to form a, living, synthetic cell (3, 36, 38, 39). These modules range from membranes to energy production, chromosome replication, transcription, translation, and cell division, as described in a recent review by Olivi et al. (Table 1.1) (38).

Table 1.1: Well studied modules for building a synthetic cell bottom-up. Table adapted from (38):

Module	Potential solution for bottom-up synthetic cell	Ref.
Container	Liposomes, water-in-oil droplets, coacervates	(40–43)
Chromosomal configuration	Linear or circular, single/complete, or multiple/partial chromosomes, single/multiple copies	(44–46)
Transcription and translation	*In vitro* transcription-translation systems (IVTT): PURE, TX-TL system	(47) Chapter 4
Energy	Arginine breakdown pathway, decarboxylase pathways, proton-pumping rhodopsins combined with ATP synthase	(48, 49) Chapter 5
Cell growth	Lipid biosynthesis, vesicle fusion	(7, 43, 47, 50, 51)
DNA replication	ϕ29, T7, T4, *Escherichia coli* replicative machineries	(47)
DNA segregation	Random or active partitioning, entropy-driven segregation	(52–54)
Cell devision		
Symmetry breaking	Reaction-diffusion, lipid phase separation, DNA partitioning	(52, 55–58)
Membrane deformation	FtsZ, Cdv, actin-processing proteins, Min system, microtubules, lipids, DNA origami, mechanical deformation	(59–65)
Membrane abscission	ESCRT-III, dynamin, active droplets, mechanical splitting	(64, 66–68)

To build a synthetic cell bottom-up, modules that contain the basic functions of life, such as the ones described above, must be stored on a genome-like information carrier, to pass them on to next generations. Many of these modules are being tested only in their protein- or RNA form in an *in vitro* environment, which does not take into account their ability to be properly transcribed and translated. Moreover, these modules are being tested independently from each other. So even if these modules work independently, there are no guarantees that they will also work all together. Although it is difficult to know for sure, it can be assumed that simply combining all modules and "kickstarting" a synthetic cell will not work immediately. For all modules to work inter-dependently, they might have to "get used to each other" by evolving to a state that allows all modules to co-operate.

Natural evolution is based on repeated cycles of genetic variation, expression, and selection (69). By non-perfect genome replication and subsequently cell division species gain genetic diversity. Some mutations might be disadvantageous or even detrimental to an organism, which then might grow slower or not proliferate at all. But other mutations are advantageous, making a species more robust, and giving it a better chance of survival in 'the struggle for life'. Although the environment in which a synthetic cell can proliferate will be highly controlled, it is difficult to predict exactly which conditions are ideal for this organism to grow. For that reason, and in the context of this thesis, I would like to question whether it is possible at all to build a synthetic cell without having it undergo any evolution.

Testing modules *in vivo* with the help of evolution

In this thesis, two important modules are tested in an *in vivo* setting, one responsible for gene expression and one for energy generation. Both modules, that could at some point be part of a bottom-up synthetic cell, have been extensively studied before. With these experiments, multiple obstacles that stand in the way of building a synthetic cell can be overcome. Firstly, these experiments can serve as a proof-of-principle to show that a module can functionally replace an analogous module in a living organism. By expressing a synthetic, non-native module and subsequently silencing or knocking out the host organism's native module, the functionality of the synthetic module can be put to the test. It is likely that the expressed synthetic module is less efficient compared to the native module, hampering the growth of the host organism. However, as the module is needed in a synthetic cell, it is automatically essential, at least in some conditions, and its functionality is directly coupled to growth. Hence, optimizing the performance of the synthetic module leads to faster growth of the host, and evolution will select for appropriate modules (Figure 1.1).

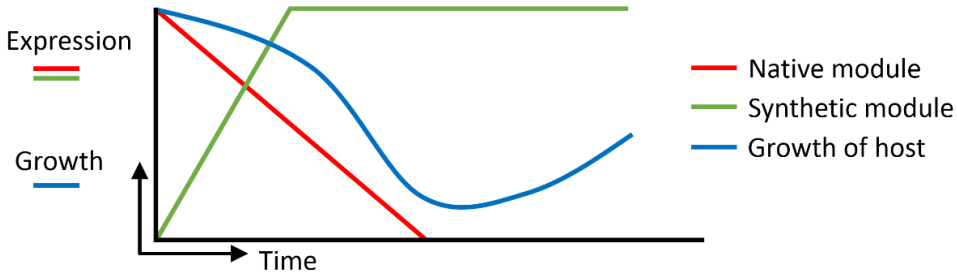


Figure 1.1: representation of an *in vivo* system for *in vitro* module optimization

This strategy for testing synthetic modules can be used for independent modules but will be even more worthwhile for testing/optimizing multiple modules together. Either subsequently, by adding a new synthetic module once the previous setting is satisfactory, or, whenever possible, in parallel, by testing multiple modules at once. Although testing multiple modules in parallel will demand a lot from the host, as multiple essential functions will be knocked down (and eventually knocked out; Figure 1.1), to be replaced by inefficient heterologous modules, it might lead to an evolutionary favorable condition. In 1932, Sewall Wright introduced the metaphor of the fitness landscape, which depicts biological fitness as peaks and valleys in a geographical landscape map of genotype space (Figure 1.2) (70, 71). In such landscapes there are multiple peaks of fitness, that are parted by deep valleys of lower evolutionary fitness. As evolution optimizes for optimal fitness, it is unlikely that an organism goes from one evolutionary fitness peak, to another, possibly higher, peak when it first leads to a loss in fitness (70).

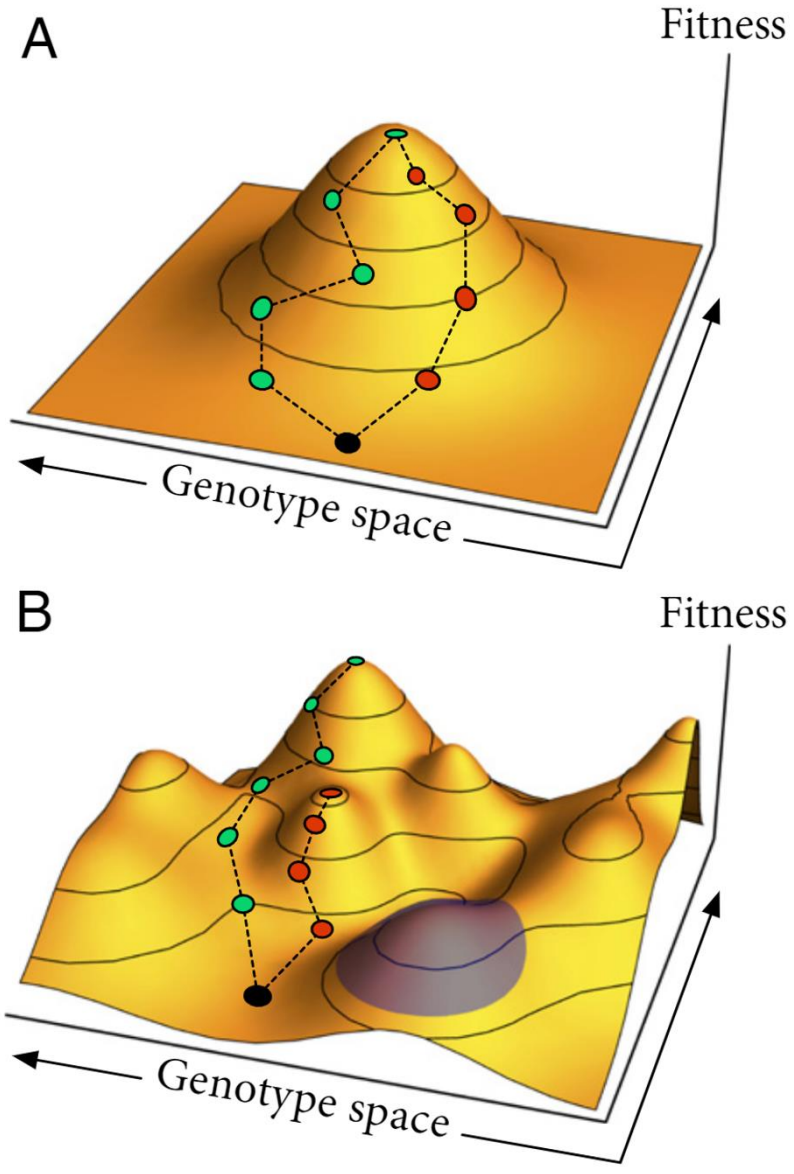


Figure 1.2: Metaphorical fitness landscapes. On the X- and Y-axis genotype space is depicted, which entails all mutations in a system. On the Z-axis fitness is depicted. A: A smooth fitness landscape with one peak. All evolutionary paths should lead to the same steady state. B: A fitness landscape with multiple peaks, separated by valleys of lower evolutionary fitness. Different evolutionary roads can lead to a different steady state. Figure from Van Cleve et al., (2015) (70).

Building the genome of a synthetic cell

Bacterial species have undergone billions of years of evolution to get to the point they are now (72, 73). This is both an advantage and a disadvantage for building a synthetic cell, bottom-up. The plethora of proteins that have evolved over these billions of years offer a gigantic toolbox for molecular biology, metabolic engineering, and synthetic biology, including for creating a synthetic cell (74–77). On the other hand, billions of years of evolution have led to a very sophisticated genome organization, that we are only just getting understanding about. Building a chromosome from scratch will most likely take a lot of trial and error. In Chapter 2 I review the current understanding of genome organization, and what we can learn from it to build a synthetic cell.

To build a chromosome that can already be tested in an *in vivo* setting, it should probably mimic a natural chromosome. For doing this in bacteria, a bacterial artificial chromosome (BAC) is an excellent candidate (78–81). A BAC is derived from the F-plasmid in *E. coli*, a naturally occurring plasmid in some strains (80, 81). BACs have their own partitioning system which allows them to stably maintain their copy number at one. Additionally, these vectors have been shown to be up to 1 Mb in size, more than what would likely be needed for a synthetic cell (23, 82). As a proof of concept, in Chapter 4 we use a BAC to test an RNA polymerase functional module which could be used in a synthetic cell. Additionally, we test the limits of genome organization, by taking one of the most genomically conserved protein complexes, and combining them in one operon.

As discussed, a BAC is a great host for a synthetic cell chromosome, however, to construct such a genome it needs additional features. In Chapter 3 we discuss a novel DNA assembly method: the RAIR pathway in *E. coli*. Although this assembly method is very efficient, the pathway has not been tested for assembly of more than four fragments and more than a few kb in size. For easier construction of a chromosome most likely Gibson assembly and assembly in yeast will be utilized (75, 83, 84) (Chapter 6). For both methods, DNA with homologous flanks is needed, but for yeast cloning at least one of the fragments needs to be able to replicate in yeast. This can be done by using a yeast artificial chromosome (YAC), which, similar to BACs can be up to 1 Mb in size (85–87). That is why in Chapter 4 we are using a BAC-YAC shuttle vector, which can be used for rapid assembly in yeast, and for *in vivo* testing of functional modules in *E. coli*.

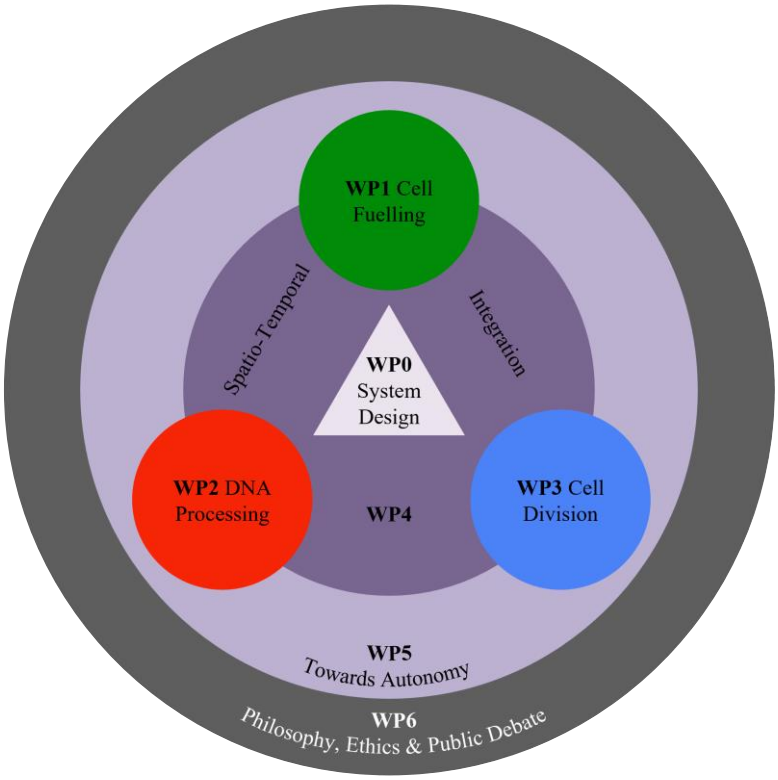


Figure 1.3: Work packages of the BaSyC project.

Outline of the thesis

This thesis describes a sub-project within the Building a Synthetic cell (BaSyC) consortium. This consortium aims to build a synthetic cell bottom-up by combining functional modules that constitute a living cell. The project is broken down into seven work packages (Figure 1.3, Table 1.2), divided between about 20 research groups at six Dutch universities and institutes. The work that is the topic of this thesis is one of the projects that have been/are being performed within work package 5: Towards autonomy. WP5 aims to test the modules designed by other work packages, combine them and work towards an autonomous synthetic cell. To do this, a genome will have to be designed that contains the genetic information for all modules. Additionally, the modules are to be tested *in vitro* and *in vivo* one by one, or in a combinatorial approach.

Table 1.2: Work packagers in the BaSyC project with description of scope.

Work Package	**Scope**
WP0 System design	Address the variables and constraints of the system by means of computational and mathematical modelling.
WP1 Cell fueling	Bottom-up engineering of functional modules that are core functions of a cell. WP1: Cell fueling, built a minimal metabolism for the synthetic cell to supply energy. WP2: DNA processing, engineer machinery for DNA replication, transcription, and subsequent RNA translation into functional proteins. WP3: Cell division, build a system to constrict the cell and divide into two daughter cells
WP2 DNA processing	
WP3 Cell division	
WP4 spatio-temporal integration	Integration of modules from WP1-3 including container control, temporal control, and intracellular spatial control.
WP5 Towards autonomy	Working towards a vesicle-based system that is able to express the modules from other WPs from a genome. Including testing of modules in a living organism.
WP6 Philosophy, ethics & public debate	Review the ethical and philosophical components of building a synthetic cell and creating public awareness.

In Chapter 2 of this thesis, the current understanding of genome organization in prokaryotic cells is reviewed. Many relevant insights have been gained on the organization of bacterial genomes in the past decades. The expression level of genes is tuned by the coding as well as the non-coding sequences. At the operon level, the order of genes may influence the expression level of these genes. One step higher, at the genome level, the location and the direction of genes and operons with respect to the chromosomal origin of replication may impact expression as well. However, new hypotheses have been brought up by recent studies in comparative genomics. These analyses have disputed some of the earlier conclusions. It is clear that, although a lot of progress has been made in this field, there are still many secrets hidden in the genome sequence. In this chapter we review the recent findings in an attempt to get a better understanding of genome organization. We discuss how gene dosage and

orientation may affect expression and evaluate the impact of large scale genome rearrangements. It is evident that a better understanding of genome organization in natural and moderately engineered systems will be crucial for the ultimate development of a functional synthetic genome.

Cloning methods to assemble smaller DNA fragments are a foundational technology needed to build synthetic genomes. In Chapter 3 the current cloning methods are discussed, and a novel cloning method is described. In molecular biology, DNA recombination is a key technique. Although many methods are currently available to do this, all of them have some shortcomings. While the most accurate techniques are relatively expensive, cheaper techniques are often less accurate, efficient, or practical. With the study presented in chapter 3 we aim to get a DNA recombination method that combines the best of both worlds: cheap, fast, and highly accurate. The RecA-independent recombination (RAIR) pathway in *Escherichia coli* DH10B proves to be a candidate for this purpose. This pathway relies on an exonuclease, a ligase, and a polymerase to repair DNA. Homologous flanks of linear DNA are first made single stranded by an exonuclease. These single stranded sequences can anneal by sequency homology, and the gaps are filled in by a polymerase, with the nicks being fixed by a ligase. To test the efficiency of the pathway, a four-fragment assembly strategy was designed aiming at recombining overlapping amplicons of a plasmid. One of the fragments contains an antibiotic resistance gene which allows to select for assembled plasmids. Additionally, two other fragments contain the two halves of the *rfp* gene, which allowed for a red fluorescence phenotype of correctly recombined plasmids. Next, we aimed to optimize the recombination strategy by overexpressing genes, of which at least some are believed to be involved in the RAIR pathway: *xthA, ligA, ligB* and *polA*. Overexpression of *xthA* led to an approximate 9-fold increase in recombination efficiency, with 98% recombination accuracy. Overexpression of *ligA* also leads to an increase of recombination efficiency but showed only 60% recombination accuracy. Additionally, the length of the homology flanks between fragments was varied. Homology regions of 50 and 30 bp resulted in a 3-fold increased recombination efficiency compared to the 20 bp flanks. Moreover, the longer the overlapping region, the higher the cloning accuracy (50 bp; 99%, 30 bp; 96% and 20 bp 89%). Overall, this technique is cheap and easy to implement in most labs and is an attractive alternative for other (more expensive) cloning methods.

Chapter 4 describes the first attempt of testing a module *in vivo*, while challenging a genomic design principle seen across all living organisms. The RNA polymerase subunits of all annotated genomes so far are scattered across the chromosome. Often, prokaryotic genes

encoding functionally related proteins are clustered in operons. The compact structure of operons allows for co-transcription of the genes, and for co-translation of the polycistronic messenger RNA to the corresponding proteins. This leads to reduced regulatory complexity and enhanced gene expression efficiency, and as such to an overall metabolic benefit for the protein production process in bacteria and archaea. As mentioned before, this is not the case for the subunits of RNA polymerase, one of the most conserved and ubiquitous protein complexes in all domains of life. To analyze the impact of this genetic organization on the fitness of *Escherichia coli*, and to serve as a proof of concept for testing modules for a synthetic cell, we designed an operon of the RNA polymerase genes. This operon was constructed using yeast assembly on a bacterial artificial chromosome (BAC) – yeast artificial chromosome (YAC) shuttle vector. In *E. coli* the copy number of the BAC is (close to) one, making it an efficient candidate for testing synthetic cell modules, as it mimics the copy number of the genome. After introduction of the BAC/YAC-encoded RNA polymerase operon, we deleted the native chromosomal genes, which led to a reduction in growth on minimal medium. However, by using adaptive laboratory evolution, the growth rate was rapidly restored to wild-type level. Hence, we show that a highly conserved genetic organization of core genes in a bacterium can be reorganized by a combination of design, construction and optimization, yielding a well-functioning synthetic genetic architecture.

In Chapter 5 we test an ATP-generating module that could be used as energy source for the synthetic cell. The selected arginine deiminase (ADI) pathway utilizes three enzymes and an antiporter to convert arginine to ornithine, thereby producing one ATP. The antiporter imports one arginine for each ornithine exported, thereby rendering the pathway practically orthogonal from cellular metabolism. We aimed to grow *Escherichia coli* using a heterologous expressed arginine deiminase pathway as sole ATP source. For this, a library approach was designed to test multiple promising variants of the pathway, under control of randomized ribosome binding sites. The library was tested by using a stringent growth-no growth selection by growing the cells anaerobically without ATP source, except for arginine, thereby coupling the arginine deiminase pathway to growth. As no growth was observed using this stringent selection method, a screening approach was developed to find active arginine deiminase pathway variants under only partially ATP-limited conditions. We found that strains harboring an active arginine deiminase pathway grew to an up to three times final cell density compared to the negative control, a finding that was confirmed by in-silico models.

In Chapter 6 a summary of the thesis is provided. Next, an overall discussion on the thesis is given, and potential future experiments are described. Finally, an outlook is given focusing on building the synthetic cell from the perspective of Work Package 5, how to test and combine modules efficiently, and use them as chassis for a synthetic cell.

Chapter 2

From natural towards synthetic genome organization in bacteria

Joep Houkes[1], Nico J. Claassens[1], John van der Oost[1]

[1] Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands

Abstract

In the past decades many relevant insights have been gained on the organization of bacterial genomes. At the level of a gene, the expression level is tuned by coding as well as non-coding sequences. One step further, at the operon level, also the order of genes may influence their relative expression levels. At a genome level, the location and the direction of genes and operons, with respect to the chromosomal origin of replication, have been reported to impact expression as well. However, recent comparative genomics analyses have disputed some earlier conclusions, and provided new hypotheses. It appears that despite impressive progress, there are still many secrets hidden in the genome sequence. Here we review recent findings that may contribute to a better understanding of the organization of the bacterial genome. We discuss how gene dosage and orientation may affect expression. In addition, we evaluate the impact of large genome-scale rearrangements and of physical chromosome structure. It is evident that a better understanding of genome organization in natural and engineered systems will be crucial for the ultimate development of a functional synthetic genome.

Introduction

The majority of bacteria store their genetic information on a single, circular chromosome, the replication of which relies on a single origin of replication (88). However, it should be noted that many variations of the bacterial genome organization have been described, including polyploidy (2-10 copies; (89)), linear chromosome (e.g., *Agrobacterium*; (90)), more than one chromosome (e.g., *Vibrio cholerae*; (88)), co-occurrence of mega-plasmids (91), and possibly (as demonstrated in some Archaea) even multiple origins of replication (92). The main focus of this review, however, will be on the most frequently occurring bacterial genome organization, i.e., a single, circular chromosome.

Gene dosage effect

In case of a typical bacterial chromosome, duplication starts by local unwinding of the double stranded DNA at a single origin of replication (*oriC*), after which the two replication forks move in both directions, eventually meeting on the opposite site at the terminus (*ter*) site of the chromosome (Figure 2.1) (93). This mechanism implies that, during replication, genes that are positioned closer to the *oriC* have temporarily a higher copy number compared to genes located further away. This phenomenon is known as gene dosage effect, and is considered an important factor in gene expression, especially in bacteria (94, 95).

During slow growth, one round of replication is finished per cell division (96, 97). During fast growth regimes, however, more than one round of replication can take place simultaneously in a single cell, a process known as multi-fork replication (Figure 2.1) (98). In *Escherichia coli*, close to the *oriC*, eight-fold increased gene copy numbers have been observed during multi-fork replication (99). Since a gene's copy number may strongly correlate with its expression, the gene dosage effect often increases expression of origin-proximal genes under optimal growth conditions.

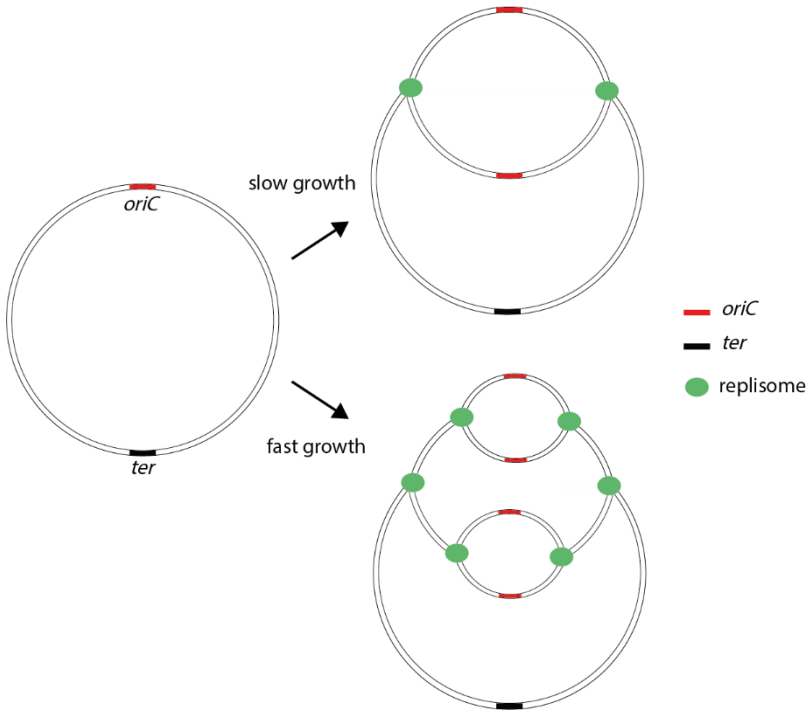


Figure 2.1: Gene dosage effect during replication of bacterial chromosome. Upon replication, regions close to the *oriC* will have a higher relative copy number (2-fold), especially during fast growth (4-fold, or even higher). This gene dosage effect may result in elevated expression of genes that are located closer to the *oriC*.

Genes that are highly expressed during exponential growth, such as genes encoding RNA polymerase, ribosomal RNA and ribosomal proteins are often found to cluster systematically around the *oriC* (98, 100–102). In this way, the expression of these growth-related genes is relatively high due to the gene dosage effect during fast growth (98). In line with this, analysis of the fast-growing bacterium *Vibrio cholerae* shows that localization of ribosomal genes close to *oriC* is important for vitality (103, 104). Relocating this ribosomal gene cluster to an *oriC*-distal position resulted in significant growth rate reductions (up to 20%) (103, 105). The reduction of growth rate was largest during exponential growth, indicating the gene dosage of these genes does indeed contribute to overexpression. Addition of a second copy of the gene cluster at the *oriC* distal position released the growth defect (103). Using *in-silico* analyses, it was found that natural chromosomal inversions and gene

translocations occur preferentially *oriC*-symmetrical (the same distance from the *oriC* as before the translocation) (106–109). It is hypothesized that this, at least partly, relates to the gene dosage effect, as large *oriC* asymmetrical inversions or translocations would alter gene expression levels in many cases.

Another factor related to gene dose, is the genomic copy number of genes. A well-known example is the prokaryotic ribosomal RNA operon (5S, 16S and 23S rRNA), the copy number of which ranges from 1 to 15 copies (110–112). A plausible explanation why bacterial and archaeal rRNA genes often occur in multiple copies whereas the genes encoding the ribosomal proteins do not, is the fact that transcription of the latter genes is followed by a second, compensatory amplification through translation. The ribosome content of cells, a key factor for the doubling time of cells, has been suggested to be limited by the number of rRNA copies in the chromosome. This hypothesis has been tested in a study in which systematic reduction of the 7 *rrn* operons from wild type *E. coli* were deleted in to generate a set of strains with 1-7 copies. Indeed, the resulting doubling times were significantly lower compared to wild-type *E. coli* (113, 114). In another study, extra copies of *rrn* operons were added to the *E. coli* chromosome, which showed that with 8 *rrn* copies the growth rate was comparable to wild type, but that more copies decreased the growth rate (115). It has been suggested that under specific growth conditions, slightly changed *rrn* operon copy numbers in *E. coli* could actually be beneficial for growth (115). The wide range of *rrn* copy numbers (1-15) that occurs in different bacterial and archaeal genomes (110–112) does not appear to correlate with host features (genome size, doubling time). Hence, most likely other control mechanisms, including promoter strength, may be responsible for compensating for different copy numbers.

Gene orientation bias

A vast majority (90%) of prokaryotes have disproportionally more genes on the leading strand compared to the lagging strand (116–118). Across all sequenced bacteria and archaea, the percentage of genes on the leading strand ranges from 45%-90% (119). A generally accepted explanation for this bias is the potential occurrence of collisions between the protein complexes responsible for replication (DNA polymerase and associated proteins) and transcription (RNA polymerase). Both enzymes use the same template, but move at different speeds, inducing head-tail collisions on the leading strand and head-on collisions on the lagging strand (118, 120–126). Whereas head-on collisions can lead to error-prone replication fork stalling, head-tail collisions are less harmful (116, 127). This effect causes

lagging strand genes to accumulate more mutations over time compared to genes on the leading strand (128).

The gene orientation bias is known for a long time. Already in 1988 it was suggested that genes with higher expression rates were predominantly located on the leading strand (120). Later, it was hypothesized that not high expression, but rather gene essentiality is the driving force for this bias (129, 130). According to this hypothesis, essential genes (involved in information processing and storage, cell cycle control, cell wall biogenesis, posttranslational modification, energy production and conversion, as well as transport and metabolism of carbohydrates, amino acids and nucleotides) are preferentially located on the leading strand (123, 131, 132). However, as approximately only 10% of genes in *E. coli* and *Bacillus subtilis* are essential, gene strand bias can only be partly explained by essential genes being preferably on the leading strand (133, 134). Additionally, the strand bias in *E. coli* is not as pronounced as in many other bacteria, with only approximately 55% of its genes (but 63-71% of its essential genes) on the leading strand (135). In *B. subtilis* on the other hand 74% of its genes (93% of its essential genes) are located on the leading strand.

In multigene operons, with longer transcripts, there is a bigger chance of a collision between the transcription complex and the replication machinery. This leads to a stronger bias for operons to be on the leading strand, as is the case in *E. coli* (125). In agreement with this, all 7 rRNA operons on the *E. coli* chromosome are located on the leading strand (102). Likewise, all ten *B. subtilis* rRNA operons are located on the leading strand (136). In addition, it should be mentioned that six of the rRNA operons of *B. subtilis* are located extremely close to the origin (within less than 200 kb).

A striking observation was made that low-GC gram-positive firmicutes (e.g. *Bacillus* and *Clostridium* species) harbouring the PolC-type DNA polymerase have on average a much higher strand bias (74%) than bacteria that do not encode *polC* (58%) (135, 137). The PolC polymerase is responsible for leading strand replication, while the lagging strand is replicated by DnaE (137). In addition, these bacteria often exhibit a strong (54 %) purine asymmetry, meaning there are more purines (A and G) on the leading strand compared to the lagging strand (122, 138). Many prokaryotes have a slight purine asymmetry (51%) (139), the hypothesis for this is that during replication, fork stalling generally leads to more mutations in the lagging strand than in the leading strand (124, 127, 128). Energetically, it is cheaper to introduce a pyrimidine to repair the mutation compared to a purine (140), it is proposed that this causes purine enrichment of the leading stand, both coding and non-coding (116). The coincidence in firmicutes of three typical genetic features (strong strand bias, purine

asymmetry and the presence of *polC*) has led to the hypothesis that PolC-based, replication of the leading strand, might proceed with higher fidelity compared to DnaE. This would mean that DnaE introduces more mutations during replication on the lagging strand, compared to PolC on the leading stand. Incorporating a pyrimidine is energetically less costly than a purine, would then cause more pyrimidines to be incorporated at the site of the mutation, which leads to enhanced purine levels on the leading stand (122). An interesting hypothesis that remains to be verified.

Experimental evidence for gene strand bias shows that inversing genes can be harmful for the cell, leading to decreased replication fork progression rates, replication fork arrests or even double strand DNA breaks (127, 141–143). Contrarily, introduction of small and large genomic inversions in *E. coli* revealed that this only has a moderate physiological effect (144). Instead, it has been reported that changing the balance between the two replication arms has a much bigger effect. Is indicated above, *E. coli* does not have a very strong leading strand bias (55% of genes on leading strand) (135). Although these inversions might not have short-term effects on the cell's physiology, it cannot be ruled out that on the long run they still may have an effect due to increased mutation rates.

In recent years, several studies (mainly in *E. coli* and *B. subtilis*) have been executed to address the effect of genomic location (distance from *ori*) and of genomic orientation (leading/lagging strand) on expression by integrating genes encoding reporters such as the green fluorescent protein (GFP). In an initial study (145), 300-fold differences were described, not due to a gene dosage effect but rather mediated by "several different features related to chromosome organization" (see below). In subsequent studies, no significant gene orientation bias was observed, but in these cases the observed expression variations was reported to correlate with the genomic position (gene dosage effect) (146, 147), as well as with the earlier observed regions that have generally higher transcription rates (e.g., ribosomal operons) (147). The latter regions have been identified by analysis of *in vivo* protein-DNA interactions in *E. coli* (extensive protein occupancy domains (EPODs) (148)), as discussed in more detail below (Chromosome structure).

A similar GFP-cassette genome-integration study was also performed in *B. subtilis* for 38 loci (149). Although leading/lagging strand bias was not specifically mentioned, we noted that the highest expression is observed for leading strand loci. It cannot be ruled out that this effect may (at least partly) be caused by read-through, as the more genes and promoters are present on the leading strand (74%) in *B. subtilis*. In another study in, 14 expression cassettes were randomly integrated across the *B. subtilis* genome, revealing a gene dosage effect of up

to 5-fold expression difference (150). In this case, no difference in expression is reported between integration on the leading and the lagging strand. The authors hypothesize this is because the proposed negative effects of genes on the lagging strand of the genome mostly occur during replication, when DNA polymerase and RNA polymerase can collide. In this study the expression levels are measured for a short term only, which might not be enough to reveal the evolutionary disadvantages from having genes expressed on the lagging strand (150).

The evolutionary pressure for genes to be on the leading strand of replication leads to the question of why there are lagging strand genes at all? If there is selection against genes on the lagging strand, why do they still exist? It might simply be that evolution led to a system that is performs well enough (under certain conditions), and not necessarily to a system with the highest possible efficiency. Over the years, some other hypotheses have been proposed on this matter. It could potentially be a balance between evolutionary pressure to and random inversion events, which leads to the current distribution of genes across both strands (151). In another study it has been hypothesized that it could be genetically favorable for stress-response (or stress-induced) genes to be positioned on the lagging strand (152). This would agree with the observation that during growth of some gram-positive bacteria (*Bacillus*, *Listeria*) on rich medium only few genes on the lagging strand are expressed, whereas under stress conditions the expression of many of the stress-response genes on the lagging strand genes is upregulated (153–156). In this way, replication rates will generally slow down during stress, while occurring replication-transcription conflicts may lead to higher mutation rates in these genes which is suspected to be beneficial for these genes that are rapidly responding to threats (152, 157). However, this hypothesis is disputed in a recent review where it was reported that genes regulated by stress-associated transcription factors are more frequently located on the leading strand (158). Hence, this matter remains elusive. It would be interesting to divide stress-related genes in two groups: those that respond either to biotic or to abiotic threats, since the former may benefit from enhanced mutation rates (arms race) whereas the fitness of the latter most likely is bigger in case of high fidelity replication.

Conservation of gene clusters

The clustering of certain genes on the genomes of different bacteria appears to be rather well conserved. This is partly due to phylogenetic relatedness, but for a big part because of evolutionary benefits of this colocalization (159, 160). Gene clusters in prokaryotes are typically composed of functionally related genes ("guilt by association" (161)). Genes of which the derived proteins have related functions in bacteria (e.g., subunits of a protein

complex, or enzymes of a metabolic pathway) are often organized in operons (162, 163). Operon-encoded genes are often transcribed from a single promoter, resulting in a polycistronic mRNA. Operons are proposed to be beneficial for bacteria as they simplify regulation of the simultaneous expression of functionally related genes, and as such they contribute to the coordinated production of associated protein in appropriate stoichiometries (164, 165).

The selection pressure to conserve gene location (and order) can be due to the interaction of gene products, as subunits in protein complexes and/or as enzymes that catalyze consecutive steps in metabolic pathways (166, 167). This property of bacterial and archaeal genomes is used by scientists to discover novel functionally related genes, e.g., genes involved in CRISPR-Cas (CRISPR-associated gene clusters were discovered like this (168, 169)) and other anti-phage defense systems (170), in metabolic pathways (171), and in bacterium-plant associations (172).

It is suggested that spatial conservation across species is high because major chromosomal rearrangements are relatively rare as they generally are not beneficial for host fitness (101, 173, 174). Chromosomal rearrangements that influence the replichore balance (a more or less similar length of the two replicating chromosome halves, from origin to terminator) lead to fitness loss and are thus selected against (174). In addition, chromosomal inversions impede the gene strand bias, and might (on the short or long term) also result in loss of fitness. This agrees with long term evolution experiments with *E. coli*, in which major genomic rearrangements are often symmetrical to the *oriC* (173). This also suggests that, at least in *E. coli* under the used conditions, leading/lagging strand bias does not play an important role.

In 2016, the synthetic genome of JCVI syn3.0 was published (175). Here, Hutchison and colleagues minimized the *Mycoplasma mycoides* genome from 1 Mbp to less than 0.5 Mbp. Additionally, the gene order of one of the eight synthetic chromosomal segments was reorganized logically, by further grouping genes with similar functions. Cells with this reorganized segment grew about as fast as native JCVI syn3.0 cells (175). The other seven segments of the JCVI syn3.0 genome were also redesigned, however, for unknown reasons these did not result in a functional cell (175).

Chromosome structure

Like the genomes of archaea and eukaryotes, the bacterial genome has to be folded in order to fit inside the cell (176, 177). Although studying chromosomal packaging appeared more difficult in bacteria than in eukaryotes, major progress has been made in recent years (177,

178). Packaging of bacterial DNA occurs via nucleoid-associated proteins, most notably, Histone-like nucleoid structuring (H-NS) proteins and structural maintenance of chromosomes (SMC) complexes (Figure 2. 2) (177). H-NS proteins, which have often been transferred by horizontal gene transfer, are small polypeptides that bind DNA, favoring AT rich sequences (179–185). Interestingly, some functionally related H-NS proteins often arise from convergent evolution, the H-NS protein from *B. subtilis*, Rok, shares no structural similarity with *E. coli* H-NS, yet performs the same function (186). Additionally, H-NS proteins from *Burkholderia* spp. and *Mycobacterium tuberculosis*, Bv3f and Lsr2 respectively, share no structural homology with Rok, and only share partial structural similarity with *E. coli* H-NS in their DNA-binding domain (185, 187). SMC complexes are protein rings that occur in all domains of life and can loop around DNA to condense it (188).

Bacterial chromosomes are partitioned into chromosome interaction domains (CIDs) at the scale of tens to hundreds of kbs (189, 190). In bacteria, these domains where first found in *Caulobacter crescentus*, which has 23 CIDs during growth in rich medium, and 29 CIDs during starvation (190, 191). The boundaries between CIDs are formed by genes of at least 2kb that are highly expressed in one of these conditions (190, 191). For example, one of the ribosomal RNA operons forms a sharp CID boundary in the exponential phase, when these genes are highly expressed. During starvation, these genes are expressed to a lesser extent, leading to a less-pronounced CID boundary (190). In *E. coli* the chromosome is divided into 31 CIDs, 22 of these have boundaries corresponding with highly expressed genes (including five of the rRNA operons), the other nine coincide with genes that contain an export signal sequence, or mobile genetic elements (192). It has been hypothesized that protein translocation is coupled to transcription-translation, which might be aided by these CIDs (193). The *B. subtilis* chromosome is partitioned into 20 CIDs, 60% of the boundaries correspond to highly expressed regions (like *sfrA*, involved in sporulation and the operon *pdhABCD*, encoding the pyruvate dehydrogenase multienzyme complex), approximately 30% of the boundaries contain nucleoid-associated protein Rok binding sites, implying that Rok, and other bacterial nucleoid-associated proteins, could be involved in pinpointing CID boundaries (189). This diversity suggests that multiple mechanisms may be responsible for defining CID boundaries, including local decompaction of active transcribed regions (191, 192).

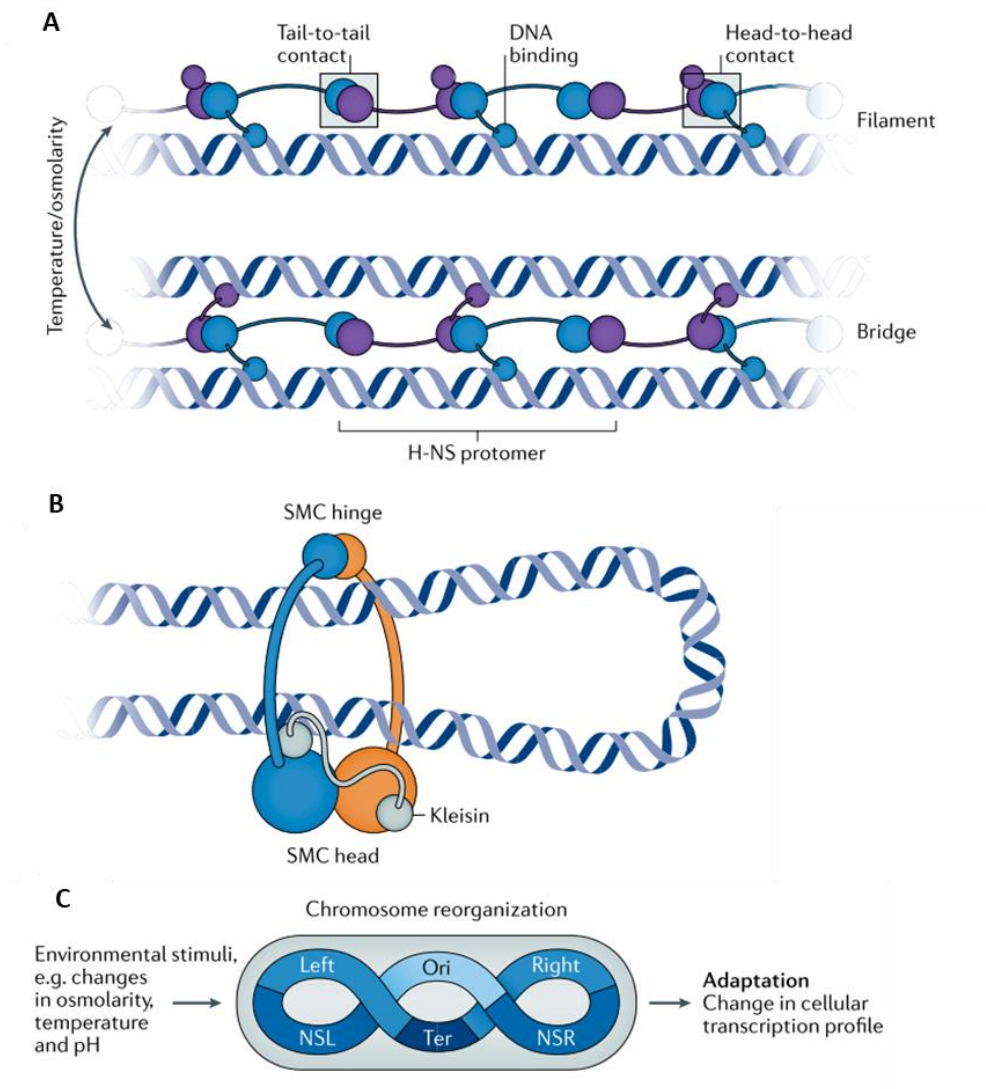


Figure 2. 2. Bacterial genome structure. A: Packaging DNA in via histon-like nucleoid proteins. B: Packaging of DNA via SMC complex. C: Chromosomes can structurally reorganize due to external stimuli, thereby changing cellular expression profile. Figure adapted from Dame et al., 2020 (177).

Global chromosome organization in bacteria is best studied in *E. coli.* Four macrodomains (Ori, Ter, Left and Right) and two non-structured regions (NSL and NSR) have been distinguished for *E. coli* where the macrodomains have reduced mobility (movement of the

chromosome within the cell) compared to the non-structured regions (194, 195). The conformation of the *E. coli* chromosome can change upon environmental changes. Certain compounds (e.g., Mg^{2+}) can directly influence the genome organization by acting as a ligand for H-NS, altering its conformation (196).

Other regions that have been associated with genome structure in *E. coli* have been termed 'extensive protein occupancy domains' (EPODs) (148). These are domains on the chromosome that are highly enriched in protein binding sites and, therefore, DNA binding proteins, of over 1 kb in size. Two classes of EPODs are discerned, transcription silent (ts)EPODS and highly expressed (he)EPODs (148). The hsEPODs are found in coding regions of highly expressed genes, including genes encoding proteins involved in transcription-translation. A total of 151 tsEPODs have been observed in non-coding regions. It is suspected that these domains serve an architectural role, but this remains elusive. Interestingly, the gene expression studies described above, in which a GFP cassette was randomly integrated into the *E. coli* genome found differences in expression that could not be explained by gene dosage effect alone (145). Additional research showed that some of the loci that were expected to result in high expression due to gene dosage effect, but turned out to result in low expression, were actually located in tsEPODS (145, 147).

Outlook

It is well known that the expression of a single gene relies on many distinct yet interconnected features, such as copy number, promoter strength, 5'-UTR, secondary structure, codon bias, 3'-UTR, and terminator) (197). As such, it is not unexpected that the overall functionality of a genome depends on even more features, both structural and regulatory. Spectacular progress has been made during the last decades in gaining insights in the important features of genome architecture. This includes the release of thousands of complete bacterial and archaeal genome sequences (198), allowing for extensive *in silico* and experimental genomics analyses that has revealed an impressive advance of our biological and molecular genetics knowledge. Many of the relevant features of genome functionality have been reported over the years, resulting in some generic features but also in many specific features (this review). Although it is very well possible that specific features that impact genome functionality did evolve independently, it seems more realistic to assume that many generic features are playing a role in genome functionality in many different bacteria. This would resemble the situation in gene architecture, a similar set of features is playing a role, but each gene appears to have a specific combination of features, with quantitative and /or qualitative variation, to reach a certain functionality. If this would also be true at the genome level, each

of these features does contribute differently to the overall functionality of a genome in a certain organism and/or under certain conditions. As described in this review, the different features that impact the genome include the aforementioned features that determine gene functionality, and additional variable such as distance of gene to origin (gene dosage), leading/lagging strand bias, and chromosome structure/gene accessibility (CID, EPOD). This would mean that genes as well as genomes are depending on a range of features, and most likely different combinations have evolved that lead to a 'functional' (not necessarily optimal) solution. Because of the combined contribution of different features, their individual contribution may vary a lot between genes within a genome, and even more so between genes between different genomes.

A consequence of this complexity would be that it does not allow for a straightforward comparison of different model systems (gene, genome), especially when different experimental approaches have been used. Revealing relevant details of the important features should serve as a basis for future projects on designing synthetic genomes. Apart from progress in biological and molecular genetics, also technological breakthroughs are crucial for the future of synthetic biology. In that respect, progress at the level of *in vitro* DNA synthesis should be mentioned (175), as well as *in vivo* DNA recombination (87, 175). All together, these developments set the stage for unprecedented synthetic biology efforts, including the bottom-up generation of a microbial genome.

Exciting recently established examples of genome engineering include: (i) splitting the genome of *E. coli* in two or more partial chromosomes (82), (ii) converting the circular *E. coli* genome to one or more linear chromosomes (199, 200), and (iii) minimizing the genomes of several bacteria by the elimination of non-essential genes. Two selected examples of minimized bacterial genomes are *M. mycoides*, reducing it from 1.0 Mbp to 0.5 Mbp (175), and *C. crescentus*, implementing major sequence edits to allow for (more efficient) synthesis of DNA building blocks, eventually reducing the genome from 4.0 Mbp to 0.8 Mbp (19). All these developments serve as a solid fundament for future bottom-up approaches to design, generate and transplant synthetic genomes to biological or abiotic compartments.

Chapter 3

Fast homology mediated *in vivo* DNA assembly in *Escherichia coli* DH10B

Joep Houkes[1], Karolis Sirmonatis[1], Belén Adiego-Pérez[1], Constantinos Patinios[1], Thijs Nieuwkoop[1], Nico J Claassens[1] and John van der Oost[1]

[1] Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands

Abstract

DNA assembly is a key technique in molecular and synthetic biology. Even though many different methods are currently available to perform this task, they all come with their shortcomings. While the most accurate techniques are relatively expensive, cheaper techniques generally are neither accurate nor efficient. The aim of this study was to get the best of both worlds: develop a cheap and fast DNA recombination method that is highly accurate. For this purpose, we selected the RecA-independent recombination (RAIR) pathway of *Escherichia coli* DH10B. This pathway relies on an exonuclease, a ligase and a polymerase to repair DNA. To test the efficiency of the pathway, a four-fragment assembly strategy was designed aiming at assembling linear, overlapping double-stranded DNA fragments into plasmid. Two of these fragments contain the two halves of a *rfp* gene, which allowed for a red fluorescence phenotype for correctly assembled plasmids. One other fragment contains an antibiotic marker for screening purposes. To optimize the assembly efficiency, genes that are believed to be involved in the RAIR pathway: *xthA*, *ligA*, *ligB* and *polA* were overexpressed. It was found that overexpression of *xthA* led to an approximate 2-fold increase in number of transformants from the assembly fragments, with 98% correct assemblies. *LigA* also leads to an increase of transformants but showed only 60% accuracy. Additionally, the length of the homology flanks between fragments was varied. Homology regions of 50 and 30 bp resulted in a 3-fold increased assembly efficiency compared to the 20 bp flanks. Moreover, the longer the overlapping region, the higher the cloning accuracy (20 bp: 89%, 30 bp: 96% and 50 bp: 99%). Overall, this technique is cheap and easy to implement in most labs, and is an attractive alternative for other (more expensive) cloning methods.

Introduction

One of the most fundamental technologies for molecular biology, metabolic engineering and synthetic biology is DNA assembly. It allows for the almost unlimited design, assembly, and modification of DNA. Due to progress in DNA cloning techniques, ground-breaking discoveries in cellular and molecular biology have been made, as well as synthetic genomes have been created (31, 201). For such a cornerstone technique in biology, cost and speed of scientific projects could be significantly enhanced by improved cloning methodologies (202). In 1973 the first DNA assembly method was developed, which was based on classical "cut and paste" by bacterial Type II restriction enzymes and phage T4 DNA ligases (203). Especially after the development of the Polymerase Chain Reaction (PCR; (77)), this has been complemented by a series of technical developments. In methods like Golden Gate and MoClo cloning, the standard Type II restriction enzymes are replaced by Type IIS restriction enzymes that cut DNA outside of their palindromic recognition sites that provides unlimited freedom in generating matching sticky ends for fusing DNA fragments. (204–206). In the bridging oligo-based cloning method, non-homologous DNA fragments are cloned with bridging oligo's using ligase chain reaction (LCR) (207, 208). During the reaction the DNA strands are melted, which allows the bridging oligo's to anneal the DNA fragments together, allowing them to be ligated. Up to 20 DNA fragments with a total construct size of up to 26 kb in size have been reported to be assembled by using LCR. In Gibson assembly (75) to-be-assembled PCR fragments are generated with overlapping ends, after which one of the DNA strands from each amplicon is partially removed by a 3'-5'exonuclease. The complementary single strand overhangs anneal, and the gaps are filled in using a DNA polymerase and subsequently annealed with a ligase. Gibson assembly has been used to assemble 5-15 DNA fragments (75). Another in vitro homology-based cloning technique is SLiCE assembly (Seamless Ligation Cloning Extract). SLICE uses an *E. coli* cell extract for homology-mediated assembly of DNA (84, 209–211).

E. coli is frequently used in molecular biology studies for various tasks, ranging from propagation of plasmid DNA to whole genome sequencing by creating large clone libraries (212). *E. coli* strain DH10B is widely used to perform these tasks as it possesses various useful qualities, such as high transformation efficiency, capability of taking up and maintaining large sized plasmids, ability to screen via LacZ-based blue-white-screening, and lack of methylation dependent restriction systems (MDRS) (213). Additionally, DH10B lacks recombinase enzymes such as the native RecA, or a phage-derived RecE/RecT. These recombinase enzymes may cause plasmid instability, either by integrating plasmid in the

genome or by inducing multimerizations and deletions of a plasmid sequence (214). However, It has been known for a long time that *E. coli* strains lacking *recA* still have recombination capacity (215).

Almost 30 years ago, the RecA-independent recombination pathway (RAIR) was discovered and used to clone DNA for the first time (214, 216, 217). It was first noticed that linear plasmids frequently circularized in *recA*- strains. Recirculation always took place at the sites that shared homology, even as low as 4 bp (214). Furthermore, it was observed that in order to get successful recombination between DNA fragments both of the inserts had to be linear and share at least 10 bp homology. Recombination efficiency was observed to increase significantly when longer overhangs were used (217, 218). Since these observations in the 90s, this promising technique has only seldomly been applied. Nonetheless, it did re-emerge as a viable option to clone DNA in recent years (202, 218–223).

The simplicity and low cost of this method makes it a very attractive option for DNA assembly. It does not require any heterologous enzymes nor an *in vitro* step before transformation. It is supported by simple and robust *in vivo* assembly of DNA fragments that share short homologous sequences (10-50 bp). Usually, a RAIR pathway protocol follows three simple steps: (i) homology generation between DNA fragments by PCR, (ii) DNA clean-up, and (iii) transformation in a RecA-deficient strain for DNA assembly (219).

Although the RAIR pathway was discovered almost three decades ago, its mechanism is still not completely clear. Analogous to the *in vitro* Gibson assembly approach, it is generally assumed that the *in vivo* RAIR pathway utilizes an exonuclease which generates single-stranded homologous ends, allowing for annealing of homologous fragments. Next, a DNA polymerase elongates the DNA ends to fill the gaps, followed by a ligase which ligates the DNA.

Further research tested the effect of six exonucleases (*xthA, recE, exoX, recBCD, sbcCD, nfo*, and *tatD*) endogenous to *E. coli*, which prefer double stranded DNA as a substrate (202). Knockout of *xthA* (*Exo III*) drastically decreased RAIR cloning efficiency (by 99.3%), whereas inhibition of any of the other exonucleases had little effect on RAIR cloning efficiency. Also, DNA polymerases involved in RAIR cloning were assessed. *E. coli* possess 5 different DNA polymerases, three non-essential (Pol II, Pol IV, and Pol V) and two essential (Pol I and Pol III). Single knockouts of the non-essential polymerases did not yield negative effects on RAIR cloning efficiency, indicating either that they are not present in the pathway, or that they have overlapping functionalities. The polymerase Pol III could not be

knocked out as it is essential for cell growth. Despite the fact that Pol I is needed for cell to grow on rich medium, only the N-terminal domain which encodes the 5' to 3' exonuclease activity is essential. A mutant *polA* gene, coding for Pol I was constructed, with the non-essential C-terminal domain deleted. Strains with this mutation had a threefold decrease in homologous recombination efficiency while comparing with wild type, suggesting that PolA to some extent is involved in RAIR pathway (202). Although the ligase that is involved in the RAIR pathway has not been elucidated, it may be that the housekeeping LigA is responsible for the final step of RAIR-repair.

Materials and Methods

Bacterial strains and culture conditions

E. coli DH10B (F– *mcrA* Δ (*mrr-hsdRMS-mcrBC*) φ80*lacZ*ΔM15 Δ*lacX74 recA1 endA1 araD139* Δ (*ara-leu*)7697 *galU galK* λ– *rpsL* (Str^R) *nupG* (224)) strain was routinely grown in LB medium (10 g/L NaCl, 10 g/L Tryptone and 5 g/L Yeast extract) or for competent cell preparation in 2xYP medium (16 g/L Peptone and 10 g/L Yeast extract) at 37 °C 200 rpm, and on solid media (LB supplemented with 15 g/L bacteriological agar) at 37 °C with appropriate antibiotics (50µg/ml kanamycin, 25µg/ml chloramphenicol).

Plasmids and assembly fragment design

Plasmids created during this study were assembled using Q5 PCR amplification (New England Biolabs) and HiFi DNA assembly (New England Biolabs), using manufacturer instructions (plasmids used in this study are listed in Supplementary table 3. 1, Primers are listed in Supplementary table 3.2). Plasmids isolations were performed by using Thermo Scientific GeneJET Plasmid Miniprep kit Using manufacturer instructions.

Fragments for assembly testing were based on plasmid pTN003 (Supplementary table 3. 1) and amplified using PCR amplification (Supplementary table 3.2). To ensure no full plasmids remained after amplification, PCR product was treated with DpnI (New England Biolabs) restriction enzyme that cleaves methylated DNA, according to manufacturer instructions. Additionally, all fragments were run on 1% agarose gel in 0.5xTAE for 20-40 minutes at 100V. DNA isolation from a gel was performed by using the Zymoclean Gel DNA Recovery Kit Following manufacturer instructions.

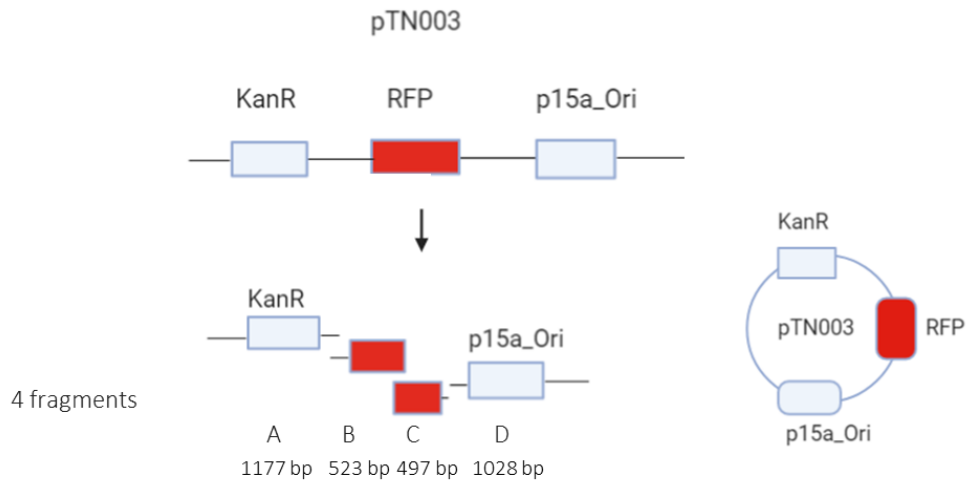


Figure 3.1: Schematic representation of plasmid assembly design. pTN003 plasmid divided into four (A, B, C & D) fragments with 50 bp homologous flanks. Fragment sizes are indicated in the figure.

Plasmid pTN003 was split into four fragments A (1177 bp), B (523 bp), C (497 bp) and D (1028 bp) that share 50 bp homology at their ends (Figure 3.1). One of the cuts was designed to be in the middle of the *rfp* gene. Thus, only correctly assembled plasmids could give red colonies. Additionally, a two-fragment assembly is possible by combining fragments A with B and C with D.

Competent cell preparation and transformation

E. coli cells were made electrocompetent by culturing at 37 °C in 2xYP (supplemented with 1 mM IPTG when appropriate) at 200 rpm until OD600 nm 0.4 was reached. The cells were then washed one time with 1 culture volume of ice-cold ddH2O and two times with 0.5 culture volumes of ice-cold 10% glycerol. Finally, the cells were suspended in ice-cold 10% glycerol to a final volume of 400 µL for each 100 mL of initial culture volume. From the first washing step onwards, the procedure was performed at 4 °C.

Electroporation was performed with 1ng of plasmid DNA in ice-cold 2 mm electroporation cuvettes at 2500 V, 200 Ω and 25 µF. Recovery in LB medium was performed immediately after at 37 °C, 750 rpm for 1 h when plasmids were transformed. During assembly experiments, recovery was performed at 30 °C, 750 rpm for 2.5 h. After recovery, the cells were plated in LB agar plates with appropriate antibiotics. Single colonies were picked and

re-suspended in 50 µL of ddH_2O for colony PCR and used to inoculate 10 mL of LB with appropriate for overnight incubation and subsequent isolation of plasmids.

Fragments were typically transformed at 100 fmol per fragment using the same transformation conditions, unless specified otherwise. After recovery, 50 µL of the cell suspension was directly plated on LB plates with kanamycin. As a negative control the cells will be transformed with all but one fragment, colonies that are found in this transformation are either from mis-assembly or full plasmids left from fragment preparation.

Results & Discussion

E. coli DH10B as a tool to clone DNA by itself

The RAIR pathway allows for *in vivo* assembly of linear fragments with short overlapping sequences (218). Despite the popular Gibson assembly method for *in vitro* recombination (75), the RAIR variant is hardly used. Still, we considered that an efficient *in vivo* assembly method would be an interesting addition to the genetic toolbox. To test the capacity of *E. coli* DH10B for DNA assembly, cells were transformed in triplicate with four DNA fragments (A, B, C and D, Figure 3.1) that were obtained by PCR amplification of different, partly overlapping parts of an RFP-encoding plasmid. After plating and overnight growth red and white colonies were counted. On average over three plates 262 red colonies were counted, indicating correct recombination, and 6 white colonies, indicating at least the kanamycin resistance gene is present, as the colony grows on kanamycin (Figure 3. 2). The negative control, transformed with only three of the four fragments showed only 3 colonies on average in triplicate.

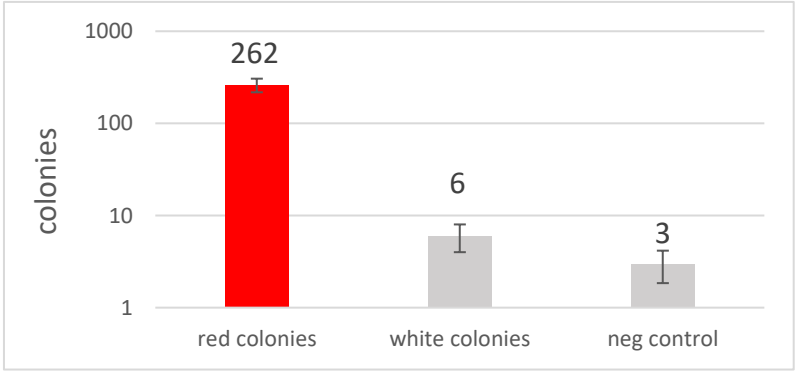


Figure 3.2: Colony count after transformation of *E. coli* DH10B with four fragments A, B, C and D in triplicate.

The mechanism responsible for this *in vivo* recombination in *E. coli* DH10B is suspected to be the RAIR pathway, as the DH10B genotype is Δ*recA* (224). It has been shown previously that knocking out exonuclease III (*xthA*) leads to a drastic decrease in recombination efficiency in the RAIR pathway (202). To further investigate its role in the RAIR pathway, and in an attempt to improve the RAIR-based assembly efficiency, plasmid pKS01 was designed to overexpress *xthA*. A ligase is also suspected to be involved in the RAIR pathway, at the time of writing, the ligase involved in the RAIR pathway has not been elucidated. Only two ligases have been found in the *E. coli* genome (225, 226). To find out whether either of these is involved in the RAIR pathway, and to improve RAIR efficiency plasmid pKS02 and pKS03 were designed to overexpress *ligA* and *ligB* respectively. Plasmid pKS04 was designed to overexpress *PolA* gene, which is speculated to play a role in the RAIR pathway (202). The candidate genes in these four plasmids were placed under control of a lacUV5 promoter (Figure 3.3). Additionally, as a negative control empty plasmid pKS05 was designed where no gene was placed under the lacUV5 promoter.

Overexpression of *xthA* improves recombination efficiency

E. coli DH10B strains harboring one of the five overexpression plasmids were made competent to assess recombination efficiency. During preparation of competent cells, 1 mM of IPTG was added to induce the expression of the candidate genes on plasmids pKS01-05. The transformation efficiency of each strain was calculated by transforming with 1 ng of plasmid DNA (pTN003) in triplicate and counting the number of colonies (Supplementary figure 3.1). Wild-type *E. coli* DH10B has the highest transformation efficiency while this is lower for the strains harboring overexpressing plasmids.

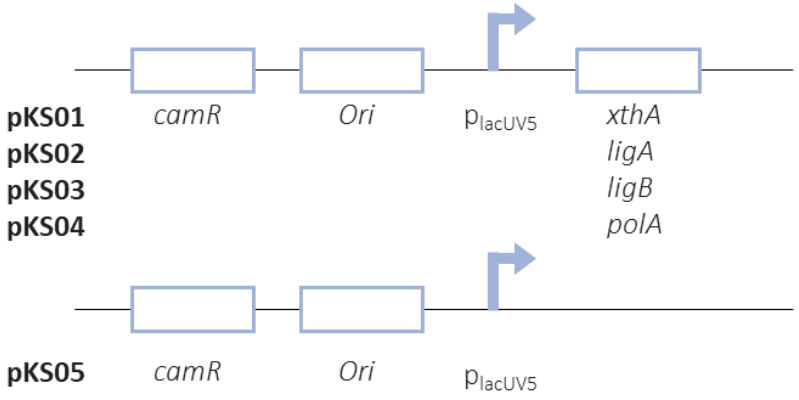


Figure 3.3: Schematic representation of overexpression plasmids. Names of the plasmids on the left, with their respective genes indicated.

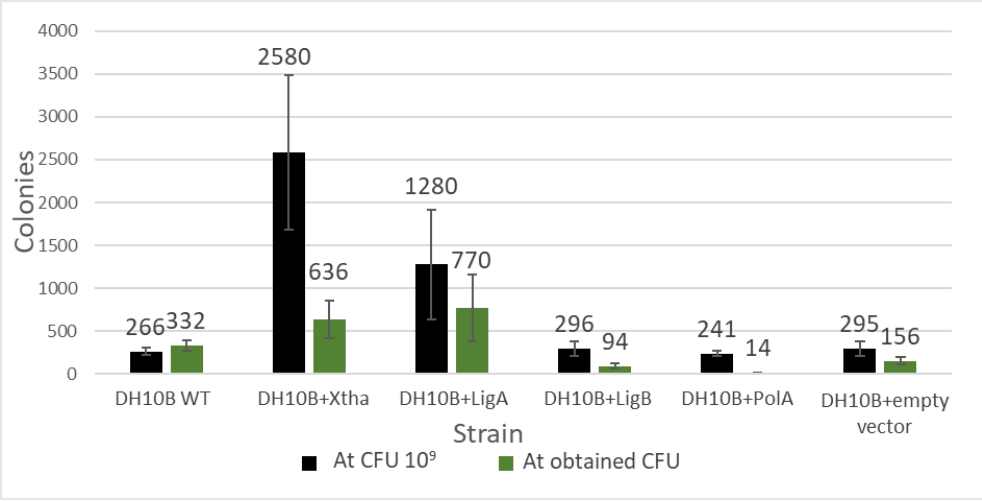


Figure 3.4: Schematic representation of overexpression plasmids. Names of the plasmids on the left, with their respective genes indicated.

The strains harboring the five plasmids were transformed with the four fragments to assess their recombination efficiency. First, the numbers of red and white colonies were counted. For the wild-type DH10B strain on average total 332 colonies were counted, and for the strain harboring the non-expressing control plasmid pKS05, in total 156 colonies were found in average (Figure 3.4). The strains with plasmids pKS01 and pKS02, harboring the Exo III/XthA gene and LigA respectively, performed much better with 636 and 770 colonies on average, respectively. Strains with pKS03 and pKS04 (ligB and polA), did not outperform the wild-type with totals of averaged 94 and 14 transformants, respectively. However, it is difficult to objectively evaluate the direct impact of overexpressed genes as transformation efficiency of DH10B containing any of the overexpression plasmids went down several folds (Supplementary figure 3.1). As we are interested in the efficiency of assembly, and not the transformation efficiency, the colonies of transformants in the assembly experiments were normalized for the transformation efficiency for whole plasmids (Figure 3.3, black bars). The strain with pKS01 (Exo III/XthA) has an approximate 9-fold higher recombination efficiency compared to WT DH10B and the empty plasmid strain. The strain harboring pKS02 (ligA) has an approximately 4-fold increase in recombination efficiency compared to the empty plasmid control. Both strains harboring pKS03 and pKS04 are comparable to the WT and empty vector controls.

Overexpression of both *xthA* ($p<0.05$) and *ligA* ($p<0.1$) increase the recombination efficiency of *E. coli* DH10B. However, when observing the red and white colonies, the strain with pKS02 harboring LigA is much less accurate compared to the wild-type (Figure 3.5). Recombination accuracy of pKS01 is comparable to the WT.

The proposed mechanism for the RAIR pathway utilizes at least three genes, a DNA exonuclease to generate single-stranded homologous 'sticky' ends, a DNA polymerase to fill in the gaps after annealing of the sticky ends, and finally a DNA ligase to ligate the strand of DNA. As both *xthA* and *ligA* overexpression enhanced the recombination efficiency, it was decided to design plasmid pKS06, expressing both genes using the placUV5 promoter (Supplementary figure 3.2). The combined overexpression of *xthA* and *ligA* led to increased assembly efficiency in DH10B. However, overexpression of both genes combined led to a low cloning accuracy (54%), comparable to efficiency for overexpression of *ligA* alone (60%) (Supplementary figure 3.3 & 3.4). This could be due to *ligA* taking part in alternative end-joining mechanism (A-EJ) (227). Until recently it was assumed that *E. coli* cannot join non-homologous ends, however, this was disproved. *E. coli* has been demonstrated to possess an A-EJ mechanism, which utilizes microhomology between two DNA strands to join them (31) LigA has been reported to be involved in this pathway. Hence, overexpression of *ligA* could have led to different recombination between fragments due to A-EH and not homologous recombination. In theory, the plasmid would only need two of the four fragments for the *E. coli* cells to survive (the fragments containing the antibiotic resistance gene and the Ori). Additionally, this mechanism could anneal blunt ends of the fragments without exonuclease activity, which would also lead to incorrect assembly and to loss of red color.

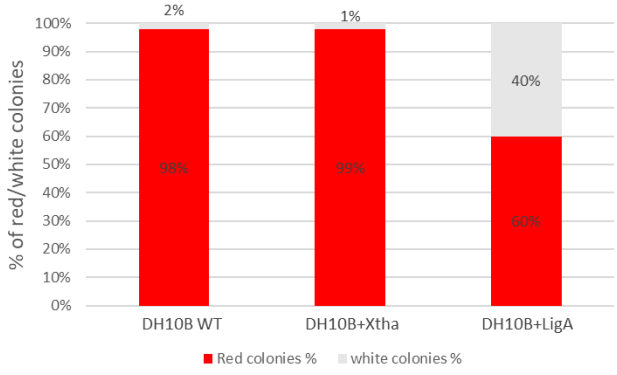


Figure 3.5: Cloning accuracy of WT DH10B and the strains harboring pKS01 and pKS02. The percentage of red colonies in red, and white colonies in gray shaded.

Recombination efficiency increases with homology length

To further assess the possibilities of the RAIR pathway as *in vivo* cloning tool in *E. coli* we set out to test the optimal length of sequence homology between the fragments. Fragments were designed with 20 and 30 bases of homology, and compared with the 50 bp homologous ends that were used previously, for both DH10B WT strain and the strain harboring pKS01 (Figure 3.6). Using homologous ends of 30 bp yields comparable total colony counts to 50 bp homology (1469 and 1572 colonies at normalized transformation efficiency respectively). At the lower end, 20 bp homologous ends lead to an approximate 3-fold decrease in total colonies obtained, with 519 colonies at normalized transformation efficiency. Recombination accuracy at various length of homology was also assessed (Supplementary figure 3.5). The longer the homologous ends the more accurate recombination is. In the strain harboring pKS01, with 50 bp homology, approximately 99 % accuracy can be reached, which goes down to 96 % for 30 bp homology, and to 89 % for 20 bp of homology. For regular cloning activities, these numbers are all high enough to work with.

In a study conducted by Jacubus and Gross (218) the cloning potential of *E. coli* DH5α was tested. Homology lengths ranging from 5bp to 30bp for two fragments DNA assembly were tested. Similar to our study, it was noticed that cloning efficiency was highest with the largest overhang (30 bp) for which 100% of the two-fragment assembled plasmids was corrected, while for the minimally needed 10 bp overhang only 25% successful assemblies was found. However, when fragment number was increased from two to three only 60% of transformants were accurately cloned at 30 bp homology. This is contradicting to our study with DH10B, which 93% of transformants were accurately assembled at 30 bp homology even from four DNA fragments. Another study (23) by Kostylev *et al.* (219), which used the *E. coli* DH5α RAIR system managed to get 100% assembly rate using a four-fragment assembly, and successful assembly of a six-fragment plasmid with 50-75 % reported accuracy.

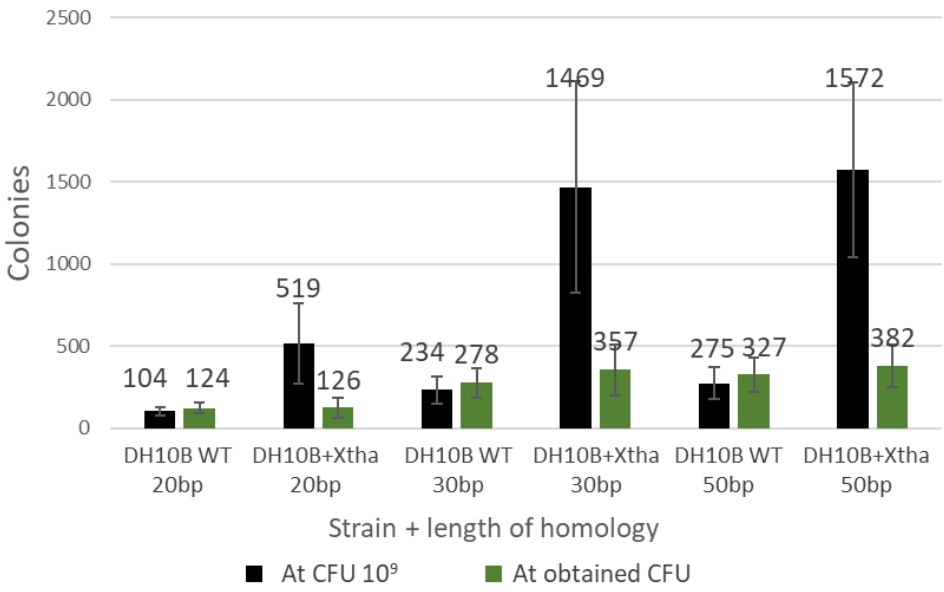


Figure 3.6: Colony counts after recombination of four fragments. Black: normalized for transformation efficiency, green: real counted colony number. Experiment was performed in triplicate.

In another study (228) multiple *E. coli* strains were compared to find the best strain for *in vivo* recombination (Table 3.1). The selected strains were tested using a two-fragment assembly and are compared with a two-fragment assembly in DH10B, and four-fragment assemblies in DH10B and DH10B pKS01. Accuracy of assemblies done by Beyer and colleagues was tested by sequencing 12 colonies (228), whereas in the here presented study we used red colony count as accuracy test. In general, it is observed that 'cloning strains' (K-strains) like TOP10, NEB5α, NEB10b and DH10B have a higher recombination efficiency compared to BL21 (DE3), a popular B-strain generally used for recombinant protein production. Presumably because of the lack of R/M systems in the cloning strains, allowing more foreign DNA to 'survive' in the cell and subsequently recombine.

Table 3.1: Assembly efficiency of multiple *E. coli* strains. Adapted from Beyer et al., (2015)

Strain	Homology length	Number of colonies	CFU	Accuracy
TOP10 (self-prepared)	32 bp	1160	$6x10^6$	11/12
TOP10 (commercial)	32 bp	233	109	10/12
NEB5a (commercial)	32 bp	283	$1\text{-}3x10^9$	11/12
NEB10b (commercial)	32 bp	1073	$1\text{-}3x10^9$	12/12
BL21 (DE3) (commercial)	32 bp	9	$1\text{-}5x10^7$	9/9
JM109 (commercial)	32 bp	92	108	12/12
DH10B (this study) 50bp	**50 bp**	**2085**	**$1.3x10^9$**	**98%**
DH10B (this study) 4-fragment assembly	**50 bp**	**266**	**$1.3x10^9$**	**98%**
DH10B pKS01 (this study) 4-fragment assembly	**50 bp**	**636**	**$2.3x10^8$**	**99%**

Conclusion

Overall, the RAIR pathway in *E. coli* is a promising recombination tool, that can be used for almost any cloning. Compared to Gibson assembly, which utilizes a similar cloning mechanism as the proposed RAIR mechanism, this method is cheaper, and for most labs readily usable. We demonstrate that the number of transformants can be improved by overexpression of XthA. However, the XthA overexpression plasmid decreases transformation efficiency. Also, the use of an expression plasmid and its antibiotic resistance restrict the antibiotics that can be used on the to-be-assembled plasmid and complicate plasmid purification of the assembled plasmids. This could be alleviated by introducing the overexpression of *xthA* from the genome. During this study we attempted to introduce *xthA* genomically at different loci, using the p_{lacUV5} promoter and a constitutive promoter, however, we did not succeed to obtain such strains. It would be of interest to further pursue this path, either by introducing a second copy of *xthA* or by replacing the native *xthA* promoter by a stronger one, to create an all-round *E. coli in vivo* cloning strain.

Supplementary information:

Supplementary table 3. 1: Plasmids used in this study.

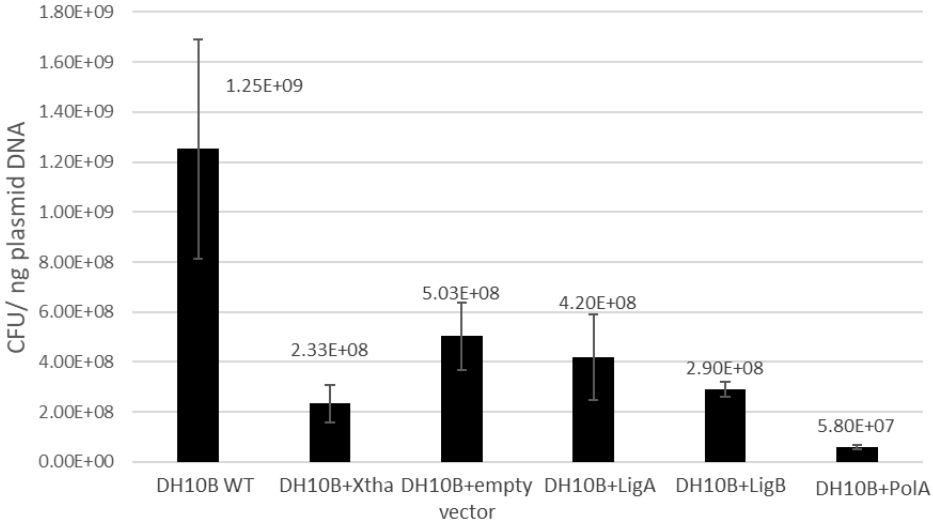
Plasmid	Description	Source
pTN003	Plasmid harbouring RFP used to make recombination fragments	This study
pKS01	Plasmid for overexpression of *xthA*	This study
pKS02	Plasmid for overexpression of *ligA*	This study
pKS03	Plasmid for overexpression of *ligB*	This study
pKS04	Plasmid for overexpression of *polA*	This study
pKS05	Empty plasmid control for overexpression plasmids	This study
pKS06	Plasmid for overexpression of *xthA* and *ligA*	This study

Supplementary table 3.2: Primers used in this study.

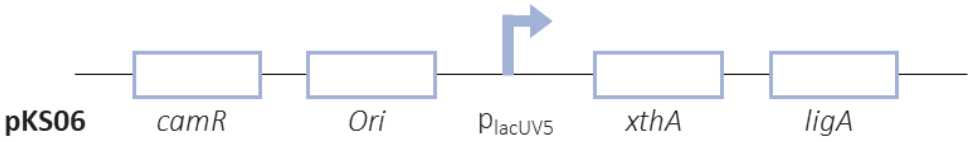
Primer name	Primer sequence	Description
Frag A fw	GGAGTTCTGAGGTCATTACTGG	Construction of fragment A
Frag A rv	CAGCAACGACTCATAGAAAGC	Construction of fragment A
Frag B fw	AAACACCACGTCGACCC	Construction of fragment B
Frag B rv	GAGCTCGCTTGGACTCC	Construction of fragment B
Frag C fw	GGAGTTCTGAGGTCATTACTGG	Construction of fragment C
Frag C rv	AACTCACCGTCTTGCAGG	Construction of fragment C
Frag D fw	GGTGGTGTTGTTACCGTTACC	Construction of fragment D
Frag D rv	GAGCTCGCTTGGACTCC	Construction of fragment D
Placuv5_fw	TTAAAAGAGACAAATTTCATGTCAACGGTACCTGCAGTC	Backbone construction of pKS plasmids; p_{Lacuv5}
Placuv5_rv	GCGCTGCATGCCTATTTGTTTTCTAATTGCGTTGCGCTC	
Ori_fw	AACAAATAGGCATGCAGCG	Backbone construction of pKS plasmids; Ori
Ori_rv	CTGACTTCAGGTGCTACATTTG	
Cat_fw	AATGTAGCACCTGAAGTCAGTAGCACCAGGCGTTTAAGG	Backbone construction of pKS plasmids; *cat*
Cat_rv	AGCAGCTCTCCTGGCTCAAAGAAACAGAAGCCACTGGAGC	
Xtha_fw	TTTGAGCCAGGAGAGCTG	pKS1 construction
Xtha_rv	ATGAAATTTGTCTCTTTTAATATCAACGGC	pKS1 construction
ligA forw	GCTCCAGTGGCTTCTGTTTCTCAGCTACCCAGCAAACG	pKS2 construction
ligA rev	AGACTGCAGGTACCGTTGACATGGAATCAATCGAACAACAACTGAC	pKS2 construction
ligB forw	GCTCCAGTGGCTTCTGTTTCCTAAGGTTCAAAACCTGTGATCTGC	pKS3 construction
ligB rev	AGACTGCAGGTACCGTTGACCCTGACTATCTACAGCGGAGG	pKS3 construction
polA forw	GCTCCAGTGGCTTCTGTTTCTTAGTGCGCCTGATCCC	pKS4 construction
polA rev	AGACTGCAGGTACCGTTGACAACAGGCACGGACATTATGG	pKS4 construction

Supplementary table 3.2 (Cont.): Primers used in this study.

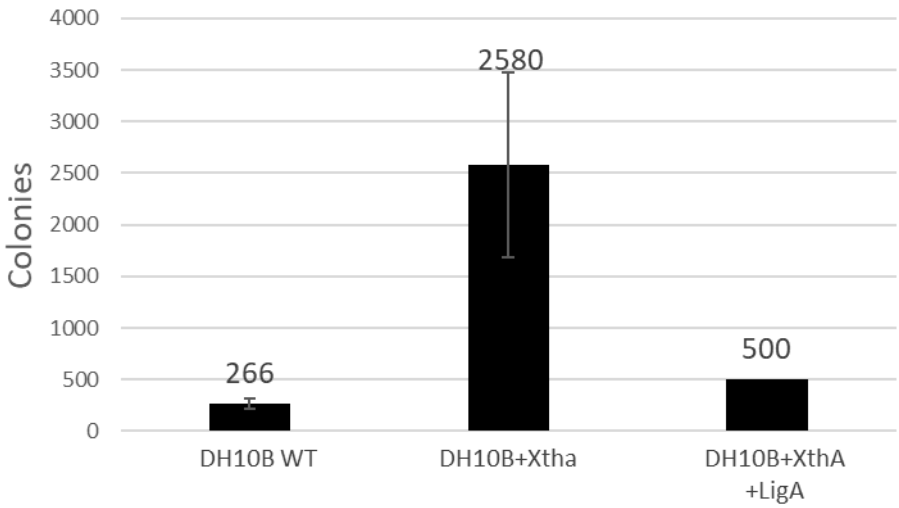
Primer name	Primer sequence	Description
20bp rv D	AGTAATGACCTCAGAACTCC	Primers used to create fragments with different length homology
20bp fw B	CTTTCTATGAGTCGTTGCTG	
20bp rv B	GTAACGGTAACAACACCACC	
20bp fw D	CGGTGAACGCTCTCTACTAG	
30bp ori D	TGTTGTTTGTCGGTGAACG	
30bp rv B	GGAGTCCTGGGTAACGGTAAC	
30bp rv D	TGATAGATCCAGTAATGACCTCAG	
30bp fw B	CAGCTGCGTGCTTTCTATGAG	
Pks6 fw	AGACTGCAGGAGCCGTTGACATGGAATCAATCG AACAACAACTGAC	pKS6 construction
Pks6 rv	GTCAACGGCTCCTGCAGTCTTTTGAGCCAGGAG AGCTG	pKS6 construction



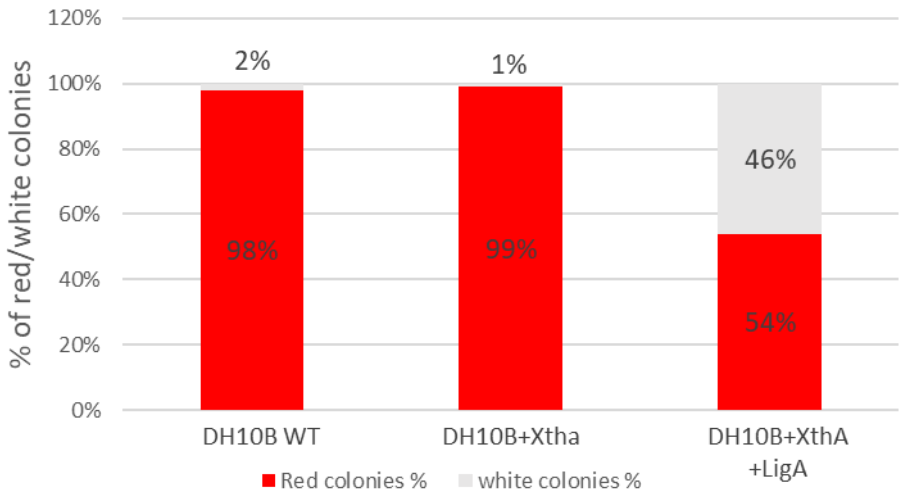
Supplementary figure 3.1: Transformation efficiency in colony forming units (CFU) per ng of pTN003 plasmid DNA for competent cells harboring different overexpression plasmids.



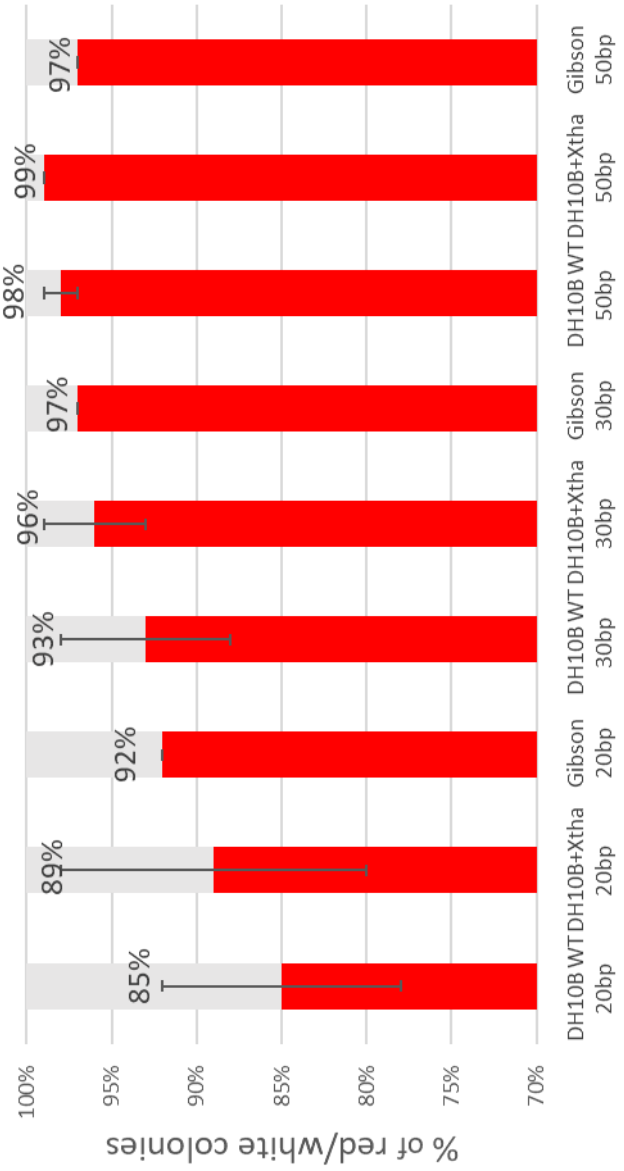
Supplementary figure 3.2: Design of pKS06 harboring *xthA* and *ligA*.



Supplementary figure 3.3: Comparing total number of transformants for combined *xthA* + *ligA* overexpression to reference strains, data for a 4-fragment assembly normalized by transformation efficiency.



Supplementary figure 3.4: Comparing assembly accuracies for combined *xthA* + *ligA* overexpression to reference strains, data for a 4-fragment assembly, the percentage of red colonies in red, and white colonies in light gray.



Supplementary figure 3.1: Assembly accuracy after 4-fragment recombination with varying homology length. the percentage of red colonies in red, and white colonies in light gray. Note that Y-axis does not start at 0.

Chapter 4

Design, construction and optimization of a synthetic RNA polymerase operon in *Escherichia coli*

Joep Houkes[1], Lorenzo Olivi[1], Zacharie Paquet[1], Nico J. Claassens[1], John van der Oost[1]

[1] Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands.

Abstract

Prokaryotic genes encoding functionally related proteins are often clustered in operons. The compact structure of operons allows for co-transcription of the genes, and for co-translation of the polycistronic messenger RNA to the corresponding proteins. This leads to reduced regulatory complexity and enhanced gene expression efficiency, and as such to an overall metabolic benefit for the protein production process in bacteria and archaea. Interestingly, the genes encoding the subunits of one of the most conserved and ubiquitous protein complexes, the RNA polymerase, are not clustered in a single operon. Rather, its genes are scattered in all known prokaryotic genomes, generally integrated in different ribosomal operons. To analyze the impact of this genetic organization on the fitness of *Escherichia coli*, we constructed a bacterial artificial chromosome harboring the genes encoding the RNA polymerase complex in a single operon. Subsequent deletion of the native chromosomal genes led to a reduced growth on minimal medium. However, by using adaptive laboratory evolution the growth rate was restored to wild-type level. Hence, we show that a highly conserved genetic organization of core genes in a bacterium can be reorganized by a combination of design, construction, and optimization, yielding a well-functioning synthetic genetic architecture.

Introduction

Operons were first described in the 1960s by Jacob and Monod as "a group of genes regulated by a single operator" (229, 230). At present, operons are generally defined as clusters of genes that are co-transcribed as a single polycistronic mRNA. As first observed in the lactose (*lac*) operon of *E. coli* (229, 230), subsequent experimental analyses of bacterial operons revealed that the clustered genes often encode proteins (or RNAs) with related functions. Indeed, comparative genomics analyses corroborated that operons are generally composed of functionally related proteins ('guilt by association'), such as enzymes of a metabolic pathway and subunits of a multi-protein complex (161, 231).

Ever since the discovery of the operon organization, the potential evolutionary forces that drive operon formation have been discussed. Several hypotheses have been suggested to explain how operons could potentially contribute to a selective advantage over individual genes: (i) operons contribute to reduction of genome size and to simplification of gene expression control (232), (ii) operons avoid energy loss through appropriate co-transcription and co-translation of functionally-related genes (233), and (iii) operons improve functional horizontal gene transfer (234–236).

In some metabolic pathways and protein complexes, uneven stoichiometries are required. In these cases, it has been demonstrated that differential transcription occurs by using multiple promoters (231, 237), while differential translation of the cistrons within the operons is achieved in multiple ways. The rates of translation initiation can be varied by tuning the strength of the Ribosome Binding Sites (RBS) and the mRNA secondary structure around the start codon, as well as by translation elongation, through modulating the codon bias (161, 238–240).

It is interesting to note that, despite being one of the most conserved protein complexes in the three domains of life, the genes coding for the subunits of the prokaryotic DNA-dependent RNA polymerase (RNAP) complex are not clustered in a single operon. At present, not a single prokaryotic genome is known in which the RNAP genes are organized as a single operon. Instead, the genes are spread throughout the genome at different loci. However, in bacteria, in archaea and even in the genomes of chloroplasts in photosynthetic eukaryotes (algae and plants), the RNAP genes are typically co-localized in distinct operons with ribosomal genes (Figure 4.1).

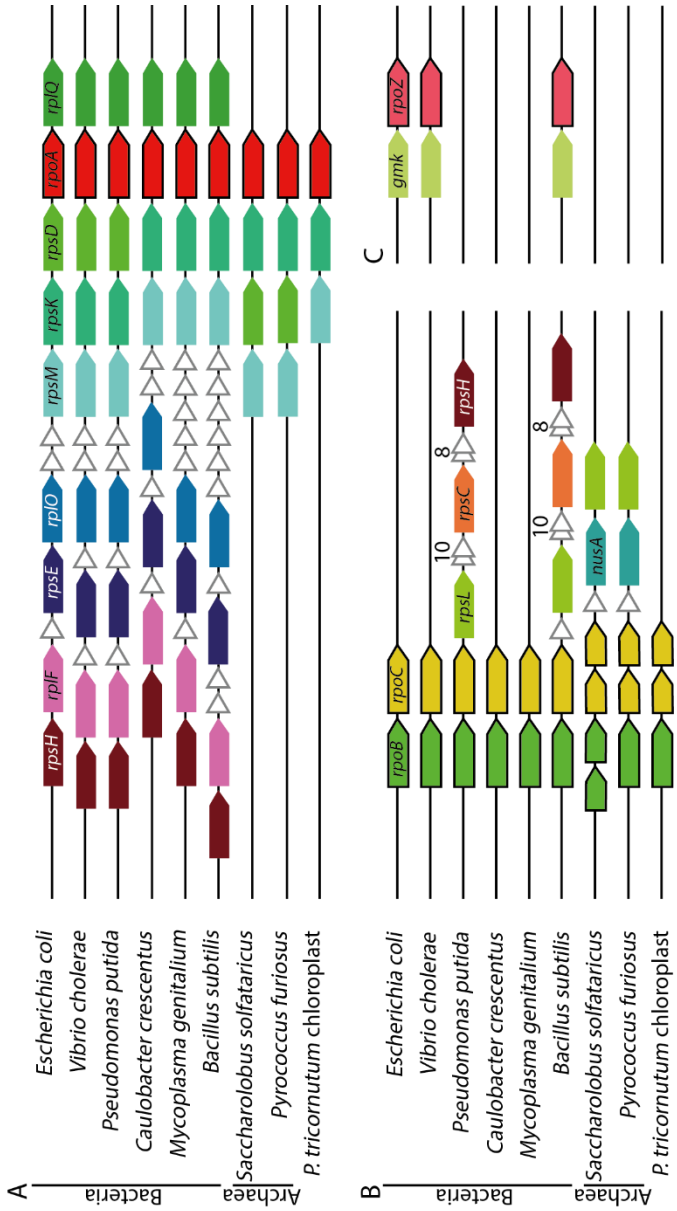


Figure 4.1: Syntheny of genes encoding core subunits of DNA-dependent RNA polymerase (RNAP). (A) α subunit encoded by *rpoA* , (B) β and β' subunits encoded by *rpoB* and *rpoC* and (C) ω subunit encoded by *rpoZ* across selected bacterial and archaeal model species, as well as the chloroplast of the microalgae *Phaeodactylum tricornutum*. Same color genes indicate conserved clustering across species. Genes indicated by white triangles are not conserved. *Rps* and *rpl* genes encode ribosomal proteins (16S and 23S subunits, respectively), *nusA* encodes a transcription termination/anti-termination protein and *gmk* encodes a guanylate kinase. Figure generated with data from STRING (string-db.org).

The bacterial RNAP core complex consists of five subunits: two copies of the α subunits and single copies of the β, β' and ω subunits (241). The *rpoA*-encoded α subunit dimer plays a key role in assembly of the RNAP complex, acting as a scaffold for assembling the β and β' subunits (242). Furthermore, the α subunit interacts with certain transcription factors to regulate transcription. The *rpoA* gene of *E. coli* and many other bacteria is co-located in an operon harboring ribosomal genes *rpsM*, *rpsK*, *rpsD* and *rplQ* (Figure 4.1A, (243, 244).

The *rpoB* and *rpoC* genes encode the structurally related β and β' subunits, respectively, that make up the hetero-dimeric core of the RNAP complex, of which the β' subunit harbors the catalytic polymerase center (245, 246). Most likely the *rpoB* and *rpoC* genes are the result of a gene duplication (245, 246). In line with this model, a single orthologous RNAP gene still exists in some phages, probably encoding a homo-dimer (246). The bacterial *rpoB* and *rpoC* genes are always clustered, often overlapping, and occasionally fused (247). In addition, a functional synthetic *rpoB-rpoC* fusion protein has been reported (248). In several bacteria the *rpoB* and *rpoC* genes reside in an operon with the ribosomal genes *rplK*, *rplA*, *rplJ* and *rplL* (Figure 4.1B, 4.2B). In *E. coli*, this operon has a complex regulation: involvement of four different promoters, regulation by multiple transcription factors and a transcriptional attenuator terminating approximately 70% of transcription just upstream of *rpoB* (Figure 4.2B) (249).

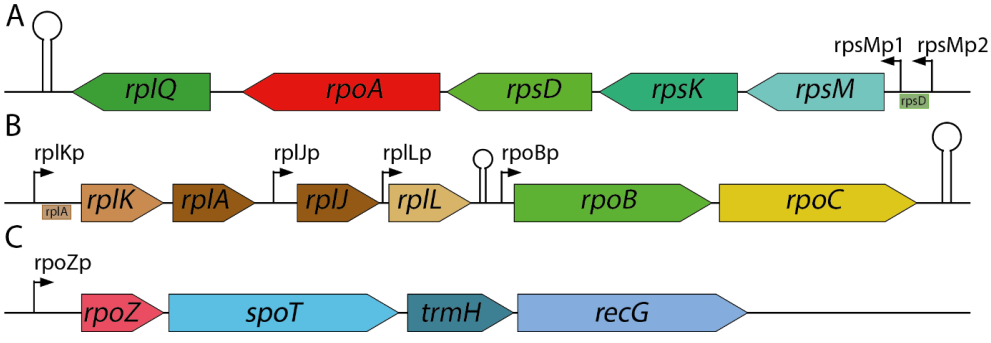


Figure 4.2. Operons of the RNAP subunits in *E. coli*. (A) operon harbouring the *rpoA* subunit of RNAP, with the ribosomal genes *rpsM*, *rpsK*, *rpsD* and *rplQ* (243, 244). (B) operon harboring the *rpoB* and *RpoC* subunits of RNAP, with the ribosomal genes *rplK*, *rplA*, *rplJ* and *rplL*. An attenuator between *rplL* and *rpoB* halts approximately 70% of transcription (249). (C) operon harboring the *rpoZ* subunit of RNAP, with *spoT*, involved in stringent stress respons (250–253), *trmH*, a tRNA methyltransferase (254) and *recG* involved in DNA repair and DNA recombination (255).

The only non-essential subunit of the bacterial RNAP core is the ω subunit, which upon knockout leads to growth retardation, but not to cell death (241, 256). The ω subunit is encoded by *rpoZ,* which in *E. coli* resides in an operon with *trmH*, *recG* and *spoT* (Figure 4.2C). TrmH is a tRNA methyltransferase, and RecG is an ATP-dependent DNA helicase which plays a critical role in DNA repair and DNA recombination (254, 255). SpoT is responsible for the synthesis and degradation of ppGpp, the effector molecule for stringent response, which enables bacterial cells to react to stress conditions by altering expression of many genes (250–253). Interestingly, the primary ppGpp binding site of the *E. coli* RNAP is located at the interface of the β' and the ω-subunits. The ω subunit plays a role in regulating ppGpp-dependent control of RNAP activity (257), and it has been reported to act as a chaperone for the RNAP subunits (258). The ω subunit binds mainly to the β' subunit, close to the active polymerase site, indicating a role in controlling the RNAP catalytic activity (245).

The RNAP $\alpha_2\beta\beta'\omega$ core forms a holoenzyme with a σ factor to initiate transcription. Bacteria have several different σ factor, each of which is responsible for transcription of a specific subset of genes (259). The housekeeping σ factor in *E. coli* is σ70, encoded by *rpoD* which controls a large number of promoters, and regulates gene expression during 'normal' growth. Six additional σ factors in *E. coli* each control the expression of a particular subset of genes, active during specific environmental conditions (260). Regulation of σ factor expression is very complex, as they are very condition-dependent, unlike the RNAP-core subunits, which are always present.

When comparing the amino acid sequences, the subunit composition, the overall structure, the molecular mechanism and, to some extent, the genomic organization of RNAPs in all domains of life, it becomes apparent they all derive from a common ancestor (261). Both the 13-subunit archaeal and the 12-subunit eukaryotic RNAP complexes contain orthologues of the bacterial RNAP β-, β'-, α- and ω-subunits. This reflects a common evolutionary history, in which the *rpoB/rpoC* gene pair encodes the catalytic β/β' hetero-dimer of an ancient RNAP variant. At a later stage in the RNAP evolution, the catalytic core was most likely supplemented with the *rpoA*-encoded α-subunit dimer, and the *rpoZ*-encoded regulatory ω-subunit. The archaeal RNAP core resembles the bacterial RNAP complex, with some additional genes encoding auxiliary subunits (261). The three basic eukaryotic RNAPs (Pol I, II, III) and the 2 plant-specific RNAPs (Pol IV, V) are all derived from the archaeal RNAP, each with specific sets of auxiliary subunits (261).

In this study, we set out to use a synthetic biology approach to test if this evolutionary-conserved scattering of the RNAP genes can be reorganized into a single operon, and how such a different architecture would impact cellular fitness. For this, we designed and constructed an operon of the RNAP core genes in *E. coli*. This synthetic operon was introduced on a bacterial artificial chromosome (BAC) and expressed in *E. coli*. Subsequently, native RNA polymerase genes on the *E. coli* chromosome were knocked out, to assess the function of the RNA polymerase operon. This led to a slightly lower growth rate on rich medium compared to wild-type *E. coli*, but, to almost complete loss of growth on minimal medium. However, by adaptive laboratory evolution (ALE) on minimal medium, we could restore growth and even improve the yield of the strain with the synthetic RNAP operon. Overall, this study demonstrates that an evolutionary-conserved operon organization of a core protein complex can be successfully reorganized, suggesting plasticity of genome organization and regulation.

Material and Methods

Strains and growth conditions

The *E. coli* DH10B strain (Invitrogen, supplementary table 4.2) was used for expression of the synthetic operon. This strain was cultured at 37 °C in LB medium (10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract), 2xYP medium (16 g/L peptone, 10 g/L yeast extract) or minimal M9+glucose medium (11.28 g/L 5x M9 salts, 0.12 g/L $MgSO_4$, 5.5 mg/L $CaCl_2$, 3.6 g/l glucose supplemented with 0.1 mL 1000x trace elements solution (50 g/L EDTA, 8.3 g/L $FeCl_3$-$6H_2O$, 0.84 g/L $ZnCl_2$, 0.13 g/L $CuCl_2$-$2H_2O$, 0.1 g/L $CoCl_2$-$2H_2O$, 0.1 g/L H_3BO_3, 16 mg/L $MnCl_2$-$4H_2O$) and 0.5mM leucine) at 180 rpm or on LB-agar plates containing 1.5% (w/v) agar (Oxoid) unless stated otherwise. The LB medium was supplemented with different antibiotics (LB/Ab) when appropriate, to final concentrations of 20 µg/mL kanamycin (Carl Roth) (Kan20), or 30 µg/mL chloramphenicol (Sigma Aldrich) (Cam30).

Yeast strain *Saccharomyces cerevisiae* CEN.PK2-1D (*Euroscarf,* supplementary table 4.2) strain was used for synthetic operon construction. This strain was cultured at 30°C in 10 mL YPD medium (20 g/L peptone, 10 g/L yeast extract, 20 g/L glucose) or in SC medium (1.9 g/L nitrogen base without amino acids, 5 g/L ammonium sulphate, 20 g/L glucose, 2g/L drop-out mix, appropriate auxotrophic marker (76 mg/L uracil, 380 mg/L leucine, 76 mg/L histidine, 76 mg/L tryptophan)) in 50 mL tubes at 180 rpm.

Preparation of electrocompetent cells

E. coli cells were made electrocompetent by culturing at 37 °C in 2xYP (supplemented with 0.01 M L-arabinose for recombination experiments), typically in 50 mL medium in 250 mL Erlenmeyer flasks, at 200 rpm until an OD600 nm of 0.4 was reached. The cells were then cooled rapidly on ice, and subsequently washed once with 1 culture volume of ice-cold ddH2O and twice with 0.5 culture volumes of ice-cold 10% glycerol. Finally, the cells were suspended in ice-cold 10% (vol/vol) glycerol to a final volume of 200 µL for each 50 mL of initial culture volume.

Electroporation of 20 µL competent cells was performed in ice-cold 2 mm electroporation cuvettes at 2500 V, 200 Ω and 25 µF (ECM 630 BTX). Immediately after electroporation, cell recovery was performed in 1 mL LB medium at 37 °C, 750 rpm for 1 h when plasmids were transformed. During recombination experiments, recovery was performed at 30 °C, 750 rpm for 2.5 h. After recovery, the cells were plated on LB/Ab agar plates. Single colonies were picked and re-suspended in 50 µL of ddH_2O for colony PCR and used to inoculate 10 mL of LB/Ab for overnight incubation and subsequent isolation of plasmids, using GeneJET Plasmid MiniPrep Kit (Thermo Fisher Scientific). The provided protocol was adjusted by initially centrifuging the cell cultures at 4700 rpm for 10 min, and introducing of an incubation (2 min) at room temperature after addition of Elution Buffer or ddH_2O (warmed to 70 °C).

Transformations of *S. cerevisiae* were performed by chemical transformation. Cells were plated from glycerol stock on YPD, a single colony was picked for overnight culturing in YPD (typically in 10 mL medium in 50 mL tubes, at 180 rpm). The culture was diluted to OD600nm 0.4 in YPD and incubated at 30°C, 200 rpm for 3 hours (typically in 50 mL medium in 250 mL Erlenmeyer flasks). The cells were then washed with 0.5 culture volume of ddH_2O. Cells were resuspended in ddH_2O to a final volume of 1 mL, aliquoted into 100 µL volumes and stored at 4°C for up to a week. To a 100 µL cell suspension, 350 µL of a transformation mix (consisting of 240 µL PEG-3350, 36 µL 1 M LiOAc, 50 µL 2 mg/mL denatured salmon sperm DNA, and 34 µL of DNA mix containing 500 ng of DNA fragments and 1 µg of backbone plasmid) was added. Next, the cells/transformation mix was heat-shocked at 42 °C for 40 min. The cells were resuspended in 1 mL YPD and 500 µL was used to inoculate 5 mL YPD for overnight recovery, to boost the recombination efficiency. The remaining 500 µL was plated on SC-agar plates with the appropriate auxotrophy markers. The success of the transformation was assessed by colony PCR and single colonies were used to inoculate 10 mL SC with the appropriate auxotrophy markers. For plasmid extraction, the

cells were resuspended in 200 µL GeneJET Plasmid MiniPrep (ThermoFisher Scientific) resuspension buffer supplemented with 3 µL 1000 U/mL Zymolase (Amsbio) and incubated for 30 min at 37 °C to digest the cell walls. The rest of the extraction was performed according to the MiniPrep protocol.

Plasmid construction

All PCR reactions for cloning purposes were performed using Q5® High Fidelity 2X Master Mix (New England Biolabs). The reactions mixtures were prepared using 1 ng of template, 25 µL Q5® High Fidelity 2X Master Mix, primers to a final concentration of 500 nM, and ddH_2O to a final volume of 50 µL (primers are listed in supplementary table 4.4). Amplification products were run on 0.7% agarose gels stained with SYBR Safe DNA Gel Stain (Invitrogen). The bands of interest were then excided, and the DNA purified using Zymoclean Gel Recovery kit (Zymo Research). Cloning was done using HiFi assembly (New England Biolabs) according to manufacturer protocol. Overhangs for HiFi assembly were added as 5' extensions of PCR-primers (list of plasmids in supplementary table 4.3, list of primers in supplementary table 4.4).

Preparation of knock-out strains

Cells were transformed with a linear knock-out fragment with 50 bp recombination flanks harboring a chloramphenicol resistance marker flanked by mutant *lox66* and *lox72* sites that upon recombination are not recognized anymore by Cre recombinase (262). After recombination, single colonies were picked, streaked on LB/Ab agar plates supplemented with isopropyl-β-D-thiogalactopyranoside (IPTG) (Fisher Scientific, catalogue) to a final concentration of 0.5 mM, and then re-suspended in ddH_2O to perform colony PCR to determine whether appropriate recombination occurred. IPTG-induced expression of Cre recombinase from the BAC vector, generating mixed colonies with chloramphenicol resistant (CmR) and sensitive (CmS) cells. These mixed colonies were re-streaked on LB/Ab plates. Single colonies were picked from these plates and re-streaked on LB agar and in LB agar/Cam30. CmS colonies were picked and re-suspended in ddH_2O to perform colony PCR, to confirm the successful excision of the chloramphenicol resistance gene.

Growth assays of knock-out strain and data analysis

After knock-out was confirmed by colony PCR, growth assays were performed to assess growth rates of the mutant strains. Precultures were prepared on LB for each strain and for wild-type DH10B. Precultures were washed 3 times with ddH_2O and diluted to OD600 0.1.

Next, 15 µL of diluted preculture of each strain and 135 µL of LB or M9+Glucose was transferred in a 96-well plate. The wells were covered with 50 µL of mineral oil (Bio-Rad), to avoid evaporation during the experiment. The plate was then incubated at 37 °C in a Biotek ELx800 absorbance microplate reader (Fisher Scientific). The provided reader control software, Gen5, was used to set a measuring protocol consisting of a cycle of 5 min of linear shaking followed by absorbance measurement at 600 nm, for at least 24 h. The data were then exported to an Excel spreadsheet. An in-house MatLab script was used to process the data, yielding strain-specific growth graphs and doubling times.

Sequencing and analysis

For genomic sequencing analysis DNA was isolated and sequenced using Illumina NovaSeq paired end 150bp. Mutation analysis was done using BreSeq (263) using the DH10B genome (NCBI ref.: NC_010473.1) as reference. In addition, Genious Prime was used to map the sequencing results to a DH10B reference genome and to validate the genomic knockouts of the four RNAP genes.

Results and discussion

Designing and building a synthetic RNAP operon

We rationally designed a synthetic RNAP operon in the order *rpoABCZ*. First, *rpoA* was introduced by Gibson assembly on bacterial artificial chromosomes (BACs), controlled by a few different promoters and RBS. One variant contained the native *rpoA* promoter, i.e. the rpsMp2 promoter of the operon, and the native *rpoA* RBS. In addition, *rpoA* was inserted downstream three constitutive promoters of different strength (weak, moderate, strong) (264), that each were combined with one of 5 RBS in a linearly increasing strength range (20, 40, 60, 80 and 98% of the predicted maximum strength), as designed by EMOPEC (265). For each combination, the native *rpoA* gene on the *E. coli* chromosome was knocked out, and a comparative growth assay on LB medium determined the best performing combination: strong constitutive promoter and RBS80 (supplementary table 4.1). Interestingly, knock-out of the native *rpoA* gene was successful only for 7 out of 24 promoter-RBS combinations, strongly suggesting that there is a certain range of *rpoA* expression levels that allows for *E. coli* survival. To allow for easy addition and selection of subsequent subunits, we decided to move the marker directly downstream of the operon. The *kanR* resistance marker was removed from the BAC-*rpoA* using the λ-red system and *cre*

recombinase. This created an addictive plasmid bale to be propagated because of the presence of the essential *rpoA* gene. We then introduced *rpoB* with 6 different RBSs (native and 5 synthetic variants) and a *kanR* gene directly downstream of *rpoA*. Deletion of chromosomal *rpoB* gene and subsequent growth assays of the 6 RBS variants, showed that the native RBS associated with *rpoB* on the BAC, resulted in the fastest growth. We aimed to continue this approach for *rpoC* as well, but several attempts to introduce *rpoC* in the operon on the BAC were not successful. To introduce *rpoC* we aimed to switch the antibiotic resistance gene in the BAC to *tetR*, to select for BAC-*rpoABC* after transformation into the double knockout strain. Unexpectedly we did not obtain any transformants harboring *rpoABC.* This prevented us from following the planned approach, to properly introduce, delete and optimize expression for each RNAP gene one-by-one.

Therefore, an alternative approach was designed. As we could not continue with the one-by-one optimization of the *rpo* genes, we decided to assemble the operon at once. Making a combinatorial library of the synthetic *rpo* operon would lead to a large collection of *E. coli* strains from which all genomic *rpo* genes would first need to be deleted and confirmed, leading to a major experimental effort. Hence, we decided to use the previously identified well performing strong promoter in combination with RBS80 for *rpoA*, and the native RBS variant for *rpoB.* In addition, without prior knowledge, we tested the RBS80 variant upstream *rpoC* as well as upstream *rpoZ*. Downstream rpoZ, we included an in-house designed synthetic terminator consisting of a stem-loop

and a T-stretch (5'-ccccgcttcggcggggttttttt) (Figure 4.3). To efficiently assemble all these parts in the relatively large BAC construct at once, we chose to further construct the BAC in *Saccharomyces cerevisiae* because of its highly efficient recombination system. For that purpose, a BAC-yeast artificial chromosome (BAC-YAC) shuttle vector was constructed. First, we PCR amplified the bacterial replication system (*sopA, sopB, sopC* and *repE*) from a BAC (pBeloBAC11 (78)), as well as the yeast centromere region from a YAC (pHLUM (266)) with a *his3* and a *ura3* selection marker. Furthermore, the RNAP genes *rpoA*, *rpoB*, *rpoC* and *rpoZ* (and the aforementioned RBS variants) were PCR-amplified from *E. coli* DH10B. To allow for eventually knocking out the native *rpo* genes, the genes encoding the λ-red system (*gam*, *bet*, *exo*) and the Cre recombinase (*cre*) were PCR amplified by using plasmid pSC020 (267) as a template. All PCR amplifications were carried out using extended primers that generated specific 50 base pair overhangs to allow for efficient homologous recombination in *S. cerevisiae*. Next, we transformed *S. cerevisiae* CEN.PK2-1D (histidine, leucine, tryptophan, uracil auxotroph) with the 8 fragments as described above

for homologous recombination and selected for correct assemblies using medium lacking histidine and uracil (Figure 4.3). The resulting colonies were demonstrated by PCR to contain the designed BAC-YAC clone. Two of these colonies were used for plasmid isolation and transformation into *E. coli* DH10B. From two *E. coli* DH10B transformants, the sequence of the obtained BAC-YAC constructs was analyzed and confirmed to be correct.

Knockout strategy

After transformation of *E. coli* DH10B with the BAC-YAC shuttle vector harboring the RNAP operon, the native genes were knocked out using λ-red recombination (Figure 4.3). For this, a repair template containing a chloramphenicol resistance gene flanked by *lox66* and *lox72* sites was used (262). The repair template was PCR-amplified using primers harboring overhangs homologous to the knockout location. Using this approach, the chromosomal *rpoA, rpoB-rpoC* and *rpoZ* genes of the BAC-YAC-containing *E. coli* strain were deleted consecutively (Figure 4.3). The successful genomic deletions were confirmed initially by PCR, and finally by genome sequencing. This confirmed that the synthetic operon could fully replace the scattered genomic *rpo* genes. To assess the growth of this newly created strain, named strain JH10B, growth assays were performed on rich medium (LB) and minimal medium (M9+glucose), respectively (Supplementary figure 4.1). It was found that on rich medium the growth rate of JH10B was slightly lower (reduced growth rate approximately 7%) compared to growth of the wild-type *E. coli* DH10B. On minimal medium, however, no growth was observed within the first two days for the engineered strain harboring the RNAP operon (Supplementary figure 4.1).

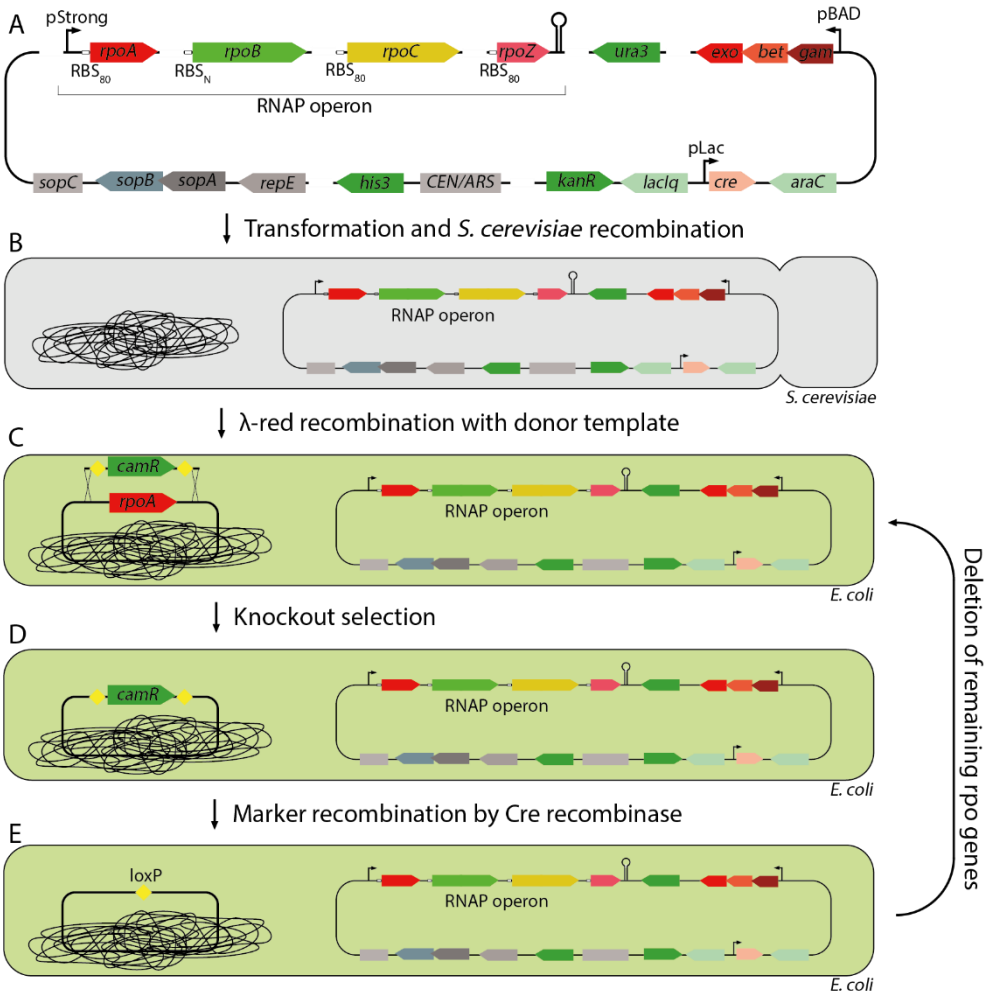


Figure 4.3: Design and construction of RNAP operon on BAC-YAC and knockout of native genes. (A) Design of BAC and parts utilized for assembly. (B) Following this, yeast recombination, isolation, and transformation into *E. coli* DH10B. (C) Knockouts are performed using λ-red recombination, (D) after which selection occurs using the antibiotic resistance marker. (E) Finally, the antibiotic resistance marker is recombined using Cre recombinase, leaving a dysfunctional *loxP* scar. Steps C-E are repeated for remaining *rpo* genes.

Evolutionary optimization

In an attempt to recover the ability of strain JH10B to grow on minimal medium, we decided to perform adaptive laboratory evolution (ALE) on M9+glucose. For this, two colonies (biological replicates a & b) were randomly selected and grown in 10 mL M9+glucose until the OD600 was at least 0.4. After this, 10 µL (0.1%) was transferred to a fresh tube with 10 mL M9+glucose, and after each passage a sample was taken for storage. After the first passage the obtained strain JH10B-ALE-1 already started growing on M9+glucose, and after 12 passages (strain JH10B-ALE-12) a plate reader experiment was done to assess growth of selected generations of the adapted strains (JH10B-ALE-1,-2,-3,-4,-8,-12) in minimal and rich medium, compared to the wild-type strain (Figure 4.4, Supplementary figure 4.2). Although the lag-phase of the wild-type is longer than the lag phases of the ALE strains, the growth rate of the evolved strains is lower (up to 23%). Already after one round of ALE, the doubling time of both biological ALE-1 replicates (strains JH10B-ALE-1a/b; 2.2/1.8 hrs respectively) was comparable to the wild-type strain (1.6 hrs), whereas after the second round of ALE the evolved strain (JH10B-ALE-2b) grows faster than the wild-type on M9+glucose medium (1.4 hrs). Additionally, the yield (final OD600) of some of the evolved strains (JH10B-ALE3b; 1.39) are substantially higher compared to the yield of the wild-type strain (0.66) (Supplementary figure 4.2). While this experiment enhanced the growth rate of JH10B on M9+glucose, growth on LB did decrease (approximately 42% for JH10B-ALEa and 35% for JH10B-ALEb) during the experiment (Supplementary figure 4.3).

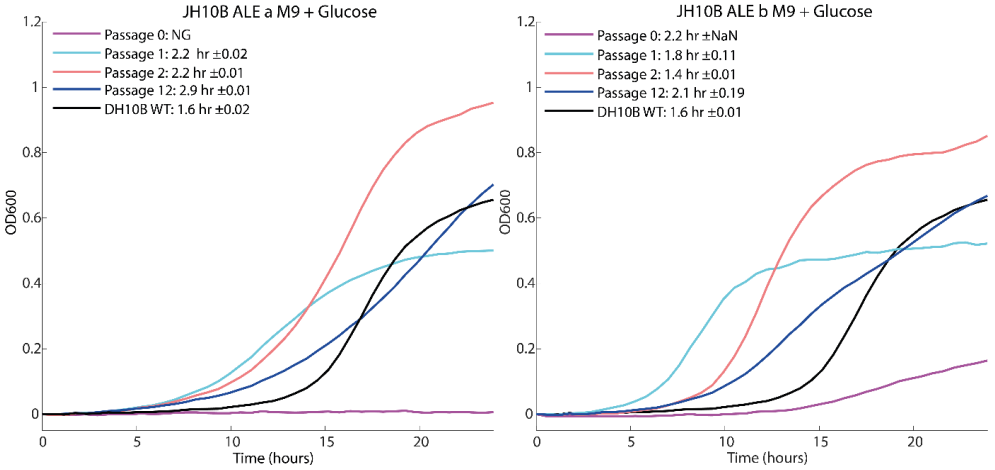


Figure 4.4: Growth assays of *E. coli* strains harboring the RNAP operon after ALE on M9 medium with glucose. Representation of each line shown in figure legend. Doubling time and standard deviation of 6 replicates are indicated.

Genome sequencing and mutational analysis

Whole-genome sequencing was performed on both biological replicates of the strains obtained after one adaptation cycle (JH10B-ALE1a and JH10B-ALE1b), with the wild-type (passage 0) as control. In total one unique mutation was found in *rpoB* in JH10B-ALE1a (T563P) and one unique mutation in *rpoC* in JH10B-ALE1b (128LDMPL duplication), compared to the parental strain. Another mutation in *rpoA* (Y68C) was found not only in both ALE1 replicates, but also in passage 0, before ALE (Table 4.1). Sequencing results indicate that this mutation appeared either during PCR amplification or during yeast recombination/replication of the fragment. Visual inspection of the crystal structure indicates that the duplication in *rpoC* is located at the interface of the DNA strand. The mutation in *rpoA* in both replicates localizes in a loop close to the binding site with the β and β' subunits.

Interestingly, but perhaps not surprisingly, during previously performed ALE experiments with *E. coli*, mutations are frequently found in the genes encoding the RNAP subunits (268–273). In one of these studies it has been concluded that mutations in the RNAP complex can generally satisfy selection of enhanced growth rates under many conditions (273). Being the most central transcriptional regulatory hub, many amino acid substitutions at relevant sites in the RNAP complex can potentially lead to differences in the host's transcriptional profile (274, 275). In one study, 80% of *E. coli* MG1655 strains that were adapted to a minimal medium with glycerol as carbon source, appeared to harbor mutations in the *rpoC* gene (272). These *rpoC* mutations appeared to lead to a 60% increase in growth rate in glycerol minimal medium after reintroduction to MG1655 while simultaneously changing the expression pattern of 900-1200 (20–27%) of its genes. At the time of writing, in ALEdb (www.ALEdb.org), a web-based platform that reports on published ALE-acquired mutations for *E. coli* contains 21738 unique mutations. Of these, 132 were found in *rpoB*, 109 in *rpoC* and 47 in *rpoA* and none in rpoZ, for a total of 288 unique mutations, 1.32% of all unique mutations from the database, while these genes make up 0.21% of the *e coli* genome (276). The mutation found in *rpoB* in this study (T563P) has been found before during ALE on minimal medium, this mutation could confer rifampicin resistance (273).

Table 4.1: mutations in *E. coli* strains harbouring RNAP operon after ALE experiment

Gene	Passage 0	Passage 1	Passage 2	Mutation & function
Biological replicate a				
intergenic	A	A	G	Downstream of putative HNH nuclease *yajD*
rpoA (in operon)	A (WT=G)	A	A	Y68C
rpoB (in operon)	A	C	C	T563P substitution
Biological replicate b				
yidZ	T (WT=C)	T	T	R297W HTH-type transcriptional regulator Involved in anaerobic NO protection
ilvE	C	C	T	A259V Aminotransferase (leucine, isoleucine, valine, phenylalanine)
lptF	T	T	C	L311P translocation of LPS from the inner to the outer membrane
rpoA (in operon)	A (WT=G)	A	A	Y68C
rpoC (in operon)		TCGATAT GCCGCTGC	TCGATAT GCCGCTGC	Duplication of LDMPL at position 128

Although selection for enhanced growth under specific conditions very often leads to mutations in RNAP genes, this often comes with a cost. While the fitness increases in the selected environments, it decreases in different environments (272, 277, 278). In the present study, ALE on M9 glucose led to faster growth in this condition (Figure 4.4), but to reduced growth on LB medium (Supplementary figure 4.3). The obtained amino acid substitutions can have different effects on the subunit, and on the RNAP complex. Some studied mutations in *rpoC* have been shown to decrease open complex half-life (272, 279), which affects the transcription initiation and elongation activity (280). These studies also show that down-

regulated genes often have promoters with stress-related σ factors, whereas up-regulated genes rather tend to have growth-related σ factors (279). It is not exactly known how mutations in RNAP genes modulate the expression of genes that lead to certain phenotypes. However, the RNAP complex can be regarded as the ultimate transcriptional regulator, allocating the cellular resources to the specific molecular functions.

Lessons for synthetic genome re-organization and modularization

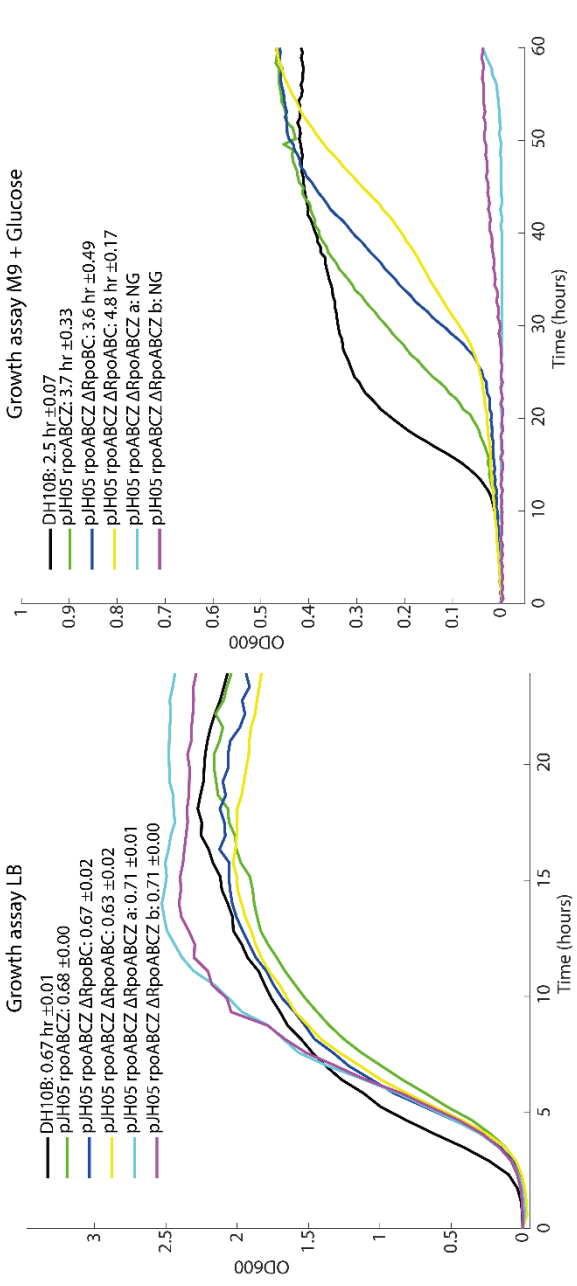
Based on the here presented successful transplantation of separate chromosomal genes to a fully functional BAC-based RNAP operon, we conclude that, at least under the tested conditions, the ubiquitous 'coupling' of the bacterial and archaeal RNA polymerase genes with the ribosomal genes is not essential for life. This is an encouraging result for the design and construction of synthetic genomes, based on partial rational combinations of components. Some attempts to construct rearranged synthetic genomes have already been made, most notably in a minimized *Mycoplasma* species JCVI-Syn3.0 (281) and in *S. cerevisiae* (282). In *S. cerevisiae*, a core set of 13 glycolysis genes, which in nature are scattered across the yeast chromosomes, were expressed from one chromosomal locus using the native promoters, after which the native genes were deleted. This led to a strain which was phenotypically similar to the wild-type strain. This demonstrated that co-localization of genes is feasible even for a eukaryotic species with multiple chromosomes. In the prokaryotic minimal cell JCVI-Syn3.0, genome reorganization towards clustering of functionally related genes was attempted as well. First, genes were split in seven different categories for co-localization (e.g., DNA repair, transcription, translation, glycolysis). Then the genome was split in eight segments, and for each segment separately all genes belonging to one category were co-localized within this segment. However, this meant that some original operon structures had to be rearranged. Hereto, a relatively random approach was followed by which promoters/RBS were manually selected without any optimization to regulate the reorganized genes. This highly randomized approach still led, maybe surprisingly, to viable cells after modularization of one out of eight segments. However, the other seven reorganized segments did not lead to viable cells. The latter result strongly suggests that a rational approach should probably be combined with a random/combinatorial approach as described in this study, to allow for selecting appropriate combinations of promoter and RBSs.

In conclusion, the current study reveals how modularization in bacteria can be performed via a step-wise introduction of genes with synthetic control elements (promoters, RBSs) and subsequent deletion of the native genes. This approach could be the basis to modularize larger parts of the genome. After introduction of a synthetic gene or operon with a small library of

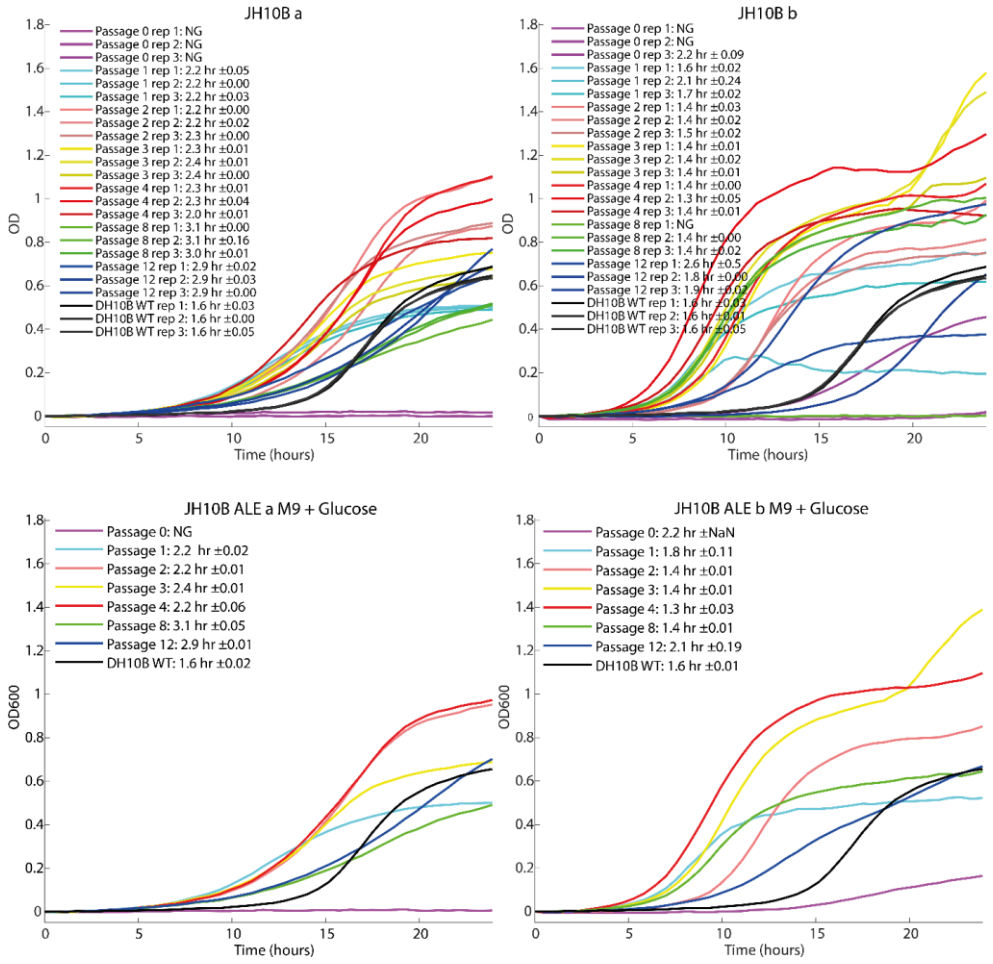
synthetic promoters or RBS, the native gene (s) can be deleted to identify viable clones, which could optionally be further improved by ALE. This strategy will take relatively long but could be speeded up lab-automation. Eventually this may lead to a fully reorganized modular genome, that could be highly beneficial for easy ‘swapping’ of modules when engineering cells towards desired applications.

Acknowledgements

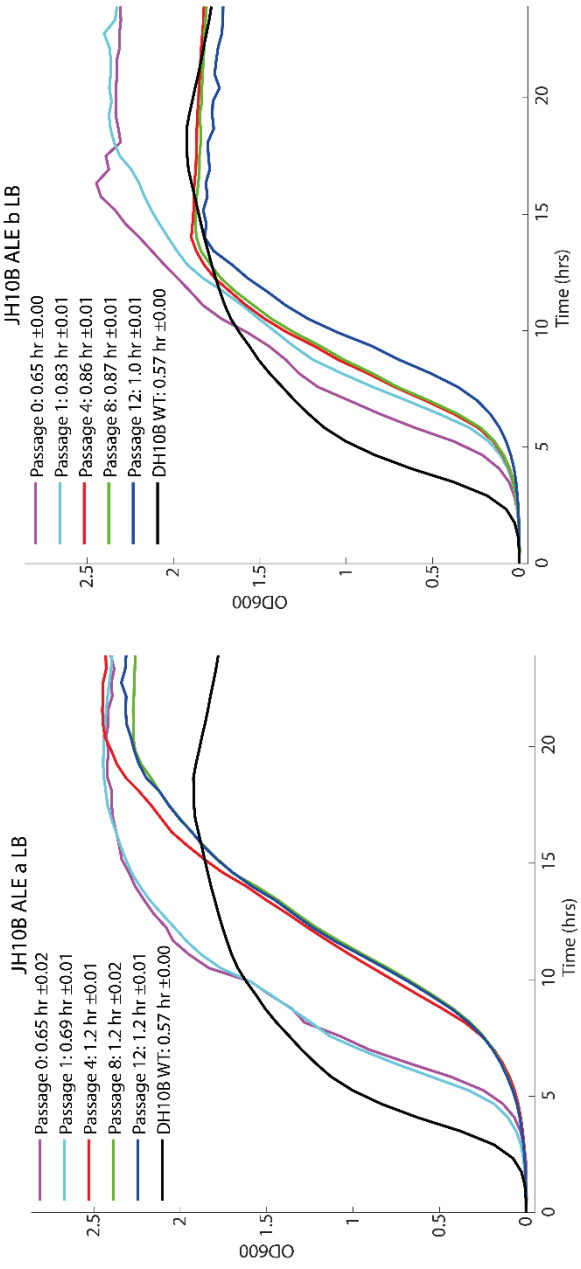
We thank dr. Sjoerd Creutzburg for supplying plasmid pSC020 and the synthetic terminator, and dr. Arren Bar-Even for supplying the MatLab script for determining growth rates.



Supplementary figure 4.1: Growth assays of *E. coli* strains harboring the RNAP operon on LB (left) and M9 medium with glucose (right). Representation of each line shown in figure legend, numbers are doubling time, time point of highest doubling time, and maximum OD reached, respectively. Lines are means of 6 technical replicates.



Supplementary figure 4.2: Growth assays of *E. coli* strains harboring the RNAP operon after ALE on M9+glucose, all replicates. Top graphs have all three biological replicates per passage consisting out of 2 technical replicates split out. Bottom graphs have biological replicates combined. Each representation of each line shown in figure legend. Doubling time and standard deviation are indicated.



Supplementary figure 4.2: Growth assays of *E. coli* strains harboring the RNAP operon after ALE on LB medium. Representation of each line shown in figure legend. Doubling time and standard deviation of 6 replicates are indicated.

Supplementary table 4.1: Combinations of promoters and RBSs tested for *rpoA*. Combination in red was not obtained, green combinations allowed complementation of the lethal *ΔrpoA* mutation, doubling times measured in the consequent growth assay are included.

	P_{Weak}	**$P_{Moderate}$**		**P_{Strong}**		**P_{Native}**	
RBS_{20}	pJL_wA20	pJL_mA20		pJL_sA20		pJL_nA20	
RBS_{40}	pJL_wA40	pJL_mA40		pJL_sA40		pJL_nA40	
RBS_{60}	pJL_wA60	pJL_mA60		pJL_sA60	0.54	pJL_nA60	
RBS_{80}	pJL_wA80	pJL_mA80	0.553	pJL_sA80	0.539	pJL_nA80	
RBS_{98}	pJL_wA98	pJL_mA98		pJL_sA98		pJL_nA98	
RBS_{N}	pJL_wAN	pJL_mAN	0.549	pJL_sAN	0.544	pJL_nAN	0.543

Supplementary table 4.2: List of strains.

Strain	Comment	Source
E. coli DH10B	Used for cloning purposes	Invitrogen
E. coli DH10B pJL_mA80 Δ*rpoA*	DH10B Δ*rpoA* harboring pJL_mA80	This work
E. coli DH10B pJL_mAN Δ*rpoA*	DH10B Δ*rpoA* harboring pJL_mAN	This work
E. coli DH10B pJL_sA60 Δ*rpoA*	DH10B Δ*rpoA* harboring pJL_sA60	This work
E. coli DH10B pJL_sA80 Δ*rpoA*	DH10B Δ*rpoA* harboring pJL_sA80	This work
E. coli DH10B pJL_sAN Δ*rpoA*	DH10B Δ*rpoA* harboring pJL_sAN	This work
E. coli DH10B pJL_nAN Δ*rpoA*	DH10B Δ*rpoA* harboring pJL_nAN	This work
E. coli DH10B pJH05 Δ*rpoA*	DH10B Δ*rpoA* harboring pJH05	This work
E. coli DH10B pJH05 ΔrpoA ΔrpoB/rpoC	DH10B Δ*rpoA* Δ*rpoB/rpoC* harboring pJH05	This work
E. coli JH10B	DH10B Δ*rpoA* Δ*rpoB/rpoC* Δ*rpoZ* harboring pJH05	This work
E. coli JH10B 1-12 a	DH10B Δ*rpoA* Δ*rpoB/rpoC* Δ*rpoZ* harboring pJH05. Evolved for 1-12 generations in M9+glucose	This work
E. coli JH10B 1-12 b	DH10B Δ*rpoA* Δ*rpoB/rpoC* Δ*rpoZ* harboring pJH05. Evolved for 1-12 generations in M9+glucose	This work
S. cerevisiae CEN.PK2-1D	MATa/α ura3-52/ura3-52 trp1-289/trp1-289 leu2-3_112/leu2-3_112 his3 Δ1/his3 Δ1 MAL2-8C/MAL2-8C SUC2/SUC2	Euroscarf

Supplementary table 4.3: List of plasmids.

Plasmid	Description and relevant characteristics	Reference
pHLUM	Harbors CEN/ARS, *his3* and *ura3* used in shuttle vectors	(266)
pBeloBAC11	Harbors BAC replication system *sopA, sopB, sopC* and *repE*	(78)
pSC020	Harbors λ-red system and *cre*	(267)
pJL_wA20	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under weak promoter and RBS_{20}	This work
pJL_wA40	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under weak promoter and RBS_{40}	This work
pJL_wA60	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under weak promoter and RBS_{60}	This work
pJL_wA98	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under weak promoter and RBS_{98}	This work
pJL_wAN	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under weak promoter and RBS_N	This work
pJL_mA20	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under medium promoter and RBS_{20}	This work
pJL_mA40	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under medium promoter and RBS_{40}	This work
pJL_mA60	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under medium promoter and RBS_{60}	This work
pJL_mA80	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under medium promoter and RBS_{80}	This work
pJL_mA98	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under medium promoter and RBS_{98}	This work
pJL_mAN	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under medium promoter and RBS_N	This work
pJL_sA20	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under strong promoter and RBS_{20}	This work
pJL_sA40	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under strong promoter and RBS_{40}	This work
pJL_sA60	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under strong promoter and RBS_{60}	This work
pJL_sA80	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under strong promoter and RBS_{80}	This work
pJL_sA98	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under strong promoter and RBS_{98}	This work
pJL_sAN	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under strong promoter and RBS_N	This work
pJL_nA20	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under native promoter and RBS_{20}	This work

Supplementary table 4.3 (Cont.): List of plasmids.

Plasmid	**Plasmid**	**Plasmid**
pJL_nA40	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under native promoter and RBS_{40}	This work
pJL_nA60	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under native promoter and RBS_{60}	This work
pJL_nA80	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under native promoter and RBS_{80}	This work
pJL_nA98	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under native promoter and RBS_{98}	This work
pJL_nAN	BeloBAC backbone, λ-red system, *cre*, *kan* and *rpoA* under native promoter and RBS_N	This work
pJH05	BeloBAC backbone, CEN/ARS, *his3*, *ura3*, λ-red system, *cre*, *kan* and RNAP operon with strong promoter, RBS_{80} *rpoA* RBS_N *rpoB* RBS_{80} *rpoC* RBS_{80} *rpoZ*	This work

Supplementary table 4.4: List of primers.

Primer name	Primer sequence	Description
RpoA fw	ATGCAGGGTTCTGTGACAGAG	Used for amplifying genes from *E. coli* genome
RpoA rv	TTACTCGTCAGCGATGCTTG	
RpoB fw	ATGGTTTACTCCTATACCGAG	
RpoB rv	TTACTCGTCTTCCAGTTCG	
rpoC fw	GTGAAAGATTTATTAAAGTTTCTG	
rpoC rv	TTACTCGTTATCAGAACCGCC	
rpoZ fw	ATGGCACGCGTAACTGTTCAG	
rpoZ rv	TTAACGACGACCTTCAGCAATAG	
rpoA+R20+spacer_fw	GACGTAATCGTCCAACTTTGAGTAGTGACACAATGCAGGGTTCTGTGACAGAG	Used for amplifying rpoA with respective RBSs
rpoA+R40+spacer_fw	GACGTAATCGTCCAACTTTGAGACATGACACAATGCAGGGTTCTGTGACAGAG	
rpoA+R60+spacer_fw	GACGTAATCGTCCAACTTTGCAAGAGGACACAATGCAGGGTTCTGTGACAGAG	
rpoA+R80+spacer_fw	GACGTAATCGTCCAACTTTGGGGAATGACACAATGCAGGGTTCTGTGACAGAG	
rpoA+R98+spacer_fw	GACGTAATCGTCCAACTTTGAGGAGTGACACAATGCAGGGTTCTGTGACAGAG	
rpoA+Rn+spacer_fw	GACGTAATCGTCCAACTTTGAGAGAGGACACAATGCAGGGTTCTGTGACAGAG	

Supplementary table 4.4 (Cont.): List of primers.

Ps+spacer_fw	AGGCCGTGCCGGCACGTTGCAATACTTGACATATCACTGTGATTCACATATAATATGCGGA CGTAATCGTCCAACTTTG	Used for amplifying *RpoA* with respective promoters
Pm+spacer_fw	AGGCCGTGCCGGCACGTTGCACCTATTGACAATTAAAGGCTAAAATGCTATAATTCCACGA CGTAATCGTCCAACTTTG	
Pw+spacer_fw	AGGCCGTGCCGGCACGTTGCTCCCTTTGATATTGCATCCCGCGTATATAATATGTCGACGT AATCGTCCAACTTTG	
Pn+spacer_fw	AGGCCGTGCCGGCACGTTGCGATCGTCGAGCTTTACTCCAAGTAAAGCTTAGTACCAAAGA GACGTAATCGTCCAACTTTG	
rbs20+rpoB_fw	TGGGCATGCGCCTGGAAAACTGGCCACCGGCAAGCATCGCTGACGAGTAAGATGGCAACCC TATGGTTTACTCCTATACCGAG	Used for amplifying *rpoB* with respective RBSs
rbs40+rpoB_fw	TGGGCATGCGCCTGGAAAACTGGCCACCGGCAAGCATCGCTGACGAGTAAAGCTGGAACCC TATGGTTTACTCCTATACCGAG	
rbs60+rpoB_fw	TGGGCATGCGCCTGGAAAACTGGCCACCGGCAAGCATCGCTGACGAGTAACAAGAGAACCC TATGGTTTACTCCTATACCGAG	
rbs80+rpoB_fw	TGGGCATGCGCCTGGAAAACTGGCCACCGGCAAGCATCGCTGACGAGTAATAGAGAAACCC TATGGTTTACTCCTATACCGAG	
rbs98+rpoB_fw	TGGGCATGCGCCTGGAAAACTGGCCACCGGCAAGCATCGCTGACGAGTAATAGAGGAACCC TATGGTTTACTCCTATACCGAG	
rbsN+rpoB_fw	TGGGCATGCGCCTGGAAAACTGGCCACCGGCAAGCATCGCTGACGAGTAACTGAGGAACCC TATGGTTTACTCCTATACCGAG	
rpoB+overlap_rv	AAAAAAACCCCGCCGAAGCGGGGCGCCCAGTAGAAGCAGCAACTGTTAATTAATTACTCGT CTTCCAGTTCG	

Supplementary table 4.4 (Cont.): List of primers.

rbs20+rpoC_fw	TGAAAGAGATTCGTTCGCTGGGTATCAACATCGAACTGGAAGACGAGTAAAGCTGCCAAAT	Used for amplifying *rpoC* with respective RBSs
rbs40+rpoC_fw	TGAAAGAGATTCGTTCGCTGGGTATCAACATCGAACTGGAAGACGAGTAAAGCTGGCAAAT	
rbs60+rpoC_fw	TGAAAGAGATTCGTTCGCTGGGTATCAACATCGAACTGGAAGACGAGTAAAGTAGGCAAAT	
rbs80+rpoC_fw	TGAAAGAGATTCGTTCGCTGGGTATCAACATCGAACTGGAAGACGAGTAAAAGGAGCAAAT	
rbs99+rpoC_fw	TGAAAGAGATTCGTTCGCTGGGTATCAACATCGAACTGGAAGACGAGTAACGGAGGCAAAT	
rbsN+rpoC_fw	TGAAAGAGATTCGTTCGCTGGGTATCAACATCGAACTGGAAGACGAGTAACGGGAGCAAAT	
rpoC+overlap_rv	AAAAAAACCCGCCGAAGCGGGGCGCCCAGTAGAAGCAGCAACTGTTAATTAATTACTCGT	
rbs20+rpoz_fw	TTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCAGCTACTTTAA	Used for amplifying *rpoZ* with respective RBSs
rbs40+rpoZ_fw	TTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCAGACATTTTAA	
rbs60+rpoZ_fw	TTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCACGAGTTTTAA	
rbs80+rpoZ_fw	TTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCAGGCACTTTAA	
rbs98+rpoZ_fw	TTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCAGGAGCTTTAA	
rbsN+rpoZ_fw	TTACTCGTTATCAGAACCGCCCAGACCTGCGTTCAGCAGTTCTGCCAGGCGAGCTTTTTAA	

Supplementary table 4.4 (Cont.): List of primers.

rpoZ+T+HR_rv	ACTCCGTTACAAAGCGAGGCTGGGTATTTCCCGGCCTTTCTGTTATCCGCAAAAAACCCC GCCGAAGCGGGGCGCCCAGTAGAAGCAGCAACTGTTAATTAATTAACGACGACCTTCAGCA ATAG	*rpoZ* reverse with terminator
pHLUM_OH fw	CCTTTTACAGCCAGTAGTGCTCGCCGCAGTCGAGCGACAGGGCGAAGCCCGATCGCTTGCC TGTAACTTACACGC	Used for amplifying *his3* and CEN/ARS
pHLUM_OH rv	GTGTGTAAGCAGAATATATAAGTGCTGTTCCCTGGTGCTTCCTCGCTCACTAATCGGTGTC ACTACATAAGAAC	
Bac_fw	TACCGCTGAAAGTTCTGCAAAG	pBelobac11 backbone amplification
Bac_rv	AACGTGCCGGCACGG	
pSC020_OH_fw	TTACCCAGGCCGTGCCGGCACGTTGCGGCCGCGGATAACAGAAAGGCCGG	Amplification of λ-red and *cre*
pSC020_OH_rv	TGAGTTTTTCTAATTGCGTTGCGCTCACTG	
Ura3_OH_fw	CATATTCCATTTTGTAATTTCGTGTCGTTTCTATTATGAATTTCATTTATCACTCAACCCT ATCTCGGTC	Amplification of *ura3*
Ura3_OH_rv	ACTCCGTTACAAAGCGAGGCTGGGTATTTCCCGGCCTTTCTGTTATCCGCTCAATTCATCA TTTTTTTTTTATTCTTTTTTTTG	
rpoA_KO_fw	GATCGTCGAGCTTTACTCCAAGTAAAGCTTAGTACCAAAGAGAGGACACATAGAGTCGGAC TTCGCGTTCGC	Amplification of KO template for *rpoA*
rpoA_KO_rv	GATGGCGCATGACCTTATCCTTCTCAGTAAAACCTTAACCTGTGATCCGGGCTAGTTATTG CTCAGCGGTGG	
rpoB_KO_fw	GTCCGCTCAATGGACAGATGGGTCGACTTGTCAGCGAGCTGAGGAACCCTTAGAGTCGGAC TTCGCGTTCGC	Amplification of KO template for *rpoB* and *rpoC*
rpoC_KO_rv	CATAAAAAAACCCGCCGAAGCGGGTTTTTACGTTATTTGCGGATTAACGAGCTAGTTATTG CTCAGCGGTGG	

Supplementary table 4.4 (Cont.): List of primers.

rpoZ_KO_fw	GATTTCAGTATCATGCCCAGTCATTTCTTCACCTGTGGAGCTTTTTAAGTTAGAGTCGGAC TTCGCGTTCGC	Amplification of KO template for *rpoZ*
rpoZ_KO_rv	ATCAGTTGATTCAGGCTTTCAAACAGATACAAGGGCGACCCGCTTTGTGAGCTAGTTATTG CTCAGCGGTGG	

Chapter 5

Introducing the orthogonal ATP-regenerating arginine deiminase pathway in *Escherichia coli*

Joep Houkes[1], Yining Li[1], Qing Yu[1], Sara Benito-Vaquerizo[2], Emma van Aalzum[2], John van der Oost[1], Nico Claassens[1]

[1]Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands

[2] Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands

Abstract

The bacterial arginine deiminase (ADI) pathway appears to be one of the simplest metabolic routes for the generation of ATP. This pathway utilizes three enzymes to convert arginine to ornithine, CO_2, and an ammonium ion, thereby generating one ATP. An additional antiporter imports the arginine substrate while exporting the ornithine product, rendering the arginine deiminase pathway independent from the core cellular metabolism. In this study, we set out to grow *Escherichia coli* using a heterologously expressed arginine deiminase pathway to support its ATP regeneration. For this purpose, a library of pathway operons was generated with different genes and ribosome binding sites. The goal was to obtain functional variants of the pathway by using a stringent growth/no-growth selection by coupling the arginine deiminase pathway to growth. As this selection did not lead to growth, a less stringent approach was used based on screening for ornithine production, which allowed to identify active arginine deiminase pathway variants in *E. coli*. When growing engineered *E. coli* cells under ATP-limited anaerobic growth on glucose, we found that strains harboring the best performing versions of the pathway had up three-fold higher biomass yields. This demonstrates that the arginine deiminase pathway can be a functional, auxiliary ATP-regenerating module to support growth.

Introduction

There are two ways in which energy can be conserved in organisms, by storing it in energy-rich bonds (e. g. ATP), or by creating an electrochemical gradient over a membrane (283). The energy pool stored in electrochemical gradients and ATP can be interconverted via the ATP synthase complex, which can operate in both directions (284). The canonical pathway in many life forms to generate electrochemical gradients is respiration. Although this pathway is very efficient in generating ATP, the high number of membrane-associated components in the respiration chain make it hard to engineer this complex pathway in cell factories or synthetic cells. A simpler one-component solution to generate electrochemical gradients is provided by light-driven, proton-pumping rhodopsins. However, rhodopsins only generate relatively limited ATP under light, and only a small amount (285). Pathways that directly generate energy in the form of ATP include routes containing substrate level phosphorylation reactions, such as glycolysis as well as several routes that have been reviewed recently (49, 286).

In synthetic biology in general, and in synthetic cells in particular, a relatively simple energy generating module may be very useful. One of the most simple and promising substrate level phosphorylation routes to engineer in synthetic systems is the arginine deiminase (ADI) pathway. The ADI pathway requires only 4 genes: *arcA, arcB, arcC* and *arcD,* which encode for three enzymes and an antiporter, respectively. The ArcD antiporter allows the pathway to be completely orthogonal from the rest metabolism of the cell, by importing the substrate (arginine) and exporting the product of the pathway (ornithine) (Figure 5.1). After import of the substrate, the ArcA arginine deiminase drives the conversion of arginine to citrulline, after which the ArcB carbamoyl transferase catalyses the conversion of citrulline and inorganic phosphate to ornithine and carbamoyl-P. Carbamate kinase (ArcC) finally catalyses the reaction of carbamoyl-P with ADP to CO_2 and NH_4, thereby generating one ATP (Figure 5.1).

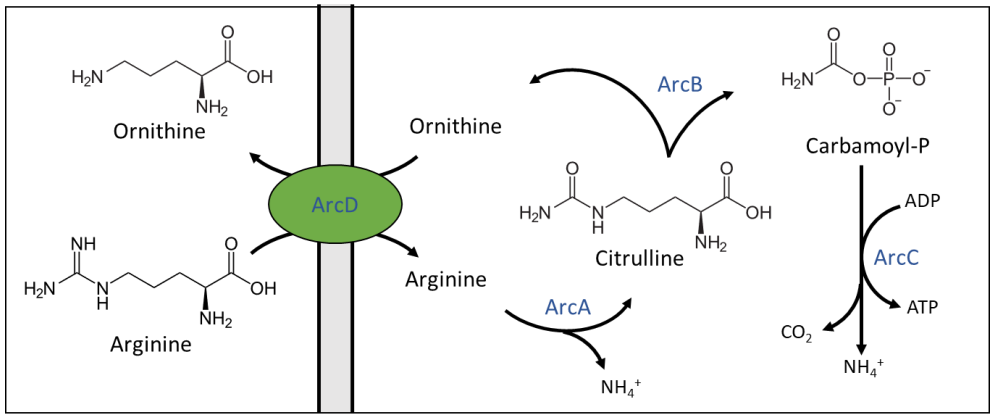


Figure 5.1: Schematic representation of the ADI pathway, protein names indicated in blue. ArcA: arginine deiminase ArcB: ornithine carbamoyltransfererase, ArcC: carbamate kinase, ArcD: arginine/ornithine antiporter.

The ADI pathway is found in several bacteria, archaea, and some anaerobic eukaryotes (287), for energy generation (288, 289) and for NH_4^+ production that may help cells surviving acidic environments (290–294). Previously the ATP-producing potential of the ADI pathway has been used to enhance the growth of the slow-growing acetogenic bacterium *Acetobacterium woodii*. Higher growth rates were obtained transplanting the genes encoding the ADI pathway from *Clostridium autoethanogenum* to *A. woodii.* Most likely by increased ATP production, a 36% increased biomass yield of the engineered acetogen was reported (295).

In this study, we introduced the ADI pathway in *E. coli* to function as an orthogonal ATP producing pathway. To increase the chance of obtaining a functional ADI pathway, we have constructed library of operons with different gene variants and ribosome binding sites to express the pathway. Initially, we aimed to select for strains harboring a functioning ADI pathway. We performed selection by using conditions in which growth of the *E. coli* host strictly requires the ATP-production from functional ADI pathway (296). Unfortunately, our initial approach was not successful. As an alternative, a screening approach was established in which we analyzed the ornithine production. When growing some of the highest ornithine producing strains under ATP-limited conditions, we did not only show that the ADI pathway is active, but also that it leads to a up to 3 times higher biomass yield, which was verified using an *in silico* model.

Materials and Methods

Bacterial strains and culture conditions

E. coli strains (Table 5.1) were routinely grown in LB medium (10 g/L tryptone, NaCl 10 g/L, yeast extract 5 g/L) supplemented with 15g/l agar for plates. 2x YP medium (Tryptone 16 g/L, yeast extract 10 g/L) was used for competent cell generation. For growth experiments, strains were grown in M9 minimal medium (KH_2PO_4 3 g/L, $Na_2HPO_4 \cdot 2H2O$ 7.52 g/L, NaCl 0.5 g/L, NH_4Cl 0.5 g/L), and supplemented with 20 mM glucose, 40 mM acetate, 40 mM pyruvate, 40 mM arginine and/or 20 mM citrulline unless specified otherwise. Minimal medium was adjusted to pH 7 using HCl. Antibiotics were added appropriately (33 μg/mL chloramphenicol (cam) or 50 μg/mL kanamycin (kan)).

Table 5.1: List of strains.

Strain	Use	Reference
E. coli BW25113	Anaerobic assays with ADI operon	(297)
E. coli SIJ488	Control for *E. coli* SIJ488 *ΔackAΔptA*	(298)
E. coli SIJ488 *ΔackAΔptA*	Knock-out strain for full ATP limiting conditions using the ADI operon	Kindly provided by the lab of dr. Arren Bar-Even
E. coli DH10B	Cloning of the ADI library	(224)
L. lactis IL1403	Amplification of ADI genes	(299)

Genetic library design and assembly

Codon harmonized gene variants were designed using the Codon Harmonizer at https://codonharmonizer.systemsbiology.nl/ (300) and ordered as G-blocks (IDT DNA, Supplementary table 5.2). All PCR reactions for cloning purposes were performed using Q5® High Fidelity 2X Master Mix (New England Biolabs). The reactions mixtures were prepared using 1 ng of template, 25 μL Q5® High Fidelity 2X Master Mix, primers to a final concentration of 500 nM, and ddH2O to a final volume of 50 μL (primers are listed in Supplementary table 5.1).

The RBS variants were ordered as primer overhangs with degenerate bases, to be integrated upon PCR amplification. Plasmids were constructed in two phases using Gibson assembly, to reach the highest possible yield (number of colonies). First, two, three-part assemblies

were done, to create plasmids with *arcA* and *arcB* or *arcC* and *arcD* respectively. After transforming *E. coli* DH10B with these libraries, a small part of the transformants were plated, the rest was used immediately pooled and grown on LB medium. A sample of the plated colonies were used to check the assembly by colony PCR (data not shown). After isolation of plasmid DNA from the library, PCR was used to amplify the gene pairs of from their respective plasmids, and another three-part assembly was done to obtain a plasmid library for the complete *arcABCD* operon. After transformation of the Gibson mix into *E. coli* DH10B and recovery, the transformants were grown on LB cam. From this culture, the library was isolated, and used to transform to *E. coli* BW25113 and *E. coli* SIJ-488 *ΔackAΔptA* a small part of the transformants were plated, the rest was used immediately grown on LB liquid medium.

Microplate reader growth assays

Aerobic precultures grown on M9 + 20 mM glucose were washed 3 times with ddH_2O and diluted to OD600 0.1 in the respective medium. Next, 15 µL of diluted preculture of each strain and 135 µL of medium was transferred in a 96-well plate. The wells were covered with 50 µL of mineral oil (Bio-Rad), to avoid evaporation during the experiment. The plate was then incubated at 37 °C in a Biotek ELx800, absorbance microplate reader (Fisher Scientific) placed in an anaerobic tent. The provided reader control software, Gen5, was used to set a measuring protocol consisting of a cycle of 5 min of fast linear shaking at 19 Hz followed by absorbance measurement at 600 nm, for at least 24 h. An in-house MatLab script was used to process the data, yielding strain-specific growth graphs and doubling times.

HPLC analyses

For amino acid analysis and fermentation product analysis high-performance liquid chromatography (HPLC) was performed. Samples were taken from the end of fermentation (48 or 72 hours depending on experiment time), and further processed using dabsyl chloride derivatization to allow for amino acid analysis. For this, 10 µL of clear culture supernatant was mixed with 100 µL 0.15M $NaHCO_3$ pH 9.0. Dabsyl chloride solution (1.3 mg/mL dabsyl chloride in acetonitrile) was prepared fresh, and 200 µL was added to the sample solution and mixed before incubating at 70°C for 20 minutes in a heat block. Samples were cooled down to room temperature and 690 µL 70% ethanol was added. The same procedure was used for standards (arginine, citrulline, ornithine ranging from 2.5 – 20 mM), norleucine was used as internal standard for all measurements. The derivatized amino acids were analyzed in a Shimadzu LC 2030 C Plus HPLC using an Agilent Poroshell 120 EC-C18 column,

250x4. 6 mm, 4 µm particles (Agilent; 690970-902) with a flow rate of 1 ml/min. For a separation, an eluent gradient was applied with 25 % eluent A: 10 mM formic acid in water and 75% eluent B: 10 mM formic acid in acetonitrile. During the run of 30 minutes the ratio A/B is linearly changed to 25/75%. At the end the ratio is brought back to 75/25 % again and kept at that ratio for ten minutes to stabilize the column for the next run. The derivatized amino acids are detected using a visible light detector at 436 nm.

Sequencing

Promising strains were further analyzed by sequencing the ADI genes. For this, strains were grown overnight on LB medium and subsequent plasmid isolation was done using GeneJET Plasmid MiniPrep Kit (Thermo Fisher Scientific). Isolated plasmids were sequenced using four sequencing primers (Supplementary table 5.1) that can bind in the operon regardless of gene variants at Macrogen Europe. Sequencing results were aligned using Benchling.

Constraint-based metabolic modelling

The genome-scale metabolic model (GEM) of *E. coli*, iJO1366 (301)was retrieved in SBML format level 3 version 1 from BiGG Models (302). Model simulations were performed using COBRApy 0. 24.0 (303)and Python 3.9.

Flux balance analysis (FBA) (304) was applied to simulate four different conditions: growth on glucose; growth on glucose and arginine; growth on glucose with ADI, and growth on glucose and arginine with ADI in anaerobic conditions. Glucose uptake rate ('EX_glc__D_e') was constrained to 8 mmol gDW^{-1} h^{-1} in every condition, and arginine uptake rate ('EX_arg__L_e') was constrained to 20 mmol gDW^{-1} h^{-1} when growth was simulated on glucose and arginine (with/without ADI). Three reactions were added to the model to simulate growth on glucose/glucose + arginine with ADI pathway: Arginine deiminase ('argA'), Citrulline transport ('TCITRe'), and Citrulline exchange reaction ('EX_citr__L_e'). For all conditions, the fluxes through putrescine/ornithine antiporter (periplasm) ('PTRCORNt7pp'), Ornithine transport via ABC system (periplasm) ('ORNabcpp'), Arginine/agmatine antiport (periplasm) ('ARGAGMt7pp'), and oxygen exchange reaction ('EX_o2_e') reactions were set to 0, and the lower bound of the ATP maintenance ('ATPM') reaction was constrained to 8.4 mmol gDW^{-1} h^{-1}. In every condition, the biomass synthesis reaction (‘BIOMASS_Ec_iJO1366_WT_53p95M’) was used as the objective function, and fluxes were computed for the maximum growth rate. The ratio between the predicted maximum growth rate (h^{-1}) and the glucose uptake rate (mmol gDW^{-1} h^{-1}) was used to calculate the biomass yields in gDW mol^{-1} of glucose.

Furthermore, Flux variability analysis (FVA) was applied in every condition to calculate the minimum and maximum production rate (mmol gDW $^{-1}$ h^{-1}) of acetate ('EX_ac_e'), ornithine ('EX_orn_e') and citrulline ('EX_citr__L_e') at 90% optimality.

Results and discussion

Synthetic ADI-operon design and library construction

In total, 11 ADI gene variants were used in this approach, the wild-type sequences of the *Lactococcus lactis* ADI genes, *arcA, arcB, arcC* and *arcD2*, as well as versions of these genes that were codon-harmonized for expression in *E. coli*, the codon-harmonized *arcB* from *Pseudomonas aeruginosa*, and *arcD2* from *Lactobacillus sakei*, and a codon-harmonized version of the latter. For each gene a small library of six RBSs using degenerate bases was designed, based on the RedLibs algorithm (305, 306). This algorithm generates a degenerate RBS design that covers a wide, linear range of predicted translation initiation rates. By combining 11 gene variants with 6 RBS variants per gene the theoretical library size is 62,208 (Figure 5.2). The library was assembled in a 2-step approach (see methods). To estimate assembly efficiency of the library a colony PCR was performed on 16 colonies of the *E. coli* BW25113 pADI_Lib transformants, which suggested 13 out of 16 colonies were correct (Supplementary figure 5.1). Four of the correct colonies were picked for overnight culturing, plasmid isolation and sequencing. All four colonies harbored the complete operon, while having diverse gene variants and RBSs, indicating that this library construction method leads to a diverse population (Supplementary table 5.3).

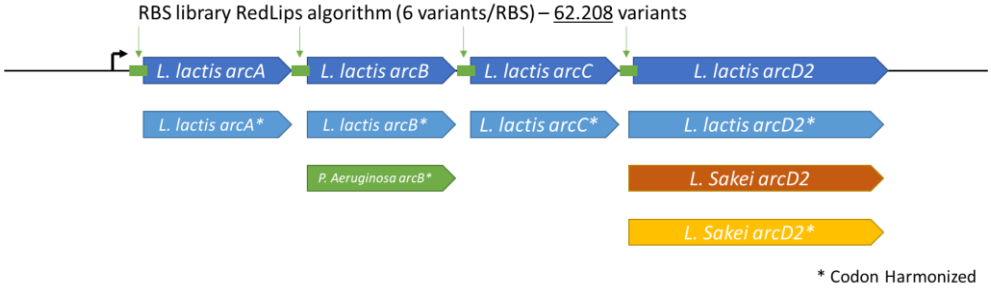


Figure 5.2: Library preparation of ADI operon. For each gene the originating organism is shown, an asterisk indicates the codon harmonized version. Each gene is preceded by an RBS designed with RedLibs (305, 306) using degenerate bases to give six variants with different predicted translation initiation strengths.

Selection of ADI operon library in full ATP-limited conditions does not result in growth

To assess the functionality of the ADI pathway as sole, functional ATP regeneration system in *E. coli*, the transformant library was tested in full ATP-limiting conditions. For this, two strictly ATP-limited conditions were tested (i) *E. coli* BW25113 pADI_lib was cultured anaerobically in M9 supplemented with arginine and acetate and (ii) *E. coli* BW25113 *ΔackAΔptA* pADI_lib was cultured anaerobically in M9 supplemented with arginine and pyruvate. In anaerobic conditions, acetate cannot be used for substrate level phosphorylation, nor for respiration. In native conditions, pyruvate can be used for substrate level phosphorylation by acetate kinase (*ackA)*, in the latter strain *ackA* is knocked out. Therefore, in both conditions, ATP can only be generated using the ADI pathway. As positive control, both libraries were grown anaerobically on M9 supplemented with arginine and glucose. Library strains were cultured anaerobically in ATP limiting conditions for up to 4 weeks, but no growth was observed by visual inspection and OD600 measurement, while growth was observed in the positive control bottles (Data not shown).

Screening of ADI operon library in partial ATP-limited conditions identifies functional clones

As we could not identify growth for the selection strains harboring the synthetic ADI operon library, we suspected that ATP-regeneration capacity of the synthetic ADI system was not sufficient to support growth in full-ATP limited conditions. As alternative approach to identify functional synthetic ADI operons, we resorted to a screening approach instead. For this, *E. coli* BW25113 pADI_lib was plated, and single colonies were picked and grown aerobically in M9 medium supplemented with glucose. After overnight culturing the strains were washed and transferred to a microtiter plate and grown anaerobically in M9 medium supplemented with arginine and glucose. Under these conditions, *E. coli* performs fermentation and part of the carbon flux from glucose is converted to acetate to produce ATP using substrate level phosphorylation during glycolysis and by acetate kinase (307). Using this pathway, *E. coli* can theoretically obtain a maximum of four ATP per glucose consumed, while under aerobic conditions the respiratory chain can yield approximately 30 ATP per glucose consumed (308). In these ATP-limiting conditions, strains that harbor a functioning ADI operon can obtain extra ATP, by converting arginine into ornithine.

In total 94 strains were picked and grown in the microtiter plate. In this screen, 84 of 94 strains showed growth over the 72-hour experiment. Among the growing strains the doubling

times varied from 1.2 to 1.8 hours. Interestingly, the highest biomass yield (measured as maximum OD_{600}) reached by these cultures varied widely, ranging from 0 to 2.0, averaging out at 0.62.

Of the tested strains, 55 produced the final product of the ADI pathway: ornithine, varying from ~0.4 mM to 21 mM (Figure 5.3). Additionally, in 49 of the samples citrulline, an ADI-pathway intermediate, was present, in amounts varying from 0.1 mM to ~ 37 mM. Citrulline is an intermediate in the ADI pathway, which is further converted to ornithine and carbamoyl-P, the latter is than converted to produce ATP. Finding citrulline in the supernatant implies that instead of ornithine, citrulline is exported by the antiporter, and no ATP is produced inside the cell (Figure 5.1). It is possible that these strains have a well-functioning arginine deiminase and antiporter, but the carbamoyl transferase or carbamate kinase enzymes are not produced, or not active.

Strikingly, the five strains that produced the most ornithine, are also the strains with the highest yield (growing to the highest OD_{600}) (Figure 5.3). These strains are ADI 83, ADI 11, ADI 39, ADI 54, and ADI 9 from left to right. These data suggest that in these strains the ADI pathway is functioning efficiently, converting arginine to ornithine, thereby producing ATP, which can be used for growth.

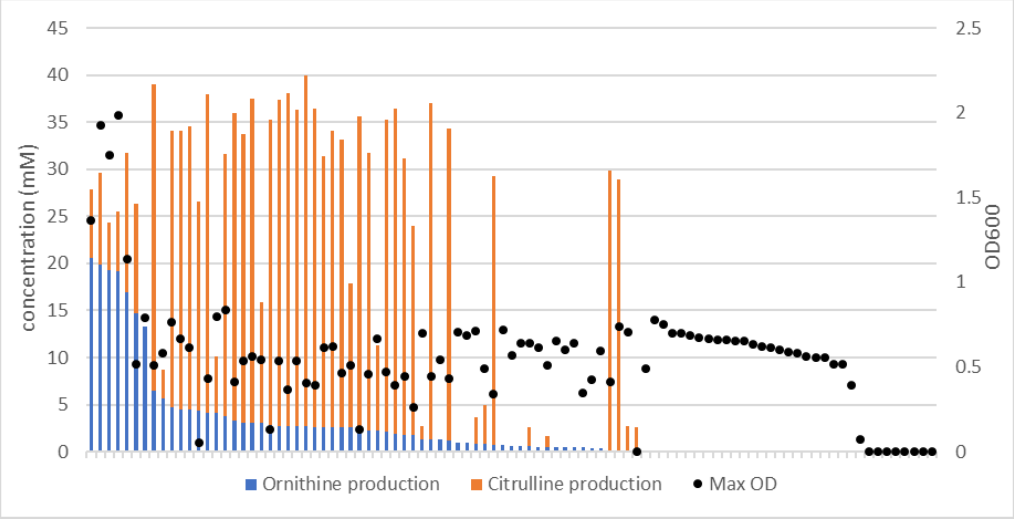


Figure 5.3: Production of ornithine and citrulline, and the maximum OD_{600} of 94 strains from ADI library. Sorted by ornithine production, then citrulline production, then OD. The five highest ornithine producers also have the highest maximum OD_{600}

Additional testing of colonies producing most ornithine and growing to highest ODs

To get more information on the ADI strains, the 5 best performing strains were selected, together with 19 randomly chosen strains. The plasmid harboring the ADI operon was isolated from all of these 24 strains, and the complete ADI operon was sequenced. We were able to obtain the full sequence of the synthetic ADI operon for 15 out of the 24 strains (Supplementary table 5.4). These 15 full sequences include the sequence of the 5 best performing strains. Of the gene variants used, all but the *L. lactis arcD* and *L. lactis arcD* harmonized variant are present in the sequenced trains. From this data, we could not find a correlation between the 5 best performing strains and gene variants. It appears that many variants can perform well.

For the 5 best-performing strains further growth experiments were performed in triplicate on M9 supplemented with glucose with and without arginine (Supplementary figure 5.2). The experiments indicate that adding arginine to the medium greatly enhances the yield of the five ADI strains, but not the yield of the empty plasmid control. For these strains the fermentation products were analyzed, as well as the production of ornithine and citrulline (Figure 5.4).

All five strains harboring the ADI pathway produce more formate and acetate when arginine is present. In the empty plasmid control no real difference is observed between growth with and without arginine. During growth on glucose without arginine, more lactate is produced. In *E. coli* fermentation, ATP is produced during the reaction from pyruvate to acetate, but not from pyruvate to lactate, which might add to the stimulatory effect of the arginine deiminase pathway on growth.

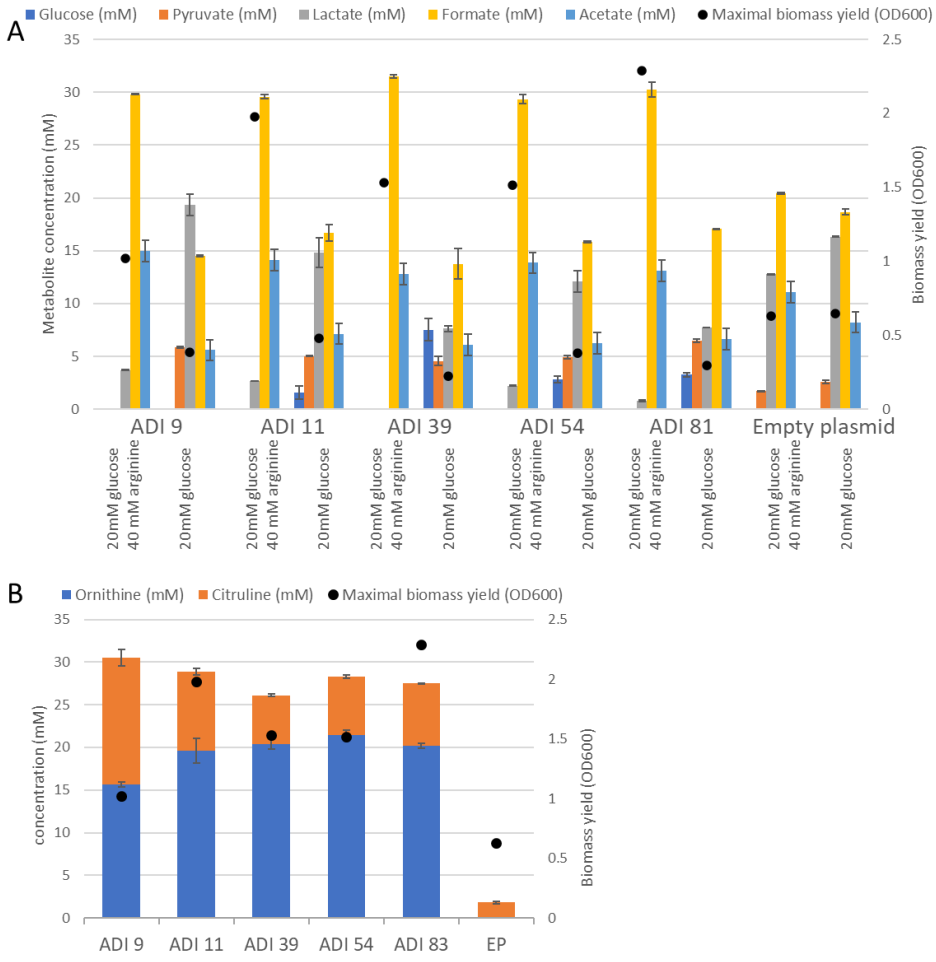


Figure 5.4: Extracellular metabolite production for best-growing strains. **A**: Fermentation product analysis of best performing colonies. Each strain was grown on M9 supplemented with glucose and M9 supplemented with glucose and arginine as indicated in the figure. Bars represent concentration of respective fermentation products; black dots represent average highest OD obtained during 48-hour fermentation in triplicate. Error bars represent standard deviation of 2 replicates **B**: Citrulline and ornithine production of best performing colonies. Strains were grown on M9 supplemented with glucose and arginine. Error bars represent standard deviation of 3 replicates.

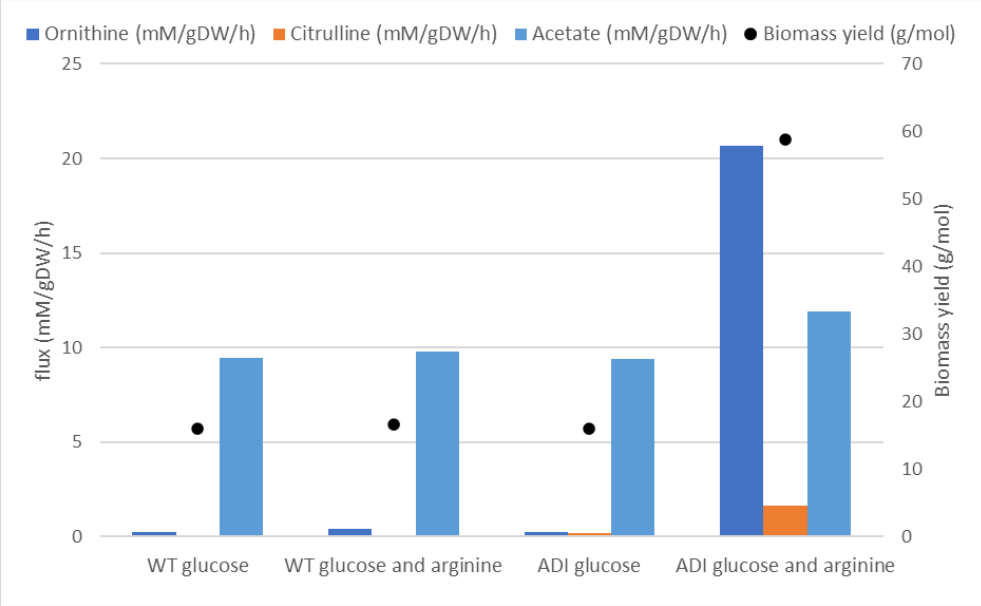


Figure 5.5: In silico simulation of *E. coli* strains with and without the ADI pathway (ADI and WT respectively) modelled on only glucose and on glucose and arginine. Biomass yields were obtained at the maximum growth rate and metabolite production rates were the maximum flux obtained at 90% optimality.

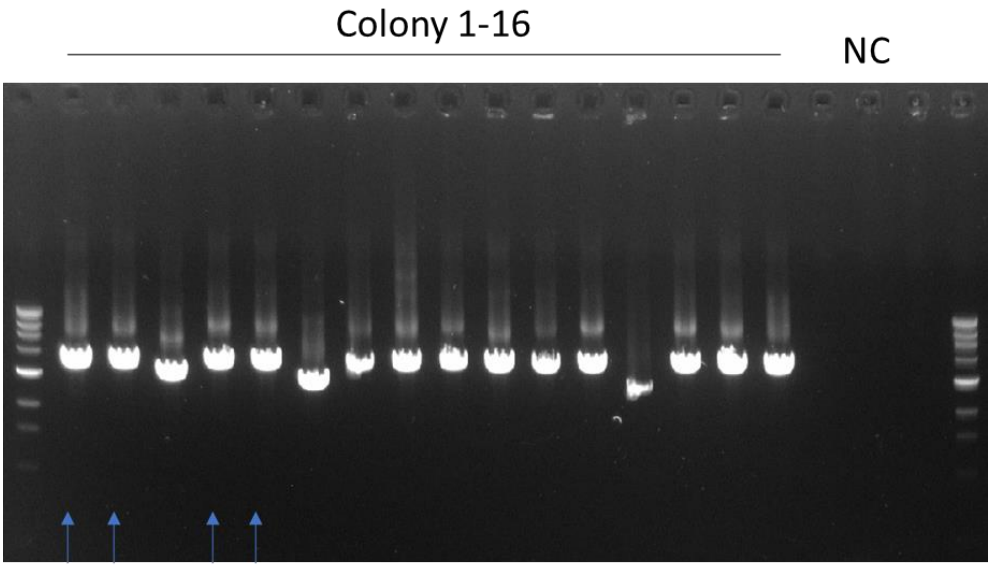
In-silico simulation of synthetic ADI pathway functionality in *E. coli*

The genome-scale metabolic model of *E. coli*, iJO1366 (301) was used to simulate the results obtained in the ADI screening assay. The model predicts 3. 5-fold higher biomass yield when the ADI pathway is present and both glucose and arginine are present (Figure 5.5). In all other cases the biomass yield remains stable. This fits well with the results obtained during the screening experiment, the best performing strains grew approximately to a 2.5 - 3.5 times higher OD600 compared to the empty plasmid strain (Figure 5.4). The model also predicts a major ornithine production and minor amounts of citrulline production, this is also seen for the best growing strains experimentally, though relatively more citrulline is still seen experimentally, suggested the ADI pathway operation may be further optimized. The model also predicts slightly increased acetate production.

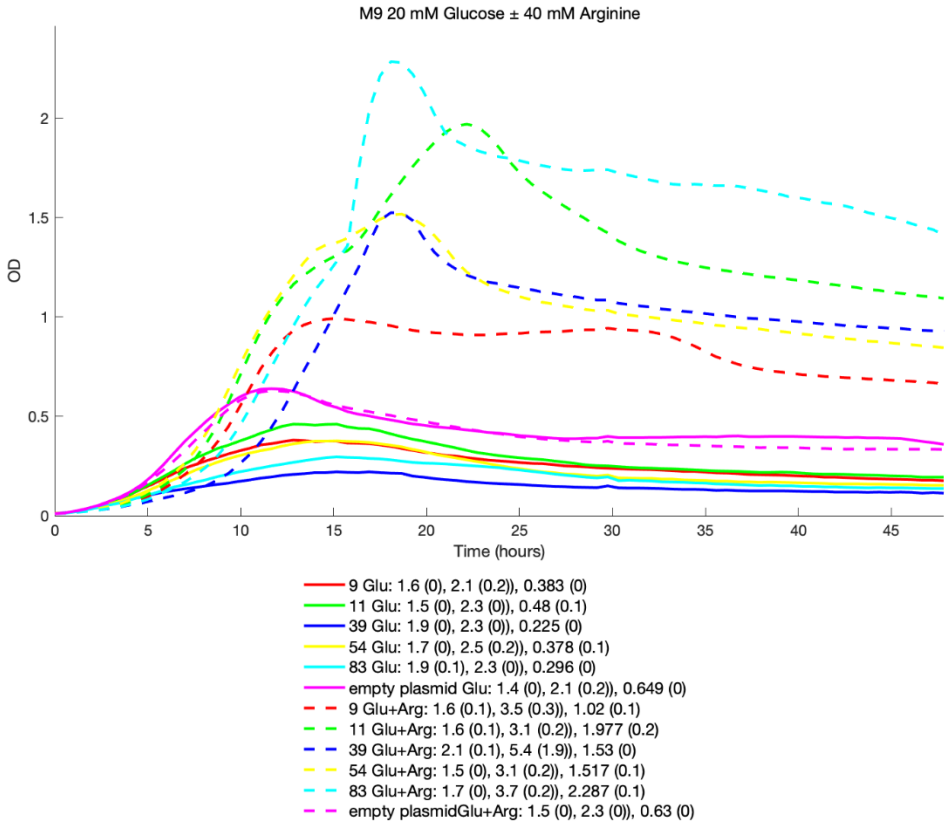
Conclusions

We here demonstrated that it is feasible to engineer the arginine deiminase pathway as an orthogonal ATP-generating pathway in *E. coli*. The pathway can be used to for additional ATP production, which has also been shown previously in *A. woodii* (295). In *E. coli* this pathway can lead to up to 3. 5 times higher yields compared to an ADI-deficient control. Although the here used ADI pathway is not sufficient for generating all ATP in *E. coli*, further improving, or evolving the current ADI operon could potentially lead to such improved functionally. Theoretically, one ATP per arginine substrate is enough to sustain *E. coli* growth, therefore, it could also be enough for fueling minimal synthetic cells in the future.

Supplementary data



Supplementary figure 5.1: Colony PCR on 16 colonies harboring ADI operon library plasmid. 13/16 appear to be the correct size. Colonies of bands shown by blue arrows were used for sequencing of the operon variants.



Supplementary figure 5.2: Growth experiment of best performing strains on M9 supplemented with glucose with and without arginine.

Supplementary table 5.1: List of primers used in this study.

Primer Name	Primer Sequence	Description
ArcA_fw	ATGAACAATGGAATTAATGTTAAC	Amplification of *arcA* from *L. lactis*
ArcA_rv	TTACAAATCTTCACGCCAAAG	
ArcB_fw	ATGACATCACCACTTATTACAAAAG	Amplification of *arcB* from *L. lactis*
ArcB_rv	TTATTTAAAATCTTCAGGAACTGC	
ArcC1_fw	ATGGTAAAACGTATTGTTGTAGC	Amplification of *arcC* from *L. lactis*
ArcC1_rv	TTAAGGAACAATTCTTGTTCC	
ArcD2_fw	ATGGAAAACAAGAAAACAAAAGG	Amplification of *arcD* from *L. lactis*
ArcD2_rv	TTAGTAGCCTAGTAATCCCCG	
LL_arcA_H_fw	ATGAATAACGGCATTAACGTG	Amplification of *L. lactis arcA* harmonized from g-block
LL_arcA_H_rv	AGTCCATGGGTTAGGCTATCTTAC	
LL_arcB_H_fw	ATGACCAGCCCGCTGATTAC	Amplification of *L. lactis arcB* harmonized from g-block
LL_arcB_H_rv	TGAGTATTGGACTCACAAACTTATTTAAAATC	
LL_arcC_H_fw	ATGGTAAAACGCATTGTGGTAG	Amplification of *L. lactis arcC* harmonized from g-block
LL_arcC_H_rv	AGTTCCGGTGTCACCGGTAG	
LL_arcD_H_fw	ATGGAAAATAAGAAAACCAAAGG	Amplification of *L. lactis arcD* harmonized from g-block
LL_arcD_H_rv	TCGGCATTAGTACCCGAGC	

Supplementary table 5.1 (Cont): List of primers used in this study.

Primer Name	Primer Sequence	Description
ArcA_fw	ATGAACAATGGAATTAATGTTAAC	Amplification of *arcA* from *L. lactis*
ArcA_rv	TTACAAATCTTCACGCCAAAG	
ArcB_fw	ATGACATCACCACTTATTACAAAAG	Amplification of *arcB* from *L. lactis*
ArcB_rv	TTATTTAAAATCTTCAGGAACTGC	
ArcC1_fw	ATGGTAAAACGTATTGTTGTAGC	Amplification of *arcC* from *L. lactis*
ArcC1_rv	TTAAGGAACAATTCTTGTTCC	
ArcD2_fw	ATGGAAAACAAGAAAACAAAAGG	Amplification of *arcD* from *L. lactis*
ArcD2_rv	TTAGTAGCCTAGTAATCCCCG	
LL_arcA_H_fw	ATGAATAACGGCATTAACGTG	Amplification of *L. lactis arcA* harmonized from g-block
LL_arcA_H_rv	AGTCCATGGGTTAGGCTATCTTAC	
LL_arcB_H_fw	ATGACCAGCCCGCTGATTAC	Amplification of *L. lactis arcB* harmonized from g-block
LL_arcB_H_rv	TGAGTATTGGACTCACAAACTTATTTAAAATC	
LL_arcC_H_fw	ATGGTAAAACGCATTGTGGTAG	Amplification of *L. lactis arcC* harmonized from g-block
LL_arcC_H_rv	AGTTCCGGTGTCACCGGTAG	
LL_arcD_H_fw	ATGGAAAATAAGAAAACCAAAGG	Amplification of *L. lactis arcD* harmonized from g-block
LL_arcD_H_rv	TCGGCATTAGTACCCGAGC	

Supplementary table 5.1 (Cont.): List of primers used in this study.

Primer Name	Primer Sequence	Description
PA_ArcB_H_fw	ATGGCTTTTAACATGCATAAC	Amplification of *P. aeruginosa* a*rcB* harmonized from g-block
PA_ArcB_H_rv	TGAGTATTGGACTCACAAACTTAAATATC	
LS_arcD_H_fw	ATGACCGAAGAAAAACCGG	Amplification of *L. sakei* a*rcD* harmonized from g-block
LS_arcD_H_rv	AAGCAGCAACTGAGTCGGC	
LS_ArcD_fw	ATGACGGAAGAAAAACCAGC	Amplification of *arcD* from *L. sakei*
LS_ArcD_rv	CTAAATAACGATTTTCCCTGAAACAAC	
pACYC_ADI_fw	TGCCGACTCAGTTGCTGC	Amplification of pACYC184 backbone
pACYC_ADI_rv	CAAAGTTGGACGATTACGTC	
L. lactis ArcA RL_fw	GACGTAATCGTCCAACTTTGDAARGGAGATTAATTAATGAACAATGGAATTAATGTTAAC	*L. lactis* a*rcA* fw primer with RedLibs RBS
L. lactis ArcB RL_fw	ATAGCCTAACCCATGGACTTAACBGGARGTTTAATATGACATCACCACTTATTACAAAAG	*L. lactis* a*rcB* fw primer with RedLibs RBS
L. lactis ArcC RL_fw	GTTTGTGAGTCCAATACTCATATGGGBGGKTATTTAATGGTAAAACGTATTGTTGTAGC	*L. lactis* a*rcC* fw primer with RedLibs RBS
L. lactis ArcD RL_fw	CTACCGGTGACACCGGAACTTYGAGGAGTATTTCDTATGGAAAACAAGAAAACAAAAGG	*L. lactis* a*rcD* fw primer with RedLibs RBS
L. lactis ArcA Harmonized RL_fw	GACGTAATCGTCCAACTTTGAATTAAGGGBGCAATWATGAATAACGGCATTAACGTG	*L. lactis* a*rcA* harmonized fw primer with RedLibs RBS
L. lactis ArcB Harmonized RL_fw	GATAGCCTAACCCATGGACTCTSAAGGAGGTTGHATATGACCAGCCCGCTGATTAC	*L. lactis* a*rcB* harmonized fw primer with RedLibs RBS

Supplementary table 5.1 (Cont.): List of primers used in this study.

Primer Name	Primer Sequence	Description
L. lactis ArcC Harmonized RL_fw	GTTTGTGAGTCCAATACTCACTGKTGGAGGTTTGTDATGGTAAAA CGCATTGTGGTAG	*L. lactis arcC* harmonized fw primer with RedLibs RBS
L. lactis ArcD Harmonized RL_fw	CTACCGGTGACACCGGAACTTAAHGGMGCAATTAATATGGAAAAT AAGAAAACCAAAG	*L. lactis arcD* harmonized fw primer with RedLibs RBS
L. sakei ArcD RL_fw	CTACCGGTGACACCGGAACTTAAGGDAGAGYTTAATATGACGGAA GAAAAACCAGC	*L. sakei arcD* fw primer with RedLibs RBS
L. sakei ArcD Harmonized RL_fw	CTACCGGTGACACCGGAACTAGDTAAGGAATAWCCTATGACCGAA GAAAAACCGG	*L. sakei arcD* harmonized fw primer with RedLibs RBS
P. aeruginosa ArcB Harmonized RL_fw	GATAGCCTAACCCATGGACTAATTGGVGMAGATATTATGGCTTTT AACATGCATAAC	*P. aeruginosa arcB* harmonized fw primer with RedLibs RBS
pTN03_fw	TATAGGATCCTAACTCGAGC	Backbone primers used in plasmid for 2-step assembly
pTN03_rv	GCCTTAATACGTTATTAGCCG	
pTN03_fw_arcAB	GTTTGTGAGTCCAATACTCATATAGGATCCTAACTCGAGC	Primers used in 2-step assembly of *arcA* and *arcB*
pTN03_rv_arcAB	CAAAGTTGGACGATTACGTCGCCTTAATACGTTATTAGCCG	
pTN03_fw_arcCD	TGCCGACTCAGTTGCTGCTTTATAGGATCCTAACTCGAGC	Primers used in 2-step assembly of *arcC* and *arcD*
pTn03_rv_arcCD	TGAGTATTGGACTCACAAACGCCTTAATACGTTATTAGCCG	

Supplementary table 5.2: G-blocks used in this study.

Name	Sequence
L. lactis arcA Harmonized	ATGAATAACGGCATTAACGTGAATAGCGAAATTGGAAAACTGAAAAGCGTACTCCTCCACCGGCCGGGCGCTGAAGTGGAAAACATTACAC CGGACACCATGAAACAGCTGCTGTTTGATGATATTCCGTATCTCAAAATTGCGCAGAAAGAGCATGATTTCTTTGCGCAGACCCTGCGCGA TAACGGCGCCGAAACCGTGTATATAGAAAACCTGGCCACCGAAGTGTTTGAAAAAAGCTCGGAAACCAAAGAAGAATTCCTGTCGCACCTG CTGCACGAAGCTGGCTACCGCCCGGGCCGGACCTATGATGGCCTGACCGAATATCTGACCAGCATGAGCACCAAAGATATGGTAGAAAAAA TATATGCGGGCGTGCGCAAAAACGAACTGGATATAAAACGCACCGCCCTGTCAGATATGGCCGGCTCGGATGCCGAAAACTATTTCTACCT CAATCCGCTGCCGAACGCTTACTTCACCCGCGACCCGCAGGCGAGCATGGGCGTAGGCATGACAATTAACAAAATGACATTCCCGGCGCGC CAGCCGGAATCACTGATTACCGAGTATGTGATGGCGAACCACCCGCGCTTCAAAGATACCCGATATGGCGCGATCGCAATCATACAACCC GCATAGAAGGCGGCGATGAACTGATTCTGAATAAAACCACCGTAGCGATAGGAGTGAGCGAACGCACCAGCTCGAAAACCATTCAGAACCT GGCGAAAGAGCTGTTTGCCAATCCGCTGAGCACCTTTGACACCGTACTGGCTGTGGAAATTCCGCATAATCATGCCATGATGCATCTGGAT ACCGTATTTACCATGATTAATCATGATCAGTTTACCGTATTCCCGGGCATTATGGATGGCGCGGGCAACATAAACGTATTCATTCTGCGGC CGGGCAAAGATGACGAAGTGGAAATTGAACATCTGACCGACCTGAAAGCCGCCCTCAAGAAAGTACTGAACCTGAGCGAACTGGACCTGAT TGAATGCGGCGCCGGCGATCCGATTGCTGCGCCGCGGGAACAGTGGAACGATGGCAGCAACACCCTGGCGATAGCCCCCGGCGAAATAGTG ACCTACGACCGCAATTATGTAACCGTGGAGCTGCTGAAAGAGCATGGCATAAAAGTGCATGAAATTCTGAGCTCAGAACTGGGTCGTGGAC GTGGTGGTGCTCGTTGCATGAGCCAGCCGCTGTGGCGCGAAGATCTGTAAGATAGCCTAACCCATGG
L. lactis ArcB Harmonized	ATGACCAGCCCGCTGATTACCAAAGCCGAAGTAAATAGCGTATTCCAGGGCCGGTCACTGCTGGCGGAAAAAGATTTCACCCCGGCGGAAA TTAATTACCTGGTGGATTTTGGCCTGCATCTGAAAGCCCTGAAACAGCAGAATATTCCGCACCACTACCTGGAAGGCAAAAACATTGCTCT GCTCTTTGCCAAAACCAGCACCCGCACCCGCGCGGCTTTTACCACCGCGGCCATTGACCTGGGCGCGCATCCGGAATACCTGGGCGCGAAT GATATTCAGCTGGGCATTAAAGAAAGCACCGAAGATACCGCGCGGGTACTGGGCAGCATGTTTGATGCGATAGAACGCCGCGGCTTTTCGC AGAAAGAAGTGGAAGATCTGGCCAAATACTCGGGCGTGCCGGTGTGGAACGGCCTGACCGATGATTGGCATCCGACCCAGATGATTGCCGA CTTTATGACCGTGAAAGAAAATTTTGGCCATCTGAAAGGCCTGACCCTGGTATACGTAGGCGATGGCCGCAATAATATGGCCAATAGCCTG ATAGTAACCGGCTCGATGCTGGGCGTAAACGTGCATATTGTGGCGCCGGATAGCCTGCATCCGACCAAAGAAGTGATGGATATTGCCAACA AATTTGCGGAAAAAAGCGGCGCGAAACCGCTGGCCACCAGCAACATTGAAGAAGGCGTGAAAGGGGCCAATATTATTTACAGCGACGTGTG GGTAAGCATGGGCGAATCGAATTGGGAAGAACGCGTGAAACTGCTGACCCCGTACCGCATAACCATGGATATGCTGAAAATGACCGGCAAC GCGGACAACGGCCAGCTGATATTTATGCACTGTCTGCCGGCCTTCCATGATACCGAAACCGAATATGGCAAAGAAATAAAAGAAAAATATG GCCTGACCGAAATGGAAGTAACCGACGAAGTATTCCGCTCGAAATATGCGCGGCAGTTTGAAGAAGCCGAAAACCGCATGCACTCGATAAA AGCCATAATGGCGGCCACCCTGGGCAACCTGTTTATACCGGCCGTGCCGGAAGATTTTAAATAAGTTTGTGAGTCCAATACTCA

Supplementary table 5.2 (Cont.): G-blocks used in this study.

Name	Sequence
L. lactis ArcC1 Harmonized	ATGGTAAAACGCATTGTGGTAGCTCTGGGCGGAAACGCGATTCTTTCAAAAGACCCGAGCGCTAAGGCGCAGAAAGAAGCCCTGGAAGTGACC GCGGAACAGCTGATGCCGCTGATTAAAGATAACGACCGACACAGATGGTGGTAACCCATGGCAACGGCCCGCAAGTGGGCAATCTGCTGCTGCAG CAGCTCGGCAGCGATTCGGAAAAAAAACCCGGCGATGCCGCTGGATACCGTGGGCGCCATGACACAGGGCGAGATTGGCTTTTGGACAGTGCAG GCCTTTACCCGGGCCATGGCCAACAACGGGCTGGAGCATTATCCGGTGATTAGCGCCGTAACCCAGACCCTGGTGGATGAAAACGATCCGGCG TTTAAAAACGCGTCGAAACCGATTGGCCCGTTTTACACCGAAGAACAGCTGCCGGAAATTAAGAAGCAGTTTCCGGATTGGGAACTGATGGAA GACGCGGGCCGGGGCTATCGGCGGGTAGTGCCGTCGCCGAAACCGCTGGAAATTATTGAAGCTAAAGCGTTTGGCCCGATACTGGAATCGGGC TCGCTCCTGATTGCGTCGGGGGGCGGCGGCATTCCCGTAATTAAAGATGGCAAGGGCTATCAGGGCGTGGAAGCTGTAATTGATAAAGATTTC TCGGCCGCGAAACTTGCGGAGATGGTGGGCGCCGACGAGCTGCTGATTCTGACCTCAGGCGGCGACGTATATCTGAATTATGGAAAAGAAAAT CAGGTGCATCTGGAAACCGTAAAAGTGTCGGAAATGCGGAAATATGTGGACGAAGAACGGTTTGCCAAAGGCAGCATGGGCCCGAAAGTAGAA GCCGCGGTATCGTTTGTGGAACGGACCGGCAAGACCGCGACCATTACCTCACTGGACAACGTGGAAAACTTTGTAAAATCGGGCTCGGGCACC CGGATTGTGCCGTAACTACCGGTGACACCGGAACT
L. lactis ArcD2 Harmonized	ATGGAAAATAAGAAAACCAAGGAATTTCACTGTTTGCCCTCCTGGCCATTATTATTTCGGGCGCCATTGGCGGCGGCGTATTCAACCTGGCG AACGACCTCGCTAACGGCAGCACCCCCGGCGGCGTAATGATTAGCTGGCTCTTTATAGGCTTTGGAATTTTCATGCTGGTACTGTCGTTTAAC CGGCTTATTACCATTCGGCCCAAACTGTCAGGCGTATCGGATTATGCGCGGGAAGGATTTGGCGACTTTATAGGCTTTCTCAGCGGCTGGGGC TACTGGATTAGCGCGTGGACCGGCACCATTGGCTTTGCCGTGCTGATGATGACCTCGGCGGACTATTTCTTCCCGTCAAAATTTGCGAACTCA AACGGCTCACTGACCGTGCTGAGCGTAATTATTGTGAGCATAATAAGCTGGATACTGATGTTTCTCGTGGACCGGGGCGTAGAAACCGCTGCT CTGGTAAACGCTGTGGTGATGATTGCCAAGCTCATTCCGCTGGTGGTATTTAGCATAACCGGAATTATACTCTTTAAAGCGGACGTATTCACC CAGCATTTTTGGCAGACCTTCACCTCAAACCTCGCGGCTGATGGAAGCGTAAAATCACTCGTGTGGCATGCTATGACCGCGGGCGGCCTCCTG GATCAGATTAAAGGGTCACTTATGGTAATGGGTATTTGTAGGAATTGAAGGGGCCGCTATGATGGGCAACCGAGCCAAGAAAAAATCA GATACCGCGAAAGCTACCATTATTGGCCTCGCCGTGCTGCTGGTAATTTATGTACTGCTGAGCCTGCTGCCGTATGGCTATATGGACCAAGCC TCGCTGGCCAACGTGAAAGCCCCGGGGCTGGTGTATATTCTGAATGAGATGGTGGGCGGGTGGGGGGGCTCGCTTATGGCCATTGGCCTCATG ATTTCACTGCTGGGCGCGTGGCTGTCATGGACCATGCTGCCGGTGGAGGCTACCCAGCAGCTGGCCGAGAAAAAACTCCTGCCGTCGTGGTTT GGCAAGCTTAACAAATATCATGCTCCGTCGAACTCGCTGCTGATTACCCAACTTATGATTCAGATATTTATTATAATTACCTATTTTGTAGCT AACGCGTATAACGTATTTATTTATATGGCCACCGCGGTAATTATGATTTGCTATGCGCTGGTGGGCGCGTATCTGCTGAAAATTGGAATTAAA GAAGGCAGCATTAAAAACATACTGATAGGCTTTTTCACATTTGCGTTTCAGGCGCTTGCTCTGTATCTGAGCGGCTGGCAGTATGTGTGGCTG GCTATGATTCTGTATACAATTGGCTTCATTCTGTTTATTAAAGCTAAAAAAGAATCGCGGCAGGCCATAAGCGTAAAAGAATGGGCCGGCATG TTCATTGTAGCTCTGCTCGGCGTGCTGGCCATTCTTGTACTGATTTGCGGGGCCAAAAGCGGCACCGCGCTGGATCTGCGAGGCCTGCTCGGG TACTAATGCCGA

Supplementary table 5.2 (Cont.): G-blocks used in this study.

Name	Sequence
L. sakei ArcD Harmonized	ATGACCGAAGAAAAACCGGCGAAAAAAATTGGTCTGTTAGCGCTGATTGCGTTAGTGATTAGCAGCTCAATTGGCAGCGGGGTCTTCGGAC TGACCTCAGATCTGGCCAGCGCGTCAGGTCCGGTACTGATGCGCGTGGGTAATTGTCGGTTTTGGGATTCTGATGCTAGCGCTGTC ATTAAACAACCTGTTAATGAAAGAACCAGAACTCGAAGGGATTTTCTCATACGCGGAAAAGGGTTTTGGTCCCTTTGCCGGTTTCATTAGC GGTTGGGGCTATTGGTTAAGCGCGTGGTTAGGGAATGTAGCGTTTGCCACCATTCTGATGAGCGCCTTAGGCTATTTCTTCCCAATTTTTA AAAGCGGGCAAAATCTGCCCAGCATCCTGGTGGCCAGCGTGCTGTCATGGTCACTGACCTATTTTGTGAACCGTGGGGTAGAAGGTGCGGC GGCCATCAACACCTTAGTAACCATTTGCAAATTAATTCCGCTCTTTGTCTTCATTATTTTTGGGATTGTACTGTTTAAAGGCCATCTCTTC ACCCAGGCGTTTTGGAACAACATGAGCAGCTCATTTGTGGCCGGTGATGTAATGTCACAGATCAAGAATTGCATGATGGTGATGATGTGGG TCTTTGTCGGCATCGAAGGTGCGAGCATGCTGTCAGCCCGTGCCGAAAAGAAGTCAGATGCGGGCAAAGCGACCATTCTGGGACTGGTAAG CCTGCTCGCCATTTATATCCTGGCCTCAGTGCTGCCATATGGGTATCTGACCCAGGATCAGCTGGCGTCAATCAAGCAGCCGGCCATGCTG TACATTTTTGAACAAATGGTGGGGACGTGGGGGGGTTACTTCATCGGCGTGGGGCTGATCATCAGCATTCTGGGGGCGTGGCTGTCATGGA CCATGCTGCCAGCCGAAACCATGTTACTGATGGCGAAGCAGAACCTGTTACCGGCGTACTTCGGCCGTGTCAATAAGAAAAAAGCGCCAAC CTTTGCGCTCGTCGTAACCGCCGGGTTAATTCAGGTGTTCCTCTTCACGCTCCTGTTCACCACCAAAGCGTATAATTTCGCGTATAGCCTG TGCACCGCGAGCATTATCGTGTGCTACATGTTAGTAGCGGCGTATCAGATTAAATATTCATGGGCCCACCTACAGGAAAAAGGAAACCGTC AGCAGCTGCTGATTGGGGTGCTGGCCCTGCTGTTTGAAATTGCCGGAATCCTGATGGCCGGCGTCAGCTACCTGCTGCTTTGCTTTATAGC CTATATTCCAGGGATTTTACTTCTATGGCCGCGCCCCGCAAAAACAACGGGCACCAGCATTTCCTGAGCAAAGGGGAATGGCTGATCACCACC ATTATTGTCATCGGCGCGATTATCGGGATTTGGCTGGTGGTGAGCGGGAAAATCGTGATTTAGTGCCGACTCAGTTGCTGCTT
P. Aeruginosa ArcB harmonized	ATGGCTTTTAACATGCATAACCGAAACCTGTTAAGCCTGATGCATCATAGCACCCGCGAACTGCGATATCTGTTAGATCTGTCACGCGATT TAAAACGCGCGAAATATACCGGCACCGAACAGCATCTGAAACGCAAAAACATAGCGCTGATTTTTGAAAAGACCTCAACACGCACCCG CTGCGCGTTTGAAGTGGCGGCGTATGATCAGGGCGCGAACGTGACCTATATTGATCCGAACTCATCTCAGATTGGCCATAAAGAAAGCATG AAAGATACCGCACGGGTATTAGGCCGGATGTATGATGCGATTGAATATCGCGGCTTTAAACAGGAGATTGTGGAAGAGTTAGCAAAATTTG CGGGCGTACCGGTGTTTAACGGCCTGACCGATGAGTATCATCCGACCCAGATGTTAGCGGATGTGCTGACCATGCGCGAACATAGCGATAA ACCGCTGCATGATATTAGCTATGCGTATCTGGGCGATGCACGCAACAACATGGGCAACTCACTGCTGCTGATTGGCGCGAAATTAGGCATG GATGTACGCATTGCGGCTCCGAAAGCATTATGGCCGCATGATGAATTTGTGGCGCAGTGCAAAAAATTTGCGGAAGAGAGCGGCGCGAAAC TGACCTTAACCGAAGACCCGAAAGAGGCGGTGAAAGGCGTGGATTTTGTGCATACCGATGTGTGGGTGTCTATGGGCGAACCGGTGGAGGC GTGGGGCGAACGGATTAAAGAGCTGCTGCCATATCAGGTGAACATGGAAATTATGAAAGCGACCGGCAACCCGCGGGCGAAATTTATGCAT TGCCTGCCGGCATTTCATAACAGCGAAACCAAAGTGGGCAAACAGATTGCGGAACAGTACCCGAACCTGGCGAACGGCATTGAAGTGACCG AAGATGTGTTTGAATCTCCATATAACATTGCGTTTGAACAGGCGGAAAACCGCATGCACACCATTAAAGCAATTTTAGTGTCTACCTTAGC GGATATTTAAGTTTGTGAGTCCAATACTCA

Supplementary table 5.3: Genes in the operon of the 4 randomly sequenced colonies after library assembly. Asterix indicates the codon harmonized version of the gene.

Colony	***arcA***	***arcB***	***arcC***	***arcD***
1	*L. lactis* *	*L. lactis*	*L. lactis*	*L. sakei**
2	*L. lactis*	*L. lactis**	*L. lactis**	*L. sakei*
4	*L. lactis**	*P. aeruginosa**	*L. lactis*	*L. sakei*
5	*L. lactis**	*L. lactis**	*L. lactis*	*L. lactis*

Supplementary table 5.4: Genes in ADI plasmids of sequenced colonies including predicted translation initiation rate (TIR) of RBS. Question marks are not (fully) sequenced. 15 out of 24 sequenced plasmids were sequenced completely. Red colonies and gene names are the five strains that reached the highest biomass yield. Asterix indicates the codon harmonized version of the gene.

Strain	arcA		arcB		arcC		ArcD	
	Gene variant	Predicted TIR	Gene variant	Predicted TIR	Gene variant	Predicted TIR	Gene variant	Predicted TIR
1	*L. lactis arcA**	0.909	*L. lactis arcB**	0.753	*?*	NA	*L. sakei arcD2**	0.239
3	*?*	NA	*?*	NA	*?*	NA	*?*	NA
4	*L. lactis arcA;*	0.066	*L. lactis arcB**	0.075	*L. lactis arcC*	0.918	*?*	NA
9	*L. lactis arcA;*	0.229	*L. lactis arcB**	0.600	*L. lactis arcC*	0.424	*L. sakei arcD2**	0.756
10	*L. lactis arcA;*	0.429	*L. lactis arcB*	0.934	*L. lactis arcC*	0.424	*L. sakei arcD2**	0.239
11	*L. lactis arcA;*	0.229	*L. lactis arcB*	0.256	*L. lactis arcC*	0.087	*L. sakei arcD2*	0.558
13	*L. lactis arcA;*	0.229	*L. lactis arcB, arcB* mix*	0.233	*L. lactis arcC*	0.249	*L. sakei arcD2**	0.239
18	*L. lactis arcA, arcA* mix*	0.429	*L. lactis arcB**	0.753	*L. lactis arcC*	0.599	*L. sakei arcD2**	0.594
20	*L. lactis arcA;*	0.066	*P. aeruginosa arcB**	0.727	*L. lactis arcC*	0.424	*L. sakei arcD2**	0.756
23	*L. lactis arcA**	0.909	*P. aeruginosa arcB**	0.073	*L. lactis arcC*	0.249	*L. sakei arcD2**	0.935

Supplementary table 5.4 (Cont.).

Strain	arcA		arcB		arcC		ArcD	
	Gene variant	**Predicted TIR**	**Gene variant**	**Predicted TIR**	**Gene variant**	**Predicted TIR**	**Gene variant**	**Predicted TIR**
27	*L. lactis arcA;*	0.229	*P. aeruginosa arcB**	0.421	*L. lactis arcC*	0.087	*L. sakei arcD2*	0.260
35	*L. lactis arcA*	0.901	*L. lactis arcB**	0.401	*L. lactis arcC*	0.753	*L. sakei arcD2**	0.079
36	*?*	NA	*?*	NA	*?*	NA	*?*	NA
39	*L. lactis arcA**	0.909	*P. aeruginosa arcB**	0.563	*L. lactis arcC*	0.599	*L. sakei arcD2*	0.558
54	*L. lactis arcA;*	0.229	*L. lactis arcB**	0.401	*L. lactis arcC*	0.599	*L. sakei arcD2**	0.239
58	*L. lactis arcA*; long deletion in the gene*	0.909	*?*	0.909	*L. lactis arcC*	0.918	*L. Sakei arcD2**	0.756
68	*L. lactis arcA**	0.909	*L. lactis arcB*	0.107	*L. lactis arcC*	0.087	*L. sakei arcD2**	0.756
73	*?*	NA	*P. Aeruginosa arcB**	0.563	*L. lactis arcC*	0.249/0.424	*Multiple*	NA
76	*L. lactis arcA;*	0.901	*L. lactis arcB**	0.897	*L. lactis arcC*	0.918	*L. sakei arcD2**	0.935
81	*?*	NA	*L. lactis arcB**	0.075	*L. lactis arcC*	0.424/ 0.918	Multiple	NA

Supplementary table 5.4 (Cont.).

arcA		**arcB**		**arcC**		**ArcD**	
Gene variant	**Predicted TIR**	**Gene variant**	**Predicted TIR**	**Gene variant**	**Predicted TIR**	**Gene variant**	**Predicted TIR**
L. lactis arcA;	0.229	*P. aeruginosa arcB**	0.250	*L. lactis arcC**	0.599	*L. sakei arcD2**	0.935
*L. lactis arcA**;	0.909	*P. aeruginosa arcB**	0.073	*L. lactis arcC*	0.599	?	NA
L. lactis arcA;	0.229	*L. lactis arcB*	0.427	*L. lactis arcC*	0.753	*L. sakei arcD2**	0.594
L. lactis arcA;	0.429	?	NA	?	NA	?	NA

Chapter 6

Summary and general discussion

English summary

Synthetic biology has seen tremendous advances in the recent years, but one question still remains unanswered: can we build a living cell from scratch? In this thesis I dive into the subject of building a bottom-up synthetic cell, from the perspective of the genome. I design functional modules that can at some point be part of a synthetic cell and use a bacterial chassis to test and optimize these modules *in vivo*. In Chapter 1 I start by introducing synthetic cells, and two main approaches that are generally separated when talking about synthetic cell, top-down and bottom-up. Some examples of top-down and minimal synthetic cells are provided, in the context of this thesis.

Next, challenges of building a synthetic cell bottom-up are discussed, and some ways to deal with these challenges are provided. Examples of modules that could be used in a bottom-up synthetic cell are given, and possible methods to test these modules *in vivo*. Evolutionary fitness landscapes are discussed, to understand why it can be beneficial to test multiple modules at once, and the evolutionary drawback of testing them one by one. Finally, the consortium of which this thesis project is part of is introduced, and an outline of the thesis is given.

Chapter 2 is a review of the current understanding of genome organization in prokaryotic cells. Many relevant insights have been gained on the organization of bacterial genomes, and its effect on the phenotype of these bacteria, in the past decades. The expression level of genes is tuned by the coding as well as the non-coding sequences. Codon bias influences expression level from within the coding sequence, and promoters, RBSs, and other regulatory elements from the non-coding sequence. At the operon level, the expression level of genes is influenced by their order in the operon. At the whole genome level, the location and the direction of genes and operons with respect to the chromosomal origin of replication (*oriC*) have a great impact on expression as well. In general, genes closer to the *oriC* have a higher copy number, especially during fast growth regimes, and therefore higher expression level. This gene dosage effect is one reason why many highly expressed (essential) genes cluster around the *oriC*. Additionally, there is a bias for genes to be on the leading strand of replication, with some organisms encoding up to 90% of genes on the leading strand. It is hypothesized that this is beneficial because it limits collisions between DNA polymerase and RNA polymerase during replication. However, new hypotheses on genome organization have been brought up by recent studies in comparative genomics, which have disputed some of the earlier conclusions. Although a lot of progress has been made in this field, there are

still many secrets hidden in the genome sequence. In this chapter we review the recent findings in an attempt to get a better understanding of genome organization. It is evident that a better understanding of genome organization in natural and partly engineered systems will be crucial for the ultimate development of a functional synthetic genome.

In Chapter 3 current cloning methods are discussed, and an effective and cheap *in vivo* recombination technique that can be used for cloning is described. In recent years great advances have been made in the field of DNA recombination and cloning. *In vitro* cloning methods like Gibson assembly and Golden Gate cloning have made it very easy to generate plasmids from multi-fragment assemblies. While these *in vitro* techniques are accurate they are expensive because of the required enzymes. An alternative is provided by *in vivo* recombination in *Saccharomyces cerevisiae*, which has become more and more efficient, especially for cloning large genome-sized DNA fragments. However, a drawback of this method for bacterial plasmid assembly is the need for yeast replicons and markers and colony formation takes ~2 days. With the study presented in chapter 3 we aim to get an accurate DNA recombination method that is the best of both worlds: cheap and fast. The RecA independent recombination (RAIR) pathway in *Escherichia coli* DH10B proves to be a candidate for this purpose. This pathway relies on an exonuclease, a ligase and a polymerase to repair DNA. Homologous flanks of linear DNA are first made single stranded by an exonuclease. These single stranded sequences can anneal by sequence homology, and the gaps are filled in by a polymerase, with the nicks being fixed by a ligase. To test the efficiency of the pathway, a four-fragment assembly strategy was designed aiming at recombining overlapping PCR amplicons of a plasmid. One of the fragments contains an antibiotic resistance gene which allows to select for assembled plasmids. Additionally, two other fragments contain the two halves of the *rfp* gene, which allowed for a red fluorescence phenotype of correctly recombined plasmids. Next, we aimed to optimize the assembly strategy by overexpressing genes that are believed to be involved the RAIR pathway: *xthA, ligA, ligB* and *polA*. Overexpression of *xthA* led to an approximate three-fold increase in recombination efficiency, with 98% recombination accuracy. Overexpression of *ligA* also leads to an increase of recombination efficiency but showed only 60% recombination accuracy. Additionally, the length of the homology flanks between fragments was varied. Homology regions of 50 and 30 bp resulted in a three-fold increased recombination efficiency compared to the 20 bp flanks. Moreover, the longer the overlapping region, the higher the cloning accuracy (50 bp; 99%, 30 bp; 96% and 20 bp 89%). Overall, this technique is cheap and easy to implement in most labs and is an attractive alternative for other (more expensive) cloning methods.

Chapter 4 describes the first attempt of testing a module *in vivo*, while challenging a genomic design principle seen across all living organisms. Often, prokaryotic genes encoding functionally related proteins are clustered in operons. The compact structure of operons allows for co-transcription of the genes, and for co-translation of the polycistronic messenger RNA to the corresponding proteins. This leads to reduced regulatory complexity and enhanced gene expression efficiency, and as such to an overall metabolic benefit for the protein production process in bacteria and archaea. The subunits of RNA polymerase, one of the most conserved and ubiquitous protein complexes, of all annotated prokaryotic genomes contradict this statement, as they are scattered throughout the chromosome. To analyze the impact of this genetic organization on the fitness of *Escherichia coli*, and to serve as a proof of concept for testing modules for a synthetic cell, we designed an operon for the RNA polymerase genes. This operon was constructed using yeast assembly on a bacterial artificial chromosome (BAC) – yeast artificial chromosome (YAC) shuttle vector. In *E. coli* the copy number of the BAC, which is (close to) one, makes it an efficient candidate for testing functional modules of a synthetic cell, as it mimics the copy number of the genome. After construction of the synthetic RNA polymerase operon, we deleted the native chromosomal genes, which led to a reduction in growth on minimal medium. By using adaptive laboratory evolution, we were able to restore the growth rate to wild-type level. Hence, we show that a highly conserved genetic organization of core genes in a bacterium can be reorganized by a combination of design, construction, and evolutionary optimization, yielding a well-functioning synthetic genetic architecture.

In Chapter 5 we test the arginine deiminase (ADI) pathway as a potential ATP generating module that could be used as energy source for the synthetic cell. The ADI pathway uses three enzymes and an antiporter to convert arginine to ornithine, thereby producing one ATP. The antiporter imports one arginine for each ornithine exported, thereby rendering the pathway orthogonal from cellular metabolism. This makes it a great candidate as ATP generating pathway for a synthetic cell. To test this, we aimed to grow *Escherichia coli* using a heterologous expressed arginine deiminase pathway as sole ATP source. For this, a library approach was designed to test multiple promising variants of the pathway, under control of randomized ribosome binding sites. The library was tested by using a stringent growth-no growth selection by growing the cells anaerobically on minimal medium with acetate as carbon source, and arginine as ATP source. Anaerobically, *E. coli* cannot use acetate to regenerate ATP. In our engineered system this has to come from arginine, which couples the function of the ADI pathway to growth. However, no growth was observed using this stringent selection method, so instead a screening approach was developed to find active

arginine deiminase pathway variants. For this method we grew strains from the library on glucose with and without arginine, and tested activity of the ADI pathway by measuring ornithine produced. We found that strains harboring an active arginine deiminase pathway grew to an up to three times higher OD600 compared to the negative control, a finding that was confirmed by *in silico* models. In conclusion, we show that the ADI pathway is active in *E. coli* and can be used to enhance growth in minimal conditions. With adaptive laboratory evolution, we could be able to improve the efficiency of this pathway for it to be used it as single ATP source.

General discussion

Building a bottom-up synthetic cell remains one of the greatest challenges in synthetic biology (1–4, 6, 9, 36, 39, 309–312). Without using specific knowledge gained in top-down genome reduction, I would argue it is not possible at all. We know that there are still many genes of unknown function, that are essential for building a synthetic cell (18, 175). Next to this, there are multiple genomic features that are largely unknown, but at least semi-essential to cells (19). When working in top-down synthetic biology there is the advantage of having an archetype to learn from, its ancestor. In bottom-up synthetic biology, not only do all the necessary genes have to be encoded on the genome, but also all the regulatory elements which make the cell from a chemical soup into a living organism. Take as an example the cell cycle, in which the DNA has to replicate and segregate, the cell has to constrict and close to form two daughter cells, all in a timely manner (38). The only way to overcome these challenges is by keeping on working on it and paying strict attention to many (regulatory) details.

Serendipity in Science

One of the greatest stimulators of science is serendipity. In the Oxford dictionary serendipity is defined as follows: "The occurrence and development of events by chance in a happy or beneficial way (313)." However, I think, especially in science it is not always 'by chance', or 'in a happy way' that new discoveries are made serendipitously, but more so by paying attention. The term serendipity was first used in 1754 by Horace Walpole, in a letter to his friend Horace Mann he explained about an unexpected discovery he had made about a lost painting. He did this by referencing an old Persian tale: "The Three Princes of Serendip" (314). In this tale, three princes are travelling through Serendip (The island we now know as Sri Lanka), and during these travels they make brilliant deductions from seemingly unrelated observations. An example from the story: The three princes are traveling and meet a merchant who has lost his camel. The three princes conclude that the camel must be lame, blind in one eye and missing a tooth. The merchant accuses the three princes of stealing his camel and has them brought forth to the king. At the king the princes can explain their observations: they saw tracks of camel hooves, but it only showed three, the fourth was being dragged, therefore, the camel was lame. The grass next to the path had been eaten only on one side, indicating the camel was blind in one eye. And little lumps of chewed grass the size of a camel tooth on the path suggested that the camel was missing one tooth (315).

In science, some of the most important breakthroughs have been made through serendipity. Perhaps the most notable example is the discovery of penicillin, the first antibiotic, by Alexander Fleming. Fleming noticed, that when growing some *Penicillium* molds, other bacteria did not grow on the 'mold juice' as he called it. Later, he made crude isolates of penicillin and tested it on many microbes. He did stability studies at different temperatures, and tested its toxicity in a laboratory mouse and rabbit (316). Interestingly, Fleming never tested infecting these laboratory animals with disease causing microbes before treating them with penicillin. Although Fleming did make the serendipitous discovery of penicillin, it's true potential was only discovered later by Howard Walter Florey and Ernst Boris Chain, with whom Fleming shared the Nobel prize in 1945 (317, 318).

Serendipity in this thesis

In Chapter 3 of this thesis we describe a novel cloning technique by recombination in *recA* deficient *Escherichia coli* strains. To do this, *E. coli* was transformed with linear DNA fragments with shared sequence homology between the ends of the DNA fragments. These fragments recombined to form a linear plasmid with antibiotic resistance and an *rfp* gene to select for correctly assembled plasmids. However, we started this project with the idea to use the recombinational superstar *Saccharomyces cerevisiae* to develop a cell-free extract-based method, avoiding the need for yeast transformation and replication elements. *S. cerevisiae* has the ability to perform homology-mediated recombination *in vivo* (319). This ability has been used extensively in synthetic biology, e.g., to create a yeast designer chromosome (87), and most notably to assemble the fragments creating the first chemically synthesized genome (33, 320) and the genome for the first synthetic cell (175).

Our idea was to use the homology-mediated recombination *in vitro*, by creating a *S. cerevisiae* cell-free extract, combining it with linear strands of DNA with homologous flanks, performing DNA isolation and subsequently transforming *E. coli* with recombined DNA. For this, various methods were tested to create an active yeast cell-free extract that could be used to recombine DNA *in vitro*. We disrupted *S. cerevisiae* cells by bead beating with metal beads and grains of sand. In total, 13 methods were tested, and the yeast cell-free extract was visually inspected by microscopy to get an estimation of the lysed cell percentage. Two linear DNA fragments were designed with 50 bp sequence homology at the flanks, that both contain half of an *rfp* gene, which is reconstituted when recombined (Figure 6.1A), similar to the fragment design in Chapter 3). The DNA fragments were mixed with the yeast lysate and incubated for 2 hours at 30 °C. Next, DNA isolation was performed and *E. coli* DH10B strain

was transformed with the isolated DNA. Colonies were counted and compared with the efficiency of lysis of each of the yeast cell-free extract preparation methods (Figure 6.1B).

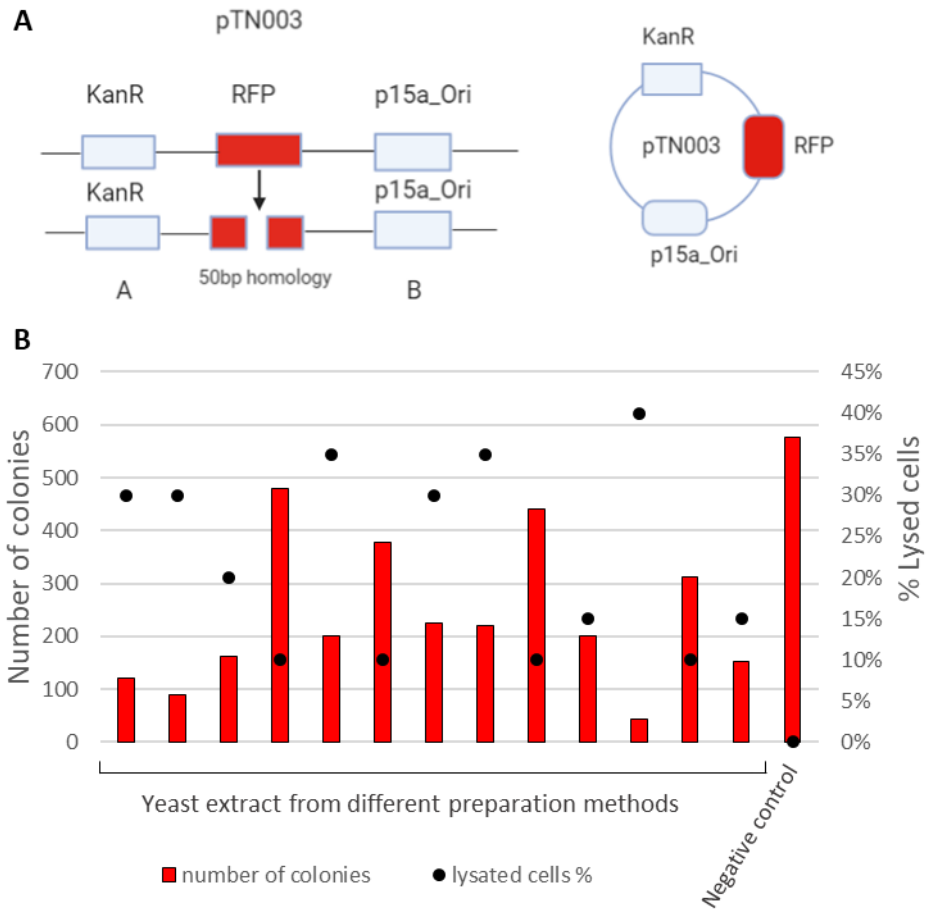


Figure 6.1: *In vitro* yeast recombination using a 2-fragment assembly and 13 yeast cell-free extract preparation methods. **A**: Fragment design, two fragments were designed that each contain half of the *rfp* gene (left). Assembly of the fragments leads to a plasmid with reconstituted *rfp* (right). **B:** graph with number of colonies for each transformation (red bars, left axis) combined with lysing efficiency of each of the yeast cell-free extract preparation methods (black dots, right axis).

Unexpectedly, we found that all of the transformations done with DNA that had been subjected to yeast cell-free extract showed less colonies, compared to the negative control (DNA incubated in recombination buffer for 2 hours at 30 degrees and subsequent isolation). Moreover, a trend was noticed that yeast cell-free extracts with higher efficiencies of lysis actually showed lower colony numbers after *E. coli* transformation. This led us to hypothesize that most likely the yeast cell-free extract was degrading the DNA, and not recombining the fragments, but the recombination was occurring in *E. coli* instead. While going through the literature we discovered that we were not the first to find this method, but we did improve it, which is very cheap and convenient but hardly used in the field (215–217).

Cloning with yeast cell-free extract

As described before, *S. cerevisiae* remains the number one organism for large-scale DNA assembly (33, 83, 87, 175, 320). Hypothetically, if the recombination mechanism could be successfully isolated from *S. cerevisiae* it could manifest as an all-round molecular cloning tool. It would allow the combination of complex recombination in yeast, with the fast growth and easy selection in *E. coli*, resulting in a fast cloning regime. In general, it is understood that yeast recombination of linear DNA with homologous regions works by 5' to 3' resection of the ends, followed by annealing of complementary ssDNA, trimming of the 3' ends and finally ligation of the nicks (321, 322). For this process it is believer there are over 30 genes involved, from exonucleases to single strand binding proteins, and from helicases to ligases (322). The highly complicated system is not yet completely understood, making it virtually impossible to isolate the proteins involved separately. For this reason, our approach was to use yeast cell-free extract, which in our hands did not lead to recombination.

Other studies have been done that used cell-free extract of *S. cerevisiae* to perform DNA recombination. A study from 1993 used a temperature sensitive topoisomerase II mutant (top2-1), to prove that this protein was not involved in *S. cerevisiae* recombination (323). In this study, two plasmids with a mutated tetracycline resistance gene were used that could recombine to form a functional tetracycline resistance. The fragments were incubated with cell-free extract at temperatures restrictive and non-restrictive to top2-1 (37°C and 30°C respectively). After incubation the recombined plasmids were transformed into *E. coli* DH5α and selected for using tetracycline. Incubation at both temperatures led to recombination mutants growing in *E. coli* DH5α, from which the authors conclude that topoisomerase II is not involved in yeast recombination (323). But could it not be possible that this recombination did not occur in the cell-free extract but actually in *E. coli*. The *E. coli* strain

that is used in the study, DH5, is a close relative to the DH10B strain used in Chapter 3 and also lacks *recA* (324). Additionally, no controls in the experiments were conducted that would disprove this theory. Interestingly, some more recent studies have argued that topoisomerase II is indeed involved in recombination (325, 326).

Some other studies have without a doubt illustrated that it is possible to use *S. cerevisiae* cell-free extract to recombine DNA. Already in 1985 recombination products of 2 plasmids in cell-free extract were shown by electrophoretic analysis (327). However, up to our knowledge, no studies have shown recombination in cell-free extract, which came even close to the efficacy of *in yeastu* cloning and could be a practical cloning method.

Future of genome design, *in vivo* or *in vitro*?

The first synthetic genomes were created in the early 2000s, when bacteriophage φX174 DNA (5.4 kb) and poliovirus cDNA (7.5 kb) were synthesized (328, 329). Both of these genomes were built from overlapping oligonucleotides that were annealed and subsequently ligated. Amplification the poliovirus DNA was done by plasmid ligation, while bacteriophage φX174 DNA was amplified by PCR amplification. Both methods resulted in infectious viral particles, paving the way for novel, ambitious genome designs to come. In recent years we have seen that various efforts have been made to create synthetic genomes (19, 33, 83, 87, 175, 282, 320, 330, 331). These can be split into full *in vitro* assemblies, when no living organism is used to aid in constructing the sequence (329, 330), but more often, a combination of chemical synthesis of oligonucleotides, which after *in vitro* assembly to form genome fragments are aided by *in vivo* recombination to construct the full genomes (19, 33, 83, 87, 175, 282, 320, 328, 331). It must be said that there is currently no reason not to use *in yeastu* assembly for constructing large (genome-sized) DNA molecules. As it seems like the constraints for this method have not been reached yet, with more and more ambitious assembly projects being finished (19, 33, 83, 175). But it is also interesting to look at fully *in vitro* constructing DNA.

In vitro oligonucleotide synthesis followed by enzymatic assembly to form long DNA molecules has seen major advances in the recent years. Probably the most used technique for simple homology-based cloning currently is Gibson Assembly (75). This has been used to generate DNA molecules of several hundred kilobases in length, and for complete synthesis of small genomes (75, 330). Another technique that has seen great advances is Golden Gate assembly, which uses type IIS restriction enzymes to create overhangs for subsequent annealing and ligation (332). This cloning method has not shown the advances in size of

DNA, rather in number of molecules, with up to 120 fragments being assembled in a one-pot reaction (333). The 120-part assembly has been made possible by large scale analysis of which 4 nucleotide overhangs, generated by the Type IIS restriction enzyme cleavage, are most efficient in assembling (333, 334).

One of the biggest constraints of fully *in vitro* construction of genomes is the existence of sequences that cannot be synthesized. These impossible to synthesize sequences consist of, among others: repeats, highly structured DNA, high GC content, low GC content, etc. A study showed that more than 75% of all available bacterial sequences are not feasible for low-cost DNA synthesis (335). Partly, this can be overcome by recoding the genome, e.g. using different synonymous codons, although this will not alter the protein sequence of expressed genes, it might have unknown effects on a protein expression (164, 165) or regulatory level (19, 336, 337).

An interesting side effect of creating constructing a genome from scratch, starting by chemical synthesis of oligonucleotides, is that it gives a large design freedom. As said before, still many constraints exist for low-cost DNA synthesis (335), but even with these constraints there are a plethora of possibilities for genome design. When building the minimal synthetic cell JCVI-syn3.0 the authors used an *in yeastu* assembly of eight mega-fragments to build the final genome (175). On these fragments the gene order was kept similar to the originating organism, but another set of eight fragments was made with 'logical' gene orders, clustering all the genes with similar function (175). Inserting one of these fragments led to a cell, which grew comparable to its ancestor, indicating the possibility of such reshuffling of genes, however, genomes with more than one reorganized fragment did not lead to growth (175). In another, more recent study the complete genome of *Caulobacter crescentus* was minimized and synthesized to create *Caulobacter ethensis-2.0* (*C. eth-2.0*) (19). In this study, the 4 Mb genome was minimized to 786 kb and the number of genes and regulatory elements reduced from 6290 to 799. Moreover, the genome was recoded, by 133,313 base substitutions, 123,562 codons were rewritten (19). Using computational biology, difficult to synthesize parts were recoded, to make sure the whole genome could be ordered in 236 blocks for a four-tier assembly (19, 335, 338). Additionally, TTG and TTA codons were completely removed from the genome, allowing for further recoding. The genome was then transplanted into *C. crescentus* leading to a merodiploid strain, with the both the native chromosome, and the *C. eth.-2.0* chromosome. Transposon mutagenesis revealed that 432 of 530 genes of *C. eth.-2.0* are functional as they bared transposons in the native genome (19).

To get a grasp of how we can build a genome for a bottom-up synthetic cell from scratch we can compare the presented genome construction methods described here, with the BAC-YAC approach used in Chapter 4 of this thesis. Of course, constructing an operon of only four genes is difficult to compare with constructing an over-500 gene chromosome, but due to similarities in assembly, the same would theoretically be possible for the approach taken in Chapter 4, by adding more (synthetic) modules to the BAC-YAC by *in yeastu* assembly. The question is whether such a large-scale genome construction would yield interpretable results, as it is not so straightforward which genes to delete to assess functionality of the modules. Additionally, an approach as radical as the *C. eth.-2.0* approach (19) might lack the possibility for tuning, which following the approach taken in Chapter 5 can be done with a library approach. A similar approach could be taken when designing a full genome from scratch, but the sheer number of possible variations of such a library would be incredibly big. As only two functional chromosomes were created after *C. eth.-2.0* assembly, it would be impossible to gather enough wide data on functionality from a library approach at this scale (19).

Alternatively, the approaches taken in this thesis are a lot less costly compared to generating a full genome from commercially available oligonucleotides. These oligonucleotides start at 7 cents (Twist Bioscience: https://twistbioscience.com/) to 11 cents (IDT DNA: https://idtdna.com) per base, meaning commercial synthesis of the *C. eth.-2.0* genome starts at $55,000. In this thesis, we use oligo's of up to 70 bases in size as primers to PCR-amplify genes or plasmid fragments from existing material. Additionally, by adding ambiguous bases in the oligonucleotide design we end up with a pool of primers to create variation in the RBS and thus the expression strength (265, 305). Furthermore, in Chapter 5 we assemble a four-gene operon of the arginine deiminase (ADI) pathway. For this, we use gene homologs from different organisms in a one-pot Gibson assembly reaction, generating more variability on the gene level (75). This led to a great variation of functionality of the operon, which we used to screen and select for the best performing strain. This method of creating a synthetic operon could be expanded to larger fragments in combination with yeast assembly and is much cheaper compared to commercially ordering oligo's. However, to create a full genome for a bottom-up synthetic cell from scratch most likely a combination of both methods is needed, as some genes cannot be simply PCR amplified. In the *C. eth.-2.0* genome, one of the design rules was to keep the same order and orientation for all the genes as compared to *C. crescentus* (19). In the JCVI-syn3.0 synthetic cell, the authors aimed to logically re-order all genes on the genome, but this only succeeded for one of the 8 genome fragments in the one-shot design the authors performed (175). For a bottom-up synthetic cell, there is no

ancestor to copy the gene order and orientation from. This means that most likely, trial and error will need to be performed with designed libraries to optimize functional modules one by one or in a combination of a few modules (Chapter 4, 5). Then, it might be possible to construct a full genome similar to JCVI-syn3.0 and *C. eth.-2.0* and use the prior knowledge to test this *in vivo* (19, 175).

Large ORFs in *E. coli rpoB* and *rpoC* genes

Another striking observation was made during our work on the RNA polymerase operon (Chapter 4). Of the genes that are part of *E. coli* RNA polymerase, *rpoB* and *rpoC* are by far the largest (4029 and 4224 bp respectively (26)). When viewing these genes with sequence analysis software it was noted that both *rpoB* and *rpoC* have relatively long, unannotated open reading frames (ORFs) in the opposite direction from the genes (Figure 6.2). These ORFs are 2889, 2412, 1542 and 1104 bp in length, much longer than would be expected by chance alone. Of the 64 possible codons, 3 encode a stop codon, on average every 20th codon or so should be a stop codon. If we assume codons are randomly present in the genome it is possible to calculate the chance of having a long stretch of codons without any of these 3 stop codons present. The chance of having a 4029 bases or 962 codons without a single stop codon is incredibly small: 8.75×10^{-21}, for 1104 bp the chance would be 2.23×10^{-8}. With these chances, having four of these large ORFs randomly, without any evolutionary benefit, is nearly impossible.

As further investigation to characterize the ORFs they were used a input for NCBI BLAST, a tool to find related sequences (339). Unfortunately, none of the ORFs shared any homology with known genes or proteins. Only hypothetical genes and proteins were found in organisms closely related to *E. coli* (e.g., *Shigella sonnei* and *Salmonella enterica*). Further investigation found that in these organisms the hypothetical genes were in the same location, as reverse ORFs in *rpoB* and *rpoC*. This can be expected for closely related organisms as genes, or the RNA polymerase are highly conserved (chapter 4). To investigate if these ORFs are transcriptionally active the transcription pattern of *rpoB* and *rpoC* was investigated using other work (340). From this study, in which *E. coli* was grown in many different (stress) conditions before total RNA was isolated and sequenced, it is clear that there is virtually no expression of these ORFs on the reverse strand, while *rpoB* and *rpoC* are some of the most expressed genes in all conditions .

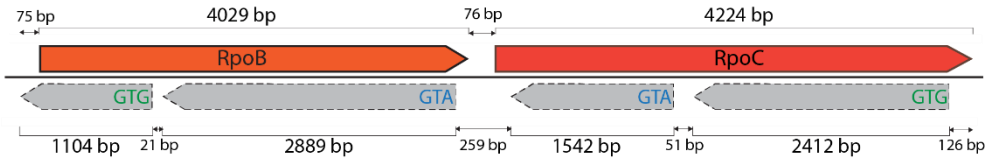


Figure 6.2: the genes *rpoB* (orange) and *rpoC* (red) situated on the genome. On the reverse strand large ORFs are indicated with gray arrows. Sizes are indicated and start codons for the unannotated ORFs are shown.

From these results we must conclude that most likely these ORFs are not genes, there is no transcription, and no homologs can be found using BLAST. This still does not explain why these large ORFs that are so unlikely are present in *rpoB* and *rpoC*. Interestingly, all 4 ORFs present, are in the same reading frame. Moreover, they are in exactly the opposite reading frame from *rpoB* and *rpoC* meaning that their codons line up. Stop codons in *E. coli* are TAA, TAG and TGA, their counterparts on the other strand are TTA, CTA and TCA, codons that code for leucine, leucine, and serine respectively. In coding sequences of *E. coli*, the TTA and CTA codons only comprise 14 % and 4 % total leucine codons, and 1.4 % and 0.4 % of all codons respectively (341). The TCA codon comprises only 12% of all serine codons, and 0.7% of total codons in *E coli* (341). This means that the total chance of randomly encountering a stop codon on the reverse strand in a gene in the opposite frame of the gene is 2.6 %. This increases the chance of encountering such ORFs to 1.46×10^{-11} for 2889 bp and 7.35×10^{-5} for 1104 bp. Which is still unlikely, but especially the latter one would be expected to occur in the *E. coli* genome. One other observation that can be made is that these ORFs do not need to bring an evolutionary advantage to remain, as long as the genes on the other strand are conserved enough, and as we have established before, they are (Chapter 4). Interestingly, in a recent study described before where the genome of *C. crescentus* was minimized and recoded to form *C. eth.-2.0* proof was given that similar ORFs, present in WT *C. crescentus* are not essential (19). As a large part of the genome was recoded (133,313 base substitutions) the *rpoC* gene on the *C. eth.-2.0* did not contain this large ORF anymore. Transposon mutagenesis analysis showed that with reconstitution of *rpoC* (without reverse strand ORFs) the native *rpoC* gene had transposons present, interrupting the ORF, which was not detrimental to the cell (19). It would be interesting to perform a similar experiment in the RNA polymerase operon that we designed in Chapter 4, to discover the essentiality of these ORFs for *E. coli*.

The path towards building a genome for a bottom-up synthetic cell.

A synthetic cell needs to have a carrier of genetic information, in most natural bacterial cells this is a single, circular chromosome with one origin of replication (342). The synthetic genomes that have been constructed so far mostly follow this logic (19, 175), or at least follow the native organism (83). Although there are some examples of using linear DNA in bacterial organisms (199, 200) or in precursor synthetic cells (343), the most obvious route would be to use a single circular chromosome. The way forward is to use functional modules, that can be tested *in vivo* and/or *in vitro*. For testing functional modules *in vivo* Chapter 4 and 5 of this thesis describe potential ways to do this. In short, the strategy we demonstrate to do this is by designing an expression construct, that preferably has a genome-like copy number, for example a bacterial artificial chromosome (BAC) (344). To optimize the functional module one can design a library of gene or expression variants (Chapter 5, (265, 305)), or use adaptive laboratory evolution by coupling the functional module to growth (Chapter 4, (268, 345)).

Another cellular mechanism that could be tested in this way for a synthetic cell could be membrane lipid synthesis. A study performed in the Building a Synthetic Cell (BaSyC) consortium designed a 7-gene lipid synthesis pathway; the Kennedy metabolic pathway (43, 346). The pathway uses fatty acyl-coA and glycerol-3-phosphate to produce phosphatidylethanolamine (PE) and phosphatidylglycerol (PG), the most abundant lipids in *E. coli* membranes. It has been used *in vitro* and in liposomes and shown to produce phospholipids in a cell-free extract (43, 347). This system could most likely be placed on a BAC, for testing it *in vivo* in *E. coli*. Although, in itself this would not give us many new insights, as these are native *E. coli* genes, producing native phospholipids. However, it could be tested if this simplified fatty acid biosynthesis could be enough to create a viable *E. coli* in which more extensive native fatty acid biosynthesis metabolism is knocked out. In addition, this module could be combined with other modules, to start building the synthetic chromosome needed for a functioning cell.

If costs are no issue, probably the best method to construct the genome is by commercially ordering oligos with homologous ends that allow for Gibson assembly into plasmids for the first tier of assembly (75). Alternatively, to save cost *in vivo* assembly in *E. coli* DH10B could be further explored as a method for fragment assembly (Chapter 3). Depending on the size of the chromosome, one or two more tiers of yeast assembly will be needed, to combine the genome fragments into a full genome (19, 84). To test this genome, it should be transplanted into an *in vivo* testing organism, for example *E. coli*, were (part of) the host

genome should now be able to be deleted. From previous top-down synthetic biology we have learned that there are many essential parts of the genome that we do not fully understand yet (18, 175). With that in mind, it seems unlikely that a synthetic chromosome will be able to take over all functions after deletion of the native *E. coli* chromosome. If that indeed is not the case, one could argue that the best way to find which genes on the synthetic genome are non-functional, or which functions are still missing by transposon mutagenesis (11, 13, 19). Identifying the dysfunctional or missing genes and restoring them should be followed by more rounds of transposon mutagenesis. This could lead to a sort of design-built-test cycle that will eventually lead to a fully functional synthetic genome (175). If, at that point we can make vesicles comparable to bacterial membranes containing a cytoplasm with enough protein crowding to ensure kick-starting the organism, it would probably be possible to build a synthetic cell.

Bibliography

1. Szostak,J.W., Bartel,D.P. and Luisi,P.L. (2001) Synthesizing life. *Nature*, **409**, 387–390.
2. Forster,A.C. and Church,G.M. (2006) Towards synthesis of a minimal cell. *Mol Syst Biol*, **2**, 45.
3. Caschera,F. and Noireaux,V. (2014) Integration of biological parts toward the synthesis of a minimal cell. *Curr Opin Chem Biol*, **22**, 85–91.
4. Porcar,M., Danchin,A., de Lorenzo,V., dos Santos,V.A., Krasnogor,N., Rasmussen,S. and Moya,A. (2011) The ten grand challenges of synthetic life. *Syst Synth Biol*, **5**, 1–9.
5. Guindani,C., Silva,L.C., Cao,S., Ivanov,T. and Landfester,K. (2022) Synthetic Cells: From Simple Bio-Inspired Modules to Sophisticated Integrated Systems. *Angewandte Chemie International Edition*, **61**, e202110855.
6. Noireaux,V., Maeda,Y.T. and Libchaber,A. (2011) Development of an artificial cell, from self-organization to computation and self-reproduction. *Proceedings of the National Academy of Sciences*, **108**, 3473–3480.
7. Xu,C., Hu,S. and Chen,X. (2016) Artificial cells: from basic science to applications. *Materials Today*, **19**, 516–532.
8. Juhas,M. (2015) On the road to synthetic life: the minimal cell and genome-scale engineering. *Crit Rev Biotechnol*, **36**, 1–8.
9. Jewett,M.C. and Forster,A.C. (2010) Update on designing and building minimal cells. *Curr Opin Biotechnol*, **21**, 697–703.
10. Akerley,B.J., Rubin,E.J., Camilli,A., Lampe,D.J., Robertson,H.M. and Mekalanos,J.J. (1998) Systematic identification of essential genes by *in vitro* mariner mutagenesis. *Proceedings of the National Academy of Sciences*, **95**, 8927–8932.
11. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Craig Venter,J. (1999) Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science (1979)*, **286**, 2165–2169.
12. Salama,N.R., Shepherd,B. and Falkow,S. (2004) Global Transposon Mutagenesis and Essential Gene Analysis of Helicobacter pylori. *J Bacteriol*, **186**, 7926–7935.
13. Judson,N. and Mekalanos,J.J. (2000) Transposon-based approaches to identify essential bacterial genes. *Trends Microbiol*, **8**, 521–526.
14. Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R., Maruf,M., Hutchison,C.A., Smith,H.O. and Venter,J.C. (2006) Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences*, **103**, 425–430.

15. Segal,E.S., Gritsenko,V., Levitan,A., Yadav,B., Dror,N., Steenwyk,J.L., Silberberg,Y., Mielich,K., Rokas,A., Gow,N.A.R., *et al.* (2018) Gene essentiality analyzed by *in vivo* transposon mutagenesis and machine learning in a stable haploid isolate of candida albicans. *mBio*, **9**, 1–21.
16. Zhu,J., Gong,R., Zhu,Q., He,Q., Xu,N., Xu,Y., Cai,M., Zhou,X., Zhang,Y. and Zhou,M. (2018) Genome-Wide Determination of Gene Essentiality by Transposon Insertion Sequencing in Yeast Pichia pastoris. *Scientific Reports 2018 8:1*, **8**, 1–13.
17. Levitan,A., Gale,A.N., Dallon,E.K., Kozan,D.W., Cunningham,K.W., Sharan,R. and Berman,J. (2020) Comparing the utility of *in vivo* transposon mutagenesis approaches in yeast species to infer gene essentiality. *Curr Genet*, **66**, 1117.
18. Breuer,M., Earnest,T.M., Merryman,C., Wise,K.S., Sun,L., Lynott,M.R., Hutchison,C.A., Smith,H.O., Lapek,J.D., Gonzalez,D.J., *et al.* (2019) Essential metabolism for a minimal cell. *Elife*, **8**.
19. Venetz,J.E., del Medico,L., Wölfle,A., Schächle,P., Bucher,Y., Appert,D., Tschan,F., Flores-Tinoco,C.E., van Kooten,M., Guennoun,R., *et al.* (2019) Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. *Proceedings of the National Academy of Sciences*, **116**, 8070–8079.
20. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L., Mori,H., *et al.* (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 10.1038/msb4100050.
21. Koonin,E. v. (2000) How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept. *Annu Rev Genomics Hum Genet*, **1**, 99–116.
22. Fang,G., Rocha,E. and Danchin,A. (2005) How Essential Are Nonessential Genes? *Mol Biol Evol*, **22**, 2147–2156.
23. Koonin,E. v. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, **1**, 127–136.
24. Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balázsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S., *et al.* (2003) Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. *J Bacteriol*, **185**, 5673–5684.
25. Serres,M.H., Gopal,S., Nahum,L.A., Liang,P., Gaasterland,T. and Riley,M. (2001) A functional update of the Escherichia coli K-12 genome. *Genome Biol*, **2**, RESEARCH0035.

26. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., *et al.* (1997) The Complete Genome Sequence of Escherichia coli K-12. *Science (1979)*, **277**, 1453–1462.
27. Iwadate,Y., Honda,H., Sato,H., Hashimoto,M. and Kato,J. (2011) Oxidative stress sensitivity of engineered Escherichia coli cells with a reduced genome. *FEMS Microbiol Lett*, **322**, 25–33.
28. Hirokawa,Y., Kawano,H., Tanaka-Masuda,K., Nakamura,N., Nakagawa,A., Ito,M., Mori,H., Oshima,T. and Ogasawara,N. (2013) Genetic manipulations restored the growth fitness of reduced-genome Escherichia coli. *J Biosci Bioeng*, **116**, 52–58.
29. Kurokawa,M. and Ying,B.-W. (2019) Experimental Challenges for Reduced Genomes: The Cell Model Escherichia coli. *Microorganisms*, **8**, 3.
30. Reuß,D.R., Altenbuchner,J., Mäder,U., Rath,H., Ischebeck,T., Sappa,P.K., Thürmer,A., Guérin,C., Nicolas,P., Steil,L., *et al.* (2017) Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism. *Genome Res*, **27**, 289–299.
31. Hutchison,C.A., Chuang,R.-Y., Noskov,V.N., Assad-Garcia,N., Deerinck,T.J., Ellisman,M.H., Gill,J., Kannan,K., Karas,B.J., Ma,L., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science (1979)*, **351**.
32. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M., *et al.* (1995) The Minimal Gene Complement of Mycoplasma genitalium. *Science (1979)*, **270**, 397–404.
33. Gibson,D.G., Benders,G.A., Andrews-Pfannkoch,C., Denisova,E.A., Baden-Tillson,H., Zaveri,J., Stockwell,T.B., Brownley,A., Thomas,D.W., Algire,M.A., *et al.* (2008) Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. *Science (1979)*, **319**, 1215–1220.
34. Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.-Y., Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M., *et al.* (2010) Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science (1979)*, **329**, 52–56.
35. Zwart,H. (2019) From primal scenes to synthetic cells. *Elife*, **8**.
36. Schwille,P., Spatz,J., Landfester,K., Bodenschatz,E., Herminghaus,S., Sourjik,V., Erb,T.J., Bastiaens,P., Lipowsky,R., Hyman,A., *et al.* (2018) MaxSynBio: Avenues Towards Creating Cells from the Bottom Up. *Angewandte Chemie International Edition*, **57**, 13382–13392.

37. Ces,O. and Elani,Y. (2019) Community building in synthetic biology. *Exp Biol Med*, **244**, 281–282.
38. Olivi,L., Berger,M., Creyghton,R.N.P., de Franceschi,N., Dekker,C., Mulder,B.M., Claassens,N.J., ten Wolde,P.R. and van der Oost,J. (2021) Towards a synthetic cell cycle. *Nature Communications 2021 12:1*, **12**, 1–11.
39. Schwille,P. (2011) Bottom-Up Synthetic Biology: Engineering in a Tinkerer's World. *Science (1979)*, **333**, 1252–1254.
40. Drobot,B., Iglesias-Artola,J.M., le Vay,K., Mayr,V., Kar,M., Kreysing,M., Mutschler,H. and Tang,T.-Y.D. (2018) Compartmentalised RNA catalysis in membrane-free coacervate protocells. *Nat Commun*, **9**, 3643.
41. Rideau,E., Dimova,R., Schwille,P., Wurm,F.R. and Landfester,K. (2018) Liposomes and polymersomes: a comparative review towards cell mimicking. *Chem Soc Rev*, **47**, 8572–8610.
42. Zubaite,G., Simutis,K., Galinis,R., Milkus,V., Kiseliovas,V. and Mazutis,L. (2017) Droplet Microfluidics Approach for Single-DNA Molecule Amplification and Condensation into DNA-Magnesium-Pyrophosphate Particles. *Micromachines (Basel)*, **8**, 62.
43. Blanken,D., Foschepoth,D., Serrão,A.C. and Danelon,C. (2020) Genetically controlled membrane synthesis in liposomes. *Nat Commun*, **11**, 4317.
44. Libicher,K., Hornberger,R., Heymann,M. and Mutschler,H. (2020) *In vitro* self-replication and multicistronic expression of large synthetic genomes. *Nat Commun*, **11**, 904.
45. Sakatani,Y., Ichihashi,N., Kazuta,Y. and Yomo,T. (2015) A transcription and translation-coupled DNA replication system using rolling-circle replication. *Sci Rep*, **5**, 10404.
46. van Nies,P., Westerlaken,I., Blanken,D., Salas,M., Mencía,M. and Danelon,C. (2018) Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat Commun*, **9**, 1583.
47. Gaut,N.J. and Adamala,K.P. (2021) Reconstituting Natural Cell Elements in Synthetic Cells. *Adv Biol*, **5**, 2000188.
48. Pols,T., Sikkema,H.R., Gaastra,B.F., Frallicciardi,J., Śmigiel,W.M., Singh,S. and Poolman,B. (2019) A synthetic metabolic network for physicochemical homeostasis. *Nat Commun*, **10**, 4239.
49. Sikkema,H.R., Gaastra,B.F., Pols,T. and Poolman,B. (2019) Cell Fuelling and Metabolic Energy Conservation in Synthetic Cells. *ChemBioChem*, **20**, 2581–2592.

50. Bhattacharya,A., Brea,R.J., Niederholtmeyer,H. and Devaraj,N.K. (2019) A minimal biochemical route towards de novo formation of synthetic phospholipid membranes. *Nat Commun*, **10**, 300.
51. Kuruma,Y., Stano,P., Ueda,T. and Luisi,P.L. (2009) A synthetic biology approach to the construction of membrane proteins in semi-synthetic minimal cells. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1788**, 567–574.
52. Jun,S. and Wright,A. (2010) Entropy as the driver of chromosome segregation. *Nat Rev Microbiol*, **8**, 600–607.
53. Huh,D. and Paulsson,J. (2011) Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, **108**, 15004–15009.
54. Gerdes,K., Howard,M. and Szardenings,F. (2010) Pushing and Pulling in Prokaryotic DNA Segregation. *Cell*, **141**, 927–942.
55. Shimobayashi,S.F., Ichikawa,M. and Taniguchi,T. (2016) Direct observations of transition dynamics from macro- to micro-phase separation in asymmetric lipid bilayers induced by externally added glycolipids. *EPL (Europhysics Letters)*, **113**, 56005.
56. Vendel,K.J.A., Tschirpke,S., Shamsi,F., Dogterom,M. and Laan,L. (2019) Minimal *in vitro* systems shed light on cell polarity. *J Cell Sci*, **132**.
57. Nance,J. and Zallen,J.A. (2011) Elaborating polarity: PAR proteins and the cytoskeleton. *Development*, **138**, 799–809.
58. Litschel,T., Ramm,B., Maas,R., Heymann,M. and Schwille,P. (2018) Beating Vesicles: Encapsulated Protein Oscillations Cause Dynamic Membrane Deformations. *Angewandte Chemie International Edition*, **57**, 16286–16290.
59. Caspi,Y. and Dekker,C. (2014) Divided we stand: splitting synthetic cells for their proliferation. *Syst Synth Biol*, **8**, 249–269.
60. Loiseau,E., Schneider,J.A.M., Keber,F.C., Pelzl,C., Massiera,G., Salbreux,G. and Bausch,A.R. (2016) Shape remodeling and blebbing of active cytoskeletal vesicles. *Sci Adv*, **2**.
61. Tsai,F.-C. and Koenderink,G.H. (2015) Shape control of lipid bilayer membranes by confined actin bundles. *Soft Matter*, **11**, 8834–8847.
62. Tanaka,S., Takiguchi,K. and Hayashi,M. (2018) Repetitive stretching of giant liposomes utilizing the nematic alignment of confined actin. *Commun Phys*, **1**, 18.
63. Sato,Y., Hiratsuka,Y., Kawamata,I., Murata,S. and Nomura,S.M. (2017) Micrometer-sized molecular robot changes its shape in response to signal molecules. *Sci Robot*, **2**.

64. Weirich,K.L., Dasbiswas,K., Witten,T.A., Vaikuntanathan,S. and Gardel,M.L. (2019) Self-organizing motors divide active liquid droplets. *Proceedings of the National Academy of Sciences*, **116**, 11125–11130.
65. Last,M.G.F., Deshpande,S. and Dekker,C. (2020) pH-Controlled Coacervate–Membrane Interactions within Liposomes. *ACS Nano*, **14**, 4487–4498.
66. Ganzinger,K.A., Merino-Salomón,A., García-Soriano,D.A., Butterfield,A.N., Litschel,T., Siedler,F. and Schwille,P. (2020) FtsZ Reorganization Facilitates Deformation of Giant Vesicles in Microfluidic Traps**. *Angewandte Chemie International Edition*, **59**, 21372–21376.
67. Schöneberg,J., Pavlin,M.R., Yan,S., Righini,M., Lee,I.-H., Carlson,L.-A., Bahrami,A.H., Goldman,D.H., Ren,X., Hummer,G., *et al.* (2018) ATP-dependent force generation and membrane scission by ESCRT-III and Vps4. *Science (1979)*, **362**, 1423–1428.
68. Schlimpert,S., Wasserstrom,S., Chandra,G., Bibb,M.J., Findlay,K.C., Flärdh,K. and Buttner,M.J. (2017) Two dynamin-like proteins stabilize FtsZ rings during Streptomyces sporulation. *Proceedings of the National Academy of Sciences*, **114**, E6176–E6183.
69. Darwin,C. (1859) On the Origin of the Species 1st ed. John Murray, London.
70. van Cleve,J. and Weissman,D.B. (2015) Measuring ruggedness in fitness landscapes. *Proceedings of the National Academy of Sciences*, **112**, 7345–7346.
71. Wright,S. (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, **1**, 356–366.
72. Brown,J.R. and Doolittle,W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Reviews*, **61**, 456–502.
73. Battistuzzi,F.U., Feijao,A. and Hedges,S.B. (2004) A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol*, **4**, 44.
74. Barrangou,R. and van der Oost,J. (2013) CRISPR-Cas Systems Barrangou,R., van der Oost,J. (eds) Springer Berlin Heidelberg, Berlin, Heidelberg.
75. Gibson,D.G., Young,L., Chuang,R.-Y., Venter,J.C., Hutchison,C.A. and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*, **6**, 343–345.
76. Yu,D., Ellis,H.M., Lee,E.-C., Jenkins,N.A., Copeland,N.G. and Court,D.L. (2000) An efficient recombination system for chromosome engineering in Escherichia coli. *Proceedings of the National Academy of Sciences*, **97**, 5978–5983.

77. Mullis,K., Faloona,F., Scharf,S., Saiki,R., Horn,G. and Erlich,H. (1986) Specific Enzymatic Amplification of DNA *In vitro*: The Polymerase Chain Reaction. *Cold Spring Harb Symp Quant Biol*, **51**, 263–273.
78. Wang,K., Boysen,C., Shizuya,H., Simon,M.I. and Hood,L. (1997) Complete Nucleotide Sequence of Two Generations of a Bacterial Artificial Chromosome Cloning Vector. *Biotechniques*, **23**, 992–994.
79. Shizuya,H. and Kouros-Mehr,H. (2001) The development and applications of the bacterial artificial chromosome cloning system. *Keio J Med*, **50**, 26–30.
80. Shizuya,H., Birren,B., Kim,U.J., Mancino,V., Slepak,T., Tachiiri,Y. and Simon,M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences*, **89**, 8794–8797.
81. O'Connor,M., Peifer,M. and Bender,W. (1989) Construction of Large DNA Segments in Escherichia coli. *Science (1979)*, **244**, 1307–1312.
82. Yoneji,T., Fujita,H., Mukai,T. and Su'etsugu,M. (2021) Grand scale genome manipulation via chromosome swapping in Escherichia coli programmed by three one megabase chromosomes. *Nucleic Acids Res*, **49**, 8407–8418.
83. Richardson,S.M., Mitchell,L.A., Stracquadanio,G., Yang,K., Dymond,J.S., DiCarlo,J.E., Lee,D., Huang,C.L.V., Chandrasegaran,S., Cai,Y., *et al.* (2017) Design of a synthetic yeast genome. *Science (1979)*, **355**, 1040–1044.
84. Chao,R., Yuan,Y. and Zhao,H. (2014) Recent advances in DNA assembly technologies. *FEMS Yeast Res*, **15**, n/a-n/a.
85. Murray,A.W. and Szostak,J.W. (1983) Construction of artificial chromosomes in yeast. *Nature*, **305**, 189–193.
86. Hsiao,C.-L. and Carbon,J. (1979) High-frequency transformation of yeast by plasmids containing the cloned yeast ARG4 gene. *Proceedings of the National Academy of Sciences*, **76**, 3829–3833.
87. Postma,E.D., Dashko,S., van Breemen,L., Taylor Parkins,S.K., van den Broek,M., Daran,J.-M. and Daran-Lapujade,P. (2021) A supernumerary designer chromosome for modular *in vivo* pathway assembly in Saccharomyces cerevisiae. *Nucleic Acids Res*, **49**, 1769–1783.
88. Jha,J.K., Baek,J.H., Venkova-Canova,T. and Chattoraj,D.K. (2012) Chromosome dynamics in multichromosome bacteria. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1819**, 826–829.

89. Soppa,J. (2014) Polyploidy in Archaea and Bacteria: About Desiccation Resistance, Giant Cell Size, Long-Term Survival, Enforcement by a Eukaryotic Host and Additional Aspects. *Microb Physiol*, **24**, 409–419.
90. Lewis,J.S., Jergic,S. and Dixon,N.E. (2016) The E. coli DNA Replication Fork. In *Journal of Bacteriology*. American Society for Microbiology (ASM), Vol. 175, pp. 31–88.
91. Lewis,J.S., Jergic,S. and Dixon,N.E. (2016) The E. coli DNA Replication Fork. In *Philosophical Transactions of the Royal Society B*. The Royal Society, Vol. 377, pp. 31–88.
92. Gao,F. (2015) Bacteria may have multiple replication origins. *Front Microbiol*, **6**, 1–4.
93. Lewis,J.S., Jergic,S. and Dixon,N.E. (2016) The E. coli DNA Replication Fork. *Enzymes (Essen)*, **39**, 31–88.
94. Chandler,M.G. and Pritchard,R.H. (1975) The effect of gene concentration and relative gene dosage on gene output inEscherichia coli. *Mol Gen Genet*, **138**, 127–141.
95. Cooper,S. and Helmstetter,C.E. (1968) Chromosome replication and the division cycle of Escherichia coli. *J Mol Biol*, **31**, 519–540.
96. Mott,M.L. and Berger,J.M. (2007) DNA replication initiation: Mechanisms and regulation in bacteria. *Nat Rev Microbiol*, **5**, 343–354.
97. Murray,H. and Koh,A. (2014) Multiple Regulatory Systems Coordinate DNA Replication with Cell Growth in Bacillus subtilis. *PLoS Genet*, **10**, e1004731.
98. Couturier,E. and Rocha,E.P.C. (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol*, **59**, 1506–1518.
99. Fossum,S., Crooke,E. and Skarstad,K. (2007) Organization of sister origins and replisomes during multifork DNA replication in Escherichia coli. *EMBO J*, **26**, 4514–4522.
100. Rocha,E.P.C. (2008) The Organization of the Bacterial Genome. *Annu Rev Genet*, **42**, 211–233.
101. Sobetzko,P., Travers,A. and Muskhelishvili,G. (2012) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc Natl Acad Sci U S A*, **109**.
102. Nomura,M. (1999) Engineering of bacterial ribosomes: Replacement of all seven Escherichia coli rRNA operons by a single plasmid-encoded operon. *Proceedings of the National Academy of Sciences*, **96**, 1820–1822.

103. Soler-Bistué,A., Aguilar-Pierlé,S., Garcia-Garcerá,M., Val,M.-E., Sismeiro,O., Varet,H., Sieira,R., Krin,E., Skovgaard,O., Comerci,D.J., *et al.* (2019) Macromolecular crowding links ribosomal protein gene dosage to growth rate in Vibrio cholerae. *bioRxiv*, 10.1101/619304.

104. Dryselius,R., Izutsu,K., Honda,T. and Iida,T. (2008) Differential replication dynamics for large and small Vibrio chromosomes affect gene dosage, expression and location. *BMC Genomics*, **9**, 559.

105. Soler-Bistué,A., Mondotte,J.A., Bland,M.J., Val,M.-E., Saleh,M.-C. and Mazel,D. (2015) Genomic Location of the Major Ribosomal Protein Gene Locus Determines Vibrio cholerae Global Growth and Infectivity. *PLoS Genet*, **11**, e1005156.

106. Campo,N., Dias,M.J., Daveran-Mingot,M.-L., Ritzenthaler,P. and Le Bourgeois,P. (2004) Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. *Mol Microbiol*, **51**, 511–522.

107. Darling,A.E., Mikló,I. and Ragan,M.A. (2008) Dynamics of Genome Rearrangement in Bacterial Populations. *PLoS Genet*, **4**, 1000128.

108. Mackiewicz,P., Mackiewicz,D., Kowalczuk,M., Cebrat,S., Eisen,J.A., Heidelberg,J., White,O. and Salzberg,S.L. (2001) Flip-flop around the origin and terminus of replication in prokaryotic genomes A response to Evidence for symmetric chromosomal inversions around the replication origin in bacteria by. *Genome Biol*, **2**.

109. Repar,J. and Warnecke,T. (2017) Non-Random Inversion Landscapes in Prokaryotic Genomes Are Shaped by Heterogeneous Selection Pressures. *Mol Biol Evol*, **34**, 1902–1911.

110. Espejo,R.T. and Plaza,N. (2018) Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA. *Front Microbiol*, **9**, 1232.

111. Stoddard,S.F., Smith,B.J., Hein,R., Roller,B.R.K. and Schmidt,T.M. (2015) rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res*, **43**, D593–D598.

112. Liao,D. (2000) Gene Conversion Drives Within Genic Sequences: Concerted Evolution of Ribosomal RNA Genes in Bacteria and Archaea. *J Mol Evol*, **51**, 305–317.

113. Kiss,A., Sain,B. and Venetianer,P. (1977) The number of rRNA genes in Escherichia coli. *FEBS Lett*, **79**, 77–79.

114. Asai,T., Condon,C., Voulgaris,J., Zaporojets,D., Shen,B., Al-Omar,M., Squires,C. and Squires,C.L. (1999) Construction and Initial Characterization of Escherichia coli

Strains with Few or No Intact Chromosomal rRNA Operons. *J Bacteriol*, **181**, 3803–3809.

115. Gyorfy,Z., Draskovits,G., Vernyik,V., Blattner,F.F., Gaal,T. and Posfai,G. (2015) Engineered ribosomal RNA operon copy-number variants of E. coli reveal the evolutionary trade-offs shaping rRNA operon number. *Nucleic Acids Res*, **43**, 1783–1794.
116. Gao,N., Lu,G., Lercher,M.J. and Chen,W.-H. (2017) Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. *Sci Rep*, **7**, 10572.
117. Mao,X., Zhang,H., Yin,Y. and Xu,Y. (2012) The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res*, **40**, 8210–8218.
118. Rocha,E.P.C. (2008) The Organization of the Bacterial Genome. *Annu Rev Genet*, **42**, 211–233.
119. Koonin,E. V. (2009) Evolution of genome architecture. *Int J Biochem Cell Biol*, **41**, 298–306.
120. Brewer,B.J. (1988) When polymerases collide: Replication and the transcriptional organization of the E. coli chromosome. *Cell*, **53**, 679–686.
121. de Carvalho,M.O. and Ferreira,H.B. (2007) Quantitative determination of gene strand bias in prokaryotic genomes. *Genomics*, **90**, 733–740.
122. Hu,J., Zhao,X. and Yu,J. (2007) Replication-associated purine asymmetry may contribute to strand-biased gene distribution. *Genomics*, **90**, 186–194.
123. Liu,B. and Alberts,B.M. (1995) Head-On Collision Between a DNA Replication Apparatus and RNA Polymerase Transcription Complex. *Science (1979)*, **267**, 1131–1137.
124. Paul,S., Million-Weaver,S., Chattopadhyay,S., Sokurenko,E. and Merrikh,H. (2013) Accelerated gene evolution through replication–transcription conflicts. *Nature*, **495**, 512–515.
125. Price,M.N. (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res*, **33**, 3224–3234.
126. Rocha,E.P.C. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol*, **10**, 393–395.
127. Sankar,T.S., Wastuwidyaningtyas,B.D., Dong,Y., Lewis,S.A. and Wang,J.D. (2016) The nature of mutations induced by replication-transcription collisions. *Nature*, **535**, 178–181.

128. Million-Weaver,S., Samadpour,A.N., Moreno-Habel,D.A., Nugent,P., Brittnacher,M.J., Weiss,E., Hayden,H.S., Miller,S.I., Liachko,I. and Merrikh,H. (2015) An underlying mechanism for the increased mutagenesis of lagging-strand genes in Bacillus subtilis. *Proceedings of the National Academy of Sciences*, **112**, E1096–E1105.

129. Rocha,E.P.C. and Danchin,A. (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet*, **34**, 377–378.

130. Lin,Y., Gao,F. and Zhang,C.-T. (2010) Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem Biophys Res Commun*, **396**, 472–476.

131. Mirkin,E. V. and Mirkin,S.M. (2007) Replication Fork Stalling at Natural Impediments. *Microbiology and Molecular Biology Reviews*, **71**, 13–35.

132. Windgassen,T.A., Wessel,S.R., Bhattacharyya,B. and Keck,J.L. (2018) Mechanisms of bacterial DNA replication restart. *Nucleic Acids Res*, **46**, 504–519.

133. Kato,J.-I. and Hashimoto,M. (2007) Construction of consecutive deletions of the Escherichia coli chromosome. *Mol Syst Biol*, **3**, 132.

134. Kobayashi,K., Ehrlich,S.D., Albertini,A., Amati,G., Andersen,K.K., Arnaud,M., Asai,K., Ashikaga,S., Aymerich,S., Bessieres,P., *et al.* (2003) Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences*, **100**, 4678–4683.

135. Rocha,E.P.C. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol*, **10**, 393–395.

136. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessières,P., Bolotin,A., Borchert,S., *et al.* (1997) The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. *Nature*, **390**, 249–256.

137. Dervyn,E., Suski,C., Daniel,R., Bruand,C., Chapuis,J., Errington,J., Jannière,L. and Ehrlich,S.D. (2001) Two Essential DNA Polymerases at the Bacterial Replication Fork. *Science (1979)*, **294**, 1716–1719.

138. Saha,S.K., Goswami,A. and Dutta,C. (2014) Association of purine asymmetry, strand-biased gene distribution and PolC within Firmicutes and beyond: a new appraisal. *BMC Genomics*, **15**, 430.

139. CHARGAFF,E. (1951) Structure and function of nucleic acids as cell constituents. *Fed Proc*, **10**, 654–9.

140. Chen,W.H., Lu,G., Bork,P., Hu,S. and Lercher,M.J. (2016) Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun*, **7**.

141. Mirkin,E. v., Castro Roa,D., Nudler,E. and Mirkin,S.M. (2006) Transcription regulatory elements are punctuation marks for DNA replication. *Proceedings of the National Academy of Sciences*, **103**, 7276–7281.
142. Vilette,D., Ehrlich,S.D. and Michel,B. (1996) Transcription-induced deletions in plasmid vectors: M13 DNA replication as a source of instability. *Mol Gen Genet*, **252**, 398–403.
143. Srivatsan,A., Tehranchi,A., MacAlpine,D.M. and Wang,J.D. (2010) Co-Orientation of Replication and Transcription Preserves Genome Integrity. *PLoS Genet*, **6**, e1000810.
144. Esnault,E., Valens,M., Espéli,O. and Boccard,F. (2007) Chromosome Structuring Limits Genome Plasticity in Escherichia coli. *PLoS Genet*, **3**, e226.
145. Bryant,J.A., Sellars,L.E., Busby,S.J.W. and Lee,D.J. (2014) Chromosome position effects on gene expression in Escherichia coli K-12. *Nucleic Acids Res*, 10.1093/nar/gku828.
146. Goormans,A.R., Snoeck,N., Decadt,H., Vermeulen,K., Peters,G., Coussement,P., van Herpe,D., Beauprez,J.J., de Maeseneire,S.L. and Soetaert,W.K. (2020) Comprehensive study on Escherichia coli genomic expression: Does position really matter? *Metab Eng*, **62**, 10–19.
147. Scholz,S.A., Diao,R., Wolfe,M.B., Fivenson,E.M., Lin,X.N. and Freddolino,P.L. (2019) High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. *Cell Syst*, **8**, 212-225.e9.
148. Vora,T., Hottes,A.K. and Tavazoie,S. (2009) Protein Occupancy Landscape of a Bacterial Genome. *Mol Cell*, **35**, 247–253.
149. Jeong,D.E., So,Y., Park,S.Y., Park,S.H. and Choi,S.K. (2018) Random knock-in expression system for high yield production of heterologous protein in Bacillus subtilis. *J Biotechnol*, **266**, 50–58.
150. Sauer,C., Syvertsson,S., Bohorquez,L.C., Cruz,R., Harwood,C.R., van Rij,T. and Hamoen,L.W. (2016) Effect of Genome Position on Heterologous Gene Expression in Bacillus subtilis: An Unbiased Analysis. *ACS Synth Biol*, **5**, 942–947.
151. Chen,X. and Zhang,J. (2013) Why Are Genes Encoded on the Lagging Strand of the Bacterial Genome? *Genome Biol Evol*, **5**, 2436–2439.
152. Lang,K.S. and Merrikh,H. (2018) The Clash of Macromolecular Titans: Replication-Transcription Conflicts in Bacteria. *Annu Rev Microbiol*, **72**, 71–88.
153. Camejo,A., Buchrieser,C., Couvé,E., Carvalho,F., Reis,O., Ferreira,P., Sousa,S., Cossart,P. and Cabanes,D. (2009) *In vivo* Transcriptional Profiling of Listeria

monocytogenes and Mutagenesis Identify New Virulence Factors Involved in Infection. *PLoS Pathog*, **5**, e1000449.

154. Scortti,M., Monzó,H.J., Lacharme-Lora,L., Lewis,D.A. and Vázquez-Boland,J.A. (2007) The PrfA virulence regulon. *Microbes Infect*, **9**, 1196–1207.
155. Mostertz,J., Scharf,C., Hecker,M. and Homuth,G. (2004) Transcriptome and proteome analysis of Bacillus subtilis gene expression in response to superoxide and peroxide stress. *Microbiology (N Y)*, **150**, 497–512.
156. Nicolas,P., Mäder,U., Dervyn,E., Rochat,T., Leduc,A., Pigeonneau,N., Bidnenko,E., Marchadier,E., Hoebeke,M., Aymerich,S., *et al.* (2012) Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. *Science (1979)*, **335**, 1103–1106.
157. Lang,K.S., Hall,A.N., Merrikh,C.N., Ragheb,M., Tabakh,H., Pollock,A.J., Woodward,J.J., Dreifus,J.E. and Merrikh,H. (2017) Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. *Cell*, **170**, 787-799.e18.
158. Schroeder,J.W., Sankar,T.S., Wang,J.D. and Simmons,L.A. (2020) The roles of replication-transcription conflict in mutagenesis and evolution of genome organization. *PLoS Genet*, **16**, e1008987.
159. Mushegian,A.R. and Koonin,E. v (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences*, **93**, 10268–10273.
160. Huynen,M., Snel,B., Lathe,W. and Bork,P. (2000) Exploitation of gene context. *Curr Opin Struct Biol*, **10**, 366–370.
161. Aravind,L. (2000) Guilt by Association: Contextual Information in Genome Analysis: Figure 1. *Genome Res*, **10**, 1074–1077.
162. Overbeek,R., Fonstein,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling.
163. Svetlitsky,D., Dagan,T. and Ziv-Ukelson,M. (2020) Discovery of multi-operon colinear syntenic blocks in microbial genomes. *Bioinformatics*, **36**, i21–i29.
164. Quax,T.E.F., Claassens,N.J., Söll,D. and van der Oost,J. (2015) Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell*, **59**, 149–161.
165. Quax,T.E.F., Wolf,Y.I., Koehorst,J.J., Wurtzel,O., van der Oost,R., Ran,W., Blombach,F., Makarova,K.S., Brouns,S.J.J., Forster,A.C., *et al.* (2013) Differential Translation Tunes Uneven Production of Operon-Encoded Proteins. *Cell Rep*, **4**, 938–944.

166. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem Sci*, **23**, 324–328.

167. Marsh,J.A., Hernández,H., Hall,Z., Ahnert,S.E., Perica,T., Robinson,C. V. and Teichmann,S.A. (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*, **153**, 461–470.

168. Jansen,R., Embden,Jan.D.A. van, Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, **43**, 1565–1575.

169. Makarova,K.S. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res*, **30**, 482–496.

170. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science (1979)*, **359**.

171. Brouns,S.J.J., Walther,J., Snijders,A.P.L., van de Werken,H.J.G., Willemen,H.L.D.M., Worm,P., de Vos,M.G.J., Andersson,A., Lundgren,M., Mazon,H.F.M., *et al.* (2006) Identification of the Missing Links in Prokaryotic Pentose Oxidation Pathways. *Journal of Biological Chemistry*, **281**, 27378–27388.

172. Levy,A., Salas Gonzalez,I., Mittelviefhaus,M., Clingenpeel,S., Herrera Paredes,S., Miao,J., Wang,K., Devescovi,G., Stillman,K., Monteiro,F., *et al.* (2018) Genomic features of bacterial adaptation to plants. *Nat Genet*, **50**, 138–150.

173. Raeside,C., Gaffé,J., Deatherage,D.E., Tenaillon,O., Briska,A.M., Ptashkin,R.N., Cruveiller,S., Médigue,C., Lenski,R.E., Barrick,J.E., *et al.* (2014) Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with Escherichia coli. *mBio*, **5**.

174. Eisen,J.A., Heidelberg,J.F., White,O. and Salzberg,S.L. Evidence for symmetric chromosomal inversions around the replication origin in bacteria.

175. Hutchison,C.A., Chuang,R.-Y., Noskov,V.N., Assad-Garcia,N., Deerinck,T.J., Ellisman,M.H., Gill,J., Kannan,K., Karas,B.J., Ma,L., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science (1979)*, **351**.

176. Cairns,J. (1963) The Chromosome of Escherichia coli. *Cold Spring Harb Symp Quant Biol*, **28**, 43–46.

177. Dame,R.T., Rashid,F.-Z.M. and Grainger,D.C. (2020) Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat Rev Genet*, **21**, 227–242.

178. Grainger,D.C. (2016) Structure and function of bacterial H-NS protein. *Biochem Soc Trans*, **44**, 1561–1569.

179. Kahramanoglou,C., Seshasayee,A.S.N., Prieto,A.I., Ibberson,D., Schmidt,S., Zimmermann,J., Benes,V., Fraser,G.M. and Luscombe,N.M. (2011) Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res*, **39**, 2073–2091.

180. Lucchini,S., Rowley,G., Goldberg,M.D., Hurd,D., Harrison,M. and Hinton,J.C.D. (2006) H-NS Mediates the Silencing of Laterally Acquired Genes in Bacteria. *PLoS Pathog*, **2**, e81.

181. Navarre,W.W., Porwollik,S., Wang,Y., McClelland,M., Rosen,H., Libby,S.J. and Fang,F.C. (2006) Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella. *Science (1979)*, **313**, 236–238.

182. Oshima,T., Ishikawa,S., Kurokawa,K., Aiba,H. and Ogasawara,N. (2006) Escherichia coli Histone-Like Protein H-NS Preferentially Binds to Horizontally Acquired DNA in Association with RNA Polymerase. *DNA Research*, **13**, 141–153.

183. Grainger,D.C., Hurd,D., Goldberg,M.D. and Busby,S.J.W. (2006) Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res*, **34**, 4642–4652.

184. Grainger,D.C. (2016) Structure and function of bacterial H-NS protein. *Biochem Soc Trans*, **44**, 1561–1569.

185. Gordon,B.R.G., Li,Y., Cote,A., Weirauch,M.T., Ding,P., Hughes,T.R., Navarre,W.W., Xia,B. and Liu,J. (2011) Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proceedings of the National Academy of Sciences*, **108**, 10690–10695.

186. Smits,W.K. and Grossman,A.D. (2010) The Transcriptional Regulator Rok Binds A+T-Rich DNA and Is Involved in Repression of a Mobile Genetic Element in Bacillus subtilis. *PLoS Genet*, **6**, e1001207.

187. Gordon,B.R.G., Li,Y., Wang,L., Sintsova,A., van Bakel,H., Tian,S., Navarre,W.W., Xia,B. and Liu,J. (2010) Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, **107**, 5154–5159.

188. Nolivos,S. and Sherratt,D. (2014) The bacterial chromosome: architecture and action of bacterial SMC and SMC-like complexes. *FEMS Microbiol Rev*, **38**, 380–392.

189. Marbouty,M., Le Gall,A., Cattoni,D.I., Cournac,A., Koh,A., Fiche,J.-B., Mozziconacci,J., Murray,H., Koszul,R. and Nollmann,M. (2015) Condensin- and

Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol Cell*, **59**, 588–602.

190. Le,T.B.K., Imakaev,M. v., Mirny,L.A. and Laub,M.T. (2013) High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science (1979)*, **342**, 731–734.
191. Le,T.B. and Laub,M.T. (2016) Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J*, **35**, 1582–1595.
192. Lioy,V.S., Cournac,A., Marbouty,M., Duigou,S., Mozziconacci,J., Espéli,O., Boccard,F. and Koszul,R. (2018) Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell*, **172**, 771-783.e18.
193. Woldringh,C.L. (2002) The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol Microbiol*, **45**, 17–29.
194. Espeli,O., Mercier,R. and Boccard,F. (2008) DNA dynamics vary according to macrodomain topography in the E. coli chromosome. *Mol Microbiol*, **68**, 1418–1427.
195. Niki,H., Yamaichi,Y. and Hiraga,S. (2000) Dynamic organization of chromosomal DNA in Escherichia coli. *Genes Dev*, **14**, 212–223.
196. van der Valk,R.A., Vreede,J., Qin,L., Moolenaar,G.F., Hofmann,A., Goosen,N. and Dame,R.T. (2017) Mechanism of environmentally driven conformational changes that modulate H-NS DNA-bridging activity. *Elife*, **6**.
197. Nieuwkoop,T., Finger-Bou,M., van der Oost,J. and Claassens,N.J. (2020) The Ongoing Quest to Crack the Genetic Code for Protein Production. *Mol Cell*, **80**, 193–209.
198. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S., *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res*, **50**, D20–D26.
199. Cui,T., Moro-oka,N., Ohsumi,K., Kodama,K., Ohshima,T., Ogasawara,N., Mori,H., Wanner,B., Niki,H. and Horiuchi,T. (2007) Escherichia coli with a linear genome. *EMBO Rep*, **8**, 181–187.
200. Liang,X., Baek,C.-H. and Katzen,F. (2013) Escherichia coli with Two Linear Chromosomes. *ACS Synth Biol*, **2**, 734–740.
201. Weber,W. and Fussenegger,M. (2012) Emerging biomedical applications of synthetic biology. *Nat Rev Genet*, **13**, 21–35.
202. Nozaki,S. and Niki,H. (2019) Exonuclease III (XthA) Enforces *In vivo* DNA Cloning of Escherichia coli To Create Cohesive Ends. *J Bacteriol*, **201**.

203. Cohen,S.N., Chang,A.C.Y., Boyer,H.W. and Helling,R.B. (1973) Construction of Biologically Functional Bacterial Plasmids *In vitro*. *Proceedings of the National Academy of Sciences*, **70**, 3240–3244.
204. Smolke,C.D. (2009) Building outside of the box: iGEM and the BioBricks Foundation. *Nat Biotechnol*, **27**, 1099–1102.
205. Engler,C., Kandzia,R. and Marillonnet,S. (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS One*, **3**, e3647.
206. Weber,E., Engler,C., Gruetzner,R., Werner,S. and Marillonnet,S. (2011) A Modular Cloning System for Standardized Assembly of Multigene Constructs. *PLoS One*, **6**, e16765.
207. Chandran,S. (2017) Rapid Assembly of DNA via Ligase Cycling Reaction (LCR). In *Methods in Molecular Biology*. Humana Press Inc., Vol. 1472, pp. 105–110.
208. Kok,S. de, Stanton,L.H., Slaby,T., Durot,M., Holmes,V.F., Patel,K.G., Platt,D., Shapland,E.B., Serber,Z., Dean,J., *et al.* (2014) Rapid and Reliable DNA Assembly via Ligase Cycling Reaction. *ACS Synth Biol*, **3**, 97–106.
209. Benoit,R.M., Ostermeier,C., Geiser,M., Li,J.S.Z., Widmer,H. and Auer,M. (2016) Seamless Insert-Plasmid Assembly at High Efficiency and Low Cost. *PLoS One*, **11**, e0153158.
210. Zhang,Y., Werling,U. and Edelmann,W. (2012) SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Res*, **40**, e55–e55.
211. Messerschmidt,K., Hochrein,L., Dehm,D., Schulz,K. and Mueller-Roeber,B. (2016) Characterizing seamless ligation cloning extract for synthetic biological applications. *Anal Biochem*, **509**, 24–32.
212. Robins-Browne,R.M., Holt,K.E., Ingle,D.J., Hocking,D.M., Yang,J. and Tauschek,M. (2016) Are Escherichia coli Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? *Front Cell Infect Microbiol*, **6**, 141.
213. Durfee,T., Nelson,R., Baldwin,S., Plunkett,G., Burland,V., Mau,B., Petrosino,J.F., Qin,X., Muzny,D.M., Ayele,M., *et al.* (2008) The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. *J Bacteriol*, **190**, 2597–2606.
214. Watson,J.F. and García-Nafría,J. (2019) *In vivo* DNA assembly using common laboratory bacteria: A re-emerging tool to simplify molecular cloning. *Journal of Biological Chemistry*, **294**, 15271–15281.
215. Fishel,R.A., James,A.A. and Kolodner,R. (1981) recA-independent general genetic recombination of plasmids. *Nature*, **294**, 184–186.

216. Oliner,J.D., Kinzler,K.W. and Vogelstein,B. (1993) *In vivo* cloning of PCR products in E.coli. *Nucleic Acids Res*, **21**, 5192–5197.
217. Bubeck,P., Winkler,M. and Bautsch,W. (1993) Rapid cloning by homologous recombination *in vivo*. *Nucleic Acids Res*, **21**, 3601–3602.
218. Jacobus,A.P. and Gross,J. (2015) Optimal Cloning of PCR Fragments by Homologous Recombination in Escherichia coli. *PLoS One*, **10**, e0119221.
219. Kostylev,M., Otwell,A.E., Richardson,R.E. and Suzuki,Y. (2015) Cloning Should Be Simple: Escherichia coli DH5α-Mediated Assembly of Multiple DNA Fragments with Short End Homologies. *PLoS One*, **10**, e0137466.
220. Huang,F., Spangler,J.R. and Huang,A.Y. (2017) *In vivo* cloning of up to 16 kb plasmids in E. coli is as simple as PCR. *PLoS One*, **12**, e0183974.
221. García-Nafría,J., Watson,J.F. and Greger,I.H. (2016) IVA cloning: A single-tube universal cloning system exploiting bacterial *In vivo* Assembly. *Sci Rep*, **6**, 27459.
222. Li,C., Wen,A., Shen,B., Lu,J., Huang,Y. and Chang,Y. (2011) FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnol*, **11**, 92.
223. Klock,H.E., Koesema,E.J., Knuth,M.W. and Lesley,S.A. (2008) Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins: Structure, Function, and Bioinformatics*, **71**, 982–994.
224. Grant,S.G.N., Jessee,J., Bloom,F.R. and Hanahan,D. (1990) Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proceedings of the National Academy of Sciences*, **87**, 4645–4649.
225. Gellert,M. (1967) Formation of covalent circles of lambda DNA by E. coli extracts. *Proceedings of the National Academy of Sciences*, **57**, 148–155.
226. Sriskanda,V. (2001) A second NAD+-dependent DNA ligase (LigB) in Escherichia coli. *Nucleic Acids Res*, **29**, 4930–4934.
227. Chayot,R., Montagne,B., Mazel,D. and Ricchetti,M. (2010) An end-joining repair mechanism in Escherichia coli. *Proceedings of the National Academy of Sciences*, **107**, 2141–2146.
228. Beyer,H.M., Gonschorek,P., Samodelov,S.L., Meier,M., Weber,W. and Zurbriggen,M.D. (2015) AQUA Cloning: A Versatile and Simple Enzyme-Free Cloning Approach. *PLoS One*, **10**, e0137652.
229. Fani,R., Brilli,M. and Liò,P. (2005) The Origin and Evolution of Operons: The Piecewise Building of the Proteobacterial Histidine Operon. *J Mol Evol*, **60**, 378–390.

230. Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**, 318–356.
231. Huynen,M., Snel,B., Lathe,W. and Bork,P. (2000) Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Res*, **10**, 1204–1210.
232. Nuñez,P.A., Romero,H., Farber,M.D. and Rocha,E.P.C. (2013) Natural Selection for Operons Depends on Genome Size. *Genome Biol Evol*, **5**, 2242–2254.
233. Sneppen,K., Pedersen,S., Krishna,S., Dodd,I. and Semsey,S. (2010) Economy of Operon Formation: Cotranscription Minimizes Shortfall in Protein Complexes. *mBio*, **1**.
234. Woese,C. (1998) The universal ancestor. *Proceedings of the National Academy of Sciences*, **95**, 6854–6859.
235. Richard Dawkins (1976) The Selfish Gene.
236. Lawrence,J. (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev*, **9**, 642–648.
237. Wade,J.T., Roa,D.C., Grainger,D.C., Hurd,D., Busby,S.J.W., Struhl,K. and Nudler,E. (2006) Extensive functional overlap between σ factors in Escherichia coli. *Nat Struct Mol Biol*, **13**, 806–814.
238. Li,G.-W., Burkhardt,D., Gross,C. and Weissman,J.S. (2014) Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*, **157**, 624–635.
239. Quax,T.E.F., Wolf,Y.I., Koehorst,J.J., Wurtzel,O., van der Oost,R., Ran,W., Blombach,F., Makarova,K.S., Brouns,S.J.J., Forster,A.C., *et al.* (2013) Differential Translation Tunes Uneven Production of Operon-Encoded Proteins. *Cell Rep*, **4**, 938–944.
240. Quax,T.E.F., Claassens,N.J., Söll,D. and van der Oost,J. (2015) Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell*, **59**, 149–161.
241. Burgess,R.R. (1969) Separation and Characterization of the Subunits of Ribonucleic Acid Polymerase. *Journal of Biological Chemistry*, **244**, 6168–6176.
242. Murakami,K., Kimura,M., Owens,J.T., Meares,C.F. and Ishihama,A. (1997) The two α subunits of Escherichia coli RNA polymerase are asymmetrically arranged and contact different halves of the DNA upstream element. *Proceedings of the National Academy of Sciences*, **94**, 1709–1714.

243. Bedwell,D., Davis,G., Gosink,M., Post,L., Nomura,M., Kestler,H., Zengel,J.M. and Lindahl,L. (1985) Nucleotide sequence of the alpha ribosomal protein operon of Escherichia coli. *Nucleic Acids Res*, **13**, 3891–3902.
244. Thomas,M.S., Bedwell,D.M. and Nomura,M. (1987) Regulation of α operon gene expression in Escherichia coli. *J Mol Biol*, **196**, 333–345.
245. Sutherland,C. and Murakami,K.S. (2018) An Introduction to the Structure and Function of the Catalytic Core Enzyme of Escherichia coli RNA Polymerase. *EcoSal Plus*, **8**, ecosalplus.ESP-0004-2018.
246. Koonin,E. v, Krupovic,M., Ishino,S. and Ishino,Y. (2020) The replication machinery of LUCA: common origin of DNA replication and transcription. *BMC Biol*, **18**, 61.
247. Zakharova,N., Paster,B.J., Wesley,I., Dewhirst,F.E., Berg,D.E. and Severinov,K. v. (1999) Fused and Overlapping rpoB and rpoC Genes in Helicobacters, Campylobacters, and Related Bacteria. *J Bacteriol*, **181**, 3857–3859.
248. Severinov,K., Mooney,R., Darst,S.A. and Landick,R. (1997) Tethering of the Large Subunits of Escherichia coli RNA Polymerase. *Journal of Biological Chemistry*, **272**, 24137–24140.
249. Ralling,G. and Linn,T. (1984) Relative activities of the transcriptional regulatory sites in the rplKAJLrpoBC gene cluster of Escherichia coli. *J Bacteriol*, **158**, 279–285.
250. Chatterji,D. and Kumar Ojha,A. (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Curr Opin Microbiol*, **4**, 160–165.
251. CASHEL,M. and GALLANT,J. (1969) Two Compounds implicated in the Function of the RC Gene of Escherichia coli. *Nature*, **221**, 838–841.
252. Durfee,T., Hansen,A.-M., Zhi,H., Blattner,F.R. and Jin,D.J. (2008) Transcription Profiling of the Stringent Response in Escherichia coli. *J Bacteriol*, **190**, 1084–1096.
253. Srivatsan,A. and Wang,J.D. (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Curr Opin Microbiol*, **11**, 100–105.
254. Persson,B. (1997) The spoU gene of Escherichia coli, the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2'-O-methyltransferase activity. *Nucleic Acids Res*, **25**, 4093–4097.
255. Whitby,M.C., Vincent,S.D. and Lloyd,R.G. (1994) Branch migration of Holliday junctions: identification of RecG protein as a junction specific DNA helicase. *EMBO J*, **13**, 5220–5228.
256. Burgess,R.R. (1971) RNA POLYMERASE. *Annu Rev Biochem*, **40**, 711–740.

257. Igarashi,K., Fujita,N. and Ishihama,A. (1989) Promoter selectivity of Escherichia coli RNA polymerase: omega factor is responsible for the ppGpp sensitivity. *Nucleic Acids Res*, **17**, 8755–8765.

258. Mukherjee,K., Nagai,H., Shimamoto,N. and Chatterji,D. (1999) GroEL is involved in activation of Escherichia coli RNA polymerase devoid of the omega subunit *in vivo*. *Eur J Biochem*, **266**, 228–235.

259. Gruber,T.M. and Gross,C.A. (2003) Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. *Annu Rev Microbiol*, **57**, 441–466.

260. Tripathi,L., Zhang,Y. and Lin,Z. (2014) Bacterial Sigma Factors as Targets for Engineered or Synthetic Transcriptional Control. *Front Bioeng Biotechnol*, **2**, 33.

261. Werner,F. and Grohmann,D. (2011) Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol*, **9**, 85–98.

262. Albert,H., Dale,E.C., Lee,E. and Ow,D.W. (1995) Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *The Plant Journal*, **7**, 649–659.

263. Deatherage,D.E. and Barrick,J.E. (2014) Identification of Mutations in Laboratory-Evolved Microbes from Next-Generation Sequencing Data Using breseq. In *Methods in molecular biology (Clifton, N.J.)*. NIH Public Access, Vol. 1151, pp. 165–188.

264. Wenk,S., Yishai,O., Lindner,S.N. and Bar-Even,A. (2018) An Engineering Approach for Rewiring Microbial Metabolism. In *Methods in Enzymology*. Academic Press, Vol. 608, pp. 329–367.

265. Bonde,M.T., Pedersen,M., Klausen,M.S., Jensen,S.I., Wulff,T., Harrison,S., Nielsen,A.T., Herrgård,M.J. and Sommer,M.O.A. (2016) Predictable tuning of protein expression in bacteria. *Nat Methods*, **13**, 233–236.

266. Mülleder,M., Capuano,F., Pir,P., Christen,S., Sauer,U., Oliver,S.G. and Ralser,M. (2012) A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nat Biotechnol*, **30**, 1176–1178.

267. Creutzburg,S.C.A. (2020) Structure-function relations of RNA molecules involved in gene expression and host defence. 10.18174/532325.

268. Choe,D., Lee,J.H., Yoo,M., Hwang,S., Sung,B.H., Cho,S., Palsson,B., Kim,S.C. and Cho,B.-K. (2019) Adaptive laboratory evolution of a genome-reduced Escherichia coli. *Nat Commun*, **10**, 935.

269. Tenaillon,O., Rodríguez-Verdugo,A., Gaut,R.L., McDonald,P., Bennett,A.F., Long,A.D. and Gaut,B.S. (2012) The Molecular Diversity of Adaptive Convergence. *Science (1979)*, **335**, 457–461.

270. Du,B., Olson,C.A., Sastry,A. V., Fang,X., Phaneuf,P. V., Chen,K., Wu,M., Szubin,R., Xu,S., Gao,Y., *et al.* (2020) Adaptive laboratory evolution of Escherichia coli under acid stress. *Microbiology (N Y)*, **166**, 141–148.

271. Mohamed,E.T., Mundhada,H., Landberg,J., Cann,I., Mackie,R.I., Nielsen,A.T., Herrgård,M.J. and Feist,A.M. (2019) Generation of an E. coli platform strain for improved sucrose utilization using adaptive laboratory evolution. *Microb Cell Fact*, **18**, 116.

272. Conrad,T.M., Frazier,M., Joyce,A.R., Cho,B.K., Knight,E.M., Lewis,N.E., Landick,R. and Palsson,B. (2010) RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. *Proc Natl Acad Sci U S A*, **107**, 20500–20505.

273. Murphy,H. and Cashel,M. (2003) Isolation of RNA Polymerase Suppressors of a (p)ppGpp Deficiency. In *Methods in Enzymology*. Academic Press Inc., Vol. 371, pp. 596–601.

274. Klein-Marcuschamer,D., Santos,C.N.S., Yu,H. and Stephanopoulos,G. (2009) Mutagenesis of the Bacterial RNA Polymerase Alpha Subunit for Improvement of Complex Phenotypes. *Appl Environ Microbiol*, **75**, 2705–2711.

275. Conrad,T.M., Lewis,N.E. and Palsson,B.Ø. (2011) Microbial laboratory evolution in the era of genome-scale science. *Mol Syst Biol*, **7**, 509.

276. Phaneuf,P. V, Gosting,D., Palsson,B.O. and Feist,A.M. (2019) ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. *Nucleic Acids Res*, **47**, D1164–D1171.

277. Cheng,K.K., Lee,B.S., Masuda,T., Ito,T., Ikeda,K., Hirayama,A., Deng,L., Dong,J., Shimizu,K., Soga,T., *et al.* (2014) Global metabolic network reorganization by adaptive mutations allows fast growth of Escherichia coli on glycerol. *Nat Commun*, **5**, 1–9.

278. Dragosits,M., Mozhayskiy,V., Quinones-Soto,S., Park,J. and Tagkopoulos,I. (2013) Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in *Escherichia coli*. *Mol Syst Biol*, **9**, 643.

279. Utrilla,J., O'Brien,E.J., Chen,K., McCloskey,D., Cheung,J., Wang,H., Armenta-Medina,D., Feist,A.M. and Palsson,B.O. (2016) Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. *Cell Syst*, **2**, 260–271.

280. Saecker,R.M., Record,M.T. and DeHaseth,P.L. (2011) Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to

Initiation-Competent Open Complexes, and Initiation of RNA Synthesis. *J Mol Biol*, **412**, 754–771.

281. Hutchison,C.A., Chuang,R.-Y., Noskov,V.N., Assad-Garcia,N., Deerinck,T.J., Ellisman,M.H., Gill,J., Kannan,K., Karas,B.J., Ma,L., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science (1979)*, **351**.
282. Kuijpers,N.G.A., Solis-Escalante,D., Luttik,M.A.H., Bisschops,M.M.M., Boonekamp,F.J., van den Broek,M., Pronk,J.T., Daran,J.-M. and Daran-Lapujade,P. (2016) Pathway swapping: Toward modular engineering of essential cellular processes. *Proceedings of the National Academy of Sciences*, **113**, 15060–15065.
283. Decker,K., Jungermann,K. and Thauer,R.K. (1970) Energy Production in Anaerobic Organisms. *Angewandte Chemie International Edition in English*, **9**, 138–158.
284. Junge,W. and Nelson,N. (2015) ATP Synthase. *Annu Rev Biochem*, **84**, 631–657.
285. Claassens,N.J., Volpers,M., Santos,V.A.P.M. dos, van der Oost,J. and de Vos,W.M. (2013) Potential of proton-pumping rhodopsins: engineering photosystems into microorganisms. *Trends Biotechnol*, **31**, 633–642.
286. Folch,P.L., Bisschops,M.M.M. and Weusthuis,R.A. (2021) Metabolic energy conservation for fermentative product formation. *Microb Biotechnol*, **14**, 829–858.
287. Zúñiga,M. (2002) Evolution of arginine deiminase (ADI) pathway genes. *Mol Phylogenet Evol*, **25**, 429–444.
288. Crow,V.L. and Thomas,T.D. (1982) Arginine metabolism in lactic streptococci. *J Bacteriol*, **150**, 1024–1032.
289. Cunin,R., Glansdorff,N., Piérard,A. and Stalon,V. (1986) Biosynthesis and metabolism of arginine in bacteria. *Microbiol Rev*, **50**, 314–352.
290. Marquis,R.E., Bender,G.R., Murray,D.R. and Wong,A. (1987) Arginine deiminase system and bacterial adaptation to acid environments. *Appl Environ Microbiol*, **53**, 198–200.
291. Casiano-Colón,A. and Marquis,R.E. (1988) Role of the arginine deiminase system in protecting oral bacteria and an enzymatic basis for acid tolerance. *Appl Environ Microbiol*, **54**, 1318–1324.
292. Degnan,B.A., Fontaine,M.C., Doebereiner,A.H., Lee,J.J., Mastroeni,P., Dougan,G., Goodacre,J.A. and Kehoe,M.A. (2000) Characterization of an Isogenic Mutant of Streptococcus pyogenes Manfredo Lacking the Ability To Make Streptococcal Acid Glycoprotein. *Infect Immun*, **68**, 2441–2448.
293. Curran,T.M., Lieou,J. and Marquis,R.E. (1995) Arginine deiminase system and acid adaptation of oral streptococci. *Appl Environ Microbiol*, **61**, 4494–4496.

294. Fulde,M., Willenborg,J., de Greeff,A., Benga,L., Smith,H.E., Valentin-Weigand,P. and Goethe,R. (2011) ArgR is an essential local transcriptional regulator of the arcABC operon in Streptococcus suis and is crucial for biological fitness in an acidic environment. *Microbiology (N Y)*, **157**, 572–582.

295. Beck,M.H., Flaiz,M., Bengelsdorf,F.R. and Dürre,P. (2020) Induced heterologous expression of the arginine deiminase pathway promotes growth advantages in the strict anaerobe Acetobacterium woodii. *Appl Microbiol Biotechnol*, **104**, 687–699.

296. Orsi,E., Claassens,N.J., Nikel,P.I. and Lindner,S.N. (2021) Growth-coupled selection of synthetic modules to accelerate cell factory development. *Nat Commun*, **12**, 5295.

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

298. Jensen,S.I., Lennen,R.M., Herrgård,M.J. and Nielsen,A.T. (2015) Seven gene deletions in seven days: Fast generation of Escherichia coli strains tolerant to acetate and osmotic stress. *Scientific Reports 2015 5:1*, **5**, 1–10.

299. Bolotin,A., Wincker,P., Mauger,S., Jaillon,O., Malarme,K., Weissenbach,J., Ehrlich,S.D. and Sorokin,A. (2001) The Complete Genome Sequence of the Lactic Acid Bacterium Lactococcus lactis ssp. lactis IL1403. *Genome Res*, **11**, 731.

300. Claassens,N.J., Siliakus,M.F., Spaans,S.K., Creutzburg,S.C.A., Nijsse,B., Schaap,P.J., Quax,T.E.F. and Van Der Oost,J. (2017) Improving heterologous membrane protein production in Escherichia coli by combining transcriptional tuning and codon usage algorithms. *PLoS One*, **12**, e0184355.

301. Orth,J.D., Conrad,T.M., Na,J., Lerman,J.A., Nam,H., Feist,A.M. and Palsson,B.Ø. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Mol Syst Biol*, **7**, 535.

302. King,Z.A., Lu,J., Dräger,A., Miller,P., Federowicz,S., Lerman,J.A., Ebrahim,A., Palsson,B.O. and Lewis,N.E. (2016) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*, **44**, D515–D522.

303. Ebrahim,A., Lerman,J.A., Palsson,B.O. and Hyduke,D.R. (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol*, **7**, 1–6.

304. Orth,J.D., Thiele,I. and Palsson,B.O. (2010) What is flux balance analysis? *Nature Biotechnology 2010 28:3*, **28**, 245–248.

305. Jeschek,M., Gerngross,D. and Panke,S. (2016) Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat Commun*, **7**, 11163.

306. Höllerer,S., Papaxanthos,L., Gumpinger,A.C., Fischer,K., Beisel,C., Borgwardt,K., Benenson,Y. and Jeschek,M. (2020) Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nature Communications 2020 11:1*, **11**, 1–15.
307. Enjalbert,B., Millard,P., Dinclaux,M., Portais,J.C. and Létisse,F. (2017) Acetate fluxes in Escherichia coli are determined by the thermodynamic control of the Pta-AckA pathway. *Scientific Reports 2017 7:1*, **7**, 1–11.
308. Nelson,D.L., Cox M. C. and Freeman,W.H. (2004) Lehninger: principles of biochemistry 4th ed.
309. Blain,J.C. and Szostak,J.W. (2014) Progress Toward Synthetic Cells. *Annu Rev Biochem*, **83**, 615–640.
310. Attwater,J. and Holliger,P. (2014) A synthetic approach to abiogenesis. *Nat Methods*, **11**, 495–498.
311. Luisi,P.L., Ferri,F. and Stano,P. (2006) Approaches to semi-synthetic minimal cells: A review. *Naturwissenschaften*, **93**, 1–13.
312. Abil,Z. and Danelon,C. (2020) Roadmap to Building a Cell: An Evolutionary Approach. *Front Bioeng Biotechnol*, **8**, 927.
313. Oxford Dictionairy Serendipity.
314. Walpole,H. (1842) The Letters of Horace Walpole: Earl of Orford: Including Numerous Letters Now First Published from the Original Manuscripts. Lea and Blanchard (ed) C. Sherman & Co. Printers, Philadelphia.
315. Remer,T.G. and Armeno,C. (1965) Serendipity and The three princes, from the Peregrinaggio of 1557 Norman, University of Oklahoma Press.
316. Fleming,A. (1929) On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol*, **10**, 226.
317. Simon,J.R. and Moore,P.D. (1987) Homologous recombination between single-stranded DNA and chromosomal genes in Saccharomyces cerevisiae. *Mol Cell Biol*, **7**, 2329–2334.
318. Simon,J.R. and Moore,P.D. (1987) Homologous recombination between single-stranded DNA and chromosomal genes in Saccharomyces cerevisiae. *Mol Cell Biol*, **7**, 2329–2334.
319. Simon,J.R. and Moore,P.D. (1987) Homologous recombination between single-stranded DNA and chromosomal genes in Saccharomyces cerevisiae. *Mol Cell Biol*, **7**, 2329.

320. Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.-Y., Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M., *et al.* (2010) Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science (1979)*, **329**, 52–56.

321. Fishman-Lobell,J., Rudin,N. and Haber,J.E. (1992) Two alternative pathways of double-strand break repair that are kinetically separable and independently modulated. *Mol Cell Biol*, **12**, 1292–1303.

322. Krogh,B.O. and Symington,L.S. (2004) Recombination Proteins in Yeast. *Annu Rev Genet*, **38**, 233–271.

323. MacLeod,A.M., Ferroni,G.D. and Unrau,P. (1993) Genetic recombination of homologous plasmids catalysed by cell-free extracts of topo-isomerase mutant strains of Saccharomyces cerevisiae. *World J Microbiol Biotechnol*, **9**, 583–586.

324. Song,Y., Lee,B.-R., Cho,S., Cho,Y.-B., Kim,S.-W., Kang,T.J., Kim,S.C. and Cho,B.-K. (2015) Determination of single nucleotide variants in Escherichia coli DH5α by using short-read sequencing. *FEMS Microbiol Lett*, **362**, 73.

325. Bergerat,A., de Massy,B., Gadelle,D., Varoutas,P.-C., Nicolas,A. and Forterre,P. (1997) An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature*, **386**, 414–417.

326. Sabourin,M. (2003) Yeast recombination pathways triggered by topoisomerase II-mediated DNA breaks. *Nucleic Acids Res*, **31**, 4373–4384.

327. Symington,L.S., Morrison,P. and Kolodner,R. (1985) Plasmid recombination intermediates generated in a Saccharomyces cerevisiae cell-free recombination system. *Mol Cell Biol*, **5**, 2361–2368.

328. Cello,J., Paul,A. v. and Wimmer,E. (2002) Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template. *Science (1979)*, **297**, 1016–1018.

329. Smith,H.O., Hutchison,C.A., Pfannkoch,C. and Venter,J.C. (2003) Generating a synthetic genome by whole genome assembly: φX174 bacteriophage from synthetic oligonucleotides. *Proceedings of the National Academy of Sciences*, **100**, 15440–15445.

330. Gibson,D.G., Smith,H.O., Hutchison,C.A., Venter,J.C. and Merryman,C. (2010) Chemical synthesis of the mouse mitochondrial genome. *Nat Methods*, **7**, 901–903.

331. Annaluru,N., Muller,H., Mitchell,L.A., Ramalingam,S., Stracquadanio,G., Richardson,S.M., Dymond,J.S., Kuang,Z., Scheifele,L.Z., Cooper,E.M., *et al.* (2014)

Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science (1979)*, **344**, 55–58.

332. Engler,C., Kandzia,R. and Marillonnet,S. (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS One*, **3**, 3647.
333. HamediRad,M., Weisberg,S., Chao,R., Lian,J. and Zhao,H. (2019) Highly Efficient Single-Pot Scarless Golden Gate Assembly. *ACS Synth Biol*, **8**, 1047–1054.
334. Potapov,V., Ong,J.L., Kucera,R.B., Langhorst,B.W., Bilotti,K., Pryor,J.M., Cantor,E.J., Canton,B., Knight,T.F., Evans,T.C., *et al.* (2018) Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. *ACS Synth Biol*, **7**, 2665–2674.
335. Christen,M., Deutsch,S. and Christen,B. (2015) Genome Calligrapher: A Web Tool for Refactoring Bacterial Genome Sequences for de Novo DNA Synthesis. *ACS Synth Biol*, **4**, 927–934.
336. Wang,K., Fredens,J., Brunner,S.F., Kim,S.H., Chia,T. and Chin,J.W. (2016) Defining synonymous codon compression schemes by genome recoding. *Nature*, **539**, 59–64.
337. Lajoie,M.J., Kosuri,S., Mosberg,J.A., Gregg,C.J., Zhang,D. and Church,G.M. (2013) Probing the Limits of Genetic Recoding in Essential Genes. *Science (1979)*, **342**, 361–363.
338. Christen,M., del Medico,L., Christen,H. and Christen,B. (2017) Genome Partitioner: A web tool for multi-level partitioning of large-scale DNA constructs for synthetic biology applications. *PLoS One*, **12**, e0177234.
339. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403–410.
340. Bhatia,R.P., Kirit,H.A., Predeus,A. v. and Bollback,J.P. (2022) Transcriptomic profiling of Escherichia coli K-12 in response to a compendium of stressors. *Sci Rep*, **12**, 8788.
341. Holcomb,D.D., Alexaki,A., Katneni,U. and Kimchi-Sarfaty,C. (2019) The Kazusa codon usage database, CoCoPUTs, and the value of up-to-date codon usage statistics. *Infection, Genetics and Evolution*, **73**, 266–268.
342. Jha,J.K., Baek,J.H., Venkova-Canova,T. and Chattoraj,D.K. (2012) Chromosome dynamics in multichromosome bacteria. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1819**, 826–829.
343. van Nies,P., Westerlaken,I., Blanken,D., Salas,M., Mencía,M. and Danelon,C. (2018) Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat Commun*, **9**, 1583.

344. Shizuya,H., Birren,B., Kim,U.J., Mancino,V., Slepak,T., Tachiiri,Y. and Simon,M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences*, **89**, 8794–8797.

345. Dragosits,M. and Mattanovich,D. (2013) Adaptive laboratory evolution – principles and applications for biotechnology. *Microb Cell Fact*, **12**, 64.

346. Scott,A., Noga,M.J., de Graaf,P., Westerlaken,I., Yildirim,E. and Danelon,C. (2016) Cell-Free Phospholipid Biosynthesis by Gene-Encoded Enzymes Reconstituted in Liposomes. *PLoS One*, **11**, e0163058.

347. Shimizu,Y., Inoue,A., Tomari,Y., Suzuki,T., Yokogawa,T., Nishikawa,K. and Ueda,T. (2001) Cell-free translation reconstituted with purified components. *Nat Biotechnol*, **19**, 751–755.

Nederlandse samenvatting

Het onderzoek voor dit proefschrift is gedaan in het kader van het BaSyC consortium: Building a Synthetic Cell (een synthetische cel bouwen). Zoals de naam al doet vermoeden, is het uiteindelijke doel van dit consortium het maken van een synthetische cel. Met een synthetische cel wordt een cel bedoeld die niet uit de natuur komt, maar door wetenschappers is gemaakt. Het bouwen van een synthetische cel kan in het algemeen op twee manieren gedaan worden: met een top-down of een bottom-up benadering. Bij de top-down benadering start men met een levend organisme, en wordt alles wat niet essentieel is voor het organisme om te overleven weggehaald. Je houdt dan een minimale cel over. Dit wordt ook wel een minimale synthetische cel genoemd. Een nadeel van deze methode is dat op deze manier niet het hele systeem begrepen wordt. In het BaSyC consortium willen we een synthetische cel via de bottom-up methode maken: hierbij worden verschillende levenloze onderdelen samengebracht, die samen een levende cel moeten vormen. Als dit lukt zou dat de eerste keer zijn.

Voor mijn proefschrift heb ik onderzoek gedaan naar de verschillende onderdelen, of modules, die samengebracht moeten worden om een synthetische cel te maken. Hoofdstuk 1 begint met een introductie waarin ik verschillende synthetische cellen die gemaakt zijn door andere wetenschappers tegen het licht houdt. Daarnaast beschrijf ik de uitdagingen die er zijn om een synthetische cel met de bottom-up methode te maken.

Een synthetische cel moet net als gewone cellen een genetische code hebben, waar alle informatie in staat zodat de cel kan verdubbelen: het genoom. Het genoom bevat de genen van het organisme, die op hun beurt de informatie bevatten voor alle eiwitten die nodig zijn om alle processen in de cel te kunnen laten plaatsvinden. Over veel van deze eiwitten en genen is redelijk veel bekend, maar er is nog veel onduidelijk over regulatie van genen, en op grotere schaal de regulatie van het hele genoom. In Hoofdstuk 2 bekijk ik de organisatie van de genomen in bacteriën, om daarvan te leren wat belangrijk is om een synthetische cel te kunnen maken.

Om onderdelen voor een synthetische cel te kunnen testen moet het stukje DNA dat de genetische informatie van het onderdeel bevat makkelijk van het ene naar het andere organisme gebracht kunnen worden. Dit doen we door DNA te knippen en te plakken, ook wel kloneren genoemd. In Hoofdstuk 3 wordt een nieuwe methode om te kloneren beschreven, die we eigenlijk per toeval ontdekt hebben. De methode is erg simpel, en daarom

makkelijk toepasbaar. We gebruiken hiervoor een stam van *Escherichia coli* bacteriën. Deze bacteriën worden in de wetenschap heel veel gebruikt. Deze bacteriën zijn heel goed in het opnemen van stukjes DNA, en als je de juiste stam neemt, kunnen ze dit ook goed aan elkaar plakken. Om stukken DNA op de juiste volgorde aan elkaar te plakken, hoef je alleen maar te zorgen dat de uiteinden van de verschillende stukken DNA dezelfde code hebben. Dit was al eerder bekend, maar in hoofdstuk 3 zijn we erachter gekomen welk gen betrokken is bij dit systeem, en laten we zien dat als we dat gen tot over expressie brengen, zodat er meer eiwit gemaakt wordt, het systeem nog beter werkt.

In Hoofdstuk 4 testen we een module die gebruikt kan worden in de synthetische cel. We testen vier essentiële genen die in de *E. coli* bacterie nodig zijn, om genen te vertalen naar eiwitten. Zonder deze genen kan de cel niet overleven. Deze eiwitten werken samen in een complex, iets wat je vaker ziet in bacteriën. Heel vaak zie je dat de genen van deze eiwitten dicht bij elkaar in het genoom zitten. Echter is dat bij deze vier genen niet zo, ze zitten erg verspreid over het genoom, en wat misschien wel interessanter is, dit is in alle bacteriën zo. Bij het testen van deze vier genen als module voor een synthetische cel hebben we ze erg dicht bij elkaar gezet. Zo konden we ook bepalen of het belangrijk is dat deze genen zo ver uit elkaar zitten. Om de module te testen hebben we de vier genen in *E. coli* gezet op een plasmide. Een plasmide is DNA wat niet het genoom is. Daarna hebben we de originele genen die ver van elkaar afzitten, van het genoom weggehaald. De cellen die we toen hadden, groeiden eigenlijk nog prima, wat betekent dat het dus niet heel veel uitmaakt of deze genen ver van elkaar afzitten of dicht bij elkaar. We merkten wel dat op een groeimedium met minder nutriënten deze cellen veel langzamer groeien, dan cellen die we niet hadden aangepast. Om dit op te lossen hebben we evolutie toegepast: we lieten de aangepaste cellen lang groeien op het groeimedium met minder nutriënten. Op deze manier hebben we cellen geselecteerd die sneller verdubbelen. Deze cellen groeien net zo snel als de niet aangepaste cellen.

Een synthetische cel heeft ook energie nodig om te kunnen groeien. De meeste organismen gebruiken één molecuul als meest voorkomende energiebron: ATP. Dit molecuul heeft een energierijke verbinding die heel makkelijk gebruikt kan worden door eiwitten om reacties te katalyseren. Eiwitten die reacties katalyseren worden ook wel enzymen genoemd. In Hoofdstuk 5 testen we een module die als energiebron gebruikt zou kunnen worden in een synthetische cel. Dit is een module die bestaat uit vier genen die vier eiwitten beschrijven, drie van die eiwitten zijn enzymen die het molecuul arginine kunnen omzetten naar het molecuul ornithine. Daarbij wordt een ATP-molecuul geproduceerd. Deze genen hebben we

uit verschillende bacteriën gehaald. We wilden graag proberen om deze genen op een plasmide in *E. coli* te zetten om te testen of deze module genoeg zou zijn voor de groei van *E. coli*. Dit bleek echter niet het geval, wanneer we alle andere energiebronnen uitzetten, groeiden de cellen niet. We hebben toen een andere strategie bedacht om te testen of deze module werkt. We hebben verschillende combinaties van varianten van de genen getest, om de optimale combinatie te vinden. Dit hebben we getest door de cellen met arginine te groeien, en de ornithine die geproduceerd wordt te meten. Hoe meer ornithine, hoe beter de module werkt. Omdat deze module zo simpel is (het bevat maar vier genen) is het nog altijd een goede kandidaat als energiebron voor een synthetische cel, alleen misschien niet als enige energiebron.

Tot slot worden in Hoofdstuk 6 verschillende resultaten uit dit proefschrift bediscussieerd. Daarnaast heb ik het over serendipiteit, hoe het mogelijk is om door goed op te letten nieuwe ontdekkingen te doen. Dit gebeurde namelijk in hoofdstuk 3 van dit proefschrift. Hier waren we op zoek naar een betere methode om DNA te kloneren, en daarvoor wilden we bakkersgist gebruiken. Bakkersgist (*Saccharomyces cerevisiae*) staat erom bekend erg goed te zijn in het kloneren van DNA. We wilden een mengsel van kapotte gistcellen mengen met DNA om het te laten kloneren. Om te testen of het DNA juist gekloneerd was, brachten we het in *E. coli* cellen. We ontdekten dat hoe minder kapotte gistcellen we hadden, hoe meer correct gekloneerd DNA we hadden. Onze negatieve controle, waarbij we geen gist gebruikten maar gewoon water, werkte zelfs het beste van allemaal. Hierdoor werden we aan het denken gezet: zou het niet de *E. coli* zijn die ons DNA kloneert, in plaats van de gist? Hierdoor ontstond hoofdstuk 3. Uiteindelijk speculeer ik over het maken van het complete genoom voor een synthetische cel, en hoe dit misschien wel het best kan.

Acknowledgements

After ~~four~~ five years I am very happy that my thesis is finally ready! But of course, I could not have done it all by myself. Thank you to everyone that has trained me, motivated me, cheered me up, explained to me, or did whatever else to help me during this time!

Allereerst natuurlijk **John**, in 2017 stuurde ik na mijn thesis bij microbiologie een e-mail of je nog een plek had voor een PhD student. Ik kreeg mail terug dat we maar eens moesten praten, en toen was het eigenlijk al rond. Het project moest in het begin echt nog van de grond komen, dat vond ik best lastig, iedere keer tijdens onze brainstorm meetings kwamen er nieuwe ideeën. Uiteindelijk ben ik heel blij met hoe het gegaan is, en heb ik veel geleerd van de vele gesprekken waar we het hebben gehad over genoom organisatie, evolutie, en eigenlijk alles wat we maar interessant vonden. Ik ben erg dankbaar dat je mij deze kans gegeven hebt!

Nico, ik kende jou al toen ik nog Master student was van onze gezamenlijke hobby: unihockey. Jij ging net voor je postdoc naar Duitsland toen ik begon bij Microbiologie. Na ongeveer een half jaar werd duidelijk dat jij mijn co-promotor zou worden, vooral toen je terugkwam bij microbiologie hadden we veel contact, en dat kwam de voortgang van mijn onderzoek zeker ten goede! Het was heel fijn om jou als begeleider te hebben, zonder de ongelimiteerde artikelen bibliotheek die in jouw hoofd lijkt te schuilen had dit proefschrift er heel anders uit gezien. Bedankt voor alles!

I would like to thank the thesis committee: **Vitor Martins dos Santos**, **Pascale Daran-Lapujade**, **Bert Poolman**, and **Lennart Schada von Borzyskowski** for taking time to read and evaluate my thesis.

Thijs and **Lorenzo**, thank you very much for being my paranymphs. It was really great having you as colleagues (and as a student, **Lorenzo**) within the small BaSyC team.

Thijs, het was heel fijn om samen te sparren over onze projecten. Jij begon een half jaar voordat ik begon met je PhD bij Microbiologie, waar we de hele tijd op hetzelfde kantoor hebben gezeten. Het waren veel goede gesprekken op kantoor, maar ook daarbuiten onder het genot van een biertje. Heel veel succes bij Scope, en laten we contact houden!

And of course, **Lorenzo**, I got to know you in the beginning of 2018 when you started your Master's thesis at Microbiology, under my supervision. Man, you were a great student! Even

though a lot of work you did was trial and error, and didn't end up in chapter 4, I could not have done it all without you. After your thesis you became a great colleague in the BaSyC team, it has been great to discuss our projects together, although talking to you I have to watch the time, because before we know it, we are talking for hours. Good luck finishing your PhD, and let's keep in touch!

Max, the one that got away, it's a pity that you can't be there at my defence. I have known you since our thesis together at Microbiology. Together with **Thijs** we were the original BaSyC Boys. It was great working with you in Microbiology and speaking Dutch with you!

Raymond, het zag er even naar uit dat jij mijn co-promotor zou worden, maar uiteindelijk paste mijn project toch beter bij Nico. Bedankt voor alle praatjes en brainstorm sessies die we, zeker in het begin, gehad hebben. **Serve**, dankjewel voor al je goede vragen tijdens presentaties.

I would never have been able to complete the work in this thesis without the technicians that keep the lab running, thanks **Steven**, **Ineke**, **Phillipe**, **Merlijn**, **Iame. Tom van de Weijer**, je bent al een tijd niet meer bij BacGen, bedankt dat je mij de weg hebt gewezen op het lab. **Rob**, bedankt voor al je hulp op het lab bij bestellingen, sequencen en je interesse in mijn onderzoek. **Ton**, bedankt dat je je kennis over HPLC met mij wilde delen, en voor al je hulp daarbij. **Guus**, wat fijn dat er eindelijk een gemeenschappelijke primer stock kwam, thanks! **Belén**, although you only became a technician after I was finished in the lab, the few times I returned it was great to see the structure you and **Rob** have brought to the BacGen lab, and to the strain- and primer- collections.

Natuurlijk wil ik ook **Willem de Vos** en **Thijs Ettema** bedanken, zonder jullie werk had ik mijn PhD niet kunnen doen bij Microbiologie. **Anja**, **Heidi** en **Hannie**, bedankt voor jullie werk achter de schermen. Altijd als er iets geregeld moet worden kunnen wij als PhD kandidaten bij jullie terecht.

I have to also thank my office mates: **Thijs**, **Max**, **Belén**, **Ismael**, **Janneke**, **Jurre** and **Lorenzo**. We had a lot of fun in the office while working hard (sometimes). Our The Floor Is Lava game led to some hilarious situations and was a great addition to PhD life. The office dinners and movie nights were great, and we should do one again soon! P.S. **Thijs** we did not forget you still have to bring a lot of cake.

Enrique, I cannot finish this acknowledgements section without mentioning you. We followed each other from Helix to Radix to Helix, for theses and PhDs. It is always great to talk with you, I wish you all the best in your career!

Marnix, you were my supervisor during my Master's thesis at Microbiology, which ultimately made me decide to pursue a PhD at Microbiology as well, thank you! **Sjoerd**, it was great having you around in the lab, you are a walking problem solver. Every time something in the lab didn't go as planned, you are the first person every BacGenner goes to. **Wen**, it was great having you as across the bench buddy in the lab, your enthusiasm is very contagious. **Mr. Mihris**, we were together at MIB since our master's thesis, your work ethic is unbelievable, and it was great having you as a colleague. **Costas**, it was great working with you in the BacGen lab, I really enjoyed our time supervising **Karolis** together with **Thijs** and **Belén**. **Suzanne**, you were the last addition to the BaSyC boys, which transformed us into the BaSyC team, it was great working with you. Of course I also want to thank all of the other colleagues that make the BacGen lab a great atmosphere to work, thank you **Adini, Anneleen, Belén, Carina, Carl, Catarina, Costas, Despoina, Eric, Eugenios, Guus, Hanne, Ioannis, Isabelle, Ismael, James, Janneke, Jeroen, Jorrit, Joyshree, Jurre, Lorenzo, Maartje, Mamou, Marnix, Max, Miguel, Mihris, Olufemi, Prarthana, Ricardo, Rob, Sjoerd, Stijn, Suzan, Teunke, Thomas, Wen, Yifan,** and anyone I might forget.

During my PhD I had the opportunity to supervise some great students, that all have had a big impact on the overall project. **Lorenzo**, thank you for your really great work on chapter 4! **Zach**, I'm convinced there is no native French speaker that speaks English as well as you. You were a great student, thanks for all your efforts in chapter 4. **Santi**, you were part of the iGEM team, your hard work added to the amazing achievement of being second place in the world during the iGEM Jamboree. It was great working such a talented team of students. **Qing**, thank you so much for your tireless work on chapter 5, I really enjoyed working with you. Thank you **Karolis**, you were probably the best supervised student, with four shared supervisors. It was great that the project you worked on, which is in chapter 3, worked out so well! **Stefan**, your thesis started at the start of the COVID pandemic, which meant all labwork could not go on, and you had to do a literature study. Despite that, your work was of great value in chapter 5 of this thesis. **Yining**, you were a great student, thanks for all your work for chapter 5.

Daniel, Yifan, Reinier, Jannie, Nancy and **Daan**, it was great being in the Microbiology Seminar Committee with you. I learned a lot from reaching out to great researchers and

learning from their work from the seminars. **Daniel**, thanks especially to you for taking the lead during the pandemic and seeing the opportunity in the situation. We got some great speakers from all over the world in this period, that were able to give their talk in an online seminar.

It was a great tradition to have fries at Ambacht once a week. Thanks for joining, **Belén, Carrie, Catarina, Enrique, Heidi, Ismael, Ivette, Janneke, Jannie, Max, Patrick, Sharon, Thijs,** and of course the great staff at **Cafetaria Het Ambacht**.

Caifang, Costas, Enrique, Ioannis, Ivette, Lot, Ran and **Thanaporn**, thank you for organizing the amazing PhD trip to Boston and New York.

Since the beginning of 2022 I am working at Xenikos in Nijmegen. Thanks, **Ypke** and **Maarten**, I am very grateful that I could start working part-time at the start, to have the time to finish writing my PhD. Thank you **Jürgen**, for being a great mentor in my new role at a biopharmaceutical company. I am learning a lot from you, and of course from all the rest of the colleagues, **André, Bram, Eric, Fokke, Lisa, Maria, Nicole, Olivia, Rebecca, Sonja, Willem** and **Ylja**.

Toen ik aan mijn PhD begon woonde ik nog op B3 in Duivendaal, ik wil iedereen die ik uit die tijd ken bedanken voor alle steun en gezelligheid! **Pjotr**, wij kenden elkaar al sinds de Master, toen we daarna ongeveer tegelijkertijd een PhD gingen doen hebben we veel ervaringen gedeeld, vaak onder het genot van een (speciaal)biertje, zeker in de corona tijd. Bedankt voor alle gesprekken en het samen klagen over onze onderzoeken. **Sierd**, Ik ben heel blij dat ik jou bij op B3 heb leren kennen. Het is altijd lachen met jou, en het is altijd gezellig samen. **Julie**, van het samen een tafel jatten, tot je eigen logeerkamer in ons huis, veel dank voor je luisterende oor tijdens de afgelopen jaren. **Yannick,** buurman, jij moet wel de meest relaxte Texelaar ter wereld zijn. Jij bent er altijd bij, en ik ben heel blij om jou als vriend te hebben. **Rogier**, kapitein, het is ieder jaar weer erg leuk om te gaan zeilen. Je bent een perfecte kapitein, en als we nog een paar keer gaan weet ik zeker dat **Lies, Gwen, Jasper, Yannick** en ik het zelf ook kunnen met jouw lesjes. **Alex, Douwe, Jasper, Pjotr, Rogier, Sierd** en **Yannick** na het NK Risk in 2018 heb ik de afgelopen jaren genoten van onze Risk middagen, en alle andere spellen. Natuurlijk wil ik de rest van de B3jaarden ook bedanken, **Alex, Douwe, Fleur, Gwen, Jaccoline, Jasper, Julie, Lies, Linda, Lisa, Marit, Pjotr, Rogier, Sierd, Thomas, Thomas, Willemien** en **Yannick.**

Nick, Rick en **Steven**, bedankt voor alle avondjes bij De Haen, en de leuke feestjes op de zaak! **Joep, Tom, Joury en Sidney,** ik kin óg al sinds de basisschoeël, wao weej same in

“On The Run” zote. Ongeluifelik det weej mekaar nog steeds zeen. Ut is altied gezellig wennier we same un pilske gaon drinke!

Netuurlik kin ik ouk idderein dae mei geit nao l’Orbitelle neet vergaete. **Floortje, Karin, Koen, Marcel, Mieke** en **Jippie** bedank veur alle gezelligheid, Kro-tjes en rosétjes de aafgeloupe jaore.

Mam, Pap, bedank veur al ógge steun, en alle interesse die geej heb getoeënd. Ik vind 't altied erg lastig um dudelik oet te legge wat ik gedaon heb de aafgeloupe jaore. 't Is zoeë fijn um beej óg in 'n werm nes thoes te kómme. Op momente det ik zelf dach det 't neet zoel lökke, heb geej mich d'r doorhaer gesleip. **Marieke** en **Siem**, 't is altied joeksig um samen wat te gaon doon, bedank veur alle gezelligheid de aafgeloupe jaore!

Lisa, jij bent altijd degene die altijd voor me klaar staat. Je voelt goed aan als het even niet meezit, en ook als ik er soms niet over wil praten weet jij wat te doen. Het zal niet altijd even makkelijk zijn geweest de afgelopen jaren, en ik ben super dankbaar dat je er altijd voor me bent. Ik moet natuurlijk ook **Palm** en **Pilsje** niet vergeten, jullie onvoorwaardelijke liefde kan nergens tegenop.

About the cover

The cover of this thesis was designed using the artificial intelligence (AI) art generator DALL·E 2. An AI art generator uses an input prompt to generate an artwork using complex algorithms that have been developed using machine learning. For the cover artwork the following prompt was used: “complete image of A complete 3D synthetic cell cut in half which shows the DNA and proteins inside on a darkblue background”. The cover shows that AI is not perfect yet, as the DNA helix turns in the incorrect direction.

DALL·E 2 can be reached via https://openai.com/dall-e-2/

Overview of completed training activities

Discipline specific activities

- Microbiology Centennial Symposium, Laboratory of Microbiology and Wageningen University and Research, Wageningen (NL) (2017)
- BaSyC Work Package 5 Meeting, BaSyC, Wageningen (NL) (2018, 2020)
- BaSyC "Kick-start" Training Program, BaSyC, Delft, Amsterdam, Groningen, Wageningen, Nijmegen (NL) (2018)
- BaSyC International Symposium, BaSyC, Delft (NL) (2018)
- BaSyC Spring Meeting, BaSyC, Amsterdam, Online (NL) (2019, 2021)
- Sound of Biotech Conference, Nederlandse Biotechnologie Vereniging, Ede (NL) (2019)
- BaSyC Fall Meeting, BaSyC, Delft, Online (NL) (2019, 2020)
- The International BioDesign Research Conference, Biodesign Research, Online (2020)
- Syncell Conference, University of New Mexico, Online (2020, 2021)

General courses

- VLAG PhD Week, VLAG, Wageningen (NL) (2017)
- PhD Workshop Carousel, WGS, Wageningen (NL) (2018)
- Adobe Illustrator Course, Laboratory of Microbiology, Wageningen (NL) (2019)
- BaSyC Summer School, BaSyC, Texel (NL) (2020)

Other activities

- Preparation of Research Proposal (2017)
- iGem Supervision (2018)
- PhD trip Microbiology & SSB, Boston, New York (US) (2019)
- Microbiology seminar committee (2018-2021)
- BaSyC meetings Bacterial Genetics group
- Bacterial Genetics group meetings
- PhD Meetings Laboratory of Microbiology

List of publications

Vlot, M., **Houkes, J.**, Lochs, S. J. A., Swarts, D. C., Zheng, P., Kunne, T., Mohanraju, P., Anders, C., Jinek, M., van der Oost, J., Dickman, M. J., & Brouns, S. J. J. (2018). Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR–Cas effector complexes. *Nucleic acids research*, 46 (2), 873-885. DOI: 10.1093/nar/gkx1264

Hebron, K. E., Li, E. Y., Arnold Egloff, S. A., Von Lersner, A. K., Taylor, C., **Houkes, J.**, Flaherty, D. K., Eskaros, A., Stricker, T.P., & Zijlstra, A. (2018). Alternative splicing of ALCAM enables tunable regulation of cell-cell adhesion through differential proteolysis. *Scientific reports*, 8 (1), 1-15. DOI: 10.1038/s41598-018-21467-x

Houkes, J., Olivi, L., Paquet Z., Claassens, N. J., van der Oost, J. (2021). Design, construction and optimization of a synthetic RNA polymerase operon in *Escherichia coli*. *Manuscript submitted*. Preprint available at bioRxiv. DOI: 10.1101/2021.11.05.467461

Houkes, J., Claassens, N. J., van der Oost, J. (2023). From natural towards synthetic genome organization in bacteria. *Manuscript in preparation.*

Houkes, J., Li, Y., Yu, Q., Benito-Vaquerizo, S., van Aalzum, E., van der Oost, J., Claassens, N. J. (2023). Introducing the orthogonal ATP-regenerating arginine deiminase pathway in *Escherichia coli*. *Manuscript in preparation.*

Houkes, J., Sirmonatis, K., Adiego-Perez, B., Patinios C., Nieuwkoop T., Claassens, N. J., van der Oost, J. (2023). Fast homology mediated *in vivo* DNA assembly in *Escherichia coli* DH10B. *Manuscript in preparation.*

About the author

Joep Houkes was born on the 30th of June, 1993, in Blerick, The Netherlands. After graduating from the Valuascollege in Venlo, he started a bachelor's degree in Biology and Medical Laboratory Research at the Hogeschool van Arnhem en Nijmegen (HAN). He performed his first internship at Synthon Biopharmaceuticals in Nijmegen, studying the effect of random and targeted integration of antibody genes on antibody production in CHO cells. For his graduation internship, he went to the USA. he worked at the department of pathology microbiology and immunology at Vanderbilt University in Nashville, under supervision of dr. Andries Zijlstra. Here, he studied the effect of ALCAM, a cell surface protein, on bone cancer metastases, using a chicken embryo model.

After obtaining his bachelor's degree from the HAN, Joep continued his studies doing a master's degree in Biotechnology, with the specialization of cellular and molecular biotechnology at Wageningen University. He performed his minor thesis at the department of Virology, under supervision of dr. Richard Kormelink. It encompassed research of Tomato Spotted Wilt Virus encoded NSs protein and potential plant interaction partners.

In 2017 he started a PhD at the department of Microbiology at Wageningen University under supervision of Prof. dr. John van der Oost and dr. Nico Claassens. Here, he worked in a consortium aiming to build a synthetic cell. His research focused on investigating genome organization, and how to use this information for building a synthetic genome, and on testing modules that could potentially be used in a Synthetic cell, in an *in vivo* environment.

Currently, Joep is working at Xenikos as a Bioprocessing Scientist for the T-Guard project. T-Guard is a break-through therapy for people suffering from acute graft-versus-host disease. It is currently in phase III clinical trial.

The research described in this thesis was financially supported by the Dutch Research Council (NWO) Gravitation Project BaSyC (024.003.019) to Prof. Dr John van der Oost.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Printed by proefschriftprinten.nl